Machine Learning I
Logistics Regression Case Study

Due Date: 8th September 2021

Problem Statement:
Build a Model with a ballpark of target
lead Conversion rate around 80%

Gayathri MNVL
Ramasubramaniam P

gayathrimnvl@hotmail.com
Prs1712@gmail.com

Summary

Analysis done using a X Education past historical data on their Lead Generation. Basic data has lot of information about how the potential learners joined the courses or not joined.

We need to find ways to get more industry professionals to Join the courses using the data/information provided from past history.

We have used Python and it's Library such as Numpy, Pandas, Seaborn and Sklearn to analysis our data and develop a model.

We have defined the following steps to get a final conclusion and recommendation to CEO of X Education.

- Data Cleaning
- Data Analysis (EDA)
- Data Preparation
- Model Development
- Model Prediction
- Final Conclusion

Data Cleaning:

We received around 9240 rows and 37 columns of raw data from X Education based on their past historical data.

Categorical and Numerical columns with missing values above 35% and unnecessary columns are dropped.

Categorical columns with missing values between 5% ~ 35% are imputed with most frequent values.

We have identified the Column 'Specialization' is important for the X Education company to make use for hot lead generation

We considered the value 'Select' as equivalent to 'NaN' and considered as missing value.

We dropped rows with missing value less than 5% and all Binary columns skewed above 95% have dropped.

Data Analysis:

We used Data toolkit procedure to analysis our given dataset. Plotted respective graph to analysis Univariate/Multivariate for Both Numerical and Categorical Variables. Used HeadMap to identify correlation matrix between the variables.

Data Preparations:

Based on the analysis, we identified columns which are irrelevant. We also analysis outliers, and those outliers handled by StandardScaler.

Using Sklearn from Python, we created dummies from the data frame to all categorical columns and Converted Binary variables into Boolean.

Model Development:

Used Sklearn and Stats Model to develop Logistic Regression. Finalized the Model, if we found no p-Value above 0.05 and VIF above 5. Iteratively performed to build model till we achieve the above conditions. (p-Value <0.05 and VIR < 5)

Model Prediction:

Based on the above Model, we applied it in our test data. Based on the ROC Graph, we have taken Optimal Cutoff as 0.35.

Calculated Accuracy, Sensitivity and Specificity respectively.

Final Conclusion:

Based on the Predictions on the test data, we finally predicted potential lead if probability is > 0.41.

Final conclusion and recommendation to X Education CEO, based on the Logistic Regression Model, and Ask in the Problem statement adhering to the 80% ballpark conversion rate, model says that Top significant features which helps to convert into hot leads as follows.

1. Last Notable Activity_Had a Phone Conversation
2. Lead Origin_Lead Add Form
3. Lead Notable Activity_Unreachable
4. Lead Source_Welingak Website
5. Last Activity_SMS Sent
6. Total Time Spend on Website.

Keeping all the above features, X Education can target those potential lead and high change to get all those to enroll into the Courses.