**Capstone Project Submission**

**Instructions:**

1. Please fill in all the required information.
2. Avoid grammatical errors.

| **Team Member’s Name, Email and Contribution:** |
| --- |
| **Member 1: Saaquib Mustafa**  **Email:** [saaquibmustafa26@gmail.com](mailto:saaquibmustafa26@gmail.com)  **Contributions:**   1. Clean and prepare the data for analysis. 2. Added Useful Codes to simplify the analysis. 3. Done Initial analysis and helped in visualization. 4. Prepared Project Summary 5. Prepared Key Notes,conclusion and PPT     **Member 2: Raja Chowdhury**  **Email:** [rajachowdhury2468@gmail.com](mailto:rajachowdhury2468@gmail.com)  **Contributions:**   1. Done the visualization for analysis. 2. Helped in analysis and Data Cleaning 3. Added Useful Codes to simplify the analysis. 4. Prepared conclusions and PPT 5. Prepared introduction and key finding     **Member 3: Sandipan Das**  **Email:** [sandipan.das202@gmail.com](mailto:sandipan.das202@gmail.com)  **Contributions:**   1. Prepared Technical Documentation 2. Helped in Data Cleaning 3. Help in Summary Preparation & PPT 4. Helped in Key Notes     **Member 4: Sahil Kolambakar**  **Email**: sahilrajkolambkar@gmail.com  **Contributions:**   1. Prepared Project Presentation 2. Helped in Data Cleaning 3. Help in Summary Preparation & Technical Documentation 4. Helped in Key Notes and Conclusion |
| **Please paste the GitHub Repo link.** |
| **Github Link:-** [Saaquib01/Capstone-ML-Linear-Regression (github.com)](https://github.com/Saaquib01/Capstone-ML-Linear-Regression) |
| **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:**  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.  You are provided with historical sales data for 1,115 Rossmann stores. The task is to forecast the "Sales" column for the test set. Note that some stores in the dataset were temporarily closed for refurbishment.  This dataset has around 1017209 observations in it with 18 columns and it is a mix between categorical and numeric values.  Analysing the data to discover key understandings (not limited to these) such as:   * On which day of the week sales is highest? * On which week of the year sales is the highest? * Are sales affected by any holiday or not? * Which store type has the maximum number of sales? * Information regarding stores who are continuing with their promo and who are not.   **Our Approach:**  Our first step was to understand the dataset of the Rossmann store and then we started exploring and analysing each column. And we found out that we need to clean our dataset before performing the exploratory data analysis in order to get more accurate outcomes. Hence took the following steps:   * We treated null values of different columns accordingly. * We removed rows where the price was equal to zero. * We handled the outliers from columns using capping method * We figured out which are the important columns. * We did dummification to convert categorical variables to numeric ones. * In Machine Learning we used algorithms such as Linear Regression, Lasso, Ridge, Decision Tree, Gradient Boosting and XG boost and fetched respective scores from all of these.   **Key Findings from EDA:**   * From the first graph it is clear that 'a' Store Model has the maximum number of sales and store counts followed by 'd' while Store Model 'b' has the least number of sales and store counts. * From the second graph it is surprising to see that store model "b" which has the least number of store counts performs quite well on average sales and customers compared to other store models. * We can see from the graph that Basic Assortment level have the maximum number of sales and store counts followed by Extended level while Extra Assortment have the least number of sales and store counts * We can infer from the graph that assortment level 'b' with least store counts have performed quite well compared to 'a'. While there is another surprising fact that assortment level 'c' has the maximum number of sales with the least number of customers. * Here we can see from the graph that there is a linear relationship between customer and sales and it is also noticeable that whenever promo was open, the store has higher sales and customers compared to the period when promo was closed, which means promo had a good impact on the sales. * We can infer from the graph that the sales are highest on the 30th followed by the 2nd and 4th dates of every month while sales are lowest on the 1st date of every month followed by the 25th and 26th dates. * Here we can infer from the graph that the sales are at maximum on Mondays while sales are zero on Sunday because it seems like stores used to remain closed on Sundays. * Here we can see that the Sales and Customers are at peak during November and December due to festive seasons like Christmas while sales are at lowest during January and May or we may say these months to be off season. * Here we can see that during the Public holidays the store made more sales compared to Easter and Christmas holidays. * Here we can see that 17.8% of the sales get affected by the school holidays which also means that around 17% of the sales are oriented towards the school students.   **Conclusion from EDA:**   * We can say that as assortment level a(basic) was most followed c(extended). We can conclude that most of the stores either used to keep basic types of products or extended ones. * As we can see a linear relationship between customer and sales whenever promo is applied it can be concluded that most of the customers came for shopping during the promo days as cost was lower on those days. * As we can see in the graph that sales were low on the initial days of the month as compared to the final days it can be assumed that people used to shop for the next month and the end of the previous month. Those products can mainly be basic necessities of a person's daily life. * It can be seen that average sales on Monday was more as compared to Sundays because mainly on holidays people prefer to do other things rather than shopping for their basic necessities or they might prefer to stay at home. * As shown in the graph, sales during the November and December months were high followed by a holiday and it can be assumed that most people might be followers of Christian religion. * It can be assumed that school holidays make a big difference in sales. It can be assumed that out of the total percentage of products a good percentage of products is meant for school students i.e. 17.8%.   **Prediction Scores:**   * Linear Regression:- 0.961247 * Lasso- 0.958184129841043 * Ridge: 0.9612453650783583 * Decision Tree-0.989807065871467 * Gradient Boosting- 0.9752728974497129 * XGboost-0.9898649570522713     **Drive Link:** https://drive.google.com/drive/folders/1TGYW0CV16kBITzrxRg2ikRPvBZHUmpBe?usp=share\_link |
|  |