APLICACIÓN DE CIENCIA DE DATOS PARA PROYECCIÓN DE SALDOS DE PRODUCTOS DE CAPTACIONES EN ENTIDAD BANCARIA_

Carlos Alberto León Gil y Mauricio Pinzón Cortes

Nota de Aceptación

Certificamos que el presente Trabajo de Grado Satisface, en alcances y calidad, todos los requisitos que demanda un Trabajo de Grado de Maestría.

David Arango Londoño
Director

Jurado

Daniel Enrique González G.
Jurado

Aprobado en cumplimiento de los requisitos exigidos por la Pontificia Universidad Javeriana Cali, para optar el título de Magister en Ciencia de Datos.

Camilo Rocha
HERNÁN CAMILO ROCHA NIÑO Ph. D.
Decano Facultad de Ingeniería y Ciencias

JUAN CARLOS MARTÍNEZ ARIAS
Director Posgrados de Ingeniería y Ciencias

Santiago de Cali, julio de 2023

Acta de Correcciones al Documento de Trabajo de Grado

Santiago de Cali, julio de 2023

Autor: Carlos Alberto León Gil y Mauricio Pinzón Cortes

Título del Trabajo de Grado: "APLICACIÓN DE CIENCIA DE DATOS PARA PROYECCIÓN DE SALDOS DE PRODUCTOS DE CAPTACIONES EN ENTIDAD BANCARIA"

Director: David Arango Londoño

Como indica el artículo 2.13 de las Directrices para Trabajo de Grado de Maestría, he verificado que el estudiante indicado arriba ha implementado todas las correcciones que los Jurados del Proyecto de Trabajo de Grado definieron que se efectuaran, como consta en el Acta de Evaluación correspondiente.

David Arango Londoño

Firma del Director del Trabajo de Grado

Santiago de Cali, 31 de julio del 2023

Doctora
Gloría Inés Alvarez V.
Directora Maestría en Ciencia de Datos
Facultad de Ingeniería y Ciencias
Pontificia Universidad Javeriana de Cali

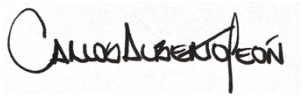
Asunto: Presentación para evaluación del proyecto aplicado

Cordial Saludo,

Con el fin de cumplir con los requisitos exigidos por la Universidad para optar por el título de Magíster en Ciencia de Datos, nos permitimos presentar a su consideración el proyecto denominado "APLICACIÓN DE CIENCIA DE DATOS PARA PROYECCIÓN DE SALDOS DE PRODUCTOS DE CAPTACIONES EN ENTIDAD BANCARIA", el cual fue realizado por los estudiantes Carlos Alberto León Gil y Mauricio Pinzón Cortes con código (s) 8972921 y 8971720 pertenecientes a la Maestría en Ciencia de Datos, bajo la dirección de David Arango Londoño.

El suscrito director del Proyecto Aplicado autoriza para que se proceda a hacer la evaluación de este proyecto, toda vez que ha revisado cuidadosamente el documento y avala que ya se encuentra listo para ser presentado y sustentado oficialmente.

Atentamente,



Carlos Alberto León Gil	Mauricio Pinzón Cortes	David Arango Londoño
C.C. 1.015.454.074 de Bogotá	C.C. 79.884.948 de Bogotá	C.C. 1.130.586.950 de Cali

Documentación anexa:
Resumen del Proyecto Aplicado en formato digital (máximo 1 página).
Una copia digital (PDF) del documento del proyecto aplicado

FICHA RESUMEN PROYECTO DE TRABAJO DE GRADO

POSIBLE TÍTULO: APLICACIÓN DE CIENCIA DE DATOS PARA PROYECCIÓN DE SALDOS DE PRODUCTOS DE CAPTACIONES EN ENTIDAD BANCARIA

1. ÁREA DE TRABAJO: Financiero
2. TIPO DE PROYECTO (Aplicado, Innovación, Investigación): Aplicado
3. ESTUDIANTE(S): Carlos Alberto León Gil y Mauricio Pinzón Cortes
4. CORREO ELECTRÓNICO: cleon00@javerianacali.edu.co, mpinzonc@javerianacali.edu.co
5. DIRECCIÓN Y TELEFONO: Calle 36 F Sur No 3B - 74 Este, Bogotá D.C Tel: (+57) 313 369 42 23, Calle 7 A Bis C No 80ª – 50 Apto: A 203, Bogotá D.C Tel: (+57) 313 629 08 90
6. DIRECTOR: David Arango Londoño
7. VINCULACIÓN DEL DIRECTOR: Temporal
8. CORREO ELECTRÓNICO DEL DIRECTOR: david.arango@javerianacali.edu.co
9. CO-DIRECTOR (Si aplica): NO
10. GRUPO O EMPRESA QUE LO AVALA (Si aplica): ENTIDAD BANCARIA
11. OTROS GRUPOS O EMPRESAS:
12. PALABRAS CLAVE (al menos 5): Ciencia de datos, Machine Learning, Previsiones, Saldos de portafolio
13. FECHA DE INICIO: Junio del 2022
14. DURACIÓN ESTIMADA (En meses): Doce meses
15. RESUMEN:

Los datos son el insumo principal de un proyecto de ciencia de datos y a su vez hoy día son el activo más importante que se tiene en cualquier sector. Los resultados de la aplicación de técnicas de ciencia de datos para obtener valor y conocimiento, permiten la mejora continua en el proceso de toma de decisiones generando valor a nivel del negocio. Actualmente en el entorno financiero, se hace necesario hacer uso de la información para la toma de decisiones de una manera más eficiente y oportuna, no solo por buenas prácticas o temas de moda sino por supervivencia. En este sentido tener la mayor cantidad de información para la toma de decisiones hace que los modelos predictivos tengan bastante relevancia.

Actualmente no se tiene definido un modelo de predicción de saldos de productos de captaciones para cuentas de ahorros y cuentas corrientes, el cual se hace necesario para poder generar estrategias en pro del mantenimiento o aumento de los saldos, con el fin de garantizar que exista el capital para realizar colocaciones y aumentar la utilidad neta del negocio.

APLICACIÓN DE CIENCIA DE DATOS PARA PROYECCIÓN DE SALDOS DE PRODUCTOS DE CAPTACIONES EN ENTIDAD BANCARIA

Carlos Alberto León Gil
Código 8972921
Mauricio Pinzón Cortes
Código 8971720

Proyecto Aplicado para optar al título de
Magister en Ciencia de Datos

Director
David Arango Londoño

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, MAYO DE 2023

TABLA DE CONTENIDO

INTRODUCCIÓN

Los datos día tras día se han vuelto una forma de aprendizaje continuo para mejorar la toma de decisiones en los diferentes sectores económicos, tal es el caso, que las entidades bancarias pueden aprender de sus clientes a partir de los datos históricos y comportamiento que se han tenido a lo largo del tiempo. Por lo cual, el análisis de los datos para este sector se puede basar en los patrones que se establecen para cierto segmento de clientes según la finalidad del negocio en la operación, por ende, la extracción de estos puede otorgar conocimiento útil al banco para mejorar sus procesos y mantener un perfilamiento del cliente; es así que, se podrían conectar los modelos predictivos, pues de acuerdo con estos modelos "tienen como principal objetivo aproximar posibles valores del futuro o desconocidos a través de los datos de los que ya se dispone".

Por lo anterior, dichos modelos pueden usar los métodos de clasificación y regresión para mejorar la exactitud con la que se miden los productos de captación que permiten obtener a las entidades bancarias recursos del público, de modo que este proyecto, pretende lograr un impacto en el desarrollo, análisis y ejecución del manejo de proyección de saldos de ciertos productos de captación en la entidad bancaria, pues al realizar el estudio de las necesidades puntuales del negocio se pueden definir las fuentes de información que se manejan en las bases de datos del banco y así, aplicar reglas de calidad en la transformación de los datos según su estructura, además manteniendo la capacidad de correr este modelo de acuerdo con la infraestructura y capacidad de las máquinas y las aplicaciones para obtener los resultados esperados para las elecciones a realizar en la mejora de procesos de captación de clientes.

El análisis, modelamiento y evaluación del modelo de datos, permite resaltar de manera pertinente los beneficios económicos al mejorar la toma de decisiones con el manejo de campañas según el perfilamiento del cliente con la predicción de sus necesidades. Finalmente, la visualización de los resultados del modelo al momento de su ejecución se podrá revisar en un tablero de control para lograr su lectura y procesamiento de manera eficaz, contando esta implementación como cambio significativo, pues ayuda a orientar de manera adecuada las ofertas de productos y aumenta la fidelización de los clientes que usan productos de captación con la entidad bancaria.

1. DEFINICIÓN DEL PROBLEMA

1.1. PLANTEAMIENTO DEL PROBLEMA

Los datos son el insumo principal de un proyecto de ciencia de datos y a su vez hoy día son el activo más importante que se tiene en cualquier sector. Los resultados de la aplicación de técnicas de ciencia de datos para obtener valor y conocimiento permiten la mejora continua en el proceso de toma de decisiones generando valor a nivel del negocio. Actualmente en el entorno financiero, se hace necesario hacer uso de la información para la toma de decisiones de una manera más eficiente y oportuna, no solo por buenas prácticas o temas de moda sino por supervivencia. En este sentido tener la mayor cantidad de información para la toma de decisiones hace que los modelos predictivos tengan bastante relevancia.

Actualmente no se tiene definido un modelo de predicción de saldos de productos de captaciones para cuentas de ahorros y cuentas corrientes, el cual se hace necesario para poder generar estrategias en pro del mantenimiento o aumento de los saldos, con el fin de garantizar que exista el capital para realizar colocaciones y aumentar la utilidad neta del negocio.

1.2. FORMULACIÓN DEL PROBLEMA

¿Cómo predecir los valores de los saldos de productos de captaciones en la entidad bancaria mediante técnicas de ciencia de datos?

2. OBJETIVOS DEL PROYECTO

2.1. OBJETIVO GENERAL

Crear un modelo predictivo en un segmento específico de clientes, por medio de la aplicación de métodos y técnicas de ciencia de datos, el cual permita determinar la proyección de saldos de productos de captaciones de una entidad bancaria, con el propósito de desarrollar estrategias para el mantenimiento y/o aumento de los saldos.

2.2. OBJETIVOS ESPECÍFICOS

I. Definir las fuentes de información mediante los datos suministrados ir la compañía, manteniendo un enfoque analítico.
II. Verificar y comprender los datos para aplicar reglas de calidad y preparación de los datos.
III. Analizar y seleccionar los métodos y técnicas que permitan garantizar el desarrollo, análisis y evaluación del modelo predictivo.
IV. Utilizar una herramienta de visualización para evidenciar y exponer los resultados del modelo predictivo a la entidad bancaria.

2.3. RESULTADOS ESPERADOS

I. Diseño, creación y desarrollo de EL - ETL para extracción, cargue y transformación de los datos, con el fin de contar con la disposición de los datos y creación de la base de datos para el cargue de la información, este como insumo y origen de datos para el modelo.
II. Testeo, análisis y selección de métodos como la metodología CRISP-DM y creación de seis algoritmos de predicción y pronósticos *(series de tiempo)* en total con el más alto grado de confiabilidad.
III. Gestión estratégica de la información con los resultados del modelo predictivo en la optimización de toma de decisiones de la entidad bancaria.
IV. Desarrollo e implementación de DASHBOARD en la herramienta Power BI para mostrar los resultados del modelo, logrando hacer uso continuo por parte de la compañía.

V. Logro de gestión del conocimiento, al dejar la documentación del modelo implementado en el proyecto para su entendimiento y usabilidad.

2.4. METODOLOGÍA Y GESTIÓN DEL TIEMPO

En la investigación se tuvo en cuenta técnicas e instrumentos para la colección de datos, basados en la metodología CRISP-DM [27] *(Ilustración 1)* con la cual se desarrollarán las cuatro fases del proyecto planteadas, estas permitieron obtener información general de la problemática que se aborda en el presente proyecto aplicado y la validación luego de poner en práctica la metodología, siguiendo esto, el proyecto se desarrolla por fases, cuatro (4) en total, que se describen en la Ilustración 2:

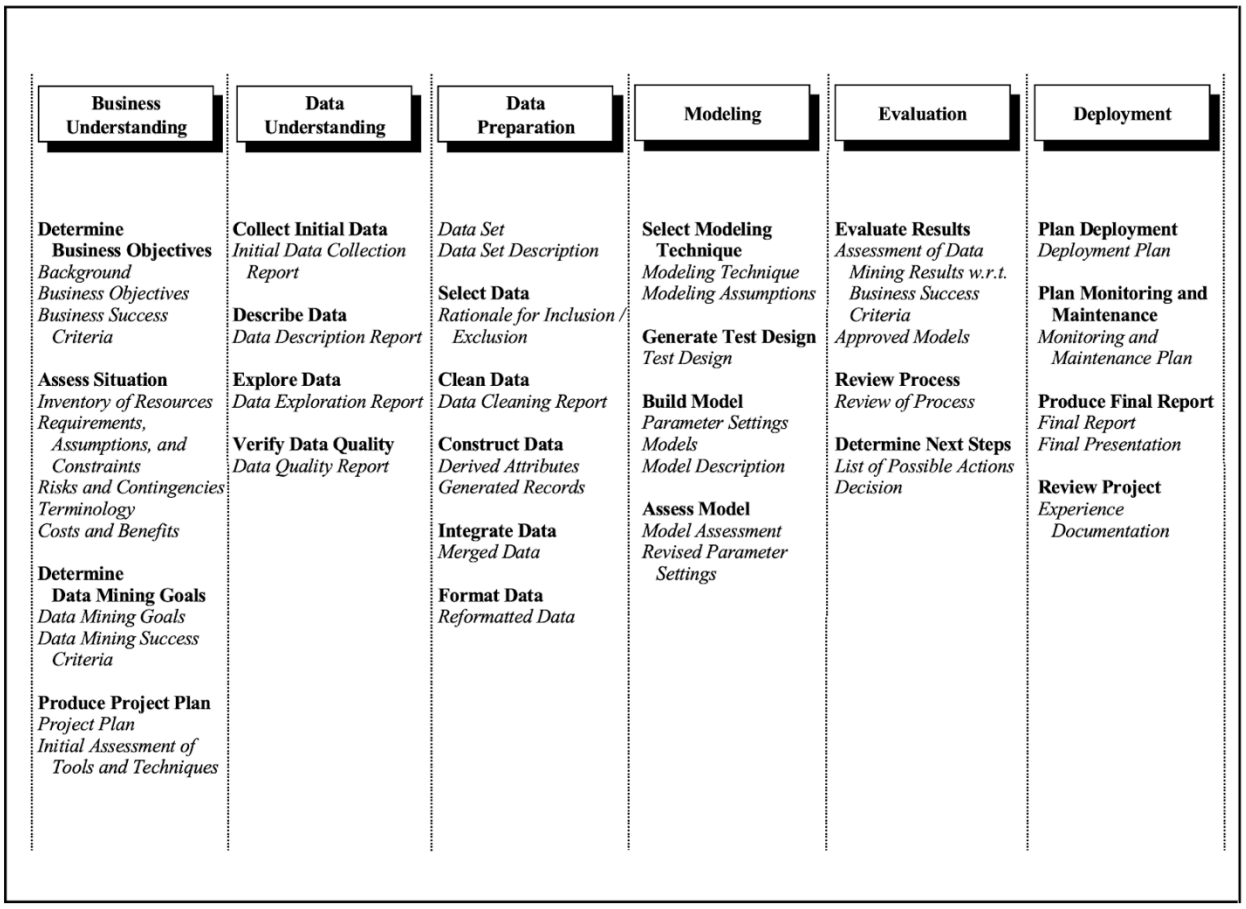


Ilustración 1: Metodología CRISP-DM - Descripción general de las tareas de CRISP-DM y sus resultados. Fuente: Wirth, R., & Hipp, J. (2000, April). – pág. 6

- La primera fase comprende la recopilación de la información para definir del problema, conociendo y entendiendo las necesidades del negocio *(Business Understanding)*, revisión documental y estadística para el manejo de criterios básicos en el enfoque de analítica en el cual se centrará el modelo.

- La segunda fase permite lograr la recolección y alistamiento de los datos, es decir, se realiza el estudio y comprensión de los datos *(Data Understanding)* y análisis de los datos para seleccionar las características de los datos *(Data Preparation)*, en pro de la elección de las variables a utilizar en el modelo de predicción como solución del proyecto aplicado.

- La tercera fase tiene el alcance principal, pues al momento de realizar el diseño del modelo, se verifican las variables con las cuales se construirá el modelo *(Modeling)*, a partir de las variables según los parámetros establecidos y tipos de datos que generan un nivel de significancia y confiabilidad de los resultados, por ende, en este apartado se probará el modelo *(Evaluation)* para evaluar que los resultados ayuden y generen el apoyo que se espera para tomar las decisiones de manera óptima basados en datos y estadísticas.

- Finalmente, una cuarta fase, dónde se realiza la validación del modelo *(Deployment)* según chequeos de los criterios de diseño y modelado de predicción, se espera concluir con la comprobación de expertos en el área de estudio tratada según los resultados obtenidos en el proceso.

