Capstone Project Submission

Team Member's Name, Email and Contribution:

Name-Potdar Vinayak Sharadchandra

Email- vinayak.potdar.vp@gmail.com

Contributions-

- Data Wrangling on both datasets.
- Null Value Analysis and Treatment
- EDA performed after merging both the provided datasets.
- Feature Engineering
- Creation of Utility functions MAPE( ) and Performance(model_name) to evaluate model performance and error values.
- Model fittings and Performance Analysis
- Model Selection and Feature Importance calculation.

Please paste the GitHub Repo link.

Github Link:- https://github.com/vinayakpotdar2114/Capstone-Project-2-Sales-Prediction-Rossmann-Sales-Prediction

Name of Project: Rossmann Sales Prediction- Capstone-2

Summary of Project:

The dataset we are working here is of a drug store named Rossman. The dataset is present in two different tables, each containing a certain type of information. The experiment we are performing here is Sales Prediction using different machine learning models. Here, we try to understand the data using Exploratory Data Analysis and then start working with the machine learning models. The major focus is the prediction of Sales of the stores and the factors affecting the sales.

Components of Project:

● Loading Libraries and Data:
● Null values Analysis and Treatment:
● Correlation and Multi Collinearity Check:
● Exploratory Data Analysis:
● Feature Engineering:
● Splitting Dataset:
● Creating Utility Functions:
● Machine Learning Model Fittings (Regression Models):
● Hyper Parameter Tuning:
● Feature Importance

Problem Statement:

We are provided with historical sales data for 1,115 Rossmann Drug stores. The task is to forecast the "Sales" column for the test set.

Approach:

Loaded Libraries and Data required for the experiment. Did a quick overview of the data. Followed this by Null values Analysis and Treatment but refrained from dropping any feature. This was followed by Correlation and Multi Collinearity Check to understand the relation between various features with respect to Sales i.e., target variable and avoid features affecting the models adversely. Performed a Exploratory Data Analysis on selected features based on the correlation matrix data. Feature Engineering was the next step where less relevant features were dropped and categorical variables were encoded. This was proceeded by splitting the dataset and fitting the ML models. We created two utility functions here MAPE( ) and Performance(arg) for performance calculation. Hyper Parameter Tuning was part of the learning process with some relevant parameter values. Feature Importance calculation was done post model selection.

Conclusion:

Among the Machine Learning Algorithms that we tested; Decision Trees Regressor turned out to be the best fitting regression model for out dataset with testing score of 0.93 which is very high. The model might be on the verge of slight overfitting.