LEAD SCORING CASE STUDY

Sanket
Saravanan Ilangovan
Sree Saradha

PROBLEM

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

There are a lot of leads generated in the initial stage, but only a few of them come out as paying customers. In the middle stage, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion.

X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score 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%.

BUSINESS OBJECTIVE

X education wants to know most promising leads.

For that they want to build a Model which identifies the hot leads.

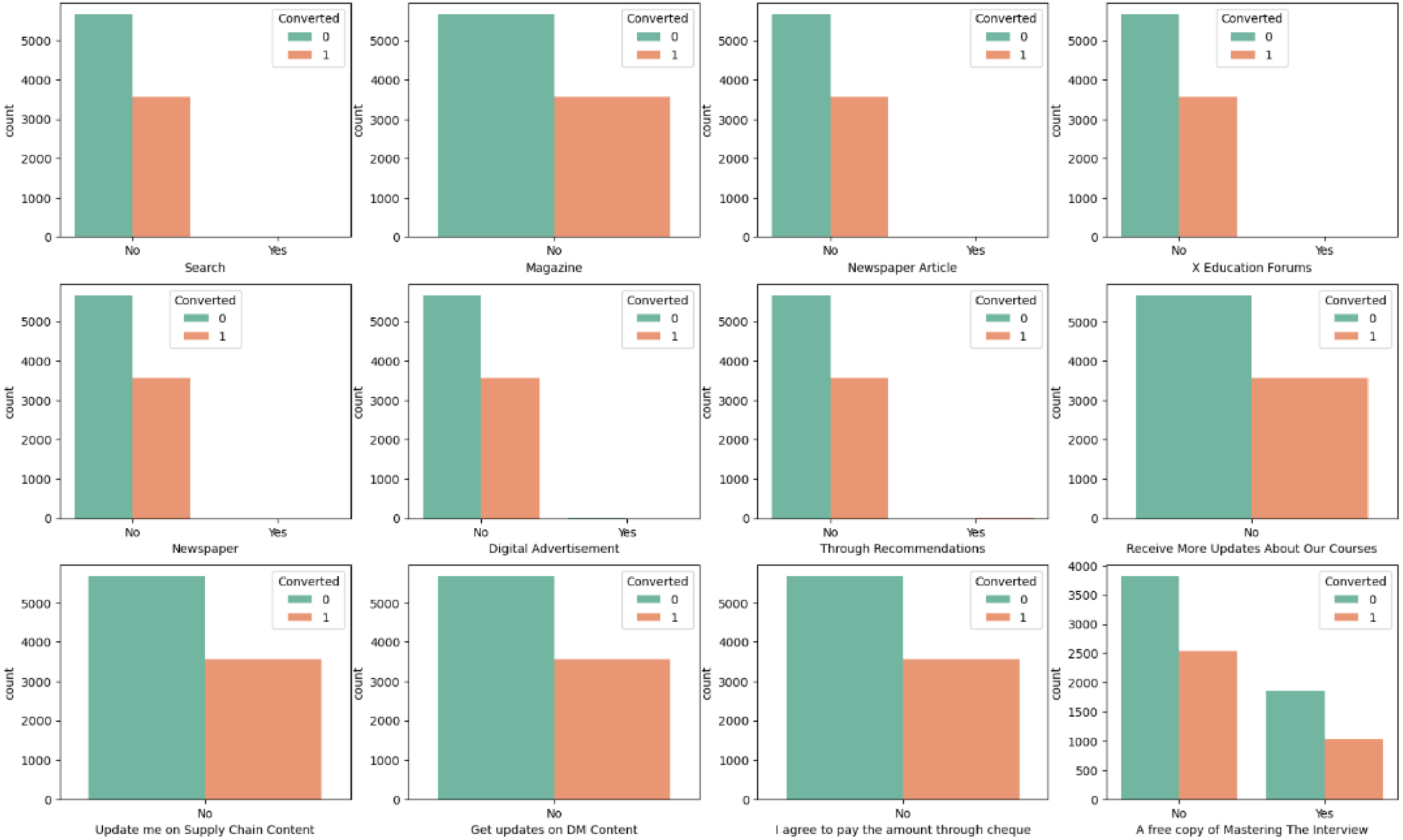
Deployment of the model for future use.

