**Capstone Project Submission**

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

| **Team Member’s Name, Email and Contribution:** |
| --- |
| Contribution:  1. Rahul Ray: (ryrahul124@gmail.com)  1. Data Wrangling:  1. Data Cleaning and Data Manipulation  2. Univariate analysis  2. Sales Analysis (Weekly)  3. Different Pair Plots  4. Store Type & Assortment Chart  5. Correlation between important features  6. Feature Selection & Feature Engineering  7. Linear Regression  8. Decision Tree Regression & Random Forest Regression  2. Gaurav Yadav:(gauravyadav7939@gmail.com)  1. Data Wrangling:  1. Data Cleaning and Data Manipulation  2. Univariate analysis  2. Sales Analysis (Time-wise & Yearly)  3. Encoding  4. Multivariate Analysis  5. Correlation between important features  6. Feature Selection & Engineering  7. Linear Regression  8. Decision Tree Regression & Random Forest Regression  3. Mohan Vishe:(mohanvishe1999@gmail.com)  1. Data Wrangling:  l. Data Cleaning and Data Manipulation  2. Univariate analysis  2. Multivariate analysis  3. Hypothesis  4. Conclusions  5. Correlation between important features  6. Time base splitting, Factor affecting and choosing Algorithm  7. Decision Tree Regression & Random Forest Regression  8. Technical Documentation and presentation    4. Shambhuraj Desai: (shambhurajdesai97@gmail.com)  1. Data Wrangling:  1. Data Cleaning and Data Manipulation  2. Univariate analysis  2. Outliers analysis  3. Filling the missing values  4. The skewness of the variables  5. Transformation of skewed data  6. PPT presentation |
| **GitHub Repo link.**  GitHub Link:-  **https://github.com/MohanVishe/Supervised-ML--Regression----Retail-Sales-Prediction** |
| **Google drive link.**  **Google Drive**  **https://drive.google.com/drive/folders/1Ezu6B68vvjh922sTeKIhgeYLEX7v8OPT?usp=sharing**  **Individual colab project**  **https://colab.research.google.com/drive/1W7VaAIH38Lp3mtXmT2yraBaejBWG\_l67?usp=share\_link**  **Team colab project**  **https://colab.research.google.com/drive/18UNkvchbr4lSPJOBXoLaXMJB36jKPUs2?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)** |
| **SUMMARY**   * **Abstract**   Client interest and demand for product changes with time. Which is important for growing business. Sales prediction is estimating the demand for a service or product for a particular time. This project contains the real-world business problem  of sales prediction with the help of machine learning models.  Our task is to predict sales and find important factors affecting demand to improve sales.  In this project, we applied machine learning techniques to a real-world problem of predicting store sales. This kind of prediction enables store managers to create effective staff schedules that increase productivity and motivation. We used the popular open-source statistical programming language Python. We used feature selection, and model selection to improve our prediction result. In view of the nature of our problem, the R2 Score is used to measure the prediction accuracy.  Keywords: Retail Sales Prediction, Machine Learning Algorithms.   * **Methodology**   In this research paper, we used machine learning regressor models to predict the major store chain, Rossman. Here we implemented three models on the training set. The selected best model is implemented on a test set to predict outlet sales.   * **Experimentation**   The experiments were conducted by developing a simulation environment in python also using visualisation. We implemented six machine learning algorithms Linear Regression, lasso, ridge, decision tree, and Random Forest. We did hyperparameter tuning to improve our model performance.   * **Results and Discussions**   The six machine learning algorithms Linear Regression, lasso, ridge, decision tree, and Random Forest were implemented on the training set. We did hyperparameter tuning to improve our model performance. Six models work with accuracy levels of 0.86, 0.99, and 0.99 respectively. The following table shows the R Square Score of all models on the training dataset.    **Comparison of R square score of Models**   | **S. No** | **Regression Model** | **R square score** | | --- | --- | --- | | **1** | Linear Regression Model | 0.835077 | | **2** | Lasso Regression Model | 0.835137 | | **3** | Ridge Regression Model | 0.835092 | | **4** | Decision Tree Model | 0.972405 | | **5** | Random Forest Model | 0.983123 | |  |  |  |   **Conclusions**  In our analysis, we initially did EDA on all the features of our dataset. We first analysed our dependent variable, 'Sales' and also transformed it. Next, we analysed categorical variables and replaced null values, we also analysed numerical variables and found out the correlation, distribution and their relationship with the dependent variable using corr() function. We also removed some numerical features with mostly 0 values and hot-encoded the categorical variables.  Next, we implemented six machine learning algorithms Linear Regression, lasso, ridge, decision tree, and Random Forest. We did hyperparameter tuning to improve our model performance.   1. The sales in the month of December are the highest sales among others. 2. The Sales are highest on Monday and start declining from Tuesday to Saturday and on Sunday Sales are almost near Zero. 3. Those Stores who take participate in the Promotion got their Sales increased. 4. Type of Store plays an important role in the opening pattern of stores. All Type ‘b’ stores never closed except for refurbishment or other reasons. 5. We can observe that most stores remain closed during State holidays. But it is interesting to note that the number of stores opened during School Holidays was more than those opened during State Holidays. 6. The R Squared scores of all Liner Regression Algorithms with or without Regularization are quite good which is 0.86. 7. The R Squared score of the Decision Tree Regressor model we got 0.97 on the test set which is also good. 8. The Random Forest regressor model performed 0.98 which is very well among the others. 9. We can say that the random forest regressor model is our optimal model and can be deployed. |
|  |