LEAD SCORING CASE STUDY

Sushmitha H R

Debasis Das Sharma

Dishu Diwan Jamnal

IITB upGrad DS C49

Introduction and Objective of the Case Study

Problem Statement

X Education is an education company, that sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google. Professionals, interested in the courses, land on their website and might browse the courses or fill up a form for the course or watch some videos. Any people wo fills up a form providing theis/her email address or phone number, is classified as a lead to the company. Moreover, the company also gets leads through past referrals. Once these leads are acquired, the Sales Team employees reach out to these leads, through phone calls, emails etc. Through this process, some of the leads get converted into actual paying customers.

*The typical lead conversion rate at X education is around **30%**.*

Inspite of a lot of leads being generated in the initial stage, only a few of them get converted. In the intermediate stage, the potential leads need to be nurtured well (i.e. educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion.

The **purpose of this case study is *to identify the most promising leads*** , i.e. the leads that are most likely to convert into paying customers.

We need to build a model wherein a lead score is assigned to each of the leads, so 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 ballpark of the target lead conversion rate has been set to be around 80%, by the CEO.

Solution Methodology

Data Sourcing , Cleaning and Preparation

- Read the Data from Source
- Convert data into clean format suitable for analysis
- Remove duplicate data
- Outlier Treatment
- Exploratory Data Analysis
- Dummy variable creation using OHE for modelling

Feature Scaling and Splitting Train and Test Sets

- Feature Scaling of Numeric data
- Splitting data into train and test set.

Model Building

- Feature Selection using RFE
- Determine the optimal model using Logistic Regression
- Calculate various metrics like accuracy, sensitivity, specificity, precision and recall and evaluate the model.

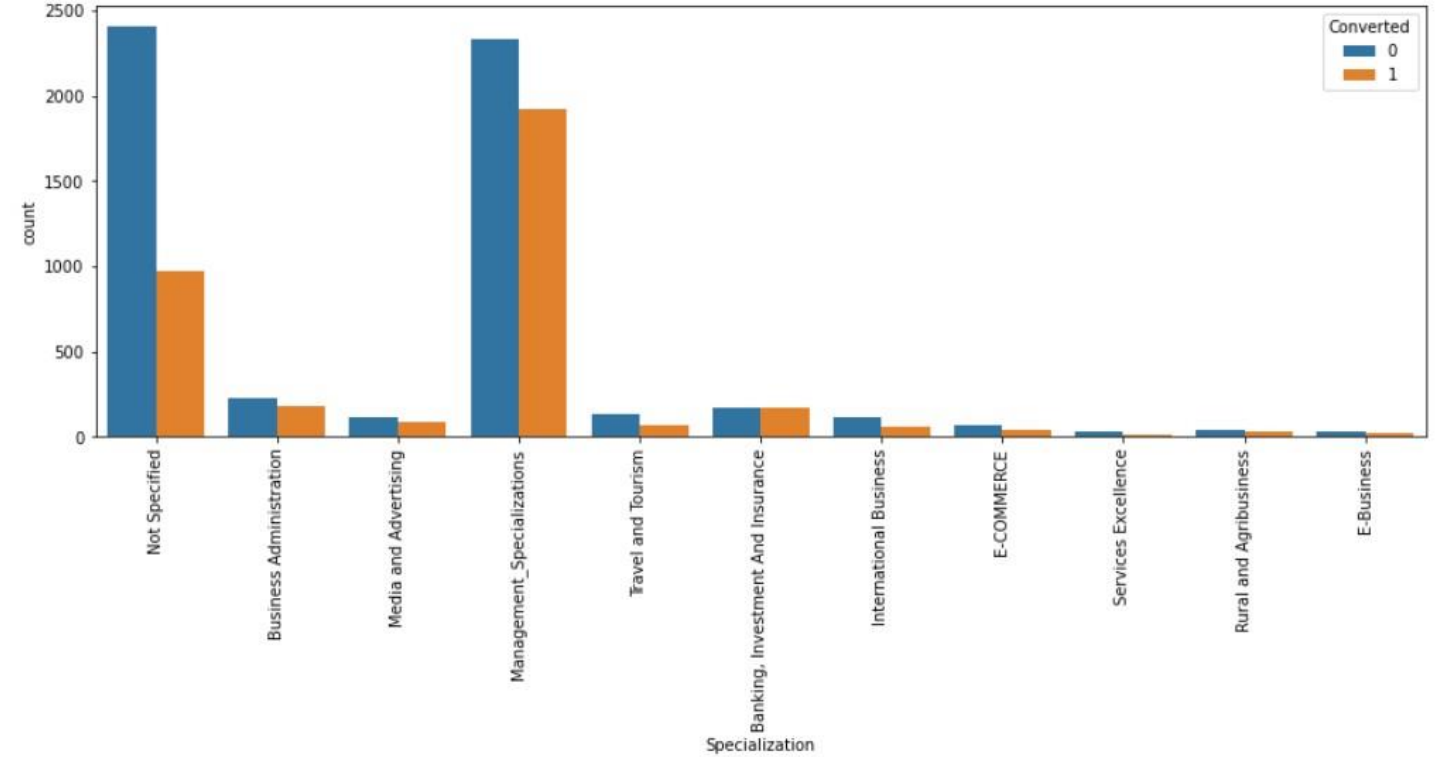
Result

- Determine the lead score and check if target final predictions amounts to 80% conversion rate.
- Evaluate the final prediction on the test set using cut off threshold from sensitivity and specificity metrics

Data Sourcing , Cleaning and Preparation

- Total Number of Rows =37, Total Number of Columns =9240.
- Data Cleaning:
  - Handling ‘Select’ values in some columns
  - Removing the “Prospect ID” and “Lead Number” which is not necessary for the analysis.
  - Dropping the columns having more than 40% as missing value
  - After checking for the value counts for some of the object type variables, we find some of the features do not have enough variance and were dropped, like, ‘Country’, ‘City’
  - Replacing NaN values with ‘Not Specified’ wherever it made business sense and combining low frequency values of some categorical variables together under ‘Others’
  - Single value features like “Magazine”, “Receive More Updates About Our Courses”, “Update me on Supply”, “Get updates on DM Content”, “I agree to pay the amount through cheque” etc. have been dropped.
- Outlier Treatment:
  - Remove top & bottom 1% of the Column Outlier values, wherever applicable
- Dummy variable creation of categorical variables for modelling
  - Splitting into binary variables for features with binary possible values
  - One Hot Encoding for variables with multiple options

