Lead Conversion case Study

Shailesh Khulbe
Rohit Gupta
Vikram Rajawat

Problem Statement

Identify the most promising leads and targeting them from the pool of generated leads that can will be joining the course which in turn will save manpower and cost involved in communication with the leads.

Approach Used

- DATA Analysis
- Data Cleaning: Missing Value Treatment, Outlier Treatment
- Univariate, Bi Variate, Multi Variate Analysis
- Final Data Preparation
- Model Building
- RFE for feature Selection
- Iterations to Select Final Model
- Metrices for Model strength
- Prediction On Test Set

Problem Solving Method

Data Import : Import data from csv file and check if data loaded correctly.

Data Preparation : Remove unwanted columns which has more than 99.5% in a single category, apply data cleaning on remaining columns such as missing value treatment, outlier treatment , binning if required and segregate numerical and categorical columns.

Analysis : Perform univariate, bi variate and multivariate to understand the range of data and how it is related to TARGET variable and decide which variables have impact on Target variab

Final Data Preparation:-Converting categorical variables to Dummy variables and making data ready for modelling.

Model Building : Building model with remaining variables, scaling the variables, using RFE , Analysing VIF and P value , Iterating to reach the final model.

Validation : Validating the model on various metrices and test data.

Data Preparation Steps

Removing Columns

- Remove columns with 40% missing values as we have 2 numerical columns meeting the criteria.
  Removing columns which have more than 99.5% in one category and ID columns

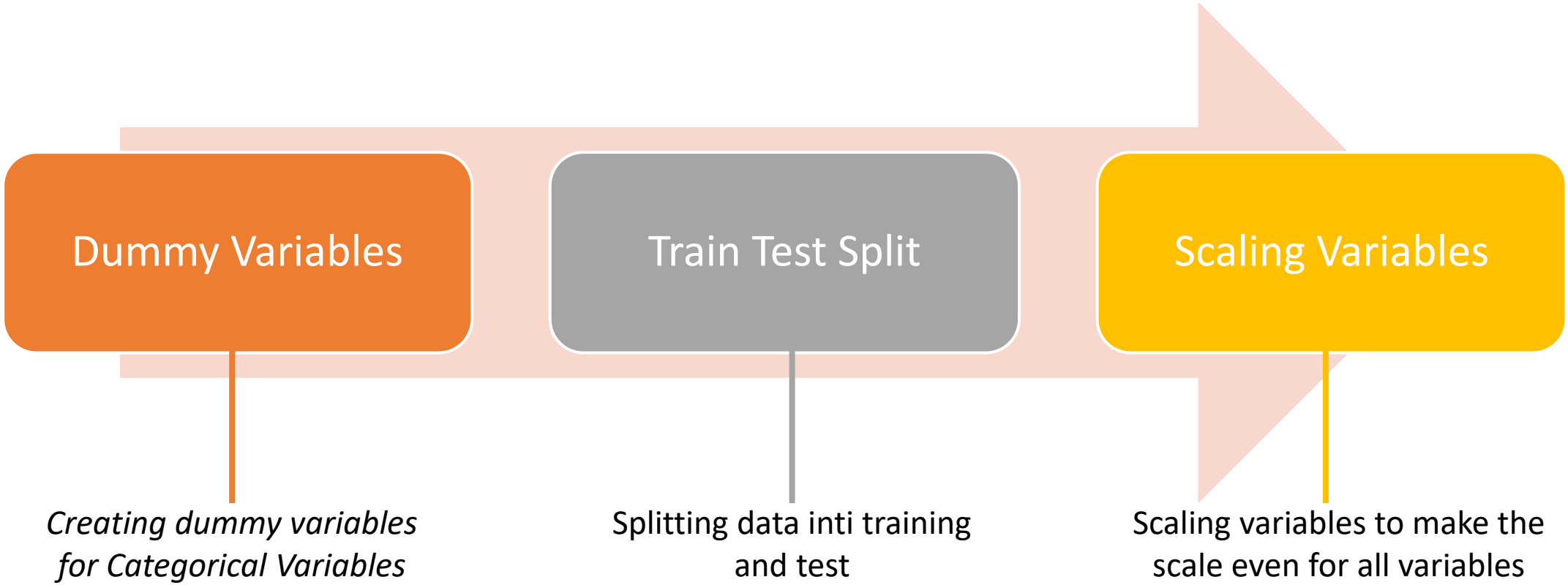
Missing Value/Outlier Treatment

- Impute missing values for remaining columns with mean or median.
- Replace with mean when variable is continuous and does not have outlier otherwise with median
- Replace with mode when variable is categorical
- Outliers are values that are different from the normal population. Replace outliers quantiles or by capping using +,- (IQR*1.5) for 25th and 75^{th} percentile or business understanding whichever is applicable.
- Delete rows for columns that have very few missing values.
- Deleting columns which have SELECT as value and missing resulting in more than 40% missing values.
- Grouping categories with less than 5% values to OTHER category.

