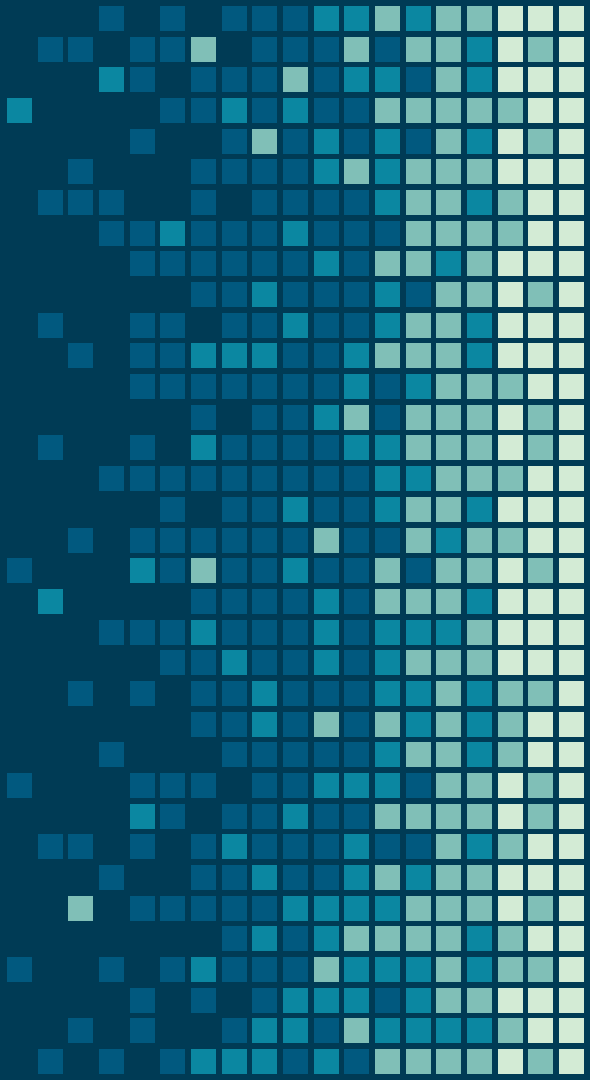
LEAD SCORING CASE STUDY.

