A CASE STUDY: LEAD SCORING

Leading scoring is one of the most popular in business. It estimate the probability how a lead – which is a potential customer would become a real customer. X Education wants to improve the performance of the process of interacting with lead, to saving the effort of the team.

A user, becoming a lead when they have accessed to X Education’s system: website, ads, surveys,...
From these customers’ behaviors, the list of leads is built and the staffs will contact to them and persuade they become the real customer. This process could take much effort than it’s necessary.
Base on existing data, we try to build a machine learning model to handle this.

1. The dataset includes many attributes. It could be numeric or categorical and contain missing values which is popular in raw surveys. And any of them might have the same meaning that noises the model.
2. First, with missing value in numeric data, to avoid negative in variance, replace them with mean values. With categorical, it can be turn to ‘Other’
3. Second, we try reduce the variety of attributes dummy categorical variables since we need to apply dummy to these attributes and as we know, many unnecessary one worse the model. In this case, we only keep the top 5 popular values, other are merged into group ‘Other’.
4. Then, we try to normalize the data to enhance the model perfomance. We apply Z-score method for 3 atributes: 'Total Time Spent on Website','TotalVisits','Page Views Per Visit'
5. After that, we have a data with 40 variables, that’s large than it should be. We try to use RFE to reduce the data’s dimension. At the end of this step, we have the 5 atrributes that is believed fit best to model

The accuracy of model is about 85%, an considerate number to believe in practice. In fact, in some case, accuracy is not the most important metric we need to count. Specificity and Sensitivity are also crucial in special situation. The result of model return a probability for each input sample. Our mission is to choose an appropriate threshold for each condition.