TRABAJO FINAL DE MÁSTER

UNIVERSIDAD DE BARCELONA

INSTITUTO DE FORMACIÓN CONTINUA

MÁSTER EN BIG DATA & DATA SCIENCE

"Desarrollo de un modelo predictivo para evaluar el riesgo de impago de un crédito en una entidad financiera"

AUTORES:

- CASTILLO COTRINA, Yossky
- FERRUFINO, Miguel
- TARBET, Hakim

TUTOR:

- ARROYO VENDRELL, Ferran

BARCELONA, SEPTIEMBRE DE 2021

Tabla de contenidos:

Introducción

El presente trabajo titulado " Desarrollo de un modelo predictivo para evaluar el riesgo de impago de un crédito en una entidad financiera" responde a la necesidad de proponer una alternativa viable para agilizar el proceso de concesión de los préstamos o créditos en las entidades financieras. En este caso concreto, se propone el uso de algoritmos Machine Learning para la predicción de la probabilidad de impago que puede tener un cliente en base a diferentes características o variables explicativas.

Dado que el proceso actual es bastante manual donde interviene un equipo humano que evalúa el nivel de dicho riesgo, con los consecuentes costes para la empresa, se propone que se usen datos históricos para entrenar un modelo predictivo y crear una solución que predice dicho riesgo en cuestión de segundos reduciendo así los tiempos, la intervención de equipos de gestión de riesgo y finalmente los presupuestos que una entidad financiera ha de asumir para llevar a cabo dichas operaciones.

En las siguientes páginas abordaremos todos los aspectos relacionados con este trabajo que lo hemos dividido en tres bloques principales, donde cada uno de ellos se compone de diferentes secciones:

Bloque I: Definiciones iniciales

En este primer bloque necesitamos definir de forma clara diferentes aspectos del trabajo, tales como:

- Sección 1: Objetivo principal del proyecto
- Sección 2: Antecedentes
- Sección 3: Punto de partida, contexto actual
- Sección 4: Stakeholders
- Sección 5: Análisis DAFO del proyecto

Bloque II: Estudio de la viabilidad económica y roles

El principal objetivo de este bloque es estudiar la viabilidad económica del proyecto, para ello necesitamos tener claro los diferentes roles del equipo de trabajo, así como el cronograma y timing del proyecto.

- Sección 6: Roles del equipo de trabajo

- Sección 7: Explicación de la gestión del equipo de trabajo
- Sección 8: Timing del proyecto
- Sección 9: Análisis económico y Payback

Bloque III: Desarrollo y conclusiones

En este último bloque es cuando desarrollamos diferentes modelos predictivos que nos permitan hacer comparativas (benchmarking) entre ellos y obtener conclusiones de las variables más importantes para este proyecto. Las secciones que se tratarán en este bloque son:

- Sección 10: Selección del dataset
- Sección 11: Fuentes de información y Benchmarking
- Sección 12: Desarrollo
- Sección 13: Contraste de hipótesis y elección de la métrica de ajuste
- Sección 14: Output del proyecto
- Sección 15: Conclusiones

1. Objetivo principal del proyecto

El objetivo principal del proyecto es el de evaluar el nivel de riesgo de impago de un crédito bancario por parte de un cliente de una entidad financiera mediante un modelo predictivo basado en Machine Learning.

Se pretende desarrollar un modelo predictivo que evalúe los datos de una solicitud de préstamo presentada por un cliente y que devuelva un valor decimal entre 0 y 1 donde 0 es bastante improbable que el cliente impague y 1 es bastante probable que el cliente impague.

Para llevar a cabo dicha valoración, el modelo tendrá que hacer un análisis de propensión a impagar basado en el histórico de todos los préstamos tanto concedidos como rechazados por la entidad financiera además de otros datos obtenidos de fuentes externas.

2. Antecedentes

Al tratarse de un sector con un nivel de confidencialidad bastante alto, es difícil saber qué entidades financieras tiene ya aplicadas técnicas de Machine Learning para la predicción del nivel de riesgo crediticio, la hipótesis que manejamos es que al menos entidades financieras top mundiales tales como Citibank, JPMorgan Chase, HSBC Holdings, BNP Pariba, Bank of America, BBVA, Santander, Crédit Agricole, Deutsche Bank, entre otros... Probablemente ya lo están usando, sobre todo porque han confirmado que usan Big Data, Machine Learning e Inteligencia Artificial para otras áreas tales como la evaluación de precios de acciones, eficiencia en los procesos internos, detección de fraude, entre otros. (Usachev, n.d.)

Se ha notado que hay varios estudios, trabajos de fin de grado y propuestas relacionadas con el análisis de riesgo crediticio usando Machine Learning, que han sido elaboradas y presentadas por alumnos y profesionales del sector. (Felman, 2018)

También se ha visto que algunas entidades financieras incluyendo, Home Credit, han creado competiciones en Kaggle para la elaboración de modelos predictivos relacionadas con el análisis de riesgo usando datos. (Group, 2018)

Aunque no sabemos a ciencia cierta qué bancos o entidades financieras han desarrollado ya sus propios modelos, creemos que estos modelos aún no han sido implementados al 100% en los procesos de evaluación de préstamos en la mayoría de ellas.

3. Punto de Partida: Contexto Actual

En base a los objetivos del proyecto y los antecedentes antes mencionados se plantea la siguiente situación actual.

En nuestro contexto actual, la población de niveles socioeconómicos altos tiene mayor penetración a créditos. Sin embargo, personas pertenecientes a sectores socioeconómicos medios y bajos buscan mejorar su calidad de vida e incrementar su poder adquisitivo accediendo a créditos bancarios para poder invertir en un vehículo, vivienda propia, negocio familiar, microempresa u otra oportunidad de inversión. A esto se suma la pandemia que estamos viviendo, y cómo las economías de muchos países se han visto negativamente impactadas. En consecuencia, todos los gobiernos y entidades financiaras están preparando programas de financiamiento a través de créditos de diversos tipos a los cuales puedan acceder todo tipo de personas y empresas.

Actualmente muchas personas buscan obtener un préstamo financiero y no lo consiguen debido a diferentes factores: Carecer de antecedentes crediticios o ser insuficientes, no tener un trabajo formal, ingresos mensuales limitados, tiempo que llevan laborando, etc. Esto genera que accedan a fuentes de financiamiento informal pagando tasas de interés exorbitantes y cuotas impagables, generando una experiencia crediticia negativa en las personas. Como propuesta de solución, se plantea desarrollar un Modelo Predictivo que evalúe diferentes parámetros, muchos de ellos que los modelos tradicionales de Scoring crediticio no evalúan, con el objetivo de promover la inclusión financiera. En el caso tradicional, la calificación de riesgo de cumplimiento crediticio que actualmente manejan las instituciones financieras carece del análisis de diferentes parámetros que terminan en la negativa de otorgar un crédito a una persona o micro empresa, esto debido a que normalmente solo analizan parámetros tradicionales y utilizan un sistema manual o semiautomático que se aplica desde hace muchos años atrás y que no incluye parámetros, que probablemente, estén afectando esa calificación crediticia. Incluso, es posible decir que los parámetros que afectan esta calificación están en constante cambio debido a que nuestros cambios socioeconómicos, políticos, ambientales, entre otros, afectan de forma directa e indirecta al riesgo que conlleva hacerse de un crédito. El mejor ejemplo en este caso es ver cómo ha afectado y sigue afectando la pandemia COVID-19 a la situación actual de toda la población y economía mundial.

Otro de los problemas que enfrenta la banca tradicional es el tiempo que se tarda en revisar cada parámetro de forma manual, esto puede tomar días, incluso semanas, un tiempo muy valioso para ambas partes, el cliente y la entidad; además del gasto en personal calificado que se debe cubrir. (Ucha, 2021)

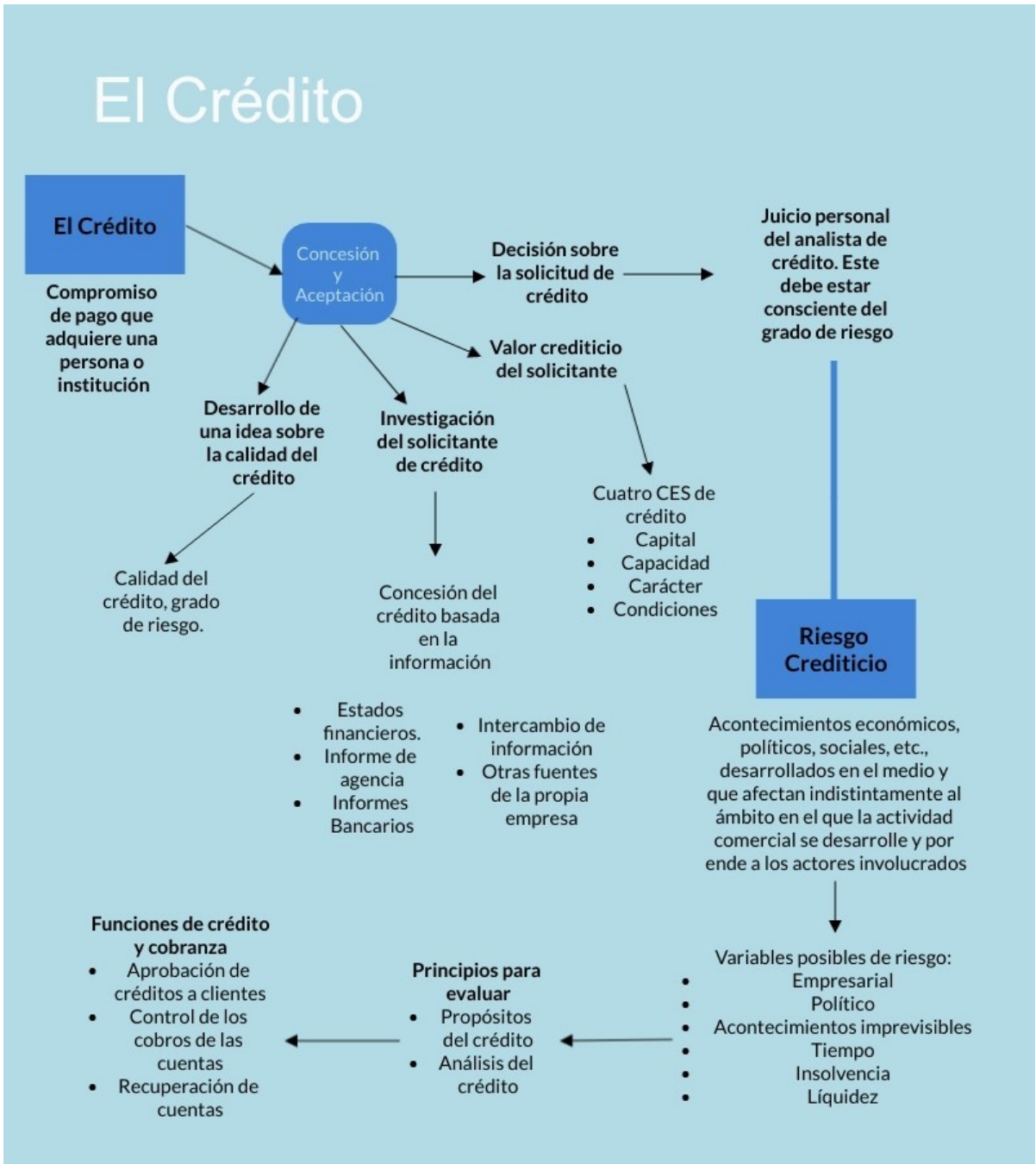


Ilustración 1 - Crédito y Riesgo Crediticio Tradicional. Fuente: elaboración propia.

Debido a estos puntos planteados, se ve óptimo aplicar un modelo predictivo que constantemente este aprendiendo gracias a la AI y Machine Learning, y que sea capaz de ir identificando estos factores determinantes en la calificación de un riesgo crediticio que continuamente se encuentran cambiando según la situación actual.

Con el desarrollo del Modelo Predictivo basado en Machine Learning, una vez insertados los parámetros de la solicitud crediticia de un cliente, el modelo generará un resultado en cuestión

de segundos. Incluso la entidad financiera puede integrarlo en sus aplicaciones tanto web como apps, para que los clientes hagan una simulación en cualquier momento, sin la necesidad de tener que desplazarse a las oficinas del banco, ni hacer largas colas, ni preparar toda la documentación que se requiere para iniciar el proceso de evaluación crediticia.

4. Stakeholders

4.1. Stakeholders directos:

- Propietarios y desarrolladores del modelo predictivo y plataforma web
- Desarrolladores de mejoras o configuraciones personalizadas del modelo predictivo
- Clientes que quieran hacer uso del modelo. En este caso pueden ser cualquier tipo de agentes de crédito u otras que requieran este tipo de predicciones.
- Proveedores de información también conocidos como buros de información crediticia.

4.2. Stakeholders indirectos:

- Clientes de las entidades financieras
- Controladores y supervisores de sistemas financieros
- Gobierno y organizaciones interesadas en las áreas de desarrollo humano, financiero y socioeconómico

5. DAFO de la Situación

	OPORTUNIDADES:	AMENAZAS:
	• **Cambios en el estilo de vida de las personas que demandan una transformación digital en todos los sectores, sin distinción.** • **Interés de entidades bancarias en invertir en proyectos de Transformación Digital.** • **Disponibilidad de datos en la web, redes sociales, institutos nacionales de estadística, etc.** • **Nuevas tecnologías que están generando una revolución digital, donde la competencia aún es limitada.**	• **Suplantación de identidad de un cliente durante el proceso de evaluación crediticia.** • **Desinterés por parte de las entidades bancarias para implementar el modelo predictivo.** • **Carencia de infraestructura de punta para la implementación del proyecto.**
FORTALEZAS:	ESTRATEGIAS (FO):	ESTRATEGIAS (FA):
• **Uso de herramientas que sustentan las bases de la Transformación Digital.** • **Integración de nuevos parámetros que la banca tradicional no evalúa.**	• Contactar a gerentes del área de Inteligencia de Negocios para promocionar la aplicación. • Subir la aplicación a Internet para medir el nivel de satisfacción de los usuarios, para maximizar los	• Implementar autenticación de dos factores para validar la identidad de un usuario. • Hacer un cuadro resumen del indicador Beneficio – Costo para mostrar la rentabilidad del proyecto.

• **Reducción de tiempo y costes en el proceso de evaluación crediticia.**	beneficios del proyecto.	
DEBILIDADES:	ESTRATEGIAS (DO):	ESTRATEGIAS (DA):
• **Carencia de un Dataset con datos actualizados en tiempo real.** • **Falta de expertise en los procesos de créditos bancarios.** • **Insuficiente presupuesto para construir la infraestructura de punta para la implementación del proyecto.**	• Consumir servicios web de institutos nacionales de estadística, implementar web scraping para extraer datos de la web. • Contactar a profesionales expertos en el sistema financiero.	• Buscar un socio que tenga infraestructura de punta para la implementación del proyecto. • Contactar a una institución financiera para que tenga un rol de CONSULTOR durante el desarrollo del proyecto, con el beneficio de obtener un descuento futuro en la implementación de la aplicación.

6. Roles del Equipo de Trabajo

- Modelado, extracción, transformación, carga de los datos, desarrollo web para outputs del modelo.

  Tarbet, Hakim

- Modelado de GLM, enriquecimiento de variables, validación y comparación de modelos

  Ferrufino, Miguel

- Exploración de datos, modelado de Random Forest, aplicación de métricas de ajuste y validacion de modelos.

  Castillo Cotrina, Yossky

7. Explicación de la Gestión del Equipo de Trabajo

7.1. Lista de actividades del equipo:

Son las acciones a realizar por un rol dentro del proceso que cumple con un objetivo:

- Obtención y exploración del dataset
- Tratamiento, preparación y clasificación de los datos
- Creación del diccionario de datos
- Limpieza y transformación de datos
- Minería de datos
- Selección de herramientas
- Selección de método de predicción
- Creación de algoritmo de machine learning
- Creación de dataset entrenamiento
- Creación de dataset de prueba
- Pruebas de eficacia del modelo
- Creación del modelo de predicción
- Entrenamiento del modelo
- Validaciones del modelo
- Desarrollo de output del sistema
- Desarrollo del entorno del sistema
- Integrar el sistema en el entorno de producción
- Monitoreo y control

7.2. Objetivos:

- Obtener un dataset adecuado para el correcto entrenamiento del modelo predictivo
- Creación de un modelo que se ajuste y pueda predecir los positivos y negativos de impago de un crédito
- Aplicar modelos supervisados al desarrollo del modelo predictivo crediticio
- Desarrollar un ambiente productivo para el output del modelo al usuario final

7.3. Diagrama de Roles:

A continuación, se presenta el diagrama de roles según el trabajo realizado a lo largo del proyecto el cual fue realizado de manera muy transversal donde cada uno de los integrantes cumplió varios roles según la etapa y tarea a realizarse.

