LEAD SCORE CASE STUDY

PROBLEM STATEMENT

- Company need help to select the most promising leads, i.e., the leads that are most likely to convert into paying customers.
- The company requires us 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 has given a ballpark of the target lead conversion rate to be around 80%.

BUSINESS OBJECTIVE

- There are quite a few goals for this 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.

ANALYSIS

- Analysis of the datasets has been done in the Python (**Jupyter notebook**)
  - Attached the notebook for the step-by-step procedure that we have done as part of this case study.
  - What have we done ??
    - Starting by importing the Leads.csv
    - Check the structure of the data (Its normal routine check) (Info, Describe, Shape, etc..).
    - Data preparation
      - (Data quality check) Finding percentage of missing values of all the columns.
      - (Handling missing values) Remove columns with high missing values.
      - Checked outliers for the numerical columns and capped them.
      - Encoding (One hot encoding).

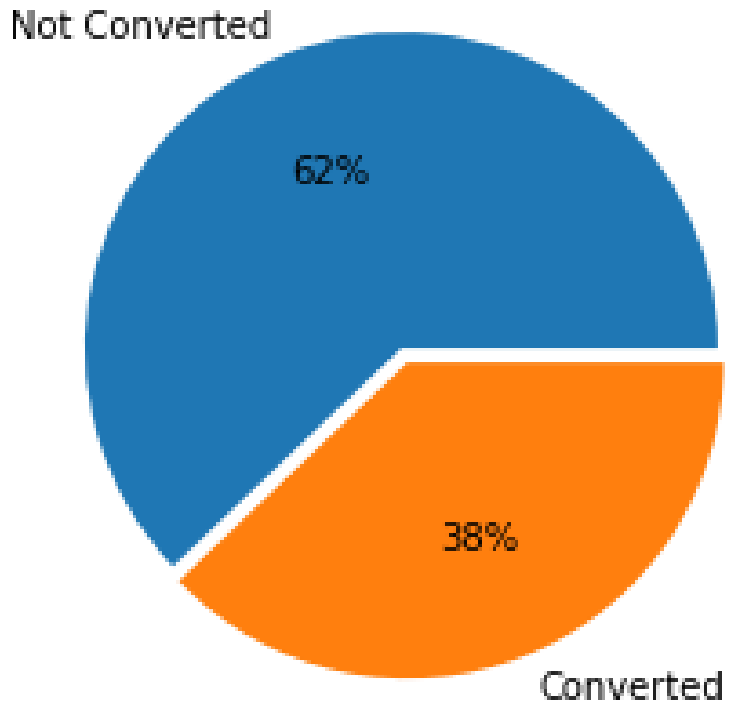
ANALYSIS CONT…

- EDA
  - Performed Univariate Analysis
  - Performed Bivariate Analysis
  - Outlier detection
  - Checking the data imbalance.
- Dummy variable creation
- Test-Train split
- Feature scaling
- Looking at correlations.
- Model Building
- Model evaluation
- Conclusion.

USED LIBRARIES

IMBALANCE RATIO

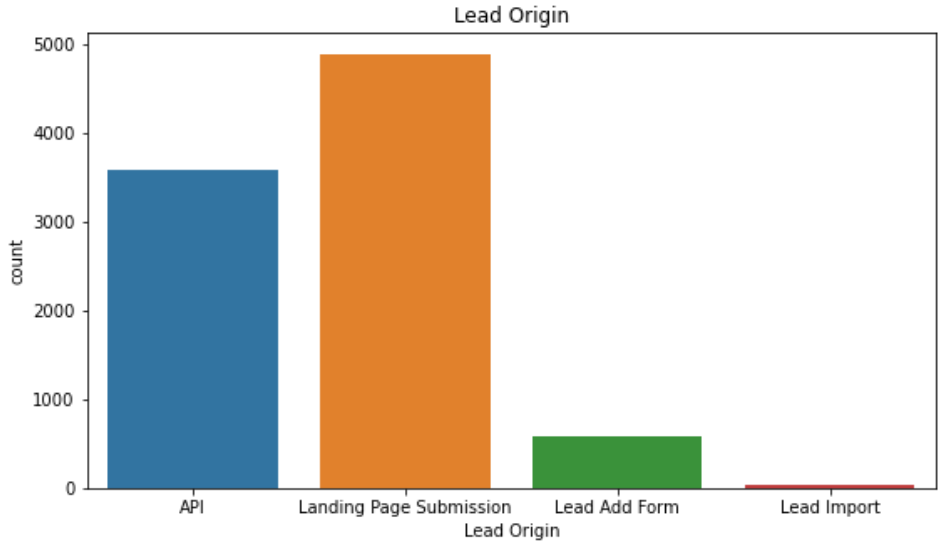
Percentage of Converted vs Non-Converted leads

