ITESO, Universidad Jesuita de Guadalajara

Proyecto: Modelo de Puntuación crediticia

Equipo:

Mugica Liparoli Juan Antonio

Enriquez Nares Diego Emilio

Brizuela Casarin Ana Sofia

Materia: Modelos de Crédito

Profesor: Rodolfo Slay Ramos

Fecha de Entrega: 24 de Septiembre del 2024

Asignatura: Modelos de Crédito **Fecha:** 09 Septiembre 2024
Docente: Rodolfo Slay Ramos

Resumen

Los modelos de calificación crediticia son herramientas estadísticas que evalúan la solvencia crediticia y determinan la probabilidad de incumplimiento de las obligaciones crediticias. Estos modelos son utilizados por agencias de crédito y los prestamistas para evaluar el riesgo de prestar dinero o extender crédito a personas y/o empresas.

Introducción.

El modelo de puntuación crediticia es una herramienta esencial en el sector financiero, utilizada para evaluar la capacidad y la disposición de los individuos y las empresas para cumplir con sus obligaciones financieras. Este sistema se basa en la recopilación de información cuantitativa y cualitativa de los solicitantes de crédito, a fin de asignar una calificación que refleja el riesgo asociado con otorgarles un crédito.

Los modelos de puntuación crediticia, como el conocido FICO score y otros sistemas personalizados, se basan en diferentes factores que se ponderan para predecir la probabilidad de incumplimiento de pago. Estos factores suelen incluir aspectos como el historial crediticio, el nivel de endeudamiento, la antigüedad de las cuentas crediticias y otros datos financieros. El objetivo de estos modelos es proporcionar una métrica objetiva que permita a las instituciones financieras tomar decisiones más informadas sobre la aprobación o rechazo de solicitudes de crédito.

El desarrollo de un modelo efectivo requiere la identificación de las variables clave que impactan significativamente en la capacidad de pago del solicitante. Además, los pesos asignados a estas variables deben reflejar su importancia relativa en la predicción del riesgo crediticio. El presente proyecto tiene como objetivo construir un modelo de puntuación crediticia que permita clasificar a los solicitantes según su historial crediticio y predecir con precisión el comportamiento futuro en el cumplimiento de sus obligaciones financieras.

Asignatura: Modelos de Crédito **Fecha:** 09 Septiembre 2024

Docente: Rodolfo Slay Ramos

Desarrollo del proyecto:

Clases y funciones realizadas:

1. Clase EDA (Análisis exploratorio de los datos):

Objetivo General: Proporcionar herramientas para realizar un análisis exploratorio de datos, lo cual es crucial para entender las características fundamentales del conjunto de datos y prepararlo para modelado posterior.

Métodos y sus Objetivos Específicos:

- __init__(self, data): Inicializar el objeto `EDA` con el DataFrame proporcionado, asegurándose de que no esté vacío para evitar errores en análisis subsiguientes.

- validate_data(self, data): Confirmar que el DataFrame no esté vacío, garantizando la integridad del análisis.

- data_summary(self): Suministrar un resumen visual y numérico del conjunto de datos para evaluar su estructura y calidad inicialmente.

2. Clase DQR (Data Quality Review)

Objetivo General: Asegurar la calidad de los datos mediante una limpieza detallada y sistemática, preparándolos para análisis predictivos eficaces.

Métodos y sus Objetivos Específicos:

- __init__(self, data): Configurar el objeto `DQR` con el conjunto de datos a limpiar.

- perform_clean(self): Ejecutar una limpieza comprensiva que incluye la estandarización de tipos de datos, manejo de valores atípicos y faltantes, y eliminación de duplicados.

3. Clase CreditScoreModel

Objetivo General: Desarrollar y evaluar un modelo predictivo para el score de crédito utilizando los datos limpios y preparados por las clases anteriores.

Métodos y sus Objetivos Específicos:

- __init__(self, data): Inicializar el modelo con el conjunto de datos ya limpio.

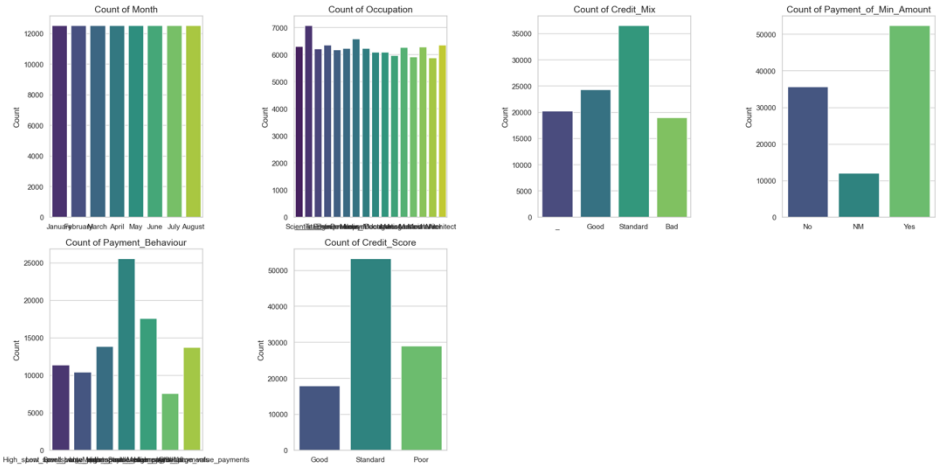
- apply_scoring(self): Implementar un sistema de puntuación basado en varias condiciones específicas que reflejan factores críticos en la determinación del score de crédito. Posteriormente, evaluar la precisión del modelo comparando las predicciones con los scores reales de crédito.