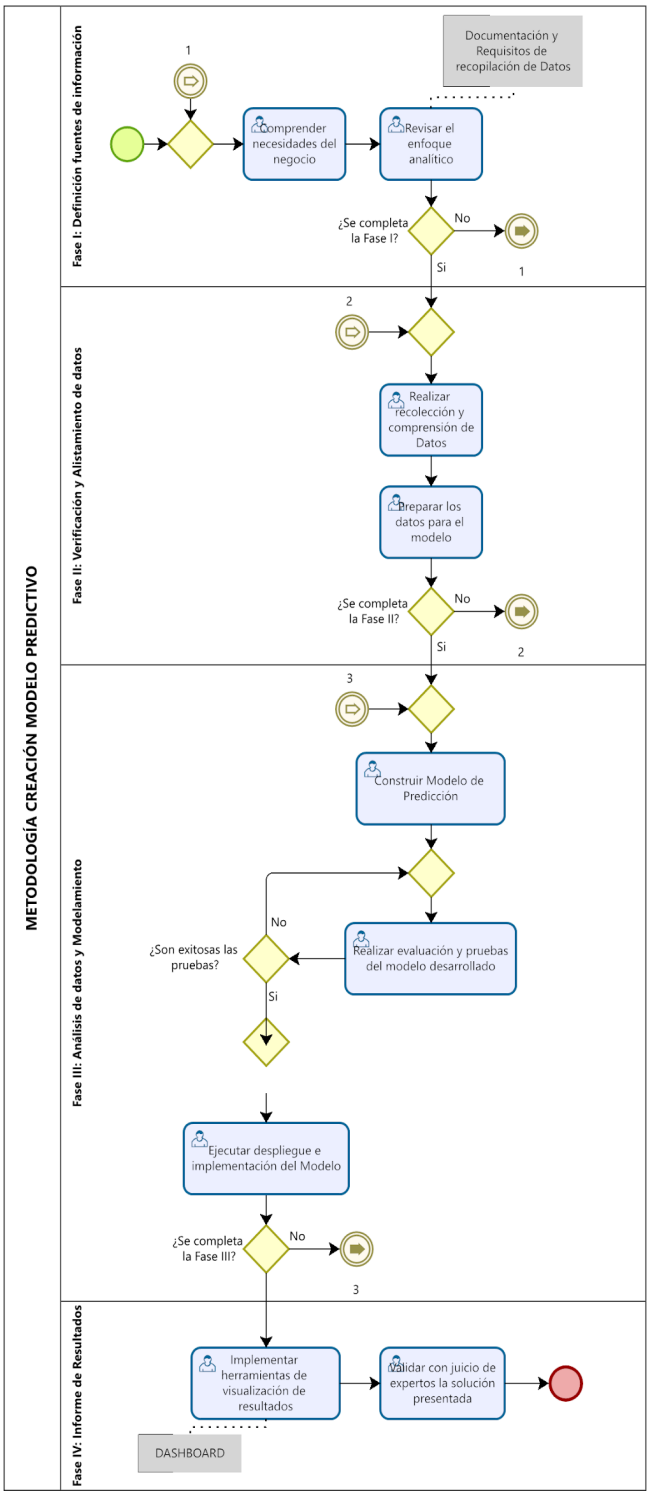


Ilustración 2: Diagrama de proceso para las fases de la metodología. Fuente: Autores

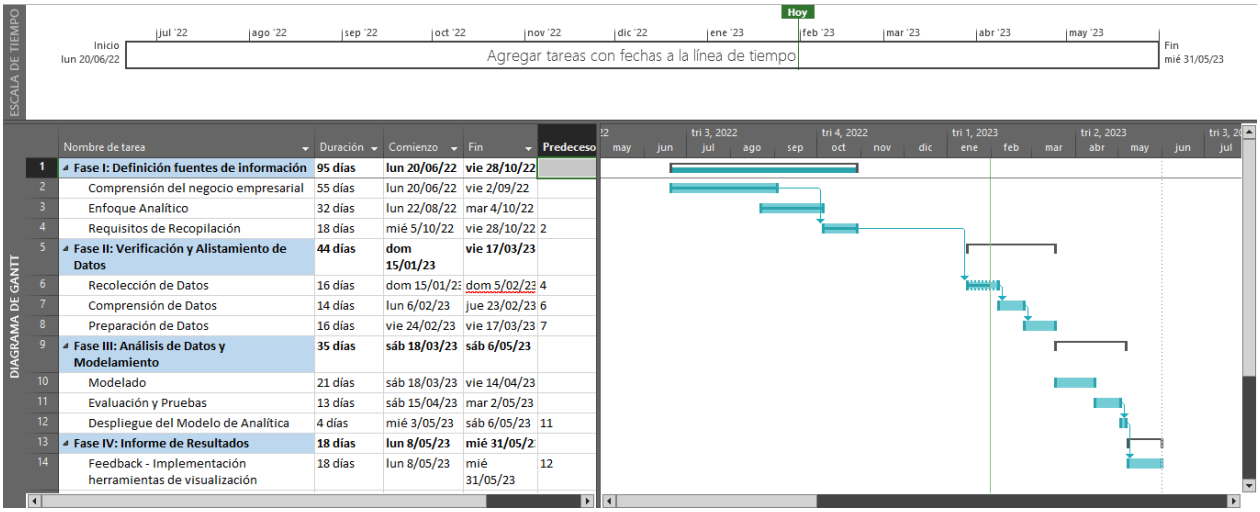


Ilustración 3: Cronograma de actividades del proyecto. Fuente: Autores.

3. MARCO DE REFERENCIA

3.1. MARCO TEÓRICO

En la toma de decisiones permanentemente se elaboran planes para el futuro. Entonces, los datos que describen las situaciones deben ser lo suficientemente acertados, precisos y con un nivel de error muy mínimo para efectuar decisiones tácticas y de gran valor, con el cual se permitan desarrollar procesos y negociaciones estratégicas. En la actualidad y como se menciona en la revista internacional de previsión en su edición de febrero del 2022, los métodos de aprendizaje automático están ganando gran popularidad en el campo de las previsiones, ya que han mostrado gran rendimiento empírico en las recientes competiciones M4 [28] en el cual (i) se aumenta significativamente el número de series, (ii) se amplía el número de métodos de previsión y (iii) se incluyen intervalos de predicción en el proceso de evaluación. Por otro lado M5 [2] se centra en la identificación del método o métodos más adecuados para diferentes tipos de situaciones que requieren predicciones y hacer estimaciones de incertidumbre.

Día a día las organizaciones generan gran cantidad de datos que requiere ser analizadas para descubrir, analizar y desarrollar estrategias cada vez más rigurosas y específicas con el cual se impacte todos los ejes de la organización. Así es que día a día se han desarrollado diferentes técnicas para abordar los casos de uso de las organizaciones una de ellas está enfocada con el aprendizaje automático, el cual consiste en una seria de técnicas estadísticas implementadas para la resolución de problemas, en los cuales los modelos estadísticos desarrollados tienen la capacidad de aprender, identificando patrones y analizar datos históricos para producir futuros estimados.

3.1.1. CONCEPCIONES Y DEFINICIONES DE LOS MODELOS PREDICTIVOS

Los modelos predictivos [3] se emplean con el objetivo de predecir un comportamiento que no se ha probado, todo esto enmarcado dentro de la primera cultura del uso de la modelización estadística el cual se centra en la construcción de algoritmos eficientes para obtener buenos modelos predictivos para pronosticar el futuro [4]. El análisis predictivo considera varias dimensiones y variedad de herramientas de creación de modelos los cuales normalmente

contienen una amplia variedad de variables explicativas a veces conocidas como variables independientes o variables predictoras así mismo se desarrollan algoritmos con el cual se puede caracterizar información histórica, que luego se puede emplear para predecir la naturaleza y probabilidad de eventos o sucesos futuros [5].

La supervisión predictiva de procesos ayuda a la identificación de los problemas empresariales antes de que se presenten y así mismo permite reasignar recursos antes de que estos no se han aprovechados, diversos enfoques de monitoreo predictivo de procesos utilizan técnicas de aprendizaje automático [6]. Estos métodos de aprendizaje automático en conjunto con técnicas de estadísticas en grandes datos multiestructurados permiten la identificación de correlaciones y relaciones causales, clasificar y predecir eventos, identificar patrones y anomalías son usados por la ciencia de datos como lo indica [7], estos métodos de aprendizaje automático

Al implementar el aprendizaje automático, no solo se consideran las capacidades de predicción de los modelos estadísticos sino también la capacidad de agregación y agrupación de los datos, dentro de los cuales se pueden identificar diversos tipos de aprendizaje algorítmico los cuales pueden ser implementados en aprendizaje automático dentro se clasifican en: i. Aprendizaje Supervisado, el cual infiere una función de datos etiquetados donde se conoce la variable objetivo [8] ii. Aprendizaje No supervisado, función de datos no etiquetados aquí no se conoce la variable objetivo [9] iii. Aprendizaje Semi-supervisado, infiere una función de datos combinados en donde algunos datos están etiquetados y otros no, en donde se puede o no conocer la variable objetivo iv. Aprendizaje reforzado, en este tipo de aprendizaje se busca que las máquinas aprendan en base a su propia experiencia, recompensado o penalizando la acción tomada por el agente – modelo estadístico - según el ambiente o entorno en donde interactúa. Hay que tener en cuenta que como las variables cambian constantemente, así mismo las decisiones tomadas por los agentes pueden cambiar [10].

3.1.2. TIPOS DE MODELOS DE REGRESIÓN MÁS UTILIZADOS

Dentro de los cuales tenemos i. Modelos de regresión, son aquellos que se utilizan para predecir valores futuros, pertenecen al grupo de modelos de aprendizaje supervisado dado que los algoritmos que lo ejecutan conocen cual es la variable objetivo y en este modelo se espera como

resultado un valor numérico [11] ii. Regresión Lineal Simple, en el cual se evalúan patrones que presentan una relación de linealidad entre una variable independiente “x” y una variable dependiente “y”. Para que exista dicha linealidad, cualquiera que sea el valor de “x” va a influir en el valor de “y” y su representación visual será una línea recta. Tiene diferentes aplicaciones, por ejemplo, estimar la estatura de un niño a medida que cumple años tomando en cuenta la estatura de los padres, conocer la cantidad de personas que tendrán empleo tomando en cuenta los valores históricos, entre otros [12]. iii. Regresión Lineal Múltiple, en este modelo existen dos o más variables independientes llamadas predictoras que ejercerán influencia sobre el valor a predecir en una única variable dependiente. Un ejemplo de este modelo es la predicción de precios de vivienda en donde se consideran varios datos asociados a la vivienda como la medida de la superficie, ubicación, estrato, cantidad de habitaciones, parqueadero, entre otros [12].

Por otro lado tenemos los iv. Modelos de Clasificación, con este tipo de modelos se busca identificar la categoría o clase a la que pertenecen los datos a evaluar. Se hace la división de los datos en el set de prueba y set de entrenamiento. La idea es que, mediante la validación del set de entrenamiento, el modelo pueda identificar de forma correcta a que clase o categoría pertenecen los datos evaluados y cuyo resultado se expresa generalmente en números binarios o texto. El que hace la operación de clasificar el algoritmo se llama clasificador. A continuación, algunos modelos de clasificación: Arboles de decisión [13], Bosques de decisión - Random Forest [14], Naive Bayes [15], K-Vecinos más cercanos y Regresión logística [16] v. Modelos de Clustering [17], este tipo de modelos no se tiene un objetivo que predecir. Se observan los datos y luego se intenta agrupar observaciones o eventos con características similares y formar diferentes grupos. Es un tipo de aprendizaje no supervisado. Pueden ser utilizados para dividir los clientes en diferentes grupos para poder ofrecerles productos basados en sus características comunes vi. Modelos de Asociación, los cuales son una clase de modelos de minería de datos en la cual se busca encontrar ítems u observaciones que se parezcan juntos en transacciones de un determinado conjunto de datos. Aquí se establecen reglas que indican dependencias entre los ítems u observaciones de dicho conjunto de datos. Estos modelos se utilizan para entender los comportamientos de compra de grupos de clientes con el objetivo de realizar ventas cruzadas, descuentos y promociones, entre otras. Uno de los algoritmos que más se utilizan en este modelo es A-Priori que utiliza elementos frecuentes en un set de datos para crear reglas que serán utilizadas para determinar qué tan fuerte o débil es la relación entre dos objetos vii. Modelos de Control, estos modelos utilizan redes neuronales para la resolución de problemas en los que se requiere que aprendan de sus experiencias observando los datos históricos y se ejecuten los

ajustes correspondientes. Este tipo de modelos puede ser utilizado para predecir la variación del precio del mercado de valores y en el área financiera para evaluar riesgos y detección de fraude.

3.2. ANTECEDENTES

El propósito de utilizar herramientas tecnológicas y de la información, permiten organizar los datos para manejarlos de manera adecuada y eficiente, consiguiendo certeza a la hora de toma de decisiones, esto puede ayudar a mejorar la predicción de algún negocio y/o sector, tal es el caso en el estudio desarrollado por Daza[18] el cual se realizó un estudio comparativo entre tres modelos predictivos de la propagación del virus con el objetivo de pronosticar los casos de COVID-19, por medio de la ejecución de Modelo Autorregresivo AR, modelo de Medias Móviles MA y el modelo Media Móvil Integrada Autorregresiva Estacional con Regresores Exógenos (SARIMAX). La evaluación de los modelos se realizó usando series temporales y pruebas de pronósticos (análisis forecasting).

En el 2019 A. Peña [19] desarrollo y presentó un modelo de pronósticos ponderado para predecir la inflación mensual en un corto plazo de tiempo mediante enfoques de datos univariados, por lo cual manejo un modelo Arima según las divisiones del gasto del IPC en Colombia y uso de técnicas de autoaprendizaje como el modelo Random Forest para pronosticar los índices de inflación, además para obtener un resultado acertado, se manejó bajo diferentes medidas de error MAPE, MAE y RMSE.

En 2019, en Perú se revisó el manejo de microcréditos al ser un componente principal en el desarrollo de la economía del país, por lo cual, al realizar un análisis basado en seis modelos de Machine Learning, como Regresión Logística (RL) [20], Random Forest (RF) [21], Support Vector Machine (SVM)[22], Artificial Neural Network (ANN), Decision Tree (dTree) y k-Nearest Neighbors (kNN), de acuerdo con [23] afirma que el modelo más asertivo que puede predecir la reducción de riesgo crediticio y mejorar cómo otorgar el microcrédito en base a variables determinadas a clientes es *Artificial Neural Network (ANN)*.

Existe un estudio que permite calcular una propensión de ahorro e inversión para identificar los potenciales clientes y generar estrategias de fidelización con entidades bancarias por medio de

campañas de la banca patrimonial. Lo anterior, se logró de acuerdo con lo que indica D. Guerrero calderon [24] en el cual manejando modelos de clasificación como Logistic Regression (LR), Linear Discriminant Analysis (LDA), k-Nearest Neighbors (kNN), Naive Bayes (NB), Random Forest Classifer (RF) y Gradient Boosting Classifer (GB), el mejor desempeño lo tuvo Random Forest (RF), con un "accuracy" de 0.936, haciendo que este modelo ayude a establecer una mejora en los saldos de los clientes que integran los Fondos de Inversión Colectiva y los Fondos de Ahorros Voluntarios de Pensiones.

En el trabajo "A note on the validity of cross-validation for evaluating autoregressive time series prediction" [25] se presentan varias teorías y un ejemplo de simulación sobre el favorable funcionamiento de la validación cruzada (Cross-validation ó CV) de K-fold en los métodos de aprendizaje automático para la predicción en los pronósticos de series de tiempo en comparación con la evaluación fuera de la muestra (out-of-sample ó OSS) y otras técnicas específicas de series temporales, como la validación cruzada no dependiente ya que estos procedimientos estándar son muy usados para el testeo de los modelos de clasificación y regresión.

Por otro lado menciona [26] que los modelos de series temporales promedio de movimiento autorregresivo, pueden considerarse como medios para transformar los datos en ruido blanco, es decir, en una secuencia de errores no corregida todo esto si los parámetros son conocidos. así mismo este puede ser calculado directamente desde las observaciones. Si se elige el modelo adecuado, no habrá autocorrelación en los errores. Para muestras grandes, los residuos de un modelo correctamente ajustado se parecen mucho a los verdaderos errores del proceso.

En resumen, los antecedentes aportaron en el desarrollo del objeto de estudio de manera significativa, pues al ver que los modelos de series temporales y de pronósticos, resaltan la importancia de la evaluación de resultados con métricas de error, donde el modelo Random Forest está en la mayoría de estos; esto hizo que tomáramos en cuenta la evaluación de resultados de tal forma en los modelos propuestos de predicción en este proyecto, por lo cual se procedió a comparar los valores reales con los pronósticos, tal es el caso que la métrica MASE dio mejor resultado para el modelo XGBOOST, donde se tuvo 0,066 de error. Adicionalmente, una ventaja en nuestra propuesta de modelo es la visualización de resultados, ya que no solo se cuenta con las gráficas que se tienen en el software R-Studio utilizado, sino que se integraron los resultados del modelamiento con Power BI para visualizarlos de manera interactiva con el usuario y así poder tomar decisiones referentes con los resultados de predicción de saldos de captación.

4. CARACTERIZACIÓN E IDENTIFICACIÓN

En el presente capítulo, se desarrolla el objetivo: Definir las fuentes de información mediante los datos suministrados ir la compañía, manteniendo un enfoque analítico, el cual aporta a la parte inicial de la aplicación para dar solución a la problemática que aqueja a la compañía y al sector bancario, por ende, se identifica el contexto y conocimiento del negocio, el cual, aporta como primera medida la generación de la solución estratégica a la problemática presentada al no conocer con exactitud las necesidades del cliente al cual se ofrecerán los productos de captación correspondientes, esto ayuda no solo a la situación deseada de nuestro objeto de estudio, sino a la construcción del modelo predictivo, ya que cada día el sector financiero busca impactar de manera sostenible y eficaz su negocio.

4.1. CARACTERIZACIÓN Y DIAGNÓSTICO ESTADO ACTUAL

En el caso de estudio, aunque el objetivo es crear un modelo predictivo para la proyección de saldos, es fundamental conocer la problemática que enfrenta el sector y especialmente la entidad bancaria, con el fin de contemplar en el acercamiento a la realidad las necesidades que entra el modelo de predicción a satisfacer para el negocio, así mismo, se podrá construir de manera acertada para obtener los resultados esperados.

