LEAD SCORING CASE STUDY.

By

- Rashmi Deepika
- Umang Shukla

PROBLEM STATEMENT

- An education company named X Education runs by selling online courses to industry professionals.

- X Education promotes the courses on its platform on numerous websites and search engines like Google. The people, when directed to the website, browse through the platform and fill some course form. The people filling the form by providing their details are called as leads. The sales team of the company then start approaching these leads through various sources. After approaching them, some leads are converted while most are not. The typical lead conversion rate at X Education is around 30%.

- X Education usually gets a lot of leads but its lead conversion rate is very poor (only 30%).

- The company wants to find the 'Hot Leads', which are the leads who are most likely to be converted. This will help the sales team in targeting selected selected segment of people and can eventually increase the conversion rate of the company

- Through Analysis, we have to help them find the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company has asked to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

SOLUTION METHODOLOGY

- Importing Necessary Libraries

- Reading and Understanding Data

- Data Cleaning
  - Checking unique values and treating them
  - Missing Value Treatment
  - Outlier Treatment
  - Treatment of irrelevant features
  - Sanity Check

- Exploratory Data Analysis (EDA)

- Data Preparation
  - Dummy Variable Encoding
  - Dataset Splitting
  - Standardizing the dataset

SOLUTION METHODOLOGY

- Model Building using statsmodels
- Model Evaluation on the Test Dataset
- Calculating the Lead Score
- Determining Feature Importance

DATA CLEANING

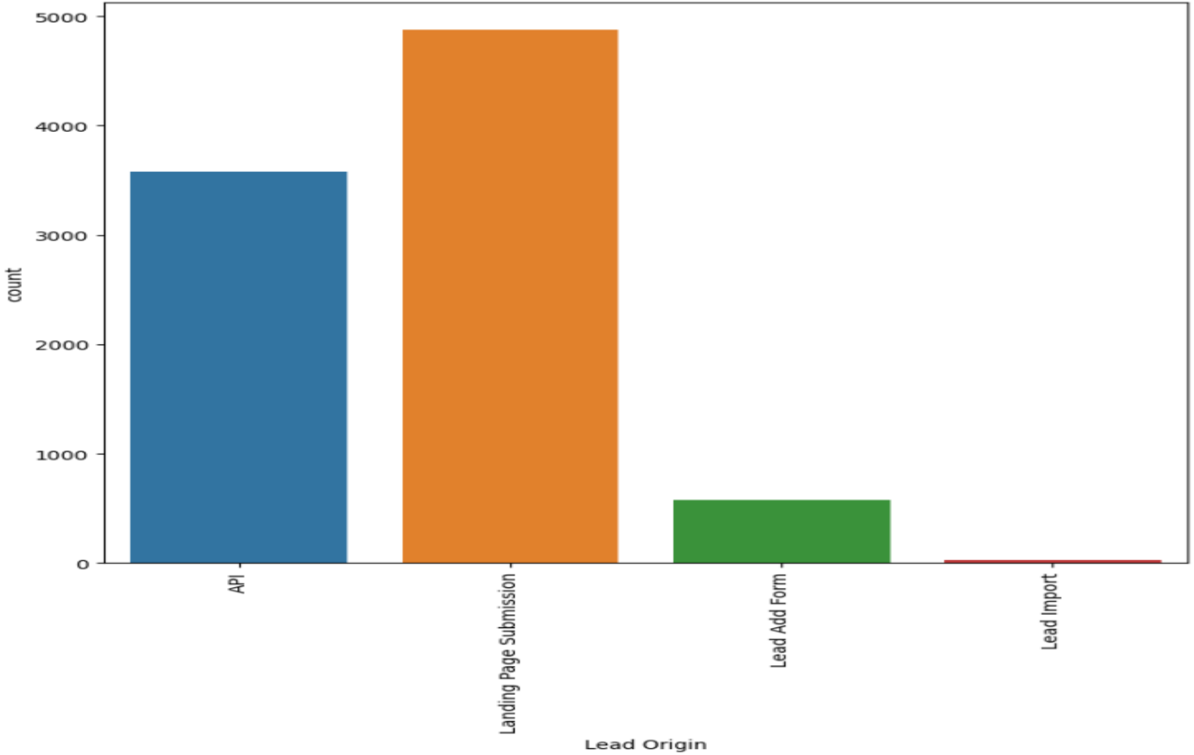
- Dropped features having missing values more than 40%
- Replaced redundant values such as “Select” in some features as Missing Values
- Then, after replacing the values, we dropped the features having more than 40% missing value again
- Dropped rows from features having very few missing values
- Imputed the NULL values of some features with their relevant values as mode & ‘Unknown’ label.
- Dropped Irrelevant features, which are single-value dominating.
- Removed extreme outliers of numerical features like ‘TotalVisits’ and ‘Page Views Per Visit’ that could affect our analysis and the model

- Some variables has been standardized and grouped like ‘Lead Source’ and ‘Last Activity’.
- Dropped similar value features like ‘Last Activity’ and ‘Last Notable Activity’.
- Dropped columns which are inaccurate like ‘Country’ and ‘City’, since Indian city are listed in non Indian Country.

