UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

Analítica de datos en una empresa de cobranzas: Modelo para maximizar el alcance del proceso operativo en la empresa de crédito y cobranzas Siccec

Ignacio Barreiro Larrea
Fausto Estrella Muñoz
Mateo Arellano García

Ingeniería Industrial

Trabajo de fin de carrera presentado como requisito
para la obtención del título de
INGENIERO INDUSTRIAL

Quito, 6 de diciembre de 2023

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

HOJA DE CALIFICACIÓN
DE TRABAJO DE FIN DE CARRERA

Analítica de datos en una empresa de cobranzas: Modelo para maximizar el alcance del proceso operativo en la empresa Siccec Crédito y Cobranzas.

Ignacio Barreiro Larrea

Fausto Estrella Muñoz

Mateo Arellano García

Nombre del profesor, Título académico **María Gabriela Baldeón, PhD.**

Quito, 6 de diciembre de 2023

© DERECHOS DE AUTOR

Nombres y apellidos: Ignacio Barreiro Larrea

Código: 00212930

Cédula de identidad: 1719929885

Nombres y apellidos: Fausto Estrella Muñoz

Código: 00212880

Cédula de identidad: 1722447321

Nombres y apellidos: Mateo Arellano García

Código: 00212090

Cédula de identidad: 1720698610

Lugar y fecha: Quito, 6 de diciembre de 2023

ACLARACIÓN PARA PUBLICACIÓN

Nota: El presente trabajo, en su totalidad o cualquiera de sus partes, no debe ser considerado como una publicación, incluso a pesar de estar disponible sin restricciones a través de un repositorio institucional. Esta declaración se alinea con las prácticas y recomendaciones presentadas por el Committee on Publication Ethics COPE descritas por Barbour et al. (2017) Discussion document on best practice for issues around theses publishing, disponible an http://bit.ly/COPETheses.

UNPUBLISHED DOCUMENT

Note: The following capstone project is available through Universidad San Francisco de Quito USFQ institutional repository. Nonetheless, this project – in whole or in part – should not be considered a publication. This statement follows the recommendations presented by the Committee on Publication Ethics COPE described by Barbour et al. (2017) Discussion document on best practice for issues around theses publishing available on http://bit.ly/COPETheses.

RESUMEN

En el contexto del sector financiero, la gestión efectiva de carteras y la recuperación de deudas son imperativos cruciales para las instituciones. Este estudio se centra en SICCEC, una entidad financiera ecuatoriana que enfrenta desafíos en la gestión de sus carteras y busca estrategias para optimizar la recuperación de deudas. Se propone la aplicación de técnicas analíticas y modelos predictivos para identificar patrones y variables que impactan en la recuperación de carteras, con el objetivo de diseñar estrategias basadas en estadística para hacer la recuperación. Se debe realizar un análisis detallado de los datos, identificando factores clave como el saldo vencido, historial de pagos, tipo de flujo utilizado en el contacto, e información demográfica que pueden influir en la recuperación. Mediante modelos predictivos, se evalúa el efecto de los distintos predictores en el porcentaje de recuperación permitiendo la adaptación de estrategias nuevas más eficaces. La implementación de estas estrategias busca optimizar los recursos institucionales al dirigir esfuerzos hacia los casos donde el porcentaje de recuperación sea mayor.

Palabras clave: Gestión de carteras, recuperación de deudas, análisis de datos, modelos

predictivos, estrategias personalizadas, optimización de recursos.

ABSTRACT

In the context of the financial sector, effective portfolio management and debt recovery are crucial imperatives for institutions. This study focuses on SICCEC, an Ecuadorian financial entity facing challenges in portfolio management and seeking strategies to optimize debt recovery. The application of analytical techniques and predictive models is proposed to identify patterns and variables influencing portfolio recovery, aiming to design statistically based strategies for enhanced recovery. A detailed analysis of data is necessary, identifying key factors such as overdue balance, payment history, and the type of contact flow used, and demographic information that can impact recovery. Through predictive models, the effect of various predictors on the recovery percentage is evaluated, allowing for the adaptation of more effective strategies. The implementation of these strategies aims to optimize institutional resources by directing efforts towards cases with a higher recovery percentage.

Keywords: Portfolio management, debt recovery, data analysis, predictive models, customized strategies, resource optimization.

Tabla de Contenidos

Introducción....9
Revisión Litería 11
Metodología 13
Desarrollo del Tema.... 15
Sample 15
Explore.... 17
Modify 24
Model 27
Assess 35
Conclusiones 37
Referencias Bibliográficas.... 40
Anexo A: Flujos de Gestión.... 43
Anexo B: Conclusiones Flujo 44
Anexo C: Variables Consideradas Reporte Ultima Gestión.... 44
Anexo D: Variables Consideradas Reporte Tipo de Flujo.... 44
Anexo E: Importancia de las características.... 44
Anexo F: Conclusión saldo vencido 47
Anexo G: Conclusiones Pagos 47
Anexo H: Conclusiones Día de Pago 48

ÍNDICE DE FIGURAS

Figura #1. Levantamiento de proceso gestión de cobranza Siccec 10

Figura #2. Fases de la metodología SEMMA 13

Figura #3. Bases de Datos y sus Características 15

Figura #4. Promedio Días Mora 18

Figura #5: Promedio Valor Pago por Acción Quito 18

Figura #6. Promedio Valor Pago por Acción Resto Ciudades 18

Figura #7. Promedio Valor Pago por Día Quito 19

Figura #8. Promedio Valor Pago por Día Resto Ciudades 19

Figura #9. Promedio Valor Compromiso vs Valor Pago en Quito 19

Figura #10. Promedio Valor Compromiso vs Valor Pago en Resto Ciudades 19

Figura #11. Promedio de Pago por Ciudad 20

Figura #12. Promedio de Porcentaje de recuperación por día de Pago 20

Figura #13. Porcentaje de Recuperación por Estado Civil 21

Figura #14. Porcentaje de recuperación por Género 21

Figura #15. Porcentaje de Flujos 22

Figura #16. Promedio Costo Total por Flujo 23

Figura #17: Promedio de Ganancia por Flujo 24

Figura #18: Tipo de datos de cada variable 25

Figura #19: Limpieza de Datos en Python 25

Figura #20: Revisión de valores nulos en Python 25

Figura #21: Conversión de datos cualitativos a cuantitativos en Python 26

Figura #22: Predictores y Variable de respuesta reporte tipo de flujo 26

Figura #23: Conversión de datos categóricos a variables dummy 27

Figura #24: Resultados de la Regresión ajustada en Python 28

Figura #25: Resultados de la Regresión logística en Python 29

Figura #26: Resultados del modelo XgBoost en Python 29

Figura #27: Resultados SVM 31

Figura #28: Resultados Algoritmo Redes neuronales 32

Figura #29: Resultado Algoritmo LIGHT GBM 33

Figura #30: Resultados algoritmo Random Forest …..... 34

Figura #31: Top 10 Importancia de las características Random Forest 34

Figura #32: Resumen Métricas Algoritmos..... 36

INTRODUCCIÓN

El mercado de la cobranza es un sector en crecimiento en Ecuador con empresas como Siccec fundada en 1998 que cuenta con más de 1000 empleados y es líder en el sector de cobranzas. Siccec ofrece una variedad de servicios de cobranza y crédito incluyendo la cobranza de préstamos y la gestión de carteras de crédito de las cuales ha logrado recuperar más de 10 mil millones de dólares realizando distintas gestiones.

Uno de los puntos fuertes de SICCEC está en la recuperación de cartera morosa donde utilizan una estrategia conocida como “campañas” para realizar las recuperaciones. Las campañas consisten en enfocar los recursos y el personal para atender en un cierto tiempo (un día por lo general) únicamente a clientes con una característica específica qué puede ser: monto de deuda especifico, ciudad específica, si el cliente es nuevo o antiguo, si el cliente tiene compromisos incumplidos, si el cliente no ha sido contactado por la empresa anteriormente.

Los supervisores del área de cartera morosa toman decisiones diarias sobre en qué clientes deben centrarse para las recuperaciones y crean una campaña específica para atender a esos clientes. Estas campañas se registran en un sistema llamado RECBLUE, donde cada gestor telefónico puede ver los clientes asignados en función de la campaña actual. Además, se notifica a un monitor el medio de contacto preferido para ese cliente, que puede ser una llamada, un SMS (Mensaje de texto), WhatsApp, IVR (Respuesta de voz interactiva) o correo electrónico, elegido por el supervisor.

Dentro del sistema RECBLUE, el gestor tiene acceso automático a la información de los clientes asignados y realiza una evaluación inicial para determinar si se han realizado gestiones previas con ese cliente. Si el cliente es nuevo, los gestores suelen intentar contactarlo mediante una llamada en la que se notifica que su crédito está vencido, proporcionando los detalles de los días y el monto adeudado, siguiendo un guion autorizado por la empresa. En

caso de que haya un contacto positivo, se establece un compromiso de pago, y se espera la respuesta del cliente para aceptar o rechazar los términos propuestos y actualizar el estado del cliente en el sistema. Si la respuesta es negativa se actualiza los datos del cliente en el sistema y el monitor realiza una gestión por un medio alternativo elegido por los supervisores.

Si no se logra el contacto, el gestor realiza una investigación en fuentes como el SRI, CNT o IESS para obtener más información del cliente. Si ninguno de estos enfoques tiene éxito, es posible organizar una visita de campo, que es la última opción que Siccec utiliza con ciertos clientes debido a su alto costo. Se puede encontrar detalles adicionales sobre el proceso en el levantamiento de proceso realizado a continuación.

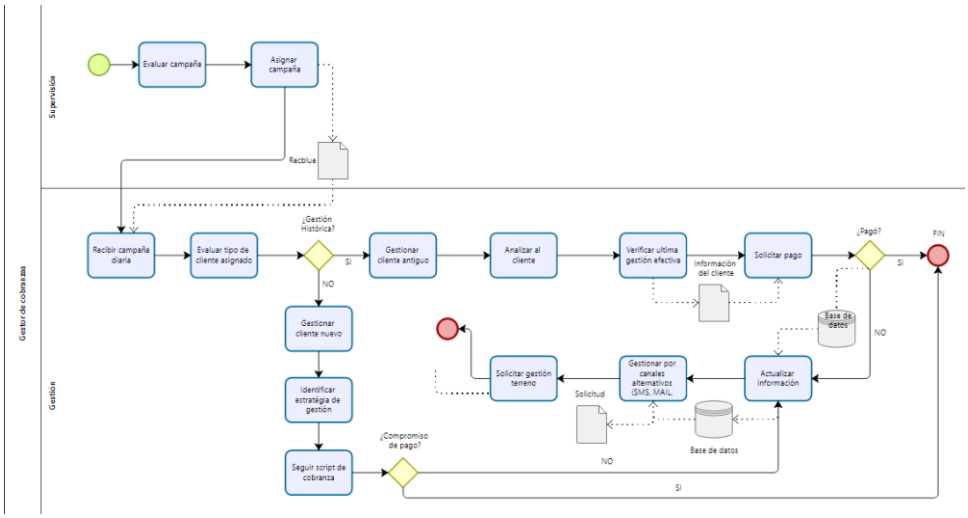


Figura 1: Levantamiento de proceso gestión de cobranza Siccec

Después de analizar el proceso descrito anteriormente hemos logrado identificar una problemática en la manera de gestionar la cartera morosa por parte de Siccec. Nos hemos dado cuenta de que no han establecido una estrategia técnica que guíe a la empresa en la gestión y aumento de contacto con deudores lo que afecta la efectividad al momento de realizar las recuperaciones. La empresa actualmente carece de un plan estratégico sólido para mejorar el proceso de recuperación de deudas. Actualmente, las campañas y los métodos de gestión se

basan únicamente en las decisiones del supervisor, que a menudo carecen de un análisis respaldado para determinar por qué se eligen ciertos medios de contacto (WhatsApp, correo electrónico, IVR, llamada y SMS) para algunos clientes en lugar de otros.

