**Capstone Project Submission**

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

|  |
| --- |
| **Team Member’s Name, Email and Contribution:** |
| D.N. Raghavendra  [Narayanaraghavendra0@gmail.com](mailto:Narayanaraghavendra0@gmail.com)  Linear regression  Lasso regression  Ridge regression  Random forest Regression  Dhanraj Tiwari[Rishutiwari020@gmail.com](mailto:Rishutiwari020@gmail.com)  EDA  Data visualization  Elastic net regression |
| **Please paste the GitHub Repo link:** |
| Github Link:- <https://github.com/Link/to/Repo>  https://github.com/narayana111100/capstone-project-2/blob/main/a\_Rossmann\_Sales\_Prediction\_Capstone\_project\_\_Raghavendra\_.ipynb |
| **Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)**   **Rossman retail store** Rossman retail store is one of the famous retail store 3000 drug store in European countries here, we have to predict the sales and competition between the store In this data we have 2 dataset files they are Rossman stores data and a sales data of Rossman here, to predict we have to do firstly the Exploratory data analysis and Visualizing the 2 data so, we have to merge up both datasets so ,that we have 1017209 rows and 9 columns in Rossman dataset and in 1115 rows and 10 columns we have to merge these 2 datasets and we have to see are there any null values. we have to import the libraries which are required .we have to find the dataset is open or closed we can use one hot encoding such that the if store is open = 1,store is closed = 0 and we can drop columns which are open in binary path .We can see that in "Promo2SinceYear", "Promo2SinceWeek", "Promo Interval" nearly 50 % of values are null. There is no point in keeping those so we are dropping these columns We will replace null values with median of columns "Competition Distance", "Competition Open Since Month", "Competition Open Since Year" then next we have to discrete variables to continuous variable and while visualising we have seen they are positively skewed ,we have to drop the duplicate values ,we can group by sales and day so, that we can find on which day it has highest sales and lowest sales and same as monthly and sales also we have to group by year and sales. Now, competition between two stores distance the visualising shows the positively skewed curve and we have done training and testing for linear regression and we use **mean squared error, root of mean squared error, mean absolute error and r2 score** and we get a linear regression model same as regularized linear regression and lasso regression and Ridge regression all of the models showed high accuracy.  Summary:  The demand for a product or service keeps changing from time to time. No business can improve its financial performance without estimating customer demand and future sales of products/services accurately. Sales forecasting refers to the process of estimating demand for or sales of a particular product over a specific period of time. In this article I will show you how **machine learning** can be used to **predict sales** on a real-world business problem . This case study solves everything right from scratch. So, you will get to see each and every phase of how in the real world a case study is solved. **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.  **Solution**  Firstly we have to import the necessary libraries or use the libraries when needed and do an EDA and then we can do visualizing the charts or graphs .we have to find out the null values and if any null values we can drop the column understanding the problem as there are 2 datasets we can merge them both into a dataset .We should group by the month and sales, group by year and values then we have to remove the variance inflation factor which year has high VIF can be removed .features are the major important factors in modelling the ML model so, features which are not can be removed or dropped.so ,now we come across the features which are required and the absolute correlation can be seen .we should also watch about outliers and detect the outliers.  Next we can start doing our modelling first of all we have to turn categorical variables to numerical variables by one hot encoding we should import library sklearn train , test ,split then we can give 75%training or 80% for training and 25% or 20% testing then we can do machine learning model.  **Lasso Regression &Ridge regression** also gives a good accuracy by importing the library ROC, AUC such that we can good accuracy **By** importing **grid search CV** we have to give the parameters as variable alpha values and CV value and train, test a and split here we can use formulas like **MSE, RMSE, R2score,adjusted R2** and we can visualise the chart. So, that we can get good accuracy if the accuracy between 0-1 it is very good that we are doing correct if it is near to o it is very good enough. |
|  |