Lead Scoring Case Study

By:- Parishi Gupta
Amanjot Singh
Parteek Malik

Problem Statement:-

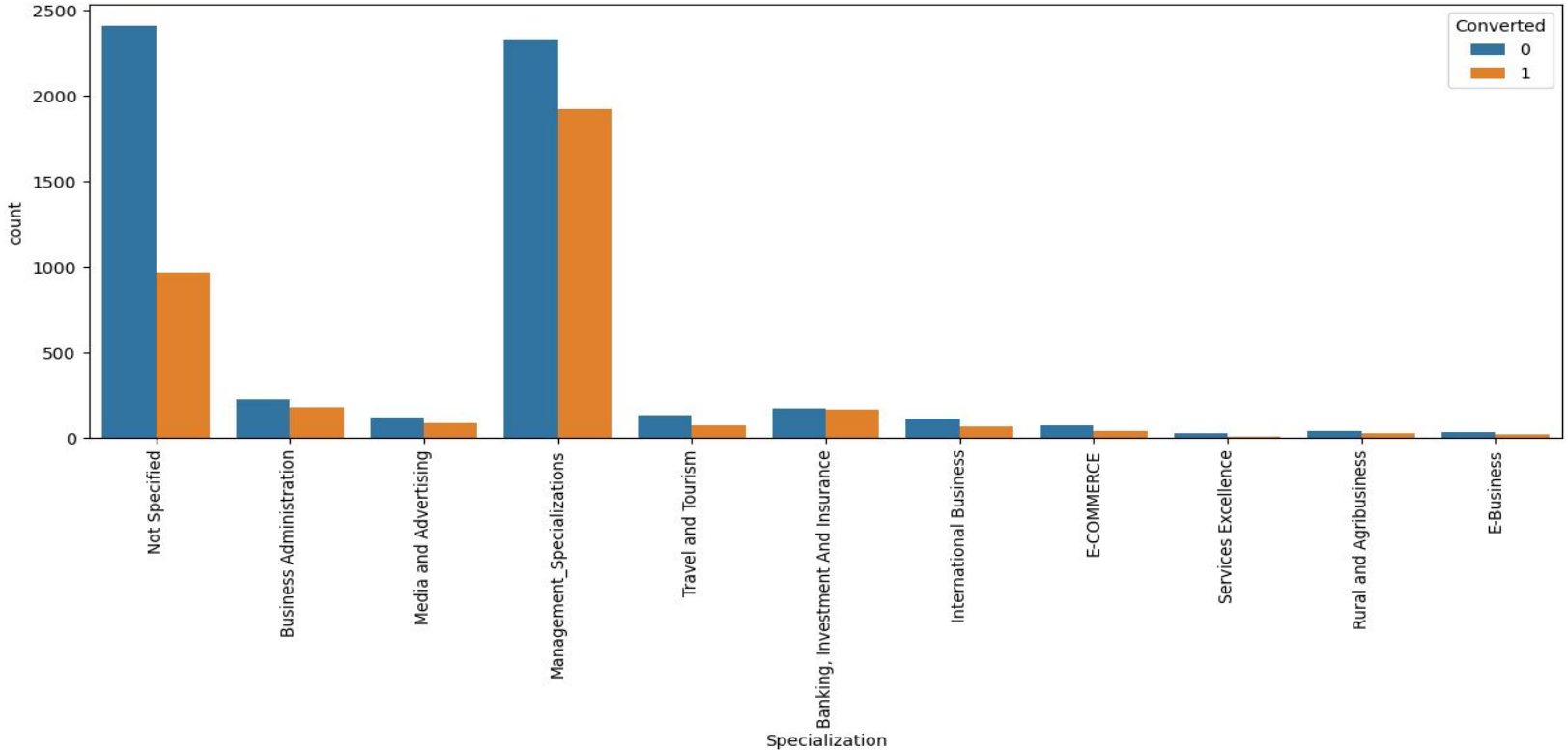
X Education sells online courses to industry professionals and Students. X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers.
The company needs a model wherein a lead score is assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.
The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%

Analysis Approach:-

- After understanding and cleaning the data using appropriate methods, EDA was done and columns having only 1 value were removed as they would affect the logistic regression model due to high data imbalance.
- Then after creating dummy variables for categorical variables and dividing the train and test data in 70:30 ratio, we went on building a logistic regression model with the help of RFE selecting 15 variables. Some variables were manually removed depending on VIF>5 and P-value<0.05
- Then by making a confusion matrix and using ROC Curve and optimal cut-off value, accuracy, sensitivity and specificity were calculated for model evaluation.
- Cut-off of 0.3 was decided based on plot of accuracy, sensitivity and specificity and prediction was done on the train data. Precision and Recall metrics values came out to be 87.8% and 91.3% respectively on the train data set.
- For test data, confusion matrix was made and with optimal cut-off value calculated the conversion probability and found out the accuracy value to be 92.78%; Sensitivity=91.98%; Specificity= 93.26%.

