Lead Scoring Case Study

Problem Statement

An education company named X Education sells online courses to industry professionals. Although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. X Education needs help to select the most promising leads, i.e., the leads that are most likely to convert into paying customers. The company requires to build a model wherein we 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%.

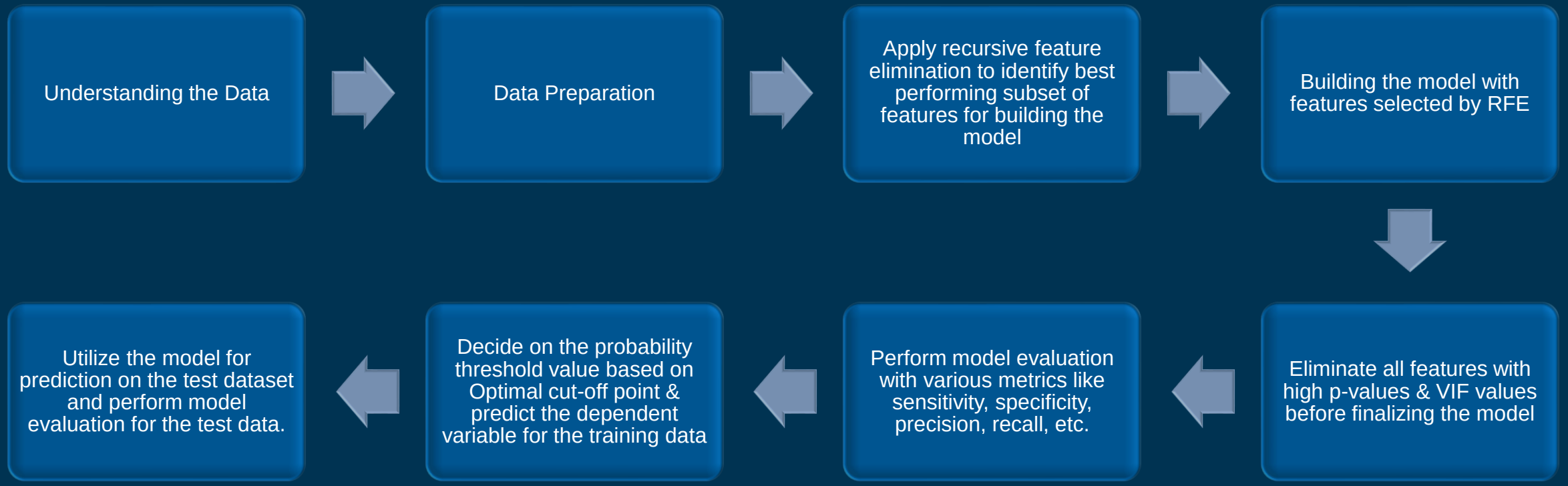
Goals of the Case Study

1. Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.
2. There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

Solution Approach

- The required Python libraries to perform data analysis/visualization are imported.
- Import the dataset and perform initial validation checks – summary, statistics, data types, etc.
- Perform analysis to identify variables with null values – threshold limit for this case study has been considered as 40%.
- Columns with null values percentages greater than 40% are excluded from the analysis.
- Validated remaining columns within acceptable limit for null values, for e.g. NaN values replaced with mean/mode values as applicable.
- Outlier analysis performed by leveraging box plot to check for outlier values.
- Univariate/Bivariate analysis performed for categorical and numerical variables.
- Train Test Split.
- Feature Scaling and Dummy Variables.
- Logistic Regression Model Building and Prediction
- ROC Curve
- Model Evaluation
- Precision and Recall
- Conclusion

