Modelo analítico predictivo para visualizar la información pública relacionada al consumo de energía eléctrica de los cotopaxenses y su influencia en su cartera vencida

Peñaherrera Sandoval, Juan Gabriel

Vicerrectorado de Investigación, Innovación y Transferencia de Tecnología

Centro de Posgrados

Maestría en Gestión en Sistemas de Información e Inteligencia de Negocios

Trabajo de titulación, previo a la obtención del título de Magíster en Gestión en Sistemas de Información e Inteligencia de Negocios

Msc. Montaluisa Yugla, Franklin Javier

30 de septiembre del 2021

Curiginal

Document Information

Analyzed document	TESIS_JUAN_P_30_09_2021.docx (D113918545)
Submitted	9/30/2021 9:41:00 PM
Submitted by	Juan Carlos Altamirano
Submitter email	jc.altamiranoc@uta.edu.ec
Similarity	0%
Analysis address	jc.altamiranoc.uta@analysis.urkund.com

Sources included in the report

SA	**Trabajo de Titulación_ Fernando Alvarez_.pdf** Document Trabajo de Titulación_ Fernando Alvarez_.pdf (D104459916)	1
W	URL: http://ve.scielo.org/scielo.php?script=sci_arttext&pid=S1316-00872008000200015&lng=es&tlng=es.Shao, Fetched: 9/30/2021 9:42:00 PM	1

Firmado electrónicamente por:
FRANKLIN JAVIER MONTALUISA YUGLA

Msc. Montaluisa Yugla, Franklin Javier

Director

C.C.: 0502166796

ESPE

UNIVERSIDAD DE LAS FUERZAS ARMADAS

INNOVACIÓN PARA LA EXCELENCIA

Vicerrectorado de Investigación, Innovación y Transferencia de Tecnología

Centro de Posgrados

CERTIFICACIÓN

Certifico que el trabajo de titulación, "**Modelo analítico predictivo para visualizar la información pública relacionada al consumo de energía eléctrica de los cotopaxenses y su influencia en su cartera vencida**" fue realizado por el señor **Peñaherrera Sandoval, Juan Gabriel** el mismo que ha sido revisado y analizado en su totalidad, por la herramienta de verificación de similitud de contenido; por lo tanto, cumple con los requisitos legales, teóricos, científicos, técnicos y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE, razón por la cual me permito acreditar y autorizar para que lo sustente púbicamente.

Sangolquí, 1 de octubre del 2021

Firmado electrónicamente por:
FRANKLIN JAVIER
MONTALUISA
YUGLA

Msc. Montaluisa Yugla, Franklin Javier

Director

C.C.: 0502166796

Vicerrectorado de Investigación, Innovación y Transferencia de Tecnología

Centro de Posgrados

RESPONSABILIDAD DE AUTORÍA

Yo, **Peñaherrera Sandoval, Juan Gabriel** con cédula de ciudadanía **0502501976**, declaro que el contenido, ideas y criterios del trabajo de titulación, "**Modelo analítico predictivo para visualizar la información pública relacionada al consumo de energía eléctrica de los cotopaxenses y su influencia en su cartera vencida**" es de mi autoría y responsabilidad, cumpliendo con los requisitos legales, teóricos, científicos, técnicos y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE, respetando los derechos intelectuales de terceros y referenciando las citas bibliográficas.

Sangolquí, 1 de octubre del 2021

Peñaherrera Sandoval, Juan Gabriel

C.C.: 0502501976

UNIVERSIDAD DE LAS FUERZAS ARMADAS

INNOVACIÓN PARA LA EXCELENCIA

Vicerrectorado de Investigación, Innovación y Transferencia de Tecnología

Centro de Posgrados

AUTORIZACIÓN DE PUBLICACIÓN

Yo, **Peñaherrera Sandoval, Juan Gabriel** con cédula de ciudadanía **0502501976**, autorizo a la Universidad de las Fuerzas Armadas ESPE publicar el trabajo de titulación, "**Modelo analítico predictivo para visualizar la información pública relacionada al consumo de energía eléctrica de los cotopaxenses y su influencia en su cartera vencida**" en el Repositorio Institucional cuyo contenido ideas y criterios son de mi responsabilidad.

Sangolquí, 1 de octubre del 2021

Peñaherrera Sandoval, Juan Gabriel

C.C.: 0502501976

Dedicatoria

Dedicado a mi esposa, compañera de vida y de grandes experiencias, así como a mi hijo, que espero servirle de ejemplo, tal como lo han sido para mí, mis padres y abuelos, respecto al sacrificio y trabajo constante.

Agradecimiento

Doy gracias a mis padres por haberme inculcado la educación, la bondad y la pasión por no rendirse.

Agradezco a mi esposa por ayudarme, acompañarme y amarme día a día a pesar de las diferentes ocupaciones y batallas.

Agradezco a mis hermanos, cuñada y suegros por estar siempre presentes apoyándome y apoyando también a mi esposa y mi hijito.

Agradezco, a mis abuelitos y tíos por ser una guía y un ejemplo en la vida.

Finalmente doy gracias a la vida por todos los amigos, profesores y a mi tutor, que me han ayudado, para que, este proyecto de tesis haya sido culminado.

Índice de contenido

Índice de tablas

Índice de figuras

Resumen

El objetivo del presente proyecto de tesis fue identificar las características relacionadas a la cartera vencida en la empresa eléctrica de Cotopaxi mediante el análisis de su problemática, la cual, es que, su cartera vencida bordea los USD 8 millones, por lo que, inicialmente, se entendió y exploró la data histórica respecto al consumo de energía eléctrica y los comportamientos de pago de los consumidores comprendidos en la provincia de Cotopaxi.

A partir del estado del arte, se seleccionaron los algoritmos que mejor se adaptaron a la data mantenida por la empresa eléctrica de Cotopaxi y se generó una comparación de modelos de Machine Learning con las características relacionadas a la cartera como son: sus clientes, su situación demográfica, sus cuentas, planillas, valores adeudados, así como, las fechas de pago, adicionalmente, se consideró también los sectores urbanos, rurales, kilovatios hora consumidos, para detectar cual modelo fue el más exacto.

De forma simultánea e iterativa, se siguieron los lineamientos de la metodología: DSR (A Design Science Research Methodology for Information Systems Research), así como, el proceso incremental y de mejora continua de CRISP-DM asociado al ciclo de vida de la analítica y operacionalización de datos, estas metodologías permitieron que en el proceso se vayan sumando características, optimizando el modelo y añadiendo conceptos como la limpieza de datos, la validación cruzada y el tuneo de hiperparámetros, arrojando así, finalmente un modelo, adaptado a la realidad de la empresa eléctrica de Cotopaxi.

Palabras Clave:

- **CARTERA VENCIDA**
- **MINERÍA DE DATOS**
- **MODELO ANALÍTICO**
- **ALGORITMO**

Abstract

The objective of this thesis project was to identify the characteristics related to past due portfolio in the Cotopaxi Electric company by analyzing its problems, which is that its past due portfolio is around USD 8 million therefore, first, historical data regarding electricity consumption and payment behaviors of consumers in the province of Cotopaxi were investigated and explored.

Based on the state of the art, the algorithms that best adapted to the data maintained by the Cotopaxi Electric company were selected and a comparison of Machine Learning models were generated with the characteristics related to the portfolio such as: their clients, their demographic situation, their accounts, statements, amounts owed, as well as payment dates, additionally, urban and rural sectors and kilowatt hours consumed were also considered to detect which model was the most accurate.

Simultaneously and iteratively, guidelines of the methodology were followed: DSR (A Design Science Research Methodology for Information Systems Research), as well as, the incremental and continuously improving process of CRISP-DM associated with the life cycle of analytics and operationalization of data. These methodologies allowed that features be added in the process, optimizing the model and adding concepts such as data cleaning, cross-validation and hyperparameter tuning, thus finally yielding a model, adapted to the current situation of the Cotopaxi Electric company.

Key words:

- **SLOW PAYER**
- **DATA MINING**
- **ANALYTICAL MODEL**
- **ALGORITHM**

Capítulo I: Introducción

Antecedentes

Se solicitó información relacionada a planillas, cuentas y clientes mantenida por la empresa eléctrica de Cotopaxi de los años 2019 y 2020, lo cual, fue aprobada mediante Oficio Nro. ELEPCOSA-PE-2021-00386-O, esta empresa posee alrededor de USD 8 millones como cartera vencida según (La Gaceta, 2021), de los cuales, como Objetivo Estratégico Institucional se plantea incrementar la eficiencia empresarial (ELEPCO S.A., 2019) orientándose al menos al 50% de esta cartera, este valor debido a que, el resto de cartera ya es muy antigua, dicha antigüedad se debe a qué, en administraciones anteriores no se llevó una correcta administración y control de varios procesos como son: la recaudación, aumento de canales de recaudación, procesos de cortes y reconexiones de energía eléctrica a cuentas impagas y sobre todo, una correcta y completa exposición y visualización de la información.

Adicionalmente, se parte de la realidad geopolítica y de la cartera vencida en el consumo de energía eléctrica de la provincia de Cotopaxi, que en comparación con las provincias de Pichincha, Guayas y Azuay, donde sus respectivas empresas disponen de una migración (aún en curso o funcionando en paralelo con sistemas legados) al sistema SAP CIS-CRM (MEER, 2018), además, de la envergadura de los Departamentos de Sistemas así como de la tecnología de cada una de estas empresas eléctricas, difiere de la realidad en Cotopaxi, por ejemplo, la Eléctrica Quito en su área de Sistemas posee alrededor de sesenta personas donde ocho de ellas están 100% dedicadas al sistema SAP y pueden obtener datos estadísticos y metadata de la propia herramienta, mientras que, la realidad en la empresa eléctrica Cotopaxi

es diferente, puesto que, en el año 2020 se tenían apenas cinco personas incluyendo al Jefe de Sistemas, (a inicios del 2020 se llegó a 8 personas con partidas provisionales), sin embargo, están distribuidas entre soporte de sistemas y redes, desarrollo y proyectos, pero no dispone de profesionales con conocimientos en SAP, ni mucho menos recursos para destinarlos 100% a dicha migración, posterior soporte y mantenimiento, tampoco posee recursos para crear un área específica para temas de Business Intelligence o de Minería de Datos.

Además, se conoce sobre la experiencia de las tres empresas eléctricas mencionadas, donde la migración a SAP ha llevado ya, más de dos años con alrededor de 8 a 11 profesionales y expertos externos (proveedores) dedicados 100% a dicha labor (CGE, 2021), personal del que no dispone la empresa eléctrica de Cotopaxi.

Sin embargo, partiendo del Objetivo Estratégico Institucional de la empresa eléctrica de Cotopaxi, dicha empresa está expandiendo sus canales de recaudación, mejorando sus procesos de cortes y reconexiones, sus procesos de convenios y acuerdos de pago, pero, no se está explotando adecuadamente las tecnologías de información, no se expone información completa al interior de la institución y no se dispone de un análisis predictivo con respecto a la cartera para poder tomar acciones proactivas, en este punto lamentablemente solo posee acciones reactivas, a pesar de que, por iniciativa propia de personal del Departamento de Sistemas de la empresa eléctrica de Cotopaxi, se han realizado ciertos esfuerzos en el ámbito analítico, con la elaboración de Dashboards (Garzón Ulloa & Chicaiza Castillo, 2017) a nivel descriptivo, sin embargo, este no es suficiente para brindar información a nivel predictivo.

El problema de investigación

Contexto del Problema

Como se ha explicado anteriormente se ha delimitado el contexto de la problemática hacia la población cotopaxense, la cual, tiene distintas realidades geopolíticas hay tanto zonas urbanas como un gran porcentaje en el sector rural y se tiene sector Costa, pero no se han realizado esfuerzos dentro del área de tecnologías de información enfocados en brindar una correcta información de forma clara y completa sobre el consumo de energía, costos, históricos y tendencias de pago individualizado de una forma abierta e intuitiva.

Adicionalmente, día a día se tiene un número creciente de consumidores y estos a su vez pueden solicitar una o varias cuentas, sean estas residenciales, comerciales o industriales, sin embargo, este proyecto de tesis se enfoca solo en las cuentas residenciales, donde si se tiene cierta información descriptiva de los consumidores, que incluso en muchos casos ni si quiera es usada, debido a que la empresa eléctrica de Cotopaxi dispone de un sistema y/o plataforma de hace más de 25 años, lo que, no lo hace intuitivo ni versátil, por lo que, no se conoce de primera mano si un cliente ha pasado por etapas de morosidad, ni mucho menos se tiene una proyección de su comportamiento de pago mediante un resumen 360 del cliente (Dashboards), en consecuencia, tampoco se conoce si es acertado entregarle otro medidor o negociar un convenio de pago a cada cliente.

Planteamiento del Problema

El hecho de que la empresa eléctrica de Cotopaxi mantenga una gran cartera vencida, hace que se tenga gasto en vez de inversión, debido a que, se necesitan esfuerzos económicos para procesos como el de visitas, cortes de energía y reconexiones, esto incluso baja el nivel en la calidad de servicio que exigen los entes de control como la ARCONEL (ARCONEL, 2020), en consecuencia, esto repercute tanto en sus operaciones y a la disminución de sus ingresos.

Agregado a lo anterior, lamentablemente en la empresa eléctrica de Cotopaxi no se ha destinado esfuerzos específicos para entender a los consumidores de energía eléctrica y visualizar dichos datos en áreas de recaudación e incluso en atención al cliente y, por ende, el ¿por qué se está teniendo una cartera vencida?, ya que, solo con esta información se tendrán insumos para poder proyectar o predecir problemas futuros y con ello permitir plantear acciones proactivas.

Por los motivos anteriormente descritos, se necesita mejorar la visualización de la información de los clientes, impulsando una solución analítica, que parta de una comparación entre los modelos predictivos que mejor se adapten a la realidad cotopaxense y que, luego de los análisis arrojen con mayor exactitud pronósticos sobre los factores o características a ser tomados en cuenta, para mejorar las decisiones con relación a la prevención de cartera vencida y recuperación de cartera de la eléctrica en mención.

Objetivos

General

Diseñar un modelo analítico predictivo que permita visualizar la información pública relacionada al consumo de energía eléctrica de los cotopaxenses y su influencia en su cartera vencida.

Específicos

OE1: Realizar un análisis de literatura sobre recomendaciones existentes para determinar técnicas de minería de datos y seleccionar la que mejor se ajuste a la solución planteada.

OE2: Desarrollar un modelo predictivo de analítica de datos que se adapte a la realidad del pago por consumo de energía eléctrica en Cotopaxi.

OE3: Validar el modelo predictivo con la información actual de la empresa eléctrica para determinar la probabilidad de que el modelo ayuda a identificar la cartera vencida.

Variables

Visualizar la información sobre el consumo de energía eléctrica en Cotopaxi mediante un modelo analítico predictivo que permita identificar su cartera vencida.

Señalamiento de variables

Variable dependiente: Identificar Cartera vencida.

Variable independiente: Visualizar la información sobre el comportamiento de pago en relación al consumo eléctrico en Cotopaxi mediante un modelo analítico predictivo.

Justificación importancia y alcance

El presente proyecto de tesis partirá de una comparación entre los modelos analíticos que mejor se adapten a la realidad de la data relacionada con la cartera vencida de la empresa eléctrica de Cotopaxi, y luego de las respectivas evaluaciones, se obtendrá el modelo analítico más eficiente y las características que permitan arrojar una mayor exactitud en la predicción de cartera, entregando de esta manera, un listado sobre los factores o parámetros (inputs) a ser tomados en cuenta por la empresa eléctrica en mención, en función de lo planteado, se pretende aportar al objetivo estratégico institucional de incrementar la eficiencia empresarial, brindando un estudio inicial que permita tomar nuevas decisiones en cuanto a la prevención sobre su cartera, ya que, actualmente la empresa eléctrica de Cotopaxi posee acciones reactivas para la recuperación de esta.

En relación a la idea anterior, se plantea indicar en que características la empresa eléctrica de Cotopaxi posee data con ruido, no limpia o faltante, para que esta empresa pueda lograr posteriormente arreglarla o mitigar dichos factores, ya que, la falta de información en estos parámetros afectará en la predicción de su cartera.

En consecuencia, se pretende entregar un insumo para que la empresa eléctrica de Cotopaxi tenga una guía y esta sea tomada en cuenta para un futuro proyecto de analítica de datos, y con el conocimiento anticipado, la empresa eléctrica en mención, tenga la opción de implementar un enfoque de seguimiento y cobro más exhaustivo a los clientes que tienen un mayor score respecto a la morosidad en el modelo a definirse posteriormente en este proyecto de tesis.

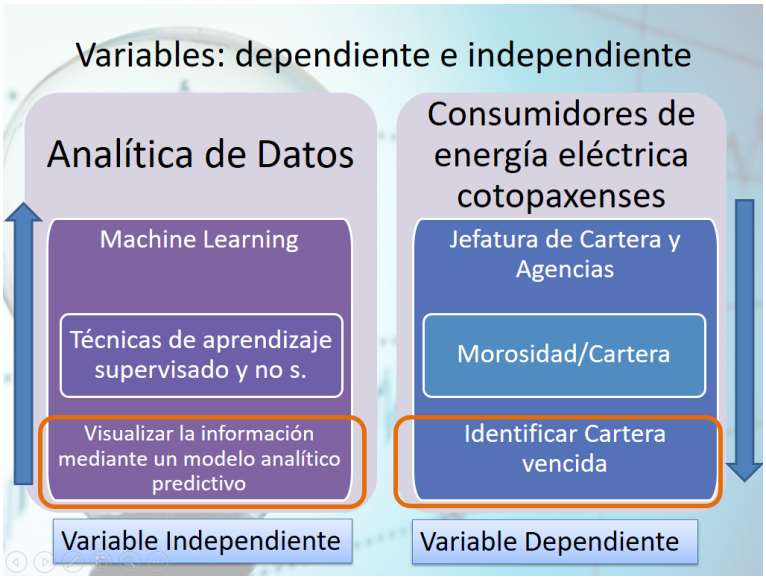
Capítulo II: Marco Teórico

Red de Categorías

Por medio de estos fundamentos se busca tener una coherencia con las variables dependientes e independientes planteadas, de una forma jerárquica, hasta llegar a los conceptos de más bajo nivel, que explicarán el objetivo de este proyecto de Tesis.

Figura 1

Variables: dependiente e independiente



Fundamentación científica de la variable dependiente

Analítica de Datos

Según (Rodríguez & Gamboa, 2019) el término Analítica de Datos (Analytics) es un sustituto más comercial de minería de datos, y tiene un propósito central, que es, la toma de decisiones a partir de la extracción de conocimiento oculto existente en la data, el cual se puede plasmar mediante ciertas pautas, para proyectarlo hacia eventos futuros. Por consiguiente, el objetivo central es la obtención de estos factores intrínsecos mediante: los algoritmos y las

técnicas que hacen uso de estos para su objetivo necesitado.

Estas técnicas de analítica de datos corresponden al Machine Learning, las máquinas de vectores de soporte, el aprendizaje profundo, la inteligencia artificial, entre otras, que pueden ser usadas para distintos objetivos, como la inferencia, y para esto usa el análisis supervisado y el no supervisado. El análisis no supervisado usa técnicas de minería de datos descriptiva o exploratoria para obtener estos factores o pautas, por lo contrario, el análisis supervisado hace uso de técnicas para comprender y predecir un evento futuro con base a los hechos, sobre los cuales, ya se cuenta con información de antemano (Rodríguez & Gamboa, 2019).

Machine Learning

Su objetivo es el uso de técnicas, para que, las computadoras aprendan, en el contexto de que, sepan identificar pautas dentro de la data. Es decir, un algoritmo que revise los datos y sea capaz de predecir eventos futuros.

Adicionalmente, (Joanybel, 2020) sostiene que su aprendizaje se debe mejorar, mediante un mejor desempeño, formando modelos de generalización y asociamiento.

Técnicas de Aprendizaje Supervisado y no Supervisado

Los algoritmos supervisados dependen de datos que estén previamente etiquetados, es decir, de aspectos o ejemplos prácticos que ya pasaron pero que, seguirán surgiendo en el futuro, es decir, el algoritmo aprende a predecir el valor de salida (Umaquinga Criollo, Suárez Zambrano, & Oña Rocha, 2018). Los algoritmos más frecuentes de aprendizaje supervisado son: Árboles de decisión, Clasificación de Naïve Bayes, Regresión por mínimos cuadrados, Regresión Logística, Support Vector Machines (SVM) y Métodos “Ensemble” (Conjuntos de clasificadores).

Para este tipo de algoritmos se provee una gran cantidad de datos, para que estos puedan determinar de qué se tratan sin haberlos etiquetado, en este caso aplica un carácter

exploratorio, por ejemplo, clustering busca agrupamientos basados en similitudes, pero nada garantiza que éstas tengan algún significado o utilidad, sin embargo, se pueden encontrar correlaciones que no se conocían de su existencia. Adicional a los algoritmos de clustering, se tienen análisis de componentes principales, descomposición en valores singulares (singular value decomposition) y Análisis de componentes independientes (Independent Component Analysis).

Modelo Analítico Predictivo

Será el resultado de usar ambos tipos de algoritmos Supervisados y no Supervisados, usando el Machine Learning adicional a técnicas analíticas y estadísticas, normalmente, los datos históricos se utilizan para crear un modelo matemático que capture las tendencias importantes.

Estos modelos pronostican un resultado en algún estado o tiempo futuro en función de los cambios en las entradas del modelo. El objetivo del modelo es evaluar la probabilidad de que un sujeto o evento similar tenga el mismo rendimiento en una muestra diferente o con características de entorno diferentes. Es decir, se buscan pautas de datos ocultos que respondan preguntas sobre el comportamiento en este estudio de los clientes y su relación con la cartera vencida.

De ahí que, se pretende mediante este proyecto de tesis dejar plasmadas las características iniciales al modelo identificado, como mejor opción, según las muestras de la data de consumidores de energía y entregar el estudio inicial para futuros procesos de re aprendizaje del modelo.

Fundamentación científica de la variable independiente

Consumidores de energía eléctrica cotopaxenses

Los consumidores de energía eléctrica de Cotopaxi poseen distinta información geopolítica, se alojan tanto en zonas urbanas como rurales tanto de la Sierra como del sector Costa y dependen de las operaciones de distribución y comercialización del servicio público de energía eléctrica por parte de la empresa eléctrica de dicha provincia, la cual, tiene entre sus obligaciones cobrar a sus consumidores finales las respectivas tarifas por el servicio eléctrico, que es uno de los principales objetivos a analizar en este proyecto de tesis, debido a su relación con la cartera vencida.

Jefatura de Cartera y Agencias

Esta área pretende mejorar sus niveles de recaudación, desea crear políticas efectivas sobre los niveles de cartera vencida, para disminuir sus demoras de cobranza en al menos un mes de facturación.

Además, el área solo posee información descriptiva del total de cartera y total de consumidores por agencia, sector y ruta, pero las usa para preparar acciones reactivas como el corte de luz, tampoco toma en cuenta aspectos como fechas pico en referencia a la afluencia del cliente, entre otros aspectos geopolíticos, como, por ejemplo, las distintas zonas dentro del sector urbano, diferencias entre sectores rurales y marginales y adicionalmente, la provincia de Cotopaxi posee parte de sus usuarios en la región Costa, por ende, aplican distintas realidades.

Morosidad / Cartera

Actualmente no se tienen tendencias respecto a factores que puedan influir en la morosidad de los consumidores y cuales patrones condicionan el comportamiento de la calidad de cartera medida a través del indicador o ratio de morosidad, por ende, no se posee técnicas

para estimar y menos bajo la estructura de datos que posee, por ejemplo, el número de agencias y su posición geográfica, la cantidad de formas de pago, o las tarifas que pueden incrementar o disminuir los valores a pagar según sectores, edades o kilovatios consumidos.

Identificar Cartera Vencida

Actualmente no se tiene visibilidad sobre la información del consumidor, por ejemplo, se le puede estar ofreciendo u otorgando un nuevo medidor a un cliente con tendencia morosa, o de un sector que es común la morosidad. No se tiene una clasificación de clientes según su comportamiento de pago y mucho menos su histórico y tendencia. Por lo que, es totalmente una tarea manual la identificación de la cartera vencida, peor aún determinar si los esfuerzos darán frutos sobre una cartera cobrable y, más aún si se está trabajando sobre una cartera incobrable.

Debido a lo anteriormente mencionado, se pretende evaluar el o los modelos analíticos que mejor se adapten a las distintas realidades geopolíticas de los consumidores de energía eléctrica en Cotopaxi, que permitan identificar la cartera vencida y sobre todo pronosticar o prevenir que futuras cuentas puedan llegar a caer en mora; para esto, se evaluará la precisión del pronóstico según el mayor porcentaje de exactitud obtenido y las características tomadas en cuenta.

Tras lo anterior, se plantea dejar un precedente o puntapié inicial que sirva de guía para que el o los modelos puedan re aprender a futuro sobre nuevas características que puedan surgir, y se las acople mediante las decisiones que se tomen con los resultados arrojados por este proyecto de tesis.

Capítulo III: Metodología

Metodología de la investigación

El presente proyecto de tesis se orienta mediante la adaptación de la metodología de investigación DSR: A Design Science Research Methodology for Information Systems Research (Schorr & Hvam, 2018), la cual, permite diseñar un artefacto de TI entre estos justamente los modelos, esto con el objetivo de resolver un problema organizacional importante y se divide en 6 pasos:

Identificar el Problema

Se debe definir el problema específico y justificar cual es el valor de su solución, al realizar esa justificación se cumple con motivar a la audiencia a perseguir dicha solución y aceptar los resultados. Después de que, el problema es identificado permite ir al siguiente paso para transformarlo en un objetivo.

Definir objetivos para la solución

Ahora que ya se tiene el conocimiento para inferir lo que es posible y realizable. Los objetivos pueden ser cuantitativos o cualitativos, los recursos requeridos para cumplir con este paso, incluyen el estado de los problemas y las soluciones actuales.

Diseño y Desarrollo

Para este proyecto de tesis en el desarrollo del modelo se usarán las mejores prácticas

de la metodología CRISP – DM (Schröer, Felix, & Gómez, 2021), con la ayuda de la investigación realizada, se debe identificar la funcionalidad deseada, su diseño y proceder a crearlo, para este paso se necesita el conocimiento de la teoría para aplicarlo hacia la solución.

Demostración

Es mostrar el uso del modelo que prueba que la inferencia funciona. Se puede usar experimentación, simulación o pruebas.

Evaluación

Se trata de observar y medir que tan bien el modelo respalda a la solución del problema. Se pueden comparar los objetivos que se plantearon para la solución versus los resultados actuales observados. Se necesita conocimientos de técnicas de análisis, para este caso se usarán métricas y las evidencias empíricas se aseverarán según comparaciones con la data disponible.

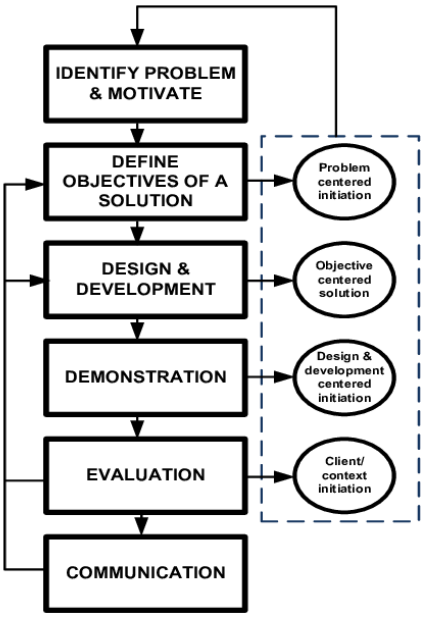
Al final de este paso se puede decidir si se debe iterar hacia atrás al paso tres (Diseño y Desarrollo) para tratar de mejorar la efectividad del modelo o continuar hacia el siguiente paso (Comunicación) y dejar las mejoras adicionales para futuros proyectos de investigación.

Comunicación

En este paso se comunica el problema y su importancia, la innovación y utilidad del modelo, la robustez del diseño y su efectividad.

Figura 2

A Design Science Research Methodology for Information Systems Research



Nota: Imagen tomada de (Schorr & Hvam, 2018)

Finalmente, se puede indicar que no es estrictamente necesario seguir esta metodología en secuencia pues está hecha para varios casos, por ejemplo, si ya se tiene definido el problema central, en el caso de una investigación se puede seguir con la definición de los objetivos, en el caso de la industria se puede pasar directo al diseño y desarrollo.

En este caso ya se tiene definido el problema, así como los objetivos, de tal manera que se procede con el paso 3 partiendo de la información que previamente ya se ha recopilado en la institución.

Preguntas de Investigación

De acuerdo a los objetivos planteados y al análisis realizado se plantean las siguientes preguntas que se irán contestando a lo largo de la presente tesis:

OE1 – RQ1.1: ¿Qué estudios existen sobre técnicas usadas en minería de datos relacionados a problemas en identificación de cartera vencida?

OE1 – RQ1.2: ¿Qué técnica de Machine Learning es la más usada en el ámbito financiero?

OE2 – RQ2.1: ¿Cuál(es) modelo(s) predictivo(s) se adapta(n) a la realidad de la empresa eléctrica de Cotopaxi?

OE2 – RQ2.2: ¿Cuáles son los patrones que tienen los consumidores con morosidad en la empresa eléctrica de Cotopaxi?

OE3 – RQ3.1: ¿Qué técnica de análisis o métrica se puede usar para evaluar el/los modelo(s) planteado(s) sobre cartera vencida?

OE3 – RQ3.2: ¿El modelo planteado permitirá identificar patrones que faculten generar grupos de consumidores con dimensiones aún no conocidas en la empresa eléctrica de Cotopaxi?

Estado del Arte

Para obtener la información actual sobre este objetivo de investigación planteado, se han usado las primeras fases de un SMS, las cuales son los criterios de inclusión y la estrategia de búsqueda, para tener un acercamiento a la realidad ecuatoriana se ha buscado en primera instancia fuentes relacionadas a tesis similares como ResearchGate, las mismas que se van referenciando dentro de este proyecto de tesis, con estas fuentes se han identificado los temas y palabras clave a buscar en Sitios como IEEE y Springer.

En cuanto a la **definición del objetivo** de investigación planteado se han usado las preguntas de investigación de la sección: ***Preguntas de Investigación***

Mediante los **criterios de inclusión y exclusión** se encuentran estudios con similares características del tema planteado, entre estos se encuentran los siguientes:

Criterios de inclusión:

- Artículos desde el año 2014
- Artículos de preferencia en idioma inglés caso contrario en español.
- Artículos relacionados a minería de datos y Machine Learning y temas de finanzas, cartera y pagos tardíos.

