Lead Score Case Study

Team Tuple :
Beon Carvalho
Safalta Kundra
Venktesh Rathi
Praveen Kumar

Rationale:

To get a basic understanding of how the logistic regression is used in solving business problems.

Background of the problem statement:

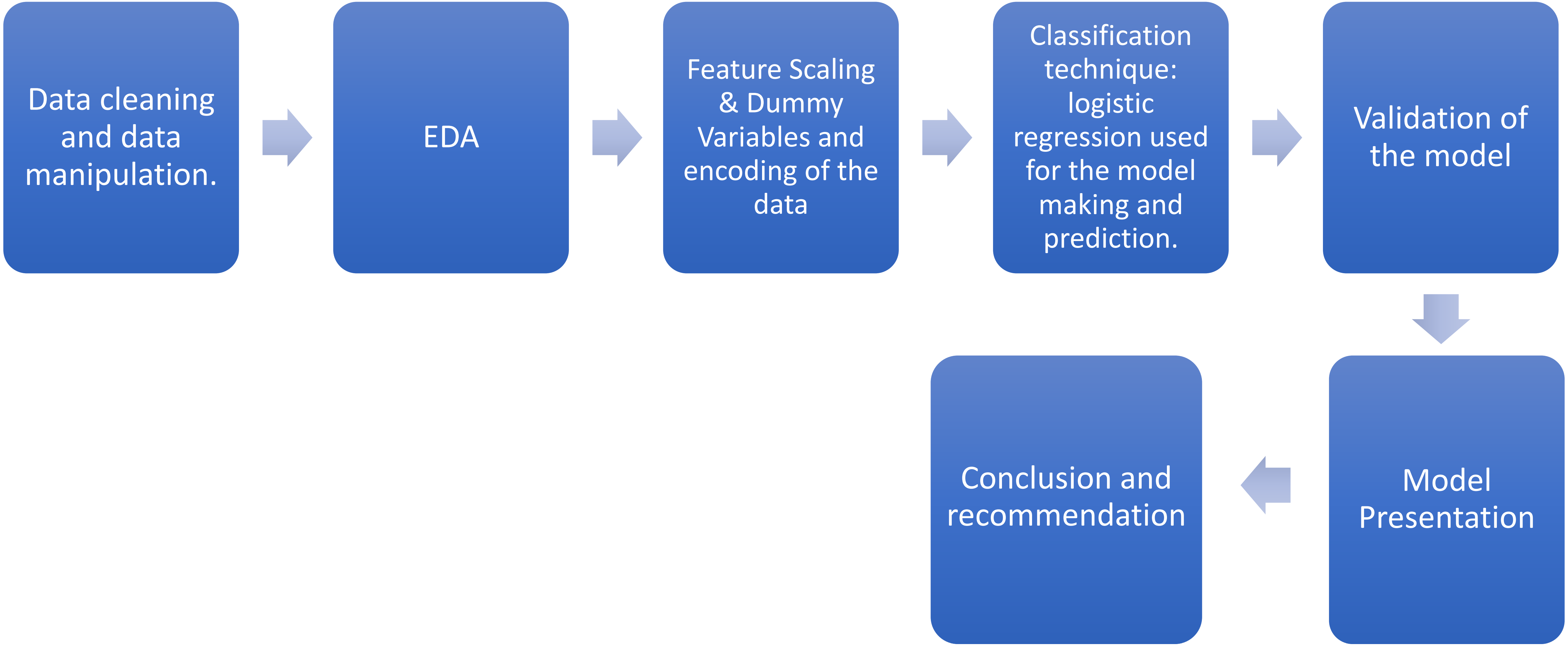
- X Education sells online courses to industry professionals.
- X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.
- To make this process more efficient, the company wishes to identify the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

.

Objective:

- X education wants to know most promising leads.
- For that they want to build a Model which identifies the hot leads.
- Deployment of the model for the future use.

Process Pipeline



Data cleaning and data manipulation.

EDA

Feature Scaling & Dummy Variables and encoding of the data

Classification technique: logistic regression used for the model making and prediction.

Validation of the model

Model Presentation

Conclusion and recommendation

Data Cleaning and Manipulation

- **Total Number of Rows =9240, Total Number of Columns =37.**Single value features like "Magazine", "Receive More Updates About Our Courses", "Update me on Supply Chain Content", "Get updates on DM Content", "I agree to pay the amount through cheque" etc. have been dropped.

- Removing the "Prospect ID" and "Lead Number" which is not necessary for the analysis.

