Resultados

Basados en los resultados podemos decir lo sisguiente:

Regresión Logística:

Este modelo identifica correctamente a 33 clientes que realmente incumplieron (verdaderos positivos).

Sin embargo, también clasifica erróneamente a 39 clientes como no incumplidores cuando en realidad sí lo hicieron (falsos negativos).

Además, etiqueta incorrectamente a 18 clientes como incumplidores cuando en realidad no lo son (falsos positivos).

Identifica correctamente a 160 clientes que no incumplieron (verdaderos negativos).

Árbol de Decisión:

Similar a la regresión logística, este modelo tiene un número similar de verdaderos positivos y falsos negativos, con 34 y 38 clientes respectivamente.

Sin embargo, tiene más falsos positivos que la regresión logística, con 45 clientes incorrectamente clasificados.

Identifica correctamente a 133 clientes que no incumplieron.

Bosques Aleatorios:

Este modelo muestra un rendimiento relativamente bueno en términos de verdaderos negativos, con 162 clientes correctamente clasificados.

Tiene un número significativo de falsos negativos, con 42 clientes incorrectamente clasificados.

Sin embargo, tiene menos falsos positivos en comparación con los otros modelos, con solo 16 clientes incorrectamente clasificados.

Identifica correctamente a 30 clientes que incumplieron.

Naive Bayes:

Este modelo muestra un buen rendimiento en términos de verdaderos positivos, con 46 clientes correctamente clasificados.

Tiene el menor número de falsos negativos entre los modelos presentados, con solo 26 clientes incorrectamente clasificados.

Sin embargo, tiene un alto número de falsos positivos, con 43 clientes incorrectamente clasificados.

Identifica correctamente a 135 clientes que no incumplieron.

## Desafios

1.- Analizar los datos de las distribuciones e identificar si hay algun valor o registros que no se deben considerar para el modelo

Solucion:

Los rangos de edad y los rangos de plazos a credito, en los ultimos niveles tienen solamente 6 y 1 integrante respectivamente, lo que puede ser optimizado ampliando los rangos de inferiores inmediatos con el fin de disminuir niveles de pocos integrantes.

2.- Investigar que es y como crear un mapa de calor para analizar la correlacion de las variables

Solucion:

Un heatmap o mapa de calor es un tipo de gráfico que se utiliza para representar información en una matriz, donde cada celda de la matriz se colorea según un valor escalar. En Python, se puede crear un heatmap utilizando la biblioteca de visualización de datos 'seaborn'. La función 'heatmap' de seaborn permite crear un heatmap a partir de una matriz de datos y colorear las celdas según un mapa de colores específico. Además de los datos, se pueden especificar opciones adicionales como el rango de los valores a representar, el tamaño del gráfico y las etiquetas del eje x e y.

Los mapas de calor (heatmap) son útiles para visualizar patrones y tendencias en datos multivariados, especialmente cuando hay muchos datos y se busca una representación resumida y, dentro de estos patrones, heatmap es perfecto para visualizar las correlaciones de columnas de un dataframe que previamente hemos podido cargar con datos procedentes de un archivo .csv o .xlsx de Excel.

Codigo del mapa de calor en la sección de abajo

3.- Crear una conclusion para cada uno de los graficos del histograma. Mirar los datos y extraer conclusiones, porque es una habilidad esencial.

Solucion:

De cada una de las graficas del histograma se pueden sacar las siguientes conclusiones:

Sexo: el 31% de los clientes del banco son clientes femeninos, mientras el 69% son clientes masculinos.

Estado Civil: el 54.8% de los clientes se encuentran solteros, mientras el 45.2% de los clientes restantes están casados o divorciados

Rango de plazos a credito: El promedio de los clientes suelen establecer plazos a credito menores a 12 meses

Rango de edad: La mayoria de los clientes que solicitan creditos, tienen entre 18 y 30 años.

Default: El 70% de los clientes son considerados buenos clientes

Tarde pero seguro, resultado del primer día, conclusiones: 1.- los hombres casados, entre los 25 y los 40 tiene el doble de posibilidades de obter un crédito que las mujeres en las mismas condiciones.

2. Sinembargo el riesgo de falta de pago es casi del triple en hombres que en mujeres y el riesgo aumenta conforme la edad se acerca más a los 25.

3.por lo tanto se podría suponer que el producto más adecuado para hombres entre; 20-30 años casados. Préstamos a largo plazo con tazas variable para generar historial crediticio . 30-50 años casados. Préstamos a corto plazo con taza de mayor rendimiento. De 50 en adelante, los riesgos con esta población los pone en un rango en el que es muy poco probable que sean aptos para recibir un crédito.