EXPLORATORY DATA ANALYSIS

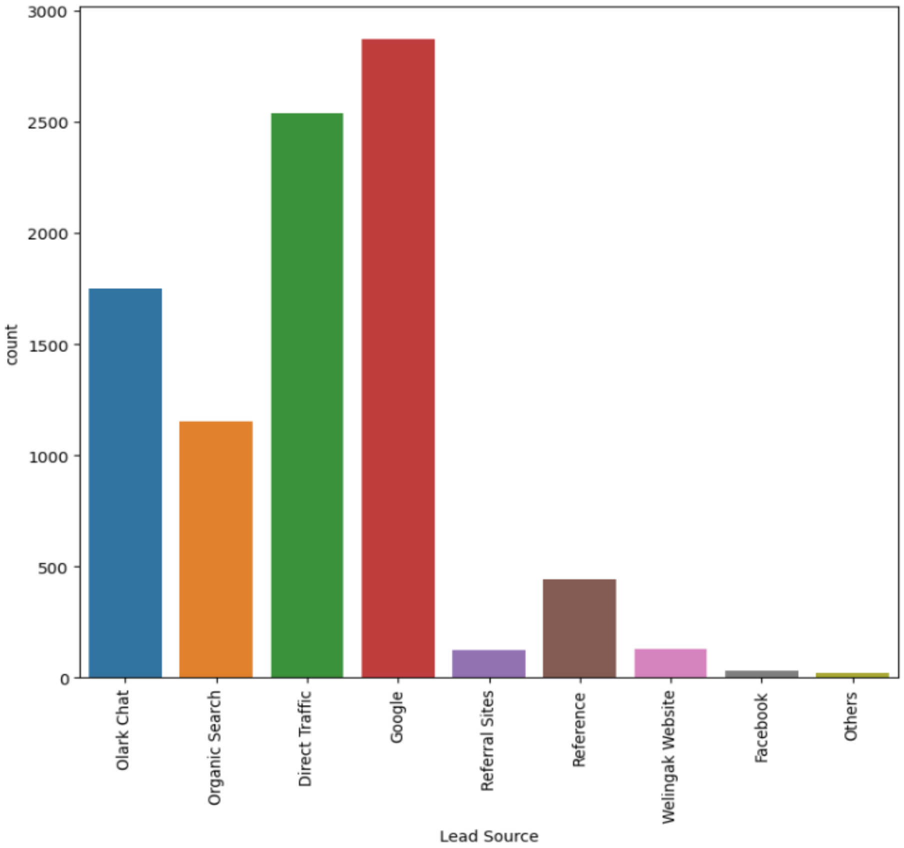
UNIVARIATE ANALYSIS

Lead Origin Distribution



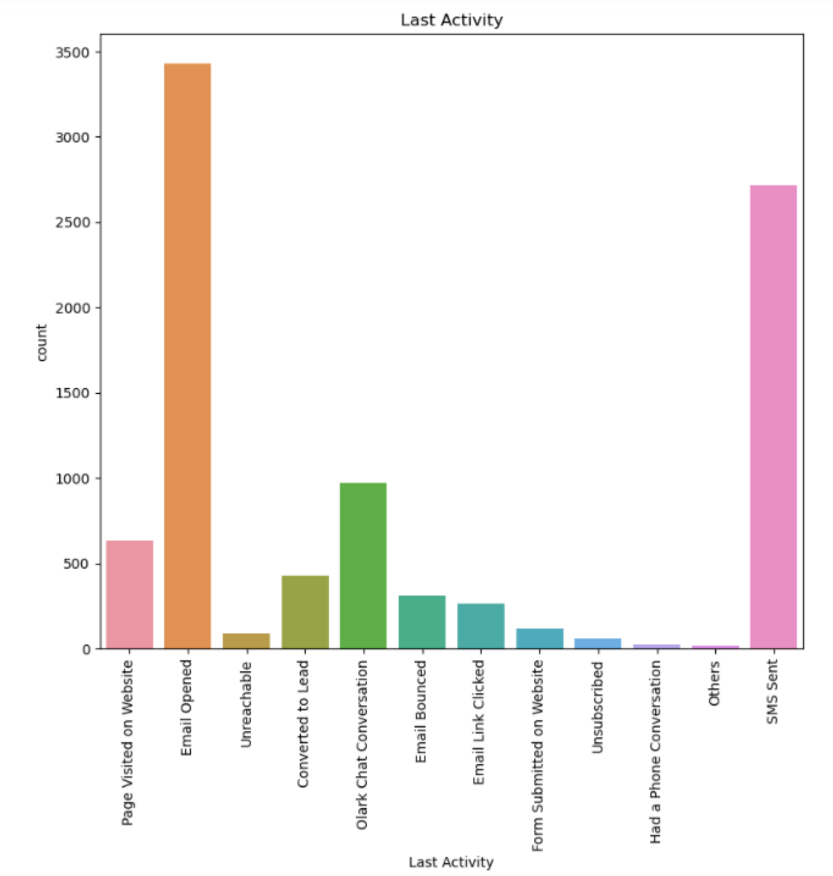
- Majority of Lead Origin is from ‘Landing Page submission’ while least from ‘Lead Import’

Lead Source Distribution



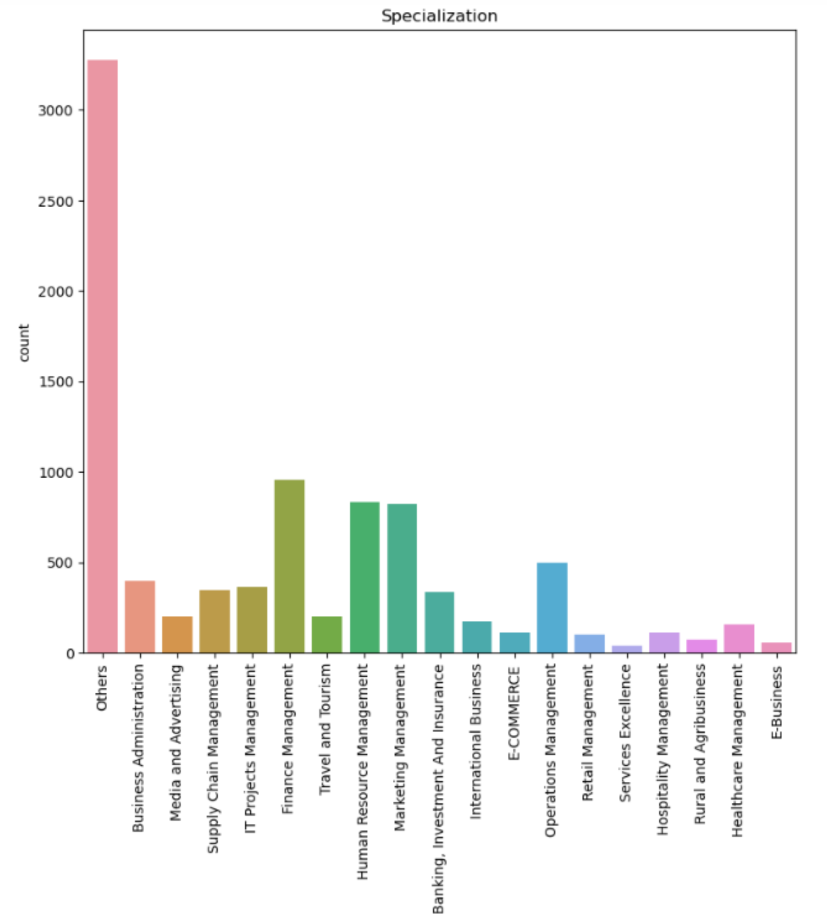
- Google is the highest source of lead , followed by Direct Traffic & Olark Chat, while least lead collection is from facebook.

Last Activity Distribution



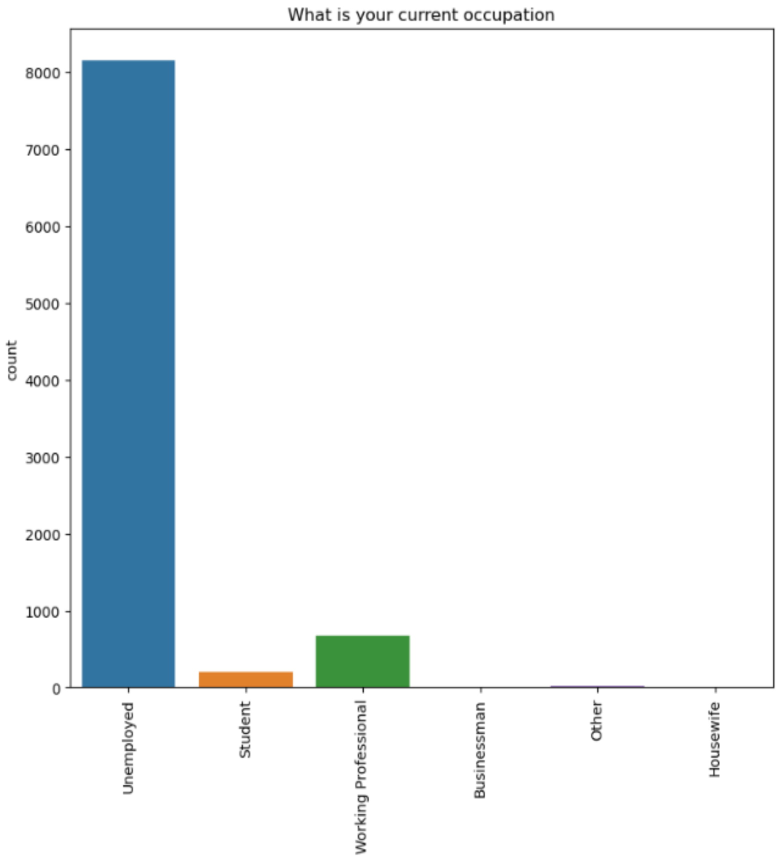
- Email is being open & SMS sent is recorded as Last Activity encountered by Leads