By
- Ameya Shukla
- Rohit Ram



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

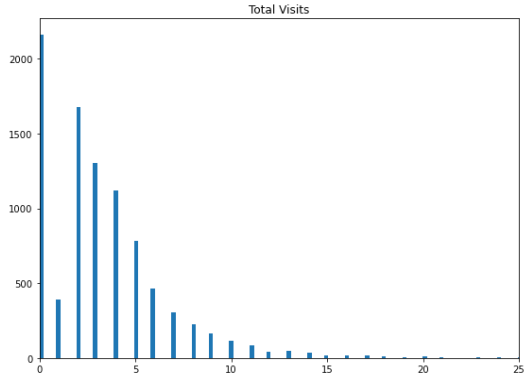
- Calculating the Lead Score

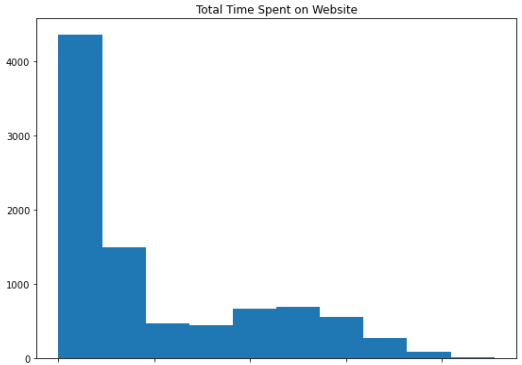
- Determining Feature Importance

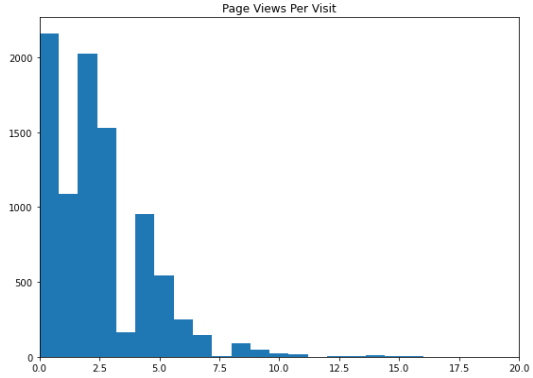
Data Cleaning

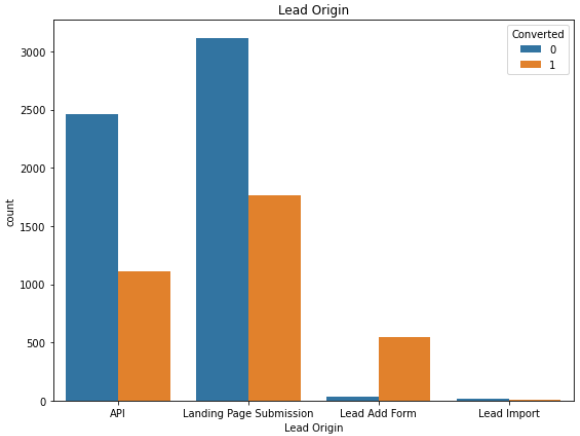
- Dropped features having missing values more than 40%
- Replaced redundant values such as “Select” in some features as Missing Values
- Then, after replacing the values, we dropped the features having more than 40% missing value again
- Dropped rows from features having very few missing values
- Imputed the NULL values of some features with their relevant values
- Dropped Irrelevant features
- Removed extreme outliers of numerical features that could affect our analysis and the model

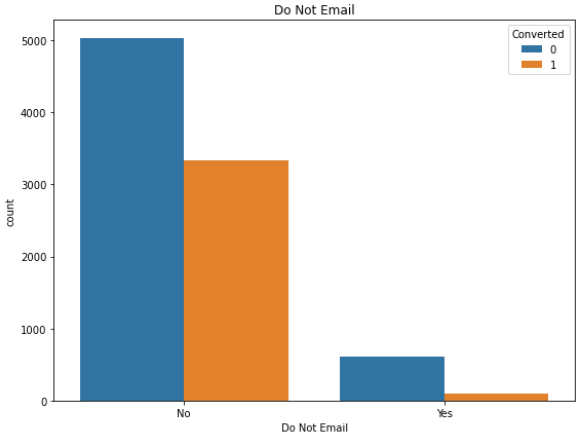
Exploratory Data Analysis (EDA)

