Lead Scoring Case Study

This Presentation is the overview of how our proposed model improved lead conversion rate from 30% to 80% of X Education company and it’s business team to optimize the efforts

by

Shashank T E

Krushil Ramani

Sanket Kamble

Content:-

- Problem Statement
  - Objectives
  - Business Need
- Analysis approach including important visuals
  - Data Cleaning
  - EDA
  - Feature Selection & Modeling
  - Interpretation
- Results
- Recommendations

Problem Statement

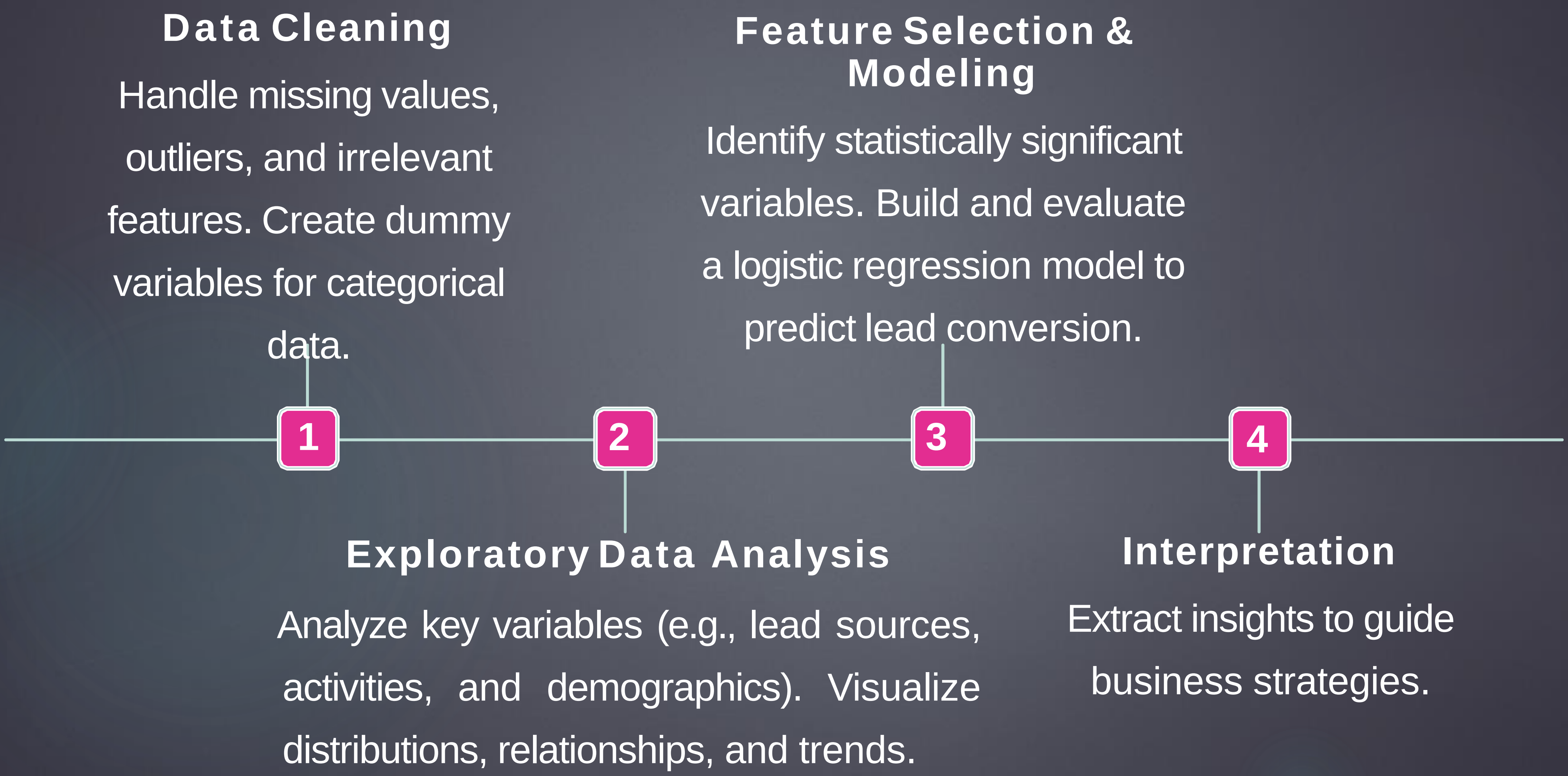
Objectives:-

- Developing a model which will improve the lead conversion rate from 30 % to 80 %.
- To identify key factors influencing lead conversion and predict the likelihood of a lead getting converted into a customer.