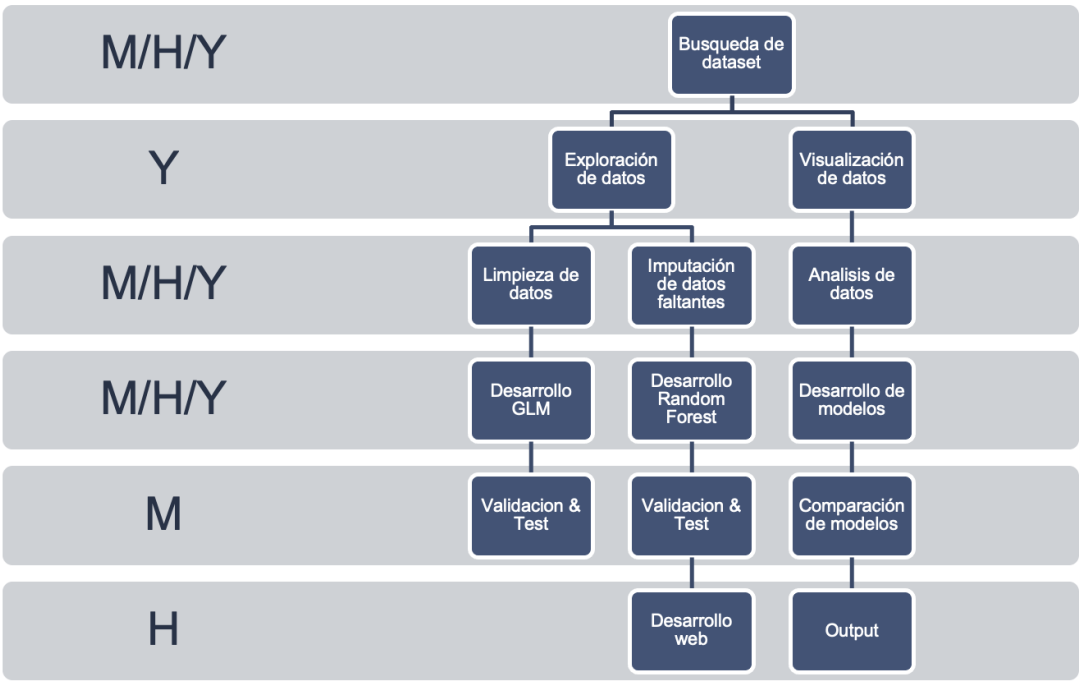


Ilustración 2 - Diagrama de Roles – Fuente: elaboración propia

Donde:

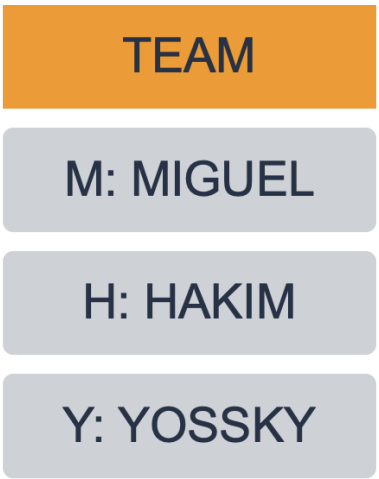


Ilustración 3 - Miembros del equipo de trabajo – Fuente: elaboración propia

8. Timing del Proyecto

8.1. Explicación de las Fases del Desarrollo.

Un proyecto de desarrollo de un modelo predictivo necesita y debe pasar por algunas fases previas al modelado de los datos y la elección del modelo ideal para el problema que estamos tratando.

En nuestro caso en el que queremos desarrollar un modelo predictivo que evalúe el riesgo de impago de un préstamo crediticio, las fases de investigación y desarrollo en las que nuestro proyecto ha de pasar para una correcta resolución del problema son:

8.1.1. Comprensión del problema y definición de la Hipótesis nula

Antes de empezar, incluso antes de explorar la base de datos, primero debemos comprender bien el problema a la que nos enfrentamos, es decir debemos conocer bien la problemática a la que se enfrenta el sector, en nuestro caso bancario.

La correcta comprensión de la problemática nos será de gran ayuda especialmente en las siguientes dos fases en la que deberemos hacer la elección de las variables relevantes para el modelo a desarrollar, así como la evaluación de las dependencias entre dichas variables.

También debemos definir una hipótesis nula sobre la cual modelaremos los datos del que disponemos, esta hipótesis nula será la base del desarrollo de nuestro modelo y es la que rechazaremos o no en base a los resultados de dicho modelo.

Es por ello que antes de empezar siquiera con la carga de los datos, primero se recomienda ahondar en la problemática a resolver y comprender bien las necesidades del sector.

8.1.2. Exploración y preparación de los datos

Una exploración inicial de los registros y variables del que disponemos en nuestra base de datos o dataset es algo crucial, cuanto más comprendemos los datos del que contamos más sencillo nos resultará resolver cualquier inconveniente que nos surja durante el desarrollo del modelo.

Una exploración de los datos nos aportará multitud de ventajas, podemos detectar los valores faltantes o nulos, los outliers, la inconsistencia en los datos, etc.… Es por ello que antes de pensar en qué tipo de modelo se ajusta mejor a nuestra problemática, primero debemos conocer bien los datos del que contamos, necesitamos que los datos sean nuestros amigos y para ello debemos tener en cuenta todas las variables del que cuenta nuestro dataset,

conocer sus valores mínimos, máximos, medias y evaluar si se ajustan a la lógica que por experiencia o por lógica deben tener.

En esta fase, es también donde aplicaremos los diferentes métodos estadísticos para la limpieza y saneamiento de los datos, todo con el objetivo de mejorar su calidad y adaptarlos para afrontar de la mejor manera el desarrollo de nuestro modelo.

8.1.3. Análisis de las variables.

En la fase de análisis de las variables, pretendemos entender de forma óptima cada una de las variables del que se compone nuestro dataset, encontrar correlaciones entre variables, conocer la importancia que tiene cada una de las variables para la predicción final de nuestra variable explicada, encontrar el método estadístico ideal para una posible reducción de variables de nuestro dataset entre otros.

La fase de análisis de variables es el que mejor punto de partida para cualquier desarrollo de un modelo predictivo, nos permite ahorrar tiempo y bastantes recursos en nuestro desarrollo final ya que con ello podemos detectar de forma prematura correlaciones entre variables o encontrar variables explicativas innecesarias que pueden sesgar nuestra predicción.

Es por ello que una vez realizada una exploración inicial y preparados los datos de nuestro dataset, es imprescindible realizar este análisis antes de ponernos con el desarrollo del propio modelo.

8.1.4. Modelado de datos y Benchmarking

Comprendemos bien el problema, y ya conocemos los datos y además de eso hemos analizado cada una de las variables y tenemos un dataset final con el que queremos modelar los datos, entonces estamos preparados para la fase 4 donde entramos en el desarrollo de nuestros modelos predictivos.

Dado que no sabemos qué modelo funciona mejor para nuestros datos, la idea a la hora de modelar es crear varios modelos predictivos y hacer una comparativa entre ellos viendo el que mejor ajuste ofrece para el perfil de los datos con el que contamos. Este benchmarking entre los modelos es la mejor garantía de conseguir un buen modelo predictivo.

Esta fase la podemos diferenciar en 3 tareas:

- Tarea 1: Seleccionar las variables más importantes para nuestro modelo, podemos apoyarnos en las técnicas de Stepwise para la selección de las variables o podemos

desarrollar nuestro propio Stepwise en caso de que el coste computacional del primero sea excesivo.

- Tarea 2: Desarrollo de diferentes modelos predictivos en base a las variables seleccionadas y las variables que por lógica o experiencia del analista consideremos que son importantes para el modelo final.
- Tarea 3: Benchmarking o comparativa de los modelos y la selección del modelo final. Sabemos que la intuición y la lógica muchas veces falla cuando tratamos con datos, por eso necesitamos que haya una comparativa de diferentes modelos para ver cuál se ajusta mejor y que represente la realidad de la mejor manera. Esta fase es a lo que llamamos Benchmarking.

8.1.5. Mantenimiento y mejora continua del modelo

Como cualquier desarrollo de software ya sea web, app o modelos de machine learning, necesita un mantenimiento periódico para la corrección de errores o revisión de registros de salida. Esta fase es algo que no se puede contabilizar en términos de tiempo ya que perdura mientras esté funcionando el modelo en producción.

Además de un mantenimiento constante, cualquier desarrollo quedaría anticuado en poco tiempo si no es mejorado de forma continua, la mejora continua es una de las mejores formas de conseguir un producto viable para la empresa.

8.2. Cronograma de Actividades a Realizar

Tal y como se comentó en la sección de **"Explicación de la Gestión del Equipo de Trabajo"** las actividades a realizar por parte del equipo de trabajo se clasifican de la siguiente manera según la fase en la que estamos:

Fase	***Título de la fase***	***Actividad a realizar***
Fase 2	Exploración y preparación de los datos	Obtención y exploración del dataset
		Tratamiento, preparación y clasificación de los datos
		Creación del diccionario de datos

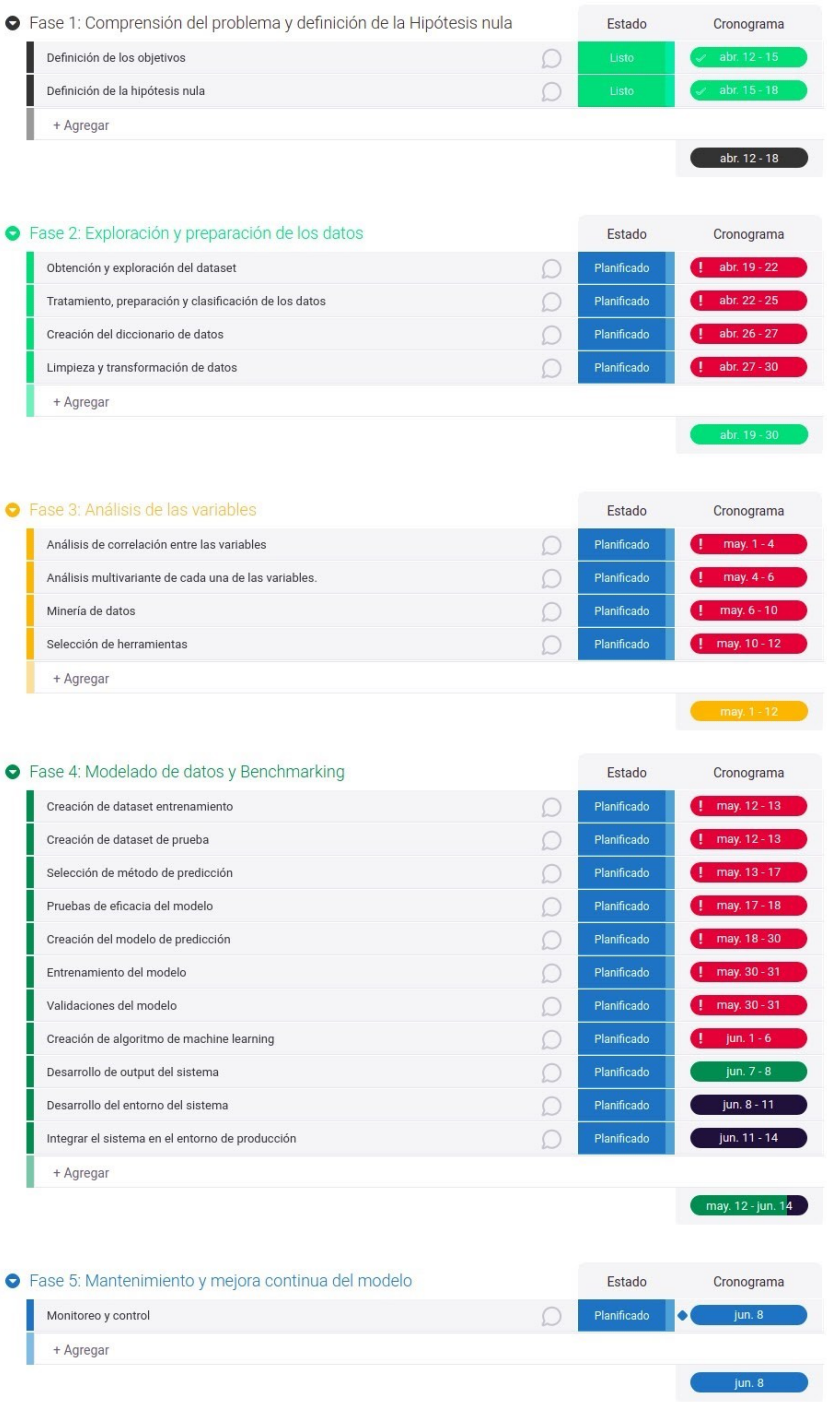
		Limpieza y transformación de datos
Fase 3	Análisis de las variables	Análisis de correlación entre las variables
		Análisis multivariante de cada una de las variables.
		Minería de datos
		Selección de herramientas
Fase 4	Modelado de datos y Benchmarking	Creación de dataset entrenamiento
		Creación de dataset de prueba
		Selección de método de predicción
		Pruebas de eficacia del modelo
		Creación del modelo de predicción
		Entrenamiento del modelo
		Validaciones del modelo
		Creación de algoritmo de machine learning
		Desarrollo de output del sistema
		Desarrollo del entorno del sistema
		Integrar el sistema en el entorno de producción
Fase 5	Mantenimiento y mejora continua del modelo	Monitoreo y control

Distribución de las tareas a realizar por parte del equipo en las diferentes fases.

En el siguiente cronograma creado con la herramienta de planificación de proyectos y gestión de tareas monday.com (https://www.significados.com, 2021), se puede ver una estimación del tiempo para completar ese trabajo en un plazo de aproximadamente 2 meses, con un equipo de 3 analistas.

Se ha intentado hacer una estimación del tiempo inicial que corresponda con la dificultad de la tarea, aunque esta estimación es algo conservadora, creemos que es un tiempo razonable para el desarrollo de un modelo predictivo al menos para hasta la fase 4 donde debemos encontrar el modelo con el mejor ajuste.

Nota: Clic aquí para ver el cronograma en tamaño completo.



Cronograma de las fases del proyecto – Fuente: elaboración propia.

9. Análisis Económico y Payback

También denominado Estudio de Viabilidad Económica, es uno de los aspectos más importantes durante el desarrollo de un proyecto porque permite saber si se debe continuar o no con el desarrollo de este.

El estudio de viabilidad de este proyecto comprende:

9.1. Determinación del Costo de Inversión

Se calcula en base a la siguiente fórmula:

$$\boldsymbol{CI} = CH + CS + CM$$

Donde:

- *CI: Costo de Inversión*
- *CH: Costo de Hardware*
- *CS: Costo de Software*
- *CM: Costo de Mobiliario*

9.1.1. Hardware:

No es necesario la compra de equipos (Laptops, estabilizadores, etc.) porque cada miembro del equipo trabajó con su propia laptop/PC personal.

9.1.2. Software:

Se trabajó con los lenguajes de programación R y Python, que poseen una licencia de código abierto.

Tabla N° 1: Determinación del Costo de Software

Descripción	**Importe (€)**
Licencia de lenguaje de programación R y Python	0.00
Total	**0.00**

Fuente: (Elaboración Propia, 2021)

9.1.3. Mobiliario:

No es necesario la compra de mobiliario (Escritorios, sillas, etc.) porque cada miembro del equipo trabajará desde casa haciendo uso de propio mobiliario.

El resumen de los costos de inversión se detalla a continuación:

Tabla N° 2: Resumen del Costo de Inversión

Costo de Inversión	**Importe (€)**
Hardware	0.00
Software	0.00
Mobiliario	0.00
Total	**0.00**

Fuente: (Elaboración Propia, 2021)

9.2. Determinación del Costo de Desarrollo

Se calcula en base a la siguiente fórmula:

$$\boldsymbol{CD} = CRH + CRM + CS$$

Donde:

- *CD: Costo de Desarrollo*
- *CRH: Costo de Recursos Humanos*
- *CRM: Costo de Recursos Materiales*
- *CS: Costo de Servicios*

9.2.1. Recursos Humanos:

Según el cronograma, el tiempo de desarrollo del proyecto inicia el 12/04/2021 y termina el 08/06/2021, lo que representa 58 días de trabajo o 464 horas hombre, considerando tres recursos asignados.

Tabla N° 3: Determinación del Costo de Recursos Humanos

Perfil	Horas-Hombre	Costo por Hora (€)	Importe (€)
Data Engineer	464	15.63	7,252.32
Data Scientist	464	15.63	7,252.32
Data Analyst	464	15.63	7,252.32
Total			**21,756.96**

Fuente: (Elaboración Propia, 2021)

(*) Cálculo del Costo por Hora:

Costo por Hora = $45000 * \frac{€}{1\ Año} * \frac{1\ Año}{12\ Meses} * \frac{1\ Mes}{240\ Horas}$

$$45000 * \frac{€}{1\ Año} * \frac{1\ Año}{12\ Meses} * \frac{1\ Mes}{240\ Horas}$$

Costo por Hora = $15.63\ \frac{€}{Hora}$

9.2.2. Recursos Materiales:

No es necesario la compra de materiales (Papel bond, lapiceros, lápices, etc.) porque cada miembro del equipo ya dispone de estos recursos que son propios.

9.2.3. Servicios:

Según el cronograma, el tiempo de desarrollo del proyecto inicia el 12/04/2021 y termina el 08/06/2021, lo que representa 58 días de trabajo o 464 horas hombre, considerando tres recursos asignados.

Tabla N° 4.1: Determinación del Costo de Energía Eléctrica

Descripción	Consumo (kWh/Hora)	N° Horas	Costo (€/kWh)	N° Recursos	Importe (€)

Energía Eléctrica	0.25	464	0.18	3	62.64
Total					**62.64**

Fuente: (Elaboración Propia, 2021)

Tabla N° 4.2: Determinación del Costo de Internet

Descripción	**Ancho de Banda (**Mb**)**	**Costo Mensual (€)**	**N° Meses**	**N° Recursos**	**Importe (€)**
Internet	600	50	2	3	300.00
Total					**300.00**

Fuente: (Elaboración Propia, 2021)

Tabla N° 4.3: Determinación del Costo de Servicios

Descripción	**Importe (€)**
Energía Eléctrica	62.64
Internet	300.00
Total	**362.64**

Fuente: (Elaboración Propia, 2021)

El resumen de los costos de desarrollo se detalla a continuación:

Tabla N° 5: Resumen del Costo de Desarrollo

Costo de Desarrollo	**Importe (€)**
Recursos Humanos	21,756.96

Recursos Materiales	0.00
Servicios	362.64
Total	**22,119.60**

Fuente: (Elaboración Propia, 2021)

9.3. Beneficios

Los beneficios son las ventajas traducidas principalmente en ahorro de tiempo y dinero, que se obtiene con la implementación del Modelo Predictivo.

9.3.1. Beneficios Tangibles:

Son aquellos resultados que se pueden cuantificar en forma inmediata después de la implementación del Modelo Predictivo, esto se traduce en ahorro de tiempo y dinero en la evaluación crediticia de un cliente.

Tabla N° 6: Determinación del Costo de Beneficios Tangibles

Descripción	**Sueldo Anual de Personal**	**% de Tiempo para Evaluación de Créditos**	**N° de Ejecutivos de Créditos**	**Importe Anual (€)**
Personal de créditos	30,000.00	60%	3	54,000.00
Total				**54,000.00**

Fuente: (Elaboración Propia, 2021)

9.3.2. Beneficios Intangibles:

- Optimización del proceso de evaluación crediticia.
- Cuantificación del riesgo de impago para reducir el número de clientes morosos.

- Mejora en el nivel de servicio al cliente.
- Mejora de la imagen institucional.
- Aprovechamiento de la tecnología de punta.

