Lead Scoring Case Study

PRESENTED BY Kapil

PROBLEM STATEMENT

An education company named X Education sells online courses to industry professionals leveraging several websites and search engines. People who browse the courses or fill up a form or watch videos about content may fill up a form or provide their contact information thus becoming leads. Leads are identified through referrals too. The company wants to improve the conversion rate for the leads. X Education has reached out to us to help them select the most promising leads. We are provided with a leads dataset containing 9000 plus datapoints.

STRATEGY

We will build a logistic regression model to assign a lead score to each of the leads. The higher lead score means that the lead will have higher likelihood of conversion.

EXECUTION

The execution steps that we followed are

1) Importing the data files and understand the data
2) Data clean up and treat outliers
3) Data Preparation for modelling
4) Model Building using RFE and manual method
5) Model Evaluation
6) Validations on the test dataset
7) Recommendations.

1) Initial Data Analysis

Leads Data

Rows
9240
Columns
37

Importing the data files and understand the data

Leads.csv

- Contains all the information of the people who have shown interest in the course by browsing the online education courses or filling up a form or watching videos about the online content etc.
- The data is about conversion rate for the leads

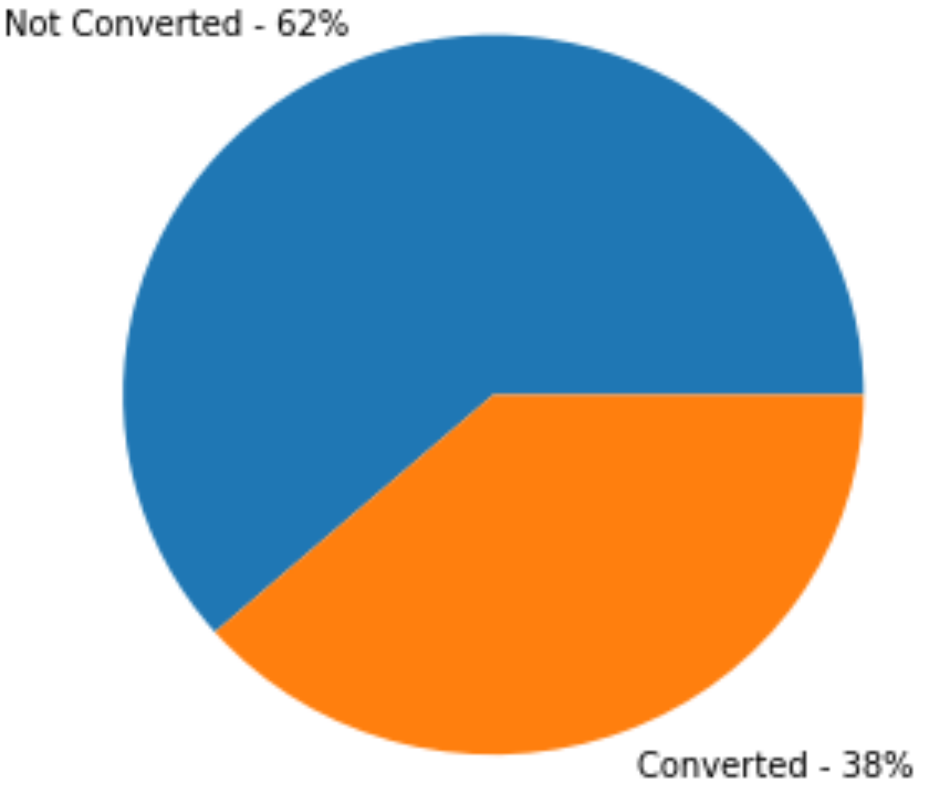
Leads Data Dictionary.xlsx

- The data dictionary which describes the meaning of the variables.

Reviewing the datasets

- Reviewed the data imports with functions like info, describe, columns and cleaned the leads data.
- Identify the columns that are irrelevant for analysis. For Eg: Last activity column is populated only when the lead is closed. This column will not be available when doing predictions and hot leads.

We have almost 38% conversion ratio and hence the data is good for modelling



2) Data Cleaning

Leads Data

Rows 9240

Columns **11**

Missing values treatment

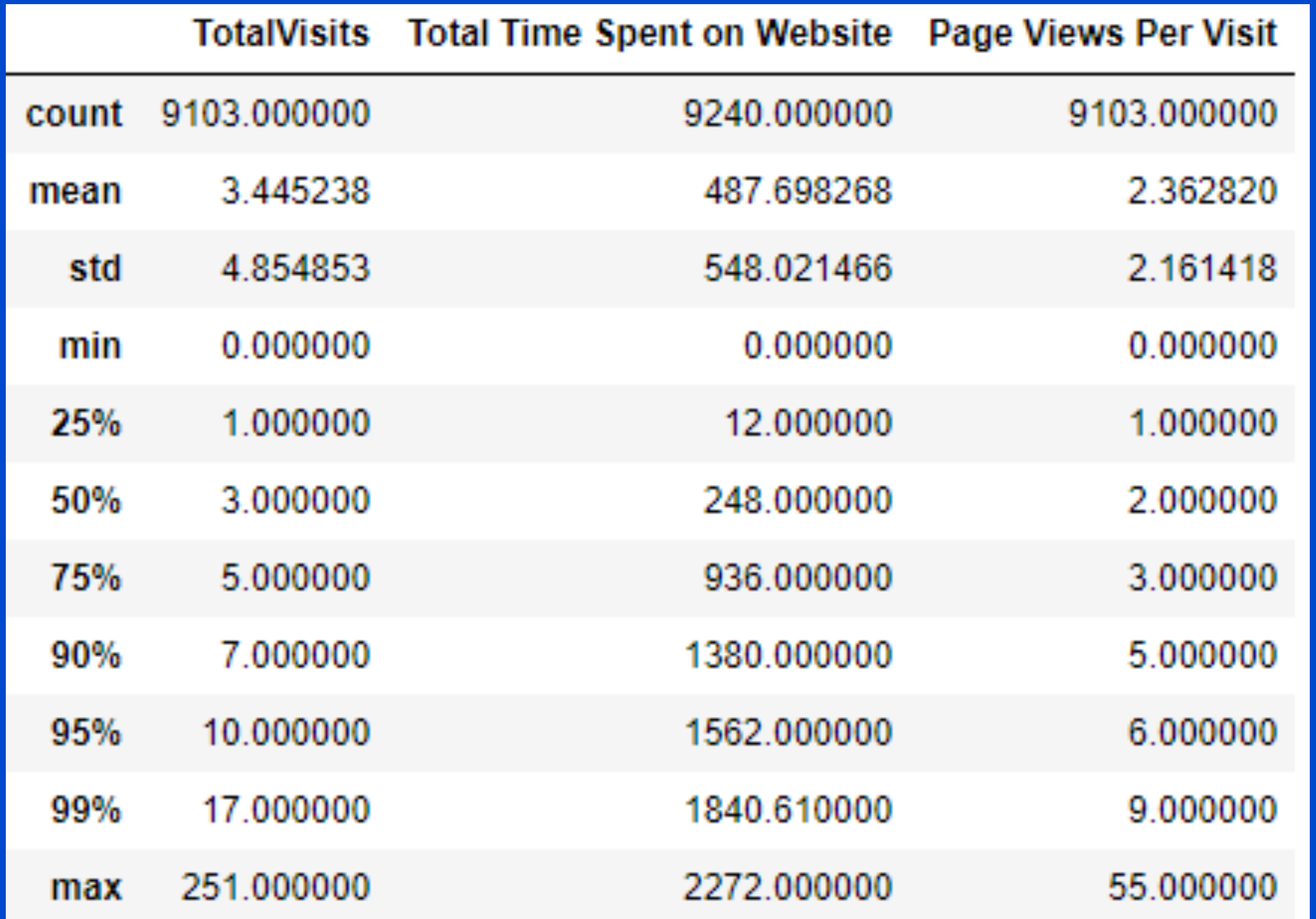
- Replaced "Select" values with Null
- Columns were dropped considering a threshold of 40% for blank values
- Numeric columns with missing values were populated with median after reviewing the quantiles

