Lead Scoring Assignment

Presentation By –

Ashwin Joshy

Hanumant Garad

Sarika Desai

DS C45

Background

- X Education , An education company named sells online courses to industry professionals
- Many interested professionals land on their website
- The company markets its courses on several websites 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
- 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%

Problem Statement

- X Education gets a lot of leads but its lead conversion rate is very poor
- 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
- We will help them to select the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- We are required to build a model wherein we need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance
- The CEO, in particular, has given a ballpark of the target lead conversion rate to be 80%.

Lead Conversion Process

1. Lead Generation
   - Advertisement on websites like Google etc.
   - Referrals
2. Visit to X Education website by these potential customers (professionals)
3. Visitors either provide Email id & Contact Details Or View videos etc.
4. Tele calling and Emailing activity to all the leads
5. ~30% leads get converted

Goals of 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.

SOLUTION APPROACH

Loading & Observing the past data provided by the Company

Univariate, Bivariate, and Heatmap for numerical and categorical columns

Performing pre-requisites for RFE and Logistic Regression

Duplicate removal, null value treatment, unnecessary column elimination, etc.

Outlier Treatment, Feature-Standardization

Selection of top features using RFE

Reduction of columns and Model re-building based on VIF and p-value

Parameters to evaluate are Accuracy, Recall, Precision, Specificity, and AUC

Automated Feature Selection

Manual Feature selection

Model Improvement

Final Model

Model Evaluation

Manual feature selection using VIF and p-value method

Selecting the final regression model based on VIF and p-value

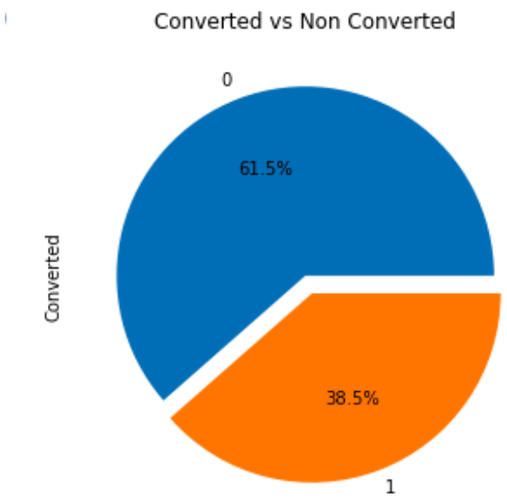
ANALYSIS

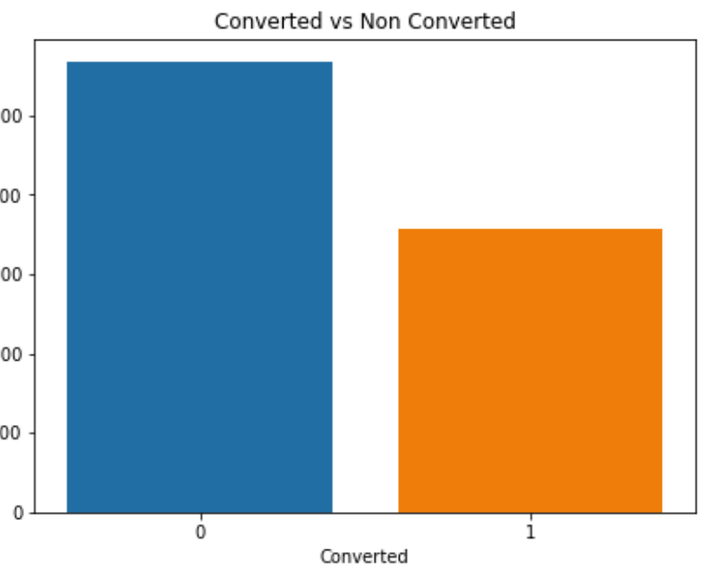
Initial Data Understanding

- The data has ~ 9k rows and 37 features
- 30 features are categorical, and 7 variables are numeric
- It seems few features such as 'TotalVisits','Total Time Spent on Website','Page Views Per Visit' has outliers
- No duplicate records in the dataset
- There are few features having single unique level e.g., "Receive More Updates About Our Courses", "Update me on Supply Chain Content"
- There are 17 variables having null values

