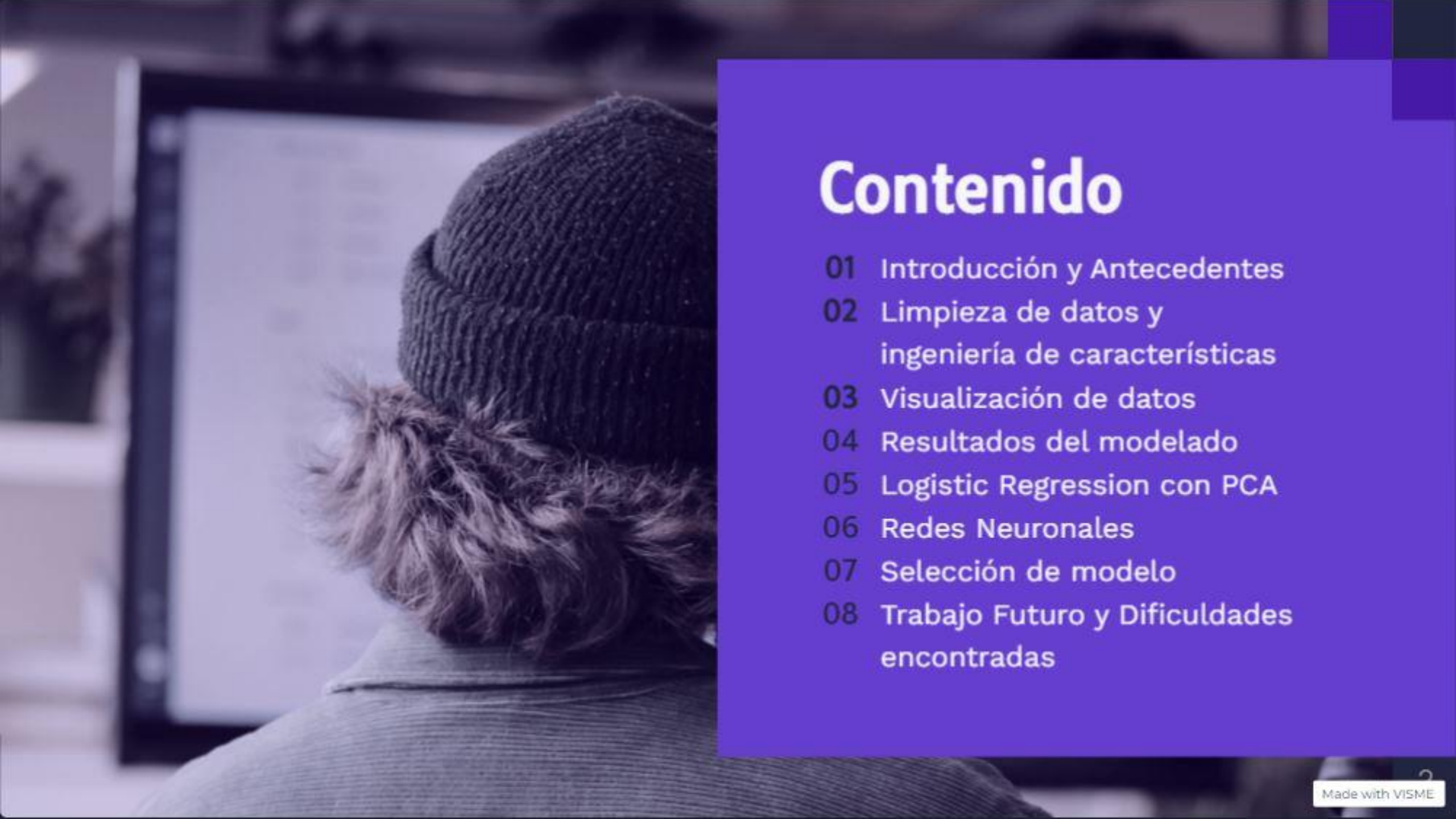
Predicting Credit Approval

Presentation

by Rodrigo Lunardi,
Data Scientist



Contenido

1.Introducción

Método: modelo predictivo basado en algoritmos de aprendizaje automático.