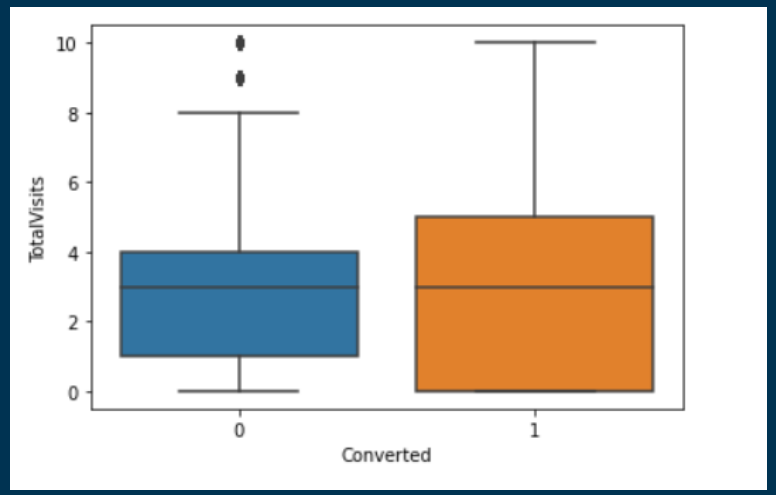
Problem Solving Methodology



Data Preparation

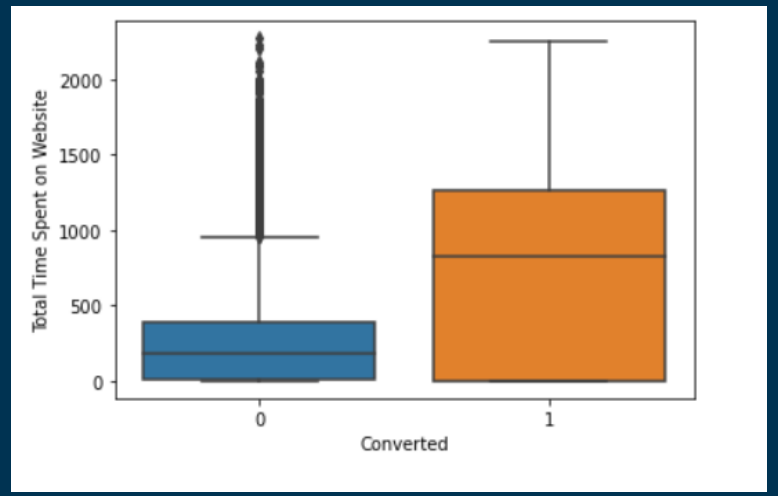
Removing rows with high number of missing values	•Rows with particular columns having high number of missing values were dropped - 'How did you hear about X Education', 'Lead Profile', 'Lead Quality', 'Asymmetrique Profile Score'
Imputing null values with mean/mode values as applicable	•Columns with null values were imputed to avoid negative impact on overall dataset - 'Country', 'Specialization'
Convert 'Select' values to NaN	•Users may not have selected any option from list hence it picked 'Select' as the default value which needs to be converted. Convert 'Select' values to NaN.
Outlier Treatment	•The outliers present in the columns 'TotalVisits' & 'Page Views Per Visit' were removed based on analysis performed.
Binary Encoding	•Convert some of the binary variables (Yes/No) to 0/1 - 'Do Not Email' & 'Do Not Call'
Dummy Encoding	•Create dummy variable for the categorical variables - 'Lead Origin', 'Lead Source', 'Last Activity', 'Specialization', 'What is your current occupation', 'City', 'Last Notable Activity'
Test-Train Split	•The original dataframe was split into train & test dataset. Train dataset was used to train the model & test dataset was used to evaluate the model.
Feature Scaling	•Scaling helps in interpretation & its important to have all variables on the same scale for the model to be easily interpretable.

