Lead score Case study

Logistic regression model

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%.

Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.
As we can see, 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, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion.

Our target is to analyze and make a logistic regression model to predict who will be converted at the end using previous data.

Steps for logistic regression model

- Collecting and reading the data
- Data treatment
    - Filling the missing values with nan or adequate value
    - Dropping columns having large number of null/missing values
    - Dropped those values with only one value.
- Exploratory data analysis
    - Univariate data analysis
    - Multivariate data analysis
- Data preparation
    - Forming adequate dummy variables for model building.
    - Splitting the data into train and test sets for training and testing the data.
    - Scaling the numerical variables using a standard scaler.
- Model building
    - Summarizing the data
    - Feature selection using RFE

    - Assessing the model with StatsModels
    - Obtaining an adequate model using the Value inflation factor method.
    - Creating Prediction
    - Model Evaluation
    - Checking precision and recall values
    - Plot ROC curve
    - Finding optimal cutoff point
    - Making prediction on test data set
- Conclusion

After the model evaluation we can conclude that variables like lead source, occupation , The reason to opt for the course etc are some very useful factors affecting the conversion probability. After the model evaluation the conversion probability is coming 80% approximately which was 30% before we applied the logistic regression model on the previous data.