Lead Scoring Case Study

DHEERAJTHEJA ALLURU
V V SATYA SAI KIRAN KUSUMANCHI

Lead Scoring Case Study

To build a Logistic Regression Model for a Company to assign Lead score for the given Leads and identify the target potential Leads that would churn.

Objective1

To improve the Lead conversion to 80 %

Objective 2

Implementation

All the work is present in Lead Scoring _CaseStudy_Submission.ipynb notebook

Exploratory Data Analysis

- Data Loading
- Data Cleaning
- Data Analysis

Data Preparation

- Create Dummies for Data
- Perform Train-Test Data Split
- Data Scaling

Model Building

- Feature Selection using RFE
- Optimal Cut-off point
- Model Evaluation

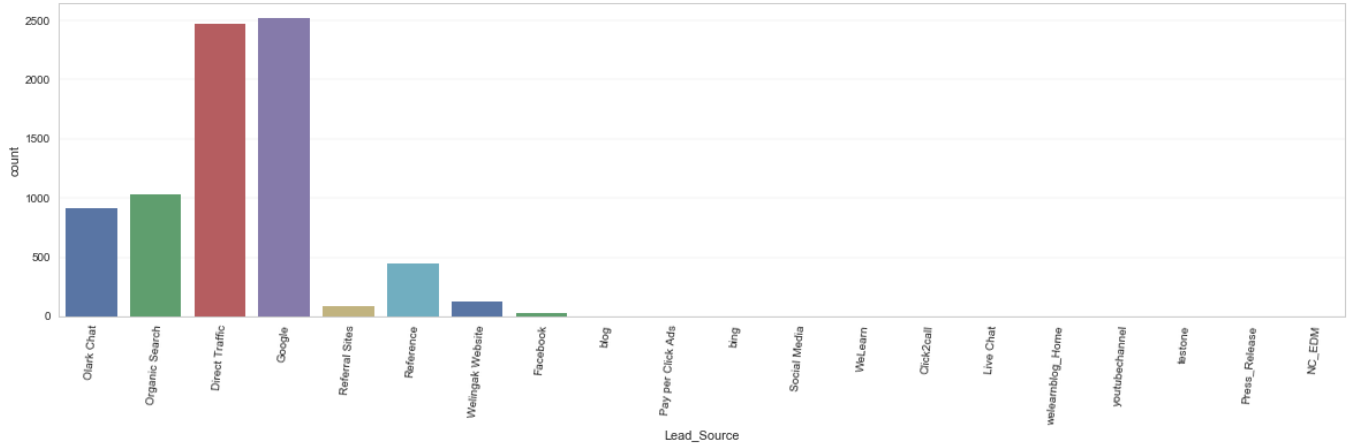
Business Value

Identification of potential leads that churn which would lead to an increase in the Business of the Company.