EDA

- **Univariate Analysis**
  - Distribution of the Target variable 'Converted'
    - ✓ 1 - customers who are successfully enrolled for online course of X Education
    - ✓ 0 - customers who are not enrolled for online course of X Education

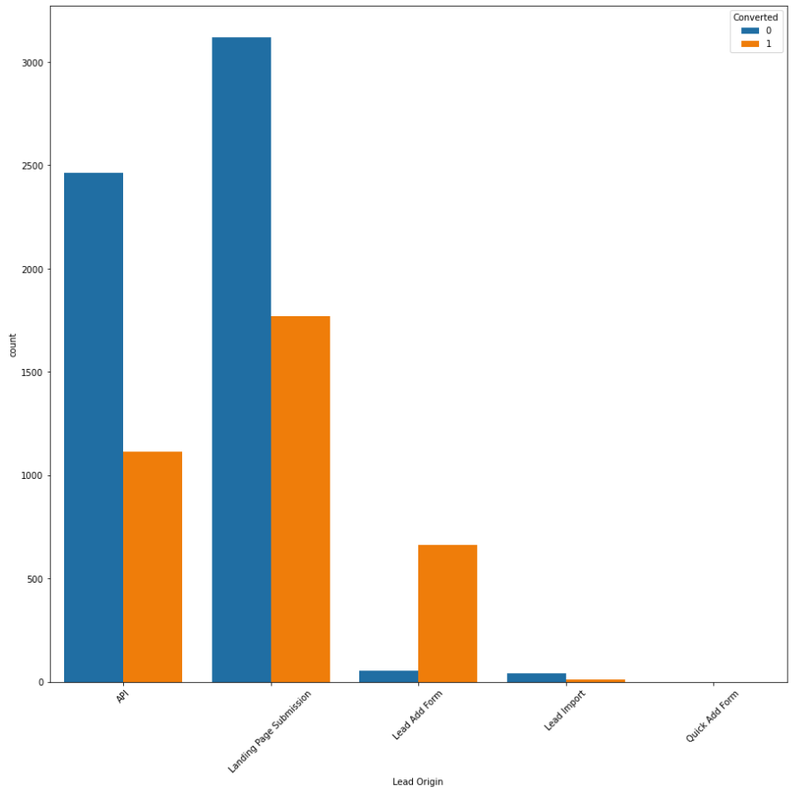


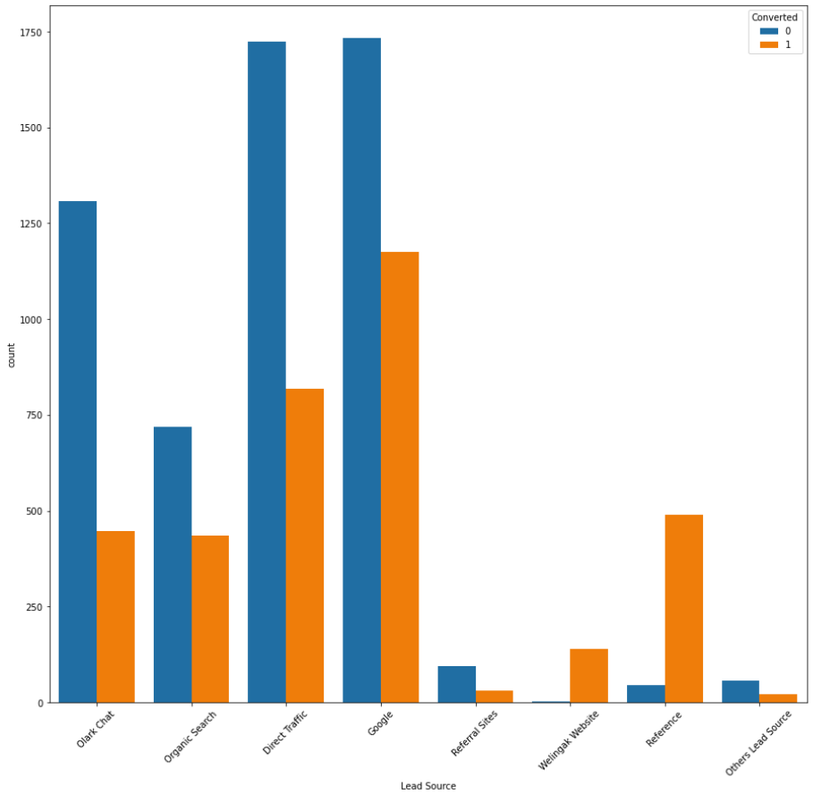


Conversion Rate is 38.5%

Univariate Analysis (Cont.)

