LEAD SCORING CASE-STUDY

PRESENTED BY –

RACHAITA MAITI
& DEEPTI SISODIYA

Problem Statement:

- X Education company sells online courses to industry professionals
- 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.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'.
- If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

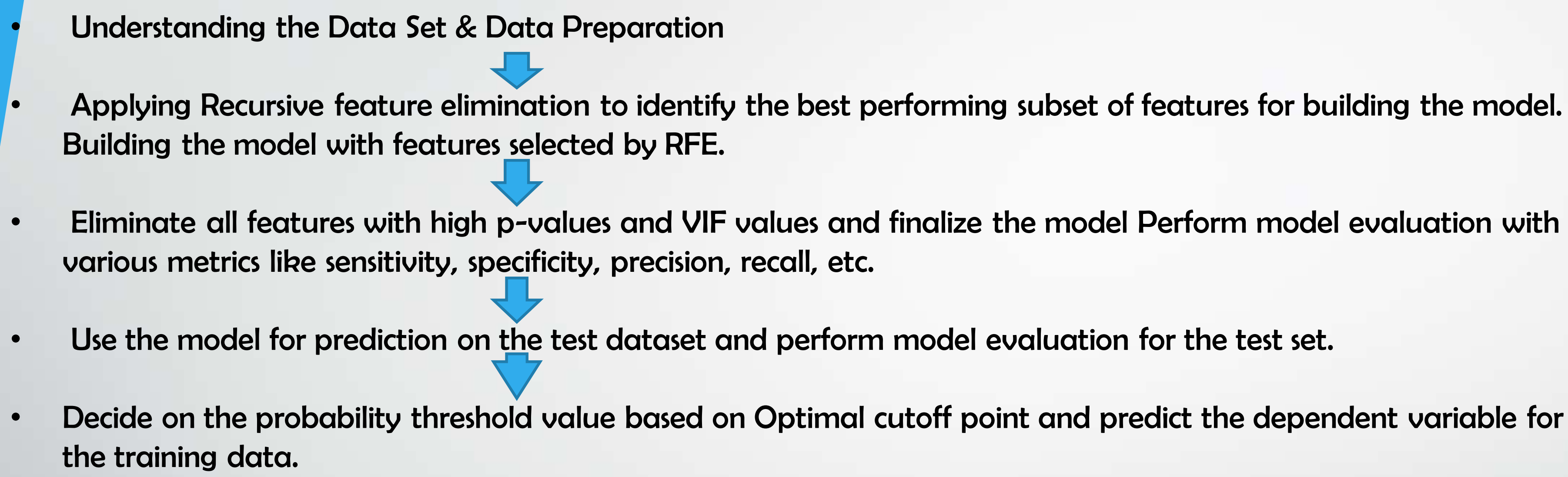
Business Objective:

- X education wants to know most promising leads.
- For that they want to build a Model which identifies the hot leads.
- Propagation of the model for the future use.

Solution Procedure :

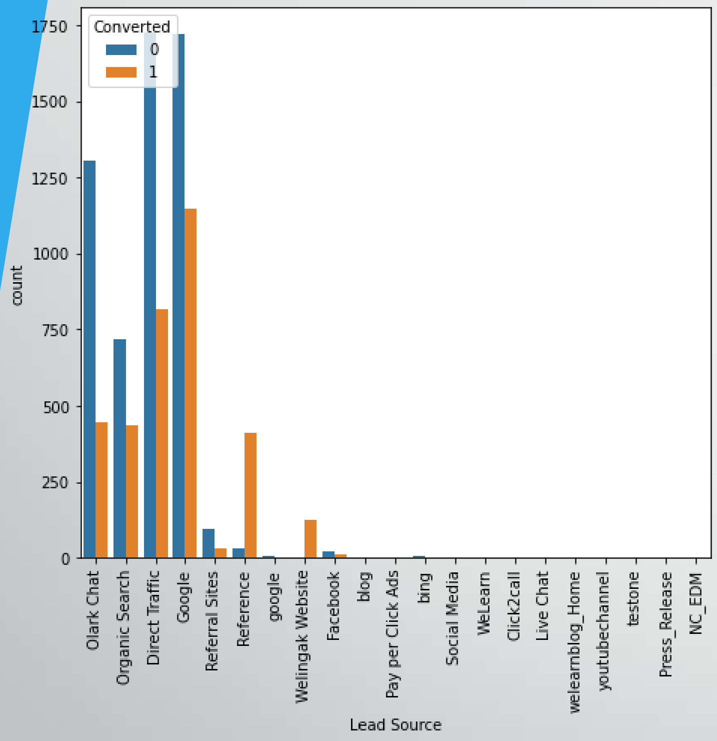
- Source the data for analysis.
- Inspecting Data and Data cleaning .
- Exploratory data analysis(EDA) for inspecting the attributes for progress.
- Feature Dummy Variables and encoding the data.
- Split dataset into Test and Train.
- Classification technique: logistic regression used for the model building and prediction.
- Evaluate the model by different measures and metrics.
- Making predictions on test set.
- Conclusions .

