

Lead Score Case Study

Submitted by:
SinchanShetty
Sincy Thomas
Niharika Kummari

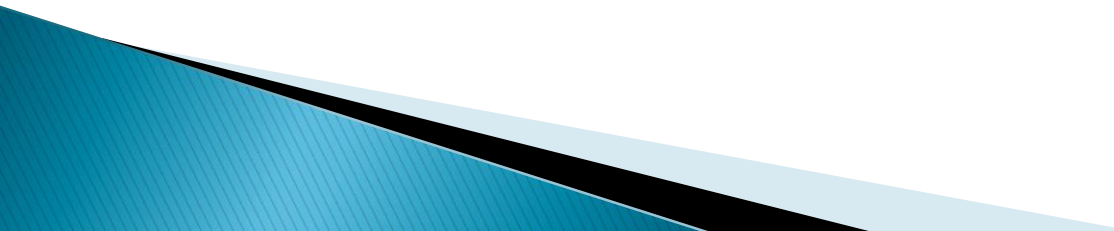
Problem Statement

- ▶ *X Education 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.*

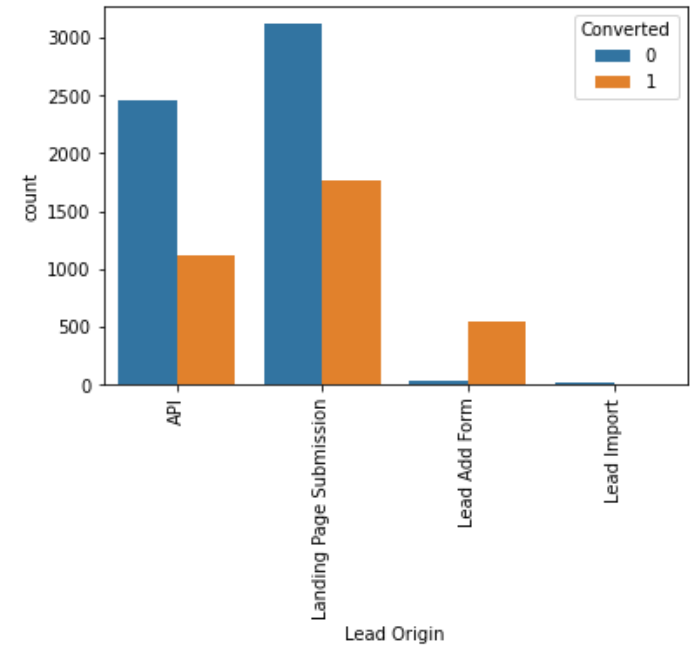
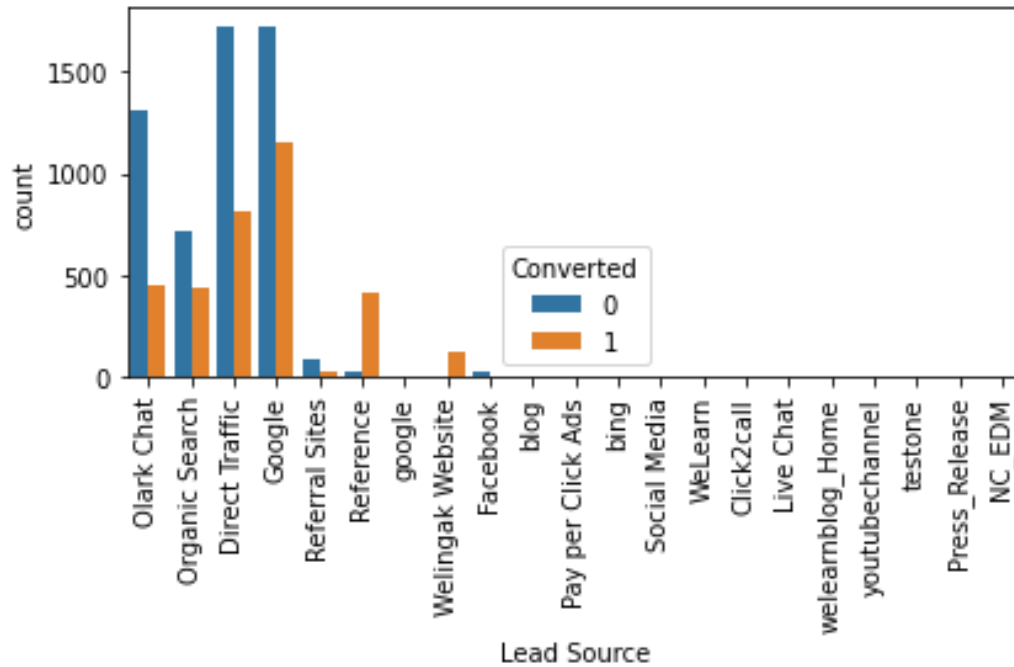
Business Objective

