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. Jayanth V (Email – arambamvj@gmail.com) : 1) Data Overview. 2) Null values treatment. 3) Mapping string column into numeric features 4) Implementation of One Hot Encoding for feature enhancement. 5) Data inference using Correlation and removal of data with minimum correlation. 6) Random Forest Implementation. 7) Decision Tree Model Implementation. 8) Project structuring. 2. **Gowthaam Kumarasamy (Email - gowthaam02@gmail.com) :** 1) Data Visualization 2) Feature Engineering. 3) Outlier Treatment.. 4) Linear Regression. 5) Lasso Regularization. 6) Hyper Parameter Tuning. 7) Cross-Validation. 8) Technical Documentation.
Please paste the GitHub Repo link.

GitHub Link:- https://github.com/jayV1999/Machine-Learning.git

Google Drive Link: https://drive.google.com/drive/folders/1_CCxlzReZxA-co3d7w9YfFr_9gnGaYLp?usp=sharing

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

We are provided with historical sales data for 1,115 Rossmann stores.

The task is to forecast the "Sales" column for the test set. There are two datasets provided for us.

At first, we performed Data cleaning by creating a data frame and removing unnecessary columns with null values, dropping the duplicate rows in this dataset, and started getting basicinsight from the cleaned dataset.

We also performed Data Visualization to get basic insights into the effect that individual columns had on the sales column.

We merged the two datasets after performing basic cleaning and went with feature selection in which the numeric features in our dataset were taken into consideration.

we have also treated the outliers in the data by removing them using Z-score implementation.

After feature selection, we tend to split the data into training and testing pairs and started implementing basic Machine Learning Models to predict the sales

The Models Were :
a) Linear Regression
b) Lasso Regularization
c) Decision Tree
d) Random Forest

After Implementation, we found out that the Decision tree model is the one that gives us maximum accuracy in terms of prediction and to further improve on that we implemented Cross-validation and hyperparameter Tuning.

The main objective of sales forecasting is to paint an accurate picture of expected sales. Sales teams aim to either hit their expected target or exceed it.
When the sales forecast is accurate, operations go smoothly and future planning for the company's growth is done efficiently.

Upon having this analysis, it can be established that given the dataset, the model developed is able to explain 92.4649% of the variations and is able to predict the sales values in a good range.