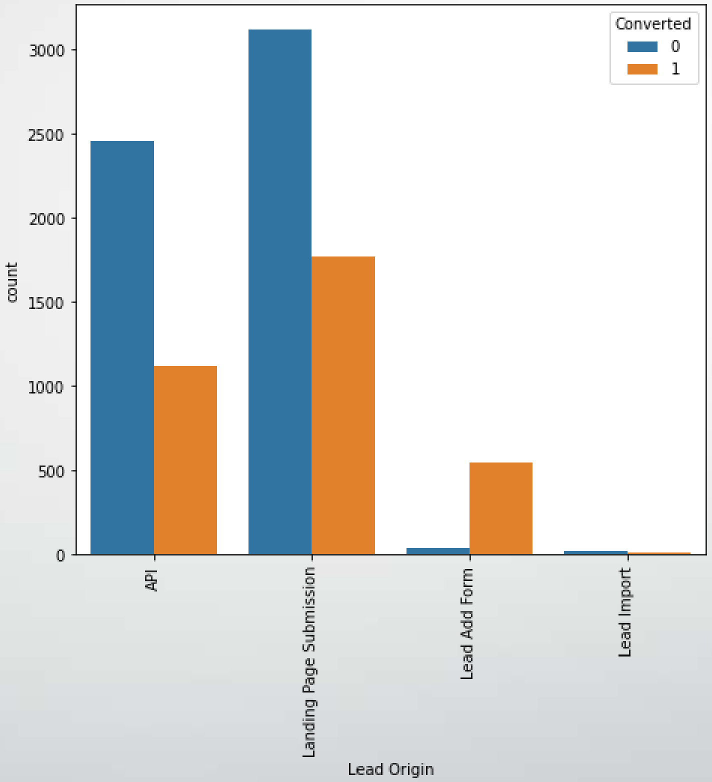
Problem Solving Methodology:



- Understanding the Data Set & Data Preparation
- Applying Recursive feature elimination to identify the best performing subset of features for building the model. Building the model with features selected by RFE.
- Eliminate all features with high p-values and VIF values and finalize the model Perform model evaluation with various metrics like sensitivity, specificity, precision, recall, etc.
- Use the model for prediction on the test dataset and perform model evaluation for the test set.
- Decide on the probability threshold value based on Optimal cutoff point and predict the dependent variable for the training data.

EDA:

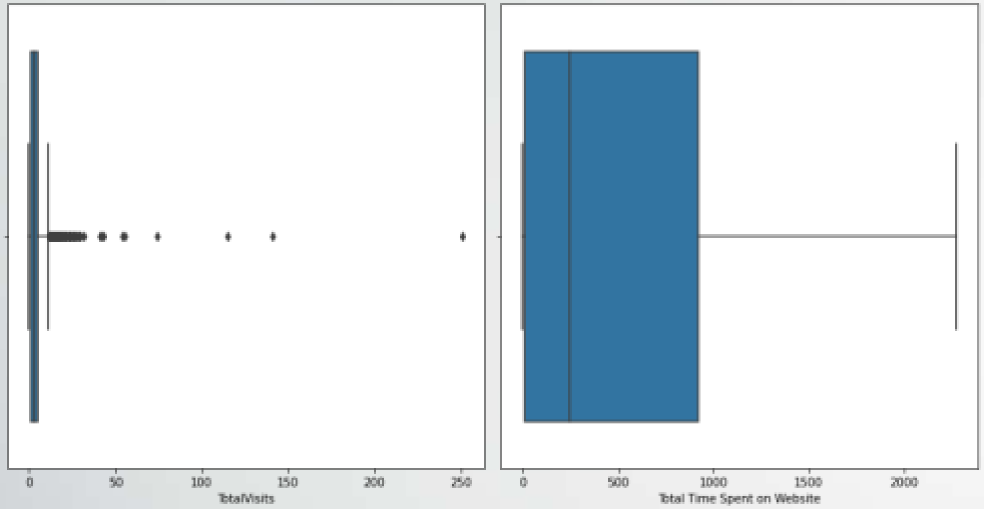


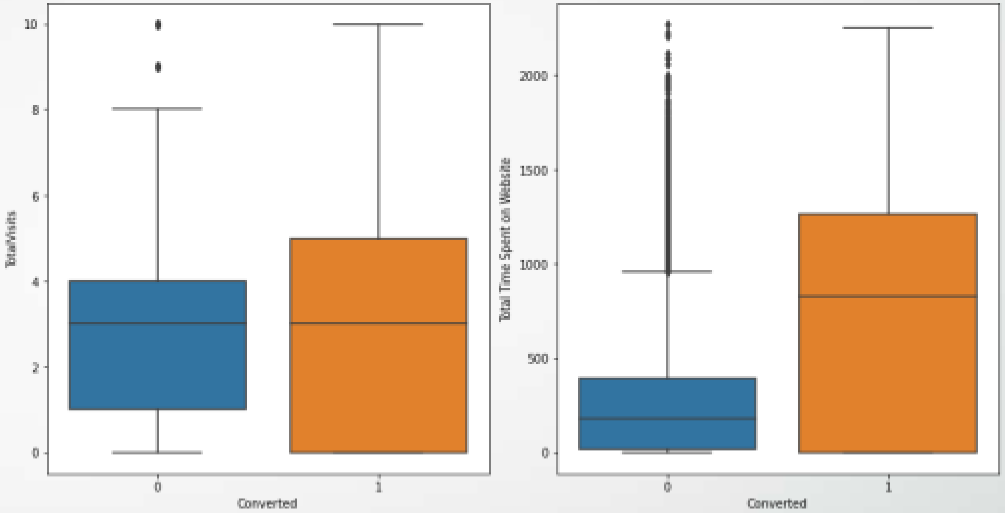


- API and Landing Page Submission has less conversion rate(~30%) but counts of the leads from them are considerable

- The count of leads from the Lead Add Form is pretty low but the conversion rate is very high

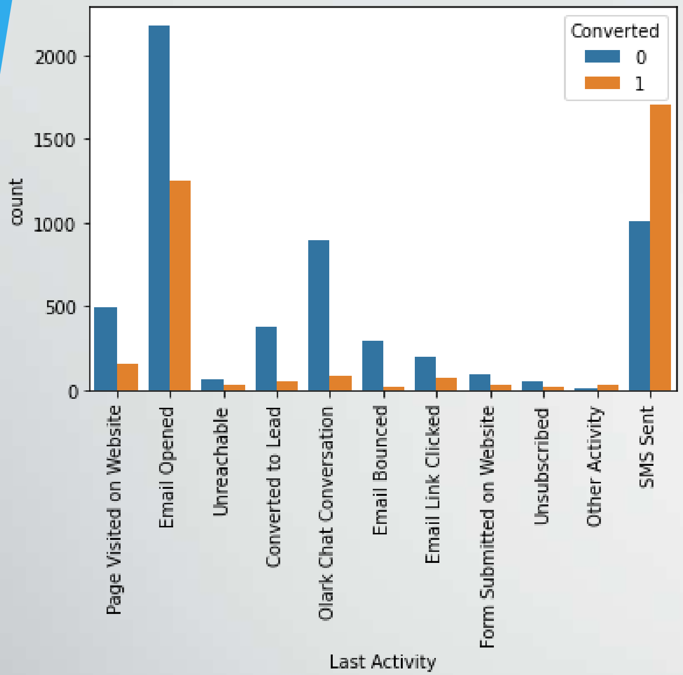
- Lead Import has very less count as well as conversion rate and hence can be ignored