Specialization Distribution



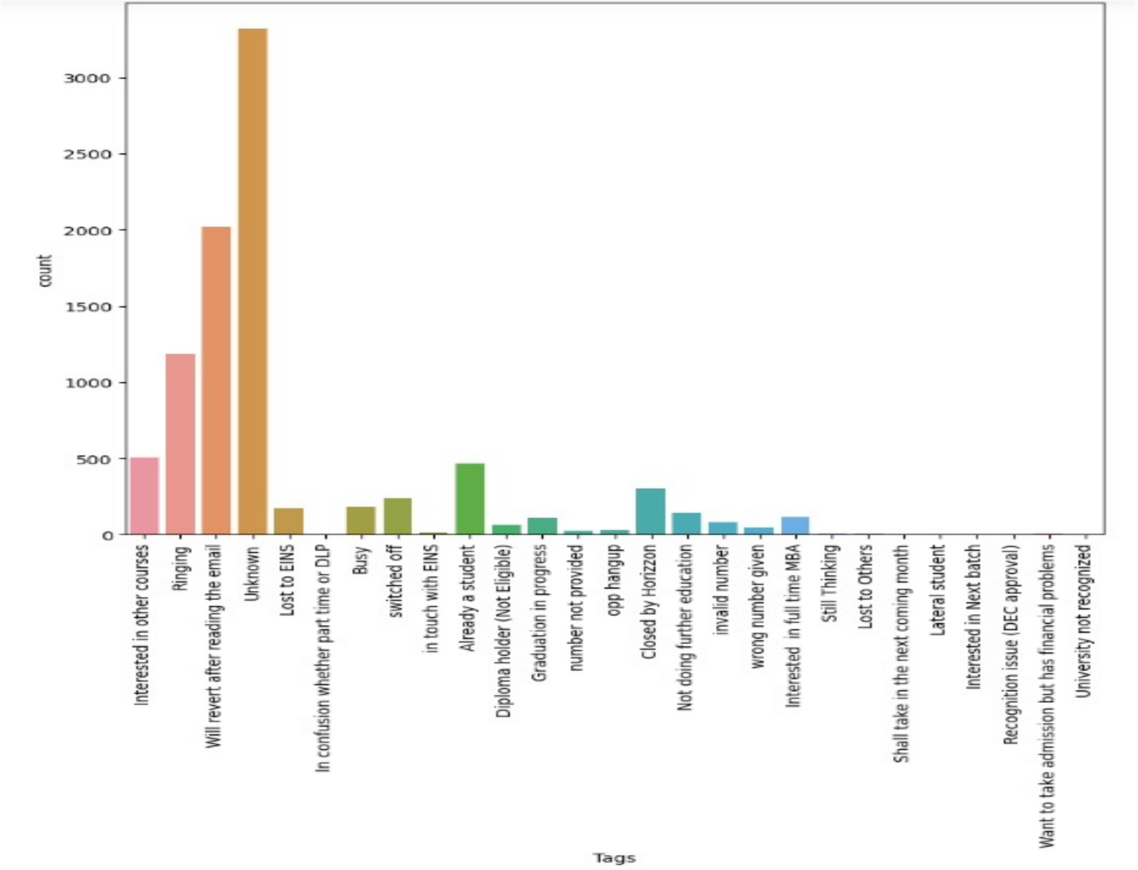
- Majority of lead is from Finannce Management, Marketing Management & HumanResource Managemnt specialization.

What is your current Occupation Distribution



- The unemployed population have approached more for the program , followed by Working professional.

Tags Distribution

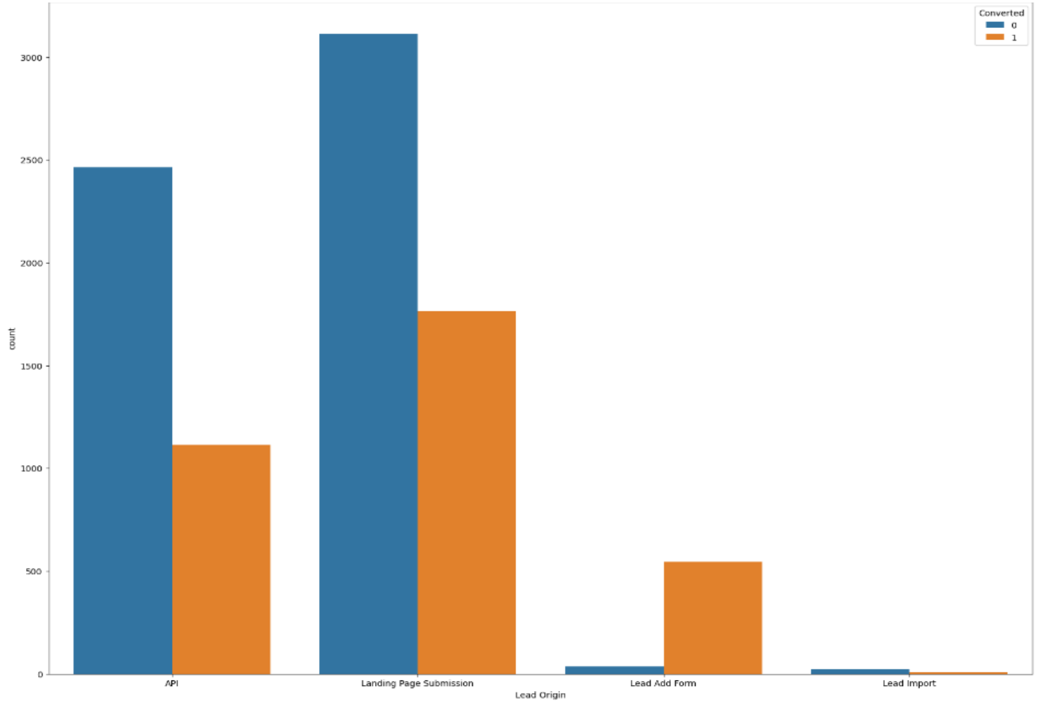


- Majority of population responded will they will revert back after reading email.