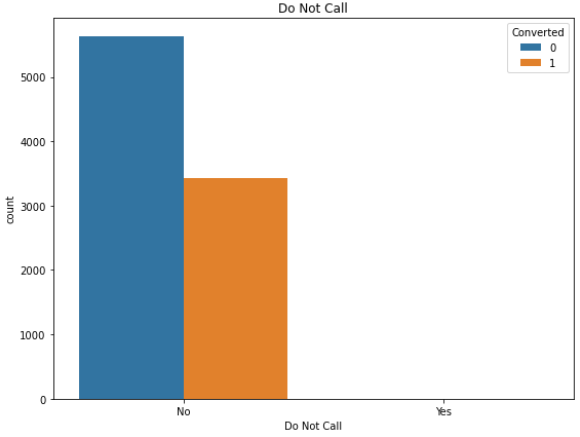


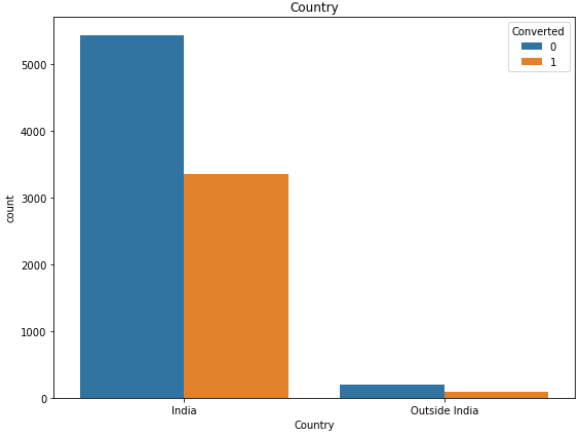


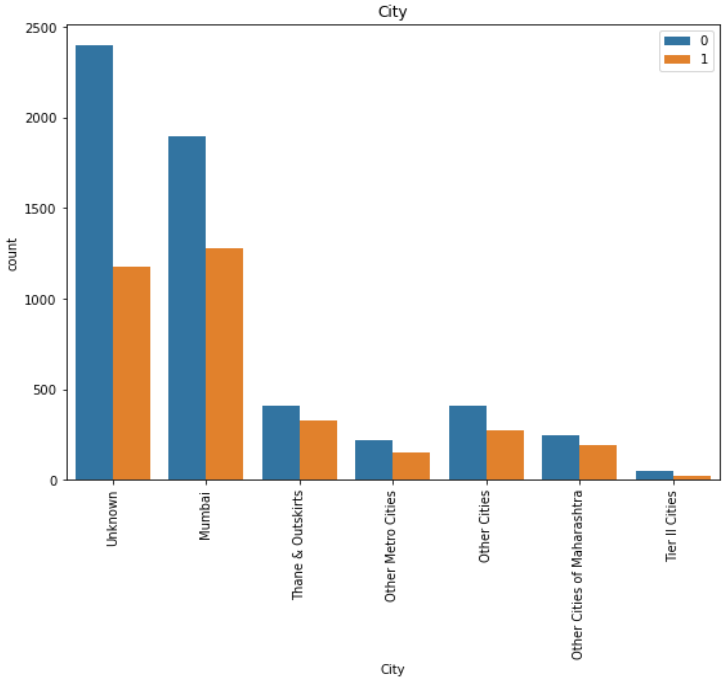


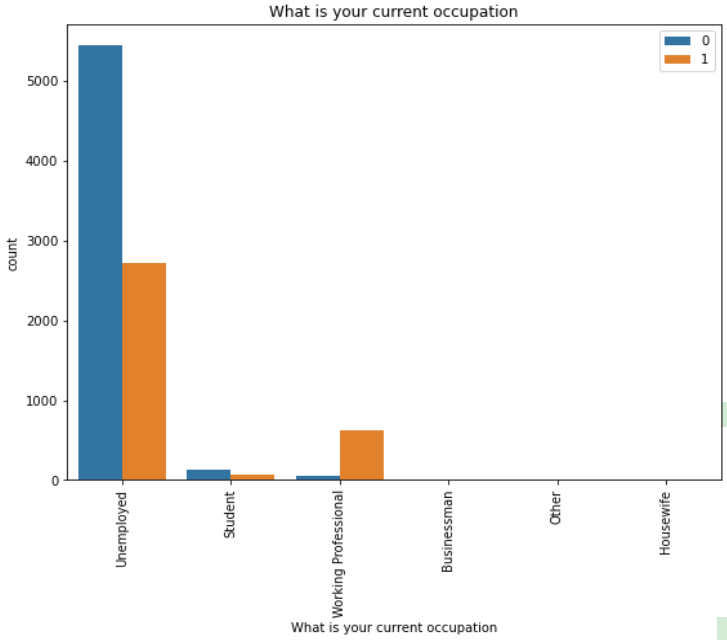

Lead Origin
Converted
0
1
count
3000
2500
2000
1500
1000
500
0
API
Landing Page Submission
Lead Add Form
Lead Import
Lead Origin