Preparación — **Entrenamiento** — **Evaluación**

Objetivo

Predecir la aprobación de crédito de los clientes.

Features

Antecedentes

Target

'Class' es una columna binaria donde 0 son los créditos no aprobados y 1 son los aprobados.

Balanceo

El target está desbalanceado. Requiere estratificar los datos.

- 0 0.3
- 1 0.7

Características

El conjunto de datos contiene 1000 filas y 21 columnas

- 7 Columnas numéricas
- 13 Columnas categóricas

Pré-Procesamiento

Limpieza de Datos

Acciones:

- Tratamiento de datos inconsistentes
- Agrupamiento de valores poco relevantes
- Apertura de característica en 02 nuevas columnas

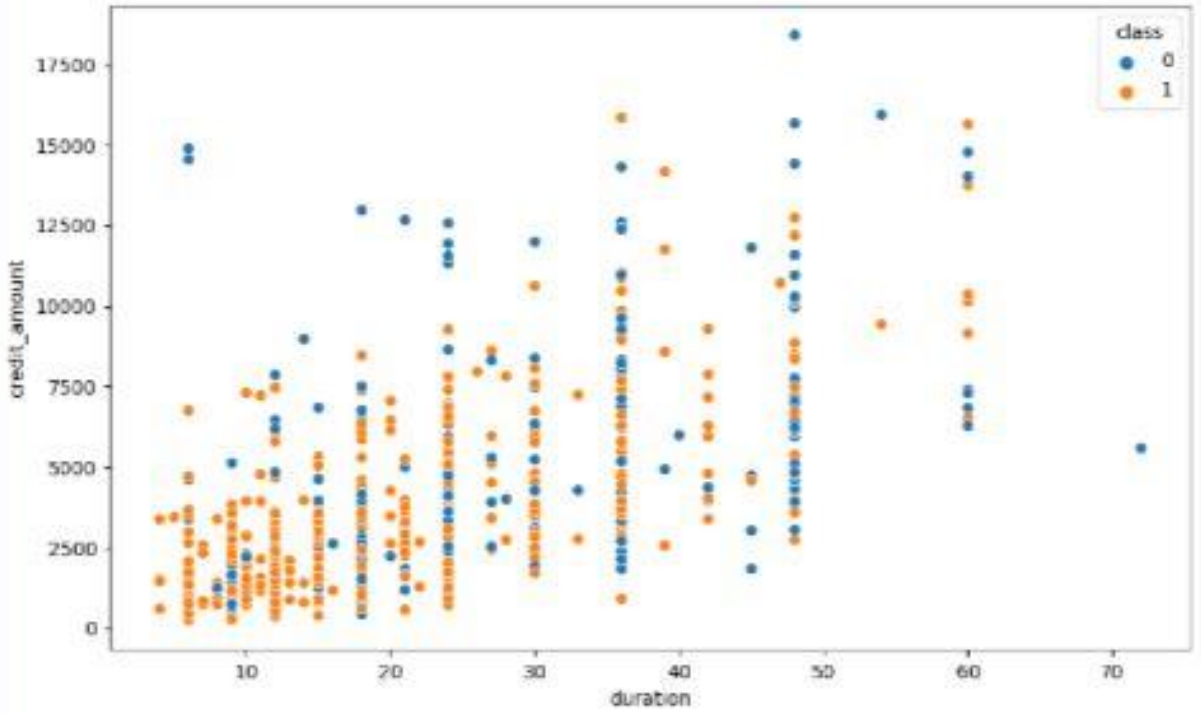
```
# Reemplazar valores inconsistentes columna 'credit_history': 'no credits/all paid' por 'no credits'
df_1['credit_history'] = df_1['credit_history'].replace('no credits/all paid', 'no credits')
df_1['credit_history'] = df_1['credit_history'].replace('all paid', 'no credits')
# Reemplazar valores con menos de 90 en la columna 'purpose' por 'other'
df_1['purpose'] = df_1['purpose'].replace('education', 'other')
df_1['purpose'] = df_1['purpose'].replace('repairs', 'other')
df_1['purpose'] = df_1['purpose'].replace('domestic appliance', 'other')
df_1['purpose'] = df_1['purpose'].replace('retraining', 'other')
# Reemplazar valores por > 100 y mantener < 100 y 'no known savings'
df_1['savings_status'] = df_1['savings_status'].replace('100<=X<500', '>100')
df_1['savings_status'] = df_1['savings_status'].replace('500<=X<1000', '>100')
df_1['savings_status'] = df_1['savings_status'].replace('>=1000', '>100')
# Dividir la columna personal_status en dos columnas: gender y marital_status
df_1[['gender', 'marital_status']] = df_1['personal_status'].str.split(' ', n=1, expand=True)
# Eliminar la columna 'personal_status'
df_1.drop('personal_status', axis=1, inplace=True)
# Reemplazar valores en la columna 'marital_status'
df_1['marital_status'] = df_1['marital_status'].replace('mar/wid', 'not single')
df_1['marital_status'] = df_1['marital_status'].replace('div/sep', 'not single')
df_1['marital_status'] = df_1['marital_status'].replace('div/dep/mar', 'not single')
# convertir la columna 'job' en una columna binaria skilled y unskilled
df_1['job']=df_1['job'].replace('unemp/unskilled non res', 'unskilled')
df_1['job']=df_1['job'].replace('unskilled resident', 'unskilled')
df_1['job']=df_1['job'].replace('high qualif/self emp/mgmt', 'skilled')
# convertir la columna objetivo en numerica
df_1['class']=df_1['class'].replace('good', 1)
df_1['class']=df_1['class'].replace('bad', 0)
```