9.4. Costos de Operación

Se calcula en base a la siguiente fórmula:

$$\boldsymbol{CO} = CRH + CIN + CEE + CME$$

Donde:

- *CO: Costo de Operación*
- *CRH: Costo de Recursos Humanos*
- *CIN: Costo de Infraestructura en la Nube*
- *CEE: Costo de Energía Eléctrica*
- *CME: Costo de Mantenimiento de Equipos*

9.4.1. Recursos Humanos:

Según nuestras estimaciones, calculamos que con la implementación del proyecto se generaría un ahorro del 75% del tiempo de evaluación de créditos. En el análisis de beneficios tangibles realizado anteriormente, calculamos que el costo anual de personal de créditos asciende a 54,000.00 euros.

Tabla N° 7: Determinación del Costo de Recursos Humanos

Descripción	**Costo Anual de Personal Pre-Implementación**	**% de Tiempo para Evaluación de Créditos Post-Implementación**	**Importe Anual (€)**
Personal de créditos	54,000.00	25%	13,500.00
Total			**13,500.00**

Fuente: (Elaboración Propia, 2021)

9.4.2. Infraestructura en la Nube:

Tabla N° 8: Determinación del Costo de Infraestructura en la Nube

Descripción	**Costo Anual ($)**	**Tipo de Cambio**	**Importe Anual (€)**
Infraestructura en la Nube: - Servidor Virtual Amazon EC2 - Procesador Virtual: 32 vCPU - Memoria RAM: 128 GB - Disco Duro: 100 GB - Sistema Operativo: Centos 8 o Debian 10	7,318.10	0.82604	6,045.04
Total			**6,045.04**

Fuente: (Elaboración Propia, 2021)

9.4.3. Energía Eléctrica:

El costo de la energía eléctrica no se considera porque en el proceso tradicional de evaluación crediticia también hay un consumo de energía eléctrica, el cual se mantiene en esta nueva implementación.

9.4.4. Mantenimiento de Equipos:

El costo de mantenimiento de equipos no se considera porque en el proceso tradicional de evaluación crediticia también hay un cronograma de mantenimiento, el cual se mantiene en esta nueva implementación.

El resumen de los costos de operación se detalla a continuación:

Tabla N° 9: Resumen del Costo de Operación

Costo de Operación	**Importe (€)**

Recursos Humanos	13,500.00
Infraestructura en la Nube	6,045.04
Energía Eléctrica	0.00
Mantenimiento de Equipos	0.00
Total	**19,545.04**

Fuente: (Elaboración Propia, 2021)

Flujo de Caja

Tabla N° 10: Flujo de Caja

Definición	**Año 0 (€)**	**Año 1 (€)**
Costo de Inversión	-0.00	
Costo de Desarrollo	-22,119.60	
Beneficios		+54,000.00
Costo de Operación		-19,545.04
TOTAL	-22,119.60	+34,454.96

Fuente: (Elaboración Propia, 2021)

Análisis de Rentabilidad

Para el análisis de rentabilidad se evaluará cuatro indicadores económicos para determinar si el proyecto es viable o no. La tasa mínima de rentabilidad exigida a la inversión es del 20% anualmente.

9.5. Valor Actual Neto (VAN):

También llamado Valor Presente Neto (VPN), compara los ingresos y egresos del proyecto en un solo momento del tiempo que por lo general es el periodo cero, es decir consiste en actualizar los costos y beneficios de un proyecto para conocer cuánto se va a ganar o perder con esa inversión. El VAN va a expresar una medida de rentabilidad del proyecto en términos absolutos netos, es decir, en n° de unidades monetarias (euros, dólares, pesos, etc.).

Fórmula N° 1: Valor Actual Neto

$$VAN = -I_0 + \sum_{t=1}^{n} \frac{F_t}{(1+k)^t}$$

Donde:

- F_t: Son los flujos de dinero en cada periodo
- I_0: Es la inversión inicial (t = 0)
- n: N° de periodos de tiempo
- k: Tasa mínima de rentabilidad exigida a la inversión (20% Anual)

Reemplazamos los valores del flujo de caja en la fórmula:

VAN = $-22{,}119.60 + \frac{34{,}454.96}{(1+0.20)^1}$

VAN = 6,592.87 euros

El Valor Actual Neto es mayor que cero, lo cual indica que la implementación del Modelo Predictivo es económicamente viable.

9.6. Tasa Interna de Retorno (TIR):

Representa la tasa de rendimiento a la cual el proyecto se hace indiferente, es decir cuando el VAN es igual a cero. Es la tasa de rentabilidad que iguala el valor actual de los beneficios y el valor actual de los costos.

Fórmula N° 2: Tasa Interna de Retorno

$$0 = -I_0 + \sum_{i=1}^{n} \frac{F_i}{(1+TIR)^i}$$

Donde:

- F_i: *Son los flujos de dinero en cada periodo*
- I_0: *Inversión inicial*
- *n: Número de periodos*

Procedemos a calcular la TIR según la fórmula anterior:

- $0 = -22{,}119.60 + \frac{34{,}454.96}{(1+TIR)^1}$
- TIR = 55.77%

La Tasa Interna de Retorno es superior a la tasa mínima de rentabilidad exigida a la inversión (20%), lo cual indica que la implementación del Modelo Predictivo es económicamente viable.

9.7. Beneficio Costo (B/C):

Es una herramienta financiera que resulta de dividir los beneficios esperados entre los costos previstos durante la vida útil del proyecto. Si el análisis de la relación B/C es mayor a la unidad entonces el proyecto es rentable. (5)

Fórmula N° 3: Beneficio Costo

$$\boldsymbol{B/C} = \frac{Beneficios\ Esperados\ Totales}{Costos\ Previstos\ Totales}$$

Procedemos a calcular el indicador B/C según fórmula anterior:

- **B/C** = $\frac{34{,}454.96}{22{,}119.60}$
- **B/C** = 1.56

El indicador B/C es mayor a la unidad, es decir los beneficios son mayores que los costos, por lo tanto, la implementación del Modelo Predictivo es económicamente viable.

9.8. Payback o Plazo de Recuperación:

Es un criterio para evaluar inversiones que se define como el periodo de tiempo requerido para recuperar el capital inicial de una inversión. (4)

Considerando que los flujos de caja son iguales durante todos los periodos, la fórmula para calcular el payback es:

Formula N° 4: Payback

$$\textbf{\textit{Payback}} = \frac{I_0}{F}$$

Donde:

- I_0*: Inversión Inicial*
- *F: Valor de los flujos de caja*

Procedemos a calcular el payback del proyecto según fórmula anterior:

- **Payback** = $\frac{22{,}119.60}{34{,}454.96}$
- **Payback** = 0.64 años

Se concluye que en este proyecto de inversión tardaremos 0.64 años (7 meses y 20 días) en recuperar el dinero desembolsado.

9.9. Viabilidad Económica del Proyecto:

La implementación del proyecto es económicamente viable porque los indicadores económicos calculados lo demuestran:

VAN:	6,592.87 euros
TIR:	55.77% > Tasa mínima de rentabilidad exigida a la inversión (20%)
B/C:	1.56 > Unidad
Payback	0.64 años (7 meses y 20 días)

10. Selección del Dataset:

Para llevar a cabo el desarrollo de este modelo, se ha elegido el uso del dataset llamado Home Credit Default Risk, proporcionado por la entidad financiera Home Credit a través de la plataforma Kaggle, una plataforma online especializada en todo lo relacionado con Data Science. (Group, 2018)

10.1. Descripción del dataset:

El dataset seleccionado está pensado especialmente para el problema que estamos tratando de resolver, la evaluación del riesgo de impago de un préstamo bancario por parte de un cliente. Es sin duda el dataset más completo de todos los que hemos encontrado. Se compone de un total de nueve archivos:

- **application_train.csv:** en este archivo de se encuentran registros de todas las solicitudes que ha recibido la entidad financiera, incluye una columna TARGET que indica si el crédito ha sido concedido o no.
- **application_test.csv:** este archivo es similar al anterior, aunque no incluye la columna TARGET, su objetivo principal es testear el modelo.
- **bureau.csv:** este archivo recoge todas las solicitudes realizadas por los clientes en otras entidades financieras.
- **bureau_balance.csv:** en este archivo encontraremos el balance mensual de todos los créditos mencionados en el archivo anterior
- **POS_CASH_balance.csv:** en este archivo encontramos los balances mensuales de todos los créditos que un cliente ha tenido con la entidad Home Credit.
- **credit_card_balance.csv:** en este fichero tenemos los balances mensuales de todas las tarjetas de crédito que un cliente tiene con la entidad Home Credit.
- **previous_application.csv:** un listado de todas las solicitudes de préstamo que un cliente ha mantenido con la entidad Home Credit
- **installments_payments.csv:** un historial de reembolsos de todos los créditos desembolsados en la entidad Home Credit.
- **HomeCredit_columns_description.csv:** un diccionario de datos de todas las columnas de todos los anteriores ficheros (Group, 2018)

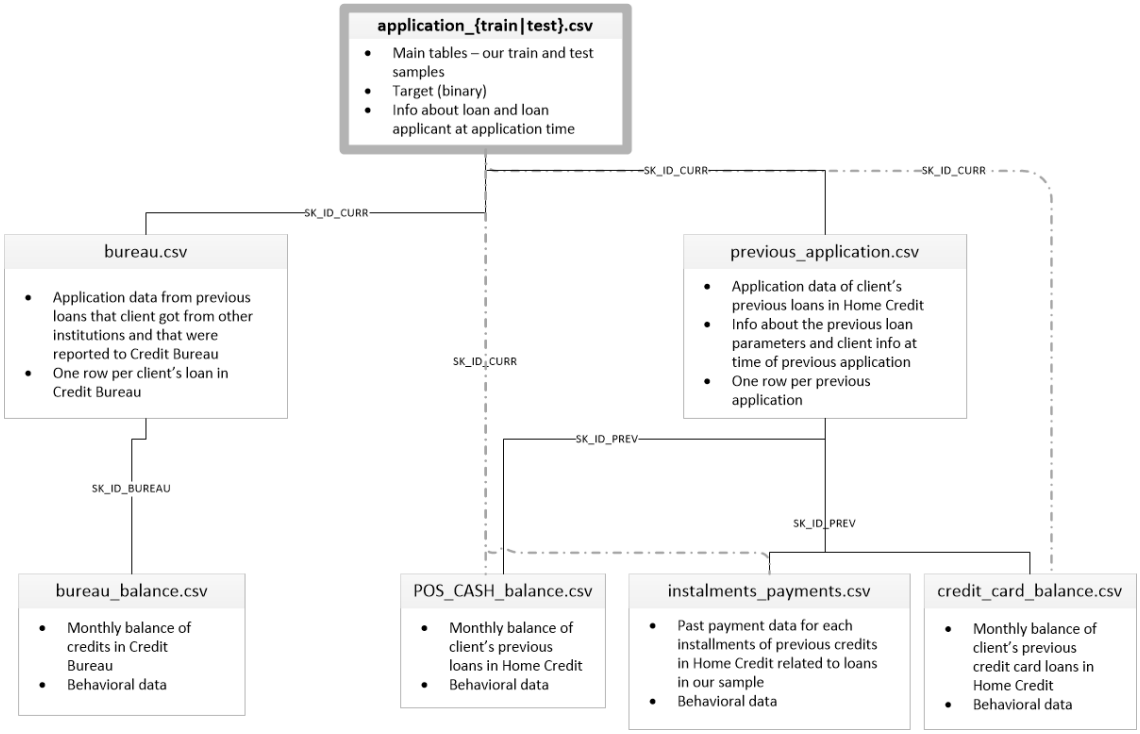
10.2. Diccionario de datos:

Se ha basado en el diccionario de datos original que se encuentra en el archivo **HomeCredit_columns_description.csv** para elaborar un nuevo diccionario de datos únicamente con las columnas con las que vamos a trabajar, también se han traducido las descripciones de esas columnas del idioma inglés al idioma español.

Se puede encontrar el nuevo diccionario de datos en el archivo "**Diccionario de Datos.xlsx**" adjunto. (Group, 2018)

10.3. Modelado de datos:

Los datos de este dataset se pueden considerar como un Modelo Relacional, dado que están perfectamente definidas todas las columnas y están relacionadas mediante ID's únicos ya sea por cliente o préstamo, en la siguiente imagen extraída de la página web de Kaggle se puede apreciar la relación de los datos.



Fuente: kaggle.com (Group, 2018) (Group, 2018)

10.4. Vista previa de los datos:

A continuación, se muestra una captura de imagen de cada set de datos:

10.4.1. Set de Datos: application_train

SK_ID_CURR <int>	TARGET <int>	NAME_CONTRACT_TYPE <chr>	CODE_GENDER <chr>	FLAG_OWN_CAR <chr>	FLAG_OWN_REALTY <chr>	CNT_CHILDREN <int>	AMT_INCOME_TOTAL <dbl>	AMT_CREDIT <dbl>	AMT_ANNUITY <dbl>
100002	1	Cash loans	M	N	Y	0	202500.00	406597.5	24700.5
100003	0	Cash loans	F	N	N	0	270000.00	1293502.5	35698.5
100004	0	Revolving loans	M	Y	Y	0	67500.00	135000.0	6750.0
100006	0	Cash loans	F	N	Y	0	135000.00	312682.5	29686.5
100007	0	Cash loans	M	N	Y	0	121500.00	513000.0	21865.5
100008	0	Cash loans	M	N	Y	0	99000.00	490495.5	27517.5
100009	0	Cash loans	F	Y	Y	1	171000.00	1560726.0	41301.0
100010	0	Cash loans	M	Y	Y	0	360000.00	1530000.0	42075.0
100011	0	Cash loans	F	N	Y	0	112500.00	1019610.0	33826.5
100012	0	Revolving loans	M	N	Y	0	135000.00	405000.0	20250.0

1-10 of 307,511 rows | 1-10 of 77 columns

Previous 1 2 3 4 5 6 ... 100 Next

Fuente: elaboración propia

10.4.2. Set de Datos: application_test

SK_ID_CURR <int>	NAME_CONTRACT_TYPE <chr>	CODE_GENDER <chr>	FLAG_OWN_CAR <chr>	FLAG_OWN_REALTY <chr>	CNT_CHILDREN <int>	AMT_INCOME_TOTAL <dbl>	AMT_CREDIT <dbl>	AMT_ANNUITY <dbl>	AMT_GOODS_PRICE <dbl>
100001	Cash loans	F	N	Y	0	135000	568800.0	20560.5	450000.0
100005	Cash loans	M	N	Y	0	99000	222768.0	17370.0	180000.0
100013	Cash loans	M	Y	Y	0	202500	663264.0	69777.0	630000.0
100028	Cash loans	F	N	Y	2	315000	1575000.0	49018.5	1575000.0
100038	Cash loans	M	Y	N	1	180000	625500.0	32067.0	625500.0
100042	Cash loans	F	Y	Y	0	270000	959688.0	34600.5	810000.0
100057	Cash loans	M	Y	Y	2	180000	499221.0	22117.5	373500.0
100065	Cash loans	M	N	Y	0	166500	180000.0	14220.0	180000.0
100066	Cash loans	F	N	Y	0	315000	364896.0	28957.5	315000.0
100067	Cash loans	F	Y	Y	1	162000	45000.0	5337.0	45000.0

1-10 of 48,744 rows | 1-10 of 76 columns

Previous 1 2 3 4 5 6 ... 100 Next

Fuente: elaboración propia

10.4.3. Set de Datos: bureau

SK_ID_CURR <int>	SK_ID_BUREAU <int>	CREDIT_ACTIVE <chr>	CREDIT_CURRENCY <chr>	DAYS_CREDIT <int>	CREDIT_DAY_OVERDUE <int>	DAYS_CREDIT_ENDDATE <dbl>	DAYS_ENDDATE_FACT <dbl>	AMT_CREDIT_MAX_OVERDUE <dbl>
215354	5714462	Closed	currency 1	-497	0	-153	-153	NA
215354	5714463	Active	currency 1	-208	0	1075	NA	NA
215354	5714464	Active	currency 1	-203	0	528	NA	NA
215354	5714465	Active	currency 1	-203	0	NA	NA	NA
215354	5714466	Active	currency 1	-629	0	1197	NA	77674.500
215354	5714467	Active	currency 1	-273	0	27460	NA	0.000
215354	5714468	Active	currency 1	-43	0	79	NA	0.000
162297	5714469	Closed	currency 1	-1896	0	-1684	-1710	14985.000
162297	5714470	Closed	currency 1	-1146	0	-811	-840	0.000
162297	5714471	Active	currency 1	-1146	0	-484	NA	0.000

1-10 of 1,716,428 rows | 1-9 of 17 columns

Previous 1 2 3 4 5 6 ... 100 Next

Fuente: elaboración propia

10.4.4. Set de Datos: bureau_balance

SK_ID_BUREAU <int>	MONTHS_BALANCE <int>	STATUS <chr>
5715448	0	C
5715448	-1	C
5715448	-2	C
5715448	-3	C
5715448	-4	C
5715448	-5	C
5715448	-6	C
5715448	-7	C
5715448	-8	C
5715448	-9	0

1-10 of 27,299,925 rows

Previous 1 2 3 4 5 6 ... 100 Next

Fuente: elaboración propia

10.4.5. Set de Datos: POS_CASH_balance

SK_ID_PREV <int>	SK_ID_CURR <int>	MONTHS_BALANCE <int>	CNT_INSTALMENT <dbl>	CNT_INSTALMENT_FUTURE <dbl>	NAME_CONTRACT_STATUS <chr>	SK_DPD <int>	SK_DPD_DEF <int>
1803195	182943	-31	48	45	Active	0	0
1715348	367990	-33	36	35	Active	0	0
1784872	397406	-32	12	9	Active	0	0
1903291	269225	-35	48	42	Active	0	0
2341044	334279	-35	36	35	Active	0	0
2207092	342166	-32	12	12	Active	0	0
1110516	204376	-38	48	43	Active	0	0
1387235	153211	-35	36	36	Active	0	0
1220500	112740	-31	12	12	Active	0	0
2371489	274851	-32	24	16	Active	0	0

