LEAD SCORING CASE STUDY

PRESENTED BY-

KOPPU THIRUPATHI

SWATI SINGH &

MARIAM ABBAS ZAIDI

PROBLEM STATEMENT:

To help select and filter the most promising leads for an online education company named X Education.

We have to build a model which assigns a lead score to each of the customers such that ones with higher lead score means they have a higher chance of converting to paying customers and the ones with lower score have lower chance.

The Business aim is to identify and acquire these potential leads called "Hot Leads". The typical lead conversion rate at X Education is around 30%. If we help them successfully identify these Hot Leads then their lead conversion rate should go upto 80%, as their sales team will be able to focus more on communicating with Hot Leads.

STRATEGY:

DATA CLEANING

- Read & understand data
- Remove null values & convert to correct data types
- Treat outliers
- Perform EDA

DATA PREPARATION

- Converting binary variables
- Creating dummies
- Performing train-test split
- Perform Scaling

MODEL BUILDING

- Feature selection(RFE)
- Building Model using Logistic Regression

MODEL EVALUATION

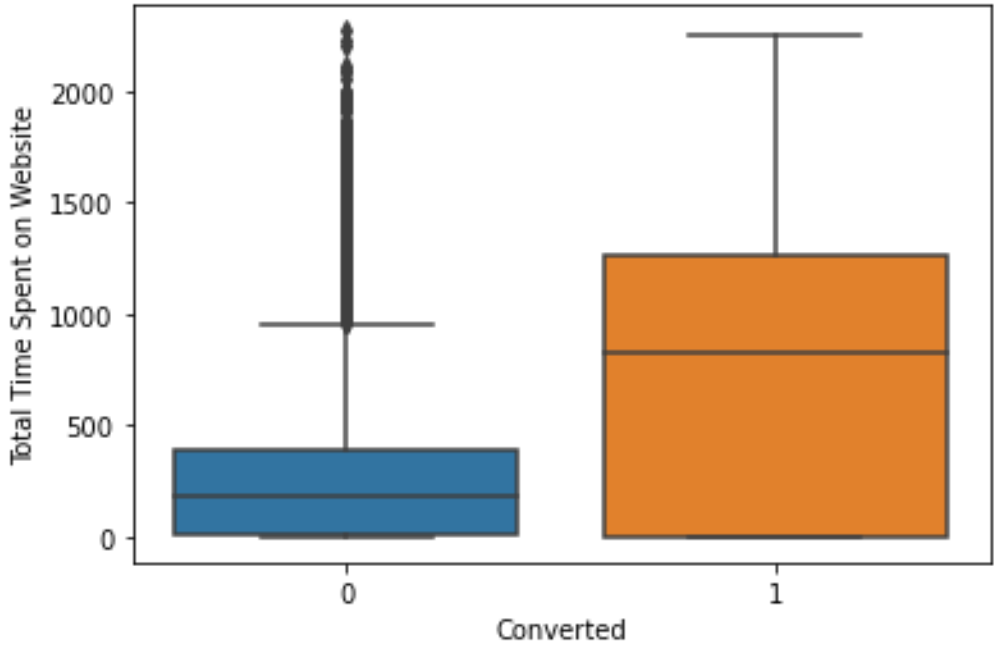
- Evaluating the confusion matrix parameters ie. Accuracy, Sensitivity, Specificity, Precision, Recall, etc on test and train data

