Lead Score Case Study
Logistic Regression

Created By:
Namia Mohamed Ali

Problem Statement

- X Education is an education company which sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google. On any given day, many professionals who are interested in the courses land on their website and browse for courses. When these people fill up a form providing their email address or phone number, they are classified to be a lead.

- X Education wants to select most promising leads that can be converted to paying customers. Now, although X Education gets a lot of leads, its lead conversion rate is very poor.

- The company has had 30% conversion rate through the whole process of turning leads into customers by approaching those leads which are to be found having interest in taking the course. The implementation process of lead generating attributes are not efficient in helping conversions

Business Objective

- The company requires a model to be built for select most promising lead. They want a strategy to distinguish between the leads which are more promising and their chances of getting converted are higher with the ones which are less promising.

- Assign 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.

- The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

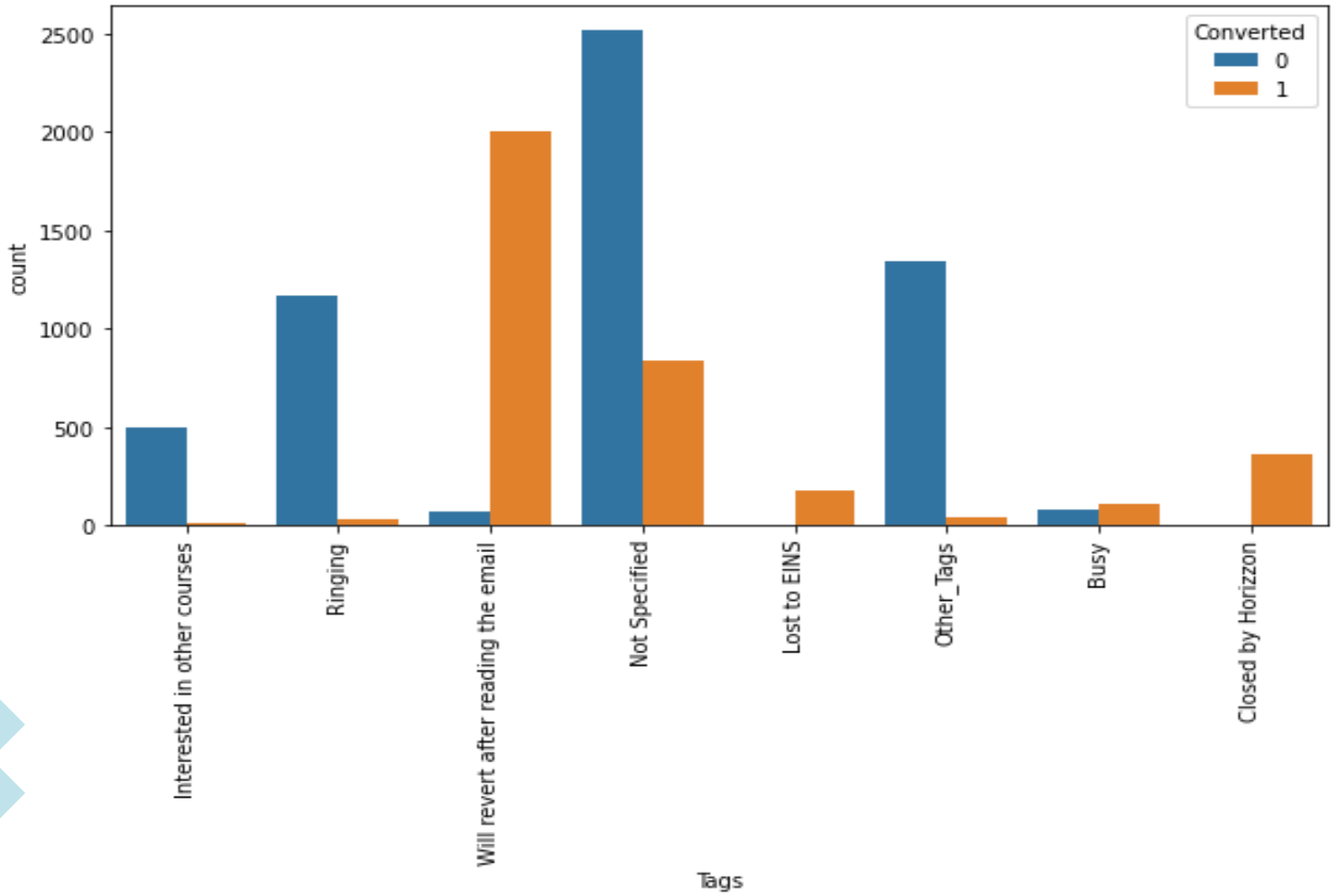
APPROACH

- Reading and understanding the data
- Data Cleaning
- EDA
- Creating Dummy Variables.
- Feature Scaling
- Splitting the data into train-test dataset
- Model Building
- Model Evaluation- Accuracy, Sensitivity, Specificity, Precision, recall
- Making predictions on test set.

Data Understanding Cleaning

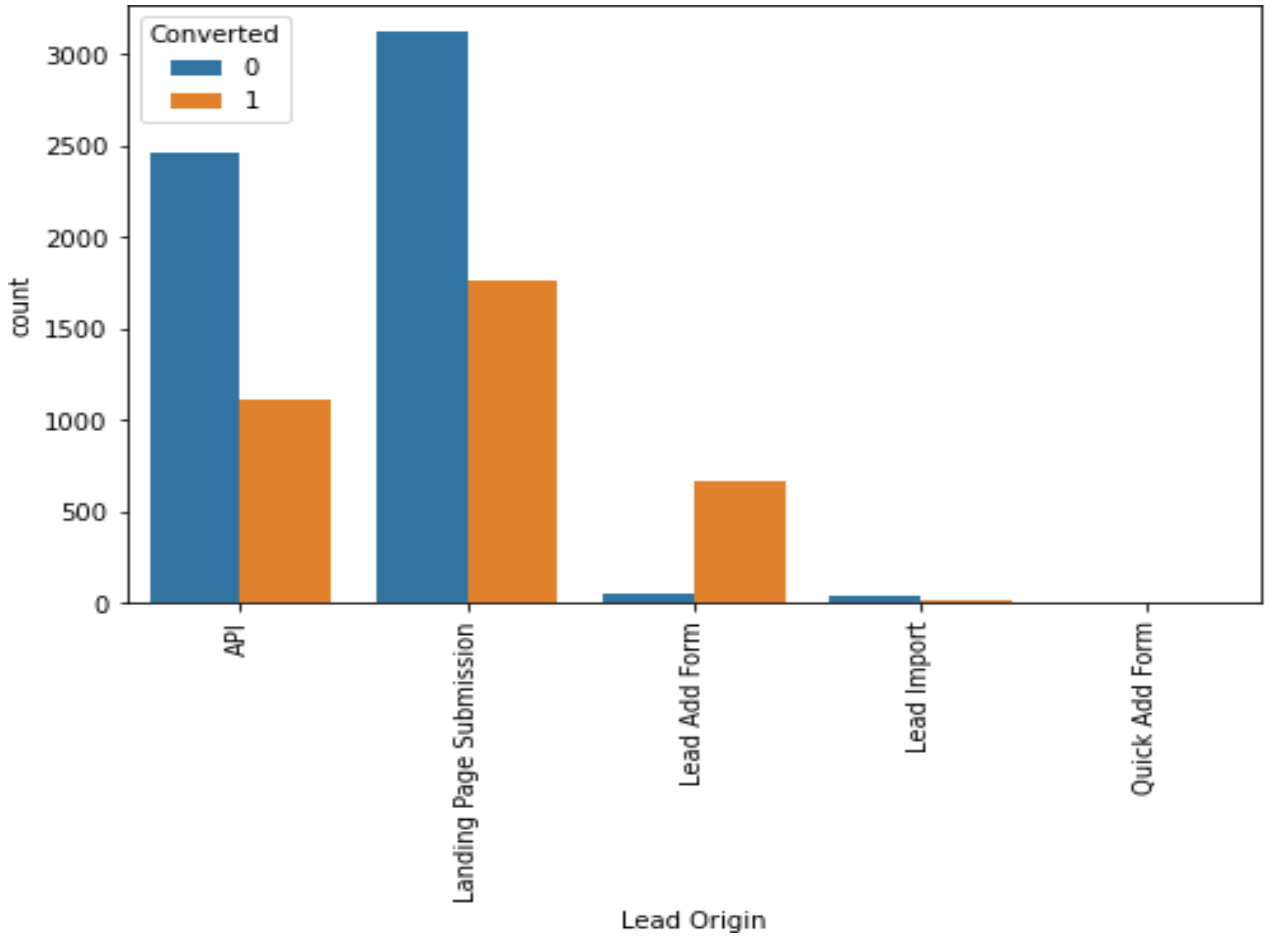
- Firstly, we analyzed the data to know more about its Attributes.
- Columns with more than 45% of null values were dropped.
- We checked some of the columns had one class with more than 90% of data. Those columns had class imbalance so, we dropped those columns.
- Lastly, we checked if any column was not relevant for our business objective so, we dropped those columns also.
- Rest of the columns were imputed with mode or we kept the null values as unknown. We performed outliers treatment on numerical columns.