- Lead Source variable has many unique levels having very less contributing % so it is better to group all that levels under single level.
- Let's consider values that shows < 1% contribution in the data
- **It seems the majority of leads comes from India ~70 and only ~4% leads comes from foreign countries so why not group all foreign countries under single value called "Foreign"**



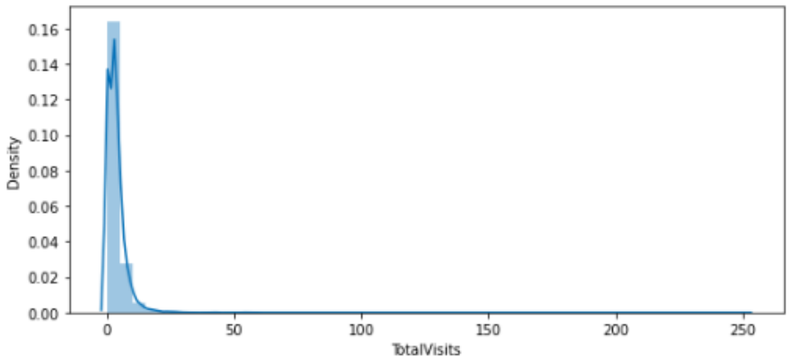


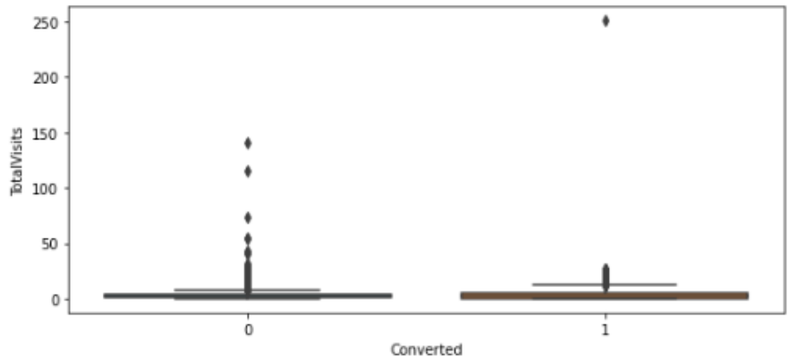
Univariate Analysis - Inferences

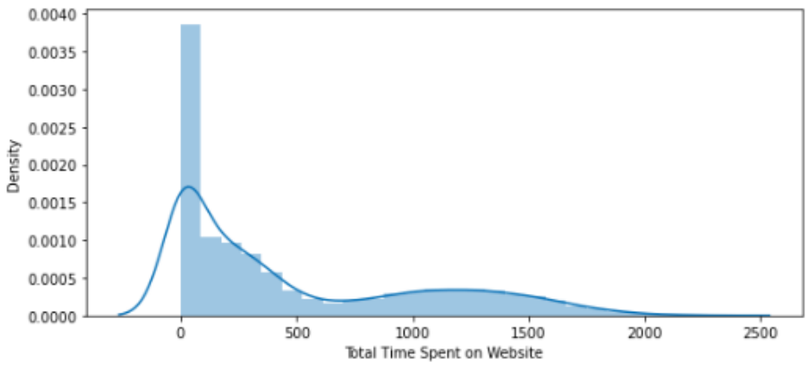
- "Lead Add Form" has highest conversion rate ~92 %
- "Quick Add Form can be ignored since it has very less values"
- "Welingak Website" is having highest conversion rate ~98 % , Others Lead Source and Google are having good conversion rate "Reference also has high conversion rate"
- Lead Conversion Rate is comparatively good for leads those wanted X Education shall send an email
- Customers whose Last Activity is "SMS Sent" has high probability of conversion.
- Slightly high conversion rate for customers who comes from India comparatively to other countries
- Working Professionals and Housewifes are more interseted for online courses and Students are relatively low because they got already enrolled for other study
- Customers whose Last Notable Activity is "SMS Sent" has high probability of conversion.