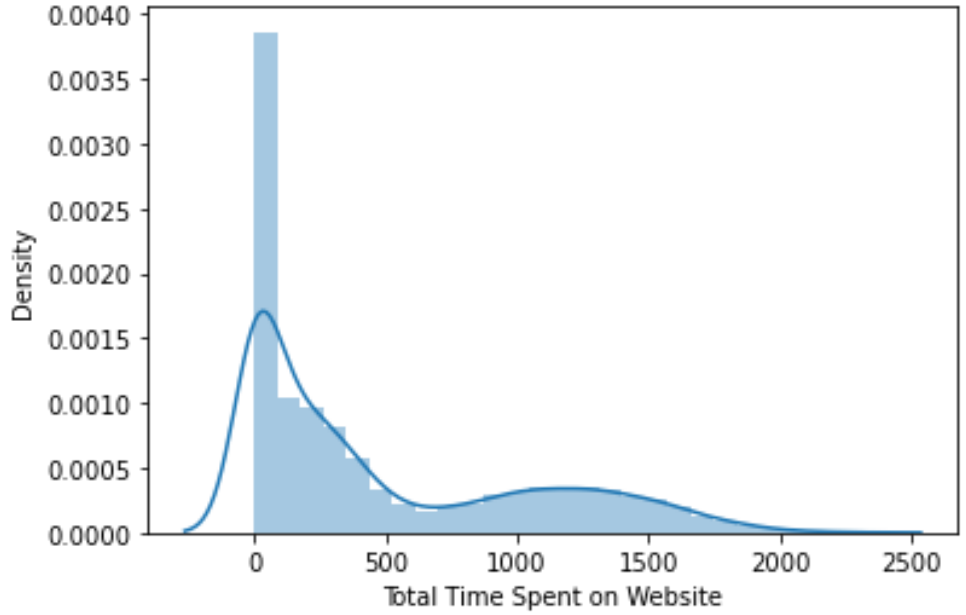
- Plotting the ROC curve

DATA CLEANING:

All columns having null values more than 40% were dropped. There are many columns where the customer has not entered any value and it shows as “Select” which is as good as null value. Since some of the columns has a high imbalance in Select and other options it is as good as getting rid of these highly imbalanced columns. The outliers have been found and removed. The missing values if less than 40% have been substituted with either mode or median values.

EXPLORATORY DATA ANALYSIS:

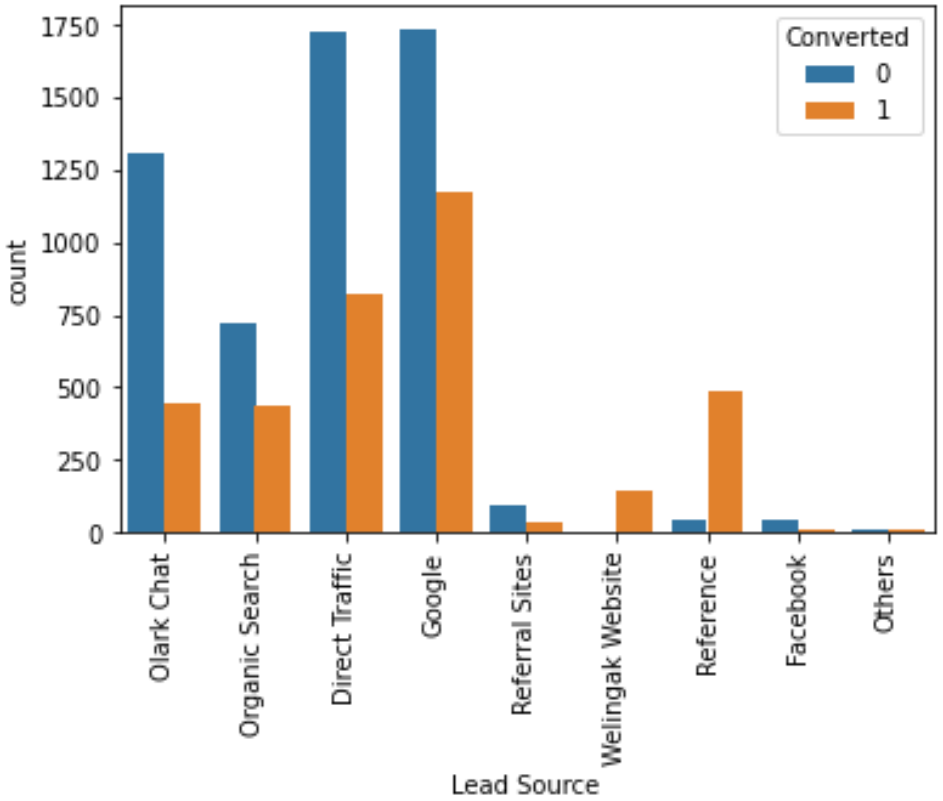




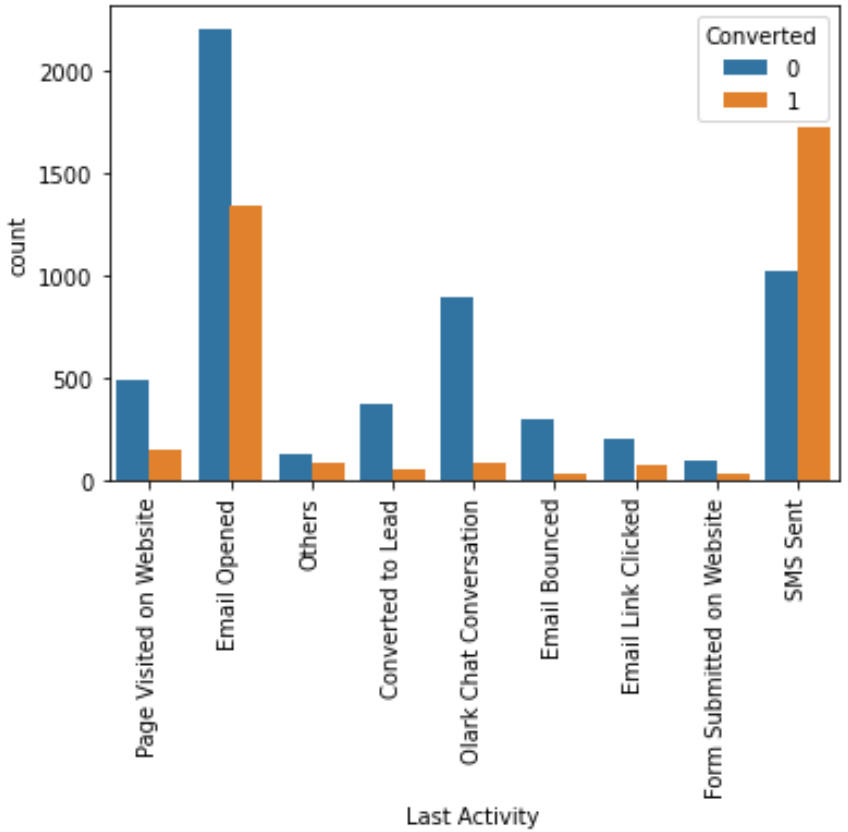
Customers who spend more time on website are likely to convert

The probability of time spent is found to be high for time between 0-300 seconds and decreases further.

EXPLORATORY DATA ANALYSIS:

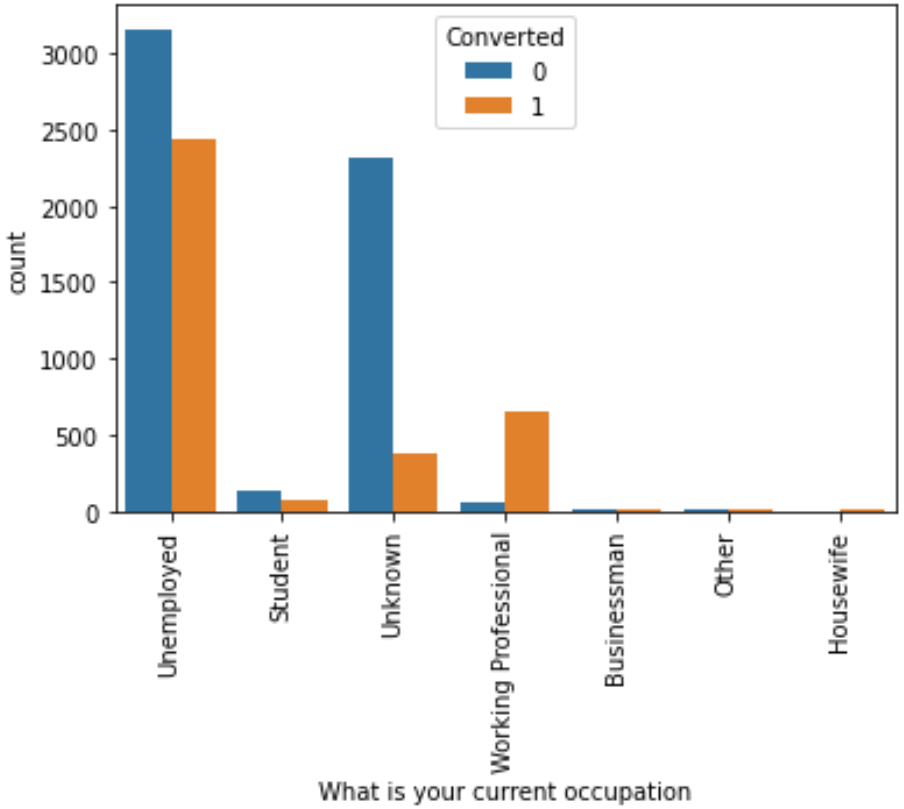


Major conversion in lead source is from Google

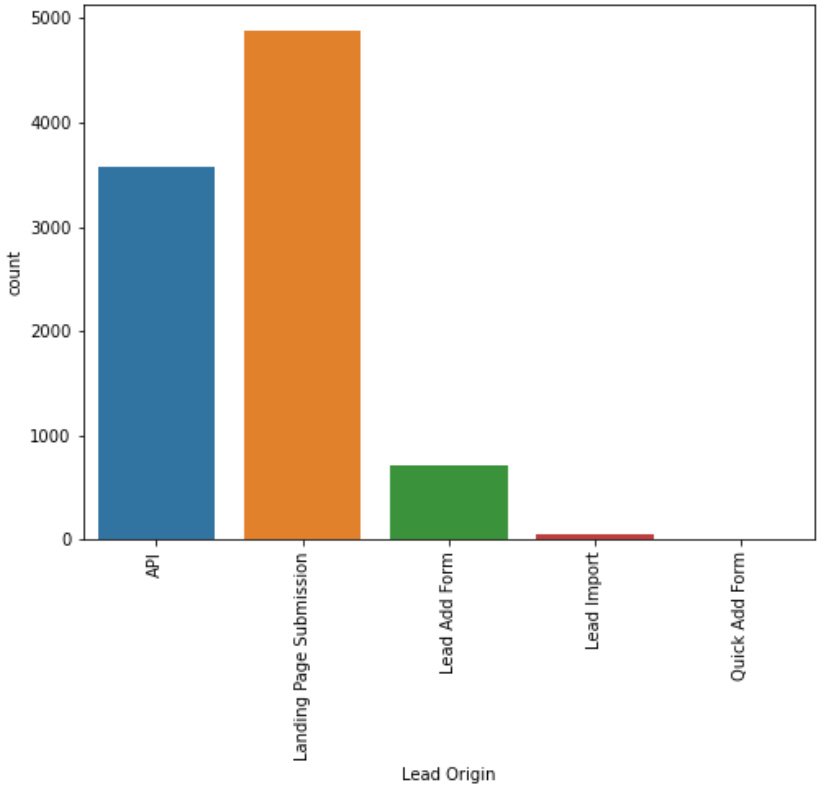


Found that it is better to target customers through email and SMS

EXPLORATORY DATA ANALYSIS:



Mostly unemployed and working professionals who want to upskill have maximum conversions



Most leads are generated by API and Landing page submission

MODEL BUILDING:

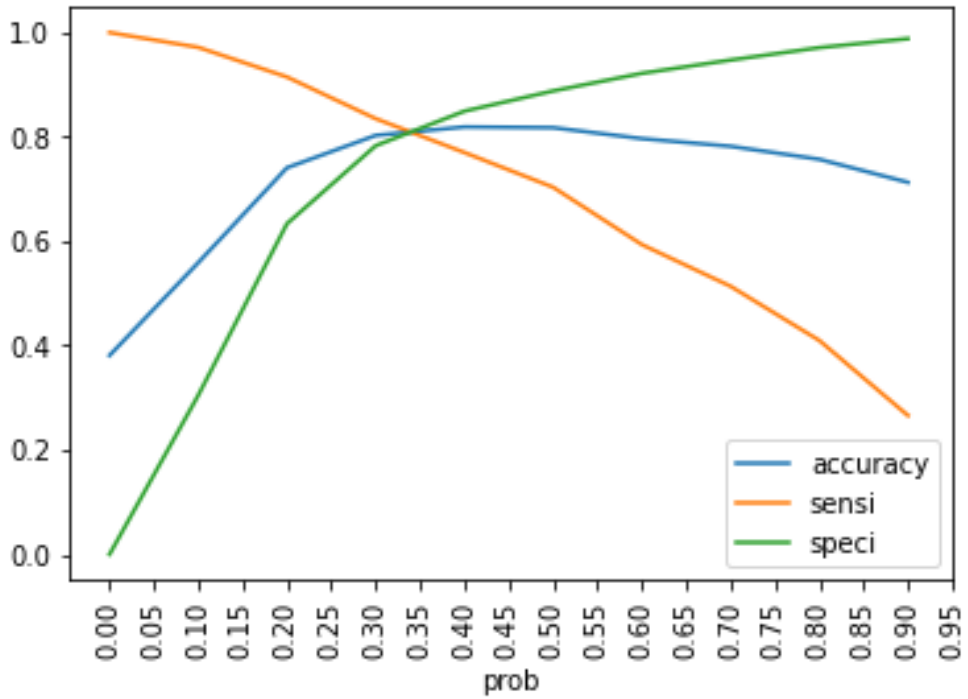
Firstly dummy variables were created for all categorical columns. Then we split the data into train and test dataset set in 70 – 30 ratio for our working.

Next we will be scaling all the features in our model so that they fall under the same scale. We do this using StandardScaler performing fit_transform on train data set.