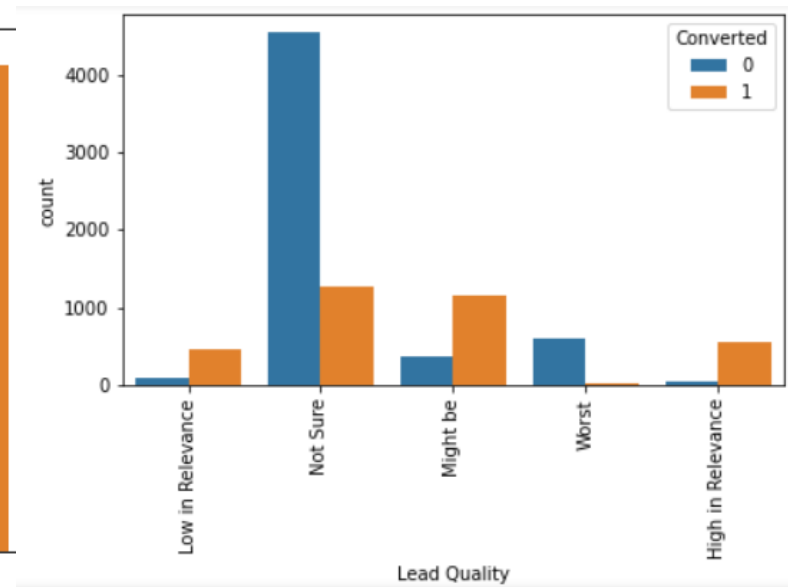
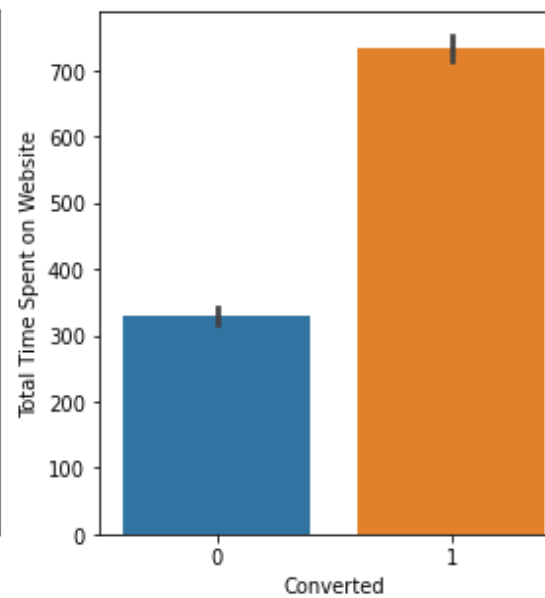
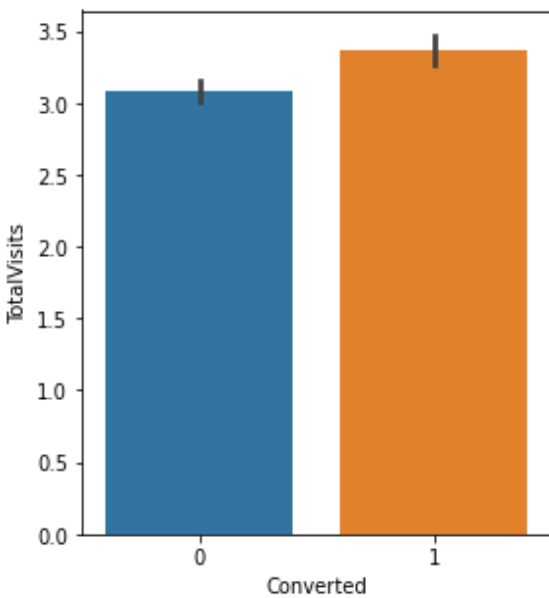
- ▶ *X education wants to know most promising leads.*
- ▶ *They want to build a Model which identifies the hot leads.*
- ▶ *TO build a model whose lead conversion rate around 80%*

