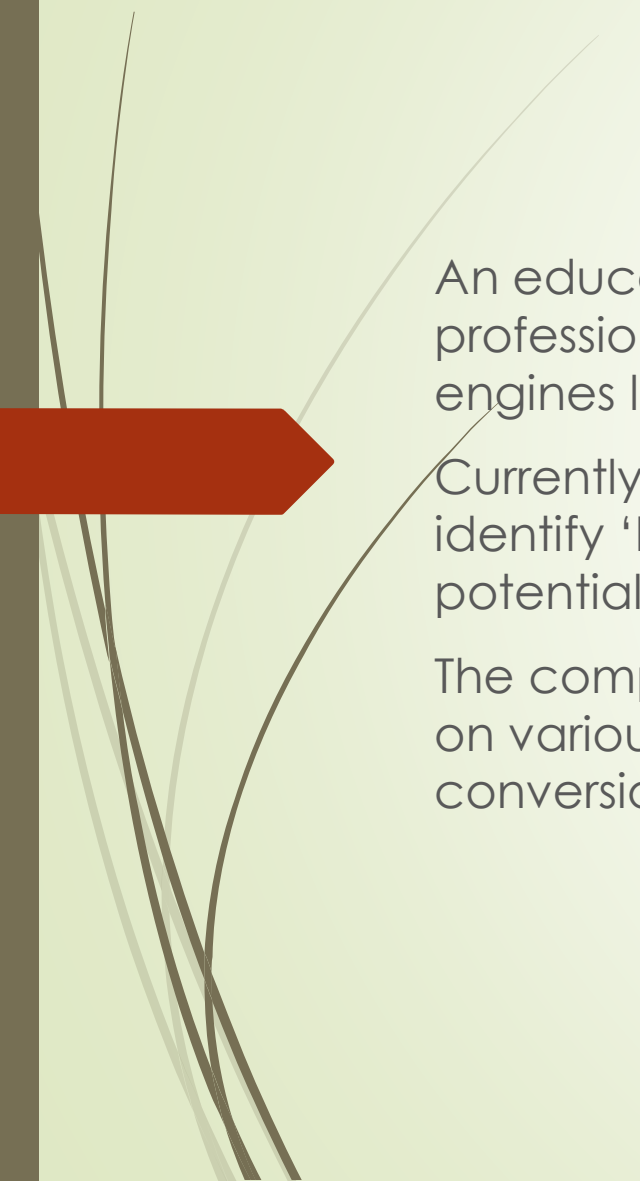# Lead Scoring Case Study

Dhruv Pathak

Nikita Vij

Rahul Kumar Tandon

Problem Statement

An education company named X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google. The company also gets leads through referrals.

Currently the lead conversion rate for the company is ~30%. The company wished to identify 'Hot Leads' so that the sales team can focus on communicating with these potential leads so as to increase the conversion rate.

The company requires a model wherein each lead is assigned a Lead Score based on various parameters and such that a lead with higher score would have higher conversion chances and the team can focus on those leads.

Approach Followed

1. **Reading and Understanding Data** : Thorough understanding of each variable contributing towards interpreting 'Hot Leads'.

2. **Data Cleaning** : Handling of Missing data, Imbalanced data, Imputation of missing values, removing less valuable features to build a light model.

3. **Exploratory Data Analysis** : Performing Univariate and Bivariate Analysis, Locating and handling any anomalies.

4. **Data Preparation** : Creating dummy variables, train-test split, feature scaling.

5. **Model Building** : Building a Logistic Regression Model using both coarse (RFE) and fine (Manual elimination) tuning on the data.

