Summary Report – Capstone Project

Lead Scoring Case Study

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. They have appointed you to help them select the most promising leads, i.e., the leads that are most likely to convert into paying customers.

Methodology –

Firstly, the data was understood carefully. According to the data, it had 37 features, out of which there were 30 categorical variables, and 7 numeric variables. After treating the imperfections in the data (omitting attributes having larger null values, and removing null records), there was around 69% of data still remaining. It was time to prepare the data for modelling.

First of all, we created dummy variables for the categorical variables, and scaled the numeric variables using normalized scaling. Then, the data set was distributed into a 70:30 training and test set ratio, and RFE (recursive feature elimination) was done to select 15 prominent features. These features were –

```
Total Time Spent on Website', 'Lead Origin_Lead Add Form', 'Lead Source_Olark Chat', 'Lead Source_Reference', 'Lead Source_Welingak Website', 'Do Not Email_Yes', 'Last Activity_Had a Phone Conversation', 'Last Activity_SMS Sent', 'What is your current occupation_Housewife', 'What is your current occupation_Student', 'What is your current occupation_Unemployed', 'What is your current occupation_Working Professional', 'Last Notable Activity_Had a Phone Conversation', 'Last Notable Activity_Modified', 'Last Notable Activity_Unreachable'
```

After selecting these features, we ran logistic regression model on it, using the IRLS (iteratively re-weighted least squares) method, to find the optimum model. After omitting 4 features from the above list (based on the p-values and VIF values), we finalized our regression model.

After the model was built, we tested it on the training set to check its accuracy, sensitivity, precision, recall etc parameters to evaluate it using an arbitrary cutoff of 0.5. After that, we used the ROC curve and sensitivity-specificity trade-off to determine the ideal cut-off point, which came out to be around 0.42.

Finally, we predicted the values for our training and test sets, and evaluated the models on both the sets, to see if the model was efficient or not.

Results

These results told us that our model was working efficiently and was able to provide valuable and accurate predictions for lead conversion, almost $4/5^{th}$ of all times. The final results are summarized in the table below -

Evaluation Parameter	*Training Set*	*Test Set*
Accuracy	78.995 %	78.922 %
Specificity	79.238 %	79.016 %
Sensitivity	78.734 %	78.820 %

Hence, using logistic regression, we were able to build a predictive model that will help enhance the Sales Funnel of the company. This was the complete methodology followed in completing the capstone project on the topic 'Lead Scoring'.

Learnings from the Project

1. How to approach a business problem using data science
2. The proper data science methodology to be followed.
3. Optimizing a regression model to the best manner possible.
4. Practical application and usage of Python and Data Science to solve real-world complex problems.