Lead score Case study

By

Harika Sadineni

Tushar Anand

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.
- . 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.

Business Objective:

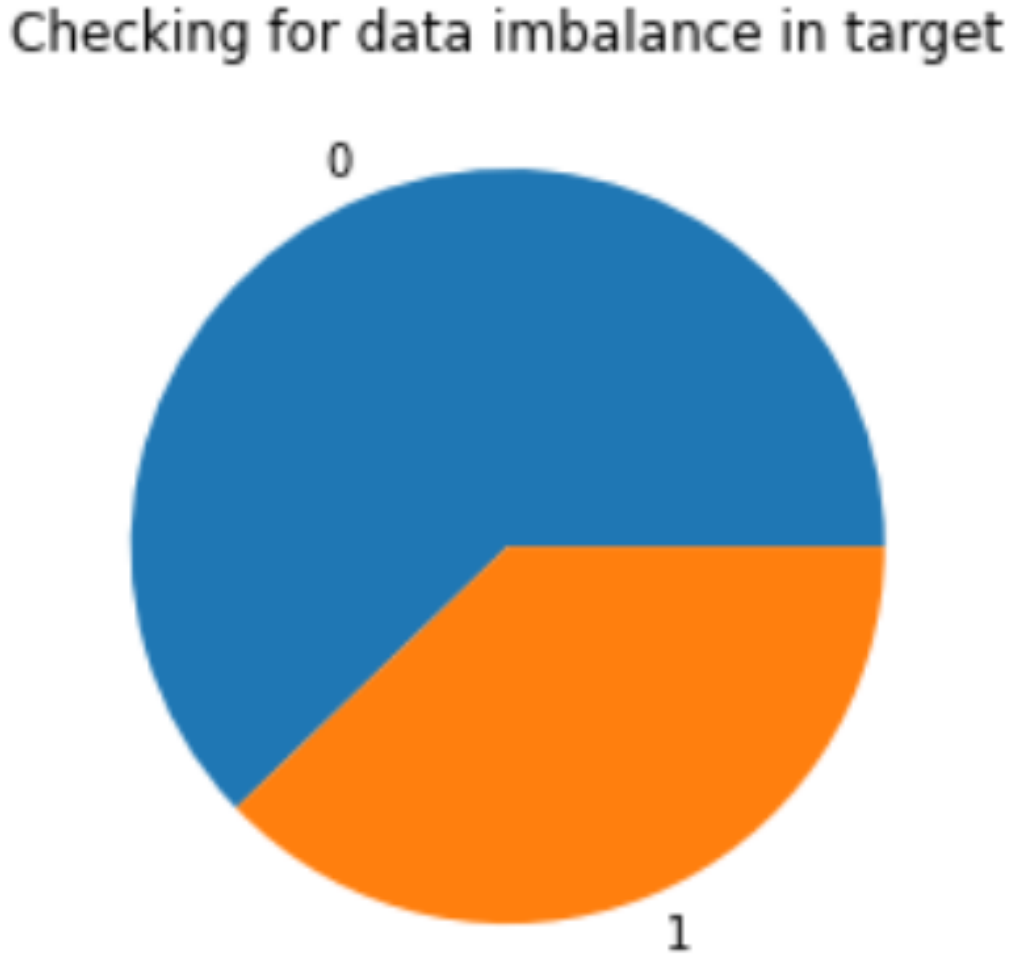
- To make this process more efficient, the company wishes to identify the most potential leads, also known as __'Hot Leads'__. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone

The flow of the Model:

1. Data cleaning and Data pre-processing:
   - Check and handle for missing values.
   - Check for duplicated values and handle them
   - Check for data imbalance and handle it.
   - Check and handle outliers.
   - Imputing missing values if necessary.
2. EDA:
   - Univariate analysis: check the distribution and value counts
   - Bivariate: check the relationship with the target variable.
   - Check multi correlation between the features.
3. Splitting the data
4. Column transformation: dummy variables and scaling
5. Model training using Logistic regression
6. Evaluation
7. Conclusions

Data pre-processing

- Index columns – Prospect ID and Lead Number are dropped as they are not useful for analysis.
- Single value features and binomial features with no as value are dropped as they are not useful for the model.
- After checking the value counts for all categorical features, the value counts less than 10 are groped as others to improve the variance.
- ‘Select’ is replaced with Nan
- Dropped the features that have more than 40 % missing values.
- The remaining missing values are imputed with random values from the respective features.
- Outliers are removed using the 99^{th} quantile.
- After Pre-processing we are left with 8863 rows and 13 columns