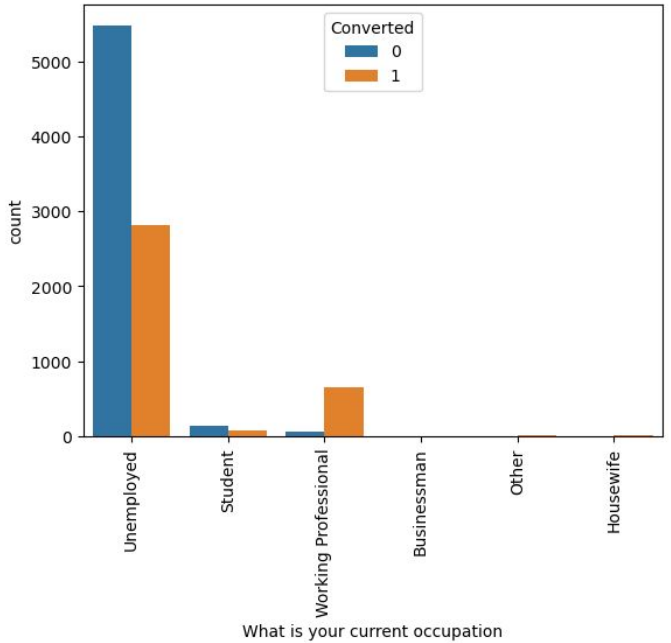
Categorical Analysis

Specialization



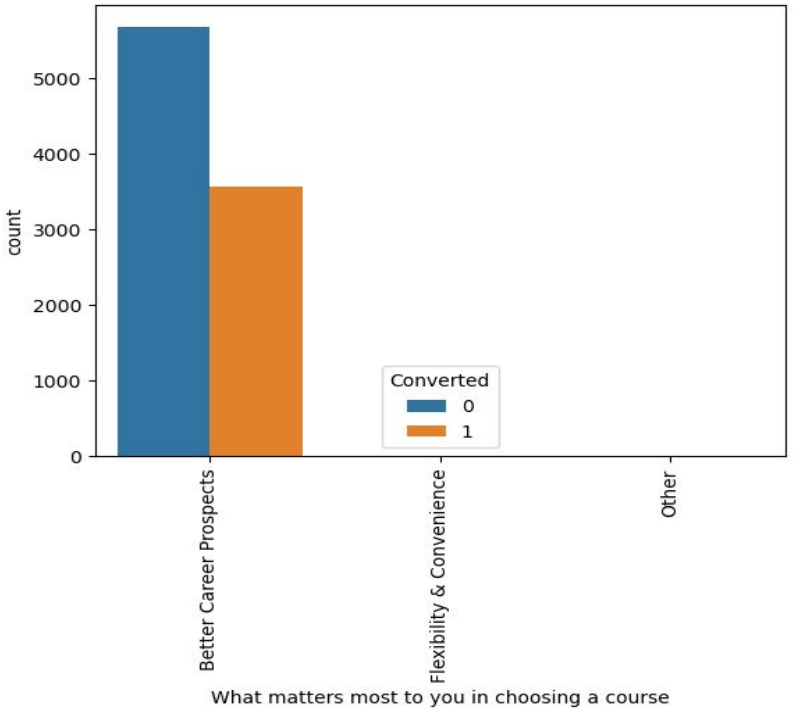
- We see that specialization with **Management** in them have higher number of lead as well as lead converted. So this is definitely a significant variable and should not be dropped.

