Lead Scoring Case Study

Presented by:

Dhairya Patel

Problem Statement

- An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.
- The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.
- Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Goals of the Case Study

- There are quite a few goals for this case study:

1. Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.
2. There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

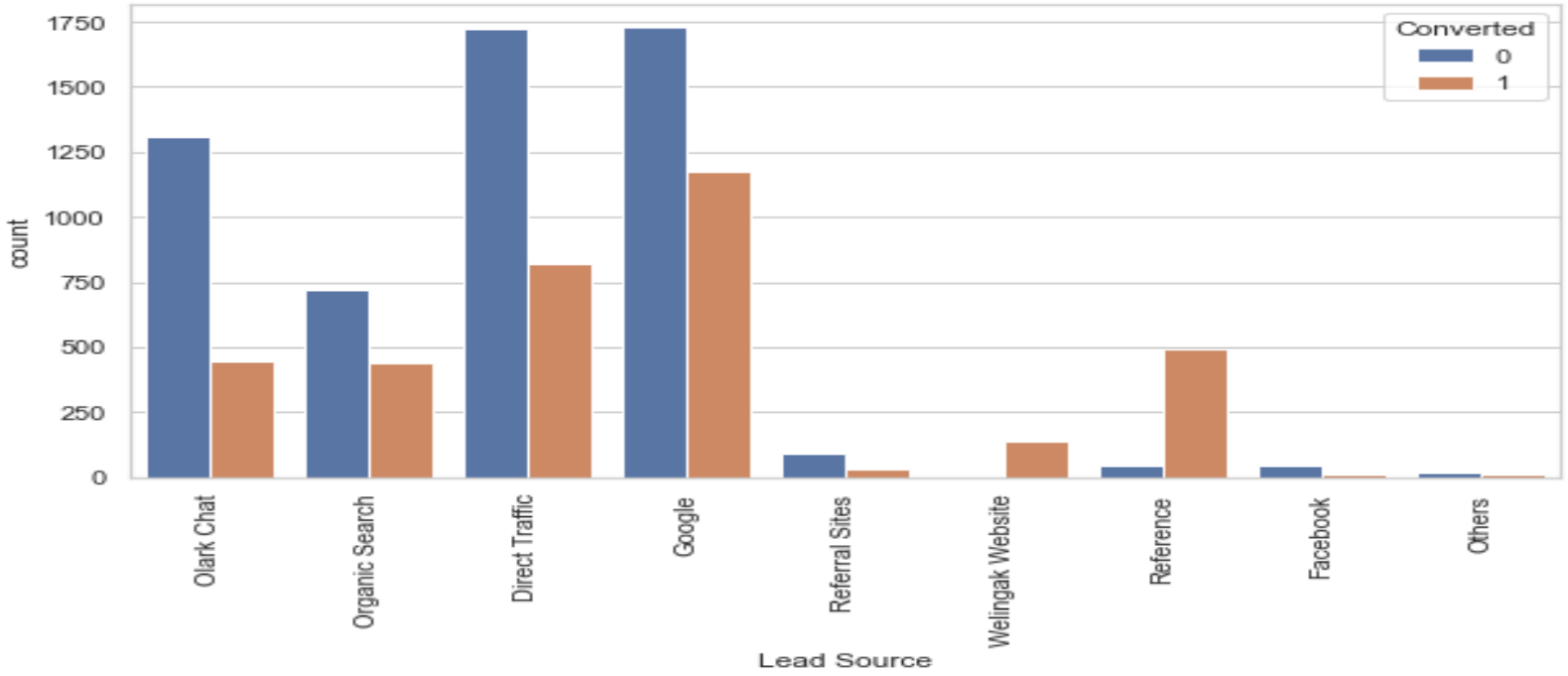
Approch

- Source the Data For Analysis
- Reading & Understanding the Data
- Data Cleaning
- EDA
- Feature Scaling
- Splitting the Data into train and test dataset
- Prepare the Data for Modeling
- Model Building
- Model Evaluation – Specificity, Sensitivity, Precision, Recall
- Making Prediction on the Test Set

Data Sourcing, Cleaning & Preparation

- Read the data from CSV file
- Data Cleaning – Handling Null Values & Removing Higher Null Values data
- Imputing the Null Values
- Removing the Redundant columns in Data
- Outlier Treatment
- Exploratory Data Analysis
- Approx. Conversion rate – 38.5 %
- Feature Standardization

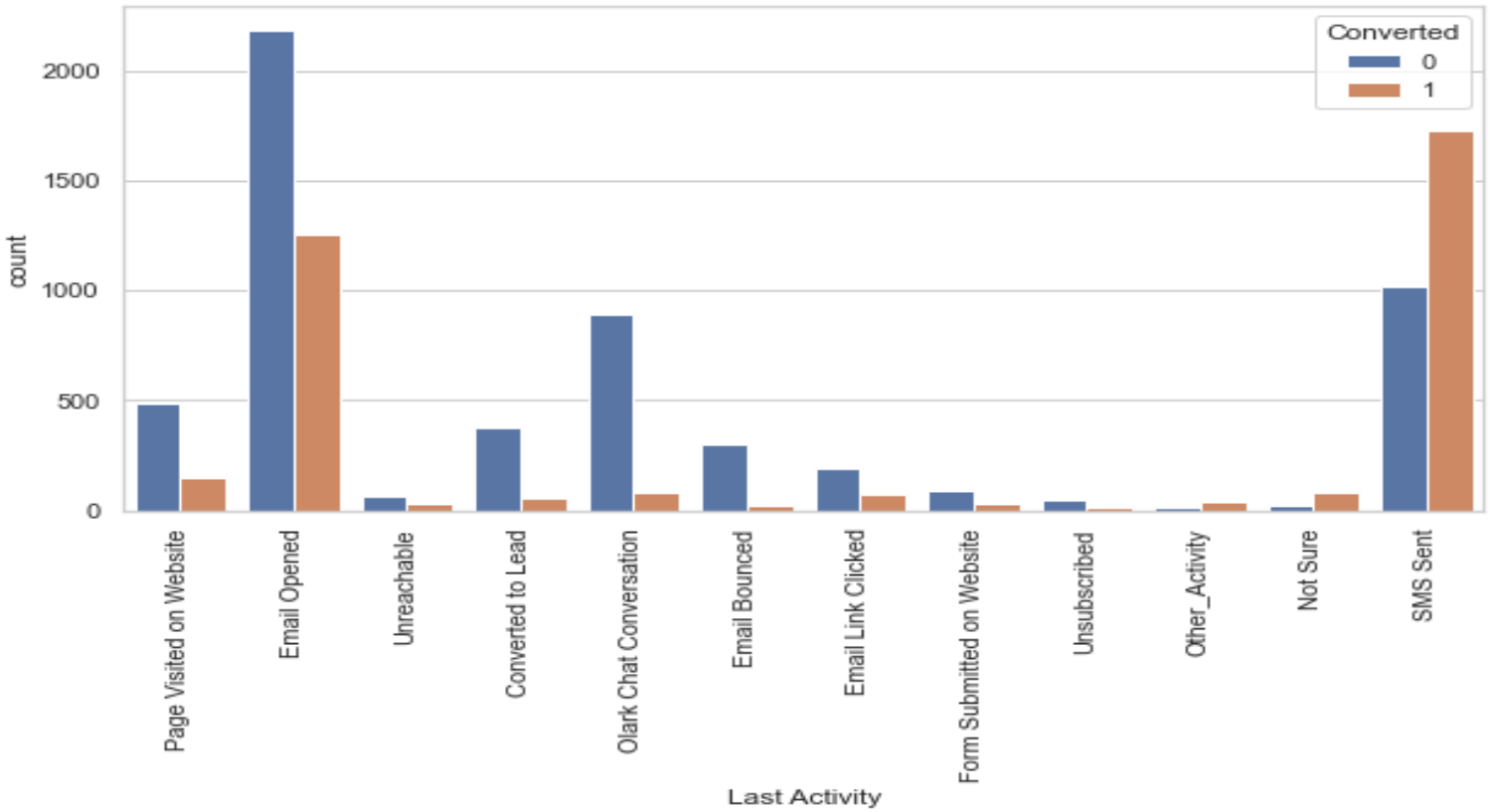
EDA – Data Analysis



Inference

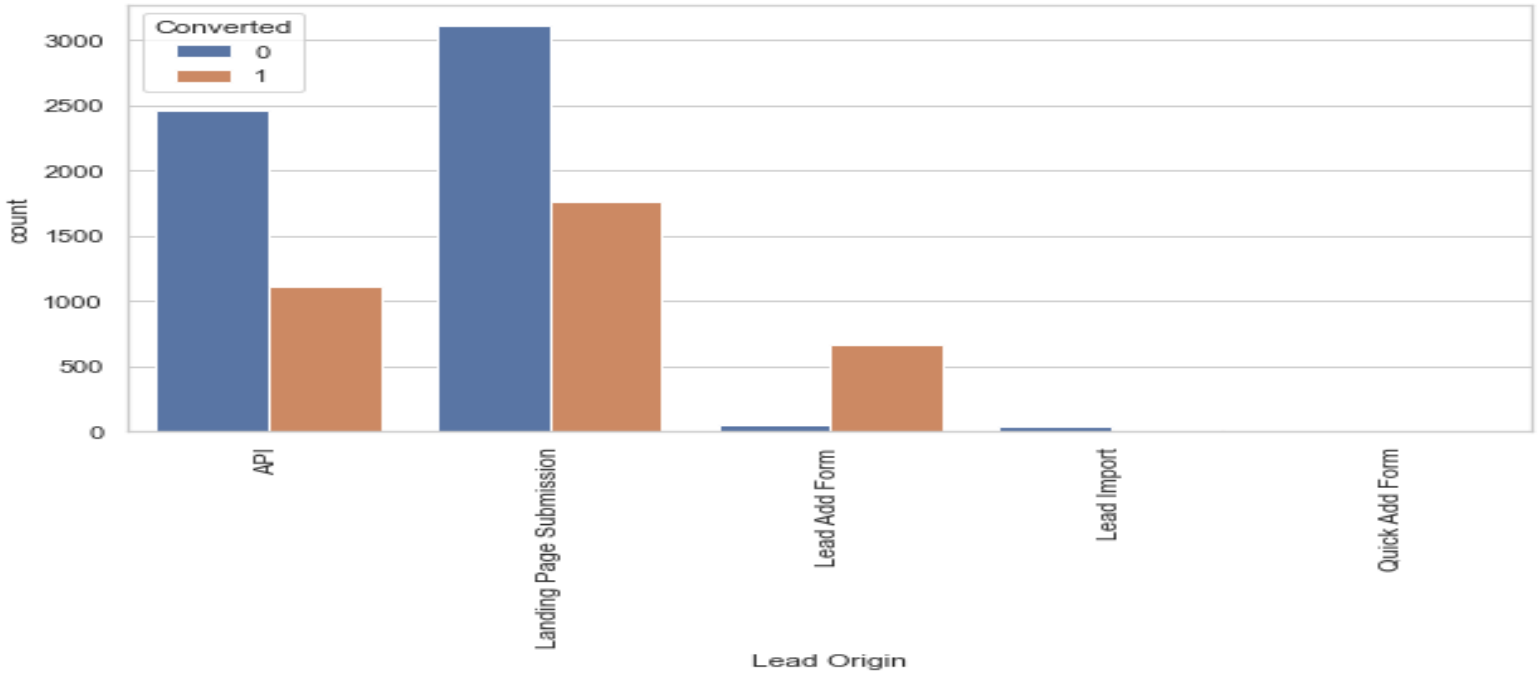
- Google and direct traffic create the greatest quantity of leads.
- The conversion rate of reference leads and leads through welingak website is high.
- To increase total lead conversion rate, focus on boosting lead conversion of olark chat, organic search, direct traffic, and google leads, as well as generating more leads from reference and welingak website.

EDA – Data Analysis



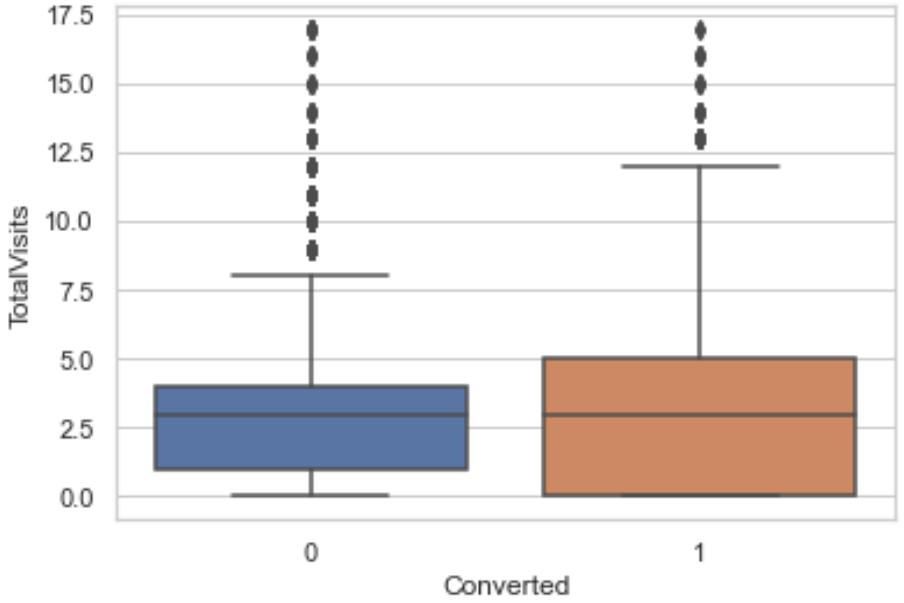
- As their Last activity, the majority of the leads have opened their email.
- The conversion rate for leads with the Last activity as SMS Sent is around 60%.

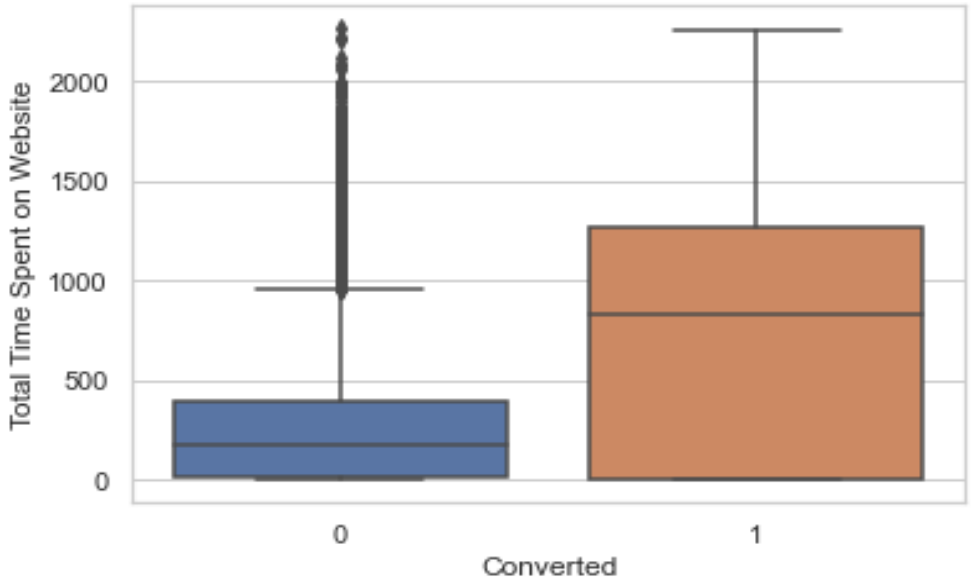
EDA – Data Analysis

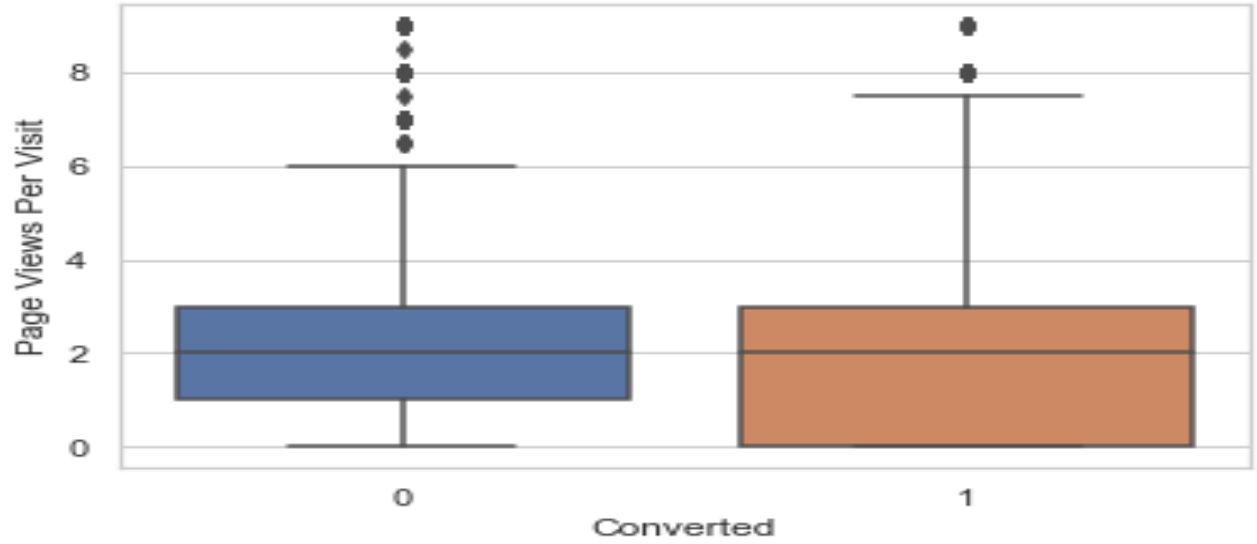


- API and Landing Page Submission generate more leads and conversions.
- The Lead Add Form has a very good conversion rate, however the lead count is not particularly high.
- Lead Import and Quick Add Form both generate a small number of leads.
- To increase total lead conversion rate, we must boost lead generation from API and Landing Page Submission origins and produce more leads via Lead Add Form.,

Outlier Treatment







Data Preparation

- Converted Binary Variable in to 0 & 1.
- Created Dummy Variables for Categorical Variables

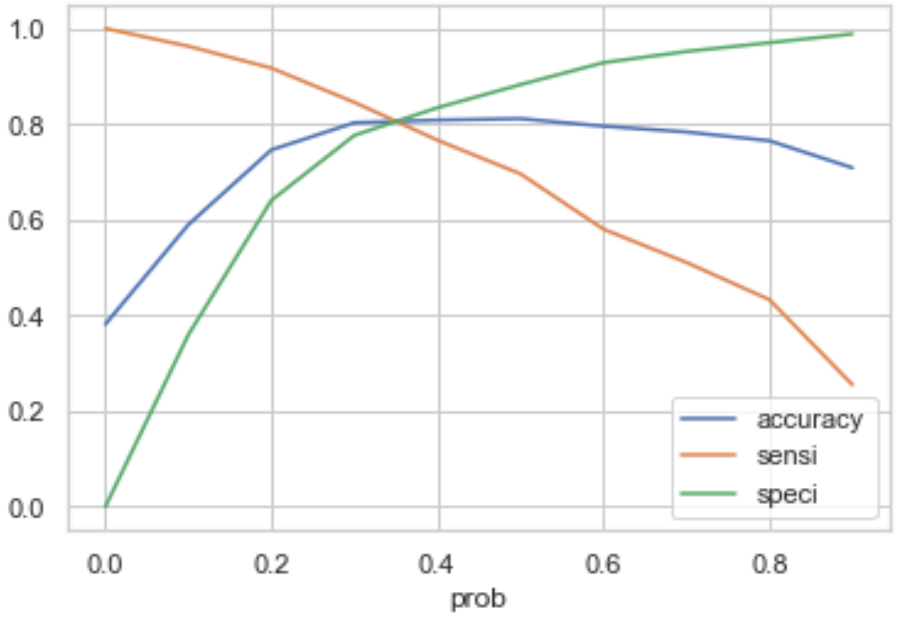
Feature Scaling & Splitting Train & Test Sets

- Feature Scaling Of Numerical Variables
- Splitting the data into Train & Test set

Model Building

- Feature Selection Using RFE
- Obtained Optimal Model using Logistic Regression
- Calculated Accuracy, Sensitivity, Specificity, Precision-Recall & Evaluate Model

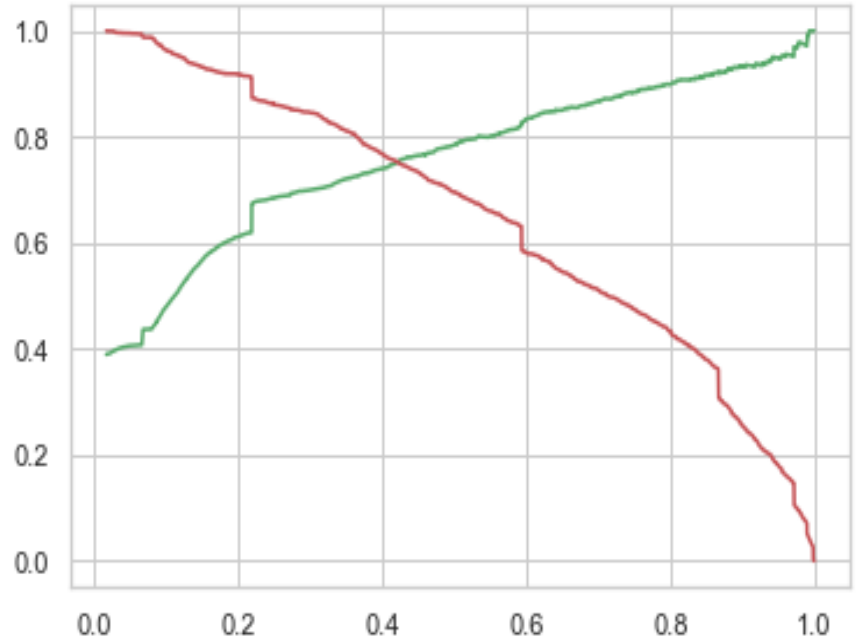
Model Evaluation – Accuracy, Sensitivity & Specificity On Train Dataset



- Accuracy - 80.68 %
- Sensitivity – 81.88 %
- Specificity - 79.95 %

- From the curve above, 0.34 is the optimum point to take it as a cutoff probability.

Model Evaluation – Precision & Recall On Train Dataset



- Precision - 71.58 %
- Recall – 81.88 %

- The graph above depicts an optimal cut-off of 0.42 based on the trade-off between Precision and Recall.

Model Evaluation – Accuracy, Sensitivity & Specificity On Test Dataset

- Accuracy - 80.05 %
- Sensitivity – 81.44 %
- Specificity - 79.16 %

Result

- Accuracy, Sensitivity, and Specificity values of the Training and test set are close to equal.
- Accuracy, Sensitivity, and Specificity values of the Training set are 80.68 %, 81.88 % and 79.95 %.
- Accuracy, Sensitivity, and Specificity values of the Test set are 80.05 %, 81.44 % and 79.16 %.
- We have made a prediction on the test set using the cut-off threshold from Sensitivity and Specificity Metrics.

Conclusion

- While we have checked both the sensitivity–specificity and Precision-Recall metrics, we have selected the optimal cut-off based on the sensitivity and specificity for making the final prediction.
- Accuracy, Sensitivity, and Specificity values of the Test set are 80.05 %, 81.44 %, and 79.16 % which are closer to the values of the training dataset.
- Hence, the Overall model seems to be good.