- After checking for the value counts for some of the object type variables, we find some of the features which has no enough variance, which we have dropped, the features are: "Do Not Call", "What matters most to you in choosing course", "Search", "Newspaper Article", "X Education Forums", "Newspaper", "Digital Advertisement" etc.

- Dropping the columns having more than 35% as missing value such as 'How did you hear about X Education' and 'Lead Profile'.

- Many columns have **“Select”** as the values such columns are
  - Specialization
  - Country
  - Current Occupation

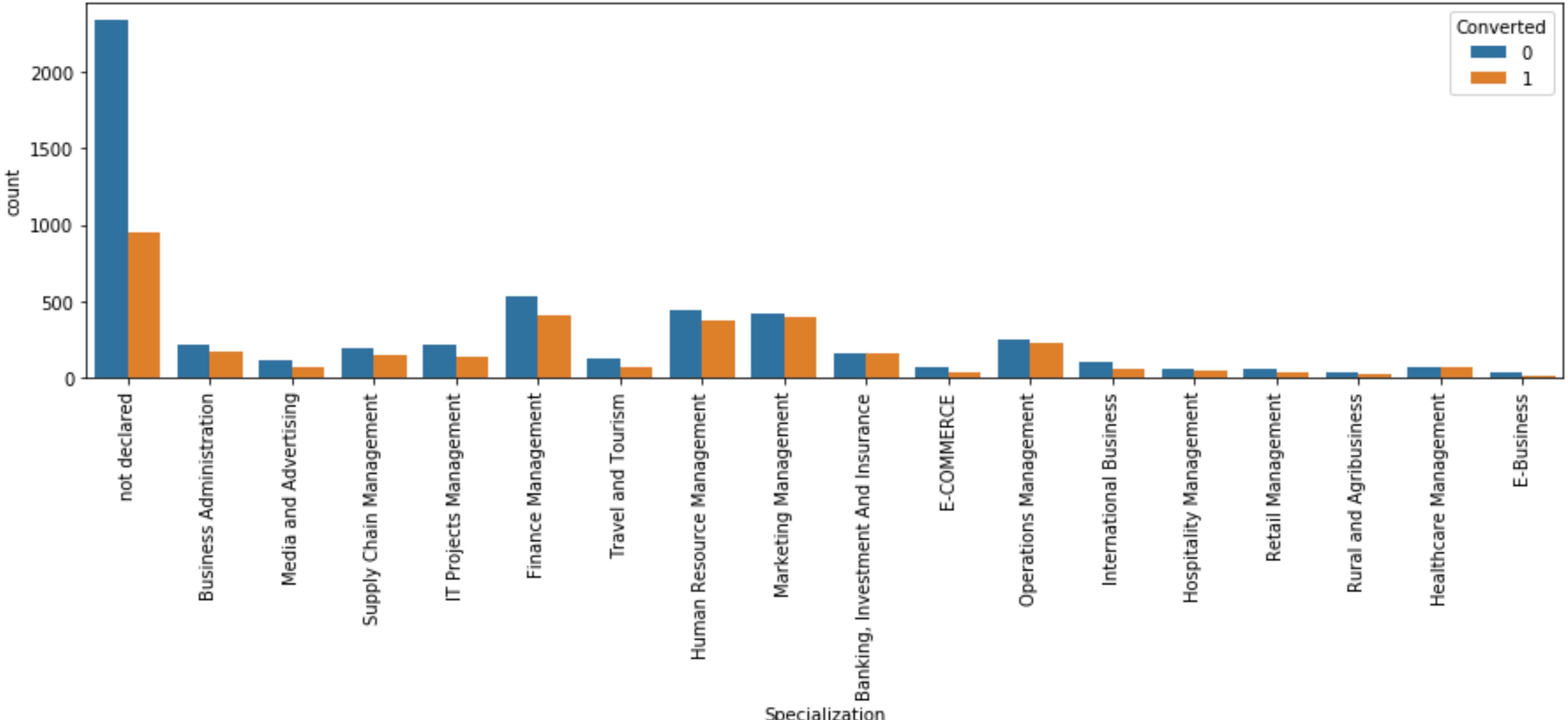
- Such values and Null values are treated as Not Declared and used for further analysis Numerical Missing values have been dropped

- Outlier Treatment of Total Visits and Page Views Per Visit.