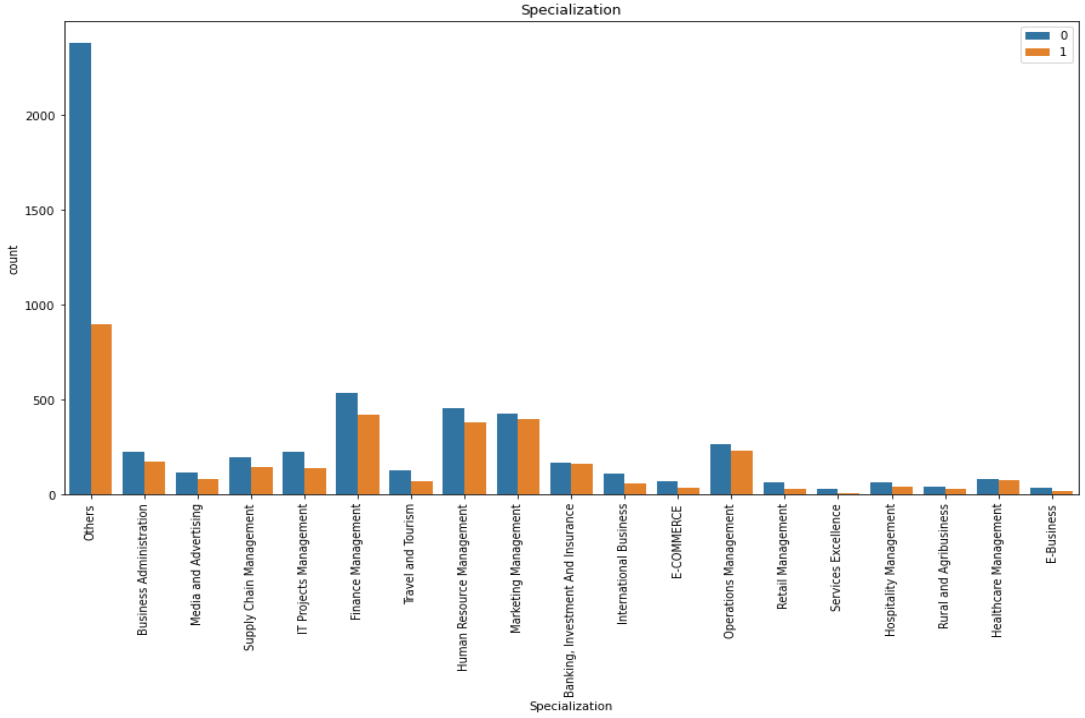

Do Not Email
Converted
0
1
count
5000
4000
3000
2000
1000
0
No
Yes
Do Not Email

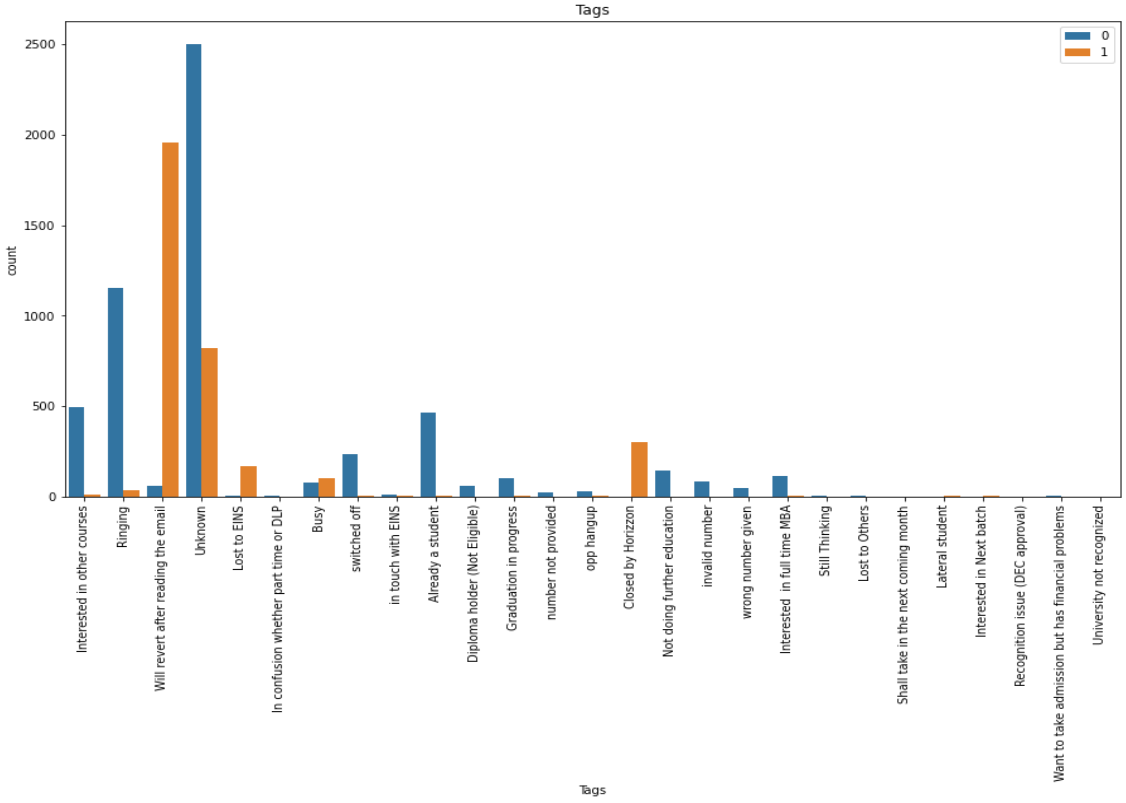

Do Not Call
Converted
0
1
count
5000
4000
3000
2000
1000
0
No
Yes
Do Not Call


Country
Converted
0
1
count
5000
4000
3000
2000
1000
0
India
Outside India
Country

