AÑO 2. No. 4
ISSN:2992-7978

Revisión narrativa: Datos e identificación de variables significativas para un modelo predictivo de portabilidad de nómina

Ma. Isabel González Becerril[1]
https://orcid.org/0009-0003-9718-6904
Dra. Rosa Gabriela Camero Berrones[2]
https://orcid.org/0000-0003-4438-1645

[1] Doctoranda, Universidad Americana de Europa
[2] Docente Investigadora, Universidad Americana de Europa

AÑO 2. No. 4
ISSN:2992-7978

Resumen

La problemática por abordar es en el sector financiero, los bancos tienen movilidad de usuarios que reciben el pago de su nómina y que pueden por derecho solicitar su migración a otra institución financiera, esta migración se considera como abandono, es conocida como portabilidad e implica perdidas generando gastos operativos adicional a generar mala reputación a la empresa, para ello se propone realizar un modelo predictivo para tomar decisiones informadas sobre estrategias de atracción, retención y fidelización y evitar estas fugas, para ello realizamos una revisión de literatura. Este estudio tuvo como objetivo realizar una revisión narrativa de literatura sobre predicción de abandono para determinar los datos que utilizan los modelos para identificar variables significativas que permitan implementar un modelo predictivo con machine learning y realizar un análisis descriptivo con los artículos recabados, las bases de datos consultadas fueron Scielo, Scholar, Siencedirect, Ieeexplore. La literatura revisada lleva a concluir que es importante identificar la información clave que sea útil para la investigación, datos que permitan el conocimiento 360° del cliente para dar atención personalizada y lograr disminuir el porcentaje de tasa de abandono. Los datos mencionados en los estudios son geográficos, demográficos, financieros, de comportamiento, se organizan y se les realiza tratamiento. Adicional, es relevante identificar las variables que contribuyan a la predicción, para eso ayuda la reducción de dimensiones, en los artículos analizados se usan técnicas de análisis de correlación, análisis de supervivencia, eliminación recursiva de atributos, análisis de componentes principales, análisis discriminante lineal.

Palabras clave: abandono, dimensionalidad, predicción

Abstract

The problem to be addressed is in the financial sector, banks have mobility of users who receive their payroll payment and who can by right request migration to another financial institution, this migration is considered abandonment, is known as portability and implies losses. generating additional operating expenses to generate a bad reputation for the company, for this purpose it is proposed to create a predictive model to make informed decisions about attraction, retention and loyalty strategies and avoid these leaks, for this we carried out a

AÑO 2. No. 4
ISSN:2992-7978

literature review. This study aimed to carry out a narrative review of literature on dropout prediction to determine the data used by the models to identify significant variables that allow implementing a predictive model with machine learning and perform a descriptive analysis with the articles collected, the databases consulted were Scielo, Scholar, Siencedirect, Ieeexplore. The reviewed literature leads to the conclusion that it is important to identify the key information that is useful for the investigation, data that allows 360° knowledge of the client to provide personalized attention and reduce the percentage of abandonment rate. The data mentioned in the studies are geographical, demographic, financial, behavioral, they are organized and processed. Additionally, it is relevant to identify the variables that contribute to the prediction, for this the reduction of dimensions helps, in the analyzed articles techniques of correlation analysis, survival analysis, recursive elimination of attributes, principal components analysis, linear discriminant analysis are used.

Keyboards: churn, dimensionality, prediction

INTRODUCCIÓN

La problemática por abordar es en el sector financiero, los bancos tienen movilidad de usuarios que reciben el pago de su nómina y que pueden por derecho solicitar su portabilidad a otra institución financiera, esta migración se considera como abandono e implica perdidas generando gastos operativos y puede generar mala reputación a la empresa cuando el propósito es impulsar la inclusion financiera con acceso y uso de servicios financieros formales, con costos bajos, regulados y que ayuden a mejorar su calidad de vida con productos bancarios acorde a sus necesidades individuales.

La portabilidad es un derecho laboral en México, mecanismo implementado desde 2010 por el Banco Central, tiene como objetivo ofrecer la libertad de elegir a los empleados y estar con el banco que mejores condiciones les ofrezca, esta oportunidad de decisión ha causado que un gran número de entidades se acerquen a este tipo de clientes ofreciéndoles diferentes beneficios, con la intención que abandonen en perjuicio de la otra entidad. “Promover mayor competencia en el sistema financiero. En particular, en el caso de cuentas de nómina, la LTOSF elimina barreras a la movilidad permitiendo que sea el empleado quien determine en qué banco quiere tener su cuenta de nómina” (Banxico, 2020, p. 54). Proceso ilustrado con la Figura 1.

AÑO 2. No. 4
ISSN:2992-7978

Figura 1

Proceso Portabilidad

Nota. La figura ilustra el proceso de portabilidad de un banco ordenante (emisor) a un banco receptor, fuente: https://www.banxico.org.mx/publicaciones-y-prensa/reportes-sobre-las-condiciones-de-competencia-enl/%7B6B2ACA7F-4D36-92C0-0E0F-7C09398F06C2%7D.pdf.

El estudio va dirigido a buscar información que sea de utilidad para el modelo predictivo de clientes con dispersión de nómina en la entidad financiera y aportará variables relevantes que permita detectar sus necesidades, hábitos de consumo y situación financiera para lograr tener conocimiento integral para otorgar el servicio centrado en el cliente y evitar su abandono extendiendo su ciclo de vida y logrando su fidelidad.

Si bien es importante atraer clientes nuevos es más rentable no perder clientes de este tipo donde su dispersión se capta de forma natural, el costo operativo-transaccional queda del lado del banco emisor generando perdida. Anteriormente al tener cautivos a los clientes permitía que las empresas otorgaran el servicio generalizado, la gestión era más enfocada a la venta de productos y servicios, si bien existe CRM no se brindaba un servicio personalizado, a través de este estudio se pretende investigar las mejores prácticas implementadas en machine learning que nos permita identificar la data necesaria, las variables significativas, los modelos con mejores resultados y las técnicas usadas para garantizar la reproducibilidad del modelo que se determine con mejor desempeño para retener clientes.