Univariate/Bivariate

- Analyse cleaned data using Histogram, box plot, pair plot etc.
- For numerical variables use histogram, box plot, scatter plot, pair plot etc.
- For categorical use value counts, bar plot, pi chart ,group by etc.
- Find variables which have high influence on TARGET VARIABLE.

Data Preparation For Modelling

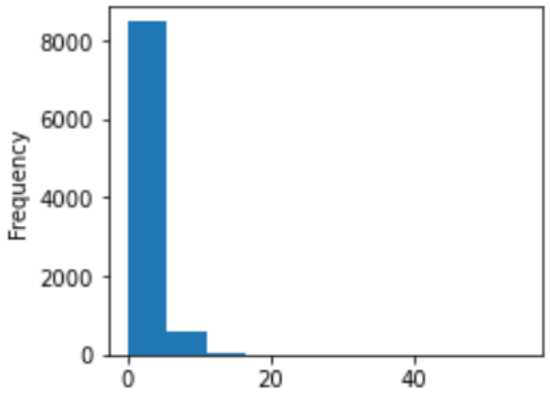


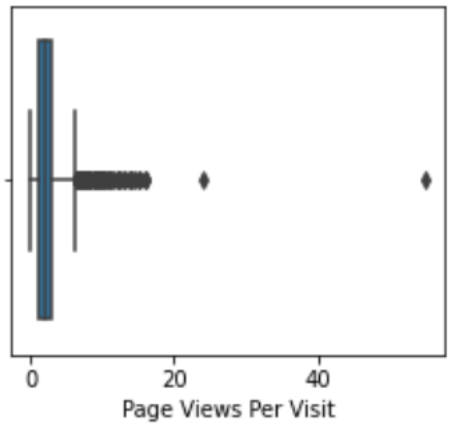
Model Building

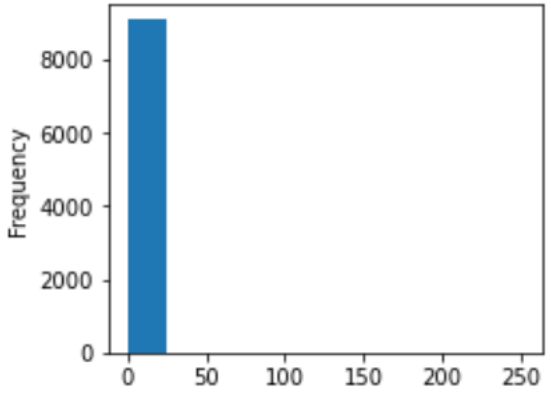
1. Use RFE for variable selection
2. Remove variables with high VIF and P value
3. Re run the model to receive a stable model
4. Validate the results on various matrices
5. Validate the result on Test data