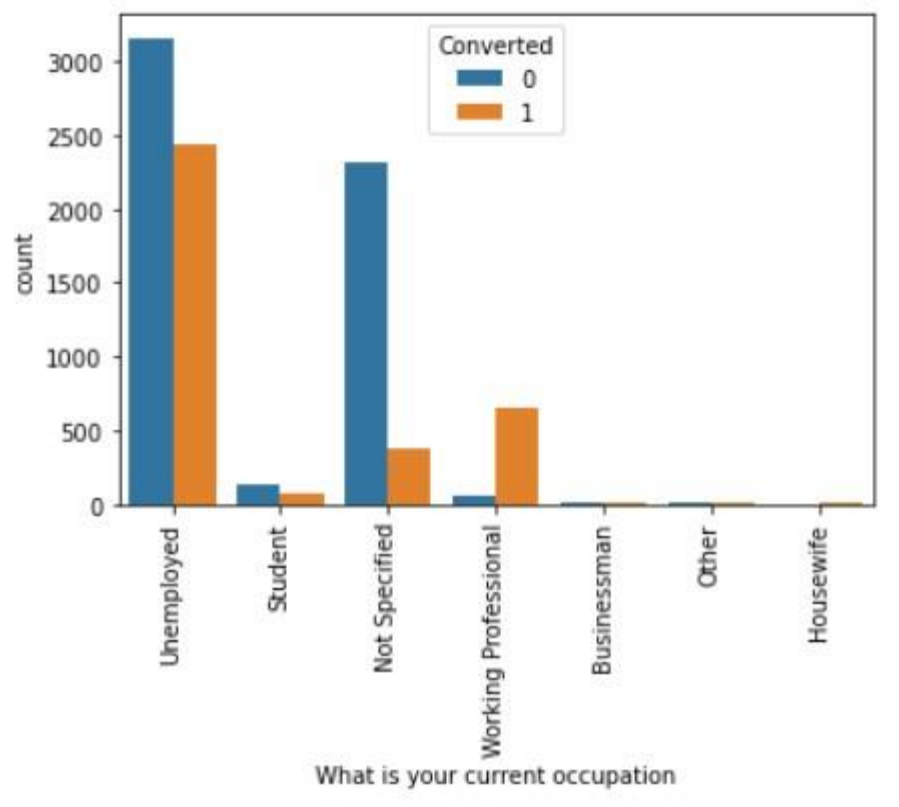
Exploratory Data Analysis



Inference

- A major section of leads have not specified the specialization. This may imply that the company is mostly being able to reach to customers who have no or bare minimum work domain knowledge.
- It may also imply that the other domains are not available in the options provided in th eform. Company may itself need to expand its options in other domains a
- Out of the specializations mentioned, a major percentage of leads come from Management sector and their conversion rate is quite high, but yet not as much as non-conversion. Comnpany may focus on this domain and try to apply better marketing stretegies to convince learners from Manangement related domain
- Banking, Investment and Assurance has a very minimum percentage of leads but the conversion rate among them is very high. Company can try to reach out to more people from this domain.

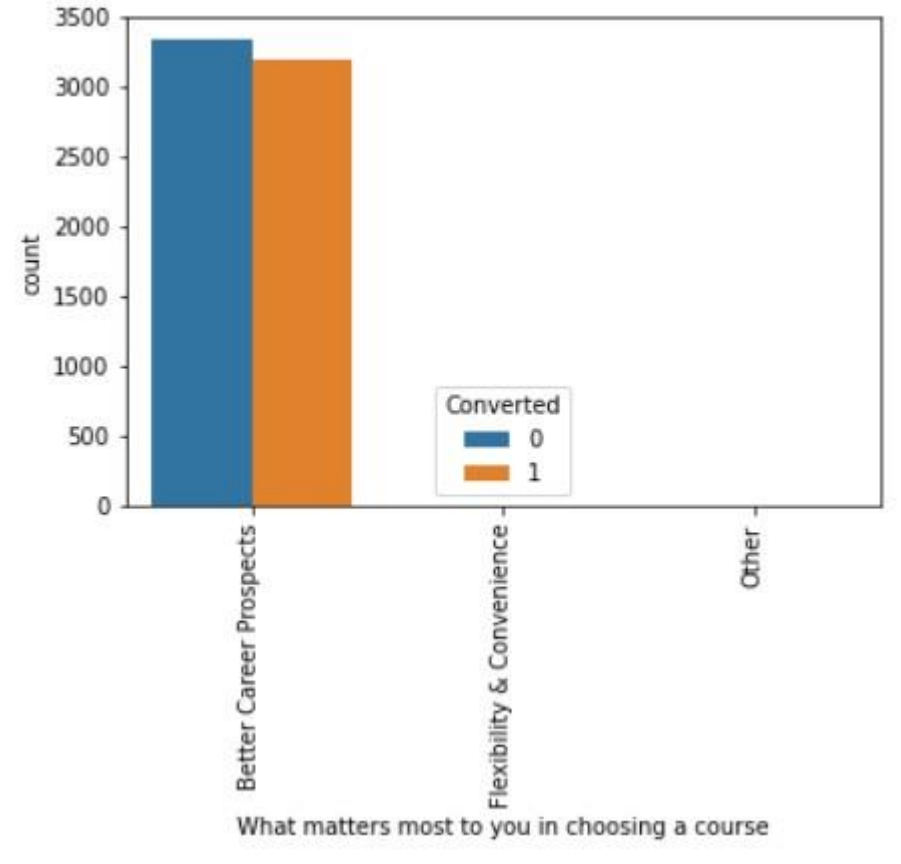
Exploratory Data Analysis



Inference

- Working Professionals going for the course have high chances of joining it..
- The highest number of leads are acquired from the Unemployed section, but their conversion rate needs to be significantly improved. Company can think of implementing promotional offers and discounts to engage the Unemployed and get them converted