5. Así mismo los productos para mujeres que parecen más adecuados serían:

20-30 casada. Préstamos a corto plazo con tazas fija, para generar historial crediticio. 30-40 casada. Préstamo a largo plazo con tazas variable. De 50 en adelante tienen la misma posición que los hombres.

• La inspección del mapa de calor revela una correlación llamativa entre las variables "estado\_civil" y "default", indicando una posible conexión entre el estado civil de los individuos y su propensión al incumplimiento de pagos. Esta asociación podría ser crucial para entender mejor los riesgos crediticios y diseñar estrategias de mitigación efectivas.- Al analizar detenidamente el patrón representado en el mapa de calor, se destaca una relación poco significativa entre la variable "rango\_edad" y "default". Esto sugiere que la edad por sí sola puede no ser un factor determinante en el comportamiento de pago, lo que podría desafiar algunas suposiciones previas sobre la relación entre la edad y la responsabilidad financiera. - La visualización revela una concentración notable de incumplimientos en ciertos rangos de plazos de crédito, señalando que la duración del crédito podría ser un factor crítico a considerar al evaluar el riesgo crediticio. Estos hallazgos subrayan la importancia de ajustar las políticas de crédito para adaptarse a diferentes plazos, y resaltan la necesidad de modelos de predicción más refinados que incorporen esta información. - La falta de correlación entre las variables "sexo" y "default" resalta que el género puede tener una influencia limitada en la capacidad de un individuo para cumplir con sus obligaciones financieras. Este hallazgo desafía los estereotipos de género relacionados con el comportamiento crediticio y sugiere que las decisiones de crédito deben basarse en factores más relevantes y predictivos.

Considero que el rango 6 de la variable “Rango Plazos Crédito” correspondiente a los créditos entre 60 y 72 meses se puede unificar con los créditos de 48 meses a 60 meses. De esta forma se podrían contabilizar en general los créditos >=48 meses.

Lo anterior también se puede realizar en la variable “Rango de edad”, al unificar los dos últimos rangos para contabilizar las personas con crédito mayores de 60 años.

Finalmente, los rangos utilizados en la variable de estado civil podrían describir de mejor forma la población si se organizan en 3 categorías: Solteros, Divorciados/Separados/Viudos y Casados, sin embargo, debido a la naturaleza de la base no se podría realizar.

Gráfico 1: La mayor cantidad de clientes del banco corresponden al sexo masculino con 690. Mientras que las mujeres representan una baja proporción con cerca del 31%. Grafico 2: Del total de clientes el 55% son “Solteros” y el 45% restante presentan uno de los siguientes estados civiles: Divorciado, separado, casado, viudo. Grafico 3: El 77% de los clientes tienen créditos que oscilan entre 1 y 24 meses. Solo el 9% de los clientes tomas créditos superiores a 36 meses. Gráfico 4: Los clientes con edades entre 18 y 40 años representan el 89% del total de clientes del banco. Mientras que las personas mayores de edad (Superior a 60 años) solo representan el 5% de los clientes.

Grafico 5: En su mayoría, los clientes del banco son considerados buenos clientes con el 70%. Aun así, se deben generar estrategias para disminuir el 30% correspondiente a clientes catalogados como “malos

3. Conclusiones

Gráfico 1: La mayor cantidad de clientes del banco corresponden al sexo masculino con 690. Mientras que las mujeres representan una baja proporción con cerca del 31%. Grafico 2: Del total de clientes el 55% son “Solteros” y el 45% restante presentan uno de los siguientes estados civiles: Divorciado, separado, casado, viudo. Grafico 3: El 77% de los clientes tienen créditos que oscilan entre 1 y 24 meses. Solo el 9% de los clientes tomas créditos superiores a 36 meses. Gráfico 4: Los clientes con edades entre 18 y 40 años representan el 89% del total de clientes del banco. Mientras que las personas mayores de edad (Superior a 60 años) solo representan el 5% de los clientes.