Criterios de exclusión:

- Artículos con temas de machine learning no relacionados a temas financieros.
- Artículos con temas de Business Intelligence sin temas de predicción.

En cuanto a la **definición de la estrategia de búsqueda, en la revisión inicial,** como se mencionó al inicio del ***Estado del Arte*** se ha buscado en primera instancia fuentes relacionadas a tesis similares existentes, lo cual, brinda una visión de qué clase de estudios buscar.

Se ha realizado posteriormente una **validación cruzada de estudios**, para verificar que los criterios de inclusión y exclusión correspondan a los estudios identificados, obteniendo así el listado inicial de documentos que sirvan de base para las siguientes fases de esta tesis.

La **Integración del Grupo de Control** lo conforman los estudios que de acuerdo a los criterios de inclusión y exclusión han aparecido en la búsqueda, con estos se procede a realizar un primer análisis su título, introducción, palabras claves y conclusiones, como se muestra en la

siguiente tabla:

Tabla 1

Tabla de validación cruzada de Grupo de Control

GC	Título	Palabras Clave
EC1	A Novel Noise-Adapted Two-Layer Ensemble Model for Credit Scoring Based on Backflow Learning	Credit Scoring Model pattern classification Predictive models support vector machines genetic algorithms Backflow Learning machine learning
EC2	Guided Fast Local Search for speeding up a financial forecasting algorithm	Forecasting predictive algorithms finance Equations Metaheuristic algorithm
EC3	An Application of "Neuro-Logit" New Modeling Tool in Corporate Financial Distress Diagnostic	Artificial neural network Financial model Forecasting Predictive models
EC4	Towards An Efficient Real-time Approach To Loan Credit Approval Using Deep Learning	Artificial neural network Deep learning Loan Credit Credit Scoring Model
EC5	An Improved Bank Credit Scoring Model: A Naïve Bayesian Approach	Credit Risk Classification Scoring artificial intelligence Bayesian bank data processing Credit Scoring Model Financial Model
EC6	Heterogeneous Ensemble for Default Prediction of Peer-to-Peer Lending in China	Credit Scoring Deep Learning Neural network

GC	Título	Palabras Clave
		Credit Scoring Model
		Credit Loan

Nota. Esta tabla muestra la validación cruzada aplicada a los estudios identificados

Para la **Construcción de la cadena de búsqueda** se analizan los estudios del grupo de control, encontrando palabras comunes y palabras independientes de cada estudio que estén de acorde al objetivo de Tesis, con ello se conformaron los contextos: Credit Data, Finance, Data Analysis y Predictive.

Tabla 2

Tabla de contextos

		EC1	EC2	EC3	EC4	EC5	EC6	# Repeticiones
Credit Data	Forecasting	1	1	1	0	1	1	5
	Credit scoring model	1	0	0	1	1	1	4
	bank data processing	0	0	0	0	1	0	1
	Credit Loan	0	0	0	1	0	1	2
Finance	Finance	0	1	1	1	1	1	5
	Business	0	0	1	1	0	1	3
	Financial model	0	0	0	0	1	0	1
	Bank credit	0	0	0	0	1	0	1
Data Analysis	Credit risk	1	0	0	0	1	1	3
	Noise adaption approach	1	0	0	0	0	0	1
	Equations	1	0	1	0	0	0	2
	Metaheuristic algorithm	0	1	0	0	0	0	1
	Pattern Classification	1	0	0	1	1	0	3
Predictive	Backflow Learning	1	0	0	0	0	0	1
	Deep Learning	0	0	0	1	0	1	2
	Genetic algorithm	1	0	0	0	1	0	2
	Bayesian	0	0	0	0	1	1	1
	Artificial neural network	1	0	1	1	0	1	4

Nota. Esta tabla muestra los contextos para la cadena de búsqueda

Una vez identificados los contextos de búsqueda se usan las palabras que se repiten para la **construcción de la cadena**:

((Forecasting OR Credit Scoring Model OR bank data processing OR credit loan)

AND (finance OR business OR bank credit)

AND (Credit Risk OR pattern classification OR Equations)

AND (bayesian OR Backflow Learning OR Deep Learning OR genetic algorithms OR Artificial neural networks))

Esta cadena de búsqueda ha sido ejecutada sobre IEEE con el filtro de años desde el 2014 en adelante con un resultado de 60 artículos entre conferencias, revistas y cursos.

https://ieeexplore.ieee.org/search/searchresult.jsp?action=search&matchBoolean=true&queryText=(((Forecasting%20OR%20Credit%20Scoring%20Model%20OR%20bank%20data%20processing%20OR%20credit%20loan)%20%0AAND%20(finance%20OR%20business%20OR%20bank%20credit)%20%0AAND%20(Credit%20Risk%20OR%20pattern%20classification%20OR%20Equations)%20%0AAND%20(bayesian%20OR%20Backflow%20Learning%20OR%20Deep%20Learning%20OR%20genetic%20algorithms%20OR%20Artificial%20neural%20networks)))&highlight=true&returnType=SEARCH&matchPubs=true&ranges=2014_2019_Year&returnFacets=ALL

Revisión de los documentos encontrados:

(Wei, Yang, Zhang, & Zhang, 2019) **A Novel Noise-Adapted Two-Layer Ensemble Model for Credit Scoring Based on Backflow Learning**

En este documento los autores indican que en problemas de clasificación como Credit Scoring (Puntuación o calificación para Créditos) los algoritmos de Machine Learning e Inteligencia Artificial se han vuelto muy importantes, el hecho de construir un modelo de aprendizaje íntegro que ya ha sido probado y será típicamente más exacto y robusto que otros

clasificadores individuales permitirá tener una importante herramienta de gestión. En este documento el modelo íntegro de reducción de ruido adaptado de dos capas (reducción de data errada) aplicado a la calificación para créditos se basa en el conocimiento mediante reflujo o Backflow, el cual, integra a cinco clasificadores base:

- Extreme gradient boosting (potenciación de gradiente extremo),
- Gradient boosting decision tree (potenciación de gradiente en árboles de decisión),
- Support vector machine (Máquinas de vectores de soporte),
- Random forest (Bosques aletaorios),
- Linear discriminant analysis (Análisis discriminante linear)

A la final se obtiene un resultado de predicción basado en la fusión de todos estos clasificadores mediante el conocimiento de reflujo o backflow eliminando o reduciendo la data errada con un mejor desempeño que otros modelos.

(Shao, Smonou, Kampouridis, & Tsang, 2014) **Guided Fast Local Search for speeding up a financial forecasting algorithm**

Este Paper trata sobre una Búsqueda local guiada que es un algoritmo meta-heurístico poderoso aplicado en el pronóstico financiero pero que tiene un gran costo en cuanto a la necesidad de infraestructura computacional y se espera lidiar con este problema mediante la combinación con una Búsqueda local rápida, el cual se estima baje este costo computacional en un impresionante 77% implementado en árboles de decisión basados en algoritmos genéticos.

(Almonayirie, 2015) **An Application of "Neuro-Logit" New Modeling Tool in Corporate Financial Distress Diagnostic**

Este artículo representa un estudio proactivo mediante la introducción de una nueva herramienta de modelado, para diagnosticar las dificultades y valores financieros y su probabilidad de ocurrencia. El modelo "Neuro-Logit" es un nuevo enfoque para el diagnóstico, predicción y pronóstico de dificultades financieras corporativas. Esta herramienta actúa como Logit (Analísis de regresión logística), pero las ecuaciones son construidas basadas en un algoritmo ANN (Red neuronal artificial), al combinarse ambos se propone el modelo Neuro-Logit, el cual reduce las limitaciones de ANN y Logit y tiene una mejor precisión de modelos Logit tradicionales. Este paper puede ser considerado como un segundo intento de un modelo de predicción supervisado de dificultades financieras y a la vez represente un enfoque empírico e innovador de modelo donde ANN es usado como una herramienta estadística.

(Abakarim, Lahby, & Attioui, 2018) **Towards An Efficient Real-time Approach To Loan Credit Approval Using Deep Learning**

En este Paper se explica que en la última década ha visto un importante aumento de la recopilación de datos, especialmente en los sectores financieros. Los bancos son de hecho uno de los mayores productores de Big Data, una realidad es que ninguna otra compañía tiene más datos de sus clientes que los bancos. Recolectar y analizar esta información es una característica clave para la toma de decisiones como la aprobación de créditos. El reto está en saber cómo construir una herramienta de explotación de data personal de forma proactiva, poderosa, responsable y ética para hacer que las propuestas de aplicación de créditos sean más relevantes y personalizadas. Para esto Machine Learning se está prometiendo como una solución adecuada

para lidiar con este problema, pero a pesar de que muchos algoritmos han sido propuestos, ninguno de estos ha tomado en consideración el paradigma del proceso en Tiempo Real. En este Paper se propone a el modelo de clasificación Binaria en tiempo real, como una solución a la aprobación de créditos, este modelo se basa en Redes Neuronales Profundas y esto permite clasificar a los aplicantes en Buenos o Malos según su riesgo, adicionalmente el modelo superó a otros en términos de precisión (accuracy) y recuperación (recall – positivos reales).

(Okesola, Okokpujie, Adewale, John, & Omoruyi, 2018) **An Improved Bank Credit Scoring Model: A Naïve Bayesian Approach**

En este Paper se habla sobre la puntuación para créditos, la cual es una herramienta usada por organizaciones para dar o negar créditos solicitados por sus clientes. Pero a pesar de que varias propuestas sobre Inteligencia Artificial han sido usadas en la puntuación para créditos, así como en la evaluación de riesgo crediticio, uno de los algoritmos mejor calificados en Minería de Datos como lo es enfoque Naive Bayes (o algoritmo bayesiano inocente) no ha sido extensamente usado, por lo que usando información demográfica como variables de entrada este Paper investiga la habilidad de que este algoritmo permita ser un clasificador en el sector.

(Li, Ding, Chen, & Yang, 2018) **Heterogeneous Ensemble for Default Prediction of Peer-to-Peer Lending in China**

En este Paper se habla sobre el concepto de peer-to-peer (P2P) lending (préstamos entre particulares) esto de alguna manera se puede relacionar con el consumo de energía eléctrica y el pago ya que no es un tema crediticio de la Banca convencional per sé, pero de igual forma habla de evitar el riesgo crediticio, mediante el uso de Conjuntos de modelos o

bosques de decisión, donde se entrenan varios modelos, cada uno de ellos con un subconjunto aleatorio de los datos. En este Paper se ha diseñado un modelo de aprendizaje múltiple con conjuntos heterogéneos aplicando la técnica de aprendizaje de potenciación de gradiente extrema (XGBoost), redes neurales profundas y regresión logística, considerando su aprendizaje de forma individual para posteriormente ser ponderados mediante una fusión lineal.

Los estudios antes mencionados responden la pregunta de investigación inicial: OE1 – RQ1.1: ¿Qué estudios existen sobre técnicas usadas en minería de datos relacionados a problemas en identificación de cartera vencida?, tomando en cuenta que el histórico crediticio puede ser comparado con el comportamiento de pago del consumo de KWH en el sentido de que genera una deuda por pagar, al igual que, un acuerdo de pago que se maneja en la empresa eléctrica de Cotopaxi, si bien, no se encontraron estudios enfocados 100% en empresas eléctricas, las técnicas de minería de datos mencionadas se pueden usar para este proyecto de tesis al estar relacionadas.

Respecto a la pregunta de estudio sobre OE1 – RQ1.2: ¿Qué técnica de Machine Learning es la más usada en el ámbito financiero?; se puede indicar que no existe una receta mágica o una sola técnica, muchos usan redes neuronales y fusiones de modelos potenciados (boosted) a medida que la data evoluciona, nuevas variables aparecen o se mejoran como con la limpieza de datos, por lo cual, se deberá realizar una comparativa de los modelos más usados y potenciados (acción a realizar en posteriores capítulos) con la data de la empresa eléctrica en mención y así poder determinar, cuál modelo es el que más se adapta.

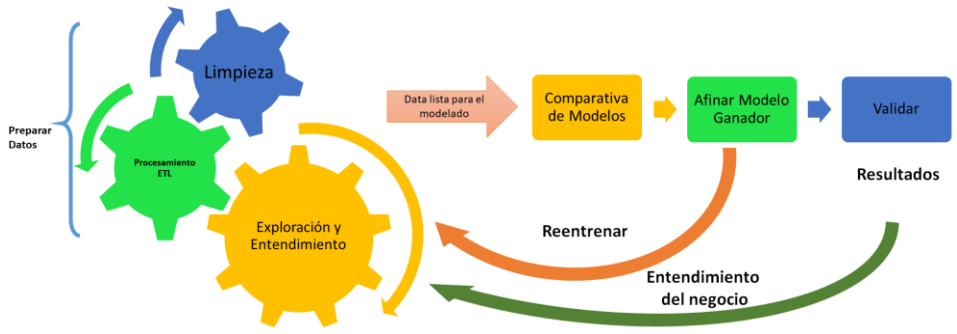
Capítulo IV: Exploración y preparación de los datos

Obtención de datos

El proceso de la analítica descriptiva y posterior análisis predictivo empieza por el entendimiento del negocio y su data, pasando por la preparación y exploración de datos, estos procesos se enfocan en obtener esas variables o características relacionadas para la necesidad o problemática planteada, que para este proyecto de tesis está relacionada con la identificación de la cartera vencida, es por ello que se parte de la metodología de CRISP-DM (Health Data Miner, 2021) así como del ciclo de vida de la analítica de datos y operacionalización de modelos (Herrera, 2019), como se muestra en las dos imágenes siguientes:

Figura 3

Ciclo de vida de la analítica de datos y operacionalización de modelos



Por lo que, basado en el entendimiento del negocio y de la data que ya dispone la empresa eléctrica de Cotopaxi y que se la tiene por medio de sus actuales esfuerzos en cuanto a reportería y dashboards o cuadros de mando gerenciales (Garzón Ulloa & Chicaiza Castillo, 2017), este proceso inicial de obtención de datos en el presente trabajo de investigación se centra en explorar los inputs relacionados a la identificación de la cartera vencida y consolidarlos en una tabla analítica tomando en cuenta este primer esfuerzo que posee la institución donde sus fuentes son:

Tabla 3

Tabla de inputs

Conceptos	Fuente	Inputs
Fuente de Datos	AS400/DB2	Sistema Comercial
Información Gerencial	QlikSense	Dashboards
Semejante a un datawarehouse	Archivos QVD	Información integrada

Nota. Esta tabla muestra las fuentes de datos para el análisis inicial

Dentro de estas fuentes es importante, tomar los primeros conceptos en los cuales se basa la actividad principal de la Institución como el objeto de estudio de esta tesis, para esto, las entidades más relevantes son: el cliente, su cuenta y medidor, su consumo y deuda, de los cuales parten los ítems que son de extrema relevancia para este estudio y posterior modelo, dentro de este ámbito o alcance, se ha identificado primero las fuentes para la obtención de los dashboards actuales en QlikSense que posee la institución y de estos se ha podido obtener los siguientes campos:

Tabla 4

Tabla de primeras características analizadas en QlikSense

Entidades/Conceptos	Características
Clientes	Estado de Cuentas
	Tipo de tarifa (Residencial)
	Agencia, sector y ruta,
	Cantón, Parroquia
	Estado de la Cuenta
	Tipo de Medidor
Información Personal	
	Clase de cliente: Urbano/Rural/Marginal
	Condición del ciudadano
	Estado Civil
Planillas	
	Deuda en dólares (relacionado a la cartera)
	Número de planillas adeudadas (relacionado a la cartera)

Entidades/Conceptos	Características
Información histórica (año, mes)	
	Valores facturados en dólares,
	KWH facturados (consumo)
	Comparación sobre Recaudación efectuada

Nota. Esta tabla muestra los primeros conceptos detectados en QlikSense

Si bien en primera instancia esta es la data que cuenta en sus Dashboards, existe más data almacenada en los archivos QVD que se generan a partir de la carga en memoria (RAM) que permite realizarla el QlikSense, la cual, se conecta hacia la Base de Datos de DB2, como, por ejemplo, la fecha de nacimiento, por lo que, se ha tenido que ir al modelo inicial de Qliksense y usar dichos conceptos o características, lo que permitirá contrarrestar con la información no arrojada en los dashboards actuales.

Estos datos han sido obtenidos mediante una consulta (o query) compuesto, propio del lenguaje SQL de QlikSense, el cual genera archivos QVD que se equiparan en cierto sentido a un Datawarehouse e igual poseen la lógica de los ETLs al tener una calendarización para su carga progresiva o incremental, mediante la opción que brinda dicho lenguaje de insertar, actualizar y borrar, con lo cual, la data permanece actualizada.

Para obtener una nueva fuente de información (staging) a partir de los modelos QVD de QlikSense que contienen las mismas tablas de la Base de Datos DB2 como se indica a continuación:

Tabla 5

Tabla de entidades mapeadas a partir de QlikSense

Entidades	Conceptos
SCEDTAV6.CUENTAS	Cuentas
SCEDTAV6.SCEF20	Medidores
SCEDTAV6.COTOPAXI	Localización

Entidades	Conceptos
SCEDTAV6.SCEL22DR	Planillas pagadas o canceladas
SCEDTAV6.SCEL2220	Planilla facturadas
SCEDTAV6.SCEL2102	Lectura de medidores
SCEDTAV6.SCEL2206	Deuda o Cartera

Nota. Esta tabla muestra las entidades obtenidas desde QlikSense

Se procedió a realizar en primer lugar la carga de la data de producción obtenida desde la empresa eléctrica de Cotopaxi mediante procesos de ETLs para generar un propio Staging, para este proyecto de tesis, donde se puedan hacer las exploraciones manuales sin afectar a la base de datos de producción, tomando en cuenta que las buenas prácticas de Ciencia de Datos nos indican que los científicos de datos trabajarán finalmente con la data real.

Posteriormente, se procede con una etapa de entendimiento de la data y una exploración manual de acuerdo al entendimiento del negocio adquirido que se lo detalla en la siguiente sección.

A partir de estos archivos QVD ha sido exportada la data hacia archivos del tipo CSV, por cada cliente, con sus cuentas, medidores y los demás datos asociados mes a mes, según las planillas de pago por el consumo de KWH, ya que según la investigación realizada y las consultas realizadas a Grupo Novatech (grupo-novatech, 2021) que es una de las empresas partner de QlikSense en Ecuador, al tener en la empresa eléctrica una versión de Qlik 2.1.1, la cual, no soporta la analítica predictiva que si lo hace QlikSense en la nube, mediante la herramienta Qlik Connectors, por no disponer de un contrato de soporte, esta no permite su actualización, ni su integración con los servicios de Qlik en la nube, por lo que, la única opción es exportar la información de su modelo en formato CSV para tener dicha información en tecnologías más actuales y potentes en relación al Machine Learning, en consecuencia, se realizó esta exportación hacia CSV.

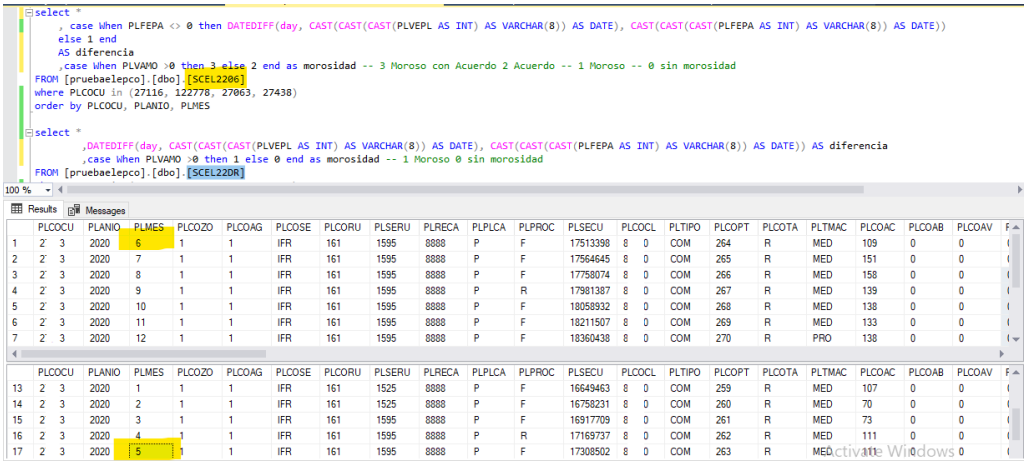
Entendimiento y selección de datos

Basados en el conocimiento de los funcionarios del área de Sistemas, del área Comercial de la empresa eléctrica de Cotopaxi y de los KPIs en los dashboard actuales de la institución, se pudo delimitar los conceptos o características que se relacionan a la cartera y al cálculo de mora, adicional, a los conceptos del lado del cliente y cuentas, que gozan de control o información relevante previo a plantear un modelo, también se ha limitado a datos residenciales, ya que, es donde existe mayor cantidad de información del propietario de las cuentas.

De esta información y conocimiento recopilado, se identifica que, además de los casos de planillas pagadas sean a tiempo o con morosidad, existen también los acuerdos de pago que cayeron o no en morosidad, por lo que, para consolidar una primera versión del tablón analítico se deben integrar las entidades: SCEL22DR (planillas pagadas), y SCEL2206 (Pagos con convenio), como preparación de la data.

Figura 4

Entendimiento de la data con exploraciones manuales

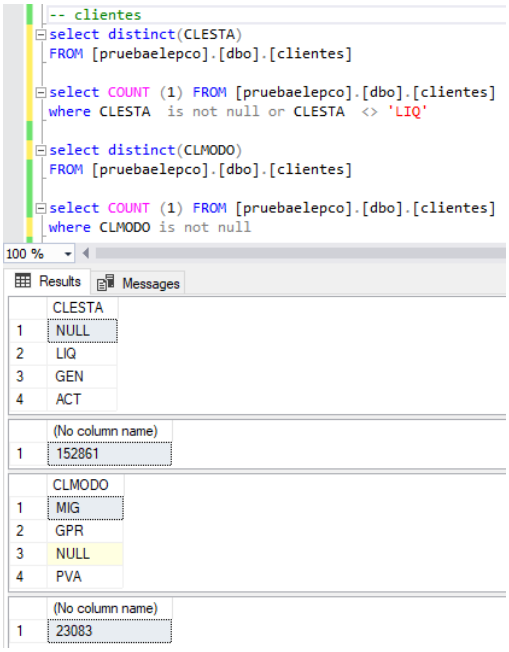


Dada esta información se entiende que un 13,70% de pagos de tarifas se han realizado

con convenios, esto se incrementa en el 2020 debido a la pandemia. Luego del entendimiento de la data, se realizan conteos manuales de cada una de las características de cada entidad, obteniendo los distintos tipos de datos para revisar que conceptos engloban a la mayoría de información, y que otros conceptos pueden estar incluso en su mayoría nulos, a estos últimos se los descarta, puesto que, la mayoría de cuentas darían valores vacíos en dichas características.

Figura 5

Conteos manuales de características encontradas



Además, se integran las entidades: SCEL22DR (planillas pagadas), y SCEL2206 (planillas vencidas con acuerdos) con la información de cuentas, medidores, clientes y se genera una primera versión del tablón estadístico con 65 características mediante procedimientos almacenados y el uso de ETLs, dentro de estas características se identifica también que si la fecha de pago es mayor a la fecha de vencimiento, ya se tiene morosidad, por lo cual, se crea mediante consultas SQL la respectiva fórmula generando un valor numérico de la siguiente manera:

Tabla 6

Tabla de primeras características analizadas

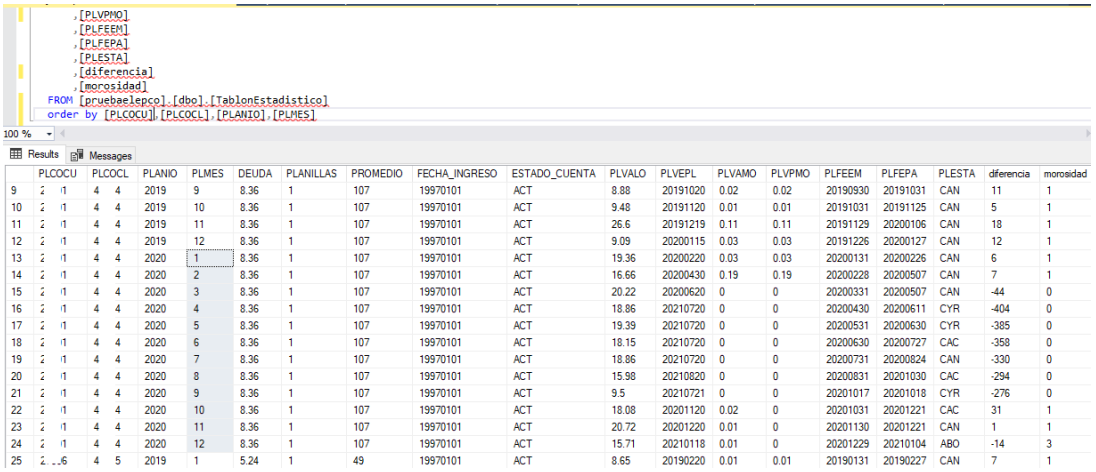
Valor Numérico	Tipo de pago
3	Pago con morosidad y con acuerdo de pago
2	Pago con acuerdo sin morosidad
1	Pago con morosidad
0	Pago sin morosidad

Nota. Esta tabla muestra los diferentes escenarios de pago

De forma adicional, se coloca la diferencia en días entre dichas fechas detectando que las fechas con anticipación de pago de más de un mes, son las que tienen acuerdos de pago, adicional a que las alojan también en la entidad SCEL2206.

Figura 6

Primera versión del tablón estadístico con diferencia de pago e indicador de morosidad



```
      ,[PLVPMO]
      ,[PLFEEM]
      ,[PLFEPA]
      ,[PLESTA]
      ,[diferencia]
      ,[morosidad]
  FROM [pruebaelepco].[dbo].[TablonEstadistico]
  order by [PLCOCU],[PLCOCL],[PLANIO],[PLMES]
```

100 % — Results | Messages

	PLCOCU	PLCOCL	PLANIO	PLMES	DEUDA	PLANILLAS	PROMEDIO	FECHA_INGRESO	ESTADO_CUENTA	PLVALO	PLVEPL	PLVAMO	PLVPMO	PLFEEM	PLFEPA	PLESTA	diferencia	morosidad
9	2	1	4	4	2019	9	8.36	1	107	19970101	ACT	8.88	20191020	0.02	0.02	20190930	20191031	CAN	11	1
10	2	1	4	4	2019	10	8.36	1	107	19970101	ACT	9.48	20191120	0.01	0.01	20191031	20191125	CAN	5	1
11	2	1	4	4	2019	11	8.36	1	107	19970101	ACT	26.6	20191219	0.11	0.11	20191129	20200106	CAN	18	1
12	2	1	4	4	2019	12	8.36	1	107	19970101	ACT	9.09	20200115	0.03	0.03	20191226	20200127	CAN	12	1
13	2	1	4	4	2020	1	8.36	1	107	19970101	ACT	19.36	20200220	0.03	0.03	20200131	20200226	CAN	6	1
14	2	1	4	4	2020	2	8.36	1	107	19970101	ACT	16.66	20200430	0.19	0.19	20200228	20200507	CAN	7	1
15	2	1	4	4	2020	3	8.36	1	107	19970101	ACT	20.22	20200620	0	0	20200331	20200507	CAN	-44	0
16	2	1	4	4	2020	4	8.36	1	107	19970101	ACT	18.86	20210720	0	0	20200430	20200611	CYR	-404	0
17	2	1	4	4	2020	5	8.36	1	107	19970101	ACT	19.39	20210720	0	0	20200531	20200630	CYR	-385	0
18	2	1	4	4	2020	6	8.36	1	107	19970101	ACT	18.15	20210720	0	0	20200630	20200727	CAC	-358	0
19	2	1	4	4	2020	7	8.36	1	107	19970101	ACT	18.86	20210720	0	0	20200731	20200824	CAN	-330	0
20	2	1	4	4	2020	8	8.36	1	107	19970101	ACT	15.98	20210820	0	0	20200831	20201030	CAC	-294	0
21	2	1	4	4	2020	9	8.36	1	107	19970101	ACT	9.5	20210721	0	0	20201017	20201018	CYR	-276	0
22	2	1	4	4	2020	10	8.36	1	107	19970101	ACT	18.08	20201120	0.02	0	20201031	20201221	CAC	31	1
23	2	1	4	4	2020	11	8.36	1	107	19970101	ACT	20.72	20201220	0.01	0	20201130	20201221	CAN	1	1
24	2	1	4	4	2020	12	8.36	1	107	19970101	ACT	15.71	20210118	0.01	0	20201229	20210104	ABO	-14	3
25	2	16	4	5	2019	1	5.24	1	49	19970101	ACT	8.65	20190220	0.01	0.01	20190131	20190227	CAN	7	1

Las columnas de morosidad, cartera vencida, diferencia entre el día de pago y la fecha de emisión anteriormente mencionadas, se crearon para no caer en el sesgo de la disponibilidad, es decir, opinar o tomar decisiones solo con la información disponible en la burbuja actual de la Base de Datos. Por lo cual, se ha enriquecido la información con las columnas calculadas que ayudan o mejoran al entendimiento de los casos y al rendimiento de

los algoritmos.

Tratamiento de los Datos

Como primer paso se debe reconocer que características o conceptos podrían no ser necesarias y removerlas del set de datos:

Por ejemplo, la cédula si es catalogada como un ID podría ser no considerada en el modelo, si bien los dos primeros dígitos de la cédula pueden indicar la provincia de procedencia de la persona, este dato no ha sido considerado en el modelo por ser información confidencial que no se la puede usar, y adicional, ya se tiene por separado las categorías de provincia, cantón, parroquia y sector, por lo que, no se la necesita.

Otro campo puede ser el nombre de la persona que no ha sido considerado en este caso, ya que, no aporta al objetivo del modelo a encontrar. En el caso de características con datos intuitivos, por ejemplo, las cuentas prepago, se descartan, ya que, como su nombre lo indica, se tuvieron que pagar previo a su uso, caso contrario el cliente no tendrá energía eléctrica, estas sumaron un total de 189 cuentas, adicional se descartan las cuentas que ya han sido liquidadas y cortadas, partiendo así de un total de 135.995 cuentas hacia una reducción inicial de: 116.359 cuentas.

Tabla 7

Tabla de ejemplo de características intuitivas

Estado de Cuentas	Tipo de estado	Descripción
PRP	Cuenta prepago	189
LIQ	Cuenta Liquidada	11428
SUD	Cuenta suspendida	17864
COR	Cuenta cortada	8019
ACT	Cuenta Activa	98495

Nota. Esta tabla muestra ejemplos de conceptos intuitivos de la data

Para columnas que no son intuitivas y que de alguna u otra forma no se relacionen con el concepto de búsqueda o modelo, se podría explorar, por ejemplo, si tienen 100% de ser únicas, entonces, bien pudiesen ser consideradas como un ID, que no ayudan con el objetivo del modelo, por ejemplo, el número de medidor, número de cuenta o número de cliente, aunque se debe tomar en cuenta estos IDs a la hora de verificar si un cliente tiene más de una cuenta.