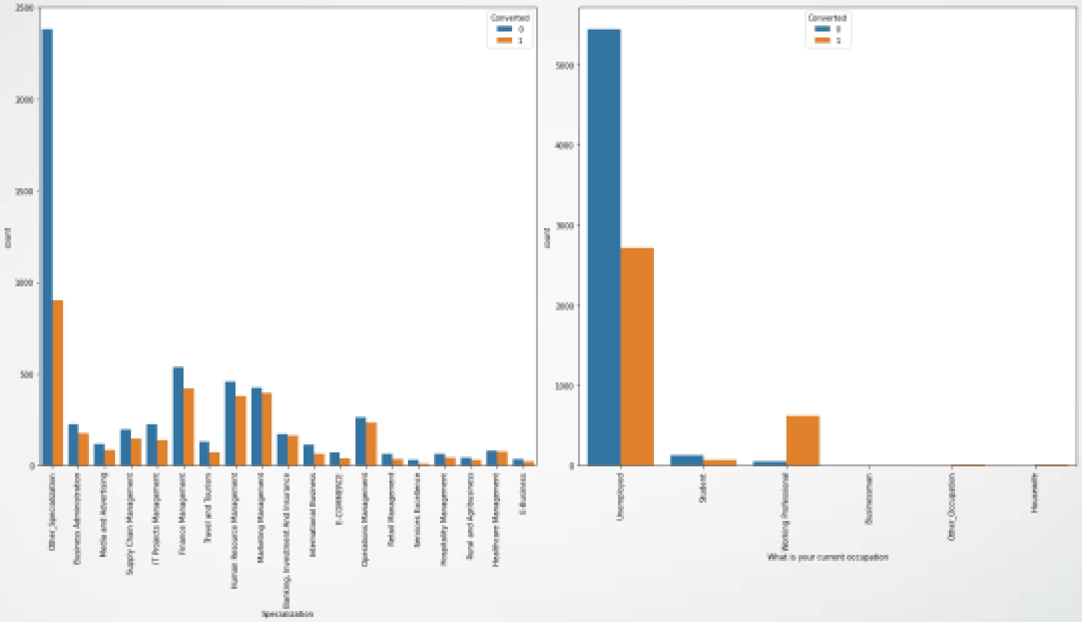




- The median of both the conversion and non-conversion are same and hence nothing conclusive can be said using this information

- Users spending more time on the website are more likely to get converted





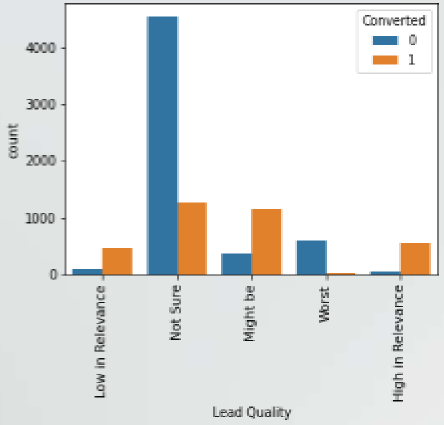
•The count of lst activity as "Email Opened" is max