REVISIÓN DE LITERATURA

La industria de las cobranzas se caracteriza por la solvencia, efectividad y sobre todo la cantidad de recursos que se dedican a la gestión de una deuda. La gestión de cobranza se hace en base a una secuencia de estrategias establecidas por expertos que definen entre otras cosas, el canal por el cuál realizar el contacto, es decir telefónico, electrónico o de terreno (Shoghi, 2019). Aparte de la creación de estrategias, el proceso de cobro está inspeccionado por gestores que reciben la información de a quién y de qué forma realizar la gestión de cobranza para obtener una mayor recuperación de cartera. El gestor debe evaluar el comportamiento pasado del deudor y la posibilidad de pago de este para poder tomar decisiones sobre cómo gestionar en un futuro (Koehler, B et al., 2022).

Optimizar el proceso de cobranza es una tarea clave en el sector financiero, ya que su eficiencia incide directamente en la liquidez y estabilidad financiera de las empresas que prestan el dinero. En un entorno cada vez más competitivo donde los pagos atrasados pueden afectar seriamente el desempeño financiero, la gestión del cobro de deudas se vuelve fundamental (Pilla, 2021). Para entender de mejor forma el comportamiento de los deudores se ha realizado múltiples estudios que analizan la morosidad. En el estudio de Espino et al, se analizó una población de 85,000 clientes de una empresa de seguros médicos para lograr predecir el riesgo de morosidad de cada cliente antes de entregarle un servicio de seguro. Se realizaron modelos predictivos de árboles de decisión con el fin de lograr tener estrategias anticipadas para los clientes con mayor riesgo de morosidad. (Espino Quiñones et al., 2018).

En un estudio realizado por Vélez et al, los autores aplican una regresión logística a una base de datos de 16,000 clientes morosos en bancos para entender en base a las características socioeconómicas de un cliente que cantidad de deuda dentro de una cartera se puede recuperar. En este caso utilizaron variables como la morosidad, el endeudamiento y los ingresos de las personas y lograron encontrar que el 50% de las carteras comercial, de microcrédito y de consumo son recuperables (Vélez et al., 2017). De igual manera la inclusión de estadística descriptiva y un análisis organizado de los datos de cobranza son útiles para evaluar el comportamiento de los deudores y entender si existen problemas. En el estudio realizado en la Universidad Peruana Unión por Romero et al, los autores realizan un análisis de datos profundo a alrededor de 500 padres de familia de estudiantes del colegio adventista José de San Martin. En este análisis lograron encontrar que alrededor del 25% de padres tenía morosidad en sus pagos y decidieron analizar las variables por las cuáles no se estaban efectuando los pagos. Los resultados obtenidos fueron que el sistema de cobranza del colegio no estaba bien estructurado y la mayoría de los padres no eran notificados de las deudas pendientes. Una vez que lograron organizar el sistema de cobranza mediante una plataforma automatizada, los porcentajes de padres con morosidad se redujeron a un 12% brindando mayor fluidez económica a la institución educativa. (Romero Carazas et al., 2021).

Por el otro lado, existe otro estudio realizado por Herrera acerca de la inclusión financiera de clientes sin historial crediticio se puede apreciar la utilización de analítica de datos para predecir si un crédito será pagado o formará parte de una cartera irrecuperable. Dentro de este estudio se utilizaron datos de 1 millón de créditos entregados entre Diciembre de 2018 a marzo de 2020 en los que se tenía información de género, tiempo de la deuda, cantidad de la deuda, fecha último pago, entre otros. Se utilizaron algoritmos denominados de clasificación que basados en los datos del crédito logren predecir si un crédito será pagado o

será irrecuperable, entre estos estaba la regresión logística, árboles de decisión, k vecinos, y redes neuronales. Una vez realizado 6 algoritmos se logró identificar que las redes neuronales tienen una mayor exactitud al predecir el pago o no de un crédito.

METODOLOGÍA

Las empresas de crédito y cobranza como Siccec manejan conjuntos de datos masivos qué deben ser procesados para mejorar la eficiencia y productividad mediante distintos proyectos. Los proyectos qué tienen estas características se los conoce como proyectos de "Data Mining o Minería de datos" en los cuales se espera obtener información valiosa en base a grandes cantidades de datos.

Existe una metodología desarrollada por SAS Institute qué ayuda a realizar una minería de datos adecuada facilitando el proceso de análisis y ayudando a encontrar patrones dentro de los datos qué nos puedan ayudar a tomar decisiones dentro de una empresa. La metodología conocida como SEMMA sigue un orden establecido en sus siglas con el acrónimo de las 5 fases qué componen la metodología: Sampling (muestreo), Exploring (exploración), Modifying (manipulación), Modeling (modelado), y Assessing (valoración).

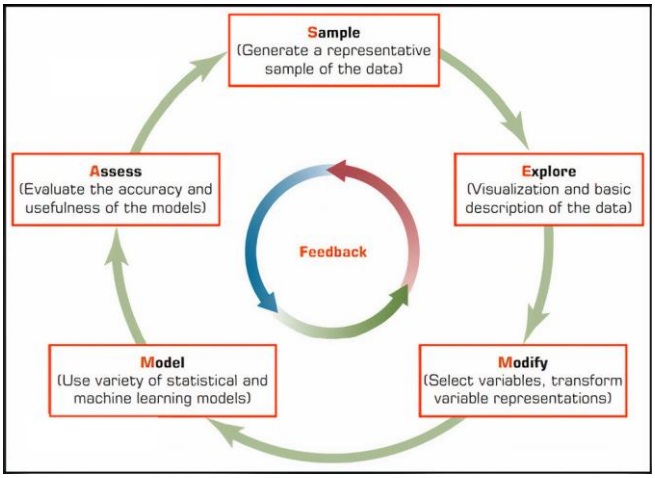


Figura 2. Fases de la metodología SEMMA (Shalda y Denle 2018)

Muestreo: Dentro de la etapa de muestreo se debe escoger una muestra dentro del conjunto de datos para poder aplicar el análisis. Es esencial qué se tenga una muestra significativa de la población y qué no sea demasiado grande para qué se logre realizar un análisis adecuado. Esta fase tiene como objetivo reducir el tiempo qué demore el análisis de los datos ya qué nos enfocamos en una muestra específica para trabajar sobre ella (Vanrel, 2011).

Exploración: La etapa de exploración tiene como objetivo el familiarizar al analista con los datos y simplificarlos de alguna manera. En esta etapa se busca relaciones, tendencias, o anomalías dentro de los datos qué deben eliminarse para lograr obtener información valiosa acerca de los datos. Es útil dentro de esta etapa visualizar los datos en gráficas para obtener información valiosa acerca de qué variables pueden ayudar a realizar el análisis qué estamos buscando hacer.

Manipulación: En la etapa de manipulación nos enfocamos en qué los datos tengan el formato adecuado para ser introducidos al modelo qué hayamos elegido. Se debe transformar cualquier dato qué no cumpla el formato qué se busca para poder facilitar el modelado más adelante.

Modelado: En la etapa de modelado se utilizan algoritmos y herramientas de software qué ayudan a predecir o arrojar un resultado acerca de lo que se está buscando. Se utilizan modelos estadísticos, redes neuronales, árboles de decisión, lógica difusa y muchos otros modelos para encontrar una solución.

Valoración: La última etapa es la valoración donde se analiza los resultados obtenidos y se realiza un análisis de bondad de los resultados para ver los resultados son útiles y verificados.

DESARROLLO DEL TEMA

SAMPLE:

La etapa de Sample es el punto de partida en la metodología SEMMA, y su objetivo principal es la selección y preparación de los datos para un análisis más profundo. En un entorno de gestión de cobranza, esta etapa se convierte en un pilar esencial para comprender el flujo de gestiones, la recuperación de deudas y la información demográfica relevante de los deudores.

En Siccec se manejan múltiples bases de datos como el Query de Pagos, Query de Asignaciones, Query de Agencia Virtual y Query de Gestiones. Cada una de estas bases de datos desempeña un papel crucial en la gestión de cobranza ya que contiene datos específicos acerca del cliente o la gestión realizada. Es por esto por lo que se debe entender qué información es importante en cada base para identificar como juntar toda la información útil y empezar a trabajar para tener dos bases unificadas. Al momento de realizar esto nos aseguramos de que únicamente se reflejen aquellas gestiones que fueron efectivas, es decir donde se recibieron pagos por parte de los deudores. En la figura 3 podemos observar el tamaño de cada una de las bases y la información que contiene de los meses de julio, agosto y septiembre.

Nombre Base de Datos	TAMAÑO (Filas x Columnas)	Detalles
QUERY_ASIGNACION_AGO_JUL_SEP.xlsx	2431 x 73	Base de datos donde se encuentran los deudores asigandos a la cartera con información demográfica de la misma.
QUERY_PAGOS_AGO_JUL_SEP.xlsx	7656 x 18	Base de datos donde se encuentran los registros de pagos de los deudores una vez hecha la gestión de cobranza.
QUERY_GESTIONES_AGO_JUL_SEP.xlsx	67032 x 54	Base de datos donde se encuentran todas las gestiones realizadas de parte de los gestores (llamadas, mail, sms, ivr, whats app).
QUERY_AGENCIA_VIRTUAL_AGO_JUL_SEP.xlsx	12031 x 17	Base de datos donde estan los envios y las respuestas de los envios tanto de sms, mail e ivr.
UltimaGestion_CONCATENADO_JUL_AGO_SEP.xlsx	7772 x 27	Base de datos en el cual con cruces de las bases explicadas anteriormente se toma en cuenta características de la última gestión.
REPORTE_CONSOLIDADO_TIPO_DE_FLUJO.xlsx	7212 x 24	Base de datos en el cual con cruces de las bases explicadas anteriormente se toma en cuenta el tipo de flujo para la gestión que llego al pago

Figura 3: Bases de Datos y sus Características.

Para los dos reportes que realizamos con el apoyo de las bases de datos que Siccec nos otorgó, en primer lugar, se realizó un cruce entre el Query de gestiones y el Query de pagos concatenando el ID de crédito y número de operación de cada gestión. El objetivo de

este cruce fue lograr extraer variables que estaban presentes en el Query de gestiones y no en el Query de pagos, por ejemplo, la fecha de la gestión. El siguiente paso fue cruzar la nueva base ya concatenada con el Query de asignaciones. Esto se realizó de igual forma con el número de operación, ID de crédito y también la fecha de la gestión que se extrajo anteriormente. El objetivo de este nuevo cruce fue obtener los datos que el cedente nos brinda de cada cliente por medio del Query de asignaciones.

Después, al tener las fechas cruzadas, consideramos únicamente los datos de clientes que hicieron un pago para poder analizar las variables que pudieron llevar a ese pago. En esta base consolidada consideramos únicamente las gestiones de los últimos 5 días laborales antes del pago para determinar cuál fue el verdadero flujo que llevó al pago. Una vez que teníamos un solo reporte consolidado de julio, agosto y septiembre con toda la información necesaria para analizar, lo dividimos en dos reportes.

