Lead Conversion Prediction

Logistic Regression Assignment

Devesh Singh & Archana Prabhakar

August, 2019

Contents

Problem Statement and Goals

Analysis Approach

Findings & Recommendations

Problem Statement and Goals

- Problem Statement
- X Education, which sells online courses to industry professionals, has a poor Lead conversion rate (30%) and would like to improve it

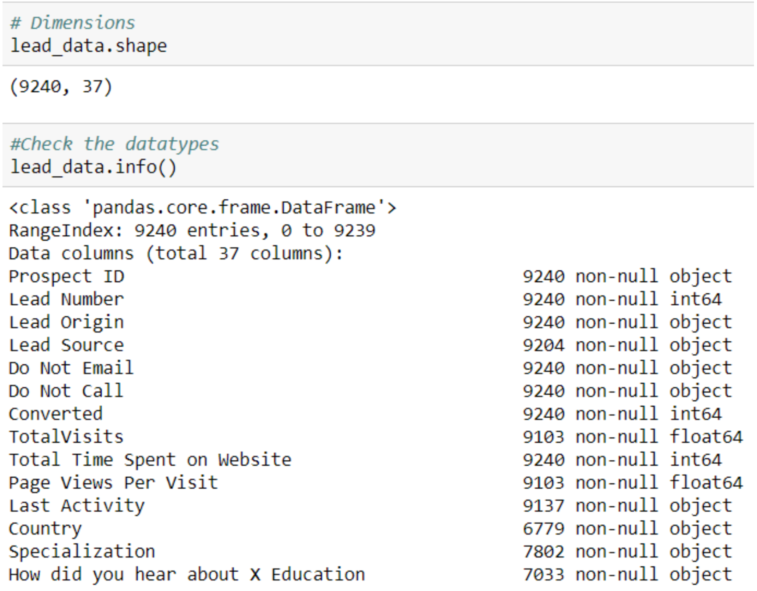
- Goals
- Build a Logistic Regression Model which can predict which leads are the “Hot” leads (has a high Lead Score)
- Provide recommendations on the what can be done to improve the lead conversion rate

Analysis Approach

- Data Understanding
- Data Cleaning and Preparation
- EDA
- Create Dummy Variables
- Test-Train Split
- Feature Scaling
- Model Building
- Feature Selection
- Model Refinement
- Check Multicollinearity
- Model Evaluation
- Make Prediction

Data Understanding

- Size
- Data Types
- Count of null values
- Basic Statistics

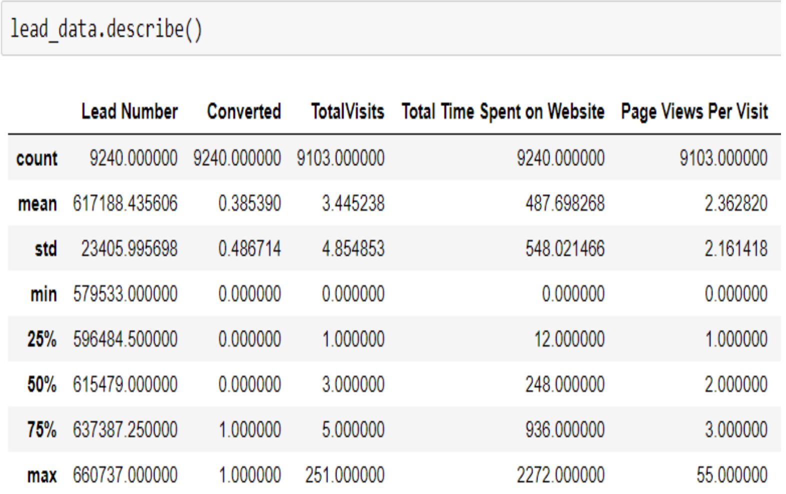


```
# Dimensions
lead_data.shape
```

```
(9240, 37)
```

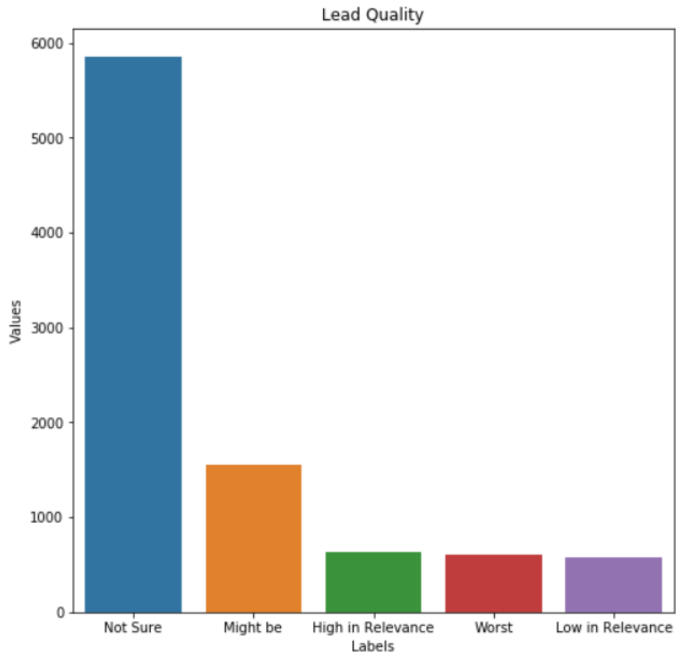
```
#Check the datatypes
lead_data.info()
```

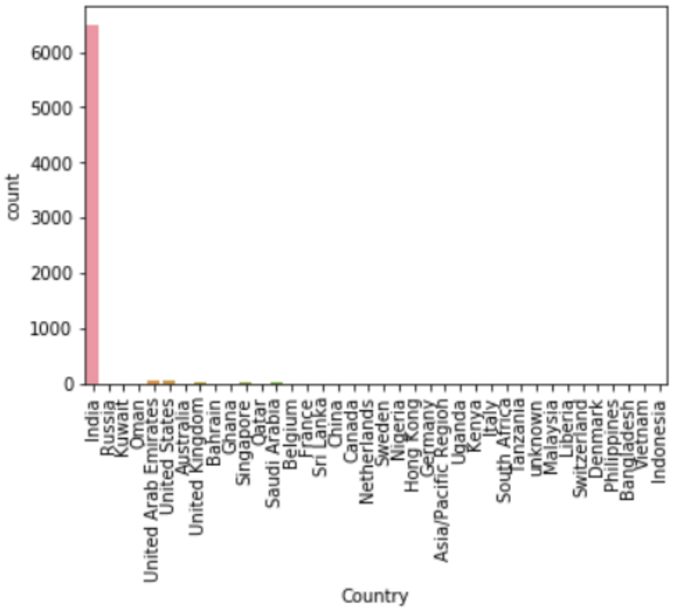
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9240 entries, 0 to 9239
Data columns (total 37 columns):
Prospect ID                                 9240 non-null object
Lead Number                                 9240 non-null int64
Lead Origin                                 9240 non-null object
Lead Source                                 9204 non-null object
Do Not Email                                9240 non-null object
Do Not Call                                 9240 non-null object
Converted                                   9240 non-null int64
TotalVisits                                 9103 non-null float64
Total Time Spent on Website                 9240 non-null int64
Page Views Per Visit                        9103 non-null float64
Last Activity                               9137 non-null object
Country                                     6779 non-null object
Specialization                              7802 non-null object
How did you hear about X Education          7033 non-null object
```

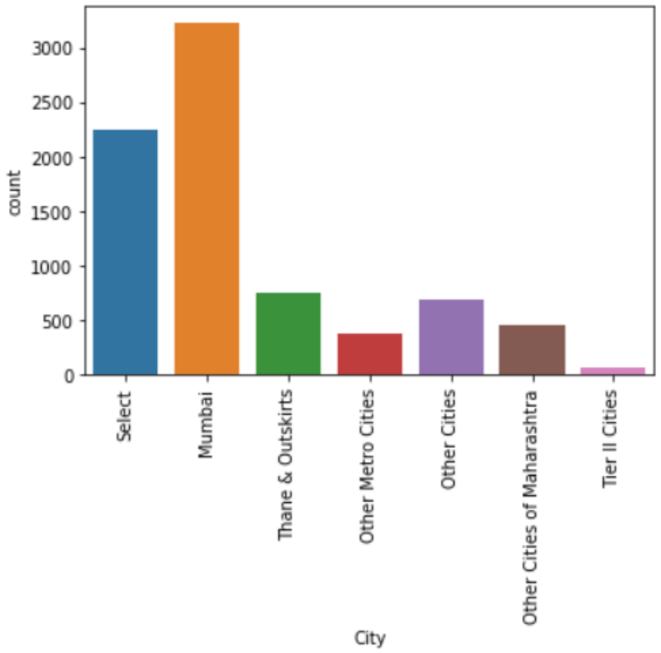