Business Need:-

- Enhance the efficiency of the sales team by prioritizing potential leads, reducing conversion time, and improving resource allocation.

Analysis Approach



1

Data Cleaning

Handle missing values, outliers, and irrelevant features. Create dummy variables for categorical data.

2

Exploratory Data Analysis

Analyze key variables (e.g., lead sources, activities, and demographics). Visualize distributions, relationships, and trends.

3

Feature Selection & Modeling

Identify statistically significant variables. Build and evaluate a logistic regression model to predict lead conversion.

4

Interpretation

Extract insights to guide business strategies.

Data Overview

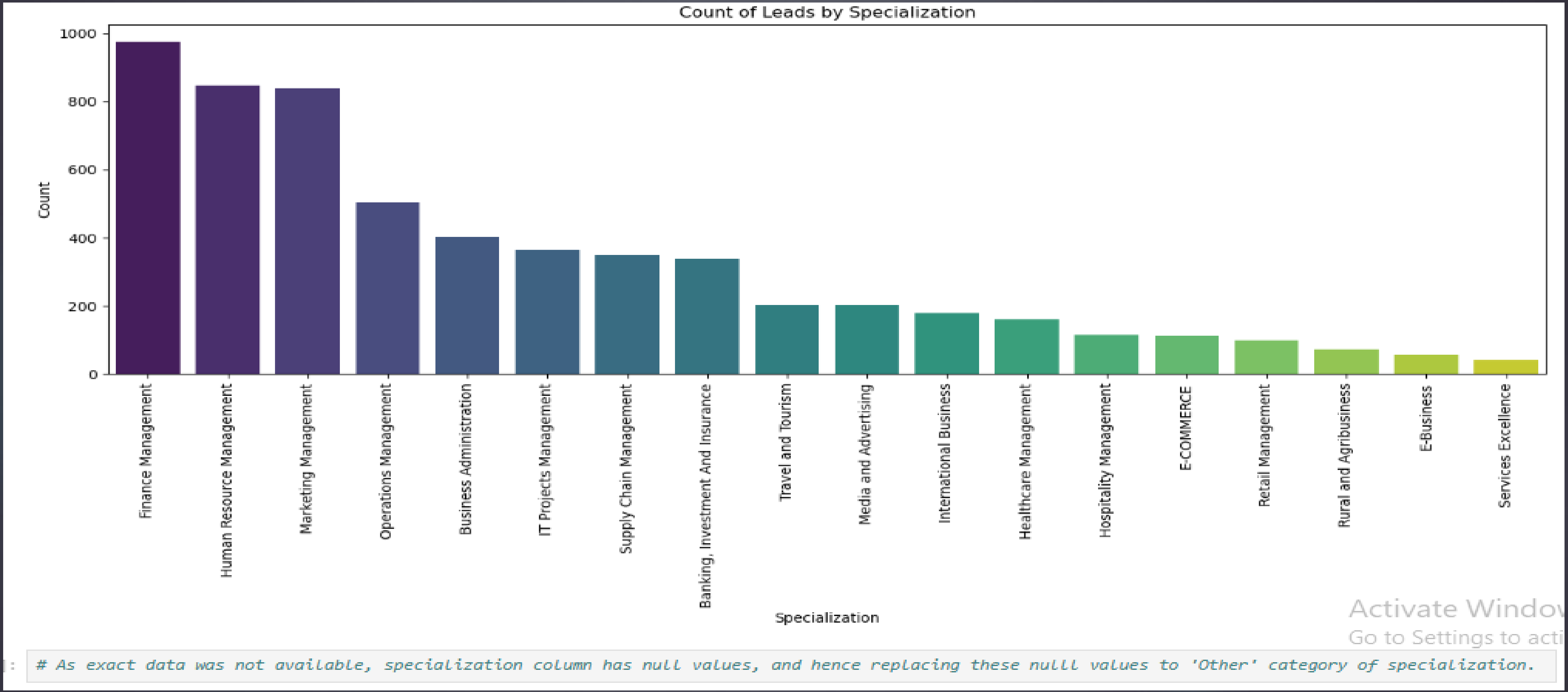
Dataset Information

- Number of records: (9240 rows, 37 columns
- Key features: Lead Source, Last Activity, Current Occupation, etc.