Para lograr el objetivo de esta articulo se realizó una revisión narrativa de literatura sobre la data que utilizan estudios con modelos predictivos. La metodología implementada fué realizar una descripción de las publicaciones científicas y en la revision narrativa de la literatura, la cual busca obtener un panorama general conciso sobre los diferentes estudios generados en modelos de deserción, con ello se pretende tener

un amplio panorama de la posible data a recabar para lograr materializar el objetivo 1 de la investigación a desarrollar la cual esta estructurada de la siguiente forma:

Objetivo General

Implementar un modelo predictivo que permita predecir la portabilidad, mediante la obtención de información del universo de clientes para su análisis de variables, diseño del modelo y evaluación de resultados para disminuir el porcentaje de tasa de abandono.

Objetivos Específicos

1. Recabar y analizar información para identificar variables significativas de clientes con probabilidad de abandono.

2. Diseñar un modelo predictivo para detectar clientes con alta probabilidad de abandono.

3. Evaluar los resultados finales del modelo predictivo para evaluar que tan bien predicen los modelos implementados.

Para este articulo se acota a documentar el estado del arte del primer objetivo el cual es identificar las variables que se pueden recabar para analizar y determinar las que sean significativas para resolver el problema de investigación.

Hipótesis

Es posible integrar información de clientes con probable portabilidad que permita analizar el comportamiento transaccional, productos contratados y hábitos de consumo para detectar patrones que motivaron su abandono e identificar variables relevantes de estos clientes que nos permita una data óptima reduciendo tiempo de cómputo y gasto

ESTADO DEL ARTE Y MARCO TEÓRICO

El presente trabajo esta dirigido a explorar los trabajos relacionados de predicción que permiten cumplir con el objetivo específico 1, el cual es recabar y analizar información para identificar variables significativas de clientes con probabilidad de abandono de nómina los cuales permitan investigar su comportamiento para segmentar y personalizar las necesidades del cliente mejorando la experiencia del

cliente, lo cual permite impulsar la innovación, tomar decisiones basadas en datos y aumentar la eficiencia operativa.

Inclusión Financiera: en la plataforma del Consejo Nacional de Inclusión Financiera "se define como el acceso y uso de servicios financieros formales (cuentas, seguros, créditos y Afores) bajo una regulación apropiada que garantice esquemas de protección al consumidor y promueva las competencias económico-financieras" (CONAIF, 2024, p. 2)

Marco Legal o Marco Normativo

Por ser el estudio en una institución financiera, está regulada y cumplen cabalmente con las disposiciones de las autoridades del sistema en términos de transparencia y privacidad de datos por lo que se encuentran protegidos tantos los recursos de los clientes como los datos personales. Se cumple la legislación mexicana definida en su Ley Federal de Protección (LFPDPPP) acorde al Capitulo 1 de Disposiciones generales

> CAPÍTULO I Disposiciones Generales
> Artículo 3.- Para los efectos de esta Ley, se entenderá por:
> " I. Aviso de Privacidad: Documento físico, electrónico o en cualquier otro formato generado por el responsable que es puesto a disposición del titular, previo al tratamiento de sus datos personales, de conformidad con el artículo 15 de la presente Ley.
> V. Datos personales: Cualquier información concerniente a una persona física identificada o identificable.
> VI. Datos personales sensibles: Aquellos datos personales que afecten a la esfera más íntima de su titular, o cuya utilización indebida pueda dar origen a discriminación o conlleve un riesgo grave para éste. En particular, se consideran sensibles aquellos que puedan revelar aspectos como origen racial o étnico, estado de salud presente y futuro, información genética, creencias religiosas, filosóficas y morales, afiliación sindical, opiniones políticas, preferencia sexual".
> El proceso de elegibilidad de datos para el algoritmo será de nómina-habientes con expediente integrado, esto garantiza que contamos con su consentimiento,

AÑO 2. No. 4
ISSN:2992-7978

"IV. Consentimiento: Manifestación de la voluntad del titular de los datos mediante la cual se efectúa el tratamiento de los mismos".

"XIII. Disociación: El procedimiento mediante el cual los datos personales no pueden asociarse al titular ni permitir, por su estructura, contenido o grado de desagregación, la identificación del mismo;" (LFPDPPP, 2010, p. 2)

METODOLOGÍA

Este documento presenta la siguiente estructura: la metodología, que describe el método utilizado en el estudio, el cual realizó una revisión narrativa de literatura con estilo libre con resultados de trabajos realizados en los años 2020, 2021 y 2023, se menciona la metodología utilizada para privacidad y protección de datos. Se continua con los resultados, que presentan la descripción de la producción científica y la revisión narrativa de literatura, que da respuesta a las preguntas de investigación. Posteriormente, se presenta la discusión y se finaliza con las conclusiones, limitaciones y futuras investigaciones.

Se realiza una revisión sistemática buscando palabras claves en las bases genéricas de datos sobre el tema de investigación, las cuales se mencionan en la tabla 1.

Tabla 1

Buscadores consultados

Base	Palabra Clave
https://scielo.org/es/	predicción abandono
https://scholar.google.es/schhp?hl=es	predicción abandono, abandono clientes
https://www.sciencedirect.com/search?	qs=prediction
https://ieeexplore.ieee.org/Xplore/home.jsp	prediction churn
https://dl.acm.org/	

Nota. Elaboración Propia

Se realiza la lectura de los trabajos encontrados y se determina de acuerdo al contenido la inclusión o exclusión para esta revisión narrativa.

AÑO 2. No. 4
ISSN:2992-7978

Variables a recopilar

Existen diferentes artículos en la comunidad científica los cuales son de predicción de deserción los cuales se comentan a continuación:

Una investigación realizada sobre predicción de deserción de clientes en una administradora de fondos, menciona dos tipos de variables que podemos recabar "La literatura sugiere usar variables sociodemográficas y de comportamiento del cliente para la ejecución del modelo de predicción. Para comprobar la utilidad de estas variables se usaron estadísticas descriptivas y además la experiencia de expertos" (Bohórquez, 2020, p. 5), las cuales pueden ser útiles al tratarse de una entidad financiera donde es similar la data recabada, elige 10 variables predictoras como edad, ingreso, promedio mensual, prima, saldo actual, empresa, categoría empresa, provincia vivienda, genero, forma de pago. Hace uso de 3 modelos, CART, bosque aleatorio, regresión logística obteniendo resultados del 88%, 93% y 91% respectivamente y determina 3 variables principales que determinan la deserción, a menor saldo es mayor la probabilidad de abandono.