También se debe tomar en cuenta los valores relacionados al consumo de KWH que, si son numéricos continuos y que, si aportan en la predicción, adicionalmente, aspectos importantes como el número de planillas atrasadas por pagar, en primera instancia estas características aparecen de acuerdo a los conceptos de negocio o experiencia en cuánto al entendimiento de la data. Pero también existe el cálculo de la correlación, para esto se puede usar herramientas como el propio Microsoft Power BI donde se tienen algoritmos de jerarquía y clúster que ya integran R de forma nativa o transparente para el usuario final (se detalla más a fondo el uso de Power BI en la sección ***Preparación y Exploración de los datos***), se puede indicar que, el coeficiente de correlación mide el grado de asociación lineal entre dos variables cuantitativas.

Al hablar solo de variables cuantitativas, en primera instancia, se está dejando de lado a las variables categóricas o cualitativas, que posteriormente serán estudiadas y tratadas para ser finalmente convertidas en nuevos valores de indicadores o Indicator Values, lo cual, se detallará en el ***capítulo 5***.

De esta primera matriz de correlación se puede observar los siguientes aspectos:

Tabla 8

Tabla de la matriz de correlación

Correlación	**Característica 1**	**Característica 2**
Significativa (+)	Planillas adeudadas actualmente	Morosidad
Significativa (+)	Valor por mora mensual	Deuda actual
Significativa (+)	Fecha de mora	Fecha de nacimiento
Moderada (+)	Deuda actual	Morosidad
Moderada (+)	Planillas adeudadas actualmente	Deuda actual
Moderada (-)	Planillas adeudadas actualmente	Fecha de Pago
Moderada (+)	Planillas adeudadas actualmente	Cartera Vencida
Débil (-)	Fecha de Pago	Morosidad
Débil (+)	Valor por mora mensual	Morosidad
Débil (+)	Planillas adeudadas actualmente	Valor por mora mensual

Nota. Esta tabla muestra correlaciones positivas y negativas

De aquí que, por ejemplo, las correlaciones negativas indican que, a mejor fecha de pago o mayor rapidez en el pago, menor cantidad de planillas adeudadas tendrá y, por ende, menor morosidad.

Tabla 9

Tabla de detalle de características

Nombre	**Descripción**
Morosidad	Planillas categorizadas según concepto de pago
Cartera vencida	Cantidad de planillas pagadas de forma tardía
PLVAMO	Valor por mora mes a mes
Planillas adeudadas actualmente	Planillas con retraso o mora con corte al último mes de obtención de esta Data
PLFEPA	Fecha en la que pagó el cliente

Nombre	Descripción
PLFEMO	Fecha en la que se ingresó la mora en el sistema debido al atraso en el pago de la planilla
Deuda actual	Deuda de las planillas no pagadas o con retraso con corte al último mes de obtención de esta Data
Fecha de Pago	Fecha de cada mes donde ha pagado el cliente
Fecha_Nacimiento	Fecha de nacimiento del cliente
Diferencia	Cantidad de días de atraso o días a tiempo, respecto a la fecha de vencimiento del pago de la planilla

Nota. Esta tabla muestra el detalle de las primeras características correlacionadas

La relación entre morosidad y cartera vencida parten de un mismo concepto, por eso su estrecha correlación o correlación fuerte de 0.91, lo cual, haría caer en un concepto de tautología que se lo detalla más a fondo en la sección ***Preparación y Exploración de los datos***.

De ahí que, si se filtran valores atípicos (el tratamiento de los valores atípicos se lo ve a más detalle en la siguiente sección) de ciertas características o se realiza una primera limpieza de datos filtrando clientes sin fecha de nacimiento o cuentas que han caído en mora, pero han pedido posteriormente convenios de pago o diferimiento, esto correlaciona a la diferencia en días respecto al no atraso en el pago, como se lo detalló al inicio de esta sección.

Dando como resultado correlaciones más estrechas, donde, por ejemplo, la fecha de nacimiento, ya tiene una correlación débil con la morosidad y la diferencia de días de atraso o días a tiempo, respecto a la fecha de vencimiento del pago de la planilla, también tiene una correlación moderada con la morosidad

Tras esta primera versión de características, se observan también cuales se relacionan dentro de los KPIs de la institución en relación a la cartera vencida, como, por ejemplo, PLCBAS (Consumo base), como se indica en la siguiente imagen:

Figura 7

Matriz de correlación con características adicionales

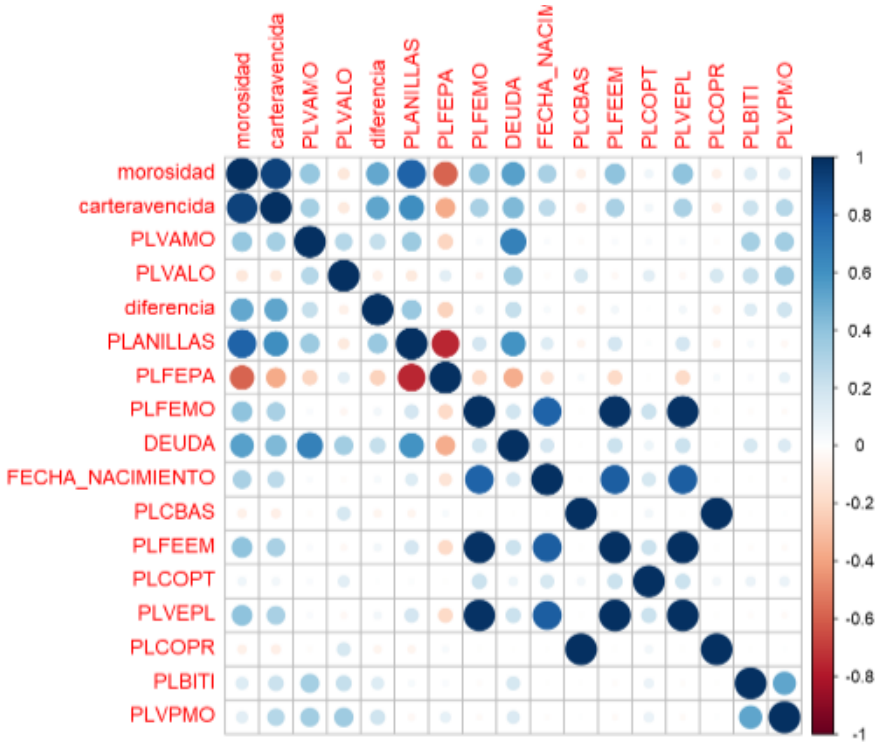


Tabla 10

Tabla de características adicionales

Nombre	Descripción
PLCBAS	Consumo base
PLFEEM	Fecha de emisión de la planilla
PLCPOT	Código del tipo de pliego tarifario
PLVEPL	Fecha de vencimiento de la planilla
PLCOPR	Consumo promedio
PLBITI	Base imponible tarifa IVA
PLVPMO	Valor pagado de la mora

Nota. Esta tabla muestra el detalle de características adicionales

Existen características como el valor de la lectura de consumo de KWH (PLLAAC), que en esta primera instancia, ya se encapsularía en la columna de diferencia de consumo del mes

anterior, con el concepto de consumo de energía activa (PLCOAC), sin embargo, si se toman en cuenta a dichos conceptos o características en este primer proceso, pero otras características como si se emitió una factura, recibo o el número de comprobante de pago no, ya que, no son necesarios o no crean cambios en los modelos, tanto por un tema de experiencia como de relación entre características, sin embargo, posteriormente se lo confirmará con aspectos como la misma correlación de Pearson.

Cabe indicar que, esta funcionalidad de correlaciones de Power BI solo compara características numéricas, no categóricas, por lo que, se irán comparando posteriormente los cálculos de las correlaciones de estas otras características, como, por ejemplo, PLTMAC que indica el tipo de medición del consumo, es decir, si ese mes fue medida la lectura de consumo eléctrico o fue promediada, que al ser un valor categórico se lo tomará en cuenta en el siguiente capítulo.

Datos Atípicos

Ya que, existe un sin número de factores, como la no medición automática debido al tipo de medidor o un sector de difícil acceso, qué, por ende, provoca una facturación con el consumo promedio, o fallas en la medición manual dependiendo del tipo de medidor, lo cual, puede originar reclamos, se tiene un sin número de cambios en los consumos y en sus valores (re facturaciones) y casos atípicos por distintos factores como las fugas de energía eléctrica que generan un gasto excesivo, por estos motivos, se recurre al último dato generado para el cálculo de la planilla o monto a pagar.

Por otra parte, se debe eliminar o reemplazar los datos nulos, ya que estos impactan de una forma no deseada al modelo, por ende, se tomará la decisión de limpiar o arreglar este tipo de valores según el caso o tipo de información.

También, en el caso de características con una mayor cantidad de registros con valores nulos que reales, se las excluye, ya que, no ayudarán al objetivo del modelo y se necesita una limpieza de datos desde el lado del control de ingreso de la misma data, lo cual, no es parte del alcance de este estudio, los campos que tienen porcentajes del máximo 20% para abajo si se pueden adaptar de mejor manera a una limpieza de datos, como se indicaba en la consultoría aplicada a la empresa eléctrica de Cotopaxi, por la empresa Le Infinite (Diagnóstico de la Calidad de Datos de la Información de Misión Crítica.), por ejemplo, mediante la media, lo cual, se lo verá más adelante.

Adicionalmente, existen diferentes tipos de tarifas donde los montos de un medidor residencial no serán nada comparables con los montos de una empresa o del sector comercial, por lo cual, en este proyecto de tesis se ha tomado en cuenta solo el tipo residencial, además, el año 2020 y aún el 2021 han tenido un evento atípico como es la pandemia, el decremento de la economía y como tal, la falta de pago de la luz eléctrica, e incluso la ayuda gubernamental respecto a la no generación de intereses o mora durante el estado de excepción, es decir, hasta agosto del 2020, donde el gobierno ecuatoriano indicó no cobrar mora, por lo que, se han tomado datos previos a la pandemia del año 2019 y el año 2020, para poder realizar las respectivas comparaciones.

De igual manera, en la etapa de exploración de los datos se descartaron los registros que siempre pagaron sin atrasos dentro de los 24 meses observados (2019 – 2020), ya que, se está buscando los patrones en las cuentas que han caído en mora, en consecuencia se ha utilizado herramientas como Microsoft Power BI, posterior a ello se ha filtrado dicha data y se ha generado una versión más limpia del Staging, este es un proceso de mejora continua donde entran los conceptos de exploración, entendimiento, procesamiento (ETL) y limpieza de data, de

forma repetitiva e incremental, como se lo indicó al inicio de este capítulo, para de esta forma, ir obteniendo una versión más limpia de la tabla analítica o tablón estadístico, preparándola para la posterior etapa de modelamiento.

También se indicó en la sección anterior que para reducir los outliers (datos atípicos) en la data, se van descartando en primera instancia las cuentas ya cortadas (11.428) o liquidadas (8.019), que son de fácil identificación y que, ya son propias de una cartera vencida no recuperable, quedándose solo con las cuentas activas donde la cartera si se puede recuperar.

Adicionalmente se observa que, de un total de 95.323 cuentas actuales ya filtradas en este universo inicial de este proyecto de tesis, al menos 5.408 cuentas, no tienen el histórico completo, es decir, los 24 meses de pago, por ende, se las descarta, y son un 5.7%, de este nuevo universo.

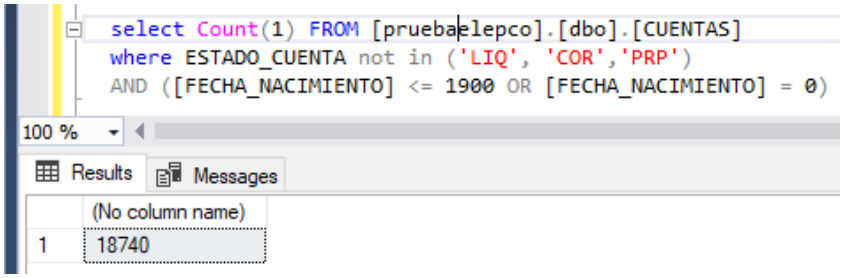
De igual manera, se realiza un conteo de cuentas que nunca hayan caído en morosidad durante el 2019 y 2020 y solo se tiene un 5% que no ha caído en atrasos en el pago, por lo cual, se justifica de una forma más notoria incluso, el tema de este proyecto de tesis de dar visibilidad sobre la cartera vencida en la cual, el 95% de este universo, alguna vez ha caído en morosidad, el otro 5% se lo descarta, como se indicó anteriormente, quedando 85.292 cuentas.

Tras lo indicado, dentro de la empresa eléctrica de Cotopaxi una de las necesidades a solventar es conocer que tan probable es que: ¿si una cuenta que ya cayó en morosidad vuelva a caer en mora el mes siguiente?, ¿cuáles son las características que indican esta tendencia?, con esta clase de aspectos y conocimiento anticipado, la empresa eléctrica en mención tendrá la opción de implementar un enfoque de seguimiento y cobro más exhaustivo a los clientes que, tienen un mayor score en el modelo a definirse posteriormente en este proyecto de tesis.

Además, un 16.1% de las cuentas no posee la fecha de nacimiento, por lo que, como se indicó en la sección de ***Tratamiento de los Datos*** para evitar outliers, en primera instancia, se aplicaron filtros como, por ejemplo, de los registros que no tienen edad, sin embargo, la limpieza de esta clase de características, se lo verá más a fondo en las siguientes secciones, con aspectos como, por ejemplo, la media para insertar data, donde esta es nula.

Figura 8

Filtro de los registros que no tienen edad



Si bien, se nota que, como un proceso interno en la empresa eléctrica de Cotopaxi mueven la fecha de vencimiento hacia una fecha muy holgada en el futuro y crea valores atípicos, esto se da cuando el cliente ha buscado un acuerdo de pago, sin embargo, en este proyecto de tesis, se lo resume en la relación de si alguien que ha caído en mora y ha llegado a este acuerdo de pago, finalmente vuelve a caer en morosidad, unos de los aspectos principales que se intenta predecir.

Preparación y exploración de los datos

Se decidió trabajar con la data en Microsoft Azure Machine Learning, debido a que, como se explicó en la sección anterior de ***Obtención de datos***, si bien, ya existe una analítica descriptiva mediante los Dashboard de QlikSense de la empresa eléctrica de Cotopaxi, la versión adquirida por dicha institución 2.1.1, no soporta la analítica predictiva y, la única opción es volver a comprar una nueva licencia y soporte por parte de la empresa eléctrica en mención.

Figura 9

Versión de QlikSense de la empresa eléctrica de Cotopaxi

Tras la anterior, cabe resaltar que Microsoft está posicionado en el cuadrante mágico de Gartner para temas de ciencia de datos y aprendizaje de máquina, por encima de otras herramientas como Knime y RapidMiner.

Figura 10

Microsoft Azure Machine Learning en el cuadro mágico de Gartner

Nota: Imagen tomada de (Information Matters, 2021)

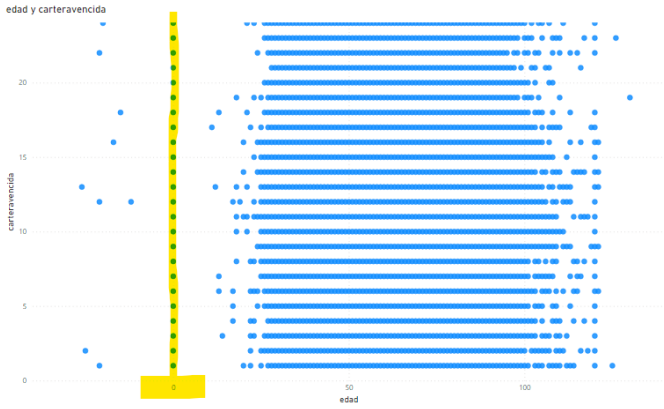
Por otra parte, para no caer en tautologías donde, por ejemplo, la misma variable de días en mora, puede dar ya la predicción buscada por una estrecha relación con el target o variable a predecir, ya que, puede implicar que a mayor número de días de mora es menos probable que el cliente pague, se realiza de manera paulatina y regresiva el Exploratory Data Analysis (EDA) como se lo indicó en la sección anterior, incluso se puede utilizar para la exploración de data, herramientas como, Microsoft Power BI y nuevamente se regresa a los procedimientos de entendimiento y selección de los datos.

Respecto a lo anterior, se ha seleccionado para la exploración inicial de datos, a Microsoft Power BI, ya que, al ser de la misma línea Microsoft, posee conexiones nativas hacia Microsoft Azure Machine Learning (Kemp, Gilley, & Maggie, 2021) y Power BI posee una interfaz gráfica más amistosa para la exploración inicial y demostración gráfica tanto para el usuario final como para quien explora la data.

Adicionalmente, dentro de Power BI, se tiene diversos tipos de gráficos para explorar la Data y entre estos, el gráfico de dispersión, de donde se puede observar que adicional a que se tienen fechas de nacimiento vacías o nulas, se tienen edades en cero (la edad fue calculada a partir de la fecha de nacimiento).

Figura 11

Exploración de la data de edad

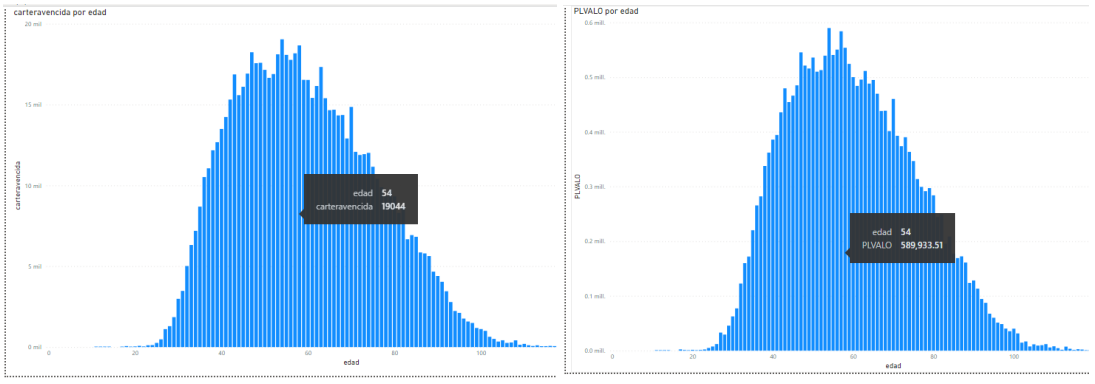


También, se tienen edades negativas, esto se da, debido a que, el ingreso de las fechas de nacimiento no ha sido validado desde los aplicativos y se tienen fechas incorrectas en la base de datos, de ahí que, se vuelve reiterativo el regresar a los procedimientos de tratamiento y la exploración de la data (Herrera, 2019).

Continuando con la exploración, se puede apreciar que existe una cantidad importante de cuentas sin edad, sin embargo, excluyendo a los que no tienen edad, se observa que, entre los 45 a los 65 años son los rangos de edad que, más se atrasan en los pagos.

Figura 12

Exploración de la morosidad y deuda por edades

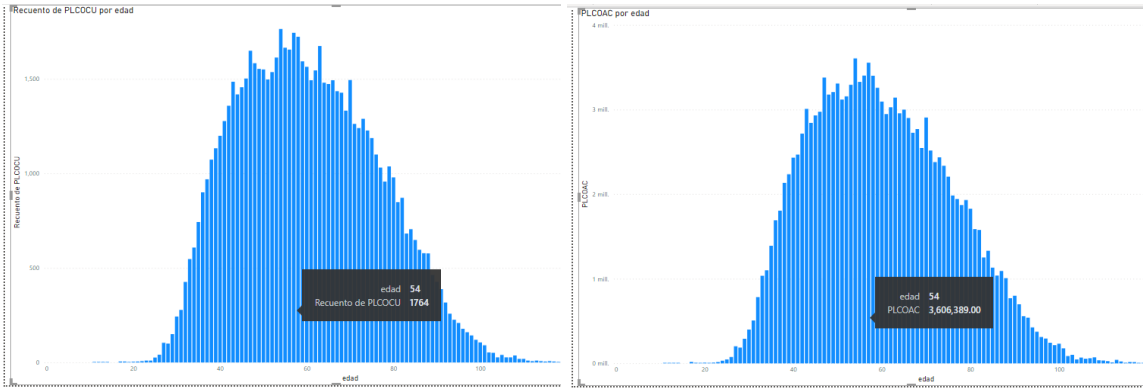


De igual manera, coincide que los clientes de 54 años son los que más valor de consumo tienen, como se indica en la imagen anterior.

Tras lo indicado, se puede concluir que la mayor cantidad de cuentas pertenece a clientes de edad de 54 años, aseveración empírica, que así lo confirma la misma data.

Figura 13

Exploración de la cantidad de cuentas y consumo de energía eléctrica activa por edades



Obviamente lo anterior, coincide con quienes más consumen energía eléctrica. Como se puede apreciar, las gráficas anteriores tienen una curva con una distribución normal casi perfecta, obviamente, en esta primera muestra se están excluyendo a las personas que no tienen registrada su fecha de nacimiento, lo cual, coincide con lo indicado en la consultoría realizada por la empresa Le Infinite a ELEPCO, sobre el no tener una data limpia.

En conclusión, es importante segmentar la información y guiarse en el dolor o búsqueda del objetivo, que, en este proyecto de tesis, son las cuentas que se han quedado en mora.

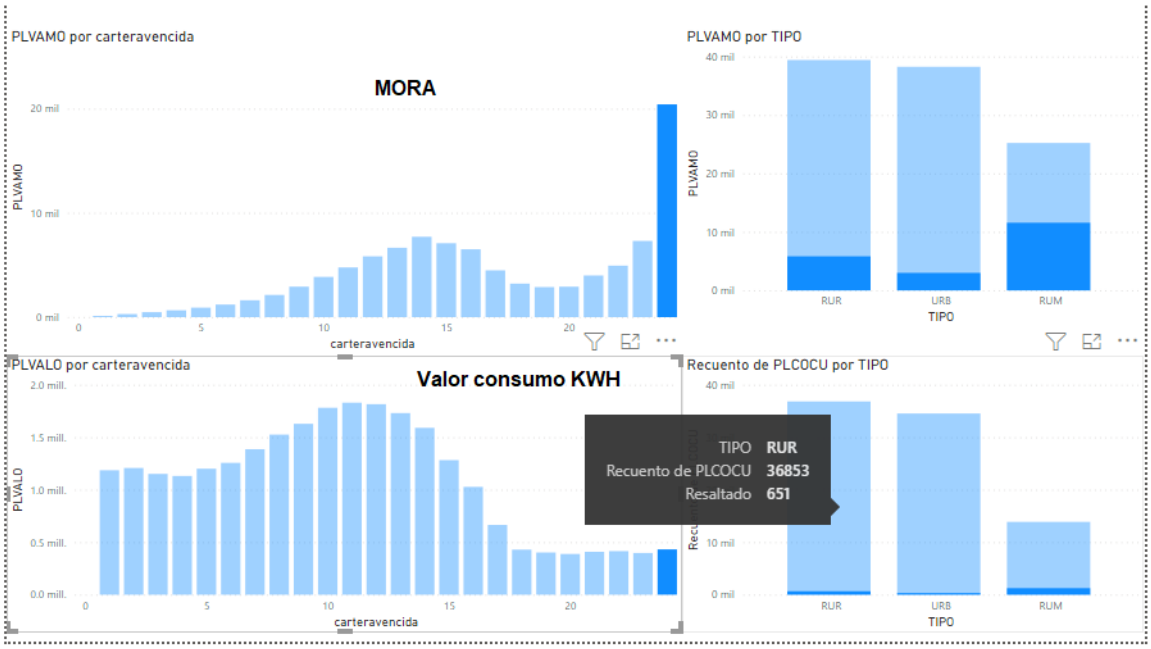
Tras lo descrito anteriormente, se debe resaltar que, el tener una data con distribución normal, ayuda en materia de machine learning, ya que, muchos algoritmos se desempeñan mejor cuando los valores de las características se aproximan a dicha distribución (Joaquín, Ajuste de una distribución, 2021), y como se lo verá, posteriormente en el siguiente capítulo de

diseño de modelos.

Adicional a lo identificado, se puede verificar también que la mayor cantidad de cuentas se las tiene registradas en el sector rural, sin embargo, en el sector marginal es donde se acumula la mayor cantidad de pagos de planillas tardías, con cuentas que en los 24 meses siempre se han atrasado, pero si se toma en cuenta el valor de los cobros por morosidad, significa que, si bien dichos valores se acumulan más en el sector marginal, por lo contrario, a nivel de la recaudación en general por consumo de KWH, este valor es pequeño en relación a la zona urbana, esto es completamente entendible, ya que, la misma demanda de una zona urbana es muy alta en relación a zonas marginales, aseveración empírica que lo confirma la data.

Figura 14

Comparación valor de la mora versus el valor de consumo

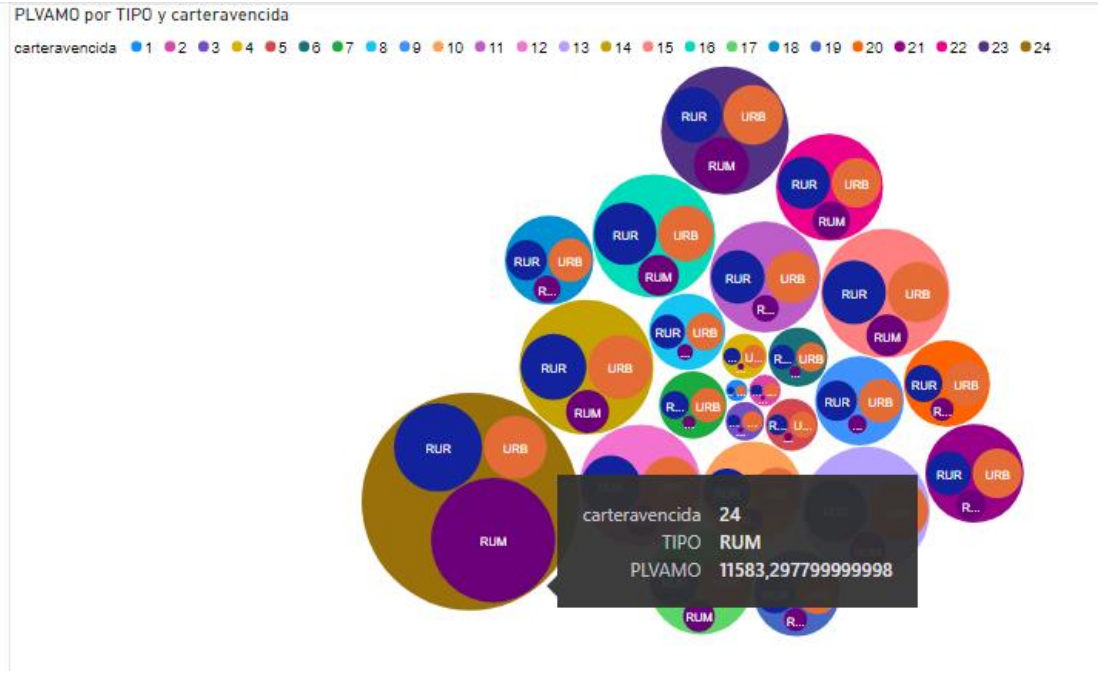


De acuerdo al análisis anterior, en cuentas con 24 meses de atraso, el valor por mora,

da un total aproximado de solo $24.500,00, ahí entran temas como, los acuerdos de pago y el diferimiento de la deuda sin generar nueva mora, en la siguiente gráfica, se muestra un total aproximado de $11.500,00 en el sector marginal, sin embargo, dicho dato no es el valor total, ya que, la empresa eléctrica en los procesos o en la programación de su sistema en AS/400 genera el valor final de mora cuando se realiza el proceso de cobro por cada cuenta.

Figura 15

Valores por mora en las distintas zonas



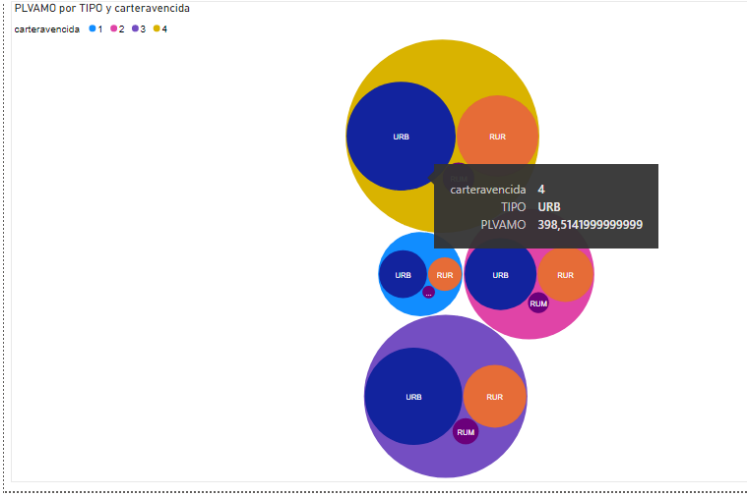
Tras lo anterior, se puede indicar que, más bien el factor predominante, no es el valor por mora, sino la recaudación por consumo de KWH, lo que, más afectaría en las planillas con atraso en su pago, ya que, este valor es mucho más grande e importante que el mora, de donde, se puede apreciar que los atrasos esporádicos, es decir, entre 1 a 4 meses dentro del 2019 y 2020, se da más en las zonas urbanas y estos, finalmente, son los que a nivel de recaudación, pesan más en la empresa eléctrica de Cotopaxi, ya que, por ejemplo, cuentas

atrasadas en un solo mes pueden generar entre 1 a 2 millones de dólares USD no cobrados a tiempo, por ende, se concluye que, un atraso, no solo significa mora sino que significa el valor del servicio principal de la empresa eléctrica no cobrado.

A partir de esta exploración inicial de la data, se concluye que, los atrasos esporádicos son más comunes en la zona urbana, como lo indica la siguiente imagen, y que, el consumo de energía activa y, por ende, el valor por consumo es siempre mayor en la zona urbana.