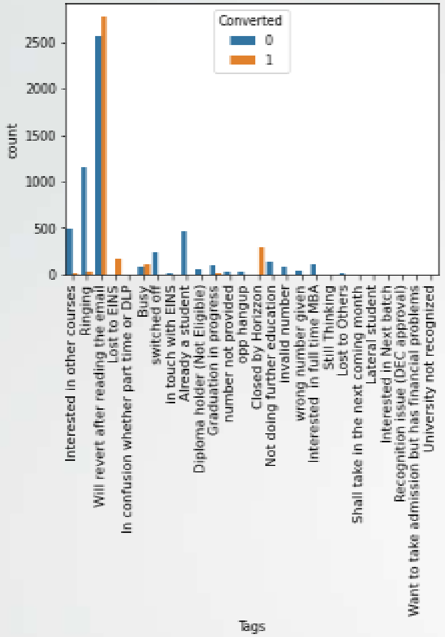
•The conversion rate of SMS sent as last activity is maximum

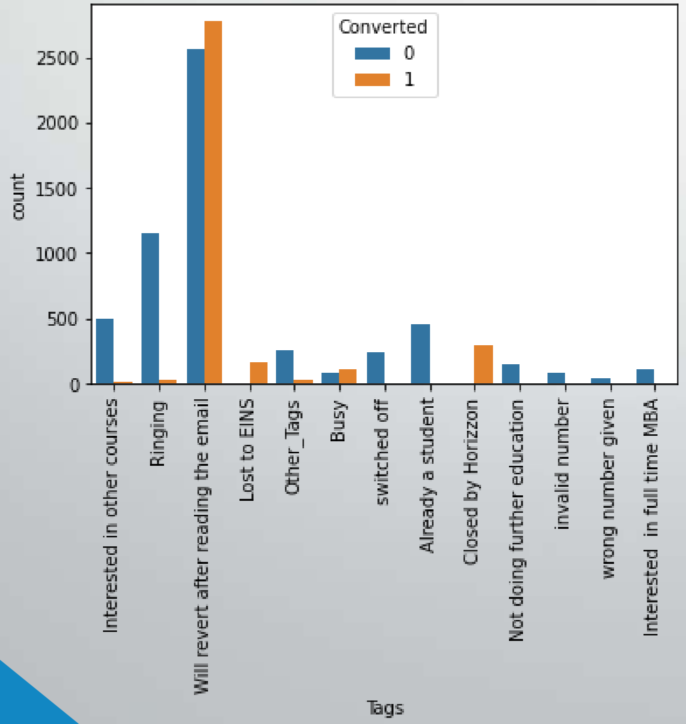
•Looking at above plot, no particular inference can be made for Specialization

•Looking at above plot, we can say that working professionals have high conversion rate

