Rossmann Sales Prediction Summary

Team Member's Name, Email and Contribution:

1. **Mujtaba Ali** – mujtwa@gmail.com

- EDA
- Clean-Up
- Feature Engineering

2. **Prateek Sachdeva** – prateeksachdeva13@gmail.com

- Pre-Processing
- Model Implementation
- Model Explainability

Please paste the GitHub Repository & Google Drive link.

Github Link:- https://github.com/mujtwa/Rossmann-Sales-Project

Google Drive Link:- https://drive.google.com/drive/folders/1F1QA8M5-VkFhgPGMuGfiSUacHr0rEkuI?usp=share_link

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

- Dirk Rossmann GmbH, commonly referred to as Rossmann, is one of the largest drug store chains in Europe with around 56,200 employees and more than 4000** stores. In 2019 Rossmann had more than €10 billion turnover in Germany, Poland, Hungary, the Czech Republic, Turkey, Albania, Kosovo and Spain. The company was founded in **1972 by Dirk Rossmann with its headquarters in Burgwedel near Hanover in Germany. The Rossmann family owns 60% of the company. The Hong Kong-based A.S. Watson Group owns 40%, which was taken over from the Dutch Kruidvat in 2004.

- 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.

- Starting with loading the data so far we have done EDA, null values treatment, encoding of categorical columns, feature selection and then model building.

- In all of these models our accuracy revolves in the range of 89 to 98%.

- And there is no such improvement in accuracy score even after hyperparameter tuning.

- DayOfWeek was found to be the most important feature, which is contributing the

highest in predicting the target variable.

- We found 89% accuracy through linear regression.

- The Regularization techniques of linear regression(Lasso, and Ridge) did not help in improving the accuracy much.

- In decision tree regression, we found 97% accuracy.

- The ensembles of decision tree i.e. Random forest gave us the highest accuracy i.e. 98% in predicting the target variable.

- However Xgboost Regressor did not perform that great by giving us 93% accuracy.

- So we found Random forest to be the best performing algorithm with an accuracy of 98%.