**Capstone Project Submission**

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

| **Team Member’s Name, Email, and Contribution:** |
| --- |
| Name: Ashutosh Sharma  Email: a[shutosh0626@gmail.com](mailto:Ashutosh0626@gmail.com)  Contribution: EDA, PPT, feature engineering, preprocessing, model building, model Evaluation  Name: Kumar Aman  Email: [amankumarss293@gmail.com](mailto:amankumarss293@gmail.com)  Contribution: EDA, Technical Documentation, Project Summary, Model building |
| **Please paste the GitHub Repo link.** |
| GitHub Link:- <https://github.com/ashutosh-sharma-xi/Rossman-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)**  **INTRODUCTION:**  **We were provided with an Open Source database namely Rossman Sales Data.**  We have to predict the sales of the Rossman Group whose stores operate in over 7 European Countries.  Understanding of Data:  There is a total of 1017209 rows and 8 features in Rossman stores data.csv and 1115 rows and 9 features in stores.csv. A few features were common in both datasets and after merging 17 columns were left. Some columns namely Promo2sinceweek, Promo2sinceyear, and PromoInterval have 49% null values while CompetetionopenSinceMonth and CompetetionOpenSinceYear have around 31% null values.  **The agenda of Our EDA was:**   * To find on which day the cumulative sales were highest. * Comparing Data distribution of various columns. * Finding Distribution of competitor distance from all stores. * Find what is the average sale per customer per store type. * Monthly distribution of sales * Anova analysis for sales over assortment.   **Key findings:**   * It can be seen that sales on Day 1 are high and it decreases till Day 4, a hike can be seen on Day 5, and Day 7 has the lowest sales. * Sales feature shows a high number of outliers. * Competition Distance is in a Pareto Distribution with most competitors within 10000 meters * Highest Per customer sale is done by Store D while Store B has the least sales. * Monthly sales distribution shows non-stationary data.   **Feature Engineering:**   * Some features had null values, and the null values were removed with 0 * Features like StateHoliday had mixed datatypes, they were converted to a single data type. * Performed binary and label encoding on other features. * Converted data to a numeric datatype. * Applied VIF * Date was showing high Multicollinearity so removed that. * Removed outliers of the Target variable with sales lower than Q1 – 1.5\*IQR.   **Data Preprocessing:**   * Defined Independent and Target Variable * Scaled independent variables * Train-Test Split; test size 20%   **Models Used:**   * OLS Regression * OLS with Grid SearchCV * L1 Regularisation- Lasso Regression * L2 Regularisation- Ridge Regression * Gradient Boosting Regressor   **Model Evaluation:**  Book1 - Excel (Product Activation Failed)  **Conclusion:**   * Some features were extremely important for the Prediction like Customer, Promo, Open, etc * As the Dataset was quite large the model generalized well and the chances of overfitting and underfitting were reduced. * In order to earn more focus company should focus more on Store d which generates the highest sales for the firm. * Gradient boosting regressor seems to be relatively very efficient with approximately 96% of r2 score * ordinary least square regression performed well with approximately 89% accuracy in both the train and test set, thus model wasn't looking highly biased or at high variance. * lasso and ridge didn't show any effective results in comparison to OLS * Company earn the most from Store type d thus it must maintain this, and should focus more on low-performing stores like store b. * On Sunday the market might be closed, resulting in Extremely fewer sales. * The Data could also be used for Time Series Analytics but as the data is from before 2015, now it is very less likely for the firm to use this in future predictions. * For around 0.65 million stores the competitor is present within a range of 5KM which shows that good competition is present in most of the cases but most of the competitors are established after the year 2000. |