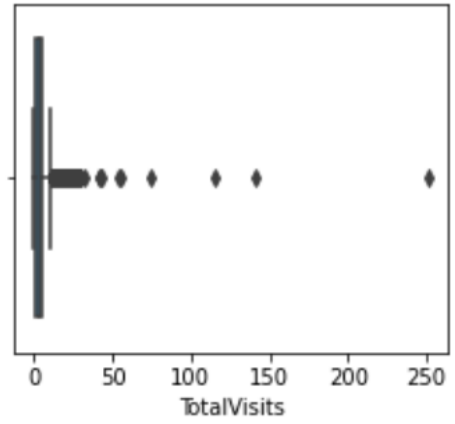
Numerical Variable Analysis

Outliers using box plot

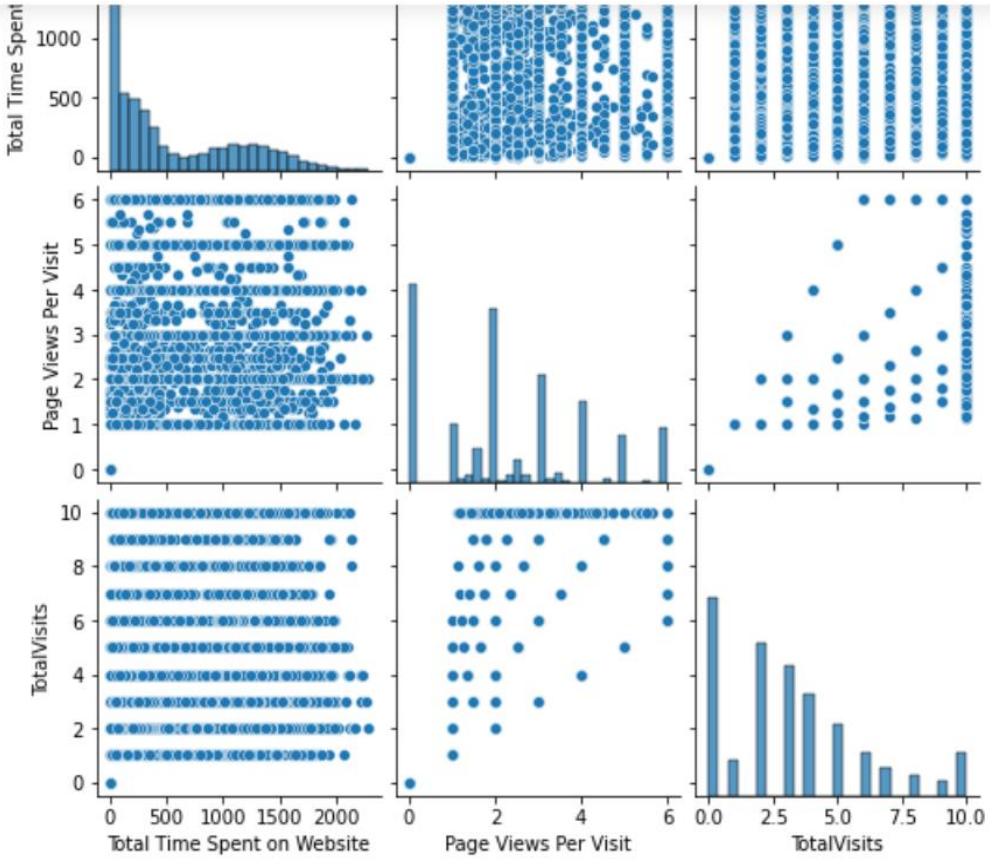


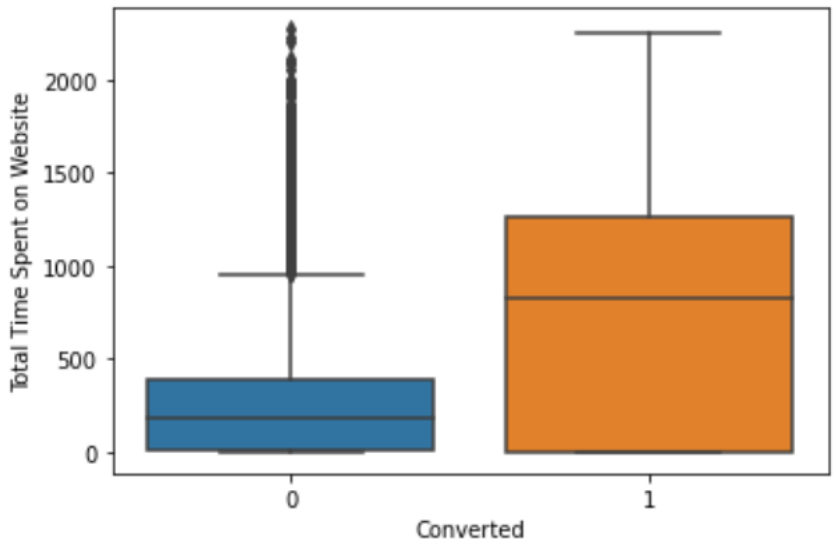