Current Occupation



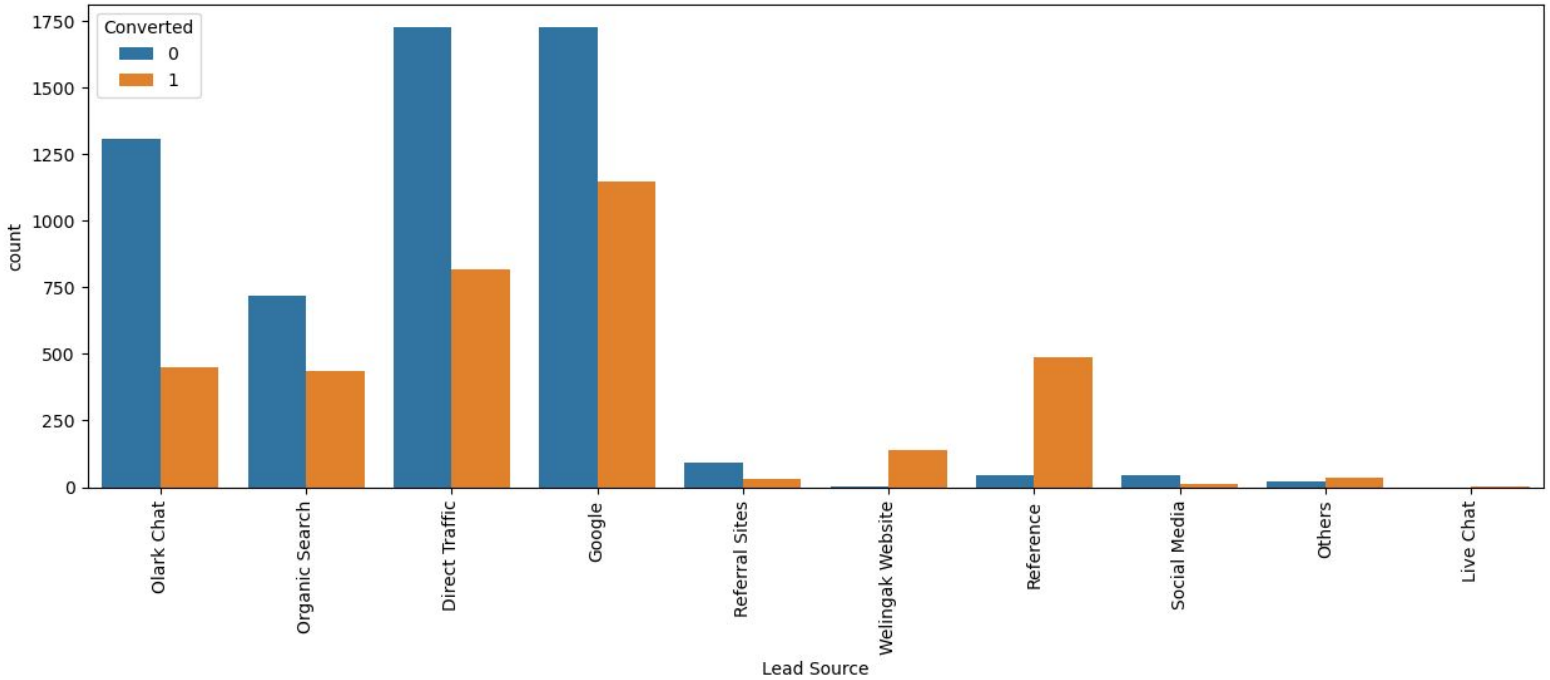
- **Working Professionals** going for the course have high chances of joining.
- **Unemployed** lead are the most in terms of Absolute numbers.