- Implementación: Usa una serie de reglas condicionales para asignar puntajes a los clientes, luego agrupa estos puntajes en categorías y calcula la precisión del modelo.

Limpieza de datos:

Realizamos unas pruebas con nuestra clase EDA donde realizamos diferentes pruebas estadísticas para ver la importancia de una buena limpieza y extracción de los datos. Observamos que podemos encontrar más datos relevantes para hacer nuestro modelo si se realiza un acertado análisis exploratorio de los datos. A continuación, mostraremos medidas estadísticas sin la limpieza de datos en comparación si es que usamos nuestra clase para limpiarlos.

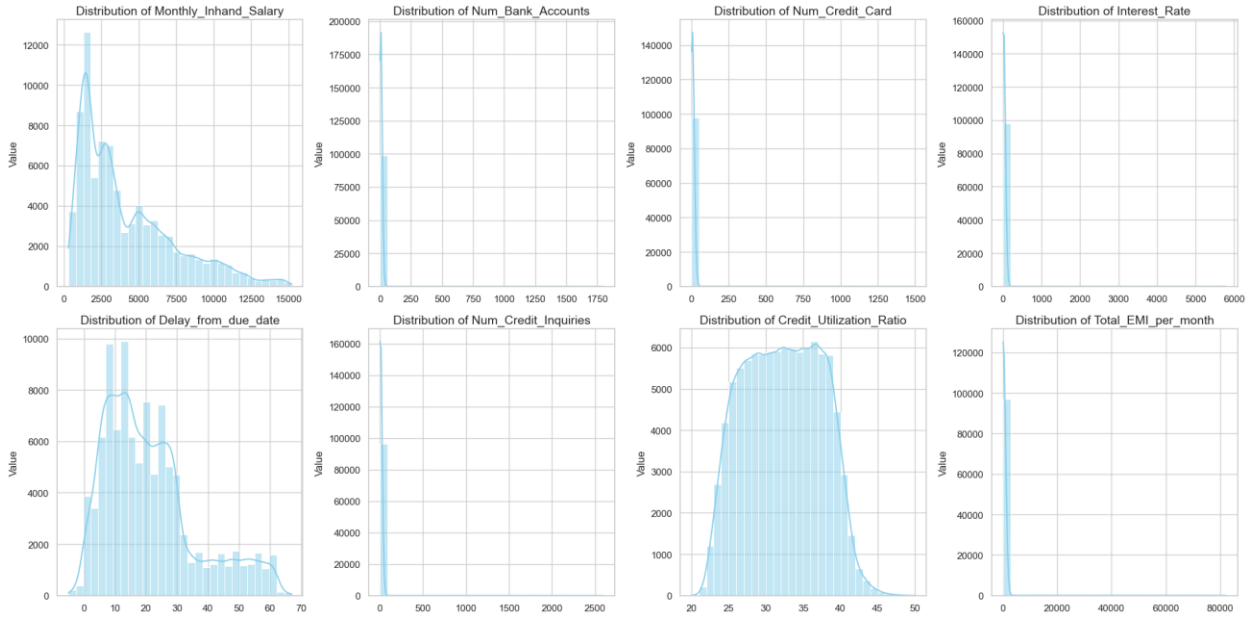
Distribución de categorías del dataset (train-2.csv):