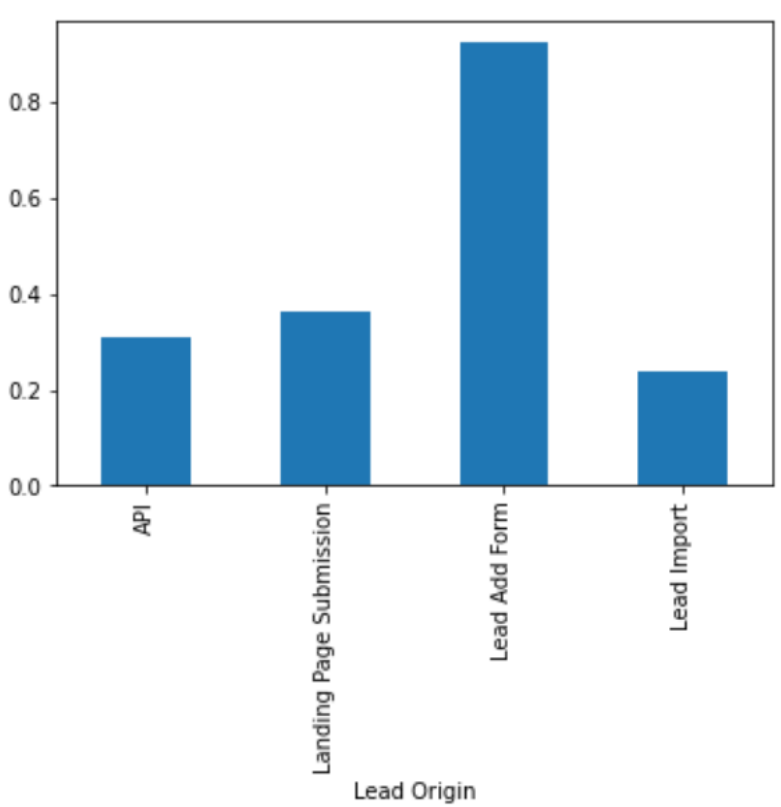


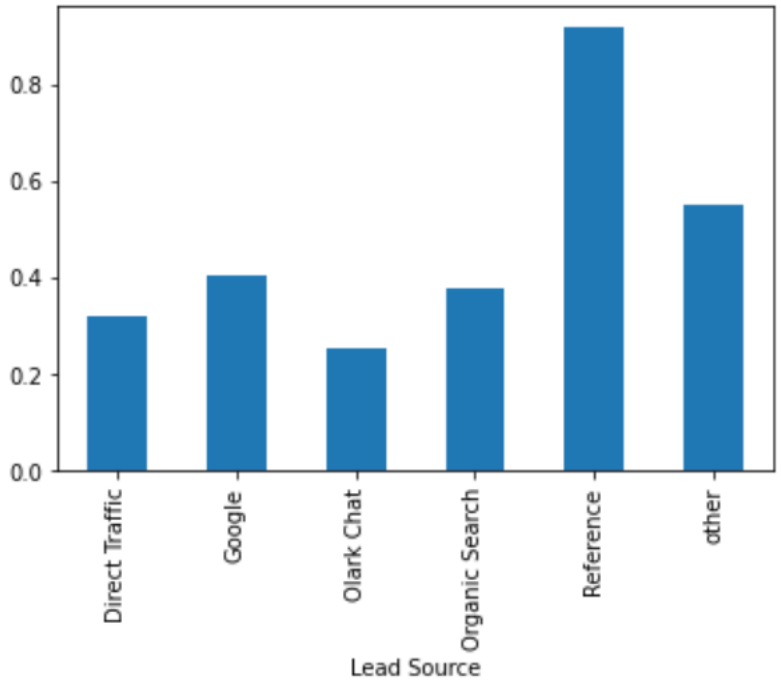
Numerical Bi-variate And Target Variable Analysis