Most important features in choosing a course



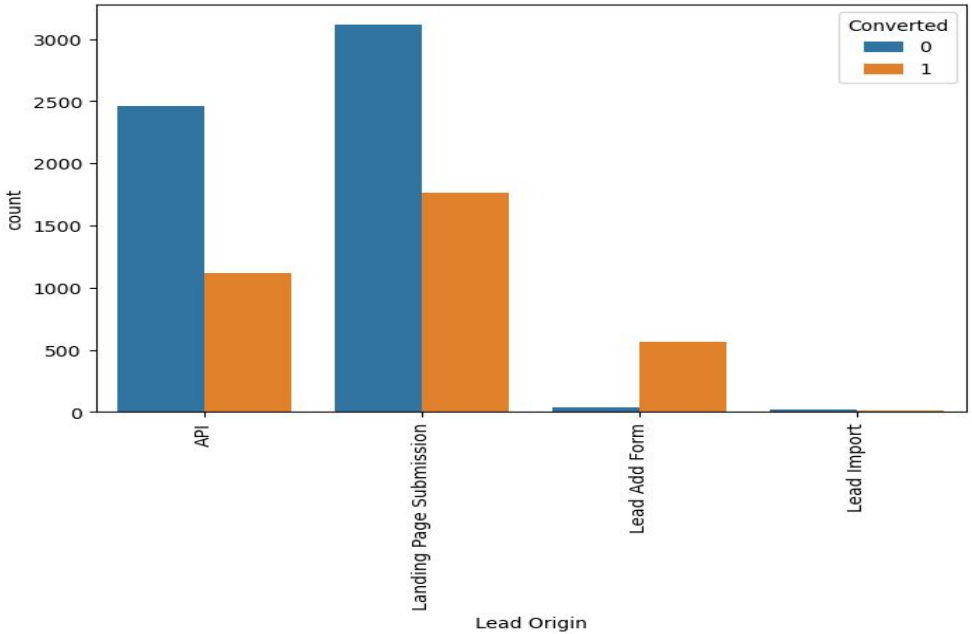
- Better Career Prospects matters most to the customers

Lead Source



- Maximum number of leads are generated by **Google** and **Direct traffic**.
- Conversion Rate of **reference leads** and leads through **Welingak website** is high.
- To improve overall lead conversion rate, focus should be on improving lead conversion of olark chat, organic search, direct traffic, and google leads and generate more leads from reference and Welingak website.

