Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:
Contributor Roles: 1. Pradeep Kumar Yadav: krpradeep0828@gmaill.com 1. Loading Libraries & Data 2. Analysing the rossmann data set 3. EDA on Rossmann Dataset 4. Conclusion of the analysis 5. Implementing supervised machine learning algorithm 6. LARS Lasso Regresion 7. Random forest with hyper parameter tuning 8. conclusion 2. Piyush M. Sonavane: piyushsonavane111@gmail.com 1. Loading Libraries & Data 2. Data Cleaning and Preparation 3. EDA on Rossmann Dataset 4. Model Implementation 5. Model Evaluation 6. Model Selection 7. Hyperparameter Tuning 8. Conclusion 3. Ganesh P. Patil: ganeshp746725@gmail.com 1. Reading and Understanding the Data 2. Creationg new variables or features for better understanding of dataset 3. Data Cleaning and Manipulations (Data Wrangling) and EDA 4. Model Building 5. Machine Learning Models Training and testing 6. Conclusion 4. Y Ishwar Rao: raoji4676@gmail.com 1. Data Loading 2. Data Wrangling a) Rossman Store Data b) Stores Data 3. Univariate Analysis 4. Bivariate Analysis 5. Machine Learning Models Training and Testing 6. Conclusions
Please paste the GitHub Repo link.

GitHub Link: https://github.com/krpradeep0828/Retail-Sales-Prediction
GitHub Link: https://github.com/piyushsonavane/Retail-Sales-Prediction
GitHub Link: https://github.com/Ganeshp30/Retail-Sales-Predictions
GitHub Link: https://github.com/Ishwar9109/Retail-Sales-Prediction

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

Retail sales is the sale of consumer goods, or final goods, by businesses to end consumers and includes in-store sales as well as online sales. Products may be durable (with a significant expected shelf life) or perishable (such as groceries). There are different retail stores in the market. They are specialty stores, department stores, supermarkets, convenience stores, and discount stores. This study is related to the sales of drug stores in European countries. These sales depend on state holiday, school holiday, day of the week and competition distance, etc.

Problem Statement:

Look at the given datasets and study the relationship between different features or trends in different features. Find the correlation between sales and other features and build an efficient machine learning model to predict future sales for given input variables. Rossmann operates over 3,000 drug stores in seven European countries. Currently, Rossmann store managers are tasked with predicting their daily sales for up to six weeks in advance. Many factors influence store sales, including promotions, competition, school and state holidays, seasonality, and location. With thousands of individual managers forecasting sales based on their specific circumstances, the accuracy of the results can vary greatly. We 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.

Approach:

We concatenated the two data sets Rossmann and Stores data to get more information about sales. We did replace null values with appropriate values like median and mode of that particular feature values. We had done univariate analysis and bivariate analysis for good understanding of the data. We converted categorical features into numerical features by using one hot encoding technique. Then we removed the features showing multicollinearity nature (removed features having VIF value more than ten). We did split the data set into train and test. We trained different algorithms using train data set. We tested algorithms like Linear Regression, Lasso, Ridge, Elastic-Net and Decision Tree Regression. We find that among these five algorithms, Decision Tree Regression giving good results.

Conclusions:

1. From correlation matrix, we can say that 'Customers' feature is highly correlated to Sales (dependent Variable).
2. The 'Month' feature was removed instead of week of year because these both features were correlated and 'Month' is less correlated with 'Sales' compared to later one.
3. In linear regression, Customers is the most influencing feature and State Holiday is at the second place.
4. In Decision Tree Regressor, Customers is the most influencing feature and Competition Distance is at the second place.
5. We find that among these five algorithms, Decision Tree Regression giving good results.

6. RMSE Comparisons (For Test dataset):

A. Linear Regression: 0.2456

B. Decision Tree Regressor: 0.157

C. Lasso Regressor: 0.286

D. Ridge Regressor: 0.245

E. Elastic Net Regressor: 0.306

7. R2 Score of test dataset:

A. Linear Regression: 0.666

B. Decision Tree Regressor: 0.863

C. Lasso Regressor: 0.544

D. Ridge Regressor: 0.666

E. Elastic Net Regressor: 0.480

8. Adjusted R2 of test dataset:

A. Linear Regression: 0.666

B. Decision Tree Regressor: 0.863

C. Lasso Regressor: 0.545

D. Ridge Regressor: 0.666

E. Elastic Net Regressor: 0.480