**Capstone Project Submission**

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

|  |
| --- |
| **Team Member’s Name, Email and Contribution:** |
| Prasanna R Kadlayyanavaramath:  [prasannark48@gmail.com](mailto:prasannark48@gmail.com)  Contributions:   1. Data Cleaning 2. EDA 3. Data Wrangling 4. Feature Engineering 5. ML Model Implementation 6. Hyper Parameter Tuning 7. Presentation work. 8. Technical document preparation. |
| **Please paste the GitHub Repo link.** |
| Github Link:- <https://github.com/Prasannark48/CapstoneProject--2-Retail-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)** |
| Sales forecasting is the process of estimating future sales. Accurate sales forecasts enable companies to make informed business decisions and predict short-term and long-term performance. Companies can base their forecasts on past sales data, industry-wide comparisons, and economic trends. Store sales are influenced by many factors, including promotions, competition, school and state holidays, seasonality, and locality. With thousands of sales records, we will be predicting sales based on their unique circumstances, the accuracy of results can be quite varied and the task is to forecast the “**Sales**” column. we are provided with historical sales data for 1,115 Rossmann stores. The task is to forecast the "Sales" column for the test set. Noting that some stores in the dataset were temporarily closed for refurbishment. Our experiment can help understand what could be the reason for the regression of such labels by feature selection, data analysis and prediction with machine learning algorithms taking into account previous trends to determine the correct Regression models.  1. First, I performed data cleaning for the raw data. Asking right question to dataset is always considered as great approach so I asked some very meaningful questions and plotted graphs and other visualization entities to get insights from data and I come with following conclusions: Sales is highly correlated to number of Customers. 2. For all stores, Promotion leads to increase in Sales and Customers both. 3. During starting days of the week like Monday,Tuesday, wednesday, friday the sales got a positive growth.But due to holiday on sunday completely it leads to zero.It gives the negative growth. So, Keeping store open on starting days of the week is must to grow the business. in peak level. 4. The most selling and crowded store type is A and D.Store type b is the least one. 5. Day of the week has a negative correlation indicating low sales as the weekends, and promo, customers and open has positive correlation. 6. CompetitionDistance showing negative correlation suggests that as the distance increases sales reduce, which was also observed through the plot earlier.   Next, we implemented 7 machine learning algorithms Linear Regression,lasso,ridge,elasticnet,decission tree, Random Forest and GradientBoosting. We did hyperparameter tuning to improve our model performance. The results of our evaluation are:   1. No overfitting is seen. 2. Random forest Regressor and Gradient Boosting gridsearchcv gives the highest R2 score of 96% and 88% recpectively for Train Set and 89% for Test set. 3. Feature Importance value for Random Forest and Gradient Boost are almost same. 4. 'CompetitionDistance','Promo','StoreType','DayOfWeek' and 'Assortment' are very important features where dependent variable depends on these features. 5. We can deploy this model.   Random forest Regressor and Gradient Boosting gridsearchcv gives the highest R2 score of 96% and 88% recpectively for Train Set and 89% for Test set.So, We are taking these ML Models for future prediction of the sales. |