Preprocessing Summary

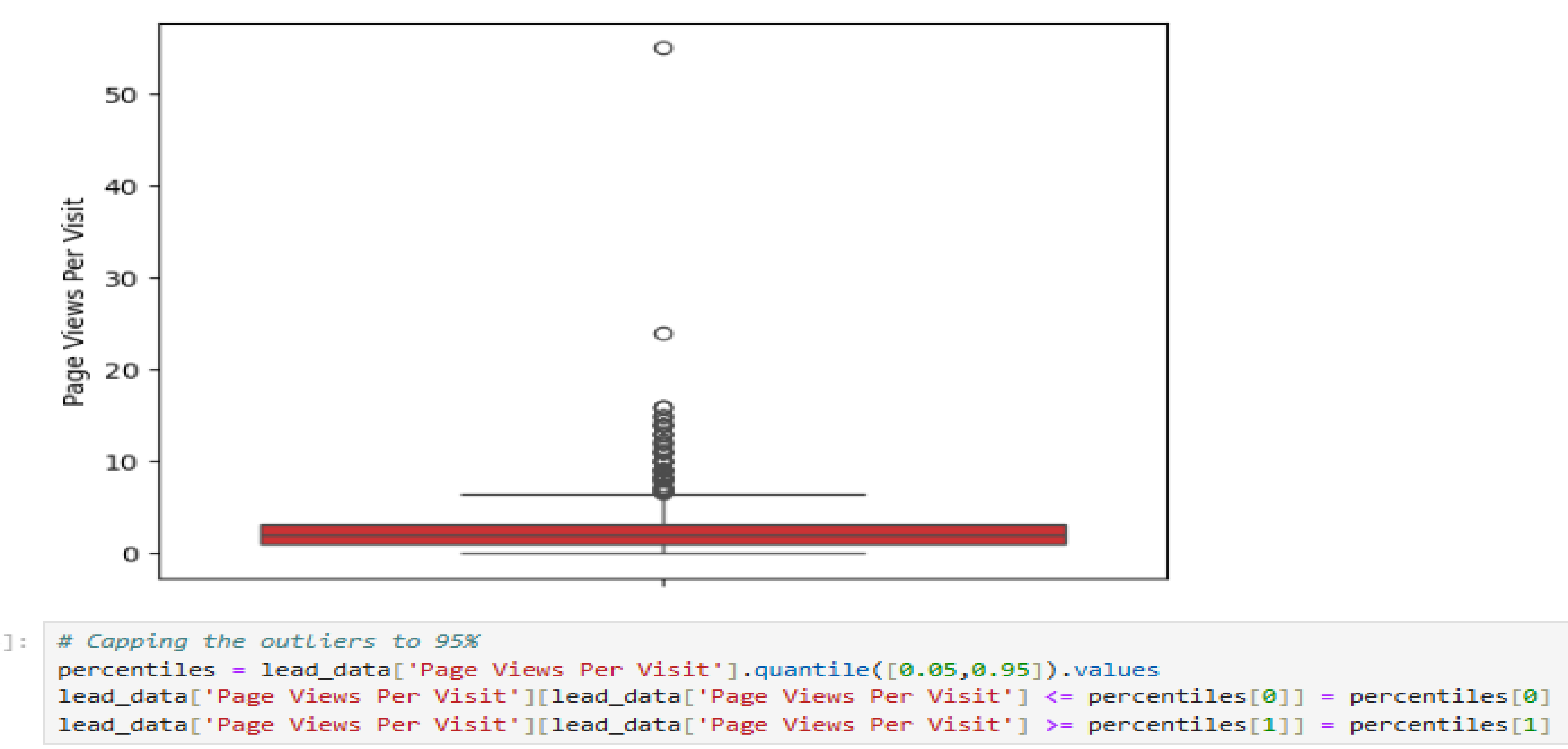
- Handled the missing values handled
- Created the dummy variables created