SOLUTION APPROACH

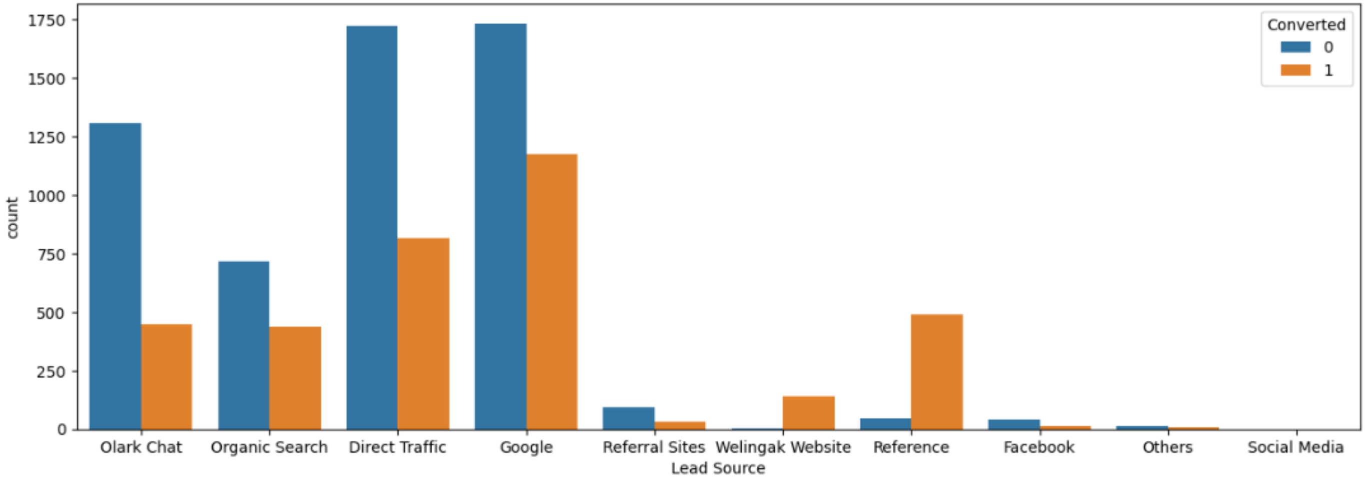
- **Data cleaning and data manipulation**
  - Check and handle duplicate data.
  - Check and handle NA values and missing values.
  - Drop columns, if it contains a large amount of missing values and not useful for the analysis.
  - Imputation of the values, if necessary.
  - Check and handle outliers in data.

- **Exploratory Data Analysis**
  - Univariate data analysis: value count, distribution of variables etc.
  - Bivariate data analysis: correlation coefficients and pattern between the variables etc.