Es importante estudiar como determinan si un cliente abandona, en el tratan de detectar que variables tienen impacto en la decision del cliente de abandonar o no a la empresa, un determinante que se menciona el trabajo investigativo es la categorización determinista del cliente y también si alguna de las variables utilizadas puede ser homologadas a variables del sector financiero como las llamadas que se realizan son el uso del servicio y trasladando al ambito financiero es su transaccionalidad, a mayor transaccionalidad se considera un cliente **activo** como define Collazos "un cliente puede considerase pasivo o activo, dependiendo si la baja es directa o no" (Meza, 2020, p. 104).

El siguiente trabajo trata de fuga de clientes en el sector de telecomunicaciones, si bien es otro ramo, existe el derecho el servicio de portabilidad a otra empresa telefónica por tanto es importante analizar el estudio propuesto por Leon, J. M., (2021):

> Según datos del Banco Mundial, la industria de las telecomunicaciones enfrenta cada año a una fuga de clientes que bordea el 30%. Estudios recientes han mostrado que tanto atributos cuantitativos: cantidad de minutos, mensajes, etc.; así como los cualitativos: edad, sexo, tipo de dispositivo tienen influencia en la fuga de clientes. En base a la literatura encontrada se definieron

AÑO 2. No. 4
ISSN:2992-7978

dos tipos de variables: demográficas y del comportamiento del consumidor las cuales son útiles para realizar la predicción de permanencia del cliente. (Leon, J. M.,, 2021, p. 2)

La portabilidad será considerada una dummy que identifica si el interés del consumidor califica como interés presente o interés ausente por realizar portabilidad numérica en corto plazo mediante su historial de portabilidades. 1: cliente migró anteriormente y 0: cliente no migró anteriormente. Solo un 7% de los clientes del estudio fueron provenientes por portabilidad de otras empresas, competidores: Claro, Movistar, Entel. Este indicador refleja la poca captación de nuevos clientes que prefieren contratar Bitel, a costo de oportunidad dejar su proveedor actual. (Leon, J. M.,, 2021, p. 78)

Al igual que existe una alta competencia en el mercado de telefonía celular lo existe para el sector financiero, existe una variabilidad de multiples ofertas de productos y servicios de varios bancos y fintech por lo cual cambiarse resulta muy fácil, el reto es fidelizar y mantener y de ser posible recuperar clientes perdidos o bien dormidos, ya que puede mantenerse activo pero sin interés en contratar o invertir. Revisando las variables que usa este articulo podemos traducir en variables que apliquen a la institución financiera como se muestra en la tabla 2.

Tabla 2

Conversión variables del sector de telecomunicaciones al sector financiero

Variables telecomunicaciones	Variables financieras
Cantidad de mensajes de texto enviados por el cliente	No. transacciones realizadas
Tiempo de permanencia	Antigüedad del cliente
Edad	Edad
Product_Type , prepago o post pago	Tiene captación y/o crédito
Sexo	Sexo
City_Name	Estado

AÑO 2. No. 4
ISSN:2992-7978

Quejas y Reclamos	Tiene quejas
Vas (Servicio de Valor Añadido)	Segmento del cliente (preferencial)
ADS (Acepta publicidad)	Push en campañas
Portabilidad (Clientes que migraron anteriormente)	Portabilidad, ¿es migrado?
Mi Bitel (Uso de la app)	Uso de la banca móvil

Nota. Elaboración Propia

Nivel de estudios, estado civil, ingresos, capacidad de pago, montos adeudados, situación laboral, cantidad de empleos, ingresos percibidos por dependencia laboral, consumos realizados y el tiempo que ha pasado el cliente sin transaccionar son variables que usaron en este estudio y enriquecen la data pensada para armar el data set para uso del modelo, de acuerdo al estado del arte recabado en este articulo, adicional se comentan las mejoras al implementar con machine learning, las cuales son niveles de retención, rentabilidad y toma de decisiones estratégicas del negocio para cuidar la relación entre los clientes y las empresa. Para la eliminación de variables no significativas utiliza el método de Regularización de Lasso y la metodología de Stepwise para eliminar las variables que no aportan de manera estadística, acorde a lo mencionado por LOAIZA “Uno de los desafíos más importantes que enfrentan las instituciones financieras es la retención de los clientes en sus productos de activos y pasivos” (CASTRO LOAIZ, 2022, p. 18), se debe buscar un total equilibrio para que el cliente tenga un balance favorable entre sus productos financieros logrando acceso a mejorar sus finanzas.

El éxito de un buen modelo con resultados precisos, consistentes y eficientes esta determinado por los datos que se recopilan y su procesamiento realizado. Una técnica de tratamiento de datos que mejora el rendimiento del modelo es binarizando como lo explora NABIPOUR (2020).

> *The nature of stock market movement has always been ambiguous for investors because of various influential factors. This study aims to significantly reduce the risk of trend prediction with machine learning and deep learning algorithms. Four stock market groups, namely diversified financials, petroleum, non-metallic minerals and basic metals from Tehran stock exchange, are*

AÑO 2. No. 4
ISSN:2992-7978

chosen for experimental evaluations. (NABIPOUR, 2020, p. 1). *This paper involves two approaches for input information. continuous data is supposed to be based on actual time series, and binary data is presented with a preprocessing step to convert continuous data to binary one with respect to each indicator nature.* (NABIPOUR, 2020, p. 4). *In this study, we use nine machine learning methods (Decision Tree, Random Forest, Adaboost, XGBoost, SVC, Naïve Bayes, KNN, Logistic Regression and ANN) and two deep learning algorithms (RNN and LSTM)* (NABIPOUR, 2020, p. 5). T*he creating deep models is different from machine learning when the input values must be three dimensional (samples, time_steps, features); so, we use a function to reshape the input values. Also, weight regularization and dropout layer are employed to prevent overfitting here* (NABIPOUR, 2020, p. 9)

En este estudio realizan dos experimentos, el primero con datos continuos y el segundo implementa una capa adicional convirtiendo datos continuos a binarios. Los resultados obtenidos detonan que con datos continuos se puede llegar a una predicción del 67% y con datos binarios a una predicción del 83%, mejorando notablemente el rendimiento al binarizar y los algoritmos RNN y LSTM son los que otorgan los mejores resultados con ambos experimentos. Para implementar en los algoritmos de deep learning se transforman los valores de entrada por valores tridimensionales.