Categorical column treatment

- Converted the binary variables (Yes/No) to 0/1
- Dropped the columns that had all values as either *Yes or No*
- Dropped the columns that were heavily skewed towards one value.
- Merged all values with < 0.5% occurrence into one category called "Other"
- Merged values with < 0.1% occurrence for "Last Notable Activity" into "Other"
- Columns that are not relevant to the model were dropped. Eg: Prospect ID, Lead No and other score variables such as tags as well.

Numeric column treatment

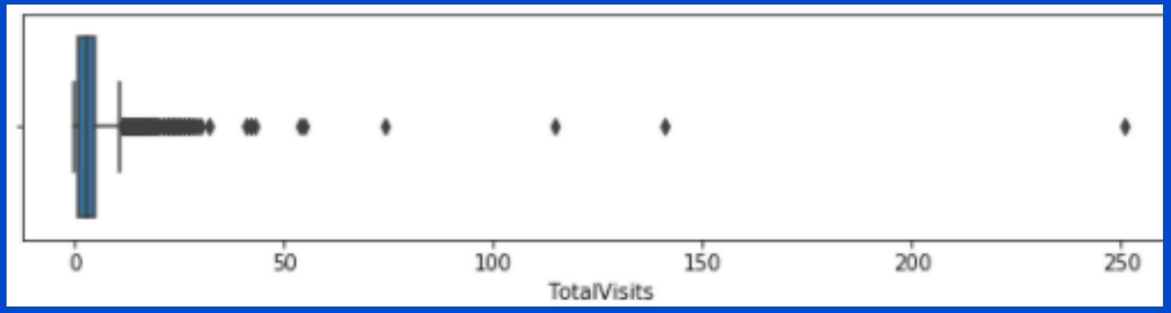
- Capping numerical variables at Higher range for 99 percentile value



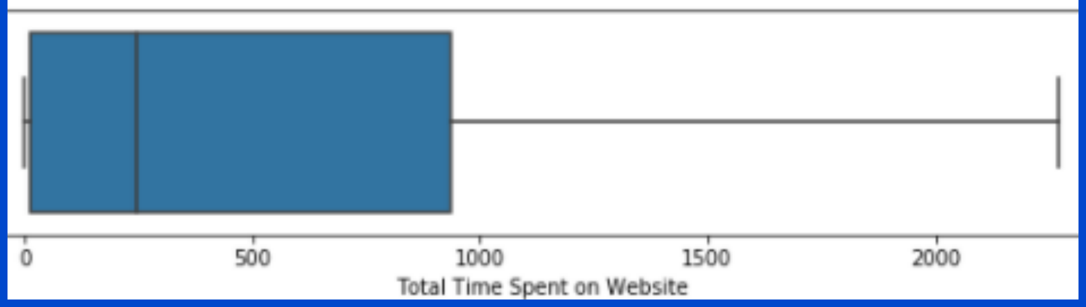
	TotalVisits	Total Time Spent on Website	Page Views Per Visit
count	9103.000000	9240.000000	9103.000000
mean	3.445238	487.698268	2.362820
std	4.854853	548.021466	2.161418
min	0.000000	0.000000	0.000000
25%	1.000000	12.000000	1.000000
50%	3.000000	248.000000	2.000000
75%	5.000000	936.000000	3.000000
90%	7.000000	1380.000000	5.000000
95%	10.000000	1562.000000	6.000000
99%	17.000000	1840.610000	9.000000
max	251.000000	2272.000000	55.000000

Outliers treatment

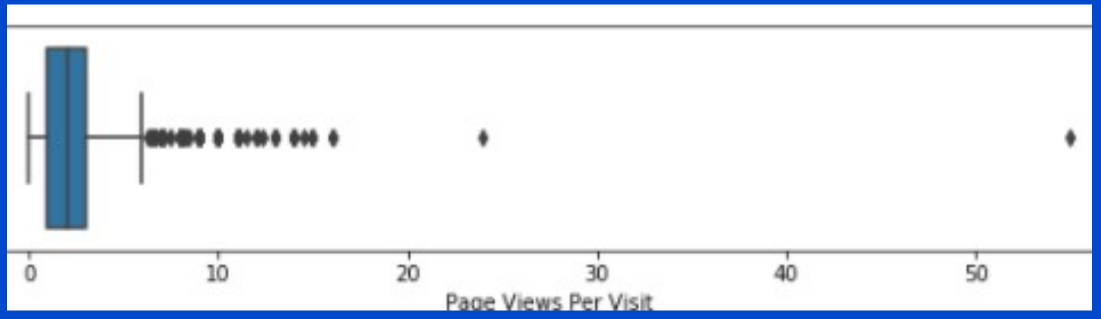
Total Visits



Total Time Spent on Website



Page Views per Visit



3) Data Preparation ⇉

	X_Test	X_Train
Rows	2772	6468
Columns	**48**	**48**

Dummy Variables

- Created dummy variables for *Lead Origin, Lead Source, Specialization, What is your current occupation* and *Last Notable Activity* columns

Test - Train Split

- Test-Train split was performed using the 70 to 30 ratios with a random state of 100.