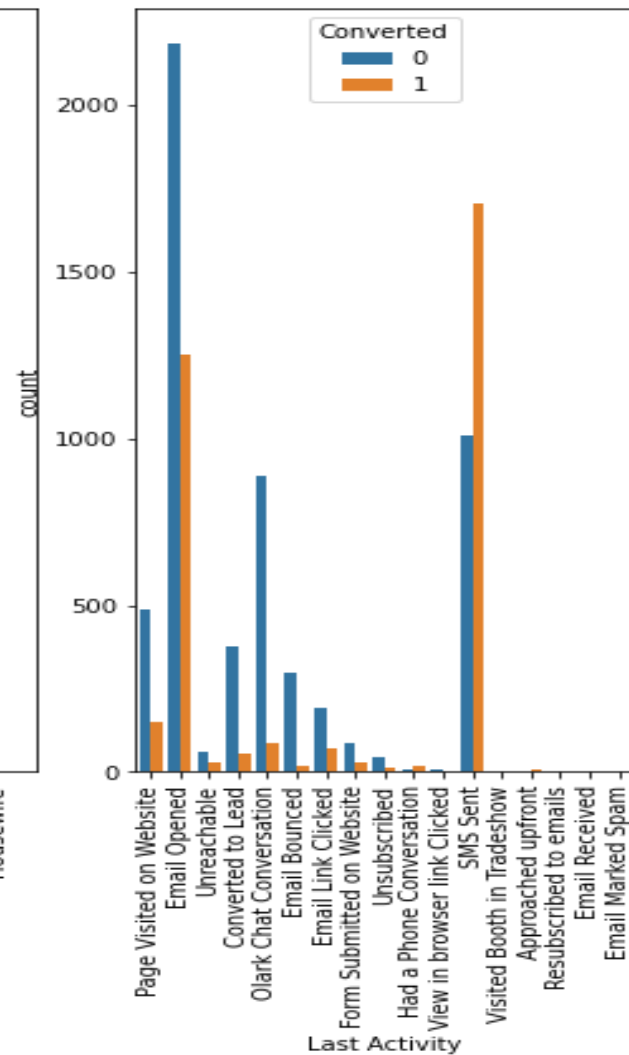
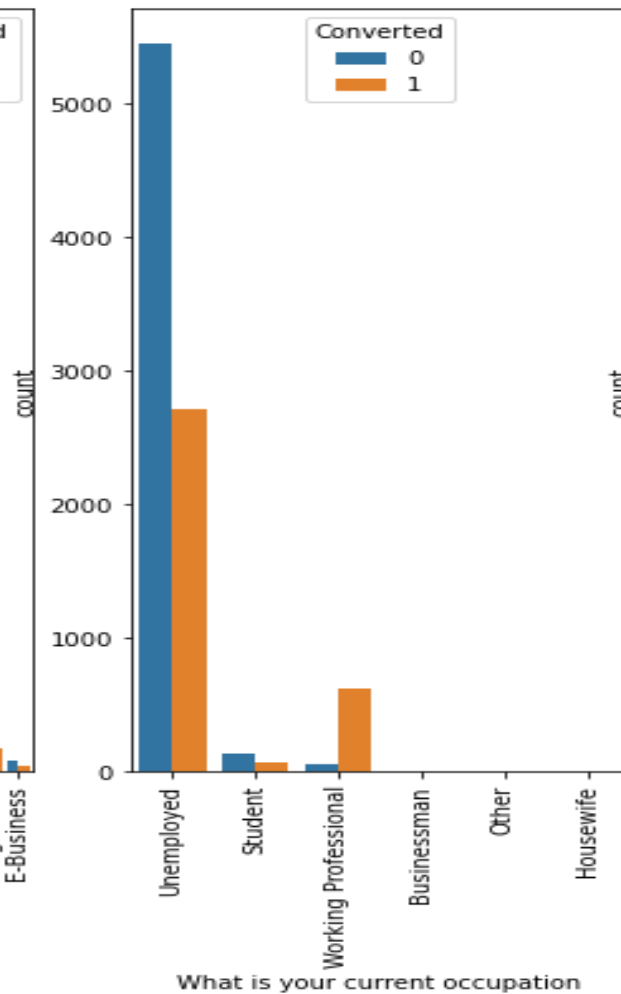