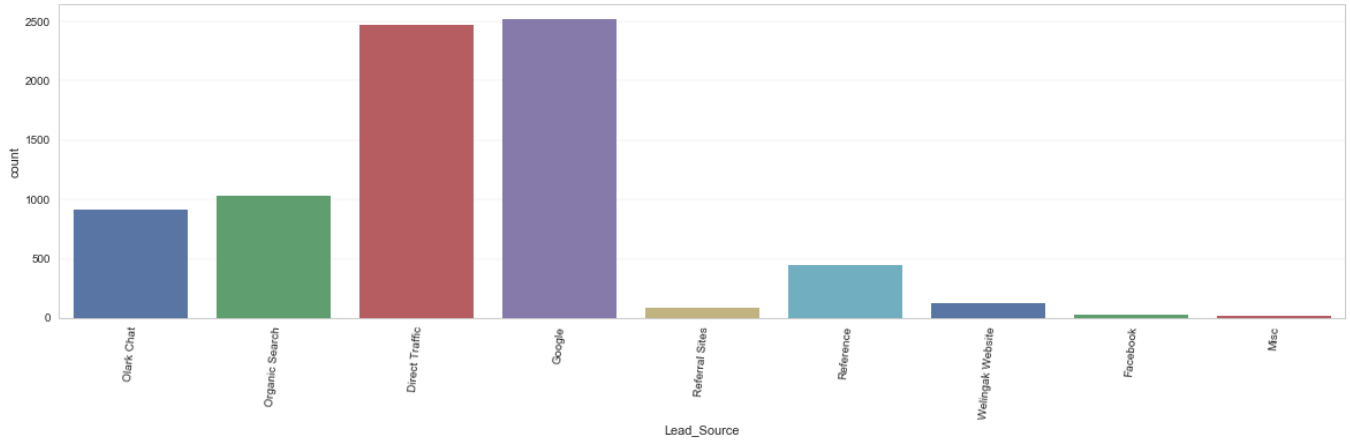
EDA – Data Cleaning

- Loaded the given dataframe using pandas.
- Cleaned the data by dropping few columns.
- Imputed data on few columns by applying mode on the column data.
- Merged few Values of columns into a single column as few columns doesn't have good amount of data.

Original values for Lead Source

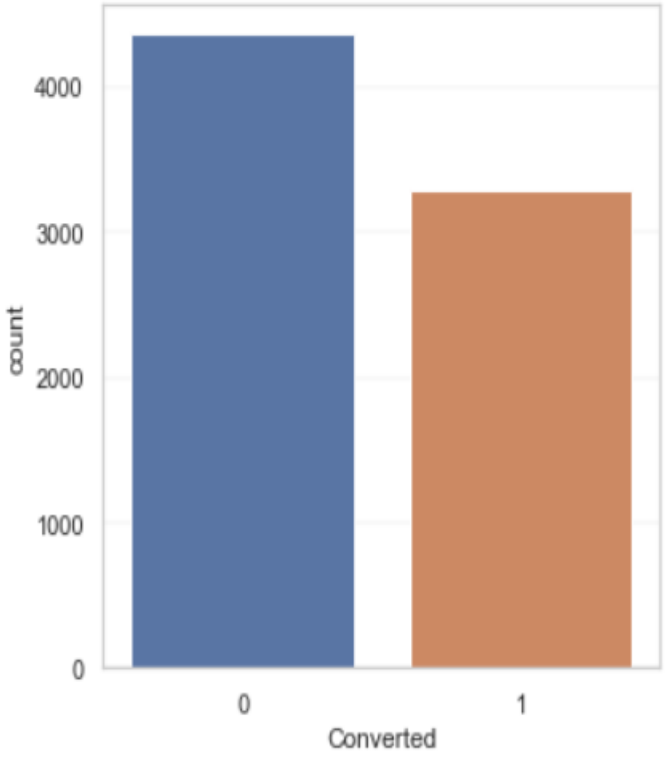


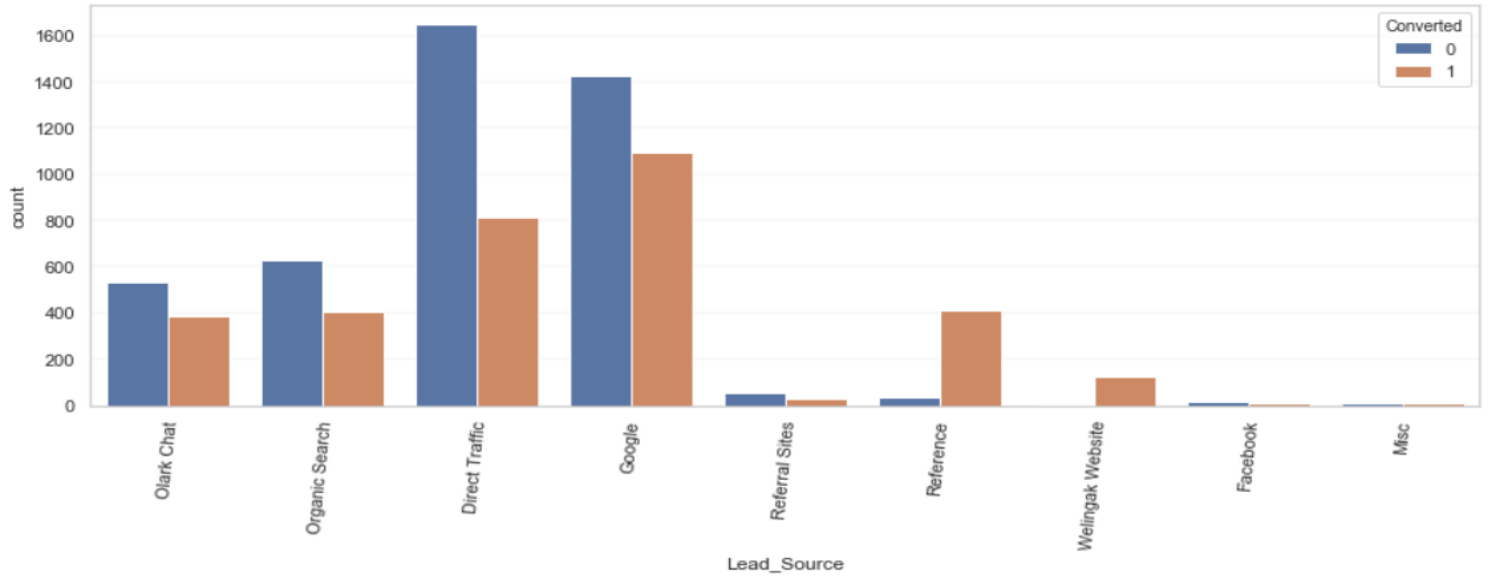
Values for Lead Source after merging into Misc

