Tarea 6: Aprendizaje supervisado

Índice

1. Introducción

El presente artículo se centra en la exploración y análisis detallado de los métodos de aprendizaje supervisado, con el propósito principal de prever la tasa de interés proyectada para un crédito.

1.1. Descripción y Contexto de los Datos

El conjunto de datos con el que a trabajar se trata de un subconjunto de la base de créditos de la plataforma 'Lending Club', correspondiente al período 2007-2018. Se segmento a únicamente los créditos que se encuentran completamente saldados y aquellos que han incurrido en incumplimiento. Lo que representa un total de 261,185 registros,divididos en dos grupos, entrenamiento y prueba. A través de un proceso de selección de variables, se decidió conservar 22 características. las cuales se encuentran estandarizadas para asegurar su comparabilidad.

2. Evaluación de Modelos de Regresión

En la cuantificación del rendimiento de los modelos de regresión, se utilizarán tres métricas de error distintas. Estas métricas proporcionan una evaluación integral de la precisión predictiva de los modelos, considerando diferentes aspectos del error de predicción. A continuación, se describen detalladamente cada una de estas métricas.

2.1. Mean Squared Error (MSE)

Es una medida del promedio de los cuadrados de los errores, es decir, la diferencia promedio al cuadrado entre los valores estimados y los valores reales.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

donde:

- n Número total de observaciones (o puntos de datos)
- y_i Valor real de la i-ésima observación.
- $\hat{y}_i$ Valor predicho por el modelo para la i-ésima observación.

Características del MSE:

No negativo: El MSE siempre será un número no negativo, ya que estamos sumando los cuadrados de las diferencias.

Sensible a los valores atípicos: Como eleva las diferencias al cuadrado, los errores más grandes tienen un efecto desproporcionadamente alto en el MSE. Esto significa que el MSE es particularmente sensible a los valores atípicos en comparación con otras métricas como el error absoluto medio (MAE).

Escala: El MSE está en las unidades de los datos al cuadrado. Esto a veces puede hacer que sea difícil de interpretar, y es por eso que a menudo se toma la raíz cuadrada del MSE para obtener la raíz del error cuadrático medio (RMSE), que está en las mismas unidades que la variable de respuesta.

Comparación entre modelos: El MSE se utiliza para comparar diferentes modelos y estimaciones. Un MSE más bajo indica un mejor ajuste entre el modelo y los datos.

2.2. Coeficiente de determinación (R^2)

Se utiliza para evaluar la bondad de ajuste de un modelo de regresión. Proporciona una medida de cuánto de la variabilidad de la variable dependiente es explicada por el modelo a través de las variables predictoras.

$$R^2 = 1 - \frac{\text{Suma de los cuadrados de los residuos (SSR)}}{\text{Suma total de los cuadrados (SST)}}$$

donde:

- SSR (Sum of Squares of Residuals) es la suma de los cuadrados de las diferencias entre los valores observados y los valores predichos por el modelo.

$$SSR = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- SST (Total Sum of Squares) es la suma de los cuadrados de las diferencias entre los valores observados y el promedio de los valores observados.

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

- y_i Valor observado real para la i-ésima observación.
- $\hat{y}_i$ es el valor predicho por el modelo para la i-ésima observación.
- $\bar{y}$ es la media de todos los valores observados.
- n es el número total de observaciones.

Ventajas y Limitaciones de R^2

Ventajas: El coeficiente R^2 es una métrica cuantitativa que ofrece una evaluación relativa de la capacidad predictiva de un modelo, expresando la fracción de la varianza total de la variable dependiente que es atribuible a las variables independientes. Un valor de R^2 cercano a 1 sugiere que el modelo tiene una capacidad explicativa casi total sobre la variación observada en torno a la media de la respuesta.

Limitaciones: A pesar de su utilidad, el R^2 carece de la capacidad de discernir entre predicciones sesgadas y no sesgadas. Un R^2 elevado no es sinónimo de predictibilidad fuera de la muestra, ya que no ajusta la bondad de ajuste basándose en la complejidad del modelo. Por tanto, un R^2 alto puede ser engañoso, particularmente en situaciones de sobreajuste, donde el modelo se ajusta excesivamente a los datos de entrenamiento a expensas de su capacidad de generalización.