Variables significativas

Una vez recabada la información de los clientes que han abandonado y de clientes que continúan dentro de la institución es de relevancia detectar las variables significativas de la data recabada.

El análisis de componentes principales (PCA) sirve para reducir las dimensiones y hacer mas legible el modelo, lo complejo es ver que dimensiones representa cada componente y su contribución ya que existe una transformación y suele no ser tan transparente explicar la composición. Wu afirma que "The Principal Component Analysis is also applied to help reduce the dimension of our data and to show the correlation of different features." (Wu, 2020, p. 1)

Agrupar características puede ser de utilidad para determinar si los clientes que abandonan presentan algún patron que pueda separar en pequeños grupos para realizar estrategias personalizadas o bien determinar algún comportamiento que separe a los clientes que solicitan su portabilidad y los que se mantienen sin solicitarla. Para ello estos cluster son definidos por MELLA - NORAMBUENA como:

> Los clústeres o conglomerados (agrupación) son un método que implica el proceso de dividir datos o poblaciones en grupos que tienen caracteres casi idénticos o patrones, pero difieren entre grupos, entre los métodos utilizados para la agrupación se encuentran: Análisis de componentes Principales (PCA).” (MELLA - NORAMBUENA, 2022, p 16)

Un trabajo realizado sobre variables que afectan el abandono estudiantes, selecciona las variables mas influyentes para entrenar el modelo de predicción, de 25 variables, se redujo con PCA y la elección del numero de componente esta en función del % de influencia acumulado que se decida elegir, en el articulo eligen 8 variables que participan en la generación del 65.21% de la información contenida en la variable dependiente, Castrillón-Gómez (2020), agrega que “Mediante un análisis estadístico multivariado, se seleccionaron aquellas más influyentes para estructurar un archivo que fue analizado por el algoritmo J48 de la plataforma Weka” (Castrillón-Gómez, 2020, p. 1)

Adicional a PCA, un articulo mas menciona varios métodos para identificar las mejores variables predictoras entre ellas evaluando la correlación a través del coeficiente de Pearson, así como el análisis de supervivencia, el cual sugiere Martínez Pérez J.R sea analizado en un contexto multivariado como:

> La deserción escolar debe ser analizada en un contexto multivariado para identificar sus causas y efectos, de ningún modo debe ser atribuida a una sola causa. Determinar la capacidad predictiva de algunos factores sobre la deserción escolar de estudiantes de Medicina, a través de un modelo de regresión logística múltiple. (Martínez Pérez J.R, 2021, p. 217)

> Para identificar los mejores predictores de este fenómeno se han utilizado varios métodos, dentro de los que se pueden citar: los análisis correlacionales, el análisis de regresión logística, el análisis de supervivencia, la minería de datos y el uso de la inteligencia artificial (Martínez Pérez J.R, 2021, p. 220)

Sobre análisis de supervivencia se encuentra el trabajo investigativo de Leon, J. M. (2021) el cual menciona dos validaciones utilizando deciles para evaluar la efectividad y éxito del modelo.

> La primera fue a través de la clasificación de las probabilidades de supervivencia pronosticadas durante un tiempo específico en orden ascendente en deciles y luego se el resultado

> obtenido con el número de clientes que abandonaron la compañía (fugados) durante este período de tiempo especificado en cada decil. La segunda manera de validar el éxito del modelo consiste en colocar las probabilidades de supervivencia pronosticadas en el mismo orden y luego comparar el número de clientes fugados hasta este tiempo especificado en cada decil. (Leon, J. M.,, 2021, p. 25)

> El análisis de supervivencia permite segmentar poblaciones en base a la duración hasta la ocurrencia de un evento. (Leon, J. M.,, 2021, p. 89)

Otro punto interesante es que Leon, J. M. usa el FIV (Factor de inflación de la varianza) para la elegibilidad de variables.

La Eliminación Recursiva de Características (RFE, por sus siglas en inglés) es una técnica de selección de características utilizada en el aprendizaje automático para mejorar la precisión y eficiencia de los modelos. Este método funciona eliminando de manera iterativa las características menos importantes y construyendo el modelo repetidamente hasta que se alcance el número deseado de características y lo uso Falla J. D. "Después de explorar en general los datos y exponer un análisis descriptivo preliminar, se utilizo, la técnica denominada Eliminación de Características Recursivas (Recursive Feature Elimination-RFE)" (Falla J. D., 2021, p. 19) para mejorar el rendimiento, obtener predictores más precisos y mejorar la comprensión del modelo implementado, todo esta en función de la data que se logre recolectar, que exista variabilidad y que no tengan redundancia las características.

Una técnica mas para reducir dimensionalidad es a través del indice Gini, como lo menciona CASTRO LOAIZA (2022) con la ganancia gini *"For attribute A , the Gain Gini − after dividing the dataset into two parts by any attribute value is calculated respectively, and the minimum value is selected as the optimal binary scheme obtained by attribute A"* (CASTRO LOAIZ, 2022, p. 11). A mayor índice de Gini menor pureza, por lo que se selecciona la variable con menor Gini ponderado. Suele seleccionar divisiones desbalanceadas, donde normalmente aísla en un nodo una clase mayoritaria y el resto de clases los clasifica en otros nodos.

Adicional a PCA, el trabajo de Chauhan Tannu (2021) mencio**na LDA que** es otra técnica de reducción de dimensionalidad, es un análisis discriminante linear que entra dentro del aprendizaje supervisado e informa cual es el discriminate más importante al momento de realizar una clasificación, es una opción también a evaluar de acuerdo a la data que se logre obtener para esta investigación

Unsupervised learning methods such as PCA, LDA are mainly applied for the dimensionality reduction. Unnecessary attributes in diabetes data set are misleading the accuracy of classifier. Hence, we can have combination of supervised and unsupervised learning for the better prediction and detection of diabetes. Zhu et al. in [23] integrated PCA, K-Mean and logistic regression.(Chauhan Tannu, 2021, p. citando a C Zu, 2019))

Explorar el indice Gini y la ganancia de información es una opción mas que menciona Lu Yifan (2022) "In our paper, we will first introduce the metrics used by the algorithms in decision trees. We will then discuss their application in the algorithms CART and ID3" (Yifan, 2022, p. 1014), para identificar los atributos mas representativos que nos lleve a obtener un data set recortado para mayor desempeño de la solución que logremos experimentar.

