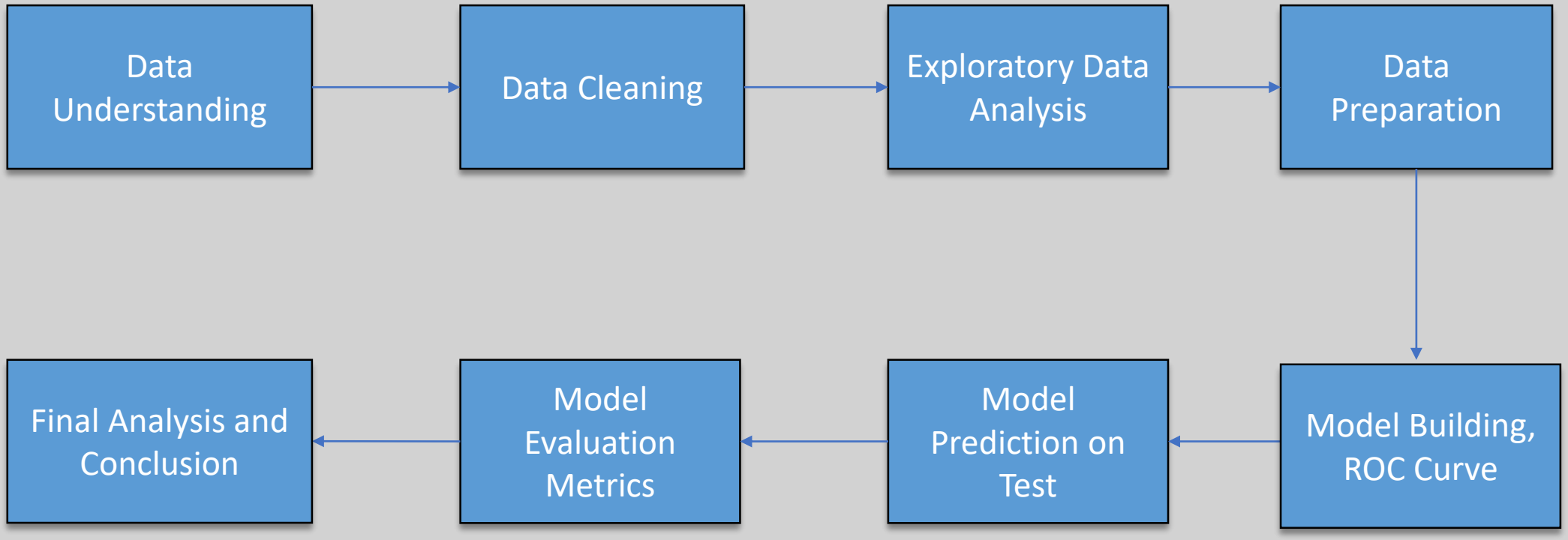
LEAD SCORING CASE STUDY

Submitted by : Snigdha Chakraborty
Abhijith Desai

Business Problem Statement

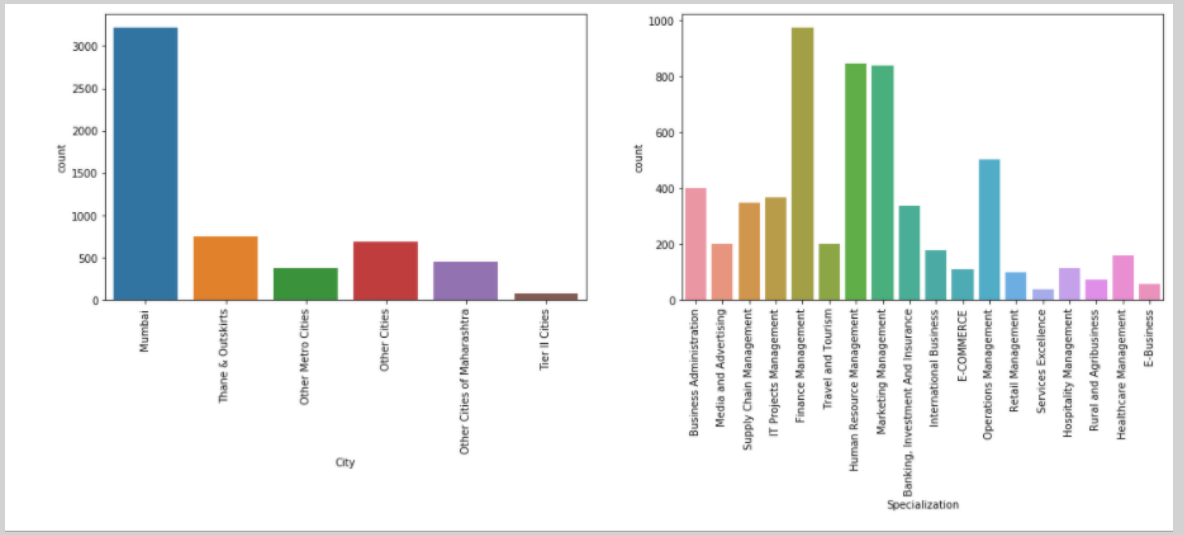
- An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses
- The company markets its courses on several websites and search engines like Google. Once these 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%
- X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you 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 has given a ballpark of the target lead conversion rate to be around 80%