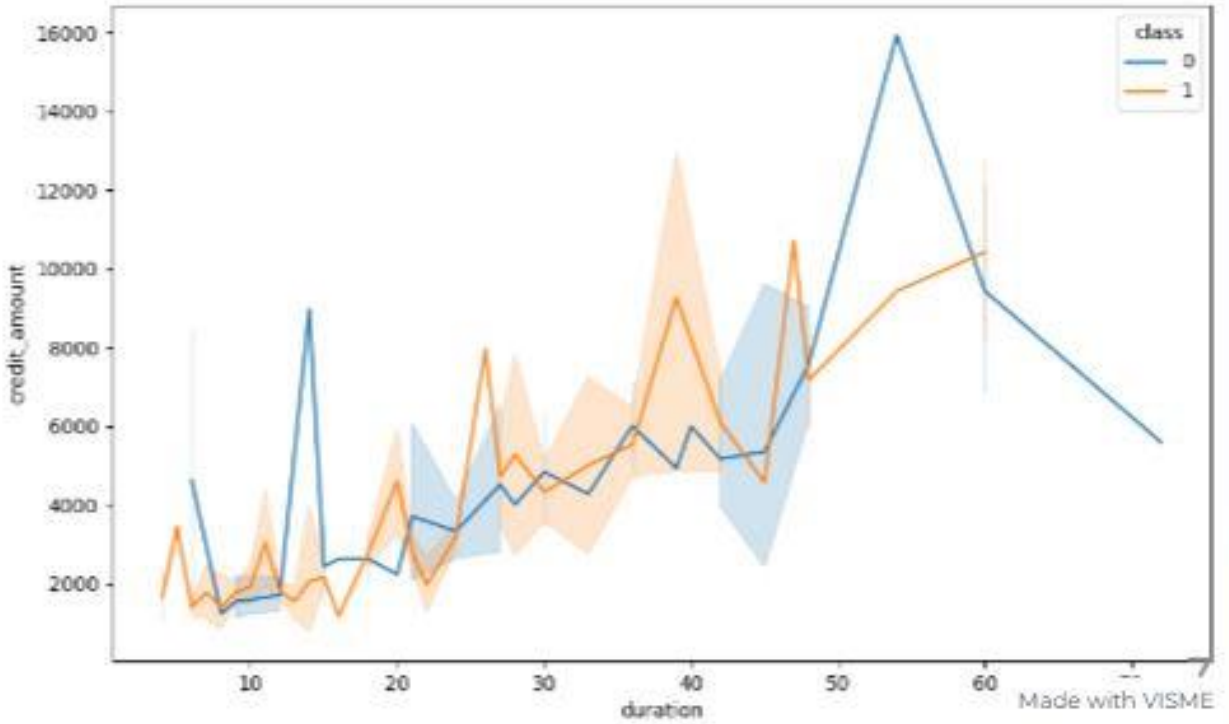
Visualización de Datos

En el gráfico de barras podemos ver que la mayoría de los créditos aprobados son para comprar televisores, seguidos por créditos para comprar autos. Los créditos más rechazados por el banco son para comprar autos.