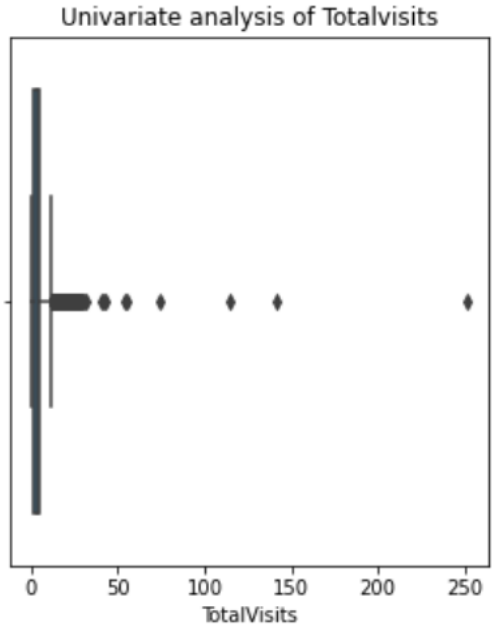


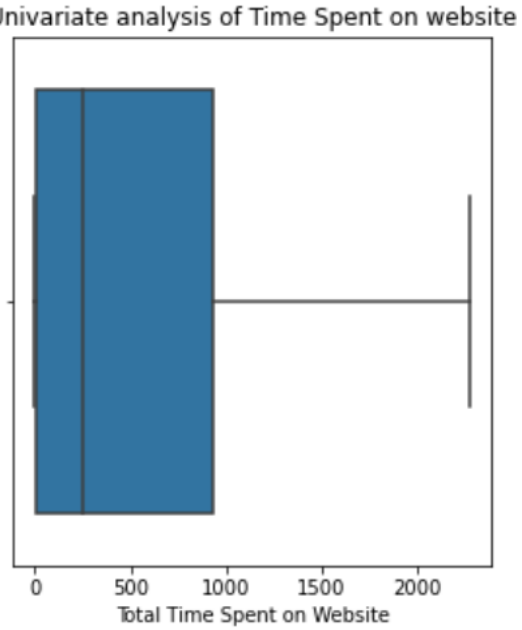
EDA

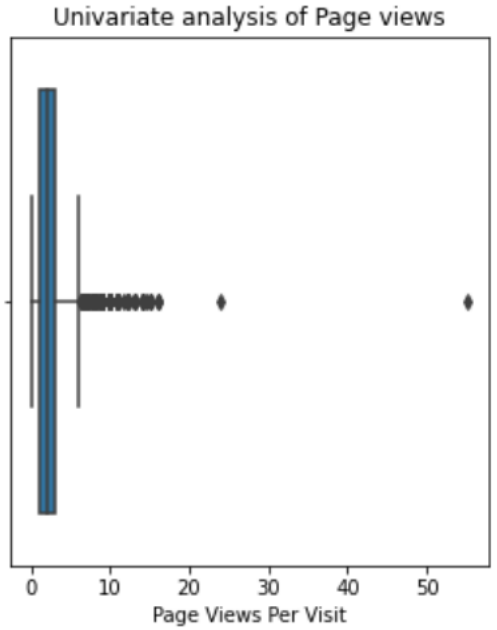
Inference:

As we can see there isn't much data imbalance in the target variable it is around 22%.