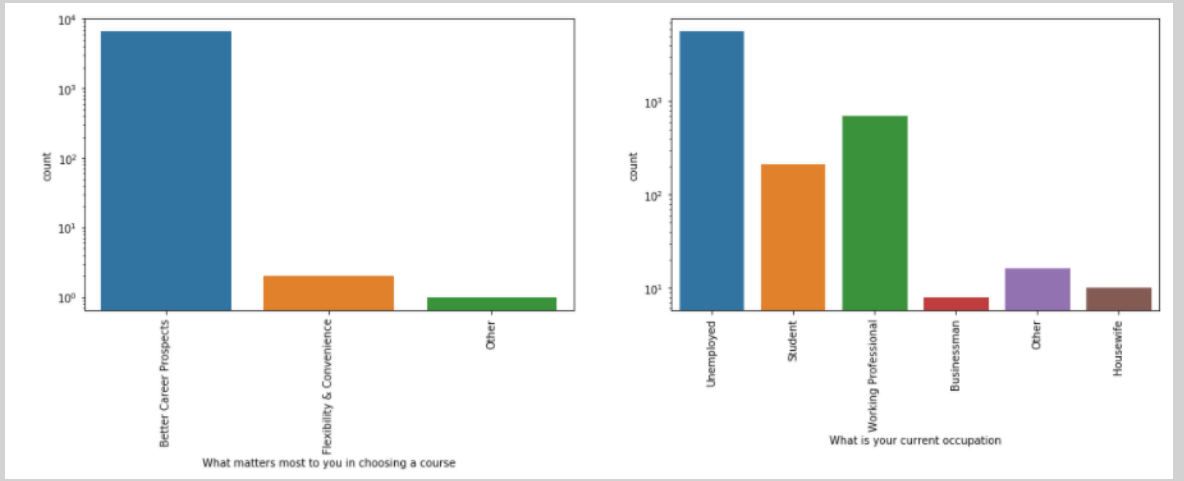
Technical Approach



Data Understanding and Cleaning

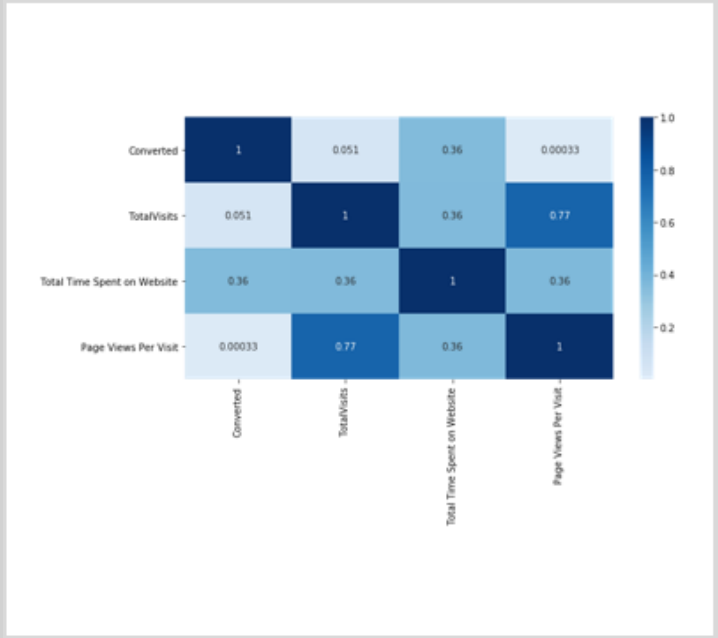
- There were ‘Select’ values in many columns which were as good as Null. Hence, replaced with NaN
- The Columns with >= 40% and rows with <=1% of missing values were dropped
- For others missing values were checked and Imputed using mean/median for numerical and mode for categorical variables such as ‘City’, ‘Specialization’ and so on. Some are shown here: -
- Low frequency items were merged into derived col as ‘Others’
- Post cleaning of data, We moved with 98.2% of original rows for EDA