EDA: Numerical Data



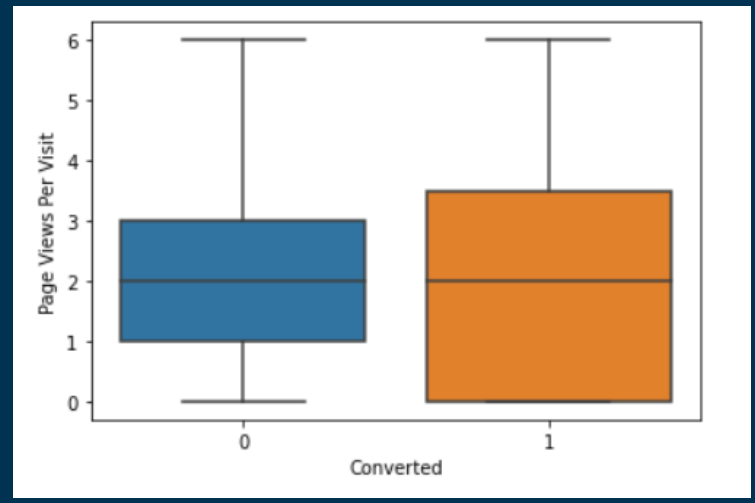
Total Visits

The median for both converted and non-converted leads are the same



Total Time Spent on Website

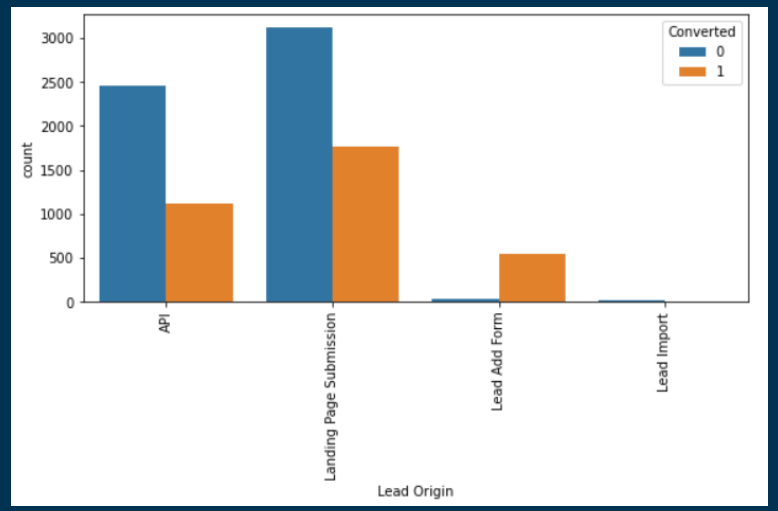
Leads spending more time on the website are more likely to be converted



Page Views Per Visit

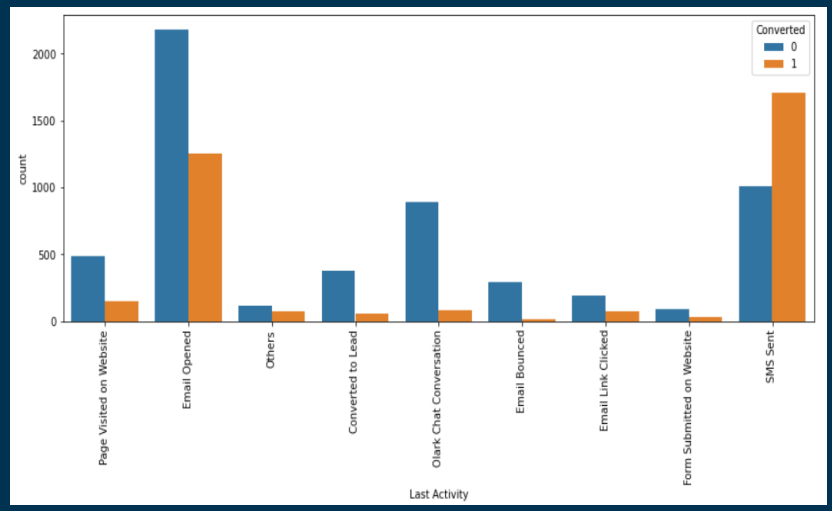
The median for both converted and non-converted leads are the same

EDA: Categorical Data



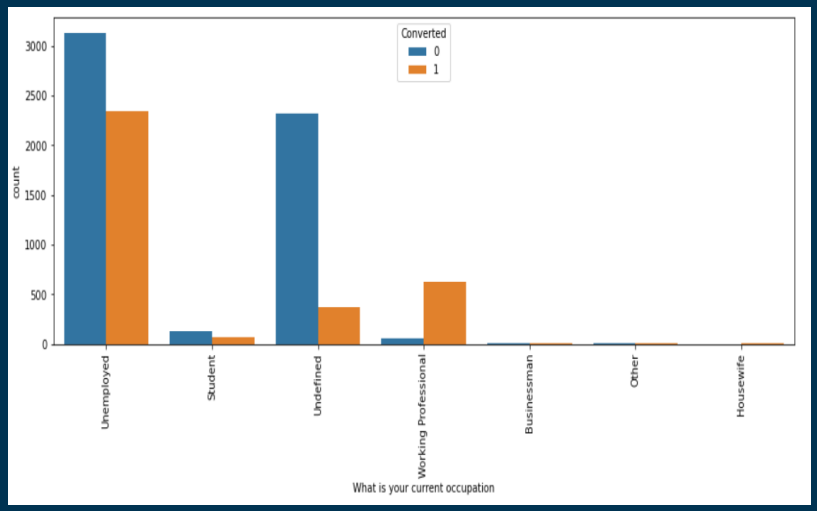
Lead Origin

The focus needs to be on overall improvement of lead conversion rate



Last Activity

Maximum leads generated have 'Email Opened' as their last activity though the conversion rate is not good



What is your current occupation

Conversion rate for 'Working Professional' are quite high though maximum leads are generated for 'Unemployed'