Figura 16

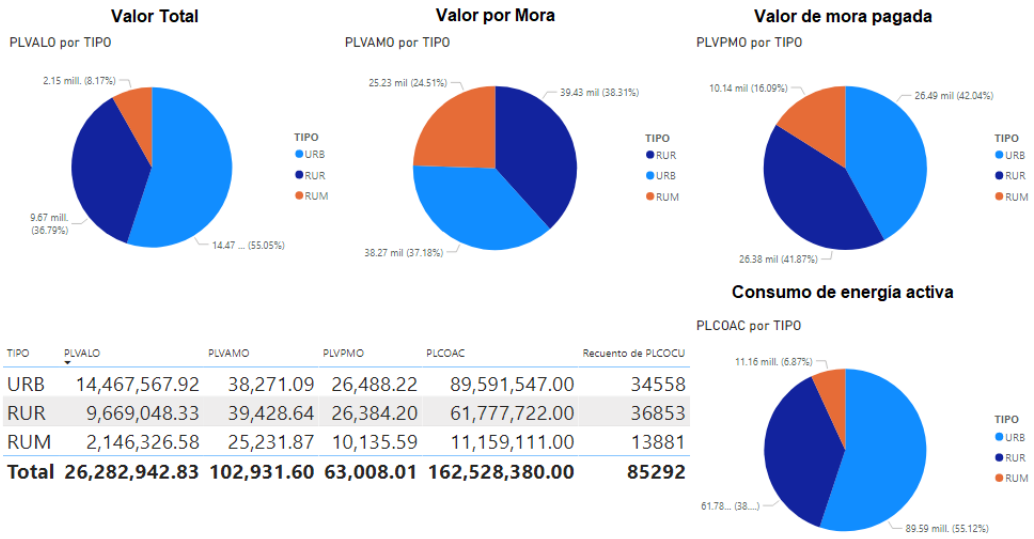
Planillas con atraso esporádico por zonas URB, RUR, RUM



Adicionalmente, como una aproximación a la realidad de la empresa eléctrica de Cotopaxi se puede indicar que del universo objeto de estudio en este proyecto de tesis, si bien, se tiene más cuentas en el sector rural, lo cual, coincide con la mayor cantidad de población en Cotopaxi, el valor de mora del sector marginal aun teniendo 24.24% menos cuentas que el sector urbano y un 26.9% porcentaje menos cuentas que el sector rural es representativo en un 24.51% frente a un 37,18% del sector urbano y un 38,31% del sector rural, sin embargo, el consumo de energía es más alto en el sector urbano, como se indica en la siguiente imagen.

Figura 17

Resumen de comparaciones por sectores urbano, rural y marginal



TIPO	PLVALO	PLVAMO	PLVPMO	PLCOAC	Recuento de PLCOCU
URB	14,467,567.92	38,271.09	26,488.22	89,591,547.00	34558
RUR	9,669,048.33	39,428.64	26,384.20	61,777,722.00	36853
RUM	2,146,326.58	25,231.87	10,135.59	11,159,111.00	13881
Total	**26,282,942.83**	**102,931.60**	**63,008.01**	**162,528,380.00**	**85292**

De acuerdo a la data analizada el valor por consumo eléctrico de las cuentas del tipo residencial en el sector urbano (55,05%) es mucho mayor que en las áreas rurales (36,79%) y marginales (8,17%), sin embargo, el valor por mora es más alto sumando los valores del sector rural (39,31%) y marginal (24,51%) en comparación al sector urbano (38,27%) y adicionalmente se puede ver como concuerda con el consumo de energía activa donde el sector urbano es el que más consume (55,12%) y por ende, su valor de pago de mora es mayor (42,04%).

Todo lo anteriormente analizado, coincide también debido a que, la facturación del consumo de energía eléctrica no tiene un precio fijo, sino que tiene un valor escalonado donde a menos consumo de KWH este se aproxima a 0,09 centavos USD y a más consumo de KWH este se aproxima a 0,10 centavos de USD (ARCONEL, 2020), es decir, el que más consume más paga, lo cual, concuerda con el mayor consumo de energía activa en las zonas urbanas.

Como se muestra en la gráfica anterior, en cuánto a valores de recaudación por consumo residencial, se tiene un 55,05% a nivel de sector urbano, 36,79% rural y solo un 8,17%

en el sector marginal, a pesar de que, la mora si es representativa en el sector marginal, el consumo de KWH de los pagos tardíos de planillas no lo es.

Tras todo lo estudiado anteriormente, este proyecto de tesis se centra o limita el alcance de su estudio, en la predicción de las cuentas con atraso esporádico, que son más difíciles de detectarlas, que cuentas donde su atraso es continuo o común, y qué, ya con la exploración realizada se detectan las zonas recurrentes en morosidad, por ende, se concluye que, en cuanto al valor económico para la empresa eléctrica de Cotopaxi, los atrasos esporádicos de la zona urbana, representan más perdida respecto a sus ingresos.

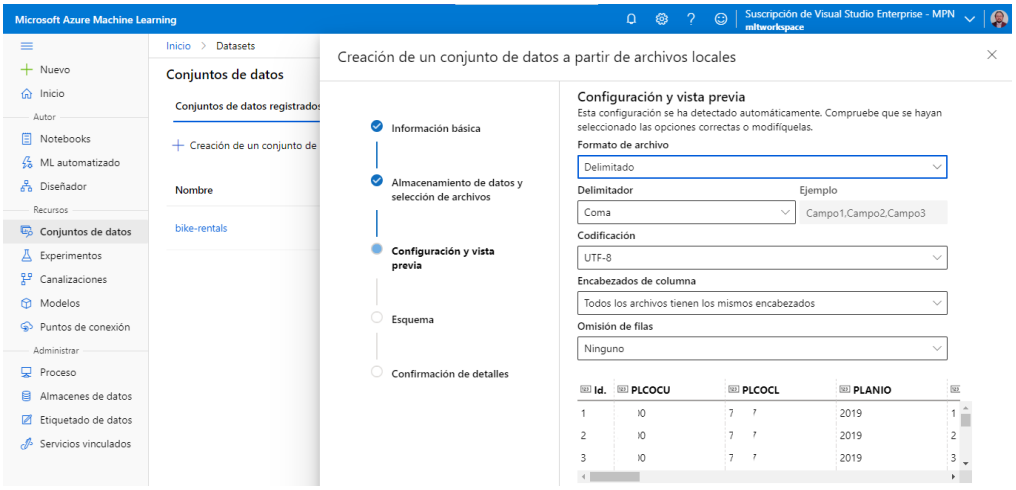
Una vez que, se ha comprendido este universo de datos para el presente proyecto de tesis y se han desarrollado procesos ETL para obtener estas 85.292 cuentas con sus 24 meses de histórico, como resultado se consigue una primera versión de la tabla analítica para explotarla con los modelos de aprendizaje de máquina, por lo que, se exporta la data mantenida en QlikSense de la empresa eléctrica en mención, hacia archivos con formatos CSV de cada una las entidades obtenidas al inicio de este capítulo.

Continuando con el proceso ETL, dicho contenido en formato CSV se cargó en un Azure Blob Storage para ser explotado o utilizado desde Azure Machine Learning. También existen opciones de Azure SQL Database, es decir, una base de datos en la nube, sin embargo, dentro de este análisis no se estará realizando cambios transaccionales a esa data, por lo que, no es necesario.

Adicionalmente, se puede mencionar que cuándo son grandes cantidades de datos (Big Data) y diferentes tipos de datos se podría utilizar Azure Data Lakes en la nube. Y, además, existen pipelines automatizados de carga, explotación y análisis de datos para MLOps (Edwards, Franks, & Coulter, 2021), lo cual, podría ser un proyecto futuro para la empresa eléctrica de Cotopaxi.

Figura 18

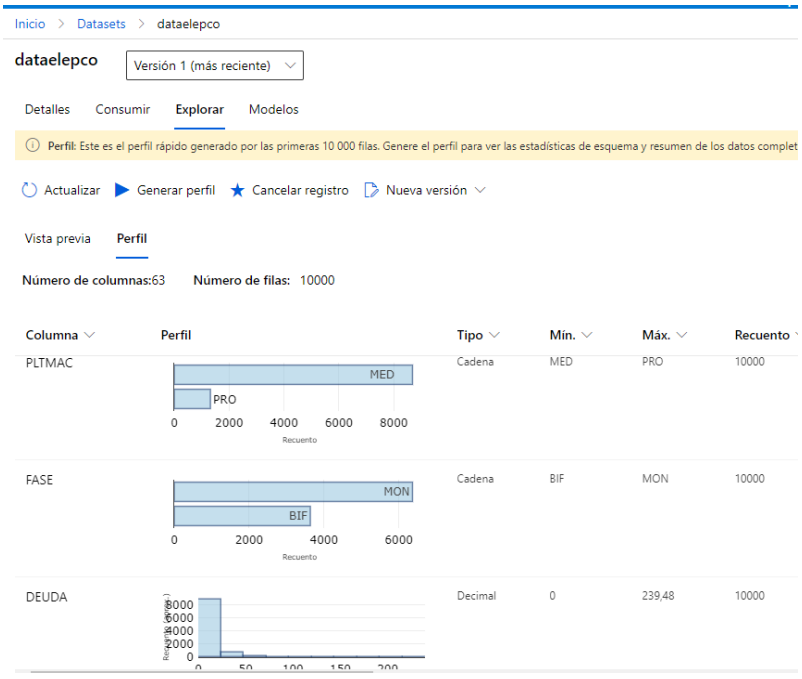
Carga de CSV en Azure Blob Storage



Cabe indicar que, Azure ML permite la exploración de la data por medio de las características de cada concepto, se tienen opciones de exploración de la data para visualizar el comportamiento de cada columna, por ejemplo, mediante histogramas, es decir, su distribución, cantidad de valores únicos, cantidad de valores nulos, entre otros.

Figura 19

Explorar data en Azure ML



También, Azure ML en cuanto a la limpieza de datos permite ir experimentando con varias opciones como la media, mediana, o, por ejemplo, mediante el análisis de componentes principales (Zhang, Gronlund, & Jason, Principal Component Analysis, 2019) para ir validando los mejores resultados según las correlaciones que se tengan hacia otras características o conceptos.

Agregación o Agrupación de características

Partiendo de esta primera versión recopilada de la data de la empresa eléctrica de Cotopaxi, a las características que varían mes a mes, se las agrupa en una sola fila de cada cuenta de los clientes, mediante un procedimiento PIVOT, por medio de ETLs de Microsoft SQL Integration Services (IDE versión community que lo puede usar la empresa eléctrica de Cotopaxi de forma gratuita), ya que, la forma en la que se pretende pronosticar o predecir posibles

clientes con morosidad es mediante la búsqueda de un patrón o varios patrones de comportamiento de acuerdo a las diferentes variables a analizar.

Para esto, se ha tomado la data mes a mes, que se tiene tanto del año 2019 como del 2020, y se la agrupa o consolida, de tal forma que, sea más fácil para el modelo identificar el comportamiento de los clientes, por lo que, al tener datos de dos años se crea una nueva columna unión de año y mes, para no perder la trazabilidad de la información histórica, también, el mismo concepto se lo agrupó, pero en formato date o fecha, para que sea de más fácil comprensión en los modelos de clasificación.

Como se indica en la siguiente imagen, una misma cuenta tiene 24 filas o registros, resumiendo lo descrito anteriormente, la idea es poder agrupar esos 24 registros mes a mes, en una sola fila mediante el proceso pivote que se describe posteriormente, sin embargo, para los análisis posteriores, resultó más comprensible dividirlos en 12 registros del 2019 y posteriormente los 12 registros del 2020.

Figura 20

Versión inicial del tablón estadístico previo las agrupaciones mediante pivote

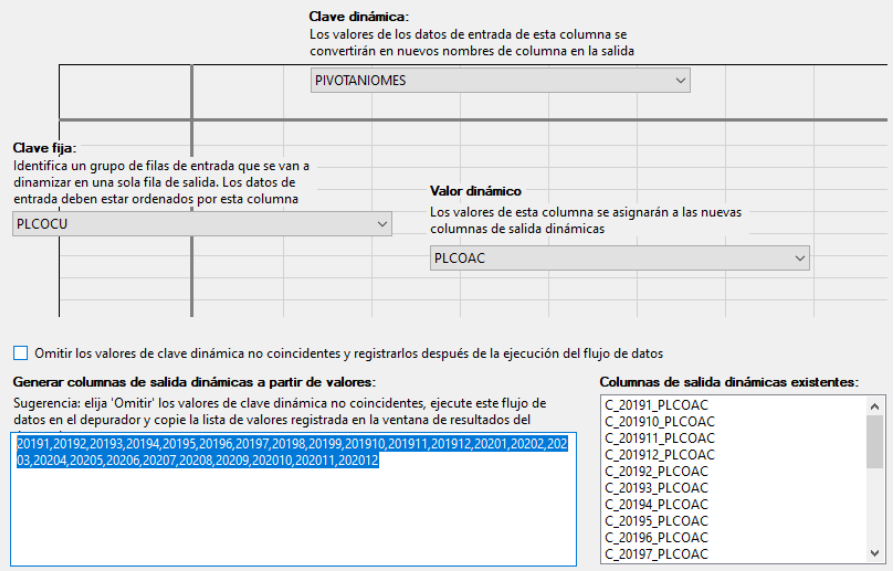
	PLCOCU	PLCOCL	PLANIO	PLMES	PIVOTANIOMES	ANIOMES	PLFEEM	PLVALO	PLVEPL	PLVAMO	PLVPMO	PLBIT0	PLFEPA	PLESTA	diferencia	morosidad
1	2	1	4	4	2019	1	20191	2019-01-01 00:00:00.000	20190131	7.73	20190220	0.03	0.03	5.32	20190311	CAN	19	1
2	2	1	4	4	2019	2	20192	2019-01-02 00:00:00.000	20190228	4.11	20190320	0	0	2.1	20190311	CAN	-9	0
3	2	1	4	4	2019	3	20193	2019-01-03 00:00:00.000	20190329	5.92	20190418	0	0	3.95	20190415	CAN	-3	0
4	2	1	4	4	2019	4	20194	2019-01-04 00:00:00.000	20190430	22.21	20190520	0	0	18.74	20190520	CAN	0	0
5	2	1	4	4	2019	5	20195	2019-01-05 00:00:00.000	20190531	7.82	20190620	0	0	11.95	20190617	CAN	-3	0
6	2	1	4	4	2019	6	20196	2019-01-06 00:00:00.000	20190630	9.14	20190720	0.01	0.01	12.79	20190722	CAN	2	1
7	2	1	4	4	2019	7	20197	2019-01-07 00:00:00.000	20190731	9	20190820	0.01	0.01	12.55	20190826	CAN	6	1
8	2	1	4	4	2019	8	20198	2019-01-08 00:00:00.000	20190830	8.34	20190919	0.01	0.01	11.35	20190926	CAN	7	1
9	2	1	4	4	2019	9	20199	2019-01-09 00:00:00.000	20190930	8.88	20191020	0.02	0.02	12.32	20191031	CAN	11	1
10	2	1	4	4	2019	10	201910	2019-01-10 00:00:00.000	20191031	9.48	20191120	0.01	0.01	13.38	20191125	CAN	5	1
11	2	1	4	4	2019	11	201911	2019-01-11 00:00:00.000	20191129	26.6	20191219	0.11	0.11	22.85	20200106	CAN	18	1
12	2	1	4	4	2019	12	201912	2019-01-12 00:00:00.000	20191226	9.09	20200115	0.03	0.03	12.68	20200127	CAN	12	1
13	2	1	4	4	2020	1	20201	2020-01-01 00:00:00.000	20200131	19.36	20200220	0.03	0.03	15.7	20200226	CAN	6	1
14	2	1	4	4	2020	2	20202	2020-01-02 00:00:00.000	20200228	16.66	20200430	0.19	0.19	13.04	20200507	CAN	7	1
15	2	1	4	4	2020	3	20203	2020-01-03 00:00:00.000	20200331	20.22	20200620	0	0	16.93	20200507	CAN	-44	0
16	2	1	4	4	2020	4	20204	2020-01-04 00:00:00.000	20200430	18.86	20210720	0	0	15.7	20200611	CYR	-404	0
17	2	1	4	4	2020	5	20205	2020-01-05 00:00:00.000	20200531	19.39	20210720	0	0	16.19	20200630	CYR	-385	0
18	2	1	4	4	2020	6	20206	2020-01-06 00:00:00.000	20200630	18.15	20210720	0	0	14.85	20200727	CAC	-358	0
19	2	1	4	4	2020	7	20207	2020-01-07 00:00:00.000	20200731	18.86	20210720	0	0	15.7	20200824	CAN	-330	0
20	2	1	4	4	2020	8	20208	2020-01-08 00:00:00.000	20200831	15.98	20210820	0	0	13.04	20201030	CAC	-294	0
21	2	1	4	4	2020	9	20209	2020-01-09 00:00:00.000	20201017	9.5	20210721	0	0	12.34	20201018	CYR	-276	0
22	2	1	4	4	2020	10	202010	2020-01-10 00:00:00.000	20201031	18.08	20201120	0.02	0	14.98	20201221	CAC	31	1
23	2	1	4	4	2020	11	202011	2020-01-11 00:00:00.000	20201130	20.72	20201220	0.01	0	17.42	20201221	CAN	1	1
24	2	1	4	4	2020	12	202012	2020-01-12 00:00:00.000	20201229	15.71	20210118	0.01	0	12.79	20210104	ABO	-14	3

Adicionalmente, se toma en cuenta el análisis de data realizado en el *Diagnóstico de la Calidad de Datos de la Información de Misión Crítica.*por la empresa Le Infinite, donde esta consultoría le indica a la empresa eléctrica de Cotopaxi que debe hacer limpieza de datos y mejorar la data como tal, ya que, sufre de falta de información, alrededor de un 25% de datos a nivel general es nulo, por lo que, se plantea realizar distintos tipos de algoritmos para los conjuntos o grupos de clientes donde sí se tiene en base de datos su información personal, por ejemplo, en los registros que si se tiene fecha de nacimiento y condición del ciudadano, que en primera instancia se presume serán de mejor adaptación a los algoritmos de clasificación.

Mientras que, de forma general se aplicará algoritmos de regresión, por ejemplo, por el hecho de no tener los ingresos de la persona, se puede enfocar en el algoritmo de regresión logística, tomando en cuenta las características que se correlacionan, relacionando así los datos históricos de pago, de consumo de KWH, así como, de consumo promedio, que se detalla más adelante. Posteriormente, se van agrupando todas las características que cambian mes a mes, por ejemplo, el consumo de energía activa, para realizar el pivote en el ETL, como se muestra en la siguiente imagen.

Figura 21

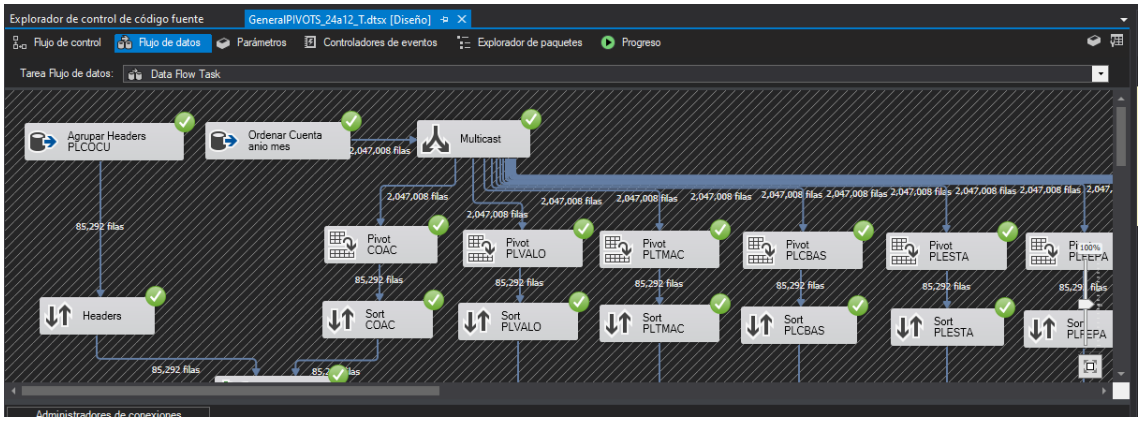
Agrupaciones mediante pivote



Luego de analizar todas las características que cambian mes a mes y que tienen valores no categóricos, en el ETL se va agregando los pivotes necesarios y se va unificando la información a la cabecera. En dicha cabecera se tienen los valores de la cuenta y cliente que son estáticos en el tiempo, como se lo indica en la siguiente figura.

Figura 22

Ejecución del proceso pivote para el agrupamiento de los 12 meses de 2019 con 85.292 filas



En adición al proceso anterior, a diferencia de algoritmos tradicionales como las redes

neuronales que, si aceptan variables categóricas, en los algoritmos de predicción o regresión, se tiene el concepto de variables Dummy o Indicator Values (Li & Lu, Convert to Indicator Values, 2021), ya que, los algoritmos tienen un mejor desempeño con data numérica, esta técnica permite crear de forma dinámica y automatizada columnas con cada una de las categorías y colocar el valor de 1 si existe y 0 si pertenece a otra categoría cada registro evaluado.

Figura 23

Ejemplo de variables categóricas hacia Variables Dummy o Indicator Values

Cliente	Estado_Civil
Persona 1	CASADO
Persona 2	EN UNION DE HECHO
Persona 3	SOLTERO
Persona 4	CASADO
Persona 5	VIUDO
Persona 6	DIVORCIADO

Cliente	Estado_Civil_CASADO	Estado_Civil_EN UNION DE HECHO	Estado_Civil_ SOLTERO	Estado_Civil_ VIUDO	Estado_Civil_ DIVORCIADO
Persona 1	1	0	0	0	0
Persona 2	0	1	0	0	0
Persona 3	0	0	1	0	0
Persona 4	1	0	0	0	0
Persona 5	0	0	0	1	0
Persona 6	0	0	0	0	1

De acuerdo a este concepto de variables dummy anteriormente indicado, se va identificando información como la clase de cliente con los conceptos: natural, privado o sector público, que, posteriormente, se los convertirá mediante Azure ML a Indicator Values, así como, el concepto PLTMAC descrito a inicios de este capítulo sobre si se hizo en un mes las mediciones de KWH o si son valores promediados.

También se adiciona, el concepto de condición del ciudadano que puede arrojar ciertas pistas sobre el comportamiento de pago, por ejemplo, se tiene datos como: analfabeto, cédula inválida por contravención, cédula inválida por expiración, discapacidad, discapacidad física

mayor de edad, discapacidad mental mayor de edad, extranjero, fallecido, interdicto, menor de edad, militar servicio activo, policía en servicio activo, residente en el exterior, estas características de tipo categóricas se las describe a continuación.

Tabla 11

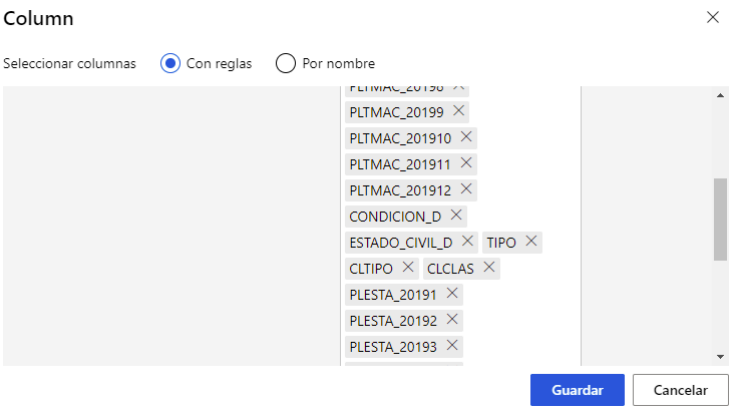
Tabla de características categóricas

Entidades	Conceptos	Tipos de datos
PLTMAC	Medición de la planilla Mes a Mes	Consumo medido, consumo promediado
PLESTA	Estado de la planilla	CAN canceladas, CNJ canjeadas, CAC canjeada y con afectación contable, CYR canjeada y re facturada, VEN vencida, GEN generada, ABO abonada
CONDICION_D	Condición del ciudadano	Analfabeto, cédula inválida por contravención, cédula inválida por expiración, discapacidad, discapacidad física mayor de edad, discapacidad mental mayor de edad, extranjero, fallecido, interdicto, menor de edad, militar servicio activo, policía en servicio activo, residente en el exterior
ESTADO_CIVIL_D	Estado Civil	CASADO, EN UNION DE HECHO, SOLTERO, NULO, VIUDO, DIVORCIADO
TIPO	Tipo de cuenta	URB urbano, RUM rural o urbano marginal, RUR rural
CLTIPO	Tipo de cliente	INS institución, COM comercial, NULL, PUB público, PAR particular, PRI privado, EMP empleado
CLCLAS	Clase de cliente	INS Institución, NULL, PAR particular, NAT persona natural

Nota. Esta tabla muestra las características categóricas

Figura 24

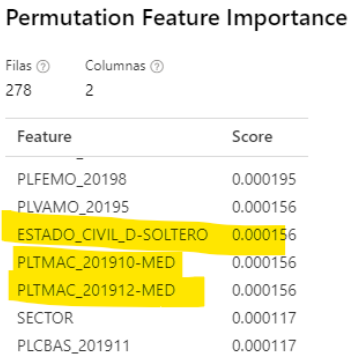
Selección de características categóricas en Azure ML



Por otro lado, mediante los experimentos en Azure ML se puede identificar la importancia de las características o features, lo cual, se lo describirá posteriormente, sin embargo, cabe recalcar que dependiendo de los tipos de algoritmos las variables categóricas tendrán más importancia al ser etiquetadas y tratadas como Indicators Values como se indica en la siguiente figura con la permutación de características (Zhao, 2016), algoritmo que permite calcular la importancia de conceptos determinando la sensibilidad del modelo por medio de permutaciones (cambios entre características) de forma aleatoria.

Figura 25

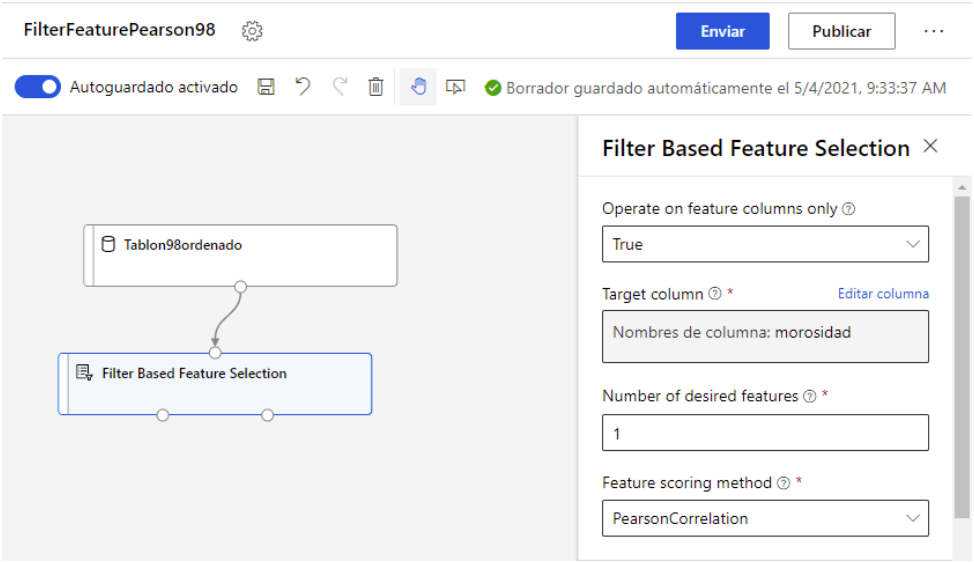
Importancia de características



Adicional, se tienen aspectos importantes para el concepto de cartera vencida que van cambiando mes a mes como: la diferencia de consumo de KWH, si pagó o no con mora, las cuales que partiendo de filas se agruparon a columnas también mediante un proceso de pivote y se evalúan de forma similar a lo realizado en las etapas previas de ***Entendimiento y Selección de Datos*** mediante las correlaciones de Pearson (Li & Lu, Filter Based Feature Selection, 2020), filtrando las características más importantes, experimentando y volviendo a entrenar el modelo.

Figura 26

Selección de características mediante la correlación de Pearson



Análisis de características

Una vez agregado o agrupado todos los valores históricos de una misma cuenta, hacia una misma fila, se procedió a no tomar en cuenta las columnas que son parte de la evidencia de pago y mora de la columna destino o target a predecir, es decir, se desea en primera instancia predecir el pago a tiempo o en mora de diciembre 2019, por ende, las siguientes características que ya son propias de este pago, no se tomarán en cuenta en el entrenamiento para no crear una tautología:

Tabla 12

Tabla de características descartadas para la inferencia

CARACTERÍSTICAS	DESCRIPCIÓN
PLVPMO_201912	Valor pagado en mora
PLFEMO_201912	Fecha en la que se ingresó la mora en el sistema debido al atraso en el pago de la planilla
PLVAMO_201912	Valor calculado de la mora
PLESTA_201912	Estado de la planilla

CARACTERÍSTICAS	DESCRIPCIÓN
morosidad_201912	Planillas categorizadas según concepto de pago a tiempo o con morosidad
carteravancida_201912	Cantidad de planillas pagadas de forma tardía
diferencia_201912	Cantidad de días de atraso o días a tiempo, respecto a la fecha de vencimiento del pago de la planilla

Nota. Esta tabla muestra las características no tomadas en cuenta para la predicción

En Azure ML los algoritmos como, por ejemplo, la misma permutación de características, el uso de la metodología CRISP-DM y experimentar con los conceptos, permiten contrarrestar los conocimientos del negocio y de su data, para descartar las características que no tiene un impacto en los modelos. Adicional a lo anterior, se van excluyendo los conceptos que arrojan valores negativos.

Con lo cual, se comprueba que las fechas son muy importantes, por lo que, yendo a la teoría de categorización, se utilizan las fechas como, fecha de pago y de mora, así como fecha de nacimiento de donde se puede agrupar la información en los llamados Bins, Intervalos o grupos en este ejemplo de edad más pequeños o enfocados a un universo específico de la información (Li & Lu, Group Data into Bins module, 2020) que, a futuro la empresa eléctrica de Cotopaxi los puede emplear según sus necesidades u objetivos.

