Lead Score Case Study

By:- Sucheta, Suryakant Chaubey, Sudhir Kumar

The Problem Statement

Company

An education company named X Education sells online courses to industry professionals.

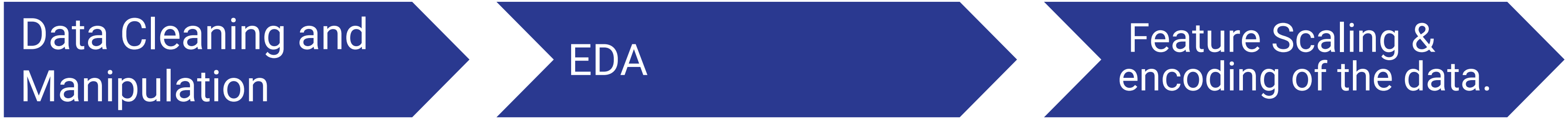
Context

On any given day, many professionals who are interested in the courses land on their website and browse for courses.

Problem statement

There are a lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom. In the middle stage, we need to nurture the potential leads well in order to get a higher lead conversion.

Solution Methodologies Part 1



Data Cleaning and Manipulation

EDA

Feature Scaling & encoding of the data.

- handle NA values and missing values.
- Imputation of the values.
- handle outliers in data.
- handle duplicate data.

Univariate Analysis

value count, distribution of variable etc.

Multivariate analysis

correlation coefficients and pattern between the variables etc.

Feature Scaling & Dummy Variables and encoding of the data.

Solution Methodologies Part 2

Classification Technique

Logistic Regression

Logistic regression Model is used for the model making and prediction

Model Validation

Validate the model on Train Test Data.

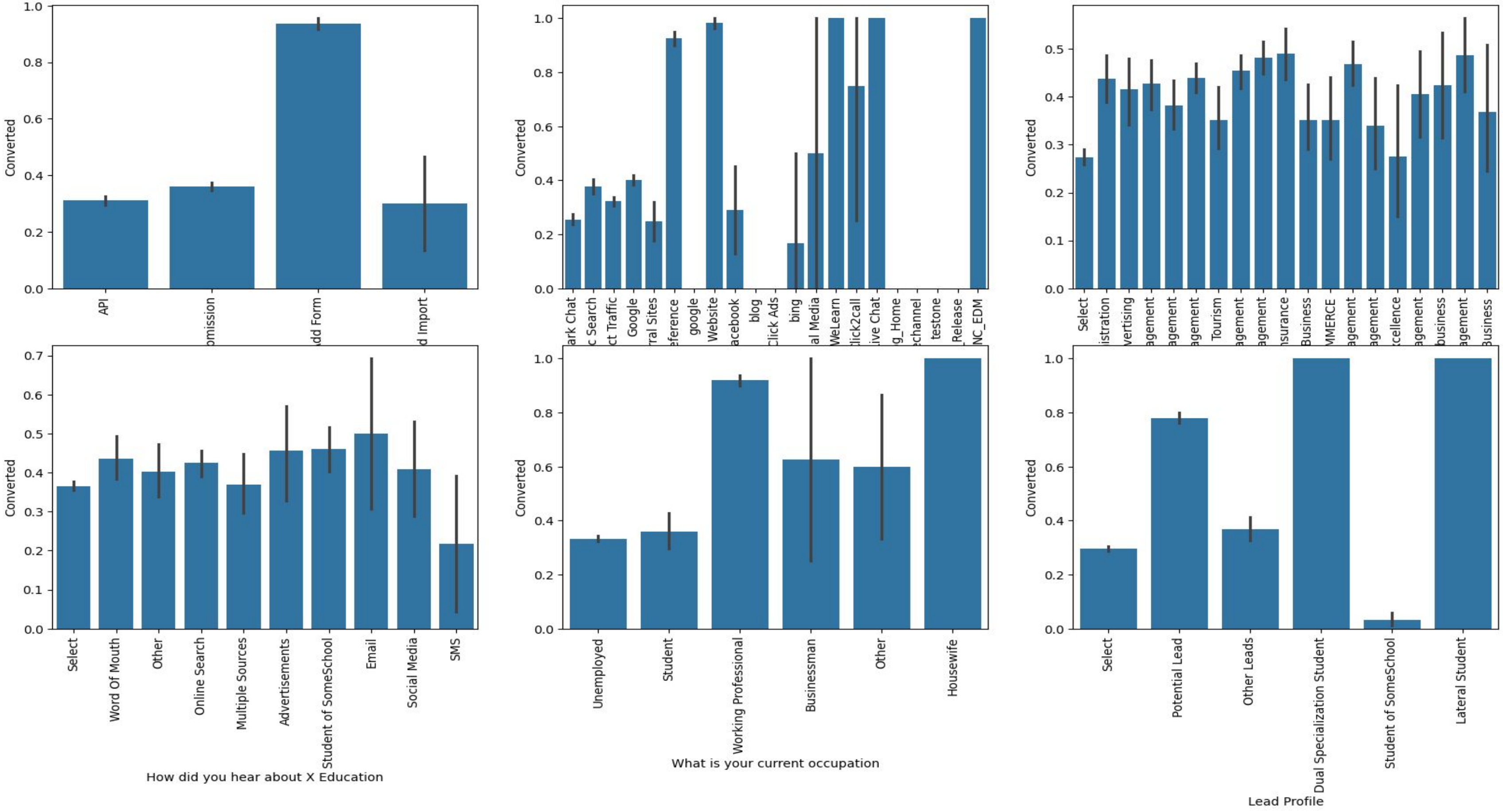
Checking VIF and P Score on Train and Test Data.