4.1.1. IDENTIFICACIÓN DE LAS CAUSAS GENERALES MEDIANTE UN DIAGRAMA DE ISHIKAWA

Por lo anterior, se procedió a realizar la identificación de las causas generales mediante un Diagrama de Ishikawa manejando métodos de estratificación para examinar de manera directa el por qué se requiere un modelo para predecir la proyección de saldos en este tipo de productos de captación tales como CDT, CDAT, cuentas de ahorro, cuentas corrientes, entre otros. Para este caso, se realizó la selección de causas mediante una lluvia de ideas para definir el punto a mejorar con la solución que otorga el manejo de un modelo predictivo y poner atención especial en las fuentes de variabilidad

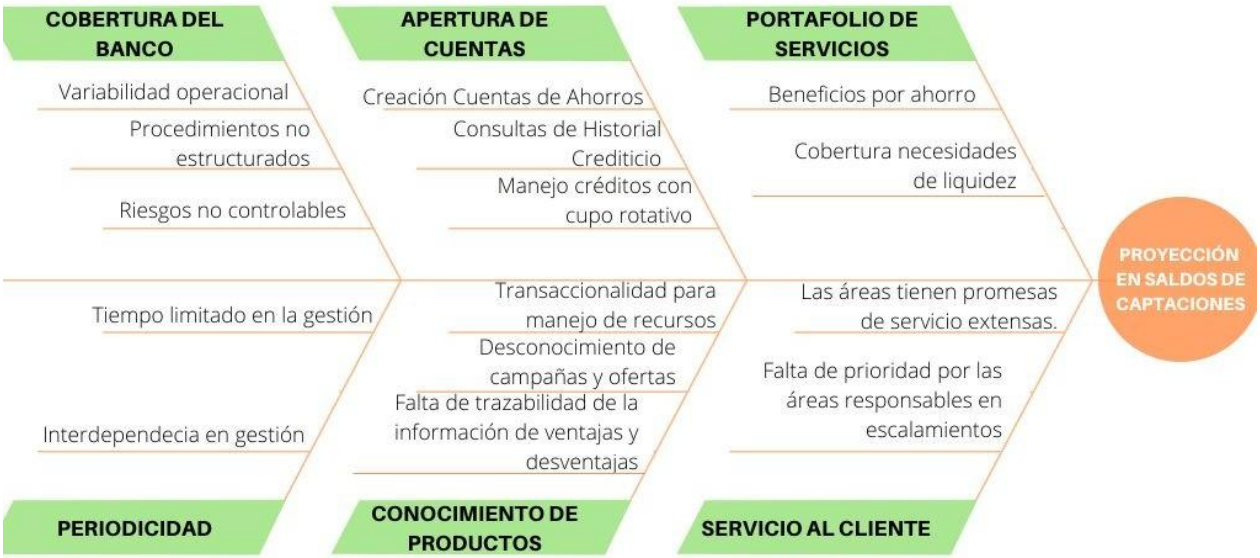


Ilustración 4: Diagrama de Ishikawa para proyección en saldos de captaciones. Fuente: Autores

La causa potencial en la que radica la problemática central es que no se cuenta con un perfilamiento del cliente que adquiere los productos, tal es el caso, como se evidencia en la Ilustración 3 que la falta de trazabilidad de la información y desconocimiento de campañas y ofertas, hacen que los clientes posiblemente potenciales no conozcan los productos que pueden adquirir con la entidad bancaria y esto aumenta la falta de fidelización por parte de los clientes.

4.1.2. MATRIZ DOFA

La finalidad del método DOFA permite que las empresas analicen su capacidad de adaptabilidad y posible evolución en la mejora continua, de acuerdo con [28] este análisis DOFA se basa en “la planeación estratégica que lleve a la empresa a integrar procesos que se anticipen o minimicen las amenazas del medio, el fortalecimiento de las debilidades de la empresa, el potenciamiento de las fortalezas internas y el real aprovechamiento de las oportunidades”. Por ende, para una compañía es fundamental potencializar la comunidad de Analítica de Datos que se tiene, así se pueden alcanzar los objetivos en nuestra visión al querer ser referente en innovación y servicio, logrando esto al mejorar la toma de decisiones empresariales a partir de los datos de nuestros clientes y las operaciones que se manejan día tras día en las áreas de trabajo, especialmente en Business Intelligence para la proyección de saldos de productos de captaciones.

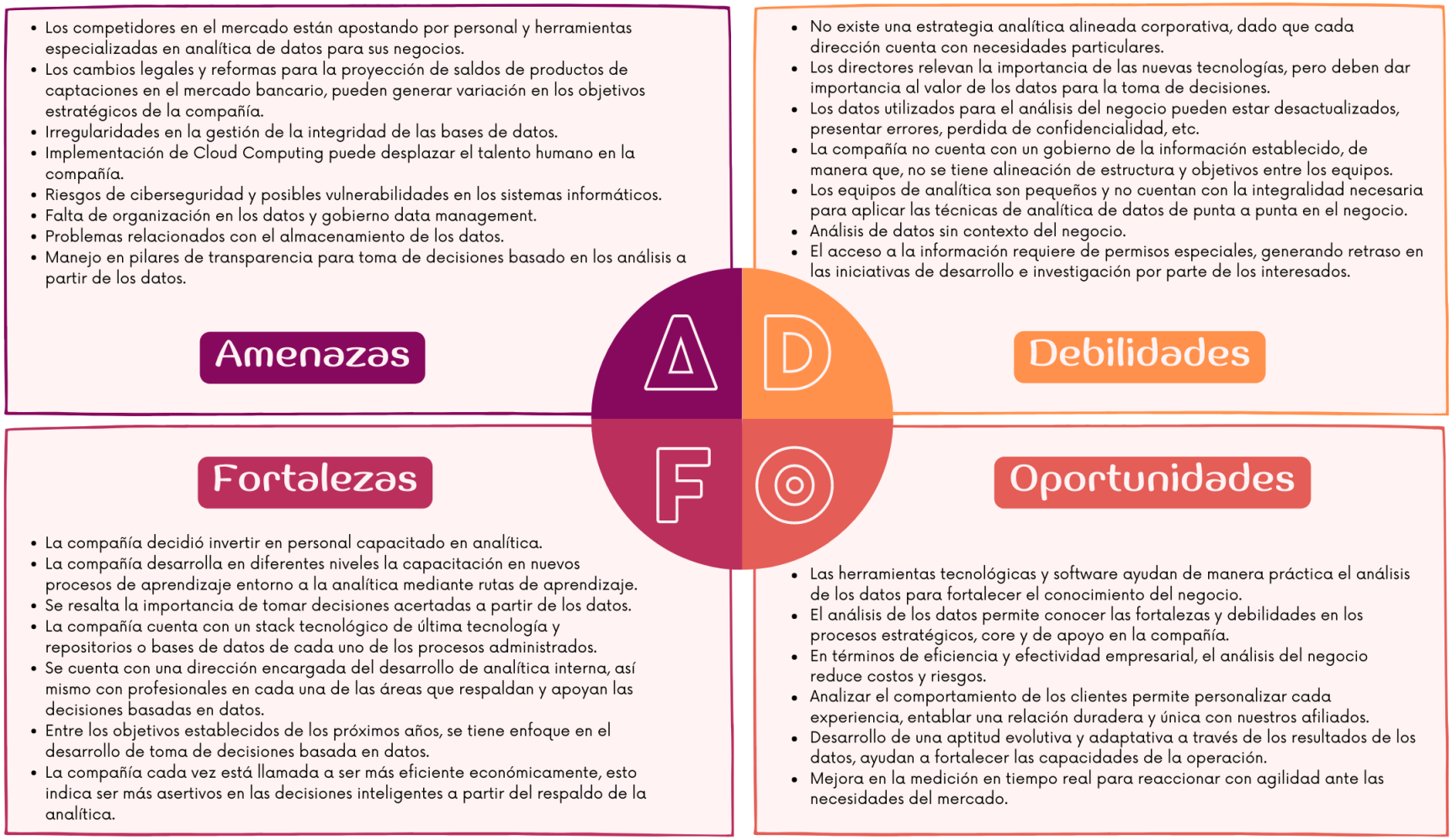


Ilustración 5: Matriz DOFA. Fuente: Autores.

4.2. COMPRENSIÓN DEL NEGOCIO

Actualmente los diversos productos financieros administrados en la compañía deben contar con diversos mecanismos de estimación que permitan mitigar los riesgos financieros asociados, para ello se requiere diseñar este proyecto con el cual se pueda identificar la relevancia de las variables en pro al negocio, este será prueba definitiva para validar si estas son de vital importancia para determinar cómo focalizar las estrategias cuando se requiera identificar en un cierto periodo de tiempo y cuál es el valor estimado del saldo de captación. La organización y la dirección de Business Intelligence son los principales afectados con la incertidumbre, la cual se asume debido a que no se cuentan con métodos de estimación a partir de los comportamientos históricos de los clientes, con la cual se permita proyectar los saldos en los siguientes periodos que permitan tomar decisiones estratégicas y de valor para los diversos productos administrados.

4.2.1. PILA TECNOLÓGICA

Dentro de la revisión de las necesidades del negocio, desde el equipo de proyecto aplicado se determina la importancia de manejar una pila tecnológica, en la cual se identificó que por el tipo de negocio en el área de Business Intelligence, se debe tener en cuenta las perspectivas del profesional de inteligencia de negocio, así mismo como el creador del contenido y el consumo de la información para proyección de saldos de productos de captaciones, a continuación, se describen cada una de estas pilas tecnológicas:

Tabla 1: Pila Tecnológica para manejo del modelo predictivo . Fuente: Autores.

Etapa de flujo de trabajo	Stack Tecnológico	Características
Ingestión de datos	DataStage	• Debido a que los datos se almacenan en diferentes formatos y en diferentes ubicaciones se utiliza herramienta para la extracción, transformación y carga para crear un repositorio único que pueda almacenar todos estos datos mediante ETL.

Capa de análisis de datos	Python, R, Power BI, Excel y Looker	• En esta capa se desarrolla el proceso de análisis de datos, utilizando herramientas para ayudar al negocio a responder las preguntas, segmentando datos, creando tableros, desarrollando datos procesables, por medio del desarrollo de analítica avanzada. • Análisis y exploración de datos (EDA) se utilizan las herramientas Python y R de acuerdo con las preferencias del equipo. • El desarrollo de tableros BI, se utiliza como herramienta oficial dentro de la compañía Power BI y para los rastreos de las interacciones de nuestros clientes dentro de la web looker. • Excel se usa para el desarrollo de informes en la mayor parte de los informes es usado.
Capa de repositorio de datos	**Sistemas de gestión de base de datos relacional:** Microsoft SQL Server, MySQL, Oracle, PostgreSQL. **Servidores de bases de datos:** NoSQL y MongoDB	• El procesamiento de datos optimiza los datos para el desarrollo de análisis más eficiente y produce un motor de análisis para ejecutar diversas consultas. • Estas herramientas se utilizan por cada uno de los equipos de desarrollo de cada una de las aplicaciones del CORE y solo en algunas pocas se pueden generar consultas directas por el equipo de analítica ya que estas BD están alojadas en producción y puede generar un riesgo en la perdida de información • Los sistemas de bases de datos relacionales y servidores se usan de acuerdo con cada una de las aplicaciones administradas, para el acceso a esta información se realiza con canalización directa con los equipos de desarrollo generando acuerdos de servicios para la toma de la información

Capa de procesamiento de datos	Datawarehouse.	• El repositorio de los datos está conformado por varios gestores de datos y Almacenes de datos.
Capa de fuente de datos	Datawarehouse.	• Esta capa también denominada capa de recopilación de datos, el cual incluye varios datos recopilados de dispositivos de res y servicios comerciales

4.2.2. HOJA DE RUTA

La elaboración de una hoja de ruta adecuada para lograr una planeación estratégica con el objetivo de poder llevar a cabo el presente proyecto aplicado de ciencia de datos en la compañía, permitirá garantizar que efectivamente se manejen los cambios necesarios para adoptar las prácticas requeridas y medir los avances alineados con los objetivos de la compañía. Por lo cual, la hoja de ruta propuesta se establece en tres (3) grandes hitos:

4.2.2.1. Definición Alcance y Prioridades

De acuerdo con los resultados obtenidos en el análisis DOFA, la compañía se debe enfocar en estrategias adaptativas como la creación de un modelo de gobierno establecido para fijar estándares y pilares internos para la recolección, almacenamiento, procesamiento y eliminación de los datos según las necesidades de las áreas, así mismo para fortalecer las capacidades del talento humano de la compañía. Además, se deberá implementar políticas en manejo de datos como estrategia defensiva a las amenazas del entorno digital en la proyección de saldos.

4.2.2.2. Manejo y acceso a Datos

La compañía debe enfocar sus esfuerzos en mejorar los puntos críticos y debilidades como lo son establecer lineamientos de arquitectura y manejo de bases de datos para tener backups con data confiable, resaltando la importancia de la precisión y calidad de los datos para llegar al mercado objetivo en la proyección de saldos de productos de captaciones. Así mismo, se debe introducir cambios y diversificar la oferta de los productos, por medio de estrategias comerciales para

mejorar la percepción del cliente frente al servicio, a partir de los análisis de los datos luego de procesos de preprocesamiento, preparación y modelamiento predictivo del proyecto.

4.2.2.3. Desarrollo e Implementación

Finalmente, como parte de las estrategias de Gestión del Cambio para incorporar en la compañía el modelo predictivo para proyección de saldos de productos de captaciones, se debe crear un Comité de ética y Cultura de los datos con el objetivo de alinear nuestra apuesta empresarial está en crear las políticas y reglas de gobierno para garantizar la confianza, transparencia, prácticas justas y cumplimiento de la ley en privacidad y protección de los datos en la dirección Business Intelligence.

5. MATERIALES Y MÉTODOS

En el presente capítulo, se evidencia la metodología utilizada acorde al objetivo: Verificar y comprender los datos para aplicar reglas de calidad y preparación de los datos, pues, ante la especificación del sector a evaluar, se reconoce cada una de las variables que permiten tener conocimiento de las necesidades de la entidad bancaria, principalmente para manejo de proyección de saldos en productos de captación, a continuación se especifica el estudio y comprensión de datos para el proyecto aplicado:

5.1. EXPLORACIÓN Y COMPRENSIÓN DE LA INFORMACIÓN

Con el objetivo de crear un modelo especifico que cumpla con la solución a las necesidades del negocio, se debe familiarizar con los datos teniendo presente los objetivos de la dirección Business Intelligence como guía para el proceso de análisis de datos para la proyección de saldos de productos de captaciones, por ende, se debe contemplar:

5.1.1. IDENTIFICACIÓN DE FUENTES

Por el diseño y estructura con el que se administran diversos niveles de información de los clientes, se han implementado en diversas fuentes de información y en diferentes tecnologías los detalles necesarios para conocer a detalle el comportamiento transaccional de cada cliente, dentro de los cuales tenemos para el Almacén de Datos Operacional (ODS) y Almacén de Datos (DWH) variadas aplicaciones, archivos, bases de datos y diversas fuentes de información corporativa. Por esta razón, el objetivo de este capítulo es vital para delimitar el tipo de información y variables se puede desarrollar el modelo a partir de la compresión de los datos.

5.1.2. DELIMITACIÓN DEL PROBLEMA A NIVEL DE NEGOCIO Y TÉCNICO

Si bien el negocio ha planteado una serie de hipótesis en el cual dentro de las herramientas, técnicas y modelos que la ciencia de datos ofrece, se puede delimitar la necesidad y aterrizar las posibles soluciones, por lo cual, se requiere la participación de las personas directamente

relacionadas con este ramo dentro de la compañía así mismo, con los encargados del soporte de las aplicaciones y administradores de la composición de la información y características permitirá conocer al detalle las diversas y posibles soluciones a la problemática.

5.1.3. IDENTIFICACIÓN DE INFRAESTRUCTURA

Las fuentes de información poseen características como precisión, actualidad, complejidad, volumen, variedad, fiabilidad y disponibilidad como por ejemplo retiros, transferencias, pagos realizados por los clientes, saldos y movimientos para cada cliente, fechas, tipo de movimientos, códigos de oficinas, esto solo por nombrar alguno de ellos. con lo que es importante cubrir técnicamente el acceso a las diversas fuentes de información y sus características para desarrollar la extracción.

5.1.4. VERIFICACIÓN DE HIPÓTESIS

Conociendo el efecto de causan los datos en la toma de decisiones, puede asumirse que usar técnicas de ciencia de datos y modelos de pronósticos, ayudan a predecir valores de saldos de captaciones en la entidad bancaria.

Dado el planteamiento de la hipótesis en torno al desarrollo del problema, es indispensable desarrollar la fase de exploración y comprensión de la información para asegurar que se puedan contrastar, probar, analizar y comprender la estructura de la información para contestar las respuestas necesarias con el fin de descubrir patrones, detectar anomalías y comprobar supuestos.

5.1.5. VALIDACIÓN DE HIPÓTESIS

Una vez se seleccionadas las variables de las diversas fuentes de información se procede con el desarrollo del proceso del viaje de los datos desde las fuentes origen asegurando la integridad, precisión y validación correspondiente al proceso en curso y disponibilizado para la automatización correspondiente al proyecto

5.1.6. DESARROLLO DE LA PROBLEMÁTICA

Una vez probados los supuestos que negocio requiere solucionar a través de las diversas técnicas y métodos de ciencia de datos, se procede al desarrollo y puesta en producción de la solución con el cual negocio pueda estimar los valores de las captaciones de su portafolio con el propósito de diseñar estrategias de negocios basados en datos.

5.2. ESPECIFICACIÓN FUNCIONAL DE PROCESO (ETL- ELT)

5.2.1. EXTRACCIÓN

El proceso inicia con la Extracción de información de las tablas maestras de cuentas de ahorros del producto de captación de la entidad bancaria que se encuentran en el ODS (Operational Data Storage).

5.2.2. TRANSFORMACIÓN

La información se lleva a la zona de stage, luego se realiza la limpieza y transformación para llevar los datos a la zona de Operacional o Datawarehouse según corresponda. Las sábanas de datos e información histórica debe ir a la zona operacional y las tablas de hechos y dimensiones que se creen en el proceso deben quedar en el Datawarehouse.

5.2.3. CONSOLIDACIÓN

Finalmente la información se debe disponer en la zona operacional o datawarehouse, si es necesario se debe diseñar un cubo multidimensional o modelo tabular para que sea fuente para la extracción de datos de parte del usuario final y/o fuente para el diseño de tableros en la herramienta de visualización.

5.2.4. DEFINICIÓN DE ENTREGABLES

Tabla 6: Entregables definidos con la base de datos creada para el modelamiento.

Nombre Entregable	**Tipo Entregable**	**Descripción entregable**
FCT_SALDOS_CUENTAS_AHORROS	Tabla	Tabla de hechos con información de cuentas de ahorros
FCT_SALDOS_CUENTAS_CORRIENTES	Tabla	Tabla de hechos con información de cuentas de corrientes
Cubo Saldos Captaciones	Cubo SSAS	Cubo que contiene la información referente a Cuentas Ahorros y Cuentas Corrientes
Tabla_Saldos_Proyeccion	Tabla	Tabla que contiene la información de los saldos históricos y su proyección
Tablero Saldos – Proyección	Tablero Power BI	Tableros que muestran el comportamiento de los saldos de Cuentas Ahorros y Cuentas Corrientes y su proyección.

