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**
• Kumar Abhinav	Kumarabhinavthakur274@gmail.com	Data Cleaning, Feature Engineering, Modelling, Recommendations , Technical Doc.
• Saumya Dash	Saumyadash9@gamil.com	Data Manipulation, Feature Engineering, Modelling, Data Visualization, Ppt.
Please paste the GitHub Repo link.		
Github Link:- **https://github.com/kumarabhinavthakur274/Kumar_Abhinav_Rossmann_Sales_Prediction_Capstone_Project**		
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, is one of the largest drug store chains in Europe operates in 7 countries with more than 3000 stores. We are provided with historical sales data for 1,115 stores and we 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. Considering all these factors in mind we will prepare some machine learning models out of which we will try to achieve maximum training and testing accuracy and select the best models among them.

We have two datasets with us that is Rossman Sales dataset which contains details of daily sales in different stores and Store dataset which has the details of stores like store type , assortment , competition distance, promo etc. We imported these two datasets in colab notebook and then we checked for null vales ,we realized that first dataset had no null values whatsoever while second dataset had null values in 6 columns, we replaced Null values with 0 , median and mode as per the need. We eliminated some instances where the store was closed since closed shops signifies no sales for particular day. We also eliminated some instances where sales were 0 even if the store was open , these instances were 0.0067% of the total dataset so can be eliminated to avoid bias. After all the cleaning operation is done we moved to data Exploratory data analysis part where we initially did the univariate analysis where we did necessary transformation in continuous features after we did some bivariate analysis and multivariate to check the relation between the dependent variable Sales and other dependent variables , we drew some valuable insights as well which might proved to be useful for us . After that we did some feature engineering we did binary encoding for categorical features and also we manipulated some features to numerical one so that it can come handy during regression.

Now we came to the most important part modelling, we started with feature selection we determined dependent variable that is square root transformation of sales and independent variables which is well treated before putting into the model. Next we did test train split we kept 30% of given data for testing rest others we used for training. Then we used minmax scaler for feature scaling . Finally we started putting our data into various models. At first we put our data in linear regression model , we came up with 86% accuracy for both training and testing dataset , our model performance was good but we can improve the accuracy by solving the overfitting problem if any, for this we applied lasso and ridge regression where we tuned the hyperparameter using gridsearchcv , Our model performed same as earlier with best parameter as 10^-15 (or even less) which signifies that our model was not overfitting. So now to improve the accuracy we applied baseline model that is decision tree regressor in which our model performance improved significantly to 95% for both training and testing sets but this was not it , we applied random forest regression algorithm after that and this this we achieved 99% accuracy during training and 97% accuracy during testing. With this we came to the conclusion that random forest was the best model to be used for the sales prediction . Finally we plotted line plot of predicted values and actual values to verify our result we also printed a data frame for predicted values and actual values. Finally we gave our conclusion and recommendations based on our analysis.