LEAD SCORE CASE STUDY

SUBMITTED BY,

DHARANI ARUMUGAM

PROBLEM STATEMENT

- An X education 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, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

GOALS

- 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.

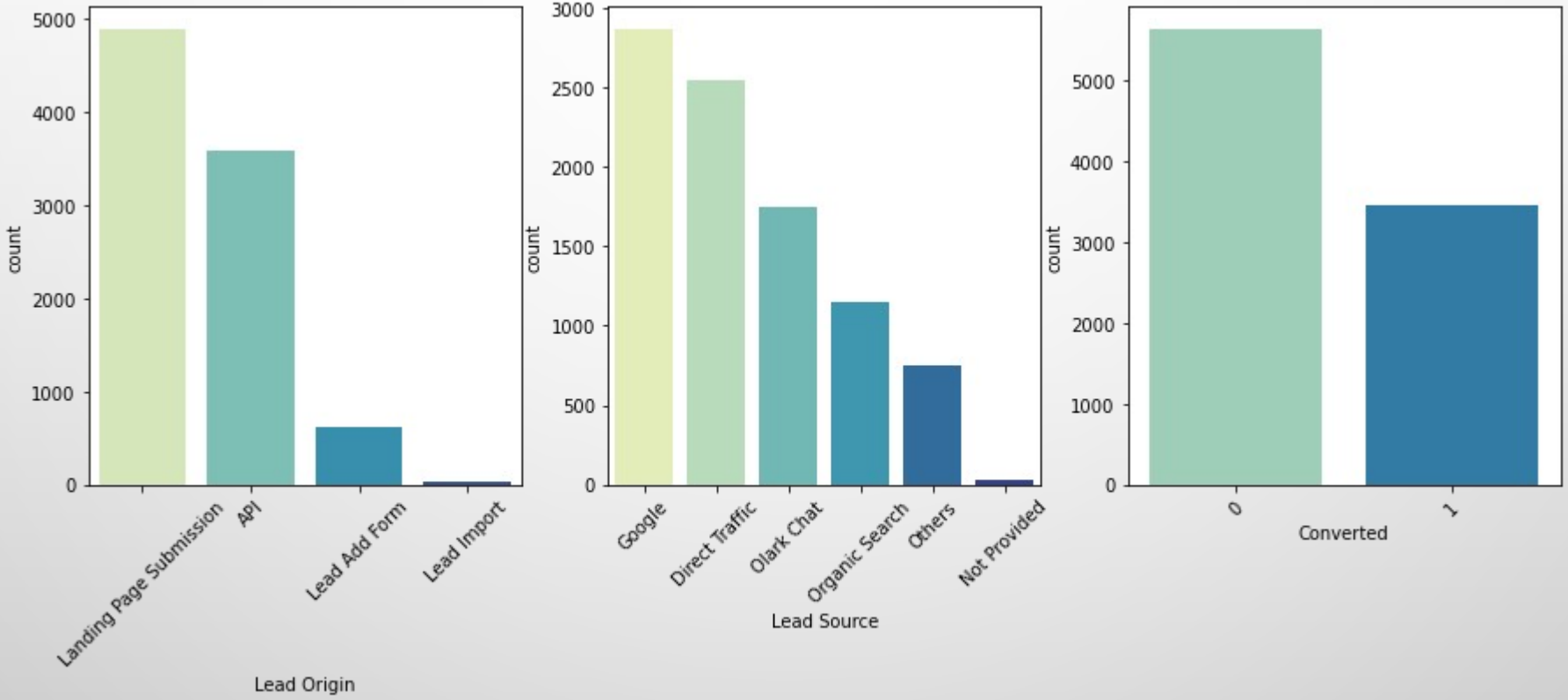
BUSINESS OBJECTIVE

- X education wants to know the potential hot leads
- For that they want to build a ml model to identify them
- Focus on the hot leads