5.2.5. DIAGRAMAS DE ARQUITECTURA

Diagrama de Arquitectura AS – IS

El diagrama AS – IS del flujo del proyecto es en donde se especifica la situación actual y la realidad del proceso con sus errores y aciertos.

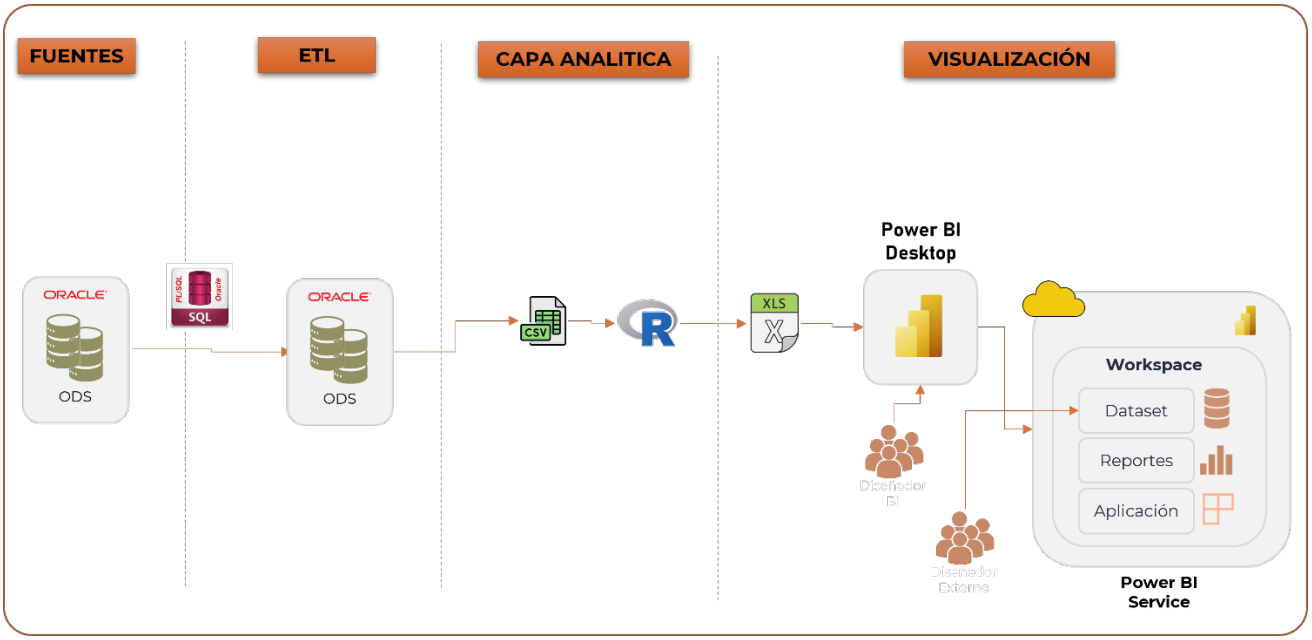


Ilustración 6: Diagrama de Arquitectura AS-IS

Diagrama de Arquitectura TO-BE

Muestra la arquitectura a la que se pretende llegar al final de la evolución del proceso o la solución a implementar.

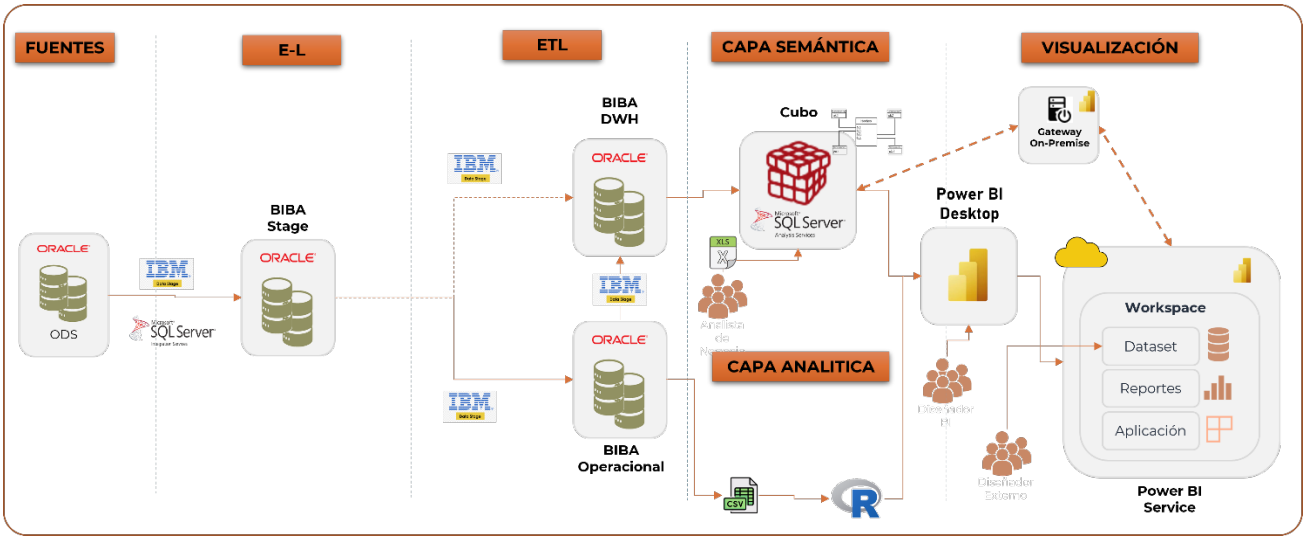


Ilustración 7: Diagrama de Arquitectura TO-BE

5.2.6. MODELO DE DATOS DE PROCESO

Se ha desarrollado un modelo de copo de nieve para estructurar y organizar la información de manera eficiente. Por la naturaleza sensible y confidencial de los datos involucrados, no se puede mostrar el modelo detallado por el riesgo potencial de fuga de información. La seguridad y protección de los datos son de suma importancia, y se han implementado medidas adecuadas para salvaguardar la integridad y privacidad de la información.

5.2.7. CAPAS TECNOLÓGICAS DEL PROCESO

A continuación, se detallan las capas tecnológicas sobre la implementación del DWH:

Tabla 7: Capas tecnológicas de la aplicación.

Capas Tecnológicas	**Capas lógicas aplicación**			
	Integración	**Modelamiento**	**Presentación**	**Cliente**
herramienta o Software	(SSIS - DataStage)	(SSAS)	Power BI, Excel	Power BI, Excel
Sistema Operativo	Windows	Windows	Windows	Windows
Servidor de Aplicación	ARGEL.	ARGEL.		N/A

5.2.8. POLÍTICAS DE ACCESO A LA BASE DE DATOS BIBA - ORACLE

Prerrequisitos:

- Contar con un equipo ubicado dentro de la organización con una versión de Oracle cliente 12c instalada y PL/SQL la última versión autorizada por la entidad bancaria.
- Actualizar el TNSname (TNSname.ora es un archivo de configuración utilizado por Oracle Database) dentro de cada equipo local que requiera la conexión.
- El usuario debe estar creado dentro de la base de datos.
- El usuario deberá contar con una IP fija.

Accesos:

- El usuario podrá acceder por medio de su usuario de red.
- La asignación del Password se realizará por parte del DBA administrador de la base de datos.
- Una vez cumpla con los prerrequisitos y cuente con la contraseña, acceda desde la versión que tenga instalada de PL/SQL a las dos bases de datos de consumo: PDBIOP y PDBIDW.

5.3. ANÁLISIS DE LOS DATOS Y SELECCIÓN DE CARACTERISTICAS

Es preciso tener presente que para llevar a cabo el enfoque del proyecto y predecir el valor futuro de una captación de clientes específicos, los datos a evaluar y analizar son datos personales, ya que se pueden identificar y asociar los datos específicos a una persona dando cuenta de la individualidad en la sociedad.

Los datos con los que se trabaja, al ser datos personales están divididos en Datos Públicos como lo son por ejemplo nivel educativo, estado civil, género, si tiene personas a cargo menor de 18 años, tipo de vivienda, estrato, profesión y ciudad de residencia y Datos Semi Privados, como datos financieros por el manejo del saldo que tiene con la compañía, datos de las oficinas donde el clientes realiza sus procesos de retiro de saldos, manejo de declaración de renta, y otros datos personales como la edad del cliente. [30]

5.3.1. COLECCIÓN DE DATOS

La extracción de los datos desde las diversas fuentes de información de la compañía financiera, comprende una serie de actividades con el objetivo de garantizar una buena calidad de estos, por ende, dentro de estas actividades el primer paso para entender la dinámica transaccional con la cual de almacena la información es la identificación de las fuentes, para ello en la tabla 2 se aprecian las principales fuentes y tablas de información en la cual se procede a realizar la adquisición de la información a partir de las reglas objetivo.

Tabla 2: Bases y fuentes de información para comprensión de los datos.

Tabla	Fuente
USR_BIDW.FCT_CUENTAS_AHORROS	USR_BIBAST.GBL_ODS_ATAHORROS
USR_BIDW.FCT_CUENTAS_CORRIENTES	USR_BIBAST.GBL_ODS_ATCTACTES
USR_BIDW.GBL_DIM_CODIGOS_CIIU	USR_BIDW.FCT_CUENTAS_AHORROS
USR_BIDW.GBL_DIM_PARTICIPANTE	USR_BIDW.FCT_CUENTAS_AHORROS
USR_BIDW.GBL_DIM_TIEMPO	USR_BIDW.FCT_CUENTAS_AHORROS
USR_BIDW.NEG_DIM_ESTADO_CUENTA	USR_BIDW.FCT_CUENTAS_AHORROS
USR_BIDW.NEG_DIM_MOTIVO_CANCELACION	USR_BIDW.FCT_CUENTAS_AHORROS
USR_BIDW.NEG_DIM_SUBPRODUCTOS	USR_BIDW.FCT_CUENTAS_AHORROS
USR_BIDW.NEG_DIM_TIPO_CUENTA	USR_BIDW.FCT_CUENTAS_AHORROS
GBL_DIM_ACT_ECONOMICA	
GBL_DIM_EDAD	
GBL_DIM_ESTADO_CIVIL	
GBL_DIM_GENERO	
GBL_DIM_NIVEL_ESTUDIO	
GBL_DIM_OCUPACION	
GBL_DIM_SEGMENTO	
GBL_DIM_SUB_SEGMENTO	
GBL_DIM_TIPO_CLIENTE	

Las principales consideraciones que se tienen en cuenta dentro de la adquisición de los datos son las fuentes de las variables que se utilizan y son requeridas a nivel significativo para el modelamiento de los datos, y como bien se sabe, algunos de los conceptos para la recopilación incluyen el hecho de que estos se pueden recopilar de una amplia cantidad, número y tipos de fuentes de datos, de hecho, la mayor parte de la información se extrae de bases relacionales.

A continuación, la tabla 3 describe mediante un diccionario de datos las variables a utilizar, las cuales fueron identificadas dentro de la comprensión empresarial:

Tabla 3: Diccionario de datos. Fuente: Autores

Variable	Tipo	Definición
ID	dbl	Identificación del registro dentro del set de datos.
ID_ASIGNACION	dbl	Identificador de la unidad comercial de asignación del cliente según todos sus productos.
ID_COD_CIIU	dbl	Identificación de la actividad económica de cada cliente.
ID_COD_CIUDAD	dbl	Identificador ciudad origen a la cual pertenece el cliente.
ID_COD_OCUPACION	dbl	Identificador ocupación del cliente.
ID_COD_SUBPRO	dbl	Identificador de la línea de producto a la cual esta matriculada la cuenta.
ID_DECLARA_RENTA	dbl	Booleano declara renta (Si/No).
ID_EDAD	dbl	Edad del cliente.
ID_ESTRATO	dbl	Identificador del estrato al cual pertenece el cliente dentro de la entidad financiera.
ID_MAR_STATUS	dbl	Identificador estado marital del cliente.
ID_NIVEL_EDUCATIVO	dbl	Nivel de educación del cliente: Preescolar, Primaria, segundaria, universitaria y sin formación académica.
ID_OFICINA	dbl	Identificador de la oficina y está dado por la estructura organizacional de la entidad financiera.
ID_PER_CARGO_MN18	dbl	Número de personas a cargo del cliente menores o iguales a 18 años de edad bajo su responsabilidad.
ID_PER_CARGO_MY18	dbl	Número de personas a cargo del cliente mayores de 18 años de edad bajo su responsabilidad.
ID_REGION	dbl	Identificador de la región de emisión del producto.
ID_SEG_COMERCIAL	dbl	Identificador del segmento comercial registrado en la afiliación inicial.
ID_SEX	dbl	Identificador del sexo del cliente.

ID_TIPO_IDEN	dbl	Identificador tipo de identificación del cliente.
ID_TIPO_VIVIENDA	dbl	Identificador del tipo de vivienda del cliente.
ID_ZONA	dbl	Identificador de la zona a la cual pertenece la oficina de emisión del producto financiero.
SALDO_CAPTACION	dbl	Saldo final al cierre del periodo.
AAHO_FUENTE	String	Identificador del código tipo de cuenta 336 Ahorros y 339 Corriente.
AAHO_NUM_CUENTA	String	Identificador consecutivo de la cuenta de ahorro de la persona en la base de datos.
AAHO_COD_SUB_PRO	String	Identificador del código del subproducto.
AAHO_ESTADO	String	Indica si la cuenta se encuentra activa.
AAHO_COD_CIIU	String	Identificación de la ciudad registrada en la vinculación inicial a la entidad bancaría
AAHO_OFICINA	String	Identificación oficina dentro de la entidad bancaria.
AAHO_REGION	String	Identificación de la región a la cual pertenece la oficina en la entidad bancaría.
AAHO_ZONA	String	Identificación de la zona donde se encuentra la oficina en la entidad bancaria.
AAHO_TIPO_SUBPRODUC	String	Indica el tipo de subproducto a la cual pertenece la cuenta bancaria registrada.
AAHO_INAC_MOVIMIEN	String	Indica si la cuenta tiene movimiento reciente
SALDO_FECHA	Numérico	Indica el valor del saldo en pesos colombianos (COP) a la fecha de corte.
BB_TIPO_CLPJ	String	Indica el tipo de persona N (Natural) J (Jurídica).
BIRTHDATE	Date	Fecha de nacimiento del cliente.
ASIGNACION	String	Nombre de oficina asignada al cliente.

5.3.2. DESCRIPCIÓN DE LOS DATOS

En las bases de datos transaccionales en la entidad financiera hay varios registros y atributos para procesar información, pero, en las consideraciones y delimitación del proyecto, se realizaron

actividades primarias relacionadas con el descubrimiento de las principales fuentes de información necesarias para predicción de los saldos de productos de captación.

El origen de los datos se relaciona en las fuentes de datos con los saldos y movimientos por cliente, por ende, en la actividad de investigación y descubrimiento de variables de interés, se identificaron 89 variables numéricas y categóricas con 55.772 registros correspondientes a la transaccionalidad de 64 meses de clientes con saldos mayor o igual a 100.000 COP al último día de cada mes, debido a las reglas de negocio establecidas para estimar saldos a cierre. Si bien 64 de estas variables están relacionadas con el saldo final de cada cuenta al último día de cada mes, las 25 variables adicionales relacionan la descripción del cliente: edad, género, nivel educativo, estatus marital, personas a cargo, personas a cargo, tipo de vivienda, estrato, entre otros.

Al relacionar el comportamiento a nivel de negocio, se definieron en el proceso de parametrización para el desarrollo de los modelos las siguientes variables: código de subproducto, oficina, regional, zona, segmento comercial y declaración de renta.

5.3.3. EXPLORACIÓN DE DATOS

Para realizar la exploración de los datos es importante conocer algunas medidas que permiten identificar la composición de las variables numéricas del set de datos recuperado, en la tabla 4 se podrá identificar los tipos de datos, registros con nulos, cantidad de valores únicos, mínimos, promedio y máximo para cada una de las variables identificadas.

Sin embargo, este proceso se limita a verificar qué factores son importantes para las predicciones del modelo [31] por lo cual, ayuda a evaluar qué tan aceptable es el modelo en sí, por eso si son consistentes y tienen influencia las variables con los resultados esperados, harán que la compañía se sienta segura y tenga la confianza para tomar decisiones en el manejo de saldos de productos de captación.

5.3.4. COMPRENSIÓN DE LOS DATOS

Actualmente se estiman métricas de comportamiento de los clientes en la compañía, como la proporción de clientes que retiran sus saldos en diversos periodos de tiempo, composición poblacional con diverso nivel de captación de los productos, entre otras. El cual permite ver a través de descriptivos los principales comportamientos de los productos y al tratarse de la administración de los recursos es importante estimar en la entrada el comportamiento de los clientes de acuerdo con sus características y comportamiento financiero dentro de la entidad. Con el fin de obtener una comprensión de la información, es importante desarrollar una comprensión de datos, el cual implica estudiar más de cerca los datos, con esto identificamos tendencias dentro de la información como se evidencia a continuación:

Tabla 4: Exploración de datos.