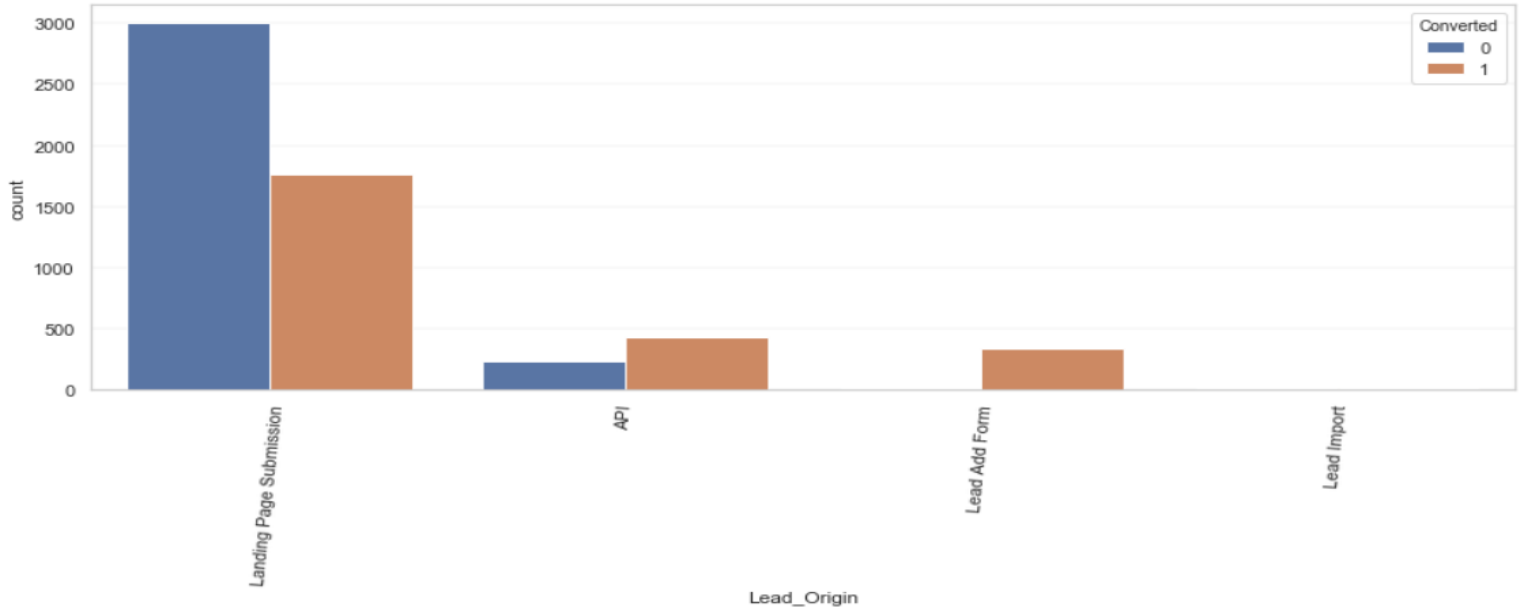


EDA – Univariate Analysis

- Performed the Univariate Analysis based on the Target Variable ‘Converted’.
- Presented are few of the columns and their distribution with respect to Converted variable.