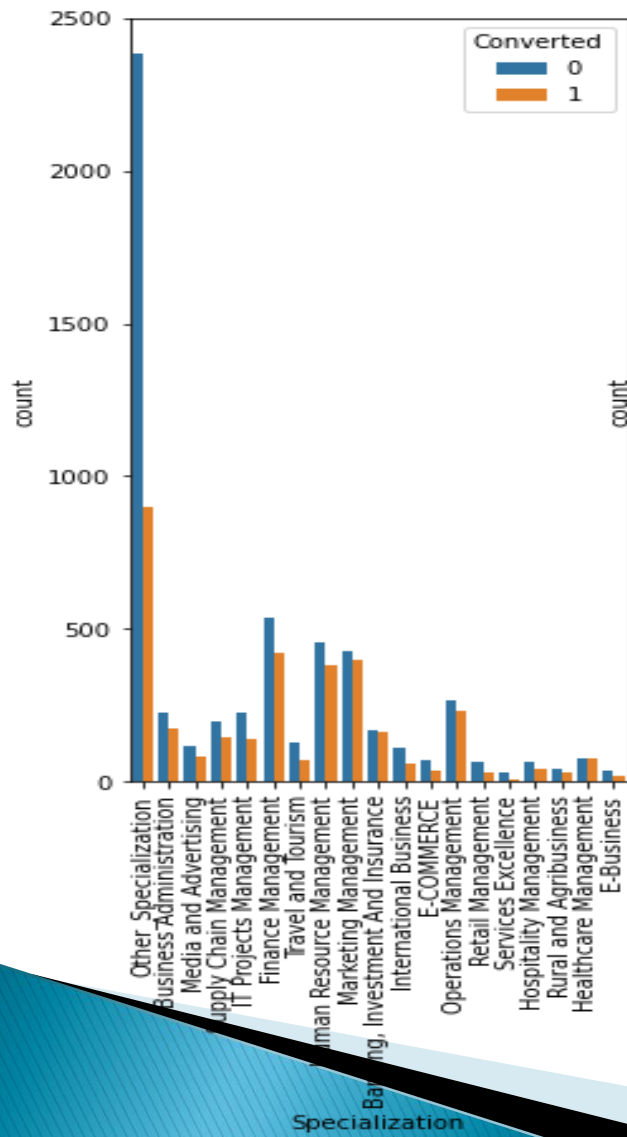
Steps in Analysis