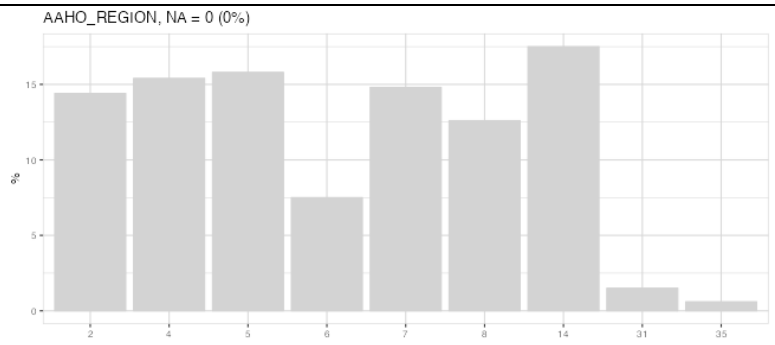
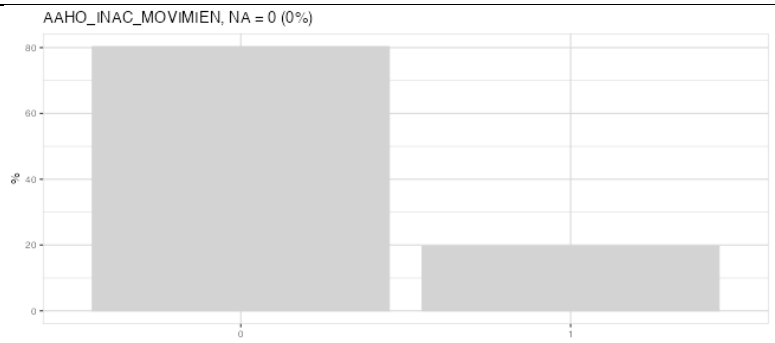
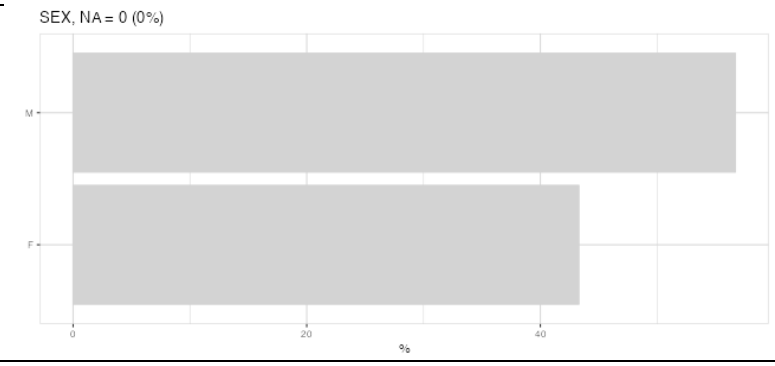
variable	type	na	na_pct	unique	min	mean	max
SEX	chr	0	0	2			
SALDO_202304	dbl	0	0	55540	50071.5	8026258.77	1720785278.47
SALDO_202303	dbl	0	0	55542	50071	8114875.91	1709086550.47
SALDO_202302	dbl	0	0	55540	50070.5	8259896.47	1687393783.47
SALDO_202301	dbl	0	0	55540	50070	8522503.96	1734314769.06
SALDO_202212	dbl	0	0	55538	50069.5	9005287.24	1893964211.97
SALDO_202211	dbl	0	0	55529	50001.19	8697451.48	1819572314.06
SALDO_202210	dbl	0	0	55535	50068.5	8742129.48	3285041198.55
SALDO_202209	dbl	0	0	55536	50068	8950497.28	3213396631.55
SALDO_202208	dbl	0	0	55538	50067.5	9204621	2765004721.05
SALDO_202207	dbl	0	0	55534	50103.06	9599895.79	2970030681.55
SALDO_202206	dbl	0	0	55542	50082.5	9816963.23	3612925782.05
SALDO_202205	dbl	0	0	55531	50043.05	9468863.1	3764749039.55
SALDO_202204	dbl	0	0	55538	50081.5	9593218.88	4332988609.56
SALDO_202203	dbl	0	0	55534	50081	9497230.94	4323444838.06
SALDO_202202	dbl	0	0	55540	50099.06	9614140.75	5635212815.06
SALDO_202201	dbl	0	0	55533	50097.06	9618807.4	5850250658.56
SALDO_202112	dbl	0	0	55538	50095.06	9801727.42	5646921079.06
SALDO_202111	dbl	0	0	55538	50007.74	9654496.11	5620309784.56
SALDO_202110	dbl	0	0	55539	50052.14	9357778.44	4578590130.06
SALDO_202109	dbl	0	0	55536	50089.06	9346746.53	4074876804.9
SALDO_202108	dbl	0	0	55541	50001	9369133.88	4801713709.06
SALDO_202107	dbl	0	0	55534	50085.06	9504134.25	4339008298.18
SALDO_202106	dbl	0	0	55535	50083.06	9548385.35	4106710418.46
SALDO_202105	dbl	0	0	55542	50081.06	9208218.29	6711322475.07

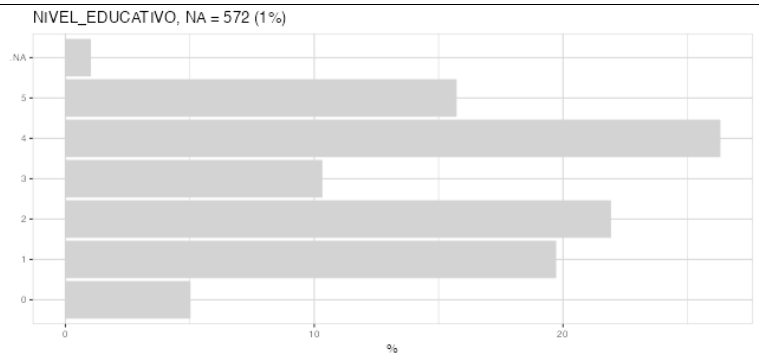
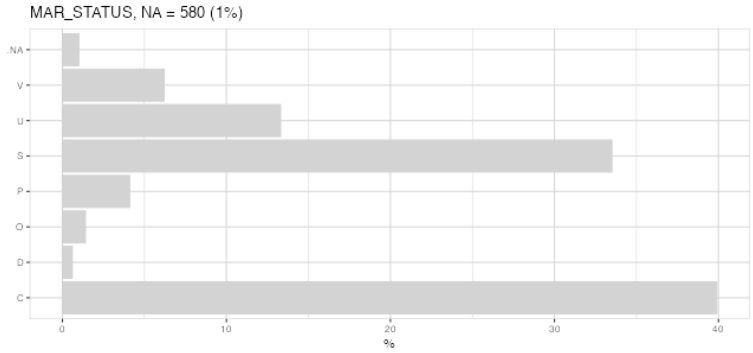
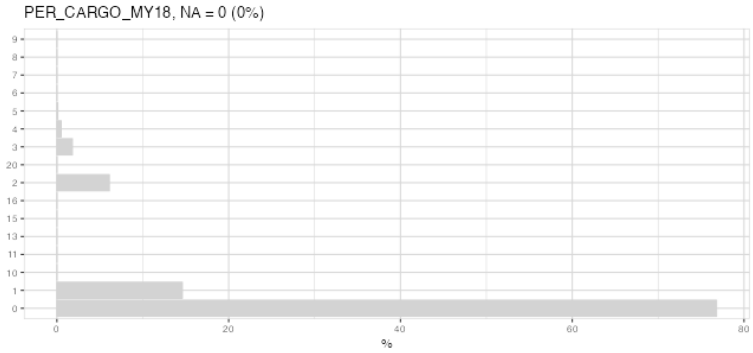
variable	type	na	na_pct	unique	min	mean	max
SALDO_202104	dbl	0	0	55531	50079.06	9040642.17	2980727050.28
SALDO_202103	dbl	0	0	55535	50077.06	8868203.54	3374810284.28
SALDO_202102	dbl	0	0	55530	50018.38	8863049.87	3423510085.78
SALDO_202101	dbl	0	0	55531	50052.73	8833217.17	4118873708.78
SALDO_202012	dbl	0	0	55525	50071.06	8897342.94	3909664608.28
SALDO_202011	dbl	0	0	55531	50120.5	8609323.05	3682453018.78
SALDO_202010	dbl	0	0	55531	50118.5	8344808.84	3662110503.78
SALDO_202009	dbl	0	0	55539	50115.5	8338189.31	3563356990.78
SALDO_202008	dbl	0	0	55527	50111.5	8252669.56	1898141600.28
SALDO_202007	dbl	0	0	55532	50107	8353151.88	2944017344.2
SALDO_202006	dbl	0	0	55535	50102.5	8248829.36	2844343449.76
SALDO_202005	dbl	0	0	55526	50098.5	7785487.12	2745635962.26
SALDO_202004	dbl	0	0	55520	50094	7706467.28	2714095689.04
SALDO_202003	dbl	0	0	55506	50090	7295715.39	2624538754.07
SALDO_202002	dbl	0	0	55507	50085.5	6951772.41	2557630714.17
SALDO_202001	dbl	0	0	55514	50010.5	6929359.15	2508192494.04
SALDO_201912	dbl	0	0	55512	50069.91	7067610.08	2427934209.88
SALDO_201911	dbl	0	0	55504	50072.5	6694676.2	2089868731.78
SALDO_201910	dbl	0	0	55502	50005.32	6565132.3	2443143154.78
SALDO_201909	dbl	0	0	55501	50049.5	6573129.56	1949025419.04
SALDO_201908	dbl	0	0	55504	50059.5	6684777.12	2214914673.44
SALDO_201907	dbl	0	0	55496	50055.5	6650419.13	1873471051.42
SALDO_201906	dbl	0	0	55500	50051	6749521.75	1852792425.82
SALDO_201905	dbl	0	0	55476	50030.5	6294128	1689556212.07
SALDO_201904	dbl	0	0	55470	50024.04	6297498.68	1660293762.35
SALDO_201903	dbl	0	0	55473	50038.5	6350103.55	1684560265.28
SALDO_201902	dbl	0	0	55488	50034	6366439.22	2084769341.78
SALDO_201901	dbl	0	0	55478	50030	6356192.76	2016752670.43
SALDO_201812	dbl	0	0	55471	50018.5	6449742.8	1601106483.78

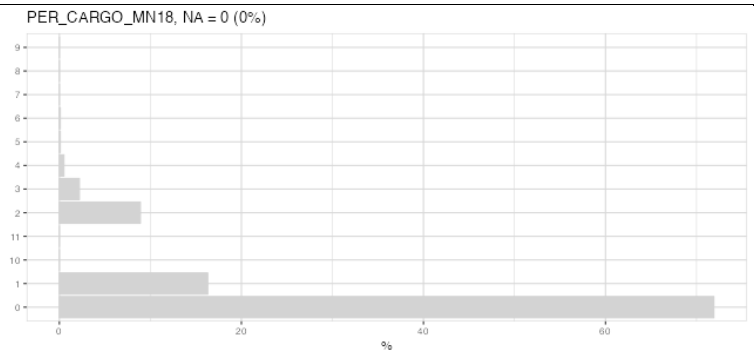
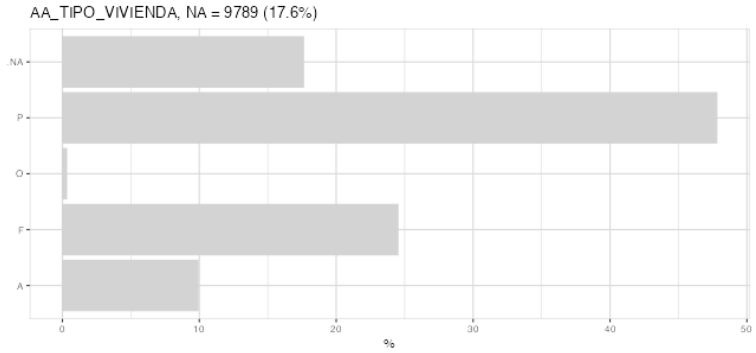
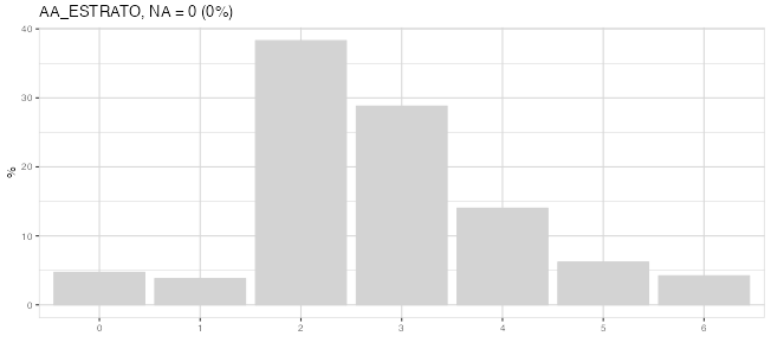
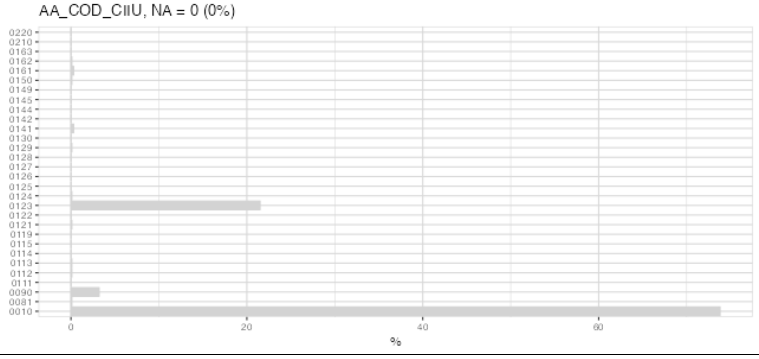
variable	type	na	na_pct	unique	min	mean	max
SALDO_201811	dbl	0	0	55475	50007.72	6113205.66	1536169660.28
SALDO_201810	dbl	0	0	55471	50017	6008623.19	1669323585.28
SALDO_201809	dbl	0	0	55489	50012.5	6085052.73	1490621707.48
SALDO_201808	dbl	0	0	55469	50002.42	6163414.54	2762302290.86
SALDO_201807	dbl	0	0	55474	50007.45	6161495.75	3201134099.77
SALDO_201806	dbl	0	0	55470	50014	6238325.65	2935701995.27
SALDO_201805	dbl	0	0	55465	50023.5	5829248.77	2603945789
SALDO_201804	dbl	0	0	55453	50009	5831196.76	2598054278.39
SALDO_201803	dbl	0	0	55469	50004	5922953.54	7040386431.78
SALDO_201802	dbl	0	0	55465	50010.5	5810449.35	6793510034.18
SALDO_201801	dbl	0	0	55441	50000.5	5869757.91	14615343457.76
PER_CARGO_MY18	chr	0	0	16			
PER_CARGO_MN18	chr	0	0	12			
NIVEL_EDUCATIVO	chr	572	1	7			
MAR_STATUS	chr	580	1	8			
COD_CIUD_DIR_PPAL	chr	37	0	1070			
BIRTHDATE	chr	0	0	20924			
BB_TIPO_CLPJ	chr	0	0	1			
BB_SEG_COMERCIAL	chr	0	0	20			
ASIGNACION	chr	2	0	739			
AAHO_ZONA	dbl	0	0	10			
AAHO_REGION	dbl	0	0	9			
AAHO_OFICINA	dbl	0	0	745			
AAHO_NUM_CUENTA	dbl	0	0	55772			
AAHO_INAC_MOVIMIEN	dbl	0	0	2			
AAHO_FUENTE	dbl	0	0	1			
AAHO_ESTADO	dbl	0	0	1			
AAHO_COD_SUB_PRO	dbl	0	0	27			
AAHO_COD_CIIU	dbl	5	0	375			

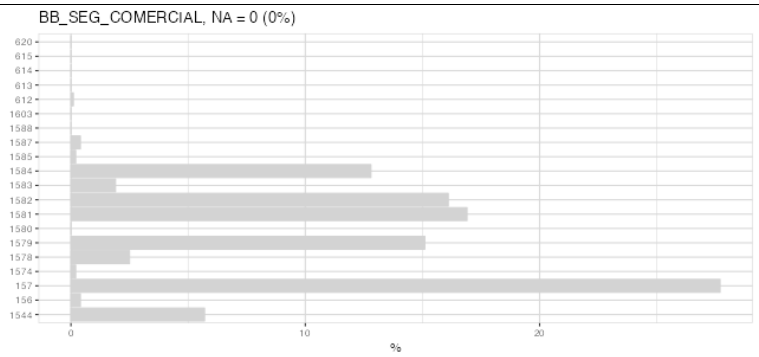
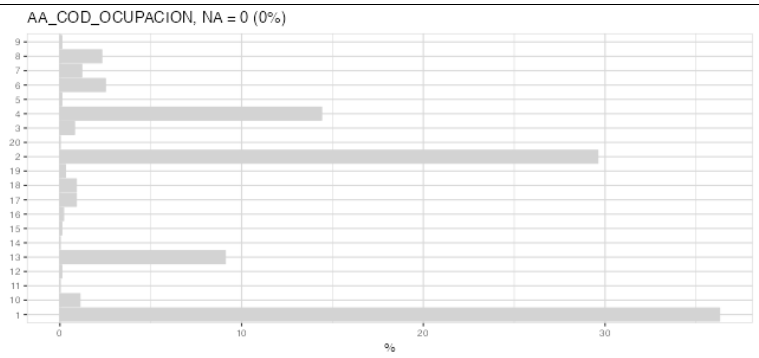
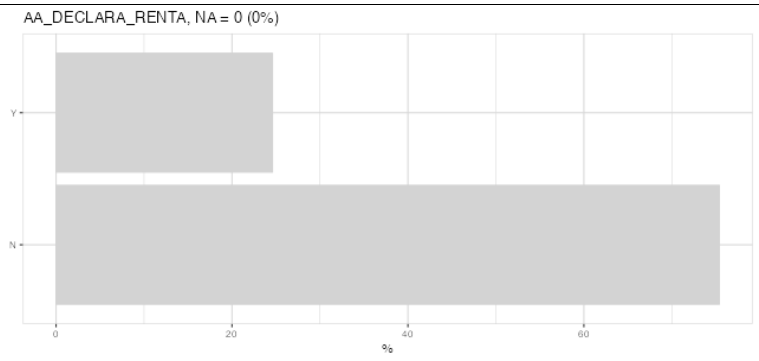
variable	type	na	na_pct	unique	min	mean	max
AA_TIPO_VIVIENDA	chr	9789	17.6	5			
AA_ESTRATO	dbl	0	0	7	0		6
AA_DECLARA_RENTA	chr	0	0	2			
AA_COD_OCUPACION	chr	0	0	20			
AA_COD_CIIU	chr	0	0	391			