EXPLORATORY DATA ANALYSIS

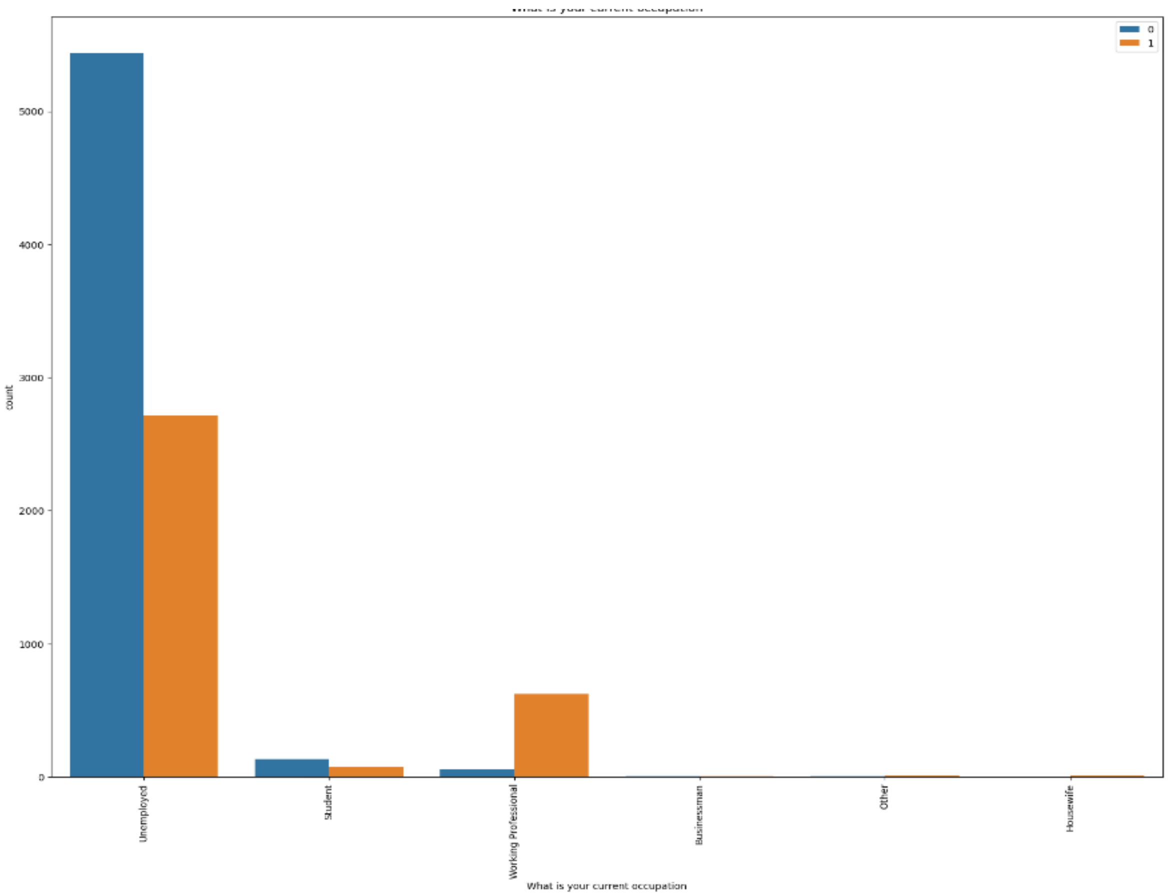
(BIVARIATE ANALYSIS)

Relation between Lead Origin & Conversion



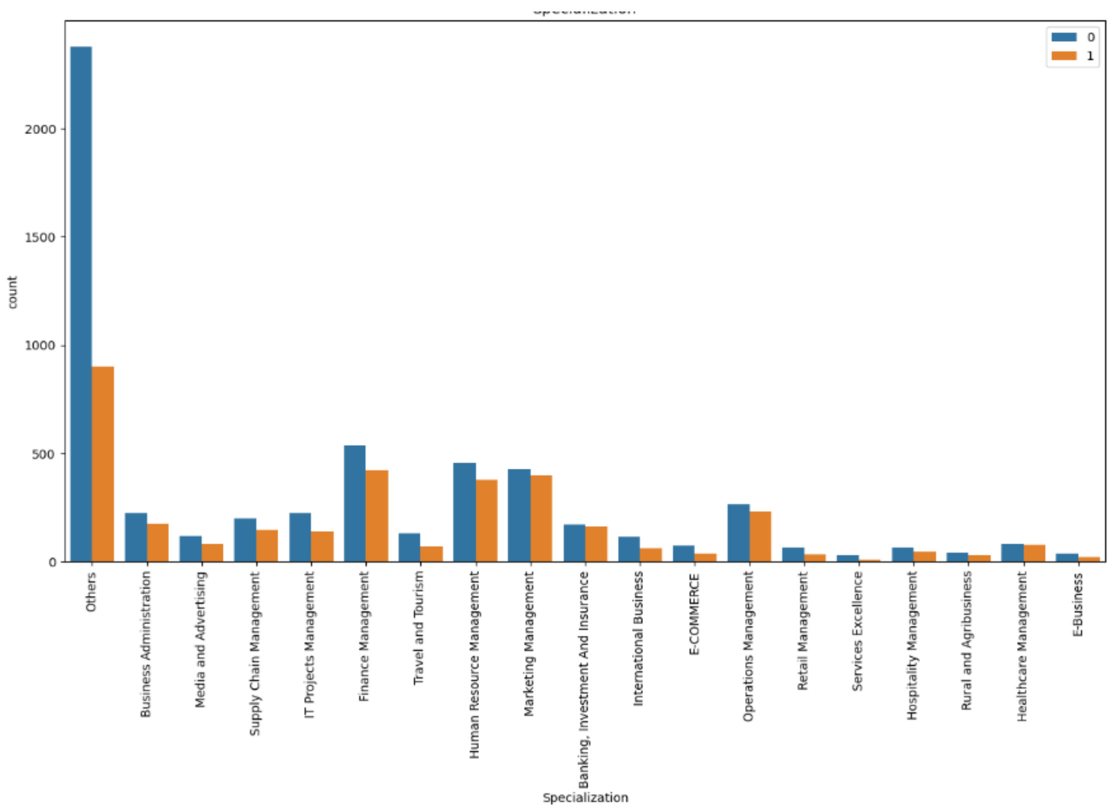
- Relation between Lead conversion with Lead Origin
- Here highest lead conversion is received from ‘Landing Page submission’
- From ‘Lead and form’ origin , conversion rate is higher than non conversion population.