Categorical Variable Analysis With Target Variable





Multivariate Analysis With Target Variable

First Iterations Results

Dep. Variable:	Converted	**No. Observations:**	6467
Model:	GLM	**Df Residuals:**	6451
Model Family:	Binomial	**Df Model:**	15
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1970.3
Date:	Sun, 26 Feb 2023	**Deviance:**	3940.6
Time:	21:50:46	**Pearson chi2:**	9.04e+03
No. Iterations:	8	**Pseudo R-squ. (CS):**	0.5169
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-4.8976	0.347	-14.131	0.000	-5.577	-4.218
Do Not Email	-1.6431	0.183	-8.997	0.000	-2.001	-1.285
Total Time Spent on Website	1.2842	0.050	25.617	0.000	1.186	1.382
Lead Origin_Lead Add Form	3.7761	0.246	15.380	0.000	3.295	4.257
Lead Source_Olark Chat	1.0281	0.119	8.672	0.000	0.796	1.260
Lead Source_other	0.6424	0.227	2.827	0.005	0.197	1.088
Last Activity_Olark Chat Conversation	-1.0988	0.180	-6.121	0.000	-1.451	-0.747
Last Activity_SMS Sent	0.5725	0.171	3.353	0.001	0.238	0.907
What is your current occupation_Working Professional	2.7517	0.244	11.259	0.000	2.273	3.231
What is your current occupation_other	1.4302	0.295	4.850	0.000	0.852	2.008
Tags_Closed by Horizzon	8.8183	0.805	10.951	0.000	7.240	10.397
Tags_Ringing	-0.3814	0.412	-0.925	0.355	-1.189	0.427
Tags_Will revert after reading the email	4.2518	0.343	12.382	0.000	3.579	4.925
Tags_other	3.0785	0.354	8.705	0.000	2.385	3.772
Last Notable Activity_Modified	-0.6881	0.110	-6.274	0.000	-0.903	-0.473
Last Notable Activity_SMS Sent	1.3961	0.201	6.943	0.000	1.002	1.790

Final Iterations Results

Dep. Variable:	Converted	**No. Observations:**	6467
Model:	GLM	**Df Residuals:**	6453
Model Family:	Binomial	**Df Model:**	13
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1995.1
Date:	Sun, 26 Feb 2023	**Deviance:**	3990.3
Time:	22:02:30	**Pearson chi2:**	8.50e+03
No. Iterations:	8	**Pseudo R-squ. (CS):**	0.5132
Covariance Type:	nonrobust		

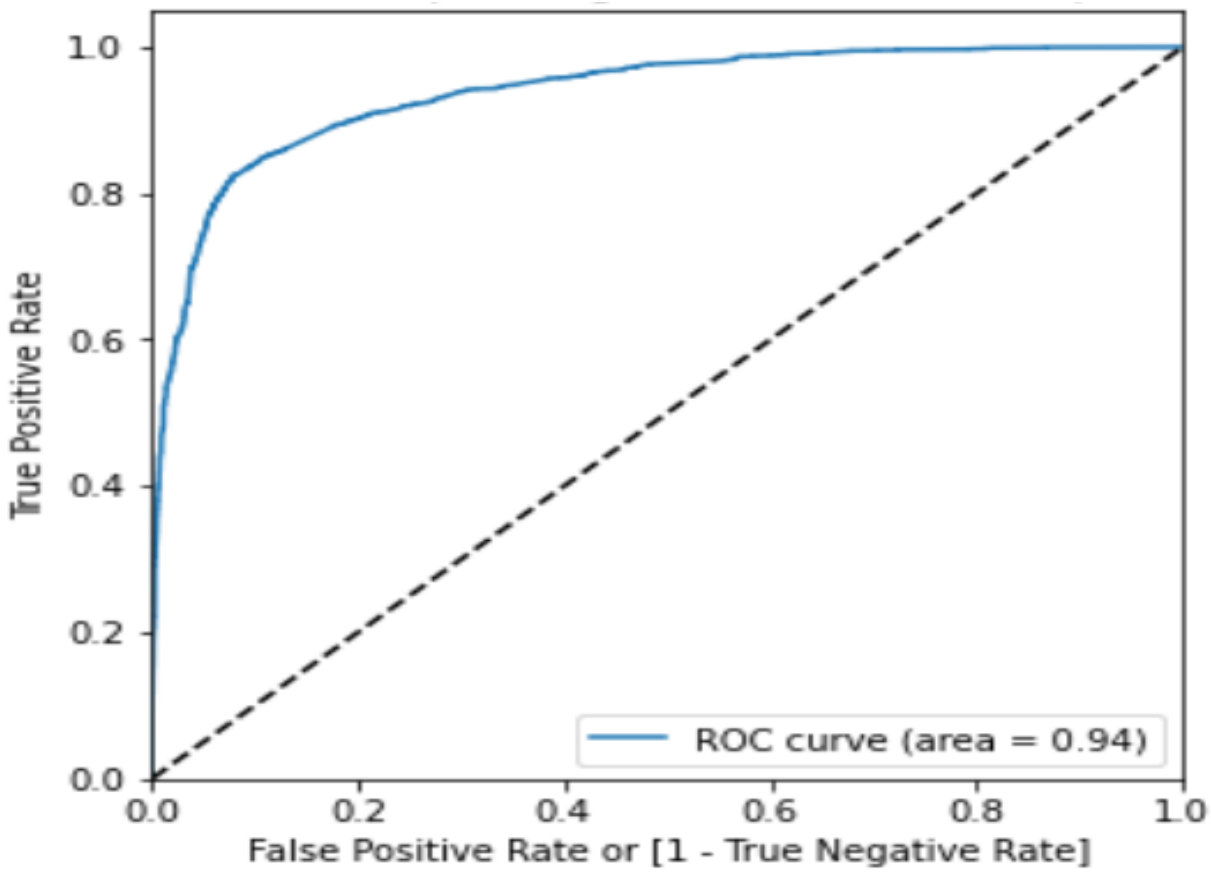
	coef	std err	z	P>|z|	[0.025	0.975]
const	-4.9468	0.205	-24.134	0.000	-5.349	-4.545
Do Not Email	-1.5872	0.182	-8.711	0.000	-1.944	-1.230
Total Time Spent on Website	1.2902	0.050	25.746	0.000	1.192	1.388
Lead Origin_Lead Add Form	3.7703	0.244	15.421	0.000	3.291	4.250
Lead Source_Olark Chat	1.0381	0.117	8.884	0.000	0.809	1.267
Lead Source_other	0.6020	0.226	2.669	0.008	0.160	1.044
Last Activity_Olark Chat Conversation	-0.9141	0.178	-5.122	0.000	-1.264	-0.564
Last Activity_SMS Sent	1.5774	0.089	17.665	0.000	1.402	1.752
What is your current occupation_Working Professional	2.7666	0.245	11.291	0.000	2.286	3.247
What is your current occupation_other	1.4875	0.295	5.042	0.000	0.909	2.066
Tags_Closed by Horizzon	9.0541	0.758	11.940	0.000	7.568	10.540
Tags_Will revert after reading the email	4.3959	0.200	21.960	0.000	4.004	4.788
Tags_other	3.2939	0.217	15.203	0.000	2.869	3.719
Last Notable Activity_Modified	-1.1251	0.093	-12.100	0.000	-1.307	-0.943

