Capstone Three Final Report

Store Sales Times Series Forecasting

Carl Riemann

September 27, 2024

Corporación La Favorita is one of Ecuador's largest retail chains. With years of sales data available, building a reliable predictive model can help the company make data-driven decisions for future business strategies.

The objective of this project is to develop a time series forecasting model that predicts daily sales for product family of Corporación La Favorita using historical sales data from 2013 to 2017. The predictions aim to support the company's planning, on product sales.

A graph of sales by month

Description automatically generated

The datasets used for this project includes daily sales data from the company's various stores across Ecuador, spanning from January 1, 2013, to August 15, 2017. There are a total of 6 data frames containing features such as store location, product family, promotions, and sales volumes, along with additional factors like oil prices and holidays.

A graph of sales by product family

Description automatically generated

To prepare the dataset for modeling, the data wrangling process involved several key transformations. First, I focused on cleaning the dataset by removing irrelevant columns and filtering out data before 2015. Additionally, I grouped the data by date and product family to capture seasonal sales trends. New features, such as holiday and promotion effects, were also incorporated to better understand the factors influencing sales fluctuations.

A graph of a sales graph

Description automatically generated with medium confidence

During the EDA I visualized sales trends over time using line plots, which revealed clear seasonal patterns and spikes around holidays. I also explored the distribution of sales across different product families and stores through histograms and box plots, identifying outliers and variations in sales performance. I identified that the 2016 earthquake in Ecuador caused sales to increase, maybe from peoples panic and overstocking on products. After the EDA I created the ARIMA model, the 'sales' column as the target variable, along with 'date' and 'family' as primary grouping factors to capture seasonality and trends at the product family level.

The ARIMA model performed particularly well in predicting sales for categories with relatively stable demand, such as AUTOMOTIVE and HARDWARE. However, for families like BEVERAGES and GROCERY I, larger errors indicate that external factors or a more complex model may be required to improve accuracy. Categories in the middle range, such as CLEANING and FROZEN FOODS, show reasonable performance, but there is still room for improvement in capturing their sales trends more accurately.

A table of numbers and words

Description automatically generated with medium confidence