Relation between Current Occupation & Conversion



- Unemployed population have highest conversion rate.
- While among working professional, conversion rate is higher than non conversion

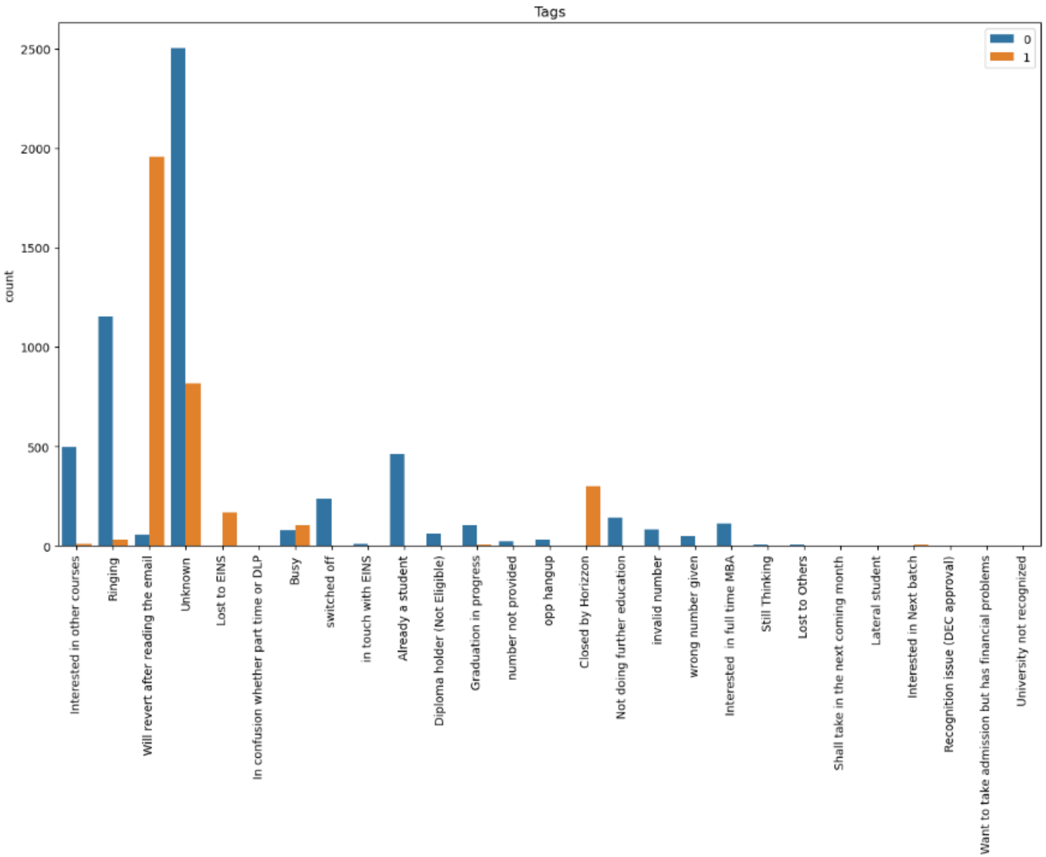
Relation between Specialization & Conversion



- Higher conversion is among Finance, HR and marketing management

- While from others category conversion rate is less than half on non conversion rate.

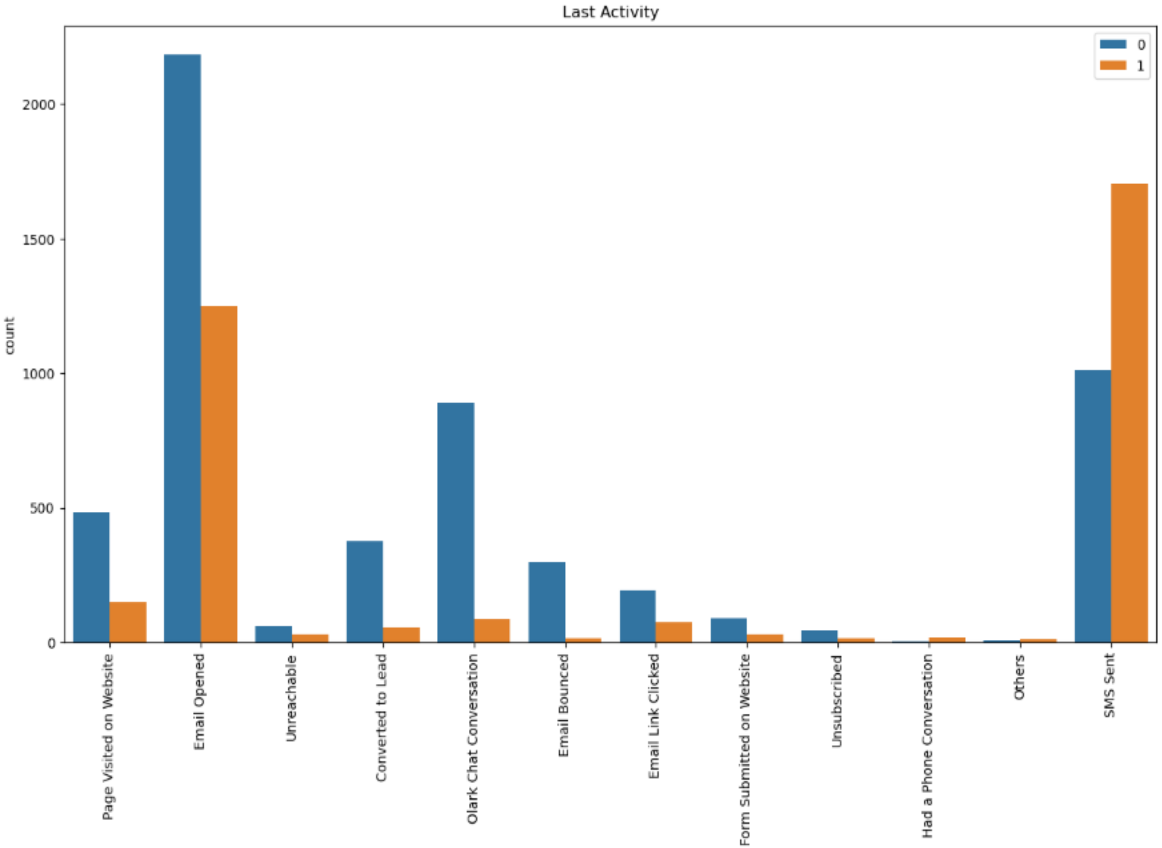
Relation between Tags & Conversion



- Only 6 tag categories seem to have good conversion wrt to total population

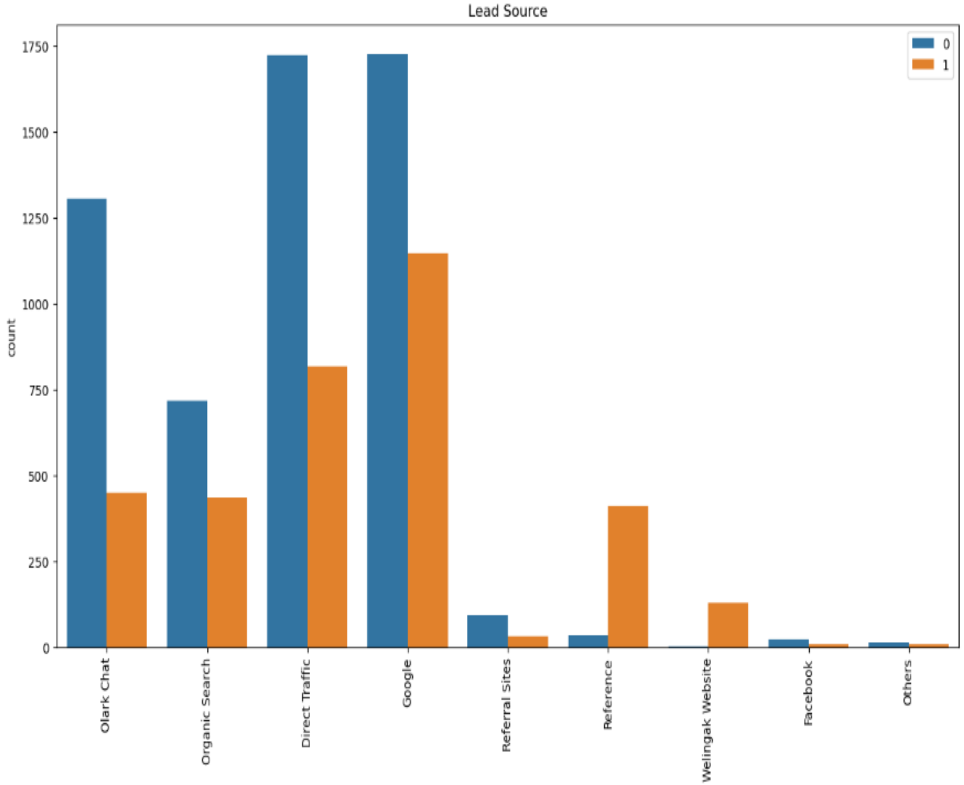
- 'Will revert after reading mail' category have the highest conversion, so it looks largest population of this category responded positively.