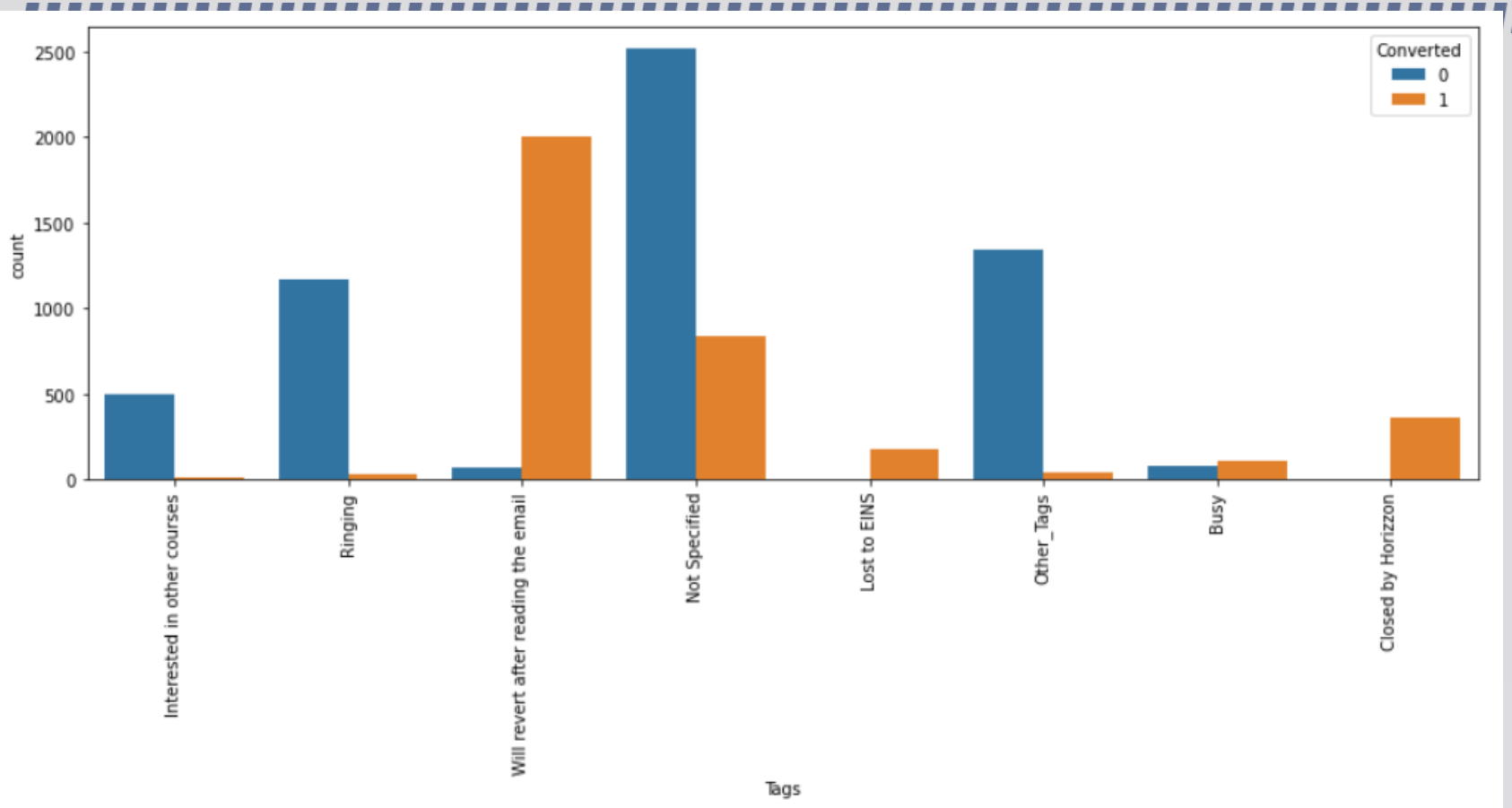
Exploratory Data Analysis



Inference

- A major 71 % of leads arelooking for a better job prspect, whereas a significant 29% of leads havenot specified anything at all. The job market being so volatile and competitive, the company can strategise to tie up with companies and promote job placement help features