Exploratory Data Analysis



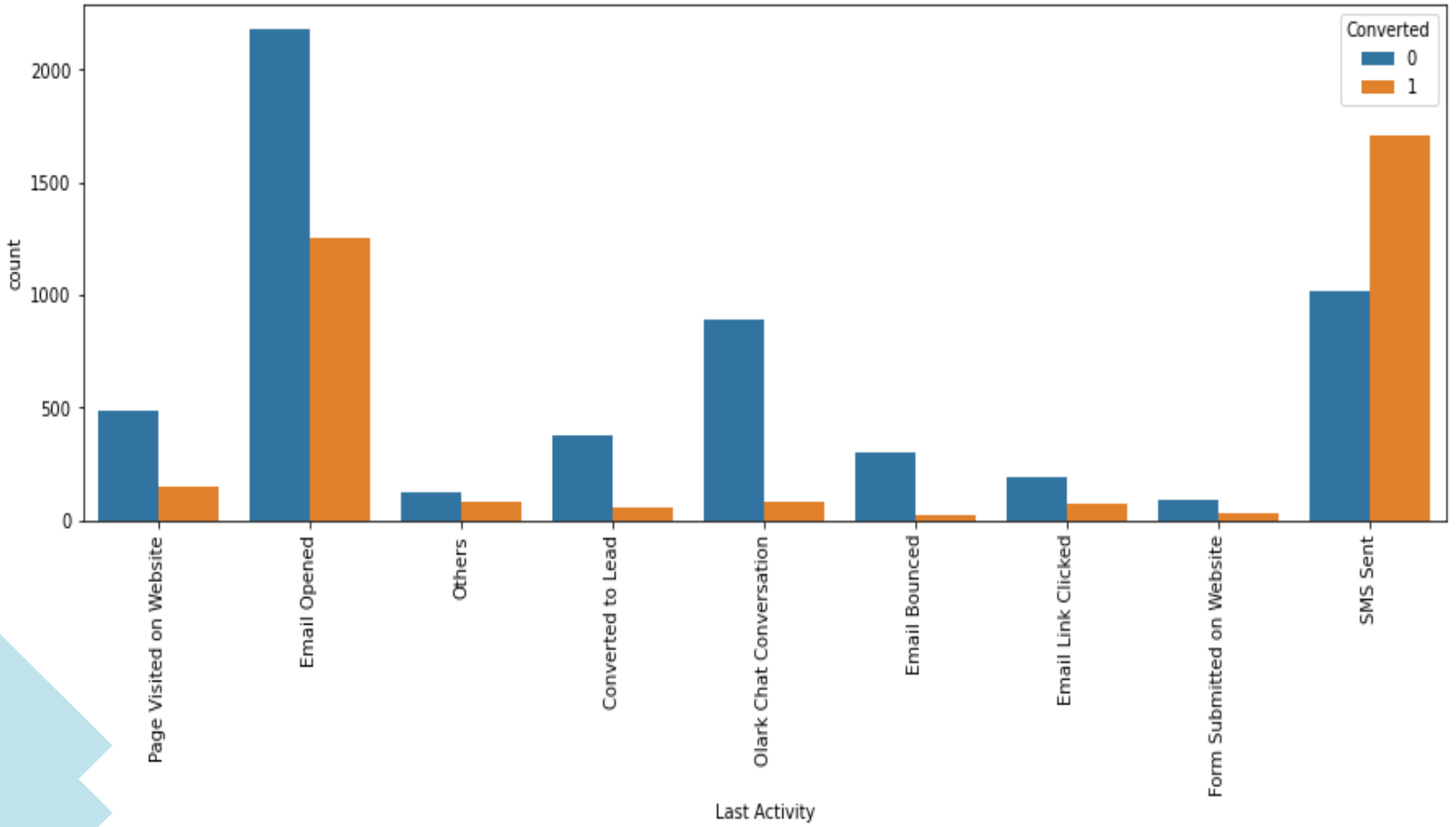
Tags represents the current status of the customers, We can clearly see that customers who spend has their tag status as closed by horizon followed by lost to EINS has higher rate of conversion. There are more leads who has tag will revert after email with higher chances of conversion.

