Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:

Prabhat Patel (prabhatpatel51@gmail.com)

Contribution- Analyzing and deriving new attributes from the given data set. Insights related to sales on a weekly, monthly basis. Stores in operation with types analysis. Implemented Linear Regression , Random Forest Regressor and Tuned the Random Forest Regressor using RandomizedSearchCV.

Annayan Bose (annayanbose@gmail.com)

Visualizing observations in the graphical representation holiday and distance effect and Observed some of the key factors such as outliers that were impacting sales prediction. Implemented models linear regression, Decision tree, Random Forest Regressor and tuned Random Forest Regressor using GridSearchCV.

Please paste the GitHub Repo link.

Github Link:- https://github.com/AnnayanB/SupervisedML_RetailSalePrediction.git

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

PROBLEM STATEMENT

The data set that was provided to us was of Rossman that operates over 3,000 drug stores in 7 European countries. Rossman store challenge was to predict 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.

So we were provided with historical sales data for 1,115 Rossmann stores and the task was to forecast the "Sales" column for the test set.

APPROACH

Exploratory Data Analysis:-

We performed exploratory data analysis with python to get insights from the data to observe following things:-

- There were more sales on Monday, probably because shops generally remain closed on Sundays which had the lowest sales in a week.
- The positive effect of promotion on Customers and Sales is observable.
- Most stores have competition distance within the range of 0 to 10 kms and had more sales than stores far away probably indicating competition in busy locations vs remote locations.
- Store type B though being few in number had the highest sales average. The reasons include all three kinds of assortments specially assortment level b which is only available at type b stores and being open on sundays as well.
- The outliers in the dataset showed justifiable behaviour. The outliers were either of store type b or had promotion going on which increased sales.

Models Implementation:-

1. Linear Regression
Fair Accuracy with low train and test score.
2. Decision Tree Regressor
Good train and test score.
3.Random Forest Regressor
Very good train and test score with good accuracy.
4. Random Forest Regressor with hyperparameter tuning
Very good train and test score with good accuracy.

Conclusions:-

- Decision tree was chosen as baseline model considering our features were mostly categorical with few having continuous importance.
- Random Forest shows improvement of 3.508% as compared to Decision tree.
- Random Forest Tuned Model gave the best results and only 0.023% improvement was seen from the basic random forest model which indicates that all the trends and patterns that could be captured by these models without overfitting were done and maximum level of performance achievable by the model was achieved.

Model Evaluation Parameter table

	MAE_train	MSE_train	RMSE_train	R2_train	Adj_r2_train	train_score	MAE_test	MSE_test	RMSE_test	R2_test	Adj_r2_test	test_score
Linear Regression	0.3777	0.2536	0.5036	0.7464	0.7464	0.7464	0.3796	0.2464	0.4964	0.7384	0.7382	0.7384
DecisionTreeRegressor	0.0246	0.0034	0.0582	0.9966	0.9966	0.9966	0.1872	0.0658	0.2565	0.9302	0.9301	0.9302
Random Forest Regressor	0.0455	0.0039	0.0626	0.9961	0.9961	0.9961	0.1424	0.0355	0.1884	0.9623	0.9623	0.9623
Random Forest Regressor tuned	0.0610	0.0069	0.0832	0.9931	0.9931	0.9931	0.1419	0.0353	0.1879	0.9625	0.9625	0.9625