2.3. Mean Absolute Error (MAE)

El Error Absoluto Medio (MAE, por sus siglas en inglés Mean Absolute Error) es una métrica utilizada para evaluar la calidad de modelos en tareas de regresión. Representa el promedio de las diferencias absolutas entre los valores predichos y los valores observados, sin considerar la dirección de esos errores (es decir, sin tomar en cuenta si los valores están por encima o por debajo de la línea de mejor ajuste).

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

donde:

- n Número total de observaciones.
- y_i es el valor real de la i-ésima observación
- $\hat{y}_i$ es el valor predicho por el modelo para la i-ésima observación,

Características del MAE

No negativo: Al igual que el MSE, el MAE siempre será un número no negativo, ya que se toma el valor absoluto de las diferencias.

Escalable: El MAE está en las mismas unidades que la variable de respuesta, lo cual facilita su interpretación.

Menos sensible a valores atípicos: A diferencia del MSE, el MAE no eleva los errores al cuadrado, por lo que los valores atípicos tienen menos influencia en el resultado final. Esto hace que el MAE sea una métrica más robusta en presencia de valores atípicos.

Interpretación intuitiva: El MAE puede interpretarse como el error promedio en las predicciones del modelo.

3. Elección del modelo

Para seleccionar el modelo más adecuado, se llevará a cabo una evaluación exhaustiva de diversos modelos utilizando todas las variables disponibles. Este proceso inicial de exploración tiene como objetivo determinar la eficiencia de cada modelo en relación con nuestro conjunto de datos específico. A continuación, se presentan y describen detalladamente el resultado promedio de cada modelo con un numero de evaluaciones de $n = 3$.

Modelo	MSE Promedio	MAE Promedio	R2 Promedio
XGBoost	1.365	0.874	0.935
LightGBM	2.983	1.311	0.857
RandomForest	4.425	1.473	0.788
ExtraTrees	5.506	1.759	0.736
BayesianRidge	5.805	1.871	0.722
Ridge	5.805	1.872	0.722
Huber	5.845	1.856	0.720
GradientBoosting	6.330	1.950	0.696
KNN	9.267	2.359	0.555
DecisionTree	9.841	2.116	0.528
AdaBoost	12.481	2.948	0.401
OrthogonalMatchingPursuit	12.517	2.824	0.399
PassiveAggressive	12.873	2.909	0.382
ElasticNet	19.805	3.605	0.050
Lasso	20.844	3.699	0.000
LassoLars	20.844	3.699	0.000
Dummy	20.844	3.699	0.000
Lars	6759.560	6.945	-323.292

Tomando en cuenta los resultados obtenidos en nuestras evaluaciones, hemos seleccionado XGBoost como el modelo a implementar.

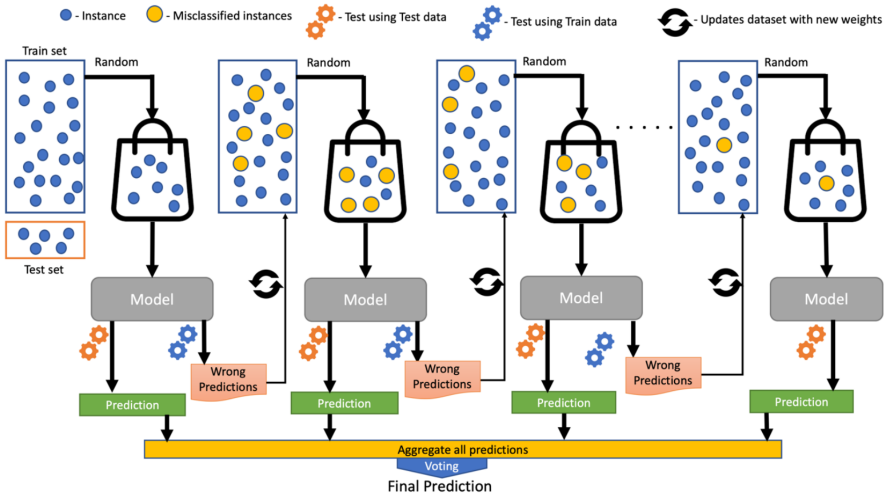
4. XGBoost

4.1. Descripción

XGBoost (eXtreme Gradient Boosting) es un algoritmo de aprendizaje supervisado que se implementa como un conjunto optimizado de árboles de decisión. Es una extensión del algoritmo Gradient Boosting Machine (GBM), diseñada para mejorar la velocidad y la eficiencia del modelo.