1-10 of 10,001,358 rows

Previous 1 2 3 4 5 6 ... 100 Next

Fuente: elaboración propia

10.4.6. Set de Datos: credit_card_balance

SK_ID_PREV <int>	SK_ID_CURR <int>	MONTHS_BALANCE <int>	AMT_BALANCE <dbl>	AMT_CREDIT_LIMIT_ACTUAL <int>	AMT_DRAWINGS_ATM_CURRENT <dbl>	AMT_DRAWINGS_CURRENT <dbl>	AMT_DRAWINGS_OTHER_CURRENT <dbl>
2562384	378907	-6	56.970	135000	0	877.500	0
2582071	363914	-1	63975.555	45000	2250	2250.000	0
1740877	371185	-7	31815.225	450000	0	0.000	0
1389973	337855	-4	236572.110	225000	2250	2250.000	0
1891521	126868	-1	453919.455	450000	0	11547.000	0
2646502	380010	-7	82903.815	270000	0	0.000	0
1079071	171320	-6	353451.645	585000	67500	67500.000	0
2095912	118650	-7	47962.125	45000	45000	45000.000	0
2181852	367360	-4	291543.075	292500	90000	289339.425	0
1235299	203885	-5	201261.195	225000	76500	111026.700	0

1-10 of 3,840,312 rows | 1-8 of 23 columns

Previous 1 2 3 4 5 6 ... 100 Next

Fuente: elaboración propia

10.4.7. Set de Datos: previous_application

SK_ID_PREV <int>	SK_ID_CURR <int>	NAME_CONTRACT_TYPE <chr>	AMT_ANNUITY <dbl>	AMT_APPLICATION <dbl>	AMT_CREDIT <dbl>	AMT_DOWN_PAYMENT <dbl>	AMT_GOODS_PRICE <dbl>	WEEKDAY_APPR_PROCESS_START <chr>
2030495	271877	Consumer loans	1730.430	17145.000	17145.000	0.000	17145.000	SATURDAY
2802425	108129	Cash loans	25188.615	607500.000	679671.000	NA	607500.000	THURSDAY
2523466	122040	Cash loans	15060.735	112500.000	136444.500	NA	112500.000	TUESDAY
2819243	176158	Cash loans	47041.335	450000.000	470790.000	NA	450000.000	MONDAY
1784265	202054	Cash loans	31924.395	337500.000	404055.000	NA	337500.000	THURSDAY
1383531	199383	Cash loans	23703.930	315000.000	340573.500	NA	315000.000	SATURDAY
2315218	175704	Cash loans	NA	0.000	0.000	NA	NA	TUESDAY
1656711	296299	Cash loans	NA	0.000	0.000	NA	NA	MONDAY
2367563	342292	Cash loans	NA	0.000	0.000	NA	NA	MONDAY
2579447	334349	Cash loans	NA	0.000	0.000	NA	NA	SATURDAY

1-10 of 1,670,214 rows | 1-9 of 37 columns

Previous 1 2 3 4 5 6 ... 100 Next

Fuente: elaboración propia

10.4.8. Set de Datos: installments_payments

SK_ID_PREV <int>	SK_ID_CURR <int>	NUM_INSTALMENT_VERSION <dbl>	NUM_INSTALMENT_NUMBER <int>	DAYS_INSTALMENT <dbl>	DAYS_ENTRY_PAYMENT <dbl>	AMT_INSTALMENT <dbl>	AMT_PAYMENT <dbl>
1054186	161674	1	6	-1180	-1187	6948.360	6948.360
1330831	151639	0	34	-2156	-2156	1716.525	1716.525
2085231	193053	2	1	-63	-63	25425.000	25425.000
2452527	199697	1	3	-2418	-2426	24350.130	24350.130
2714724	167756	1	2	-1383	-1366	2165.040	2160.585
1137312	164489	1	12	-1384	-1417	5970.375	5970.375
2234264	184693	4	11	-349	-352	29432.295	29432.295
1818599	111420	2	4	-968	-994	17862.165	17862.165
2723183	112102	0	14	-197	-197	70.740	70.740
1413990	109741	1	4	-570	-609	14308.470	14308.470

1-10 of 13,605,401 rows

Previous 1 2 3 4 5 6 ... 100 Next

Fuente: elaboración propia

10.5. Tratamiento de datos del dataset:

En la exploración inicial que hemos realizado sobre los datos contenidos en el Dataset, se observa que hay algunas columnas que no aportan información de valor para la resolución de nuestro problema además observamos algunos datos nulos o incompletos.

Aunque el conjunto de datos en si es bastante completo y con una calidad del dato bastante elevada, es necesario hacer sobre ello; y así se hará; un tratamiento para adecuarlos a nuestro modelo. Algunas de las tareas que se van a llevar a cabo sobre el conjunto de datos son:

- Transformación de los datos al tipo adecuado
- Normalización y codificación de los datos

En el momento del desarrollo del problema, se hará un análisis previo de cada una de las variables definidas en el Diccionario de datos adjunto.

11. Fuentes de Información y Benchmarking

11.1. Fuentes externas de información:

Las fuentes de información es algo crucial para cualquier análisis estadístico, y lo es aún más cuando se trata de un sector como el que tratamos en este trabajo. En nuestro dado que el trabajo lo estamos haciendo sobre un dataset ya final, no necesitamos incorporar fuentes externas para el desarrollo de un modelo predictivo dado que el dataset escogido ya de por sí ofrece una cantidad y variedad de datos inmensa y no sólo hablamos del número de observaciones que son millones, sino en la cantidad de variables predictores que son unas de 122 variables diferentes que van desde datos simples como la edad de cliente hasta datos ya más completos acerca de puntuaciones que tiene un cliente en fuentes externas a la entidad que ha elaborado el dataset.

Por ese motivo, consideramos que no necesitamos añadir nuevas fuentes para completar la información de este dataset.

11.2. Benchmarking:

A lo largo del desarrollo de los diferentes modelos que se han realizado para este trabajo, se han hecho comparativas de absolutamente todas las acciones y no sólo de qué modelo es el que tiene mejor ajuste, incluso se han realizado comparativas de una variable antes y después de sufrir una imputación.

La tarea de benchmarking ha demostrado ser clave para garantizar un resultado óptimo y por eso ha de ser el centro de todas las acciones, en nuestro caso se han realizado las siguientes comparativas:

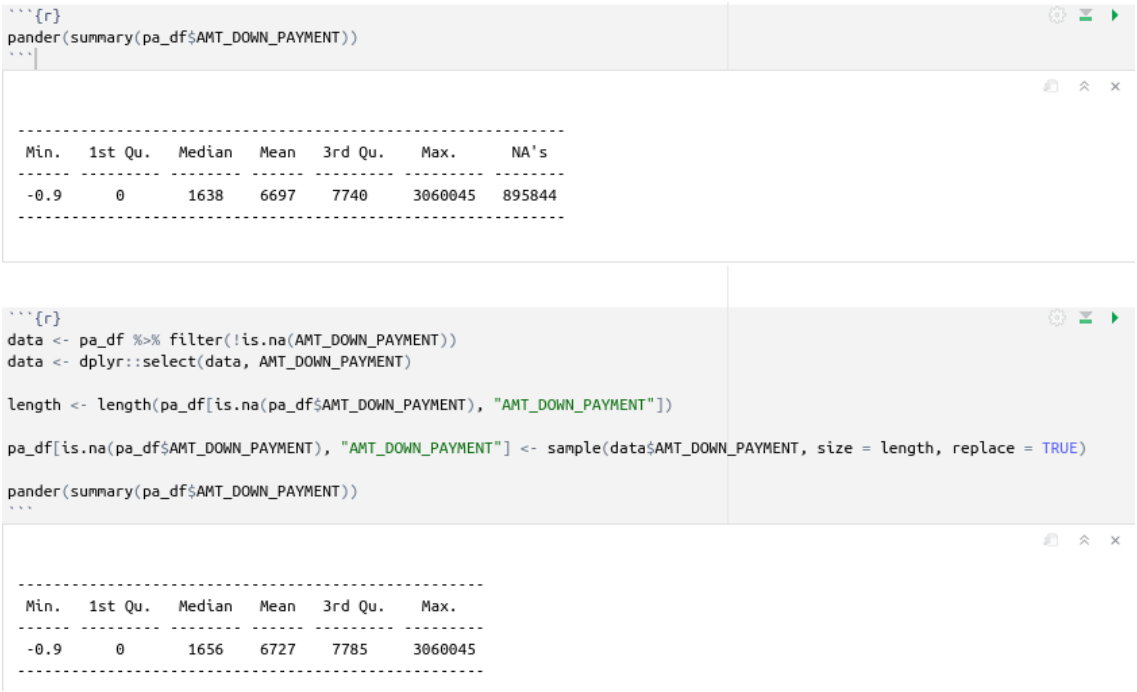
11.2.1. Benchmarking de variables:

Los datos de este dataset no son limpios, hay infinidad de valores restantes en diferentes variables, la gran mayoría de ellas son variables numéricas, por lo que hemos tenido que hacer imputaciones en su mayoría generando un valor aleatorio en base a los valores que si existen.

Para que esta imputación de variables sea aceptable desde nuestro punto de vista, hemos considerado que no debe afectar en gran medida al resultado del summary de lo que era antes de dicha imputación. Por lo que hemos basado nuestra estrategia de imputación en

vigilar el output del summary de cualquier variable a imputar antes y después de imputarla y ver que no ha habido grandes variaciones en métricas básicas como el mínimo, máximo, el primer y tercer cuartil, la media o la mediana.

Por ejemplo:



```r
pander(summary(pa_df$AMT_DOWN_PAYMENT))
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
-0.9	0	1638	6697	7740	3060045	895844

```r
data <- pa_df %>% filter(!is.na(AMT_DOWN_PAYMENT))
data <- dplyr::select(data, AMT_DOWN_PAYMENT)

length <- length(pa_df[is.na(pa_df$AMT_DOWN_PAYMENT), "AMT_DOWN_PAYMENT"])

pa_df[is.na(pa_df$AMT_DOWN_PAYMENT), "AMT_DOWN_PAYMENT"] <- sample(data$AMT_DOWN_PAYMENT, size = length, replace = TRUE)

pander(summary(pa_df$AMT_DOWN_PAYMENT))
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-0.9	0	1656	6727	7785	3060045

Benchmarking de antes y después de una imputación de una variable – Fuente: elaboración propia

Como se puede ver en este ejemplo, siempre mantenemos vigiladas las métricas estadísticas básicas de cualquier variable a imputar para asegurarnos de que dicha imputación de los valores faltantes no tenga un gran afecto sobre la media (o cualquier otra métrica) de la variable. Esto lo conseguimos gracias a un benchmarking de las variables.

11.2.2. Benchmarking entre modelos:

Elegir el modelo adecuado no sería posible sin hacer comparativas entre varios modelos, esto es algo que es de obligada realización, por ese motivo en la sección de **"La mejor especificación"** se ha realizado benchmarking usando las diferentes técnicas Stepwise, que ya de por si es un benchmarking entre todas las combinaciones de modelos que pueden formar nuestras variables predictoras, y se ha optado el que arroja el valor AIC más bajo entre las 3 técnicas step posibles:

- **Backward:** que parte de un modelo con todas las variables y va quitando variables hasta encontrar el que tiene el menor coeficiente AIC
- **Forward:** que parte de un modelo vacío y va agregando variables hasta encontrar también el modelo con el menor coeficiente AIC
- **Stepwise:** que es una mezcla entre las dos técnicas anteriores, va quitando y añadiendo variables constantemente hasta encontrar el modelo con el menor coeficiente AIC.

En nuestro caso, todas estas técnicas nos han arrojado el mismo modelo, pero en algunos casos no siempre coinciden, por lo que tenemos una posibilidad de encontrar el modelo ideal siempre observando el coeficiente AIC o BIC (si optamos por este último). Este benchmarking nos asegura que el modelo final elegido es el ideal.

En la sección de “La super mejor especificación”, una vez definido el modelo con el mejor ajuste, se procede a validar la función link para saber cuál de ellas es la que arroja mejor ajuste, logit, probit o cloglog.

Aunque siempre nos arroja pequeñas diferencias en el coeficiente AIC, esta comparativa nos asegura que la elección que hemos realizado sea la correcta y ha sido fundada sobre algún razonamiento que hemos establecido.

12. Hipótesis y métrica de ajuste

12.1. Hipótesis inicial:

El desarrollo del modelo predictivo predice con eficiencia la probabilidad de impago de un cliente y con un umbral por defecto del 50% no se reduce la morosidad en una entidad financiera.

12.2. Métrica de ajuste:

Antes de decidir cuál va a ser nuestra métrica de ajuste, primero debemos sentar las bases de lo que son los falsos negativos y falsos positivos en nuestro caso.

Recordando que el objetivo de nuestro modelo es evaluar el riesgo de impago de un préstamo, lo cual un positivo para nosotros sería un **imapagador** (mal cliente) y un negativo es un pagador (buen cliente), por lo que para nuestro caso:

- Un falso negativo: es una persona que el modelo cree que va a pagar, pero realmente no tiene capacidad de pago.
- Un falso positivo: es una persona que el modelo cree que no va a poder pagar, pero en realidad es una persona que si tiene esa capacidad.

Una vez claras estas definiciones, nos enfrentamos al siguiente dilema: ¿qué es lo más importante en nuestro caso?

- Reducir los falsos negativos, es decir que el modelo sea muy preciso y por lo tanto restrictivo con la concesión de los préstamos, lo que se traduce en conceder menos préstamos, pero a clientes buenos.
- Reducir los falsos positivos, es decir conceder el máximo número de préstamos sin importar ya que puede que haya buenos clientes dentro de los falsos positivos, esto se traduce en mayor cantidad de préstamos concedidos.

Para tomar una decisión de cuál es la mejor estrategia, hay que tener en cuenta también el coste por adquisición de un cliente:

- Reduciendo los falsos negativos, tendríamos un coste por cliente bastante elevado, ya que con la misma inversión captaremos mucho menos clientes, pero buenos clientes.

- Reduciendo los falsos positivos, tendríamos un coste por cliente bastante bajo ya que el número de clientes que captaremos con una inversión similar es mucho mayor que en el caso de clientes.

Este tema realmente depende de cada empresa y de su estrategia, pero en nuestro caso lo creemos que la elección no debe ser una u otra, sino ambas y para ello vamos a apoyarnos en una variable muy importante dentro del modelo que es **AMT_CREDIT** (Amount Credit, el monto de crédito). Aquí su summary:

```
```{r}
pander(summary(application_train$AMT_CREDIT))
```
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
45000	270000	513531	599026	808650	4050000

Summary de la variable AMT_CREDIT dentro del dataset application_train.csv - Fuente: elaboración propia

Es importante destacar que la moneda con la que se trabaja en este dataset es el yuan chino (de ahí las cifras elevadas).

Nuestra idea adoptar una medida de ajuste u otra según el monto solicitado ya que después ver esas métricas hemos coincidido que el dataset que se nos presenta se trata de créditos de consumo por lo que no tiene ningún sentido ser restrictivos para importes bastante pequeños, que son la mayoría tal y como se ven en el summary superior.

Pero si debemos ser algo restrictivos cuando se trata de importes elevados, es por ello que hemos establecido el **3º Cuartil como límite (808.650)** para adoptar una u otra métrica de ajuste.

- Si el importe solicitado es menor que el 3º cuartil, entonces buscaremos reducir al máximo los falsos positivos concediendo el mayor número de créditos posibles.
- Si el importe solicitado es mayor que el 3º cuartil, entonces buscaremos mayor precisión dando preferencia a clientes fiable que puedan asumir el coste del préstamo.

La elección de este límite se fundamenta sobre todo en una revisión a la variable **AMT_INCOME_TOTAL**, que es el total de ingresos de un cliente. Donde el valor medio es

de 168.798, que representa aproximadamente un 20.87% del límite establecido, es decir una quinta parte de del límite establecido.

```{r}
pander(summary(application_train$AMT_INCOME_TOTAL))
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
25650	112500	147150	168798	202500	1.17e+08

Summary de la variable AMT_INCOME_TOTAL dentro del dataset application_train.csv - Fuente: elaboración propia

Además de esta variable, nos hemos fijado en la tasa de interés, al ser mayoritariamente préstamos de consumo, la tasa de interés es bastante elevada por lo que el ROI de un préstamo es bastante elevado. Lo que nos permite concluir que, concediendo bastantes préstamos (incluyendo a malos clientes, falsos negativos) tendremos un elevado ROI final debido a la tasa de interés elevada que estamos aplicando a cada uno de los préstamos concedidos.

12.3. Elección final de la métrica de ajuste:

Finalmente, y para este trabajo consideramos que es fundamental aumentar el RECALL, ya que con ello evitamos grandes pérdidas para la entidad financiera al reducir la cantidad de clientes morosos.

Aunque tenemos claro que esta elección va a tener un efecto negativo sobre la PRECISION, por ese motivo haremos una comparativa de diferentes modelos predictivos (GLM vs Random Forest) donde buscaremos un equilibrio entre ambas métricas, pero siempre dando preferencia al RECALL.