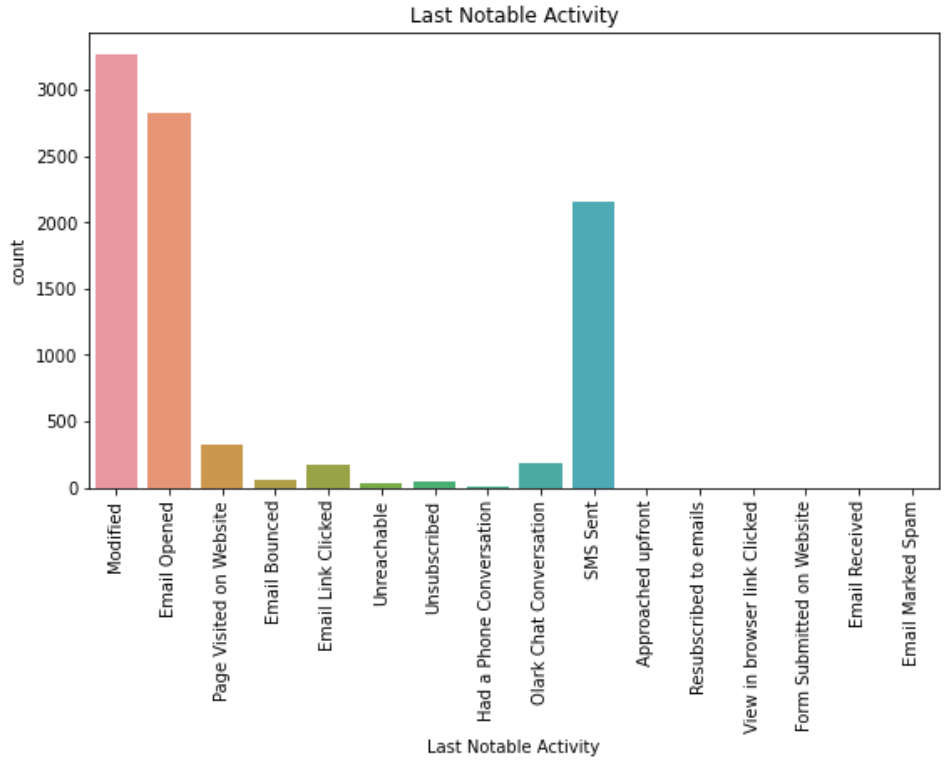


Imbalance Ratio:
62: 38

UNIVARIATE ANALYSIS

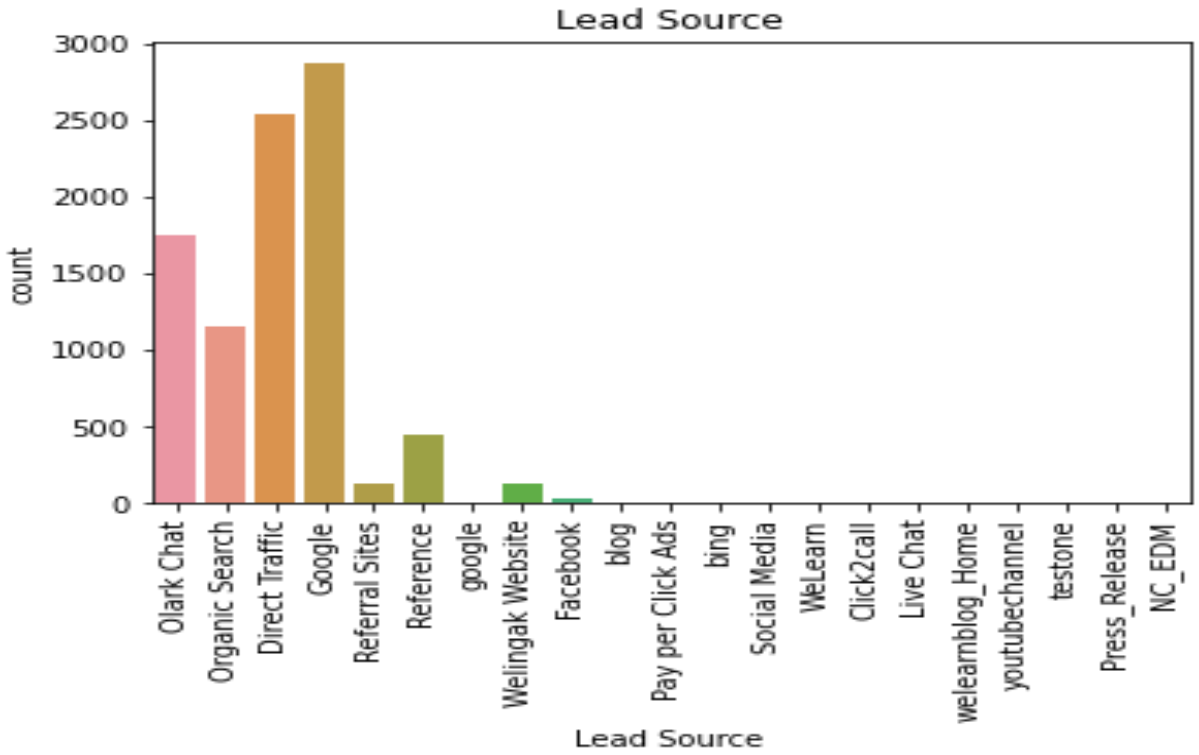
LEAD ORIGIN && LAST NOTABLE ACTIVITY