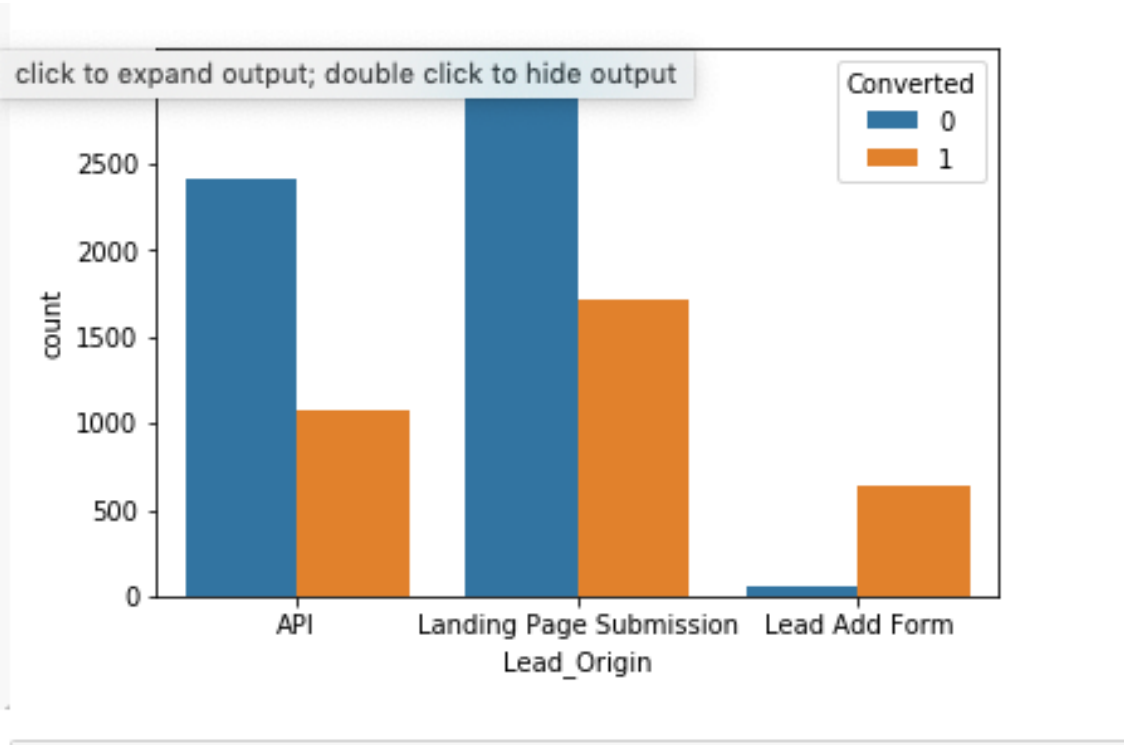
- Observed that major part of null values in "Page Views Per Visit", "TotalVisits" are Converted. So, imputing median values of them to null values.

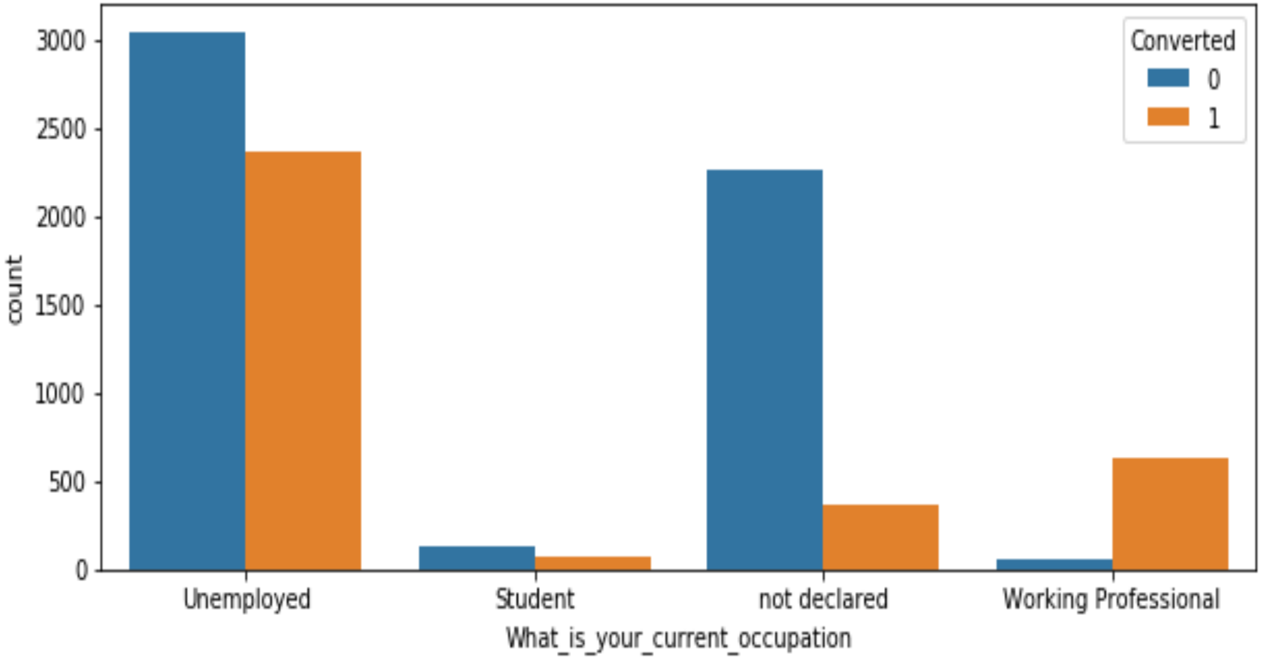
EDA (Exploratory Data Analysis)