13. Desarrollo

13.1. Comparativas de Modelos Matemáticos Tradicionales Vs. Modelos de Machine Learning

A continuación, se analiza los siguientes indicadores:

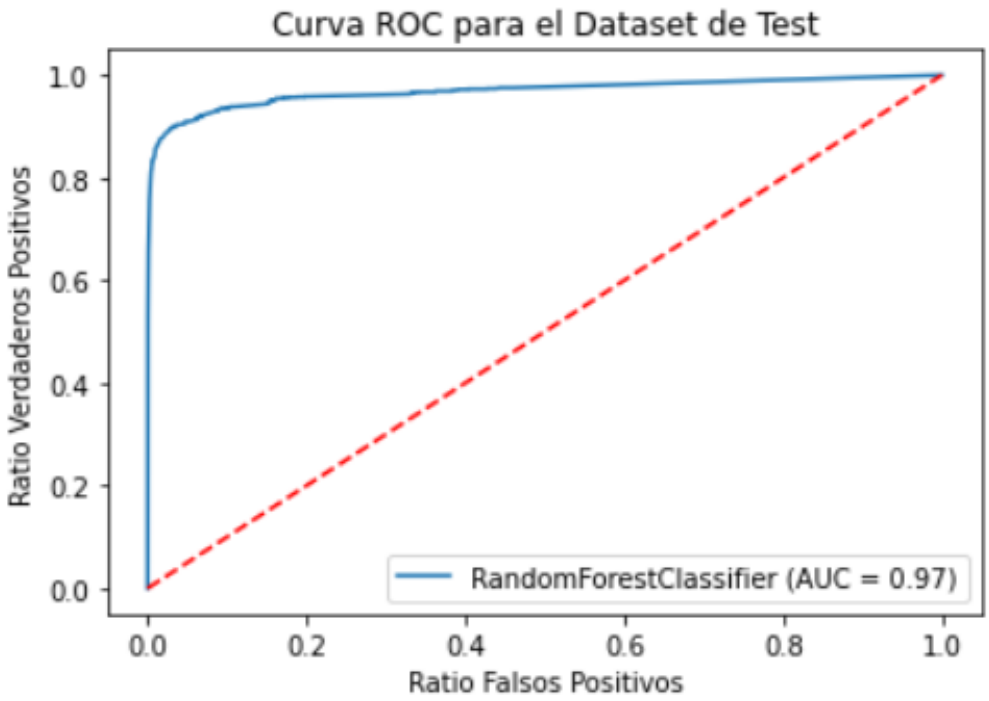
- Curva ROC (Receiver Operating Characteristic)
- Recuperación (Recall)

Coeficiente Kappa de Cohen

13.1.1. Curva ROC (Receiver Operating Characteristic)

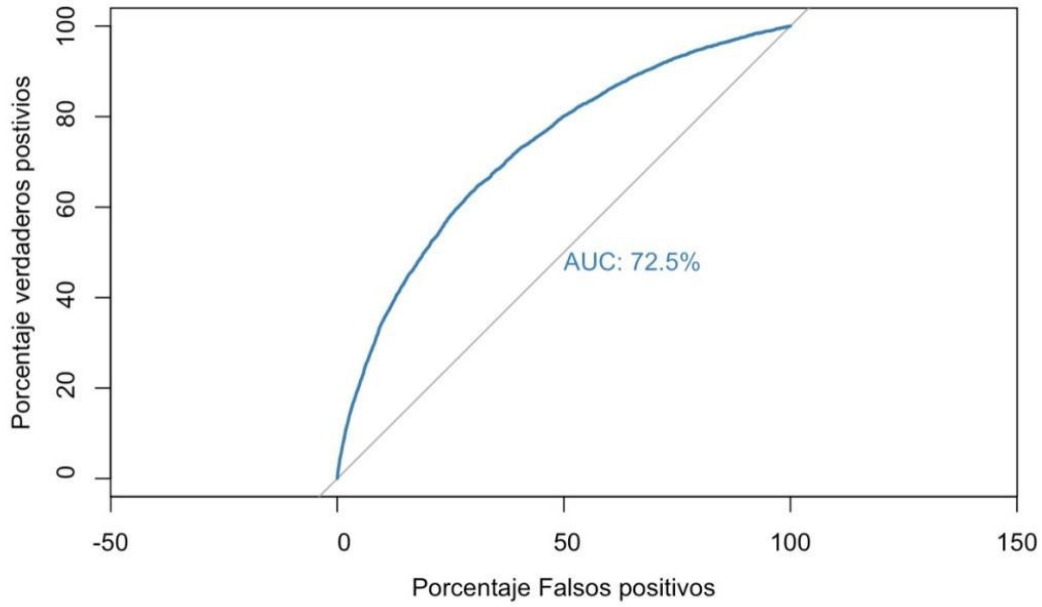
Se observa que el indicador AUC (Area Under the ROC Curve) del modelo Random Forest es del 0.97, superior al AUC del Modelo Lineal General. Tener un AUC del 0.97 significa que las predicciones se están clasificando bastante bien, independientemente del umbral definido.

Imagen: Curva ROC del modelo Random Forest



Fuente: elaboración propia

Imagen: Curva ROC del Modelo Lineal General



Fuente: elaboración propia

13.1.2. Recuperación (Recall)

Se observa que el modelo Random Forest tiene un recall del 88% y un accuracy del 98%, mientras que el Modelo Lineal General tiene un recall y accuracy del 95.6 y 67% respectivamente. El accuracy del Modelo Lineal General es bajo, ya que se está equivocando en el 33% de las predicciones.

El recall del modelo Random Forest es bueno, ya que solo el 12% de impagos no se está detectando. Sin embargo, es inferior al recall del modelo GLM, esto se debe a que se ha decidido definir un umbral del 22.5% para los no impagos, superior al umbral del 8% del modelo GLM, con el fin de permitir a la institución financiera contar con una cartera de clientes más amplia y corriendo un riesgo bajo de impago.

Imagen: Matriz de confusión y reporte de clasificación del modelo Random Forest con un umbral del 22.5%

```
Matríz de confusión:
 [[13850   217]
 [  156  1093]]

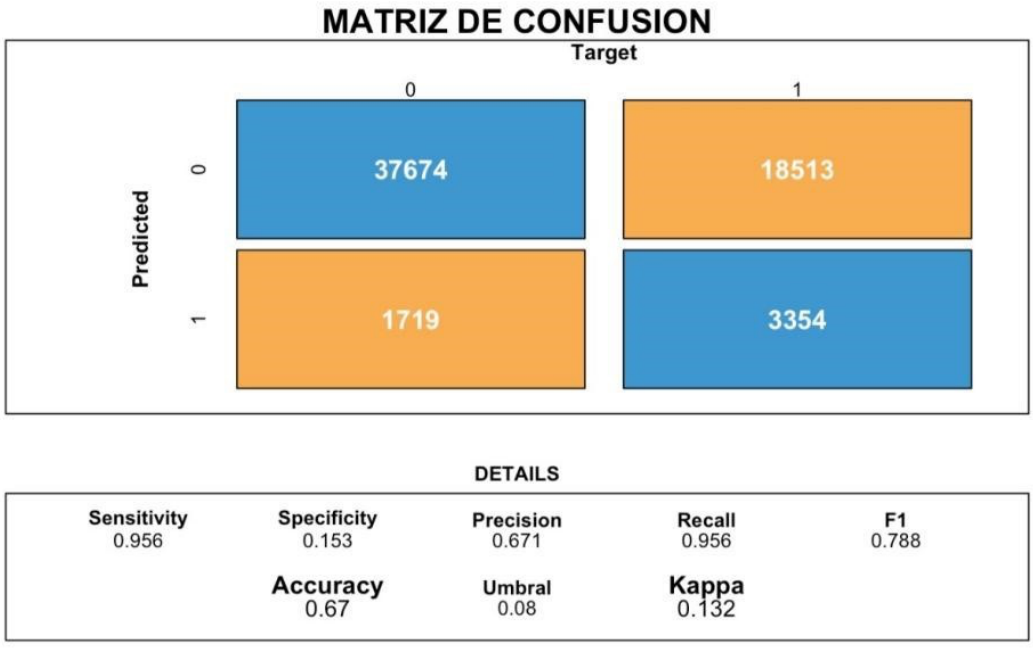
Reporte de clasificación:
              precision    recall  f1-score   support

           0       0.99      0.98      0.99     14067
           1       0.83      0.88      0.85      1249

    accuracy                           0.98     15316
   macro avg       0.91      0.93      0.92     15316
weighted avg       0.98      0.98      0.98     15316
```

Fuente: elaboración propia

Imagen: Matriz de confusión y reporte de clasificación del Modelo Lineal General con un umbral del 8%



Fuente: elaboración propia

13.1.3. Coeficiente Kappa de Cohen

El coeficiente Kappa de Cohen del modelo Random Forest es del 0.87, muy superior al coeficiente del Modelo Lineal General, además está bastante cercano a uno (1), lo cual significa que existe una fuerte concordancia entre los valores observados y predecidos.

Tabla: Coeficientes Kappa de Cohen del Modelo Lineal General y Random Forest

Modelo	Coeficiente Kappa de Cohen
Modelo Random Forest	0.87
Modelo Lineal General (GLM)	0.13

Fuente: elaboración propia

13.1.4. Conclusión

De acuerdo con el análisis de los tres indicadores:

- Curva ROC (Receiver Operating Characteristic)
- Recuperación (Recall)
- Coeficiente Kappa de Cohen

Concluimos que el mejor modelo es **Random Forest.**

13.2. Gráficas y Explicaciones Justificadas de los pasos dados

A continuación, se describe los pasos realizados para el desarrollo del Modelo Random Forest:

13.2.1. Preparación de los datos

Se realizó la imputación de valores faltantes y la conversión de datos a numéricos:

Imagen: Imputación de valores faltantes

```
> Imputación de valores faltantes en la variable AMT_GOODS_PRICE, considerando que solo aplica a préstamos de consumo, imputaremos el valor cero en estos créditos con la premisa que son de otro tipo:

```{r}

summary(application_train_df$AMT_GOODS_PRICE)
```

   Min. 1st Qu. Median   Mean 3rd Qu.   Max.   NA's
  40500 238500 450000 538396 679500 4050000   278

```{r}

application_train_df[is.na(application_train_df$AMT_GOODS_PRICE), "AMT_GOODS_PRICE"] <- 0
```

```{r}

summary(application_train_df$AMT_GOODS_PRICE)
```

   Min. 1st Qu. Median   Mean 3rd Qu.   Max.
      0 238500 450000 537909 679500 4050000
```

Fuente: elaboración propia

Imagen: Conversión de datos a numéricos

```
[4]  le = LabelEncoder()
     df_app["NAME_CONTRACT_TYPE_NUM"] = le.fit_transform(df_app["NAME_CONTRACT_TYPE"])

     le = LabelEncoder()
     df_app["CODE_GENDER_NUM"] = le.fit_transform(df_app["CODE_GENDER"])

     le = LabelEncoder()
     df_app["FLAG_OWN_CAR_NUM"] = le.fit_transform(df_app["FLAG_OWN_CAR"])

     le = LabelEncoder()
     df_app["FLAG_OWN_REALTY_NUM"] = le.fit_transform(df_app["FLAG_OWN_REALTY"])
```

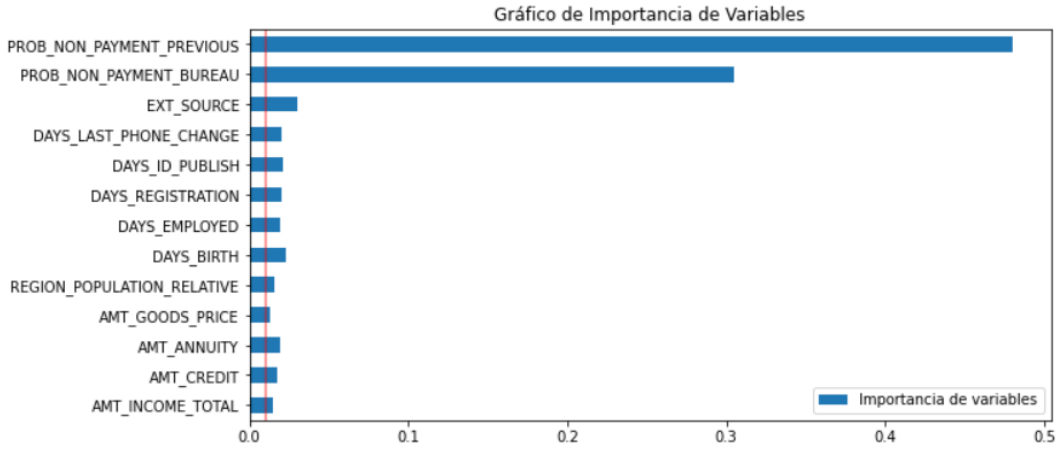
Fuente: elaboración propia

13.2.2. Entrenamiento del modelo

Se realizó la evaluación del modelo Random Forest con distintos parámetros con el objetivo de encontrar los valores óptimos y obtener la mejor especificación.

El siguiente gráfico representa las variables más relevantes en el modelo:

Imagen: Gráfico de Importancia de Variables



Fuente: elaboración propia

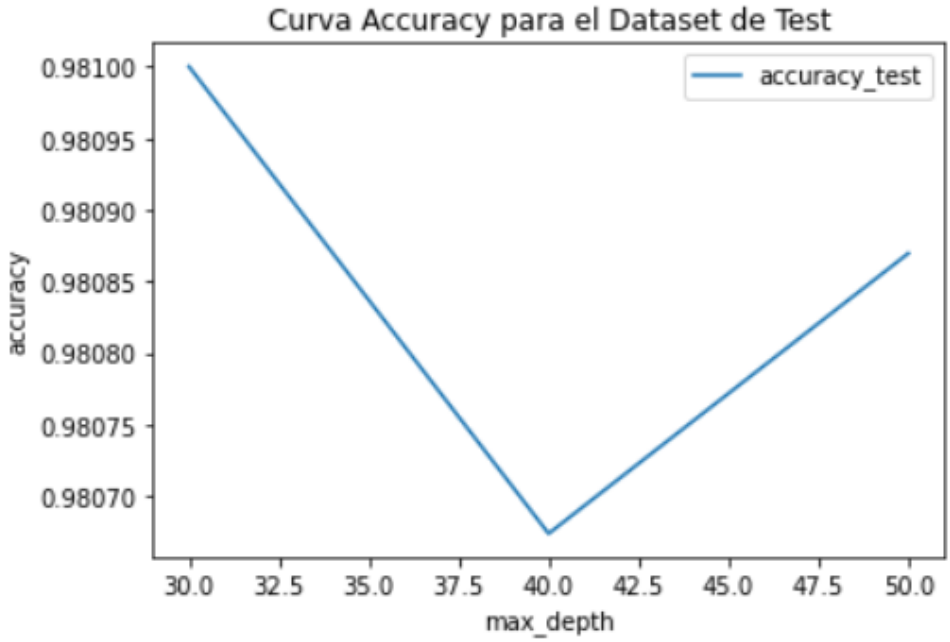
Luego de encontrar la mejor especificación, se procedió a construir nuevamente el modelo Randon Forest con diferentes parámetros para encontrar los valores óptimos, el resultado fue el siguiente:

Imagen: Accuracy para el dataset de entreno y prueba para diferentes parámetros

	n_estimators	max_depth	accuracy_train	accuracy_test
0	50	30	0.999680	0.981000
1	50	40	0.999790	0.980674
2	50	50	0.999790	0.980870
3	75	30	0.999873	0.980804
4	75	40	0.999962	0.980543
5	75	50	0.999962	0.980609

Fuente: elaboración propia

Image: Curva Accuracy para el Dataset de Prueba



FUENTE:

Fuente: elaboración propia

Después se observar los resultados, se seleccionó los mejores parámetros para el dataset de prueba:

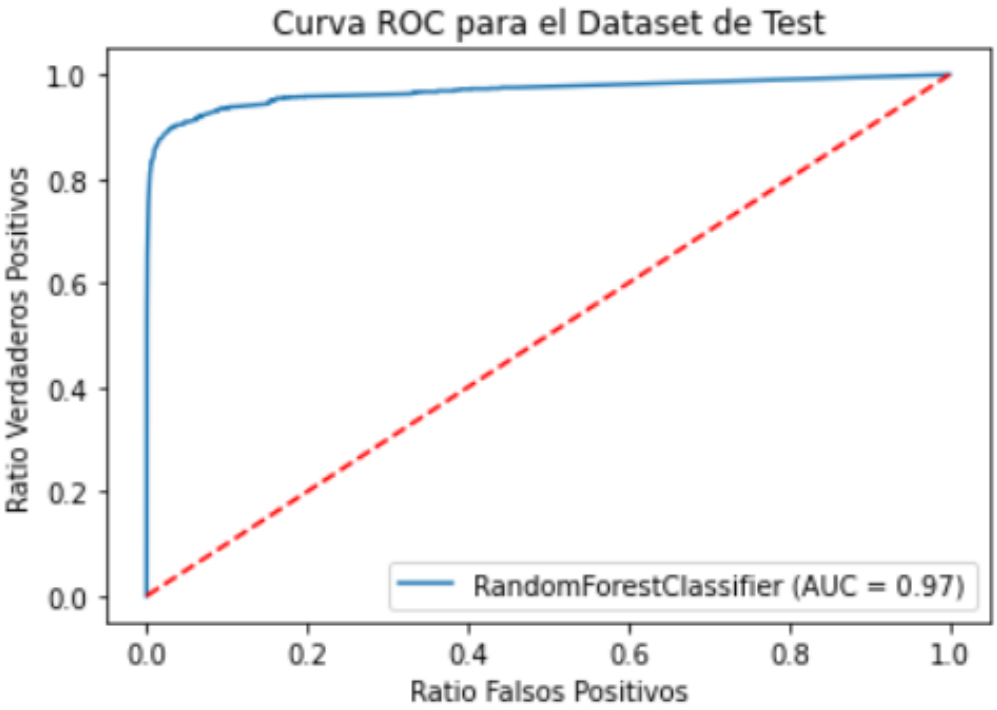
- n_estimators = 50
- max_depth = 30

13.2.3. Evaluación del modelo:

Se aplicó las siguientes métricas:

a. **Curva ROC (Receiver Operating Characteristic):** Se observa que el indicador AUC (Area Under the ROC Curve) del modelo Random Forest es del 0.97, lo cual significa que las predicciones se están clasificando bastante bien, independientemente del umbral definido.

Imagen: Curva ROC del modelo Random Forest



Fuente: elaboración propia

b. **Recuperación (Recall):** Según el siguiente reporte de clasificación, se observa que para predecir NO IMPAGOS (Valor cero) el modelo está acertando muy bien, sin embargo, en la predicción de IMPAGOS (Valor uno) se observa una **precision = 93%** y un **recall = 83%**, es decir el 17% de impagos no se está detectando. Teniendo en cuenta que nuestro objetivo es reducir los falsos negativos, procederemos a definir un umbral para ser más estrictos para que una predicción sea cero (No impago), de esta forma conseguiremos reducir los falsos negativos y aumentar el recall.