Feature Selection using RFE

Running RFE with 15 variables as output

```
logreg = LogisticRegression()

# Running RFE with 15 variables as output

rfe = RFE(logreg, 15)
rfe = rfe.fit(X_train, y_train)
```

```
# List of columns selected by RFE

cols = X_train.columns[rfe.support_]
cols
```

```
Index(['Do Not Email', 'Total Time Spent on Website',
       'Lead Origin_Landing Page Submission', 'Lead Origin_Lead Add Form',
       'Lead Source_Olark Chat', 'Lead Source_Welingak Website',
       'Last Activity_Others', 'Last Activity_SMS Sent',
       'Specialization_Undefined', 'What is your current occupation_Housewife',
       'What is your current occupation_Undefined',
       'What is your current occupation_Working Professional',
       'Last Notable Activity_Had a Phone Conversation',
       'Last Notable Activity_Modified',
       'Last Notable Activity_Olark Chat Conversation'],
      dtype='object')
```

Building the Model

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.0191	0.127	0.151	0.880	-0.230	0.268
Do Not Email	-1.6609	0.184	-9.043	0.000	-2.021	-1.301
Total Time Spent on Website	1.1167	0.041	26.957	0.000	1.036	1.198
Lead Origin_Landing Page Submission	-1.0257	0.129	-7.978	0.000	-1.278	-0.774
Lead Origin_Lead Add Form	3.1149	0.235	13.239	0.000	2.654	3.576
Lead Source_Olark Chat	1.1455	0.124	9.256	0.000	0.903	1.388
Lead Source_Welingak Website	2.4917	0.756	3.294	0.001	1.009	3.974
Last Activity_Others	1.2780	0.225	5.685	0.000	0.837	1.719
Last Activity_SMS Sent	1.3212	0.077	17.200	0.000	1.171	1.472
Specialization_Undefined	-0.9729	0.126	-7.704	0.000	-1.220	-0.725
What is your current occupation_Undefined	-1.1404	0.090	-12.691	0.000	-1.317	-0.964
What is your current occupation_Working Professional	2.3885	0.195	12.271	0.000	2.007	2.770
Last Notable Activity_Modified	-0.9810	0.081	-12.176	0.000	-1.139	-0.823
Last Notable Activity_Olark Chat Conversation	-1.1825	0.336	-3.517	0.000	-1.842	-0.523

	Features	VIF
8	Specialization_Undefined	2.33
4	Lead Source_Olark Chat	1.93
2	Lead Origin_Landing Page Submission	1.89
11	Last Notable Activity_Modified	1.69
9	What is your current occupation_Undefined	1.63
3	Lead Origin_Lead Add Form	1.61
7	Last Activity_SMS Sent	1.56
5	Lead Source_Welingak Website	1.37
1	Total Time Spent on Website	1.30
10	What is your current occupation_Working Profes...	1.19
0	Do Not Email	1.15
12	Last Notable Activity_Olark Chat Conversation	1.09
6	Last Activity_Others	1.05

Generalized Linear Regression Model from statsmodels is utilized to build the Logistic Regression Model. The model is built initially with 15 variables selected by RFE. Unwanted features are dropped sequentially after checking p-values & VIF followed by building the model multiple times. The final model passes both significance & the multi-collinearity test.

Conversion Probability & Predicted Column

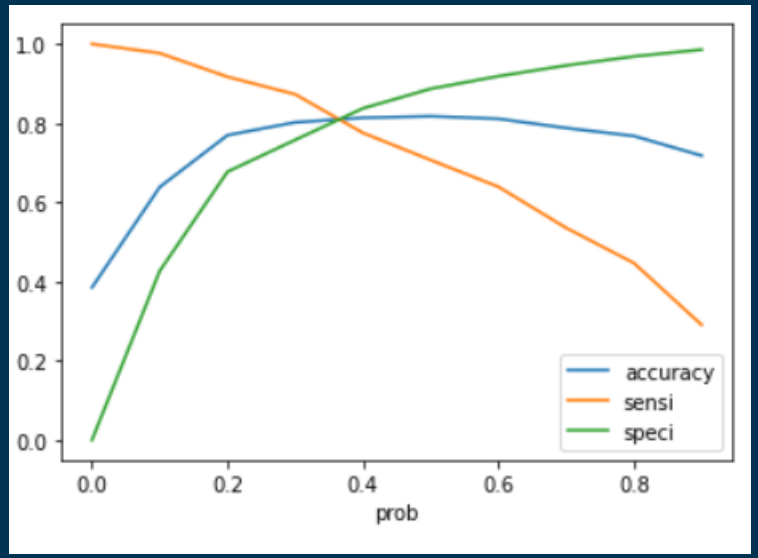
Top 5 records with actual Converted flag & predicted probabilities