Specialization to the converted rate

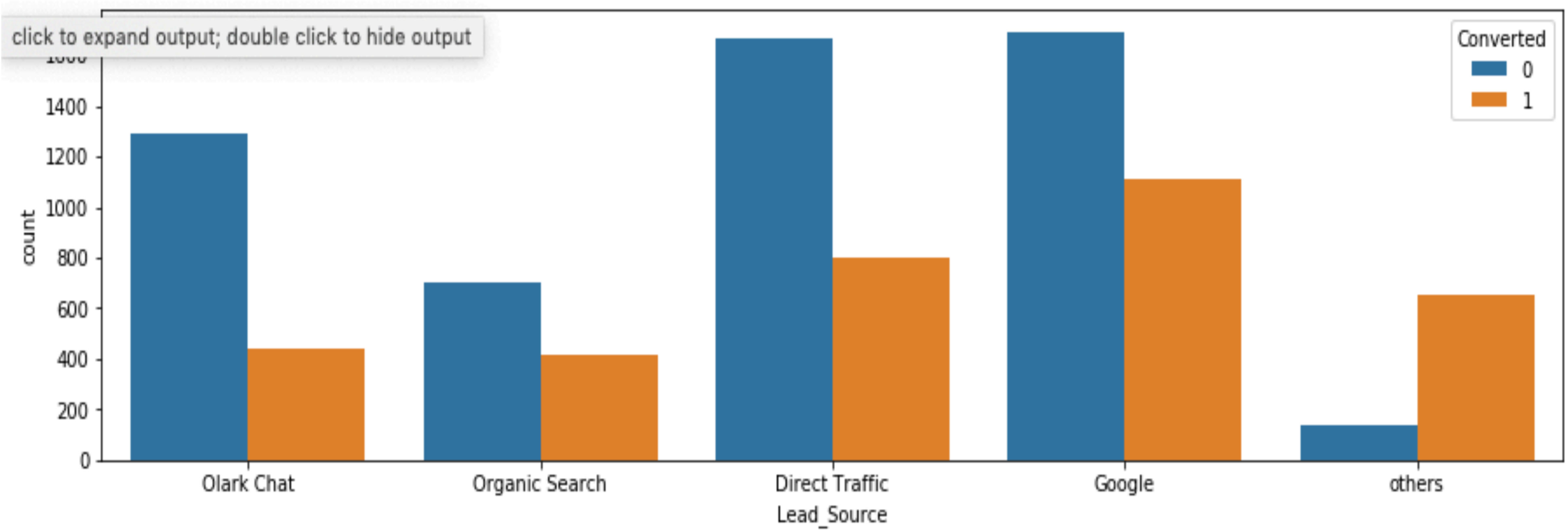


EDA Cont.


click to expand output; double click to hide output
Converted
0
1
count
2500
2000
1500
1000
500
0
API
Landing Page Submission
Lead Add Form
Lead_Origin


Converted
0
1
count
3000
2500
2000
1500
1000
500
0
Unemployed
Student
not declared
Working Professional
What_is_your_current_occupation

EDA Cont.


click to expand output; double click to hide output
Converted
0
1
count
1400
1200
1000
800
600
400
200
0
Olark Chat
Organic Search
Direct Traffic
Google
others
Lead_Source

Data Conversion

- Numerical Variables are Normalized.
- Dummy Variables are created for object type variables
- Total Rows for Analysis: 8792
- Total Columns for Analysis: 42

Model Building

- Splitting the Data into Training and Testing Sets
- As you know, 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 30 variables as output
- Building Model
- Removing the variable whose p- value is greater than 0.05 and vif value is greater than 0.02
- PREDICTIONS ON TEST DATA SET
- Overall accuracy 81%