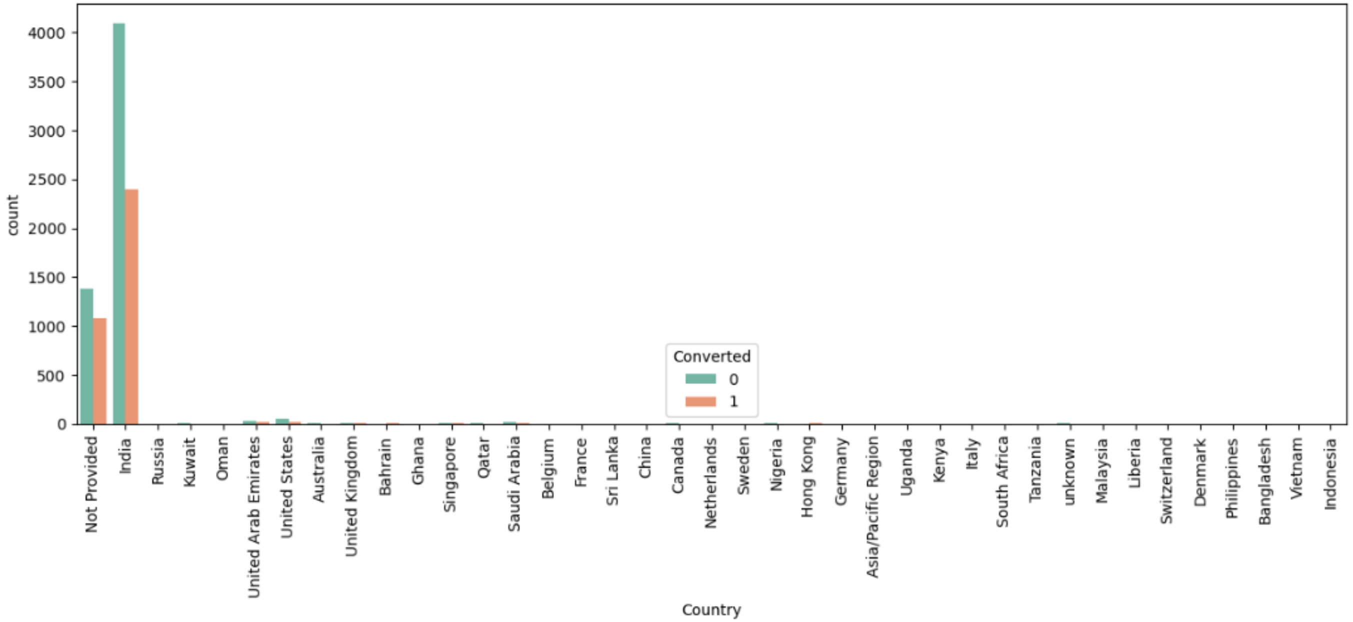
- Feature Scaling & Dummy Variables and encoding of the data.
- Classification technique: logistic regression is used for the model making and prediction.
- Validation of the model
- Model Presentation
- Conclusions and recommendations.



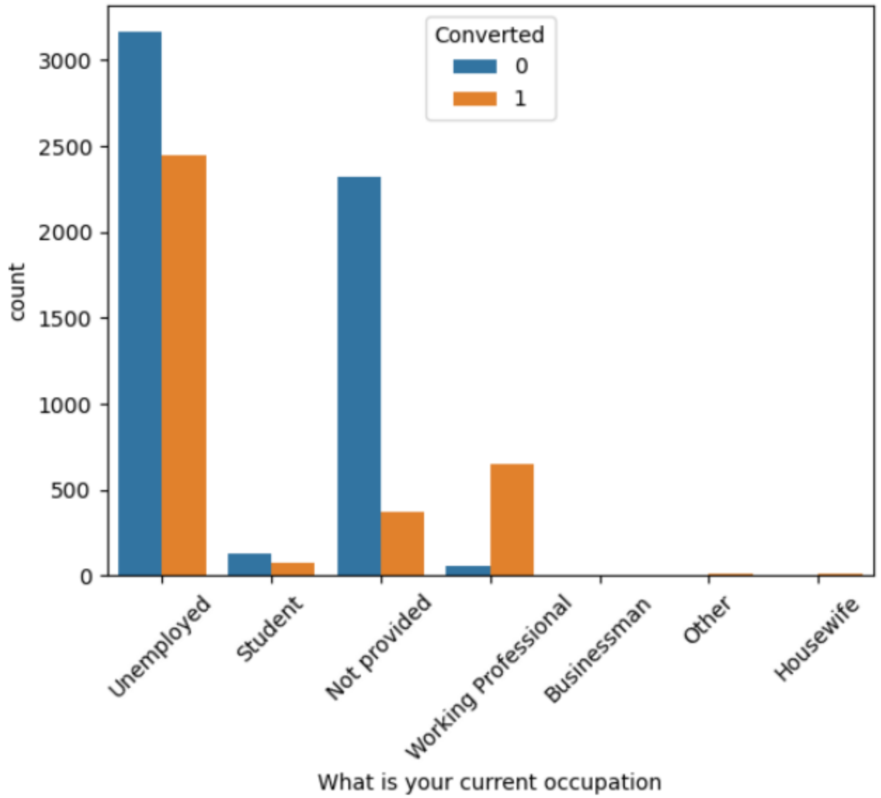
For all these columns except 'A free copy of Mastering The Interview' data is highly imbalanced

