Summary of Lead Scoring case study

Problem Statement:

What is a Lead:

There is an education company named X. This company sells online courses to industry professionals. The company markets their products on different platforms for example google or other search platform. When these people land on company's website they might browse the courses, 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.

What is a Hot Lead:

Once the initial pool of leads is acquired, then sales team start making calls, writing emails, etc. From the initial pool of leads they want to nurture the leads. These leads are called Hot leads. When the hot lead opt for the admission then they are called converted lead

What is the Aim of this case study:

The conversion rate of Hot leads to converted leads is 30% for X education. So using the data provided by the company we have improve this conversion rate.

What is the result of case study:

With the data provided above we have managed to create a model that is able to predict 80% people who are highly likely to be converted.

Problem Approach:

Solution:

The company provided us with approximately 9000 records. Of which 38 are converted leads and 68% are not. A total of 22 features are there which are the parameters of those converted and non converted leads.

The solution is divided into 5 steps:

Step 1: Loading Data

Initial look at the data and shape. And Identifying the conversion rate.

Step 2: Cleaning Data

The second step is data cleaning. There was a lot of missing data. A total of 13 columns had more than 15 % of missing data. So initially we removed the columns. Our thought process was that if don't get a better model by removing these many columns, then we will try to use techniques like Mean, Median, Mode or random to fill the values. But since in the end we achieved 80% in the end. So we decided not to ingest fake values in the data.

Also There were 4 other columns having approx. 1 % of missing values. So we straight away removed the rows.

Now our data was left with categorical values and non standard values. Before converting to dummy variables we removed "Prospect IDS" and "Lead Number", because they were unique and not important. Then we created dummy variables for all the categorical values.

After converting dummy variables we divided the data into Train, Test split. And then standardised the training values.

We also looked at the outliers in the above data and found that “Total Visits” and “Page vies per visit” had outliers. There is no much data spread except a few points which are lying outside the scope, therefore we decided not to remove outliers.

Step 3: Applying PCA

Since we have 71 columns to analyse the best possible solution is to use PCA. And after applying PCA and analysing the scree plot we found that around 10 variables were able to explain 90 % of the variance in the data.

When we mapped these principal components with original columns we found that categorical variables “do not email” and “do not call” while numerical variables “total visits” and “page view per visit” were leading for explaining the data.

Step 4: Applied Logistic Regression

On the data above we applied logistic regression model with default setting using sklearn.

Step 5: Analysing data

We ran our model on the training dataset to see the accuracy, recall, precision and f1 score. We found that for training data the values were 78%, 67%, 75%, 71% respectively. While our test results were 80%, 69%, 73%, 71%.

Final Result:

Our Accuracy on test score reached 80%