```
lead_data.describe()
```

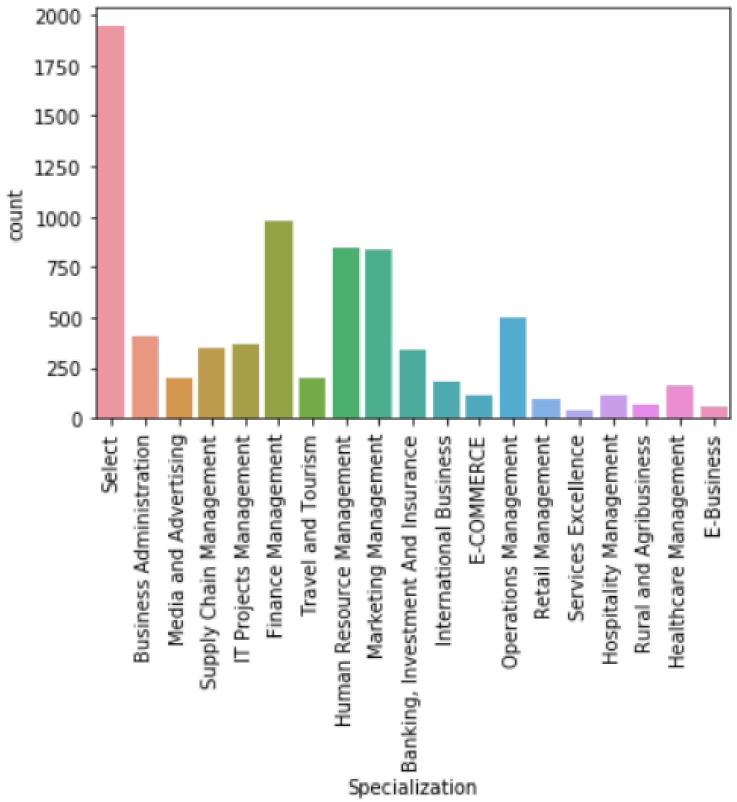
	Lead Number	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit
count	9240.000000	9240.000000	9103.000000	9240.000000	9103.000000
mean	617188.435606	0.385390	3.445238	487.698268	2.362820
std	23405.995698	0.486714	4.854853	548.021466	2.161418
min	579533.000000	0.000000	0.000000	0.000000	0.000000
25%	596484.500000	0.000000	1.000000	12.000000	1.000000
50%	615479.000000	0.000000	3.000000	248.000000	2.000000
75%	637387.250000	1.000000	5.000000	936.000000	3.000000
max	660737.000000	1.000000	251.000000	2272.000000	55.000000

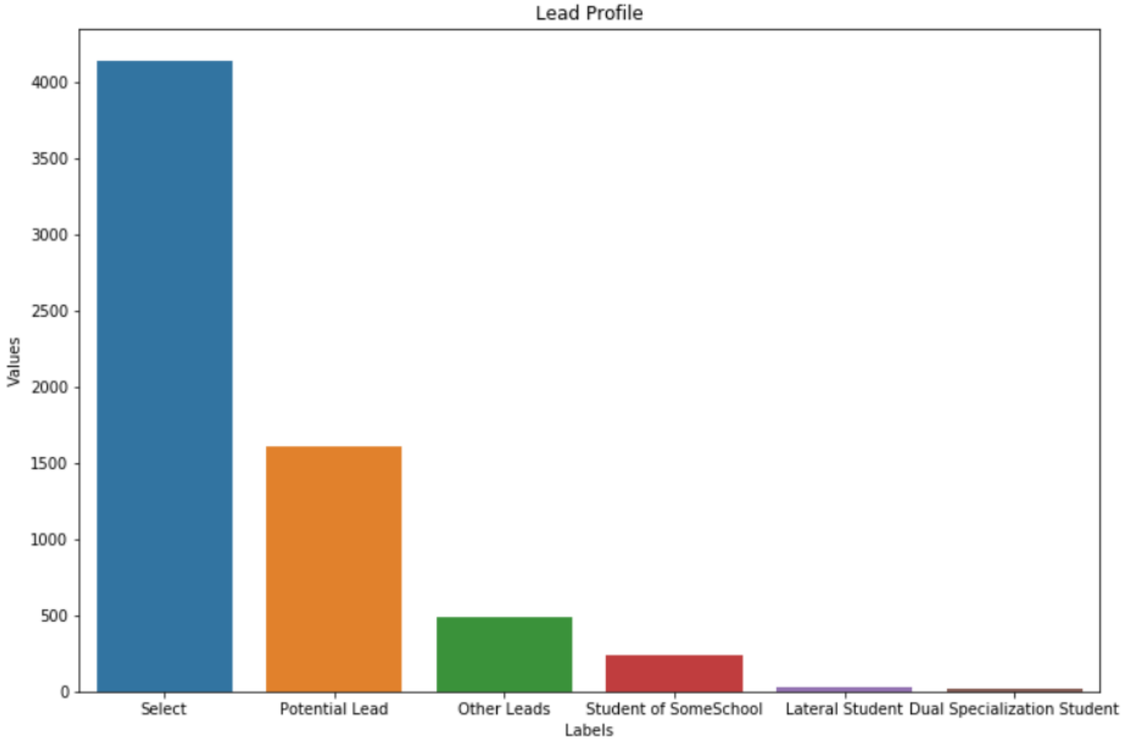






Data Quality Checks (1)

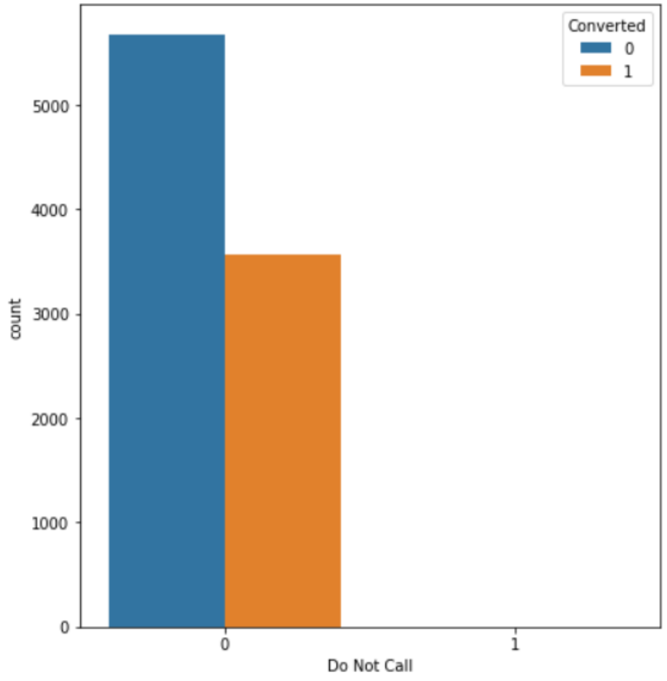


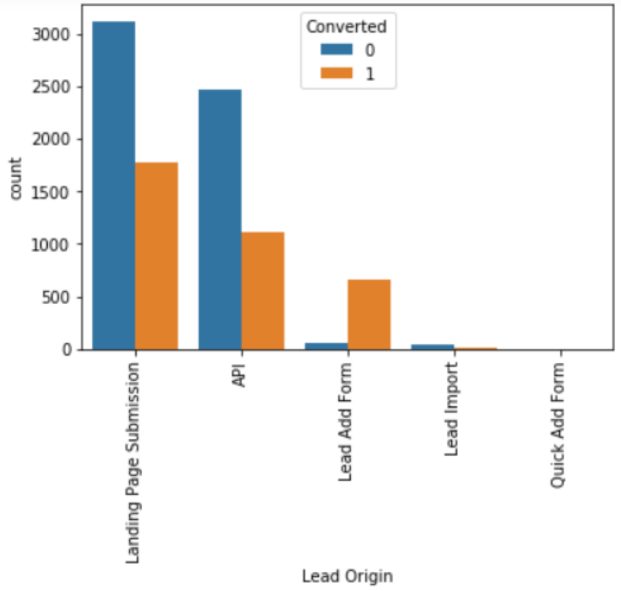