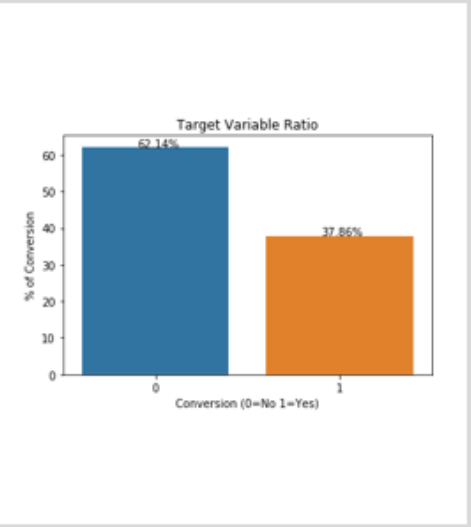


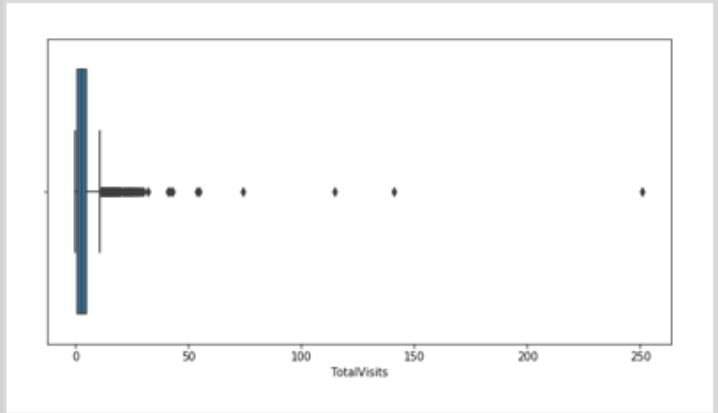


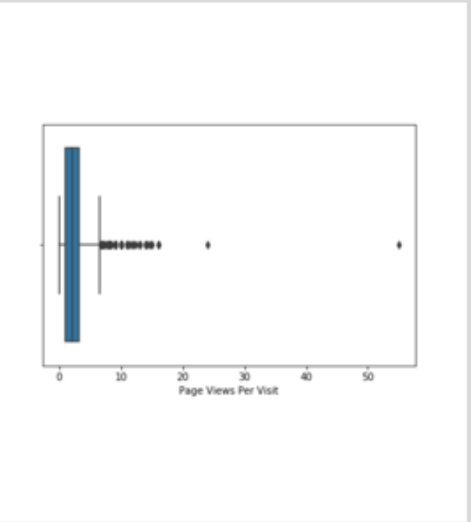
Exploratory Data Analysis

- As part of Data Imbalance check we observed that target variable ratio is 62:38 i.e., ~38% of conversion is available. Hence, data is balanced
- Heat Map helped to confirm correlation
- Univariate and Bivariate used for data spread analysis and relation with target variable
- Outliers were detected and handled using soft capping at 5% and 95%
- Based on the EDA and data inspection we concluded that there are columns not adding any information to the model, hence we can drop them for further analysis
- After all the process of data cleaning we managed to retain 98% of data which is good for model building









Data Preparation

- Dummy variables were created for categorical values
- Variables with binomial values were converted into 0 and 1
- The data was split into train - test
- Numerical Features of training dataset were scaled using standardized scaler
  - The train data was scales using fit_transform()
  - The test data was split using transform()
- The model building was started with final 58 variables

Model Building (RFE)

- Initial Model was build using all features, but stats represented that there were many in-significant variables
- Hence, used RFE to select top 18 variables to start our first model
- With consecutive iteration variables with high p-value >=0.05 and VIF (highly correlated) were subsequently dropped from model building
- Final model was built with 11 variables with VIF and p-value in accepted range
- Prediction was done with 1 if prob >=0.5 else 0
- The model gave an overall accuracy of 89%, sensitivity = 82% and specificity = 93%

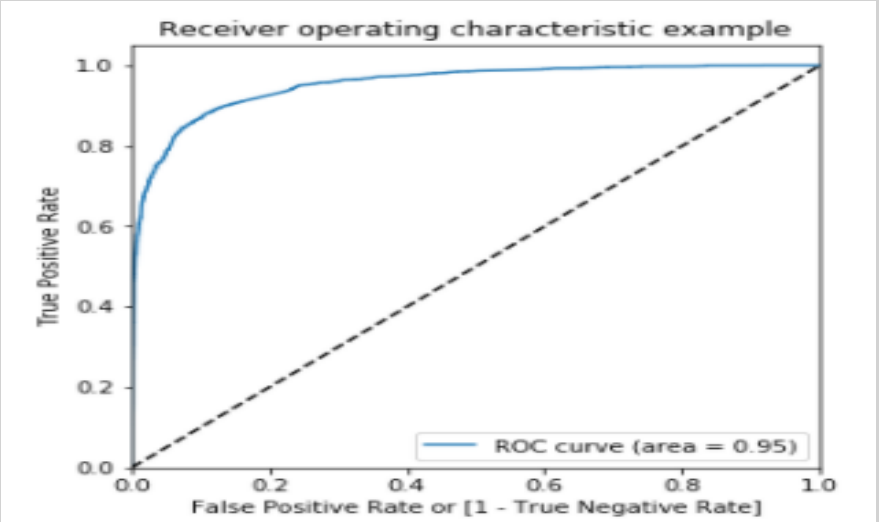
Generalized Linear Model Regression Results

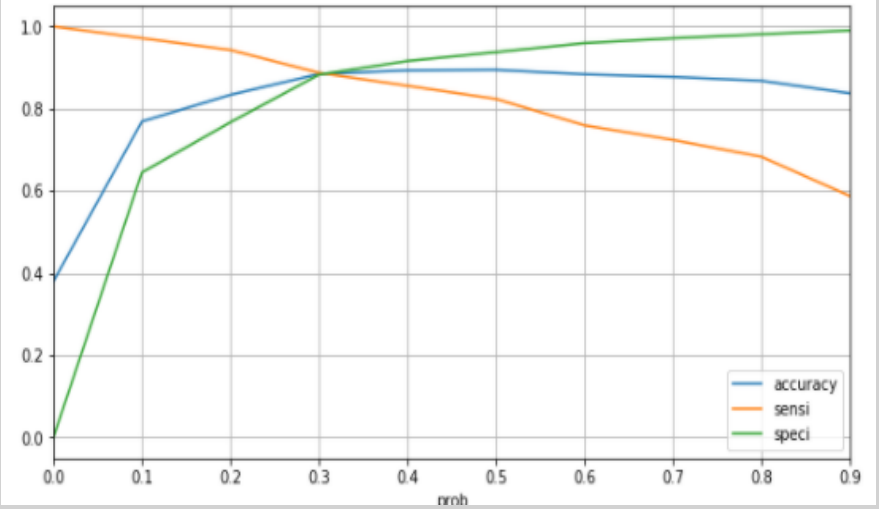
Dep. Variable:	Converted	**No. Observations:**	7259
Model:	GLM	**Df Residuals:**	7247
Model Family:	Binomial	**Df Model:**	11
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1912.2
Date:	Fri, 04 Dec 2020	**Deviance:**	3824.3
Time:	02:04:55	**Pearson chi2:**	9.97e+03
No. Iterations:	8		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-4.9610	0.196	-25.284	0.000	-5.346	-4.576
Do Not Email	-1.5832	0.200	-7.899	0.000	-1.976	-1.190
Total Time Spent on Website	1.1692	0.049	23.865	0.000	1.073	1.265
Lead Source_Olark Chat	1.3847	0.122	11.364	0.000	1.146	1.624
Lead Source_Reference	4.3873	0.268	16.346	0.000	3.861	4.913
Lead Source_Welingak Website	6.8363	1.015	6.736	0.000	4.847	8.825
Last Activity_Olark Chat Conversation	-1.5990	0.187	-8.553	0.000	-1.965	-1.233
Last Activity_Others	1.8784	0.616	3.049	0.002	0.671	3.086
Last Activity_SMS Sent	1.3181	0.091	14.535	0.000	1.140	1.496
Specialization_Travel and Tourism	-0.7478	0.300	-2.491	0.013	-1.336	-0.159
Tags_Others	3.3432	0.188	17.764	0.000	2.974	3.712
Tags_Will revert after reading the email	7.5027	0.244	30.765	0.000	7.025	7.981