Tabla 5: Exploración de datos para variables de interés

variable = AAHO_REGION type = double na = 0 of 55 772 (0%) unique = 9 2 = 8 029 (14.4%) 4 = 8 616 (15.4%) 5 = 8 814 (15.8%) 6 = 4 171 (7.5%) 7 = 8 255 (14.8%) 8 = 7 007 (12.6%) 14 = 9 736 (17.5%) 31 = 823 (1.5%) 35 = 321 (0.6%)	
variable = AAHO_INAC_MOVIMIEN type = double na = 0 of 55 772 (0%) unique = 2 0 = 44 764 (80.3%) 1 = 11 008 (19.7%)	
variable = SEX type = character na = 0 of 55 772 (0%) unique = 2 F = 24 132 (43.3%) M = 31 640 (56.7%)	

variable = NIVEL_EDUCATIVO type = character na = 572 of 55 772 (1%) unique = 7 0 = 2 792 (5%) 1 = 10 988 (19.7%) 2 = 12 198 (21.9%) 3 = 5 768 (10.3%) 4 = 14 675 (26.3%) 5 = 8 779 (15.7%) NA = 572 (1%)	
variable = MAR_STATUS type = character na = 580 of 55 772 (1%) unique = 8 C = 22 261 (39.9%) D = 309 (0.6%) O = 766 (1.4%) P = 2 280 (4.1%) S = 18 691 (33.5%) U = 7 442 (13.3%) V = 3 443 (6.2%) NA = 580 (1%)	
variable = PER_CARGO_MY18 type = character na = 0 of 55 772 (0%) unique = 16 0 = 42 857 (76.8%) 1 = 8 145 (14.6%) 10 = 3 (0%) 11 = 1 (0%) 13 = 1 (0%) 15 = 1 (0%) 16 = 1 (0%) 2 = 3 384 (6.1%) 20 = 1 (0%) 3 = 1 010 (1.8%) ...	

variable = PER_CARGO_MN18 type = character na = 0 of 55 772 (0%) unique = 12 0 = 40 100 (71.9%) 1 = 9 068 (16.3%) 10 = 4 (0%) 11 = 4 (0%) 2 = 4 989 (8.9%) 3 = 1 209 (2.2%) 4 = 278 (0.5%) 5 = 70 (0.1%) 6 = 36 (0.1%) 7 = 7 (0%) ...	
variable = AA_TIPO_VIVIENDA type = character na = 9 789 of 55 772 (17.6%) unique = 5 A = 5 506 (9.9%) F = 13 643 (24.5%) O = 195 (0.3%) P = 26 639 (47.8%) NA = 9 789 (17.6%)	
variable = AA_ESTRATO type = double na = 0 of 55 772 (0%) unique = 7 0 = 2 619 (4.7%) 1 = 2 142 (3.8%) 2 = 21 365 (38.3%) 3 = 16 061 (28.8%) 4 = 7 832 (14%) 5 = 3 434 (6.2%) 6 = 2 319 (4.2%)	
variable = AA_COD_CIIU type = character na = 0 of 55 772 (0%) unique = 391 0010 = 35 671 (64%) 0081 = 51 (0.1%) 0090 = 1 529 (2.7%) 0111 = 22 (0%) 0112 = 28 (0.1%) 0113 = 32 (0.1%) 0114 = 2 (0%) 0115 = 6 (0%)	

0119 = 21 (0%) 0121 = 25 (0%) ...	
variable = BB_SEG_COMERCIAL type = character na = 0 of 55 772 (0%) unique = 20 1544 = 3 186 (5.7%) 156 = 220 (0.4%) 157 = 15 459 (27.7%) 1574 = 92 (0.2%) 1578 = 1 371 (2.5%) 1579 = 8 418 (15.1%) 1580 = 7 (0%) 1581 = 9 417 (16.9%) 1582 = 8 952 (16.1%) 1583 = 1 077 (1.9%) ...	
variable = AA_COD_OCUPACION type = character na = 0 of 55 772 (0%) unique = 20 1 = 20 227 (36.3%) 10 = 619 (1.1%) 11 = 1 (0%) 12 = 78 (0.1%) 13 = 5 076 (9.1%) 14 = 1 (0%) 15 = 43 (0.1%) 16 = 108 (0.2%) 17 = 490 (0.9%) 18 = 506 (0.9%) ...	
variable = AA_DECLARA_RENTA type = character na = 0 of 55 772 (0%) unique = 2 N = 42 059 (75.4%) Y = 13 713 (24.6%)	

6. DESARROLLO Y ANÁLISIS DE MODELOS

En este capítulo se evidencia el desarrollo de las predicciones de saldos con las técnicas seleccionadas acorde con el objetivo específico: Analizar y seleccionar los métodos y técnicas que permitan garantizar el desarrollo, análisis y evaluación del modelo predictivo.

Mientras que los mercados están cada vez más competitivos y las expectativas de los clientes son más difíciles de satisfacer, es indispensable conocer de primera mano las necesidades y posibles factores que se deberán contemplar para obtener la eficiencia en el uso de los productos en la entidad bancaria, por lo cual, se requiere que la información sea confiable y tener claridad de la conducta de los productos, ya que hoy en día la analítica de datos se ha vuelto un factor imperante para poder predecir el comportamiento del cliente para las empresas que manejan este tipo de servicios y productos financieros.

A continuación, se mostrarán las diferentes técnicas, métodos y/o modelos usados para el objeto de estudio y su respectiva simulación según sus métricas, esto diseñado para conocer el manejo de cada una de las variables en el proceso de predicción de saldos de productos de captaciones de la entidad bancaria, pues aquí, se podrá observar cómo cada uno de los modelos contribuirá más que otro o la suma de todos a mejorar la eficiencia de los resultados según características evidenciadas al agruparlas por las regionales a nivel nacional donde se emiten estos productos.

6.1. ESTÁNDARES DE DESARROLLO

- **Objetivo:** Establecer los estándares definidos para desarrollar soluciones para mejorar la toma de decisiones por medio de modelos de predicción y/o pronóstico, así mismo, incluir las buenas prácticas y asegurar la consistencia del desarrollo, documentación y entendimiento de los procesos desarrollados logrando una mayor eficiencia en el mantenimiento y escalabilidad de estos.

- **Alcance:** Los estándares definidos aplican para todos los nuevos requerimientos sobre las tecnologías manejadas, así como el ajuste de los procesos ya existentes.

- **Manejo de ambientes:** A continuación, se presentan los aspectos para tener en cuenta

correspondiente al uso de los diferentes ambientes.

- Todas las soluciones que impliquen desarrollo deben construirse sobre servidores, herramientas y bases de datos establecidas para ambientes de desarrollo de acuerdo con los lineamientos de arquitectura TI en la entidad bancaria.
- Todos los desarrollos deben pasar por un proceso de validación y calidad con el fin de garantizar su funcionamiento y la correcta construcción del Software, para esto debe ser desplegado el paquete a validar en ambientes que se tienen para QA.
- El acceso por parte de los desarrolladores a las bases de datos para consulta o validaciones se debe realizar utilizando los usuarios definidos y/o personales según sea el caso para cada ambiente:
  Para Producción: Usuario Personal
  Para QA: Usuario Genérico
  Para Desarrollo: Usuario Genérico
- Los usuarios de aplicación en ambiente de producción solo deben ser utilizados por los procesos productivos (ETL y Gestión de Reportes), no está bien que se utilice un usuario productivo para temas de desarrollo y/o validaciones de actualizaciones (Updates), ya que pueden cambiar la lógica en el modelo de datos.
- Todos los objetos de bases de datos y/o archivos creados en fases de Desarrollo y QA como parte de pruebas propias del proceso, deben ser depurados o truncados una vez el proceso sea puesto en producción, esto con el fin de no cargar las plataformas de desarrollo con información obsoleta (tener en cuenta que el espacio es limitado).
- Dentro de las tareas de cada desarrollo se debe ejecutar la generación de su correspondiente Backup incluyendo documentación, artefactos y configuraciones, de manera recurrente (se recomienda diariamente), con el objetivo de evitar la pérdida de información.

6.1.1. LINEAMIENTOS GENERALES PARA LA CREACIÓN DE OBJETOS DE BASES DE DATOS

- Antes de crear cualquier objeto final (fact, dim, tbl), se debe garantizar que en el DWH ya no existe un objeto que cumpla con el mismo objetivo o en su defecto si se puede complementar el objeto existente, con el fin de evitar la redundancia de información o tener diversos objetos para un mismo concepto.
- Los objetos que hagan parte del modelo final de la solución (Fact y dimensiones), debe

crearse en la base de datos asignada a la zona de DWH.

- Los objetos de tipo temporal deben ser creados en la base de datos asignada a la zona de Stage y estos deben ser borrados o truncados al final de la ejecución del proceso (no debe permanecer ningún objeto temporal con información de los procesos luego de su ejecución).
- Los objetos que no hagan parte de modelos, pero tampoco hacen parte de temporales, ej. Sábanas de datos o Bases de información para un fin específico, deben ser creados en la base de datos Operacional.
- El nombre de los objetos debe describir claramente su contenido con el fin de que se identifique fácilmente su uso y objetivo.
- Evitar en el nombramiento de los objetos el uso de artículos (el, la, los, las, un, una, etc.) o preposiciones (de, desde, en, con, entre, etc.) ni conjunciones (y, o, etc.).
- Para la separación de las palabras que componen el nombre de los objetos sobre la base de datos debe utilizarse el carácter underline "_".
- A nivel de BD, los objetos deben tener un comentario breve indicando su proceso y su fuente.
- El idioma definido para nombrar los objetos y sus atributos es el español, a excepción de los casos en donde para facilitar el entendimiento se considere el uso del inglés.
- Los nombres de las tablas de hechos deben ser descriptivos, en lo posible usar palabras completas del hecho representado.
- El nombre de las dimensiones debe ser descriptivo, en singular, en lo posible usar palabras completas de las entidades representadas, así como su correspondiente tipo (T0, T1, T2).

6.1.2. HERRAMIENTA DE INTEGRACIÓN

Las siguientes herramientas son las utilizadas y vigentes para el desarrollo de procesos de integración de datos (ETL / ELT /EL).

Herramienta	Versión	Objetivo	Descripción
SSIS	2010	EL	Integrador para extraer información desde las diferentes fuentes y/o File Server hacia zona de Stage.

SSIS	2017	ETL	Procesamiento de todo el flujo, lógica y reglas de negocio del desarrollo partiendo desde zona de Stage hacia la consolidación en el DWH.
DATASTAGE	11.5	ETL	Integrador para la construcción de procesos completos de ETL.

6.2. SSIS – ESTÁNDARES DE DESARROLLO

A continuación, se presentan los estándares de desarrollo a tener en cuenta si se requiere trabajar con SQL Server Integration Services

6.2.1. NOMBRAMIENTO DE PROYECTOS Y PAQUETES

El nombramiento del Proyecto será el igual al código de servicio o nombre de asignado desde el inventario de procesos de GBI.

Ej.

0001_CLI_MODELO_MORA

Para el nombramiento de paquetes se debe tener en cuenta la siguiente nomenclatura según sea el caso.

Paquete padre o principal → PP_<NÚMERO_PROCESO>_<ACCION_A_REALIZAR>

Paquetes Hijos → PH_<CÓDIGO_PROCESO>_<ACCION_A_REALIZAR>

Ej.

PP_0001_GENERA_MODELO

PH_0001_EXTRACCIONES_ODS

PH_0001_CARGA_DIMENSIONES

PH_0001_CONSOLIDA_FACT

6.2.2. MANEJO DE CONEXIONES

- Todas las conexiones en SSIS deben ser creadas de tipo Proyecto.
- El nombramiento de las conexiones se debe manejar teniendo en cuenta el siguiente estándar:

Prefijo: CNX_
Tipo de Conexión: (ORC, SQL, EXCEL, Para archivos su extensión (TXT, CSV, XML))
Nombre Personalizado: NOMBRE_DB/SERVER
Ejemplo:

Tecnología	BD / Archivo	Nombramiento
ORACLE	CRM_DB	CNX_ORC_CRM_DB
FILE	CUENTAS_MORA_YYYYMMDD.CSV	CNX_CSV_CUENTAS_MORA
SQL SERVER	CRM_DB	CNX_SQL_CRM_DB

6.2.3. MANEJO DE PARÁMETROS Y VARIABLES

Parámetros de Proyecto

Prefijo: PARAM_
Tipo de Dato: (STR, INT, DT, DTTM, BOOL, DEC, DOUB, FLT, NUM)_
Nombre Personalizado: NOMBREPARAMETRO
Ej. Si se quiere agregar un parámetro de proyecto llamado nombre proceso de tipo String el nombre del parámetro seria el siguiente: *PARAM_STR_NOMBRE_PROCESO*

Parámetros de Paquete

Prefijo: PARAM_
Tipo de Dato: (str, int, dt, dttm, bool, dec, doub, flt, num)_
Nombre Personalizado: Nombre_Parametro (Cammel Case)
Ej. Si se quiere agregar un parámetro de paquete llamado nombre proceso de tipo String el nombre del parámetro seria el siguiente: *PARAM_str_Nombre_Proceso*

Variables de Paquete

Prefijo: VAR_
Tipo de Dato: (str, int, dt, dttm, bool, dec, doub, obj)_
Nombre Personalizado: Nombre_Variable (Cammel Case)
Ej. Si se quiere agregar una variable llamada nombre empresa de tipo String el nombre de la

variable seria el siguiente: *VAR_str_Nombre_Empresa*

6.3. DATASTAGE – ESTÁNDARES DE DESARROLLO

Los estándares definidos aplican para todos los nuevos requerimientos sobre Datastage así como el ajuste de los procesos ya existentes.

6.3.1. ESTRUCTURA DE LA SOLUCIÓN

Los Jobs/Secuencias del proceso deben ser creados en el Proyecto, bajo la carpeta "Jobs", según su dominio del proceso y código de servicio del proceso (teniendo en cuenta el inventario de procesos).

BIBA → Jobs → <DOMINIO> → <CODIGO_SERVICIO>

Ejemplo:

BIBA → Jobs → 01_CLIENTES → 0001_CLI_MODELO_MORA

6.3.2. MANEJO DE CONEXIONES

Todas las conexiones a las diferentes fuentes de información que utilicen los procesos en DataStage deben ser creadas en la carpeta "Data Connection" dentro del proyecto. Debe considerar que si existe una conexión a la fuente que requiere debe utilizar la existente, no debe existir más de una conexión para la misma fuente de información. El componente a utilizar debe ser de tipo "Conexión de Datos".

Para el nombramiento de las conexiones se debe manejar el siguiente estándar:

CNX_<ABREV_TECNOLOGÍA>_<NOMBRE_FUENTE>
Ej. *CNX_ORA_MDM*

6.3.3. NOMBRAMIENTO DE OBJETOS

Carpeta de Jobs para los procesos

Se debe manejar una carpeta para agrupar los diferentes tipos de Jobs con el fin de organizar mejor el proyecto según objetivo:

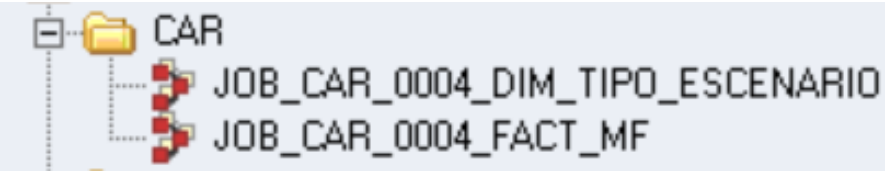
- **EXT**: Trabajos para extracción desde la fuente a Stage
- **TRA**: Trabajos para transformación de datos (estandarización, validación de duplicados, formatear los datos según solicitud).
- **CAR**: Trabajos para cargue de información
- **ACT**: Trabajos para actualización de datos
- **LIM**: Trabajos de limpieza o borrado de objetos (BD o FS)

Nomenclatura para Jobs:
Considerar las acciones indicadas para establecer su correspondiente prefijo de esta.

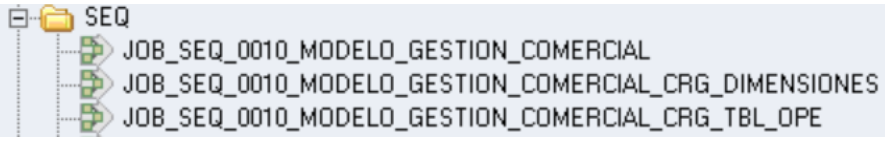
JOB_<ACCIÓN>_<COD_SERVICIO>_< DESCRIPCIÓN_OBJETIVO>
Cuando
Acción = EXT, TRA, CAR, ACT, LIM

Código Servicio = El código asignado desde el inventario de procesos. ej. 0001, 0002, etc.

Ejemplos:



Nomenclatura para Secuencias:
SEQ_<COD_SERVICIO>_< DESCRIPCIÓN_OBJETIVO>
Ejemplos:



6.4. SSAS – ESTÁNDARES DE DESARROLLO EN GBI

Los estándares definidos aplican para todos los nuevos requerimientos sobre SSAS así como el ajuste de los procesos ya existentes.

Nombramiento proyecto

Separado por guion bajo "_":

<COD_SERVICIO>_<ABREV_PROCESO>_CUBO_<NOMBRE_MODELO>

Ejemplo:

0001_CLI_CUBO_MORA

Data Sources

CNX_<ABREV_TECNOLOGÍA>_<NOMBRE_FUENTE>

Abreviaturas:

Oracle → ORC

SQL Server → SQL

Ejemplo:

CNX_ORC_PDBIDW

Nombramiento Vistas

Se recomienda contar con una vista de origen de datos por cada tabla de hechos con sus dimensiones, para mejorar la organización y diseño de la solución.

Separado por espacio:

VW <Nombre modelo>

Ejemplo:

VW Mora

Nombramiento Dimensiones

Separado por espacio:

Dim <Nombre dimensión>

Ejemplo:

Dim Tiempo

Nombramiento Cubos

Separado por espacio:

Cubo <Nombre Cubo>

Ejemplo:

Cubo Mora

6.5. TÉCNICAS Y/O MODELOS PARA ANÁLISIS DE SERIES DE TIEMPO

De acuerdo con la necesidad de la comparación de efectividad y eficacia en el uso de ciertos modelos de predicción, es indispensable traer a colación el significado de un Pronóstico el cual *"Es una estimación cuantitativa o cualitativa de uno o varios factores (variables) que conforman un evento futuro, con base en información actual o del pasado"* por ende, en el objeto de estudio es razonable tener presente que los patrones de saldos de productos de captaciones podrán seguir ocurriendo en el futuro. Por lo cual, se elaborará un pronóstico con métodos de Series de Tiempo, ya que los datos los históricos se restringen a valores pasados de la variable que estamos pronosticando, extrapolando sus valores, la cual es el Saldo de Captación según la Regionales a nivel nacional. [32] Dentro de este orden de ideas, se establecen las siguientes técnicas y/o modelos para el análisis por Series de Tiempo:

6.5.1. PROPHET W/ REGRESSORS

En el pronóstico de datos por series de tiempo, se encuentra el modelo Prophet el cual es similar a un modelo aditivo, es decir, que los efectos de factores individuales se agrupan para el modelamiento de los datos, por lo cual las tendencias no lineales se ajustan a las estacionalidades en tiempo, lanzado por el equipo Core Data Science de Facebook.