Relation between Last Activity & Conversion



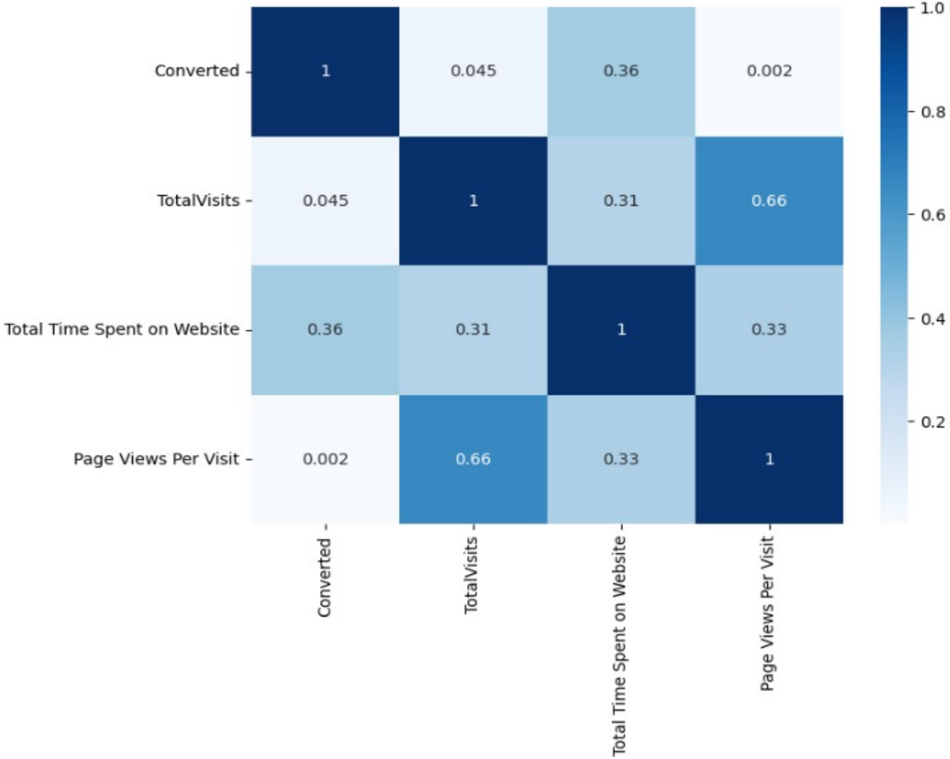
- SMS Sent category have highest conversion rate followed by email opened.

Relation between Lead Source & Conversion



Google source population have highest Conversion, followed by Direct Traffic

CORRELATION



- Page views per visit & Total visits have some good correlation

DATA PREPARATION

- Creation of Dummy variable from categorical value like 'Lead Origin', 'Lead Source', 'Last Activity', 'Specialization',' What is your current occupation' and 'Tags' to get separate feature value wise.

- Selection of target variable is 'Converted'

- Split train & test data with 70% & 30% volume respectively.

- Scaling numerical value like 'TotalVisits','Total Time Spent on Website','Page Views Per Visit'.

MODEL BUILDING

Working on train data:

- Selected most imported 15 features out of variable remains after data cleaning through Recursive Feature Elimination from sklearn.

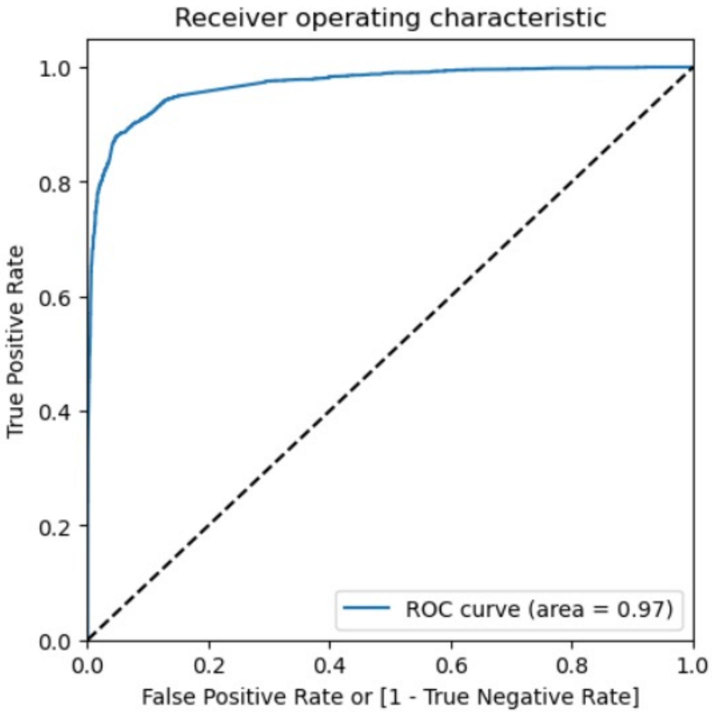
- Recursively build stable model with p value <0.05 and VIF <5 .

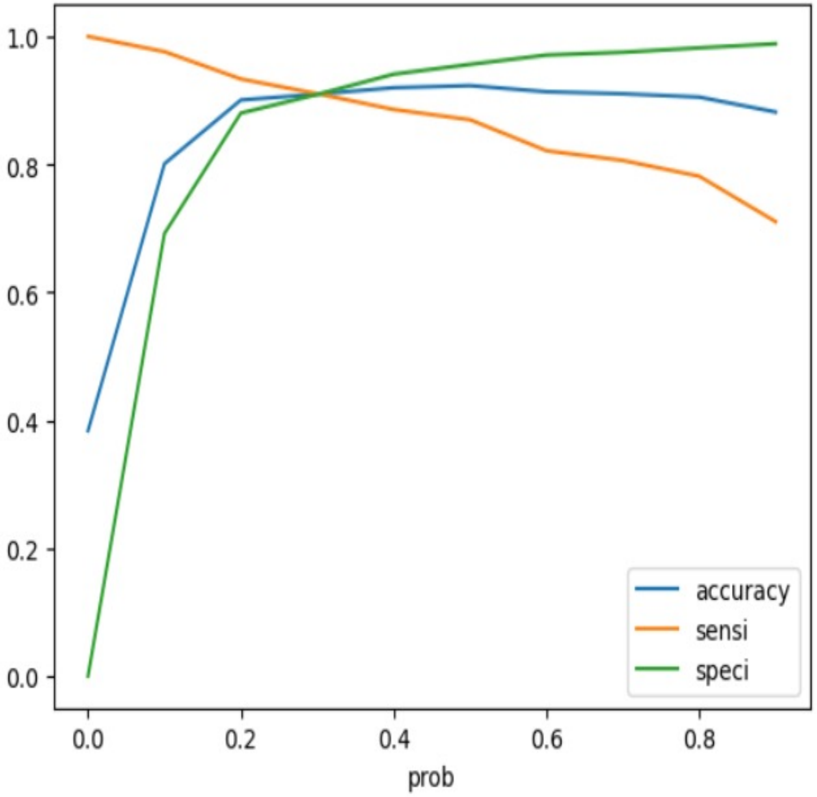
- On 5th Iteration our model is in optimal condition, hence finalizing this .

