
# DATATHON

Josefina Amicone
Mercedes Mendez
Denise Jones

Detalle del Trabajo Realizado

1. **Introducción**

El presente informe detalla el proceso de desarrollo de un modelo de predicción de ingresos basado en datos históricos. Se aplicaron técnicas avanzadas de preprocesamiento, imputación de valores faltantes, manejo de outliers y selección de hiperparámetros para obtener el mejor desempeño posible. El objetivo principal fue minimizar el error relativo medio (MAPE) en la validación del modelo, aprovechando las capacidades de XGBoost para capturar relaciones complejas en los datos y mejorar la precisión de las predicciones.

2. **Fundamentación del Modelo Elegido**

Para la predicción de ingresos, se optó por emplear el modelo XGBoost (Extreme Gradient Boosting) debido a su capacidad para modelar relaciones no lineales, su eficiencia computacional y la incorporación de técnicas de regularización que ayudan a prevenir el sobreajuste. Además, XGBoost se destaca en la optimización de hiperparámetros y en el manejo de grandes volúmenes de datos, lo que lo convierte en una alternativa robusta y precisa para problemas de regresión.

Se realizó un Grid Search con validación cruzada de cinco pliegues para encontrar la mejor combinación de hiperparámetros que minimizara el MAPE.

3. **Metodología para Imputación de Valores Faltantes**

Uno de los principales desafíos en el conjunto de datos fue la presencia de valores faltantes en las variables predictoras. Para abordar este problema se evaluaron dos estrategias:

- **KNN Imputer:** Rellena valores faltantes utilizando los k vecinos más cercanos.
- **Iterative Imputer:** Modela cada variable con valores faltantes como una función de las demás mediante regresión iterativa.

Se determinó que el uso de Iterative Imputer proporcionaba mejores resultados en términos de precisión, ya que aprovecha las relaciones entre variables para una imputación más informada.

4. **Manejo de Outliers**

Para evitar que los valores extremos afecten negativamente al modelo, se aplicó la técnica del Z-score a las variables numéricas. Se eliminaron aquellos registros cuyos valores Z absolutos superaran el umbral de 3, asegurando la remoción de datos atípicos sin alterar la estructura general del conjunto.

5. **Evaluación y Resultados**

Tras entrenar el modelo XGBoost con los mejores hiperparámetros obtenidos mediante Grid Search, se evaluó su desempeño en el conjunto de validación utilizando el MAPE, que mide el error porcentual promedio entre los valores reales y las predicciones.

El modelo final logró un MAPE de aproximadamente 12.29%, lo que indica una precisión aceptable para la predicción de ingresos en el conjunto de datos analizado.

6. **Conclusión**

Se implementó un modelo de regresión basado en XGBoost con un sólido preprocesamiento de los datos. La combinación de imputación iterativa, eliminación de outliers y optimización de hiperparámetros permitió explotar las capacidades de XGBoost para modelar relaciones complejas en los datos, obteniendo un rendimiento óptimo.

Finalmente, el modelo se utilizó para generar predicciones sobre el conjunto de prueba, y los resultados fueron guardados en un archivo .CSV para su posterior análisis.

| MODELO | MAPE | Nombre archivo |
|---|---|---|
| Original | 13.08 | Predicciones2 |
| XGBOOST | 12.69 | predicciones-gxboost |
| Outliers | 12.66 | prediccionesOUT |
| Hiperparametros | 13.08 | prediccionesHIPER |
| Transformacion de variables | 13.22 | PrediccionesTRA |
| Outliers con mejor imputacion | 12.64 | prediccionesOUTmej |
| outliers, hiperparametros e imputación mejorada | 12.35 | prediccionesOUTymas |
| outliers, hiperparametros con xgboost e imputación mejorada | 12.29 | prediccionesOUTmejXB |