**Capstone Project Submission**

| **Team Member’s Name, Email and Contribution:** |
| --- |
| | **Name** | **Email** | | --- | --- | | **Ajit Varpe** | [ajitvarpe8652@gmail.com](mailto:ajitvarpe8652@gmail.com) | | **Harisha Chennozwala** | tasharisha@gmail.com | | **Niharika Soni** | niharika.soni1851@gmail.com | | **Satya Prakash** | [satya.prakash009@gmail.com](mailto:satya.prakash009@gmail.com) |   **Contributor Roles:**  **1. Ajit Varpe**  **1. Data Preparation.**  1.1 Feature Engineering - create new variable, avg sales, avg  customers,transform state holidays.  **2. Exploratory Data Exploration.**  2.1 Sales by Promotion #Promo  2.2 Sales By Promotion  2.3 Open  2.4 Assortment  2.5 Sales by Promotion #Promo  **3. Sales Prediction.**  3.1 K-Nearest Neighbor Regression  3.2 Model Selection  3.3 Feature Importance  **4. Others** :  Adding some of the conclusions, Worked on Project Summary.  **2. Harisha Chennozwala**   1. **Data Preparation.**   1.1 Data Loading.  1.2 Dealing With Missing Values.  1.3 Count missing values in each dataset  1.4 Remove features with high percentages of missing values.  **2. Exploratory Data Exploration.**  2.1 Sales Distribution - Avg Sales,Sales Percent change-drop date column  2.2 Customers Distribution.  2.3 School holiday  2.4 Sales by store type- merge with avg store sales & customers.  **3. Sales Prediction**  3.1 Linear Regression - Ordinary Least Squares  3.2 Facebook Prophet Model  **4. Others** :  Adding some of the conclusions, worked on PPT & technical  documentation .    **3. Niharika soni**  **1. Data Preparation.**  1.1 Drop Subsets Of Data Which Might Cause Bias.  **2. Exploratory Data Exploration.**  2.1 Correlation Heatmap.  2.2 Sales over days of month and month of year for different assortments  2.3 Sales Per Day.  2.4 Sales vs Number Of Customers.  2.5 Customer and Sales according to State Holiday  2.6 Converting StateHoliday to numerical values  2.7 Frequency of Sales Distribution  **3. Sales Prediction.**  3.1 Data-Splitting-(Train -Test data splitting)  3.2 Bayesian Ridge Regression.  3.3 LARS Lasso Regression.  **4. Others** :  Adding some of the conclusions, worked on PPT & technical  documentation .  **5. Satya Prakash**   1. **Data Preparation.**   1.1 Replace missing values in features with low percentages of missing  values.  1 .2 Date Extraction.  1.3 Joining Tables.   1. **Exploratory Data Exploration**   2.1 Seasonal wise distribution  2.2 Effect of promotion on sales  2.3 Notice that test\_df has only year=2015, and months 8 & 9  2.4 Risk Analysis   1. **Sales Prediction.**   3.1 Decision Tree Regression.  3.2 Random Forest Regression.  3.3 Facebook Prophet Model |
| **Please paste the GitHub Repo link.** |
| **Github Link:- https://github.com/Team-AlmaBetter/Retail-Sales-Prediction.git**  **Drive Link:- https://drive.google.com/drive/folders/1Z\_GeLBTEsMo7uLjQqCvKl3ES2Dc6FohI?usp=sharing** |
| **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:-**  Predicting sales performance is one of the key challenges every business faces. It is important for ﬁrms to predict customer demands to offer the right product at the right time and at the right place. The importance of this issue is underlined by the fact that ﬁguratively a bazillion consulting ﬁrms are on the market trying to offer sales forecasting services to businesses of all sizes. Some of these ﬁrms rely on advanced data analytics techniques, the kind of which we will be covering in this Project.  Rossmann is the largest drugstore in Germany. Moreover, it operates over 3,000 drugstores in 7 European countries. In 2015, Rossmann store managers are tasked with predicting their daily sales for up to six weeks in advance. Store sales are inﬂuencedby 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.  Explore and analyze the data to discover key factors responsible for engagement and success Sales.  **So as the first step we had checked the sanity of data:**.  In this we had to remove all the null values as well as garbage values so that we can get appropriate data.  So as the second step we divided our task in Five sections:   * Analysis of Rossmann Store Data. * Analysis of Store Data. * Analysis of Datasets and performed Different EDA Operation(Exploratory Data Analysis )to understand the depth of data . * Applied Different Machine learning Model to Test the Data and Performed Analysis on Model. * Appropriate Model Selection and Deriving conclusions based on all operations that we have performed.  1. **Analysis of Rossmann Store Data :-**   In this part we have seen that store, day of week, date, sales, customers, open, promo, state holiday, school holiday these Columns are present in Rossmann data set and how different Columns Affect Sales and promo.  Generally Checking data sets and getting overview of data that present in data.  **Conclusion :-**: how different Columns Data Affect Data sales.   1. **Analysis of Store Data :-**   In this part we have seen different types of columns present in the dataset such as store, storetype, assortment, competition distance, competition open since month, promo2, promo2 since week, promo2 since year, promo interval.  **Conclusion :-**  How Different Data and Columns Affect values.   1. **Analysis of Datasets and performed Different EDA Operation :-**   In this part we have performed different EDA operations to retrieve more information from the datasets.  **Conclusion :-** collected different information retrieved from the datasets and visualisation.   1. **Applied Different Machine learning Model :-**   In this part we have performed different Machine Learning operations to retrieve more information from the datasets.  **Conclusion :-**  Collected different information after the model getting trained like model score, accuracy.   1. **Appropriate Model Selection :-**   In this part we have Analysed different models and selected appropriate models for final model selection which have maximum model accuracy and minimum time required for execution.  **Conclusion :-**  we made a different math model for the Rossmann store sales dataset to forecast the future sales. Random forest Regression model brings us a basic understanding of how the math model works, while Facebook prophet model calculates the best solid result.  **Approach:-**  ● Importing dataset of Playstore and reviews dataset in google colab notebook.  ● Understanding the NaN, missing data, duplicates data using .info, .describe,  columns, .value\_counts(), shape, Grouping dataset  ● Merging dataset  ● Performing different EDA operation to retrieve more information form the  datasets.  **Challenges:**   1. Handling large amount of sales data (10,17,210 observations on 13 variables) 2. Prediction of sales for individual stores(out of 1115) and most stores have different patterns of sales. A single model cannot fit all stores.   **Learnings:**  1) Exploring large datasets using visualisation tools.  2) Learn the application of Time Series, Random Forest.. |
|  |