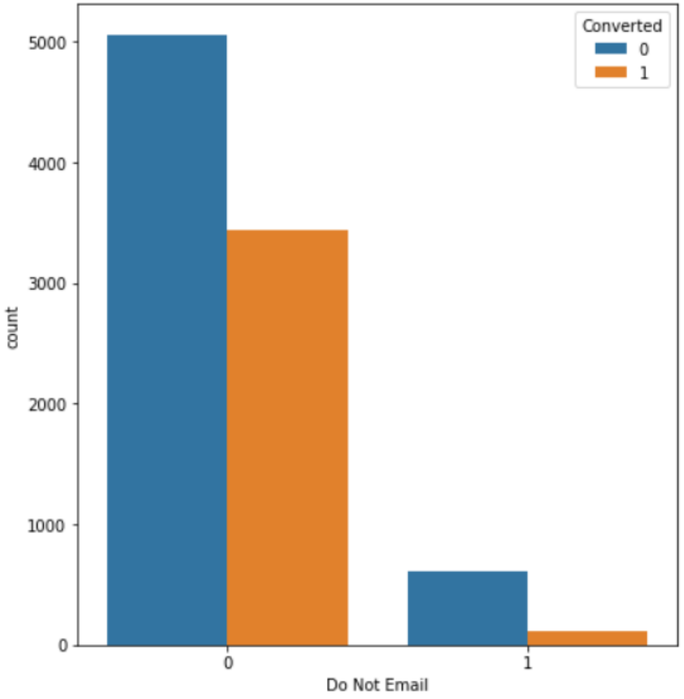
Data Quality Checks (2)

Data Cleaning and Preparation

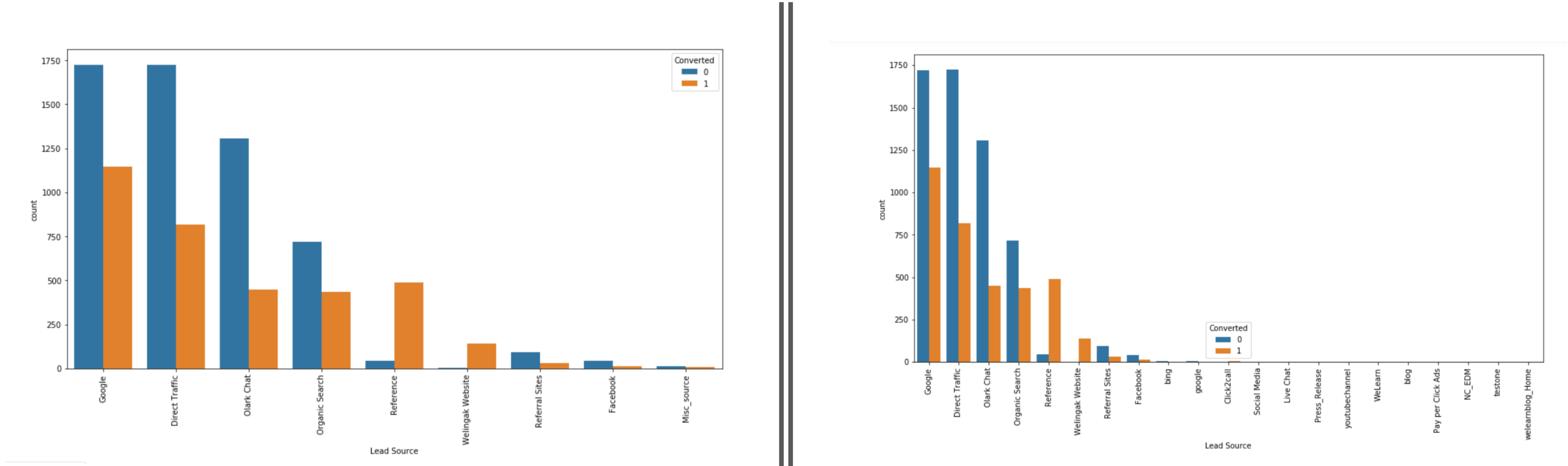
- Based on the various data quality checks, following actions have been taken -
  - Handle missing values
    - Drop Columns where >30% of missing data (E.g. Assymetrique value columns)
    - Impute with mode() where appropriate (E.g. Lead Source, Occupation)
    - Impute with mean() where appropriate (E.g. Total visits)
    - Impute with “Unknown” where appropriate (E.g. City, Specialization)
  - Drop columns which provide no additional information/variance to the model building process (E.g. Country, What Matters most)