ROC and Optimal Cut-Off

- ROC shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity)
- The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test
- The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test
- Optimal cutoff probability is that prob where we get balanced sensitivity and specificity
- From the curve, 0.3 is the optimum point to take it as a cutoff probability
- ROC Curve area = 0.95 indicates an accurate model build



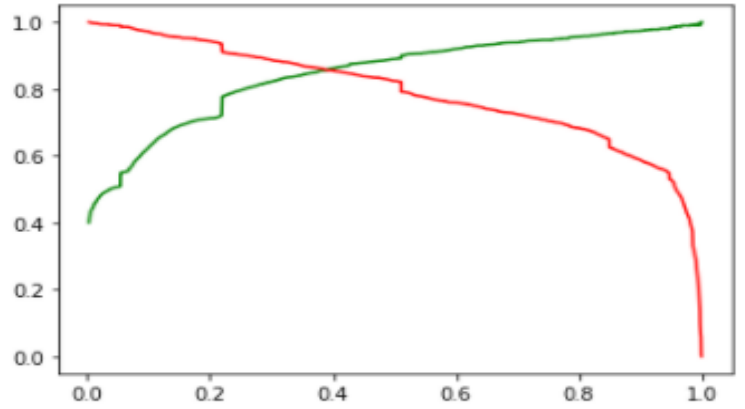


Predictions and Model Summary

- Final model with optimal threshold gave us following performance matrix

	Accuracy	Sensitivity/ Recall	Specificity
Test	87.93	85.42	89.44
Train	88.43	88.72	88.25

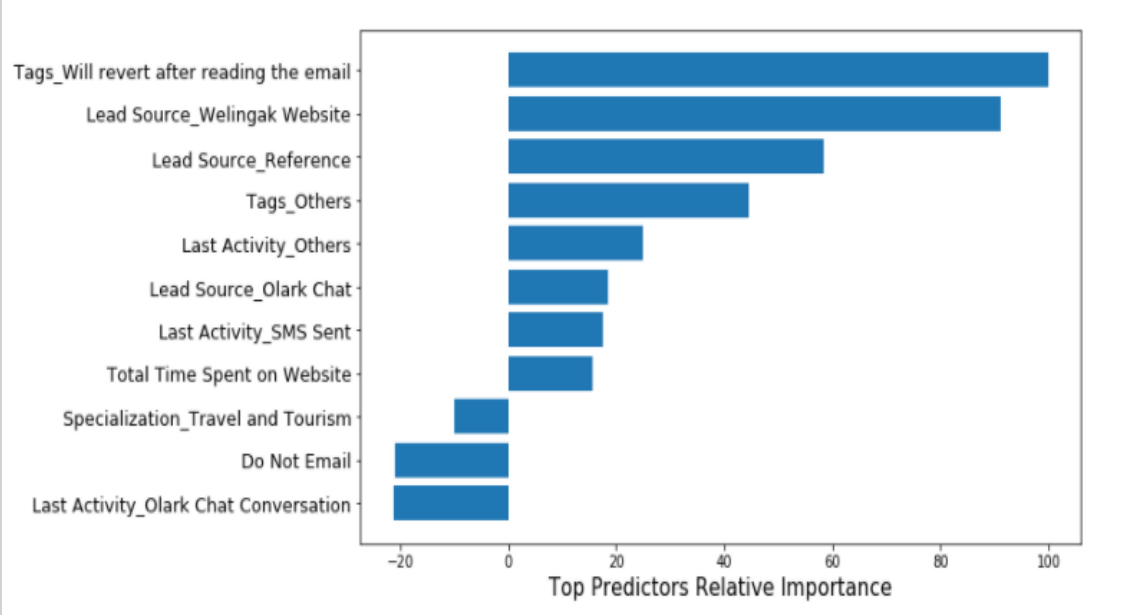
- 85% of Recall value indicates that our model can predict 85% of actual conversion cases correctly
- 83% of Precision value indicates that 83% of the conversions that our model predicted is converted
- Assigned Lead score will help to maximize the lead conversion by choosing high lead score value
- The score will also help sales team to decide conversion techniques as per their business needs