MODEL EVALUATION

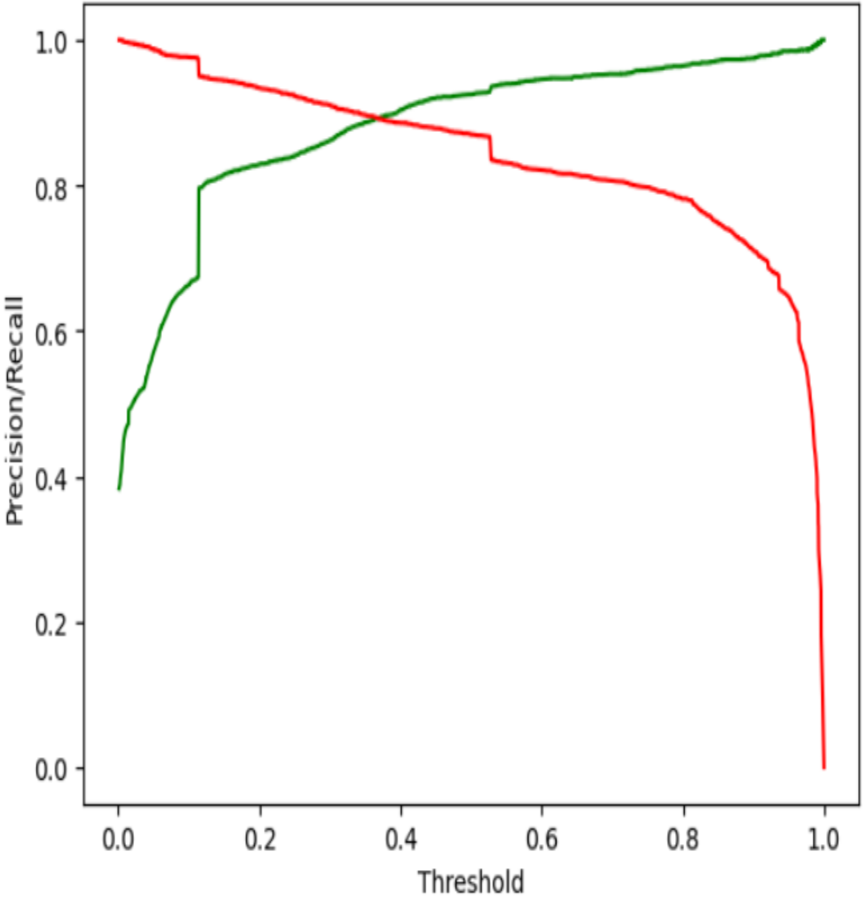
- Train data set is being created with 92% accuracy with random cut off value of 0.5.

- Respective ROC curve is being shown in figure, where curve is highly into left upper portion which Is the indication of good model.





Optimal cut off value is being found as 0.3 is being found from Confusion matrix, with intersection of accuracy, sensitivity & specificity.



- Optimal cut off vale is being found as 0.4 is being found from Precision Recall curve.

TRAIN DATASET EVALUATION

- Predictions were made on the train dataset for the optimal probability cut-off of 0.3
- Confusion Matrix -
  ```
  array([[3557, 355],
         [ 218, 2216]])
  ```
- Accuracy – 90.9%
- Sensitivity - 86.9%
- Specificity – 95.6%

TEST DATASET EVALUATION

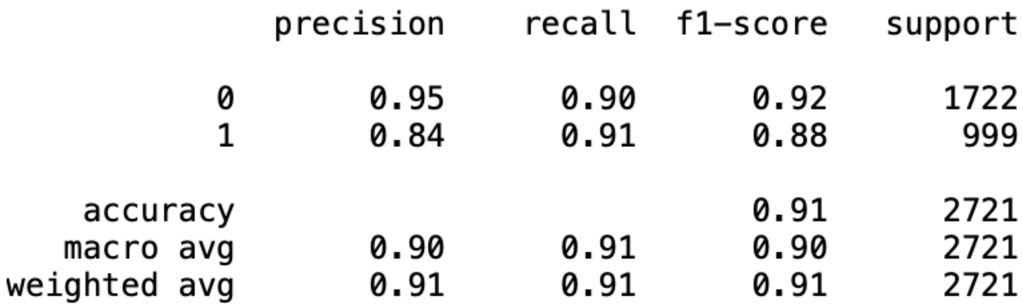
- Predictions were made on the test dataset for the optimal probability cut-off of 0.3
- Confusion Matrix -

```
array([[1554,  168],
       [  85,  914]])
```

- Accuracy – 90.7%
- Sensitivity – 91.4%
- Specificity – 90.2%

MODEL EVALUATION

- Classification Report



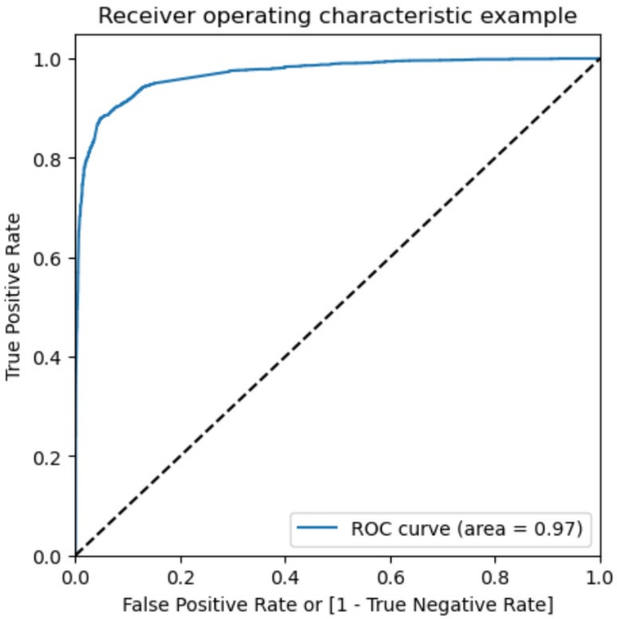
```
              precision    recall  f1-score   support

           0       0.95      0.90      0.92      1722
           1       0.84      0.91      0.88       999

    accuracy                           0.91      2721
   macro avg       0.90      0.91      0.90      2721
weighted avg       0.91      0.91      0.91      2721
```