  - Within a column, merge values which have no significant number of rows (E.g. Lead Source values like Social Media, bing etc. merged into Misc_Sources)
  - Map binary values to 1 and 0 (E.g. Do Not Call, Search)
  - Clean-up of values (E.g. Google and google)


Converted
0
1
count
5000
4000
3000
2000
1000
0
0
1
Do Not Call


Converted
0
1
count
3000
2500
2000
1500
1000
500
0
Landing Page Submission
API
Lead Add Form
Lead Import
Quick Add Form
Lead Origin

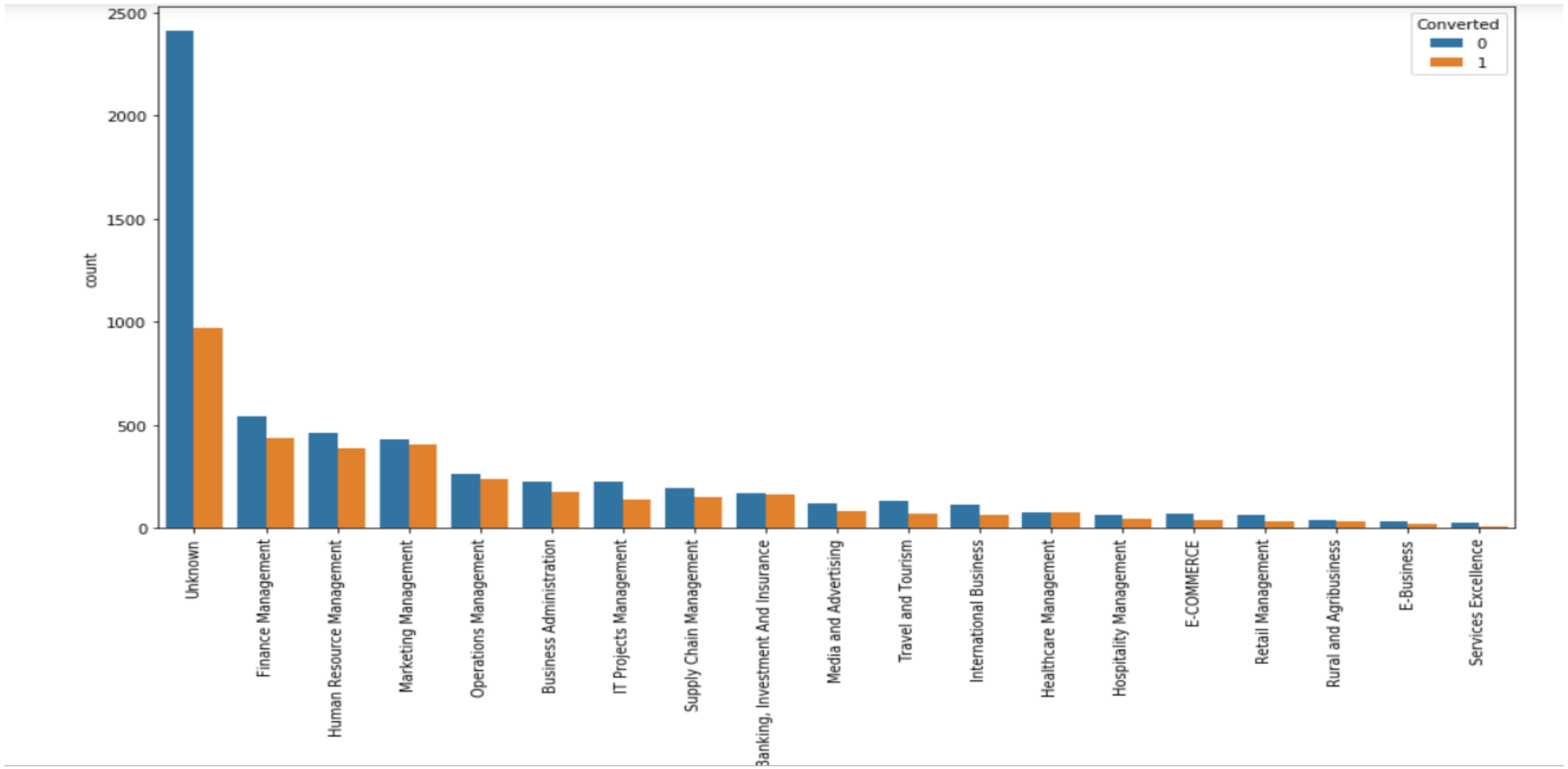

Converted
0
1
count
5000
4000
3000
2000
1000
0
0
1
Do Not Email

EDA (1)



Lead Source before and after feature engineering

EDA (2)


Converted
0
1
count
2500
2000
1500
1000
500
0
Unknown
Finance Management
Human Resource Management
Marketing Management
Operations Management