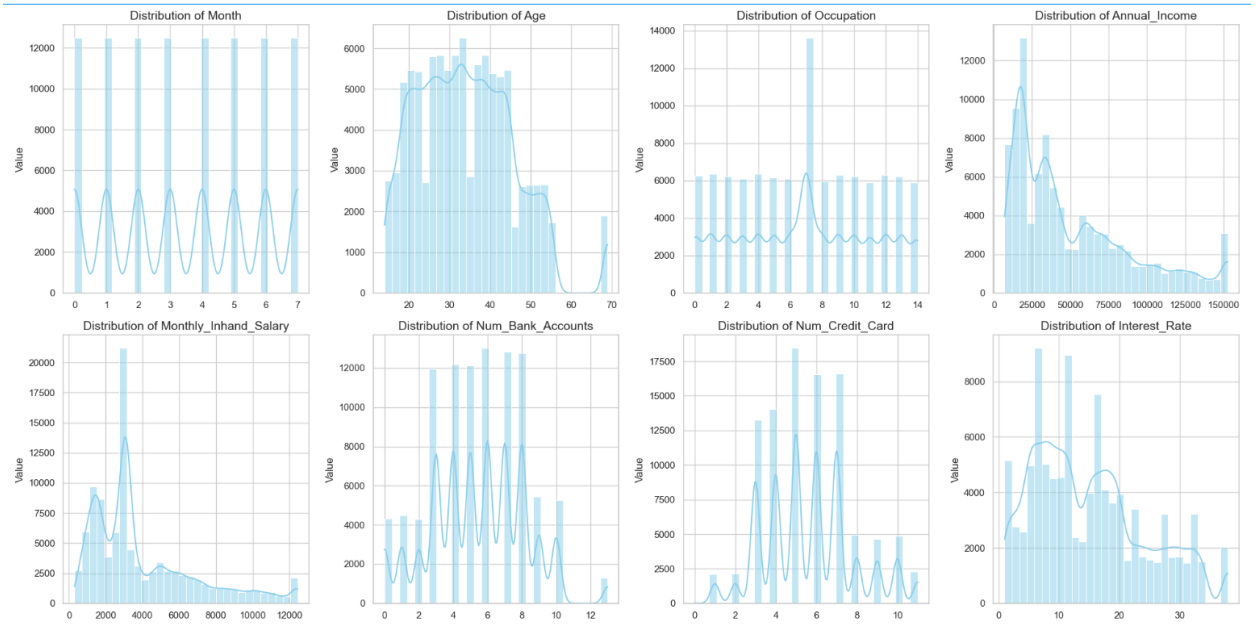
Asignatura: Modelos de Crédito **Fecha:** 09 Septiembre 2024

Docente: Rodolfo Slay Ramos

Distribución del dataset sin limpieza (Sin limpieza en los datos):



Distribución del dataset con limpieza (con limpieza en los datos):



La limpieza de datos es fundamental ya que quita valores atípicos para los diferentes tipos de objetos de nuestro dataset. Convertimos la mayoría de las columnas a numéricas y eso nos permitió tener más histogramas a analizar para tomar mejores decisiones mediante el comportamiento de nuestros datos. Saber si en una variable de nuestros datos tienen una distribución cercana a normal o si tienen la cola más pegada a la izquierda o hacia la derecha, conocer los datos es fundamental

para tomar mejores decisiones de clasificación para un modelo deseado como es nuestro caso de score de crédito.

Diagrama de caja y bigotes (sin limpieza en los datos):

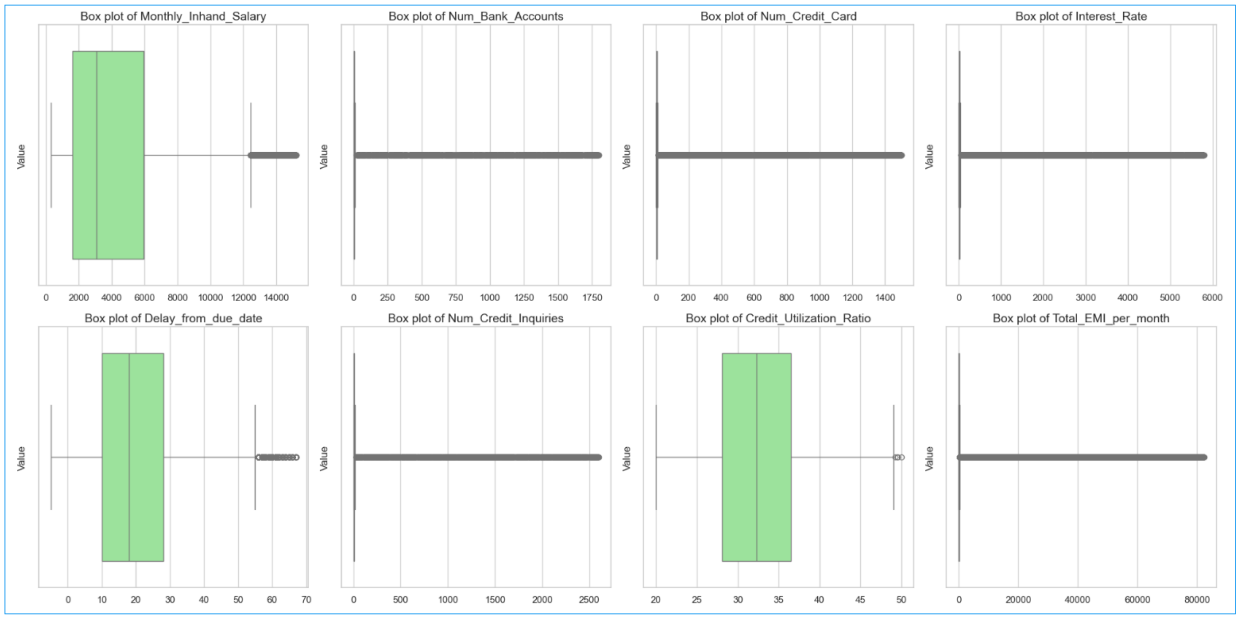
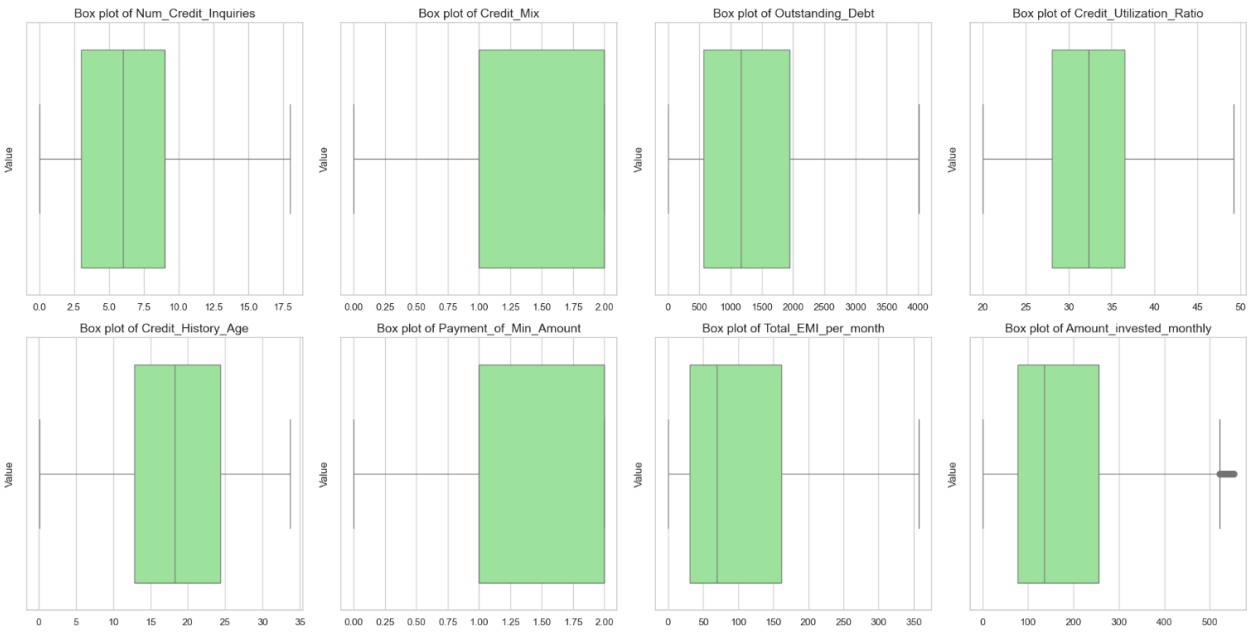