Imagen: Reporte de clasificación del modelo Random Forest

Reporte de Clasificación para el Dataset de Test:

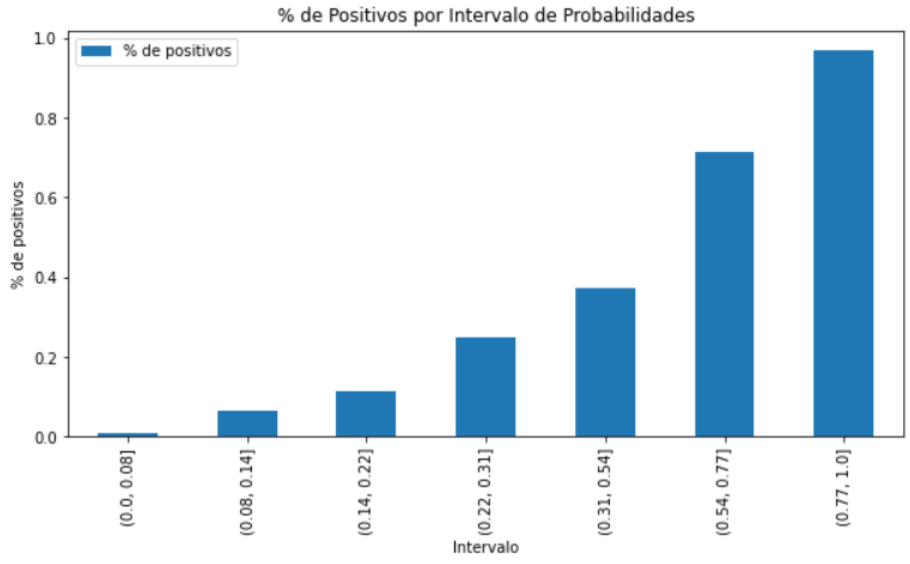
	precision	recall	f1-score	support
0	0.99	0.99	0.99	14067
1	0.93	0.83	0.88	1249
accuracy			0.98	15316
macro avg	0.96	0.91	0.93	15316
weighted avg	0.98	0.98	0.98	15316

Fuente: elaboración propia

c. **Coeficiente Kappa de Cohen:** El coeficiente Kappa de Cohen es 0.87, bastante cercano a uno (1), lo cual significa que existe una fuerte concordancia entre los valores observados y predichos.

d. **Cálculo del % de positivos por intervalo de probabilidades:** En el gráfico se observa que a partir del 22% de probabilidad de impago, el porcentaje de positivos se incrementa notablemente.

Imagen: Porcentaje de positivos por intervalo de probabilidades



Fuente: elaboración propia

e. **Cálculo del multiplicador para alcanzar el % de positivos de todo el dataset:** En el siguiente cuadro se presenta los multiplicadores para cada intervalo de probabilidades de impago.

   Por ejemplo, si una probabilidad cae en el intervalo (0.08, 0.14], donde el porcentaje de positivos es 6.40%, se concluye que está a 0.79 veces de la media de positivos del dataset (8.10%), es decir la probabilidad que el cliente impague es baja.

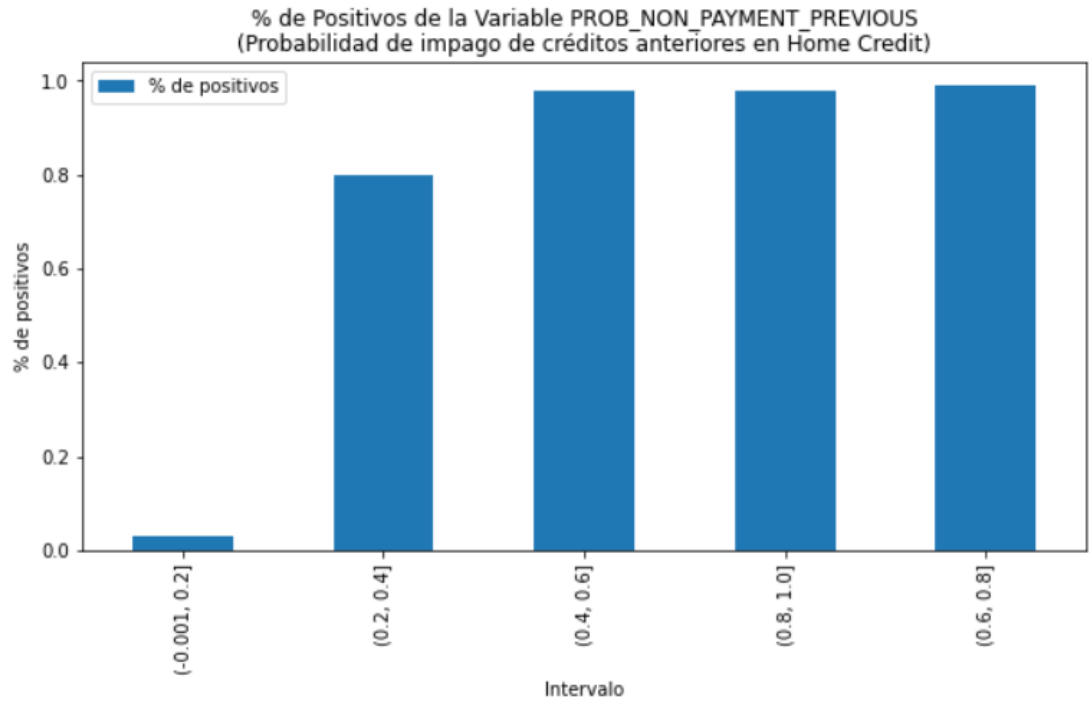
Imagen: Multiplicador referente a la media de positivos del dataset

	Intervalo	% Positivos Intervalo	% Positivos Total	Multiplicador
0	(0.0, 0.08]	0.8	8.1	0.10
1	(0.08, 0.14]	6.4	8.1	0.79
2	(0.14, 0.22]	11.3	8.1	1.40
3	(0.22, 0.31]	25.0	8.1	3.09
4	(0.31, 0.54]	37.3	8.1	4.60
5	(0.54, 0.77]	71.3	8.1	8.80
6	(0.77, 1.0]	97.0	8.1	11.98

Fuente: elaboración propia

f. **Cálculo del porcentaje de positivos para la variable PROB_NON_PAYMENT_PREVIOUS:** En el siguiente gráfico se observa que a partir del 20% de probabilidad de impago en créditos anteriores en Home Credit, existe un alto porcentaje de positivos (Impagos).

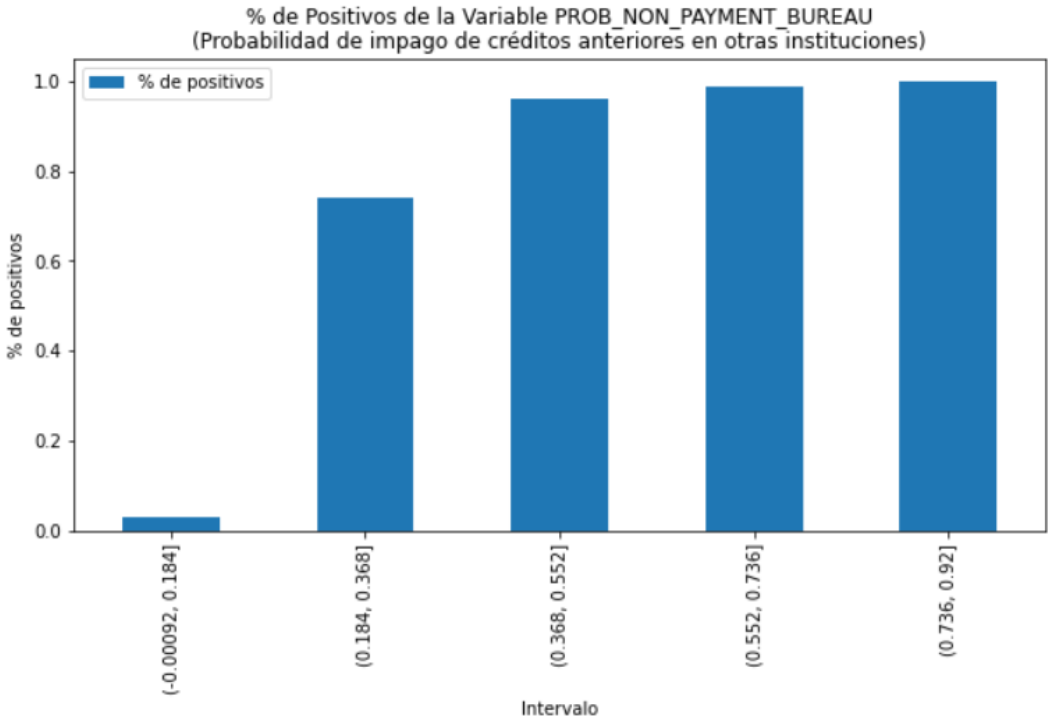
Imagen: Porcentaje de positivos para la variable PROB_NON_PAYMENT_PREVIOUS



Fuente: elaboración propia

g. **Cálculo del porcentaje de positivos para la variable PROB_NON_PAYMENT_BUREAU:** En el siguiente gráfico se observa que a partir del 18.4% de probabilidad de impago en créditos anteriores en otras instituciones, existe un alto porcentaje de positivos (Impagos).

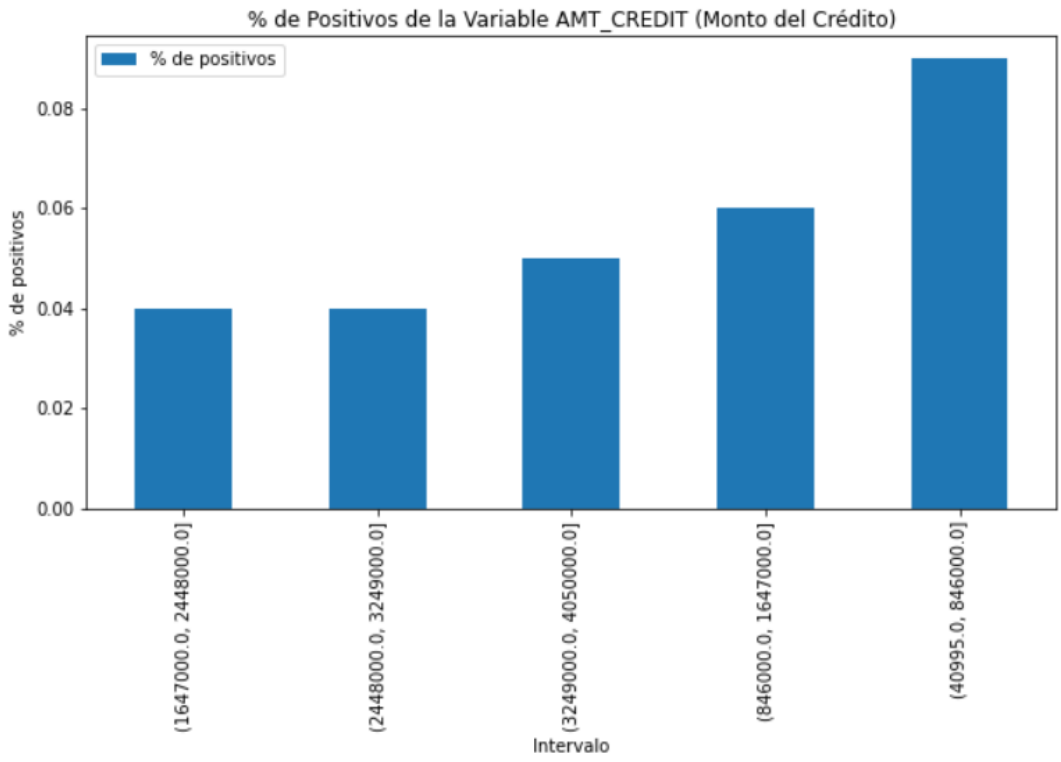
Imagen: Porcentaje de positivos para la variable PROB_NON_PAYMENT_BUREAU



Fuente: elaboración propia

h. **Cálculo del porcentaje de positivos para la variable AMT_CREDIT:** En el gráfico se observa que los montos comprendidos entre 40,995.0 y 1,647,000.0 de yuanes, tienen un alto porcentaje de positivos (impagos).

Imagen: Porcentaje de positivos para la variable AMT_CREDIT

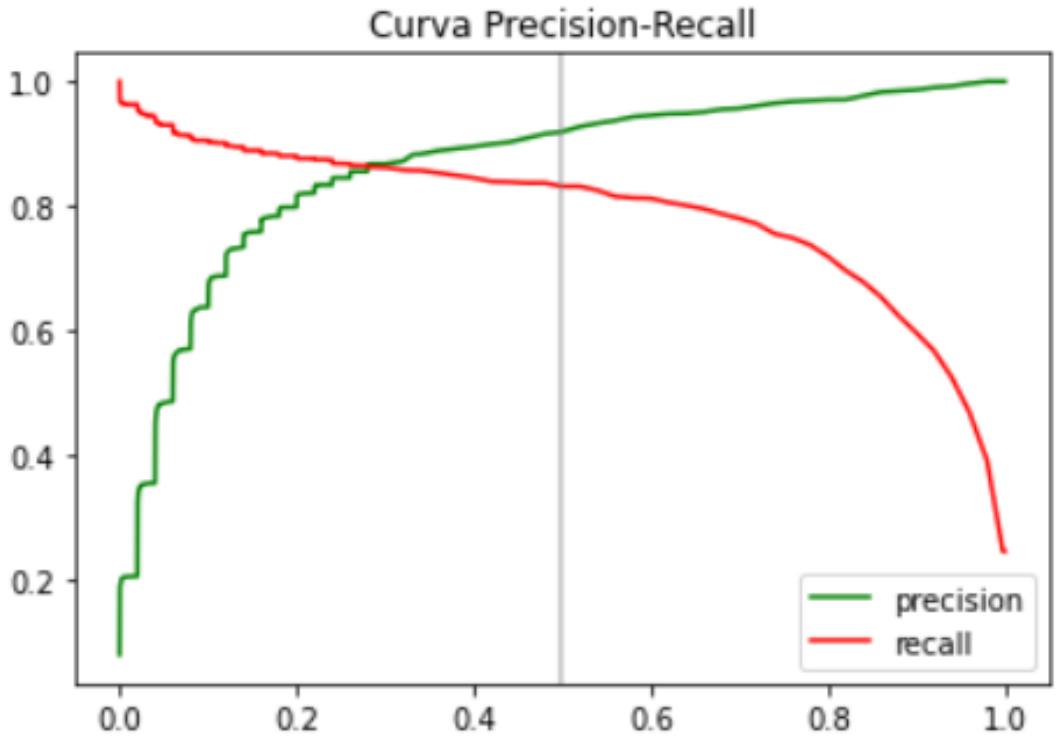


Fuente: elaboración propia

13.2.4. Definición de umbral:

Graficamos las curvas de precisión y recall para diferentes umbrales:

Imagen: Curva Precision-Recall



Fuente: elaboración propia

Nuestro objetivo es reducir los falsos negativos, es decir buscamos maximizar el recall. Según la gráfica **Curva Precision-Recall** debemos disminuir el límite de 50% y teniendo en cuenta el punto de corte, validaremos con 20.0%, 22.5% y 25.0%:

- **Generamos la matriz de confusión y reporte de clasificación con un umbral del 20.0%:**

Imagen: Matriz de confusión y reporte de clasificación del modelo Random Forest con un umbral del 20.0%

```
Matríz de confusión:
 [[13822   245]
 [  154  1095]]

Reporte de clasificación:
              precision    recall  f1-score   support

           0       0.99      0.98      0.99     14067
           1       0.82      0.88      0.85      1249

    accuracy                           0.97     15316
   macro avg       0.90      0.93      0.92     15316
weighted avg       0.97      0.97      0.97     15316
```

Fuente: elaboración propia

- **Generamos la matriz de confusión y reporte de clasificación con un umbral del 22.5%:**

Imagen: Matriz de confusión y reporte de clasificación del modelo Random Forest con un umbral del 22.5%

```
Matríz de confusión:
 [[13850   217]
 [  156  1093]]

Reporte de clasificación:
              precision    recall  f1-score   support

           0       0.99      0.98      0.99     14067
           1       0.83      0.88      0.85      1249

    accuracy                           0.98     15316
   macro avg       0.91      0.93      0.92     15316
weighted avg       0.98      0.98      0.98     15316
```

Fuente: elaboración propia

- **Generamos la matriz de confusión y reporte de clasificación con un umbral del 25.0%:**

Imagen: Matriz de confusión y reporte de clasificación del modelo Random Forest con un umbral del 25.0%

```
Matríz de confusión:
 [[13869   198]
 [  165  1084]]

Reporte de clasificación:
              precision    recall  f1-score   support

           0       0.99      0.99      0.99     14067
           1       0.85      0.87      0.86      1249

    accuracy                           0.98     15316
   macro avg       0.92      0.93      0.92     15316
weighted avg       0.98      0.98      0.98     15316
```

Fuente: elaboración propia

Análisis de umbrales:

Considerando nuestro objetivo de reducir los falsos negativos y analizando cada uno de los umbrales anteriores, concluimos que el mejor umbral es del 22.5%:

a. Cuando el modelo predice un IMPAGO (Valor 1), tiene una **precision = 83%** y un **recall = 88%**, es decir solo un 12% de impagos no se están detectando.
b. Cuando el modelo predice un NO IMPAGO (Valor 0), hay un 99% de probabilidad que el modelo acierte (**accuracy = 99%**).
c. El modelo tiene un accuracy del 98%, lo cual garantiza la fiabilidad del modelo Random Forest.
d. Finalmente, la reducción de los falsos negativos tiene un impacto en la reducción de clientes morosos, a cambio perdemos posibles clientes, pero lo compensamos con un mayor número de buenos clientes. Sin embargo, podríamos ser más estrictos y definir un umbral aún más bajo, lo cual causaría una predicción mayor de IMPAGOS (1), y sí, eso sería correr menos riesgos, pero a cambio de perder muchos buenos clientes (NO IMPAGOS).

14. Output del Proyecto

El modelo predictivo se implementará a través de una aplicación web, que se desarrollará en la plataforma Net Core con C# y se publicará en la nube usando el proveedor de servicios AZURE.

La aplicación web mostrará una pantalla donde el usuario ingresará sus datos personales, monto del crédito, tipo de crédito y otros datos; luego de procesar su solicitud se mostrará el resultado donde se indicará si califica o no al crédito, tasa de interés, número de cuotas, etc.

También se está considerando la posibilidad, en caso no califique al monto que solicita, indicarle el máximo monto al que sí aplicaría, conjuntamente con el número de cuotas y otros datos importantes para el cliente.

15. Conclusiones

15.1. Outputs:

15.1.1. Soluciones Planteadas y Objetivos Conseguibles (Explicación del modelo propuesto en la empresa a aplicar en producción):

El abanico de potenciales soluciones que podemos desarrollar en base a las predicciones generadas por nuestro modelo es bastante amplio. Dependiendo del objetivo que queremos conseguir y del público objetivo a quién queremos enfocarnos podemos algunas de las siguientes soluciones:

- Una aplicación (web o móvil) enfocada a los clientes de una entidad financiera que ofrezca a los usuarios la posibilidad de evaluar su solicitud de préstamo, así como consejos en base a las diferentes variables que nuestro modelo predictivo tiene en cuenta en su predicción, con el objetivo de optimizar al máximo su solicitud de préstamo minimizando así el riesgo de un rechazo por parte de las entidades financieras.
- Otra solución es una aplicación (web o móvil) enfocada a los departamentos de riesgo de las entidades financieras cuyo objetivo es servir de soporte en su evaluación del nivel de riesgo de impago de una solicitud de préstamo. También puede ser usada para ajustar el porcentaje de interés a cada solicitud en base al riesgo de las mismas de esta forma las entidades pueden ofrecer la opción de interés dinámico dependiendo del riesgo de cada solicitud.