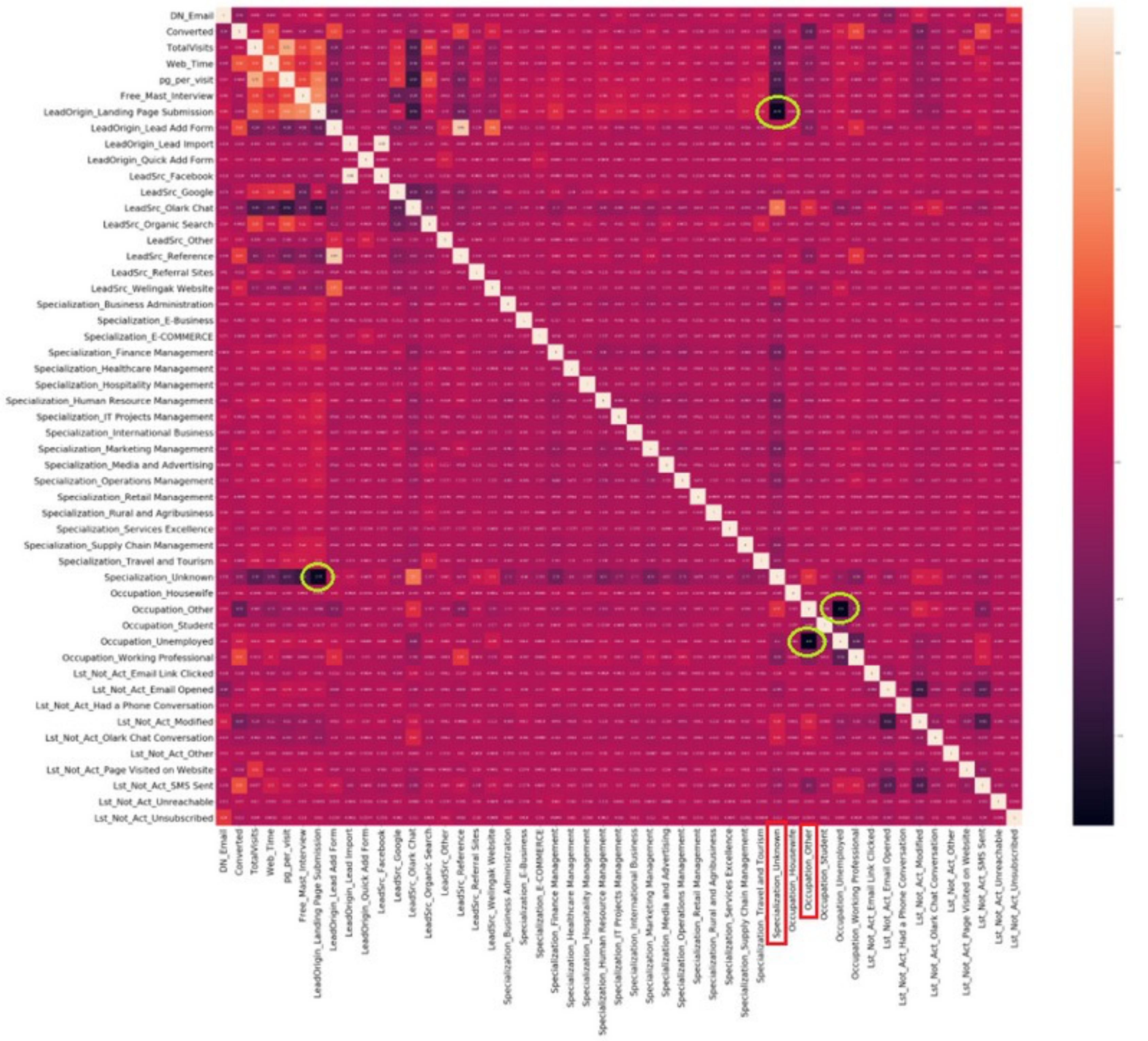
Feature Scaling

- Standardized scaling was used to soften some coefficients that are closer to the optimum than the others. Columns scaled are TotalVisits, Web_Time, pg_per_visit.

Correlation Matrix

- The Correlation Matrix indicates a very high correlation of 'Specialization_Unknown' with most of the numeric columns.
- It also shows a very high correlation of 'Occupation_Other' with the 'Occupation_Unemployed' columns.
- Dropping columns based on high co-relations and the fact that those were created by us for handling the null values.

Correlation Matrix



4) Model Building:

Inclutions as per RFE

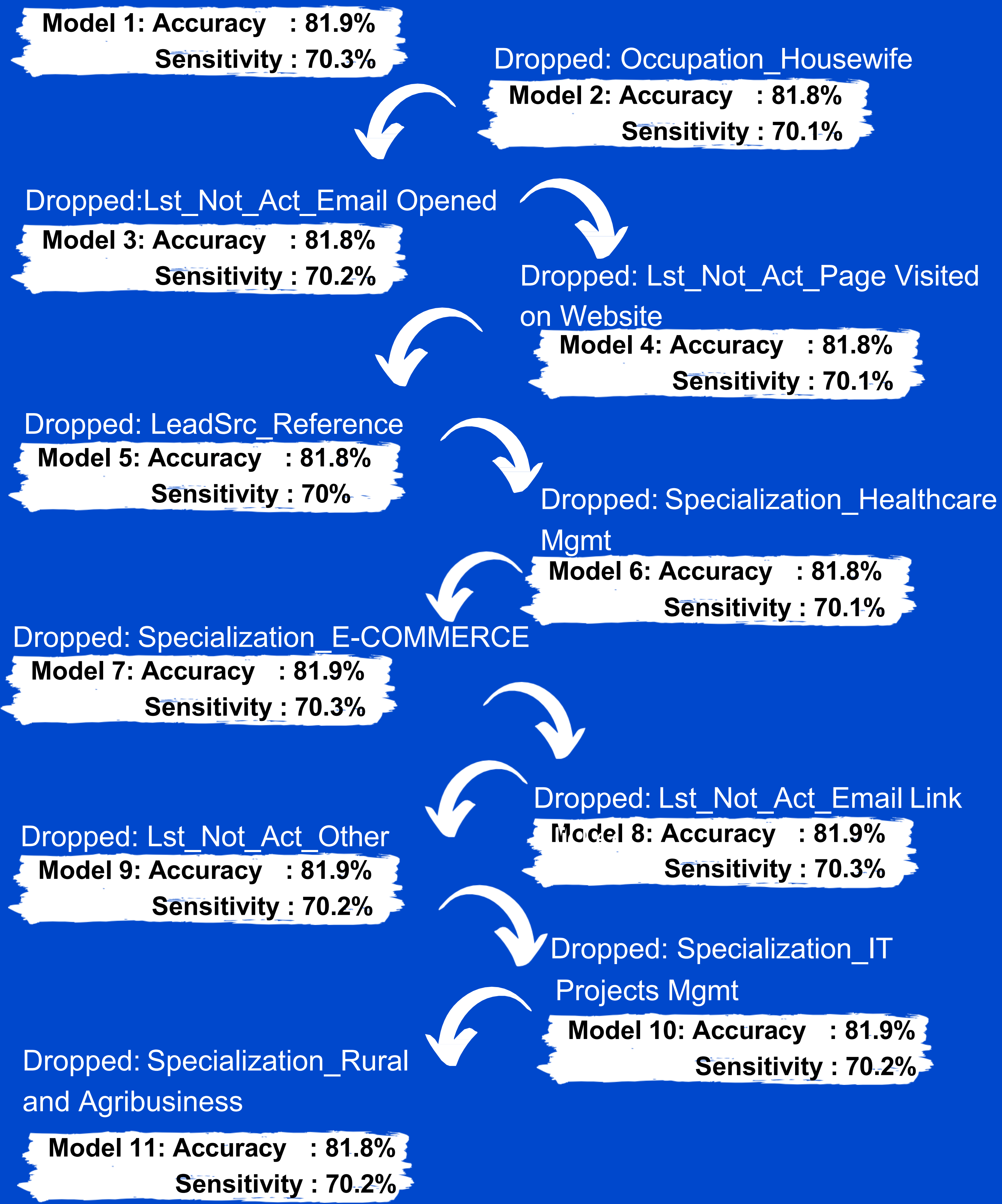
RFE used to derive top 25 columns for the first model

DN_Email Web_Time
LeadOrigin_Landing Page Submission
LeadOrigin_Lead Add Form
LeadSrc_Olark Chat
LeadSrc_Reference
LeadSrc_Welingak Website
Specialization_E-COMMERCE
Specialization_Finance Management
Specialization_Healthcare Management
Specialization_IT Projects Management
Specialization_Rural and Agribusiness
Occupation_Housewife
Occupation_Student
Occupation_Unemployed
Occupation_Working Professional
Lst_Not_Act_Email Link Clicked
Lst_Not_Act_Email Opened
Lst_Not_Act_Had a Phone Conversation
Lst_Not_Act_Modified
Lst_Not_Act_Olark Chat Conversation
Lst_Not_Act_Other
Lst_Not_Act_Page Visited on Website
Lst_Not_Act_SMS Sent
Lst_Not_Act_Unreachable

Column Elimination Criteria

- Logistical Model Summary Results were derived along with the VIF value
- The columns were dropped one by one based on the below criteria
  - P-Value should be under 0.05
  - VIF should be under 5
  - Accuracy, Sensitivity and Specificity should be more than 75%
- Finally, the 11th iteration model was finalized with 15 variables.
- All variables had a good value of VIF and P-Value. So, no more variables were dropped.

Manual selection with StatsModels (Dep. Variable :: Converted)



Model 1: Accuracy : 81.9%
Sensitivity : 70.3%

Dropped: Occupation_Housewife
Model 2: Accuracy : 81.8%
Sensitivity : 70.1%

Dropped:Lst_Not_Act_Email Opened
Model 3: Accuracy : 81.8%
Sensitivity : 70.2%

Dropped: Lst_Not_Act_Page Visited on Website
Model 4: Accuracy : 81.8%
Sensitivity : 70.1%

Dropped: LeadSrc_Reference
Model 5: Accuracy : 81.8%
Sensitivity : 70%

Dropped: Specialization_Healthcare Mgmt
Model 6: Accuracy : 81.8%
Sensitivity : 70.1%

Dropped: Specialization_E-COMMERCE
Model 7: Accuracy : 81.9%
Sensitivity : 70.3%

Dropped: Lst_Not_Act_Email Link
Model 8: Accuracy : 81.9%
Sensitivity : 70.3%

Dropped: Lst_Not_Act_Other
Model 9: Accuracy : 81.9%
Sensitivity : 70.2%

Dropped: Specialization_IT Projects Mgmt
Model 10: Accuracy : 81.9%
Sensitivity : 70.2%

Dropped: Specialization_Rural and Agribusiness
Model 11: Accuracy : 81.8%
Sensitivity : 70.2%

4) Model Building - Contd.