Diagrama de caja y bigotes (con limpieza en los datos):

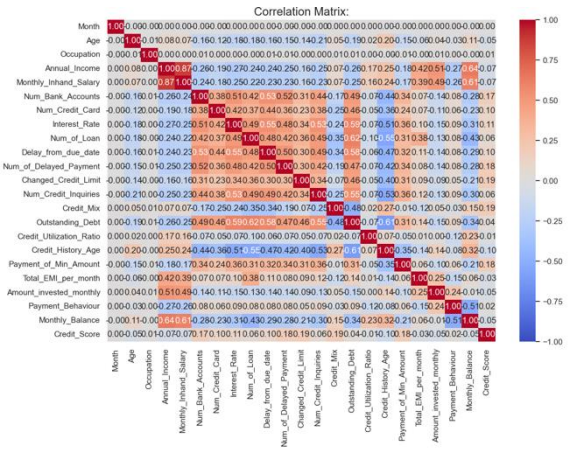


Al igual que con la prueba estadística anterior, vemos que sin limpieza de datos hay pocas categorías que nos pueden mostrar información clave para segmentar los datos.

Asignatura: Modelos de Crédito **Fecha:** 09 Septiembre 2024

Docente: Rodolfo Slay Ramos

Matriz de correlación sin limpieza vs Matriz de correlación con limpieza:



La matriz de correlación nos ayuda a entender que variables tienen más similitud entre sí y eso nos ayuda a escoger de mejor manera las variables a utilizar para calcular el score de crédito de nuestro modelo.

Estrategias basadas en la correlación:

- *Minimizar la multicolinealidad:* Al desarrollar un modelo de scoring de crédito, es deseable seleccionar variables que no solo estén significativamente correlacionadas con la variable objetivo (el score de crédito), sino que también sean lo más independientes posible entre sí para evitar la multicolinealidad.
- *Maximizar la relevancia del modelo:* Por otro lado, deseas incluir variables que tengan una fuerte relación con el score de crédito para asegurarte de que el modelo sea relevante y pueda capturar con precisión los factores que afectan el score.