Checking for outliers

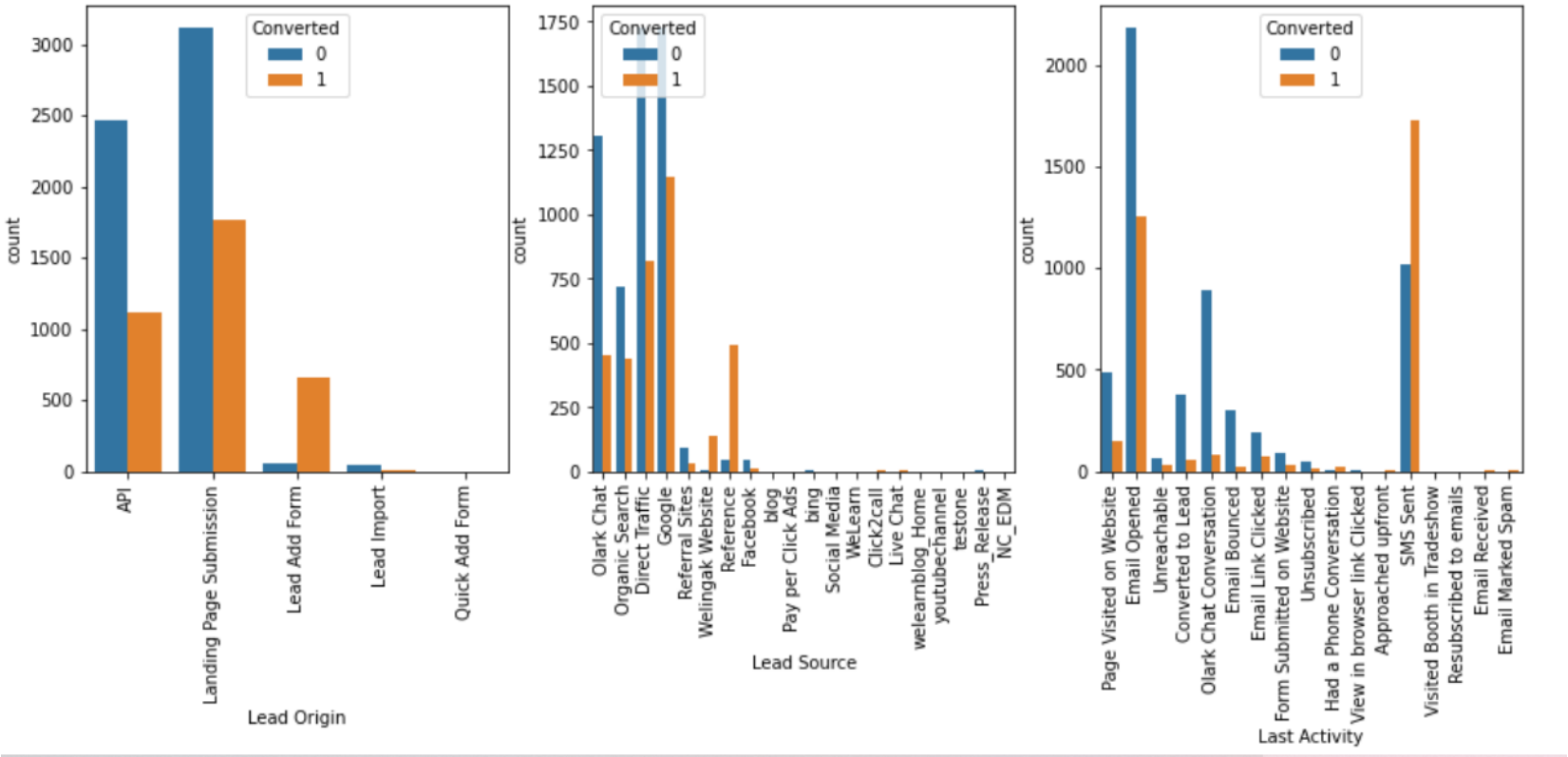






As we can see there are outliers in total visits and page views per visit.