Estrategias para el modelamiento

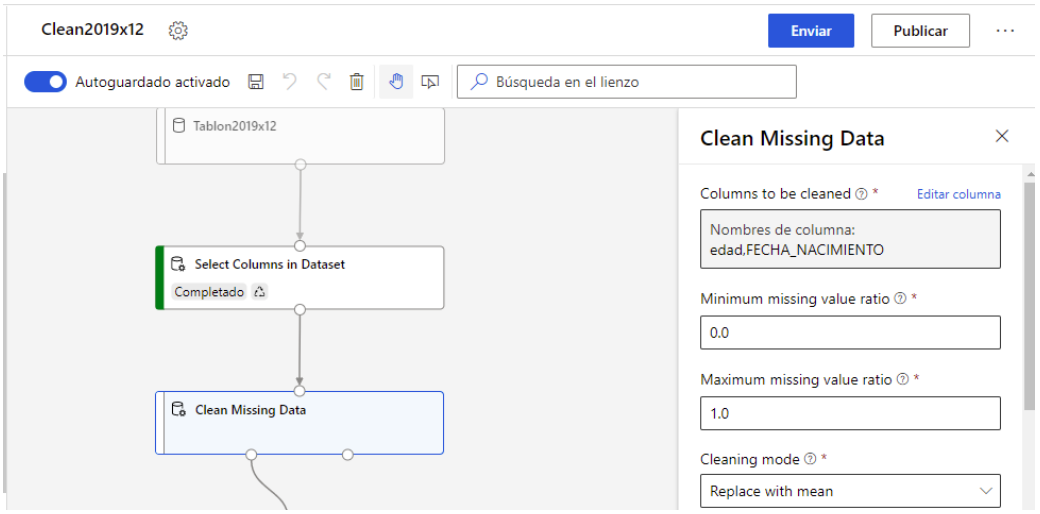
Según la data obtenida y el problema a resolver, se puede deducir que la mejor alternativa son los modelos supervisados, ya que, se conocen las variables y los resultados históricos, si bien la consultoría de la empresa Le Infinite, se dio como un paso previo a un proyecto de Business Intelligence en el 2013, esta no tuvo un enfoque hacia análisis predictivos,

sin embargo, si se ha tenido un proyecto interno de generación de DashBoards (Garzón Ulloa & Chicaiza Castillo, 2017) gerenciales con la data actual, es decir, sin Data Cleaning (Ilyas & Chu, 2019), pero se ha podido observar en la data disponible que el porcentaje de campos nulos coincide con lo estipulado (mayor al 25%), por lo que, no se tiene en absolutamente todos los clientes, los datos como, la edad, por lo que, se ha ido limpiando dicha data, que podría crear sesgos o inconsistencias en el modelo.

Tras lo anteriormente indicado, se ha obtenido la data histórica necesitada para las inferencias o análisis predictivo, planteando diferentes escenarios donde se aplican de varios algoritmos, además, se pudo experimentar con el Data Cleaning aplicando por ejemplo, la media y de esta manera, se ha podido apreciar, cómo estos cambios le dan una mayor importancia a las características que antes estaban vacías o incompletas, si bien se muestran en la siguiente imagen dos características similares, las de fecha de nacimiento y edad, posteriormente se indicará cuál de las dos características es más importante para el modelo mediante el filtrado de características.

Figura 27

Módulo de Data Cleaning de Azure ML

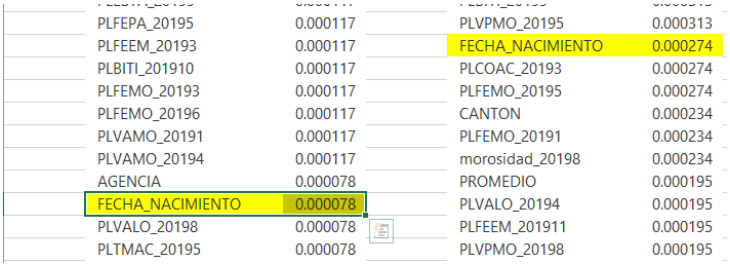


De acuerdo a la importancia de características, se usa la fecha de nacimiento y no la edad, a nivel de entendimiento es más fácil usar la edad (de forma simplificada) para una persona, por el contrario, no así para los modelos, ya que, usan la fecha en formato numérico.

Además, como se indica en la siguiente figura de resultados, la limpieza de datos optimiza estos para el modelo, es decir, mejora su importancia o impacto hacia el modelo, de ahí la necesidad de tener un correcto control desde el origen de la data, es decir, desde su alimentación con los sistemas transaccionales.

Figura 28

Optimización de características mediante Data Cleaning



PLFEPA_20195	0.000117	PLVPMO_20195	0.000313
PLFEEM_20193	0.000117	FECHA_NACIMIENTO	0.000274
PLBITI_201910	0.000117	PLCOAC_20193	0.000274
PLFEMO_20193	0.000117	PLFEMO_20195	0.000274
PLFEMO_20196	0.000117	CANTON	0.000234
PLVAMO_20191	0.000117	PLFEMO_20191	0.000234
PLVAMO_20194	0.000117	morosidad_20198	0.000234
AGENCIA	0.000078	PROMEDIO	0.000195
FECHA_NACIMIENTO	0.000078	PLVALO_20194	0.000195
PLVALO_20198	0.000078	PLFEEM_201911	0.000195
PLTMAC_20195	0.000078	PLVPMO_20198	0.000195

De acuerdo, a la data disponible y al alcance planteado se pretende predecir si un cliente dejará de ser moroso, se volverá moroso, o continuará siendo moroso, de acuerdo a las distintas variables que se han podido identificar en primera instancia del análisis de fuentes y sus características.

Tras lo anterior, cabe indicar que dentro de los modelos supervisados los algoritmos más comunes son los de clasificación, que permiten predecir las categorías, en este caso de moroso o no moroso, adicionalmente, dentro del universo de datos de la empresa eléctrica de Cotopaxi, se puede usar por ejemplo, la regresión logística (1 y 0 clasificación binaria) ya que, no se disponen datos como en una entidad financiera, relacionados a los ingresos económicos de los clientes, a este universo de data se aplicarán modelos con esta clase de algoritmos de

clasificación, de acuerdo al análisis realizado en la sección del ***Estado del Arte***, dentro de este ámbito de estudio, se podrían usar los algoritmos de:

- Máquina de Vectores de Soporte (Two-Class Support Vector Machine)
- Promedio Perceptron (Two-Class Averaged Perceptron)
- Regresión logística
- Regresión de Red Neural
- Árbol de decisión potenciado (Boosted Decision Tree)

Estos modelos se detallarán a mayor profundidad en el siguiente capítulo.

Capítulo V: Diseño de modelos

Desarrollo de Modelos

Como se había indicado al final del capítulo anterior para diseñar los distintos modelos, se tomará ventaja de la tecnología de Microsoft Azure Machine Learning (Mungi, Gronlund, & Hansen, 2021), con la robustez que posee la infraestructura Microsoft Azure en la Nube, la cual, permite tener computadores con un número de Cores y RAM muy elevado, comparado con la infraestructura que se puede adquirir en Servidores On Premises, el procesamiento es mucho más rápido y efectivo, además, Azure ML permite realizar experimentos con la data, su preparación o limpieza y posterior diseño de entrenamientos con un número grande de algoritmos e incluso integraciones a R y Python (este último lenguaje usado en un módulo posterior en este proyecto de tesis).

Adicionalmente, para la evaluación de los modelos se aprovechará su integración al framework de Machine Learning scikit learn, como se indica en (Edwards, Gilley, Gronlund, &

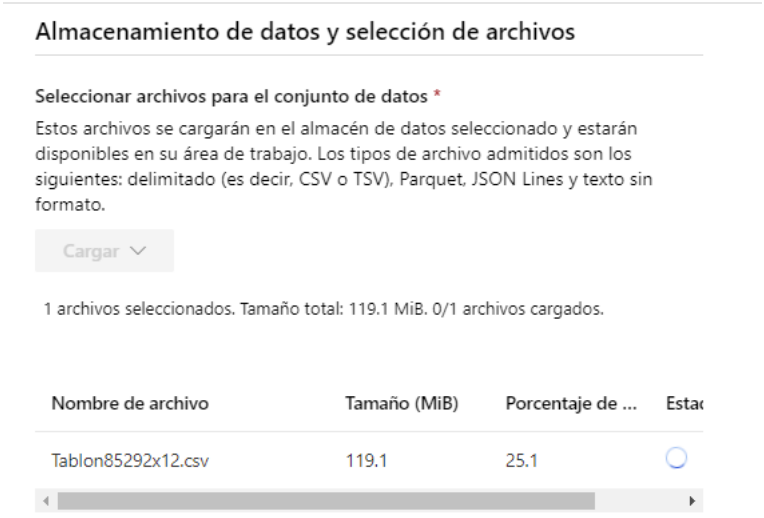
Coulter, 2021) escrito en lenguaje Python, que permite la aplicación de métricas estadísticas a los modelos a plantearse.

El costo de los recursos Microsoft Azure tanto de los recursos de procesamiento como de alojamiento de la data dependerán de su tamaño y carga, para la cantidad de datos de ELEPCO que es 119,1: MB iniciales, lo cual, siempre se irá incrementando con el aumento de la data transaccional para los siguientes años, referente a este peso en megas, en recursos dedicados podrían rondar los $0,37 por hora, mientras que en recursos que sean de pruebas, es decir, que se pueden apagar y restaurar solo bajo demanda como es en el caso para este proyecto de tesis, los recursos pueden tener una prioridad baja y podrían bajar a $0,07 la hora.

Para cubrir estos costos, se aprovecha el convenio de la empresa Microsoft Partner en la cual, actualmente estoy laborando, ya que este proyecto de tesis es un estudio y punta pie inicial del cual, cuyos valores y resultados los puede usar la empresa eléctrica de Cotopaxi como insumo para un proyecto de MLOps a futuro.

Figura 29

Carga del dataset de 85.292 cuentas por año (2019 – 12 meses).



Adicional a lo anterior, otra de las ventajas de usar Microsoft Azure Machine Learning, es que, aprovecha su infraestructura en la nube y permite publicar los modelos como servicios REST API, exponiendo así la lógica del modelo para ser consumida desde cualquier lugar geográfico con las debidas seguridades, o también se puede realizar un despliegue (deployment) para consumos de grandes volúmenes de datos en Batch, es decir, por lotes.

Las etapas para realizar el diseño de un modelo, su entrenamiento y pruebas se describen en las siguientes secciones:

Creación de un espacio de Trabajo

Un workspace (espacio de trabajo) es un lugar en la nube de Microsoft Azure donde se puede administrar la data, los recursos computacionales, los modelos e incluso el código relacionado a los experimentos de Machine Learning.

Los recursos computacionales incluyen:

- Instancias computacionales para trabajar con la data y con los modelos.
- Clústeres de cómputo donde se tienen N máquinas virtuales para procesar experimentos bajo demanda.
- Clústeres de Inferencia que son las máquinas que alojan a los servicios de predicción que usan los modelos previamente entrenados.

Creación del set de datos (dataset)

Pueden existir diversos tipos de fuentes de datos incluso compartidos entre diferentes vendors de sistemas de gestión de bases de dato, archivos planos u hojas de cálculo, e incluso se puede aplicar una suerte de ETLs mediante Microsoft Azure DataFactory (nube) y/o Microsoft Integration Services (Oo Premises), sin embargo, por las mismas restricciones

institucionales, no se podría conectar desde la nube directamente hacia la base de datos de la empresa eléctrica de Cotopaxi, por lo que, se ha obtenido un set de datos, histórico con las clases o categorías descritas en la sección ***Obtención de datos***, mediante una base de datos Staging para almacenarlos de forma temporal y así poder usarlos en este proyecto de tesis, posteriormente se trasladó la data hacia la nube mediante CSV.

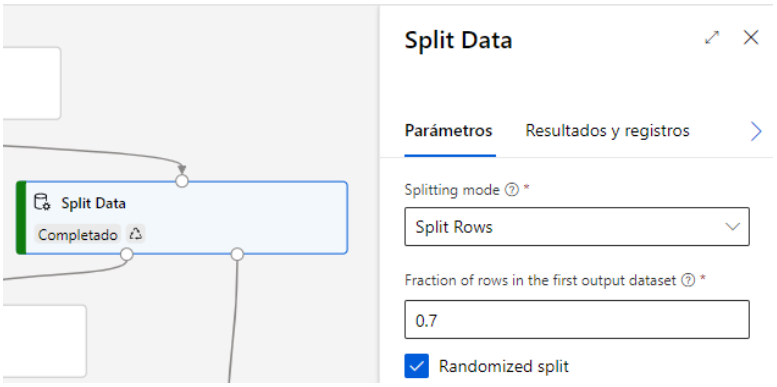
Diseñar el flujo para entrenar los modelos

De acuerdo al tipo de datos y al objetivo, se enfoca este estudio en modelos supervisados, tal como, lo indica el ***Estado del Arte*** de este proyecto de tesis los modelos que más se adaptan a la realidad de la empresa eléctrica objeto de estudio son: Boosted Decision Tree, Support Vector Machine, Ramdom Forest, redes neuronales y regresiones logísticas, por lo cual, para la siguiente sección se seleccionan estos algoritmos, así como las clases o columnas relacionadas al objetivo de predicción o inferencia, que parten del ***capítulo 4***.

Dependiendo de la cantidad de data se puede usar comúnmente la metodología Train Test Split como sostiene (Pawluszek-Filipiak & Borkowski, 2020) donde se usa un 70% de los datos en el entrenamiento y el 30% restante sirve para comparar la predicción del modelo, también se lo puede hacer en una proporción 90-10, pero dada la cantidad de data se ha elegido para este modelamiento: 70-30 como se indica en (Zhang, Gronlund, & Lu, Split Data using Split Rows, 2020).

Figura 30

Split data en Azure ML



Adicionalmente, se puede estratificar la data, es decir, verificar que tanto el primer grupo de datos usado para el entrenamiento como el segundo grupo de datos usado para las pruebas o verificaciones de predicción, tengan el mismo porcentaje de la o las columnas objetivo (target) del modelo. En este caso, solo se estratifica la característica de respuesta o la variable a predecir, esto quiere decir que a pesar de la división 70/30, se mantiene el mismo porcentaje del universo completo de morosos y no morosos.

Comparativa de modelos

Siguiendo con el proceso de modelamiento, para contestar a la pregunta de investigación: OE2 – RQ2.1: ¿Cuál(es) modelo(s) predictivo(s) se adapta(n) a la realidad de la empresa eléctrica de Cotopaxi?, se ha partido del mismo ***Marco Teórico*** donde se indicó que: el análisis supervisado hace uso de técnicas para comprender y predecir el comportamiento de un evento futuro con base en hechos que ya pasaron y sobre los cuales se cuenta con información precisa.

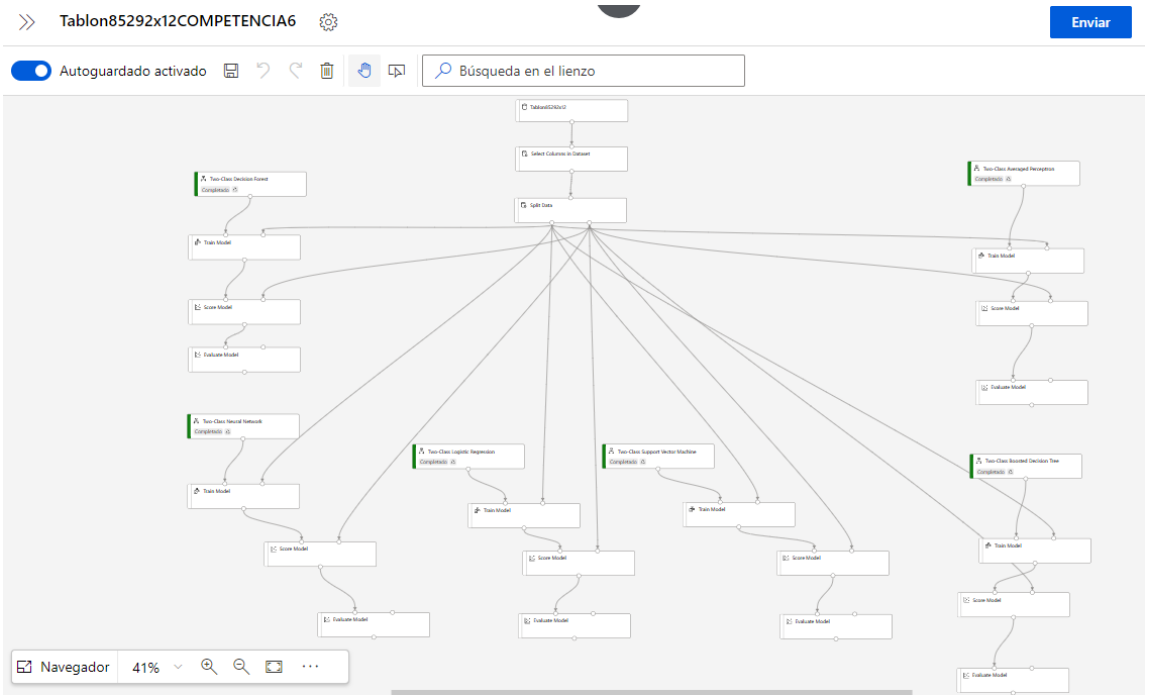
En el ***capítulo 2*** se mencionó qué: los algoritmos supervisados dependen de datos que estén previamente etiquetados, es decir, de aspectos o ejemplos prácticos que ya han pasado

pero que seguirán surgiendo en el futuro, es decir, como lo sostiene (Rodríguez & Gamboa, 2019), el algoritmo aprende a predecir el valor de salida. Unos de los algoritmos más frecuentes de aprendizaje supervisado son: Árboles de decisión, Regresión Logística, Support Vector Machines (SVM) y Métodos “Ensemble” (conjuntos de clasificadores), como lo indican (Wei, Yang, Zhang, & Zhang, 2019), a su vez, estos son los algoritmos que se han usado en este proyecto de tesis por medio de Azure Machine Learning y que se los detalla a continuación.

Los resultados de la ejecución del entrenamiento y las métricas aplicadas se describen en las siguientes secciones, según los cuales se puede elegir el mejor modelo que se usará para la realidad de la data en la empresa eléctrica.

Figura 31

Comparativa de modelos



Tras lo indicado anteriormente, se usaron los siguientes modelos sobre la data total de 85.292 cuentas partiendo del objetivo de identificar la cartera vencida, la cual, se resume en

moroso y no moroso, por ende, son dos clases o una clasificación binaria:

- Two-class Neural Network
- Two-class Logistic Regression
- Two-Class Boosted Decision Tree
- Two-Class Support Vector Machine
- Two-Class Averaged Perceptron
- Two-Class Decision Forest

En esta primera parte mediante la competencia de modelos se han enviado todas las características encontradas y analizadas en el ***capítulo 4***, hacia todos los modelos y también los datos de entrenamiento del tablón estadístico anteriormente obtenido.

Siguiendo con el proceso, se procede a realizar el resumen de cada modelo y la comparación de resultados obtenidos, que en primera instancia se lo realizó con una data de 30.000 cuentas y cuando ya se concluyó la obtención de datos con los filtros y exclusiones complementados a los ETLs, se la realizó nuevamente ya con las 85.292 cuentas finales, realizando así una comparación versus el re-entrenamiento de modelos, esto se resumirá en una tabla en la sección de ***Factores de evaluación.***

Two-class Neural Network

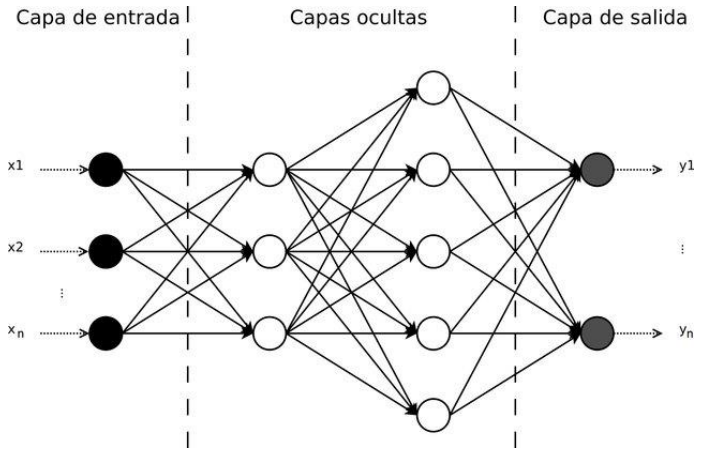
Como sostienen (Li, Lu, & Baccam, Two-Class Neural Network module, 2020), este es un modelo de aprendizaje supervisado que permite usar una red neuronal para calcular un objetivo, en este caso se la ha usado para una clasificación binaria, es decir, que solo tiene dos valores (moroso y no moroso), esta red neural se basa en capas:

- Capa de entrada (inputs)

- Una o varias capas ocultas intermedias que permiten el aprendizaje y otras capas sucesivas que permiten la obtención del sentido semántico de los valores de entrada.
- Y la capa de salida (moroso o no moroso)
- Todos los nodos de una capa se conectan mediante las aristas ponderadas hacia los nodos de la capa siguiente.
- Para calcular la salida se van dando pesos o ponderaciones en las distintas capas y se van sumando dichos valores ponderados.

Figura 32

Capas de una Red Neuronal

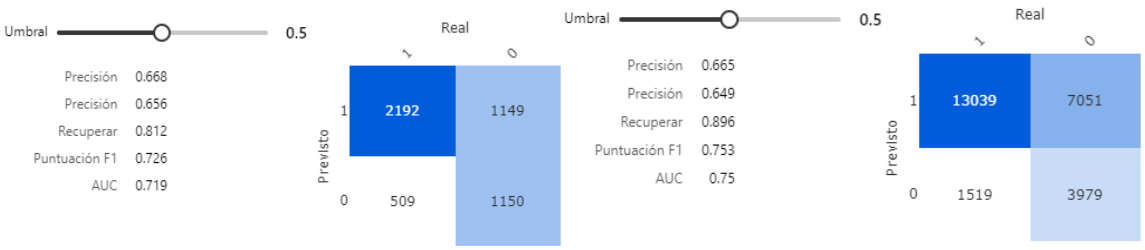


Nota: Imagen tomada de (Alvarado & F, 2017)

En el caso de Azure Machine Learning el número de nodos de la capa de entrada es igual al número de características de los datos de entrenamiento. Se indica el detalle de los resultados obtenidos, sin embargo, en la siguiente sección se realizará un resumen y comparación por cada modelo.

Figura 33

Red neuronal con 30.000 cuentas y 85.292 cuentas finales



Para este modelo se puede apreciar como los verdaderos positivos pasaron de un 44% hacia un 51% en el segundo entrenamiento (7% más), mientras que el porcentaje de falsos negativos y falsos positivos que es de un 33%, se ha incrementado en un 0,33%, por lo cual, su exactitud (accuracy) y su precisión han disminuido, indicando que este modelo no es el más adecuado para las características usadas.

Two-class Logistic Regression

Como indican (Li, Lu, Coulter, Baccam, & Gilley, Two-Class Logistic Regression module, 2020), la regresión logística es una técnica estadística muy conocida que se usa para inferir la probabilidad de una salida, se adapta a la problemática que compete este proyecto de tesis, ya que, es muy popular en clasificaciones, como en este caso, de moroso y no moroso. Este algoritmo ajusta los datos a una función logística.

La regresión logística usa regularizaciones que son una serie de límites para optimizar el modelo, es decir, para que pueda generalizar de mejor manera la data.

- Lasso (L1) se puede usar para a modelos dispersos.
- Ridge (L2) se usa cuando los datos no son dispersos.
- Regularización elástica neta que combina ambos.

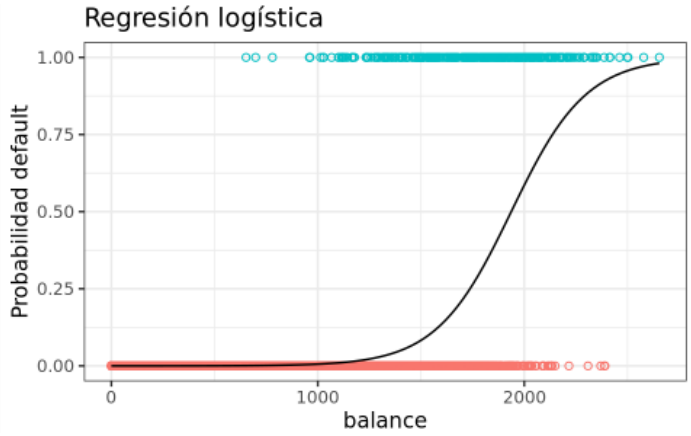
En este caso se usó L2 dado que en el capítulo anterior se realizaron algunas técnicas de data cleaning para mitigar la dispersión de la Data.

Se debe tomar en cuenta que si no se usa la regularización se caería en el overfitting es decir, sobre ajustar el modelo, este sobre ajuste hace que, el modelo se adapte perfecto a los datos de entrenamiento, pero sea deficiente para la nueva data real al inferir.

También se debe tomar en cuenta que este algoritmo a nivel de Azure Machine Learning ya trae embebido la normalización de la data, teniendo así un modelo optimizado para ser usado.

Figura 34

Regresión logística binaria

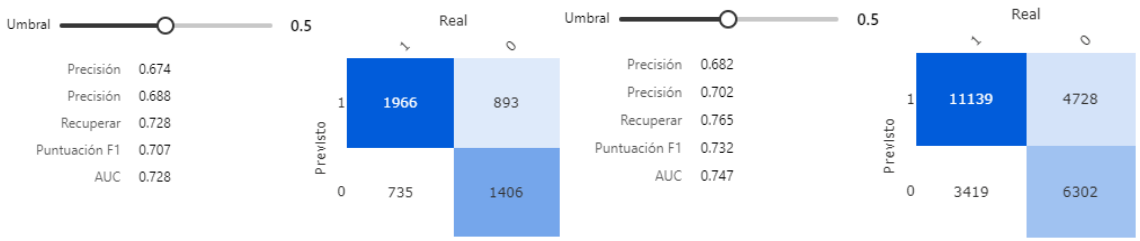


Nota: Imagen tomada de (Joaquín, Regresión logística simple y múltiple, 2020)

A continuación, se muestran los resultados obtenidos de los dos datasets mediante la técnica de regresión logística binaria:

Figura 35

Regresión Logística con 30.000 cuentas y con 85.292 cuentas finales



Como se puede apreciar los verdaderos positivos pasaron de un 39% hacia un 43,5% en el segundo entrenamiento (4% más), mientras que el porcentaje de falsos negativos y falsos positivos disminuyó en un 0,7%, lo cual, genera mayor cantidad de verdaderos positivos y verdaderos negativos, por ende, su exactitud (accuracy) y su precisión aumentaron, pero apenas llegan a un 68% y 70% respectivamente, por lo cual, no es el mejor modelo a seleccionar para este conjunto de características seleccionadas.

Two-Class Boosted Decision Tree

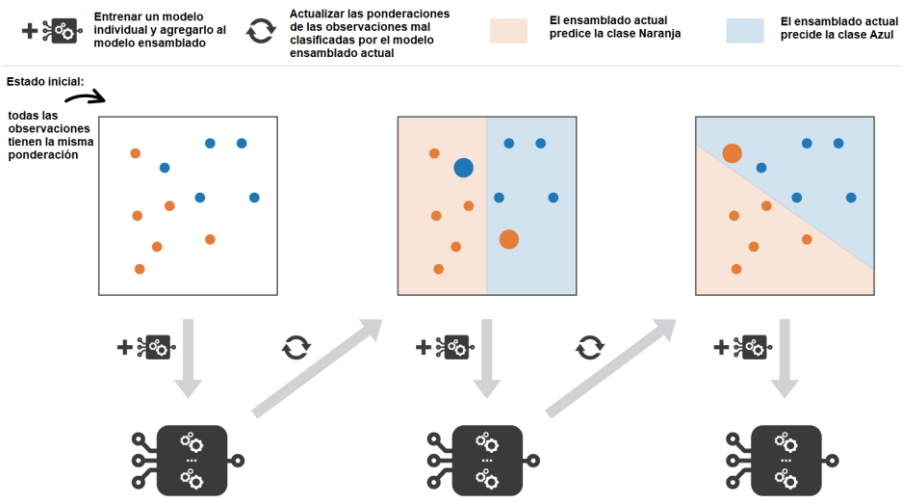
Como sostienen (Li, Lu, Coulter, Baccam, & Gilley, Two-Class Boosted Decision Tree module, 2020), este método usa el algoritmo de árboles de decisión ampliados o potenciados, donde el segundo árbol corrige los errores del primero, el tercer árbol corrige los errores del segundo y también del primer árbol y así sucesivamente, al juntarse las predicciones se establece un resultado más preciso.

A diferencia del método Bagging que se basa en votaciones sobre el método con mejor resultados, el método Boosted se enfoca en los errores, ambos son métodos ensamblados (ensemble) ya que, ejecutan múltiples algoritmos. Pero en el caso de métodos Boosted se ejecutan los árboles en serie y los pesos o ponderaciones son ajustados en base al aprendizaje

del modelo anterior enfocado en sus errores, tendiendo así mejores predicciones que al ejecutar un árbol individual.

Figura 36

Método Boosted



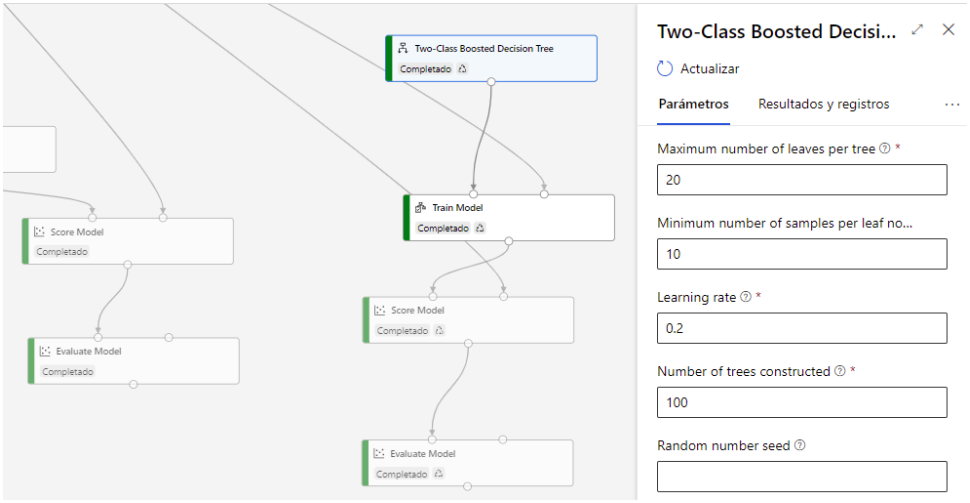
Nota: Imagen tomada de (Rocca, 2019)

Una desventaja podría ser que usa más recursos en memoria y su duración, por ejemplo, el entrenamiento con 85.292 cuentas y 100 árboles llegó a durar más de 3 horas en relación a otros algoritmos que duran alrededor de 1 hora.

Mientras se configure el uso de más árboles se tendrá una mayor cobertura, pero aumentará el tiempo de entrenamiento. En este caso se usó 100 árboles, tomando en cuenta que se tiene más de 250 características iniciales entre las que varían durante los 11 meses a evaluar y las características que se tiene en modo estático.

