Leading Score Summary

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

X Education needs help to select the most promising leads, i.e. the leads that are most likely to convert into paying customers. A model is required to be built wherein a lead score is assigned 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%

1. Data Reading and Understanding:

Here we tried to get the look and feel of the data, we observed following things

- Number of rows and columns
- Data types of each columns
- Checking first few rows how data looks
- Checking how the data is spread.
- Checking for duplicates, if any

2. Data Cleaning:

Here we checked for discrepancies in the dataset

- Checking for any column names correction
- Checking for null values and imputing them with appropriate methods
- We used mode imputation for categorical columns.
- We used mean imputation for numerical columns, if there is no skewness in data.
- We used median imputation for numerical columns, if there is skewness in the data.

3. Data Visualization and Outliers Treatment:

- We performed univariate analysis on categorical column to see which columns makes moresense and removed those columns whose variance is nearly zero.
- We performed bivariate analysis on categorical columns to see how they vary w.r.t Convertedcolumn.
- We performed univariate analysis on numerical columns by plotting box plots to see are thereany outliers in the data or not.
- We performed bivariate analysis on numerical columns with Converted column to see howthe leads are related to these columns.
- In this step we also plotted the correlation matrix to identify the columns which arecorrelated.

4. Feature Scaling

- At this stage our data was very clean and no outliers. We know that logistic regression takes theinput parameters as numerical values. Hence, we converted all the categorical columns to numerical.
- Columns which have only two levels “Yes” and “No” were converted to numerical using binarymapping.
- Columns which have more than two levels were converted to dummies using pd.get_dummies function.

5. Model Building

- Firstly, RFE was done to attain the top 15 relevant variables. Later the rest of the variables were removed manually depending on the VIF values and p-value (The variables with VIF < 5 and p-value < 0.05 were kept).

6. Model Evaluation:

- A confusion matrix was made. Later on the optimum cut off value (using ROC curve) was used to find the accuracy, sensitivity and specificity which came to be around 81.05% each.

7. Prediction

- Prediction was done on the test data frame and with an optimum cut off as 0.35 with accuracy, sensitivity and specificity of 80%.

Conclusion

It was found that the variables that mattered the most in the potential buyers are (In descending order) :

1. The total time spend on the Website.
2. Total number of visits.
3. When the lead source was:
   a. Google
   b. Direct traffic
   c. Organic search
   d. Welingak website
4. When the last activity was:
   a. SMS
   b. Olark chat conversation
5. When the lead origin is Lead add format.
6. 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.