Visualización de Datos

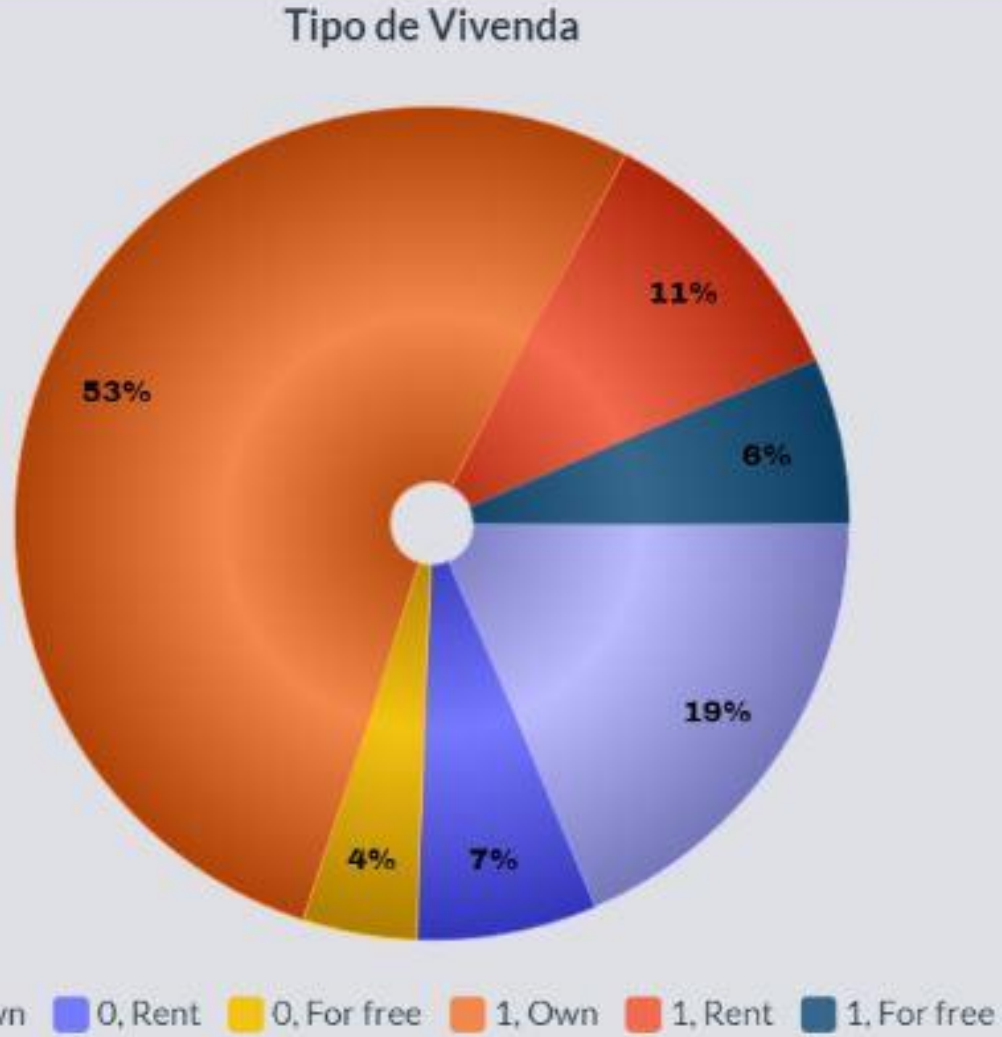
En los gráficos abajo podemos ver que la mayoría de los créditos rechazados tenían una duración alta y los montos aprobados eran bajos y tenían una duración baja.