	Converted	Converted_prob	Prospect ID
0	0	0.089000	3009
1	0	0.135292	1012
2	0	0.386896	9226
3	1	0.732897	4750
4	1	0.830160	7987

Column 'predicted' created with 1 Converted_prob > 0.5 else 0

	Converted	Converted_prob	Prospect ID	predicted
0	0	0.089000	3009	0
1	0	0.135292	1012	0
2	0	0.386896	9226	0
3	1	0.732897	4750	1
4	1	0.830160	7987	1

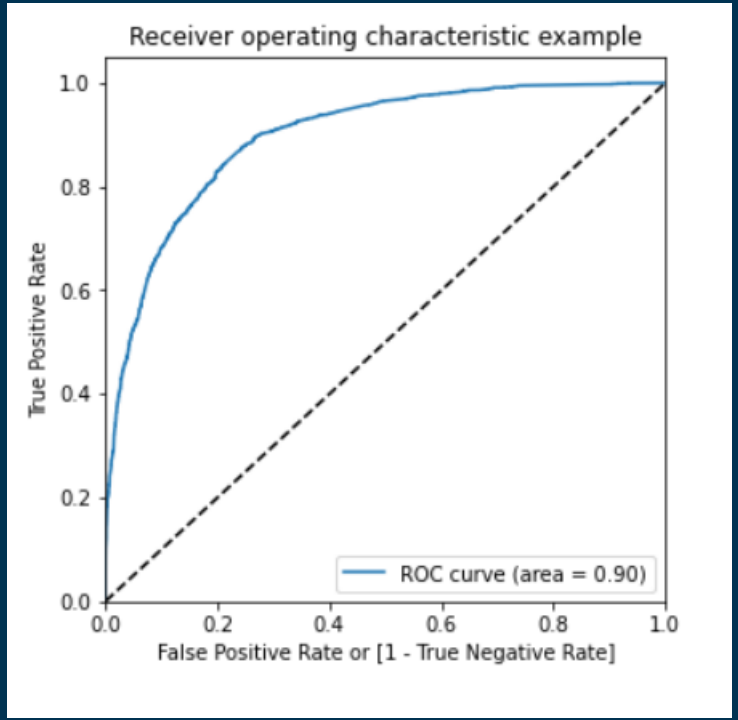
Finding Optimal Probability Threshold



Accuracy, Sensitivity & Specificity were calculated for various values of probability threshold & plotted to the above graph. From the above curve, 0.35 is found to be the optimum point for cutoff probability.

At this threshold, all 3 metrics were found to be above 80% which appears to be expected value.

Plotting the ROC Curve



It shows the trade-off between sensitivity and specificity (any increase in sensitivity will be accompanied by decrease in specificity)

Model Performance

Train Set	
Accuracy	81%
Sensitivity	82%
Specificity	80%

Test Set	
Accuracy	80%
Sensitivity	80%
Specificity	81%

The sensitivity value post model building process is greater than 80% which is expected value.
Post execution of the model on test data, the parameters remain closer to respective values calculated using trained set. Hence, the overall model seems to be good and appears highly stable.

Lead Score Calculation

	Prospect ID	Converted	Converted_prob	final_predicted	Lead_Score
0	3271	0	0.059255	0	6
1	1490	1	0.970388	1	97
2	7936	0	0.050456	0	5
3	4216	1	0.761459	1	76
4	3830	0	0.057147	0	6

- Lead score is calculated for all the leads in the dataframe.
- The Conversion Probability is multiplied by 100 to obtain the lead score for each lead.
- Higher the lead score, higher is the probability of a lead getting converted & vice-versa.
- Formula for lead score calculation: *(Lead Score = 100 * Conversion Probability)*

Conclusion

After trying out several models, our final model has following characteristics:

- All p-values are very close to zero.
- VIFs for all features are very low, there is hardly any multicollinearity present.
- Accuracy, Sensitivity and Specificity values of test set are around 80%, 80% and 81% which are approximately closer to the respective values calculated using trained set.
- We have checked both Sensitivity-Specificity as well as Precision and Recall Metrics, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction.
- Hence, the overall model seems to be good.

Recommendation

- Collect data often and run the model to get updated with potential leads. The best time to contact the potential leads is within few hours after the lead shows interest in the courses.
- Along with phone calls, its good practice to send email communication to the potential leads to provide specific information.
- Avoiding phone calls and leveraging other mediums like google advertisements or mailers would help to keep in touch with leads. This would be efficient option and help save lot of time.
- Focusing on hot leads will increase the chances of obtaining more value to the business as the number of people we contact are less but the conversion rate is higher.

Thank You