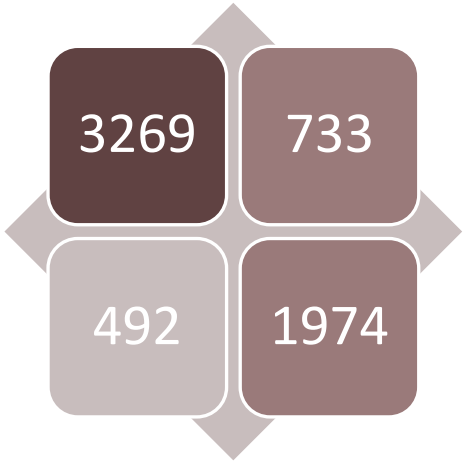
RFE method automatically eliminates unwanted features and keeps 15 of them for further assessment.

The rest of the features are checked for high p-values and high VIF and further eliminated.

MODEL EVALUATION (on Train Data-set):



The optimal cut-off probability based on Accuracy, Sensitivity and Specificity is 0.35

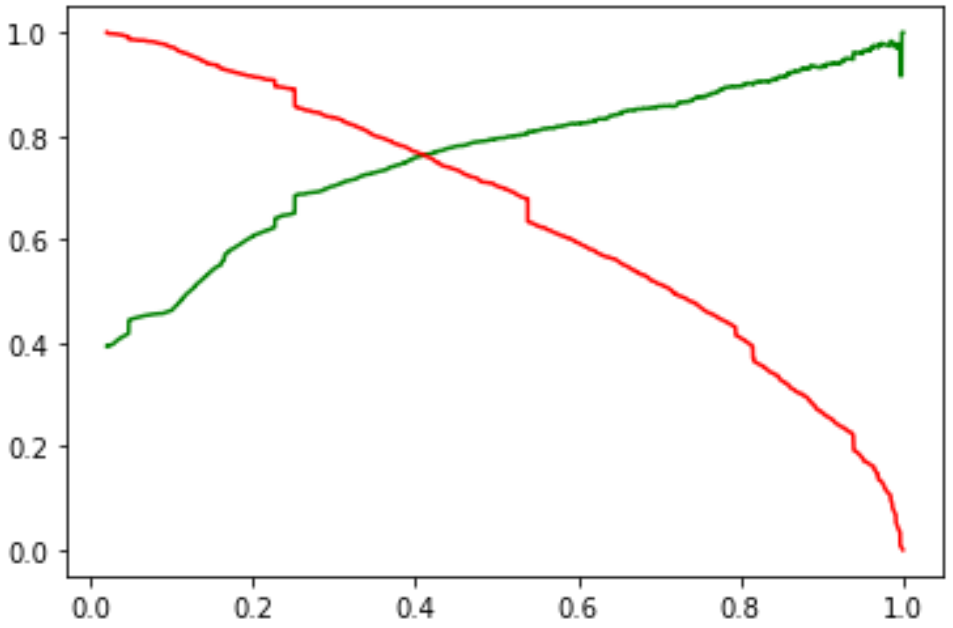


ACCURACY - 81.06%

SENSITIVITY - 80.05%

SPECIFICITY - 81.7%

MODEL EVALUATION (on Train Data-set):



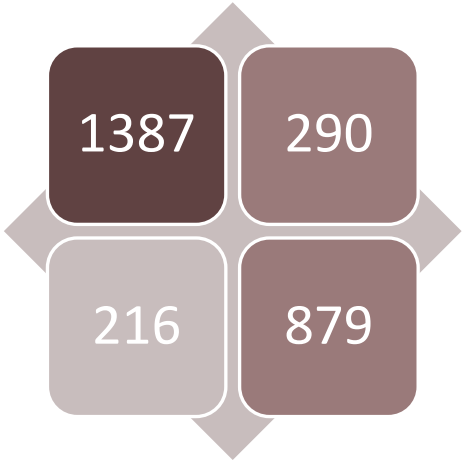
The optimal cut-off probability based on Precision and recall is 0.42

3269	733
492	1974

PRECISION – 72.92%

RECALL – 80.04 %

MODEL EVALUATION (on Test Data-set):



ACCURACY - 81.75%

SENSITIVITY - 80.0%

SPECIFICITY – 82.71%

PRECISION – 75.19%

RECALL – 80.27 %

CONCLUSION:

The difference between the test and train datasets performance metrics is very less, hence, our final model is performing well.

The top 3 features that can account for lead conversion are:

- The customers who fill Add form
- Working Professionals
- Total time spend on website

High sensitivity means that most of the leads who are likely to convert are correctly predicted, whereas high specificity ensures that most of the leads who are not likely to convert are correctly predicted. Hence we have achieved high sensitivity 80% which is required of us to do.