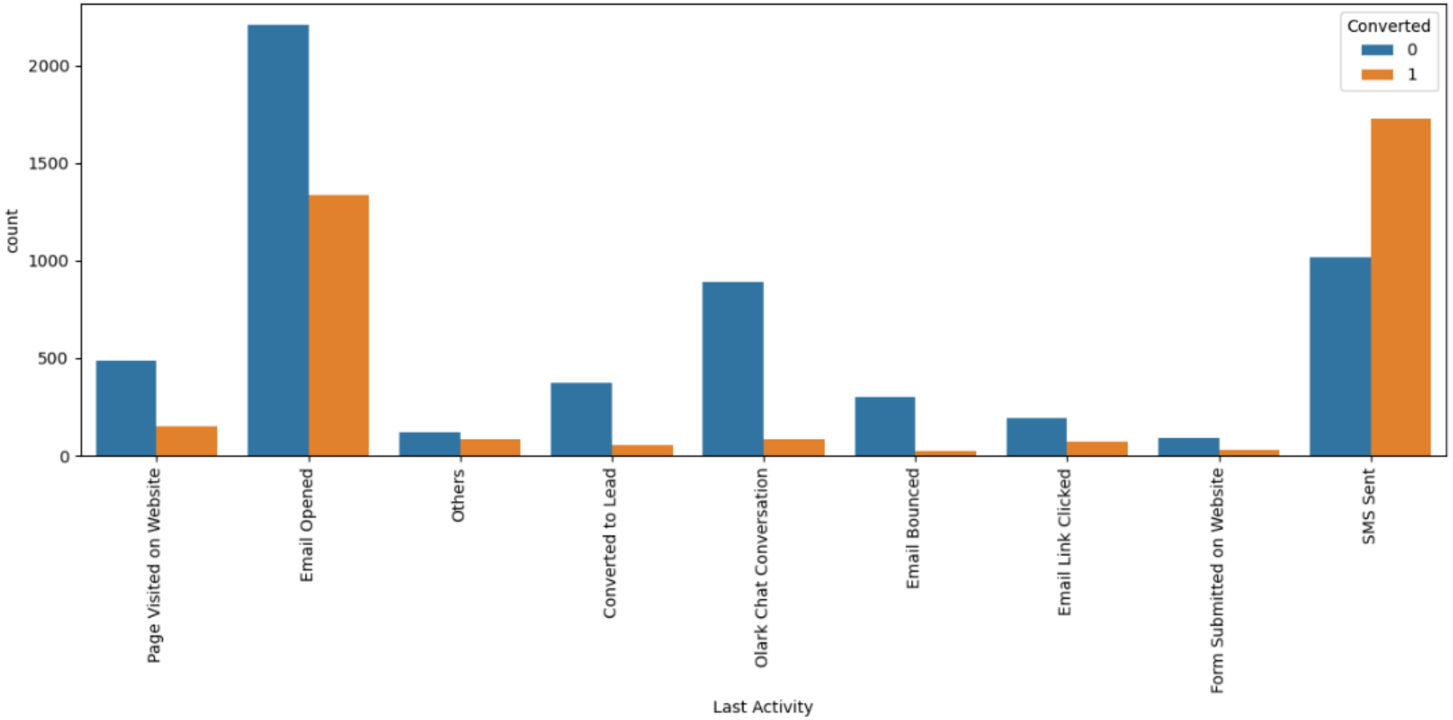


Maximum Leads are generated by Google and Direct Traffic.
Leads from Reference and Welingak Website are high.