La metodología para garantizar la privacidad y protección de datos personales de clientes con dispersión de nómina tiene 3 medidas que se ilustran con la Figura 2 y tiene como propósito que el cliente no sea identificable a través de la data usada para el modelo a través de los cuales se garantiza que :

- No se identifica a la persona y
- No es identificable a través de una manipulación en los datos

Figura 2

Medidas para evitar identificabilidad

Anonimización Cifrado	Binarización Convertir campos categóricos en binarios.	Disociación El acceso solo es a datos necesarios para fines estadísticos
		
https://blog.powerdata.es/el-valor-de-la-gestion-de-datos/enmascaramiento-de-datos-vs-encriptacion-de-datos-cual-le-conviene-mas-a-tu-negocio	https://blogmapfre.com/innovacion/el-big-data-y-su-posible-aplicacion-para-los-seguros-personales/	https://protecciondedatospersonalesonline.com/2019/01/16/datos-anonimos-y-pseudonimos/

Nota. Figura que contiene las técnicas que garantizan la identificación de un cliente protegiendo sus datos personales. Fuente: adaptado de vínculos de internet contenidos en Figura.

RESULTADOS

El planteamiento del proyecto es utilizar un método **cuantitativo**, lograr definir las variables que nos permitan obtener conocimiento para lo cual se realizo una búsqueda de artículos con la problemática de deserción para validar la data a recabar que permitan generar valor a través de un modelo predictivo, de acuerdo con el alcance es un **estudio correlacional**, donde se tiene una variable a predecir la cual estaremos revisando la intensidad de relación entre las variables predictoras por lo que los trabajos realizados dan una buena pauta de inicio para integrar la data para el modelo. La tabla 3 ilustra los resultados obtenidos para los datos a recopilar.

Tabla 3

Revisión de Literatura Información relevante de variables a recabar para clientes con abandono.

Cita	Autor	Año	Sector	Estrategia	Datos	Resultados
[1]	María Bohórquez , Joyce Torys , Milton Paredes Aguirre (2020)	2020	Financiero	Árboles de decisión, Random Forest y Regresión Logistica	Edad, Ingreso, Promedio Mensual, Saldo Actual, Nómina o independiente, Genero, Canal, transacciones	AUC, **Random Forest 93%**, Regresión Logistíca 91% y CART 88%
[2]	MOJTABA NABIPOUR1 , POOYAN NAYYERI2 , HAMED JABANI3 , SHAHAB S., AMIR MOSAVI	2020	Financiero	Decision Tree, Random Forest, Adaboost, XGBoost, SVC, Ingenuo Bayes, K-NN, RL y ANN)) y 2 (RNN) y Long short-term memory (LSTM)	**datos continuos y datos binarios** antes de usar los algoritmos, mejora resultados	RNN y LSTM superiores
[16]	Aldo Richard Meza Rodríguez, Jorge Chue Gallardo (2020)	2020	Telefónica	**Adaboost**, el cual se desarrolla a través de **aprendizaje adaptativo, regresión logística**	80,000 registros, clasificación un cliente puede considerase pasivo o activo, dependiendo si la baja es directa o no	**93%** de ACUC
[17]	Bernachea Collazos & Chilet Paisig & Guzmán Fernandez & Inche Contreras / Leon Munive (2021)	2021	Telefonía	tipos de variables: **demográficas** y del **comportamiento** del consumidor las cuales son útiles para realizar la predicción de permanencia del cliente	8,572 clientes, el comportamiento del cliente, las percepciones del cliente,la demografía, el macro entorno del cliente.	regresión logística, **88%** ACUC
[19]	YELTSIN ALEXANDER CASTRO LOAIZA (2022)	2022	Financiero	**comportamiento transaccional** histórico, variables **sociodemográficas**, **nivel de ingresos**, asociación con el banco y **nivel de riesgo** en el sistema financiero, variables con **reducción de pureza (Gini)**	400 variables, de las cuales 383 son cuantitativas y el complemento cualitativas	R logistica Stepwise, R logistica R Lasso, R logistica, Arbol de Desiciones, Bosque Aleatorio **94%**

Nota. Elaboración propia con artículos leídos.

Es vital mejorar la eficacia y precision de los modelos a través de reducir la dimensionalidad en la data recabada, al realizar reducción del número de caracteristicas se elimina el ruido y las caracteristicas redundantes que no generan valor, se ahorra tiempo y capacidad de computo al tener datos más cortos en columnas elegidas como principales, al tener menos columnas mejora la capacidad de análisis para generar valor en la toma de decisiones, es importante conservar las características esenciales de los datos por lo que las técnicas PCA y LDA son útiles para concentrar en componentes principales muchas de las dimensiones relevantes.

La tabla 4, concentra las técnicas para elegibilidad de variables predictoras con mayor relevancia.

Tabla 4

Técnicas para elegir caracteristicas relevantes que sean predictoras de abandono.