Este modelo se ajusta en el objeto de estudio, ya que se verifica la proyección de saldos basado en la estacionalidad mensual. Además, tiene un punto a favor al ser resistente en los cambios de tendencia, al tener tolerancia gradual en datos faltantes y manejar correctamente los valores atípicos en los datos históricos. El Algoritmo Prophet se ajusta a la tendencia, la estacionalidad y los días festivos, es decir, detecta automáticamente el cambio de la función de tendencia línea o de crecimiento, modela bajo las series de Fourier y se puede decidir sí tomar en cuenta los festivos en el proceso de modelamiento cuando se evalúa las variables bajo series de tiempo. Adicionalmente, de acuerdo con [6] estos parámetros se combinan de la siguiente forma:

$$y(t) = g(t) + s(t) + h(t) + e(t)$$

- **y(t)** - es el pronóstico que se está evaluando.

- **g(t)** - se refiere a la tendencia *(cambios durante un largo período de tiempo)*

- **s(t)** – entendiéndolo como la estacionalidad *(cambios periódicos o de corto plazo)*

- **h(t)** - son los efectos de los días festivos o atípicos, esto es para cuando se realiza un pronóstico de ventas.

- **e(t)** - se refiere a los cambios incondicionales que son específicos de una empresa, una persona o una circunstancia. También se le llama *término de error.*

En términos de precisión y velocidad este algoritmo tiene un rango muy alto, ya que tiene fácil descomposición, logrando extraer los coeficientes de cada una de las componentes del modelo.

6.5.2. GLMNET

Una de las técnicas que se ajusta a modelos lineales generalizados es el GMLNET, este es un algoritmo rápido, el cual se adapta a modelos de regresión lineal, logística, multinomial, de Poisson de Cox. El GMLNET es una técnica que incluye métodos para predicción y trazado, y funciones para validación cruzada, sus autores son Jerome Friedman, Trevor Hastie, Rob Tibshirani, Balasubramanian Narasimhan, Kenneth Tay y Noah Simon, con la contribución de Junyang Qian.

Adicionalmente, este modelo calcula la máxima verosimilitud penalizada, utiliza el descenso cíclico de coordenadas hacia la convergencia, puede manejar formatos de matriz de entrada dispersos, así como restricciones de rango en los coeficientes, este modelo GLMNET maneja un conjunto de subrutinas de Fortran, que permiten una ejecución muy rápida. Según [33] este modelo resuelve:

$$\min_{\beta_0,\beta} \frac{1}{N} \sum_{i=1}^{N} w_i l(y_i, \beta_0 + \beta^T x_i) + \lambda \left[(1-\alpha)\|\beta\|_2^2/2 + \alpha\|\beta\|_1\right]$$

- λ: Los valores de lambda cubren la gama de posibles soluciones, por ende, controla la fuerza general de la penalización.

- l(y, η): Contribución de probabilidad logarítmica negativa para la observación i, es decir, para el caso Gaussiano seria $\frac{1}{2}$ (y – η)²

- α: Controla la penalización neta elástica.

Se resalta que este modelo usa en su paquete de funcionamiento modelo lineal Gaussiano o "mínimos cuadrados".

6.5.3. XGBOOST

XGBoost es la abreviatura de las siglas "Extreme Gradient Boosting" lo que significa refuerzo de gradientes extremo, siendo este un método de aprendizaje automático supervisado para clasificación y regresión. Es importante resaltar que este modelo se basa en árboles de decisión y supone una mejora sobre otros métodos, maneja un ensamblado en la clasificación como bosque aleatorio y refuerzo de gradientes.

Para la clasificación del modelo, los valores se calculan generalmente utilizando el registro de momios y probabilidades, las salidas de los árboles de decisión se convierte en las nuevas entradas del dataset, este proceso se repite hasta que los residuales dejan de reducirse, por medio de manejo de valores perdidos y regularización para evitar sesgos en su uso. Uno de los puntos a favor de este modelo, es que el usuario define la extensión de los árboles, haciendo alusión a lo indicado por [34] el algoritmo XGBOOST funciona así:

- Se obtiene un árbol inicial F_0 para predecir la variable objetivo "y", el resultado se asocia con un residual (y - F_0).

- Se obtiene un nuevo árbol h_1 que ajusta al error del paso previo.

- Los resultados de F_0 y h_1 se combinan para obtener el árbol F_1, donde el error cuadrático medio de F_1 será menor que el de F_0:

$$F_1(x) < -F_0(x) + h_1(x)$$

- Este proceso se sigue iterativamente hasta que el error es minimizado lo más posible de la siguiente forma:

$$F_m(x) < -F_{m-1}(x) + h_m(x)$$

Es importante resaltar que una de las ventajas del modelo XGBOOST al igual que el modelo Prophet puede manejar valores perdidos y/o atípicos y sus resultados en términos de precisión y velocidad, dado el caso el modelo de refuerzo de gradientes extremo utiliza parámetros como Regularización, Corte, Boceto de cuantil ponderado, Aprendizaje paralelo, Búsqueda de divisiones sensible a la escasez, Acceso sensible al caché y Bloques de cómputo fuera de núcleo que ayuda a optimizar los resultados del modelo a un mayor rendimiento.

6.5.4. RANDOM FOREST – RANGER

RANGER es una de las técnicas específicas para el modelo Random Forest *(Bosque Aleatorio),* el cual crea gran cantidad de árboles de decisión para la clasificación de las predicciones primero de forma individual y luego combina todas las predicciones adaptándose no solo a la clasificación sino también a las regresiones.

RANGER es una de las formas en las cuales se ajusta el modelo, pues este método de estimación y predicción permite la combinación de las decisiones, por dos modos: clasificación y regresión, manejando parámetros de ajuste:

- **mtry:** Número de Predictores seleccionados aleatoriamente (tipo: entero, predeterminado: ver más abajo) dependiendo del número de columnas.

- **trees:** Número de Árboles (tipo: entero, predeterminado: 500L)

- **min_n:** Tamaño mínimo de nodo (tipo: entero, predeterminado: ver más abajo) dependiendo del modo, es decir, para la regresión es valor predeterminado es 5, pero para la clasificación es 10.

RANGER es una implementación rápida de los modelos Random Forest, lo cual permite:

- Iniciar con un conjunto de entrenamiento que tiene n observaciones
- La variable de interés Y
- Variables predictoras $X_1, X_2 \dots X_p$

Donde sí se quiere predecir la variable de interés Y para un caso en el que se tiene la información de las variables predictoras $X_1, X_2 \dots X_p$ para un periodo de tiempo determinado, se deben tomar cada uno de los árboles B creados para predecir la variable Y, así se tendrán las diferentes predicciones $\hat{Y}_1, \hat{Y}_2, \dots, \hat{Y}_B$ para tener una predicción combinada ya sea de clasificación o regresión. Así se unifican las predicciones, dependiendo del modo:

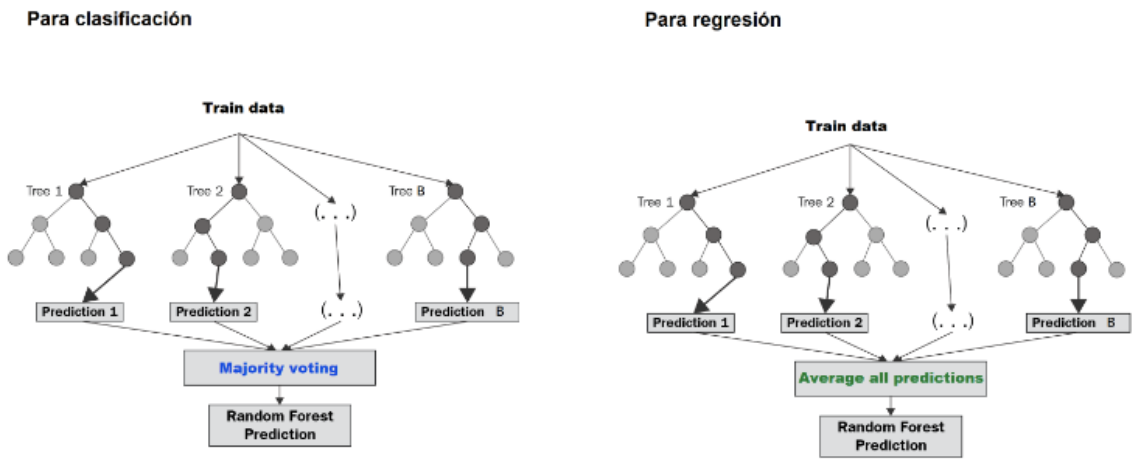


Ilustración 8: Unificación de las B predicciones con el motor RANGER del modelo Random Forest. Fuente: https://fhernanb.github.io/libro_mod_pred/rand-forests.html

Este motor se ajusta para manejar valores atípicos hacer agrupaciones, ayudando que al utilizar varios métodos de optimización como los anteriormente mencionados, se puedan producción resultados eficaces en la predicción, facilitando una estimación de los valores con mayor importancia en la clasificación.

6.5.5. KERNLAB

KERNLAB es un motor que se ajusta a un modelo de máquina de vectores de soporte, para la clasificación, el modelo intenta maximizar el ancho del margen entre clases, pero para el otro modo que es la regresión, el modelo optimiza una función de pérdida robusta que solo se ve afectada por residuos del modelo muy grandes, además tiene la ventaja de que se usa código optimizado para calcular varias expresiones del núcleo, manejando parámetros de ajuste:

- **cost:** Costo (tipo: doble, predeterminado: 1.0)

- **rbf_sigma:** Función de base radial sigma (tipo: doble, predeterminado: ver a continuación)

- **margin:** Margen de insensibilidad (tipo: doble, predeterminado: 0,1)

De acuerdo con [35] en el caso de una función “kernel” de base radial (Gaussiana) también se puede establecer en la cadena "automático", KERNLAB lo estima a partir de los datos utilizando método heurístico en 'sigest' para calcular un buen valor 'sigma' para el RBF gaussiano o el kernel de Laplace, a partir de los datos (predeterminado = "automático"), pues no existe un valor predeterminado para el parámetro del núcleo de la función de base radial, en resumen, este método utiliza números aleatorios, por lo que, sin establecer la semilla antes del ajuste, el modelo no será reproducible.

Se debe tener en cuenta que el motor KERNLAB no estima naturalmente las probabilidades de clase, Para producirlos, los valores de decisión del modelo se convierten en probabilidades utilizando la escala de Platt. Por lo cual, las probabilidades en la salida del modelo se obtienen una para cada modelo creado en la clasificación de pares:

$$k\frac{(k-1)}{2}$$

El método de pareja o acoplamiento por pares permite discriminar entre pares de clases y selecciona la clase con mayor cantidad de decisiones en la clasificación, pues implementa varias técnicas para combinar estas probabilidades, manteniendo el uso de técnicas de descomposición y solución de la clasificación multiclase.

6.5.6. PROPHET W/ XGBOOST ERRORS

Al crear un modelo integrado basado en la combinación de los métodos Prophet y Extreme Gradient Boost (XGBoost), genera un modelo hibrido que integra los dos algoritmos únicos en función de sus especialidades. Prophet se aplica al análisis de series temporales según los valores monetarios de los saldos históricos para caracterizar su impacto en la periodicidad de la serie en un periodo mensual, y XGBoost, basado en la optimización bayesiana mejorada (BayesOpt), se utiliza para describir los efectos de los saldos por las variables de las regionales.

El modelo híbrido se construye mediante un nuevo método de ponderación optimizado. El rendimiento del modelo propuesto se compara con el de los modelos individuales, logrando la predicción en base a la tendencia (Prophet) y la estacionalidad (XGBoost) modelando errores residuales.

Dado que estamos trabajando con datos mensuales, no es necesario calcular la estacionalidad diaria ni semanal. Por lo tanto, establecemos los parámetros correspondientes en FALSO en la especificación del modelo para ambos modelos, manteniendo el modelado bajo las series de Fourier. Lo anterior crea una forma de generar una especificación de un modelo PROPHET potenciado antes del ajuste y permite crear el modelo utilizando diferentes paquetes, así pues, el modo siempre será “Regresión”.

6.6. MÉTRICAS USADAS EN LA EVALUACIÓN DE LOS MODELOS

Es imperativo recalcar que, en el proceso de pronóstico, el pronóstico de datos usando métodos de Series de Tiempo reunidos como es el caso se produce cuando: ***1.*** Inicialmente se utilizan diferentes parámetros y especificaciones que tienen una interpretación humana directa, seguido de esto ***2.*** El rendimiento de la predicción se evalúa en el modelo y, si surge algún problema *(rendimiento deficiente),* los resultados del modelo hacen que el cliente *(usuario del área)* deba intervenir para realizar un reentrenamiento del modelo, finalmente ***3.*** El analista de datos puede ajustar el modelo correctamente en función del feedback recibido por el cliente. [36]

Sin embargo, es necesario evaluar no solo los modelos por los índices de predicción que puedan otorgar, sino también validar los resultados generados para realizar la predicción con modelos estadísticos guiados por error, por lo cual, es válido verificar las diferentes métricas usadas para series de tiempo, ya que en los pronósticos se pueden presentar errores eventualmente, estas medidas también ayudan a tomar la decisión del modelo que mejor pronostica los valores esperados, algunas de estas son:

6.6.1. MAE

El error absoluto medio el cual es definido como la diferencia entre el pronóstico y el valor real en los resultados esperados, es una medida común del error de pronóstico en análisis de series de tiempo, puede tener un sesgo hacia ítems de mayor volumen y normalmente es inadecuado para medir ítems con baja demanda.

Se calcula siendo y_t igual al valor real y f_t igual al valor predicho en el periodo t., su fórmula es:

$$MAE_t = \frac{\sum_{t=1}^{n} |y_t - f_t|}{N}$$

Teniendo en cuenta lo indicado por [37] algunas consideraciones al utilizar el MAE son que este arroja un número en las mismas unidades que la variable de salida, es decir, para el caso de estudio la diferencia absoluta en los saldos, por lo que es fácil de interpretar cuando trabajamos con un producto como este, donde la métrica depende de la magnitud de los saldos estudiados.

6.6.2. MAPE

El MAPE es el error de porcentaje medio absoluto, es decir, entrega la desviación en términos porcentuales y no en unidades como las anteriores medidas que son de las variables, por lo cual, es el promedio del error absoluto o diferencia entre el riesgo real y el pronóstico, indicado como un porcentaje de los valores reales.

$$M = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

El error de porcentaje medio absoluto maneja sus valores en este tipo para evitar que los errores positivos y negativos se cancelen entre sí y utiliza errores relativos para permitirle comparar la precisión de previsión entre métodos de serie de tiempo.

6.6.3. MASE

El error de escala absoluta media o MASE, fue propuesto por Hyndman y Koehler [38] como una medida general de error al realizar una predicción de series temporales, esta métrica no depende de la escala de los datos y tampoco presenta problemas con las predicciones nulas como es el caso de la métrica anterior "MAPE" el cual se indefine en esta situación, para calcular esta medida se utiliza la siguiente formula:

$$e_i = |\hat{y}_i - y_i|$$

$$q_i = \frac{e_i}{\frac{1}{n-1} * \sum_{t=2}^{n}|y_t - y_{t-1}|}$$

$$MASE = \frac{1}{n}\sum_{i=1}^{n} q_i$$

La métrica MASE es menos sensible a valores atípicos, se puede utilizar para series de datos debido a la ocurrencia de valores infinitos e indefinidos, de acuerdo con [39] cuando el resultado de esta métrica es mayor que uno, indica que las estimaciones son peores, en promedio, que dentro de la muestra de un solo paso por estimaciones del método o modelo ingenuo.

6.6.4. SMAPE

Por sus siglas en ingles *Symmteric Mean Absolute Percent Error, la* métrica del error medio absoluto simétrico porcentual propuesto en la mayoría de las evaluaciones, se evalúa por:

$$SAMPE_s = \frac{1}{n}\sum_{i=1}^{n}\frac{|\mathrm{X}_t - F_t|}{(\mathrm{X}_t + F_t)/2} * 100$$

El criterio para evaluar la calidad de predicción del filtro propuesto se realiza a través del índice SMAPE, pues conforme a lo descrito por [40] donde, t es la medición del tiempo, n es el tamaño del conjunto de la prueba, s cada serie temporal, X_t y F_t son los reales y los valores de tiempo previsto serie en el tiempo t, respectivamente. El SMAPEs de cada serie se calcula el error simétrico absoluto en porcentaje entre el real X_t y su correspondiente valor pronóstico, en todos las t observaciones del tamaño de muestra n para cada serie de tiempo s.

6.6.5. RMSE

La raíz del error cuadrático medio (Root-mean-square error, RMSE) es una métrica que mide la distancia entre los valores predichos y los valores observados o reales en un modelo de regresión, es decir, de los valores predichos $\hat{y}_i$ para i veces la regresión de la variable dependiente y_i con variables observadas n veces se calcula para n diferentes predicciones como la raíz cuadrda de la media de los cuadrados de las desviaciones, por lo cual, esta métrica se calcula como:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}$$

- $\hat{y}_i$: Valor predicho para la i-ésima observación en el conjunto de datos
- y_i: Valor observado para la i-ésima observación en el conjunto de datos
- n: Tamaño de la muestra

De acuerdo con [41] se calcula esta métrica en R haciendo uso de la función rmse(), a la cual se le pasa como parámetros los valores observados y los valores predichos del conjunto de test, que es aquel sobre el que se ha realizado la predicción.

6.6.6. RSQ

El coeficiente de determinación (R2 or R-squared) puede interpretarse como la proporción de la varianza de la variable dependiente que es predecible a partir de las variables independientes. En conformidad con [42] esta métrica explica cuánta variabilidad de un factor puede ser causada por su relación con otro factor relacionado. Esta correlación, conocida como "bondad de ajuste", se representa como un valor entre 0 y 1,0, por eso en sus valores el peor valor = 0 donde indica que el modelo no explica la variabilidad de los datos y el mejor valor =1 indica un ajuste perfecto.

