LEAD SCORING CASE STUDY

Group Members

Soumya Saridena

Devika R

Problem Statement

- 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.
- 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%.
- Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.
- 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.

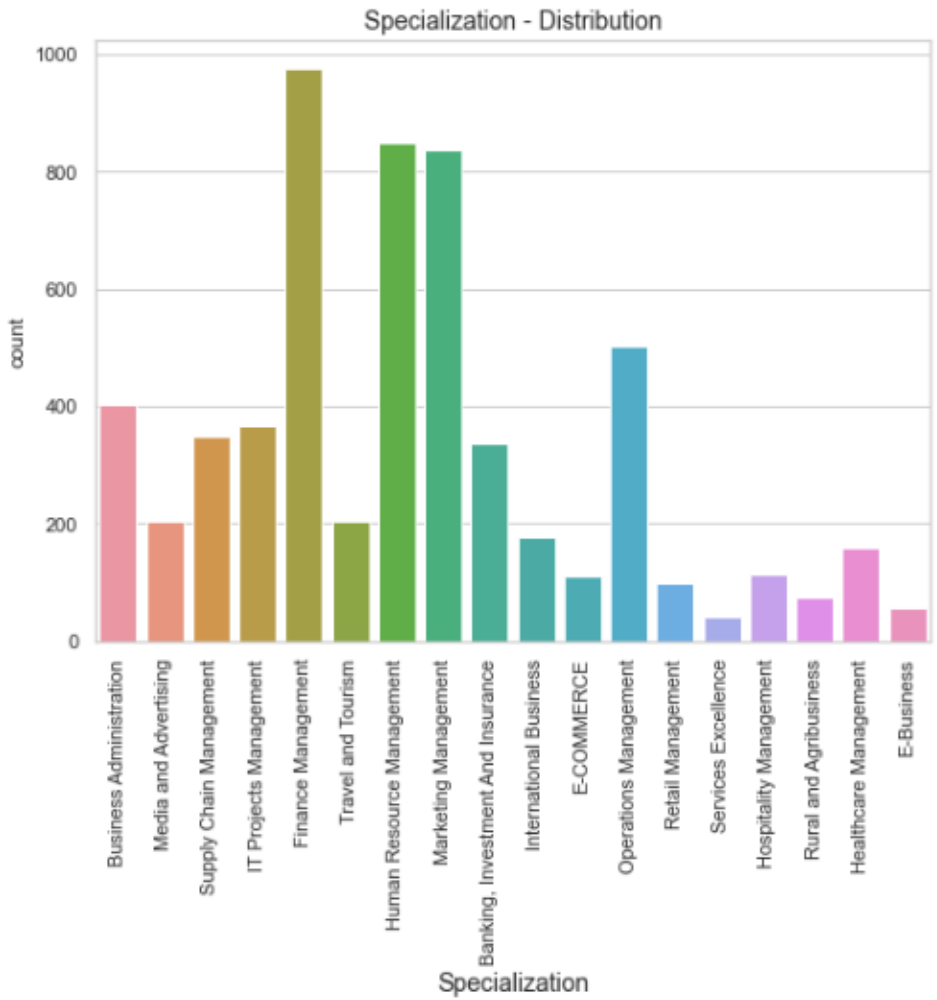
Goals of the Case Study

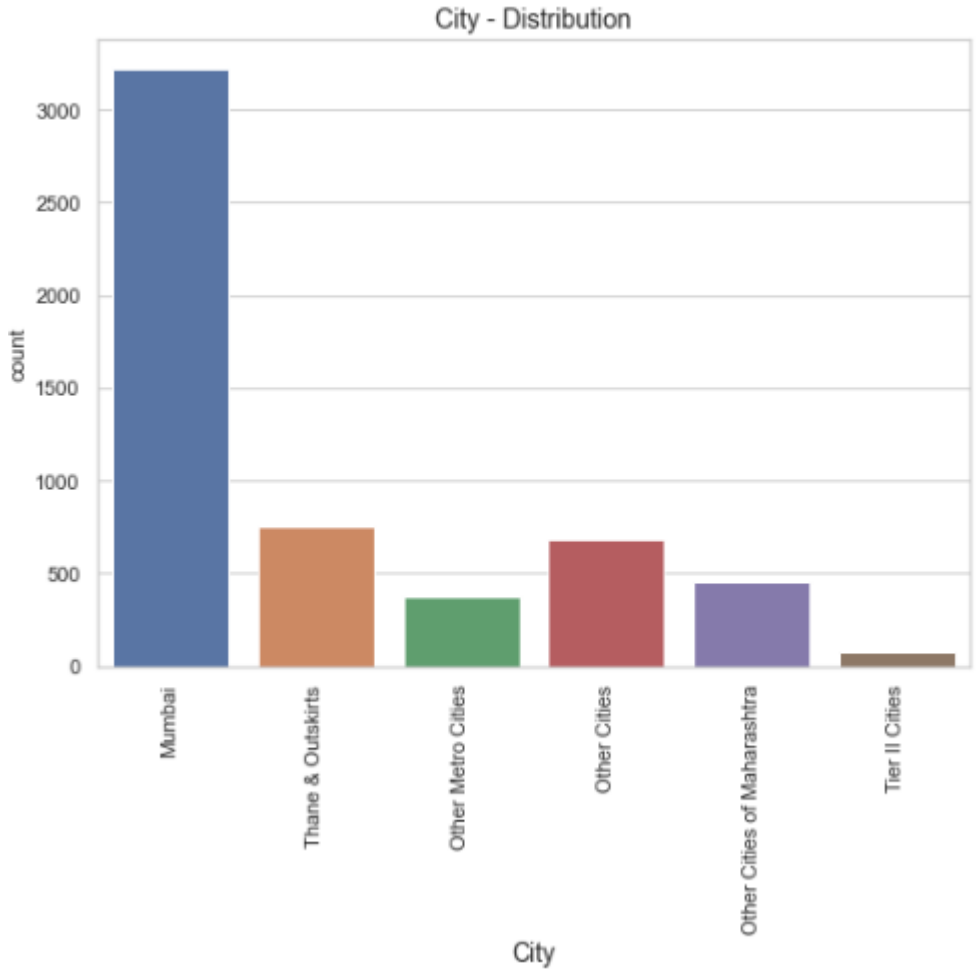
- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.
- There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

Solution Approach

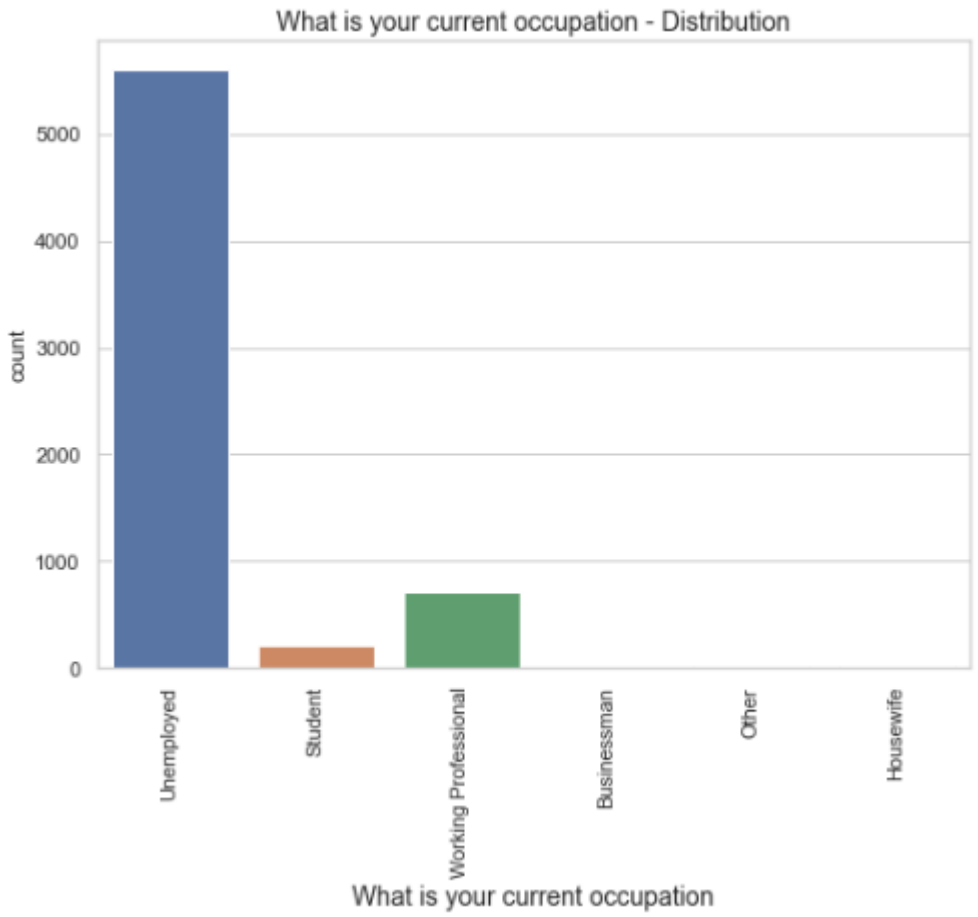
- Data Reading and Data Understanding
- Data cleaning - Missing value treatment, handling delimiters, dropping insignificant columns, imputing the null values, data type correction
- Data visualization(EDA) and outlier treatment
- Data preparation i.e., creating dummy variables and doing Train-Test split
- Feature Scaling – using Standard Scaler here
- Model Building – Algorithm used is GLM - Logistic Regression Model – adopted RFE and VIF for model building
- Model Predictions on test set
- Final Observations and Conclusions

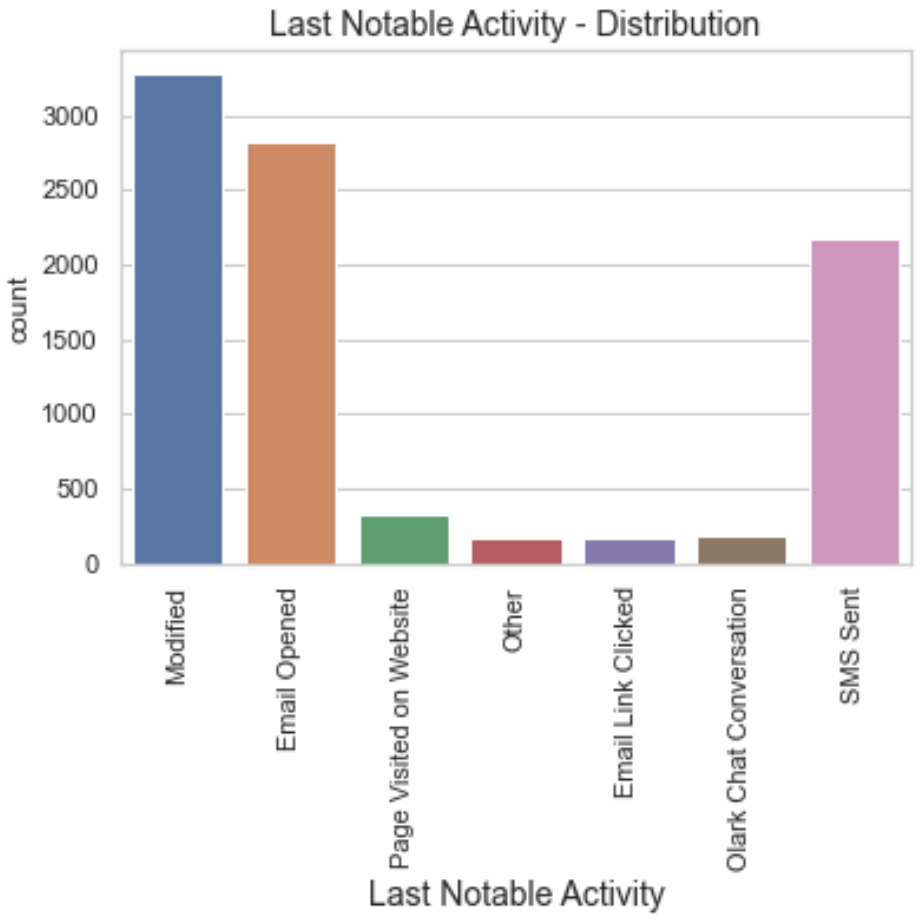
EDA – Univariate Analysis



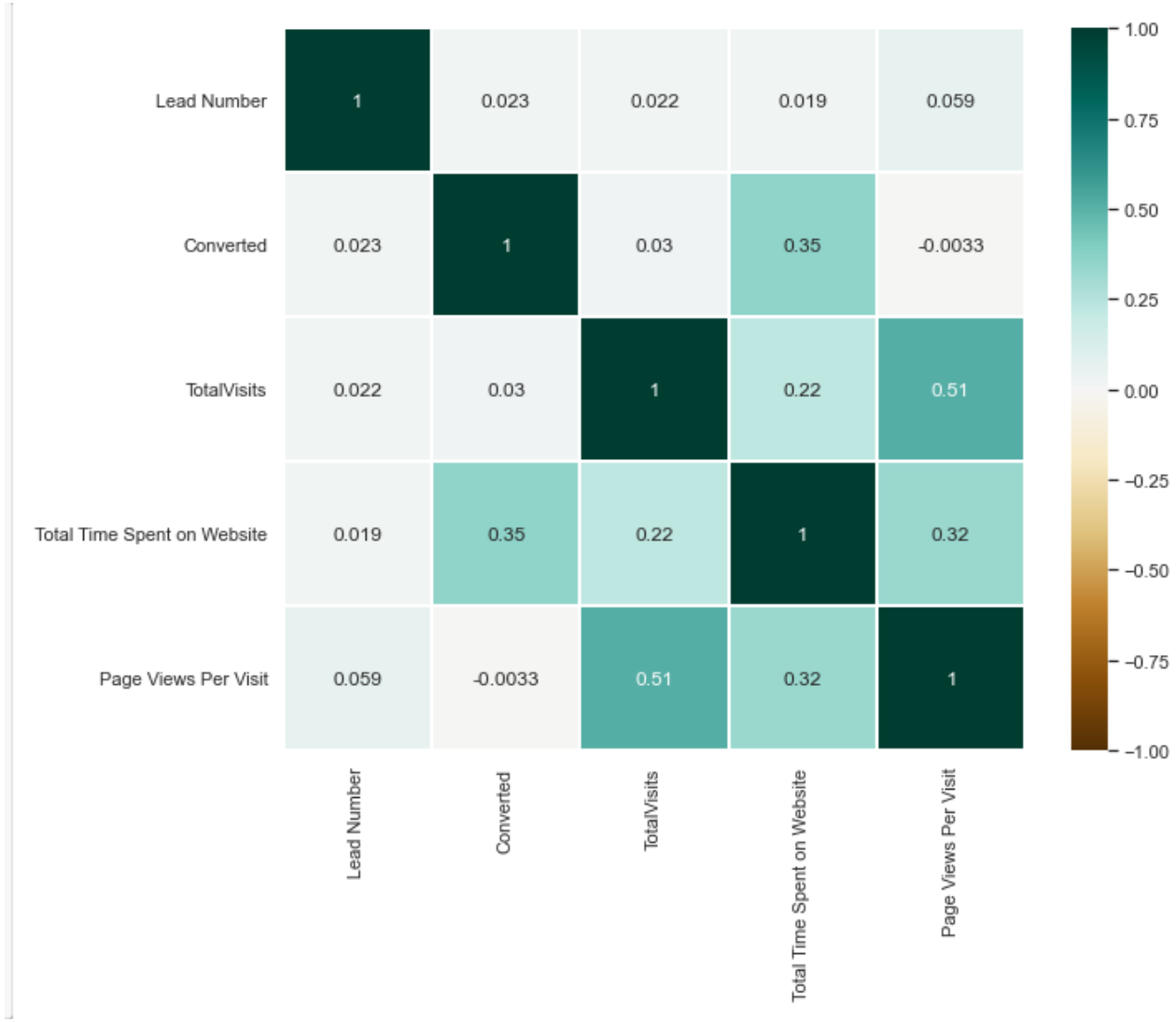


EDA - Univariate Analysis

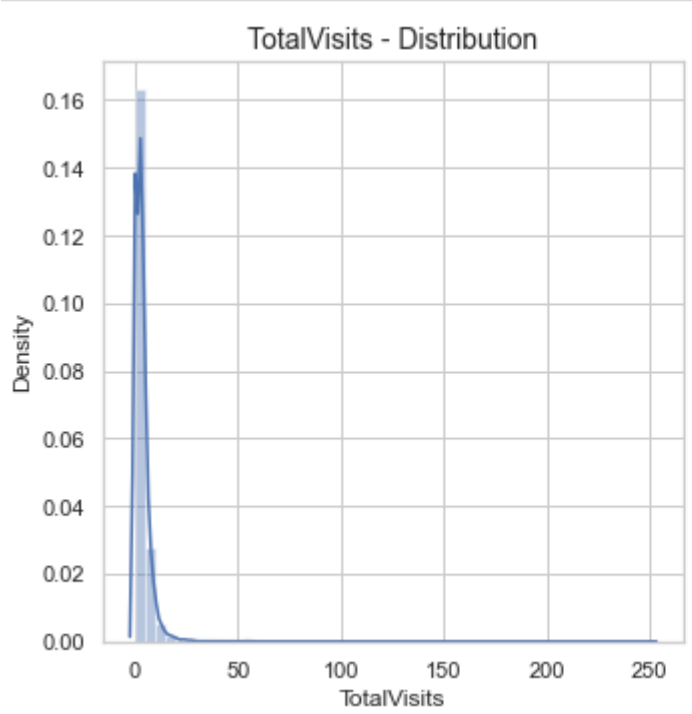


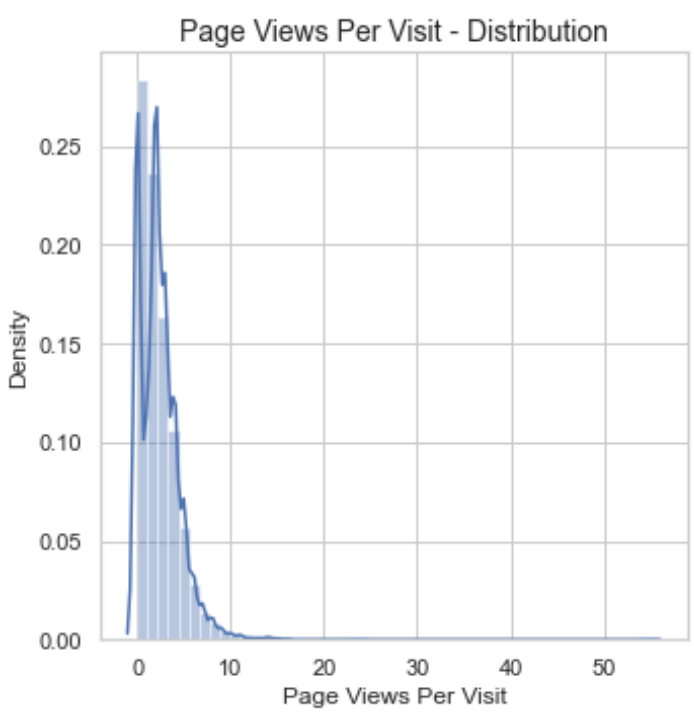


EDA – Correlation Analysis

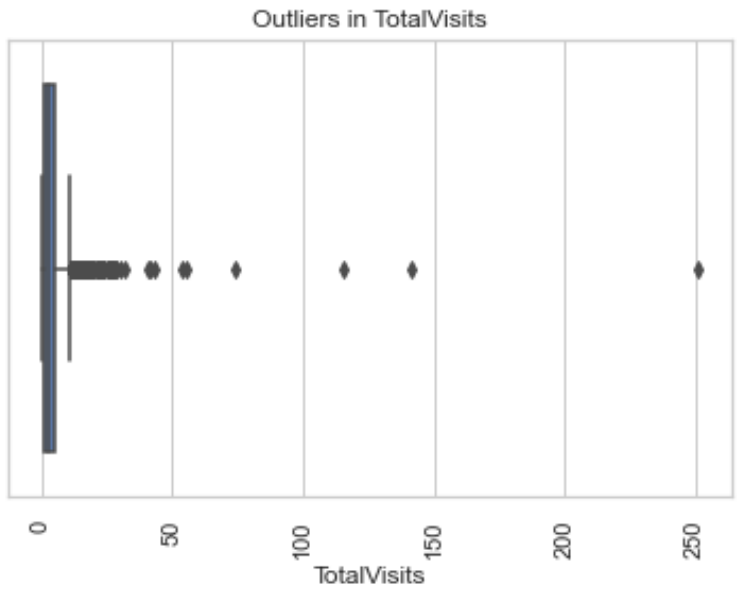


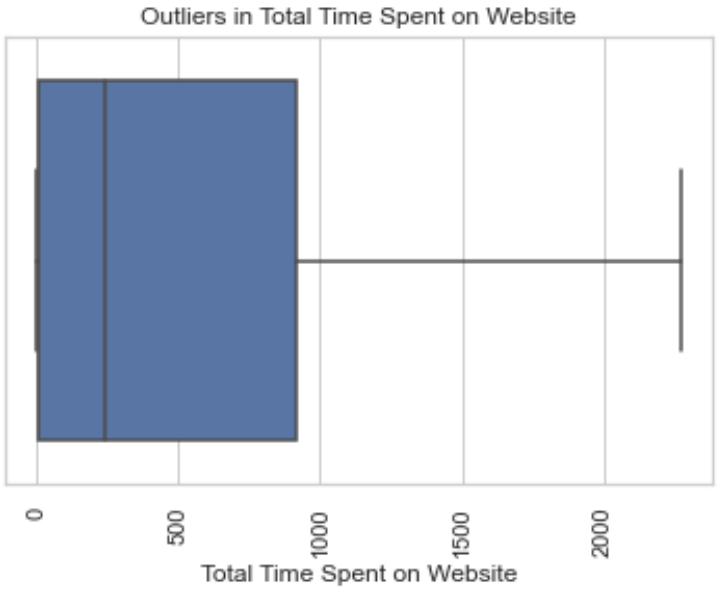
EDA – Visualizing skewness in our data

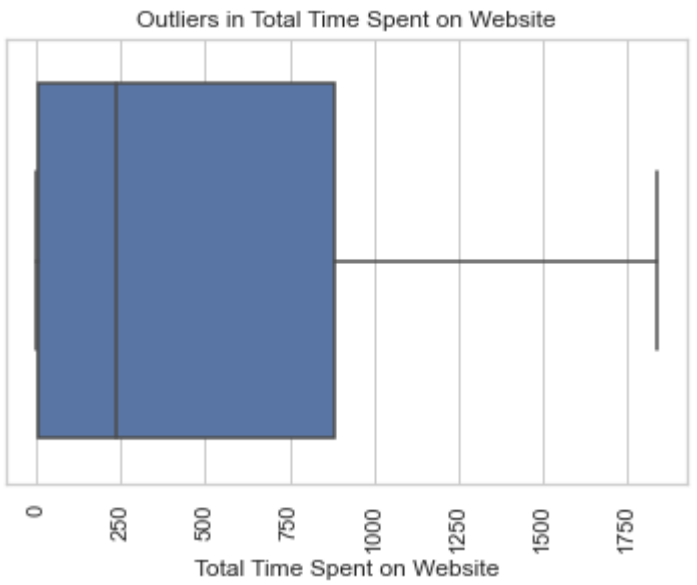




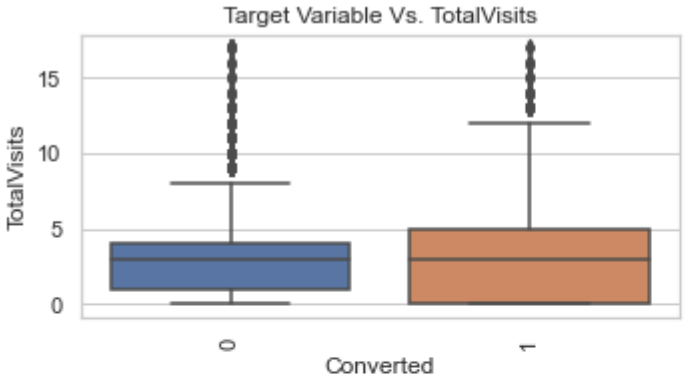
EDA – Numerical features (Outliers)

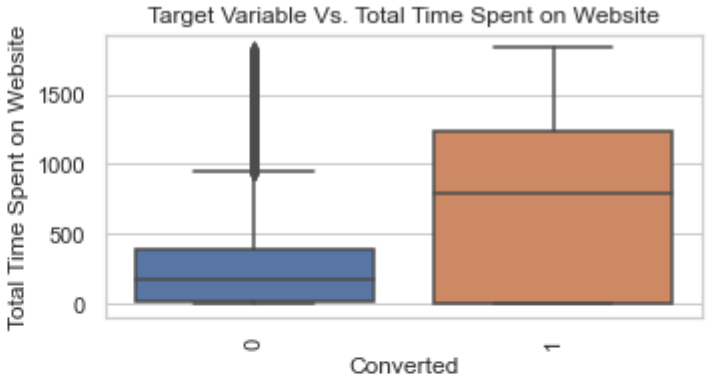


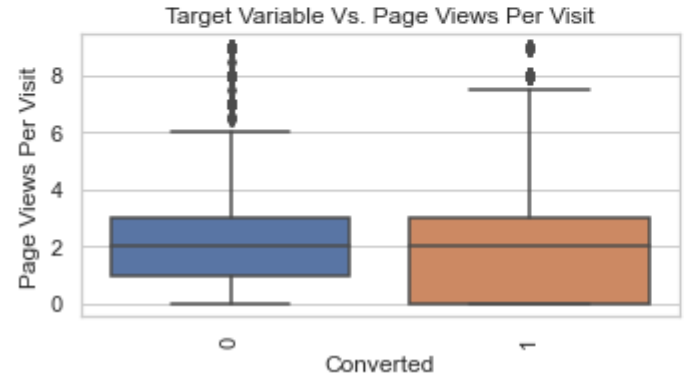




EDA – Bivariate Analysis







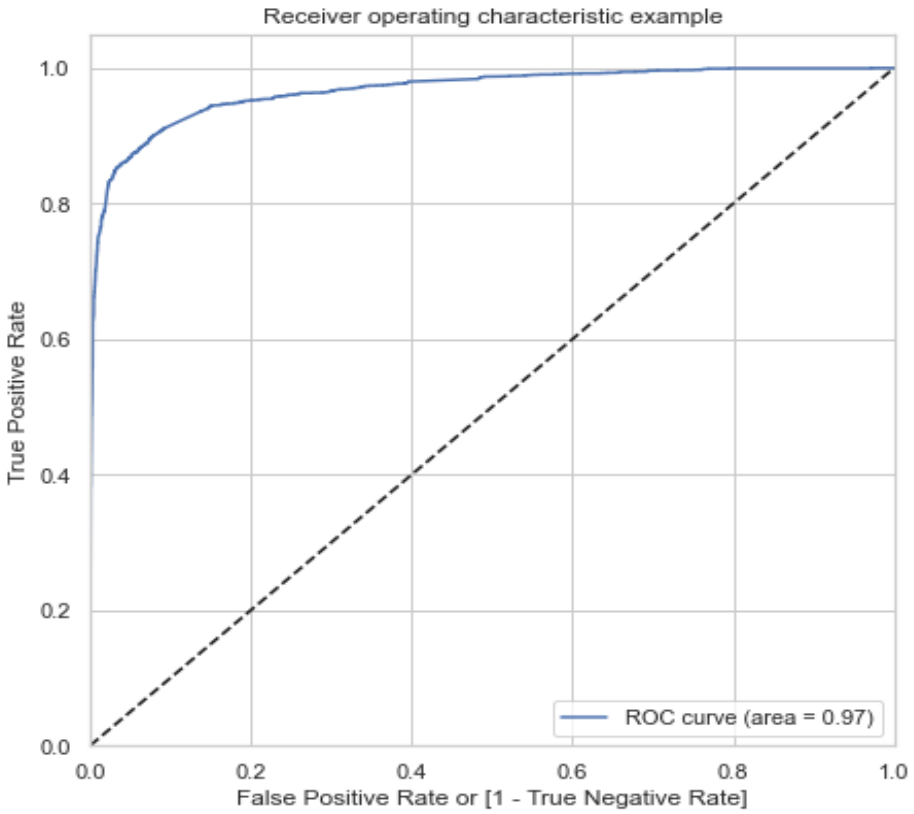
Important Inferences drawn by performing EDA

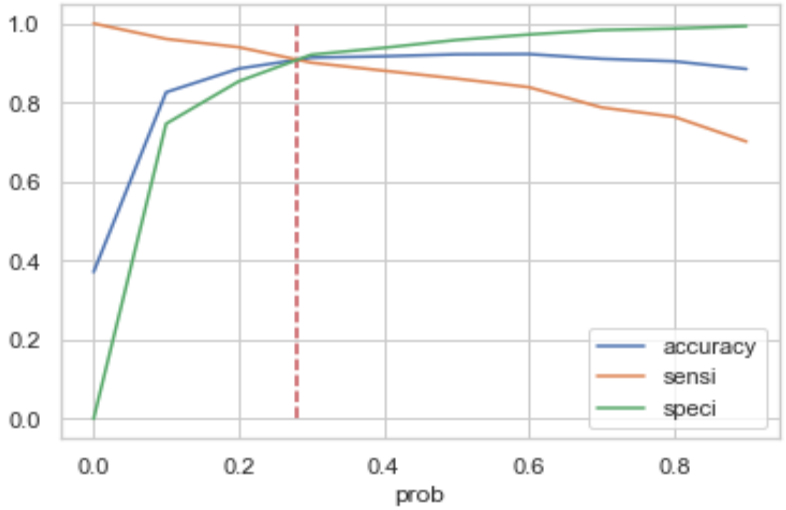
- In order to increase overall lead conversion rate, we must pay more attention to increasing lead conversion rate of Customers coming through API and Landing Page Submission and increasing lead generation from Lead Add Form.
- Lead sources from "Google" have the highest conversion probabilities. Leads with source "Reference" have the highest probability of converting.
- In order to increase total lead conversion rates, we must concentrate more on increasing lead conversion rates for customers whose most recent action was Email Opened and increase the number of leads from customers whose most recent activity was SMS Sent.
- The least likely leads to convert are those with a rural and agricultural business focus.
- The leads who reply after reading the email and others should receive more attention because they are potential prospects and have a higher conversion rate.
- Mumbai is where the majority of the leads come from. Mumbai city residents should be targeted more since they are potential leads

Model Building

- Building our first multivariate logistic regression model using the GLM (Generalized Linear Models) method of the library stats models. 'Binomial()' in the 'family' argument tells stats models that it needs to fit a logit curve to a binomial data (i.e., in which the target will have just two classes.
- Used Recursive Feature Elimination module and choose a limited set of 15 features, we then use manual feature elimination (i.e., manually eliminating features based on observing the p-values and VIFs)
- Re-trained the models by eliminating the features which had a VIF > 5
- Goodness of fit/accuracy from the final model is 91.24% on train set
- Validation on test dataset – accuracy is 92.06%
- When you plot the true positive rate against the false positive rate, you get a graph which shows the trade-off between them, and this curve is known as the ROC curve. A good ROC curve is the one which touches the upper-left corner of the graph; so higher the area under the curve of an ROC curve, the better i`s your model.
- To capture these errors, and to evaluate how well the model is, we used something known as the 'Confusion Matrix'. This brings us to two of the most commonly used metrics to evaluate a classification model: Sensitivity and Specificity
- Predictions on test data set