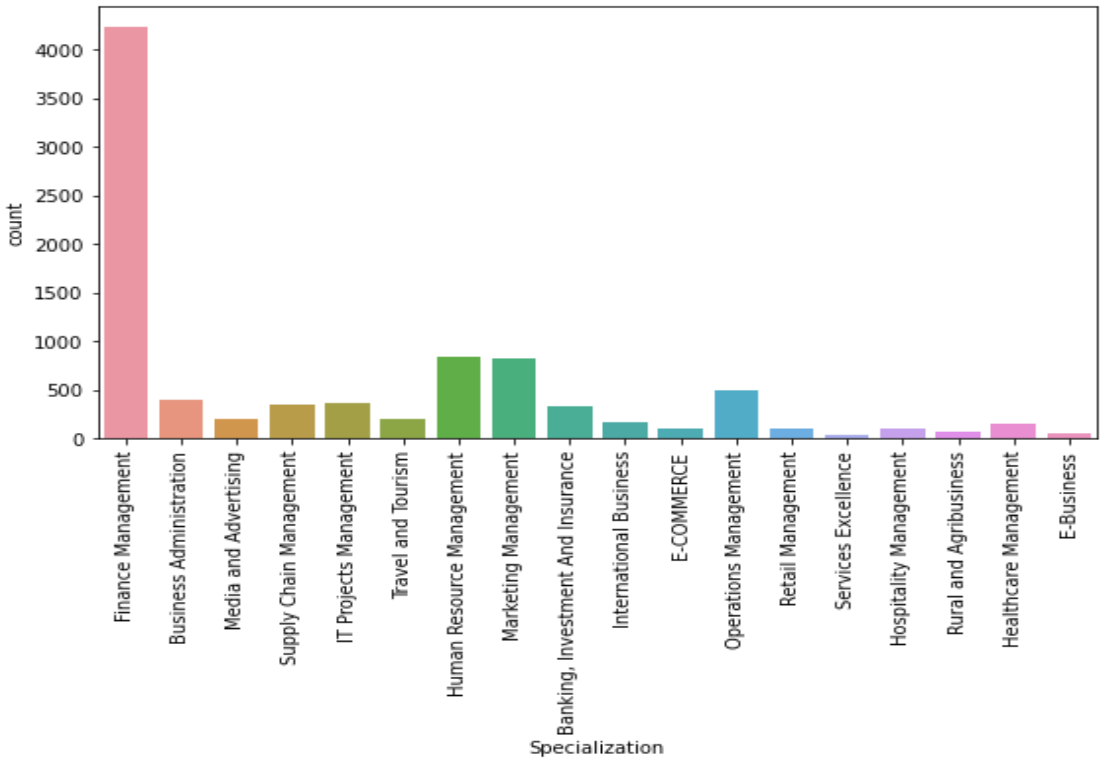
- The lead origin `Landing Page Submissions` are more compared to API and lead add form and lead imports are very less.

