Forecast The Weekly Sales for The Drug Store Rossmann

Problem Statement

This project looks into what kind of insights can the drug store Rossmann gain from their historical sales, promotion, school/state holidays and competitors data. Also it looks into how they can use them to optimize the store operation and management in order to generate more sales revenue. Plus how can they use these data sets to forecast the weekly sales for each store with a certain level of accuracy?

Context

Rossmann operates over 3,000 drug stores in 7 European countries. Currently, Rossmann store managers are tasked with predicting their weekly sales for up to eight weeks in advance. Store sales are influenced by many factors, including promotions, competition, school and state holidays, seasonality, and locality. With thousands of individual managers predicting sales based on their unique circumstances, the accuracy of results can be quite varied. Therefore the company's data science team is on a new mission to create a unified modeling method for managers to forecast the weekly results with higher accuracy. The management team also required an overall report with feasible strategies for the company to understand the general performance of all stores and to come up with a way to optimize future sales performance. Lastly an individual store performance report needs to be provided to each store manager.

Data Source

The datasets used in this study are sourced from https://www.kaggle.com/c/rossmann-store-sales/data and consists of two datasets: train.csv, store.csv

- train.csv containing 1,017,209 data samples, is composed of a total of 9 variables collected from 1115 Rossmann stores over 942 days from 01/01/2013 to 07/31/2015, with 7 numerical variables and 2 categorical variables.
- store.csv containing 1115 data samples, is composed of a total of 10 variables collected from 1115 Rossmann stores, with 6 numerical variables and 3 categorical variables.

Criteria For Success

- Determine the key features that influence sales and provide feasible strategies to the management team for future business planning.
- Build a unified forecasting model such as time series modeling for each store based on the historical sales data, or with additional key variables if necessary , ensuring the model performs better than the average method (forecasting with the average value of the historical data).
- Generate reports containing information of overall performance of 1115 stores and individual performance reports for each store.

Constraints & Scope

- Building an unified forecasting model for all the 1115 stores. Some of the stores stayed closed for a few months or up to more than half a year due to renovation and generated zero weekly sales over this period. Having zero values in the observations will add difficulties in producing time series models with high accuracy.
- There were random spikes observed in daily sales and customers in some stores, regardless of running a promotion or not, being a state/school holiday or not. The information in the current two data sets are very limited in terms of understanding the spikes and could be challenging to inference the causes of them.
- Identify the key features as additional/exogenous variables that go into the time series model. There will be 18 variables if we combined train.csv and store.csv together. Filtering out which variables are influential in predicting sales from all the variables can be time consuming.

Approach

Multiple steps will be taken to build a predictive model for this project as well as to analyze the resulting predictions.

1. The train.csv and store.csv will be imported and cleaned via Python. Mis-represented data types will be correct to the right type. The missing values in store.csv will be handled with the optimal imputation techniques based on each variables' distribution and their correlations with other variables.
2. The daily sales in train.csv will be aggregated to weekly sales for each store for further use in modeling.

3. In the EDA (exploratory data analysis) part, the daily sales, customers, open days, promotions, and holidays of 1115 stores will be aggregated to a total sum over the 942 days. New metrics as average daily sales, average daily customers and sales per customer will be created in order to analyze store performance fairly as some stores were closed for a certain period of month due to renovation, so the total open days varies between stores.
4. After finishing step 3, we will combine the dataset with store.csv, and conduct the EDA based on this combined data set. After the EDA, an overall performance report will be generated for the Rossmann's management team.
5. For the individual store report we will provide codes enabling reports creation by only inputting a store ID.
6. For the modeling part, the weekly sales data we prepared in step 2 will be used as observations, split in training and testing sets. Ensure the testing set contains 8 weekly sales data. Time series models such as AR, MA, ARIMA or auto ARIMA will be compared with the average method and the one with best performance will be chosen as the final forecasting model.The forecasting results for future eight weeks will be provided with 95% confidence interval. We will also provide codes to simplify the whole process for managers by only inputting a store ID.

Deliverables

The final draft of the project will be presented in the form of a slide deck and formal project report to analyze the overall store performance. Jupyter Notebooks for overall analyzing, individual store reporting and individual store modeling will be delivered detailing each step taken and code written for the analysis and modeling of the project. A Github repository and an interactive dashboard for the project will be created as well.