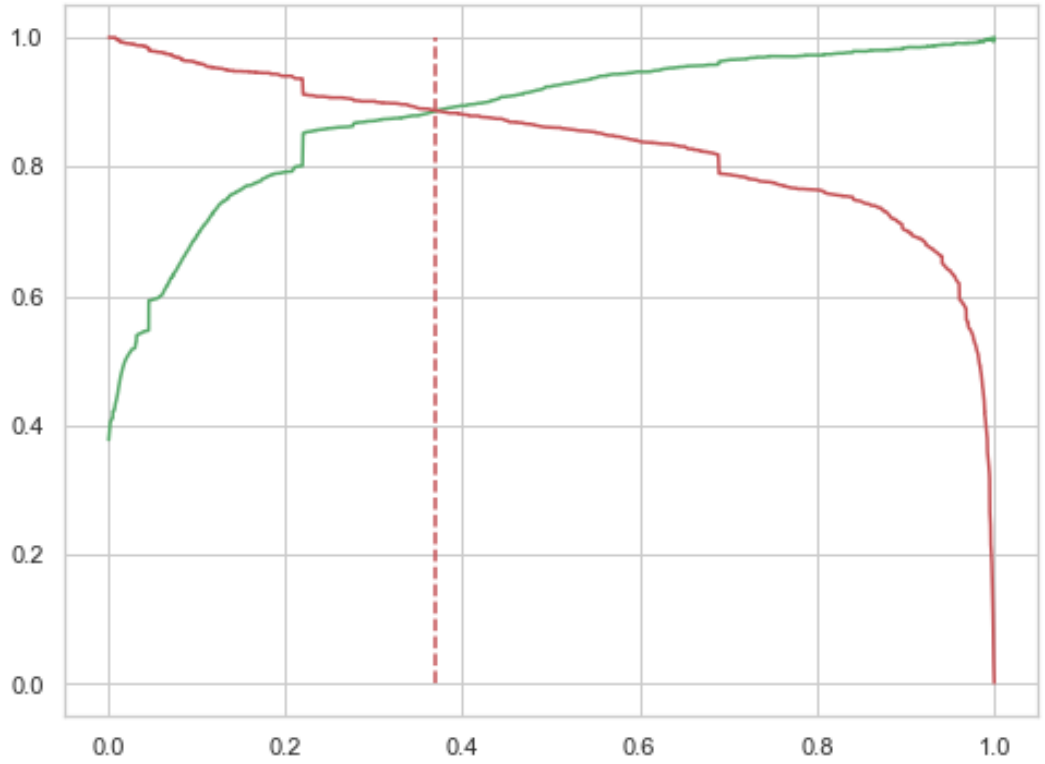
ROC Curve





- From the 1st graph, Area under the curve obtained is – 0.97
- From the 2nd graph, 0.28 is the optimum point to take it as a cutoff probability

Trade off curve between precision and recall



Precision and Recall are traded off at 0.37.
Therefore, it is safe to assume that any Prospect Lead with a Conversion Probability of more than 37% qualifies as a hot Lead.

Final Model Evaluation Metrics – Summary

- After evaluating the model using the train dataset and making predictions on the test dataset, the derived metrics are as follows:
  - Accuracy on train and test data respectively: 91.25% and 92.06%
  - Sensitivity on train and test data respectively: 90.23% and 88.95%
  - Specificity on train and test data respectively: 91.85% and 94.04%
  - Precision on train and test data respectively: 87% and 90%
  - Recall on train and test data respectively: 90% and 89%

Recommendations

- To maximize conversion and minimize pointless phone calls when the company has "limited time and resources," it should contact "Hot leads," or those leads who have more than 80% of conversion probability.
- The company should contact all of the "potential leads" when it has "ample resources and time" available. However, since it has plenty of time on its hands, it should also concentrate on clients with lower conversion rates in order to raise the lead conversion rate as a whole.

Conclusion

- So, the model evaluation on the train set is complete and the model seems to be doing a good job. The metrics seem to hold on the test dataset as well. So, it looks like we have created a decent model for the lead scoring dataset as the metrics are decent for both the training and test datasets.