6. **Model Evaluation** : Testing out the final model on the test data and evaluating various important metrics.
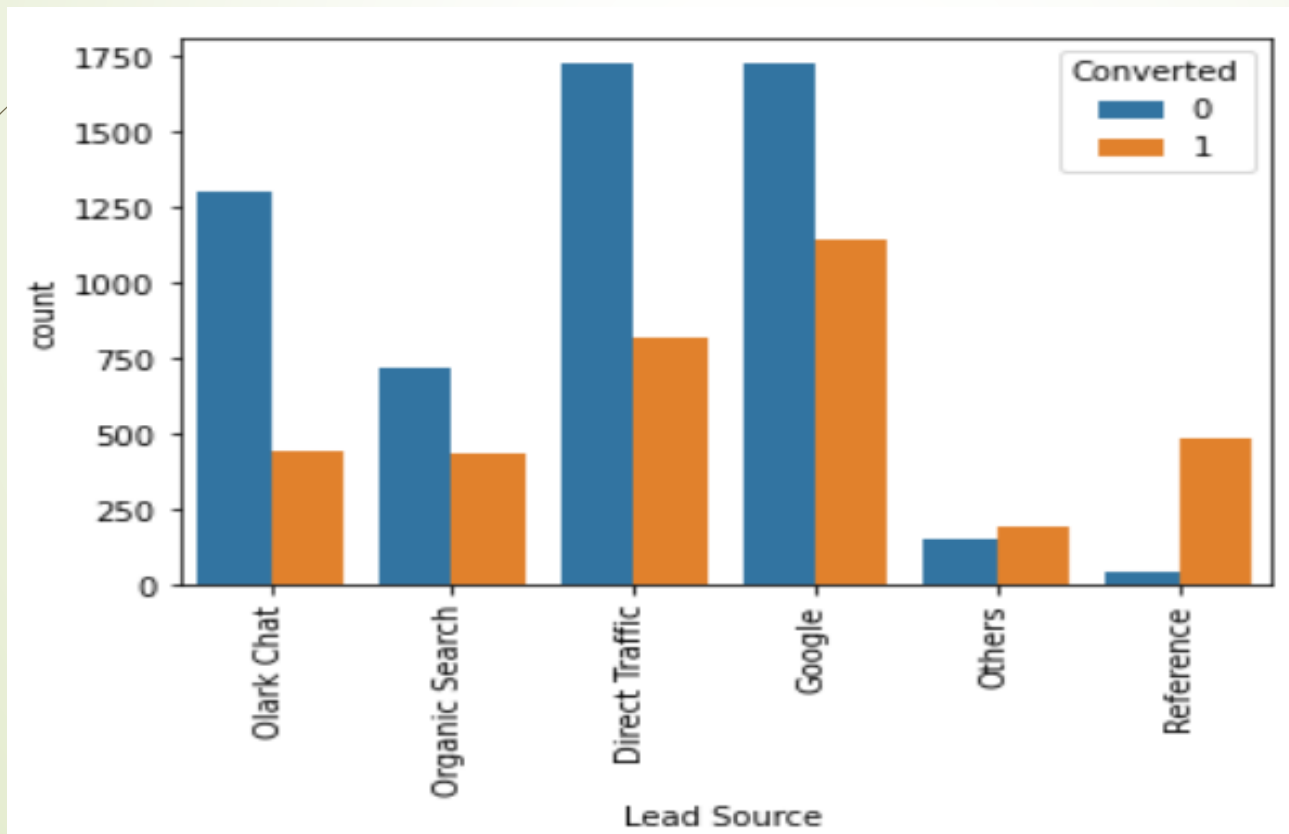
Important outcomes

LEAD ORIGIN :

Maximum leads are originating through API and Landing Form Submisssion however leads originating from Add Forms are potential leads.
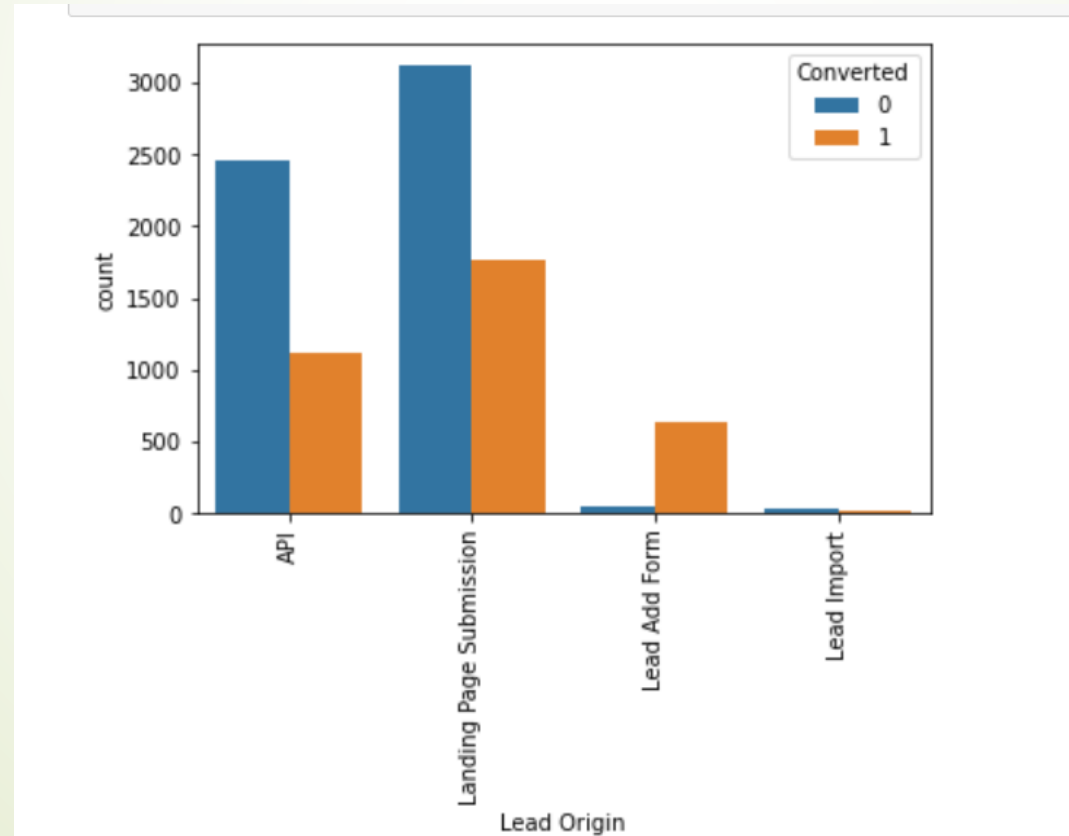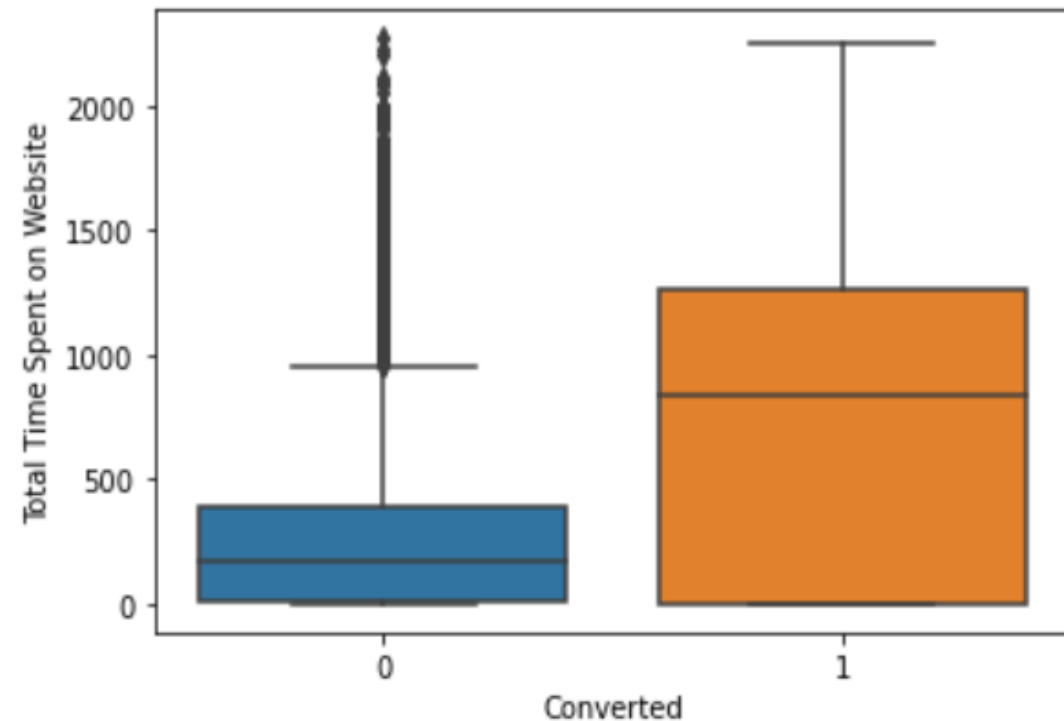
Important outcomes

LEAD SOURCE :

Maximum leads are generated through Google and Direct Traffic but the conversion is not as high as through past referrals.

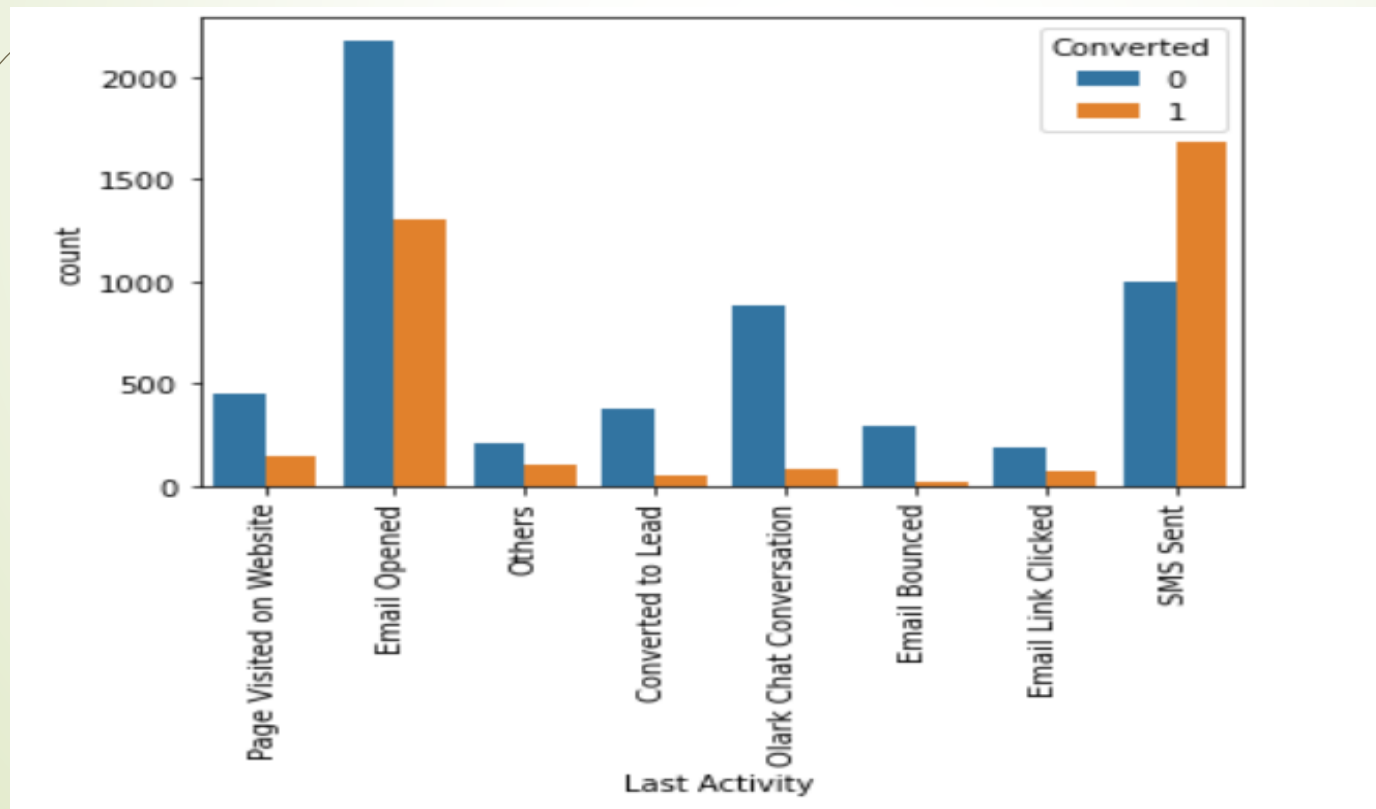# Important outcomes

TOTAL TIME SPENT ON WEBSITE :

It has been observed that people who are spending a considerable time on the company's website are highly likely to get converted into paying customers.
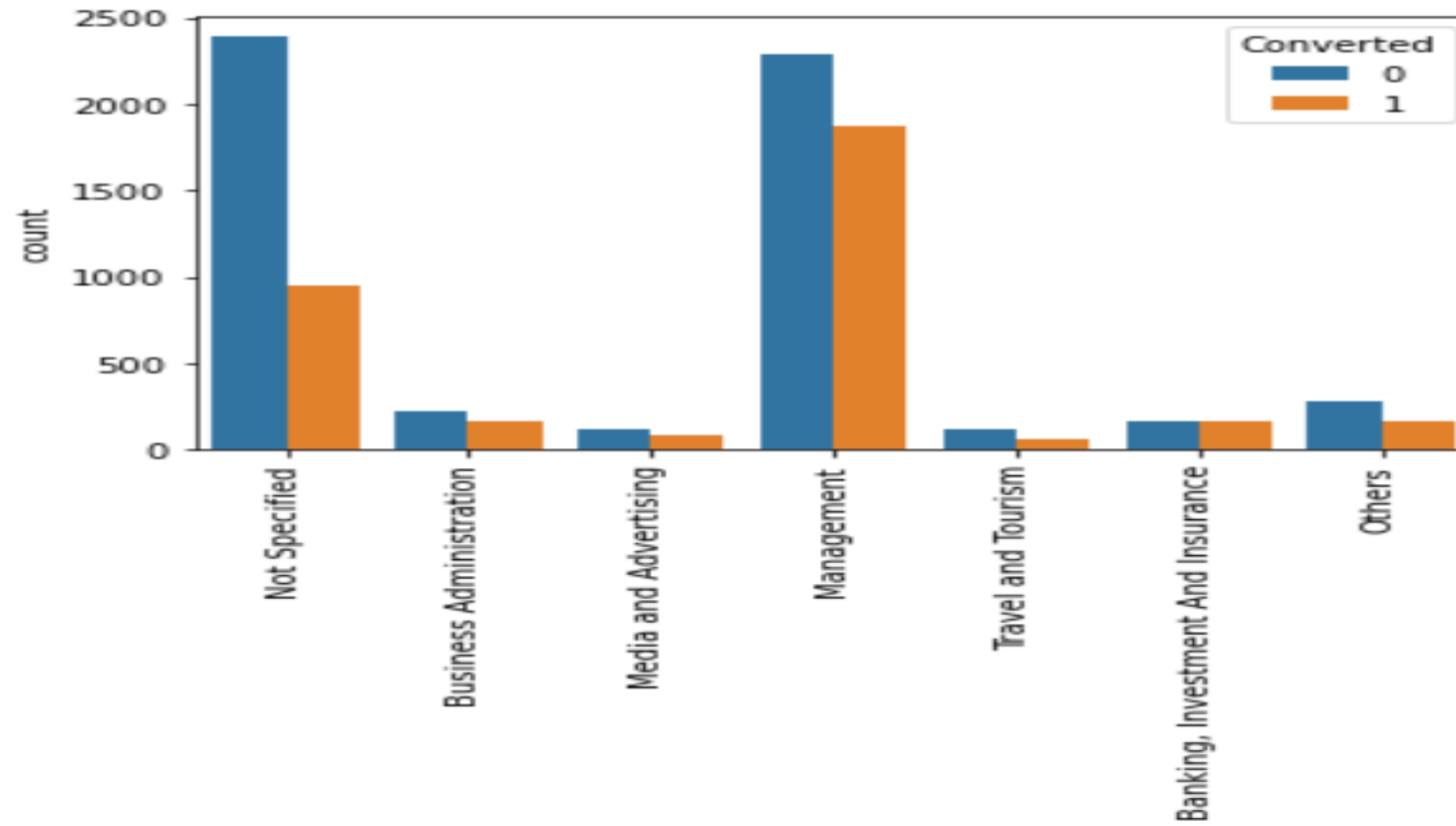
Important outcomes

LAST ACTIVITY :
It has been observed that leads with last activity marked as Email Opened and SMS Sent have converted the most in the past.
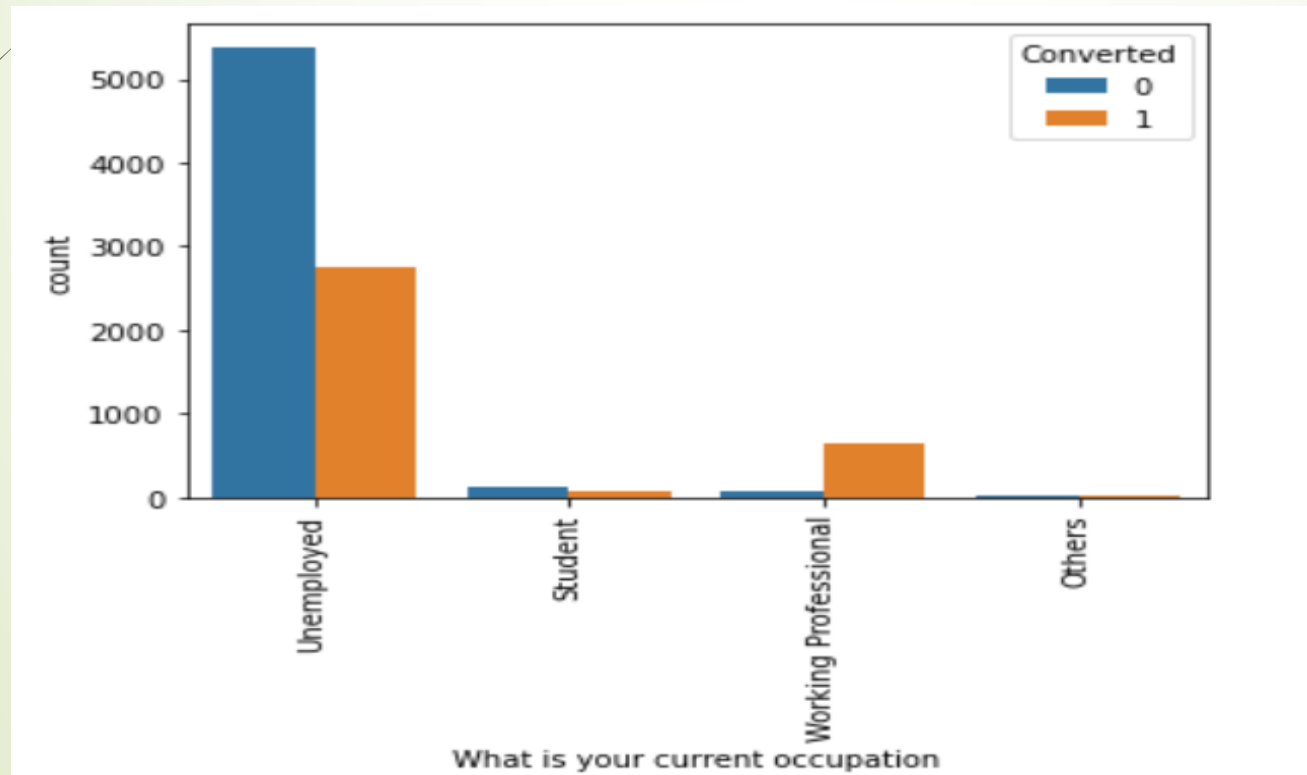
Important outcomes

SPECIALISATION :
Management sector(Finance, HR, Marketing, Operations, IT, Healthcare, Retail) has a considerable number of leads generated and also consists of Potential Leads.
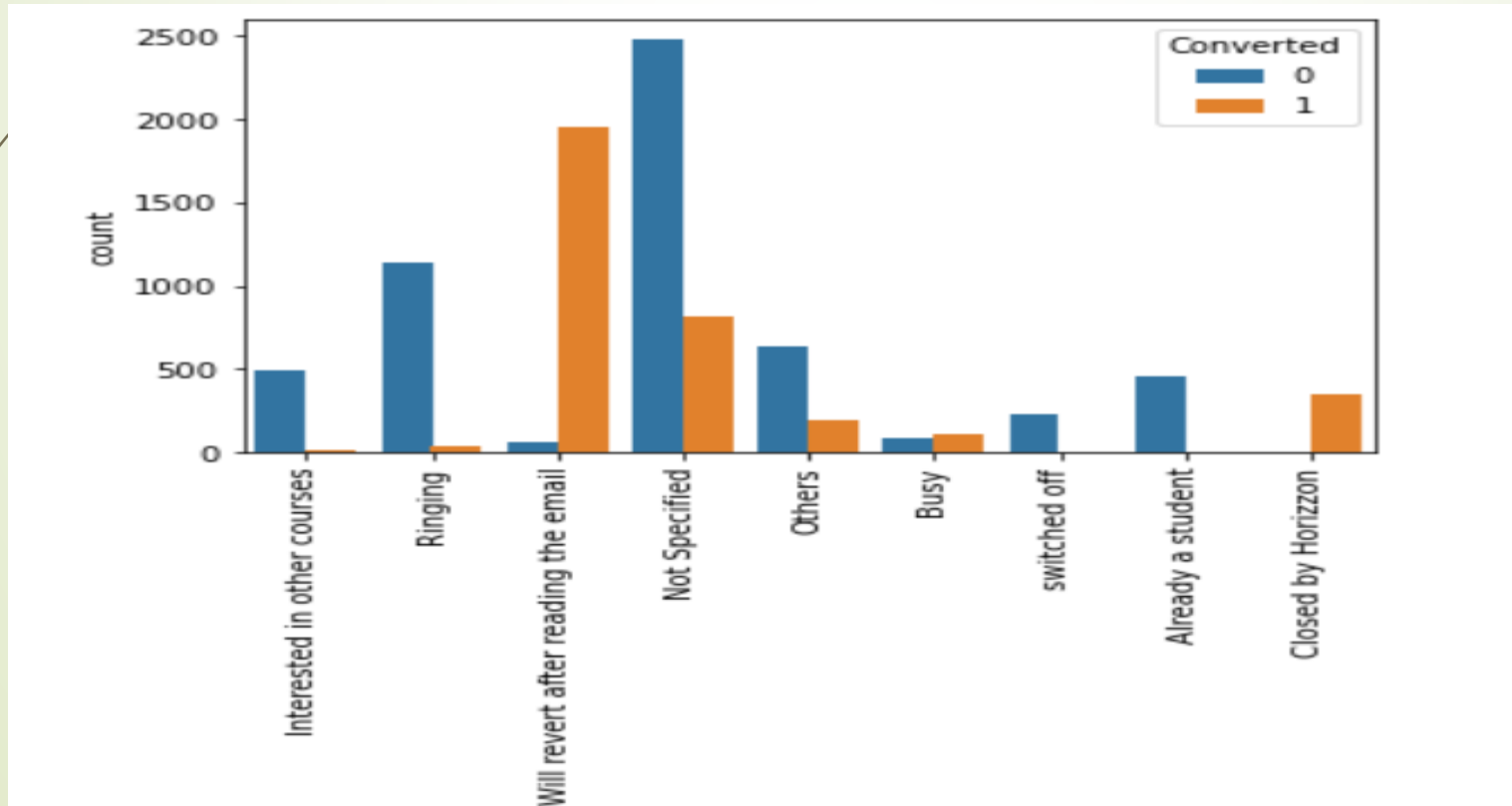
Important outcomes

CURRENT OCCUPATION :
There is a lot of traffic generating from the Unemployed sector but they do not turn out to be potential customers. On the other hand Working Professionals seem to be doing well for the organisation.

Important outcomes

TAGS ASSIGNED TO LEADS :
Maximum potential leads have been observed where the tag assigned to the lead were 'Will revert after reading the email' and 'Closed by Horizzon'.

Important outcomes

Based on the features mentioned previously the Logistic Regression Model was built with ROC Curve value of 0.96 and with an optimal cut-off value of 0.28.

Receiver operating characteristic example
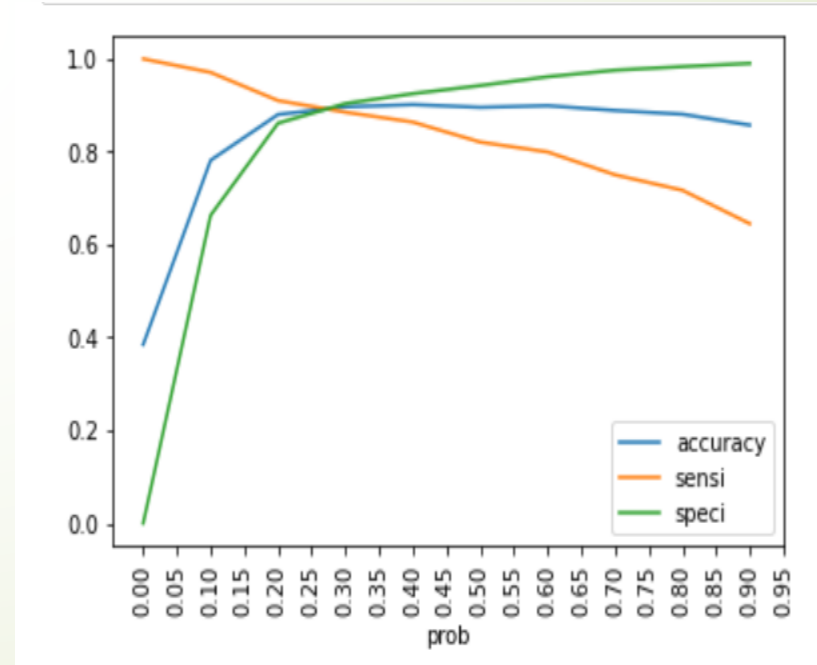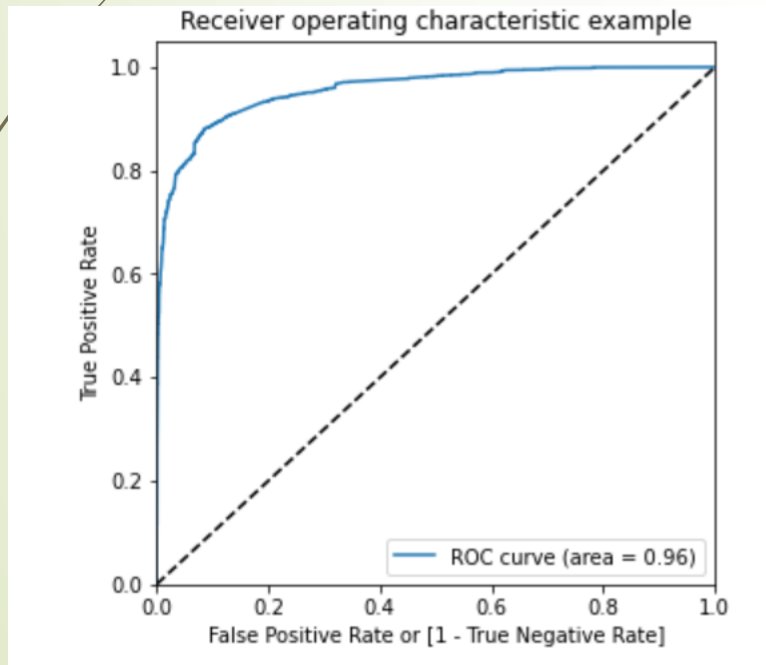
Important Metrics Values

TRAIN DATA:
1. Accuracy – 89.5%
2. Sensitivity – 88.8%
3. Specificity – 89.9%

TEST DATA:
1. Accuracy – 90%
2. Sensitivity – 88.6%
3. Specificity – 90.8%

Inferences:

1.The organisation should focus on improving their website both UI and UX wise. They can also make it more informative as it has been observed that leads who spend a lot of time on the website are hot leads.

2. Leads generated through past referrals are hot leads.

3. Leads with Tags : Email Opened, Will revert after reading the mail and SMS Sent seem to resulting in a conversion.

4. Leads closed by horizzon have a very strong chance of actually becoming a customer.

5. The company should focus more on Working Professionals specially at manager levels as there is a high chance of them getting converted as they could be looking for a boost in their current role.

6. Leads generated by Add Forms are also hot leads.