Visualización de Datos

En el gráfico de donuts podemos ver que la mayoría de los clientes buenos o malos tienen casa propia. Esto indica que aquellos que tienen casa propia son más propensos a tomar créditos.



Métricas de Clasificación

Confusion Matrix

Representa las predicciones del modelo en comparación con los valores reales. Las filas representan las clases reales, mientras que las columnas representan las clases predichas. La matriz de confusión proporciona información sobre los verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos.

Precision

Mide la capacidad del modelo para identificar correctamente los casos positivos. Una alta precisión indica que hay pocos falsos positivos.

Recall

Mide la capacidad del modelo para encontrar todos los casos positivos. Un alto recall indica que hay pocos falsos negativos..

F1

Es una métrica que combina la precisión y el recall en un solo valor. Está calculada mediante la media armónica de ambos. La puntuación de F1 es útil cuando hay un desequilibrio entre las clases o cuando tanto la precisión como el recall son igualmente importantes.

Accuracy

Es la proporción de predicciones correctas (verdaderos positivos más verdaderos negativos) respecto al total de predicciones. Es una métrica común, pero puede ser engañosa si las clases están desequilibradas.

ROC AUC

La curva ROC muestra la relación entre la tasa de verdaderos positivos y la tasa de falsos positivos para diferentes umbrales de clasificación. El AUC es una medida del rendimiento global del modelo. Un AUC más cercano a 1 indica un mejor rendimiento en la clasificación.

Resultados de los Modelos de Clasificación

El modelo que muestra el mejor rendimiento general y un equilibrio aceptable entre las métricas es la Regresión Logística. también tiene un buen equilibrio entre el rendimiento en los conjuntos de entrenamiento y prueba, lo que sugiere una capacidad de generalización adecuada.

	Decision Tree		Bagging Tree		Random Forest		KNN Clasifier		Logistic Regression		XG Boost		LightGBM		Gradient Boost	
TP/FP	25	50	13	62	20	55	15	60	**25**	**50**	25	50	25	50	21	54
FN/TN	31	144	5	170	15	160	9	166	**17**	**158**	19	156	19	156	16	159
Precision	0.74		0.73		0.74		0.73		**0.75**		0.75		0.75		0.74	
Recall	0.82		0.97		0.91		0.94		**0.90**		0.89		0.89		0.90	
F1	0.78		0.83		0.82		0.82		**0.82**		0.81		0.81		0.81	
Accuracy Test	0.67		0.73		0.72		0.72		**0.73**		0.72		0.72		0.72	
Accuracy Train	0.82		0.92		0.94		1.0		**0.76**		0.91		0.86		0.88	
ROC AUC	0.57		0.57		0.59		0.57		**0.61**		0.61		0.61		0.59	

Principal Components Análysis

PCA

Es una técnica de reducción de dimensionalidad utilizada en el campo del aprendizaje automático y la estadística.

Objetivo

Reducir la dimensionalidad de los datos al eliminar las redundancias y la información no relevante.

Processing

Los datos de entrenamiento ya preparados para los modelos después de la técnica PCA com 85% de varianza retenida pasaron de 52 características para 20.

Logistic Regression con PCA

El proceso de reducción de la dimensionalidad con PCA mejoró los resultados de la Regresión Logística. Esto significa que el PCA ayudó a capturar las relaciones relevantes en los datos

SIN PCA

TP/FP	25	50
FN/TN	17	158
Precision	0.75	
Recall	0.90	
F1	0.82	
Accuracy Test	0.73	
Accuracy Train	0.76	
ROC AUC	0.61	

CON PCA

TP/FP	37	38
FN/TN	18	157
Precision	0.80	
Recall	0.89	
F1	0.84	
Accuracy Test	0.77	
Accuracy Train	0.76	
ROC AUC	0.69	

Redes Neuronales con PCA

Unsupervised Models

Modelo 1

TP/FP	25	50
FN/TN	28	147
Precision	0.74	
Recall	0.84	
F1	0.79	
Accuracy Test	0.68	
ROC AUC	0.58	