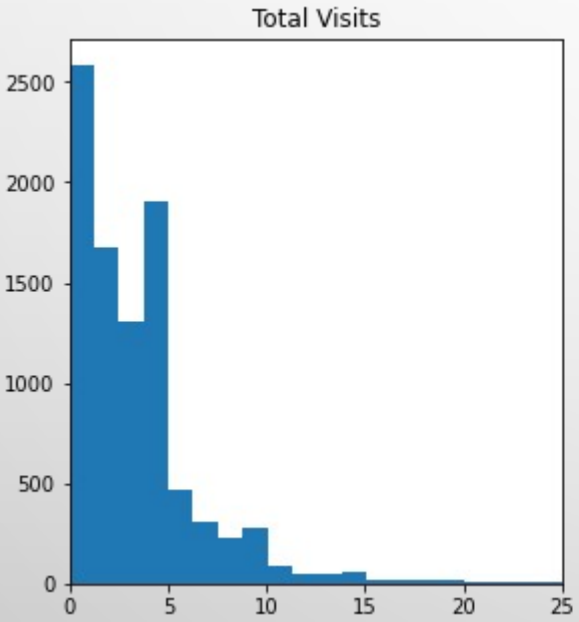
SOLUTION APPROACH

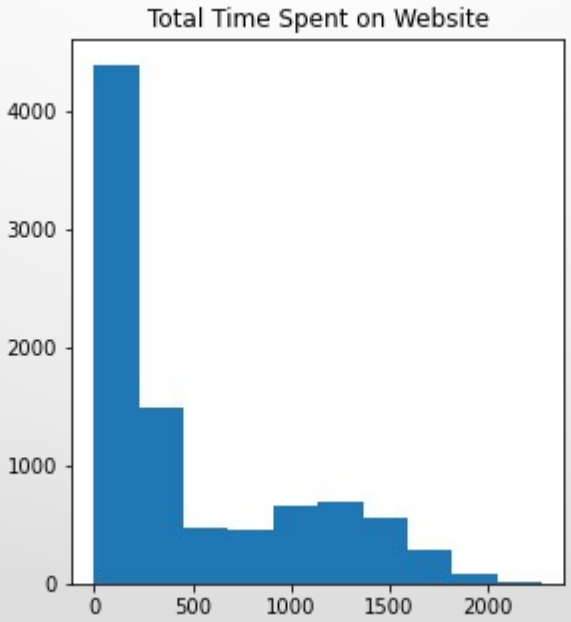
1. Data cleaning and manipulation
   - ‘Select’ values to changes as np.Nan
   - Drop columns with high missing values
   - Drop skewed categorical colmns
   - Combine categiries which has low percentage
   - Perform imputation on data
2. EDA
   - Uni-variate Analysis
   - Bi-Variate Analysis
   - Outlier Analysis
   - Handling Outliers
3. Generate dummies
4. Split train-test and perform Scaling
5. Model Building
6. Making prediction
7. Model evaluation