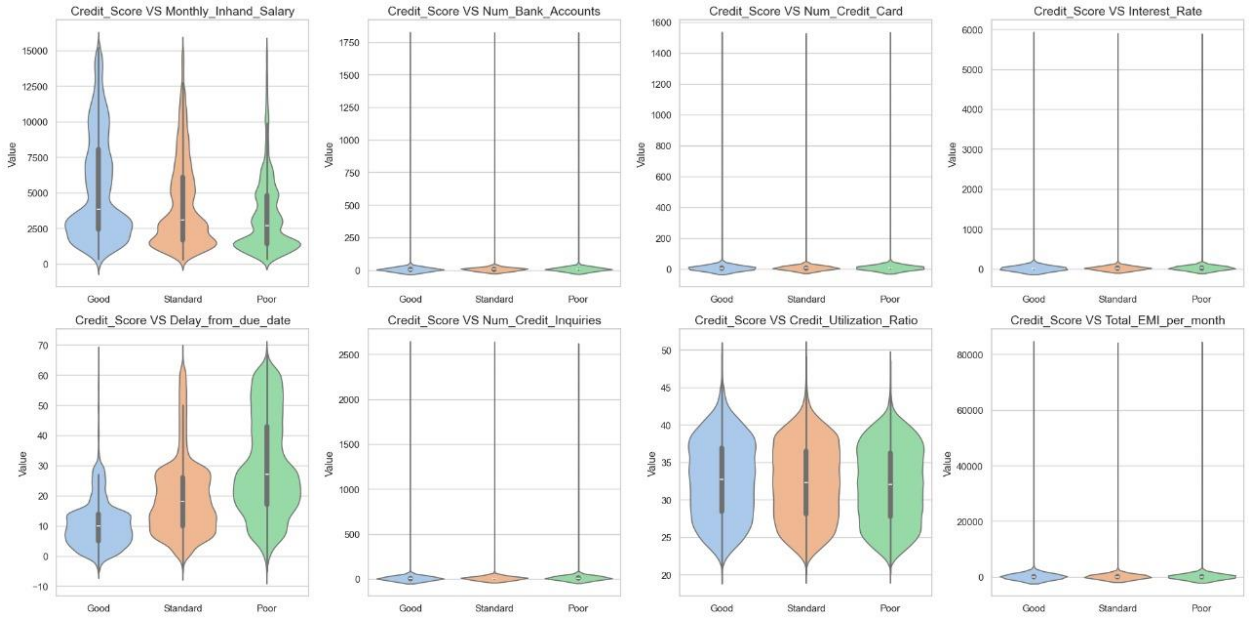
Al contar con herramientas básicas para crear nuestro modelo de scoring es interesante explorar de forma exhaustiva cómo se comportan los diferentes registros y variables de nuestros datos para tomar mejores decisiones.

Asignatura: Modelos de Crédito

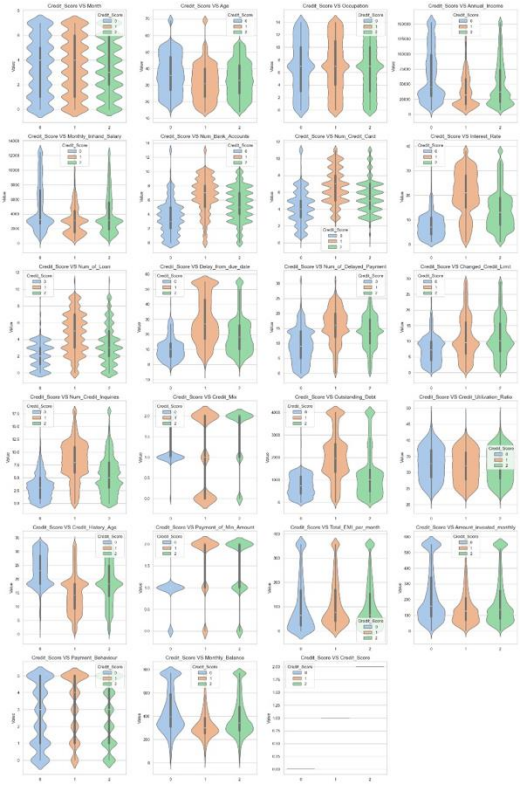
Fecha: 09 Septiembre 2024

Docente: Rodolfo Slay Ramos

Gráfica de violín (sin limpieza en los datos):



Gráfica de violín (conlimpieza en los datos):



Antes de la Limpieza de Datos:

1. **Variabilidad y Extremos:**

- Variables como `Monthly_Inhand_Salary`, `Interest_Rate`, y `Total_EMI_per_month` muestran una gran variabilidad en sus valores, especialmente en los scores de crédito 'Good' y 'Standard'. Esto podría indicar presencia de valores atípicos extremos o errores en los datos.

- `Num_Credit_Inquiries` y `Credit_Utilization_Ratio` también presentan formas de distribuciones que sugieren extremos altos, lo cual puede afectar la calidad del modelo si estos valores no son tratados adecuadamente.

2. **Distribuciones Inconsistentes:**

- `Delay_from_due_date` muestra valores negativos en los tres niveles de score de crédito, lo cual podría ser indicativo de errores de entrada o codificación en los datos.

3. **Correlaciones Potenciales:**

- `Num_of_Loan` y `Num_Bank_Accounts` parecen tener distribuciones más estrechas en los scores de crédito 'Good', sugiriendo una posible relación entre un menor número de préstamos o cuentas y un mejor score de crédito.