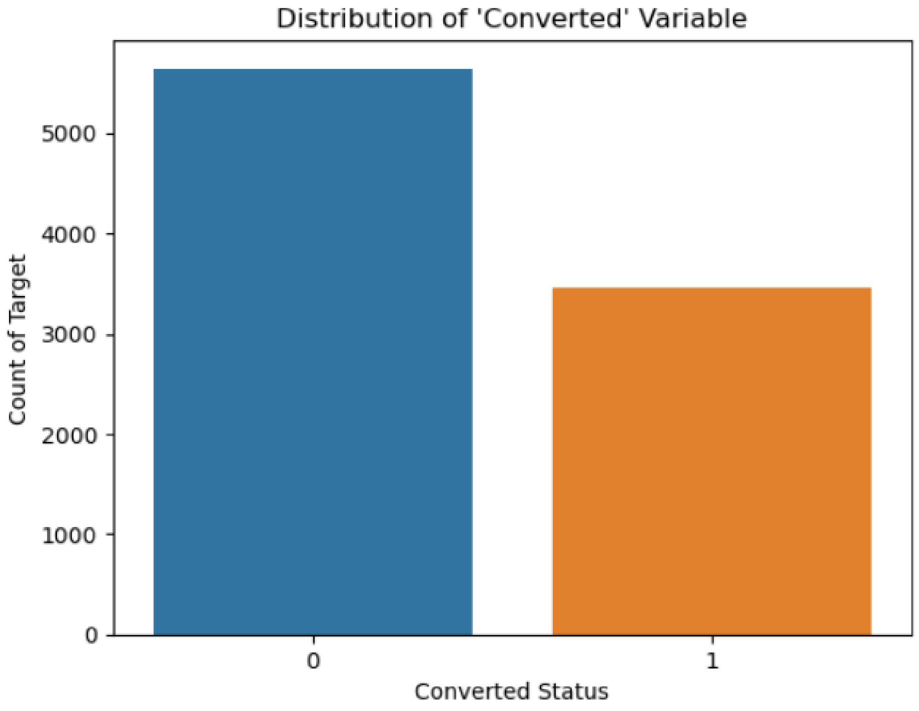
Most of the data are from India, we cannot draw any inference from this.



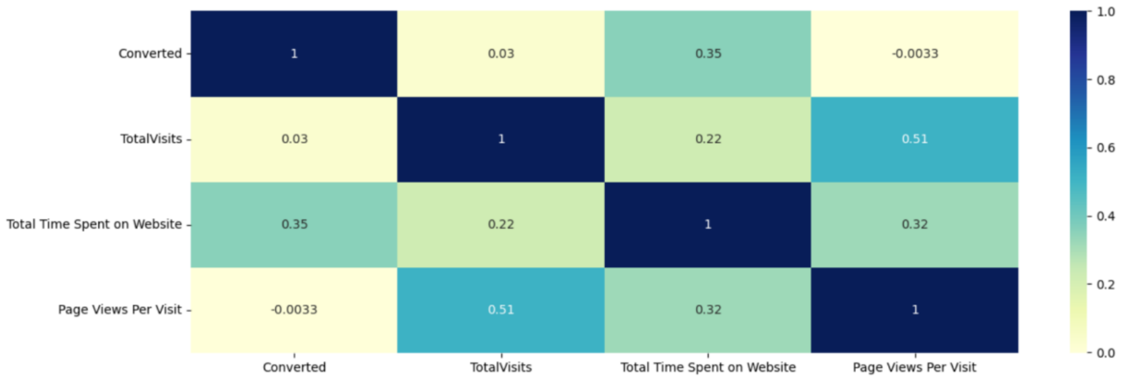


Maximum leads generated are unemployed and their conversion rate is more than 50%. Conversion rate of working professionals is very high.

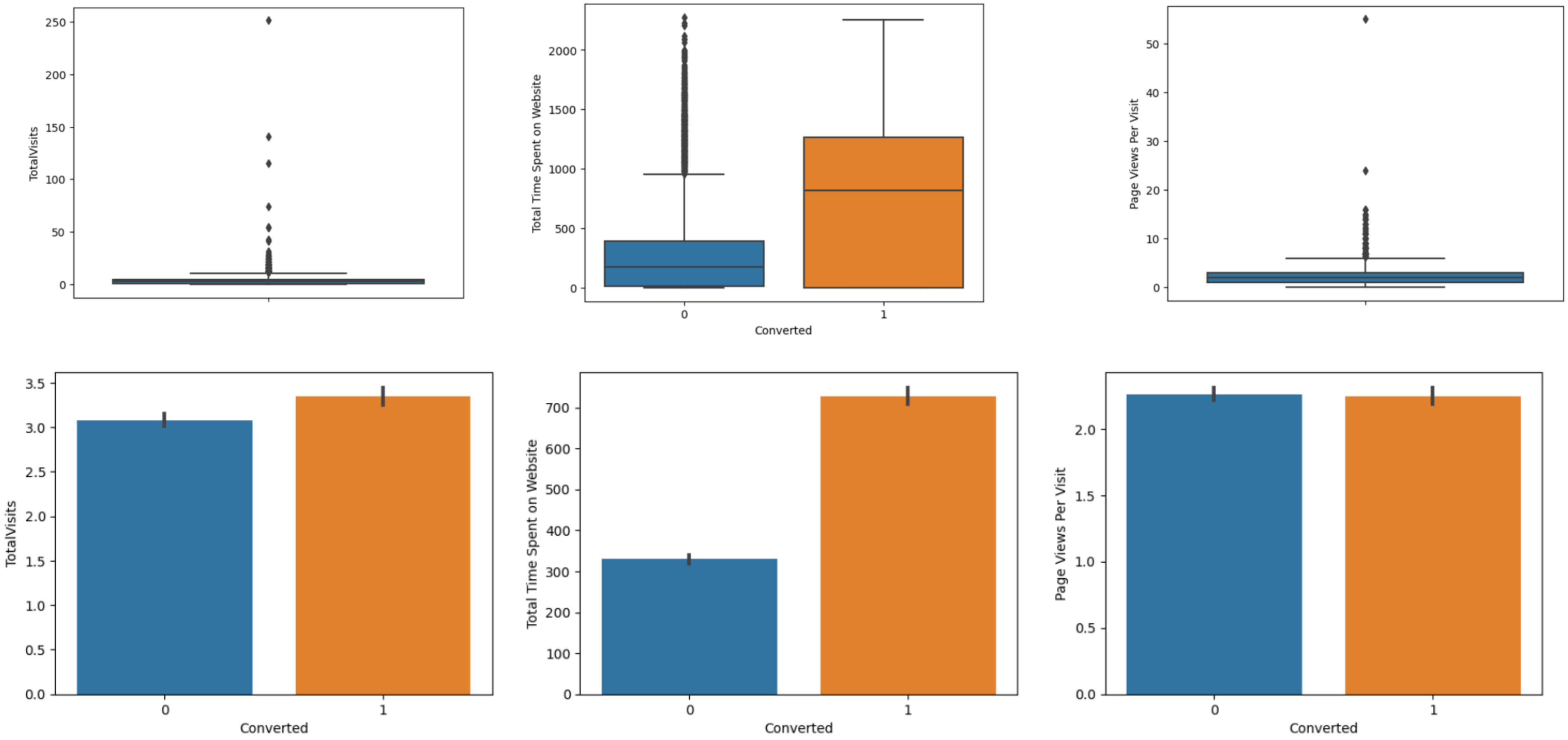
Maximum leads are generated having “Last activity” as “Email opened” but the conversion rate is not too good. “SMS sent” as the last activity has a high conversion rate



Currently, the lead Conversion rate is 38% only



The heatmap displays a matrix with colors that indicate the degree of correlation between the features.
Based on this Multicollinearity among the features is avoided.



The conversion rate is high for Total Visits, Total Time Spent on Website and Page Views Per Visit

DATA CONVERSION

- Numerical Variables are Normalised
- Dummy Variables are created for object-type variables
- Total Rows for Analysis: 9103
- Total Features for Analysis: 23

After finalizing the Model we are left with

- Total Features: 11

SCALING

The machine learning models assign weights to the independent variables according to their data points and conclusions for output. In that case, if the difference between the data points is high, the model will need to provide more significant weight to the farther points, and in the final results, the model with a large weight value assigned to undeserving features is often unstable. This means the model can produce poor results or can perform poorly during learning.

In our model, we have done scaling using standardized method

MODEL BUILDING

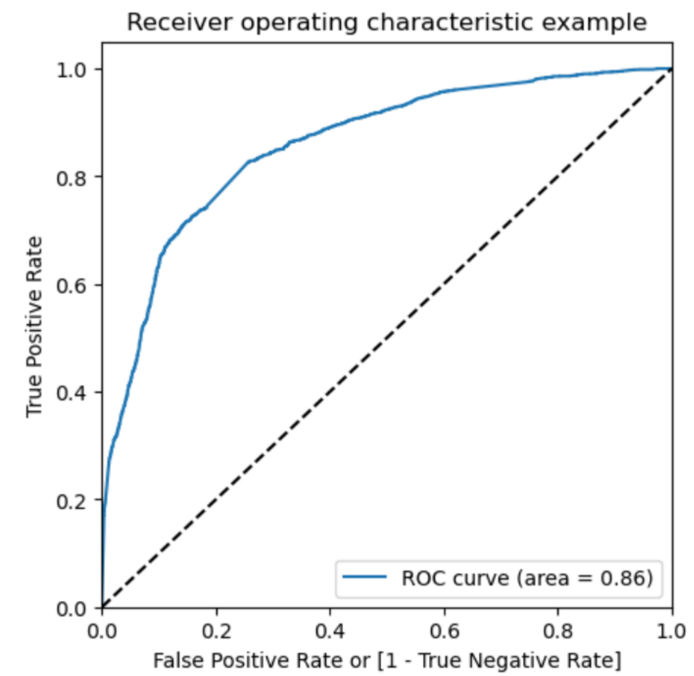
- Splitting the Data into Training and Testing Sets
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio
- Use RFE for feature Selection
- Running RFE with 15 variables as output
- Building Model by removing the variable whose p-value is greater than 0.05 and VIF value is greater than 5
- Predictions on a test data set
- Based on the model evaluation factors recall, precision, accuracy, sensitivity, and specificity final model arrived

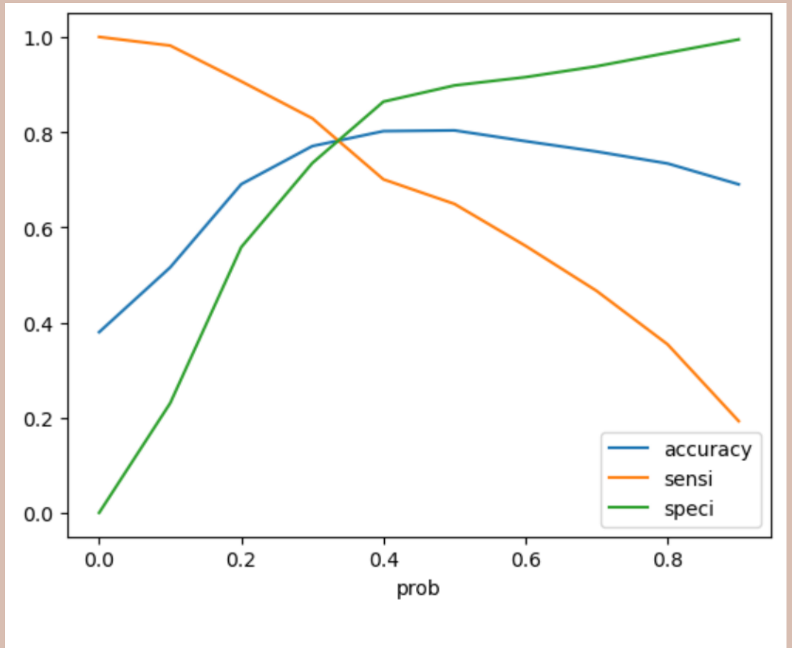
ROC CURVE

Finding Optimal Cut-off Point

The optimal cut-off probability is that Probability where we get balanced sensitivity and specificity

From the second graph, it is visible that the optimal cut-off is at 0.30





From the plot above, we can take 0.3 as the optimum point to take it as cutoff probability

The ROC curve should be valued close to 1, We are getting a good value of 0.86 indicating a good predictive model.

VALIDATION OF THE MODEL

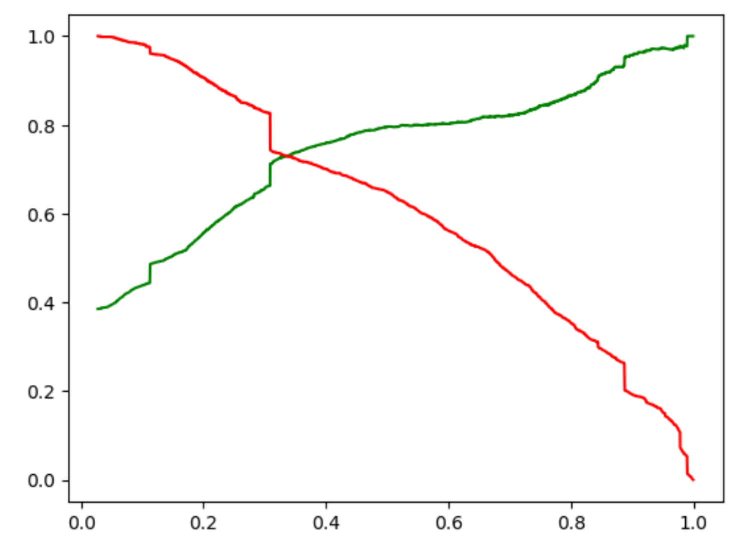
Train Data

- -Accuracy : 77.05%
- -Sensitivity :64.9%
- -Specificity : 89.8%

Test Data

- Accuracy : 77.52%
- Sensitivity :64.9%
- Specificity : 89.8%

Precision and Recall Trade-off



CONCLUSION

It was found that the features that mattered the most are

- Lead Origin_Lead _Add Form
- What is the current Occupation_Working Professional
- What is the current occupation_unemployed
- What is your current occupation_student
- Total Time Spent on Website

Recommendation based on the final model are

To Increase our Lead conversion rates

- Focus on features with positive coefficients for targeted marketing strategies.
- Develop strategies to attract high quality leads from top performing lead sources.
- Optimize communication channels based on lead engagement impact.
- Engage Working Professionals with tailored messaging.
- More budget/spend can be done on **Welingak Website** in terms of advertising, etc.
- Incentives/discounts for providing references that convert to lead, encourage providing more references.
- Working professionals to be aggressively targeted as they have high conversion rate and will have a better financial situation to pay higher fees too.

To Identify areas of improvement

- Analyse negative coefficients in specialization offerings.
- Review the landing page submission process for areas of improvement.

THANK YOU