Figura 37

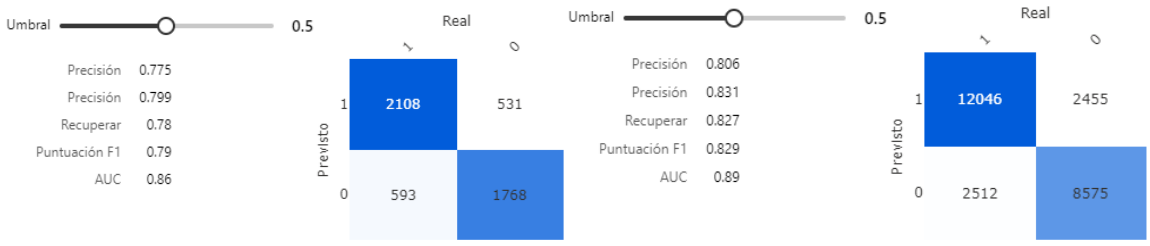
Árbol de Decisión Potenciado



A continuación, se muestran los resultados obtenidos de los dos datasets mediante la técnica de árboles de decisión potenciados:

Figura 38

Árbol de decisión potenciado con 30.000 cuentas y 85.292 cuentas finales



En este caso se puede apreciar como el porcentaje de falsos positivos y falsos negativos es mucho más pequeño en relación a los dos modelos anteriores adicionalmente, este porcentaje baja de un 25% a un 20%, esto significa un 5% más de valores verdaderos positivos y verdaderos negativos, por ende, sus métricas están por encima de los modelos anteriores.

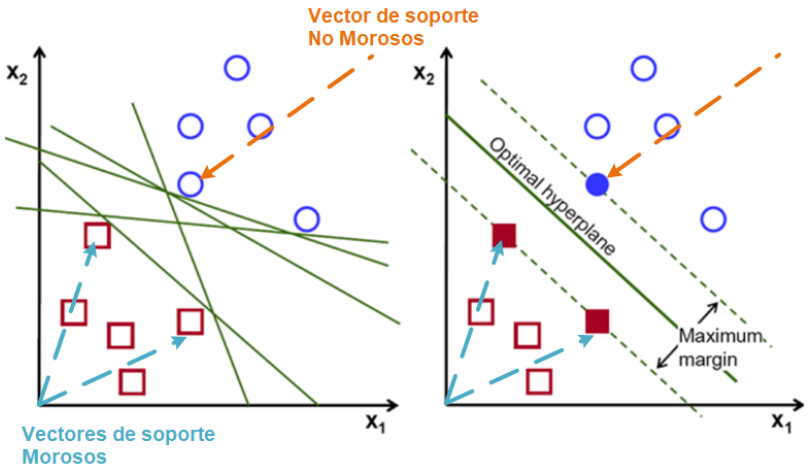
Two-Class Support Vector Machine

Como indican (Li, Lu, Coulter, Baccam, & Gilley, Two-Class Support Vector Machine module, 2020), las máquinas de vectores de soporte pueden ser usadas para la predicción basadas tanto en variables continuas (como el consumo de energía eléctrica) o variables categóricas (por ejemplo, si un mes fue medido el consumo o si fue promediado, o el estado civil del dueño de la cuenta), que son usadas tanto para clasificaciones como para regresiones.

Este algoritmo analiza estas características y reconoce patrones en un plano de características multidimensionales denominado hiperplano como lo indica (Ippolito, 2019). Todos los ejemplos de entradas (input) se representan como puntos en este plano y se mapean a categorías de salida, de tal manera que las categorías se dividen por un margen (gap), la exactitud dependerá de que tan eficiente se pueda realizar dicha distinción.

Figura 39

Support Vector Machine

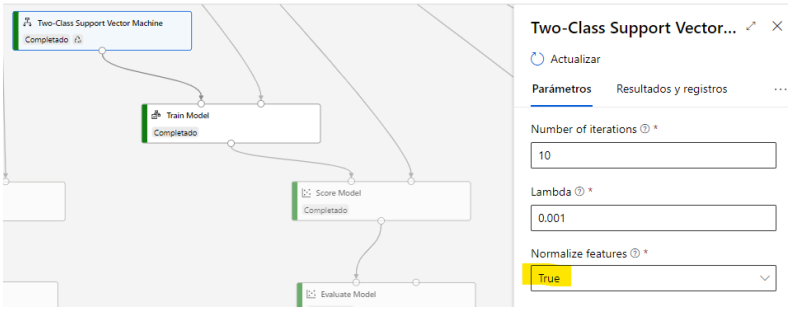


Nota imagen tomada de (Ippolito, 2019) donde las categorías de salida se dividen por un margen que permite reflejar la exactitud de la distinción de categorizar ambas salidas, por ejemplo, morosos y no morosos.

Adicional, se puede elegir si se normaliza la data de entrada, lo cual, es igualar distintas características, por ejemplo, en el caso de clases numéricas, llevarlas hacia valores comunes, ya que, va a ser de distinta escala los valores de consumo de KWH, versus los costos o la edad, cuando se ocupan ciertas métricas como el error absoluto medio, este utiliza la misma escala que los datos que se están midiendo, por lo tanto, no se puede usar para hacer comparaciones entre series que usan diferentes escalas, por ende, es ahí donde entra la normalización según indican (Baccam, Sharkey, Botsikas, & Gronlund, 2020).

Figura 40

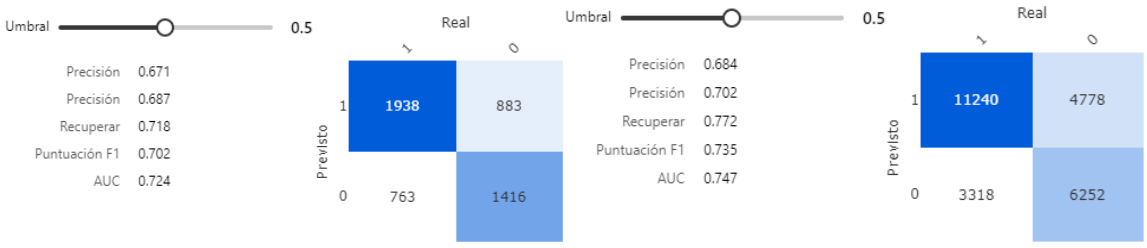
Normalizar el modelo Support Vector Machine



A continuación, se muestran los resultados obtenidos de los dos datasets mediante la máquina de vector de soporte.

Figura 41

Máquina de vectores de soporte con 30.000 cuentas y con las 85.292 cuentas finales



Como se puede apreciar los verdaderos positivos pasaron de un 38,76% hacia un 43,92% en el segundo entrenamiento (5% más), mientras que el porcentaje de falsos negativos

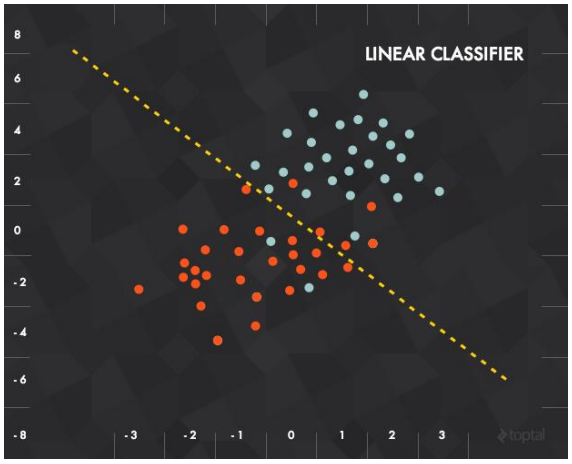
y falsos positivos disminuyó en un 1,2%, lo cual, genera mayor cantidad de verdaderos positivos y verdaderos negativos, por ende, su exactitud (accuracy) y su precisión aumentaron, pero apenas llegan a un 68,4% y 70,2% respectivamente, por lo cual, no es el mejor modelo a seleccionar para este conjunto de características seleccionadas.

Two-class Averaged Perceptron

Como sostienen (Li, Lu, Simpson, & Baccam, 2020) el modelo perceptrón promedio, es una versión previa y sencilla de las redes neuronales, donde las características de entrada se mapean hacia las distintas salidas posibles, por medio de una función lineal, en este caso las salidas son dos: morosos y no morosos, por ende, son más aptos para aprender sobre patrones separables de forma lineal, como se indica a continuación.

Figura 42

Clasificador lineal (Perceptrón)



Nota imagen tomada de (Vasilev, 2017)

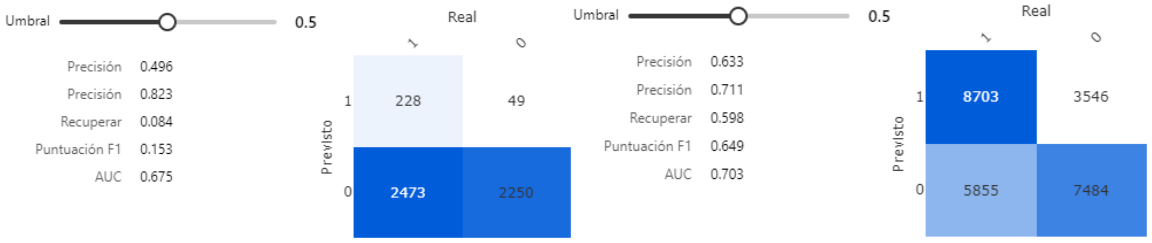
Adicionalmente, el ser Perceptrón promedio incurre en una mejora al modelo Perceptrón simple en el sentido de que los inputs son combinados con un conjunto de ponderaciones que se derivan del vector de características (Feature Vector).

También se debe tomar en cuenta que este algoritmo a nivel de Azure Machine Learning ya trae embebido la normalización de la data, teniendo así, un modelo optimizado para ser utilizado.

A continuación, se muestran los resultados obtenidos de los dos datasets mediante el perceptrón promedio.

Figura 43

Perceptrón promedio con 30.000 cuentas y con las 85.292 cuentas finales



En este modelo se da el caso de que los falsos negativos 49,46% son más que los verdaderos negativos (45%) por lo que en primera instancia se tuvo una exactitud (accuracy) y su precisión nada alentadora, sin embargo, en el segundo entrenamiento, disminuye el porcentaje de falsos negativos y falsos positivos a un 36,74% y el de verdaderos positivos y negativos aumenta hacia un 63,23%, por lo cual, su exactitud (accuracy) 63,3% y su área bajo la curva AUC 70,3% aumentaron, pero su precisión (71,1%) disminuyó debido a qué se considera como principal factor los verdaderos positivos tomando también en consideración los falsos positivos que crecieron en el segundo entrenamiento, por lo cual, no es el mejor modelo a seleccionar para este conjunto de características seleccionadas.

Two-Class Decision Forest

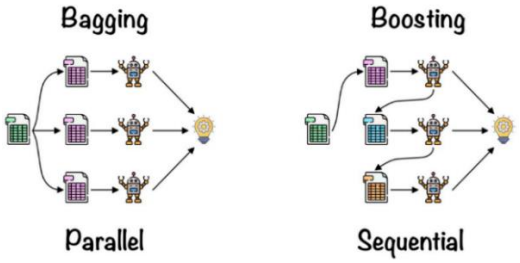
Como indican (Li, Lu, Baccam, & JiayueHu, Two-Class Decision Forest module, 2020), este modelo se basa en el algoritmo de bosques de decisión y al igual que el Boosted Decision

Tree, es un modelo de aprendizaje de ensamblados, apto para clasificaciones. Al ser un modelo ensamblado parte de que, se obtiene mejores resultados combinando modelos, es así que, se tiene una mejor cobertura de casos y exactitud que los modelos individuales.

En este caso se usa el método de votación (Bagging), que selecciona de entre varios árboles al de mejor ponderación.

Figura 44

Bagging vs Boosted

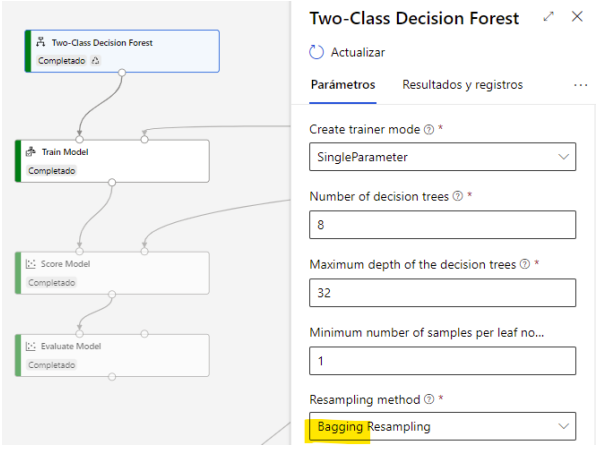


Nota imagen tomada de (R, 2021)

Estos árboles de decisión son más eficientes en los cálculos y uso de memoria por lo que, se los puede usar con grandes cantidades de datos.

Figura 45

Método Bagging Azure ML

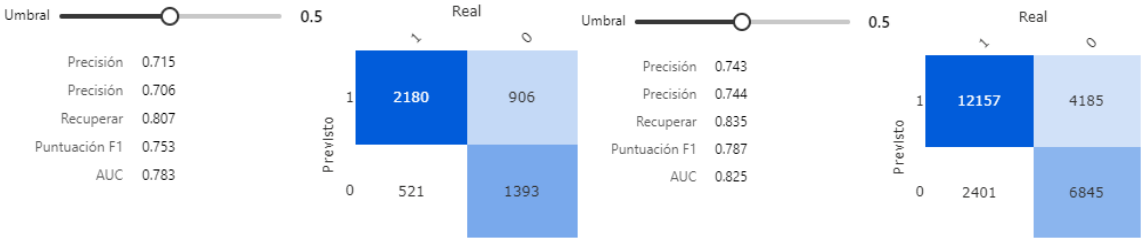


Si bien a nivel de Azure Machine learning se tiene un módulo de selección de características, que va en relación a su correlación con la etiqueta destino (target) que se desea inferir, la cual, es en este proyecto de tesis es: moroso y no moroso, este bosque de decisión trae embebido dicho módulo de selección de características.

Se adjuntan los resultados del modelo.

Figura 46

Árbol de decisión con 30.000 cuentas y con 85.292 cuentas finales



Como se indica en el concepto y características del modelo, este bosque de decisión se adapta a data no limpia, al uso de un número grande de características y a data con varios tipos de distribución, por ende, se verifica que la proporción entre falsos positivos y falsos negativos hacia los verdaderos positivos y verdaderos negativos bajó de un 39,94% hacia un 34,66%, es decir un 5,28% más verdaderos positivos y verdaderos negativos, lo cual, permite confirmar como el modelo mejora, si bien, este modelo tiene una exactitud del 74,3% y precisión del 74,4% es el segundo modelo más preciso luego del Boosted Decision Tree.

Factores de Evaluación

Dependiendo de los tipos de modelos, existen diversos tipos de factores o métricas de desempeño que permiten evaluar los mejores algoritmos que se adapten a la realidad de la data, para este proyecto de tesis y su problemática, se usaron los modelos de clasificación binaria, cuyas métricas se detallan a continuación:

Métricas en Modelos de Clasificación binaria

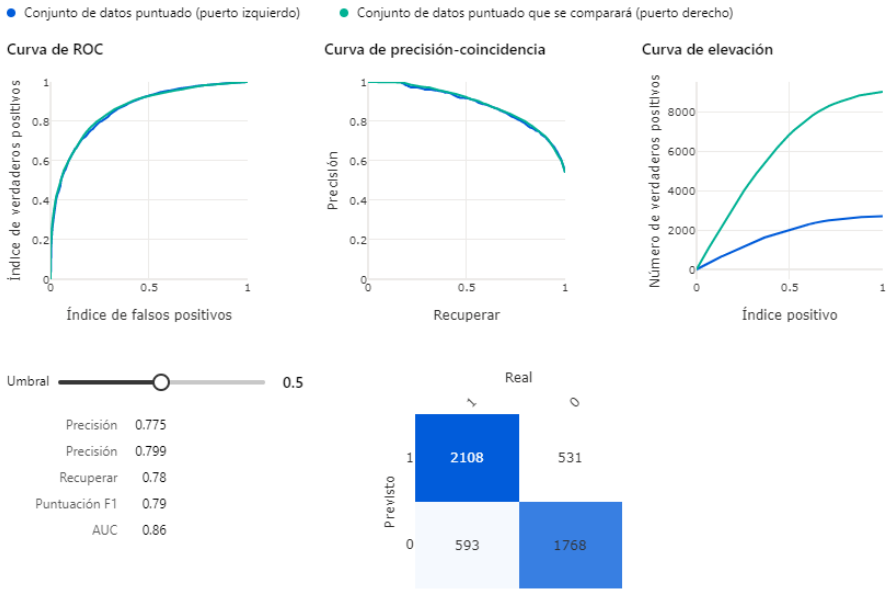
Dentro de este tipo de métricas según indica en (MS Azure ML, 2019), parecería que la exactitud es la más intuitiva, pero no debe ser nunca, la única forma de medir, cuál es el mejor modelo, ya que, por ejemplo, si se tiene el caso de que, un grupo de clientes o sector de clase alta, donde solo un ínfimo porcentaje (2%) de la población deja vencer sus deudas, se puede crear un modelo (erróneo) que siempre brinde el resultado de no morosidad, por lo que, si solo se aplica como métrica la exactitud, este modelo tendría el 98% de eficacia, pero no es real, ni útil, para todo el universo de la población, es por eso que, se usan las otras métricas como la precisión y la coincidencia (recuperación) para medir esta clase de modelos.

Como lo sustentan (Gregory, Quintanilla, Cartacio, & Gronlund, 2020), la tecnología de Microsoft Azure Machine Learning, esta no hace una distinción per se entre modelos binarios y multi-nomial o multi-clases, sino que más bien, trata a la parte verdadera y a la parte falsa como partes independientes, como si se tuvieran varias clasificaciones y se utilizan los sufijos micro, macro y ponderado, este es el equivalente a obtener una métrica de la parte verdadera y por separado de la falsa y así obtener el promedio.

Según (Gregory, Quintanilla, Cartacio, & Gronlund, 2020) AUC es el área bajo la curva, la cual, brinda la calidad de las inferencias en el modelo, mientras más grande el área mejor es su calidad, la curva representa a la característica operativa del receptor (ROC), que es la relación entre las proporciones de verdaderos positivos (TPR) del eje Y, y falsos positivos (FPR) eje X, a medida que cambia el umbral de decisión. Sin embargo, en la clasificación binaria el valor devuelto es el valor de AUC de la clase positiva más presumible. AUC permite comparar distintos tipos de modelos como se indica en la imagen siguiente.

Figura 47

AUC - área bajo la curva



Adicionalmente, también se tiene las métricas de exactitud, precisión recuperación y puntación F1, como indican (Gregory, Quintanilla, Cartacio, & Gronlund, 2020), Exactitud (Accuracy) es la proporción de predicciones correctas, es decir, verdaderos positivos más verdaderos negativos sobre el total de predicciones; Precisión es la habilidad del modelo de evitar etiquetar casos negativos como positivos y se calcula mediante el número de verdaderos positivos sobre el número de verdaderos positivos más el número de falsos positivos; Recuperación o coincidencia (Recall) es la capacidad del modelo de hallar los casos positivos, es decir, la fracción de casos clasificados como positivos que son los casos reales u observados, su fórmula es el número de verdaderos positivos sobre el número de verdaderos positivos más el número de falsos negativos y Puntuación F1 es la métrica que combina a las de precisión y de exhaustividad (recall), es decir, es la media harmónica de estas dos, es una buena métrica balanceada de ambos: falsos positivos y falsos negativos, pero, no se considera verdaderos negativos dentro del conteo (Gregory, Quintanilla, Cartacio, & Gronlund, 2020).

En las 5 métricas anteriores el rango de medición es entre 0 y 1, siendo 1 el mejor, con lo cual, se responde a la pregunta de investigación: OE3 – RQ3.1: ¿Qué técnica de análisis o métrica se puede usar para evaluar el/los modelo(s) planteado(s) sobre cartera vencida?

Resumen de la competencia de modelos

Partiendo de las métricas anteriormente descritas, se puede confirmar que el modelo ganador fue el Two-class Boosted Decision Tree, resumiendo los resultados en la siguiente tabla y respondiendo así también a la pregunta de investigación OE2 – RQ2.1: ¿Cuál(es) modelo(s) predictivo(s) se adapta(n) a la realidad de la empresa eléctrica de Cotopaxi?

Tabla 13

Tabla de comparación modelos con 30.000 y 85.292 cuentas

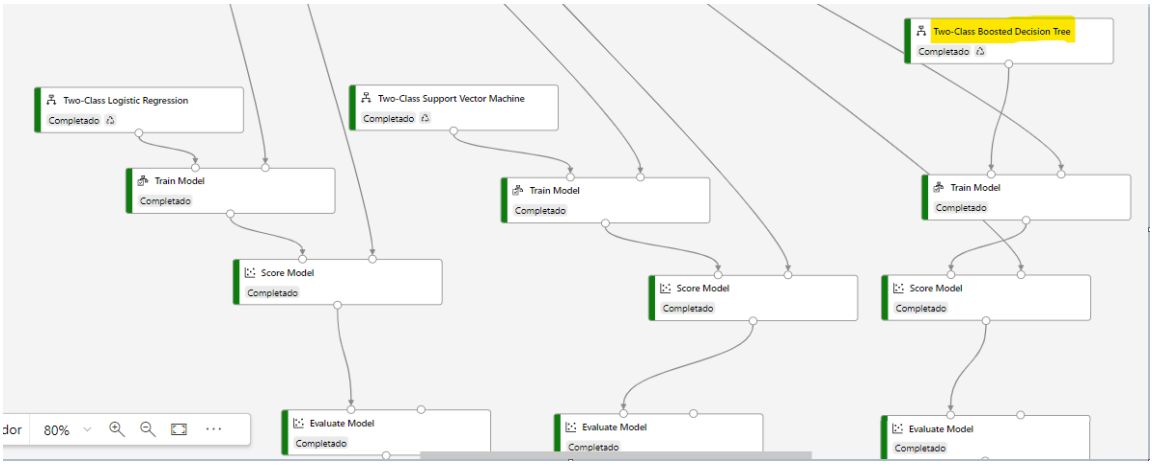
	Exactitud	**Precisión**	**Coincidencia**	**F1**	**AUC**
Two-class Neural Network	0.668	0.656	0.812	0.726	0.719
	0.665	0.649	0.896	0.753	0.75
Two-class Logistic Regression	0.674	0.688	0.728	0.707	0.728
	0.682	0.702	0.765	0.732	0.747
Two-Class Boosted Decision Tree	0.775	0.799	0.78	0.79	0.86
	0.806	**0.831**	**0.827**	**0.829**	**0.89**
Two-Class Support Vector Machine	0.671	0.687	0.718	0.702	0.724
	0.684	0.702	0.772	0.735	0.747
Two-Class Averaged Perceptron	0.496	0.823	0.084	0.153	0.675
	0.633	0.711	0.598	0.649	0.703
Two-Class Decision Forest	0.715	0.706	0.807	0.753	0.783
	0.743	0.744	0.835	0.787	0.825

Nota. Esta tabla muestra el resumen de resultados de la competencia de modelos

Como se pudo apreciar en la tabla resumen y se explicó bajo cada uno de los resultados en la sección anterior de la ***Comparativa de modelos***, tanto el modelo potenciado de árboles de decisión como el bosque de decisión resultaron ser el primero y segundo más precisos respectivamente, de la tabla anterior se puede resumir que los mismos tiene una exactitud mayor que los otros modelos, esta métrica se basa en las predicciones correctas versus las falsas, adicionalmente, evitan de mejor manera etiquetar casos negativos como positivos (precisos) y poseen una área bajo la curva mucho más grande que los anteriores modelos por ende poseen un mejor desempeño a la hora de inferir valores realmente verdaderos y también poseen un Score F1 mayor, el cual, balancea falsos positivos y falsos negativos comparados con los verdaderos positivos, por lo que, se puede afirmar que dada la gran cantidad de características de la data de la empresa eléctrica de Cotopaxi, es apropiado el modelo de bosques de decisión, pero, ya que, ambos modelos ganadores son métodos ensamblados que ejecutan múltiples algoritmos, el modelo de árboles de decisión potenciado tiene la ventaja de que, considera los errores resultantes del modelo anterior y los usa para mejorarla ejecución del modelo siguiente (ensemble), lo cual, lo hace más preciso.

Figura 48

Modelo ganador Two-class Boosted Decision Tree



Pasos posteriores del Desarrollo de Modelos

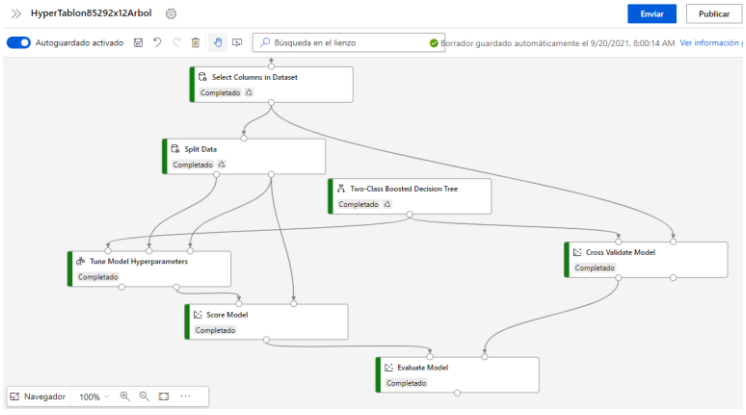
Afinamiento del modelo ganador

Como se puede apreciar en la siguiente figura del histórico de versiones del diseñador de experimentos en Azure ML, el proceso de modelamiento no termina con verificar cual fuel modelo ganador, por lo contrario, se continua con su afinamiento u optimización de acuerdo a la data que se tiene, tal como, se lo indica en la metodología CRISP-DM (Health Data Miner, 2021) y en el ciclo de vida de la analítica de datos y operacionalización de modelos (Herrera, 2019).

Agregado a lo anterior, mediante cada uno de los experimentos, estos aprendizajes van arrojando valores de exactitud, precisión, coincidencia, puntuación F1 y áreas bajo la curva más acertados u optimizados, de acuerdo a las características más limpias o parámetros más precisos en relación a la data que se vaya agregando como insumo para el modelo. Por ende, para llegar a parámetros más precisos se fueron comparando dos procesos de analítica de datos, la validación cruzada y el afinamiento de hiperparámetros como se lo indica en la siguiente imagen.

Figura 49

Afinamiento de hiperparámetros versus validación cruzada



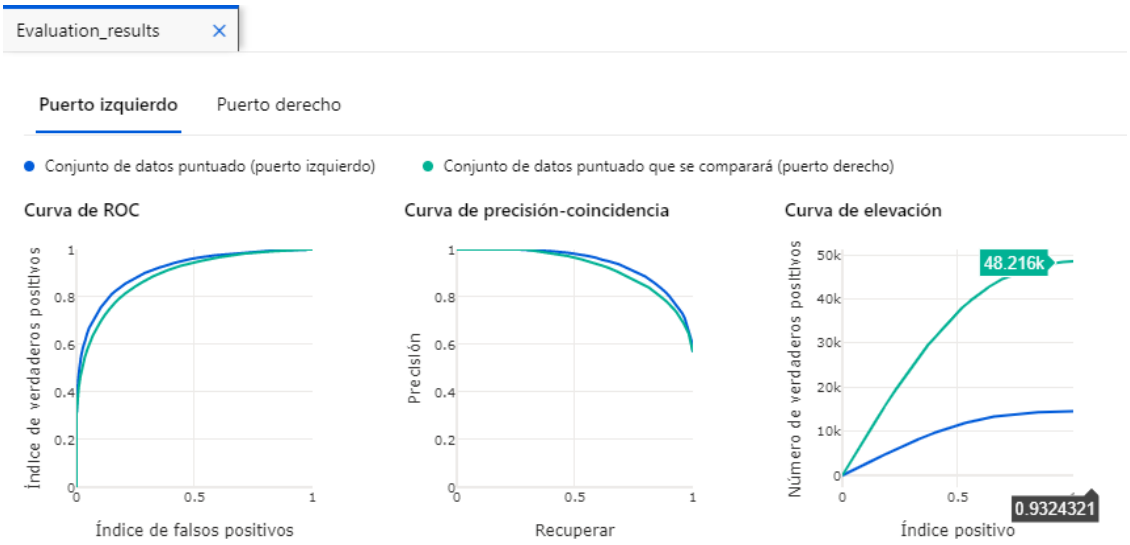
Según (Li & Lu, Cross Validate Model, 2020), la validación cruzada de una forma aleatoria divide a la data en agrupaciones, a diferencia del Split, que lo realiza una sola vez, como se lo indicó en la sección ***Diseñar el flujo para entrenar los modelos***, esto, además de que, previene el overfitting o el sobre entrenamiento para solo una porción de los datos, permite determinar, cuan sensitivo es el modelo respecto a la variabilidad de la data.

Por otro lado, como indican (Li & Lu, Tune Model Hyperparameters, 2020), el afinamiento de hiperparámetros permite ir comparando distintas configuraciones del modelo ganador hasta encontrar el mejor nivel de exactitud respecto a la data y a la configuración del modelo ganador.

En la sección de ***Factores de evaluación*** se explicó la métrica de Exactitud que, en este caso, se la toma en cuenta en el desempeño de la clasificación binaria, mientras que, en cuanto al desempeño de la calidad en regresiones se toma en cuenta el Coeficiente de determinación (R^2). Este coeficiente indica que proporción de la varianza entre los valores inferidos y los valores reales (observados de la muestra), es explicado por el modelo, es decir, qué tan apegado a los valores reales está la variable dependiente inferida por el modelo. Posteriormente, se analizan los resultados que se los realizaron de forma similar en la competencia de modelos tanto con 30.000 cuentas iniciales como con las 85.292 cuentas más limpias.

Figura 50

Resultados hiperparámetros versus validación cruzada



Como se puede apreciar en la curva ROC (característica operativa del receptor) el área bajo la curva de color azul es más grande y también en la curva de precisión-coincidencia (Precision-recall) donde el modelo más acertado es el de la entrada izquierda que usa Hiperparámetros y dispone de mayor habilidad para detectar valores positivos, así como evitar falsos positivos y falsos negativos, sin embargo, se puede apreciar en la curva de elevación (beneficios acumulativos) que el modelo con validación cruzada es más eficiente para descubrir los valores positivos que el modelo con hiperparámetros, por lo cual, se validó la forma de que ambos modelos se complementen o integren como se indica posteriormente.