Merging Data



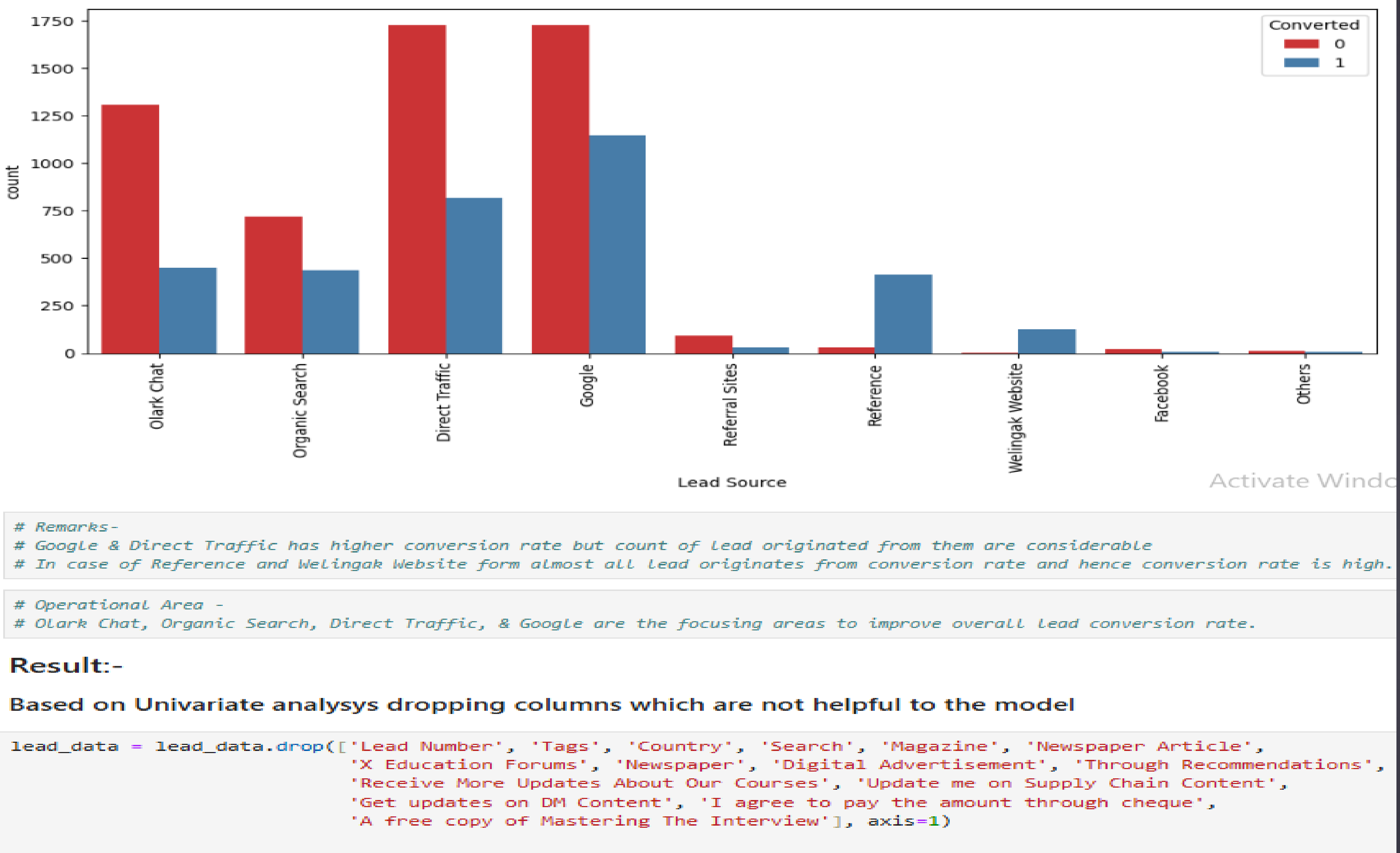
```
# As exact data was not available, specialization column has null values, and hence replacing these nulll values to 'Other' category of specialization.
```

Capping the outliers



```
# Capping the outliers to 95%
percentiles = lead_data['Page Views Per Visit'].quantile([0.05,0.95]).values
lead_data['Page Views Per Visit'][lead_data['Page Views Per Visit'] <= percentiles[0]] = percentiles[0]
lead_data['Page Views Per Visit'][lead_data['Page Views Per Visit'] >= percentiles[1]] = percentiles[1]
```

Exploratory Data Analysis (EDA)



```
# Remarks-
# Google & Direct Traffic has higher conversion rate but count of lead originated from them are considerable
# In case of Reference and Welingak Website form almost all lead originates from conversion rate and hence conversion rate is high.
```

```
# Operational Area -
# Olark Chat, Organic Search, Direct Traffic, & Google are the focusing areas to improve overall lead conversion rate.
```