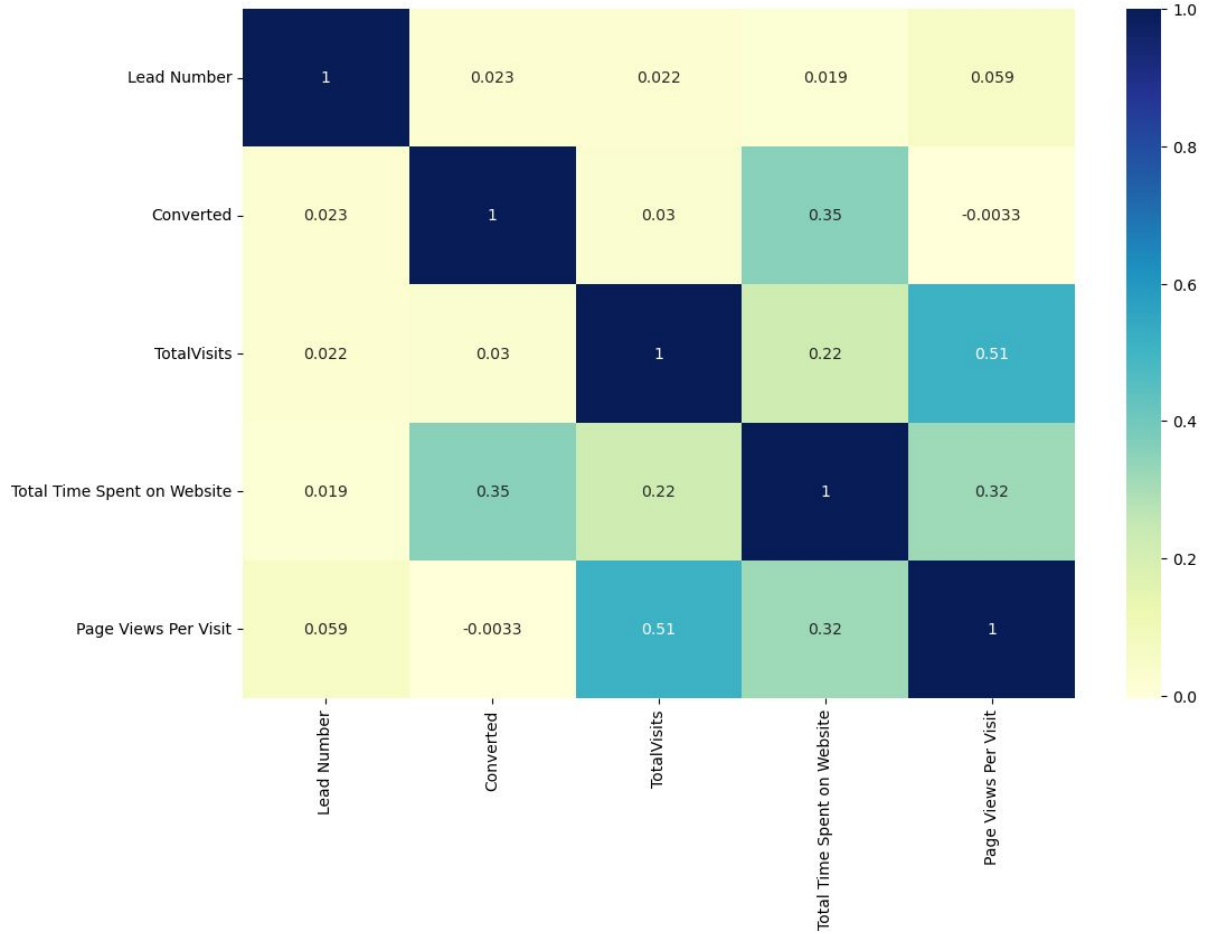
Lead Origin



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

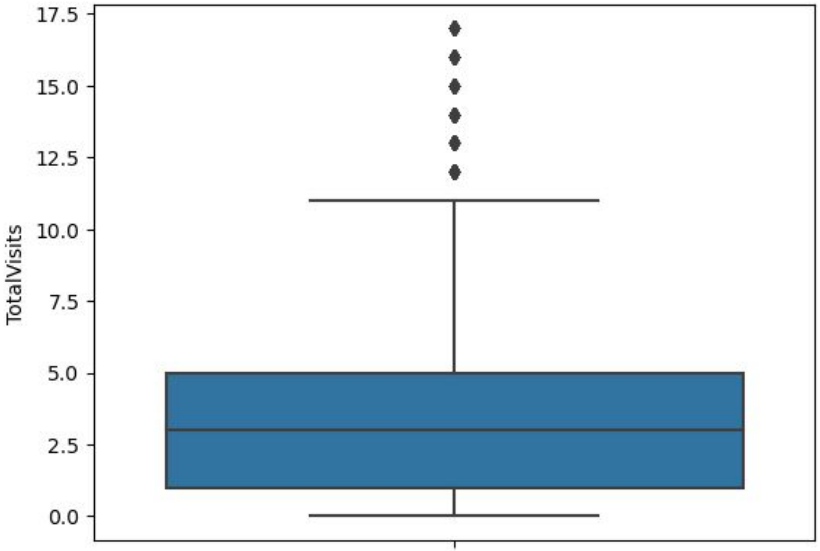
Numerical Analysis

Highest Correlating numerical variables

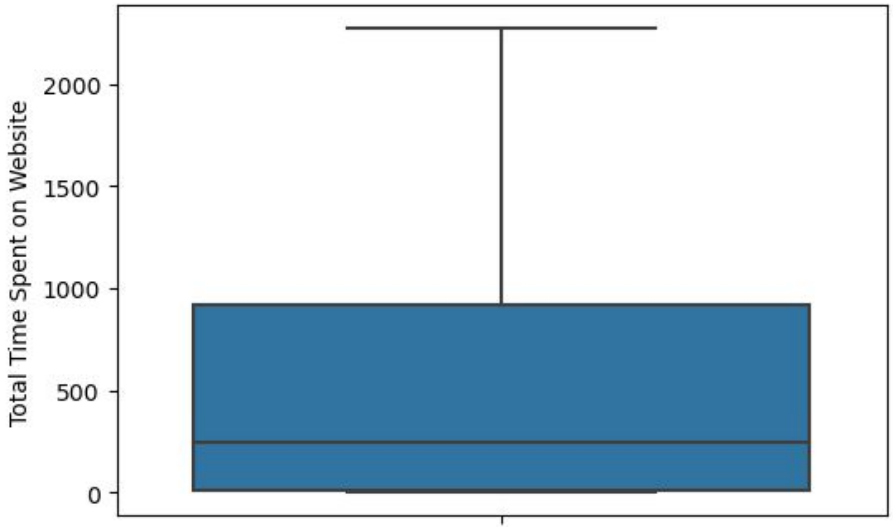


- "**Total time spent on Website**" shows the highest correlation with "Converted" variable
- *'Total Visits'* and *'Pages Views Per Visit'* also shows high correlation