Validations Results

```
In [270]: # Let's see the sensitivity of our logistic regression model
          TP / float(TP+FN)

Out[270]: 0.8085782366957903

In [271]: # Let us calculate specificity
          TN / float(TN+FP)

Out[271]: 0.9278298303367941

In [272]: # Calculate false postive rate - predicting Converted when customer does not convert
          print(FP/ float(TN+FP))

          0.07217016966320587

In [273]: # positive predictive value
          print (TP / float(TP+FP))

          0.8772080999569152

In [274]: # Negative predictive value
          print (TN / float(TN+ FN))

          0.88374336710082
```

AMT_ANNUITY

ROC Curve


True Positive Rate
False Positive Rate or [1 - True Negative Rate]
ROC curve (area = 0.94)
0.0
0.2
0.4
0.6
0.8
1.0

Prediction On test Set

```
# Let's check the overall accuracy.
metrics.accuracy_score(y_pred_final.Converted, y_pred_final.final_predicted)
```

```
0.8643578643578643
```

```
# Let's see the sensitivity of our logistic regression model
TP / float(TP+FN)
```

```
0.836852207293666
```

```
# Let us calculate specificity
TN / float(TN+FP)
```

```
0.8809248554913295
```

Result in Business Terms

	coef	Absolute
const	-4.3972	4.3972
Tags_Will revert after reading the email	4.3024	4.3024
Lead Origin_Lead Add Form	4.2262	4.2262
Tags_other	3.7069	3.7069
What is your current occupation_Working Professional	2.9169	2.9169
Last Activity_Olark Chat Conversation	-2.1973	2.1973
Last Activity_Page Visited on Website	-1.4706	1.4706
Last Activity_other	-1.4646	1.4646
Total Time Spent on Website	1.2784	1.2784
Do Not Email	-1.2691	1.2691
What is your current occupation_other	1.2638	1.2638
Lead Source_Olark Chat	0.9581	0.9581
Last Notable Activity_other	0.6534	0.6534
Lead Source_other	0.3445	0.3445

- Overall we see 13 variables coming the final model equation. Few top variables of these are

1) Tags assigned to the customer indicating the current status of the lead
2) Lead Origin – The origin identifier with which the customer was identified. So, one which is added by add form is is contributing
3) Current Occupation Working Profession is adding a lot of value