•Number of Unemployed leads are more than any other category



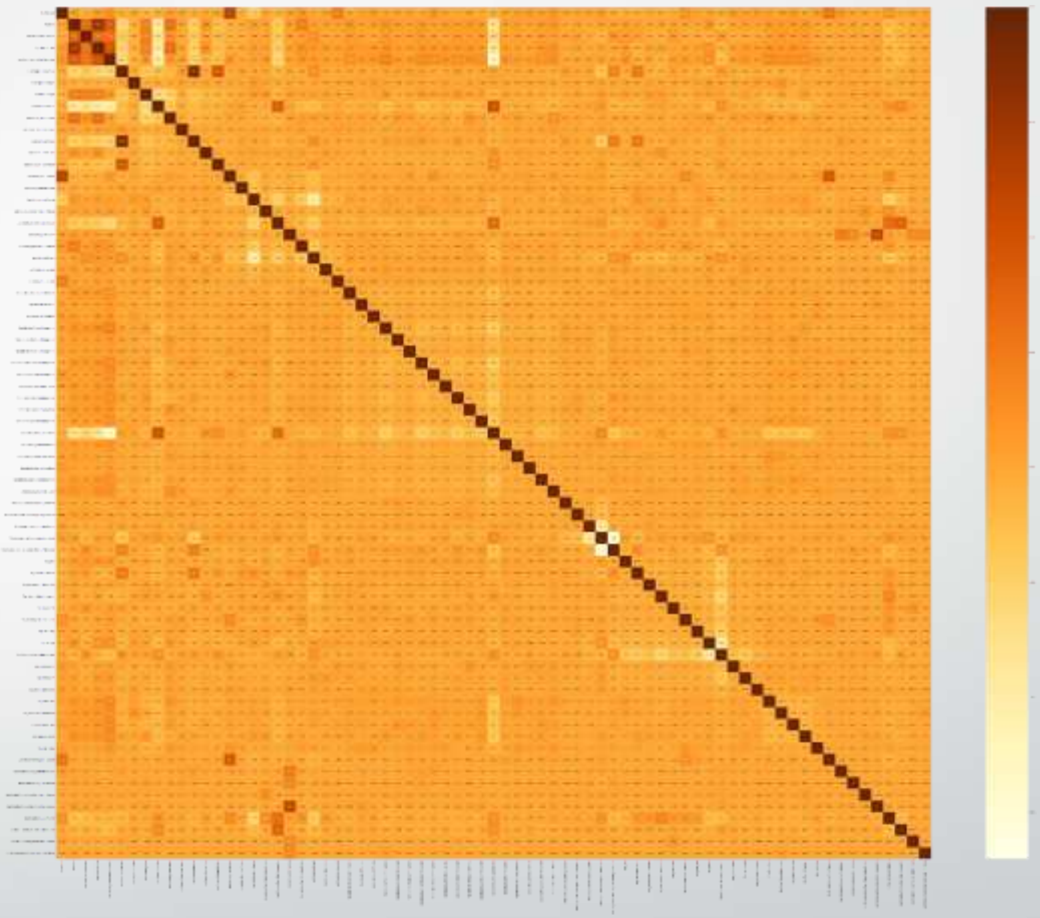




- 'Will revert after reading the email' and 'Closed by Horizzon' have high conversion rate

Correlation:

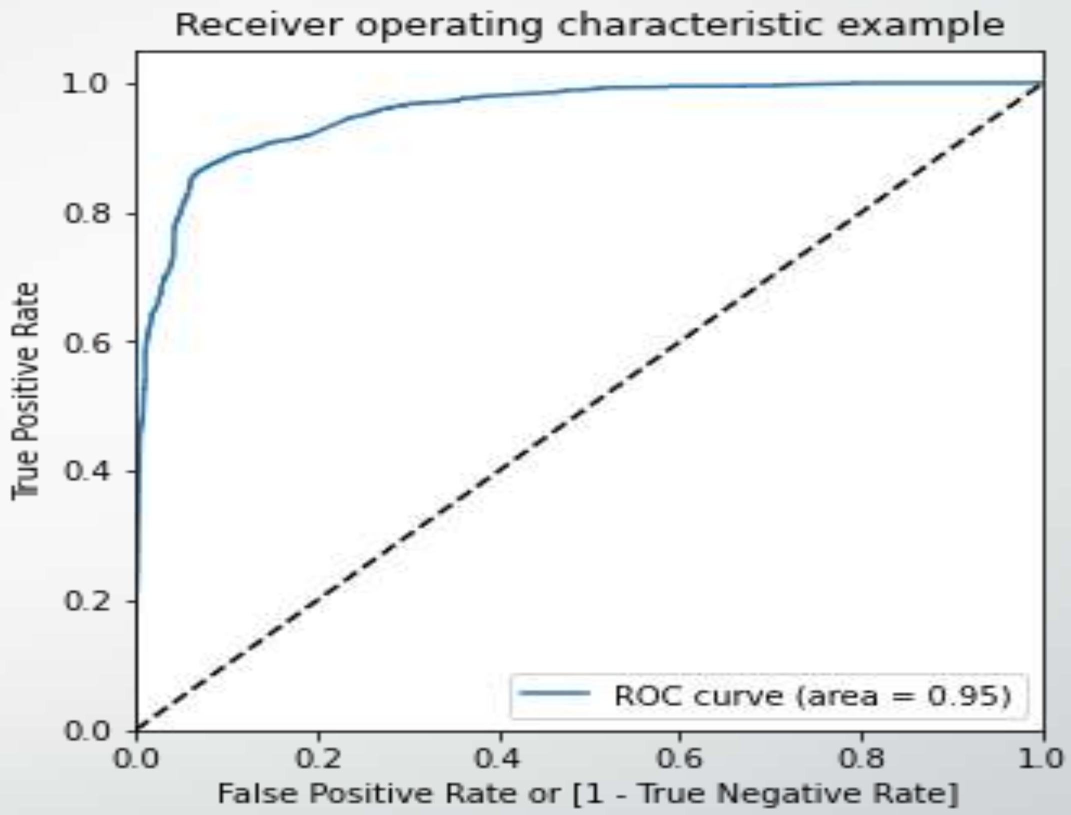
- After dropping those high correlations features, we plotted again a heatmap to check and it was confirmed that those highly correlated variables were dropped.
- There are still few left, but we will check them after creating our model to verify how much they are impacting, as from the plot on the right it is not quite understandable which variable is having high correlation.