Aunque las soluciones anteriores pueden resolver cuestiones específicas, todas ellas tienen en su base las consultas al modelo mediante una API donde se especifican los valores de una solicitud y devuelve como resultado el porcentaje de impago en cuestión de segundos.

Esto abre infinidad de posibles aplicaciones reales desde la integración de los propios sistemas de las entidades financieras hasta consultas de riesgo de impago de micro-préstamos y financiamiento de compras de bajo valor.

15.1.2. Inversión y Retorno:

La implementación del proyecto es económicamente viable porque los indicadores económicos calculados lo demuestran:

Tabla N° 11: Resumen de Indicadores Económicos

Indicador	Resultado
VAN	6,592.87 euros
TIR	55.77% > Tasa mínima de rentabilidad exigida a la inversión (20%)
B/C	1.56 > Unidad
Payback	0.64 años (7 meses y 20 días)

Fuente: elaboración propia

15.2. Aplicación Real:

15.2.1. Descripción de la aplicación:

En nuestro caso, hemos optado por desarrollar la segunda opción, una aplicación web, que toma como entrada valores de las variables más importantes para nuestro modelo y posteriormente consulta a una API (desarrollada en Python) que nos devuelve la probabilidad de impago de ese cliente como un valor continuo (porcentaje) (https://www.heroku.com/, s.f.). La aplicación le hemos dado el nombre de **"Credit App"** y puede ser accesible desde el siguiente enlace: https://credit-app-tfm.herokuapp.com/

Imagen: Aplicación Credit App, aplicación predictiva de riesgo de impago de un préstamo crediticio

Credit App

Modelo predictivo que evalúa el riesgo de impago de un préstamo bancario, por favor indique la siguiente información para hacer una nueva predicción.

Datos del crédito

Probabilidad en solicitudes anteriores

0,094545

Probabilidad en solicitudes en otras entidades

0,083359

Probabilidad de fuente externa

0,503761

Días de antelación en el que cliente cambió de teléfono

-963

Días de antelación en que el cliente cambió su DNI

-2994

Días de antelación en que el cliente cambió su registro

-4985

Anterioridad en el trabajo (en días)

63830

Edad del cliente en días al momento de la solicitud

-16038

Población de la región donde vive

673551

Total del patrimonio del solicitante

5382551

Couta anual del préstamo

27125

Cantidad Solicitada

599575

Total de ingresos

168641

Ver predicción

Fuente: elaboración propia

15.2.2. Tecnología usada para el desarrollo y puesta en producción de la aplicación:

Para el desarrollo de esta aplicación, se ha usado principalmente el lenguaje de programación Python con apoyo de los siguientes lenguajes y librerías:

- HTML + Bootstrap, para el desarrollo de la página inicial.
- Flask: Librería de Python que nos permite crear una web a partir de Python.
- Joblib: Librería de Python que nos permite leer el modelo .joblib generado en el Colab Notebook.
- Gunicorn: Librería de Python que nos permite desplegar la aplicación en producción.

Para poner en producción la aplicación desarrollada, hemos optado por usar la versión gratuita de Heroku, una plataforma en Cloud que nos ofrece un espacio gratuito donde alojar nuestra aplicación para que sea accesible vía Internet. Obviamente esta solución es simplemente para este trabajo ya que si se tratara de una solución industrial lo lógico es tener un servidor más profesional para su puesta en marcha.

15.2.3. Entrada de los datos para la predicción (Input):

Esta aplicación que se propone tiene la entrada en forma de un formulario donde los campos a solicitar corresponden con las variables más importantes que nos ha arrojado nuestro modelo de Random Forest durante el proceso de desarrollo del modelo y que son las siguientes.

- Probabilidad en solicitudes anteriores
- Probabilidad en solicitudes en otras entidades
- Probabilidad de fuente externa
- Días de antelación en el que cliente cambió de teléfono
- Días de antelación en que el cliente cambió su DNI
- Días de antelación en que el cliente cambió su registro
- Anterioridad en el trabajo (en días)
- Edad del cliente en días al momento de la solicitud
- Población de la región donde vive
- Total del patrimonio del solicitante
- Cuota anual del préstamo
- Cantidad Solicitada (monto del crédito)
- Total de ingresos del solicitante

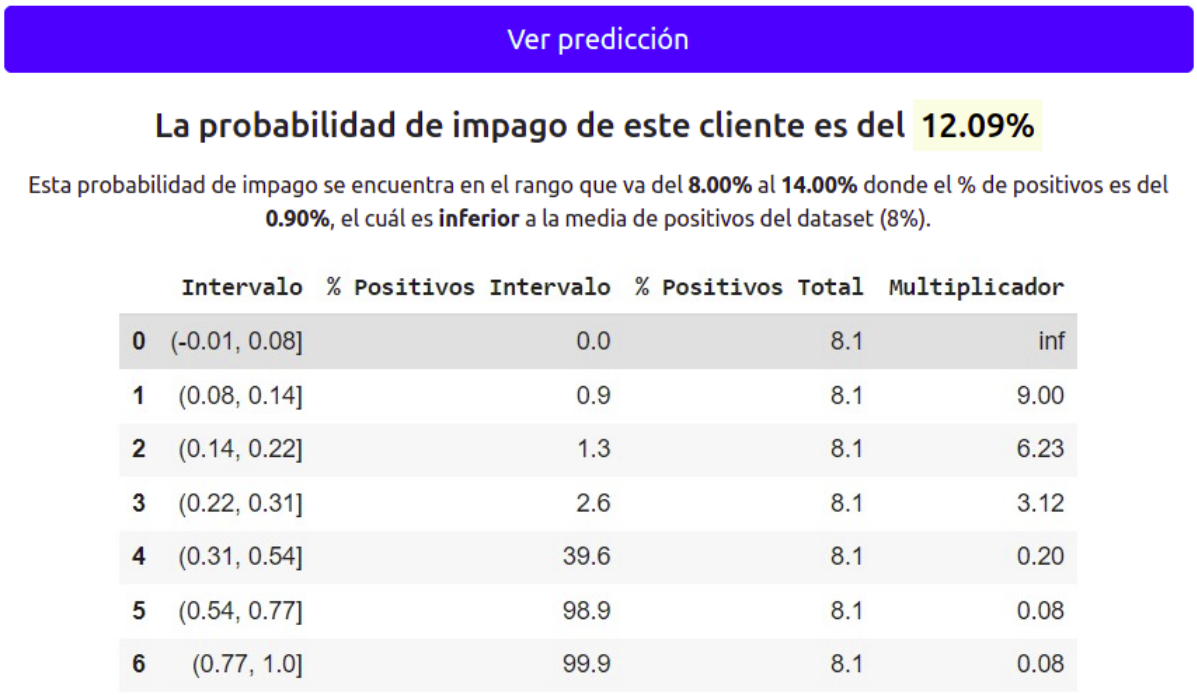
Por defecto, estos campos vienen rellenados con los valores medios del dataset de cada una de las variables, esto nos sirve para ilustrar el tipo de dato que se espera recibir en cada variable.

15.2.4. Salida de la predicción (Output):

Una vez metidos los datos de la solicitud que queremos evaluar, la aplicación nos arroja la predicción de impago de ese cliente en un porcentaje que va de 0% a 100% donde el 0% representa que no hay ningún riesgo de impago por parte del cliente y 100% representa que con toda probabilidad el riesgo de impago es máximo.

También incluye una tabla de rangos donde se puede ver la distribución de los porcentajes de los positivos con respecto a la media (8%) de esta forma el usuario sabrá tendrá una comparativa del resultado arrojado por la aplicación.

Imagen: Aplicación Credit App: output.



	Intervalo	% Positivos Intervalo	% Positivos Total	Multiplicador
0	(-0.01, 0.08]	0.0	8.1	inf
1	(0.08, 0.14]	0.9	8.1	9.00
2	(0.14, 0.22]	1.3	8.1	6.23
3	(0.22, 0.31]	2.6	8.1	3.12
4	(0.31, 0.54]	39.6	8.1	0.20
5	(0.54, 0.77]	98.9	8.1	0.08
6	(0.77, 1.0]	99.9	8.1	0.08

Fuente: elaboración propia

15.3. Una Visualización y Explicación a modo de resumen para presentar a un Directivo

15.3.1. El problema:

El problema principal se basa en los criterios de evaluación y proceso para el otorgamiento de un crédito financiero y lo complejo que puede llegar a convertirse este proceso.

Entre los criterios que toma en cuenta una entidad financiera se encuentra el perfil completo del solicitante del crédito, plenamente identificado y que incluye su información de riesgo. Esta última se encuentra directamente relacionada con los créditos anteriores o su historial de pagos, entre los principales criterios.

Se incluye también el propósito de la solicitud, acotado por el tipo de producto ofrecido por la entidad financiera. La situación económica del solicitante es importante para determinar si es posible que incumpla con sus pagos a lo cual se ha definido como IMPAGO.

En general, las entidades financieras necesitan saber a quién están por darle un crédito, y deben establecer los procedimientos para obtener la información necesaria que les permita construir un perfil del solicitante. El acceso a centrales de riesgo, por ejemplo, es clave para conocer el historial crediticio del solicitante y su verdadero nivel de riesgo.

Acceder a toda esta información y evaluarla con procedimientos convencionales puede requerir mucho análisis y tiempo por parte de los ejecutivos de créditos y al mismo tiempo esto puede recaer en errores humanos o subjetividad en el análisis. (www.esan.edu.pe, 2016)

El riesgo de crédito es el más 'sistémico' de todos, es decir, el que más problemas puede crear en la economía entera, por la máxima interconectividad con otros jugadores y el sector real. Hay un riesgo de lo que se llama la "ruptura de cadena de pagos" y de "contagio", por la falta de confianza generalizada que puede crear. (Belaunde, 2012)

15.3.2. Solución tradicional:

Los Departamentos de Riesgos Crediticios son, normalmente, los que se encargan de evaluar los posibles créditos y minimizar los riesgos para permitir una buena rentabilidad. Aplican diferentes soluciones al problema y se presentan algunas de ellas:

Evaluación de crédito tradicional

- Crear sistemas estándares de evaluación de créditos
- Realizar estudios de segmento
- Detectar aquellos créditos con riesgos superior a lo normal para hacerles seguimiento mas minucioso
- Preparar un sin numero de análisis para futuros ejecutivos de cuentas
- Realizar estudios sectoriales
- Riesgo de la cartera
- Riesgo por cliente
- Posición respecto al destino
- Requerimiento de información

Toda esta evaluación normalmente se realiza de forma manual, con diferentes fuentes de información y creación de Excel y algunos análisis estáticos y automatizados dentro de estas herramientas tradicionales (www.esan.edu.pe, 2016)

15.3.3. Solución propuesta:

La solución propuesta se trata de un Modelo Predictivo de Aprendizaje Automático que considere todos los criterios anteriormente mencionados, es decir, el perfil completo del solicitante del crédito, plenamente identificado y que incluye su información de riesgo. Sus créditos anteriores y su historial de pagos. Sus tarjetas de crédito y compras. Prestamos en efectivo y créditos previos con la entidad que evalúa el crédito. La información de buros de crédito que incluya información de créditos previos en cualquier otra entidad financiera.

Con todos estos datos, se crea y entrena un modelo capaz de aprender y predecir la probabilidad de impago de una persona o entidad que quiera se acreedora de un crédito financiero.

De esta manera se propone que, a través de un Modelo Predictivo de Aprendizaje Automático, las entidades financieras apliquen Big Data y Data Science para ahorrar muchas horas hombre, procesos tradicionales morosos y subjetivos, para obtener un procedimiento de evaluación de riesgo crediticio mucho mas eficiente y objetivo para así otorgar créditos de manera efectiva, con riesgos controlados y un menor tiempo posible para la satisfacción de los clientes.

15.3.4. Proceso de solución:

El modelo predictivo final recibe diferentes variables con las cuales es capaz de predecir si una persona que quiere obtener un crédito financiero, impagará esa deuda adquirida en base a datos históricos personales y financieros. El modelo evalúa si el individuo es un buen prospecto de recibir un crédito (préstamo de dinero) y para definir aquello es necesario evaluar si esta persona será capaz de pagar ese crédito o incumplirá su deuda e impagará el crédito. La forma en la que este modelo predice es devolviendo como salida una probabilidad del 0 al 1 de forma continua donde 0 no impaga y 1 impaga. Además, el modelo presenta una tabla de multiplicadores donde se muestra dentro de que rango se encuentra la probabilidad de impago de un cliente y cual es el porcentaje de impagos que tiene ese rango de probabilidades. Algo muy útil para la evaluación final por parte del ejecutivo de créditos.

15.3.5. Muestra de resultados:

Se presentan 5 predicciones con diferentes características y cambios en los criterios más relevantes que pueden afectar en la probabilidad de impago de un crédito:

- **Predicción con buen historial crediticio:** Esta predicción se realizó basándonos en buena calificación de probabilidad de solicitudes de crédito anteriores y en otras entidades. Estas dos variables son las más significativas del modelo.
    a. Probabilidad de solicitudes anteriores: 0.02
    b. Probabilidad de solicitudes en otras entidades (Probabilidad de impago según buros de crédito): 0.04

Datos del crédito

Probabilidad en solicitudes anteriores: 0.04

Probabilidad en solicitudes en otras entidades: 0.02

Probabilidad de fuente externa: 0.503761

Días de antelación en el que cliente cambió de teléfono: -963

Días de antelación en que el cliente cambió su DNI: -2994

Días de antelación en que el cliente cambió su registro: -4985

Anterioridad en el trabajo (en días): 63830

Edad del cliente en días al momento de la solicitud: -16038

Población de la región donde vive: 673551

Total del patrimonio del solicitante: 5382551

Couta anual del préstamo: 27125

Cantidad Solicitada: 599575

Total de ingresos: 168641

Ver predicción

La probabilidad de impago de este cliente es del 6.00%

Esta probabilidad de impago se encuentra en el rango que va del **0.00%** al **8.00%** donde el % de positivos es del **0.00%**, el cuál es **inferior** a la media de positivos del dataset (8%).

	Intervalo	% Positivos Intervalo	% Positivos Total	Multiplicador
0	(-0.01, 0.08]	0.0	8.1	inf
1	(0.08, 0.14]	0.9	8.1	9.00
2	(0.14, 0.22]	1.3	8.1	6.23
3	(0.22, 0.31]	2.6	8.1	3.12
4	(0.31, 0.54]	39.6	8.1	0.20
5	(0.54, 0.77]	98.9	8.1	0.08
6	(0.77, 1.0]	99.9	8.1	0.08

Fuente: elaboración propia

En esta predicción podemos observar cuan importantes son las probabilidades de solicitudes anteriores y de los buros crediticios. Gracias a que tenemos probabilidades bajas en ellos obtenemos como resultado un 6% de probabilidad de impago con estos datos. También podemos observar que esa probabilidad cae dentro del rango del primer intervalo donde según el análisis de datos el porcentaje de impagos es 0%.

- **Predicción con historial crediticio medio y fuente externa: con calificación regular:** Esta predicción se realizo basándonos una calificación de probabilidad de solicitudes de crédito anteriores y en otras entidades media (ni buena ni mala) y la probabilidad de una fuente externa con una calificación regular. La variable de fuente externa es la 3ra variable mas significativa del modelo.
    c. Probabilidad de solicitudes anteriores: 0.094545
    d. Probabilidad de solicitudes en otras entidades (Probabilidad de impago según buros de crédito): 0.083359
    e. Probabilidad de fuente externa: 0.503761

Credit App

Modelo predictivo que evalúa el riesgo de impago de un préstamo bancario, por favor indique la siguiente información para hacer una nueva predicción.

Datos del crédito

Probabilidad en solicitudes anteriores

0.094545

Probabilidad en solicitudes en otras entidades

0.083359

Probabilidad de fuente externa

0.503761

Días de antelación en el que cliente cambió de teléfono

-963

Días de antelación en que el cliente cambió su DNI

-2994

Días de antelación en que el cliente cambió su registro

-4985

Anterioridad en el trabajo (en días)

63830

Edad del cliente en días al momento de la solicitud

-16038

Población de la región donde vive

673551

Total del patrimonio del solicitante

5382551

Couta anual del préstamo

27125

Cantidad Solicitada

599575

Total de ingresos

168641

Ver predicción

La probabilidad de impago de este cliente es del 12.09%

Esta probabilidad de impago se encuentra en el rango que va del **8.00%** al **14.00%** donde el % de positivos es del **0.90%**, el cuál es **inferior** a la media de positivos del dataset (8%).

	Intervalo	% Positivos Intervalo	% Positivos Total	Multiplicador
0	(-0.01, 0.08]	0.0	8.1	inf
1	(0.08, 0.14]	0.9	8.1	9.00
2	(0.14, 0.22]	1.3	8.1	6.23
3	(0.22, 0.31]	2.6	8.1	3.12
4	(0.31, 0.54]	39.6	8.1	0.20
5	(0.54, 0.77]	98.9	8.1	0.08
6	(0.77, 1.0]	99.9	8.1	0.08

Fuente: elaboración propia

En esta predicción podemos observar cómo afecta la probabilidad de la fuente externa cuando nuestras primeras dos variables son la media. Obtenemos una probabilidad de impago superior a la anterior que cae dentro del rango del segundo intervalo donde según el análisis de datos el porcentaje de impagos es 0.9% que sigue siendo por debajo de la media de imagadores.

- **Predicción media con antigüedad en el trabajo baja y personas mayores:** Esta predicción se realizó basándonos una media de las primeras variables significativas, poca antigüedad laboral y persona de la 3ra edad.
    f. Antigüedad de trabajo: 90 días
    g. Edad del cliente: 21900 días (60 años)

Credit App

Modelo predictivo que evalúa el riesgo de impago de un préstamo bancario, por favor indique la siguiente información para hacer una nueva predicción.