Grafico 5: En su mayoría, los clientes del banco son considerados buenos clientes con el 70%. Aun así, se deben generar estrategias para disminuir el 30% correspondiente a clientes c

atalogados como “malos

Luego intenté hacer una regresión lineal, comprobando antes que se cumpla el requisito de linealidad def check\_linearity(df\_banco, sexo, default): #Regresion X = sm.add\_constant(df\_banco[sexo]) y = df\_banco[default] model = sm.OLS(y, X).fit()

residuos = model.resid

durbin\_watson = sm.stats.stattools.durbin\_watson(residuos) print("Estadístico de Durbin-Watson:", durbin\_watson) if durbin\_watson < 1.5 or durbin\_watson > 2.5: print("Los residuos muestran evidencia de autocorrelación.") else: print("No hay evidencia de autocorrelación en los residuos.") check\_linearity(df\_banco, ['sexo'], 'default') El estadístico de Durbin-Watson de 2.04 nos indica que nuestros datos cumplen con el criterio de linealidad. Este paso lo hice innecesariamente porque ya sale automáticamente en la tabla de regresión. Luego realicé la regresión: def test\_regresion(df\_banco, sexo, default): X = sm.add\_constant(df\_banco[sexo])

y = df\_banco[default]

model = sm.OLS(y, X).fit() print(model.summary()) test\_regresion(df\_banco, 'sexo', 'default') Si bien el P indica que la variable sexo es estadísticamente significativa en la variabilidad de "default", el modelo de regresión lineal solo explica un 0.6% de la variabilidad de "default" en función del sexo. La regresion deberia ser enriquecida con otras variables. Luego recordé que hay que chequear la homocedasticidad para la regresión lineal. La prueba de Breusch-Pagan me indicó heterocedasticidad así que quizás hay que realizarle ajustes adicionales al modelo. ¡Si alguien hace una regresión teniendo en cuenta otra variable dependiente e independiente, bienvenido sea

1. Rango de plazos de créditos y Rango de edad pueden ser modificadas.

• Rango de plazos de crédito – ya que el rango 6 que tiene solo un valor, se podría ampliar el rango 5 para que lo contenga. • Rango de edad - el rango 5 se puede ampliar para que contenga el rango 6 2. variables = ['sexo', 'estado\_civil', 'rango\_plazos\_credito', 'rango\_edad', 'default'] plt.figure(figsize=(10, 9)) plt.title('Mapas de Calor') for i, variable in enumerate(variables): plt.subplot(3, 2, i + 1) tabla\_contingencia = pd.crosstab(index=df\_banco[variable], columns='count') sns.heatmap(tabla\_contingencia, annot=True, cmap='coolwarm', fmt='g', cbar=False) plt.xlabel(variable, fontsize=20) plt.ylabel('Cantidad', fontsize=10) plt.show()

3. • sexo: la mayoría de los clientes del banco son hombres.

• estado civil: la mayoría de los clientes son solteros. • rango de meses: los plazos preferidos son de 12 meses a 24 meses sobre el resto. • rango de edad: la mayoría de clientes está entre 18-50 años. • default: la mayoría de los clientes están catalogados como "Good customer"

## ****Resultados****

Basados en los resultados podemos decir lo sisguiente: **Regresión Logística:**

* Este modelo identifica correctamente a 33 clientes que realmente incumplieron (verdaderos positivos).
* Sin embargo, también clasifica erróneamente a 39 clientes como no incumplidores cuando en realidad sí lo hicieron (falsos negativos).
* Además, etiqueta incorrectamente a 18 clientes como incumplidores cuando en realidad no lo son (falsos positivos).
* Identifica correctamente a 160 clientes que no incumplieron (verdaderos negativos).

**Árbol de Decisión:**

* Similar a la regresión logística, este modelo tiene un número similar de verdaderos positivos y falsos negativos, con 34 y 38 clientes respectivamente.
* Sin embargo, tiene más falsos positivos que la regresión logística, con 45 clientes incorrectamente clasificados.
* Identifica correctamente a 133 clientes que no incumplieron.

**Bosques Aleatorios:**

* Este modelo muestra un rendimiento relativamente bueno en términos de verdaderos negativos, con 162 clientes correctamente clasificados.
* Tiene un número significativo de falsos negativos, con 42 clientes incorrectamente clasificados.
* Sin embargo, tiene menos falsos positivos en comparación con los otros modelos, con solo 16 clientes incorrectamente clasificados.
* Identifica correctamente a 30 clientes que incumplieron.

**Naive Bayes:**

* Este modelo muestra un buen rendimiento en términos de verdaderos positivos, con 46 clientes correctamente clasificados.
* Tiene el menor número de falsos negativos entre los modelos presentados, con solo 26 clientes incorrectamente clasificados.
* Sin embargo, tiene un alto número de falsos positivos, con 43 clientes incorrectamente clasificados.
* Identifica correctamente a 135 clientes que no incumplieron