Primera entrega de proyecto

Carlos Esteban Cossio Gonzalez

Materia: Modelos y simulación I

Profesor: Raúl Ramos Pollán

UNIVERSIDAD
DE ANTIOQUIA
1 8 0 3

Universidad de Antioquia

Facultad de ingeniería

Colombia – Medellín – 2023

Descripción del problema

La venta de productos gastronómicos se ha convertido en un tema muy apetecido para los modelos predictivos ya que hoy en día se busca relacionar la óptima producción con la compraventa de productos. Si se hacen predicciones empíricas demasiado altas, los comerciantes de productos se quedan con un exceso de productos perecederos, lo cual es un problema en términos de desperdicio de comida. Por otra parte, si se hacen estimaciones bajas los artículos que son populares se agotan con facilidad, lo cual significa perdidas económicas para el supermercado. El problema que se busca solucionar es la predicción de la cantidad justa de productos adecuados en el momento adecuado.

Base de datos

Hablando de la base de datos que se utilizará, la cual proviene de una competición de la página 'Kaggle', es un compendio de varias tablas, en total son 8 archivos con extensión .csv. Los datos de entrenamiento comprenden fechas, información sobre la tienda y el artículo, si dicho artículo estaba siendo promocionado, así como las ventas unitarias. Además, entre los 8 archivos se encuentras otros data sets que pueden ayudar a la construcción del modelo requerido.

- **train.csv:** Incluye la variable objetivo unit_sales por fecha (date), store_nbr, item_nbr y onpromotion, unit_sales puede ser un dato entero o flotante, valores negativos indican una devolución. 16% de los datos en onpromotion son NaN. No hay datos para ventas ceros.
- **test.csv:** Datos de prueba, con las combinaciones de date, store_nbr, item_nbr y onpromotion que se van a predecir, parte del ejercicio consiste en predecir las ventas de un artículo nuevo basándose en productos similares.
- **stores.csv:** Almacena metadatos, store_nbr, city, state, type, cluster, la variable cluster es una agrupación de tiendas similares.
- **items.csv:** Metadatos de los items, item_nbr, family, class y perishable. Los ítems marcados como perecederos tienen una puntuación de 1.25; de lo contrario, el peso es 1.0.
- **transactions.csv:** El recuento de transacciones de ventas para cada fecha, los campos son: date, store_nbr, transactions. Solo se incluye para el período de tiempo de los datos de entrenamiento.
- **holidays_events.csv:** Datos de días feriados: date, type, locale, locale_name, description, transferred. Un feriado que se transfiere oficialmente cae en ese día calendario, pero el gobierno lo movió a otra fecha. Un día transferido se parece más a un día normal que a un día festivo. Para encontrar el día en que realmente se celebró, busque la fila correspondiente donde el tipo es Transferencia. Por ejemplo, el feriado Independencia de

Guayaquil se transfirió del 2012-10-09 al 2012-10-12, lo que significa que se celebró el 2012-10-12.

Métricas

Los modelos son evaluados en función de la métrica de evaluación; Error Logarítmico Cuadrado Medio Ponderado Normalizado (NWRMSLE), calculado de la siguiente manera:

$$\boldsymbol{NWRMSLE} = \sqrt{\frac{\sum_{i=1}^{n} w_i \left(ln(\hat{y}i + 1) - ln(y_i + 1)\right)^2}{\sum_{i=1}^{n} w_i}}$$

Donde para la fila i, $\hat{y}_i$ es el valor predicho de ventas unitarias de un artículo y y_i es el valor real de ventas unitarias; n es el número total de filas en el conjunto de pruebas.

Los pesos, w_i , se pueden encontrar en el archivo items.csv. Los artículos perecederos tienen un peso de 1.25, mientras que todos los demás artículos tienen un peso de 1.00.

Para la métrica de negocio tenemos las siguientes consideraciones:

Precisión de las Predicciones: Evalúa cuán cerca están las predicciones del valor real de ventas unitarias. Se penalizan más los errores en productos perecederos debido a su mayor peso.

Optimización de la Producción: La métrica incentiva a los modelos a hacer predicciones precisas para evitar excesos de productos perecederos, lo que minimiza el desperdicio de comida.

Satisfacción del Cliente: La métrica también incentiva a los modelos a predecir suficientes ventas para evitar agotar productos populares y, por lo tanto, reducir las pérdidas económicas para el supermercado.

En definitiva, el NWRMSLE es una métrica es adecuada cuando se predicen valores en un amplio rango. Además, evita penalizar diferencias grandes en las predicciones cuando tanto el número predicho como el número real son grandes. Predecir 5 cuando el valor real es 50 (10%) se penaliza más que predecir 500 cuando el valor real es 545 (90%)

Desempeño

Mantener el NWRMSLE por encima de un 80% garantiza que las predicciones sean mínimamente y suficiente precisas para pronosticar las ventas unitarias por producto lo que se traduce en un equilibrio óptimo entre el costo de producción y las

ganancias económicas. Si el modelo predictivo se mantiene por encima de un 80%, entonces el modelo es adecuado para su implementación.

Para cada ID en el conjunto de pruebas se debe predecir las ventas unitarias. Dado que la métrica utiliza $ln(y+1)$, se validan las presentaciones para garantizar que no haya predicciones negativas.

Bibliografía

- Kaggle. (2017). Favorita Grocery Sales Forecasting. Retrieved August 2023, from https://www.kaggle.com/competitions/favorita-grocery-sales-forecasting/overview