Para ajustar un dataset de entrenamiento utilizando XGBoost, se realiza una predicción inicial. Los residuales se calculan en función del valor predicho y de los valores observados. Se crea un árbol de decisión con los residuales utilizando una puntuación de similitud de los residuales. Se calcula la similitud de los datos de una hoja, así como la ganancia de similitud de la división posterior. Se comparan las ganancias para determinar una entidad y un umbral para un nodo. El valor de salida de cada hoja también se calcula mediante los residuales. Para la clasificación, los valores se calculan generalmente utilizando el registro de momios y probabilidades. La salida del árbol se convierte en el nuevo residual para el dataset, que se utiliza para construir otro árbol. Este proceso se repite hasta que los residuales dejan de reducirse, o bien el número de veces especificado. Cada árbol subsiguiente aprende a partir de los árboles anteriores y no tiene asignado el mismo peso.

En general XGBoost mejora los árboles de decisión al utilizar el principio de 'boosting', un enfoque de aprendizaje de conjunto donde nuevos modelos se añaden para corregir los errores cometidos por los modelos existentes.



Modelo de Árbol de Decisión

Un modelo de árbol de decisión representa una serie de decisiones basadas en las características de entrada que conducen a una predicción de valor continuo para la regresión.

Gradient Boosting

El Gradient Boosting construye iterativamente nuevos modelos para corregir los errores del modelo anterior. El modelo final es una suma ponderada de los árboles de decisión.

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i)$$

donde:

- $\hat{y}_i$ es la predicción para la observación.
- x_i son las características de entrada.

- f_k es el k-ésimo árbol de decisión.
- K es el número total de árboles

Función de Pérdida y Gradiente

XGBoost minimiza una función de pérdida personalizable $L(y_i, \hat{y}_i)$, que mide la diferencia entre las predicciones y los valores reales. Durante el entrenamiento, XGBoost utiliza el gradiente de esta función de pérdida para actualizar el modelo de manera que reduce el error de predicción.

Regularización
Utiliza regularización en la función de pérdida para controlar la complejidad del modelo, lo que ayuda a prevenir el sobreajuste:

$$\text{Obj}(\theta) = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

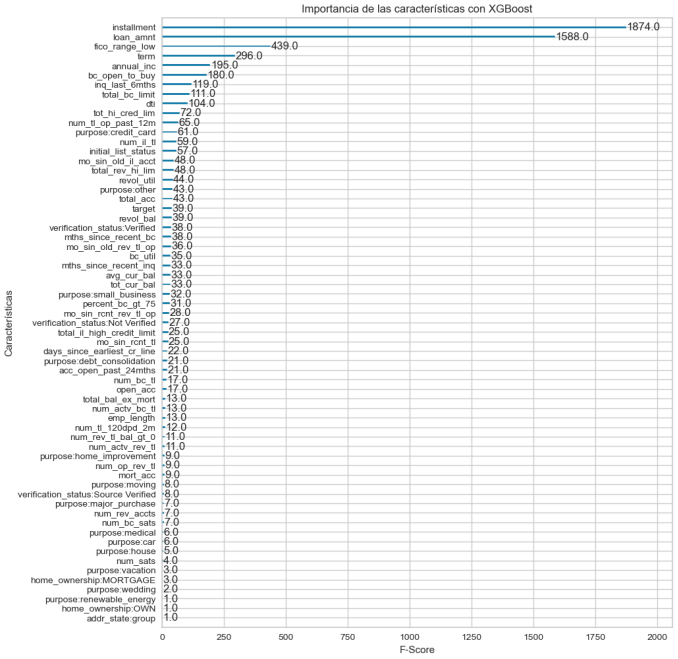
donde Ω es la función de regularización que penaliza la complejidad de los árboles.