EDA – Outlier Treatment & Scaling

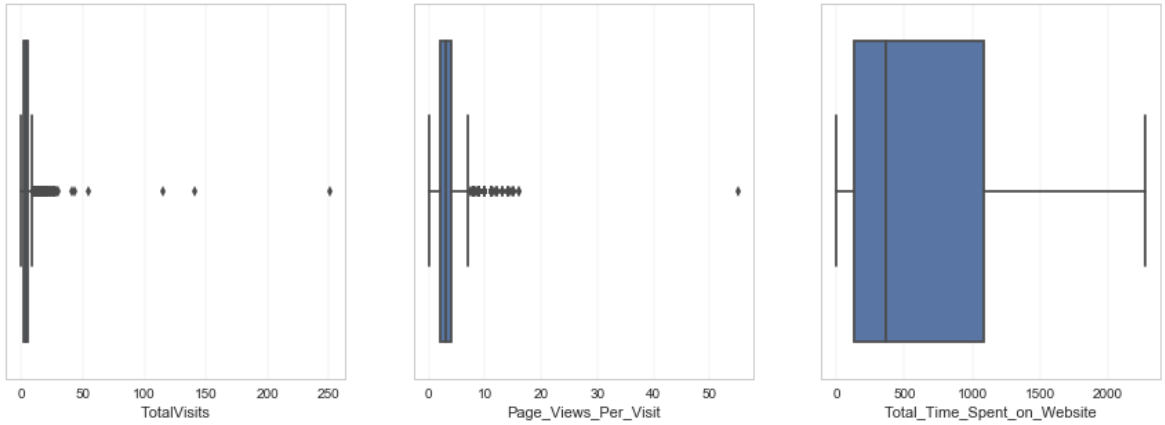
Outlier Treatment:

- Identified the outliers in the numerical data columns.
- Addressed the outliers for the columns – "Total Visits", "Page Views Per Visit", and "Total Time Spent on Website"

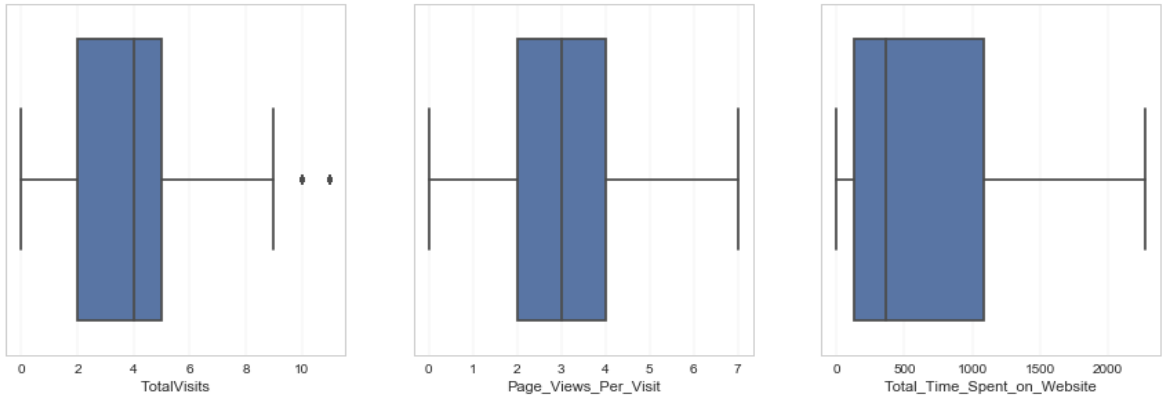
Scaling:

- Created Dummies for the Categorical variables in the data.
- Performed the Train-test Data Split on the data.
- Performed the Feature Scaling.