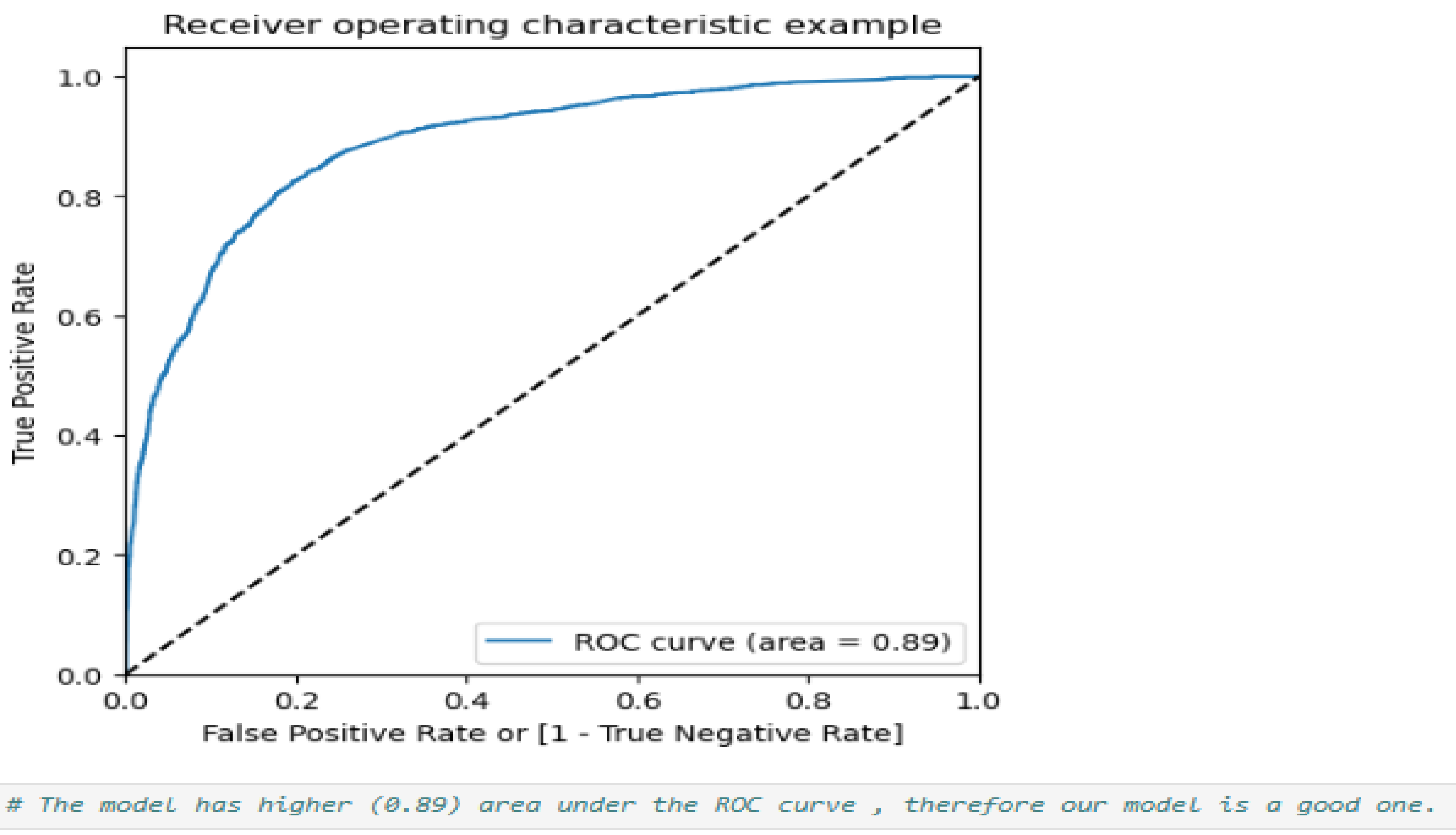
Result:-

Based on Univariate analysys dropping columns which are not helpful to the model

```
lead_data = lead_data.drop(['Lead Number', 'Tags', 'Country', 'Search', 'Magazine', 'Newspaper Article',
                            'X Education Forums', 'Newspaper', 'Digital Advertisement', 'Through Recommendations',
                            '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',
                            'A free copy of Mastering The Interview'], axis=1)
```