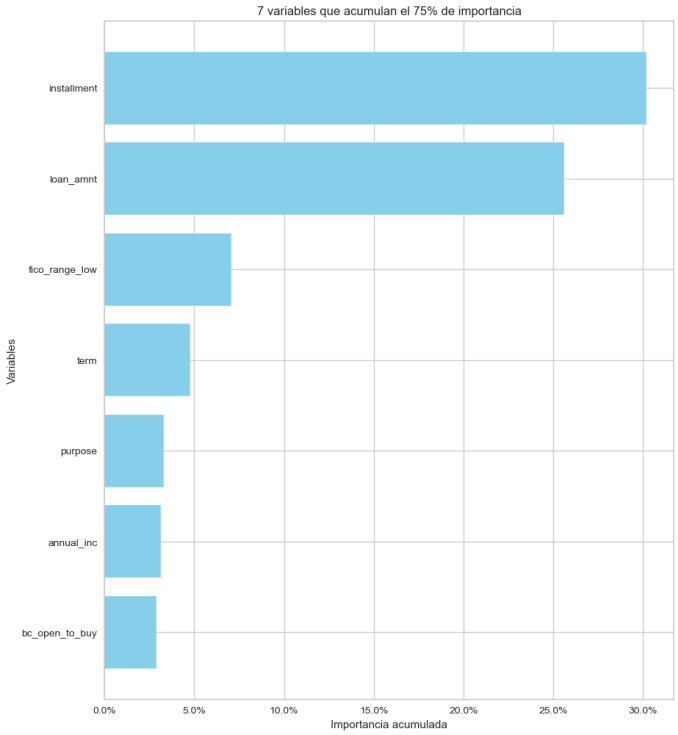
4.2. Optimización de XGBoost

La optimización del modelo XGBoost se ha realizado en dos etapas clave para mejorar la eficiencia y la eficacia del modelo de regresión. La primera etapa se centró en la selección de características, y la segunda en la afinación de hiperparámetros.

4.2.1. Selección de Características

Se implementó un proceso de selección de características basado en la importancia asignada por el modelo XGBoost. Se identificaron y conservaron las características que suman el 75 % de la importancia total, permitiendo que el modelo se concentre en la información más relevante. Esta estrategia de reducción dimensional no solo mejora la velocidad de entrenamiento del modelo, sino que también puede contribuir a la generalización al evitar el sobreajuste a las características menos significativas.





4.2.2. Ajuste de Hiperparámetros

Posteriormente, se llevó a cabo una búsqueda exhaustiva de hiperparámetros utilizando la técnica de GridSearch. Este método sistemático explora una cuadrícula de valores de hiperparámetros y evalúa el rendimiento del modelo para cada combinación utilizando una estrategia de validación cruzada. El objetivo es identificar el conjunto de hiperparámetros que resulta en la mejor precisión predictiva del modelo.

Los hiperparámetros ajustados incluyen:

1. **'learning_rate'**:
   - También conocido como tasa de aprendizaje o 'eta' .
   - Determina el tamaño del "paso"que el modelo toma al actualizar los pesos con cada árbol añadido.
   - Valores más bajos hacen que el modelo sea más robusto a la varianza del conjunto de entrenamiento, pero pueden requerir más árboles (`n_estimators`) para converger a la mejor solución.
   - Valores más altos pueden conducir a un aprendizaje más rápido, pero también a un mayor riesgo de sobreajuste.
2. **'max_depth'**:
   - Representa la profundidad máxima permitida para cada árbol.
   - Controla la complejidad del modelo. Árboles más profundos pueden capturar relaciones más complejas, pero también pueden sobreajustarse.
3. **'subsample'**:
   - Fracción de muestras que se utilizarán para entrenar cada árbol.
   - Si es menor que 1, el algoritmo seleccionará aleatoriamente un subconjunto de los datos antes de construir cada árbol, lo que puede ayudar a prevenir el sobreajuste y añadir aleatoriedad al proceso.

4. **‘colsample_bytree‘**:
    - Es la fracción de características que se utilizarán para entrenar cada árbol.
    - Un valor menor que 1 significa que cada árbol no utilizará todas las columnas al entrenar, lo que ayuda a evitar el sobreajuste y mejora la velocidad del entrenamiento.
5. **‘reg_lambda‘** (Regularización L2 o Ridge):
    - Es el término de regularización L2 en la función de coste, que penaliza los pesos más grandes.
    - Ayuda a controlar el sobreajuste al penalizar modelos más complejos.
6. **‘reg_alpha‘** (Regularización L1 o Lasso):
    - Es el término de regularización L1, que añade una penalización equivalente al valor absoluto de la magnitud de los coeficientes.
    - Puede resultar en modelos más escasos (con coeficientes cero), lo que puede ser beneficioso si se sospecha que algunas características no son informativas.
7. **‘n_estimators‘**:
    - Es el número de árboles de decisión que se construirán en el modelo.
    - Un número mayor de árboles puede aumentar el rendimiento del modelo hasta cierto punto, pero también puede llevar a un tiempo de entrenamiento más largo y posiblemente a un sobreajuste si no se combina con técnicas de regularización o paradas tempranas.

