**Retail Sales Prediction**

**Capstone Project**

|  |
| --- |
| **Team Member’s Name, Email and Contribution:** |
| Contribution:  1. Gaurav Yadav:(gauravyadav7939@gmail.com)  1. Data Wrangling:  1. Data Cleaning and Data Manipulation  2. Univariate analysis  2. Sales Analysis (Monthly & Yearly)  3. Correlation between important features  4. Feature Selection & Engineering  5. Linear Regression  6. Lasso and Ridge Regression  7. Decision Tree Regression  8. Random Forest Regression  2. 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 & Random Forest  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. Factor affecting and choosing Model  7. Decision Tree & Random Forest  8. Technical Documentation    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. Skewness of the variables  5. Transformation of skewed data  6. PPT presentation |
| **GitHub Repo Link: https://github.com/RahulI5/Retail-Sales-Prediction** |
| **Google Drive Link:** **https://drive.google.com/drive/folders/1mVguOcWz-dC3C1Uankss73C-SyUA6a6b?usp=share\_link** |
| **Summary and its Components: Approaches and Conclusions.** |
| **Retail Sales Prediction**  In this project, we applied machine learning techniques to a real word problem to predict 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, 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 and also using visualization. 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, Random Forest were implemented on the training and testing set. We did hyperparameter tuning to improve our model performance. The following table shows the accuracy of three models on the training dataset.    **Comparison of Accuracy of Models**   |  |  |  | | --- | --- | --- | | **S. No** | **Regression Model** | **Accuracy** | | **1** | Linear Regression Model | 0.867895 | | **2** | Lasso Regression Model | 0.867701 | | **3** | Ridge Regression Model | 0.867891 | | **4** | Decision Tree Model | 0.972821 | | **5** | Random Forest Model | 0.983467 | | **6** | Random Forest with Hyper Parameter Tuning | 0.983123 |   **Conclusions**  In our analysis, we initially did EDA on all the features of our dataset. We first analyzed our dependent variable, 'Sales' and also transformed it. Next, we analyzed categorical variable and replaced null values, we also analysed numerical variable, found out the correlation, distribution and their relationship with the dependent variable using corr() function. We also removed some numerical features who had mostly 0 values and hot encoded the categorical variables.  Next, we implemented six machine learning algorithms Linear Regression, Lasso, Ridge, Decision tree, Random Forest. We did hyperparameter tuning into 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 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 of the stores remain closed during State holidays. But it is interesting to note that the number of stores opened during School Holidays was more than that were opened during State Holidays. 6. The R Squared scores of all Linear Regression Algorithm 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 deploy. |
|  |