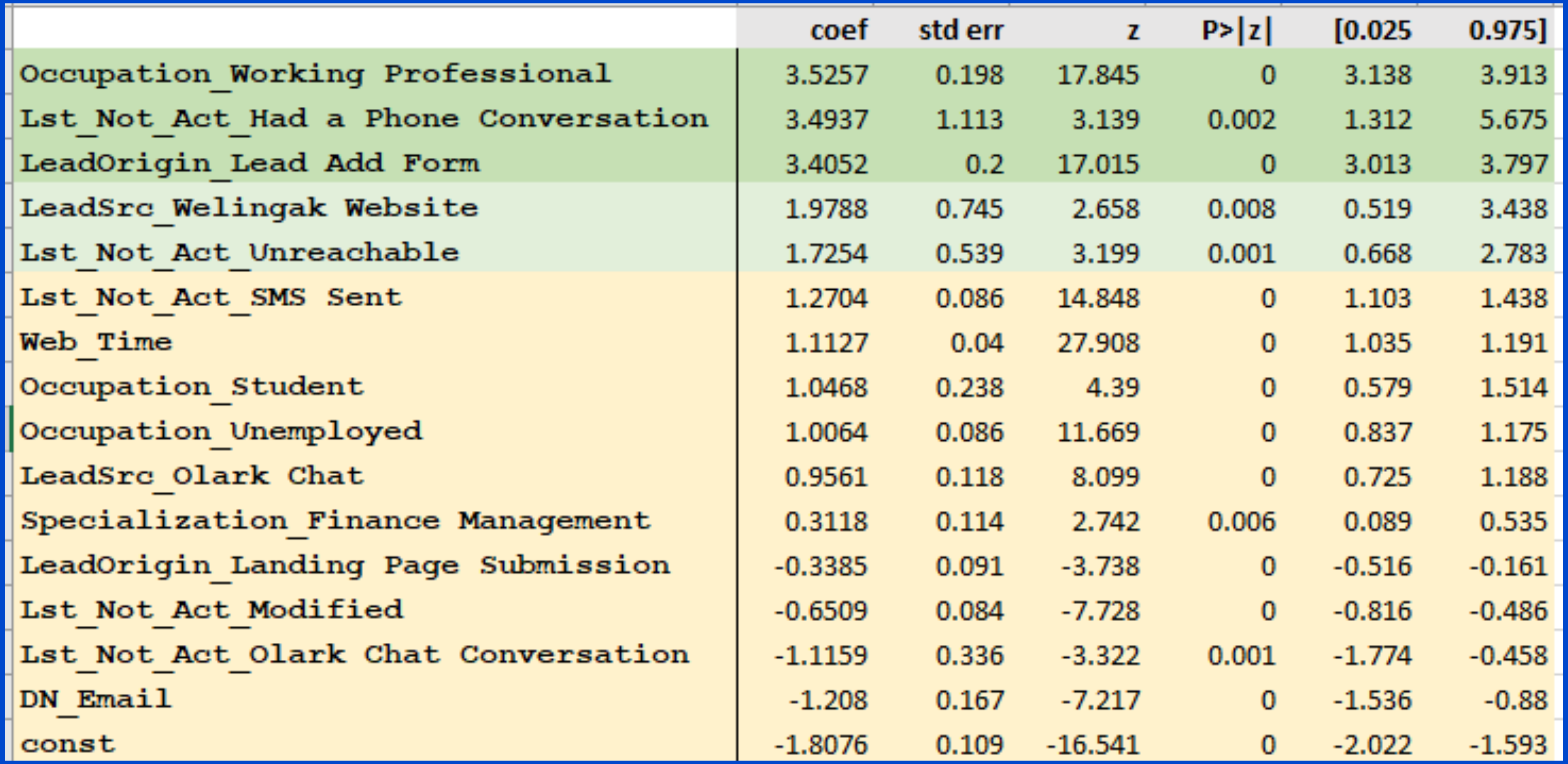
Model Summary - Technical Aspects

- All the P values are less than **0.05**
- All the VIF values are less than **5**
- The accuracy with a **0.5** probability cut off is around **82%**
- The Sensitivity with a **0.5** probability cut off is around **70%**
- The model has about **15** columns

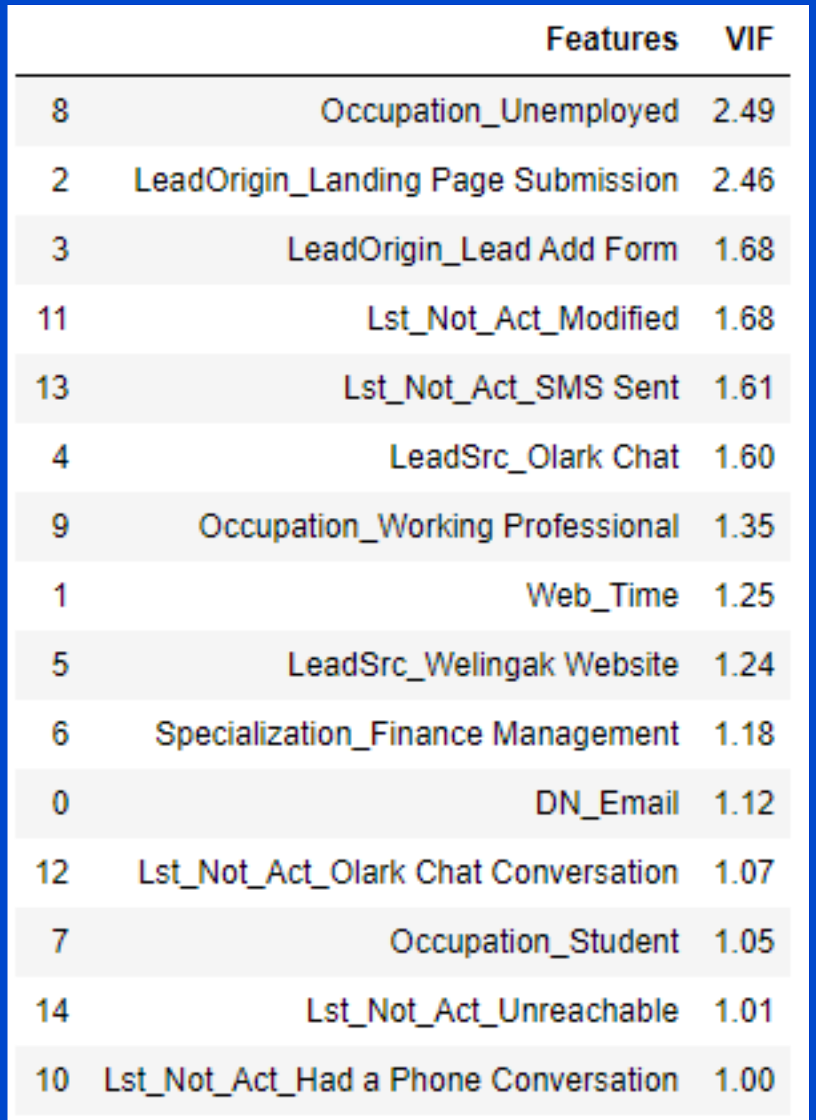
Model Summary - Business Aspects

We can now see that following are the top 4 factors which are significant factors with high coefficient and hence influence the probability of the conversion of the customers:

- "What is your current occupation" answered as "Working Professional"
- While the case is in progress, the "Last Notable Activity" was "Had a Phone Conversation" indicating the customer is worth calling again
- "Lead Origin" is through "Lead Add Form"
- "Lead Source" is through "Welingak Website"



	coef	std err	z	P>|z|	[0.025	0.975]
Occupation_Working Professional	3.5257	0.198	17.845	0	3.138	3.913
Lst_Not_Act_Had a Phone Conversation	3.4937	1.113	3.139	0.002	1.312	5.675
LeadOrigin_Lead Add Form	3.4052	0.2	17.015	0	3.013	3.797
LeadSrc_Welingak Website	1.9788	0.745	2.658	0.008	0.519	3.438
Lst_Not_Act_Unreachable	1.7254	0.539	3.199	0.001	0.668	2.783
Lst_Not_Act_SMS Sent	1.2704	0.086	14.848	0	1.103	1.438
Web_Time	1.1127	0.04	27.908	0	1.035	1.191
Occupation_Student	1.0468	0.238	4.39	0	0.579	1.514
Occupation_Unemployed	1.0064	0.086	11.669	0	0.837	1.175
LeadSrc_Olark Chat	0.9561	0.118	8.099	0	0.725	1.188
Specialization_Finance Management	0.3118	0.114	2.742	0.006	0.089	0.535
LeadOrigin_Landing Page Submission	-0.3385	0.091	-3.738	0	-0.516	-0.161
Lst_Not_Act_Modified	-0.6509	0.084	-7.728	0	-0.816	-0.486
Lst_Not_Act_Olark Chat Conversation	-1.1159	0.336	-3.322	0.001	-1.774	-0.458
DN_Email	-1.208	0.167	-7.217	0	-1.536	-0.88
const	-1.8076	0.109	-16.541	0	-2.022	-1.593

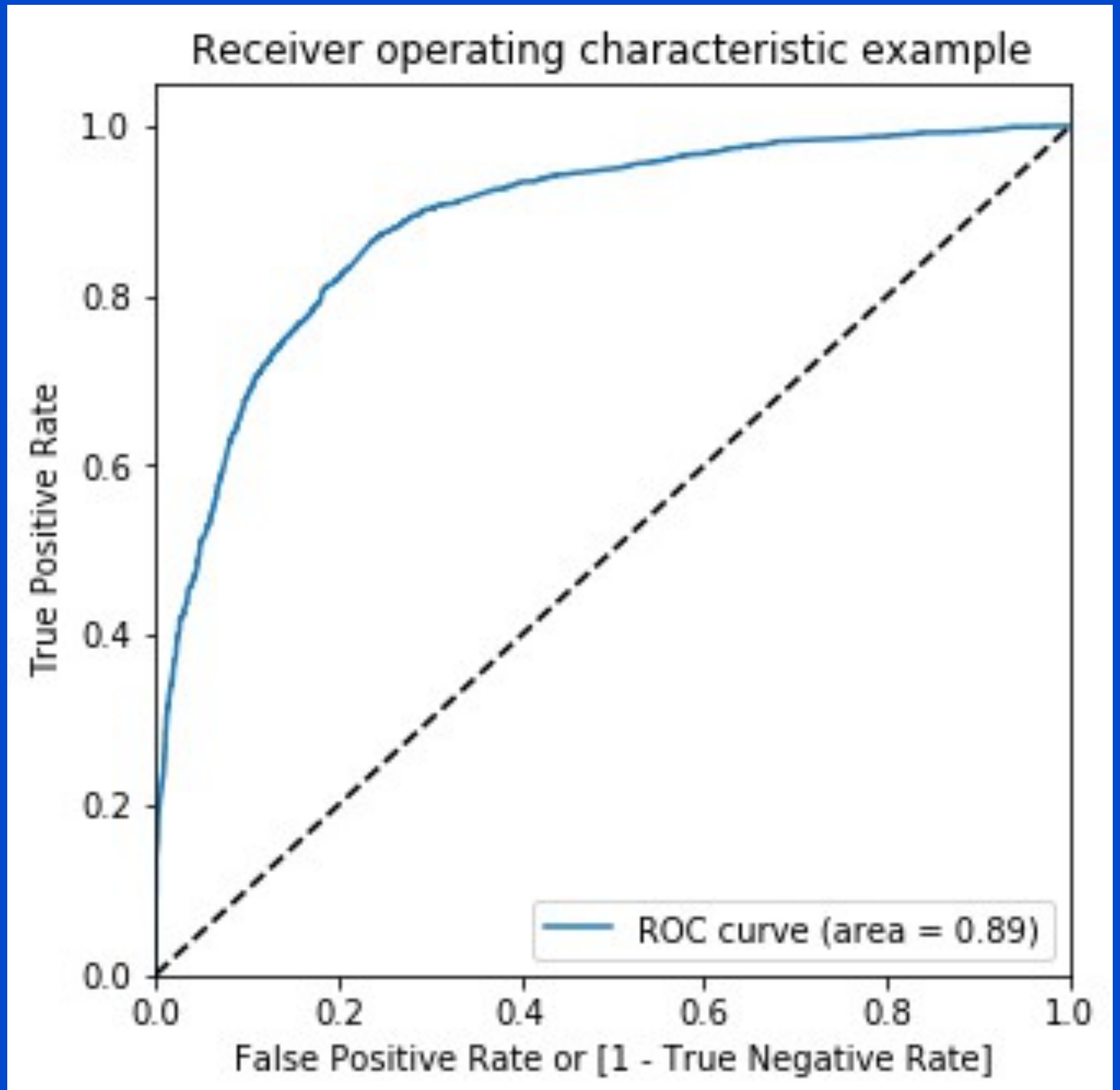


	Features	VIF
8	Occupation_Unemployed	2.49
2	LeadOrigin_Landing Page Submission	2.46
3	LeadOrigin_Lead Add Form	1.68
11	Lst_Not_Act_Modified	1.68
13	Lst_Not_Act_SMS Sent	1.61
4	LeadSrc_Olark Chat	1.60
9	Occupation_Working Professional	1.35
1	Web_Time	1.25
5	LeadSrc_Welingak Website	1.24
6	Specialization_Finance Management	1.18
0	DN_Email	1.12
12	Lst_Not_Act_Olark Chat Conversation	1.07
7	Occupation_Student	1.05
14	Lst_Not_Act_Unreachable	1.01
10	Lst_Not_Act_Had a Phone Conversation	1.00