Cita	Autor	Año	Sector	Tratamiento Variables	Muestra	Resultados
[3]	Wangwei Wu, Noviembre (2020)	2020	Deportes	**PCA** para reducir la dimensiónde datos y mostrar la correlación de diferentes características	evaluaciones psicosociales con 5,340 registros	Random Forest, Injury, **PCA,** NB A
[13]	Mella Norambuena (2022)	2022	Educación	Evaluar cuando se tenga la data que sean datos cuantitativos, normalizados, sin datos extremos y sin **multicolinealidad**, métodos utilizados para la agrupación Análisis de componentes Principales (**PCA**)	500 participantes, variables analíticas, sociodemográficas y **sociocognitivas** Incluye minería de datos, aprendizaje automático y estadísticas metodologías	análisis de regresión lineal, red neuronal, **distancias Levenshtein** y KNN
[14]	Martínez Perez J. (2021)	2021	Educación	variables mas significativas asociadas a la deserción escolar (bivariado) y capacidad de estas variables para predecir la deserción. Identificar mejores predictores utilizando los **análisis correlacionales, análisis de regresión logística, análisis de supervivencia, la minería de datos y el uso de la inteligencia artificial**.	87 participantes. Variables cuantitativas relacionadas con la deserción, Coeficiente de correlación lineal de Pearson para las variables que mostraron asociación con la deserción	regresión logistica (analisis multivariado), probabilidad de abandono **87.75%**
[15]	Castrillón Gómez & Sarache & Ruiz Herrera (2020)	2020	Educación	variables principales con tecnicas de mineria de datos, algoritmo de **clasificación bayesiana**. Establecer influencia de cada variable independiente vs dependiente, seleccionar influyentes con analisis de componentes **PCA**	410 participantes, variables agrupadas en personales, económicas, sociales, familiares y académicas	**correlación significativa** y detectar no existan variables con **alta correlación** con respecto a variable independiente
[17]	Bernachea Collazos & Chilet Paisig & Guzmán Fernandez & Inche Contreras / Leon Munive (2021)	2021	Telefonía	**Análisis de multicolinealidad, El análisis de supervivencia** permite segmentar poblaciones en base a la duración hasta la ocurrencia de un evento.	8,572 clientes, el comportamiento del cliente, las percepciones del cliente,la demografía, el macro entorno del cliente.	regresión logística, **88% ACUC**

[18]	Jésus David Falla Arango (2021)	2021	Telecomunicaciones	Uso de **selección de variables** con el método empaquetado eliminación recursiva de atributos (**RFE**)	El conjunto de datos tiene 5.889.720 registros , con 58 variables predictivas y 1 variable de respuesta (churn = si/no).	(8) RL, Kvecinos, Bosques Aleatorios, Análisis Discriminante Lineal, Naive Bayes, Perceptrón Multicapa, Gradient Boosting Machine y **XGBoost** con la mejor ROC AUC **78.54**
[19]	YELTSIN ALEXANDER CASTRO LOAIZA (2022)	2022	Financiero	**comportamiento transaccional** histórico, variables **sociodemográficas**, **nivel de ingresos**, asociación con el banco y **nivel de riesgo** en el sistema financiero, variables con **reducción de pureza (Gini)**	400 variables, de las cuales 383 son cuantitativas y el complemento cualitativas	R logistica Stepwise, R logistica R Lasso, R logistica, Arbol de Desiciones, Bosque Aleatorio **94%**
[22]	Chauhan Tannu, Rawat Surbhi, Malik Samrath, Singh Purshpa	2021	Salud	Métodos de aprendizaje no supervisado como **PCA y LDA** son aplicados para reducir deimensionalidad	Aprendizaje supervisado, aprendizaje no supervisado y aprendizaje profundo	Naïve Bayes, SVM, árbol de decisión, combinados con PCA y K-medias, K-Means dan buenos resultados. Aprendizaje profundo
[24]	Lu Yifan & Ye Tianle, Zheng Jiali, 2022	2022	Automotriz	Decision trees, nos permite realizar exploración de datos de forma rápida y eficaz	rama de probabilidad, índice de Gini y ganancia de información	Predicción de compra de auto con **CART,I D3** y **C4.5**

Nota. Elaboración propia con artículos leídos.

DISCUSIÓN

Si comparamos las portabilidades que se dan en el sector financiero con el sector de telecomunicaciones encontramos grandes similitudes por lo que es de precisar que es relevante analizar artículos con temática en ese sector, este derecho que tienen los clientes incremento la transparencia y competencia entre empresas, beneficiando a los clientes con mejor atención en cuanto a tiempo, calidad y precios justos centrados en mejorar su experiencia y satisfacción acorde a sus necesidades a través de la personalización, la tabla 5 contiene convergencias entre ambos sectores.

Tabla 5

Comparación entre sector financiero y sector de telecomunicaciones.

Rubro	Telecomunicaciones	Financiero

Movilidad, regulación y transparencia.	Cambiar de proveedor es posible con las reglas de portabilidad emitidas por el Instituto Federal de Telecomunicaciones.	En el caso del sector financiero, Banco de México emitió ley de portabilidad como un derecho laboral .
Fidelizar.	Incremento en beneficios con descuentos, planes y servicios integrales para lograr alargar el tiempo de preferencia del cliente.	Innovar a través de la personalización de servicios y productos personalizados, con mejores tasas en crédito y beneficios en cuentas de captación para atraer, retener y fidelizar al cliente.
Calidad	Mejores sustanciales en tiempos de respuesta cortos en procesos, soluciones basadas en mejorar la experiencia del cliente	Soluciones acorde a necesidades del cliente, gestión integral agilizando el acompañamiento del cliente con procesos eficientes.
Adopción de tecnología	Desarrollo de canales digitales como el móvil para renovar contratos y atraer clientes nuevos.	Incremento de canales digitales, adicionales al ATM o la banca por internet, el teléfono móvil es el que más se ha desarrollado para atención de autoservicio haciendo más ágil los procesos.

Nota. Elaboración propia.

Se obtiene información de clientes con abandono para saber que tipos de datos recabar, encontrando de tipo geográficos, de interacción, de riesgos, financieros, psicográficos [1] así como usar datos continuos y datos binarios antes de usar los algoritmos como mejora de resultados [2]. En adición un punto relevante que se comenta es el clasificar el cliente como pasivo o activo, para esta investigación como

AÑO 2. No. 4
ISSN:2992-7978

pasivo es el cliente que solicita su portabilidad [16] y otro punto a considerar es validar si se puede obtener data del comportamiento del cliente para predecir su permanencia [17], en otro trabajo realizado de Predicción del abandono de tarjeta habiente, se considera comportamiento de pago de obligaciones, variables socio demográficas, experiencia con otros productos de colocación y nivel de ingresos [19].