Exploratory Data Analysis



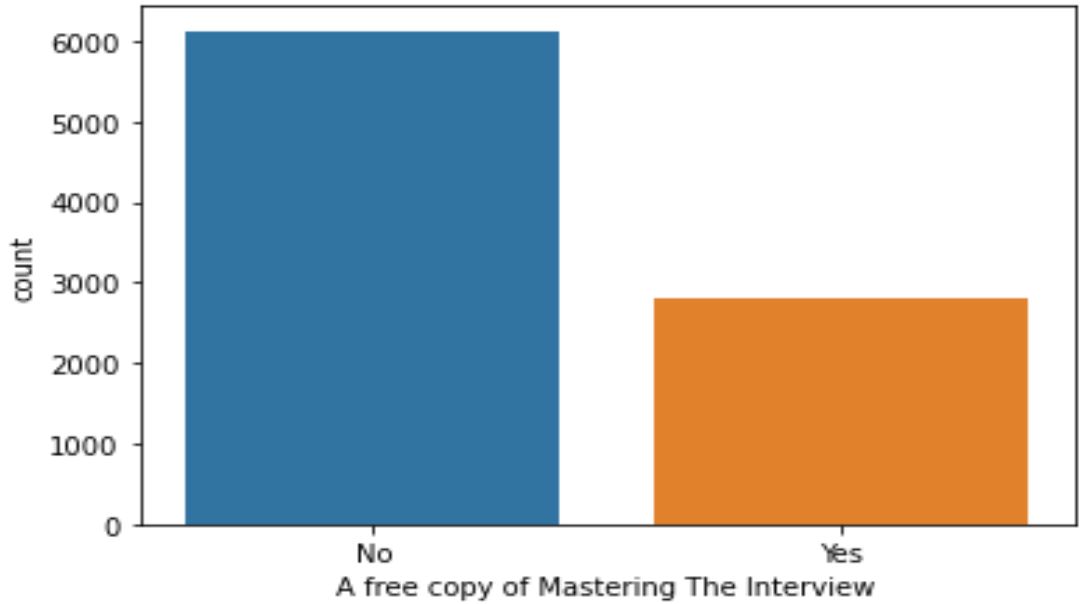
- We can clearly see that Lead Add Form has higher percentage of converted leads and therefore business should focus on generating more leads from this category.
- API and Landing Page Submission get a lot of leads but their conversion rate is less. Business should focus on improving conversion rate of these categories.

Exploratory Data Analysis

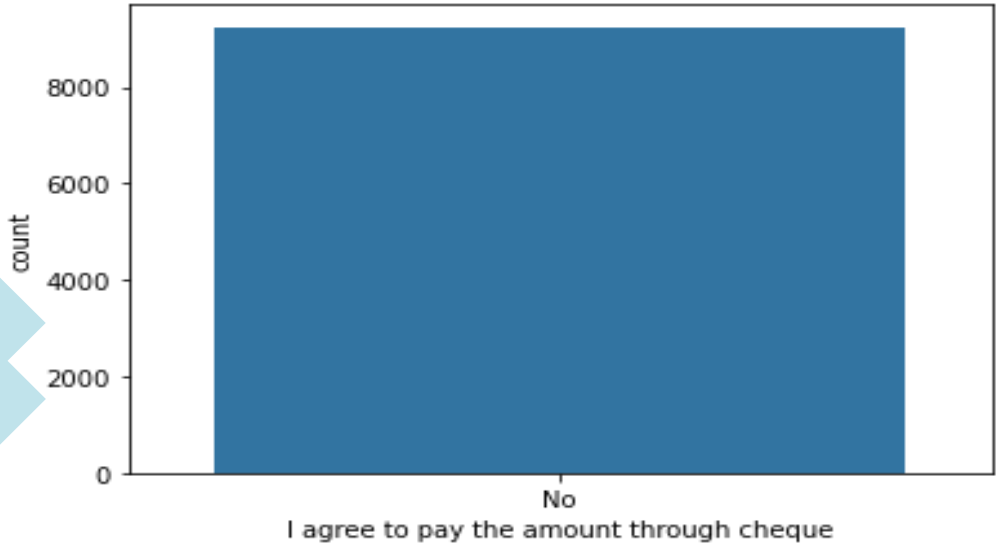


- From the plot it is evident that leads whose Last Activity is 'SMS sent' has higher conversion rate therefore business should focus on generating more leads from this category.
- We can also see that customers who opens the email as their last activity also has higher conversion rate.

Imbalance in Data



- A free copy of mastering the data has high data imbalance and hence dropped

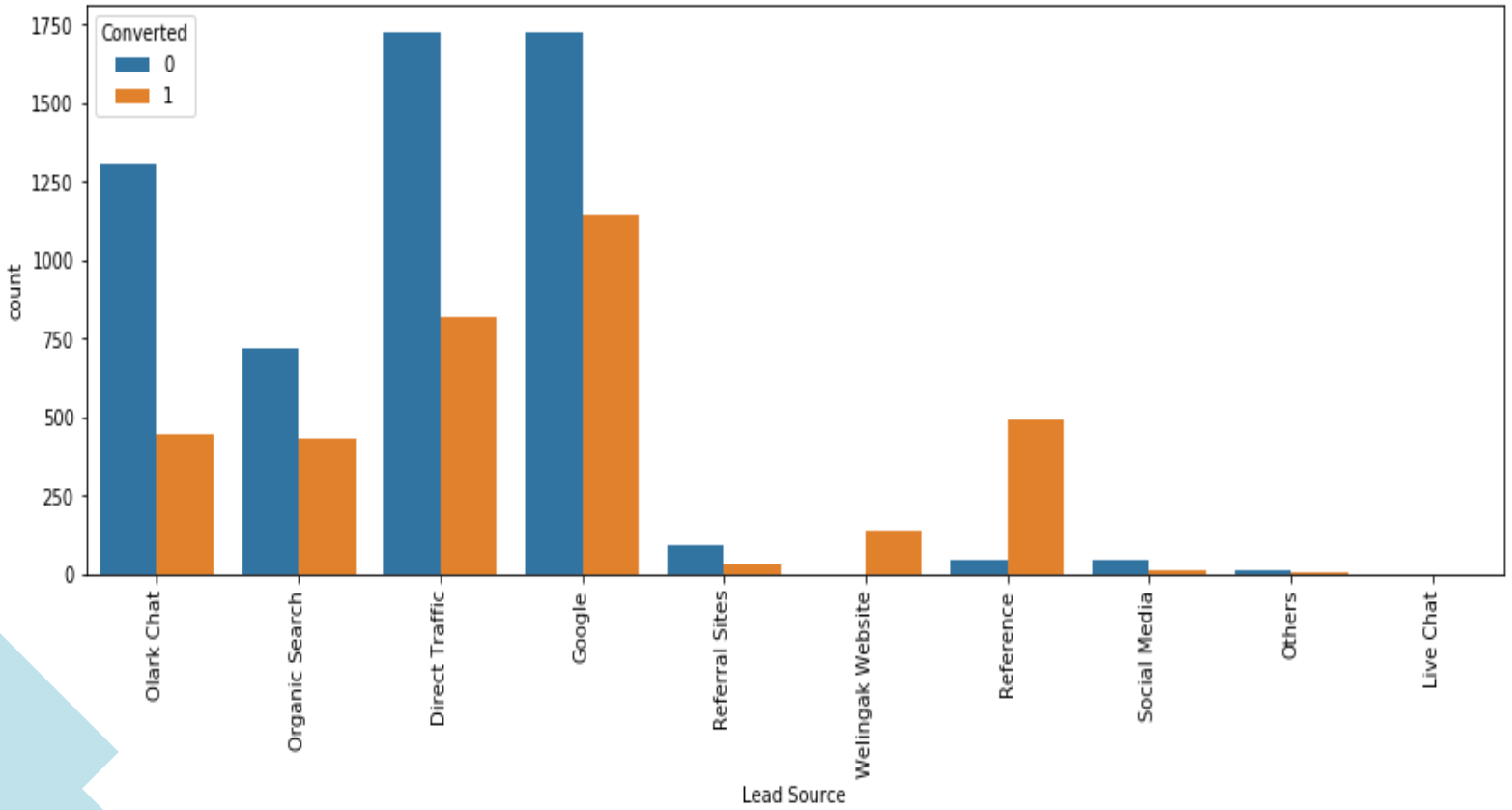


- We can also see that customers are not willing to pay amount through cheque and this attribute cannot be used for further inference.

Correlation

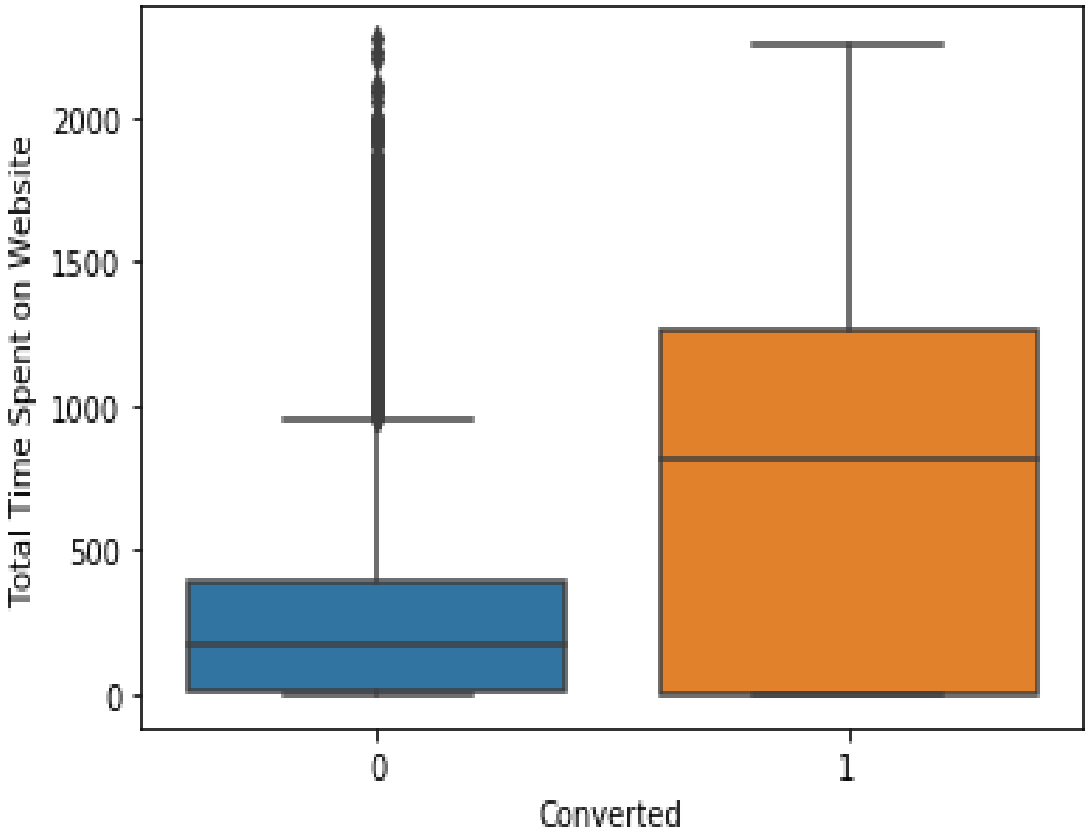
- We can see that there is some correlation between pages views per visit and total visit, which can be further analyzed using VIF.

Important variables and their Impact



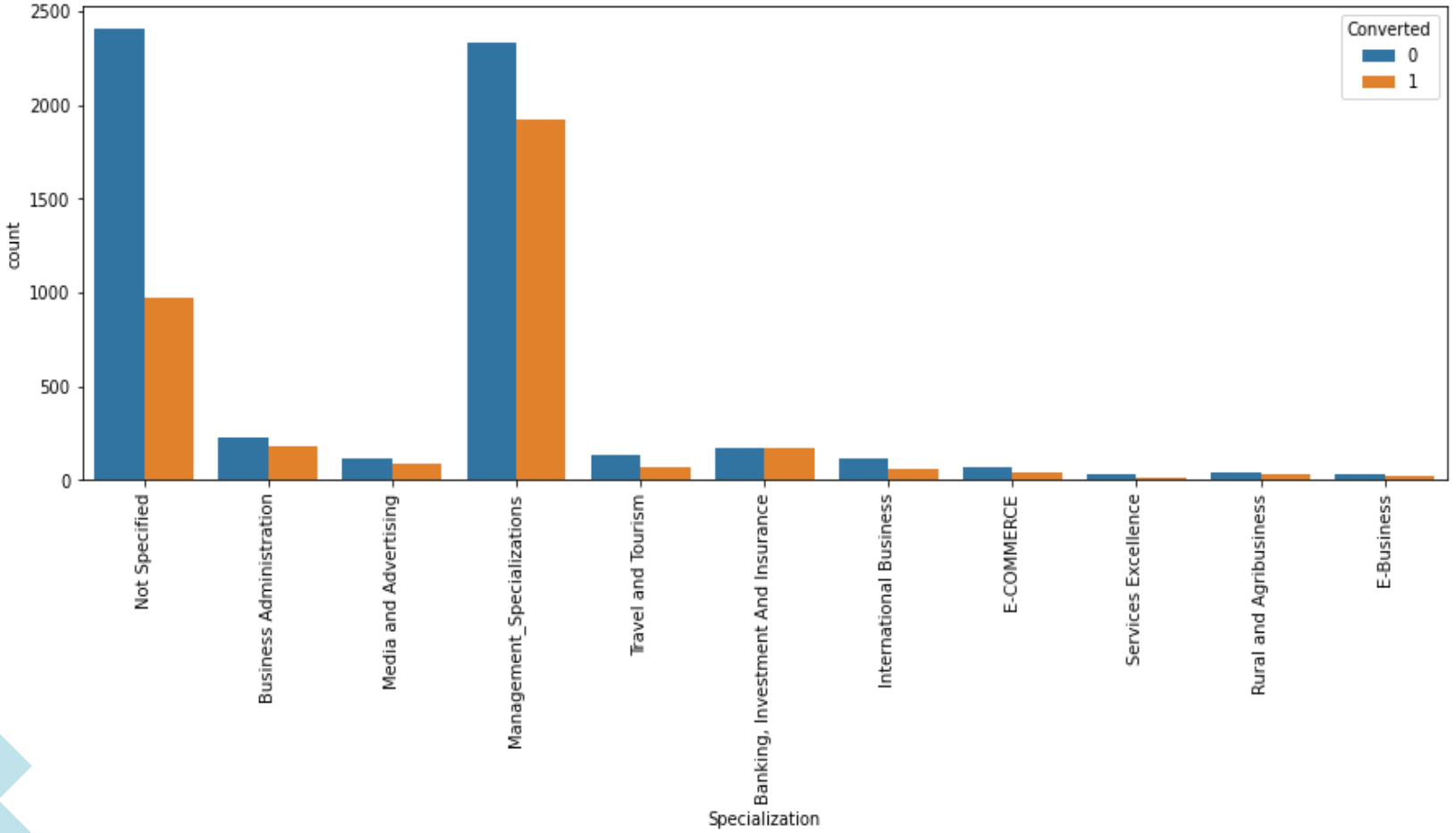
- Goggle and direct traffic has high number of leads but conversion rate is low.
- Welingak Website and reference has higher converted leads but total number of leads is low.
- Business should focus on improving conversion rate of leads from google and direct traffic. They should also try to get more leads from reference and welingak website.