Tabla 14

Tabla de comparación entre hiperparámetros versus validación cruzada

	Exactitud	**Precisión**	**Coincidencia**	**F1**	**AUC**
Hiperparámetros	0.821	0.839	0.827	0.833	0.904
	0.829	**0.851**	**0.847**	**0.849**	**0.913**
Validación Cruzada	0.783	0.800	0.799	0.799	0.863
	0.805	0.831	0.826	0.828	0.891
Comparación con el árbol sin ninguna optimización con 30.000 y 85.292 cuentas más limpias:					
Two-Class Boosted Decision Tree	0.775	0.799	0.78	0.79	0.86
	0.806	0.831	0.827	0.829	0.89

Nota. Esta tabla muestra los resultados de aplicar hiperparámetros y validación cruzada

Como se puede apreciar en la tabla anterior, la mejora en las métricas del modelo confirma que la optimización mediante hiperparámetros perfecciona el modelo, sin embargo, se hace una adaptación adicional ajustando el modelo de hiperparámetros para que este contenga también internamente la opción de validación cruzada, en este caso se baja un poco el mejoramiento en cuánto a las métricas, pero, se garantiza que no exista overfitting (Li & Lu, Tune Model Hyperparameters, 2020), esto se logra sin ingresar la data dividida mediante el Split en el tercer input del módulo, sino dejándolo libre para que el mismo módulo realice la validación cruzada.

Figura 51

Comparación de hiperparámetros con validación cruzada y con Split 70/30

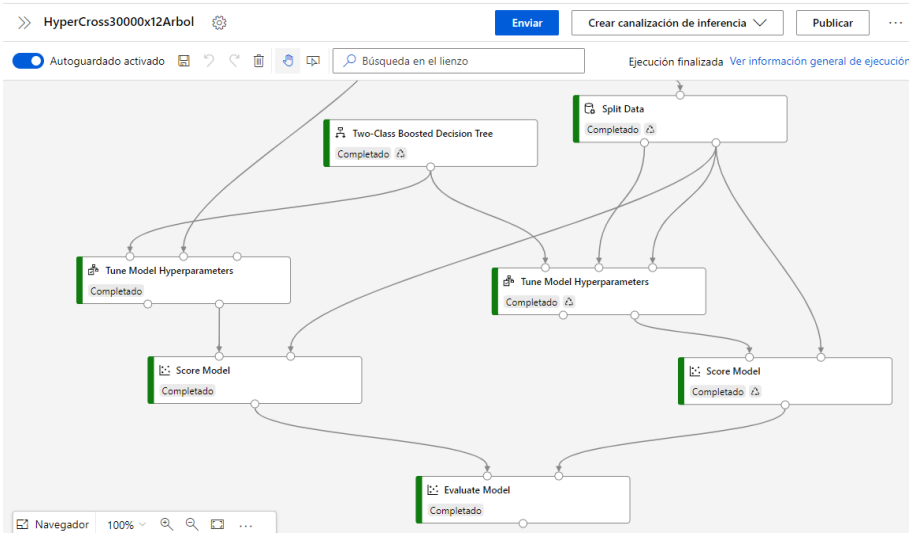
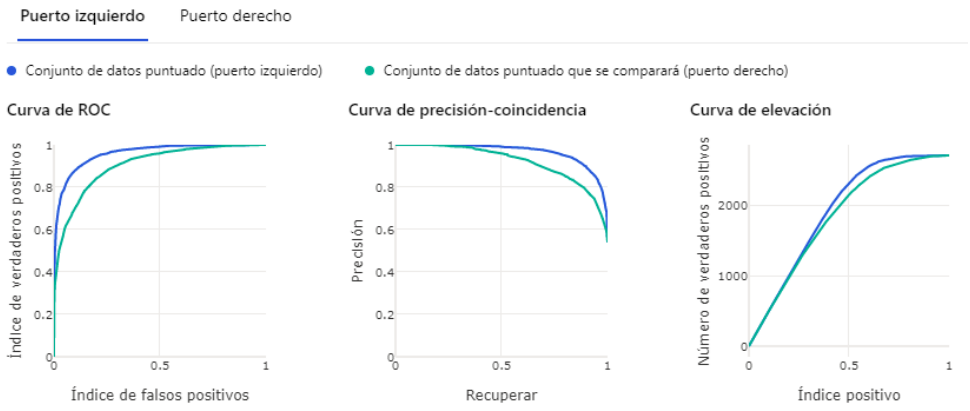


Figura 52

Comparación de resultados de hiperparámetros con validación cruzada y con Split 70/30



Como se indicó anteriormente se buscó complementar el modelo con hiperparámetros más la evaluación cruzada, de esta forma la curva de elevación (beneficios acumulativos) es más eficiente en cuánto a la búsqueda de valores positivos, teniendo como objetivo no equivocarse dando falsos negativos, que en realidad son valores positivos.

Tabla 15

Tabla de hiperparámetros con validación cruzada y con Split 70/30 con 30.000 y 85.292 cuentas

	Exactitud	**Precisión**	**Coincidencia**	**F1**	**AUC**
Hiperparámetros con validación	**0.889**	**0.901**	**0.892**	**0.897**	**0.959**
cruzada	0.829	0.852	0.848	0.85	0.913
Hiperparámetros con un Split	0.821	0.839	0.827	0.833	0.904
70/30	0.829	0.851	0.847	0.849	0.913

Nota. Esta tabla muestra los resultados de validación cruzada versus Split 70/30

Como se puede apreciar en la tabla anterior, el modelo con Validación Cruzada es mucho más preciso, con un 95.9% de área bajo la curva, pero con el universo de 30.000 cuentas inicial (aún sin limpieza de datos), el mismo modelo baja de 91.3% y a 85.2% en precisión, con las 85.292 cuentas residenciales finales, pero aun así sigue siendo mejor que el mismo proceso de hiperparámetros sin validación cruzada, como se indica en el modelo anterior solo con el SPLIT de 70/30.

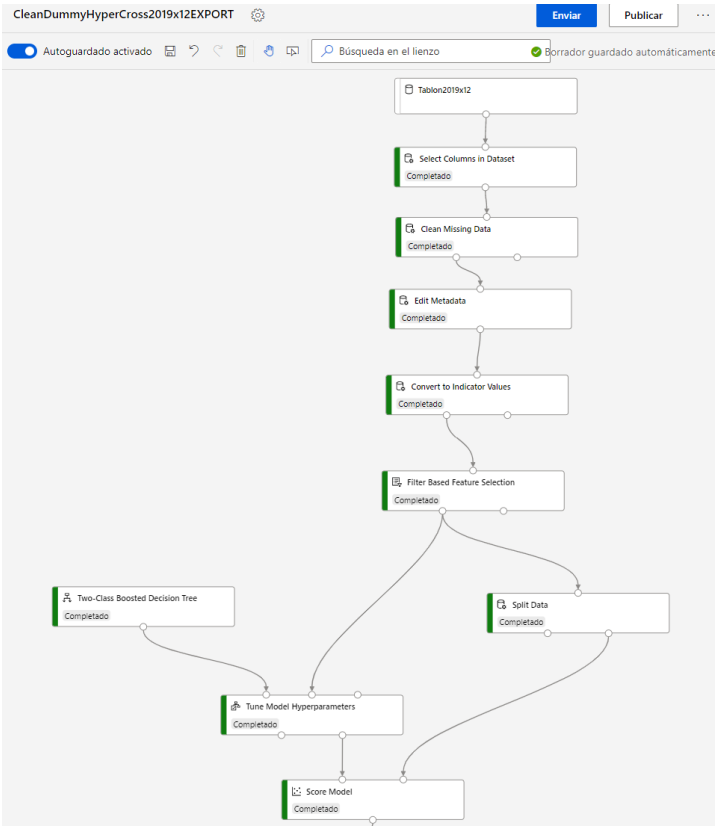
Una vez que, se determinó que la mejor optimización del modelo es mediante el afinamiento de Hiperparámetros y su opción interna de validación cruzada evitando así el overfitting, se probó con la normalización de datos, sin embargo, el modelo no tuvo una mejora importante, nada más un 0.3% en la coincidencia (recall), no así en las demás métricas, por lo que, concuerda con la teoría de modelos, indicada por la propia Microsoft mediante (Zhang, Howell, Gronlund, Lu, & Gilley, 2019) de no normalizar de una manera indiscriminada, ya que, algunos algoritmos ya usan la normalización por debajo.

Adicionalmente, se complementó con la exclusión de características que pertenecen a diciembre 2019, para no interferir con la inferencia y se incluyeron los módulos de limpieza de

data, además de la conversión a Indicator Values que se detallaron en secciones anteriores y la selección de características de mayor importancia según la correlación de Pearson respecto al target a inferir, el cual es, la cartera vencida de diciembre 2019, como se lo indica en la siguiente figura del modelo final obtenido.

Figura 53

Modelo final obtenido



Crear una canalización de predicción o inferencia

Una vez diseñado el modelo que más se adaptó a la necesidad y realidad de la data, se genera una nueva canalización o pipeline, que encapsula la limpieza de datos, la normalización o preparación de la data y el modelo de entrenamiento realizado previamente, para que este nuevo flujo pueda recibir nueva data, e inferir o predecir el objetivo para el cual fue creado, es

decir, se tiene un experimento o pipeline de entrenamiento y uno de predicción a ser publicado, lo que se describe en cada una de las secciones de modelamiento siguientes.

Como se indica en (Microsoft, 2019), al pasar el modelo hacia una canalización de inferencia en tiempo real, este adquiere el input de un Servicio Web para poder ser consumido como su nombre lo indica, en tiempo real, ingresando los nuevos valores para inferir o predecir el valor deseado.

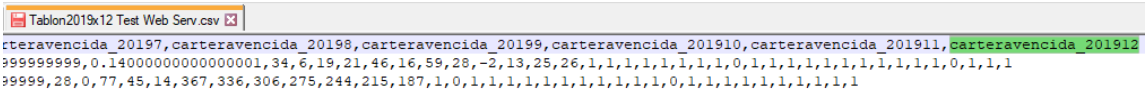
En este caso los pasos de Limpieza de datos y la conversión Indicators Values se los ha almacenado en set de datos de transformación (TD-) mientras que el modelo ha sido almacenado en un módulo de set de datos (MD-).

Adicionalmente, se reemplaza el tablón estadístico que sirvió de input para al anterior modelo de entrenamiento, por el ingreso manual, con lo cual, se probará la inferencia, este ingreso manual, lleva el mismo formato csv que el tablón, para este caso, el módulo de evaluación ya no es necesario, al predecir nueva información.

Ya que se desea predecir la cartera vencida en este caso del mes 12, dicho campo no se lo envía carteravencida_201912, justamente para que sea el modelo de inferencia, quien, valga la redundancia, lo infiera, en base a todo el entrenamiento anteriormente descrito, por ende, en la prueba manual no se lo envía.

Figura 54

Data Manual sin la columna a predecir



Pruebas manuales del modelo de inferencia

Siguiendo el proceso de modelamiento, para que el resultado sea más liviano se puede

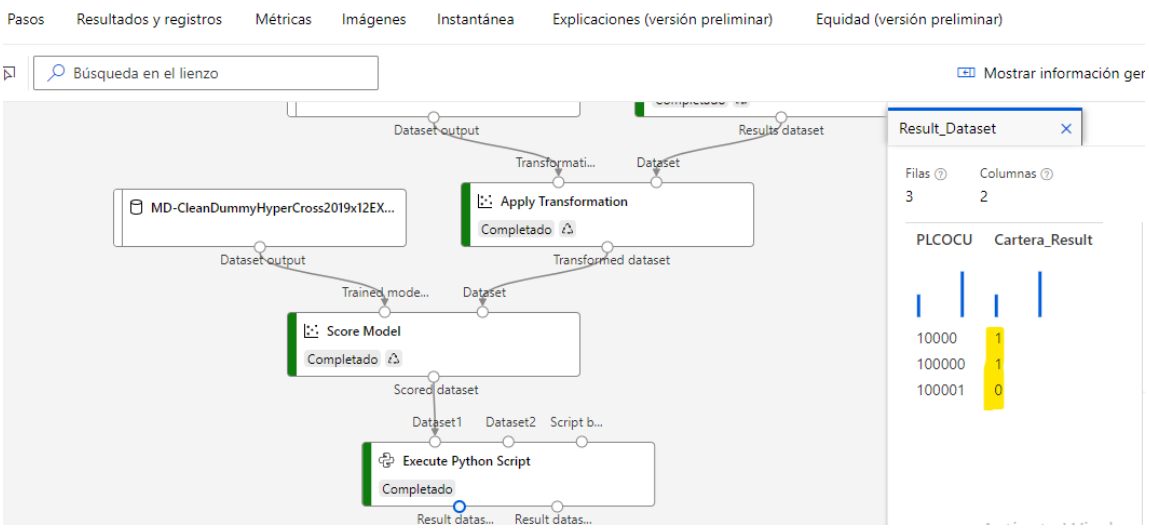
limitar la canalización de inferencia para que devuelva solo lo necesario, por ejemplo, el código de cuenta y su valor predicho de cartera vencida en el mes 12, dicha tarea se la hizo mediante lenguaje Python por medio del siguiente este script siguiendo los ejemplos de (Azure Machine Learning, 2019)

```
import pandas as pd
def azureml_main(dataframe1 = None, dataframe2 = None):
    scored_results = dataframe1[['PLCOCU', 'Scored Labels']]
    scored_results.rename(columns={'Scored Labels':'Cartera_Result'},
                          inplace=True)
    return scored_results
```

Donde, PLCOCU es el código de la cuenta y el valor predicho se lo ha llamado Cartera_Result, se ejecuta manualmente el modelo de inferencia y se puede comparar las predicciones o inferencias realizadas versus la data que se tiene, como se indica en la siguiente imagen.

Figura 55

Cartera, resultado de la inferencia en Azure ML



Crear una canalización de predicción o inferencia por lotes

Al tratarse de miles de cuentas y cientos de características o conceptos (número de columnas), es más eficiente el consumo del modelo por medio de procesos Batch o por lotes, por lo cual, luego de realizadas las pruebas se ha observado que es más eficiente y fácil su consumo por lotes. Por lo cual, se redactaron solo como conocimiento adquirido las secciones ***Crear una canalización de predicción o inferencia*** y ***Pruebas manuales del modelo de inferencia*** que son muy similares al traspaso hacia la canalización de predicción por lotes y el código Python igual se lo reutiliza.

Como indican (Li, O'Brien, Gronlund, & Coulter, 2021), al pasar el modelo hacia una canalización de inferencia por lotes, este debe recibir un nuevo input mediante un nuevo conjunto de datos (DataSet) para poder ser consumido como su nombre lo indica por lotes, ingresando los nuevos valores para inferir o predecir el valor deseado, se reemplaza el tablón estadístico que sirvió de input para el anterior modelo de entrenamiento, con lo cual, se probará la inferencia y lleva el mismo formato csv que el tablón, también el módulo de evaluación ya no es necesario al predecir nueva información, ya que, el entrenamiento previamente ya fue realizado.

Como se lo explicó anteriormente en el consumo vía Web Service, ya que, se desea predecir la cartera vencida en este caso del mes 12, dicho campo no se lo envía carteravencida_201912, justamente para que sea el modelo de inferencia, quien, valga la redundancia lo infiera, en base a todo el entrenamiento anteriormente descrito, por ende, no se lo envía, el nuevo dataset es ahora un parámetro de la nueva canalización o pipeline de inferencia por lotes.

Similar a la sección ***Crear una canalización de predicción o inferencia***, en este caso la

canalización por lotes, encapsula los pasos de Limpieza de datos y la conversión Indicators Values se los ha almacenado en set de datos de transformación (TD-) mientras que el modelo ha sido almacenado en un módulo de set de datos (MD-).

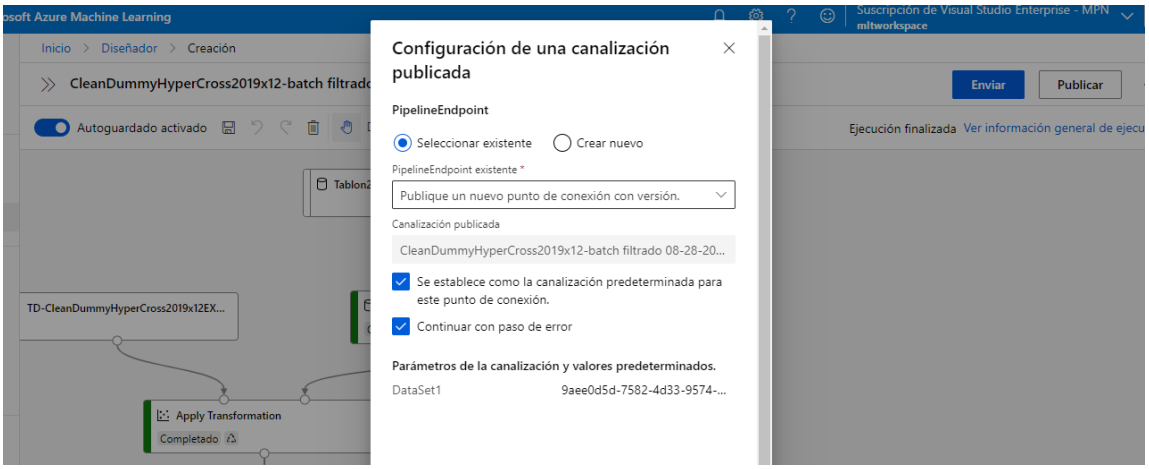
Publicación

Como se indica en (Microsoft Learn, 2019), para estos experimentos se seleccionó desplegar los modelos como servicios Web Rest API, para ser consumidos con nueva data partiendo de las secciones ***Crear una canalización de predicción o inferencia*** y ***Pruebas manuales del modelo de inferencia***, de donde se estimaba que, por tratarse de un proyecto de tesis, se usaría la tecnología de instancia de contenedores Azure de baja prioridad, aunque, para ambientes productivos se tiene la tecnología de contenedores orquestados mediante Kubernetes que mejora su escalabilidad, obviamente los costos Kubernetes son superiores a las instancias de contenedores Docker de Azure (ACI) para ambientes de pruebas.

Sin embargo, al tratarse de miles de cuentas, se pudo apreciar que, es más rápido el consumo del modelo por medio de procesos Batch o por lotes, por lo cual, luego de realizadas las pruebas, se ha observado que es más eficiente y fácil, su consumo mediante dicha opción.

Figura 56

Publicar canalización por lotes en Azure ML



El consumo por Batch o por Lotes se lo detallará en el siguiente capítulo de resultados.

Capítulo VI: Resultados

Características más importantes

Para responder a la pregunta de investigación: OE2 – RQ2.2: ¿Cuáles son los patrones que tienen los consumidores con morosidad en la empresa eléctrica de Cotopaxi?, una vez analizado el modelo ganador, adicional al entendimiento de los datos del ***capítulo 4***, se puede determinar las siguientes características más importantes a tomar en cuenta y por las cuales, el modelo se ha guiado para su proceso de inferencia.

Tabla 16

Tabla de características de la empresa eléctrica de Cotopaxi más importantes

Características	Descripción	Observación
Cartera Vencida	Cartera vencida de los meses previos al mes a inferir	El patrón de comportamiento en meses anteriores respecto a estas 5 primeras características
PLFEMO	Fecha en la que se calculó la mora de los meses previos al mes a inferir	

Características	Descripción	Observación
PLVPMO	Valor pagado por mora de los meses previos al mes a inferir	es uno de las factores más importantes para la predicción
Diferencia	Tiempo que se demoró en pagar en los meses anteriores al mes predicho	Se deriva de la fecha de pago menos la fecha de vencimiento
PLVALO	El valor a pagar en el mes a predecir y el valor pagado de los meses anteriores	Si es un valor excesivo, interfiere en la fecha del pago dependiendo de las otras características relacionadas
DEUDA	Sumatoria total de la deuda	Si tiene ya una deuda y otra acumulada interfiere en su comportamiento de pago
MORA TOTAL	Sumatoria total de la mora	
PLCOAC	Consumo de energía activa	Diferencia del consumo anterior menos el actual a lo largo del tiempo
PLCOSE	Código del sector	Sector, parroquia y cantón donde se encuentra el medidor, si este por ejemplo pertenece a la zona urbano marginal
PARROQUIA		
CANTON		
CLFEIN	Fecha de ingreso o creación de la cuenta	Esta característica va relacionada con los conceptos de deuda y mora total, porque los puede ir acarreando data desde hace años atrás
Fecha de Nacimiento	La edad se había derivado de esta característica solo para un mejor entendimiento humano, sin embargo, el modelo entiende mejor con la fecha de nacimiento en formato numérico	Se pudo detectar que la mayoría de dueños de cuentas, es de, las personas entre los 54 años, esta característica pasó por los procesos de limpieza de datos
Promedio	Promedio de consumo de KWH	Se relaciona con PLCOAC en cuanto a la diferencia de consumo de KWH

Características	Descripción	Observación
PLTMAC	Variable categórica mensual que indica si se midió o promedió el consumo eléctrico	
PLESTA	Estado de la planilla: CAN canceladas, CNJ canjeadas, CAC canjeada y con afectación contable, CYR canjeada y re facturada, VEN vencida, GEN generada, ABO abonada	Se puede ver el historial de comportamiento de pago con esta característica, por ejemplo, si siempre deja vencer su planilla
TIPO	Indica si el cliente está en un sector, urbano, rural, rural o urbano marginal.	Según los análisis realizados la zona con mayor cartera vencida es la zona marginal
CLCLAS	Persona Natural, Institución, Particular	
CLTIPO	Es el tipo de cliente si es particular o comercial	Se tienen 34 registros que deberían pasar por data cleaning con el CLTIPO: EMP (empleado) equivalente a un 0,3%, pero existen alrededor de 400 empleados en la eléctrica, por lo que, 34 no es correcto.
CONDICION_D	Condición del ciudadano posee los estados de si tiene discapacidad mental, si es residente en el exterior, menor de edad, militar, policía, analfabeto, fallecido entre otros	Sin embargo, tiene una menor importancia ya que el 30% de cuentas no dispone de esa información
Estado_Civil_D	Esta categoría tiene los estados de Soltero, Viudo, Divorciado en unión de hecho o casado	De forma similar un 30% de datos es nulo

Nota. Esta tabla muestra las características más importantes para la predicción

Las características de CLTIPO, CONDICION_D y Estado_Civil_D, así como, la fecha de nacimiento, deberían pasar por procesos de limpieza de datos, desde los ingresos transaccionales o bien igualar su información, mediante la data que se pueda obtener de la

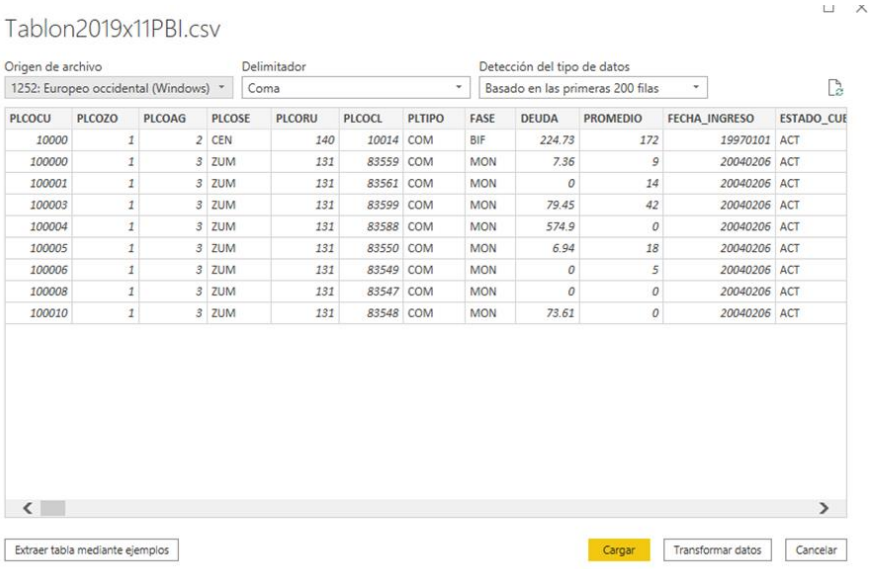
consultoría de catastros que dispone la empresa eléctrica de Cotopaxi.

Consumo del servicio de predicción

Como se ha indicado en anteriores capítulos, se ha utilizado también Microsoft Power BI por sus conexiones nativas hacia Azure ML y para mostrar las gráficas de usuario final, por lo que, se han tomado datos de los 11 meses del 2019 (enero - noviembre) y se ha dejado vacía la columna de cartera vencida del mes 12 (diciembre) a ser predicha, bajo el consumo del servicio REST API mediante la conexión a su ENDPOINT.

Figura 57

Datos 2019 meses de enero a noviembre



PLCOCU	PLCOZO	PLCOAG	PLCOSE	PLCORU	PLCOCL	PLTIPO	FASE	DEUDA	PROMEDIO	FECHA_INGRESO	ESTADO_CUE
10000	1	2	CEN	140	10014	COM	BIF	224.73	172	19970101	ACT
100000	1	3	ZUM	131	83559	COM	MON	7.36	9	20040206	ACT
100001	1	3	ZUM	131	83561	COM	MON	0	14	20040206	ACT
100003	1	3	ZUM	131	83599	COM	MON	79.45	42	20040206	ACT
100004	1	3	ZUM	131	83588	COM	MON	574.9	0	20040206	ACT
100005	1	3	ZUM	131	83550	COM	MON	6.94	18	20040206	ACT
100006	1	3	ZUM	131	83549	COM	MON	0	5	20040206	ACT
100008	1	3	ZUM	131	83547	COM	MON	0	0	20040206	ACT
100010	1	3	ZUM	131	83548	COM	MON	73.61	0	20040206	ACT

Mediante la carga de datos se usó la opción de transformación para realizar el consumo del EndPoint mediante la funcionalidad de Microsoft Azure Machine Learning de Power BI.

Sin embargo, como se explicó en las secciones ***Crear una canalización de predicción o inferencia por lotes*** y ***Publicación***, luego de realizadas las pruebas se identificó que, es más óptimo y fácil el consumo del modelo por medio de la canalización de inferencia por lotes o

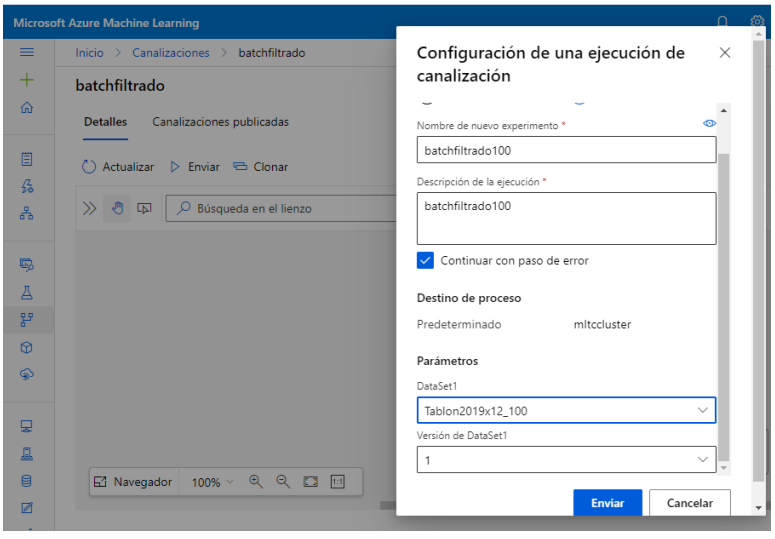
Batch, debido al gran número de conceptos o características y al número de cuentas o clientes a consultar, por lo cual, esta opción ha sido la seleccionada en este proyecto de tesis y se la describe en la siguiente sección.

Consumo por lotes

Una vez que se obtuvo la canalización por lotes publicada, se la consumió por medio de la inclusión del nuevo conjunto de datos o DataSet, en este caso se creó un conjunto de datos con 100 cuentas de ejemplo, sin la columna de cartera vencida de diciembre 2019 para ser predicha, ni ninguna columna o concepto que pueda dañar la inferencia, por ejemplo, no se incluye si pagó o no con mora o conceptos que vienen posterior a la inferencia del pago o atraso, en la vida real, como se explicó en la sección de ***Análisis de características***.

Figura 58

Dataset 2019



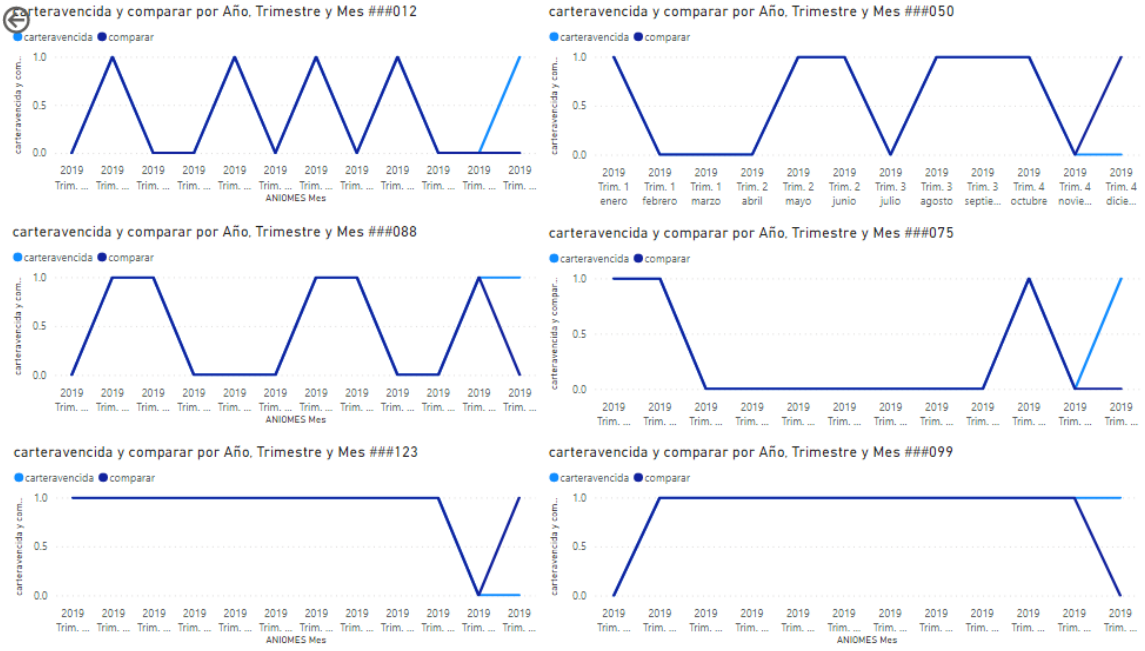
Resultados con data 2019

Dado este primer set de datos, la columna de predicción en referencia al valor real de

cartera vencida de diciembre 2019 tiene un 10% de columnas con error de inferencia, esto va dentro de las métricas obtenidas en el modelo, las cuales, arrojaron una exactitud del 82,9% y una precisión del 85,2%, también el ejemplo de 100 cuentas, coincide con que sus dueños, están dentro del rango de edad, en promedio de 54 años, en cuanto a la mayor cantidad de cartera vencida, similar a lo descubierto en la sección: ***Preparación y exploración de los datos***.