5) Model Evaluation⇉

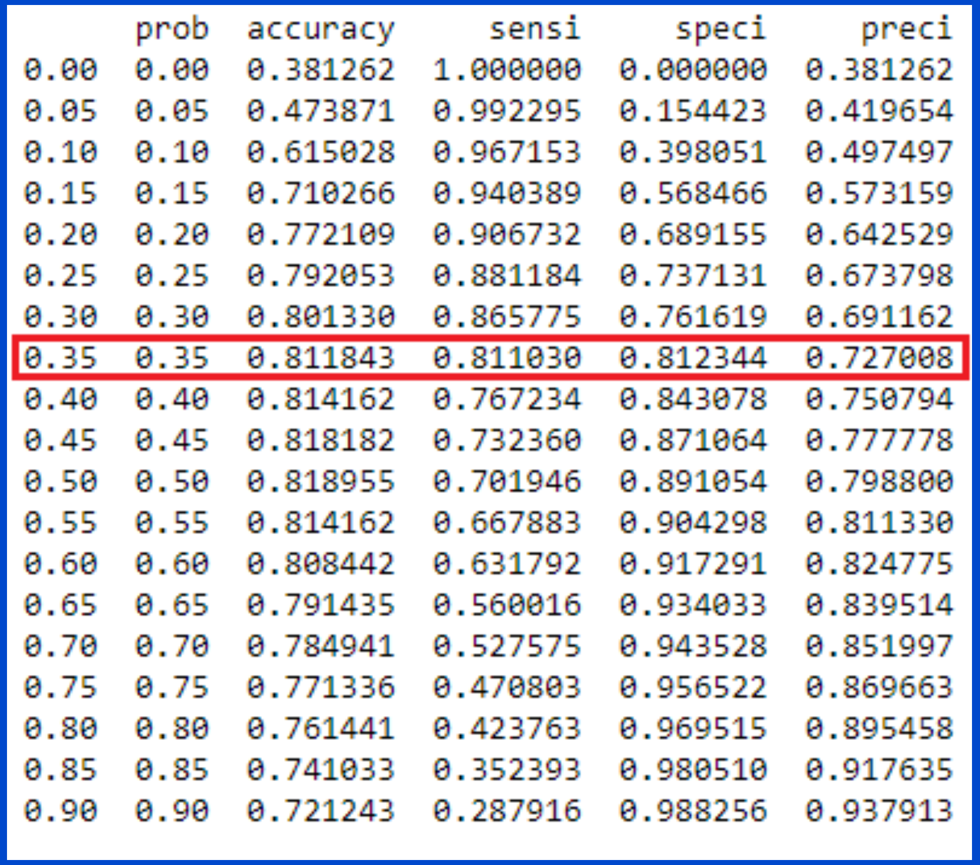
Plotting the ROC Curve

- The ROC Curve for this model is following the left-hand border and then the top border of the ROC space verty closely.
- This indicates a good accuracy for the test.
- The area under the curve is around 0.89 which is a good value to take this model forward for analysis.

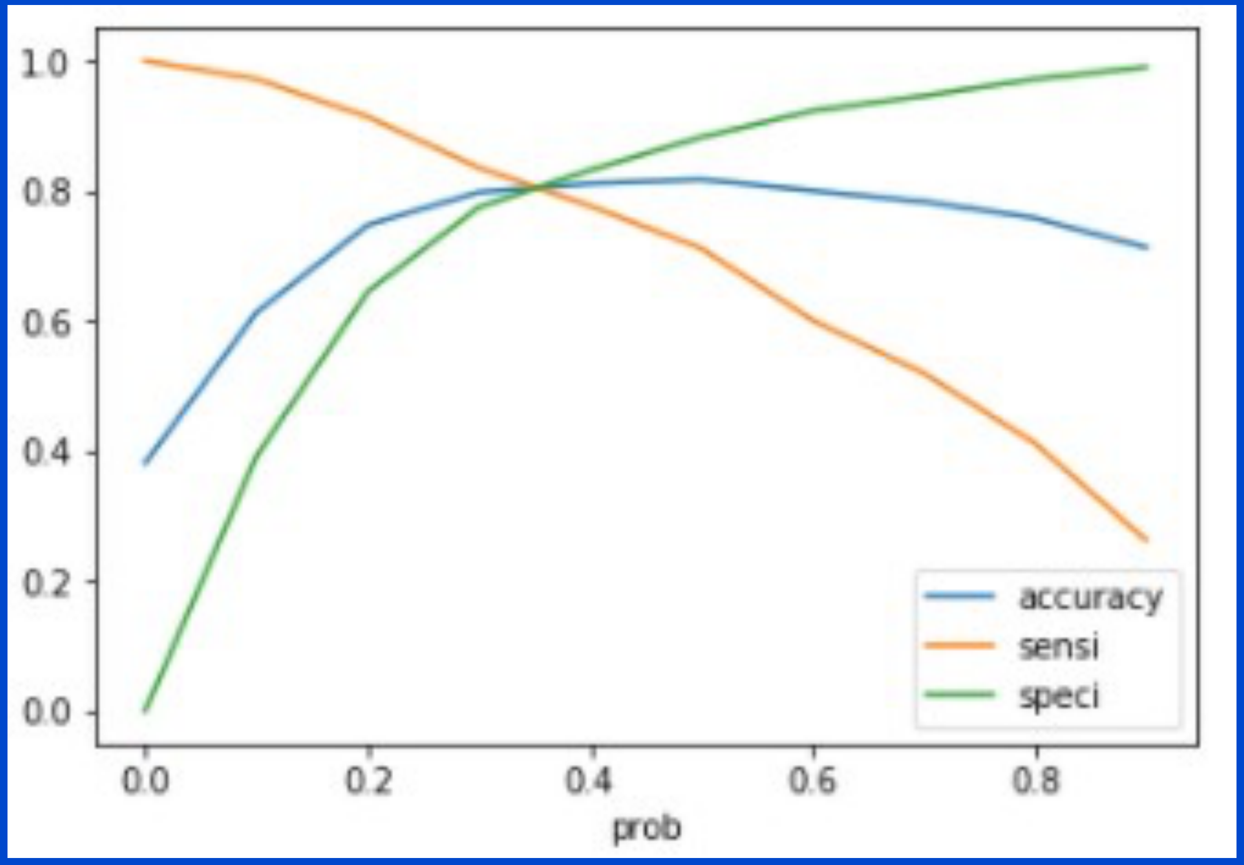


Finding Optimal Cutoff Point

- From the curve and the table displayed above, 0.35 is the optimum point to take it as a cutoff probability.