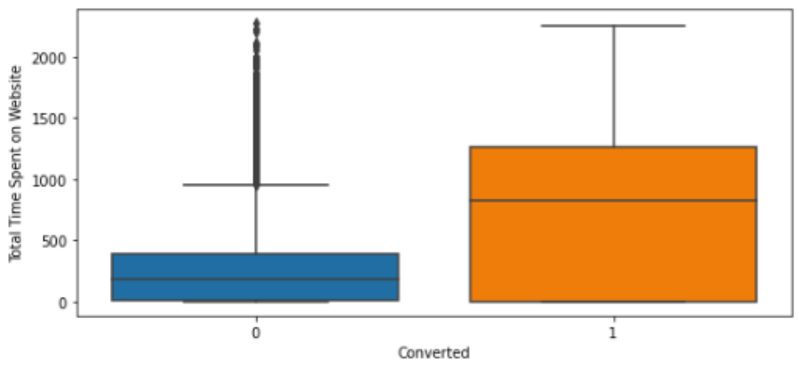
Outlier Handling

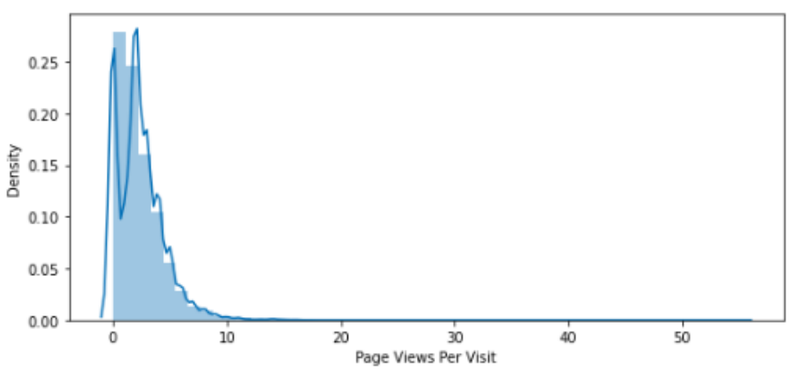
Data with outliers

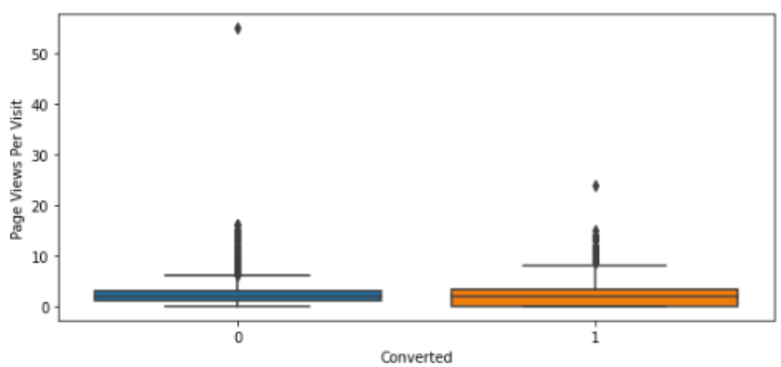




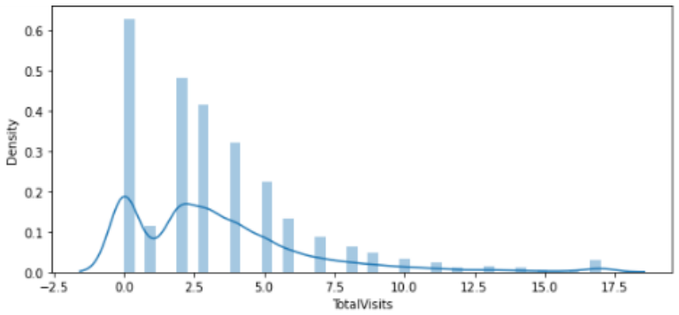


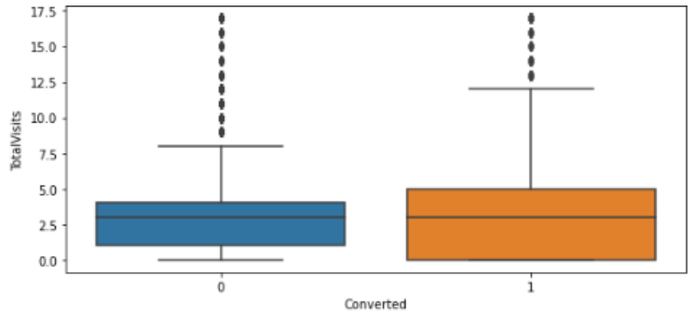


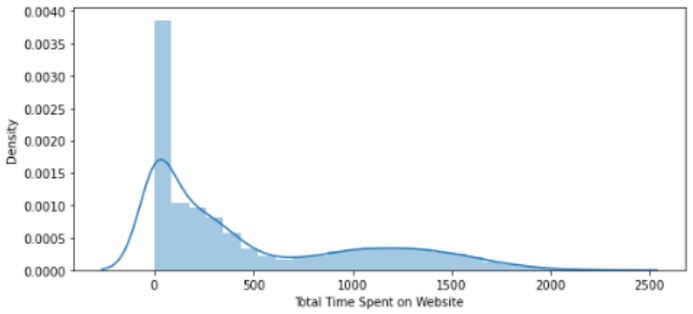


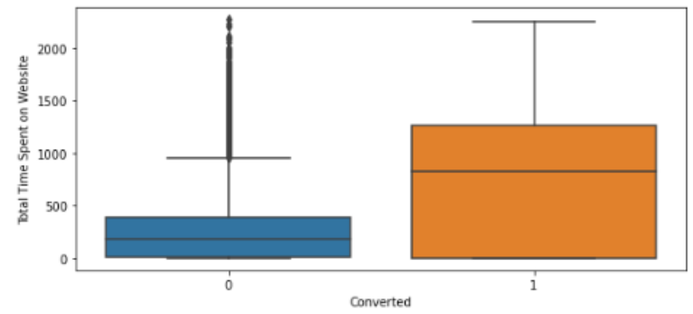