Total Visits

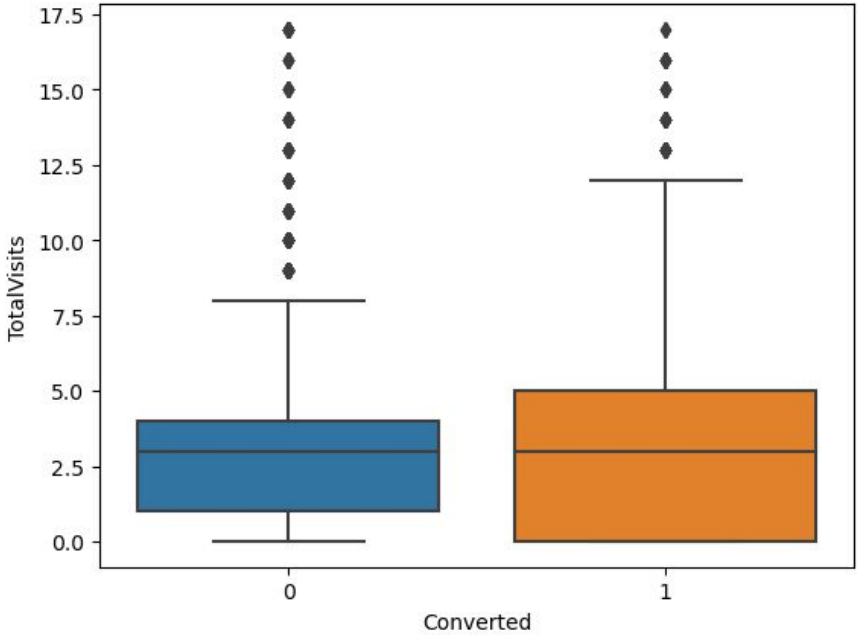


Total Time Spent on Website



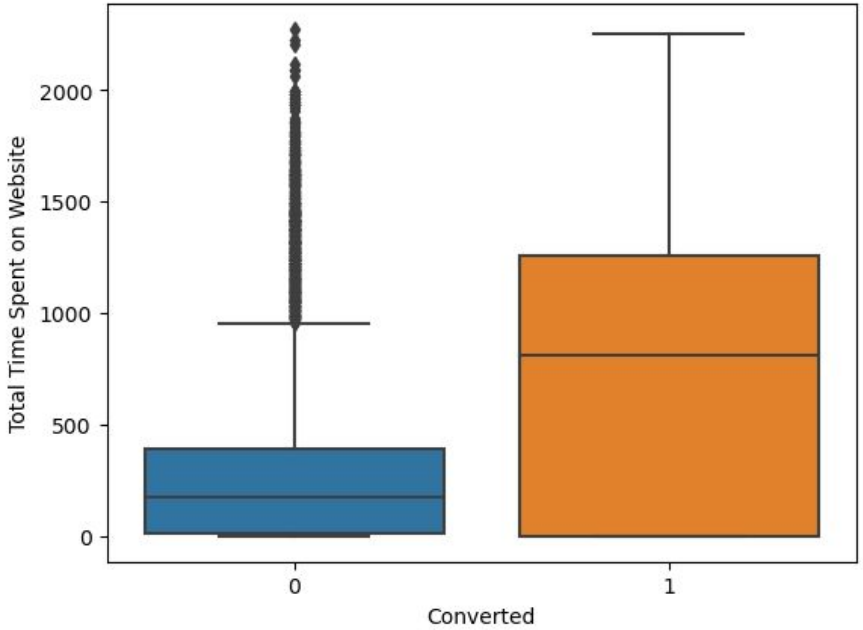
- Here, we don't have any outliers therefore no treatment is required.

Conversion rate of Total Visits



- Median of both cases are same in case of Total Visits
- But upper and lower limit of successful leads has a large gap between them