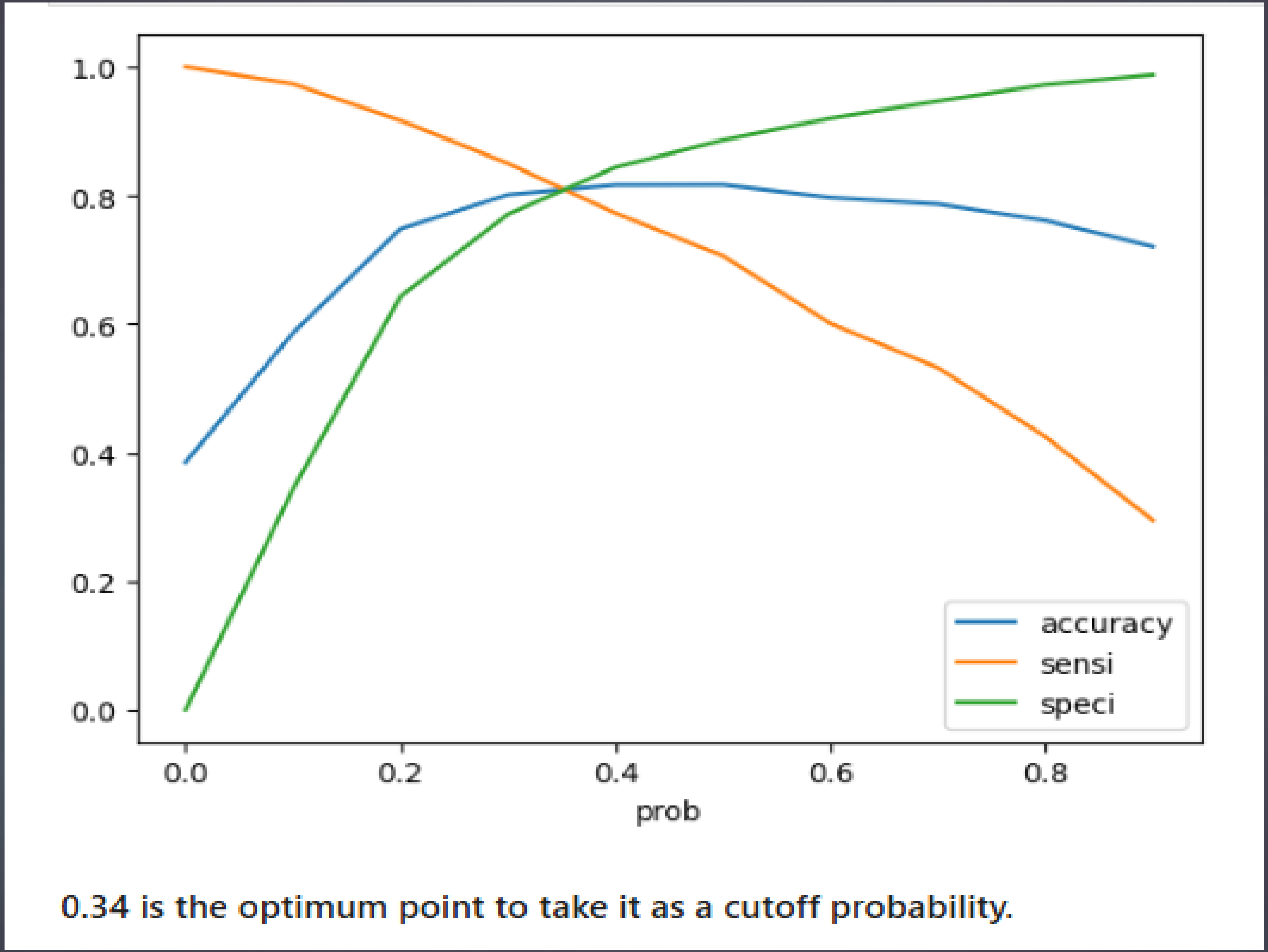
Final Model:-

With the P-values of all variables is equals to 0 (<0.05) and VIF values are low (<5) for all the variables, finalized the model with 12 variables in it.