El primer reporte (UltimaGestion_CONCATENADO_JUL_AGO_SEP.xlsx) que utilizamos fue un reporte que contiene todas las variables relacionadas a las características de la deuda para realizar un estudio de qué factores pueden afectar en el porcentaje de recuperación. Una de las variables que incluimos en este reporte fue la Acción, que indica si la gestión se realiza de forma telefónica, a través de un canal electrónico o una gestión masiva. Otro campo que utilizamos es el producto, esto quiere decir si es que el cliente utilizó MasterCard o visa para entender si existe alguna preferencia de los clientes al momento de pagar sus deudas. Por otro lado, incluimos la ciudad homologada que quiere decir Quito o el resto de las ciudades del Ecuador, días mora del deudor, el valor del pago que realizó el deudor, saldo vencido, entre otras.

Creamos un segundo reporte (QUERY_TIPO_FLUJO_JUL_AGO_SEP.xlsx), con el objetivo de conocer el tipo de flujo que se utilizó dentro de la gestión que quiere decir cuál de

los 5 canales (Llamadas, SMS, Mail, WhatsApp o IVR) o combinaciones de estos se usó. En este caso la información se presentaba verticalmente por lo que tuvimos que realizar un pivote a las gestiones de acuerdo con las fechas para tener la combinación de canales que se usó para conseguir el pago. En esta etapa también creamos nuevas variables útiles para nuestro informe en forma de indicadores críticos como la recuperación (que implica dividir el pago realizado por el cliente entre el saldo vencido), el costo de la gestión (una suma de los costos unitarios para cada tipo de gestión) y la ganancia (que se obtiene restando el honorario del pago, 7% del pago, al costo del flujo). Estos indicadores son fundamentales para evaluar el rendimiento y la rentabilidad de las gestiones de cobranza para poder comparar los distintos tipos de flujo en base a los resultados de estos nuevos campos.

EXPLORE:

Analizamos e interpretamos de manera gráfica los datos de las bases de datos obtenidas en la etapa de Sample que cuentan con datos de los meses de julio, agosto y septiembre. Para ello utilizamos la herramienta Power BI para entender la relación entre variables que teníamos dentro de la base de datos y crear reportes interactivos para poder filtrar de mejor manera nuestra información. En la etapa de exploración analizamos el separar los datos en Quito y resto de ciudades ya que alrededor del 50% de datos que maneja Siccec son de la ciudad de Quito y el otro 50% es la suma de gestiones en el resto del país.

En la Figura 4 podemos observar el promedio de días mora tanto para la ciudad de Quito como para el resto de las ciudades, en este caso al trabajar con promedios se puede observar que los clientes en ambas ciudades tienen en promedio 400 días mora.

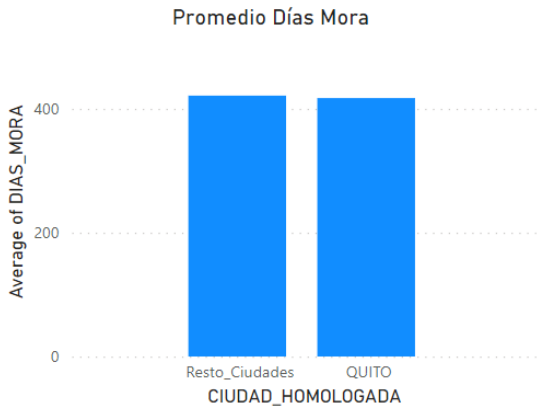


Figura 4: Promedio Días Mora

Por otro lado, podemos observar los promedios de valor pago para cada acción en Quito en la figura 5 y para el resto de las ciudades en la figura 6. En el caso de Quito podemos observar que el mayor promedio de pago se da cuando se notifica al cliente por medio de canales electrónico. En segundo lugar, tenemos la gestión y por ultimo las gestiones telefónicas "hacer llamada", mientras que para el resto de las ciudades el primer lugar coincide con Quito, mientras que ahora el segundo promedio más alto es las gestiones telefónicas y por último la gestión masiva. Al trabajar con promedios de los valores de pago es más fácil compararlos, dado que, si analizamos la Suma, las gestiones telefónicas serían considerablemente superiores al resto al ser las más utilizadas.

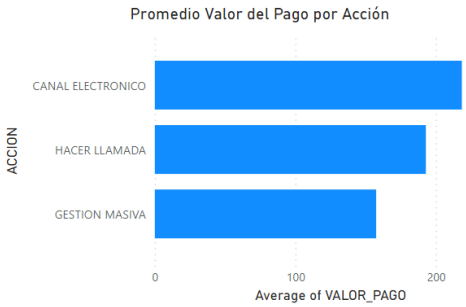


Figura 5: Promedio Valor Pago Por Acción Quito

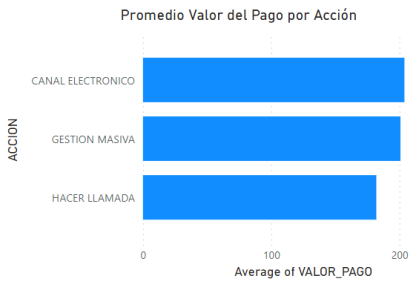


Figura 6: Promedio Valor Pago por Resto Ciudades

También tenemos el promedio de los valores pagados para Julio, agosto y septiembre en la figura 7 para Quito y en la figura 8 se observa para el resto de las ciudades. Realizamos

este gráfico ya que nos proporciona información relevante para determinar en promedio cual día pagan más los deudores mostrándonos que el promedio de valores pagados crece a finales del mes mostrándonos que gran parte de los compromisos de pagos están para dicha fecha.

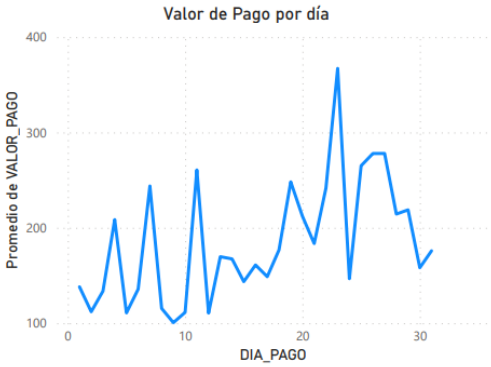


Figura 7: Promedio Jul-Sep. Valor de por Pago por Día Quito

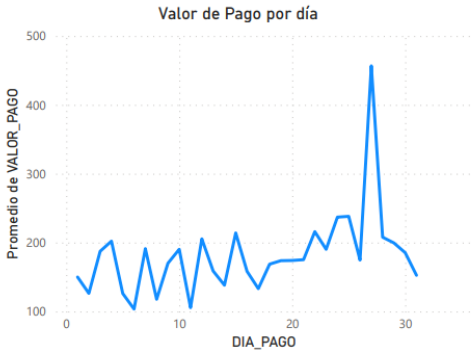


Figura 8: Promedio Jul-Sep. Valor de Pago Día Resto Ciudades

Por otro lado, podemos observar en las figuras 9 y 10, que tanto para Quito como para el resto de las ciudades el valor de compromiso promedio es considerablemente superior al valor promedio de pago, y esto nos indica que la cartera estudiada es una cartera castigada en la cual existe una alta tasa de incumplimiento de pagos y existe la necesidad de tener nuevas estrategias para incrementar la recuperación.

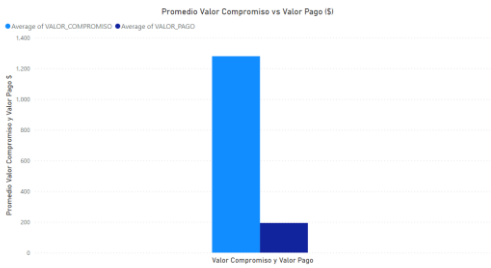


Figura 9: Promedio Valor Compromiso vs Valor Pago Quito

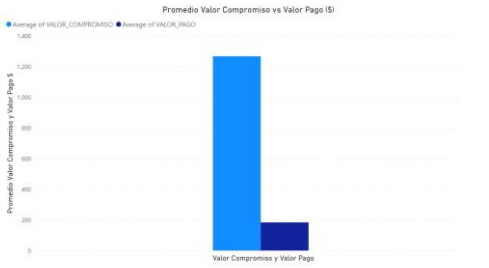


Figura 10: Promedio Valor Compromiso vs Valor pago Resto Ciudades

También, podemos observar en la Figura 11 que el promedio del pago realizado por cada cliente en Quito es ligeramente superior que el pago en el resto de las ciudades. Pero en

general los valores similares nos indican que no existe un gran desequilibrio en relación con los pagos de los clientes en Quito y el resto del país.

Figura 11: Promedio de Pago por Ciudad

Por otro lado, para el reporte de tipo de flujo, podemos analizar otras variables, como lo son el día de pago, estado civil de los deudores, y también el género de cada deudor. Por ello, realizamos dichas gráficas, lo cual podemos observar a continuación, en la figura 12, se puede observar que existe un incremento en recuperación cuando se aproxima el fin de mes.

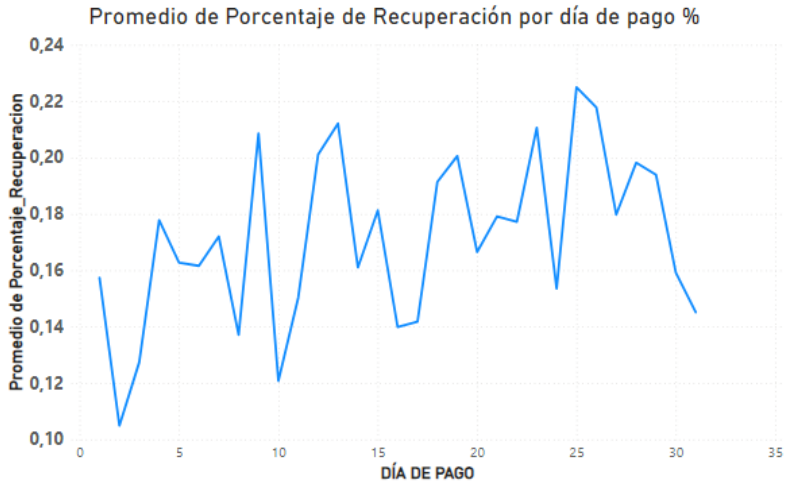


Figura 12: Promedio de Porcentaje de recuperación por día de Pago

A continuación, en la figura 13, podemos observar el porcentaje de recuperación por estado civil, es evidente que el divorciado tiene un mayor promedio, el cual el segundo es el soltero, en tercer lugar, el casado, y en los últimos lugares aquellas personas con unión libre o

viudos.

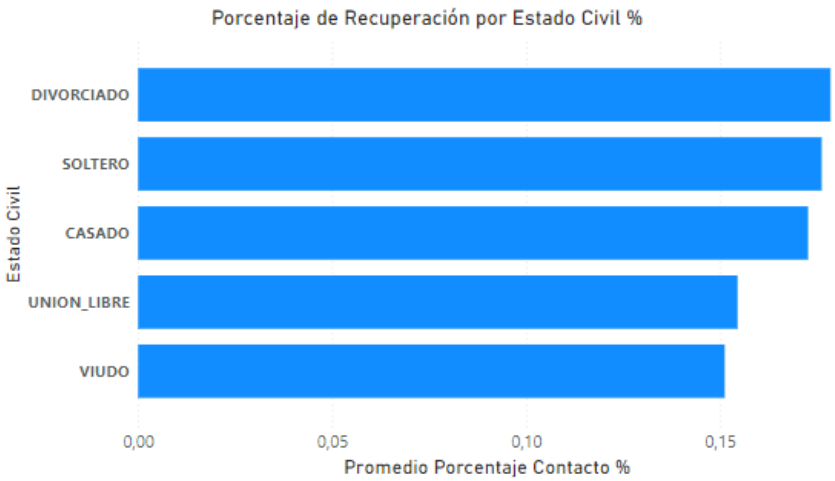


Figura 13: Porcentaje de Recuperación por Estado Civil.