Business Administration
IT Projects Management
Supply Chain Management
Banking, Investment And Insurance
Media and Advertising
Travel and Tourism
International Business
Healthcare Management
Hospitality Management
E-COMMERCE
Retail Management
Rural and Agribusiness
E-Business
Services Excellence

EDA (3)

Next Steps

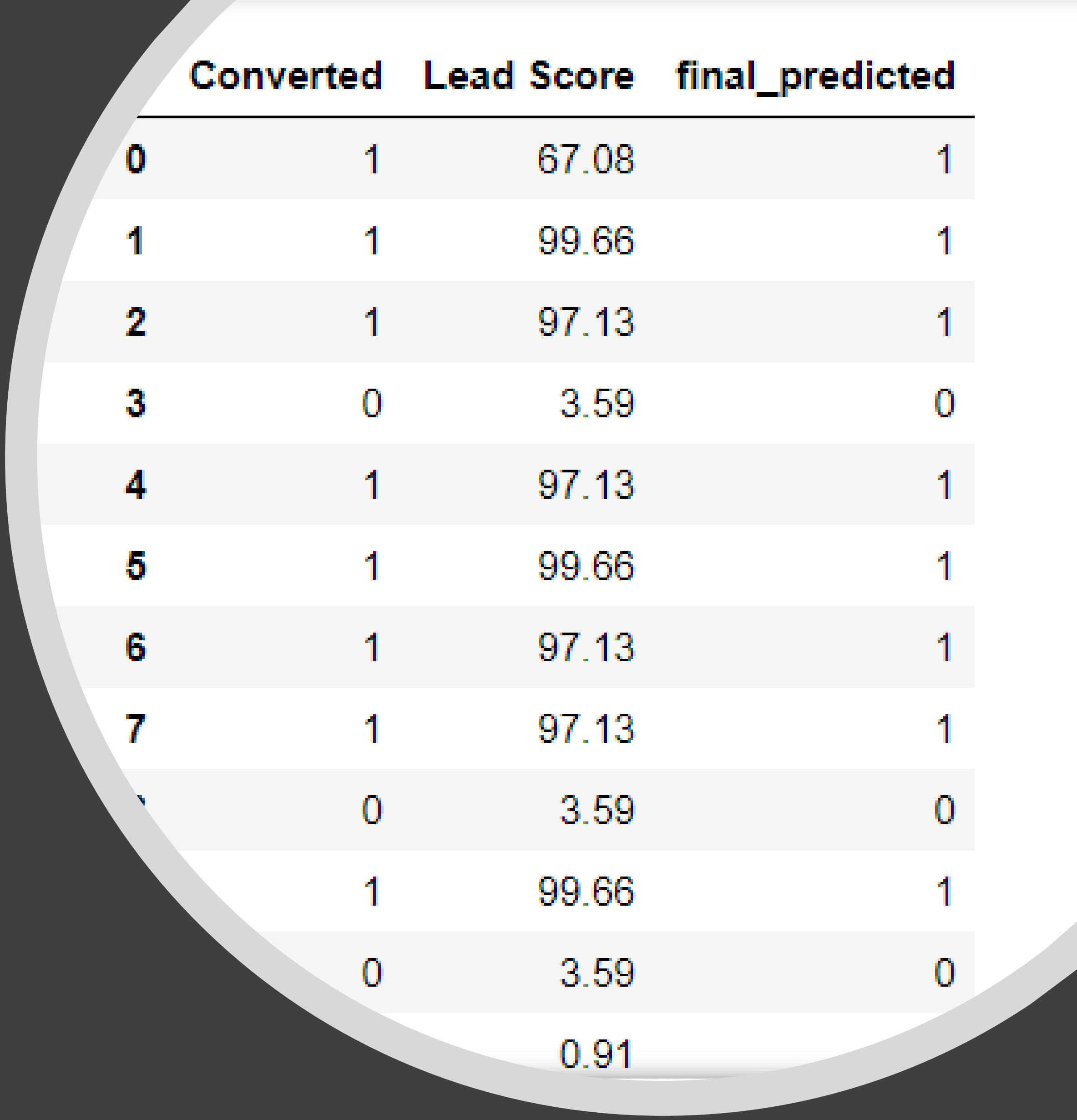
- Dummy variables were created for all the categorical variables
- Outlier check performed and no outliers found
- Test-Train split of 70-30 was done
- Features were scaled using the Standard Scaler
- Initial logistic regression model was built
- 15 features were selected using RFE
- Model was refined based on the p-values
- Multi-collinearity check with VIF

- Metrics (Train data)