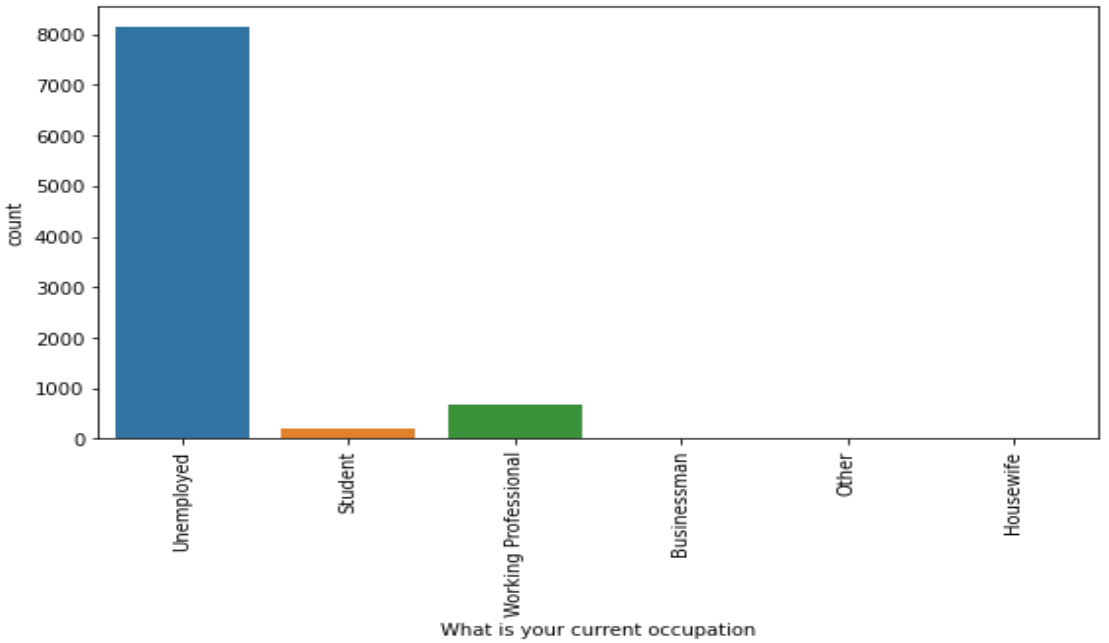
LEAD SOURCE



- Google & Direct traffic are the major leading sources

SPECIALIZATION && WHAT IS YOUR CURRENT OCCUPATION

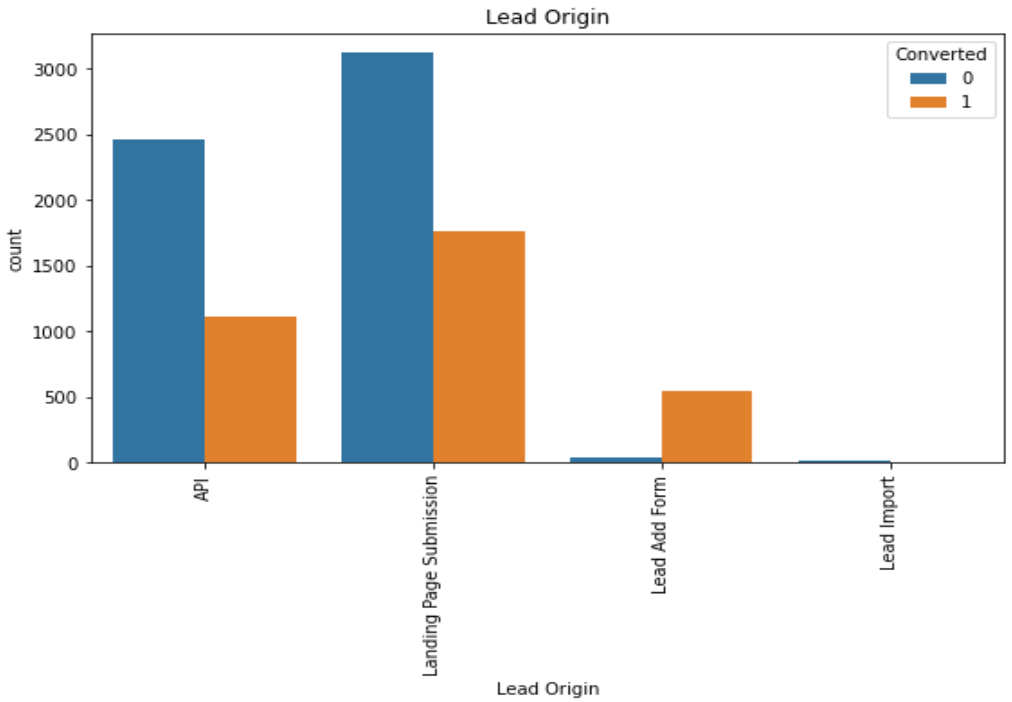


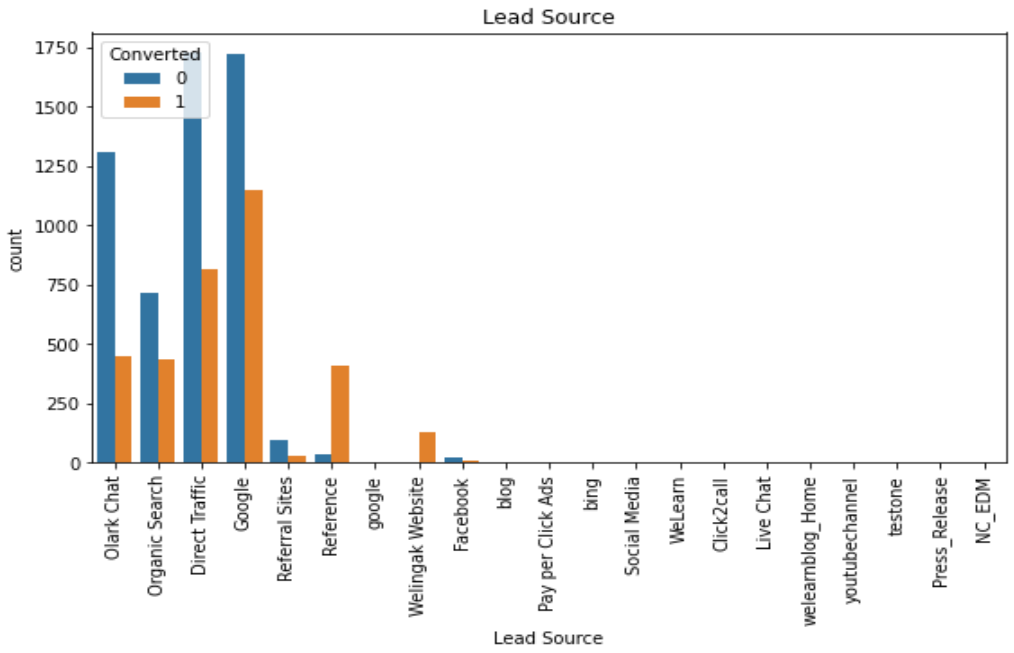


- Most of the customers are from finance management specialization before.
- Majority of the customers are unemployed and there are no customers from businessman/Others/Housewife, very less are students and some of them are working professionals