Por otro lado, también es importante observar el porcentaje de recuperación en promedio para cada género, en la figura 14, podemos visualizar que el género que tiene mayor recuperación en promedio es el masculino, con casi el 20%.

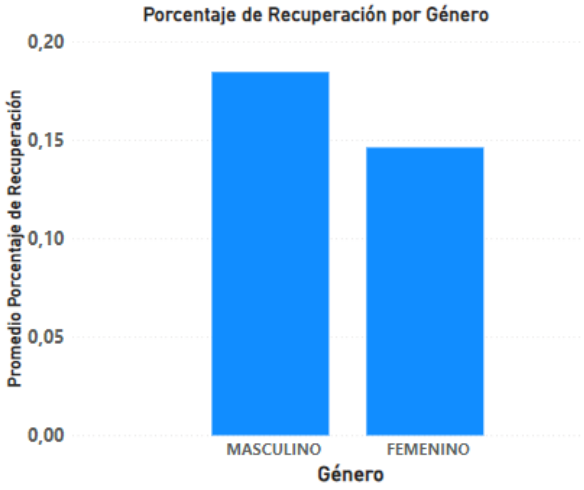


Figura 14: Porcentaje de recuperación por Género

Otro análisis fundamental en la exploración es el detalle de los flujos utilizados para realizar la gestión. Todas las posibles combinaciones de los flujos entre SMS, WhatsApp, correo, IVR y llamada pueden verse en el anexo B.

En la figura 15 podemos observar el porcentaje de recuperación por flujo, en el cual, el flujo 21 cuenta con un promedio de superior al resto de flujos (9.33%), dicho flujo hace referencia a la combinación de gestiones entre IVR y Llamada telefónica.

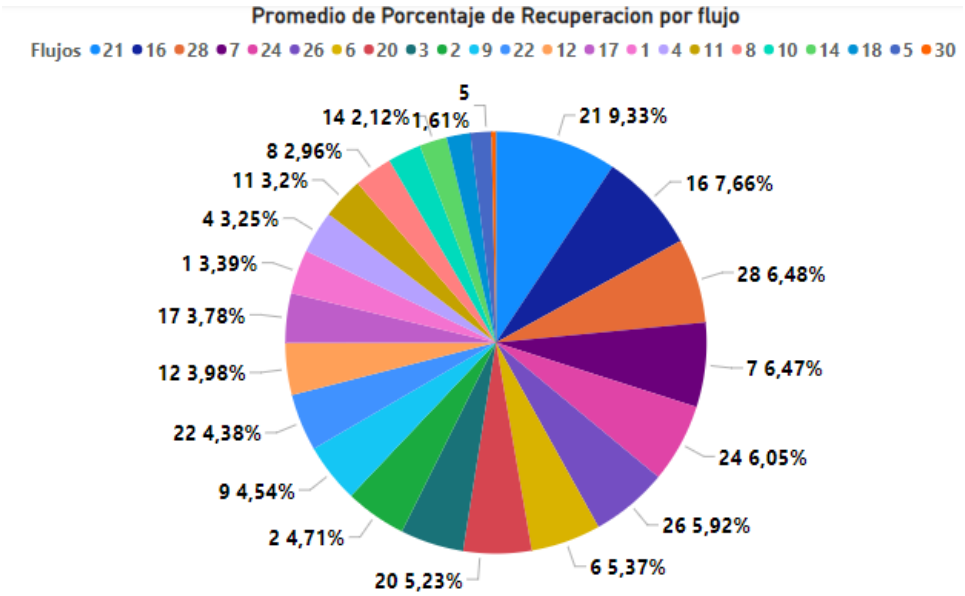


Figura 15: Porcentaje de Flujos

Es importante resaltar y comprobar, que mientras más canales tiene el flujo, existe un costo mayor. Sin embargo, hay que entender que, para minimizar el costo, la mejor alternativa no es siempre minimizar los canales sino ver la relación entre el costo y la recuperación. En la figura 16 observamos el costo promedio por cada flujo en centavos de dólar. Se observa que el flujo con mayor costo en promedio es el flujo 26, que cuenta con los canales: SMS, LLAMADA, MAIL y WHATSAPP.

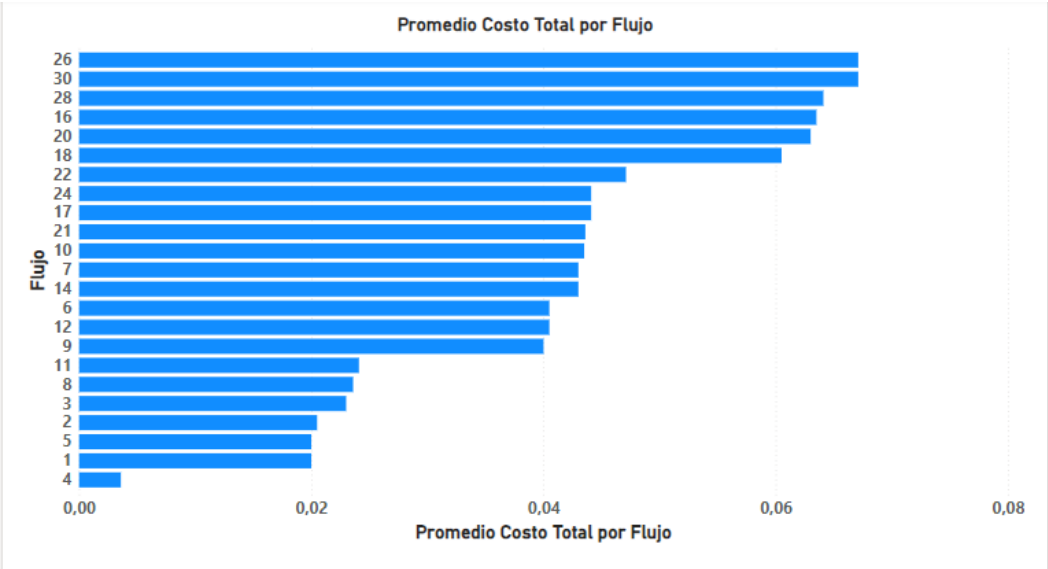


Figura 16: Promedio Costo Total por Flujo

Finalmente exploramos la ganancia promedio con la que se queda Siccec por cada flujo que es la resta entre el honorario y el costo total por flujo, tal como se observa en la figura 17. Logramos ver que flujo 16 es superior al resto ya que el costo total por cada gestión no es muy elevado. Es importante considerar que este análisis se realizó solo con las personas que pagaron una deuda pendiente y no consideramos todos los casos en donde se invirtió en contactar a una persona y no se obtuvo respuesta. Esto se debe a que es más importante entender primeramente que se hizo bien en los casos positivos para poder luego expandir las estrategias a los clientes que no han pagado aún. Por lo tanto, la ganancia con mayor ganancia en promedio es el flujo 16 que cuenta con los siguientes canales: LLAMADA, SMS y WHATSAPP.

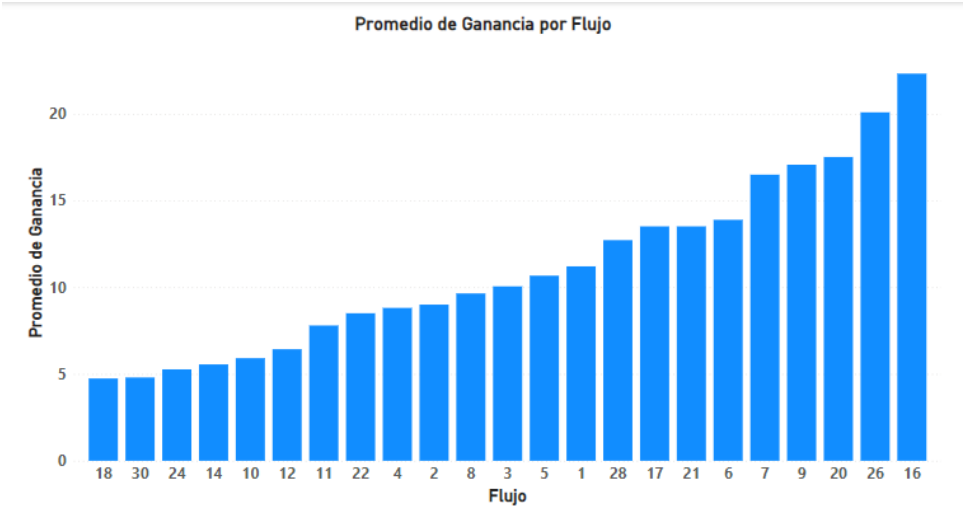


Figura 17: Promedio de Ganancia por Flujo

MODIFY:

En la etapa de modificación realizamos una limpieza de datos que aporte en la obtención de nuestro objetivo. Buscamos realizar una regresión a la base de datos utilizando las columnas que nos pudieran presentar información relevante de como reconocer cuales son aquellas características que debe tener un cliente para calificar como un mejor pagador. Para esto empezamos convirtiendo los datos de días mora y el saldo vencido de los clientes en 4 rangos en base a los cuartiles de cada columna, para que pueda encajar de mejor manera en la regresión como una variable categórica. Luego utilizamos Python dentro de Jupyter Notebook para realizar un EDA (Exploratory Data Analysis) en el cuál identificamos el tamaño de la base de datos (Figura 18) y eliminamos las columnas que no calificaban para este propósito como fechas, cedente, motivo de no pago, etc. (Figura 19).

```
ID_DEUDOR                  int64
ID_CREDITO                 int64
NUMERO_OPERACION           int64
ANIO                       int64
MES                        int64
DIA_PAGO                   int64
VALOR_PAGO               float64
Concat                    object
Saldo_Vencido            float64
Porcentaje_Recuperacion  float64
Honorario                float64
Costo_Total_Flujo        float64
Ganancia                 float64
ACCION                    object
RESPUESTA                 object
DIAGNOSTICO               object
CONTACTO                  object
FECHA_COMPROMISO          object
VALOR_COMPROMISO         float64
FECHA_GESTION             object
MOTIVO_NO_PAGO            object
CIUDAD                    object
CIUDAD_HOMOLOGADA         object
DIAS_MORA                  int64
DIAS_MORA_PAGOS            int64
PRODUCTO_CR               object
SALDO_ORIGINAL           float64
ULTIMA_GESTION            object
INTENSIDAD               float64
ULTIMA_FECHA_PAGO         object
FECHA_VENCIMIENTO        float64
GRUPO_PRODUCTO            object
CEDENTE                   object
DIAS_MORA_PAGOS_RANGOS    object
Saldo_Vencido_Rangos      object
```

Figura 18: Tipo de datos de cada variable

	VALOR_PAGO	Porcentaje_Recuperacion	ACCION	CIUDAD_HOMOLOGADA	PRODUCTO_CR	DIAS_MORA_PAGOS_RANGOS	Saldo_Vencido_
0	230.41	0.27	HACER_LLAMADA	QUITO	MASTERCARD	Rango_1a121	Rango_644.76a
1	58.00	0.11	HACER_LLAMADA	Resto_Ciudades	VISA	Rango_890_en_adelante	Rango_30.58
2	400.00	0.06	HACER_LLAMADA	QUITO	MASTERCARD	Rango_890_en_adelante	Rango_2908.45_en_
3	400.00	0.06	CANAL_ELECTRONICO	QUITO	MASTERCARD	Rango_890_en_adelante	Rango_2908.45_en_
4	400.00	0.06	HACER_LLAMADA	QUITO	MASTERCARD	Rango_890_en_adelante	Rango_2908.45_en_

Figura 19: Limpieza de Datos en Python

Una vez que realizamos la limpieza, procedimos a verificar que los datos que nos quedan no cuentan con valores nulos que podrían afectar al momento de realizar nuestro modelo de regresión. Esto se puede apreciar en la Figura 20.

```
VALOR_PAGO                False
Porcentaje_Recuperacion   False
ACCION                    False
CIUDAD_HOMOLOGADA         False
PRODUCTO_CR               False
DIAS_MORA_PAGOS_RANGOS    False
Saldo_Vencido_Rangos      False
dtype: bool
```

Figura 20: Revisión de valores nulos en Python

En vista de que las columnas con características que podrían influenciar el porcentaje de recuperación de una cartera no contaban con nulos, pudimos preparar el marco de datos para realizar la regresión.

La regresión que seleccionamos es una regresión de mínimos cuadrados ordinarios ya que este tipo de regresión nos permitió describir la relación entre una o más variables independientes cuantitativas y nuestra variable dependiente en base al p-value y sus coeficientes. Para esto, convertimos todas las columnas que contaban con datos cualitativos a cuantitativos dejando todo el "Dataframe" con datos numéricos para poder ya realizar la regresión de mínimos cuadrados ordinarios.

VALOR_PAGO	Porcentaje_Recuperacion	ACCION	CIUDAD_HOMOLOGADA	PRODUCTO_CR	DIAS_MORA_PAGOS_RANGOS	Saldo_Vencido_Rangos
230.41	0.27	HACER_LLAMADA	QUITO	MASTERCARD	Rango_1a121	Rango_644.76a1219.27
58.00	0.11	HACER_LLAMADA	Resto_Ciudades	VISA	Rango_890_en_adelante	Rango_30.58a644.75
400.00	0.06	HACER_LLAMADA	QUITO	MASTERCARD	Rango_890_en_adelante	Rango_2908.45_en_adelante
400.00	0.06	CANAL_ELECTRONICO	QUITO	MASTERCARD	Rango_890_en_adelante	Rango_2908.45_en_adelante
400.00	0.06	HACER_LLAMADA	QUITO	MASTERCARD	Rango_890_en_adelante	Rango_2908.45_en_adelante

Figura 21: Conversión de datos cualitativos a cuantitativos en Python

Para realizar el reporte de tipo de flujo utilizamos la misma base de datos que ya teníamos limpia hasta el momento y aumentamos algunas columnas que vimos que nos podían aportar con nuestro análisis y no se consideraron en la parte de la última gestión como: el género del deudor, el estado civil, la ciudad en donde reside y el tipo de flujo que se utilizó para mantener el contacto. El tipo de flujo se lo paso de numérico a categórico.

```
DIA_PAGO                int64
PAGO                  float64
Genero_Deudor          object
Estado_Civil           object
Ciudad                 object
Saldo_Vencido         float64
Porcentaje_Recuperacion float64
flujo_1                 uint8
```

Figura 22: Predictores y Variable de respuesta reporte tipo de flujo.

MODEL:

Reporte Última Gestión:

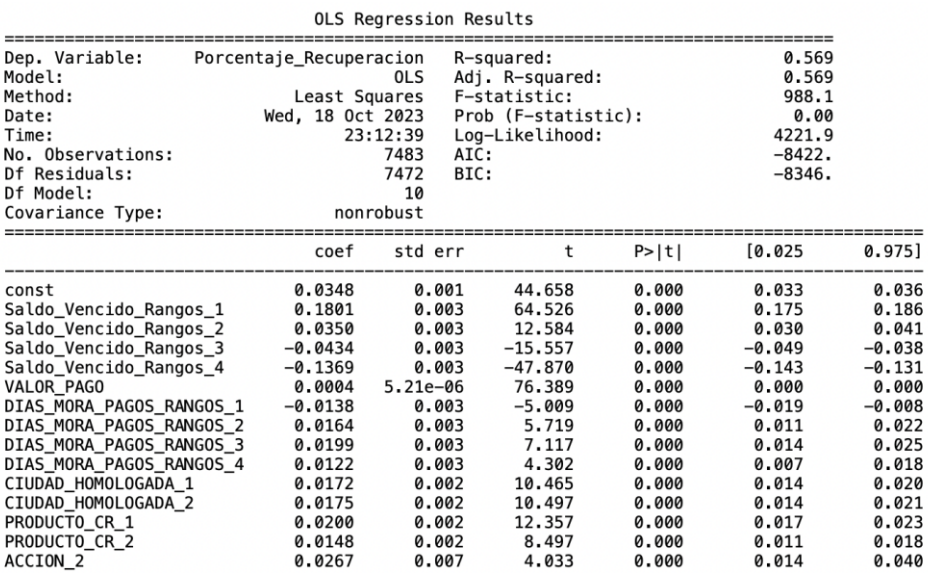
La cuarta etapa de la metodología SEMMA es la etapa de Modelar donde realizamos una regresión. Para poder realizar la regresión convertimos los predictores categóricos en variables dummies. Esto se realizó para que cada categoría dentro de cada predictor se represente por una variable dummy y de esta manera se pueda identificar a detalle, cuál categoría es la que más influye en el porcentaje de recuperación de un cliente.

	Saldo_Vencido_Rangos_1	Saldo_Vencido_Rangos_2	Saldo_Vencido_Rangos_3	Saldo_Vencido_Rangos_4	VALOR_PAGO	DIAS_MORA_PAGOS_RANGOS_1
0	0	1	0	0	230.41	1
1	1	0	0	0	58.00	0
2	0	0	0	1	400.00	0
3	0	0	0	1	400.00	0
4	0	0	0	1	400.00	0

Figura 23: Conversión de datos categóricos a variables dummy

Con esta base de datos ya preparada, se realizó la regresión para determinar cuáles variables predictoras son estadísticamente significativas y tienen un impacto en la variable de respuesta, que en este caso es el Porcentaje de Recuperación de la cartera. Se encontró aquellas variables que no son estadísticamente significativas como Acción 1 y Acción 3 (Hacer llamada y Gestión Masiva, respectivamente) y se procedió a descartar estas variables.

Este resultado es coherente debido a que a todos los deudores se les llama y las gestiones masivas son como lo dice su nombre – masivas. Con estos datos descartados, se realizó la regresión nuevamente, únicamente con los predictores que si son estadísticamente significativos para poder analizar sus coeficientes.



OLS Regression Results

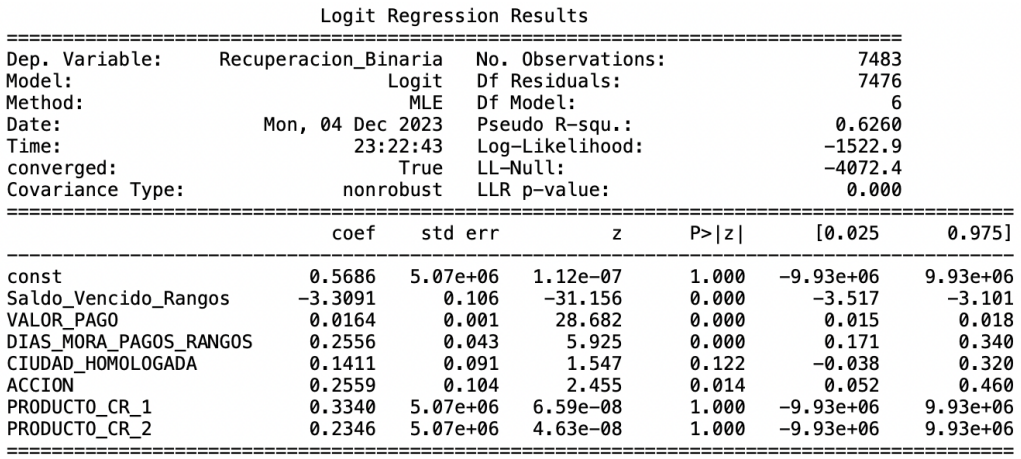
Dep. Variable:	Porcentaje_Recuperacion	R-squared:	0.569
Model:	OLS	Adj. R-squared:	0.569
Method:	Least Squares	F-statistic:	988.1
Date:	Wed, 18 Oct 2023	Prob (F-statistic):	0.00
Time:	23:12:39	Log-Likelihood:	4221.9
No. Observations:	7483	AIC:	-8422.
Df Residuals:	7472	BIC:	-8346.
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0348	0.001	44.658	0.000	0.033	0.036
Saldo_Vencido_Rangos_1	0.1801	0.003	64.526	0.000	0.175	0.186
Saldo_Vencido_Rangos_2	0.0350	0.003	12.584	0.000	0.030	0.041
Saldo_Vencido_Rangos_3	-0.0434	0.003	-15.557	0.000	-0.049	-0.038
Saldo_Vencido_Rangos_4	-0.1369	0.003	-47.870	0.000	-0.143	-0.131
VALOR_PAGO	0.0004	5.21e-06	76.389	0.000	0.000	0.000
DIAS_MORA_PAGOS_RANGOS_1	-0.0138	0.003	-5.009	0.000	-0.019	-0.008
DIAS_MORA_PAGOS_RANGOS_2	0.0164	0.003	5.719	0.000	0.011	0.022
DIAS_MORA_PAGOS_RANGOS_3	0.0199	0.003	7.117	0.000	0.014	0.025
DIAS_MORA_PAGOS_RANGOS_4	0.0122	0.003	4.302	0.000	0.007	0.018
CIUDAD_HOMOLOGADA_1	0.0172	0.002	10.465	0.000	0.014	0.020
CIUDAD_HOMOLOGADA_2	0.0175	0.002	10.497	0.000	0.014	0.021
PRODUCTO_CR_1	0.0200	0.002	12.357	0.000	0.017	0.023
PRODUCTO_CR_2	0.0148	0.002	8.497	0.000	0.011	0.018
ACCION_2	0.0267	0.007	4.033	0.000	0.014	0.040

Figura 24: Resultados de la Regresión ajustada en Python

En base a los coeficientes pudimos determinar que aquellas personas que tienen un saldo vencido más bajo son más propensas a pagar el valor de la deuda y conforme el saldo aumente, disminuye la probabilidad de pago. El valor pago tiene un coeficiente positivo y un error estándar muy bajo indicando que mientras mayor sea el valor pagado, mayor es la recuperación. Con respecto a días mora, es interesante ver como aquellas personas que tienen días mora de 312 a 889 indican tener la mayor probabilidad de pago. Finalmente, la acción del canal electrónico nos indica que aquellas personas que tuvieron un acercamiento por medio de un mail, un SMS, un IVR o un WhatsApp son 2.7% más probables a pagar la deuda que aquellos que no fueron contactados por medio de un canal electrónico.

Debido al bajo resultado del R cuadrado que nos ofreció la regresión de mínimos cuadrados ordinarios, se realizó una regresión logística y un modelo XGBoost para intentar elevar el R cuadrado del modelo e interpretar de manera comparativa los resultados.



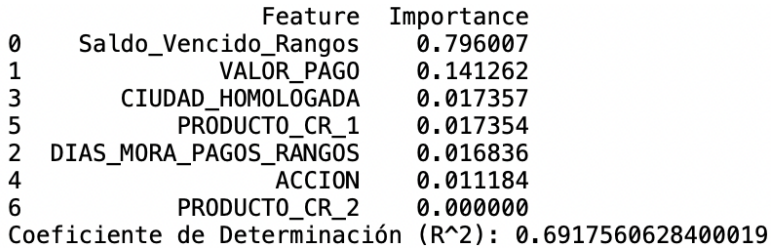
Logit Regression Results

Dep. Variable:	Recuperacion_Binaria	No. Observations:	7483
Model:	Logit	Df Residuals:	7476
Method:	MLE	Df Model:	6
Date:	Mon, 04 Dec 2023	Pseudo R-squ.:	0.6260
Time:	23:22:43	Log-Likelihood:	-1522.9
converged:	True	LL-Null:	-4072.4
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.5686	5.07e+06	1.12e-07	1.000	-9.93e+06	9.93e+06
Saldo_Vencido_Rangos	-3.3091	0.106	-31.156	0.000	-3.517	-3.101
VALOR_PAGO	0.0164	0.001	28.682	0.000	0.015	0.018
DIAS_MORA_PAGOS_RANGOS	0.2556	0.043	5.925	0.000	0.171	0.340
CIUDAD_HOMOLOGADA	0.1411	0.091	1.547	0.122	-0.038	0.320
ACCION	0.2559	0.104	2.455	0.014	0.052	0.460
PRODUCTO_CR_1	0.3340	5.07e+06	6.59e-08	1.000	-9.93e+06	9.93e+06
PRODUCTO_CR_2	0.2346	5.07e+06	4.63e-08	1.000	-9.93e+06	9.93e+06

Figura 25: Resultados de la Regresión logística en Python

En el caso de la regresión logística, no se convirtió los datos a variables dummy a excepción de la variable predictora Producto_CR debido a que es una variable categórica nominal y se mantuvo el resto en su naturaleza numérica y categórica ordinal. El R cuadrado fue superior al anterior obteniendo un valor de 62.60%. La ciudad y el producto tienen un p-value mayor a 0.05 indicando que no son estadísticamente significativos a diferencia del resto de variables. Con respecto a los coeficientes, el que más destaca es el coeficiente de Saldo Vencido con un -3.31. Esto indica que mientras más aumente el saldo vencido, la probabilidad de recuperación disminuye.



	Feature	Importance
0	Saldo_Vencido_Rangos	0.796007
1	VALOR_PAGO	0.141262
3	CIUDAD_HOMOLOGADA	0.017357
5	PRODUCTO_CR_1	0.017354
2	DIAS_MORA_PAGOS_RANGOS	0.016836
4	ACCION	0.011184
6	PRODUCTO_CR_2	0.000000

Coeficiente de Determinación (R^2): 0.6917560628400019

Figura 26: Resultados del modelo XGBoost en Python

El modelo fue separado entre un conjunto de entrenamiento y de prueba (80/20) y corrido con una optimización de hiperparámetros para encontrar por medio de un grid_search la combinación óptima de hiperparámetros a través de la validación cruzada con el fin de

obtener una mejor precisión en el modelo. Esto nos presentó un R cuadrado del 69.18% siendo el resultado más elevado de los tres modelos. Adicionalmente, en este modelo matemático se aplicó el *Feature Importance* – sin convertir las variables a variables dummy – que nos permite determinar cuáles variables afectan mayormente a la variable de respuesta. Inmediatamente podemos ver que el Saldo Vencido y el Valor Pago son las dos variables más significativas con respecto a su impacto dentro de la variable de respuesta, el Porcentaje de Recuperación, muy por encima de las otras cuatro variables. (Wang, et al., 2022).

Reporte Tipo de Flujo:

Para el reporte en el que analizamos el tipo de flujo de una gestión, se realizó un total de 4 algoritmos. En primer lugar, realizamos un Vector Machine (SVM), en segundo lugar, realizamos redes neuronales, en tercer lugar, realizamos un LIGHT GBM, y finalmente, realizamos un Random Forest. Los predictores utilizados para los 4 casos fueron: Día de Pago, monto de pago, el género del deudor, el estado civil, la ciudad, el tipo de flujo, y el saldo vencido. Mientras que la variable de respuesta fue el porcentaje de recuperación.

En primera instancia, se separa los campos a considerar en predictores y la variable de respuesta, mencionados anteriormente. Después, separamos las variables categóricas de las numéricas, y con ello, el modelo estaría listo para cada uno de los algoritmos que se presentan a continuación, con un resumen y sus respectivos resultados.

Support Vector Machine (SVM):

El support vector machine, el cual en español es máquina de vectores de soporte, es una técnica de aprendizaje supervisado utilizada principalmente para clasificación o regresión (Vapnik, 1995). Para dicho algoritmo, se estandarizó las variables numéricas, y se convirtieron las categóricas en representaciones binarias permitiendo así el procesamiento de datos mixtos de forma estructurada para algoritmos de aprendizaje automático.

En primera instancia, se divide los datos en un 80% para el set de entrenamiento y un 20% para el set de prueba. Después se construye el modelo de regresión de vectores de soporte utilizando el pipeline el cual es un flujo de trabajo, que realiza dos movimiento, el primero es un preprocesamiento de datos mencionado anteriormente, y el segundo se emplea el regresor SVM, el cual cuenta con un kernel RBF, que es una función de base radial, que cuenta con un hiperparámetro de regularización de (C=100), un gamma=scale, el cual es el coeficiente del Kernel, finalmente se ajusta utilizando los datos de entrenamiento (Vapnik, 1995). Por lo cual al realizar el modelo obtuvimos los siguientes resultados:

```
R² en conjunto de prueba: 0.6279
RMSE en conjunto de prueba: 0.1509
MAE en conjunto de prueba: 0.0950
MSE en conjunto de prueba: 0.0228
```

Figura 27: Resultados SVM

Redes Neuronales:

Este algoritmo básicamente realiza una serie de pasos para construir y entrenar un modelo de regresión utilizando la librería Keras en Python. Primero, convierte las variables categóricas en representaciones numéricas, permitiendo al modelo trabajar con estas variables. Luego, divide los datos en conjuntos de entrenamiento y prueba para poder construir el modelo secuencialmente, añadiendo capas neuronales ocultas con activación ReLU para aprender patrones en los datos. Además, se aplica regularización con dropout para evitar el sobreajuste. La capa de salida utiliza una función lineal para la regresión (González, 2023).

El modelo se compila utilizando el error cuadrático medio como función de pérdida y el optimizador Adam con una tasa de aprendizaje de 0.001. A continuación, se entrena el modelo con 100 epochs y un tamaño de lote de 32, utilizando los datos de entrenamiento y validando con los datos de prueba. Finalmente, se realizan predicciones en el conjunto de prueba utilizando el modelo entrenado. El cual obtuvimos los siguientes resultados:

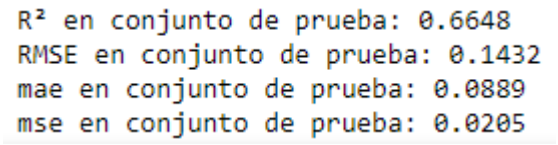


Figura 28: Resultados Algoritmo Redes neuronales

Light GBM:

Este algoritmo realiza varias operaciones para entrenar un modelo de regresión utilizando LightGBM, una biblioteca especializada en algoritmos de aumento de gradientes para tareas de regresión. Primero, divide los datos en dos grupos: uno para entrenamiento y otro para probar el modelo. Luego, configura los parámetros del modelo LightGBM, como el tipo de técnica de aumento (boosting_type), el objetivo del modelo (en este caso, realizar una regresión), y las métricas que se utilizarán para evaluar el rendimiento del modelo, como el error cuadrático medio (rmse) y el error absoluto medio (mae) (Deng, 2021).

Después, crea conjuntos de datos específicos de LightGBM (lgb.Dataset) para entrenamiento y evaluación. Entrena el modelo con 1000 iteraciones (`num_round`) y utiliza estos datos de entrenamiento y evaluación para ajustar los hiperparámetros. Finalmente, utiliza el modelo entrenado para realizar predicciones sobre el conjunto de datos de prueba

(X_test), proporcionando estimaciones numéricas para las respuestas que el modelo prevé basándose en los datos que nunca había visto antes. Los resultados de las métricas que obtuvimos fueron:

```
R² en conjunto de prueba: 0.9383
RMSE en conjunto de prueba: 0.0614
MAE en conjunto de prueba: 0.0214
MSE en conjunto de prueba: 0.0038
```

Figura 29: Resultado Algoritmo LIGHT GBM

Random Forest:

Este algoritmo se encarga de analizar la importancia de las características en un modelo de Random Forest para la tarea de regresión. Primero, divide los datos en conjuntos de entrenamiento y prueba para evaluar el modelo. Luego, se crea un modelo de Random Forest con ciertas configuraciones predefinidas, como el número de árboles (n_estimators) y los criterios de división de los nodos en los árboles. Este modelo se ajusta utilizando los datos de entrenamiento (Cifuentes, 2022).

Después, se extrae la importancia de cada característica del modelo entrenado. Esta importancia se calcula evaluando cuánto contribuye cada característica al rendimiento general del modelo. Posteriormente, se realiza un cálculo para expresar esta importancia como un porcentaje del total, lo que permite comprender qué características tienen un mayor impacto en las predicciones del modelo (Cifuentes, 2022).

Finalmente, organizamos la información en una tabla para mostrar la importancia de cada característica, lo que facilita la visualización y comprensión de qué características son más relevantes para el modelo de Random Forest en términos de su contribución a las

predicciones. En la figura 30 podemos observar las 10 características con mayor efecto en el porcentaje de recuperación y en el Anexo E se pueden observar todos los predictores organizados en orden en base a su importancia en la variable de respuesta.

Feature	Importance (%)
Saldo_Vencido	55,86469531
PAGO	41,25171654
DIA_PAGO	0,849780993
flujo_9	0,425894487
Ciudad_SANTO DOMINGO DE LOS TSACHILAS	0,194380088
Estado_Civil_DIVORCIADO	0,165192951
flujo_3	0,14019976
flujo_1	0,110525222
Genero_Deudor_MASCULINO	0,108161241
Estado_Civil_CASADO	0,104577668

Figura 30: Top 10 Importancia de las características Random Forest

Los resultados del modelo de random forest:

```
Métricas del modelo:
R² en conjunto de entrenamiento: 0.99
R² en conjunto de prueba: 0.94
MAE en conjunto de entrenamiento: 0.01
MAE en conjunto de prueba: 0.01
RMSE en conjunto de entrenamiento: 0.02
RMSE en conjunto de prueba: 0.06
MSE en conjunto de entrenamiento: 0.00
MSE en conjunto de prueba: 0.00
```

Figura 31: Resultados algoritmo Random Forest

ASSESS:

Reporte Ultima Gestión:

Analizando los resultados de los tres diferentes modelos aplicados se pudo destacar cuales modelos presentan mejores resultados acorde a la precisión del modelo correspondiente y que información se repite entre los modelos resaltando datos relevantes y variables que poseen un gran peso con respecto a la variable de respuesta.

Inicialmente, se realizó una regresión lineal de mínimos cuadrados ordinarios, con la cual se obtuvo un R cuadrado del 56.9% y los coeficientes de cada variable al ser transformados a variables dummy. Buscando mejorar la precisión del modelo se realizó una regresión logística en la cual el pseudo R cuadrado subió al 62.6% y nos ofreció información similar a la regresión anterior con respecto a las variables predictoras. Ambos modelos resaltaron el coeficiente más elevado perteneciente al Saldo Vencido en el cuál las dos regresiones nos muestran que mientras mayor sea el saldo vencido, menor es la probabilidad de recuperación de la cartera.

Finalmente se realizó un modelo XGBoost que nos ofreció una precisión del modelo del 95.19% tras una optimización de los hiperparámetros y realizando un *Feature Importance* una vez más se mostró que el variable predictor con mayor impacto en la variable de respuesta es el Saldo Vencido. El uso de los tres modelos nos ayudó a afinar resultados debido a que cada modelo presentó una precisión mayor al anterior y adicionalmente, nos ofrecieron una manera de garantizar la confiabilidad de los resultados que nos ofrecía cada modelo debido a que la información que se interpreta de los tres se repetía.

Reporte tipo de Flujo:

Analizando la figura 32, los resultados de los distintos modelos evaluados proporcionan una visión crítica del rendimiento de los algoritmos para la empresa Siccec

crédito y cobranza. Entre los algoritmos probados, las máquinas de vectores de soporte (SVM) y las redes neuronales proporcionaron un rendimiento aceptable con un R cuadrado de aproximadamente 62% y 66%. Aunque tienen cierto poder predictivo, los errores representados por MAE y RMSE son relativamente altos, lo que indica una precisión predictiva moderada. Sin embargo, son los modelos Light GBM y Random Forest los que muestran un rendimiento sobresaliente.

Los modelos tienen excelentes R-cuadrados de casi el 94%, lo que indica una gran capacidad para predecir con precisión las tasas de recuperación. Además, tanto Light GBM como Random Forest exhiben errores muy bajos (MAE, RMSE y MSE), lo que destaca su precisión en la estimación de recuperaciones importantes para las decisiones financieras y financieras en la gestión de activos, carteras y cobranzas, sobre todo en la toma de decisiones estratégicas. Este alto nivel de precisión tiene implicaciones importantes para las empresas, ya que impacta directamente en las estrategias de gestión de riesgos y la optimización del retorno. Teniendo en cuenta estos resultados, se recomienda centrarse en el desarrollo y optimización de modelos basados en Light GBM o Random Forest, ya que son capaces de proporcionar predicciones precisas en el contexto específico de la empresa Siccec crédito y cobranza.

Modelo	R^2	MAE	RMSE	MSE
Vector Machine (SVM)	0,6279	0,095	0,1509	0,0228
Redes Neuronales	0,6648	0,0889	0,1432	0,0205
Light GBM	0,9383	0,0214	0,0614	0,0038
Random Forest	0,941	0,01	0,061	0,001

FIGURA 32: RESUMEN MÉTRICAS ALGORITMOS

CONCLUSIONES

Con base en lo recopilado, hemos sacado conclusiones parciales que explicaremos a continuación. Primeramente, logramos entender la distribución de los datos que maneja Siccec y como tienen un mercado sumamente grande en la capital equivalente a la suma del resto ciudades del país pero que poco a poco está creciendo en distintas ciudades. Por otro lado, hemos logrado demostrar mostrar el efecto que tiene el tamaño de la deuda (Saldo Vencido) por encima de las otras variables en la recuperación de la cartera y cuáles son las características que más afectan al porcentaje de recuperación para que Siccec priorice la atención a clientes con estas características.

Como conclusiones a Siccec y haciendo referencia al anexo F, el análisis muestra que los saldos vencidos son el factor determinante para la recuperación. Las tendencias muestran que cuanto menor es el saldo predeterminado, mayor es la tasa de recuperación. Por lo tanto, se recomienda centrarse en los deudores con saldos más pequeños para aumentar las tasas de recuperación. Esta estrategia podría contribuir al logro de las metas financieras de Siccec y aumentar la rentabilidad ya que las empresas que usan el servicio de Siccec muchas veces analizan el rendimiento de Siccec en base al porcentaje de recuperación.

Haciendo referencia al anexo G, el segundo predictor más influyente es el monto del pago. Los datos muestran que los pagos más altos conducen a tasas de recuperación más altas, mientras que los pagos más bajos están asociados con tasas de recuperación más bajas. Se recomienda que Siccec se dirija a aquellos clientes con un historial de pagos más largo, especialmente en carteras vencidas ya que mejorará enormemente la velocidad de recuperación y la empresa no gastará tanto dinero en intentar de atender a clientes que históricamente no han pagado sus deudas.

Haciendo referencia al anexo H, el análisis de los días de pago muestra que las tasas de pago más altas se logran en la cuarta semana del mes (del 21 al 28). Se recomienda planificar sus obligaciones de pago y estrategias de cobranza de la semana para aprovechar esta tendencia para aumentar la eficiencia de cobranza y por lo tanto aumentar sus tasas de cobranza.

Las investigaciones muestran que el flujo 9, que combina llamadas e IVR, tiene el mayor impacto en la recuperación, con importantes beneficios. A pesar de los costos relativamente altos, este proceso muestra la mayor eficiencia de recuperación. Se aconsejó a Siccec centrar sus recursos en este flujo ganador y en los 5 mejores que se demostraron con estadística para no dar uso a aquellos flujos que no muestran una buena recuperación y que exista una variación entre los que se sabe ahora son los más eficientes.

El análisis demográfico revela patrones interesantes: Santo Domingo es la ciudad de más rápido crecimiento, seguida de Esmeraldas y luego Quito. En términos de género, los hombres tienen una tasa de recuperación más alta que las mujeres. Además, cabe señalar que las personas divorciadas tienen la tasa más alta de recuperación de estado civil, seguidas de las casadas y solteras. Estos datos pueden no sonar importantes, pero al momento de que Siccec filtra los clientes es de suma importancia entender que características estos deben tener para ser atendidos con la idea de aumentar el porcentaje de recuperación.

Para maximizar la efectividad de las estrategias de cobranza, es importante comprender completamente el impacto de variables clave como los saldos vencidos y los montos de los pagos. Se debe identificar perfiles de clientes con características influyentes que permitirá adaptar campañas estratégicas específicas a estos grupos, acelerando la consecución de sus objetivos. De igual forma los modelos estadísticos deben perfeccionarse con datos más recientes y evaluarse inmediatamente su relevancia para ajustar dinámicamente las estrategias manteniendo una base de datos actualizada para tener

conclusiones en tiempo real y ver si las características más influyentes en la recuperación van cambiando con el tiempo.

Es fundamental pasar de métodos intuitivos a métodos estadísticos sólidos para respaldar la toma de decisiones mediante análisis predictivos, y esto se puede conseguir con la utilización de los modelos presentados. Además, es crucial facilitar reuniones periódicas entre el equipo de estadísticas y los supervisores para evaluar el desempeño de la campaña y brindar oportunidades para la mejora continua sabiendo que clientes y que características específicas buscar al momento de gestionar el pago de una deuda.

REFERENCIAS BIBLIOGRÁFICAS

Agbemava, E., Nyarko, I. K., Adade, T. C., & Bediako, A. K. (2016). Logistic regression analysis of predictors of loan defaults by customers of non-traditional banks in Ghana. European Scientific Journal, 12(1).

Bakar, N. M. A., & Tahir, I. M. (2009). Applying multiple linear regression and neural network to predict bank performance. International Business Research, 2(4), 176-183.

Bonifaz Yambay, J., & Verdezoto Días, R. (2013). Diseño de un modelo de cobranzas de créditos de consumo otorgados por el sistema financiero y viabilización del aplicativo informativo (SAC) para disminuir el índice de morosidad en cuentas por cobrar en cartera de consumo en la ciudad de Guayaquil.

Campos Chahua, O.A., & Gamarra Lujan, S.F. (2023). Procesos de cobranzas en la recuperación de créditos asignados en un estudio jurídico de San Isidro, 2021.

Castro Marín, M. C. (2022). Morosidad de la cartera de crédito y rentabilidad de las cooperativas de ahorro y crédito del Ecuador en tiempos de COVID-19 (Bachelor's thesis).

Cifuentes, N. (2022). Modelo predictivo de la probabilidad de aumento de los días de mora para usuarios de tarjeta de crédito.

Cornejo Espinoza, S.A. (2017). La evasión tributaria y su impacto en la recaudación fiscal en el Perú.

Deng, Y., Li, D., Yang, L., Tang, J., & Zhao, J. (2021, January). Analysis and prediction of bank user churn based on ensemble learning algorithm. In 2021 IEEE International Conference on Power Electronics, Computer Applications (ICPECA) (pp. 288-291). IEEE.

Espino Quiñones, L., & García Torres, M. E. (2018). Aplicación de minería de datos basado en árboles de decisión para predecir el riesgo de morosidad de los clientes en la empresa de seguros Oncosalud SAC, 2018.

Gómez Muñoz, L. (2012). Análisis de créditos estudiantiles en la Universidad de los Estudiantes

González, V. H. B. Redes Neuronales y su Aplicación en Modelos de Cobranza.

Huang, J., Chai, J., & Cho, S. (2020). Deep learning in finance and banking: A literature review and classification. Frontiers of Business Research in China, 14(1), 1-8

Koehler, B., & Fromm, H. (2022). Analysis of a Debt Collection Process Using Bayesian Networks.

Monzón Mérida, M. R. (2022). Diseño de investigación del desarrollo de un sistema de gestión de cobranza para préstamos en mora de una entidad financiera en la ciudad de Guatemala (Doctoral dissertation, Universidad de San Carlos de Guatemala).

Negrete, E. M. (2011). El crédito comercial y la cobranza.

PILLA, P. (2021) Escuela de Administración de Empresas Análisis de la cartera ... - pucesa, Análisis de la cartera vencida en el proceso de concesión y recaudación en las cooperativas de ahorro y crédito del cantón Pelileo: Caso Rhumy Wara Ltda. Available at: https://repositorio.pucesa.edu.ec/bitstream/123456789/3172/1/77333.pdf (Accedido: 03 October 2023).

Romero Carazas, R., Torres Barrera, W., & Vásquez Villanueva, C. A. (2021). Propuesta de gestión de cobranza sistematizada para controlar la morosidad en Instituciones de Educación Básica. Gestión Joven, 22(4).

Sharda, R., Delen, D., Turban, E. (2018). Big data Intelligence, Analytics, and Data Science: A Managerial Perspective. 04. Pearson Education. New Jersey. ISBN: 9780134633282.

Shmueli, G., Bruce, P., Gedeck, P. & Patel, N. (2019), Data Mining for Business Analytics: Concepts, Techniques, and Applications in Python. Wiley

Shoghi, A. (2019). Debt collection industry: machine learning approach. Journal of Money and Economy, 14(4), 453-473.

Stone, Mary, and John Rasp. "Tradeoffs in the choice between logit and OLS for accounting choice studies." Accounting review (1991): 170-187.

Vanrell, J. A. (2011). Un modelo de proceso para proyectos de explotación de información. México: Universidad Tecnológica Nacional (FRBA).

Vapnik, V., & Cortes, C. (1995). Support-vector networks. Machine Learning, 20(3), 273–297.

Vélez, L. D. D., López, C. P. C., & Hoyos, O. M. G. (2017). Diseño de un modelo de scoring para la gestión eficiente de la cartera en una agencia de cobranzas. Escenarios: empresa y territorio, 6(7).

Wang, K., Li, M., Cheng, J., Zhou, X., & Li, G. (2022). Research on personal credit risk evaluation based on XGBoost. Procedia computer science, 199, 1128-1135.