	Prospect ID	Lead Number	Lead_Score
1814	571b5c8e-a5b2-4d57-8574-f2ffb06fdeff	579533	99
307	63926df7-6111-4983-83d9-0ccf2cb5f66d	643546	99
330	12fcbd80-953e-40fb-aaf2-96f5ff7b2961	642612	99
1065	da1da914-737c-4e1f-b66f-d8771694068c	608261	99
1071	d5f11394-7eca-4edb-8979-665b101e0d56	608100	99
317	d8ada66c-170f-41a9-b406-729a88b70778	643151	99
1096	4b77ad5c-95af-443b-9700-8874f26bda7a	607221	99
310	84bec27c-7b3e-4012-91e1-4aadb5a58c7a	643401	99
1112	623bc6c9-9184-4437-b38f-d374be49d1a3	606508	99
1050	396821ef-fd9e-4fec-862b-82c38b43c618	608709	99
301	0fe5cf59-d155-41dc-ba39-4ef62426aa17	643918	99
1116	cbefca85-bd57-43f1-8c8e-b477385aa69e	606316	99
299	e0032a65-ae4e-4221-b8c2-7650899f73ae	643977	99
1123	408173b5-4b49-4244-b4e7-6985ecb78726	606035	99
1129	61de6ddd-47c3-4931-bd2c-75c28c53e647	605652	99
1131	05bc57eb-929b-477a-991e-9d688813ac99	605341	99

Top features for Potential Leads

- Top predictors based on their relative coefficient values :-
  - Tags
    - Will revert after reading the email, Others
  - Lead Source
    - Welingak Website, Reference, Olark Chat
  - Last Activity
    - Others, SMS Sent, Olark Chat Conversation
  - Total Time Spent on Website
  - Specialization
    - Travel and Tourism
  - Do Not Email
- Top 3 variables for potential conversion:-
  - Tags_Will revert after reading the email
  - Lead Source_Welingak Website
  - Lead Source_Reference



Conclusion

- The model helped us to achieve a target lead conversion rate > 80%
- Interpretation using Log Odds: -
- We can also use Log Odds formula along with our final model variable coefficients to find conversation state of any new leads
- Log Odds Formula →

$$\ln(P1-P) = \beta 0X0 + \beta 1X1 + \beta 2X2 + \ldots\ldots\ldots\ldots + \beta nXn$$

Where:

- B0, β1, β2 ……. are coefficient and
- X1, X2, X3 ……. are respective variables

	coef	std err	z	P>|z|	[0.025	0.975]
const	-4.9610	0.196	-25.284	0.000	-5.346	-4.576
Do Not Email	-1.5832	0.200	-7.899	0.000	-1.976	-1.190
Total Time Spent on Website	1.1692	0.049	23.865	0.000	1.073	1.265
Lead Source_Olark Chat	1.3847	0.122	11.364	0.000	1.146	1.624
Lead Source_Reference	4.3873	0.268	16.346	0.000	3.861	4.913
Lead Source_Welingak Website	6.8363	1.015	6.736	0.000	4.847	8.825
Last Activity_Olark Chat Conversation	-1.5990	0.187	-8.553	0.000	-1.965	-1.233
Last Activity_Others	1.8784	0.616	3.049	0.002	0.671	3.086
Last Activity_SMS Sent	1.3181	0.091	14.535	0.000	1.140	1.496
Specialization_Travel and Tourism	-0.7478	0.300	-2.491	0.013	-1.336	-0.159
Tags_Others	3.3432	0.188	17.764	0.000	2.974	3.712
Tags_Will revert after reading the email	7.5027	0.244	30.765	0.000	7.025	7.981