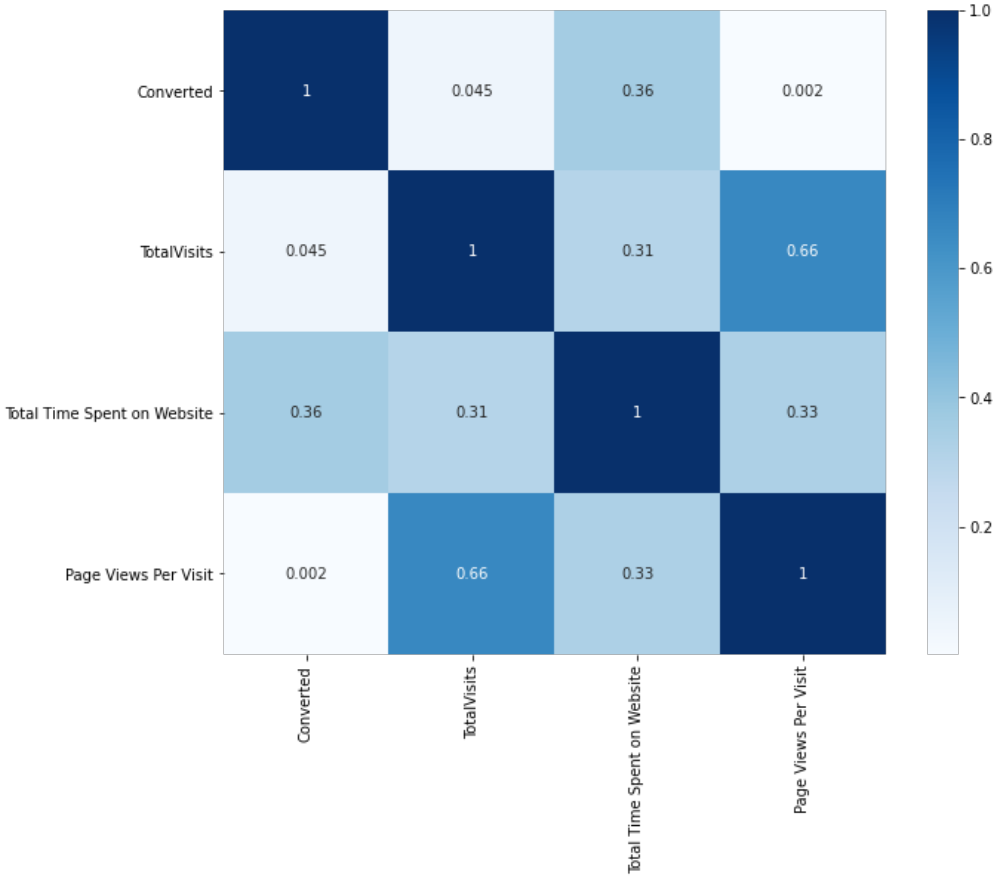

City
count
0
1
2500
2000
1500
1000
500
0
Unknown
Mumbai
Thane & Outskirts
Other Metro Cities
Other Cities
Other Cities of Maharashtra
Tier II Cities
City


What is your current occupation
count
0
1
5000
4000
3000
2000
1000
0
Unemployed
Student
Working Professional
Businessman
Other
Housewife
What is your current occupation