PCA [3] [13] [15] es una buena técnica que se menciona para reducir y seleccionar variables influyentes, otro análisis que se menciona es el análisis de supervivencia [14] [17] vale la pena explorar si es de utilidad para la problemática planteada. Eliminación recursiva de atributos (RFE) es utilizada en el artículo [18], habrá que explorar la metodología con la que obtiene las variables mínimas. Otro trabajo realizado de abandono de clientes con tarjeta de crédito recaba data transaccional, socio demográfica, nivel de ingresos, nivel de riesgos y la elección de variables principales las realiza con reducción de pureza (Gini) [19]. El contenido del artículo de Chauhan [22] adicional a PCA también usan LDA para reducir dimensionalidad, en adición el artículo de Lu Yifan [24] que usa árboles para identificar probabilidades, índice de Gini y ganancia.

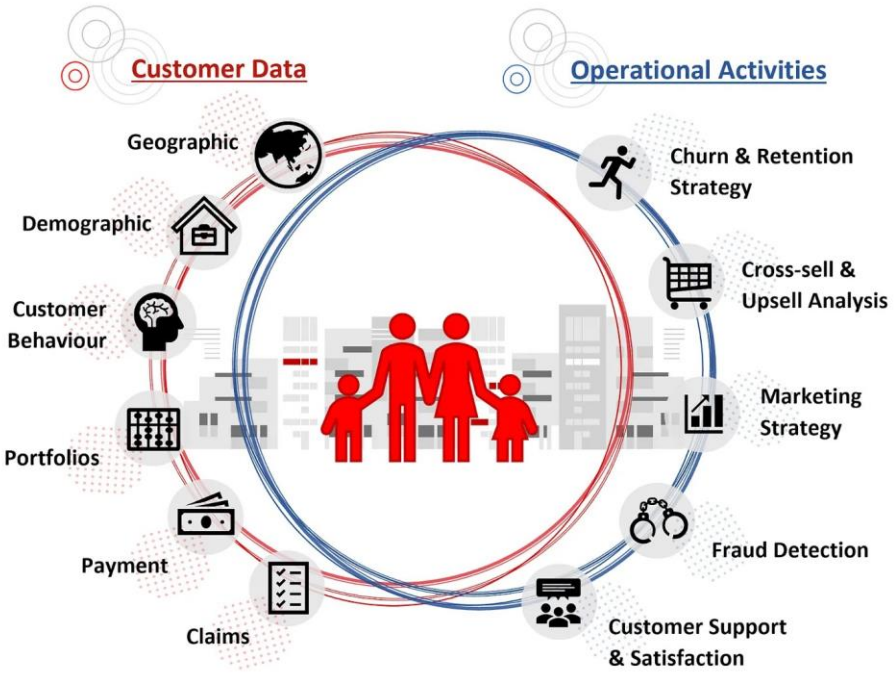
CONCLUSIONES

Recopilar datos

De acuerdo a la revisión literaria lo adecuado es tener variedad de datos que enriquezcan el modelo de forma unificada que permita tener una vista panorámica del cliente con sus datos demográficos, historial de productos contratados, historial de compras y pagos, interacciones digitales y físicas, indicadores de satisfacción, eventos significativos, para lograrlo se tiene que unificar la data de los diferentes sistemas y canales donde interactúa el cliente. La figura 3, permite visualizar el concepto 360° del cliente.

Figura 3

Vista 360° del cliente



Nota. La figura converge la actividad transaccional con los datos financieros y operativos del cliente, permitiendo tener una visión integral, fuente: https://www.atoti.io/articles/customer-360-how-it-can-be-achieved-with-atoti/

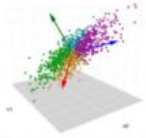
Identificar variables significativas.

Realizar la selección de atributos para reducir la data obtenida reduce la complejidad del modelo, un modelo más simple es más fácil de entender y explicar, con la revisión narrativa encontramos diferentes métodos para realizarlo como lo muestra la figura 4

Figura 4

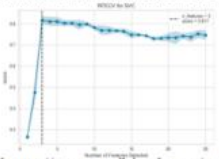
Métodos para identificar variables significativas que ayuden a fortalece el desempeño del modelo de riesgos

PCA
Transforma variables correlacionadas en nuevas variables, simplificando los datos.



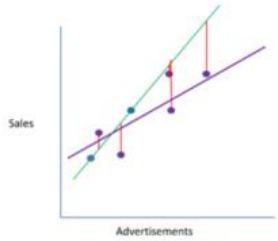
https://medium.com/@ilyurek/principal-component-analysis-pca-a-practical-guide-58dea99dd93

RFE
Elimina iterativamente seleccionando las menos importantes



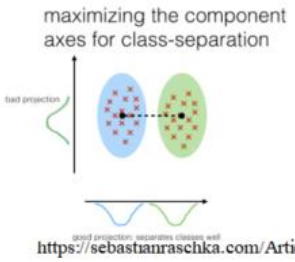
https://www.scikit-yb.org/en/latest/api/model_selection/rfecv.html

Regularización Lasso
Agrega un termino de penalizacion a la función de pérdida



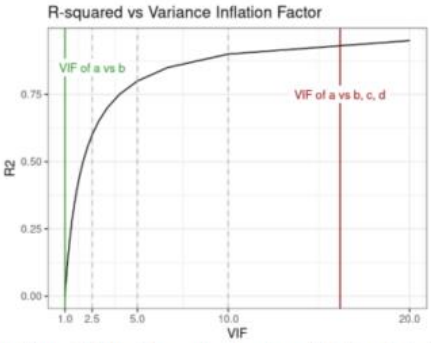
https://www.linkedin.com/pulse/lasso-regression-clearly-explained-bhabani-shankear-basak/

LDA
Maximiza la separación de clases, proyectando en menor dimensión



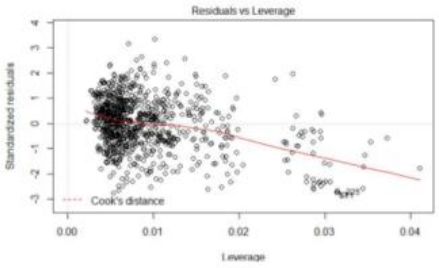
https://sebastianraschka.com/Articles/2014_python_lda.html

VIF
Determina la colinealidad, un VIF alto indica multiconealidad



https://www.blasbenito.com/post/variance-inflation-factor/

Metodología de Stepwise
Usa criterios estadísticos para optimizar la selección de variables



https://predictivemodeler.com/tag/stepwise-regression/

Nota. Adaptado de vínculos de internet contenidos en Figura (2024), recuperados el 04 de noviembre 2024.