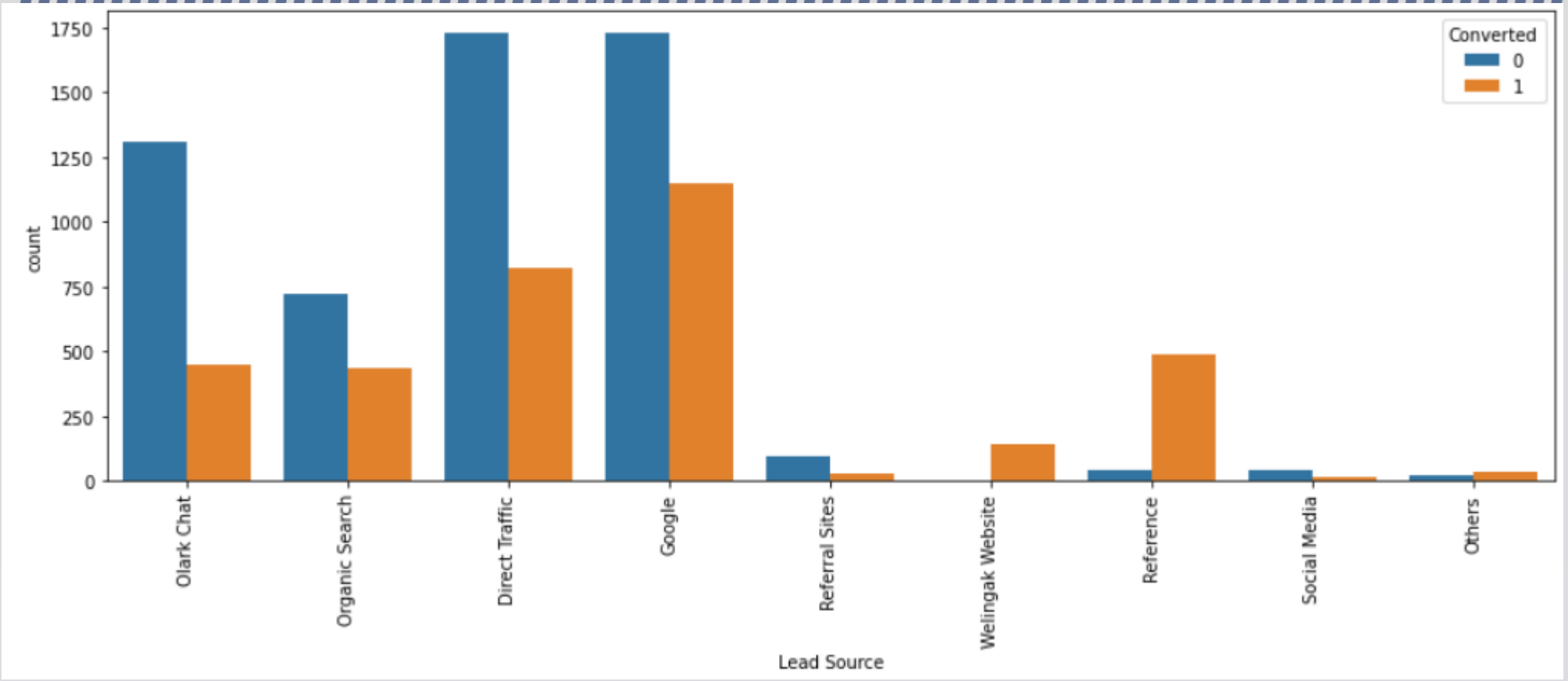
Exploratory Data Analysis



Inference

- Leads who committed to revert after reading the email, have a very successful conversion rate, with mostly around 95% of the leads being actually converted to paying customers. Email may be a preferred way of communicatio nof the offers and courses by the company
- 'Lost to EINs' and 'Closed by Horizzon' tags have very less number of leads but outof those, almost all converted. Company can focus on those channels

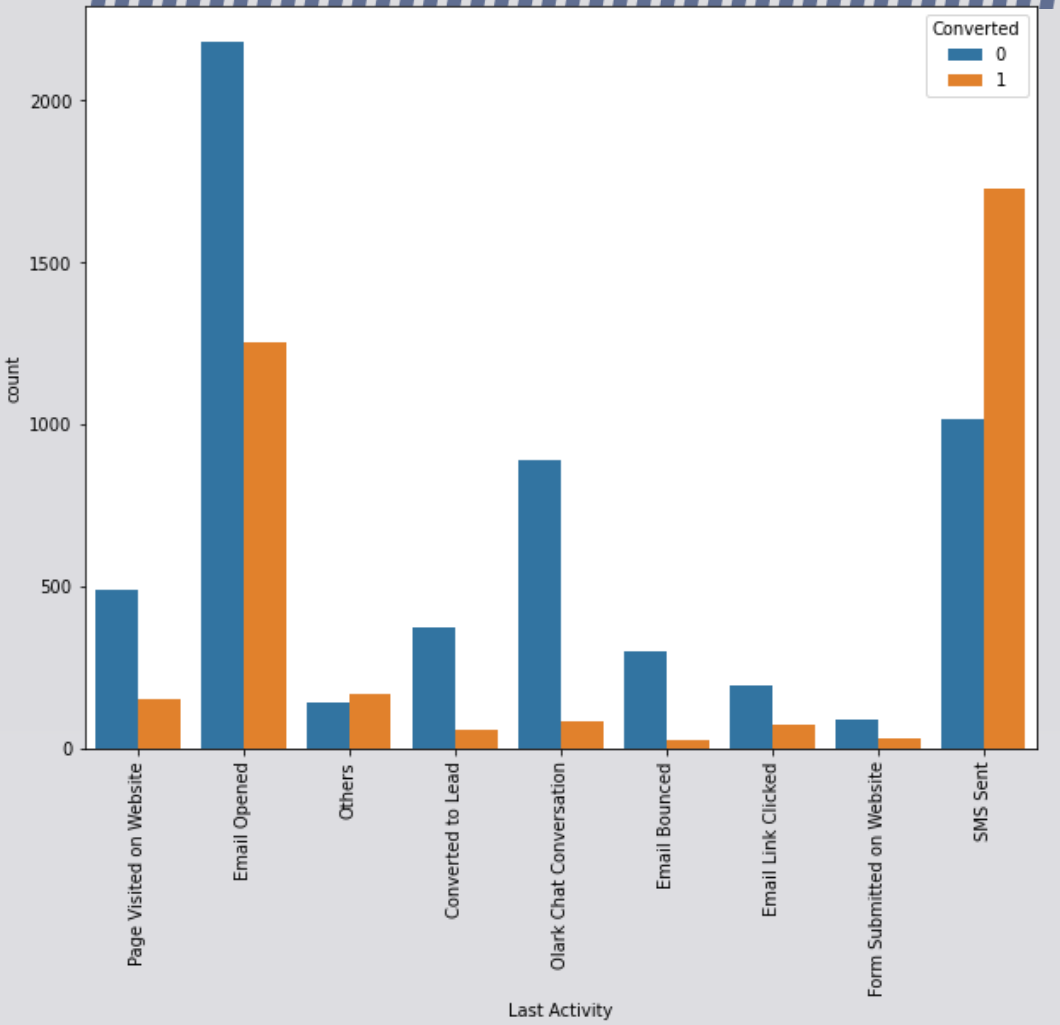
Exploratory Data Analysis



Inference

- Maximum number of leads are generated by Google and Direct traffic. But their conversion rate has to significantly improved. Clearly these are to most common channels to attract the prospective customers but company needs to strategise well to actually convert them to customers
- Conversion Rate of reference leads and leads through welingak website is high, but number of leads acquired is very low.