Specialization
0
1
2000
1500
1000
500
0
count
Others
Business Administration
Media and Advertising
Supply Chain Management
IT Projects Management
Finance Management
Travel and Tourism
Human Resource Management
Marketing Management
Banking, Investment And Insurance
International Business
E-COMMERCE
Operations Management
Retail Management
Services Excellence
Hospitality Management
Rural and Agribusiness
Healthcare Management
E-Business
Specialization


Tags
0
1
count
2500
2000
1500
1000
500
0
Interested in other courses
Ringing
Will revert after reading the email
Unknown
Lost to EINS
In confusion whether part time or DLP
Busy
switched off
in touch with EINS
Already a student
Diploma holder (Not Eligible)
Graduation in progress
number not provided
opp hangup
Closed by Horizzon
Not doing further education
invalid number
wrong number given
Interested in full time MBA
Still Thinking
Lost to Others
Shall take in the next coming month
Lateral student
Interested in Next batch
Recognition issue (DEC approval)
Want to take admission but has financial problems
University not recognized
Tags

Correlation Heatmap

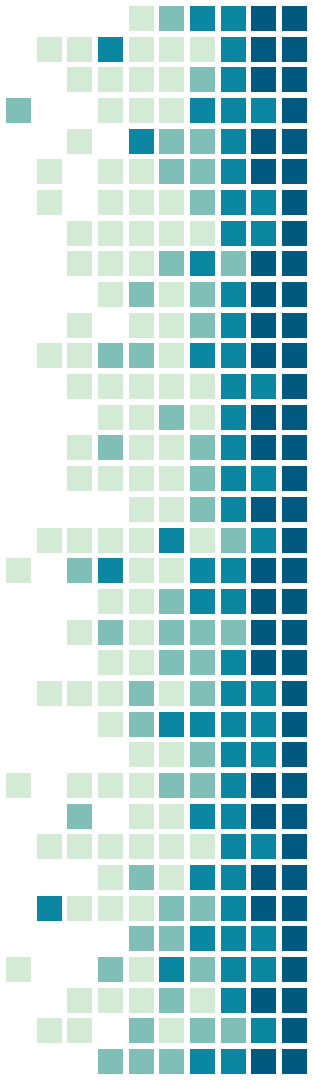


Data Preparation

- Created Dummy Variables for Categorical columns by performing One-Hot Encoding.
- Split the dataset into Train and Test dataset into 70% and 30% respectively.
- Standardized the the dataset using Standard Scaler to bring all the features of the dataset to same magnitude and to avoid overfitting as well.

Model Building

- Created the classification model using statsmodel library
- Built the model on 3 iterations
- The 1st model was built on all the features of the prepared dataset
- Then, the top 15 features for the model were selected by performing Recursive Feature Elimination
- Then, the 2nd model was built on the top 15 features. The features with high p-value from the 2nd model were dropped.
- Then, the 3rd model was built on the updated set of features. The features of 3rd model had very low p-value and VIF scores
- The 3rd model was the final and the most efficient model.



Model Evaluation

- Predictions were made on the test dataset for the optimal probability cut-off of 0.2
- Confusion Matrix -
```
array([[1561,  161],
       [ 130,  869]])
```
- Accuracy – 89.3%
- Sensitivity - 86.9%
- Specificity - 90.7%

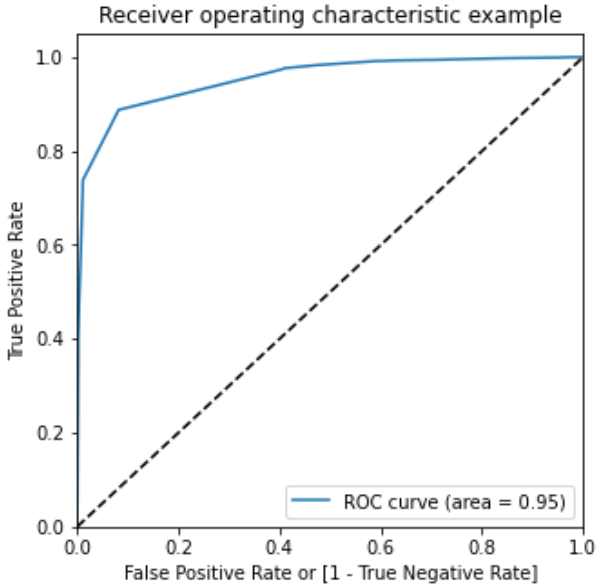
Model Evaluation

- Classification Report

	precision	recall	f1-score	support
0	0.92	0.91	0.91	1722
1	0.84	0.87	0.86	999
accuracy			0.89	2721
macro avg	0.88	0.89	0.89	2721
weighted avg	0.89	0.89	0.89	2721