Plotting the ROC Curve

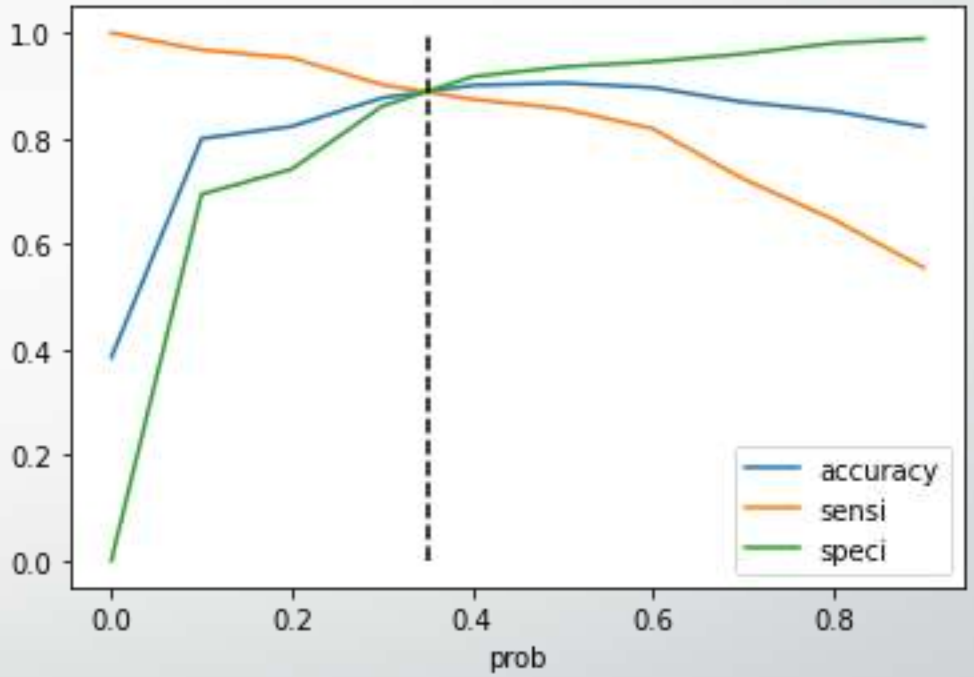
An ROC curve

- shows tradeoff between sensitivity and specificity (increase in one will cause decrease in other).
- The closer the curve follows the y-axis and then the top border of the ROC space, means more area under the curve and the more accurate the test.
- The closer the curve comes to the 45-degree diagonal of the ROC space i.e. the reference line, means less area and the less accurate is the test.



