Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Member's Name, Email and Contribution: (Individual project)

Member : Harshad Savle
Email: harshad.savle@gmail.com
Tasks :

1. Clean and prepare the data for analysis.
2. Done Initial analysis and visualization.
3. Prepared Project Summary
4. Prepared Key Notes and conclusion
5. Done the visualization for analysis.
6. Added Useful Codes to simplify the analysis.
7. Prepared conclusions and PPT
8. Prepared introduction and key finding
9. Prepared Technical Documentation
10. Prepared Project Presentation

Please paste the GitHub Repo link.

GitHub Link: https://github.com/harshadsavle/store_sales_prediction

Google Drive link:
https://drive.google.com/drive/folders/1kHiv9B_3TGT0ywPmNT9QyeP2NxyHrgPe?usp=sharing

Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)

Problem Statement:

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.

You are provided with historical sales data for 1,115 Rossmann stores. The task is to forecast the "Sales" column for the test set. Note that some stores in the dataset were temporarily closed for refurbishment.

This dataset has around 1017209 observations in it with 18 columns and it is a mix between categorical and numeric values.

Analysing the data to discover key understandings (not limited to these) such as:

- On which day of the week sales is highest?
- On which week of the year sales is the highest?
- Are sales affected by any holiday or not?
- Which store type has the maximum number of sales?
- Information regarding stores who are continuing with their promo and who are not.

Our Approach:

Our first step was to understand the dataset of Rossmann store and then we started exploring and analysing each column. And we found out that we need to clean our dataset before performing the exploratory data analysis in order to get more accurate outcomes. Hence took the following steps:

- We treated null values of different columns accordingly. We saw that there are lots of null values present in the 'CompetitionOpenSinceMonth', 'CompetitionOpenSinceYear', 'Promo2SinceWeek', 'Promo2SinceYear'. 'PromoInterval' so we had drop that columns because if we treated those values it will manipulate data.

- We find out the outliers from columns using plotting the boxplot graph. If we have a look at the data and apply some domain knowledge, we easily understand that these values are not outliers. For example: If a store giving discount on some products which means number of sales increase so when we compline our data and plot box plot on yearly based data so some values shown as outliers because in some month or an specific occasions store giving some discounts.

- In Machine Learning we used algorithms such as Linear Regression, Lasso Regression, Ridge Regression, Decision Tree, Random Forest, Gradient Boosting and XG boost and fetched respective scores from all of these using the evaluation metrics such as MAE, MSE, RMSE, RMSPE, R2, ADJUSTED R2.

- We figured out which are the important columns. Our dataset was mostly influenced by customers and sales

- We made dummies to convert categorical variables to numeric ones.

- We did hyperparameter tuning to fetch respective scores from the evaluation metrics such as a MAE, MSE, RMSE, RMSPE, R2, ADJUSTED R2

Conclusion from EDA:

We can say that as assortment level b(extra) was most followed c(extended). We can conclude that most of the stores either used to keep extra mix types of products or extended ones.

Given that there is a linear link between customers and sales whenever a promotion is used, it can be deduced that the majority of customers came on sale days since the prices were lower.

Since the graph shows that sales were lower on the first days of the month than on the last days, it can be argued that individuals tended to shop for the end of the current month and the beginning of the following one. Those items might mostly be considered to be daily necessity.

It can be seen that average sales on Monday was more as compared to Sundays because mainly on holidays people prefer to do other things rather than shopping for their basic necessities or they might prefer to stay at home.

It can be assumed that school holidays make big difference in sales. It can be assumed that out of the total percentage of products a good percentage of products is meant for school students i.e. 17.8%.

Where competition was higher, sales was higher and vice versa.

The graph indicates that sales were particularly strong in the months of November and December, which were followed by a holiday, therefore it can be inferred that the majority of individuals are likely Christians.

Feature importance:

After comparing all features of the dataset we can figure out that customers have the highest importance followed by promo.

Conclusion from Machine Learning:

Performing various regression techniques, we can observe that XGboost Regression model have the better performance (with R2 : 0.988409) but after applying hyperparameter tuning on all our models we finally came to the conclusion that Random Forest Regression model have even higher performance (with R2 :0.994091) among the other models, as Random Forest Regression can handle large datasets efficiently and it's algorithm provides a higher level of accuracy in predicting outcomes over any other regression algorithm