SUMMARY

Name : Raja Chowdhury

Gmail:- rajachowdhury2468@gmail.com

GitHub Link:- https://github.com/RajaChowdhury/Retail-Sales-Prediction---Capstone-Project.git

Problem Statement:

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. So, we are assigned to make the best predicting model.

Our Approach:

Our first step was to understand the dataset and we started exploring and analysing each column. And we found out that we need to clean our dataset so we did treatments like replacing null values, handling outliers etc. After cleaning our dataset, we did some visualisation to find out some insights and know the business trend. Then we did feature engineering like feature selection, removing skewness from the dependent variable, minmax scaler to set numerical variable into the range of 0 to 1, created dummies for categorical columns and checked correlation among variables. Then our dataset was ready for ML Modelling and we applied different regression techniques to check the model performance and in order to improve the performance we used hyperparameter tuning. After selecting our model on the basis of hyper-tuned performance we tried to understand the importance of different features in our model.

Conclusions:

- Linear relationship has been found among customer, sales and promo. And it has been seen that most of the customers came for shopping during the promo days as the cost was lower on those days.
- Average sales on weekdays were more as compared to weekends because promos were provided to the customers during weekdays to increase the sales and not during weekends and reason might be that store use to remain close on Sundays.
- Sales during November and December month was higher compared to other months and that can be due to festive season in western European countries.
- Mostly competitor stores weren't that far and the stores were densely located near each other and also sales were higher when competition was nearer.
- School holidays also influence the sales as 17.8% of the sales gets affected by the school holidays which also means that around 17% of the sales are oriented from the school students.
- Performing various regression techniques, we found 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.