Ventajas al seleccionar atributos relevantes:

1) Reduce el sobre entrenamiento: Menos datos redundantes significan menos oportunidades para tomar decisiones sobre la base de ruido.

2) Mejora la precisión: Menos datos engañosos se convierten en una mejora en la exactitud del modelo.

3) Reduce el tiempo de entrenamiento: Menos datos significa que los algoritmos aprenden más rápidamente.

Queda por abordar fuentes externas que permitan completar la data y generar valor como puede ser su perfil económico, su historial crediticio en otras instituciones para saber su capacidad actual, si cuenta con educación financiera para determinar su madurez en cuanto a su flujo de efectivo.

REFERENCIAS

[17] (Leon, J. M., 2021, pp. 2, 78, 25, 89). Técnica de Machine Learning para el cálculo de la probabilidad de fuga de los clientes de la empresa Bitel [Tesis de Licenciatura, Universidad ESAN. Facultad de Ingeniería]. Repositorio Institucional Universidad ESAN. https://hdl.handle.net/20.500.12640/2929.

[1] (Bohórquez, 2020, p. 5). Bohórquez, María, Torys, Joyce, Paredes, Milton, (2020). MODELOS DE PREDICCIÓN DE DESERCIÓN DE CLIENTES PARA UNA ADMINISTRADORA DE FONDOS ECUATORIANA. DOI - 10.46677/compendium.v7i1.777

[15] (Castrillón-Gómez, 2020, p. 1) Castrillón-Gómez O. D., Sarache W. y Ruiz-Herrera S. (2020). Predicción de las principales variables que conllevan al abandono estudiantil por medio de técnicas de minería de datos. Recuperado de http://dx.doi.org/10.4067/S0718-50062020000600217

[19] (CASTRO LOAIZ, 2022, pp. 11, 18) CASTRO LOAIZA Y. A. (2022). Predicción del abandono de tarjetahabiente aplicado en una institución financiera ecuatoriana. Recuperado de http://www.dspace.espol.edu.ec/handle/123456789/56497

[22] (Chauhan Tannu, 2021, p. citando a C Zu, 2019). T. Chauhan, S. Rawat, S. Malik and P. Singh, "Supervised and Unsupervised Machine Learning based Review on Diabetes Care," 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2021, pp. 581-585, doi: 10.1109/ICACCS51430.2021.9442021. keywords: {Support vector machines;Machine learning algorithms;Supervised learning;Stroke (medical condition);Prediction algorithms;Diabetes;Decision trees;Machine Learning;Supervised;Unsupervised;Diabetes;Decision Tree;Prediction},

[18] (Falla, J. D. 2021, p. 19) Falla, J. D. (2021). Predicción de abandono de clientes en telecomunicaciones mediante el aprendizaje automático. Recuperado de: http://hdl.handle.net/20.500.12010/22247.

[14] (Martínez Pérez J.R, 2021, p. 217) . Martínez Pérez J.R. , Pérez Leyva E. H., Ferrás Fernández Y., Bermúdez Cordoví L. L.(2021). Análisis predictivo de la deserción estudiantil en la carrera de Medicina. Recuperado de http://scielo.sld.cu/pdf/edu/v13n3/2077-2874-edu-13-03-217.pdf

[13] (MELLA - NORAMBUENA, 2022, p 16) Mella-Norambuena, J., Badilla-Quintana, M. G., & López Angulo, Y. Modelos preditivos baseados no uso de analítica da aprendizagem no ensino superior: uma revisão sistemática. Texto Livre, 15, e36310. https://doi.org/10.35699/1983-3652.2022.36310

[16] (Meza, 2020, p. 104). Meza, A.; Chue, J. (2020). Uso del algoritmo Adaboost y la regresión logística para la predicción de fuga de clientes en una empresa de telefonía móvil.Natura@economía 5(2):102-117 (2020). http://dx.doi.org/10.21704/ne.v5i2.1610

[2] (NABIPOUR, 2020, p. 4) M. Nabipour, P. Nayyeri, H. Jabani, S. S. and A. Mosavi, "Predicting Stock Market Trends Using Machine Learning and Deep Learning Algorithms Via Continuous and Binary Data; a Comparative Analysis," in IEEE Access, vol. 8, pp. 150199-150212, 2020, doi: 10.1109/ACCESS.2020.3015966.

[3] (Wu, 2020, p. 1) Wangwei Wu, (2020), Injury Analysis Based on Machine Learning in NBA Data, recuperado de DOI: 10.4236/jdaip.2020.84017

[24] (Yifan, 2022, p. 1014). Lu Yifan, Ye Tianle, Jiali Zheng (2022). Decision Tree Algorithm in Machine Learning. 2022 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA), recuperado de DOI: 10.1109/A EECA55500.2022.9918857

(LFPDPPP, 2010, p. 2) LEY FEDERAL DE PROTECCIÓN DE DATOS PERSONALES EN POSESIÓN DE LOS PARTICULARES. DOF 05-07-2010, recuperado de https://www.diputados.gob.mx/LeyesBiblio/pdf/LFPDPPP.pdf el 25 de octubre de 2024.

(Banxico, 2020, p. 54). *https://www.banxico.org.mx/publicaciones-y-prensa/reportes-sobre-las-condiciones-de-competencia-enl/%7B6B2ACA7F-4D36-92C0-0E0F-7C09398F06C2%7D.pdf*.

(CONAIF, 2024, p. 2) Política Nacional de Inclusión Financiera, https://www.pnif.mx/acerca/, recuperado el 25 de octubre de 2024.

BIBLIOGRAFIA

- Arnal, J., Del Rincón, D. y Latorre, A. (1996). Bases metodológicas de la investigación educativa. Barcelona: GR92.
- McMillan, J.H. y Schumaker, S. (2005). Investigación educativa. Madrid: PearsonAddison Wesley.
- Rodríguez, G., Gil, J. y García, E. (1999). Metodología de la investigación cualitativa. Archidona, MA: Aljibe
- Zavaleta Osmar (2021). La inclusión financiera en México, retos y oportunidades https://egade.tec.mx/es/egade-ideas/opinion/la-inclusion-financiera-en-mexico-retos-y-oportunidades.