ML – Regression

ROSSMANN RETAIL SALES PREDICTION

Almabetter, Bengaluru

Team Members	
Varsha Rani	**Vivek Chandrakant Pawar**
Rabista Parween	**Tushar Gaikwad**

1. Problem Statement:

2. Rossmann operates over 3,000 drug stores in 7 European countries. Currently, Rossmann store managers are tasked with predicting their daily sales for up to six 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.

3. Introduction:

We are given two data sets :

Rossmann Stores Data.csv - historical data including Sales

store.csv - supplemental information about the stores

4. Objective:

The objective of this project is to forecast the "Sales" column for the test set.

5. Steps Involved:

Importing Libraries

We have used the following Libraries :

- Pandas
- Numpy
- Matplotlib
- Seaborn
- Sklearn

Downloading and Checking Data

We have loaded the given dataset and created dataframe. Checked the data and merged both the dataset **"Rossmann Stores Data.csv"** and **"store.csv"** for further implementations.

Data Manipulation

We splitted Date column into Year-Month-Day and week of year.

Exploratory Data Analysis

- When stores are closed
- Effect of running promotional ads on sales
- Sales according to days of week
- Promotional Ads and customer count
- Percentage shop running continuous promotions
- Different store types
- Total sales per store type
- Total number of customers per store type
- Average sales per store type
- Average customer per store type

- Average customer spending time per store type
- Assortments and store types
- Sales Analysis

Feature Engineering

Training Model

Data Modeling

- Decision Tree Regression
- Gradient Boosting
- Random Forest
- XG Boost

Validating Model

Hyper Parameter Tunning

Conclusion

Conclusion on EDA:

- Over those two years, 172817 is the number of times that different stores closed on given days.
- From those closed events, 2263 times occurred because there was a school holiday.
- For Closed Event 30140 times it occurred because of either a bank holiday or easter or Christmas.
- After reading the description of the this task, Rossman clearly stated that they were undergoing refurbishments sometimes and had to close. Most probably those were the times this event was happening.
- The best solution here is to get rid of closed stores and prevent the models to train on them and get false guidance

- Analysis clearly states that if we run promotional advertisement then there is jump in average sales, So we can say running promotion is beneficial for stores.
- If we look over the years, there is a slight increase Year over Year but we don't see any major change from 2013 to 2015 and we actually see a very similar pattern in the months over the years with major spikes first around Easter period in March and April then in Summer in May, June and July and then finally around the Christmas period in November and December.
- For Sunday since a very few stores opens on Sundays (only 33);if anyone needs anything urgently and don't have the time to get it during the week, he will have to do some distance to get to the open ones even if it's not close to his house. This means that those 33 open stores on Sunday actually accounts for the potential demand if all Rossman Stores were closed on Sundays. This clearly shows us how important it is for stores to be opened on Sundays.
- Store type a has the highest number of stores, sales and customers from the 4 different store types. But this doesn't mean it's the best performing Store type.
- When looking at the average sales and number of customers, actually it is Store type b who was the highest average Sales and highest average Number of Customers. One assumption could be that if b has only 17 stores but such a high amount of average sales and customers, whereas a would be smaller in size but much more present.
- Surprisingly it is Store Type d who has the highest average spending per Customer, this is probably explained by an average competition distance higher than the rest which means each customer will buy more since he knows there isn't a lot of similar shops around.

- We can clearly see here that most of the stores have either a assortment type or c assortment type.
- Interestingly enough Store Type d which has the highest Sales per customer average actually has mostly c assortment type, this is most probably the reason for having this high average in Sales per customer. Having variety in stores always increases the customers spending pattern.

Conclusion on Data Modeling:

- We can understand from this project the flexibility and robustness of a decision tree based model like RandomForest which helped us predict the Store Sales of Rossman based on attributes that defines each store and its surroundings
- As we can see, it always delivers a good predicition score while not having a lot of modifications and difficulties capturing the patterns hidden in the data. Fortunately we had a train set that was large enough for it to converge but in general RandomForest performs not so bad on small sets since its resampling method (bagging) and its random production of trees allow the bias to remain not so high and in this case always performs good on unseen data where as XGboost has tendency to overfit if not gently and smartly tuned.
- I believe using hyperparameter optimization techniques like Gridsearch and RandomizedSearch is crucial to any Machine Learning problem since it allows the algorithm to not just limit itself on its defaulted parameters but to discover new opportunities of combining those parameters to reach a local optima while training on the data.