

Project Report on Ratings Prediction



BY:

MAYANK SHUKLA

A solid orange horizontal bar at the bottom of the page.

Background & Introduction

- Product reviews are becoming more important with the evolution of traditional brick and mortar retail stores to online shopping. Consumers are posting reviews directly on product pages in real time. With the vast amount of consumer reviews, this creates an opportunity to see how the market reacts to a specific product. We will be attempting to see if we can predict the sentiment of a product review using machine learning tools, particularly the Support Vector Machine.

Business Problem

We have a client who has a website where people write different reviews for technical products. Now they are adding a new feature to their website i.e. The reviewer will have to add stars(rating) as well with the review. The rating is out 5 stars and it only has 5 options available 1 star, 2 stars, 3 stars, 4 stars, 5 stars. Now they want to predict ratings for the reviews which were written in the past and they don't have a rating. So, we have to build an application which can predict the rating by seeing the review.

Approach to Data Cleaning

- * First we check the information of the given dataset because it tells that how many rows and columns are present in our dataset and data type of the columns whether they are object, integer or float.
- Drop duplicates rows if present in dataset.+Then we check for the null values present in our dataset.
- If null values are present then fill it via mean, median or mode. Or also you can remove that rows but kindly check it properly.
- After that we check the summary statistics of our dataset. This part tells about the statistics of our dataset i.e. mean, median, max value ,min values and also it tell whether outliers are present in our dataset or not.

Approach to Data Cleaning (Cntd...)

*We also check the correlation of our dataset to check the correlation of the columns with each other. If columns are highly correlated with each other let's say 90% or above then remove those columns to avoid multi coli-nearity problem.


*We extract data from date column and make new columns like day, month and year to see the outcomes with our target column that is label.

*We delete the pcircle column because it has only one unique value that tells that collected data is only for one circle.

* We cannot remove outliers because more than 20% of our data are removed.



Visualization.

- * We plot correlation matrix via heatmap to see the correlation of the columns with other columns.
 - * We also visualize the correlation of columns with target column via bar graph to see which column is highly correlated with target column.
 - * We see the number of defaulter and non defaulter customers with the help of count plot.
 - * We plot histogram to displays the shape and spread of continuous sample data.
 - * We also see the customers labels i.e defaluter/Non-defaulter according to date and month with count plot.
 - * We also see the distribution of the data with the help of distribution plot whether it is left skewed or right skewed.
- 

Modelling part

*We know that this is classification problem so we use accuracy score, classification report and confusion matrix as our evaluation matrix. We also see the AUC score and also plot the AUC_ROC curve for our final model.

* As we know this dataset is imbalance so we don't too much focus on accuracy score . We see the precision and recall value along with f1_score.

*First we see the result without doing any sampling technique and for that I use Logistic Regression with K-Fold cross validation and hyper-parameter tuning.

*Secondly we used Support Vector Machine Classifier Model and then we used Decision Tree model for better implementations.

*We also use Random Forest Classifier as our evaluation model without using hyper-parameter tuning because our dataset is too large and it takes more than hour to give the result.

Conclusion

- Looks like all the models performed very well ($>90\%$), and we will use the **Support Vector Machine Classifier** since it has the highest accuracy level at 93.94%