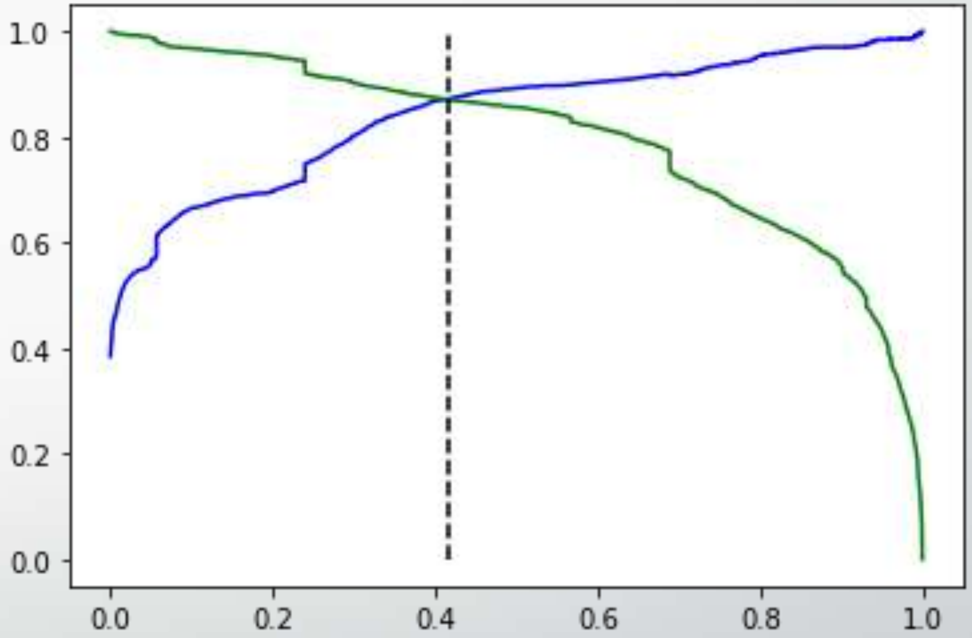
Finding the optimal cutoff point

- Now, we have created range of points for which we will find the accuracy, sensitivity and specificity for each points and analyze which point to chose for probability cutoff.



Precision-Recall Trade off

We used this cutoff point to create a new column in our final dataset for predicting the outcomes.
After this we did another type of evaluation which is by checking Precision and Recall
In Sensitivity-Specificity-Accuracy plot 0.35 probability looks optimal.
In Precision-Recall Curve 0.41 looks optimal.
We are taking 0.35 is the optimum point as a cutoff probability and assigning Lead Score in training data.



Model Metrics Running model on features selected we get following metrics:

1. Train Data:

❖ Confusion Metrics:

- Accuracy: 90.5%
- Sensitivity: 85.65%
- Specificity: 93.54%
- Precision: 89.26%
- Recall: 85.65%

	Not Converted Leads	Converted Leads
Not Converted Leads	3653	252
Converted Leads	351	2095

2. Test Data:

❖ Confusion Metrics:

- Accuracy: 77.23%
- Sensitivity: 93%
- Specificity: 68.34%
- Precision: 63%
- Recall: 93%

	Not Converted Leads	Converted Leads
Not Converted Leads	1185	549
Converted Leads	71	918

The Model seems to predict the Conversion Rate very well. We should be able to help the education company select the most promising Leads or the Hot Leads

Conclusion:

- The logistic regression model predicts the probability of the target variable having a certain value, rather than predicting the value of the target variable directly. Then a cutoff of the probability is used to obtain the predicted value of the target variable.

- Here, the logistic regression model is used to predict the probability of conversion of a customer.

- The optimum cut-off is chosen to be 0.35 i.e. any lead with greater than 0.35 probability of converting is predicted as Hot Lead (customer will convert) and any lead with 0.35 or less probability of converting is predicted as Cold Lead (customer will not convert)

- Our final Logistic Regression Model is built with 20 features.

- Features used in the final model are ['Total Time Spent on Website', 'Lead Origin_Lead Add Form', 'Lead Source_Olark Chat', 'Lead Source_Welingak Website', 'Last Activity_Email Bounced', 'Last Activity_SMS Sent', 'Tags_Closed by Horizzon', 'Tags_Lost', 'Tags_No phone number‘, 'Tags_Others', 'Tags_Will revert after reading the email', 'Last Notable Activity_Modified', 'Last Notable Activity_Olark Chat Conversation"]

- The top three categorical/dummy variables in the final model are ‘Tags_Lost to EINS’, ‘Tags_Closed by Horizzon’, and ‘Total Time Spent on Website’ with respect to the absolute value of their coefficient factors.

- The final model has a Sensitivity of 0.93, which means the model is able to predict 93% of customers out of all the converted customers, (Positive conversion) correctly.

- The final model has a Precision of 0.63, which means 63% of predicted hot leads are True Hot Leads.

THANK YOU