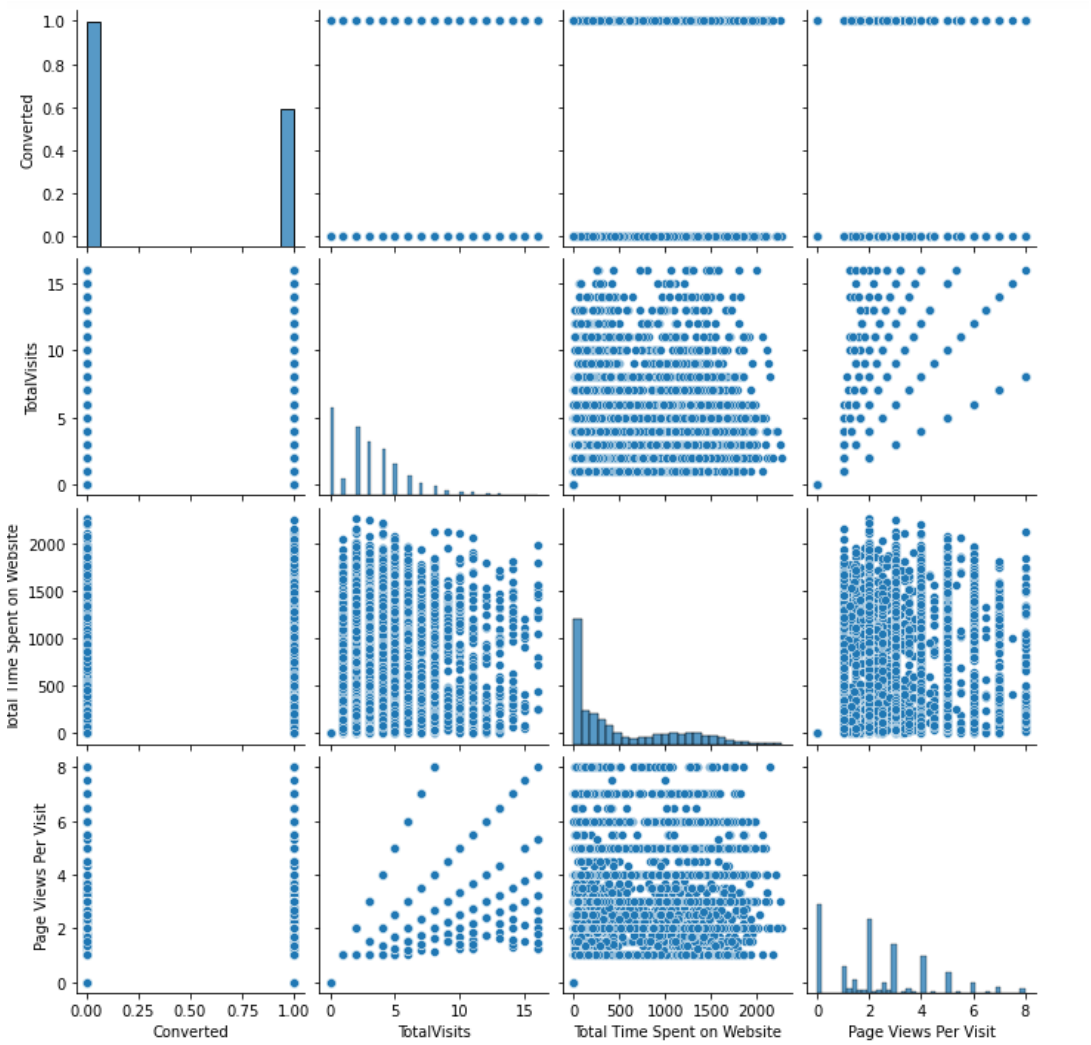
Bi variate: Categorical relation with the target



Inference;

- Most of the lead origin is from the landing page
- Lead source is majorly from google search
- Last activity Email or SMS have a higher conversion rate.

