Reporte Final - Grupo 17

Systeam

Integrantes

Del Rio, Juan Sebastián - 103337

Brizuela, Sebastián - 105288

Agha Zadeh Dehdeh, Lucía - 106905

Fecha de entrega: 02 de Noviembre de 2023

Introducción

Se realizó un análisis exploratorio sobre un dataset de reservas de hoteles, que tiene 31 columnas y 61913 filas, dónde se determinó a través de análisis y visualizaciones cuáles serían las variables relevantes. Luego se pasó al preprocesamiento de datos donde se detectó columnas con datos faltantes además outliers a los cuales se les aplicó su correspondiente tratamiento. Se exploraron distintos modelos de clasificación en la búsqueda de mejores métricas, para predecir si la reserva fue cancelada o no. Se inició con la construcción de árboles de decisión dónde se realizaron modificaciones en sus hiper-parámetros de forma arbitraria y utilizando la técnica “RandomSearchCV” y la técnica de poda en el proceso. Posteriormente se implementaron los modelos Knn y SVC y modelos de ensamblado: Random Forest y XGBoost, para luego entrenar modelos de ensamblado tipo híbrido: Voting y Stacking. Finalmente cómo modelo final se construyeron redes neuronales de clasificación con distintas arquitecturas.

Cuadro de resultados

CHP	Modelo	F1 Score	Precision Test	Recall Test	Accuracy	Kaggle
2	**Árbol**	0.85184	0.84518	0.83861	0.84570	0.84082
3	**Stacking** *(Mejor)*	0.86116	0.86498	0.86884	0.86379	0.86193
4	**Red Neuronal**	0.82974	0.76997	0.89957	0.81459	0.80829

Árbol: Se utilizó un Árbol de Decisión (DecisionTreeClassifier) en el cual se probaron diferentes técnicas para conseguir las mejores métricas. Primero se realizó una búsqueda de mejores hiper-parámetros arbitrarios, luego una búsqueda de hiper-parámetros arbitrarios pero con poda y por último se utilizó la técnica “RandomSearchCV” el cual dio el mejor modelo de predicción.

Stacking: Ensamble que entrena modelos base para combinar predicciones y luego usar un modelo meta para la estimación final, en este caso se usaron como modelos base al Random Forest y al Knn, y como modelo meta al XGBoost.

Red Neuronal: Consiste en una red neuronal con 6 capas de 100 neuronas, activación Relu para las capas ocultas, regularización L1 (0.01) y L2 (0.01) para controlar el sobreajuste, una dropout para mayor regularización y una capa de salida con activación sigmoide para la clasificación binaria, con optimizador=Adam, learning_rate=0.001, epochs=200 y batch_size=150.

Conclusiones Generales

Con base en el análisis exploratorio realizado en el dataset de reservas de hoteles, se han obtenido diversas conclusiones. En primer lugar, se realizó un análisis de los datos a los cuales se les aplicó un preprocesamiento el cual a la hora de entrenar modelos aportó un buena base para conseguir buenas predicciones, incluyendo la gestión de datos faltantes y la identificación y tratamiento de valores atípicos. En términos de simplicidad y velocidad de entrenamiento, XGBoost se destacó como el modelo más sencillo y rápido de entrenar; y fue el segundo mejor resultado obtenido en Kaggle. El modelo Stacking demostró ser el más efectivo en la competición de Kaggle.

En resumen, el análisis exploratorio y las decisiones tomadas en el preprocesamiento fueron importantes para mejorar las predicciones de los modelos. A pesar de que nuestros modelos los consideramos buenos para ser los primeros que realizamos, también es importante seguir explorando y experimentando con diferentes enfoques para lograr mejoras en la performance predictiva de nuestros modelos.

Las mejoras que se podrían haber realizado para mejorar los resultados de los modelos son continuar con la búsqueda de mejores hiper-parámetros, aplicar la técnica de reducción de dimensionalidad podría también representar una mejora significativa en el rendimiento y eficiencia del modelo, investigar proyectos de clasificación binaria para poder observar qué mejoras de hiperparametros aplica, investigar y poner a prueba distintas técnicas de regularización.

Tareas realizadas

Se realizaron todas las tareas siempre en grupo por reuniones en meet por lo tanto no hubo una división clara de las mismas (al menos de 2 a 3 reuniones semanales de aproximadamente 3 horas cada una). Sin embargo, cada uno por su cuenta le dedicó cierta cantidad de tiempo extra para el entrenamiento de modelos.