XGBoost Documentation — xgboost 2.0.2 documentation. (n.d.). https://xgboost.readthedocs.io/en/stable/

Yang, C., Lu, T., Li, B., & Lu, X. (2023). When Less Is More? Deep Reinforcement Learning-Based Optimization of Debt Collection. Deep Reinforcement

Learning-Based Optimization of Debt Collection (Junio 23, 2023).

Zevallos Correa, E. D. (2018). Modelo de control interno en la Gestión de Riesgo de morosidad en la empresa de préstamos “Inversiones & Préstamos Cruz de Mayo SAC.

ANEXOS:
ANEXO A: FLUJOS DE GESTIÓN

# FLUJO	GESTIONES				
1	llamada				
2	sms				
3	wpp				
4	mail				
5	ivr				
6	llamada	sms			
7	wpp	llamada			
8	llamada	mail			
9	ivr	llamada			
10	sms	wpp			
11	mail	sms			
12	sms	ivr			
13	mail	wpp			
14	wpp	ivr			
15	ivr	mail			
16	llamada	sms	wpp		
17	mail	sms	llamada		
18	llamada	sms	ivr		
19	mail	wpp	llamada		
20	llamada	wpp	ivr		
21	ivr	mail	llamada		
22	sms	wpp	mail		
23	ivr	wpp	sms		
24	sms	mail	ivr		
25	ivr	mail	wpp		
26	llamada	sms	wpp	mail	
27	ivr	wpp	sms	llamada	
28	llamada	sms	mail	ivr	
29	ivr	mail	wpp	llamada	
30	sms	wpp	mail	ivr	
31	llamada	sms	wpp	mail	ivr

ANEXO B: CONCLUSIONES FLUJO

Flujo	Canales	Costo	Ganancia Promedio
flujo_9	Llamada-IVR	$0,040	$17,07
flujo_3	WPP	$0,023	$10,05
flujo_1	Llamada	$0,020	$11,21
flujo_5	IVR	$0,020	$10.67
flujo_6	Llamada-SMS	$0,040	$13,89

ANEXO C: VARIABLES CONSIDERADAS REPORTE ULTIMAGESTION_CONCATENADO_JUL_AGO_SEP.

Campos considerados	Predictor	Variable de respuesta
VALOR_PAGO	X	
Saldo_Vencido	X	
ACCION	X	
CIUDAD_HOMOLOGADA	X	
DIAS_MORA_PAGOS	X	
PRODUCTO_CR	X	
Porcentaje_Recuperacion		X

ANEXO D: VARIABLES CONSIDERADAS REPORTE QQUERY_TIPO_FLUJO_JUL_AGO_SEP

Campos considerados	Predictor	Variable de respuesta
DIA_PAGO	X	
PAGO	X	
genero_deudor_DESC	X	
estado_civil_DESC	X	
ciudad_cr	X	
flujo	X	
Promedio_Saldo_Vencido	X	
Porcentaje_Recuperacion		X

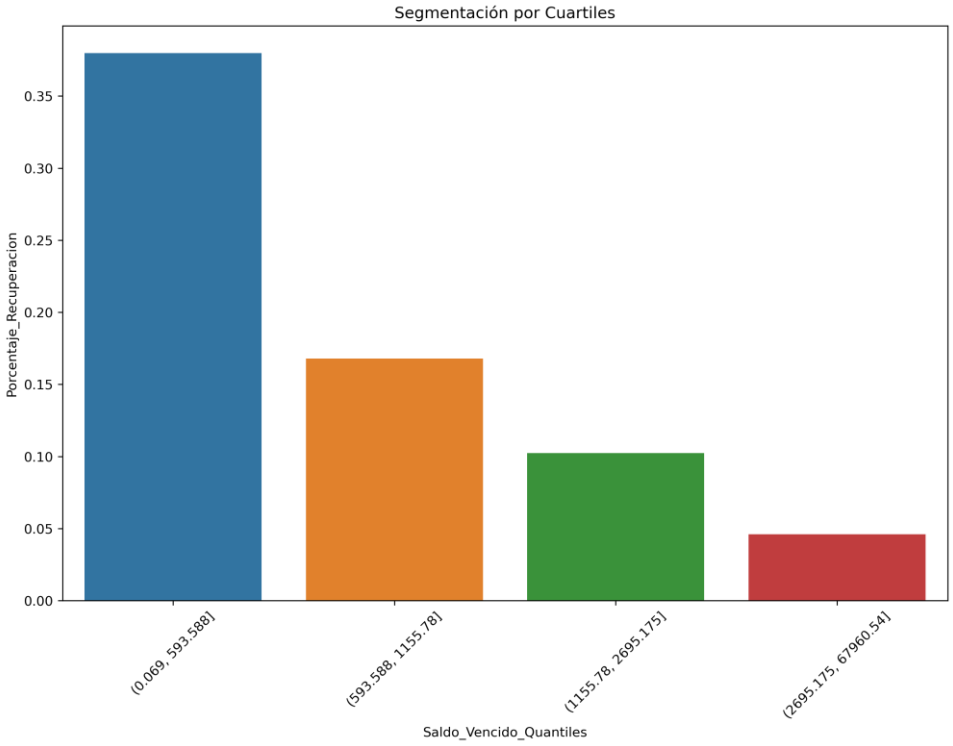
ANEXO E: IMPORTANCIA DE LAS CARACTERÍSTICAS

Feature	Importance (%)
Saldo_Vencido	55,86469531
PAGO	41,25171654
DIA_PAGO	0,849780993

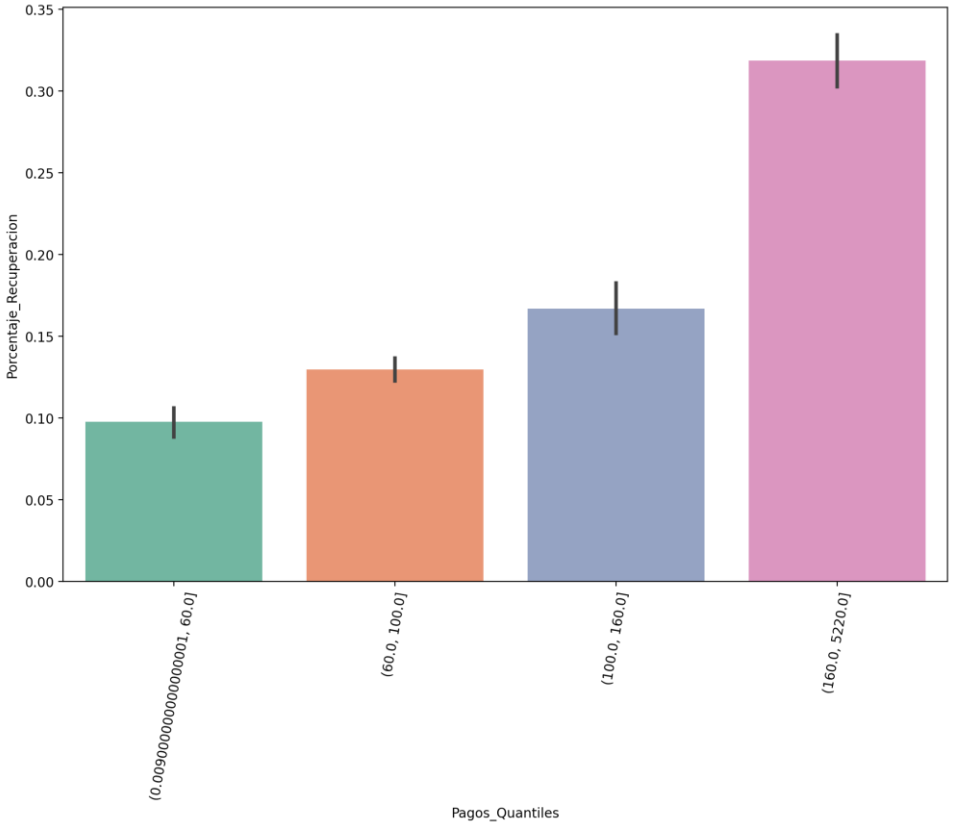
flujo_9	0,425894487
Ciudad_SANTO DOMINGO DE LOS TSACHILAS	0,194380088
Estado_Civil_DIVORCIADO	0,16519S2951
flujo_3	0,14019976
flujo_1	0,110525222
Genero_Deudor_MASCULINO	0,108161241
Estado_Civil_CASADO	0,104577668
Genero_Deudor_FEMENINO	0,102590029
Estado_Civil_SOLTERO	0,089513001
Ciudad_ESMERALDAS	0,071884433
flujo_5	0,070463248
Ciudad_IBARRA	0,061197496
Ciudad_QUITO	0,054658566
flujo_6	0,034769184
Ciudad_CUENCA	0,028956744
Ciudad_CAYAMBE	0,027785031
Ciudad_GUAYAQUIL	0,025705666
flujo_7	0,022405483
Ciudad_RIOBAMBA	0,01635748
Ciudad_TULCAN	0,016181594
Ciudad_MANTA	0,01106955
flujo_2	0,009945628
Ciudad_AMBATO	0,008783052
Estado_Civil_UNION_LIBRE	0,00797309
Estado_Civil_VIUDO	0,007756109
Ciudad_LATACUNGA	0,007550106
flujo_11	0,004424697
Ciudad_DURAN	0,004296033
Ciudad_LOJA	0,004057855
flujo_17	0,003710572
Ciudad_TENA	0,003537538
flujo_21	0,002636972
Ciudad_BABAHOYO	0,002550542
Ciudad_PLAYAS	0,002438025
Ciudad_BALZAR	0,001816342
Ciudad_LA LIBERTAD	0,001130615
Ciudad_QUEVEDO	0,000985176
Ciudad_COTACACHI	0,000854037
flujo_14	0,000733608
Ciudad_ATACAMES	0,000729346
Ciudad_SAMBORONDON	0,000562995

Ciudad_OTAVALO	0,000528749
flujo_16	0,000474634
Ciudad_SANTA CRUZ	0,000457766
Ciudad_PORTOVIEJO	0,000432678
Ciudad_LAGO AGRIO	0,000409448
Ciudad_MACHALA	0,000372609
flujo_20	0,000368117
Ciudad_LA TRONCAL	0,000333257
Ciudad_MILAGRO	0,000239958
Ciudad_SALINAS	0,000210889
flujo_10	0,000188162
Ciudad_EL CARMEN	0,000175258
Ciudad_PASAJE	0,000163865
Ciudad_CHONE	0,000147305
flujo_28	0,000108545
Ciudad_ORELLANA	9,1265E-05
flujo_8	7,27846E-05
Ciudad_ZARUMA	6,8529E-05
Ciudad_HUAQUILLAS	6,39549E-05
Ciudad_ZAMORA	5,71666E-05
flujo_4	5,60787E-05
flujo_22	4,16916E-05
Ciudad_QUININDE	3,86999E-05
Ciudad_GUARANDA	3,00535E-05
Ciudad_DAULE	2,21539E-05
flujo_12	1,17339E-05
Ciudad_ARENILLAS	5,50523E-06
flujo_18	4,68752E-06
flujo_26	3,46691E-06
flujo_30	1,71857E-06
Ciudad_SALCEDO	1,22811E-06
Ciudad_GUALACEO	8,72505E-07
Ciudad_SANGOLQUI	5,31939E-07
flujo_24	2,82479E-07

ANEXO F: CONCLUSION SALDO VENCIDO



ANEXO G: CONCLUSIONES PAGOS



ANEXO H: CONCLUSIONES DIA DE PAGO