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(X_i - y_i)^2}{\sum_{i=1}^{m}(Y - y_i)^2}$$

Las métricas usadas en la evaluación de los modelos permiten evaluar y comparará los modelos de series de tiempo y regresión propuestos y analizados en el proceso de modelamiento, utilizando los datos temporales de los saldos para predicción del saldo captación de clientes. Adicionalmente, ya que el modelo tiene dichas variables (Saldos de captación y Fechas), la primera entrada da orden a la serie temporal por el tiempo determinado, y la segunda, es la generación de agrupaciones por Regionales a nivel nacional para obtener el saldo de captaciones por regional.

Al no llegar a un consenso sobre una métrica estándar para las evaluaciones de los modelos, en los análisis de regresión se han empleado estos estudios de aprendizaje automático, permitiendo que se puedan evaluar entre sí, revisando la mejor predicción de acuerdo con su margen de error, por lo cual los datos de las métricas evaluadas en los modelos empleados en el caso de estudio fueron las siguientes:

Tabla 4: Métricas de error según cada tipo de modelo de predicción en series de tiempo utilizado. fuente: Autores

.MODEL_ID	.MODEL_DESC	.TYPE	MAE	MAPE	MASE	SMAPE	RMSE	RSQ
1	PROPHET W/ REGRESSORS	Test	1376,528	226,064	0,192	41,753	1558,497	0,983
2	GLMNET	Test	1563,851	249,618	0,218	44,055	1751,129	0,980
3	XGBOOST	Test	472,117	8,244	0,066	7,644	763,435	0,989
4	RANGER	Test	924,702	139,324	0,129	33,717	1138,782	0,986
5	KERNLAB	Test	1189,149	159,866	0,166	37,908	1332,501	0,981
6	PROPHET W/ XGBOOST ERRORS	Test	1201,463	86,172	0,167	33,044	1445,756	0,981

7. VISUALIZACIÓN DE RESULTADOS

En este capítulo se hace hincapié al cumplimiento del objetivo específico: Utilizar una herramienta de visualización para evidenciar y exponer los resultados del modelo predictivo a la entidad bancaria.

Es imperativo en cualquier proceso de aplicación con métodos y técnicas de analítica tener la validación y confirmación de los resultados obtenidos, ya que estos procedimientos están estrechamente relacionados con el mundo real en cada sector donde se implemente. Por lo cual, para exponer los resultados del modelo en visualizaciones de fácil acceso y entendimiento para los usuarios finales se utiliza el software **Power BI.** En la entidad bancaria objeto de estudio, se utiliza el escenario BI Empresarial y están en la implementación del escenario BI de autoservicio administrado. Para el uso del escenario BI Empresarial, se cuenta con una arquitectura de desarrollo de flujos de trabajo en Sql Server Integration Services, Datawarehouse en Oracle, modelo de datos y capa semántica en Sql Server Analysis Services.

Power BI más que una herramienta es una suite de soluciones diversas enfocadas en la inteligencia de negocio, de acuerdo con [43] esta herramienta de Business Intelligence permite realizar **procesamiento de grandes conjuntos de datos de una manera sencilla, ya que se puede formular, consultar** y analizar los datos en el momento, forma y cantidad que precisan los analistas de datos, por lo cual, el data set que se cree como insumo podrá evaluarse de tal forma que permita la interacción del usuario de manera precisa para contribuir con la toma de decisiones según las necesidades del negocio.

7.1. SELECCIÓN DE LA HERRAMIENTA DE VISUALIZACIÓN

Se selecciona Power BI, entre otras herramientas, por las siguientes razones:

- Interfaz intuitiva y fácil de usar
- Integración con una amplia gama de fuentes de datos
- Herramientas de limpieza de datos y transformación de datos, como Power Query
- Amplias opciones de visualización de datos y capacidad para crear visualizaciones personalizadas

- Potentes capacidades de análisis y modelado de datos
- Integración con otras herramientas de Microsoft, como Excel y SharePoint
- Acceso seguro a los informes y paneles en línea o fuera de línea
- Capacidad para compartir informes y paneles con otros usuarios, tanto internos como externos a la organización
- Opciones flexibles de licencias y precios para adaptarse a las necesidades de la organización
- Actualizaciones y mejoras regulares por parte de Microsoft

Es importante resaltar que Power BI tiene la capacidad para poder unificar los procesos analíticos en un solo informe o panel, se resalta la flexibilidad que tiene para la extracción de la información, además que se puede gestionar los procesos de optimización, limpieza, transformación y combinación de los datos. Es de gran ayuda que Power BI pueda tener diversos orígenes de datos tanto de entornos locales como en nube, esto puede ayudar a tener accesos desde la intranet de las compañía, resultados desde CRM (365, SalesForce, etc), para medición de conversión desde Google Analytics, desde conexión directa de Bases de Datos on-premise/local o en nube, en fin, permite convertir la información en informes gráficos interactivos con y para el usuario, que proporcionan herramientas y funciones para su personalización y tener información en tiempo real para gestionar decisiones estratégicas del negocio.

Acorde a [44] términos robustos y confortables de Power BI hacen mención a tener una construcción sólida, usable, amigable, muy cómodo para el análisis de información, dirigido a todos inclusive a los no informáticos, Power BI reúne los datos y los procesa, convirtiéndose en información clara, a menudo utilizando gráficas y tablas visualmente convincentes y fáciles de procesar, por ende, permite a los usuarios generar y compartir instantáneas claras y útiles de lo que está sucediendo en el negocio.

7.2. USO DE DATOS Y CREACIÓN DE LOS VISUALES

Se utilizan los datos proporcionados de la siguiente manera:

- Partiendo de los datos se crea la tabla calendario que nos sirve para tener la línea de tiempo.

- En Power BI utilizamos la gráfica de líneas para poder observar la tendencia de los datos históricos y la previsión.

Ilustración 9 Visualización de comportamiento historio de saldos de captación.

- Se descargan del Markeplace de Power BI dos visualizaciones, Timeline, que me sirve para filtrar los años y Pulse Chart que se utiliza para animar el comportamiento de los datos históricos.

Ilustración 10 Comportamiento dinámico de saldos de captación durante 64 meses

- Se agregan dos filtros, uno por cada Región y otro para seleccionar uno o varios modelos.

- Se agregan botones con cada uno de los modelos para que cuando se haga clic sobre uno de ellos muestre los datos históricos además de la previsión de un año con su respectivo umbral de mínimo y máximo.

Ilustración 11 Visualización de saldos captación para la parametrización realizada en negocio

8. CONCLUSIONES

El uso de los pronósticos dentro de las organizaciones día a día permite tomar decisiones informadas a las personas relacionadas con los diversos procesos organizacionales, debido a que las series temporales por su naturaleza permiten evidenciar el comportamiento o desempeño de los diversos procesos relacionados.

Por esta razón la gerencia de Business Intelligence de la entidad bancaria tiene la gran necesidad de identificar los diversos comportamiento que tienen los clientes, todo esto con fin de desarrollar estrategias que permitan el posicionamiento de la compañía, de sus productos y de identificar las principales necesidades, gustos y preferencias con la entidad bancaria.

A través del desarrollo de este proyecto logramos identificar el comportamiento de cada una de las fases con las cual se desarrollan la gran parte de los proyectos de ciencia de datos y la cual se utilizó para el desarrollo del proyecto, CRISP-DM (Cross Industry Standard Process for Data Mining) este marco de trabajo nos permitió desarrollar la lógica procedimental para desarrollar los objetivos planteados del proyecto, de esta manera se logra identificar las principales necesidades de negocio y como los datos disponibles dentro de la organización comienza hacer un activo tan importante. Si bien la gran cantidad de datos administrados por la entidad financiera puede llegar a ser tan abrumador es importante darle contexto desde el conocimiento de los principales usuarios y colaboraciones con el objetivo de saber el papel que desempeñan, estas dos actividades relacionadas permitió desarrollar la fase de extracción, transformación y carga (ETL) de los datos para realizar el procesamiento correspondiente.

La fase de modelamiento de los datos y la respectiva evaluación de los modelos seleccionados, se realizaron con el software R ya que mediante los diversos avances y colaboraciones de usuarios permite desarrollar análisis avanzados y desarrollo de canalizaciones mucho más eficientes y fluidas. Finalmente con la disposición de una herramienta visual en este caso mediante Power BI la entidad bancaria tiene la posibilidad de ver, analizar, comprender y ejecutar estrategias que permitan cumplir los objetivos organizacionales basadas en toma de decisiones informadas, Gestión de Riesgos, Planificación e innovación y desarrollo de una estrategia empresarial basada en datos.

9. TRABAJOS FUTUROS

De acuerdo con el desarrollo del proyecto, se evidencia en la etapa de ejecución, específicamente en el capítulo 6 Desarrollo y Análisis del modelo, donde se pueden identificar las diferentes mejoras a lograr en el continuo refinamiento y reentrenamiento del modelo, pues al analizar las diferentes variables puede que las métricas de error varíen en el modelo predictivo y ayudaría a tener otros puntos de vista por ejemplo no solo desde la capacidad y estrategias de captación de cada oficina a nivel nacional como se identificó en el proyecto aplicado.

Adicionalmente, se puede tener la información de los datos en mayor proporción y así mismo aumentar los algoritmos de aprendizaje, ya que así mejorara la probabilidad de rendimiento del modelo, pues se podría tomar más años a nivel histórico y más casos el modelo podrá predecir e identificar correctamente. Es importante entender qué características generan mayor impacto en las decisiones que toma el algoritmo y como ello se conecta con las necesidades de la entidad financiera para establecer las campañas estratégicas que permitan generar el aumento en los saldos de productos de captaciones de clientes.

Finalmente, se debe seguir investigando sobre el comportamiento de las variables y cómo sería su desempeño cuando pasan por el modelo de predicción, así pues, se podrían estar incurriendo en el conocimiento de qué variables son las que ayudan a focalizar las campañas segmentadas según las necesidades del cliente, por ejemplo generar estrategias de tráfico para lograr la captación de saldos de clientes por su edad, género, estrato socioeconómico, entre otras. Así mismo, implementar KPI`s que permitan identificar si está siendo efectiva la estrategia de captación de clientes, para evaluar la precisión del modelo.

10. REFERENCIAS BIBLIOGRÁFICAS

[1] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "The M4 Competition: 100,000 time series and 61 forecasting methods," *Int. J. Forecast.*, vol. 36, no. 1, pp. 54–74, 2020.

[2] M. O. F. Center, "The M5 competition." 2020.

[3] G. Ellis, "Chapter 13 - Model Development and Verification," G. B. T.-C. S. D. G. (Fourth E. Ellis, Ed. Boston: Butterworth-Heinemann, 2012, pp. 261–282.

[4] L. Breiman, "Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author)," *Stat. Sci.*, vol. 16, no. 3, pp. 199–231, Aug. 2001.

[5] C. McCue, "7 - Predictive Analytics," C. B. T.-D. M. and P. A. McCue, Ed. Burlington: Butterworth-Heinemann, 2007, pp. 117–141.

[6] B. Kang, D. Kim, and S.-H. Kang, "Real-time business process monitoring method for prediction of abnormal termination using KNNI-based LOF prediction," *Expert Syst. Appl.*, vol. 39, no. 5, pp. 6061–6068, 2012.

[7] M. Das, R. Cui, D. R. Campbell, G. Agrawal, and R. Ramnath, "Towards methods for systematic research on big data," in *2015 IEEE International Conference on Big Data (Big Data)*, 2015, pp. 2072–2081.

[8] D. Bzdok, M. Krzywinski, and N. Altman, "Machine learning: supervised methods," *Nat. Methods*, vol. 15, no. 1, p. 5, 2018.

[9] R. Gentleman and V. J. Carey, "Unsupervised machine learning," in *Bioconductor case studies*, Springer, 2008, pp. 137–157.

[10] B. Averbeck and J. P. O'Doherty, "Reinforcement-learning in fronto-striatal circuits," *Neuropsychopharmacology*, vol. 47, no. 1, pp. 147–162, 2022.

[11] A. C. Davison and C.-L. Tsai, "Regression model diagnostics," *Int. Stat. Rev. Int. Stat.*, pp. 337–353, 1992.

[12] O. O. Aalen, "A linear regression model for the analysis of life times," *Stat. Med.*, vol. 8, no. 8, pp. 907–925, 1989.

[13] A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody, and S. D. Brown, "An introduction to decision tree modeling," *J. Chemom. A J. Chemom. Soc.*, vol. 18, no. 6, pp. 275–285, 2004.

[14] M. Pal, "Random forest classifier for remote sensing classification," *Int. J. Remote Sens.*, vol. 26, no. 1, pp. 217–222, 2005.

[15] G. I. Webb, E. Keogh, and R. Miikkulainen, "Naïve Bayes.," *Encycl. Mach. Learn.*, vol. 15, pp. 713–714, 2010.

[16] A. Neumann, J. Holstein, J.-R. Le Gall, and E. Lepage, “Measuring performance in health care: Case-mix adjustment by boosted decision trees,” *Artif. Intell. Med.*, vol. 32, no. 2, pp. 97–113, 2004.

[17] D. J. Strauss, “A model for clustering,” *Biometrika*, vol. 62, no. 2, pp. 467–475, 1975.

[18] K. S. Daza Rosado and M. A. García Reyes, “Predicción, análisis y pronóstico de COVID-19 utilizando un modelo de Machine Learning basado en el análisis forecasting sobre series temporales.” Universidad de Guayaquil. Facultad de Ciencias Matemáticas y Físicas …, 2021.

[19] A. C. Peña Ordóñez, “Pronóstico de la inflación colombiana: una aproximación desde un modelo Arima desagregado y Machine Learning,” 2019.

[20] M. P. LaValley, “Logistic regression,” *Circulation*, vol. 117, no. 18, pp. 2395–2399, 2008.

[21] M. Khalilia, S. Chakraborty, and M. Popescu, “Predicting disease risks from highly imbalanced data using random forest,” *BMC Med. Inform. Decis. Mak.*, vol. 11, no. 1, pp. 1–13, 2011.

[22] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, “Support vector machines,” *IEEE Intell. Syst. their Appl.*, vol. 13, no. 4, pp. 18–28, 1998.

[23] M. R. Aceituno Rojo, “Modelo predictivo de análisis de riesgo crediticio usando Machine Learning en una entidad del sector microfinanciero,” 2019.

[24] D. G. Guerrero Calderon and O. I. Castro Buitrago, “MODELO PREDICTIVO DE PROPENSIÓN DE AHORRO E INVERSIÓN EN PRODUCTOS DE BANCA PATRIMONIAL.”

[25] C. Bergmeir, R. J. Hyndman, and B. Koo, “A note on the validity of cross-validation for evaluating autoregressive time series prediction,” *Comput. Stat. Data Anal.*, vol. 120, pp. 70–83, 2018.

[26] G. E. P. Box and D. A. Pierce, “Distribution of Residual Autocorrelations in Autoregressive-Integrated Moving Average Time Series Models,” *J. Am. Stat. Assoc.*, vol. 65, no. 332, pp. 1509–1526, Dec. 1970.

[27] R. Wirth and J. Hipp, “CRISP-DM: Towards a standard process model for data mining,” in *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, Manchester, 2000, pp. 29–39.

[28] J. A. Correa, “El método DOFA, un método muy utilizado para diagnóstico de vulnerabilidad y planeación estratégica,” *El Prism*, pp. 1–7, 2010.

[29] N. Hernández Lotero, “Clasificación de los datos personales e implicaciones legales,” 2018.

[30] P. Biecek and T. Burzykowski, *Explanatory model analysis: explore, explain, and examine predictive models*. CRC Press, 2021.

[31] F. Villarreal, “Introducción a los Modelos de Pronósticos,” *Univ. Nac. del Sur*, pp. 1–121, 2016.
[32] M. J. C. Cabay, E. F. M. Villa, M. E. P. Tixi, and B. M. B. Erazo, “Series temporales para el índice Diferencial Normalizado de Vegetación mediante una Red Neuronal Artificial de corto y largo plazo, y el algoritmo Prophet,” *Polo del Conocimiento*, vol. 7, no. 8, pp. 823–841, 2022.
[33] T. Hastie and J. Qian, “Glmnet vignette,” *Retrieved June*, vol. 9, no. 2016, pp. 1–30, 2014.
[34] J. J. Espinosa-Zúñiga, “Aplicación de algoritmos Random Forest y XGBoost en una base de solicitudes de tarjetas de crédito,” *Ingeniería, investigación y tecnología*, vol. 21, no. 3, 2020.
[35] A. Karatzoglou, A. Smola, K. Hornik, M. A. Karatzoglou, S. SparseM, and L. Yes, “The kernlab package,” *Kernel-Based Machine Learning Lab. R package version 0.9.-22. Available online: https://cran. r-project. org/web/packages/kernlab (accessed on 4 November 2015)*, 2007.
[36] M. J. A. Shohan, M. O. Faruque, and S. Y. Foo, “Forecasting of electric load using a hybrid LSTM-neural prophet model,” *Energies (Basel)*, vol. 15, no. 6, p. 2158, 2022.
[37] L. A. Gutiérrez González, “Predicción de múltiples series de tiempo univariadas a través de diversos modelos predictivos y meta-learning aplicado en la industrial del retail,” 2020.
[38] R. J. Hyndman, “Another look at forecast-accuracy metrics for intermittent demand,” *Foresight: The International Journal of Applied Forecasting*, vol. 4, no. 4, pp. 43–46, 2006.
[39] P. Nieto Figueroa and J. Vélez Correa, “Validación de medidas de evaluación para el pronóstico de la tasa de cambio en Colombia,” 2016.
[40] C. R. Rivero, J. Pucheta, J. Baumgartner, M. Herrera, H. D. Patiño, and V. Sauchelli, “Modelado bayesiano de un filtro autorregresivo no lineal basado en redes neuronales para el pronóstico de series temporales de lluvia acumulada mensual,” in *Anales del Congreso Nacional del Agua (CONAGUA 2011)*, 2011.
[41] S. Frutos Serrano, “Comparación entre XGBoost y Regresión Lineal Múltiple para la predicción de la evolución del precio de las acciones,” 2022.
[42] D. Chicco, M. J. Warrens, and G. Jurman, “The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation,” *PeerJ Comput Sci*, vol. 7, p. e623, 2021.
[43] G. BACA, “Evaluación de Proyectos, cuarta edición, editorial McGraw-Hill.” México, 2001.
[44] R. A. D. Vásquez, J. L. A. Espinoza, and M. A. C. Cabrera, “Power bi como herramienta de apoyo a la toma de decisiones,” *Universidad y Sociedad*, vol. 14, no. S3, pp. 195–207, 2022.