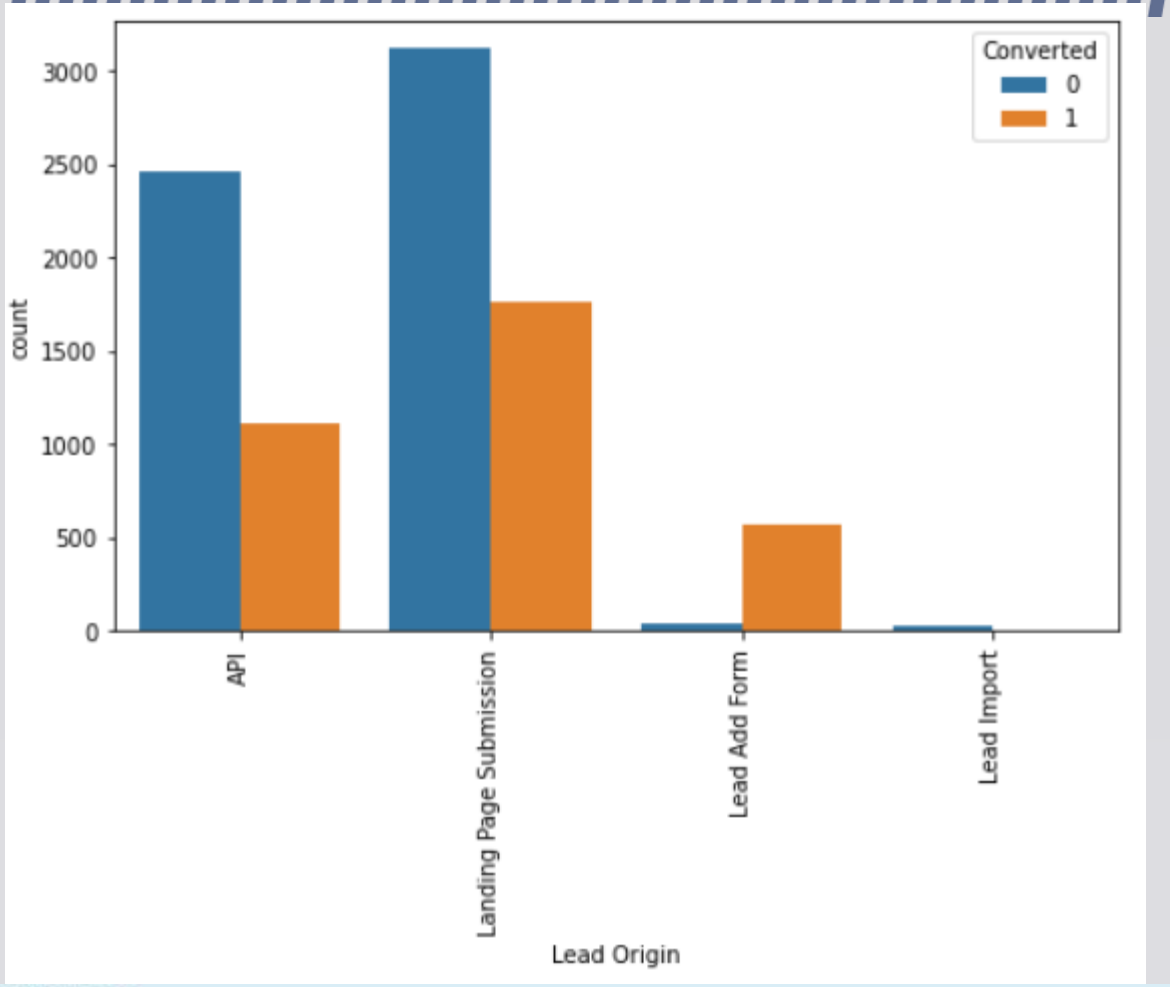
Exploratory Data Analysis



Inference

- Email and SMS are clearly the channels that attract most number of prospects.
- Interested folks are actually opening the emails. BUt after going through the details in the emails, the prospects are not convinced enough to get converted to customers.
- On the contrary, people going through teh details in SMS, are more likely to get converted.

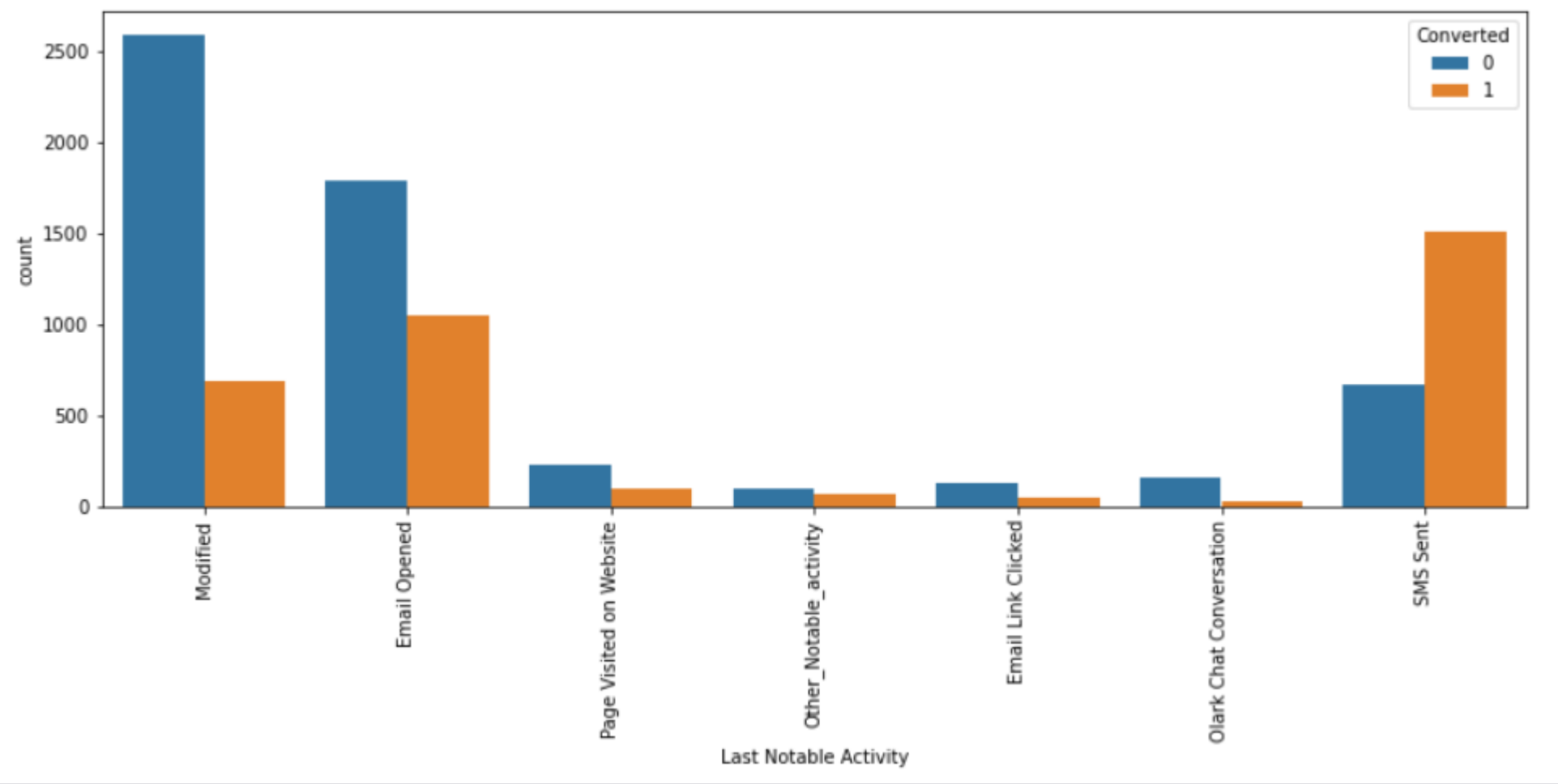
Exploratory Data Analysis



Inference

- API and Landing Page Submission bring higher number of leads as well as conversion.
- Lead Add Form has a very high conversion rate but count of leads are not very high.
- Lead Import and Lead Add Form get very few leads.
- In order to improve overall lead conversion rate, we have to improve lead converion of API and Landing Page Submission origin and generate more leads from Lead Add Form.

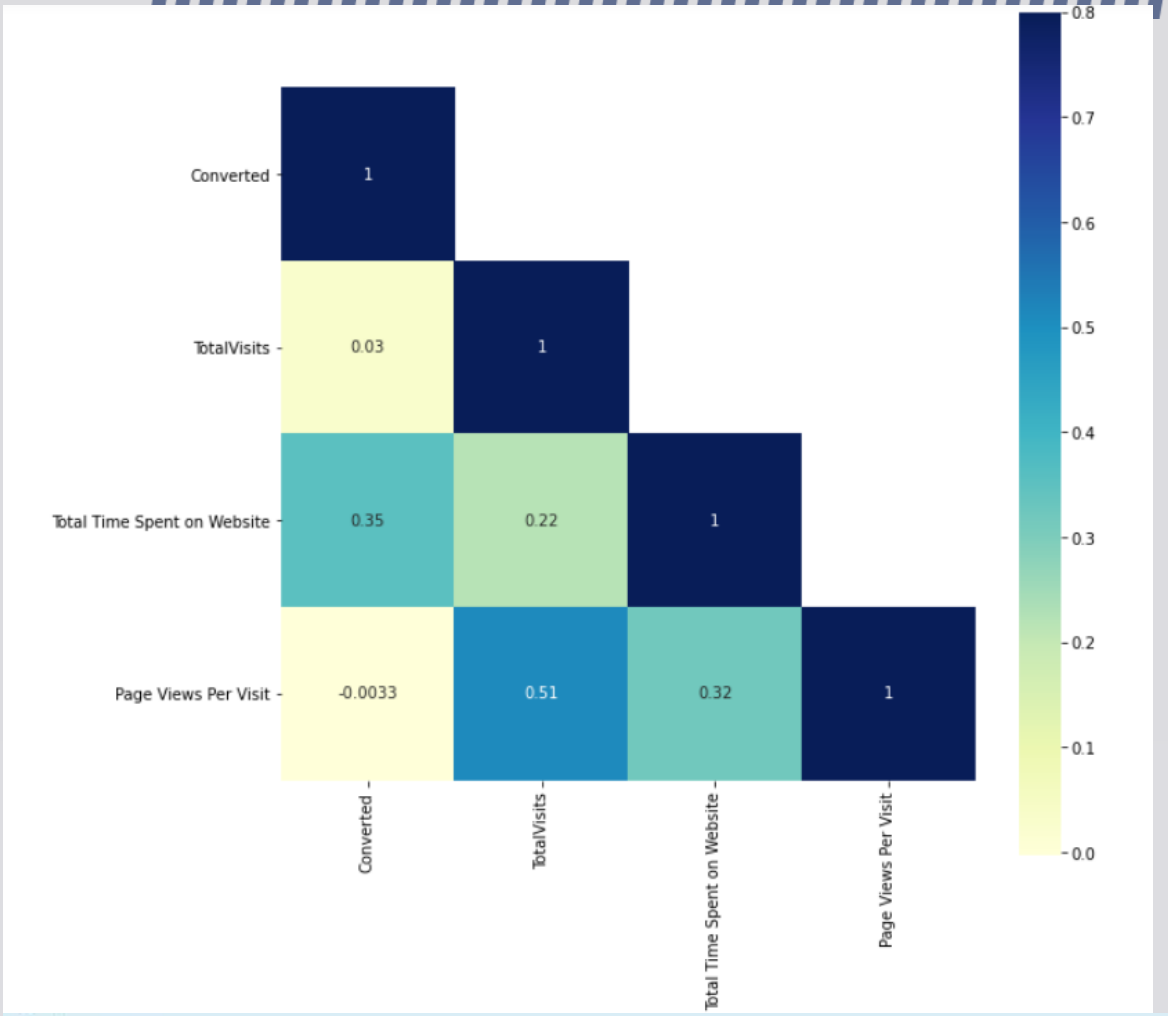
Exploratory Data Analysis



Inference

- The last noted activity of most leads are either modified, or opened an email or a SMS had been sent to them.
- We can once again observe that the highest successful conversion rate is among those to whom SMS was sent.

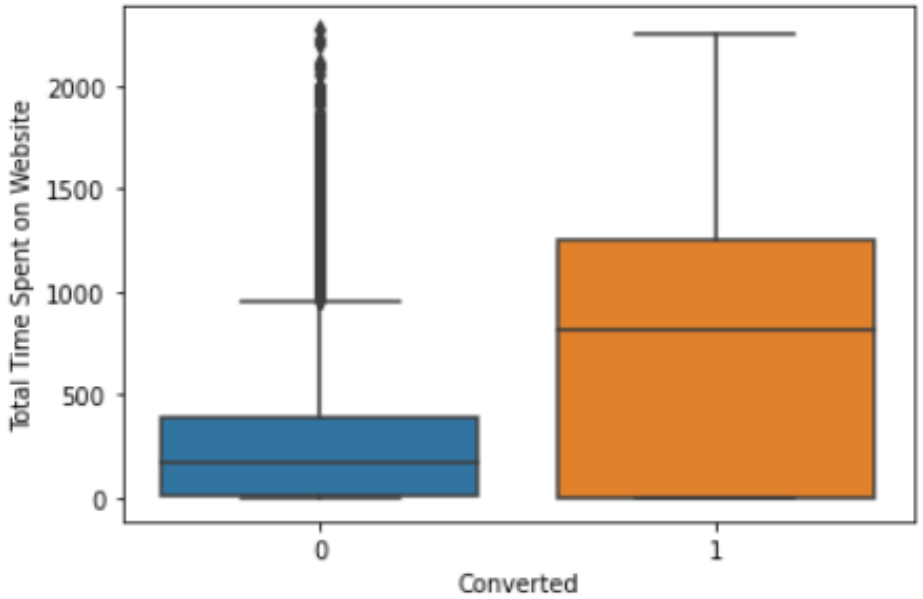
Exploratory Data Analysis



Inference

- We can observe that total time spent on website has a pretty significant relation with conversion, than other aspects
- Website being a good indicator, company may invest to make the website more lucrative to attract customers