BI-VARIATE ANALYSIS

LEAD ORIGIN, LEAD SOURCE VS CONVERTED

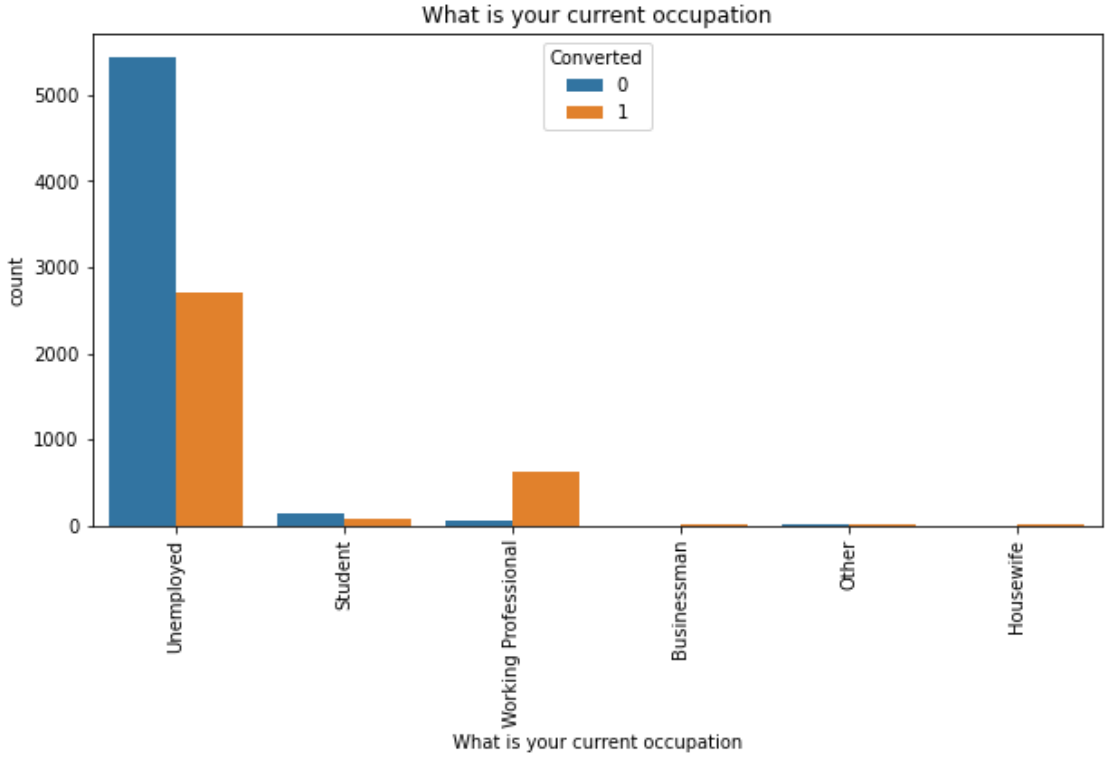




API and Landing Page Submission has less conversion rate.

- The count of leads from the Lead add form is low but the conversion rate is very high
- Lead Import has very less count as well as conversion rate and hence can be ignored
- The count of leads from the Google and Direct Traffic is maximum
- The conversion rate of the leads from Reference and Welingak Website is maximum when compared to not converted leads.