  - Accuracy – 0.90
  - Sensitivity – 0.89
  - Specificity – 0.90
  - AUC (Based on the RoC curve) – 0.96
  - Precision – 0.85
  - Recall – 0.89

Model Evaluation

Findings

- Metrics (Test Data)
  - Accuracy – 0.90
  - Sensitivity – 0.90
  - Specificity – 0.90
- Top 3 Variables
  - Lead Source -Welingk_Website
  - Last Activity – SMS Sent, Will revert after reading email
  - Tags – Lost to EINS and Closed by Horizonn



	Converted	Lead Score	final_predicted
0	1	67.08	1
1	1	99.66	1
2	1	97.13	1
3	0	3.59	0
4	1	97.13	1
5	1	99.66	1
6	1	97.13	1
7	1	97.13	1
	0	3.59	0
	1	99.66	1
	0	3.59	0
		0.91	

Final Recommendations

- Sales team should aggressively reach out to potential candidates via call/email
- Ensure that communication is kept on via SMS as well
- Since a lot of people seem to indicate that they will revert after reading the email, there should be more aggressive follow-up once the emails are sent
- Since the Welingk Website seems to be a huge source of leads, the digital advertising of the same can be increased to ensure more traffic on the site
- All leads that are recently updated have more potential for conversion versus the inactive ones, hence the sales team should focus on such leads
- X Education also seems to perform well when the Leads are closed by Horizonn, so Horizonn should be more actively engaged in pursuing leads
- They also seem to be losing business to EINS, this can be researched to find the potential causes for the loss of business