Tras aplicar el algoritmo de búsqueda, la selección de Hiperparámetros es:

Hiperparámetro	**Valor**
colsample_bytree	0.5758
learning_rate	0.0673
max_depth	5
n_estimators	998
reg_alpha	24.0463
reg_lambda	34.2127
subsample	0.6778

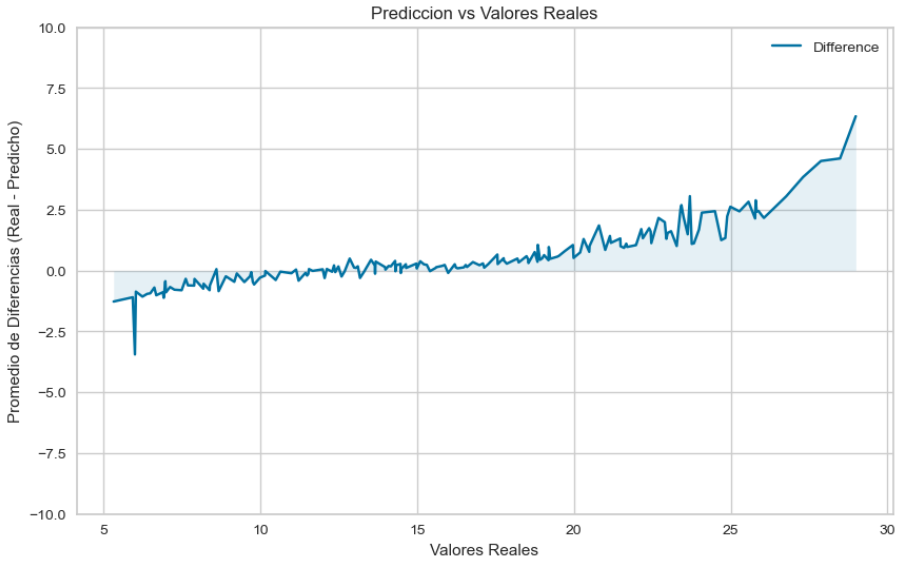
5. Resultados

El modelo implementado exhibió un rendimiento consistente, alcanzando una precisión cercana al 93 % en la validación cruzada con los conjuntos de entrenamiento y prueba. Este nivel de eficiencia sugiere que el modelo posee una capacidad robusta y fiable para la predicción de tasas de interés en nuevos créditos.

Datos	R2
Entrenamiento	0.9334
Prueba	0.936

El análisis adicional revela que el modelo tiene una tendencia a subestimar las tasas de interés en créditos cuyas tasas reales exceden el 25 %. Esta inclinación podría atribuirse a una muestra insuficiente de datos en ese segmento específico de tasas de interés.

Por otra parte, para créditos con tasas de interés que oscilan entre el 5.32 % y el 21.1 % —un rango que comprende el 95.09 % del conjunto de prueba— el modelo mostró un error promedio de menos de un punto porcentual. Este margen de error se considera mínimo, lo que evidencia un rendimiento sobresaliente del modelo para la mayoría de los datos evaluados.



bins	Error_Promedio	Observaciones	% Observaciones	% Observaciones Acumuladas
(5.32, 7.95]	-0.75	8726	16.70	16.97
(7.95, 10.58]	-0.40	7404	14.17	31.38
(10.58, 13.21]	0.00	11717	22.43	54.17
(13.21, 15.84]	0.13	10498	20.10	74.59
(15.84, 18.47]	0.25	6578	12.59	87.39
(18.47, 21.1]	0.65	3961	7.58	95.09
(21.1, 23.73]	1.54	1549	2.97	98.11
(23.73, 26.36]	2.34	920	1.76	99.89
(26.36, 28.99]	3.93	54	0.10	100.00

6. Referencias

- https://pro.arcgis.com/es/pro-app/latest/tool-reference/geoai/how-xgboost-works.htm
- https://es.wikipedia.org/wiki/Coeficiente_de_determinaci%C3%B3n
- https://en.wikipedia.org/wiki/Mean_squared_error
- https://es.wikipedia.org/wiki/Error_absoluto_medio