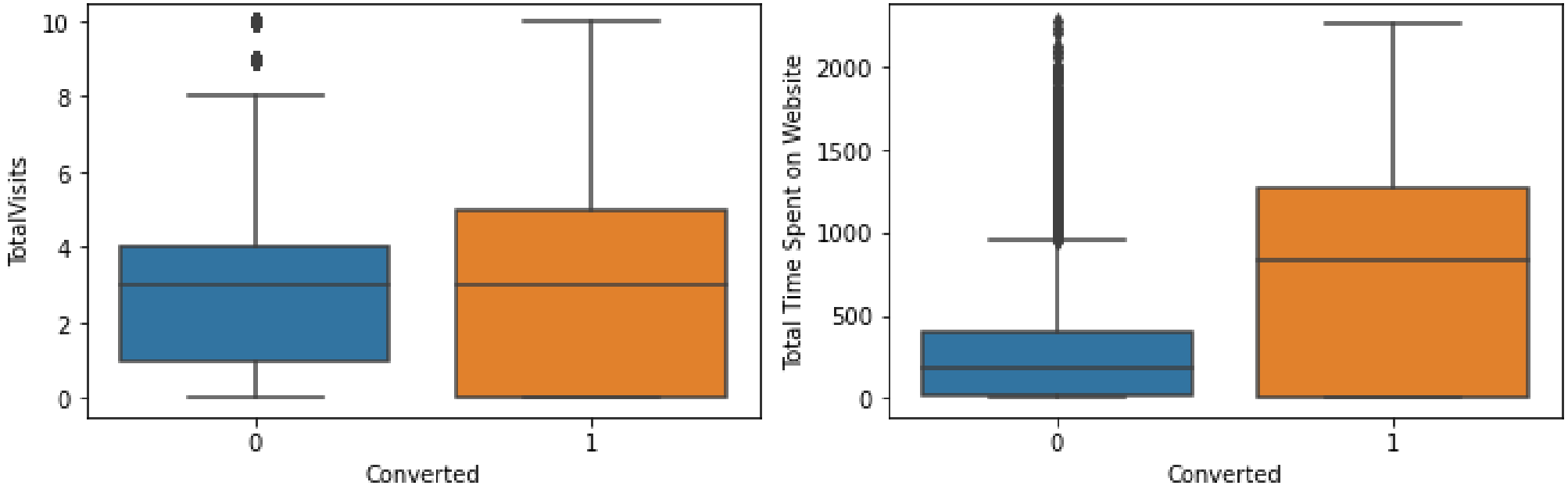
WHAT IS YOUR CURRENT OCCUPATION VS CONVERTED



- Majority of the customers are unemployed (both leads and non-leads) and there are very few customers from businessman/Others/Housewife, very less are students and some of them are working professionals

- Working professionals has highest conversion rate

TOTAL VISITS", "TOTAL TIME SPENT ON WEBSITE VS CONVERTED



- The median of both the conversion and non-conversion are same and hence nothing conclusive can be said using this information
- Users spending more time on the website are more likely to get converted

