Project Proposal:
Time Series - Predict Unit Sales for Walmart

Springboard Data Science Track: Independent Capstone 3 Project
By: Brianna Abdalla

Problem:

How can Walmart predict daily unit sales for ten of their stores in California, Texas, and Wisconsin with 95% confidence for the next 28-day selling cycle, starting from 6/19/2016?

Description:

Walmart is a multinational retail corporation that has transformed the way people shop. Founded in 1962 by Sam Walton, the company has grown to become one of the largest and most influential retailers in the world. Headquartered in Bentonville, Arkansas, Walmart operates a vast network of stores, encompassing hypermarkets, discount department stores, and grocery stores.

Walmart's success is rooted in its commitment to offering customers everyday low prices. The company leverages its immense buying power and efficient supply chain to negotiate favorable deals with suppliers, allowing them to pass on cost savings to consumers. This strategy has resonated with shoppers, making Walmart a go-to destination for a wide range of products, including groceries, electronics, apparel, home goods, and more.

Having a strong logistics infrastructure is essential to Walmart's success. It is imperative that Walmart estimates unit sales for each store to forecast demand and ensure the stores are well supplied. Unit sales can also be a means of estimating revenues for Walmart and provide leaders with valuable insights on the company's future.

Walmart generously hosted their data for the M5 forecasting competition on Kaggle. The competition includes three datasets from Walmart: calendar information, historical daily unit sales, and finally sell prices. The data includes variables that may contribute to unit sales such as the food SNAP program dates, special holidays, day of the week and more. The goal is to merge and use these datasets to create a time series that forecasts daily unit sales for the next 28-day selling cycle and evaluates using weighted root mean squared error.

The goals for this project vary slightly from the M5 competition. The aim is to provide 28-day unit sales forecasts for each of the ten stores using MAPE for preliminary models and RMSSE for the final model. This approach targets use for store managers and gives them high level predictions for unit sales.

Criteria for Success:

The Walmart M5 datasets will be cleaned and merged. Underlying relationships with unit sales will be identified. The data will be tested on preliminary models starting with the Naïve approach and moving into ARIMA based models. The final model might include advanced models such as Profit or lightGBM. The final model selection will have the lowest weighted root mean squared error and provide predictions within a 95% confidence interval for the next 28 day selling cycle.

Scope of the Solution Space:

Each store will have a high level prediction for total unit sales. Other information derived from data exploration will be presented to managers for insights including an interactive dashboard with store level details.

Constraints within Solution Space:

- The merged datasets are very large and require a large amount of computational power.
- There are outliers in the data that could prove difficult for certain time series models.
- The sell price variable is the average sell price for each item across one week of sales. Predictions for sell price may not be as useful due to items showing unit sales but having a sell price of zero. In other words, sell price predictions would be a rough estimate of true revenues.

Stakeholders

- Walmart store managers
- CEO and high level management
- Buyer and Planners

Key Data Sources:

- calendar.csv - Contains information about the dates on which the products are sold.
- sales_train_validation.csv - Contains the historical daily unit sales data per product and store [d_1 - d_1913]
- sell_prices.csv - Contains information about the price of the products sold per store and date.
- Link to data in M5 competition: https://www.kaggle.com/competitions/m5-forecasting-accuracy/data

Problem Approach Plan:

1. Load, clean and merge data.
2. Explore how different variables effect unit sales
3. Explore, test, and visualize time series components including decomposition, stationarity, ACF and PACF.
4. Begin preliminary modeling on the top store. This includes creating train test splits and preprocessing for these models: Naïve, Linear, Exponential Smoothing, and ARIMA based models.
5. Select preliminary model based on MAPE and continue to final model forecasting for each store.
6. The final model selection may include an advanced model such as Profit or LightGBM and be graded on the lowest RMSSE score.
7. Create an interactive dashboard on Tableau that provides insights to store managers.

Deliverables:

- PDF Proposal
- GitHub repository:
  - Notebooks for each step in the data science process.
  - A file of the final model
  - A file of final model metrics
  - A file for x and y variables from final model
- An interactive Tableau dashboard
- Presentation Slide Deck
- PDF Project Final Report