Module on Machine Learning I- Lead Scoring Case Study

Agenda – To work on the following problem Statement:

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%.

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.

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 a 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%.

Objective:

To build a logistic regression model to assign the score between 0 and 100 to each of the leads which can be used by the company to target potential leads. Also, as per future requirement changes, we need to handle these as well.

The steps are used:

Data cleaning.

Preparation of the model.

EDA, train and test the data set.

Gini Model and Specificity, Accuracy and Sensitivity.

Model Outcome:

- Model outcome for lead score above 35.
  Higher the lead score, the more chance the lead customer to get converted.
  Average lead score of converted leads 68, of non-converted leads 15.
  Optimum probability cutoff is 0.5

Features of model:

- Lead origin and time spent on websites
- Lead Quality and source
- Efforts made to filter out the specialization-others.
- Area under the curve ROC.
- The observation on train and test data which was based on accuracy, sensitivity and specificity.
- The coefficients of the data was measured.
- Lead Add form higest conversion.
  Total time spent on website: People with most total time have high conversion rate.
  Lead Quality: Efforts to be made to correct relevance of person in charge of lead.

- **Train Data from observation**

Sensitivity	**Accuracy**	**Specificity**
91.11	**95.72**	**94.50**
90.78	**84.18**	**94.60**

- **Test Data**

Sensitivity	**Accuracy**	**Specificity**
91.11	**95.72**	**94.50**
90.78	**84.18**	**94.60**