Prior to Outlier Treatment



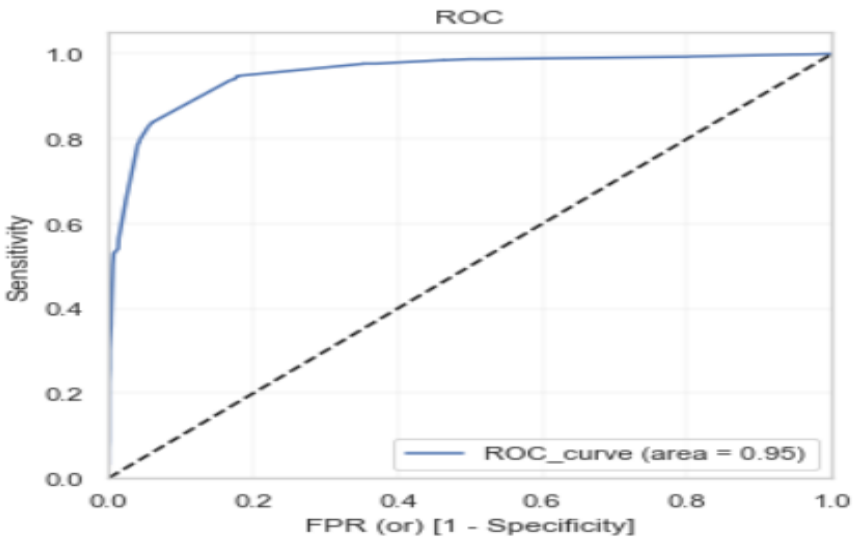
Post Outlier Treatment



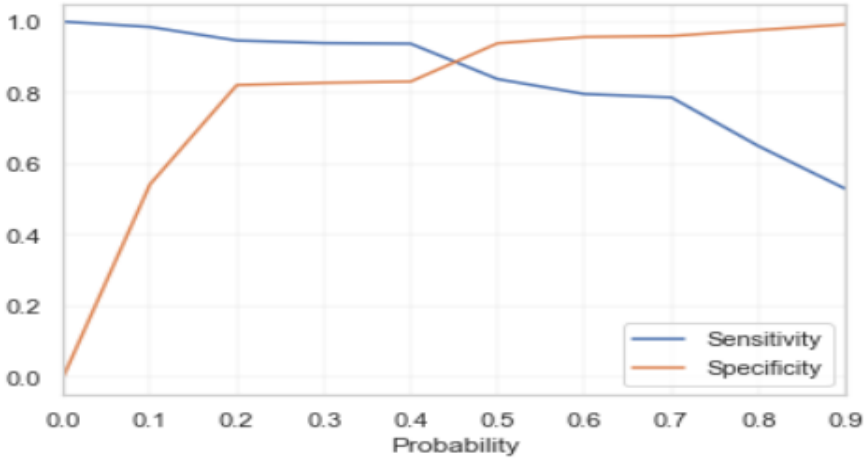
Model Building

- Feature selection done by applying RFE on the Train data.
- Dropped few features based on statsmodels outcome which have high p-value.
- Calculated VIF for the dataset to check if any features have high VIF values/multicollinearity.
- Plotted ROC Curve to find the AUC (Area Under Curve) and identified the value to be : 0.95 (95%), which is a good value.
- Optimal Cut-off point (probability) can be identified from the adjacent plot drawn where the intersection(balance) between Sensitivity and Specificity is noticed. Optimal Cut-Off probability here is : **0.45**
- Made predictions on the Test dataset to obtain the final model. And calculated the metrics of the model.

ROC Curve



Optimal Cut-off Point Identification



Conclusion

Objective of this Exercise:

X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Conclusion:

Here if we see as per the motto we need to predict as many Actual "1's" as 1 and Actual "0's" as 0 and the target from CEO of X Education is to get 80 % lead conversion. So this means we need to have "Sensitivity" of our model to be at least 80 %, here we have got the 92.49% (i.e greater than 80%) and as per the first ask we need to predict correct 1's and 0's i.e accuracy is also good here with 86.94%.

And the other metrics are also well in place:

- Sensitivity - 92.49
- Specificity – 82.76
- Accuracy - 86.94
- Precision – 80.14

Thank You