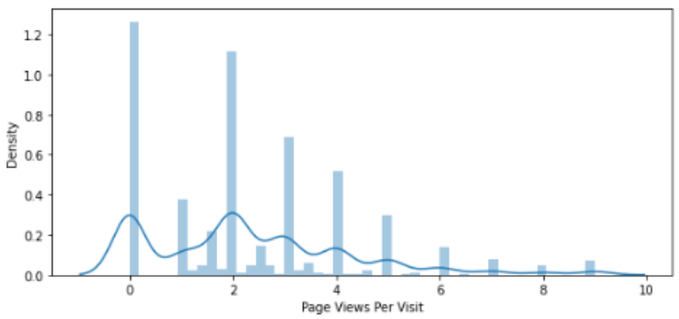
Outliers capped at 99 percentile

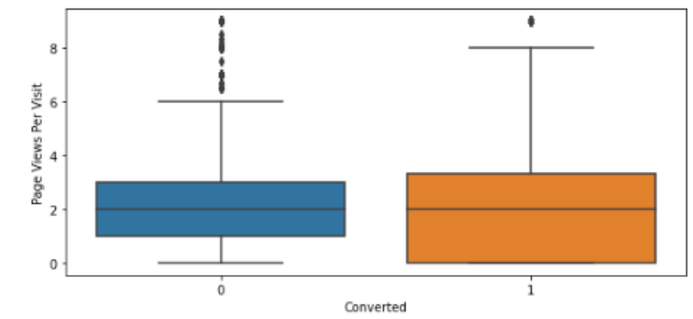








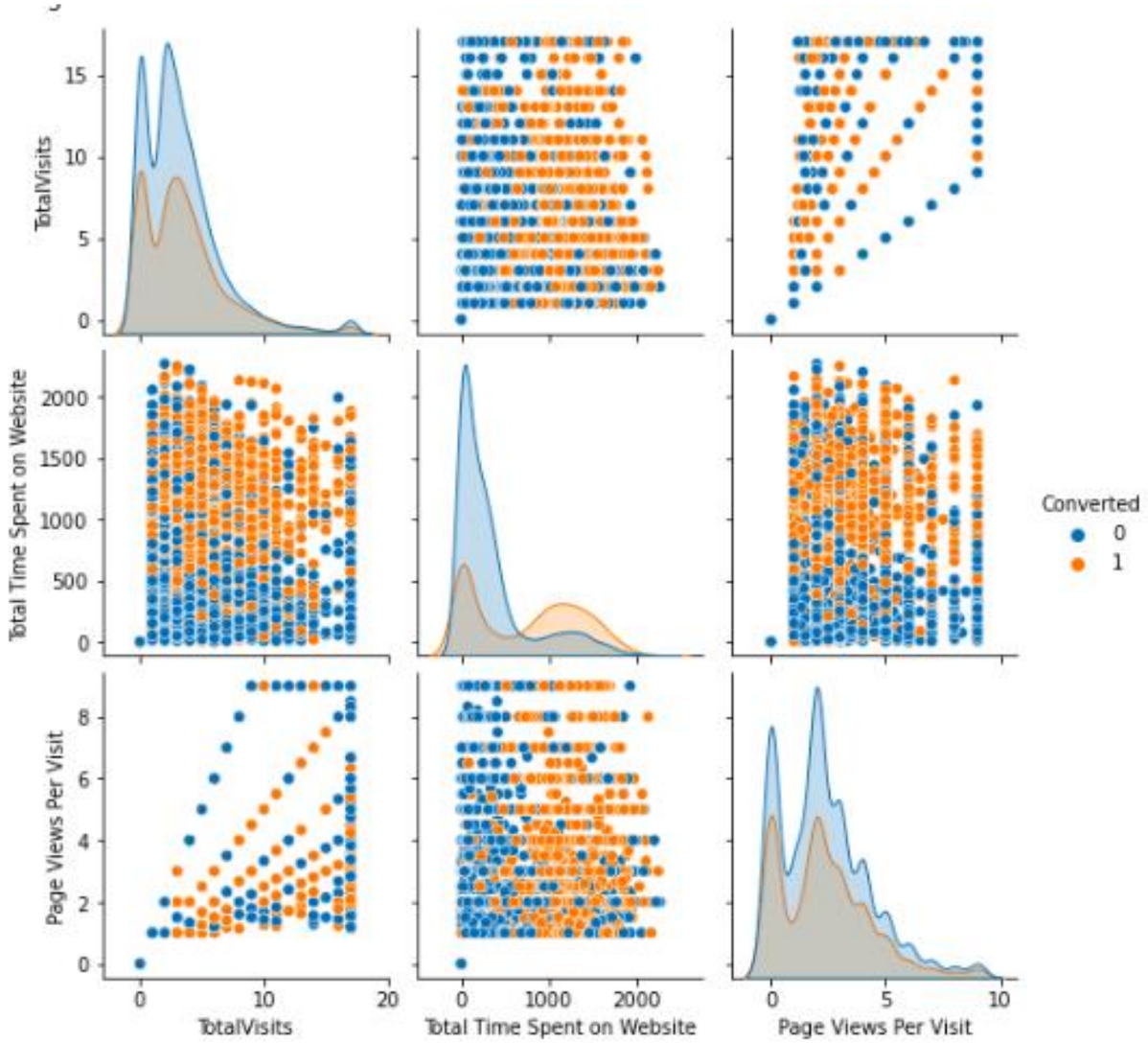




Outlier Handling Observations

- Maximum Total Visits on the websites slightly increase the probability of conversion.
- people who are spending more time on website have more probability of conversion than people who spent less time on website.
- 1-3 pages per visit gets equal chance of conversion

Bivariate Analysis



Bivariate Analysis

- There is good correlation between "TotalVisits" and "Page Views Per Visit" features. We should drop one of the feature to avoid multicollinearity