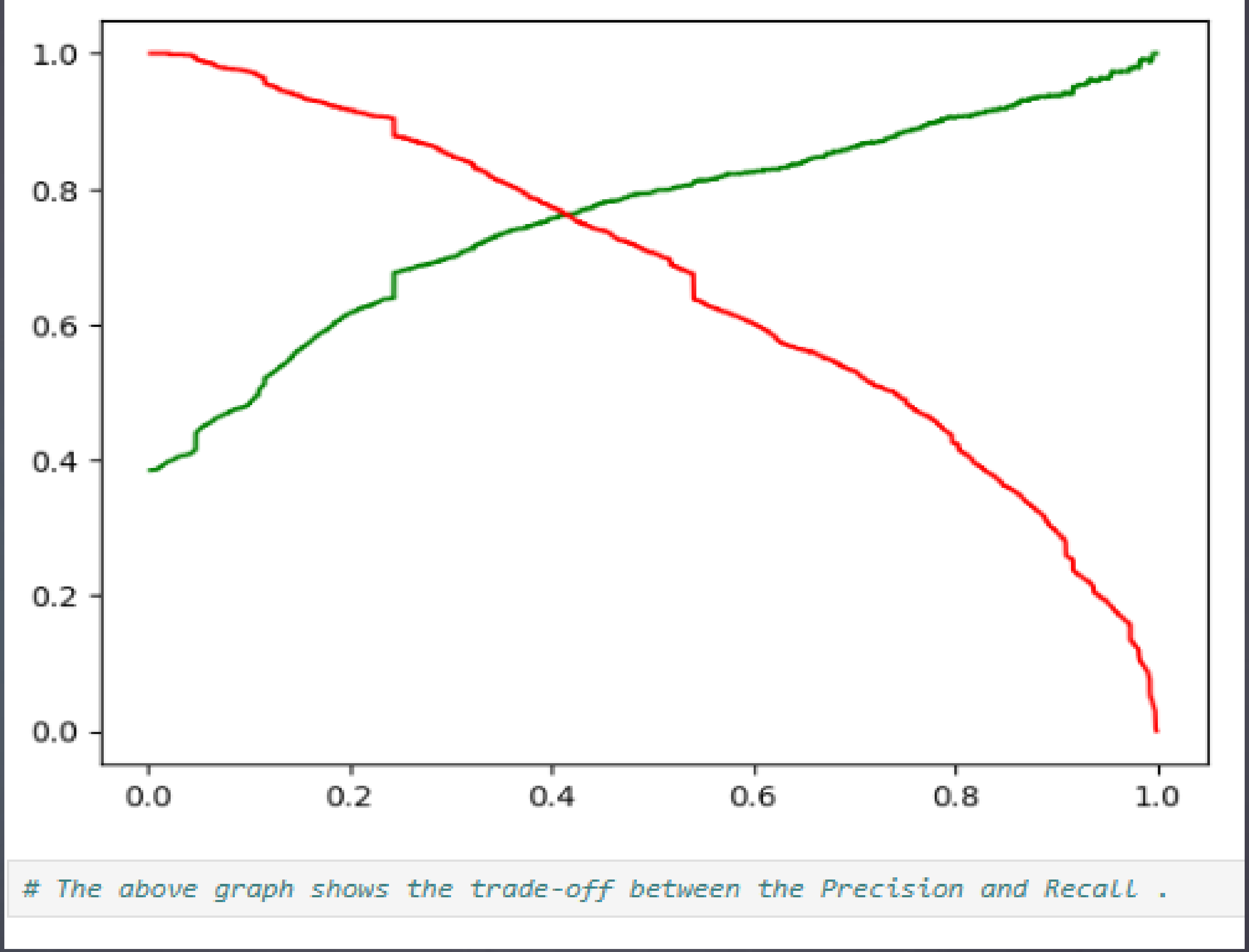
```
# The model has higher (0.89) area under the ROC curve , therefore our model is a good one.
```