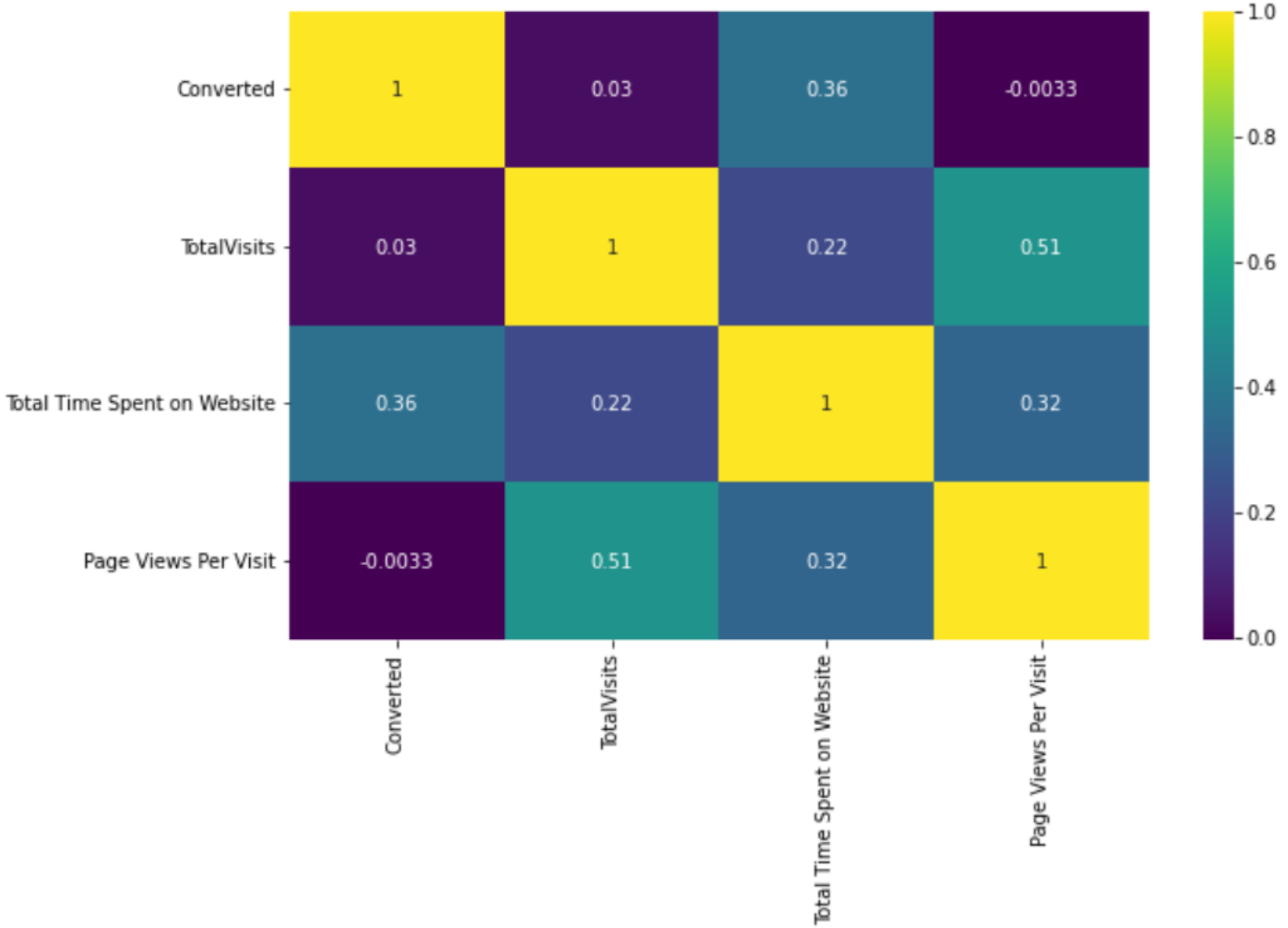
Bi variate: Numerical Features with target



Inference;

- The values of numerical features are right skewed
- There is a slightly linear relationship between total visits and page views
- Total time spent is showing any pattern. But has more value count.

Correlation map



Inference;

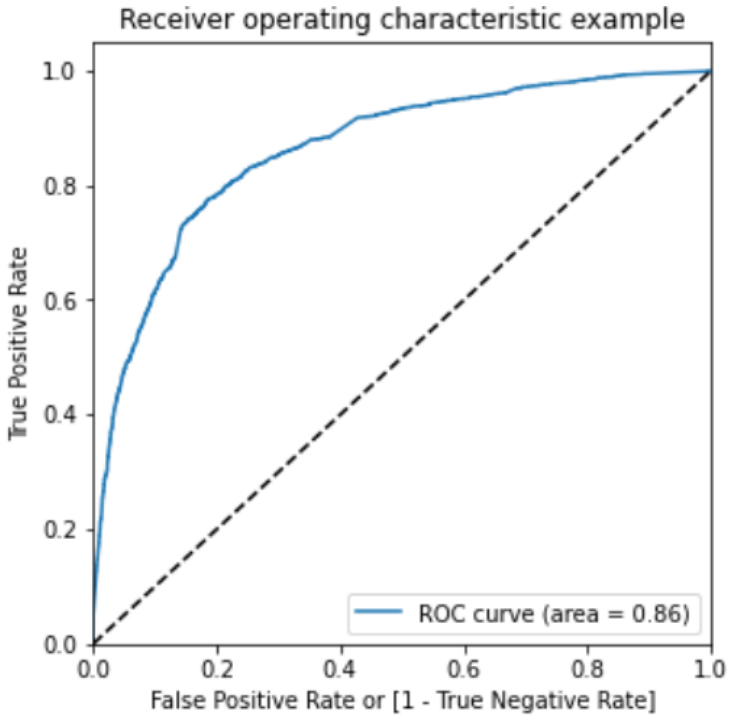
- Total time spent has a greater positive correlation with the target.
- Total visits and page views have a good correlation
- Total time spent and total visits also have a positive correlation.