Datos del crédito

Probabilidad en solicitudes anteriores: 0.094545

Probabilidad en solicitudes en otras entidades: 0.083359

Probabilidad de fuente externa: 0.503761

Días de antelación en el que cliente cambió de teléfono: -963

Días de antelación en que el cliente cambió su DNI: -2994

Días de antelación en que el cliente cambió su registro: -4985

Anterioridad en el trabajo (en días): 90

Edad del cliente en días al momento de la solicitud: -21900

Población de la región donde vive: 673551

Total del patrimonio del solicitante: 5382551

Couta anual del préstamo: 27125

Cantidad Solicitada: 599575

Total de ingresos: 168641

Ver predicción

La probabilidad de impago de este cliente es del 18.00%

Esta probabilidad de impago se encuentra en el rango que va del **14.00%** al **22.00%** donde el % de positivos es del **1.30%**, el cuál es **inferior** a la media de positivos del dataset (8%).

	Intervalo	% Positivos Intervalo	% Positivos Total	Multiplicador
0	(-0.01, 0.08]	0.0	8.1	inf
1	(0.08, 0.14]	0.9	8.1	9.00
2	(0.14, 0.22]	1.3	8.1	6.23
3	(0.22, 0.31]	2.6	8.1	3.12
4	(0.31, 0.54]	39.6	8.1	0.20
5	(0.54, 0.77]	98.9	8.1	0.08
6	(0.77, 1.0]	99.9	8.1	0.08

Fuente: elaboración propia

En esta predicción podemos observar como afecta la antigüedad laboral y la edad del individuo. Obtenemos una probabilidad de impago superior a la anterior que cae dentro del rango del tercer intervalo donde según el análisis de datos el porcentaje de impagos es 1.3% que ya puede significar un riesgo considerable dependiendo el tipo de créditos que se solicita.

- **Predicción basada en la anterior, con historial crediticio media baja (peor que la media):** Esta predicción se realizó basándonos en la anterior, pero observando cómo afecta las variables significativas
    h. Probabilidad de solicitudes anteriores: 0.15
    i. Probabilidad de solicitudes en otras entidades (Probabilidad de impago según buros de crédito): 0.20
    j. Antigüedad de trabajo: 90 días
    k. Edad del cliente: 21900 días (60 años)

Credit App

Modelo predictivo que evalúa el riesgo de impago de un préstamo bancario, por favor indique la siguiente información para hacer una nueva predicción.

Datos del crédito

Probabilidad en solicitudes anteriores
0.15

Probabilidad en solicitudes en otras entidades
0.20

Probabilidad de fuente externa
0.503761

Días de antelación en el que cliente cambió de teléfono
-963

Días de antelación en que el cliente cambió su DNI
-2994

Días de antelación en que el cliente cambió su registro
-4985

Anterioridad en el trabajo (en días)
90

Edad del cliente en días al momento de la solicitud
-21900

Población de la región donde vive
673551

Total del patrimonio del solicitante
5382551

Couta anual del préstamo
27125

Cantidad Solicitada
599575

Total de ingresos
1686

Ver predicción

La probabilidad de impago de este cliente es del 50.00%

Esta probabilidad de impago se encuentra en el rango que va del **31.00%** al **54.00%** donde el % de positivos es del **39.60%**, el cuál es **superior** a la media de positivos del dataset (8%).

	Intervalo	% Positivos Intervalo	% Positivos Total	Multiplicador
0	(-0.01, 0.08]	0.0	8.1	inf
1	(0.08, 0.14]	0.9	8.1	9.00
2	(0.14, 0.22]	1.3	8.1	6.23
3	(0.22, 0.31]	2.6	8.1	3.12
4	(0.31, 0.54]	39.6	8.1	0.20
5	(0.54, 0.77]	98.9	8.1	0.08
6	(0.77, 1.0]	99.9	8.1	0.08

Fuente: elaboración propia

En esta predicción podemos observar como afecta la probabilidad de impago en solicitudes anteriores y otras entidades de manera considerable a un individuo que contaba con variables medianamente significativas calificadas como moderadamente riesgosas y lo convierte en un individuo con alta probabilidad de impago. Además, observamos nuestros rangos de probabilidades y se observa que cae dentro del rango del quinto intervalo donde según el análisis de datos el porcentaje de impagos es 39.6% que ya significa un riesgo alto de impago.

- **Predicción basada en la anterior, con un monto dentro del rango con más % de impago: E**sta predicción se realizó basándonos en la anterior, pero introduciendo un monto de crédito dentro del rango de montos que más porcentaje de impago según el análisis comparativo de variables significativas vs % de impagos
    l. Probabilidad de solicitudes anteriores: 0.20
    m. Probabilidad de solicitudes en otras entidades (Probabilidad de impago según buros de crédito): 0.20
    n. Antigüedad de trabajo: 90 días
    o. Edad del cliente: 21900 días (60 años)
    p. Monto del crédito: 830000 (Rango de montos con mas % impagadores)

Credit App

Modelo predictivo que evalúa el riesgo de impago de un préstamo bancario, por favor indique la siguiente información para hacer una nueva predicción.

Datos del crédito

Probabilidad en solicitudes anteriores: 0.20

Probabilidad en solicitudes en otras entidades: 0.20

Probabilidad de fuente externa: 0.503761

Días de antelación en el que cliente cambió de teléfono: -963

Días de antelación en que el cliente cambió su DNI: -2994

Días de antelación en que el cliente cambió su registro: -4985

Anterioridad en el trabajo (en días): 90

Edad del cliente en días al momento de la solicitud: -21900

Población de la región donde vive: 673551

Total del patrimonio del solicitante: 5382551

Couta anual del préstamo: 27125

Cantidad Solicitada: 830000

Total de ingresos: 1686

Ver predicción

La probabilidad de impago de este cliente es del 66.00%

Esta probabilidad de impago se encuentra en el rango que va del **54.00%** al **77.00%** donde el % de positivos es del **98.90%**, el cuál es **superior** a la media de positivos del dataset (8%).

	Intervalo	% Positivos Intervalo	% Positivos Total	Multiplicador
0	(-0.01, 0.08]	0.0	8.1	inf
1	(0.08, 0.14]	0.9	8.1	9.00
2	(0.14, 0.22]	1.3	8.1	6.23
3	(0.22, 0.31]	2.6	8.1	3.12
4	(0.31, 0.54]	39.6	8.1	0.20
5	(0.54, 0.77]	98.9	8.1	0.08
6	(0.77, 1.0]	99.9	8.1	0.08

Fuente: elaboración propia

En esta predicción podemos observar como un individuo que, con variables de mala calificación, y una variable moderadamente significativa (monto del crédito) dentro un rango donde el porcentaje de impagadores es alto, lo convierte en un individuo con una probabilidad de impago prácticamente inaceptable. Según el análisis los créditos entre 40,000 y 849.000 son montos donde los clientes incurren en un porcentaje más alto de impago. En el rango de probabilidades se observa que cae dentro del rango del sexto intervalo donde según el análisis de datos el porcentaje de impagos es 98.9% riesgo posiblemente inaceptable. Curiosamente, el análisis de impagos vs montos de crédito nos muestra que montos mayores a ese rango tienen menos porcentaje de impagos.

15.4. Reflexión Final sobre problemas encontrados y soluciones llevadas a cabo:

15.4.1. Problemas y soluciones:

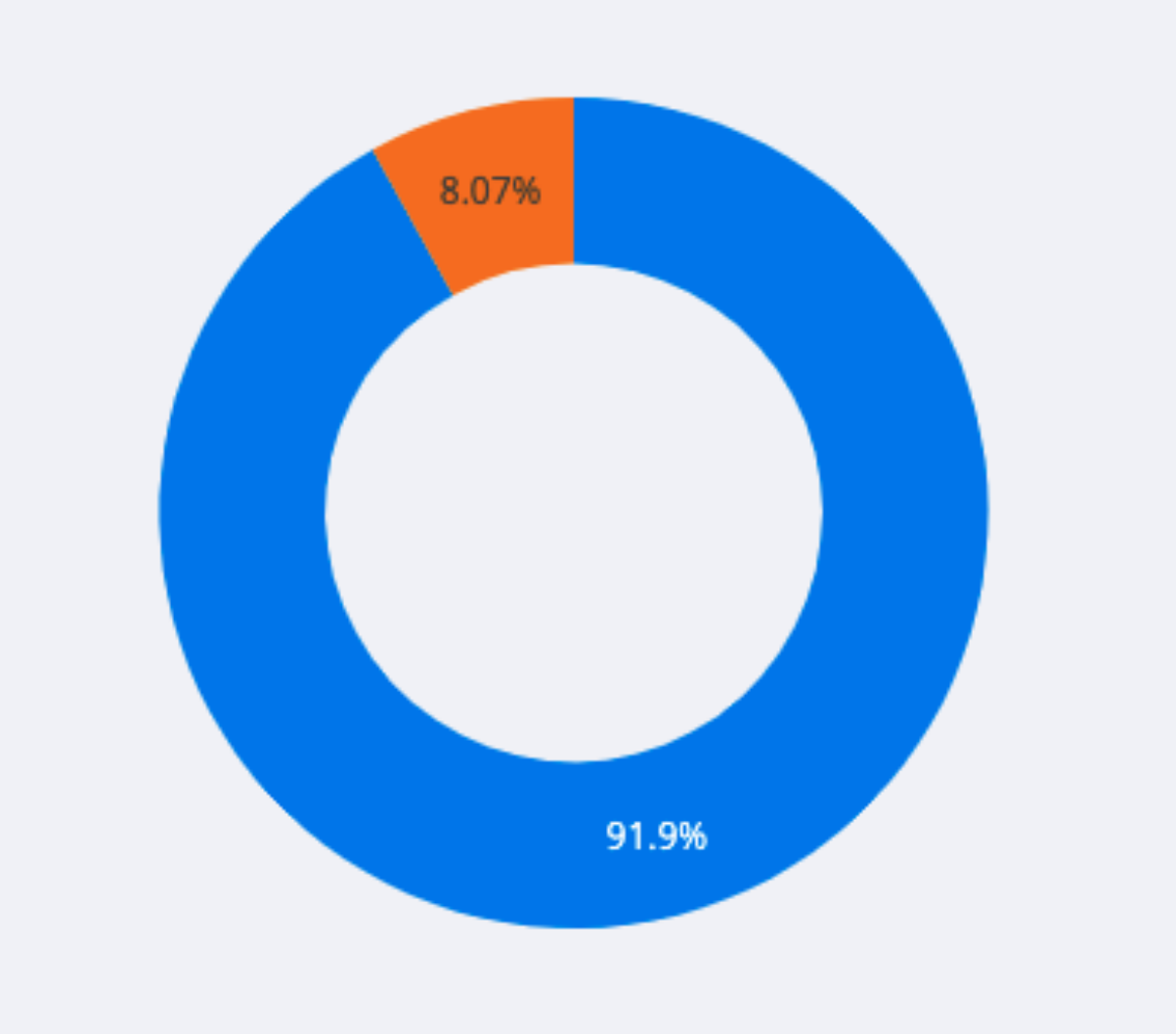
Se presentaron dificultades de diferentes características que fueron las siguientes:

Recolección de datos: Se tuvo dificultades en obtener los datos y la cantidad de datos necesaria para construir el modelo ya que estos datos financieros y personales son de alta confidencialidad y las entidades financieras y buros de crédito tienen políticas estrictas en cuanto a compartir esta información. Para ello se realizó una búsqueda exhaustiva de datos referidos, se compararon algunos datasets encontrados y finalmente llegamos a Home Credit que publicó un conjunto de datasets con la cantidad y calidad necesaria de información a través de Kaggle con diferentes fuentes de información y mucha data histórica (2.7 GB) (Group, 2018)

Datos faltantes: Se ha encontrado varios valores faltantes en diferentes columnas, tanto del dataset principal como en los datasets históricos. Esto se debe a que muchos individuos no contaban con historial de tarjetas de crédito u otra información. Para este problema aplicamos técnicas de imputación de datos según el caso.

Reducción de variables: El conjunto de datasets incluyendo los datos históricos sumaba una cantidad muy interesante de variables; sin embargo, no todas aportaban al modelo por lo cual también se aplicó técnicas de reducción de variables según su aporte e importancia en el modelo.

Dataset desbalanceado: Uno de los problemas más interesantes del proyecto fue que el dataset era un dataset desbalanceado; donde el 92% de los casos eran personas que no incurrían en impago y el 8% personas que impagaron un crédito. Eso significa que no se podía usar la precisión como parámetro de validación del modelo. En cambio, la solución para este problema fue basarnos en un umbral aceptable para el modelo y técnicas de validación como curva ROC, AUC, matrices de confusión, evaluación del recall y finalmente el coeficiente Kappa que nos mostraba la concordancia de datos.



Azul = 0; Naranja = 1 - Fuente: elaboración propia

Output del modelo: Posteriormente a superar todas las anteriores dificultades, también se nos presentó el problema de como presentar los resultados del modelo. Para ello se decidió utilizar el servicio de Heroku donde, a través de un aplicativo web, montamos la aplicación para poder

ser utilizada por nuestros usuarios finales y predecir de manera sencilla la probabilidad de impago de diferentes individuos.

Credit App

Modelo predictivo que evalúa el riesgo de impago de un préstamo bancario, por favor indique la siguiente información para hacer una nueva predicción.

Datos del crédito

Probabilidad en solicitudes anteriores: 0.20

Probabilidad en solicitudes en otras entidades: 0.20

Probabilidad de fuente externa: 0.503761

Días de antelación en el que cliente cambió de teléfono: -963

Días de antelación en que el cliente cambió su DNI: -2994

Días de antelación en que el cliente cambió su registro: -4985

Anterioridad en el trabajo (en días): 90

Edad del cliente en días al momento de la solicitud: -21900

Población de la región donde vive: 673551

Total del patrimonio del solicitante: 5382551

Couta anual del préstamo: 27125

Cantidad Solicitada: 830000

Total de ingresos: 1686

Ver predicción

La probabilidad de impago de este cliente es del 66.00%

Esta probabilidad de impago se encuentra en el rango que va del **54.00%** al **77.00%** donde el % de positivos es del **98.90%**, el cuál es **superior** a la media de positivos del dataset (8%).

	Intervalo	% Positivos Intervalo	% Positivos Total	Multiplicador
0	(-0.01, 0.08]	0.0	8.1	inf
1	(0.08, 0.14]	0.9	8.1	9.00
2	(0.14, 0.22]	1.3	8.1	6.23
3	(0.22, 0.31]	2.6	8.1	3.12
4	(0.31, 0.54]	39.6	8.1	0.20
5	(0.54, 0.77]	98.9	8.1	0.08
6	(0.77, 1.0]	99.9	8.1	0.08

Fuente: elaboración propia

15.5. Conclusiones:

Después de la construcción de dos modelos de diferentes características y metodología, como el GLM y Random Forest, hemos logrado una clasificación binaria relativamente aceptable con GLM y una muy buena clasificación con Random Forest. Las métricas consideradas fueron Recall, curva ROC, AUC, matrices de confusión y coeficientes Kappa para la evaluación, validación y comparación de modelos ya que se contaba con un dataset desbalanceado y no se podía considerar el accuracy de los modelos como medida.

Es crucial primeramente la recolección de datos, exploración de los datos y finalmente la limpieza de datos para proceder de forma correcta con la creación del modelo predictivo. Añadido a eso, es importante considerar que un proyecto de Machine Learning debe contar con diferentes etapas para crear un ambiente productivo completo; entre ellas las más importantes como el desarrollo del modelo, pruebas, validaciones, monitoreo, entrenamiento automatizado, versionamiento, actualizaciones, outputs y puesta en producción.

El modelo predictivo logra evaluar muchos criterios importantes para medir el nivel de riesgo de impago de un crédito bancario y clasifica, según una probabilidad, el riesgo de impago de este. También es importante considerar que existen otros criterios como políticos, de cartera, sociales entre otros también deben ser considerados al momento de evaluar el riesgo crediticio; sin embargo, el modelo toma en cuenta una cantidad considerable de criterios personales y financieros que lo hace un modelo que puede ser implementado para convertir este proceso a un proceso rápido, confiable y automatizado que aprenda constantemente a partir de los datos internos y externos con los que pueda contar la entidad que lo aplique.

Concluimos finalmente que el modelo predictivo cumple con el objetivo principal de predecir el riesgo de impago de un crédito de manera eficiente en base a datos históricos y personales de un cliente a través de una interfaz web y es aplicable a un entorno real con datos reales.

Considerando nuestra hipótesis inicial que consiste en reducir la morosidad en una entidad financiera y analizando diferentes umbrales, concluimos que el mejor umbral es del 22.5%, ya que nos permite lograr la reducción de los falsos negativos que tiene un impacto directo en la reducción de clientes morosos, perdiendo potenciales clientes, pero compensándolo con un mayor número de buenos clientes. En otro caso, podríamos ser más restrictivos y definir un umbral aún más bajo, lo cual causaría una predicción mayor de IMPAGOS (1) corriendo menos riesgo, pero reduciendo la cartera de clientes impactando de forma directa en los ingresos económicos de la entidad financiera.

16. Bibliografía

1. Usachev D. https://fayrix.com/blog/machine-learning-in-finance. [Online]. Disponible en: https://fayrix.com/blog/machine-learning-in-finance.

2. Felman M. https://github.com/marcelofelman/case-studies/blob/master/C%C3%B3mo%20predecir%20riesgo%20financiero%20con%20Machine%20Learning.md. [Online]; 2018. Disponible en: https://github.com/marcelofelman/case-studies/blob/master/C%C3%B3mo%20predecir%20riesgo%20financiero%20con%20Machine%20Learning.md.

3. Group HC. www.kaggle.com. [Online]; 2018. Disponible en: https://www.kaggle.com/c/home-credit-default-risk.

4. Ucha AP. https://economipedia.com/. [Online]; 2021. Acceso 6 de Marzode 2021. Disponible en: https://economipedia.com/author/a-peiro.

7. Morales VV. https://economipedia.com. [Online]. Disponible en: https://economipedia.com/definiciones/payback.html.

8. https://www.heroku.com/. [Online]. Disponible en: https://www.heroku.com/.

9. www.esan.edu.pe. https://www.esan.edu.pe/. [Online]; 2016. Disponible en: https://www.esan.edu.pe/apuntes-empresariales/2016/12/criterios-de-evaluacion-para-el-otorgamiento-de-creditos/.

10. Belaunde G. http://blogs.gestion.pe. [Online]; 2012. Disponible en: http://blogs.gestion.pe/riesgosfinancieros/2012/01/gestionar-el-riesgo-de-credito.html.

5. https://www.significados.com. https://www.significados.com/costo-beneficio/. [Online]; 2021. Disponible en: https://www.significados.com/costo-beneficio/.

6. https://monday.com. [Online]; 2021. Disponible en: https://monday.com/lang/es/.