Modelo 2

TP/FP	30	45
FN/TN	23	152
Precision	0.77	
Recall	0.86	
F1	0.81	
Accuracy Test	0.72	
ROC AUC	0.63	

Modelo 3

TP/FP	33	42
FN/TN	25	150
Precision	0.78	
Recall	0.85	
F1	0.81	
Accuracy Test	0.73	
ROC AUC	0.64	

Redes Neuronales

El primer modelo no tuvo regularización, lo cual puede hacer que esté más propenso a sobre ajustar o tener un rendimiento subóptimo en datos nuevos. Esto se evidencia en las métricas, donde la precisión, el recall y el F1-score son relativamente bajos en comparación con los otros modelos. Además, el área bajo la curva ROC también es el más bajo.

El segundo modelo utilizó la técnica de Early Stopping, lo cual puede ayudar a evitar el sobreajuste. Se observa que este modelo mejora en términos de precisión, recall, F1-score, área bajo la curva ROC y pérdida en comparación con el primer modelo. Sin embargo, aún hay margen de mejora.

El tercer modelo utilizó la regularización L2 (Ridge), que puede ayudar a controlar la complejidad del modelo y reducir el sobreajuste. Este modelo muestra mejoras en todas las métricas en comparación con los anteriores. La precisión, recall, F1-score y área bajo la curva ROC son más altos, y la pérdida es relativamente más baja.

Teniendo en cuenta las métricas y las técnicas utilizadas, el tercer modelo con regularización L2 parece ser el mejor de los tres para el proyecto de predicción de clientes buenos para tomar un crédito o no.

Supervised o Unsupervised

Logistic Regression

TP/FP	37	38
FN/TN	18	157
Precision	0.80	
Recall	0.89	
F1	0.84	
Accuracy Test	0.77	
Accuracy Train	0.76	
ROC AUC	0.69	

Redes Neuronales

TP/FP	33	42
FN/TN	25	150
Precision	0.78	
Recall	0.85	
F1	0.81	
Accuracy Test	0.73	
ROC AUC	0.64	

Selección del Modelo

Al comparar las métricas, observamos que el modelo de Regresión Logística tiene una precisión, recall y F1-score ligeramente superiores en comparación con el modelo de Red Neuronal. Además, el modelo de Regresión Logística muestra un mejor desempeño en términos de Accuracy tanto en el conjunto de prueba como en el conjunto de entrenamiento. En cuanto al área bajo la curva ROC (ROC AUC), ambos modelos muestran valores similares, pero el modelo de Regresión Logística tiene una ligera ventaja.

Dado que el objetivo es predecir si una persona es buena o no para tomar un crédito, el modelo supervisado (Regresión Logística) parece ser más adecuado en este caso. Proporciona un mejor equilibrio entre precisión, recall y F1-score, y tiene un rendimiento superior en términos de Accuracy en ambos conjuntos de datos. Además, al ser un modelo supervisado, se beneficia de utilizar etiquetas de clase conocidas para entrenar y hacer predicciones precisas.

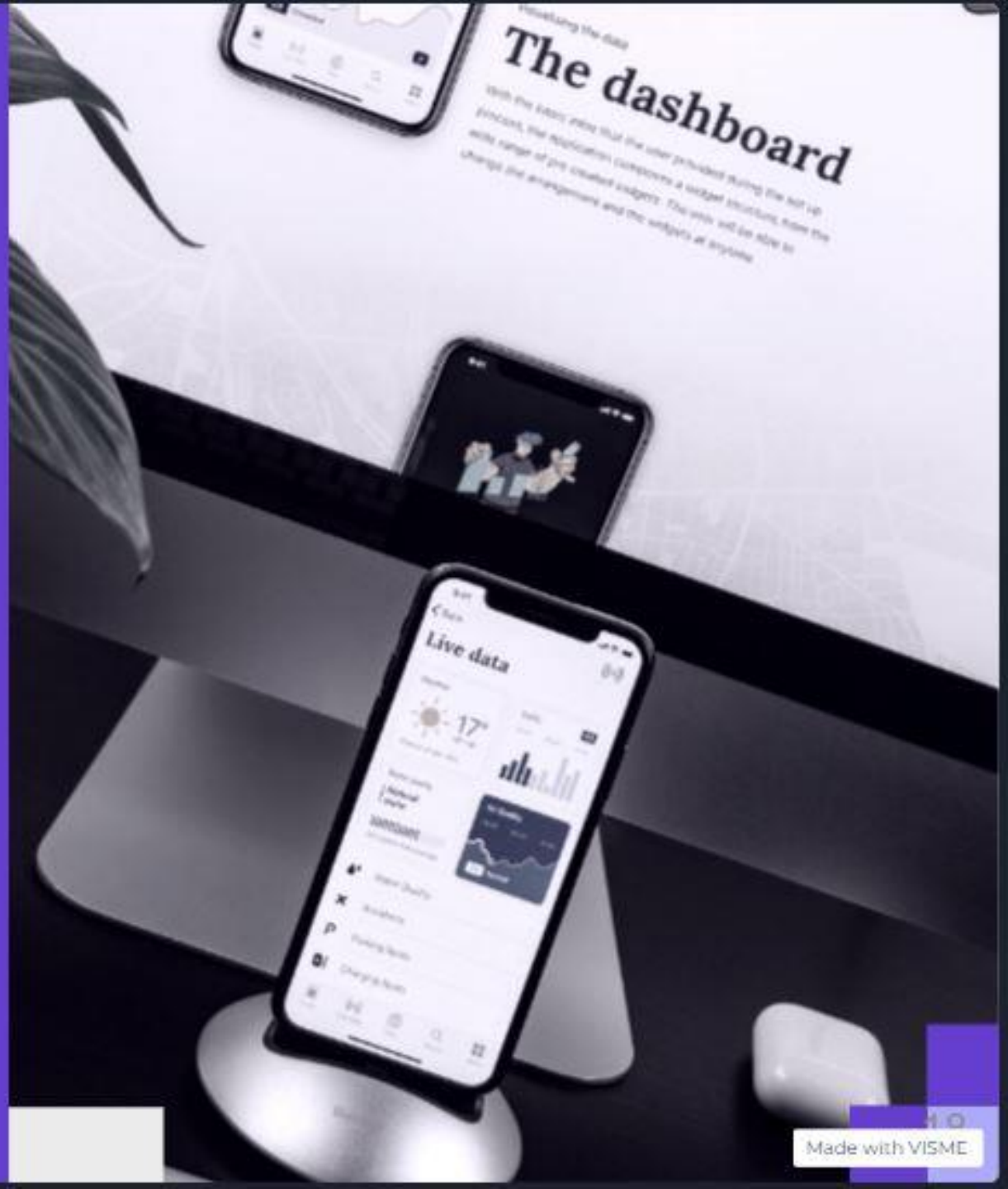
Trabajo Futuro y Dificuldades Encontradas

Trabajo Futuro

Recopilación de más datos: Si es posible, es recomendable aumentar el tamaño del dataset recopilando más datos. Un dataset más grande podría proporcionar una representación más completa de los patrones y relaciones en los datos, lo cual podría mejorar el rendimiento de los modelos.

Exploración adicional de las variables categóricas: Dado que hay 13 columnas categóricas en el dataset, sería beneficioso realizar un análisis más detallado de estas variables.

Considerar otros modelos y técnicas: Además de los modelos supervisados utilizados, podría ser beneficioso explorar otros algoritmos de aprendizaje automático y técnicas