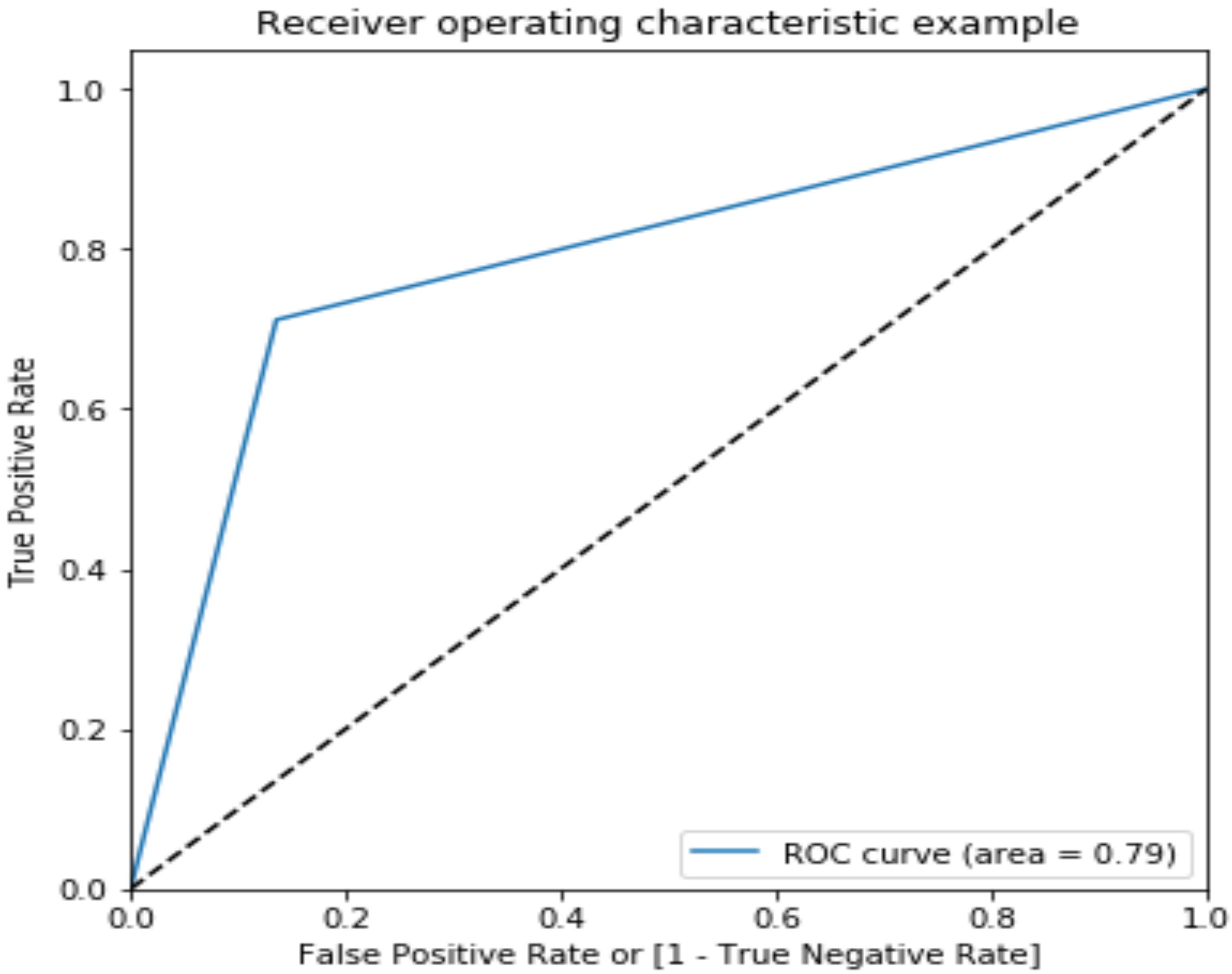
Model Results

- Confusion Matrix

```
[[1421, 223]
  [287, 707]]
```

- Accuracy --- 80.67 %
- Specificity --- 71.13 %
- Sensitivity/TPR/Recall --- 86.44 %
- FPR --- 28.87 %
- Precision --- 83.2 %

ROC Curve - Finding Optimal Cutoff Point



Conclusions

- Main Variables that contribute to analysis are Lead Origin, Lead Source & Occupation.
- Specialization and Total time spent also predict the conversion rate.
- Concrete conclusion cannot be made but suggestions can be given as the data is very less.