Conclusion and Recommendation

The Final Output and Recommendation from the Logistic Regression Model.

EDA

Categorical Bivariate Analysis



Data Converstion

Numerical Variables are Normalised

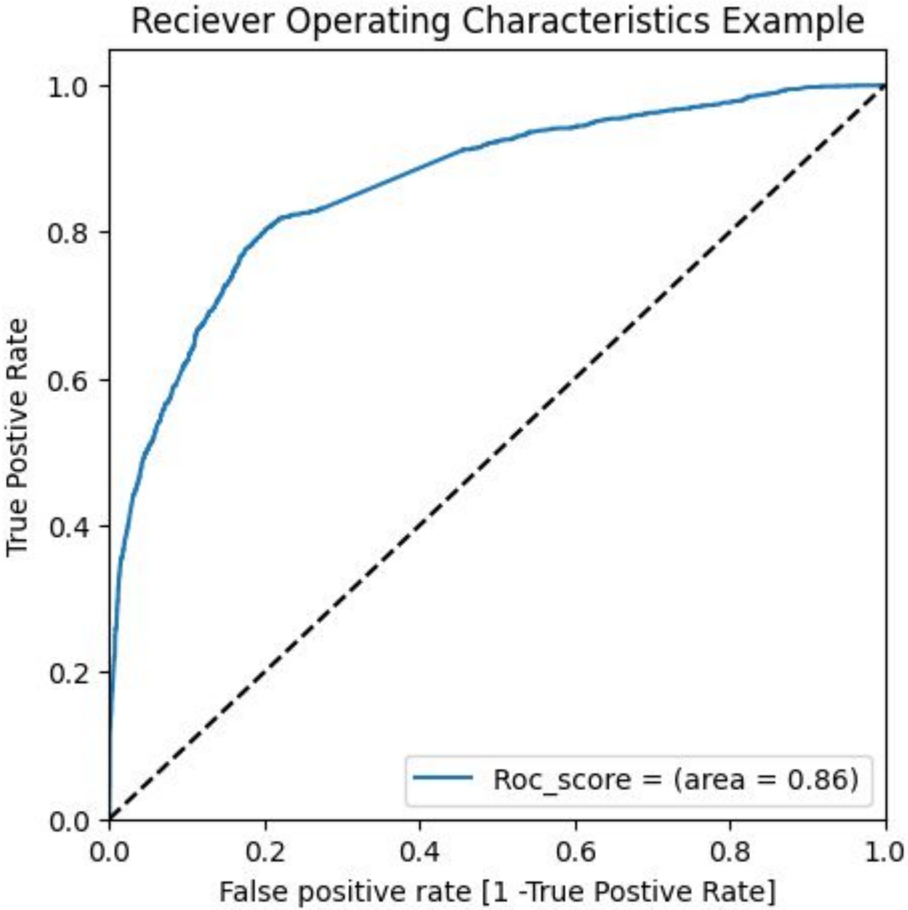
Dummy Variables are created for object type variables

Total Rows for Analysis: 9240

Model Building

- Splitting the Data into Training and Testing Sets
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- Running RFE with 15 variables as output
- Building Model by removing the variable and keep the variable where (p- value < 0.05) and (vif value < 5).
- Predictions on test data set
- Overall accuracy 80%

ROC Curve


Reciever Operating Characteristics Example
True Postive Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate [1 -True Postive Rate]
Roc_score = (area = 0.86)

Conclusion

It was found that the variables that mattered the most in the potential buyers are :

- The total time spend on the Website. 2.
- When the lead source was:
- Google
    - Direct traffic
    - Organic search
    - Welingak website

Conclusion Part 2

- When the lead origin is Lead add format. 4.
- When their current occupation is as a working professional.

Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses