Leads Scoring-Analysis

Motive

- Analysis is carried out for X Education in an effort to attract more business professionals to their courses. We learned a lot from the fundamental data on how potential customers use the site, how long they stay there, how they got there, and the conversion rate

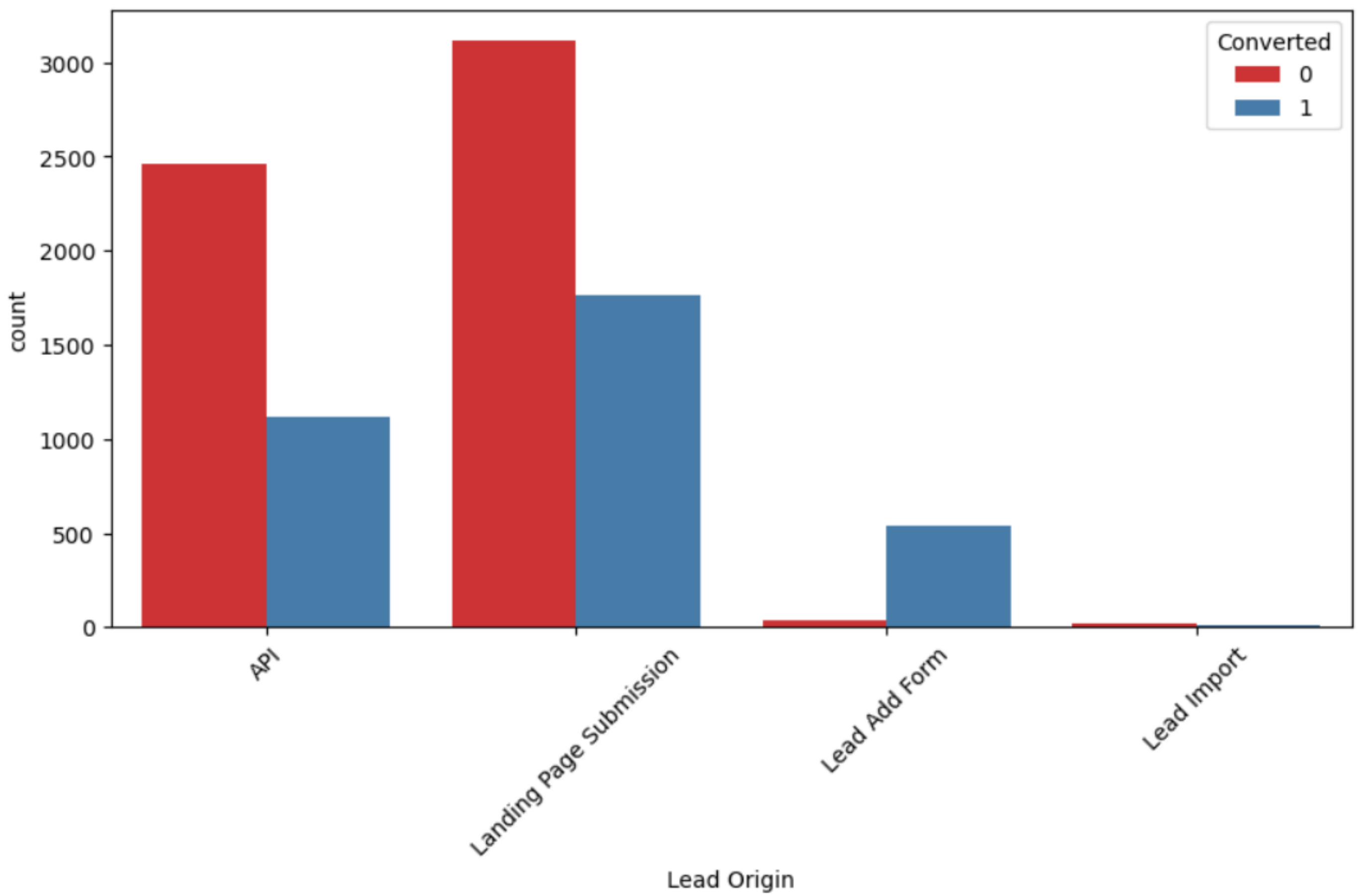
Analysis of Leads Dataset

- Inside the analysis it is found that the present of the the ‘select’ as a member present inside the many columns. That leads to no interpret any kind of value so first it would be converted into null.
- In further steps of the data analysis we calculate the amount of the null values present inside the data set and try to rectified it or remove the columns as not required for the analysis.

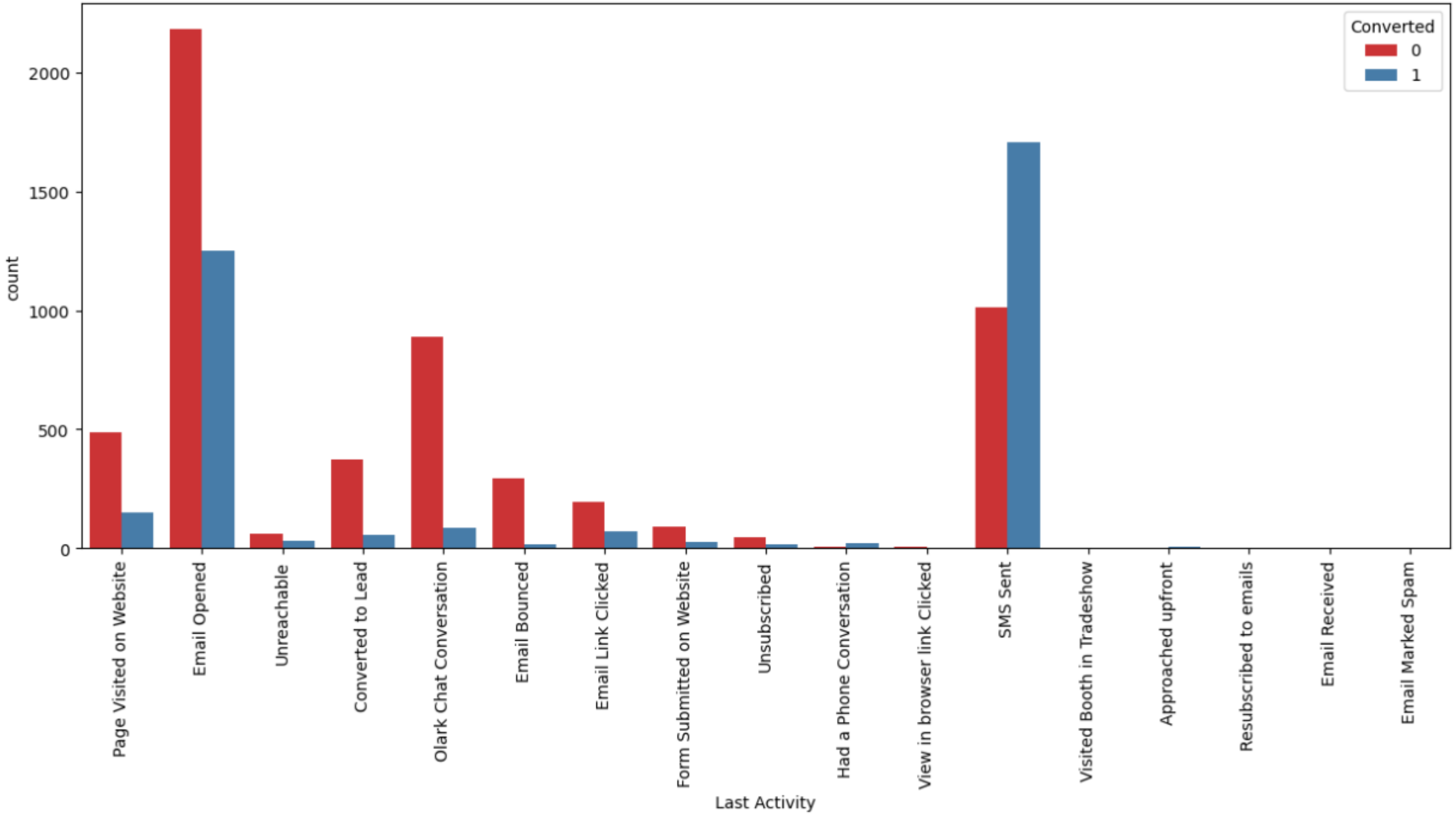
Exploratory Data Analysis

- Checks for the duplicate values and remove them if the columns are not important for the analysis.
- The further steps performing the Univariate and Bivariate data analysis by which the analysis are some of the variables are as shown-

Analysis for **Lead Origin** column



Analysis for **'Last Activity'** Column



- Based on the univariate analysis we found that many of the columns are not adding any information to the model hence we drop the following columns.
    - 'Lead Number
    - 'Tags'
    - 'Country'
    - 'Search'
    - 'Magazine'
    - 'Newspaper Article'
    - 'X Education Forums'
    - 'Newspaper'
    - 'Digital Advertisement'
    - 'Through Recommendations'
    - 'Receive More Updates About Our Courses'
    - 'Update me on Supply Chain Content'
    - 'Get updates on DM Content'
    - 'I agree to pay the amount through cheque'
    - 'A free copy of Mastering The Interview'

Data Preparation for Lead Datest

- Convert Binary variables (yes/no) to 1/0
- Creating dummy variables for the categorical columns
  - 'Lead Origin‘
  - 'Lead Source'
  - 'Last Activity'
  - 'Specialization'
  - 'What is your current occupation'
  - 'City‘
  - 'Last Notable Activity'

- Splitting data into Train and Test sets
  - X_train
  - X_test
  - y_train
  - y_test

```
X_train, X_test, y_train, y_test = train_test_split(X, y, train_size=0.7, test_size=0.3, random_state=100)
```

  - **Scaling the Feature**
  - **Feature selection using RFE**

```
from sklearn.linear_model import LogisticRegression
logreg = LogisticRegression()

from sklearn.feature_selection import RFE
rfe = RFE(logreg, step = 20)            # running RFE with 20 variables as output so that it would be easier for analysis
rfe = rfe.fit(X_train, y_train)
```

Model Building

- We create the model by taking the all model at one count and removing the columns by using there **Pvalue** and if we stuck where **Pvalue** is not applicable at that time we consider the **VIF** values based on which we perform the elimination of the column.

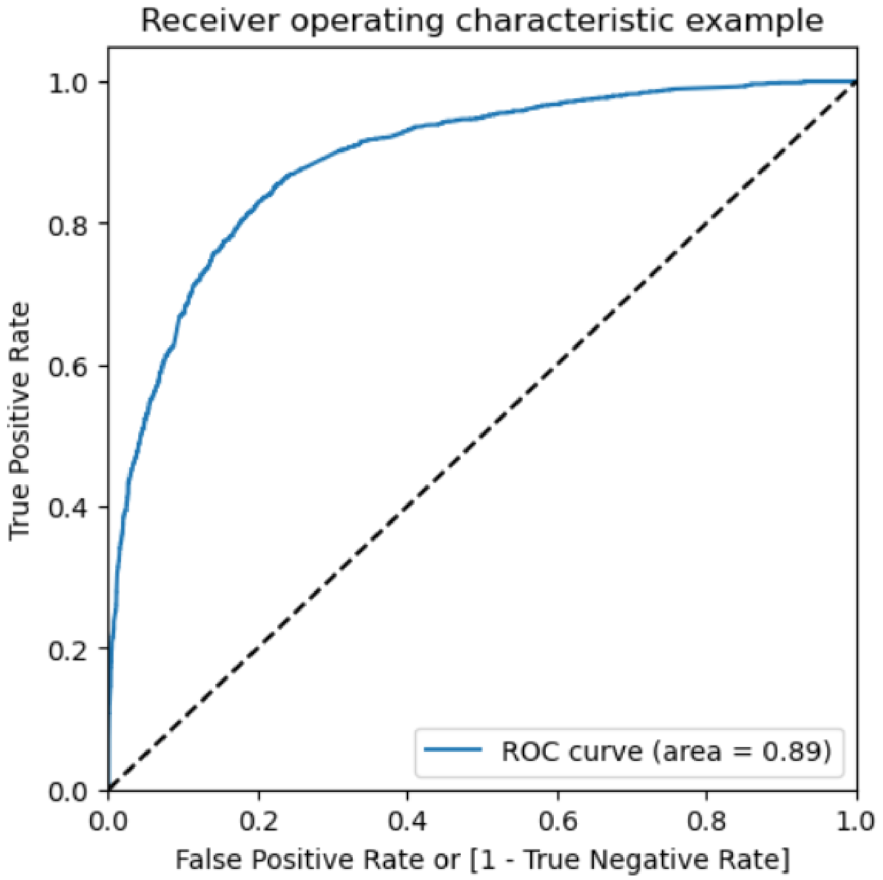
- So after repeating the process at **model-9** we see that we can not get the more optimize model apart from it so we consider **model-9** as or final model for the analysis.

Predictions

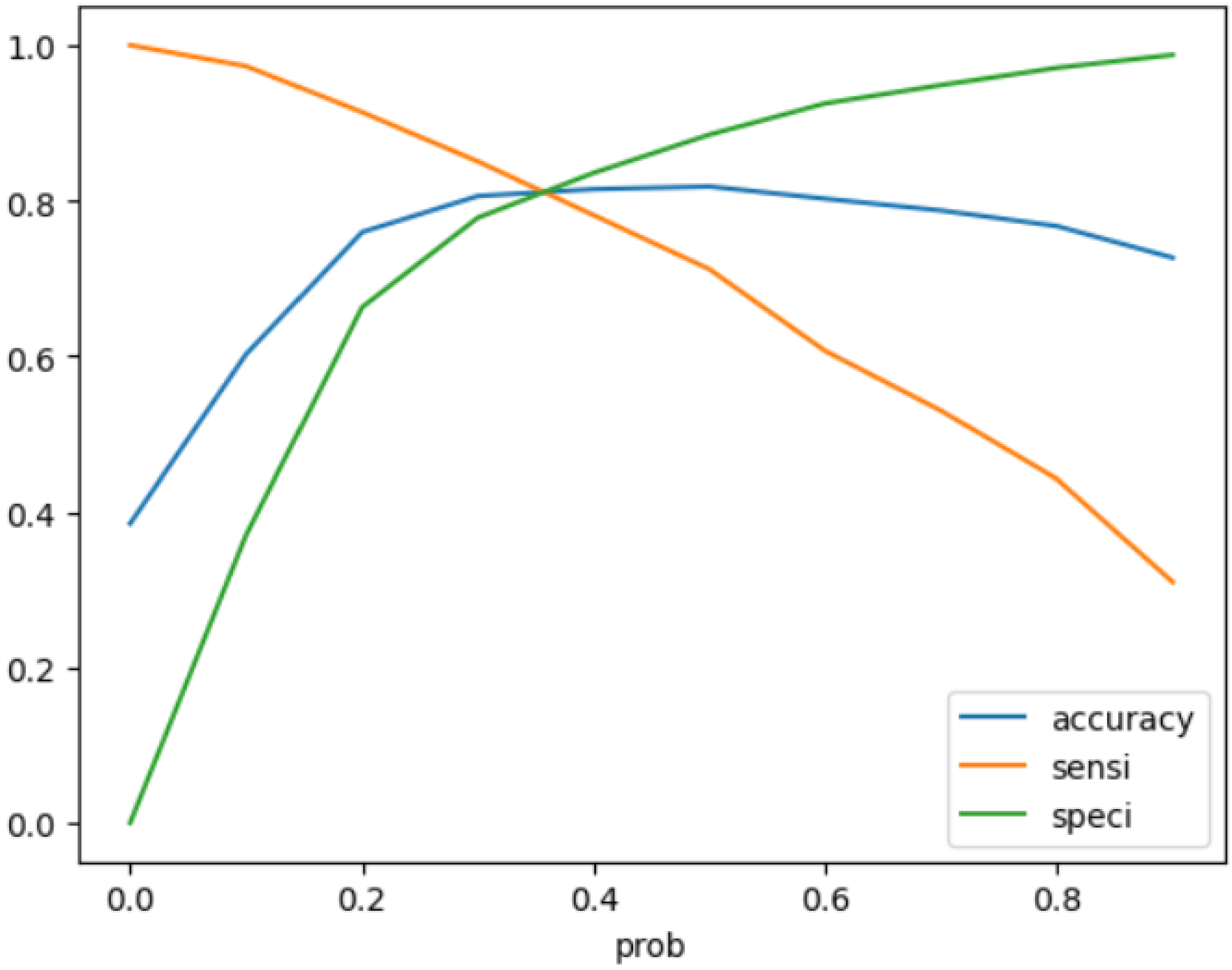
- Predictions over the **Train set**.
  - Choosing an arbitrary cut-off probability point of 0.5 to find the predicted labels.
  - Creating new column 'predicted' with 1 if Converted_Prob > 0.5 else 0
  - Calculate the confusion matrix.
  - On Train set we found the following conclusions
    - specificity was good (~88%) but our sensitivity was only 70%

    - sensitivity of 70% and this was mainly because of cut-off points to 0.5 arbitrary chosen, so this cut-off has to modify, and to get a decent vale of the sensitivity we analyse the **ROC curve**

ROC Curve for the **Train-Set**



- Plot of **accuracy sensitivity** and **specificity** for various probabilities



Model Evaluation

- We get the following results for the evaluation of the model.
  - Accuracy : 0.8110533774208786
  - Sensitivity : 0.8254292722812756
  - Specificity : 0.8020486555697823
  - False Positive rate : 0.19795134443021767
  - Positive Predictive Value : 0.7231375358166189
  - Negative Predictive Value : 0.8800224782242203

Final Observations

- Train Data –
  - Accuracy : 81.0 %
  - Sensitivity : 81.7 %
  - Specificity : 80.5 %

- Test Data –
  - Accuracy : 80.4 %
  - Sensitivity : 80.4 %
  - Specificity : 80.5 %

Some Results to Improve

- 1) The company should make calls to the leads coming from the lead sources "Welingak Websites" and "Reference" as these are more likely to get converted.
- 2) The company should make calls to the leads who are the "working professionals" as they are more likely to get converted.
- 3) The company should make calls to the leads who spent "more time on the websites" as these are more likely to get converted.
- 4) The company should make calls to the leads coming from the lead sources "Olark Chat" as these are more likely to get converted.
- 5) The company should make calls to the leads whose last activity was SMS Sent as they are more likely to get converted.
- 6) The company should not make calls to the leads whose last activity was "Olark Chat Conversation" as they are not likely to get converted.
- 7) The company should not make calls to the leads whose lead origin is "Landing Page Submission" as they are not likely to get converted.