With this lead scoring case study, we had to create a machine learning model based on logistic regression to help identify leads most likely to convert to paying customers for an online education company.

The approach began with exploring and cleaning the data. Here we looked at missing value extensively. A notable point was that a number of datapoints were marked as “select” instead of the required values. We considered these as missing values and proceeded to deal with them as such.

We discovered that there could be cases where neither deleting not imputing the data would be very accurate, such as in the case of the current occupation column. We resolved this by simply letting the missing values remain as is and assigning them a value “missing”.

Additionally, a lot of columns contained binary yes/no data. We converted these to numerical 1/0 values for ease of analysis.

Upon further inspection of the data, we discovered that for a number of columns, a single value made up a large percentage of column values. We determined that these columns would not add value to the analysis and model building as they would end up skewing the data, and so they were removed.

There were also certain categorical columns with a large number of options for the data. This could have led to many dummy variables and also a chance of overfitting. In order to generalise the data, we created buckets such as “others” to group data points with lower percentage occurrence.

Once the data cleaning was done, we proceeded to create the model. Basic data preparation steps like splitting into train-test sets and scaling numerical variables were done before this.

To create the model, we used Recursive Feature Elimination (RFE) as a starting point to get a list of coefficients that work. What we discovered was that with this set of coefficients itself we were able to get a model with p-value and VIF within the required parameters. This could be due to the thorough data cleaning and preparation steps that left us with crisp data points for modelling.

We focussed on optimising sensitivity for the model. This is because the goal of the exercise was to correctly identify as many true positives i.e. converted leads as possible. With this in mind, we determined a cut-off point to obtain a conversion rate above 80% as was asked by the CEO. We further ascertained the quality of the model with ROC curve.

The next step was to run the model on the test data to see if we were able to achieve similar results. We got a similar accuracy of 79% and sensitivity of 82%, indication our model can generalise the predictions well.

Using the predicted conversion probability that our model generated, we were able to calculate and assign scores to each of the leads indicating their likelihood of converting.

Our model found that the probability of the leads converting was based on the time they spent on the website, how many times they visited the website and if they were working professionals. Leads brought in through filling the form and from the import list as well as those sent to the site from the Olark chat platform also showed a higher tendency to convert.