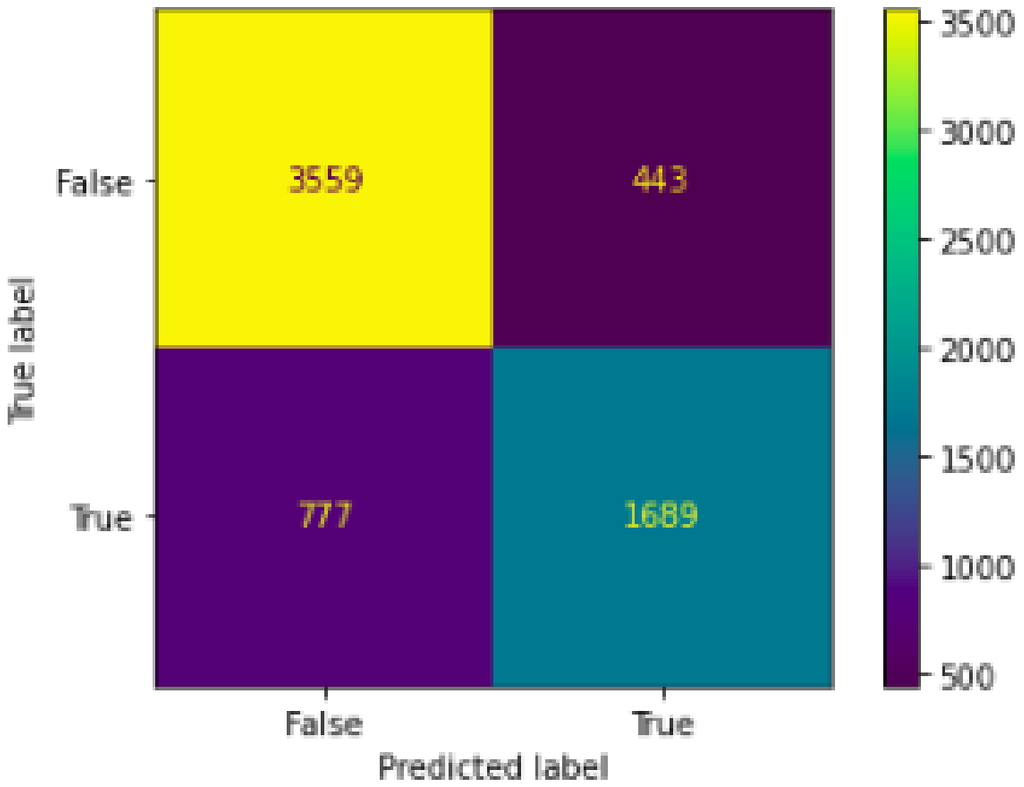
Model Building

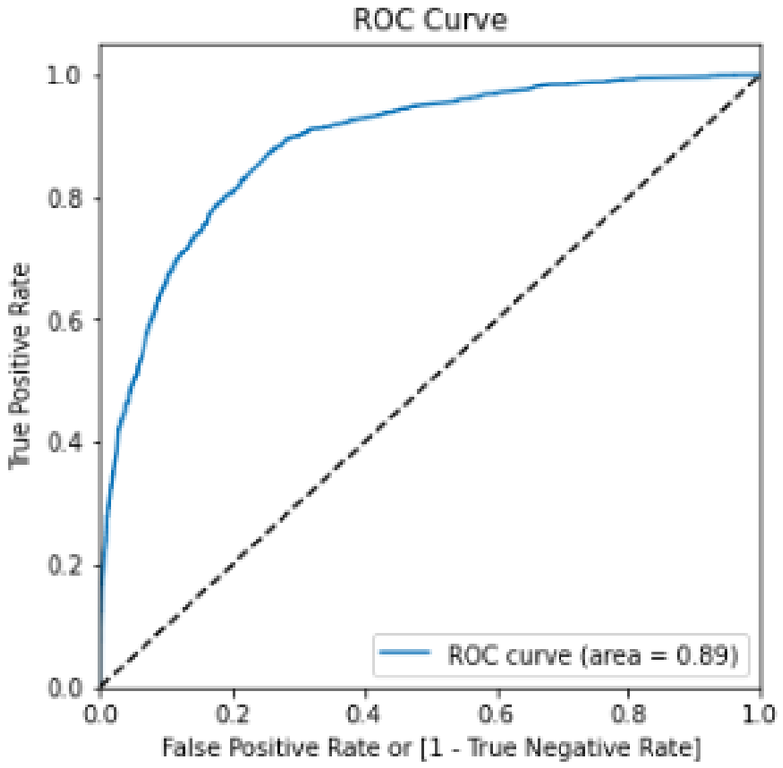
Feature Selection using RFE

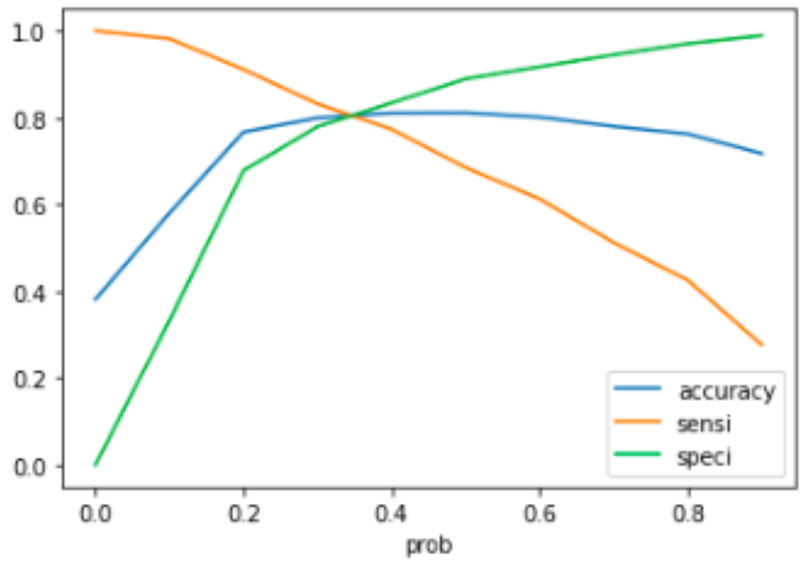
- **Model 1**
  - VIF of all features is below 5, that shows there is no multicollinearity among the independent variables
  - p-value for "What is your current occupation_Other" and "What is your current occupation_Housewife"
  - However, we will drop "What is your current occupation_Housewife" feature first and rebuild the model
- **Model 2**
  - In model 2 we will drop "What is your current occupation_Other" variable and re-build model
- **Model -3**
  - After dropping "What is your current occupation_Other" variable model has p value < .05 and VIF <5
- **Therefore, we will be selecting model 3 as our final model**

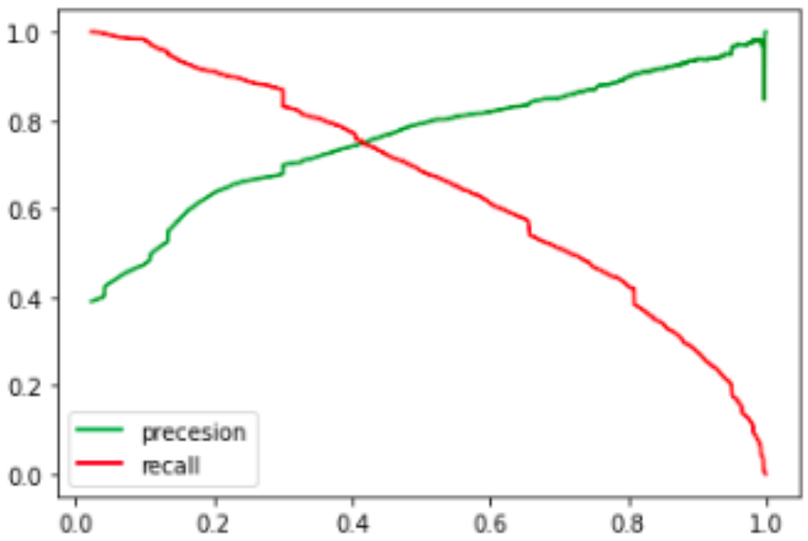
Model Evaluation

Confusion Matrix









Logistic Regression Final Model Parameters
Area under Curve = 0.89
Final cut-off = 0.35

Model Evaluation

- Training Data Set
- Overall Accuracy of the model – **80.57**
- Overall Sensitivity of the model – **80.54**
- Overall Specificity of the model – **80.58**
- **Area under curve - 0.89**

- **Test Data set**
- Overall Accuracy of the model – **81.93**
- Overall Sensitivity of the model – **81.55**
- Overall Specificity of the model – **82.17**
- **Area under curve - 0.89**

Inferences

Significant variables to which contribute most towards the probability of a lead getting converted:

- Lead Source_Welingak Website(coeff 5.14)
- What is your current occupation_Working Professional (coeff 3.5)
- Lead Source_Reference(coeff 3.2)
- Last Notable Activity_SMS Sent(coeff 1.5)
- What is your current occupation_Student(coeff 1.15)

Recommendation

- X Education Company needs to focus on following key aspects to improve the overall conversion rate:
  - Leads who gathered information about X Education from ‘Welingak’ website should be targeted
  - Target the leads got from referral program
  - Focus on leads whose last notable actvity is SMS sent that helps in higher conversion.
  - Target the working professionals as this will increase the conversion rate
  - Target the students as this will increase the conversion rate.
  - Nurture ~ 708 Hot leads identified by model to increase conversion rate

THANK YOU