Exploratory Data Analysis



Inference

We got a reconfirmation that Leads spending more time on the website are more likely to be converted. Website should be made more engaging to make leads spend more time./li>

Feature Scaling and Splitting Train and Test Sets

- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- We standardized the numerical variables using Standard scaler

Model Building

- Started with RFE for Feature Selection
  - Running RFE with 20 variables as output to identify the top 20 performing features
- Building Model by removing the variable whose p-value is greater than 0.05 and VIF value is greater than 5, one at a step
- Predictions on test data set
- Overall accuracy 92%

Model Building

- Started with RFE for Feature Selection
  - Running RFE with 20 variables as output to identify the top 20 performing features
- Building Model by removing the variable whose p-value is greater than 0.05 and VIF value is greater than 5, one at a step
- Predictions on test data set
- Overall accuracy 92%

Model Evaluation on Train Set

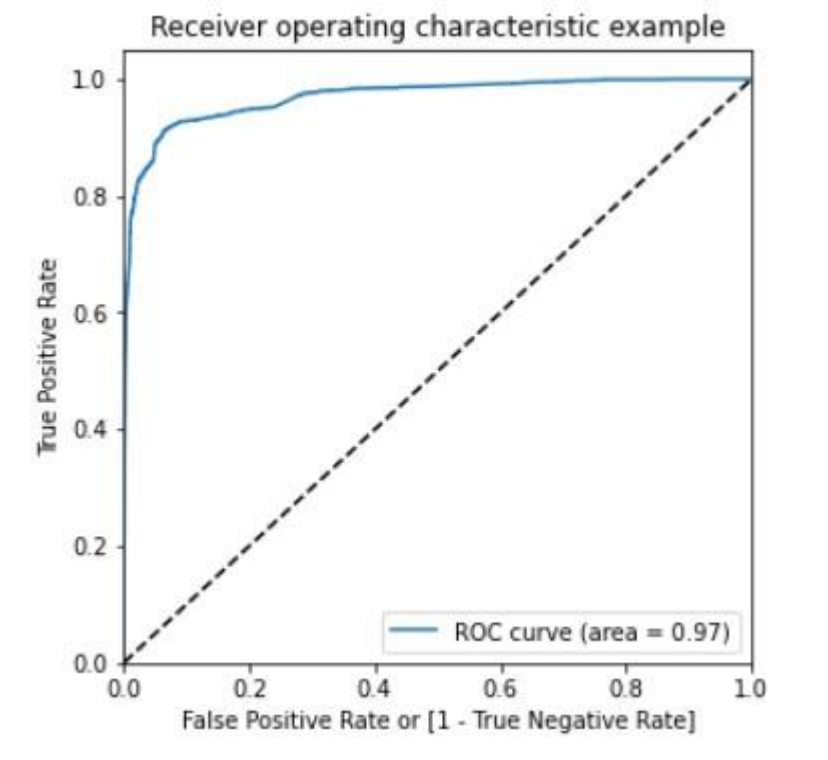
Confusion Matrix

3705	177
336	2049

Evaluation Statistics

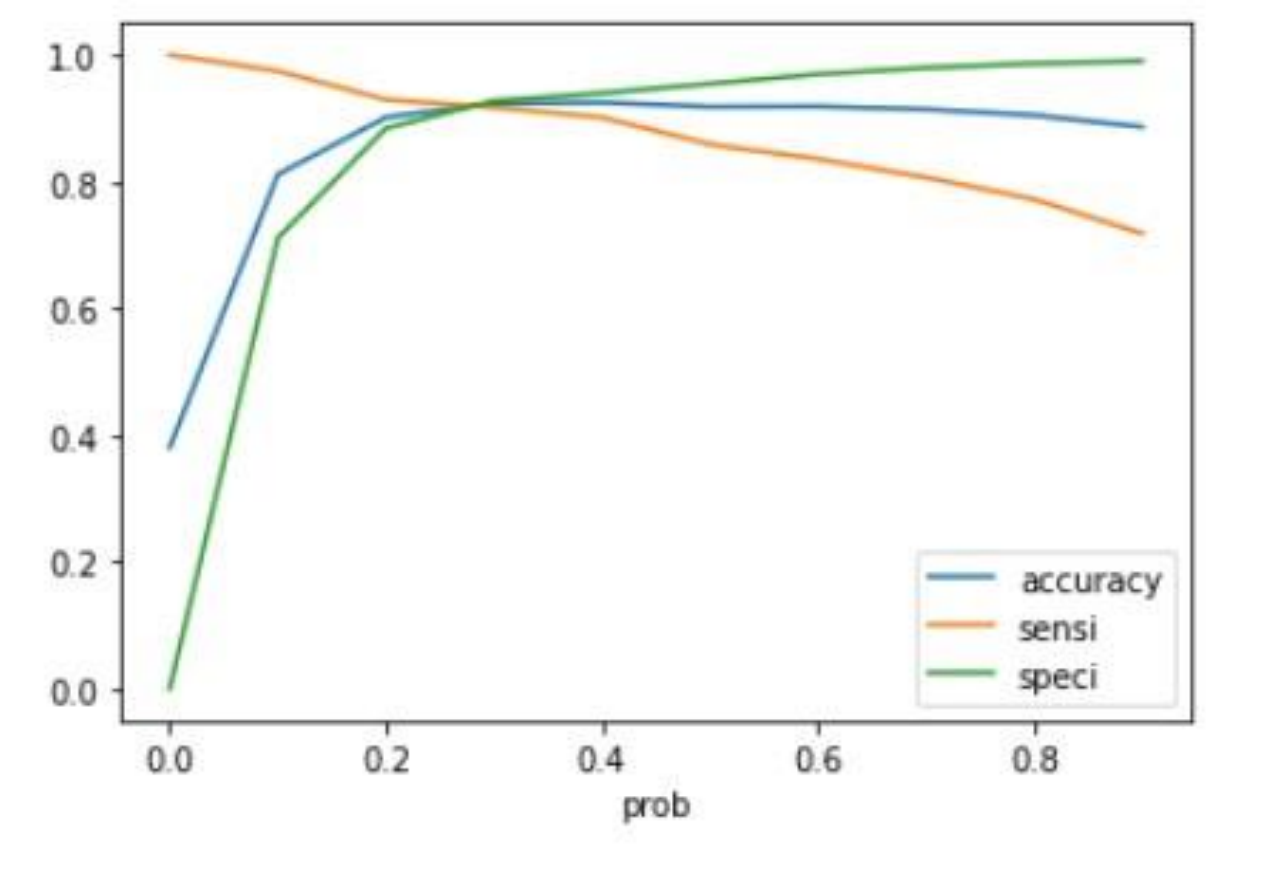
Accuracy	0.92
Sensitivity	0.86
Specificity	0.95
False Positive Rate	0.05
Positive Predictive Value	0.92
Negative Predictive Value	0.92

ROC Curve



The ROC Curve should be a value close to 1. We are getting a good value of 0.97 indicating a good predictive model.

