Objetivo del Proyecto

El objetivo principal de este proyecto es desarrollar un modelo predictivo que permita a Corporación Favorita anticipar y gestionar la demanda de productos en sus numerosas ubicaciones, utilizando como variable objetivo la cantidad de ventas unitarias por producto. El preprocesado de datos una fase para garantizar la calidad y la relevancia de la información utilizada para construir el modelo.

Progreso Alcanzado

Se han revisado los diferentes archivos .csv en búsqueda de datos nulos o no asignados. Dentro de esta limpieza de datos, se encontró que el archivo 'oil.csv' fue el único que cuanta con las características de valores 'NaN', en específico, la cantidad fue un 3.5%.

En la búsqueda de una comprensión más profunda y precisa de la demanda, se llevó a cabo un proceso para seleccionar tiendas de cada estado ecuatoriano y establecer un periodo de tiempo relevante. El objetivo de prever las ventas unitarias de todos los productos en tiendas específicas durante fechas del año 2016.

Por otra parte, se hizo una conversión de fechas en los dataframes respectivos con el motivo de poder trabajar con datos que sean de tipo float y no string. Igualmente se borraron columnas que no son importantes para el análisis del problema, gracias a la exploración de datos se pudo hacer esto.

Pipeline de Preprocesamiento

Para la construcción del modelo predictivo, se implementó un pipeline de preprocesamiento que integra dos clases de transformadores personalizados. Estos transformadores, 'MergeDataTransformer' y 'SplitDataTransformer', consolidan la información y la preparación de datos para el entrenamiento y evaluación del modelo.

MergeDataTransformer

Tiene como objetivo unir información clave de varios DataFrames con el DataFrame de entrenamiento train_store_state. Se realiza la fusión con las tiendas, datos de petróleo (oil), y eventos festivos (holidays_e). Este proceso consolida diversas fuentes de información en un solo conjunto de datos, proporcionando una base más completa para el análisis y modelado.

Algunas acciones que se realizaron:

- Ordenación de los datos por fecha para establecer una secuencia temporal.
- Manejo de valores nulos en el precio del petróleo (dcoilwtico) mediante un proceso de relleno basado en el valor posterior (backward fill).
- Llenado de valores nulos en las columnas 'type' y 'locale' con 'Normal'.

- Codificación one-hot de columnas categóricas seleccionadas, como 'store_nbr', 'item_nbr', 'onpromotion', 'city', 'type', y 'locale'.

SplitDataTransformer

La clase SplitDataTransformer se centra en dividir el conjunto de datos preparado en conjuntos de entrenamiento y prueba. Este proceso es crucial para evaluar la capacidad predictiva del modelo de manera efectiva. Las acciones realizadas por este transformador incluyen:

- Extracción de la columna 'unit_sales' como el objetivo 'y' y eliminación de esta columna de las características 'X'.
- División de los datos en conjuntos de entrenamiento y prueba, siguiendo una proporción predefinida (90% de entrenamiento, 10% de prueba).
- Establecimiento de la columna 'date' como el índice en ambos conjuntos.

X_train y y_train se utilizarán para entrenar el modelo, por otra parte, X_test y y_test se usan para medir el rendimiento (conjunto de prueba). Esta separación nos permite medir el desempeño de los modelos en datos no vistos antes.

Métricas Utilizadas

Según la competencia en Kaggle: los pesos, w_i , se pueden encontrar en el archivo items.csv (consultar la página de Datos). Los artículos perecederos tienen un peso de 1.25, mientras que todos los demás artículos tienen un peso de 1.00.

- Error Cuadrático Medio (MSE): Esta métrica proporciona una medida de la magnitud promedio de los errores cuadráticos entre las predicciones y los valores reales. Se calcula tanto en el conjunto de entrenamiento como en el conjunto de prueba.
- Error Absoluto Medio (MAE): Mide la magnitud promedio de los errores absolutos entre las predicciones y los valores reales. Similar al MSE, se calcula tanto en el conjunto de entrenamiento como en el conjunto de prueba.
- NWRMSLE (Normalized Weighted Root Mean Squared Logarithmic Error): Esta métrica es especialmente relevante en el contexto de la gestión de la demanda. Considera la naturaleza logarítmica de las ventas y aplica ponderaciones según el ítem específico. Esto es fundamental, ya que algunos productos pueden tener una influencia más significativa en el rendimiento general del modelo.

Los modelos utilizados para las primeras predicciones fueron: Regresión Lineal, ElasticNet, Lasso, Ridge. Estos tres últimos dieron resultados bastantes parecidos, con un error absoluto medio de aproximadamente 5 ventas unitarias por producto.

Conclusión

El proyecto de preprocesado de datos para el modelo predictivo de Corporación Favorita ha logrado avances significativos, pero se reconoce la necesidad visualizar los resultados de una forma más general. Además de tomar el modelo ‘ElasticNet’ y realizar predicciones de todo el dataframe de entrenamiento ya que en esta segunda entrega se utilizaron 78000 datos para entrenar los distintos modelos usados.