La importancia de reducir el riesgo crediticio

En este proyecto, como científicos de datos, hemos sido convocados por una institución financiera alemana para construir un modelo de machine learning que evalúe con mayor precisión la probabilidad de incumplimiento crediticio de sus clientes.

@Víctor Carracedo

https://data.ceibe.eu/Credit_Scoring

GitHub

civvic1/Credit_Scoring

Contribute to civvic1/Credit_Scoring development by creating an account on GitHub.

Preprocesamiento de Datos

1 Limpieza de datos

Realizamos una exhaustiva limpieza de datos para asegurar la calidad y consistencia del conjunto de datos.

2 Manejo de valores faltantes

Implementamos estrategias para manejar valores faltantes en los datos de manera efectiva y evitar sesgos en el modelo.

3 Codificación de variables categóricas

Transformamos variables categóricas en variables numéricas para que el modelo pueda procesarlas correctamente.

4 Normalización y escalado de datos

Ajustamos las escalas de las variables para asegurar una comparación justa y precisa al momento de entrenar el modelo.

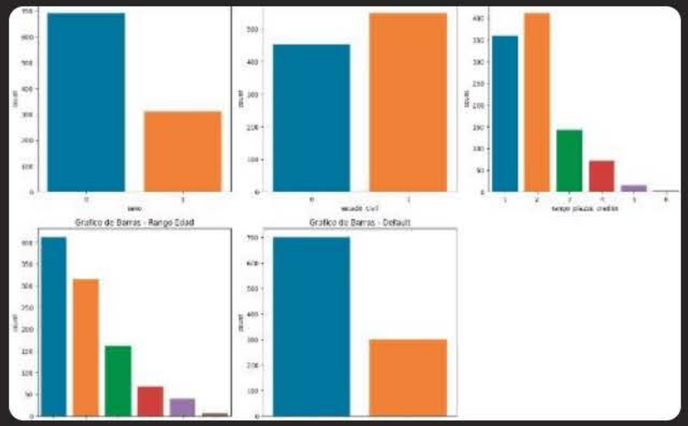
Exploración de Datos

Análisis detallado

Exploramos el conjunto de datos proporcionado en busca de patrones y características clave.

Relaciones entre variables

Utilizamos visualizaciones, como el mapa de correlacion, para comprender las relaciones entre las variables y detectar posibles influencias.



Selección de características

Identificamos las características más relevantes para el modelo y descartamos aquellas que no aportan información significativa.

Construcción de Modelos

Regresión Logística

Exploramos este modelo clásico para predecir la solvencia crediticia.

Árboles de Decisión

Utilizamos esta técnica de aprendizaje automático para representar las diferentes decisiones basadas en las características de los clientes.

Naive Bayes

Aplicamos este algoritmo de clasificación probabilístico para determinar la probabilidad de incumplimiento crediticio.

Evaluación y Selección del Modelo

1

Precisión

Utilizamos métricas como la precisión para evaluar la capacidad del modelo para realizar predicciones correctas.

2

Recall

Medimos el recall para evaluar la capacidad del modelo para encontrar todos los casos de incumplimiento crediticio.

3

Área bajo la curva ROC

Calculamos el área bajo la curva ROC para evaluar la calidad del modelo y su capacidad de discriminar entre las clases.

4

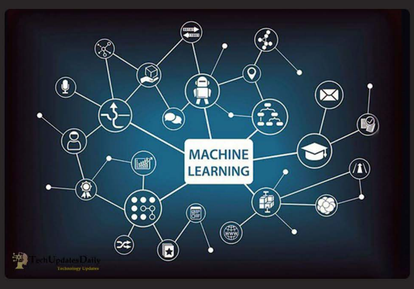
F1-score

Consideramos el F1-score que combina la precisión y el recall para seleccionar el modelo con el mejor rendimiento.

Conclusiones

Selección del Modelo

Según las métricas de evaluación (Accuracy, Precision, Recall, F1-Score, AUC-ROC), se observa que el modelo de Regresión Logística y el Árbol de Decisión tienen un desempeño similar, mientras que el modelo Naive Bayes tiene un rendimiento ligeramente superior en este caso particular.



Preprocesamiento de Datos

Se realizó un extenso preprocesamiento de datos, que incluyó la codificación de variables categóricas, la normalización/escalado de datos y la eliminación de valores nulos y duplicados.

Balance de Clases

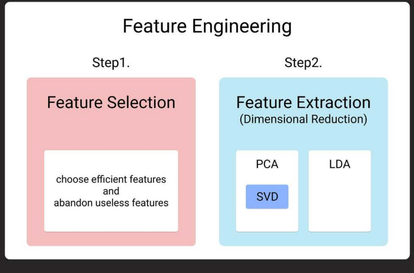
Se utilizó la técnica de oversampling (SMOTE) para abordar el desbalance en la clase objetivo (default). Esto ayudó a mejorar la capacidad del modelo para predecir la clase minoritaria.

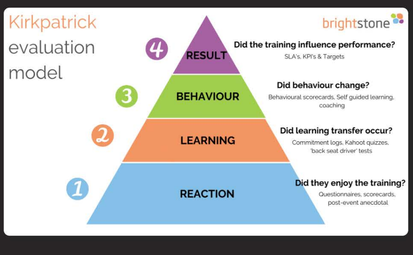
Exploración de Datos

Se realizó un análisis exploratorio de datos que incluyó gráficos de barras para algunas variables clave, así como un mapa de calor de correlaciones para comprender las relaciones entre las variables.

Feature Engineering

Se crearon nuevas características (sexo, estado civil, rangos de edad, plazos de crédito y valores de crédito) mediante el proceso de feature engineering para proporcionar información adicional y mejorar el rendimiento del modelo.





Resultados del Modelo

Los resultados del modelo fueron evaluados en términos de precisión, recall, F1-score y AUC-ROC. Se observó que los modelos tuvieron un buen rendimiento, pero la elección del modelo final dependerá de las necesidades específicas de la institución financiera (por ejemplo, si se prioriza la precisión, el recall, etc.).

Presentación

En la presentación, se puede destacar el proceso de preprocesamiento, feature engineering y el análisis exploratorio de datos. Además, es crucial explicar las métricas seleccionadas y justificar la elección del modelo final. También se puede proporcionar información sobre la interpretación de los resultados y las implicaciones prácticas para la institución financiera.