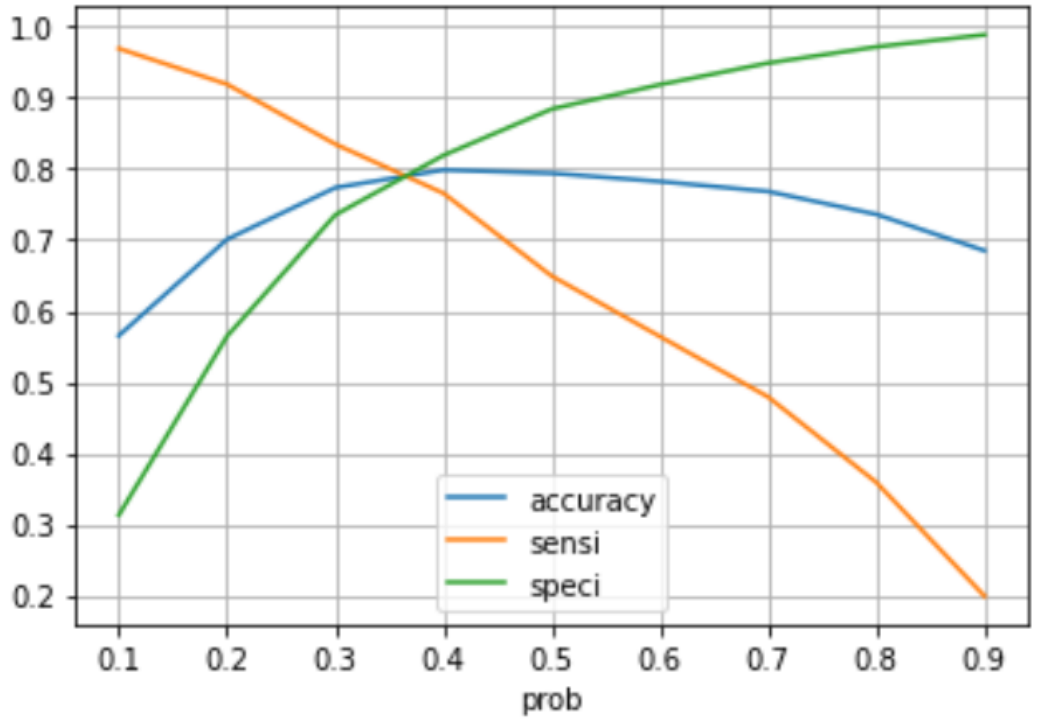
Splitting the data and column transformation

- The is split in 70 – 30 ratio
- The column transformer is created
- The numeric values are scaled using Standard scalar
- The categorical dummy variables are created using OneHotEncoder
- Total rows for analysis: 8863,
- Total columns for analysis: 114

Model Building

- A total of four models were trained.
- Model1: The first model was trained using all the features. We were able to achieve accuracy up to 81% but a very recall of 69% and precision of 76%.
- Model 2: The feature selection was done Using the RFE. 15 features were selected and used for the next model in which we were able to achieve an accuracy of 80% and still a very low recall of 67% and precision of 76%.
- Model 3: Feature importance is calculated using a Random forest algorithm and the top 20 features were used for the next model. The accuracy score got dropped so insignificant and high VIF features were dropped from the model.
- Model4: Features with good significance and low VIF were used to achieve an accuracy of 79%.





ROC and Cut-off curve

- We got 0.86 area under ROC which a good value.
- The Cut-off curve shows the optimum cut-off point to meet the sensitivity, specificity and accuracy.
- For our model we got 0.37 as the optimum cut-off point.

Conclusions

1. The company can set their Lead score cut-off at 41 i.e., any lead score above 41 can be considered potential and below it can be neglected.

2. The following are the features which matter the most to predicting potential customers:

1. Total Time Spent on Website
2. Last Activity
   a. SMS Sent
   b. Olark chat
   c. Modified
3. Total Visits
4. Lead Origin
   a. Landing page
   b.Lead add form
5. Page Views per visit
6. Lead Source
   a. Welinggak Website
   b. Google
7. Current occupation
   a. Working professionals

The company should focus on these features to increase potential leads. As per our EDA, the company can improve its performance in some of the features. For example, the company concentrate more on working professionals and also should try to get more referral lead sources as the conversion rate is high when the lead is from a referral.