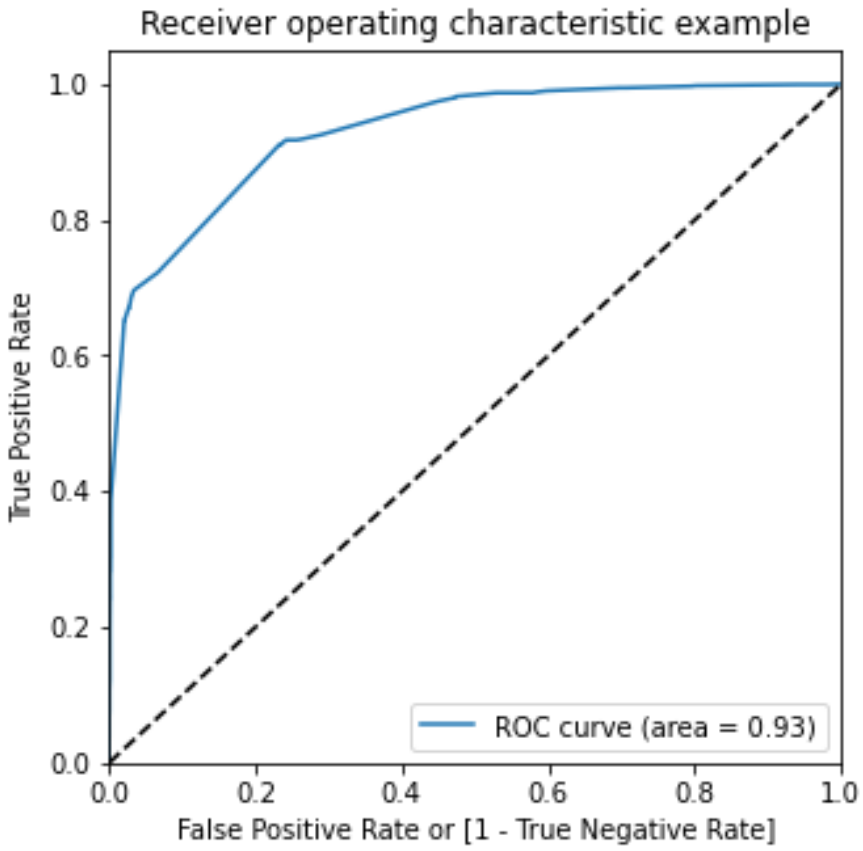
OVERALL SUMMARY

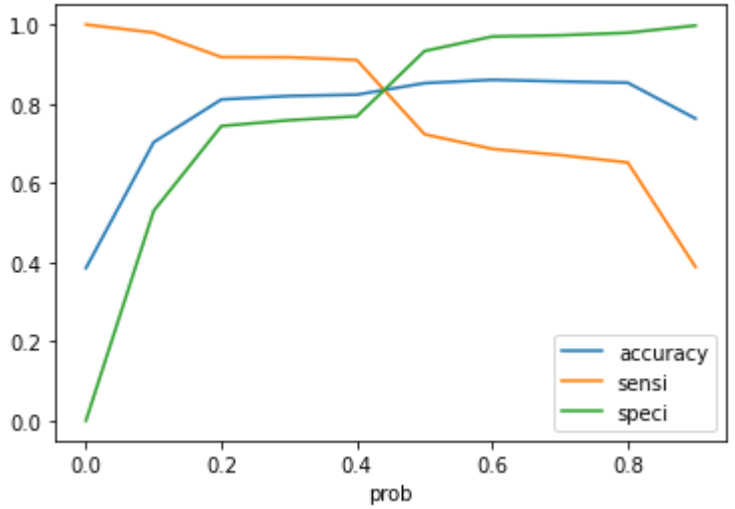
- We also observed that there are multiple columns which contains data of a single value only. As these columns do not contribute towards any inference, we can remove them from further analysis
- Websites can be made more appealing to increase the time of the Users on websites
- We should focus on increasing the conversion rate of those having last activity as Email Opened by making a call to those leads and try to increase the count of the ones having last activity as SMS sent
- To improve the overall lead conversion rate,
  - We need to focus on increasing the conversion rate of 'API' and 'Landing Page Submission' Lead Origins and increasing the number of leads from 'Lead Add Form'
  - We need to focus on increasing the conversion rate of 'Google', 'Olark Chat', 'Organic Search', 'Direct Traffic' and also increasing the number of leads from 'Reference' and 'Welingak Website'
  - We need to increase the number of Working Professional leads by reaching out to them through different social sites such as LinkedIn etc. and also on increasing the conversion rate of Unemployed leads

MODEL BUILDING & MODEL EVALUATION

- 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.
- Model Evaluation (With probability cutoff of 0.5)
  - Accuracy – 85.2%
  - Sensitivity – 72.8%
  - Specificity – 93%

ROC CURVE AND OPTIMUM VALUE OF THE CUTOFF





- Finding Optimal Cut off Point
- 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 0.41

FINAL OBSERVATION.

- Comparing the values obtained for train and test sets
  - Train Set:
  - - Accuracy: 81.7%
  - - Sensitivity: 91.7%
  - - Specificity: 75.6%

  - Test Set:
  - - Accuracy: 80.0%
  - - Sensitivity: 91.7%
  - - Specificity: 75%

CONCLUSION

- Here, the logistic regression model is used to predict the probability of conversion of a customer.
- The logistic regression model predicts the probability of the target variable having a certain value, rather than predicting the value of the target variable directly. Then a cutoff of the probability is used to obtain the predicted value of the target variable.
- Optimum cut off is chosen to be 0.41 i.e., any lead with greater than 0.41 probability of converting is predicted as Hot Lead (customer will convert) and any lead with 0.41 or less probability of converting is predicted as Cold Lead (customer will not convert)
- The top three features in the final model are:
  - 'Tags_Closed by Horizzon'
  - Tags_Lost to EINS',
  - 'Tags_Will revert after reading the email'
- The final model has Sensitivity of 0.917, this means the model is able to predict 91.7% customers out of all the converted customers, (Positive conversion) correctly.

THANK YOU