Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:		
Name	**Email**	**Contribution**
• **Saumya Dash**	**saumyadash9@gmail.com**	**Data Cleaning, Data Visualization, Feature Engineering, Modelling and PPT.**
• **Kumar Abhinav**	**kumarabhinavthakur274@gamil.com**	**Data Manipulation, Feature Engineering, Modelling, Recommendations and Technical Documentation.**
Please paste the GitHub Repo link.		
Github Link:-		
Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)		

Rossman Retail Sales Prediction

Rossman is a drug Store which operates over 3,000 stores in & European Countries. The task is to forecast the "Sales" column for the test set.

We are provided with two data sets Rossman Store Dataset including the Sales and Stores dataset as CSV files. We imported these two dataset in google colaboratory and converted it into data frame (df_sales and df_stores). Each row of the df_sales (Rossman Stores df) has attributes namely 'Store', 'DayOfWeek', 'Date', 'Sales', 'Customers' , 'Open' , 'Promo' , 'Stateholiday' , 'SchoolHoliday'. Similarly for df_stores (Store df) 'Store', 'StoreType', 'Assortment' , 'Competition Distance', 'Competition Open Since Month', 'Competition Open Since Year', 'Promo2', 'Promo2 Since Week', 'Promo2 Since Year' and 'Promo Interval'.

Then we performed various data cleaning operation such as dropping or substituting null values as per requirement, eliminated some irrelevant instances, extracted day, months and year from the date column and mapped some instances having similar values with different notations. We added some features to gain further insights.
After all the cleaning operation, we did data visualization and made some data transformation wherever required. We did Univariate Analysis, Bivariate Analysis and Multivariate Analysis to get some insights from the dataset and made some changes in the features suitable for Feature Engineering and Modelling. Moving towards Feature Engineering, we did One Hot Encoding for some categorical features and treated Multicollinearity using Heatmap and Variance Inflation Factor algorithm. We then proceeded with determining the Independent variables best suited to apply on the model for predicting our dependent variable i.e. 'Sales' column. We split the whole dataset in 70:30 ratio into Training dataset and Testing dataset and applied fit transform into the independent variables for standardization. Then we applied different Machine Learning algorithms such as Linear Regression, Lasso Regression, Ridge Regression, Decision Tree and Random Forest and fitted our dataset into them one by one. We used Hyperparameter Tuning in Lasso Regression by using Grid Search CV and tried to determine the best Alpha value to increase the accuracy of our model. Random Forest came out to be the best model among all with highest training(99%) and testing(97%) score. We determined some evaluation metrics for each models such as Root Mean Squared Error, Mean Absolute Percentage Error, R2 Score. We also determined the feature importance as per Random Forest and predicted our Target Variable.