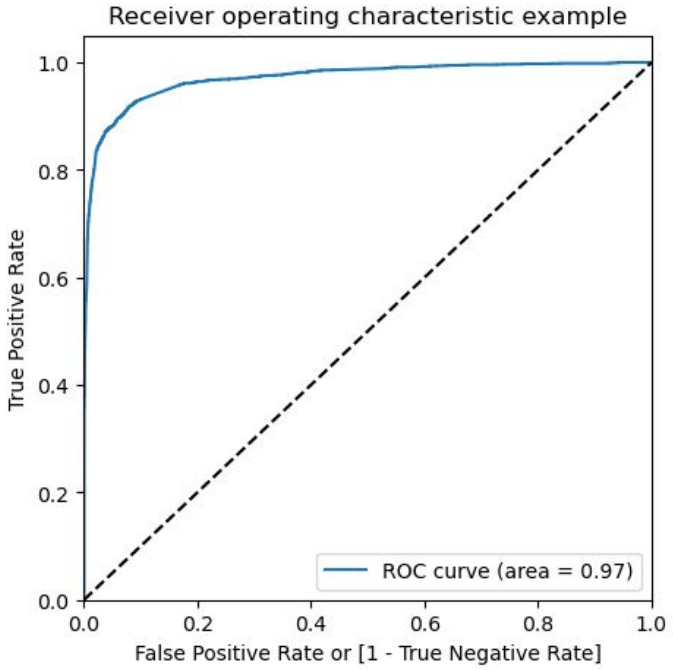
Conversion rate of Total Time Spent on Website



- Leads spending more time on the website are more likely to be converted.

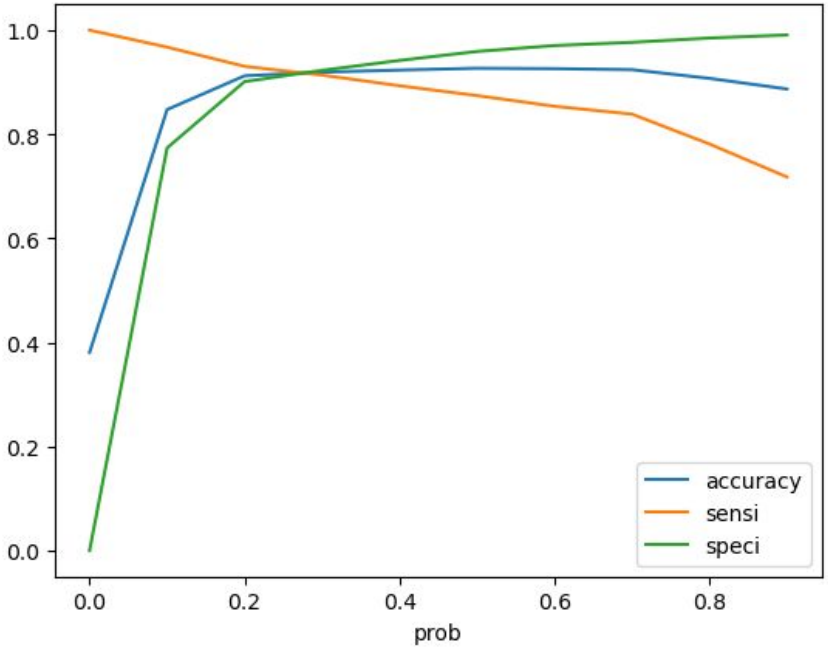
Model Evaluation Parameters

ROC Curve



- **ROC Curve** has area = **0.97**, indicating a good predictive model.

Accuracy, Sensitivity and Specificity Plot for various probabilities



- From the curve above, **0.3** is the optimum point to take it as a cutoff probability.

Final Observation:

Train Data:

Accuracy : **92.29%**

Sensitivity : 91.70%

Specificity : 92.66%

Test Data:

Accuracy : **92.78%**

Sensitivity : 91.98%

Specificity : 93.26%

- Accuracy, Sensitivity and Specificity of Train Data is around **92 percent** based on the cut-off probability value of **0.3**(Using ROC Curve)

- Final model predicts the Test Data's leads conversion rate with Accuracy, Sensitivity and Specificity of **92.78%**, **91.98%**, **93.26%** respectively

Insights:-

Most important variables for potential buyers and high lead conversion:-

Management Specializations
The total time spent on the Website.
Total number of visits.
SMS and Olark Chat conversation

Lead Sources:-

- Reference
- Welingak website
- Google Search and Direct traffic

Lead Origin:-

- Lead Add form and Landing Page Submission

Recommendations:-

- X education could focus on engaging the customers in spending more time on the website and increasing the no. of visits by marketing its course aggressively on various social media platforms.
- Also the lead sources like Google search and direct traffic suggests the need for robust SEO Optimization.
- Management Specializations seems to be more popular among customers, hence the company can launch various offers for these courses and make it as its forte.
- Lastly, it could provide cashback for every referral the enrolled customers make to their friends.