Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:
1. **Mohd Taufique**: taufiquemohd2@gmail.com • Data Exploration & Variables Identification. • Data Cleaning- Remove Duplicates, Handled & imputed Missing/Null Values. • Performed Exploratory Data Analysis (EDA) • Numerical Variable Analysis • ML Modelling and Conclusions • Presentation, Technical Documentation, Project Summary 2. **Azhar Ali**: aliazhar007007@gmail.com • Data Understanding- How data looks, how big the data is. • Data Wrangling- Removing duplicates and tackled null values • Categorical Variable Analysis • ML Modelling and Conclusions • Presentation, Technical Documentation 3. **Pushpam Raghuvanshi**: raghuvanshipushpam1991@gmail.com • Data Cleaning- Imputing null values, Changing datatype, Filtering of records • Feature Engineering, Variable Identification • Numerical Variable Analysis, Categorical Variable Analysis • ML Modelling and Conclusions • Presentation, Technical Documentation, Project Summary
Please paste the GitHub Repo link.
Mohd Taufique GitHub Link: - https://github.com/MOHD-TAUFIQUE/Retail-Sales-Prediction_ML-Regression-Project **Azhar Ali GitHub Link: -** https://github.com/Azhar-ali7/Regression-Project-Retail-Sales-Prediction **Pushpam Raghuvanshi GitHub Link: -** https://github.com/pushpam-raghuvanshi/rossmann_sales_prediction.git

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

Rossmann operates over 3,000 drug stores in 7 European countries. Currently, Rossmann stores 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 the results can be varied.

The problem statement was to forecast the "Sales" column for the test data. For the purpose we are provided with historical sales data for 1,115 Rossmann stores. We provided two datasets **i. Rossmann Stores Data.csv –** gives historical data including sales and **ii. Store.csv –** supplement information about the stores.

The first step in the analysis involved understanding the data, exploring the data, identifying the variables, and then performed data cleaning like removing the duplicates, anomaly correction, changing datatype of columns, looking for any null values and tackling them, filtering records, outlier detections and ways to dealing them. For the purpose, we import necessary python libraries, load the datasets, and used various pandas and NumPy built in functions.

The second step involved analyzing the different numerical and categorical features and show the analyzed result using different visualization charts like bar graph, clustered bar chart, box plot, pie chart, distplot, scatterplot etc. For this purpose, we used data visualization libraries – seaborn and matplotlib.

The third step involved feature engineering – dealing with categorical variables, multicollinearity, and at last after all these Data preprocessing steps we go for applying the Machine Learning algorithm to build a model that will predict the sales for the test data. For the purpose, we used Supervised ML models like Linear Regression, Ridge, Lasso, Decision Tree and Random Forest etc. and calculated evaluation metrics to check the performance.

The Final step involved is summing up the key observations and insights developed during the analysis and ML Model selection. Some key takeaways were: the relation between the sales and customers is sort of linear, i.e., the sales are increasing with the increasing number of customers which is obvious. Sales is more when Promo/Offers are running on stores, Highest sales recorded in December due to Christmas and New Year. Store with Assortment level 'b' has the highest sales. At last, based on the evaluation metrics we select the Random Forest Hyperparameter Tuned Model as it is giving the best accuracy compared to other regression models.