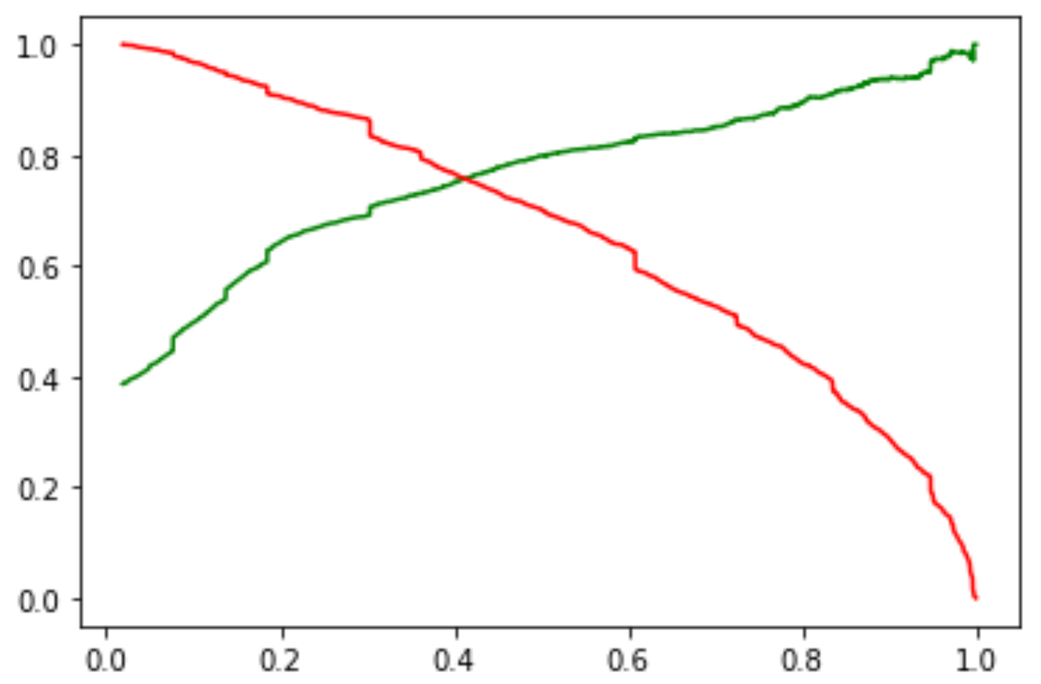


```
      prob  accuracy     sensi     speci     preci
0.00  0.00  0.381262  1.000000  0.000000  0.381262
0.05  0.05  0.473871  0.992295  0.154423  0.419654
0.10  0.10  0.615028  0.967153  0.398051  0.497497
0.15  0.15  0.710266  0.940389  0.568466  0.573159
0.20  0.20  0.772109  0.906732  0.689155  0.642529
0.25  0.25  0.792053  0.881184  0.737131  0.673798
0.30  0.30  0.801330  0.865775  0.761619  0.691162
0.35  0.35  0.811843  0.811030  0.812344  0.727008
0.40  0.40  0.814162  0.767234  0.843078  0.750794
0.45  0.45  0.818182  0.732360  0.871064  0.777778
0.50  0.50  0.818955  0.701946  0.891054  0.798800
0.55  0.55  0.814162  0.667883  0.904298  0.811330
0.60  0.60  0.808442  0.631792  0.917291  0.824775
0.65  0.65  0.791435  0.560016  0.934033  0.839514
0.70  0.70  0.784941  0.527575  0.943528  0.851997
0.75  0.75  0.771336  0.470803  0.956522  0.869663
0.80  0.80  0.761441  0.423763  0.969515  0.895458
0.85  0.85  0.741033  0.352393  0.980510  0.917635
0.90  0.90  0.721243  0.287916  0.988256  0.937913
```



Precision and Recall

- The Cut-off is recommended to be at 0.4 based on the precision and recall trade off.
- However the recall value (also known as Sensitivity) will be less than 80% at this value which does not meet the business requirements.
- Hence we will use the cut-off as 0.35 based on the Sensitivity and Specificity curve



6) Metrics on Model

Metrics on test set:

Sensitivity	: 81%
Specificity	: 81.87%
False Positive Rate	: 18.13%
Positive Predictive Rate	: 74.47%
Negative Predictive Rate	: 86.84%

Metrics on train set:

Sensitivity	: 81.10%
Specificity	: 81.23%
False Positive Rate	: 18.76%
Positive Predictive Rate	: 72.70%
Negative Predictive Rate	: 87.46%

Conclusion:

As per the above results with the probability cut off on 0.35, the test and train data set parameters appear to be almost the same without much deviations. Hence the model is good for predictions.

7) Recommendations

Our final model has about 15 columns and most of these are from the 3 categories mentioned below.

- “What is your current occupation”
- “Last Notable Activity”
- “Lead Origin”

Hence, these are the primary columns which contribute towards the lead getting converted.

Columns with high values of coefficient which mainly contribute towards the lead getting converted:

1) “Occupation_Working Professional”
2) “Lst_Not_Act_Had a Phone Conversation”
3) “LeadOrigin_Lead Add Form”

The above variables are derived as dummy variables from the following set of original categorical columns.

1) “What is your current occupation” : Answered as “Working Professional”
2) "Last Notable Activity” : registered as “Had a Phone Conversation”
3) “Lead Origin: : identified as “Lead Add Form”

The lead score has been calculated by multiplying the probability with 100.

Sales team should consider any lead with lead score greater than 35.