Important variables and their Impact



We can clearly see that customers who spend more time on site have higher conversion rate. This can be because whoever is actually interested in enrolling into some course does there research and hence spends more time on the site. Thus the websites should be more engaging

Important variables and their Impact



•We can see that Management has highest number of leads as well as well as converted leads. This is a significant column and we cannot drop it.

•Business should focus more on customers of management specialization.

Model Building and evaluation

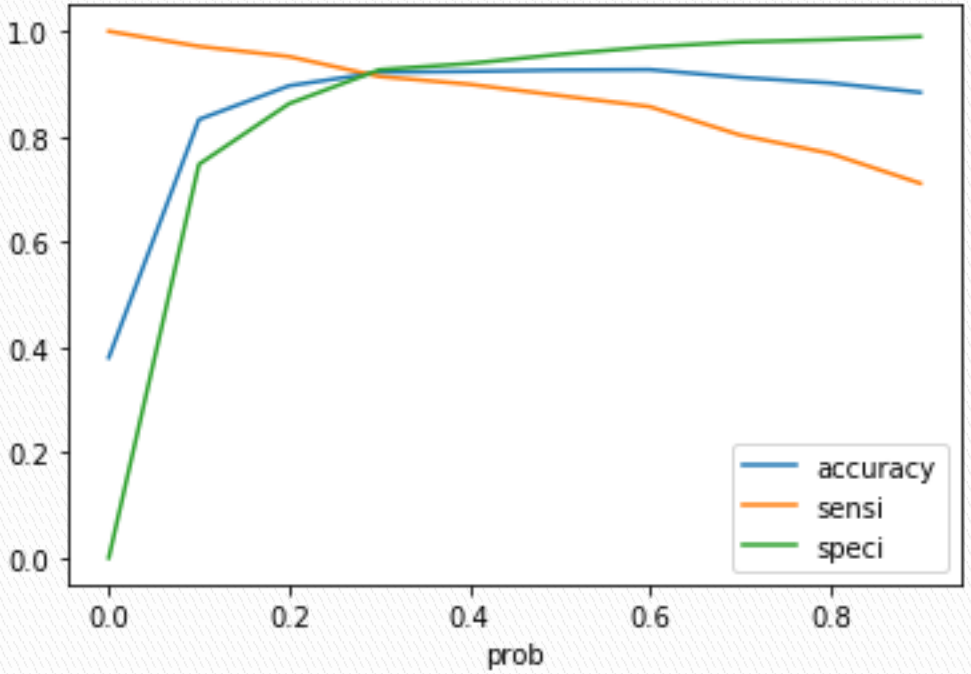
- The model was built using the stats model and RFE was done to attain the top 15 relevant variables through feature Selection. Later the rest of the variables were removed manually depending on the VIF values and p-value (The variables with VIF < 5 and p-value < 0.05 were kept).

- After plotting the ROC curve we found that our optimal threshold is 0.3. Then we calculated the metrics at this value. They are as follows: sensitivity: 91.41%, specificity: 92.68%, Accuracy:92.2%

- Prediction was done on the test data frame and with an optimum cut off 0.3. Accuracy: 92.45%, Sensitivity: 92.05%, Specificity: 92.69%

- As we can see the evaluation metrics are good when we run model on test data also. Model seems to predict conversion rate very well.

ROC Curve



Accuracy, Sensitivity and Specificity

- ROC curve is a trade off between True Positive Rate and False Positive Rate. A good ROC curve has a value close to 1. And our model's value of ROC is 0.97 which is a very good value.
- After plotting the ROC curve we found that our optimal threshold is 0.3. Then we calculated the metrics at this value. They are as follows:
- sensitivity: 91.41%
- specificity: 92.68%
- Accuracy:92.2%

Recommendations:

1.) Target leads who belong to management specialization as this category has high conversion rate.

2.) Target leads that spend a lot of time on X-education site i.e. total time spent is high.

3.) Target leads who are working professionals as the conversion rate is high.

4.) Focus on leads that come through reference.

5.) Avoid unemployed leads as they might not have the budget to enrol for the course.

6.) Avoid approaching student as they are already studying so they won't be interested but still they can be informed about the options they have for their future.

.