Model Evaluation

- ROC Curve for the test dataset
- 95% 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

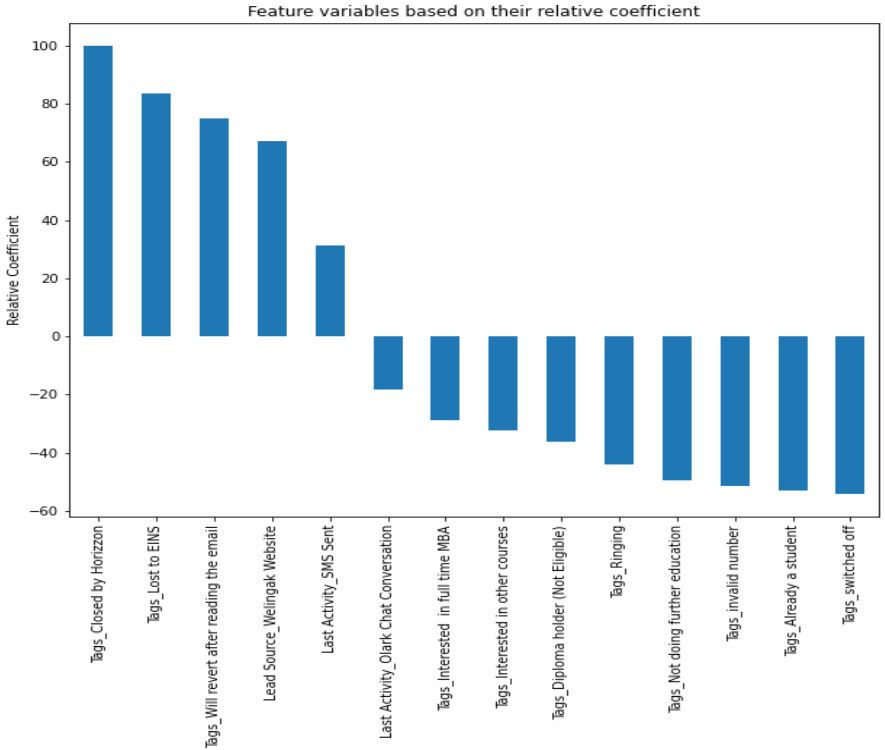
Determining Feature Importance

Coefficients of the features given by the model :

Lead Source_Welingak Website	4.33
Last Activity_Olark Chat Conversation	-1.18
Last Activity_SMS Sent	2.00
Tags_Already a student	-3.42
Tags_Closed by Horizzon	6.43
Tags_Diploma holder (Not Eligible)	-2.32
Tags_Interested in full time MBA	-1.86
Tags_Interested in other courses	-2.07
Tags_Lost to EINS	5.38
Tags_Not doing further education	-3.18
Tags_Ringing	-2.82
Tags_Will revert after reading the email	4.81
Tags_invalid number	-3.31
Tags_switched off	-3.49

Determining Feature Importance

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



Determining Feature Importance

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

	index	0
4	Tags_Closed by Horizzon	100.00
8	Tags_Lost to EINS	83.69
11	Tags_Will revert after reading the email	74.77

CONCLUSION

We decided the final model (`Model 3`) 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 muliticollinearity** among the selected features.
- At the optimal probability cut-off value of `0.2`, the overall accuracy of the model on the test and the train dataset is `0.89` and `0.90` respectively.

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

CONCLUSION

The features 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_Already a student
- Tags_invalid number
- Tags_Not doing further education
- Tags_Ringing
- Tags_Diploma holder (Not Eligible)
- Tags_Interested in other courses
- Tags_Interested in full time MBA
- Last Activity_Olark Chat Conversation

THANK YOU