- ▶ *Import data and necessary libraries.*
 - ▶ *Clean and prepare the data.*
 - ▶ *Exploratory Data Analysis.*
 - ▶ *Dummify and Feature Scaling.*
 - ▶ *Splitting the data into Test and Train dataset.*
 - ▶ *Building a logistic Regression model and calculate Lead Score.*
 - ▶ *Evaluating the model by using different metrics – Specificity and Sensitivity or Precision and Recall.*
 - ▶ *Applying the best model in Test data based on the Sensitivity and Specificity Metrics.*
- 

EDA

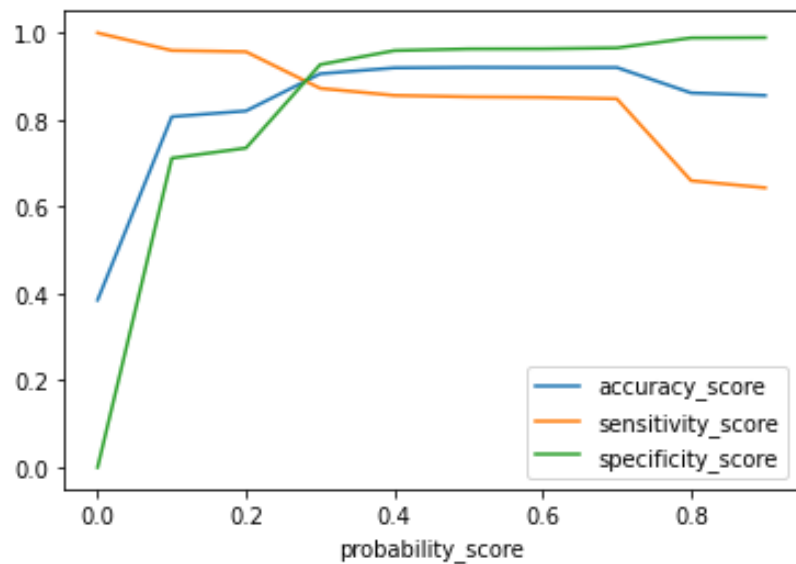






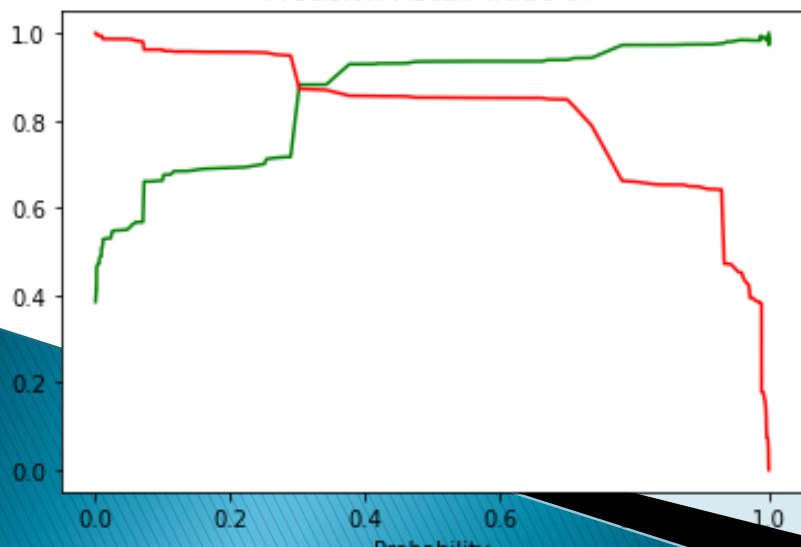
Model Building

- ▶ *Split into train and test set*
- ▶ *Scaling the variables of train set*
- ▶ *Build the first model*
- ▶ *Use RFE and VIF to remove less relevant variable*
- ▶ *Build next model*
- ▶ *Less p value and vif select the model*
- ▶ *Predict using train set*
- ▶ *Evaluate accuracy and other metric*
- ▶ *Predict using test set*
- ▶ *Precision and recall analysis on test prediction*



