**Capstone Project Submission**

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

|  |
| --- |
| **Team Member’s Name, Email and Contribution:** |
| This project was completed individually.  **Name**: Midhun R  **E-mail ID**: [connect.midhunr@gmail.com](mailto:connect.midhunr@gmail.com)  **Contributions**:   * Importing Libraries and Defining Functions * Reading Data * Data Inspection * Data Cleaning   + Handling Missing Data   + Merging Dataframes   + Conversion of Column Datatype   + Manipulation of Columns * Exploratory Data Analysis * Feature Engineering   + Feature Selection   + Handling Multicollinearity   + Handling Outliers   + Feature Transformation   + Categorical Feature Encoding * Modelling   + Train-Test Split   + Feature Scaling   + Model Training     - Linear Regression     - Ridge Regression     - Lasso Regression     - Elastic Net Regression     - Decision Tree Regression     - Random Forest Regression     - XGBoost Regression   + Model Comparison * Conclusion   + Conclusions from EDA   + Conclusions from Modelling |
| **Please paste the GitHub Repo link.** |
| GitHub Link: - <https://github.com/connect-midhunr/rossmann-sales-prediction> |
| **Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)** |
| Rossmann operates throughout Europe, with thousands of managers predicting company's daily sales. In this project, I have attempted to analyze the retail sales dataset of Rossmann stores and build a predictive model to forecast the sales over the next 6 weeks. It’s crucial that these projections be accurate since important decisions are driven by the revenue generated by the company.  Two datasets were given: one with store data and the other with sales data. After libraries were imported and the drive was mounted, data from the CSV files were read and stored in Pandas data frames. Store dataset contains 1115 records across 10 features with information about each store and sales dataset contains 1017209 records across 9 features with information regarding sales in stores from January 2013 to July 2015. The common feature among them is store ID.  Then data inspection was conducted and found that none of the datasets had duplicate rows but six columns in store dataset have missing values. Some columns also required conversion of datatype. Data cleaning was done to handle these irregularities. Two datasets were merged into one and new useful columns were created from existing ones.  Exploratory data analysis was performed to visualize the data and gain insights that can be later used to increase revenue. Some relevant conclusions were drawn that can be used by store managers to improve their sales.  Feature Engineering was done to prepare the dataset for modelling. Some useless features were dropped from the dataset, dataset was checked for multicollinearity and outliers in sales, customers and competition distance data were removed. Numerical features were transformed using log and square root to attain normal distribution and categorical data were encoded to convert them into numerical data.  Dataset was split into training and testing data for both input and target variables to use in model training. Independent variables were scaled, and models were built using seven different algorithms: linear, ridge, lasso, elastic net, decision tree, random forest and XGBoost regression.  All models were evaluated and compared using R-squared value. XGBoost gave unusually high score. So, the model using random forest algorithm was selected to avoid overfitting. If model interpretability is important for the stakeholder, then decision tree must be chosen. |