Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution

Team Member	Email id
Vivek Pawar	Vivek16pawar@gmail.com
Varsh Rani	e.varsharani@gmail.com
Tushar Gaikwad	Gaikwadtushar140@gmail.com
Rabista Parween	2001rabista@gmail.com

Contribution	Responsibilities
Coding	Vivek Pawar
Presentation	Varsha Rani
Technical Documentation	Varsha Rani
Summary	Rabista Parween
Presentation Video	Individual

Git Hub Repository

Github Link :https://github.com/2001rabista/Regression_ML_Capstone

Summary

Problem Statement:

Rossmann operates over 3,000 drug stores in 7 European countries. Currently, Rossmann store managers 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. With thousands of individual managers predicting sales based on their unique circumstances, the accuracy of results can be quite varied.

Objective:

The objective of this project is to forecast the "Sales" column for the test set.

Steps Involved:

- Libraries Importing
- Data Loading And Data Inspection
- Data Manipulation
- EDA
- Feature Engineering
- Training Model
- Decision Tree Regression
- Gradient Boosting
- Random Forest
- XG Boost
- Validating Model
- Hyper Parameter Tunning

Conclusion:

Conclusion on EDA:

- Over those two years, 172817 is the number of times that different storesclosed on given days.
- From those closed events, 2263 times occurred because there was a schoolholiday.
- For Closed Event 30140 times it occurred because of either a bank holiday oreaster or Christmas.
- After reading the description of the this task, Rossman clearly stated that theywere undergoing refurbishments sometimes and had to close. Most probably those were the times this event was happening.
- The best solution here is to get rid of closed stores and prevent the models totrain on them and get false guidance.
- For Sunday since a very few stores opens on Sundays (only 33);if anyone needs anything urgently and don't have the time to get it during the week, hewill have to do some distance to get to the open ones even if it's not close to his house. This means that those 33 open stores on Sunday actually accountsfor the potential demand if all Rossman Stores were closed on Sundays. Thisclearly shows us how important it is for stores to be opened on Sundays.
- When looking at the average sales and number of customers, actually it is Store type b who was the highest average Sales and highest average Number of Customers. One assumption could be that if b has only 17 stores but such ahigh amount of average sales and customers, whereas a would be smaller in size but much more present.

Conclusion on Data Modeling:

We can understand from this project the flexibility and robustness of a

decision tree based model like RandomForest which helped us predict the Store Sales of Rossman based on attributes that defines each store and its surroundings

- As we can see, it always delivers a good predicition score while not having alot of modifications and difficulties capturing the patterns hidden in the data.Fortunately we had a train set that was large enough for it to converge but ingeneral RandomForest performs not so bad on small sets since its resamplingmethod (bagging) and its random production of trees allow the bias to remainnot so high and in this case always performs good on unseen data where as XGboost has tendency to overfit if not gently and smartly tuned.
- I believe using hyperparameter optimization techniques like Gridsearch andRandomizedSearch is crucial to any Machine Learning problem since it allows the algorithm to not just limit itself on its defaulted parameters but todiscover new opportunities of combining those parameters to reach a local optima while training on the data.

-