Finding Optimal Cutoff Point



The graph depicts an optimal cut off of 0.3 based on Accuracy, Sensitivity and Specificity

Final Stats on the Train Set using Optimal Threshold

Confusion Matrix

3601	281
198	2187

Evaluation Statistics

Accuracy	0.92
Sensitivity	0.92
Specificity	0.94
False Positive Rate	0.07
Positive Predictive Value	0.88
Negative Predictive Value	0.94
Precision	0.88
Recall	0.92

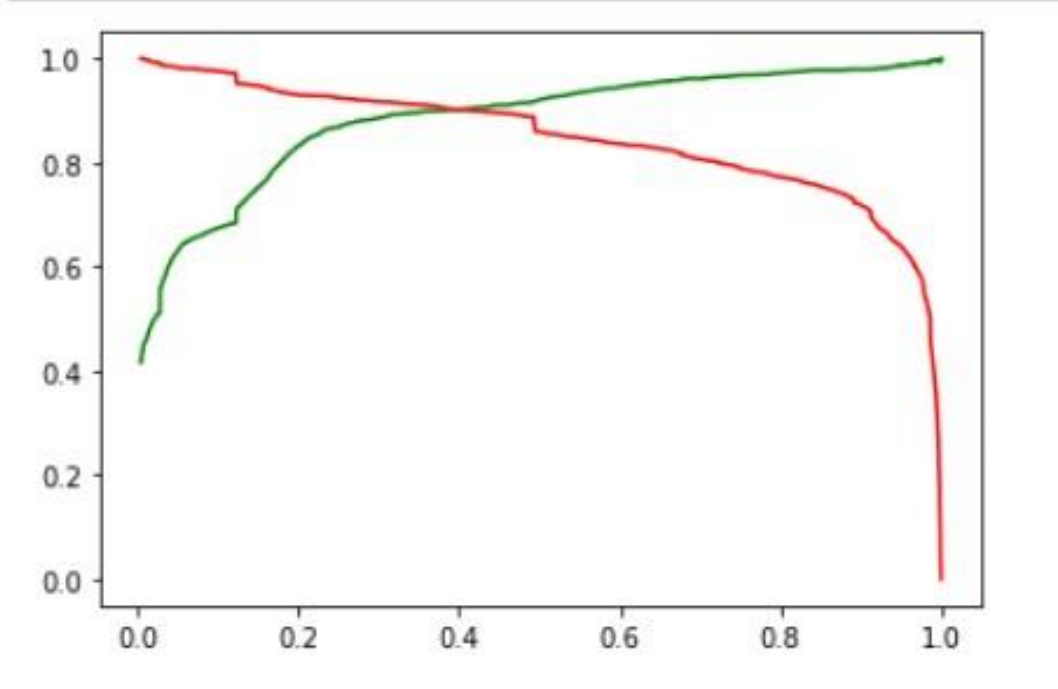
Final Stats on the Train Set using Optimal Threshold

Confusion Matrix

3601	281
198	2187

Evaluation Statistics

Accuracy	0.92
Sensitivity	0.92
Specificity	0.94
False Positive Rate	0.07
Positive Predictive Value	0.88
Negative Predictive Value	0.94
Precision	0.88
Recall	0.92



The graph depicts an optimal cut off of 0.4 based on Precision and Recall

Prediction on Test Set

Confusion Matrix

1570	106
85	925

Evaluation Statistics

Accuracy	0.92
Sensitivity	0.92
Specificity	0.93
Precision	0.89
Recall	0.92

The Model seems to predict the Conversion Rate in both train and test sets very well and we should be able to give the CEO confidence in making good calls based on this model.

Final observations

Top Three Variables

Based on the coefficient values from the final model, the following are the top three variables that contribute most towards the probability of a lead getting converted :

- Tags_Closed by Horizzon
- Tags_Lost to EINS
- Tags_Will revert after reading the email

There is a high conversion rate on leads closed by Horizzon or lost to EINS, but the total number of leads acquired through these channels, are pretty low. Company should focus on reaching more prospects here. Also prospects who have committed to revert after reading the email, are more likely to convert to actually paying customers. So the company can follow up on them in a more diligent way.

Final observations

Other Observations and Suggestions

From the EDA steps , we have also observed that :

- There is a very high conversion rate of leads coming through Welingak website, but again the number of leads acquired here is low. Company needs to promote more here
- A huge portion of leads come from the Unemployed section with a not so great conversion rate. Company can think of providing discounts or scholarships to lure them
- Working professionals have a good chance of conversion. Company can try to increase the revenue form this channel as well
- Almost everyone opting for the courses are doing so, for better job prospects and many of them are unemployed too. So good job placement aid will facilitate the conversion
- People spending more time on the website, are more likely to get converted. So the website should be made more engaging and informative.
- SMS and emails are the most successful ways of communication and should be prioritized

Final observations

Business Strategies:

In seasons when the company hires interns and has high manpower and aims to make the lead conversion more aggressive, the company may contact all the leads which have a conversion probability (value = 1) under a cut off 0.3 . From business knowledge perspective, to achieve maximum conversion, phone calls must be done to the all the people with a lead score from 40 to 100 and primarily if:

- They spend a lot of time in the website and this can be done by making the website interesting and thus bringing them back to the site.
- They are seen coming back to the website repeatedly
- Their last activity is through SMS or through Olark chat conversation
- They are working professionals

Similarly, when the company reaches its target for a quarter before a deadline and wants to focus on some new work, they need to minimize the rate of useless phone calls. In this case, they can choose a higher threshold value for Conversion Probability and may contact all the leads which have a conversion probability (value = 1 highlighted in yellow color) under column 0.7. However, the flipside here would be that, we may miss out on those leads that are actually converted but then the model wrongly predicted them as not converted. From the business perspective, the company need to focus more on hot leads with probability score higher than 90 and other methods like automated emails and SMS. This way manual calling won’t be required unless it is an emergency. The above strategy can be used but with the customers that have a very high chance of buying the course.

Thank You