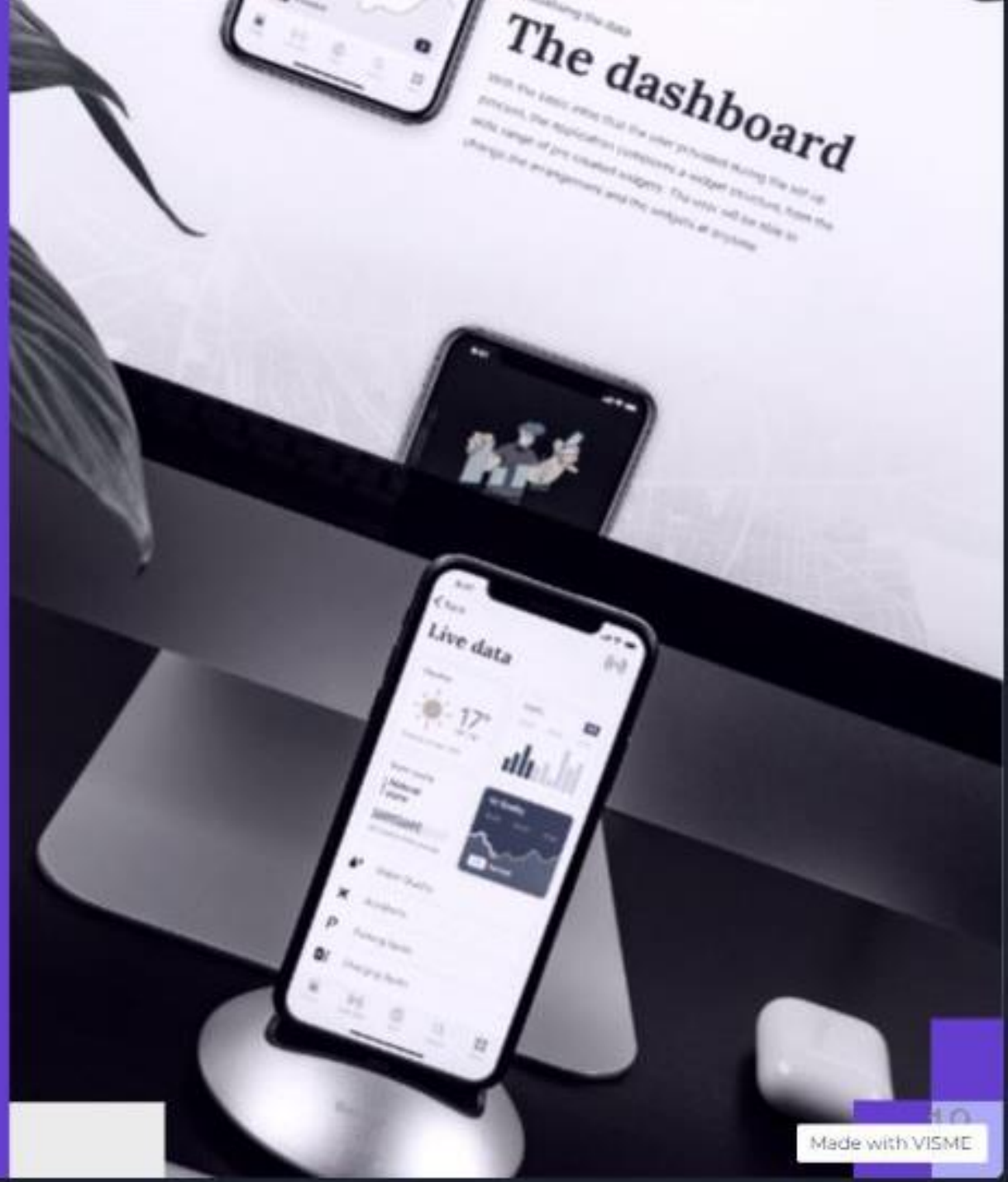


Dificuldades Encontradas

Desbalanceo de clases: Si la proporción entre las clases positivas y negativas en el dataset es muy desigual, puede generar un sesgo en los resultados y afectar el rendimiento del modelo.

Calidad y completitud de los datos: La calidad y la completitud de los datos pueden influir en la capacidad de los modelos para aprender patrones y realizar predicciones precisas. El Dataset es complexo y solamente algunas características tenían buena correlación con el Target

Interpretación de variables categóricas: Las variables categóricas pueden requirieron una codificación adecuada para ser utilizadas en los modelos. Hubo dificultades en la selección de la mejor estrategia de codificación y en la interpretación de las variables categóricas en relación con el objetivo de predicción.




Gracias!
Para más información acerca de la presentación, por favor visite
https://github.com/RodLunardi/Credit_customers.git
Made with VISME