Después de la Limpieza de Datos:

1. **Reducción de Extremos y Valores Atípicos:**

- Las distribuciones en variables como `Monthly_Inhand_Salary` y `Interest_Rate` parecen más compactas y menos afectadas por valores extremos, indicando una efectiva limpieza de datos.

2. **Corrección de Errores Evidentes:**

- `Delay_from_due_date` ya no muestra valores negativos, lo cual refleja una corrección adecuada de errores en los datos.

3. **Alineación de las Distribuciones con el Score de Crédito:**

- Variables como `Num_Credit_Card` y `Num_of_Loan` ahora muestran una variabilidad más consistente a través de los diferentes niveles de score de crédito, sugiriendo una mejor representación de la realidad tras la limpieza.

4. **Consistencia y Claridad en la Visualización:**

- Las gráficas de violín después de la limpieza muestran una representación más clara y consistente del impacto de diferentes variables en el score de crédito, lo que puede facilitar análisis predictivos más precisos.

Conclusiones Generales:

- **Importancia de la Limpieza de Datos:** El efecto de la limpieza de datos es claro al comparar ambas imágenes. La limpieza ayuda a reducir el impacto de valores atípicos y errores, lo que conduce a análisis más precisos y fiables.
- **Potenciales Indicadores de Score de Crédito:** Variables como `Num_of_Loan`, `Num_Bank_Accounts`, y `Credit_Utilization_Ratio` pueden ser indicadores clave para predecir el score de crédito, como se observa por sus distribuciones más disciplinadas después de la limpieza.
- **Mejor Visualización y Análisis:** Post-limpieza, las gráficas de violín son más uniformes y ofrecen una visualización mejorada, facilitando la interpretación y el análisis estadístico.

Variables recomendadas:

En base a un análisis exhaustivo de las variables para la calculación de score de crédito, mostraremos a continuación las variables clave para predecir mejor nuestro modelo:

- **Payment_of_Min_Amount (Cantidad mínima de pago):**

Esta variable indica si el titular de la cuenta ha realizado el pago mínimo requerido en su estado de cuenta de crédito. Es un factor importante para evaluar el comportamiento de pago del cliente y su gestión de deuda.

Categorías:

- **Yes**: El cliente ha realizado al menos el pago mínimo requerido.
- **No**: El cliente no ha realizado el pago mínimo.
- **NM** (No Mentioned): No se dispone de información sobre si se realizó el pago mínimo.

Medidas estadísticas	Valores (Count x)
Yes	52326
No	35667
NM	12007

- **Credit_Mix (Número de Tipos de Crédito):**

Refiere a la diversidad de cuentas de crédito que posee un individuo, como tarjetas de crédito, préstamos personales, hipotecas, entre otros. Un mix de crédito variado puede ser indicativo de una mayor capacidad y experiencia en el manejo del crédito.

Categorías:

- **Standard**: El cliente tiene un mix de crédito estándar, posiblemente con algunos tipos de créditos.
- **Good**: El cliente tiene un buen mix de crédito, indicativo de una gestión de crédito diversificada y responsable.
- _ (Unknown or missing): Información desconocida o no proporcionada.
- **Bad**: El cliente tiene un mix de crédito pobre, posiblemente limitado a un solo tipo de crédito o con señales de manejo inadecuado del crédito.

Categorías	Cantidad de casos (Count.x)
Standard	36,479
Good	24,337
Bad	18,989
"_"	20,195

- **Changed_Credit_Limit (Cambios en el límite de crédito):**

Indica los cambios en el límite de crédito de la cuenta del cliente. Los ajustes en el límite de crédito pueden reflejar la confianza del prestamista en la capacidad de pago del cliente o cambios en su perfil de riesgo.

Los valores numéricos reflejan el cambio en el límite de crédito desde la última evaluación, donde los valores positivos indican un aumento y los valores negativos una disminución. La presencia de un guión bajo ('_') puede indicar que no hubo cambios o que la información no está disponible.

Medidas estadísticas	Valores
Mínimo.	-6.49
Máximo.	36.97
Media.	10.39
Moda.	8.22
Valores atípicos. *	[36.97, -6.49]

Los valores atípicos fueron identificados utilizando el método del rango intercuartílico (IQR). Quartil 1: Valor que separa el 25% más debajo de los datos, Quartil 3: separa el 75% más abajo del 25% más alto. El IQR es la diferencia entre Q3 Y Q1 (Q3-Q1)

- **Num_of_Delayed_Payment (Retraso respecto a la fecha de vencimiento):**

Contabiliza la cantidad de veces que el cliente ha retrasado los pagos de sus obligaciones crediticias. Un alto número de pagos retrasados es un fuerte indicador de riesgo de crédito y dificultades financieras.

Aunque principalmente numéricos, algunos registros pueden contener errores o valores mal formados, que deben limpiarse o investigarse antes de utilizarlos en análisis predictivos.

Medidas estadísticas	Valores
Mínimo.	-3
Máximo.	4397
Media.	31.03
Moda.	19
Valores atípicos. *	[4397]

- **Num_Bank_Accounts (Numero de cuentas bancarias):**

Representa el número de cuentas bancarias que el cliente posee. Este dato puede ser indicativo de la salud financiera del cliente y su capacidad para manejar múltiples cuentas, aunque un número muy alto podría también sugerir un comportamiento de búsqueda de crédito o gestión financiera compleja.

Medidas estadísticas	Valores
Mínimo.	-1
Máximo.	1798
Media.	17.09
Moda.	6
Valores atípicos. *	[1798]

Output:

Credit Score. Nos indica cómo se encuentra el score crediticio de los clientes tomando en cuenta las variables anteriores.

Categorías	Cantidad de casos (Count.x)
Standard	53,174
Good	17,828
Bad	28,998

Es relevante saber cuántos casos hay por categoría en la variable credit_score ya que será un factor determinante a la hora de validar el accuracy de nuestro modelo. Score de crédito predicho de nuestro modelo vs valores reales de la base (Credit_score).

Las variables recomendadas que seleccionamos para el modelo muestran valores muy atípicos sin tomar en cuenta la limpieza de datos anterior, para calcular el accuracy de nuestro modelo seleccionamos las variables recomendadas ya con la limpieza de datos realizada.

Asignación de Score para las variables.

Se utilizaron las variables anteriormente mencionadas de la base de datos para categorizar en 3 diferentes niveles dependiendo su puntaje (Good, Standard, Poor) los parámetros. Con la información anterior se crearon diferentes rangos por cada variable y así generar distinta calificación a los registros. Para asignar el score de cada registro creamos condiciones y rangos de acuerdo con un análisis exploratorio de los datos para finalmente obtener su score de crédito y compararlo con la puntuación ya establecida.

Modelo:

El Modelo de Puntuación de Crédito (CSM) desarrollado en este documento tiene como objetivo principal evaluar la solvencia crediticia de individuos basándose en una serie de variables financieras y comportamentales. Esta herramienta es esencial para las instituciones financieras, ya que permite una evaluación detallada y automatizada del riesgo, facilitando la toma de decisiones en el otorgamiento de créditos y servicios financieros.

El CSM asigna una puntuación basada en varios parámetros clave. La asignación de puntos se realiza según la importancia predictiva de cada variable con respecto al riesgo de incumplimiento. Se han definido cuatro niveles de puntuación:

- **Puntuación Alta (l_sup):** 50 puntos.
- **Puntuación Media-Alta (m_sup):** 37.5 puntos.
- **Puntuación Media-Baja (m_inf):** 25 puntos.
- **Puntuación Baja (l_inf):** 12.5 puntos.

Estos puntos se asignan en función del cumplimiento de ciertos umbrales para cada variable relevante, como se detalla a continuación:

1. **Pago Mínimo Realizado (Payment_of_Min_Amount):**
    - **l_sup (50 puntos):** Si el pago mínimo realizado es menor o igual a 1.
    - **m_sup (37.5 puntos):** Si el pago mínimo realizado es menor o igual a 2.
    - **m_inf (25 puntos):** Cualquier valor superior a 2.

2. **Mezcla de Crédito (Credit_Mix):**
    - **l_sup (50 puntos):** Para una mezcla de crédito clasificada como '0' (sin créditos previos).
    - **m_sup (37.5 puntos):** Para una mezcla de crédito '1' (créditos previos diversificados).
    - **l_inf (12.5 puntos):** Para una mezcla de crédito '2' (créditos previos no diversificados).

3. **Límite de Crédito Cambiado (Changed_Credit_Limit):**
    - **l_sup (50 puntos):** Si el cambio en el límite de crédito es menor o igual a 5.76.
    - **m_sup (37.5 puntos):** Si el cambio es menor o igual a 14.66.
    - **m_inf (25 puntos):** Si el cambio es menor o igual a 24.
    - **l_inf (12.5 puntos):** Para cambios superiores a 24.

4. **Número de Pagos Retrasados (Num_of_Delayed_Payment):**
    - **l_sup (50 puntos):** Si el número de pagos retrasados es menor o igual a 9.
    - **m_sup (37.5 puntos):** Si es menor o igual a 18.
    - **m_inf (25 puntos):** Si es menor o igual a 25.
    - **l_inf (12.5 puntos):** Para cualquier valor superior a 25.

5. **Número de Cuentas Bancarias (Num_Bank_Accounts):**
    - **l_sup (50 puntos):** Si el número de cuentas bancarias es menor o igual a 3.
    - **m_sup (37.5 puntos):** Si es menor o igual a 7.
    - **m_inf (25 puntos):** Si es menor o igual a 10.
    - **l_inf (12.5 puntos):** Para cuentas superiores a 10.

El modelo utiliza una clase en Python, CSM, para procesar los datos y aplicar las puntuaciones. La clase incluye métodos para aplicar la puntuación basada en las condiciones establecidas y calcular la precisión del modelo mediante la comparación de las predicciones con los scores de crédito reales.

La validación se realiza comparando las predicciones del modelo (Our_model) con los scores de crédito reales (Credit_Score). La precisión del modelo se calcula como el porcentaje de predicciones correctas y es crucial para evaluar la efectividad del modelo en escenarios reales.

Aunque el CSM es robusto, tiene limitaciones inherentes a su diseño estático. Los pesos y umbrales fijos pueden no adaptarse bien a cambios en el mercado o en el comportamiento del consumidor. Se recomienda revisar periódicamente estos parámetros y ajustarlos basándose en análisis de datos actualizados y feedback del rendimiento del modelo. Además, la inclusión de técnicas de machine learning podría permitir ajustes dinámicos y mejorar la precisión del modelo.

Asignatura: Modelos de Crédito **Fecha:** 09 Septiembre 2024
Docente: Rodolfo Slay Ramos

Conclusión:

Este proyecto del Modelo de Puntuación Crediticia ha sido un verdadero avance hacia entender mejor cómo podemos usar la tecnología para predecir la solvencia crediticia. Hemos logrado combinar diferentes tipos de información financiera y comportamental de manera que el modelo no solo es efectivo, sino también reflexivo sobre la complejidad de las finanzas personales y empresariales.

El modelo ha demostrado ser una herramienta increíblemente útil para las instituciones financieras, simplificando sus decisiones de crédito y promoviendo un acceso más justo y transparente al crédito con una precisión impresionante del 53.34%. Sin embargo, como toda herramienta, tiene sus limitaciones. Depende mucho de los datos del pasado, que podrían no ser representativos de cambios futuros en el mercado o en el comportamiento financiero.

Es crucial que sigamos revisando y ajustando los parámetros del modelo regularmente para asegurarnos de que se mantenga relevante y preciso. En el futuro, podríamos mejorar el modelo aún más utilizando técnicas de aprendizaje automático que permitan adaptaciones dinámicas a nuevos patrones de datos y mejoren su capacidad predictiva.

Pensando a largo plazo, también podríamos considerar incluir más variables cualitativas para obtener una imagen más completa de lo que afecta la solvencia crediticia, más allá de los números fríos.

Reflexiones:

Fue retador encontrar un modelo dónde se cumpliera con los parámetros del proyecto, aun así, conseguimos dar con una combinación que completaba los requerimientos. Crear un score crediticio representa cuantificar la vida de las personas y tomar en cuenta el promedio de las personas o registros para etiquetar si son aptos o no para recibir un crédito. Los modelos existentes son una referencia con alto valor en la elaboración de esta propuesta de puntuación crediticia.

Para el primer proyecto nos basamos en nuestro criterio de forma racional clasificando las diferentes variables que nos pueden ayudar a obtener un mejor modelo de crédito, hay que entender que hoy en día el modelo FICO es el principal que se usa en Estados Unidos, en México uno similar planteado por buro de crédito, pero basándose en el modelo FICO, por ello es importante analizar las variables que toman nuestro modelo vs FICO:

Variables de nuestro modelo vs modelo fico.

Nuestro Modelo	**Modelo Fico**
Payment_of_Min_Amount	*Payment history*
Credit_Mix	*Amount owed*
Changed_Credit_Limit	*Lenght of credit history*
Num_of_Delayed_Payment	*New credit*
Num_Bank_Accounts	*Types of credit*

Las críticas más fuertes al modelo fico son las siguientes:

1. Falta de transparencia.
2. Sesgo racial.
3. No predice comportamientos futuros.
4. Ignora factores cualitativos.
5. Tiene dificultad para mejorar el puntaje.

Por estos motivos es muy relevante ir creando nuevos modelos que solucionen esas críticas y generar consciencia sobre el futuro crediticio global. Este proyecto es de gran utilidad para encontrar un modelo con menos margen de error y predecir mejor los scores de crédito. Es muy importante antes de plantear cualquier modelo hacer una limpieza y análisis de tus datos porque en el mundo real o laboral muchas veces las variables o columnas en una base de datos tienen muchos datos atípicos y si entrenamos o hacemos un modelo ignorando eso, nuestro modelo puede generarnos más ruido y predecir de peor manera.

Además, es fundamental considerar el uso de modelos de optimización en futuros proyectos de puntuación crediticia. La optimización permite ajustar los parámetros del modelo de manera precisa para minimizar errores y maximizar la precisión en la predicción de los scores de crédito. Al incorporar técnicas de optimización, como algoritmos genéticos, métodos numéricos o técnicas de aprendizaje profundo, se puede mejorar la capacidad del modelo para identificar patrones y relaciones complejas en los datos. Esto no solo refina la precisión de las predicciones, sino que también ayuda a abordar algunas de las críticas al modelo FICO, como la capacidad de incorporar factores cualitativos y predecir comportamientos futuros con mayor exactitud. Por tanto, la inclusión de modelos de optimización será clave para el desarrollo de modelos de crédito más justos y efectivos.