Figura 59

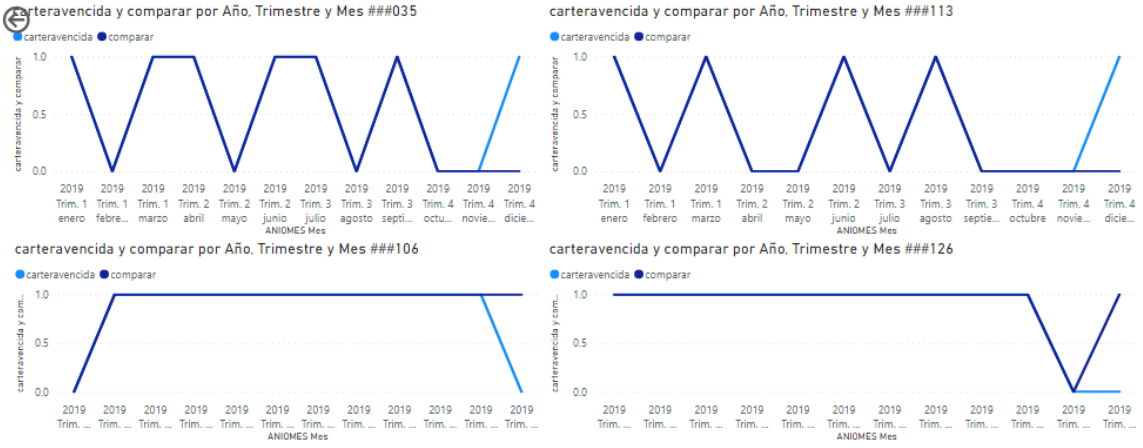
Cuentas con valores predichos errados solo en 2019



Si observamos el promedio de edad en los dueños de cuentas, en los que se equivocó el modelo, también coincide con lo indicado en las características importantes, que la media es de 54 años.

Figura 60

Cuentas con valores predichos errados en 2019 y 2020

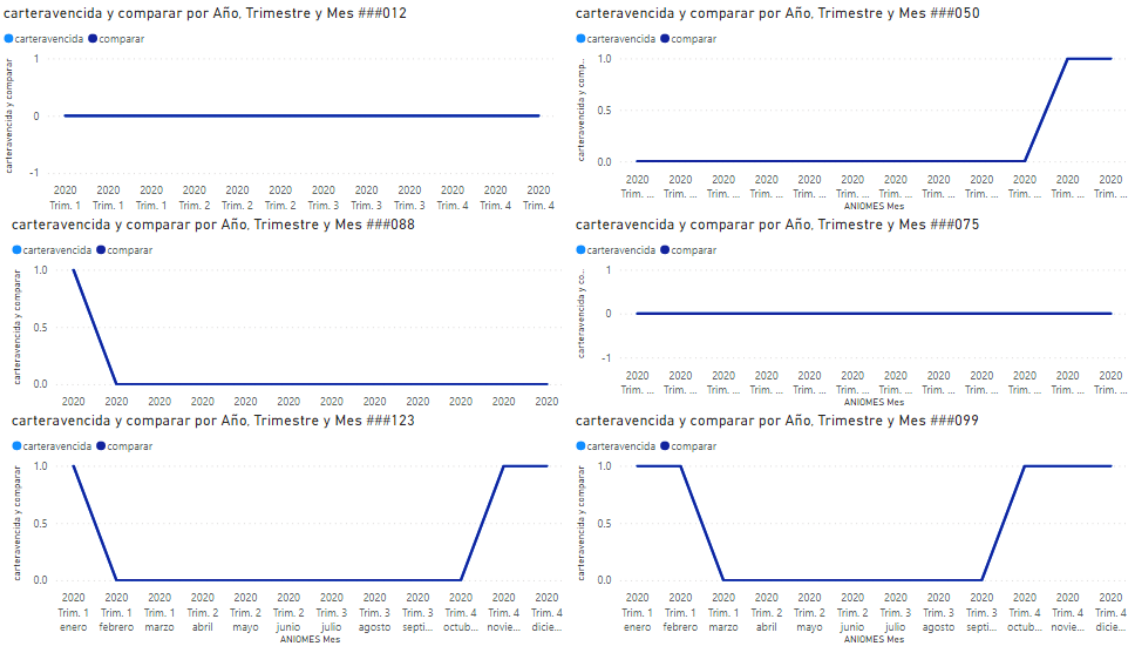


Comparación con data 2020

Posteriormente, se evalúa con las mismas cuentas del primer set de datos de 2019, pero ahora, con la data de enero a noviembre del 2020, para predecir la cartera vencida de diciembre 2020. Se realiza una gráfica similar al 2019 separando las 6 cuentas en las cuales la inferencia no se equivoca en el 2020 y se separan las 4 cuentas que la inferencia vuelve a fallar en el 2020.

Figura 61

Cuentas con valores predichos 2020 que su inferencia fue errada solo en 2019



En el caso de las cuentas ###012, ###050 y ###088 que en el 2019 tuvieron unos meses un pago puntual y otros meses atrasados de forma alternada, en el 2020 las cuentas ###012 y ###088 tienen un acuerdo de pago y pagan siempre puntual, por ende, su tendencia es plana, el caso de la cuenta ###075 es el 2019 era bastante puntual pero unos pocos meses en el 2020 igual accede a un acuerdo de pago mientras que la cuenta ###050 al final se vuelve a atrasar.

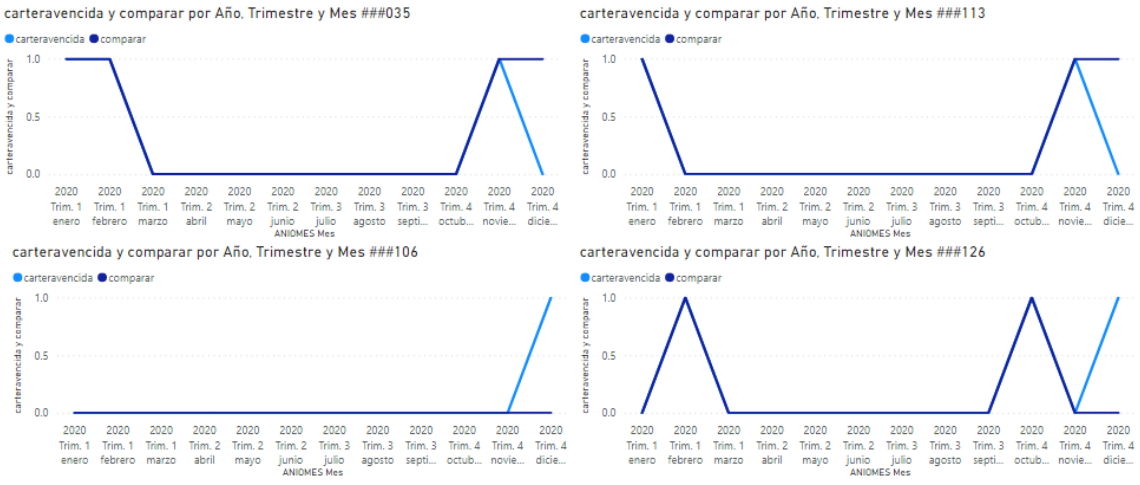
La cuenta ###099 es diferente puesto que tiene una tendencia de estar casi siempre en atrasos en el 2019 pero en el 2020 igual accede a un acuerdo de pago y ya no cae en morosidad, hasta igualarse las deudas acumuladas, pero vuelve a caer en mora de octubre 2020 en adelante, no se nota un incremento importante en el consumo de KWH, por lo que, se podría presumir es un factor externo, sin embargo, esos datos no se manejan a nivel de la empresa eléctrica de Cotopaxi y como, se tiene la historia 2019 donde casi siempre se mantuvo atrasado,

similar caso la cuenta ###123, que tiene un histórico de siempre atrasado en el 2019.

Para todos estos casos, hay que tomar en cuenta el factor externo respecto a la no generación de intereses o mora durante el estado de Excepción, es decir, hasta agosto del 2020 donde el Gobierno ecuatoriano indicó no cobrar mora, posterior a eso en los casos de las cuentas: ###099 y ###123 vuelven a su tendencia de caer siempre en mora, mientras que, la cuenta ###050 regresa a sus caídas en mora eventuales, un comportamiento similar se tiene en las cuentas donde la inferencia volvió a fallar, dichas cuentas son las ###035, ###106, ###113 y ###126, se nota en la gráfica del 2019 la tendencia a caer en mora de las cuentas: ###106 y ###126.

Figura 62

Cuentas con valores predichos errados en 2019 y 2020



Adicional, por temas de la pandemia (Covid 19) al evaluar de enero a noviembre de 2020 e inferir diciembre distintas realidades personales o familiares pudieron cambiar, mismas que no están registradas en la base de datos de la empresa eléctrica de Cotopaxi, también se debe tomar en cuenta la ayuda gubernamental respecto a la no generación de intereses o mora durante el estado de Excepción, por ende, se pasó de un 10% de error en la predicción del 2019

a un 28% con el mismo modelo en el año 2020 con el mismo subconjunto de cuentas, lo cual, dada la eventualidad mundial, es correcto, pues existieron factores externos de los cuales no posee información el modelo.

Partiendo de lo anterior, surge la necesidad en esta clase de proyectos, de implementar lo que indican (Edwards, Franks, & Coulter, 2021), que es, MLOps, es decir, flujos de trabajo (pipelines) o procesos donde el modelo siempre vaya reaprendiendo, ya que, el modelo no puede ser estático, a medida que, nuevos datos y variables van apareciendo, con lo cual, se confirma que los datos no se ajustan al modelo, sino el modelado tiene que ir evolucionando y ajustándose a las nuevas realidades, a pesar de que, puedan ser las mismas variables o características incluso.

Respecto a la pregunta de investigación: OE3 – RQ3.2: ¿El modelo planteado permitirá identificar patrones que faculten generar grupos de consumidores con dimensiones aún no conocidas en la empresa eléctrica de Cotopaxi?, dados los resultados del subconjunto de datos predicho del 2019 y 2020, se puede utilizar como un siguiente paso, técnicas de aprendizaje no supervisado con el objetivo de agrupar las tendencias o comportamiento de pago, en base a las características ya encontradas, por ejemplo, se tienen personas constantes que siempre pagan puntual o siempre pagan con mora, personas que casi siempre pagan puntual o casi siempre pagan con mora y los muy irregulares o variantes que pagan puntuales o con mora de forma alternada. Al llegar a obtener estos grupos o clúster, estos permitirán discernir de una mejor manera estas clasificaciones y otras tendencias aún no visualizadas, con las características ya encontradas en este proyecto de tesis.

Adicionalmente, ya se ha detectado una media de edad del consumidor con mayor

número de cuentas, entre los 54 años, que puede disponer de mayor cantidad de cuentas e incluso algunas de estas, estar usadas bajo arriendo, de ahí que, se pudo detectar que algunas de estas cuentas han caído en morosidad.

Conclusiones

- Si bien se tiene un mayor número de cuentas en el sector rural en Cotopaxi, el sector marginal, es donde se acumula la mayor cantidad de pagos con morosidad, por lo que, el valor por mora, es más elevado debido al número de cuentas en el sector marginal, pero el valor no cobrado, por los pagos tardíos del consumo KWH es insignificante, en comparación a los consumos del sector urbano, este es un hecho empírico, aseverado por la data, en este proyecto de tesis, por lo tanto, lo deberá tomar en cuenta la empresa eléctrica de Cotopaxi, dependiendo de sus objetivos de recuperación de cartera, de donde, se concluye que, tomando en cuenta los valores de consumo e KWH, el dolor más grande es del sector urbano.

- Uno de los patrones detectados que también es un hecho empírico, aseverado por la data, es que el mayor número de cuentas se concentra en los clientes con una media de 54 años y a su vez el valor de mayor morosidad y el valor de mayor consumo de KWH, esto se debe a que usualmente en el sector urbano a esa edad en promedio se puede llegar a tener una casa y dividirla en N departamentos de arriendo, cada una con sus medidores, por lo que, se puede concluir que la metadata faltante dentro de la base de datos es incluir campos que puedan ubicar esta clase de comportamientos, por ejemplo, si dicha cuenta pertenece a una casa o departamento en arriendo o negocio con el establecimiento arrendado incluso.

- Se puede concluir que, el usar la metodología CRISP-DM, permite ese ir y venir de la experimentación con datos, es decir, se puede jugar con la data, por indicar un término adecuado para entender el negocio mediante el análisis de sus datos y posteriormente, ir agregando conceptos o características que complementan las relaciones respecto al

objetivo o Target a inferir, repitiendo así el proceso mediante esta metodología complementada con el ciclo de vida de la analítica de datos y operacionalización de datos, y demostrado así, que, la metodología CRISP-DM, permite abrir los caminos para la implementación de un verdadero MLOps (Machine Learning más DevOps), ya que, todo proceso de mejora es incremental y continuo.

- El modelo planteado dio una precisión comprobada del 85.20%, lo cual, está dentro de los rangos de modelos efectivos, por lo cual, se puede concluir que, es un buen punto de partida el algoritmo de árboles de decisión potenciado, para inferir la cartera vencida dentro de la empresa eléctrica de Cotopaxi, sin descuidar los temas que ya han sido objetos de consultoría y de aviso para esta empresa eléctrica, como lo es, la limpieza de datos, lo cual, optimizará aún más el porcentaje de precisión de este modelo, ya que, las características con un 30% de data nula, por ejemplo, bajaron en su importancia y correlación respecto al target de cartera vencida, pues, esa falta de información hace que la inferencia no tenga siempre dicho sustento.

- Se tuvo un fuerte esfuerzo en las etapas iniciales de este proyecto de tesis, como CRISP-DM lo indica, en el entendimiento del negocio, el entendimiento de la data, su preparación y sobre todo en la limpieza de datos, tanto en temas de exclusiones por data inconsistente, fechas mal configuradas o mal ingresadas, entre otros aspectos, el tener en algunos casos hasta un 30% de datos nulos, por ende, no se debe dejar pasar por alto el hecho de contrarrestar la data que tiene la empresa eléctrica de Cotopaxi con catastros externos y sobre todo, tener conexiones a entes externos como la DINARDAP para completar los datos faltantes, adicional, al control que se pueda incorporar al sistema transaccional para un correcto ingreso de datos, de esta forma se

optimizará la precisión de los modelos.

- De acuerdo a lo analizado en este proyecto de tesis mediante Microsoft Azure Machine Learning, se concluye que, los datos no se ajustan al modelo, sino el modelado tiene que ir evolucionando y ajustándose a los nuevos datos, a pesar de que puedan ser las mismas variables o características incluso, las mismas van creciendo y cambiando en el tiempo.

Recomendaciones

- Se recomienda tomar en cuenta aspectos como MLOps, si bien, el proyecto actual logró obtener los resultados esperados, mediante MLOps, se pueden re usar los pipelines de re entrenamiento e ir mejorando el modelo de forma constante, ya que, los datos van aumentando y evolucionando día a día, y en el caso de la empresa eléctrica de Cotopaxi, se tendrán mayor cantidad de cuentas y casos o tipos de datos incluso.
- Se recomienda optimizar el uso de características (conceptos o columnas), ya que, si bien en este proyecto de tesis se usaron ciertas características mensuales (por 11 meses), se debe generar variables adicionales, como, por ejemplo, número de pagos a tiempo por trimestre, agrupando así a la información, y minimizando el número de columnas, así se optimizará la cantidad de características para exportaciones, por medio de Servicios Web, para un consumo más óptimo del modelo.
- Se pueden implementar modelos de supervivencia o cosecha, en el entendido de ir tomando en cuenta la fecha de ingreso de la cuenta y el acumulado de deuda y mora total a la fecha, obviamente este tipo de aspectos podrían requerir no solo su

implementación a nivel QVDs de QlikSense sino ya de un DataWareHouse propiamente, ya que, se podría mantener el histórico de otras características, como por cambio de estado civil, con el paso del cliente de soltero a casado u otros, pero actualmente el modelo QVD en QlikSense solo dispone del último estado civil del cliente, sin conocer cuando lo cambió, por ende, también sería importante mantener una comunicación (consumo de Web Services) con estatales externas como la DINARDAP.

- Un siguiente paso es usar técnicas de aprendizaje no supervisado para agrupar las tendencias o comportamientos de pago, por ejemplo, se tienen personas constantes, personas que casi siempre pagan puntual o casi siempre pagan con mora y los muy irregulares o variantes, estos grupos permitirán discernir más estas clasificaciones y otras tendencias con características intrínsecas, aún no visualizadas mediante los modelos supervisados que se usaron en este proyecto de tesis.

- En la competencia de modelos se usó en este proyecto de tesis ANN (Red neuronal artificial), pero no se ha usado DNN (deep neural networks), por lo que, se podría comparar si estos algoritmos tienen incluso mejores resultados que los árboles de decisión potenciados que fueron los mejores con la data actual analizada en este proyecto de tesis.

- Se recomienda cambiar dentro de la empresa eléctrica de Cotopaxi, el concepto de visitas o revisiones hacia todos los medidores de forma indiscriminada, sino más bien usar el modelo planteado de árboles de decisión potenciado, para visitar a los medidores objetivo, es decir, con un enfoque directo, detectando mediante el algoritmo indicado en este proyecto de tesis, las cuentas que estarán como posibles a

caer en mora, con al menos un mes de anterioridad al evento.

- Se recomienda a la empresa eléctrica de Cotopaxi, repotenciar el actual QlikSense con un contrato de soporte y una actualización para poder usar el componente de Qlik Connectors que permitirá complementar sus actuales esfuerzos con Machine Learning y nube, caso contrario, podría optar por herramientas como Microsoft Azure ML, la cual, posee mayores réditos en cuanto a uso de infraestructura nube, tanto volátil en ambientes de pruebas, como escalable en ambientes de producción, para realizar de una forma más eficiente los análisis de datos.

Bibliografía

Information Matters. (2021, Marzo 4). *data-driven innovation in the UK*. Retrieved from Information Matters: https://informationmatters.net/gartner-2021-magic-quadrant-data-science-machine-learning/

Abakarim, Y., Lahby, M., & Attioui, A. (2018). Towards An Efficient Real-time Approach to Loan Credit Approval Using Deep Learning. *9th International Symposium on Signal, Image, Video and Communications, ISIVC 2018 - Proceedings*, 306-313.

Almonayirie, W. E. (2015). An application of 'Neuro-Logit' new modeling tool in corporate financial distress diagnostic. *IEOM 2015 - 5th International Conference on Industrial Engineering and Operations Management, Proceeding*, 2-7.

Alvarado, M., & F, M. (2017). Pronóstico del tipo de cambio USD/MXN con redes neuronales de retroprogamación. *ISSN 1870-4069* (pp. 1-15). México: https://www.researchgate.net/publication/323985249.

ARCONEL. (2020, Junio 19). *regulaciones*. Retrieved from regulacionelectrica: https://www.regulacionelectrica.gob.ec/wp-content/uploads/downloads/2020/06/Reg-Sust-Reg-ARCONEL001-20-Directorio-res-006-20-firm.pdf

Azure Machine Learning. (2019). *Learn*. Retrieved from Create an inference pipeline: https://docs.microsoft.com/en-us/learn/modules/create-classification-model-azure-machine-learning-designer/inference-pipeline

Baccam, N., Sharkey, K., Botsikas, A., & Gronlund, C. (2020, Diciembre 18). *Data featurization in automated machine learning*. Retrieved from Azure Machine Learning: https://docs.microsoft.com/en-us/azure/machine-learning/how-to-configure-auto-features#scaling-and-normalization

CGE. (2021, Abril 12). *CentroSur*. Retrieved from CONTRALORIA GENERAL DEL ESTADO: https://www.centrosur.gob.ec/wp-content/uploads/2021/04/Informe-CGE-DPA-0032-2019.pdf

Edwards, J., Franks, L., & Coulter, D. (2021, Agosto 07). *MLOps: Model management, deployment, lineage and monitoring with Azure Machine Learning*. Retrieved from Microsoft Azure: https://docs.microsoft.com/en-us/azure/machine-learning/concept-model-management-and-deployment

Edwards, J., Gilley, S., Gronlund, C., & Coulter, D. (2021, Abril 05). *API design for machine learning software: experiences from the scikit-learn project*. Retrieved from Microsoft Azure Machine Learning: https://docs.microsoft.com/en-us/azure/machine-learning/how-to-train-scikit-learn

ELEPCO S.A. (2019, Enero). *elepcosa.* Retrieved from Plan-Estratégico-2018-2021: https://elepcosa.com.ec/wp-content/uploads/2019/01/Plan-Estrat%C3%A9gico-2018-2021.pdf

Garzón Ulloa, P. A., & Chicaiza Castillo, D. V. (2017, Agosto). *Desarrollo de una Plataforma Web de Información Gerencial para la Gestión Administrativa de una Empresa Distribuidora del Sector Eléctrico.* Ambato: PUCESA. Retrieved from Repositorio PUCESA: https://repositorio.pucesa.edu.ec/handle/123456789/2003

Gregory, C., Quintanilla, L., Cartacio, S., & Gronlund, C. (2020, Diciembre 9). *Evaluate automated machine learning experiment results*. Retrieved from Azure Machine Learning: https://docs.microsoft.com/en-us/azure/machine-learning/how-to-understand-automated-ml

grupo-novatech. (2021). *grupo-novatech*. Retrieved from qlik-vs-power-bi: https://www.grupo-novatech.com/qlik-vs-power-bi/

Health Data Miner. (2021). *IA*. Retrieved from Health Data Miner: https://healthdataminer.com/data-mining/crisp-dm-una-metodologia-para-mineria-de-datos-en-salud/

Herrera, I. (2019, Octubre 4). *SAS*. Retrieved from Blogs : https://blogs.sas.com/content/sasla/2019/10/04/operacionalizacion-de-la-analitica-como-lograr-que-los-modelos-analiticos-realmente-apoyen-el-exito-de-los-negocios/

Ilyas, I. F., & Chu, X. (2019). *Data Cleaning.* Waterloo: ACM Books #28.

Ippolito, P. P. (2019, Junio 3). *SVM: Feature Selection and Kernels*. Retrieved from towards datas cience: https://towardsdatascience.com/svm-feature-selection-and-kernels-840781cc1a6c

Joanybel, O. (2020, Marzo 22). *El Machine Learning y sus aplicaciones*. Retrieved from JoanYBel: http://www.joanybelortiz.com/aplicaciones-machine-learning-ejemplos/

Joaquín, A. R. (2020, Agosto). *Regresión logística simple y múltiple*. Retrieved from Ciencia de datos: https://www.cienciadedatos.net/documentos/27_regresion_logistica_simple_y_multiple

Joaquín, A. R. (2021, Enero). *Ajuste de una distribución*. Retrieved from Ciencia de datos: https://www.cienciadedatos.net/documentos/pystats01-ajuste-distribuciones-python.html

Kemp, S., Gilley, S., & Maggie, S. (2021, Febrero 17). *Power BI*. Retrieved from Microsoft: https://docs.microsoft.com/en-us/power-bi/connect-data/service-aml-integrate

La Gaceta. (2021, Septiembre 15). *Elepco SA tiene una cartera vencida de más de USD 8 millones*. Retrieved from la Gaceta: https://lagaceta.com.ec/elepco-sa-tiene-una-cartera-vencida-de-mas-de-usd-8-millones/

Li, B., & Lu, P. (2020, Febrero 11). *Cross Validate Model*. Retrieved from Azure Machine Learning: https://docs.microsoft.com/en-us/azure/machine-learning/algorithm-module-reference/cross-validate-model

Li, B., & Lu, P. (2020, Octubre 10). *Filter Based Feature Selection*. Retrieved from Azure Machine Learning: https://docs.microsoft.com/en-us/azure/machine-learning/algorithm-module-reference/filter-based-feature-selection

Li, B., & Lu, P. (2020, Octubre 13). *Group Data into Bins module*. Retrieved from Azure Machine Learning: https://docs.microsoft.com/en-us/azure/machine-learning/algorithm-module-reference/group-data-into-bins

Li, B., & Lu, P. (2020, Octubre 10). *Tune Model Hyperparameters*. Retrieved from Azure Machine Learning: https://docs.microsoft.com/en-us/azure/machine-learning/algorithm-module-reference/tune-model-hyperparameters

Li, B., & Lu, P. (2021, Abril 11). *Convert to Indicator Values*. Retrieved from Microsoft Azure Machine Learning: https://docs.microsoft.com/en-us/azure/machine-learning/component-reference/convert-to-indicator-values

Li, B., Lu, P., & Baccam, N. (2020, Abril 22). *Two-Class Neural Network module*. Retrieved from Azure Machine Learning: https://docs.microsoft.com/en-us/azure/machine-learning/algorithm-module-reference/two-class-neural-network

Li, B., Lu, P., Baccam, N., & JiayueHu. (2020, Septiembre 03). *Two-Class Decision Forest module*. Retrieved from Azure Machine Learning: https://docs.microsoft.com/en-us/azure/machine-learning/algorithm-module-reference/two-class-decision-forest

Li, B., Lu, P., Coulter, D., Baccam, N., & Gilley, S. (2020, Agosto 24). *Two-Class Boosted Decision Tree module*. Retrieved from Azure Machine Learning: https://docs.microsoft.com/en-us/azure/machine-learning/algorithm-module-reference/two-class-boosted-decision-tree

Li, B., Lu, P., Coulter, D., Baccam, N., & Gilley, S. (2020, Marzo 22). *Two-Class Logistic Regression module*. Retrieved from Azure Machine Learning: https://docs.microsoft.com/en-us/azure/machine-learning/algorithm-module-reference/two-class-logistic-regression

Li, B., Lu, P., Coulter, D., Baccam, N., & Gilley, S. (2020, Abril 22). *Two-Class Support Vector Machine module*. Retrieved from Azure ML: https://docs.microsoft.com/en-us/azure/machine-learning/algorithm-module-reference/two-class-support-vector-machine

Li, B., Lu, P., Simpson, D., & Baccam, N. (2020, Abril 22). *Two-Class Averaged Perceptron module*. Retrieved from Azure Machine Learning: https://docs.microsoft.com/en-us/azure/machine-learning/algorithm-module-reference/two-class-averaged-perceptron

Li, B., O'Brien, L., Gronlund, C. L., & Coulter, D. (2021, Febrero 5). *Run batch predictions using Azure Machine Learning designer*. Retrieved from Azure Machine Learning: https://docs.microsoft.com/en-us/azure/machine-learning/how-to-run-batch-predictions-designer

Li, W., Ding, S., Chen, Y., & Yang, S. (2018). Heterogeneous ensemble for default prediction of peer-to-peer lending in China. *IEEE Access*, 54396-54406.

MEER. (2018). Convenio Interinstitucional para la Adquisición e Implementación de los Productos CIS, MDM y CRM. *Convenio Interinstitucional* (p. 1). Quito: MEER.

Microsoft. (2019). *Create an inference pipeline*. Retrieved from Microsoft Learn: https://docs.microsoft.com/en-us/learn/modules/create-classification-model-azure-machine-learning-designer/inference-pipeline

Microsoft Learn. (2019). *Deploy a predictive service*. Retrieved from Microsoft Learn: https://docs.microsoft.com/en-us/learn/modules/create-classification-model-azure-machine-learning-designer/deploy-service

MS Azure ML. (2019). *Learn*. Retrieved from Evaluate a classification model: https://docs.microsoft.com/en-us/learn/modules/create-classification-model-azure-machine-learning-designer/evaluate-model

Mungi, R., Gronlund, C., & Hansen, d. (2021, Marzo 08). *What is Azure Machine Learning?* Retrieved from Azure Machine Learning: https://docs.microsoft.com/en-us/azure/machine-learning/overview-what-is-azure-ml

Okesola, O. J., Okokpujie, K. O., Adewale, A. A., John, S. N., & Omoruyi, O. (2018). An Improved Bank Credit Scoring Model: A Naïve Bayesian Approach. *Proceedings - 2017 International Conference on Computational Science and Computational Intelligence, CSCI 2017*, 228-233.

Pawluszek-Filipiak, K., & Borkowski, A. (2020). On the importance of train-test split ratio of datasets in automatic landslide detection by supervised classification. *Remote Sensing*, 2-33.

R, S. E. (2021, Junio). *Understanding Random Forest*. Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

Rocca, J. (2019, Abril 22). *Ensemble methods: bagging, boosting and stacking*. Retrieved from towards data science: https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205

Rodríguez, A., & Gamboa, E. (2019, Junio 24). *analiticamineria*. Retrieved from Pistas para su abordaje en la era de la sobreinformación: https://estadisticaun.github.io/L_Conceptual/2-3-analiticamineria-de-datos-analytics.html

Schorr, F., & Hvam, L. (2018). Design Science Research: A Suitable Approach to Scope and Research IT Service Catalogs. *EEE World Congress on Services (SERVICES)* (pp. 25-26). Banff: https://ieeexplore.ieee.org/document/8495778/.

Schröer, C., Felix, K., & Gómez, M. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. *CENTERIS - International Conference on Project MANagement / HCist - International Confer* (pp. 1-9). Wolfsburg: https://www.researchgate.net/publication/349527794_A_Systematic_Literature_Review_on_Applying_CRISP-DM_Process_Model.

Shao, M., Smonou, D., Kampouridis, M., & Tsang, E. (2014). Guided Fast Local Search for speeding up a financial forecasting algorithm. *IEEE/IAFE Conference on Computational Intelligence for Financial Engineering, Proceedings (CIFEr)*, 325-332.

Umaquinga Criollo, A. C., Suárez Zambrano, L. E., & Oña Rocha, O. R. (2018). El aprendizaje automático: Importancia, avances, técnicas y aplicaciones. Machine learning: Importance, advances, techniques and applications. *ISBN 978-958-8958-65-1 El papel de la Ingeniería en tiempos de construcción de paz y buen vivir* (pp. 176-186). Pasto: researchgate.net.

Vasilev, I. (2017). *Un Tutorial de Aprendizaje Profundo: De Perceptrones a Redes Profundas*. Retrieved from toptal: https://www.toptal.com/machine-learning/un-tutorial-de-aprendizaje-profundo-de-perceptrones-a-redes-profundas

Wei, S., Yang, D., Zhang, W., & Zhang, S. (2019). A novel noise-adapted two-layer ensemble model for credit scoring based on backflow learning. *IEEE Access*, 99217-99230.

Zhang, X., Gronlund, C., & Jason, H. (2019, Junio 05). *Principal Component Analysis*. Retrieved from Microsoft Azure: https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/principal-component-analysis

Zhang, X., Gronlund, C., & Lu, P. (2020, Mayo 3). *Split Data using Split Rows*. Retrieved from Azure Machine Learning: https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/split-data-using-split-rows

Zhang, X., Howell, J., Gronlund, C., Lu, P., & Gilley, S. (2019, Mayo 6). *Normalize Data*. Retrieved from Azure Machine Learning: https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/normalize-data

Zhao, K. (2016, Enero 12). *Permutation Feature Importance*. Retrieved from Azure AI Gallery: https://gallery.azure.ai/Experiment/Permutation-Feature-Importance-5