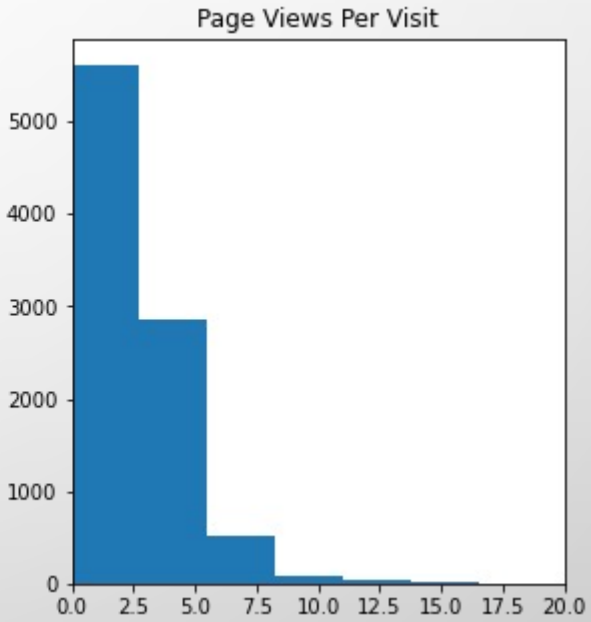
EDA – UNIVARIATE CATEGORICAL



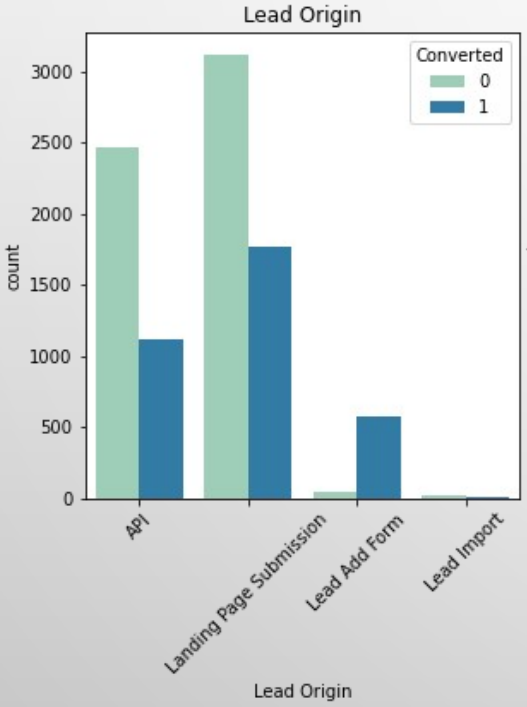
EDA – UNIVARIATE NUMERICAL

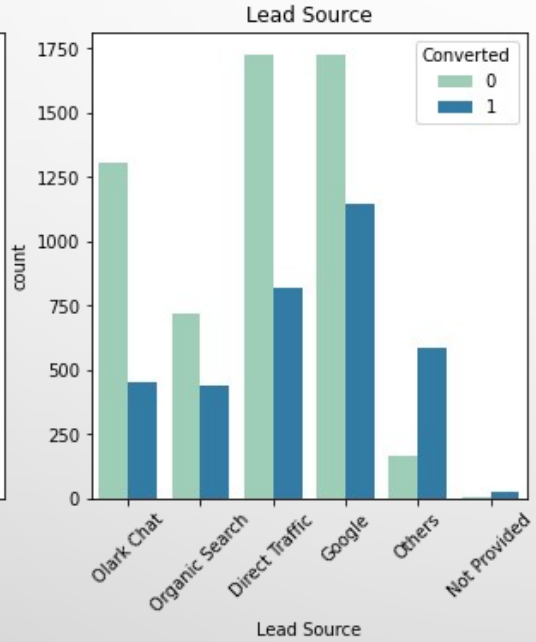


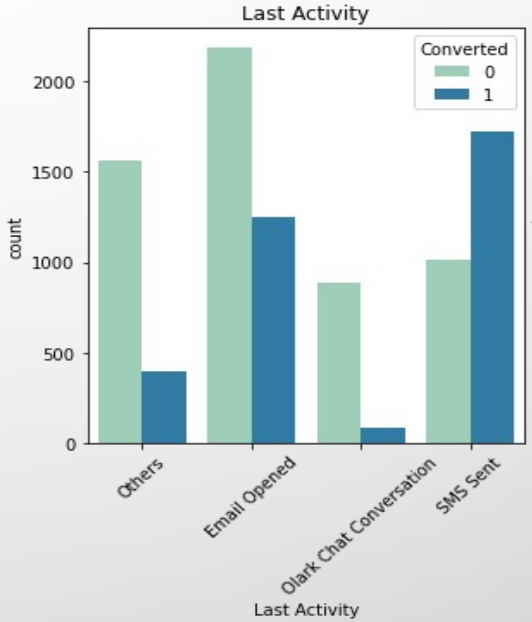


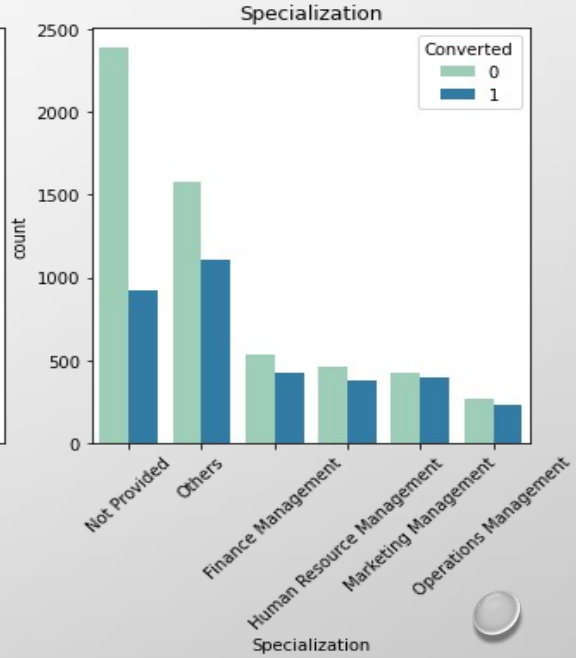


EDA – BIVARIATE



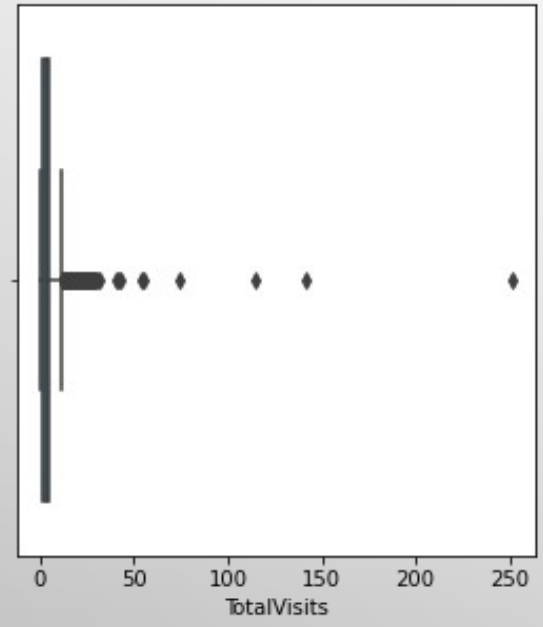


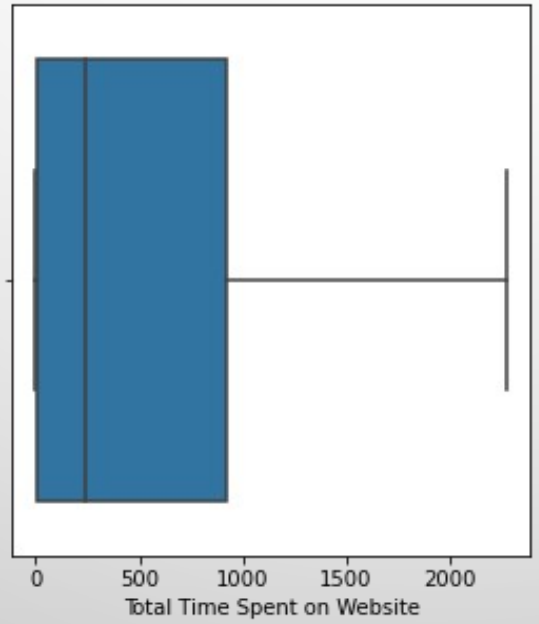


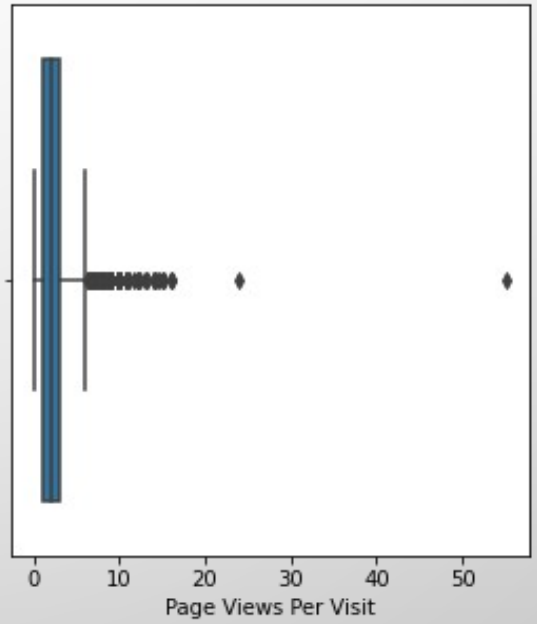


OUTLIER ANALYSIS

- Ouliter's in the data are capped with 0.95 and 0.05 range





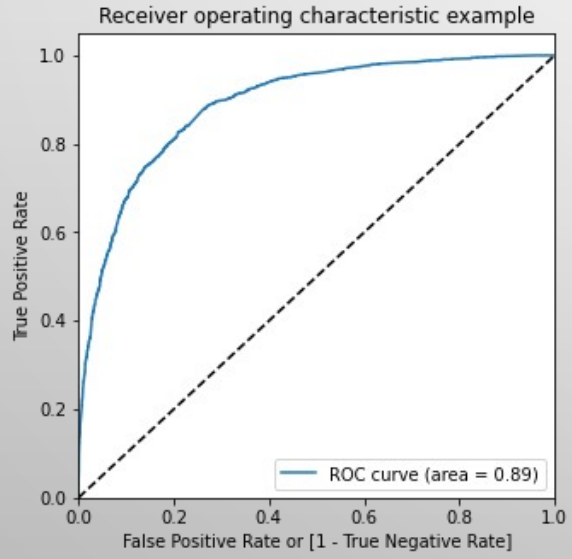


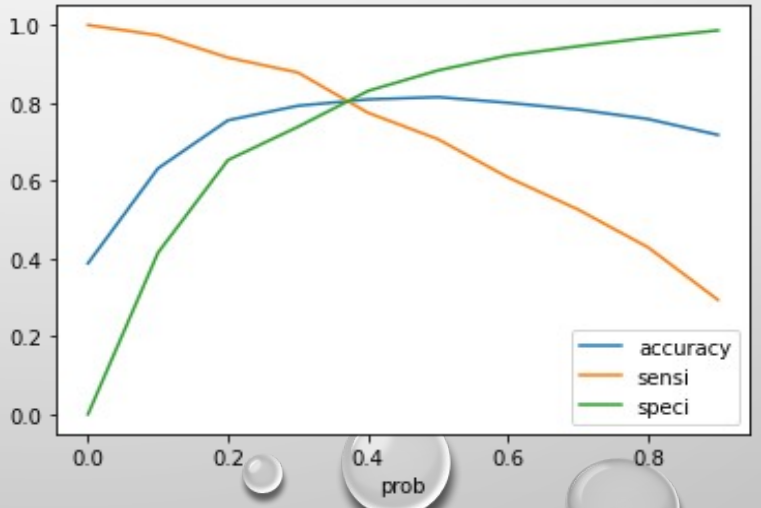
MODEL BUILDING

- Split data into training and test sets with 70: 30 ratio
- Use RFE for feature selection
- Run RFE with 15 variables as output
- Building stable model by removing variable which has p-value greater than 0.05 and VIF value greater than 5
- Predictions on test data set – calculating the lead score
- Overall accuracy is about 81%

ROC –CURE (OPTIMAL CUT-OFF)

- Find the optimal cut –off using roc curve
- Optimal cut-off is the one which has balanced sensitivity and specificity
- Here optimal cut off is 0.4





CONCLUSION

It was found that the variables that mattered the most in the potential buyers are (in descending order) : - lead origin_landing page submission

- totalvisits
- what is your current occupation_unemployed
- total time spent on website
- lead source_olark chat
- last activity_others
- last notable activity_sms sent
- last activity_olark chat conversation
- last notable activity_otheractivity
- lead origin_lead add form
- what is your current occupation_working profes
- - what is your current occupation_studentkeeping these in mind the X education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.