MODEL EVALUATION

- ROC Curve for the test dataset
- 97% of Area is covered under the curve

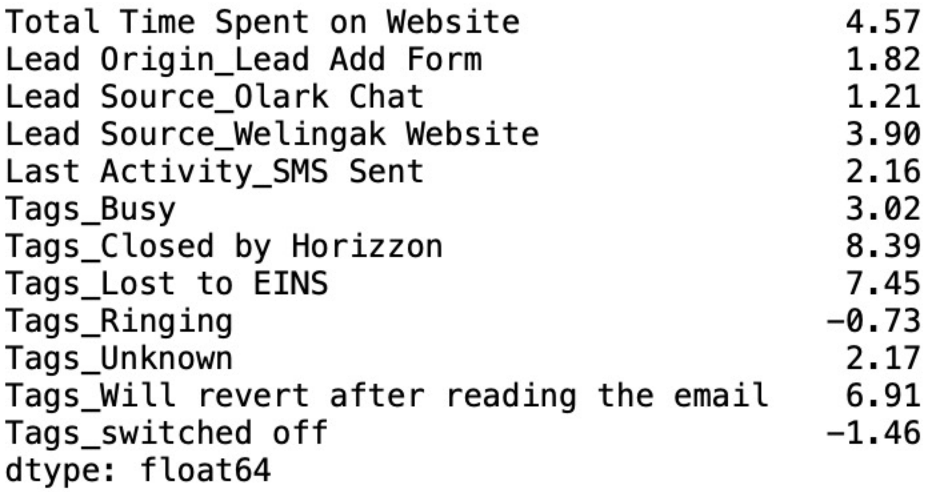


CALCULATING LEAD SCORE

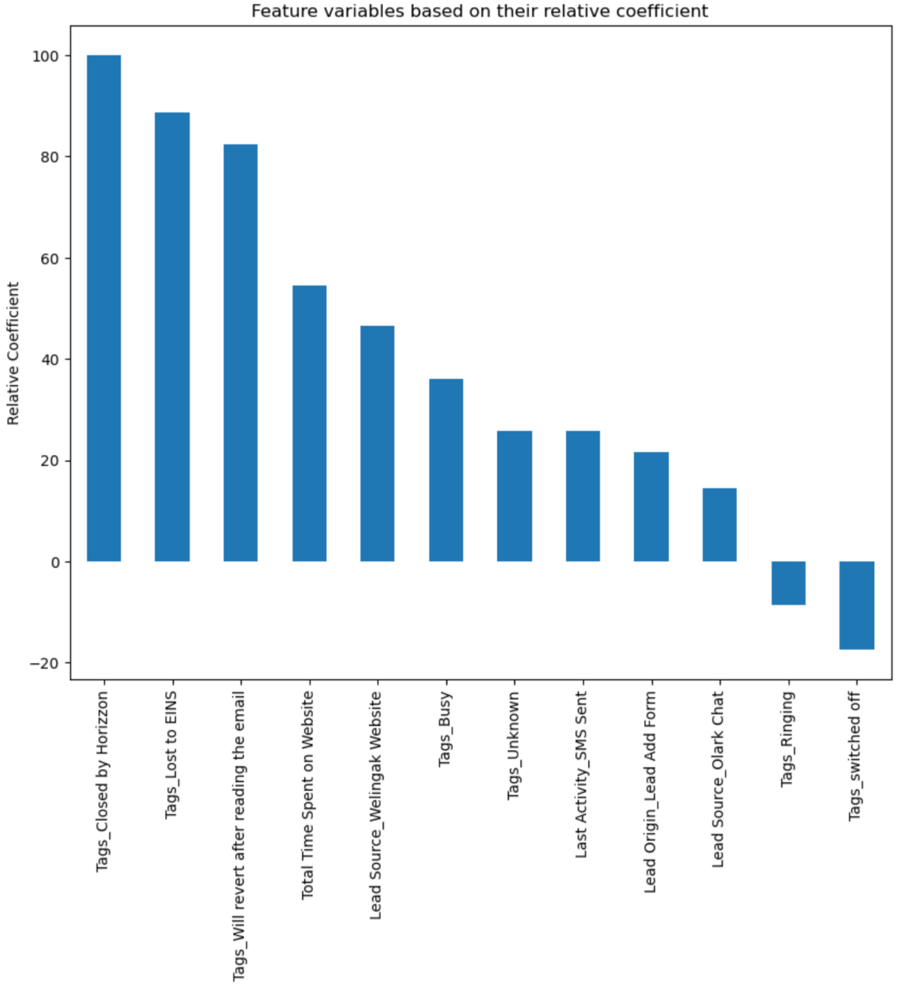
- Lead Score = 100 * probability(Conversion)
- Calculating the Lead Score for each record in the dataset to find their
- conversion probability
- The Leads are identified based upon their unique index value from the original dataset

DETERMINING FEATURE IMPORTANCE

- Coefficients of the features given by the model :

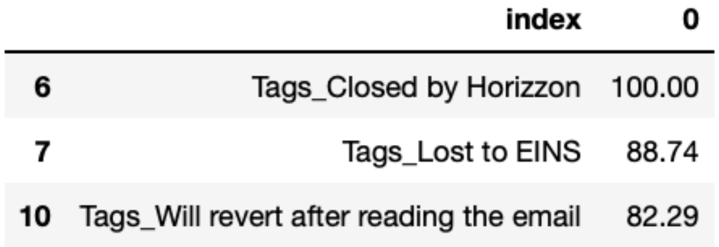


```
Total Time Spent on Website                 4.57
Lead Origin_Lead Add Form                   1.82
Lead Source_Olark Chat                      1.21
Lead Source_Welingak Website                3.90
Last Activity_SMS Sent                      2.16
Tags_Busy                                   3.02
Tags_Closed by Horizzon                     8.39
Tags_Lost to EINS                           7.45
Tags_Ringing                               -0.73
Tags_Unknown                                2.17
Tags_Will revert after reading the email    6.91
Tags_switched off                          -1.46
dtype: float64
```



- Relative coefficient value for all the features with respect to the feature with the highest coefficient

- Top 3 features which contribute most towards the probability of a lead getting converted:



	index	0
6	Tags_Closed by Horizzon	100.00
7	Tags_Lost to EINS	88.74
10	Tags_Will revert after reading the email	82.29

CONCLUSION

- **We decided the final model (Model 5) with the following characteristics:**
- The model selected features have their respective p-value < 0.05.
- The model selected features have very low VIF scores. This implies that there is almost **no mulilticollinearity** among the selected features.
- At the optimal probability cut-off value of 0.3, the overall accuracy of the model on the test and the train dataset is 0.907 and 0.909respectively.

CONCLUSION

- **The top features that contribute the most in predicting the Lead Score**
- Tags_Closed by Horizzon
- Tags_Lost to EINS
- Tags_Will revert after reading the email
- **The feature that are inversely proportional to predicting a lead score. This means that with a decrease in the values of these features, the probability of the Lead Score increases. These are the features with negative coefficient value.**
- Tags_switched off
- Tags_Ringing

THANK YOU