Optimal Cutoff Point

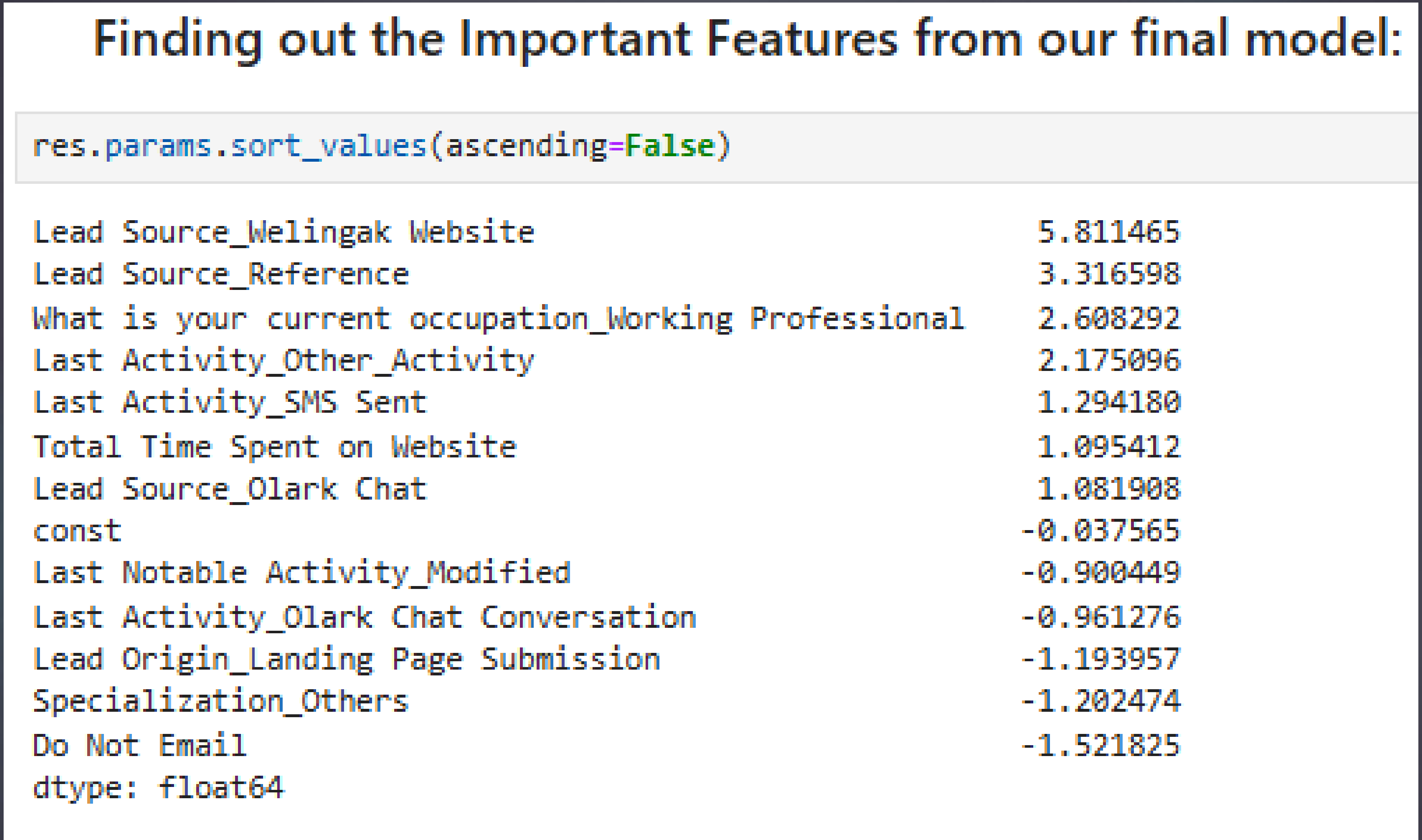


0.34 is the optimum point to take it as a cutoff probability.

Trade-off curve between precision and recall



```
# The above graph shows the trade-off between the Precision and Recall .
```



Finding out the Important Features from our final model:

```
res.params.sort_values(ascending=False)
```

```
Lead Source_Welingak Website                           5.811465
Lead Source_Reference                                  3.316598
What is your current occupation_Working Professional   2.608292
Last Activity_Other_Activity                           2.175096
Last Activity_SMS Sent                                 1.294180
Total Time Spent on Website                            1.095412
Lead Source_Olark Chat                                 1.081908
const                                                 -0.037565
Last Notable Activity_Modified                        -0.900449
Last Activity_Olark Chat Conversation                 -0.961276
Lead Origin_Landing Page Submission                   -1.193957
Specialization_Others                                 -1.202474
Do Not Email                                          -1.521825
dtype: float64
```

Results :-

Comparing the values obtained for Train & Test:

Train Data:

- Accuracy : **81.0** %
- Sensitivity : **81.7** %
- Specificity : **80.6** %

Test Data:

- Accuracy : **80.4** %
- Sensitivity : **80.4** %
- Specificity : **80.5** %

Hence model has achieved target of lead conversion rate of 80% .
It is able to give the CEO confidence in making good calls based on this model to get a higher lead conversion rate of 80%.

Recommendations

The company **should make calls** to the leads

- Coming from the lead sources "**Welingak Websites**" and "**Reference**"
- Who are the "**working professionals**"
- Who spent "**more time on the websites**"
- Coming from the lead sources "**Olark Chat**"
- Whose “**last activity was SMS Sent”**

as these are more likely to get converted.

The company **should not make calls** to the leads

- Whose last activity was "**Olark Chat Conversation**"
- Whose lead origin is "**Landing Page Submission**"
- Whose Specialization was "**Others**"
- Who chose the option of "**Do not Email**" as "**yes**"

as they are not likely to get converted.

Thank You