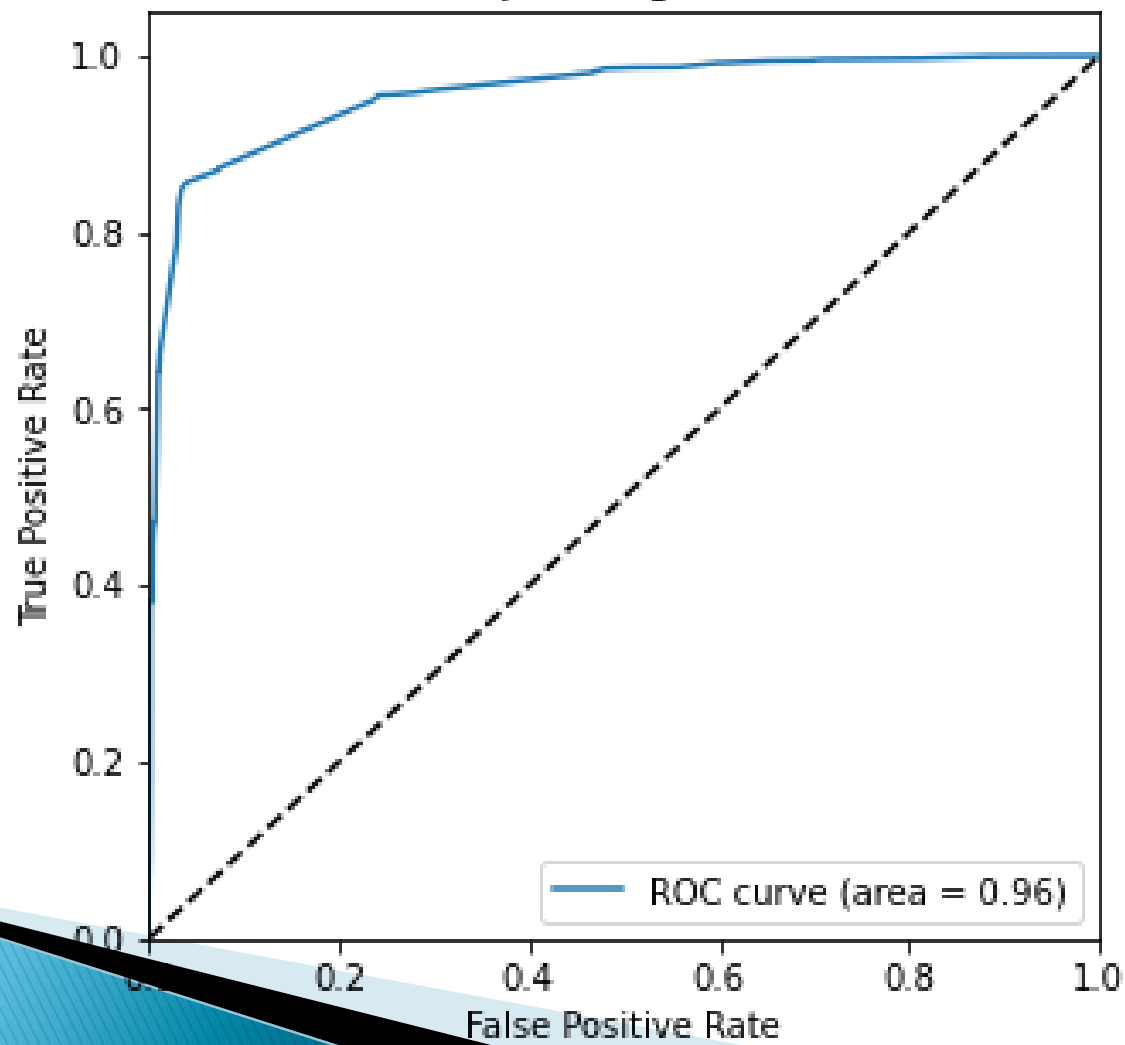
*83% Accuracy
94% sensitivity
76% specificity*

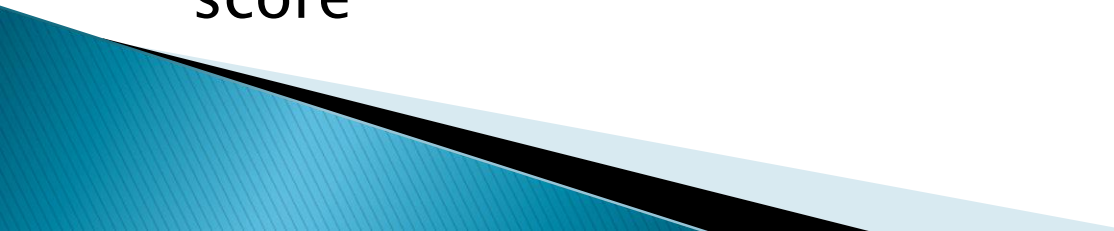
Precision-Recall Trade off



71% precision

Receiver Operating Characteristic



- ▶ 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.
 - ▶ Here, our goal is to have achieve good sensitivity score
- 

conclusion

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

cut off is chosen to be 0.27 i.e. any lead with greater than 0.27 probability of converting is predicted as Hot Lead (customer will convert) and any lead with 0.27 or less probability of converting is predicted as Cold Lead (customer will not convert).

Our final Logistic Regression Model is built with 14 features.

features used in final model are

- ▶ *['Do Not Email', 'Lead Origin_Lead Add Form', 'Lead Source_Welingak Website', 'Last Activity_SMS Sent', 'Tags_Busy', 'Tags_Closed by Horizzon', 'Tags_Lost to EINS', 'Tags_Ringing', 'Tags_Will revert after reading the email', 'Tags_switched off', 'Lead Quality_Not Sure', 'Lead Quality_Worst', 'Last Notable Activity_Modified', 'Last Notable Activity_Olark Chat Conversation']*

The final model has Sensitivity of 0.928, this means the model is able to predict 92% customers out of all the converted customers, (Positive conversion) correctly.

The final model has Precision of 0.68, this means 68% of predicted hot leads are True Hot Leads.







