Lead Scoring Case Study: Summary Report

Problem Statement:

X Education sells online courses to industry professionals. Many interested professionals visit their website and browse for courses. The company generates leads through past referrals or visitors providing their contact details. Through the current Sales process, only around 30% leads get converted. To improve this, X Education wishes to identify the most promising leads for the Sales team.
The company wants us to build a model for leads identification with around 80% probability of conversion.

Data:

The dataset provided has around 9000 data points and has attributes such as Lead Source, Total Visits, Last Activity, etc. which may or may not be useful in ultimately identifying the 'Hot' leads.

Analysis:

As the requirement is to ultimately categorize the lead conversion in 2 buckets – will convert or not convert, it is a Binary Classification problem and we can use Logistic Regression to solve it.
To begin with, we used Pandas to import the Leads dataset along with the data dictionary to understand each variable. We then began non-graphical exploratory data analysis where we handled missing values, corrected data types etc.
Post that we began with graphical EDA where we did univariate analysis. Then we performed bivariate analysis to check correlations between 2 variables using scatter plots and pair-plots. Finally did multivariate analysis using heatmap to understand the relationships better. Then to prepared data for modelling, we created dummy variables for categorical variables, split our dataset into train and test and performed feature scaling.
Then initiated with modelling process using statsmodels library. We began feature scaling with RFE where it helped us identify the top 15 variables which are influencing the model the most. We then checked the P-values to see if the selection has happened genuinely or by chance and VIF values to detect multicollinearity.
We then kept removing variables until we found a model with P-values <0.05 and VIFs < 5.
Post that made these predictions on training dataset.
To evaluate our logistic regression model, we measured beyond accuracy. We created confusion matrix, checked sensitivity vs specificity or precision vs recall and ROC curve. We also found optimum value of cut-off using Precision vs Recall. As per our model, the leads above this cut-off are likely to convert and the leads below the cut-off are more likely to not convert. We finally made these predictions on test dataset and evaluated our model.

Recommendations:

Our Sales team should focus on leads which are closed by Horizzon, lost to EINS, who will revert after reading the email, are busy, lead source is Welingak Website, last notable activity performed by the student is SMS sent and those which are

coming from add form. These are some of the factors which positively impact the lead conversion.

Whereas factors like Ringing tag, last Activity performed by the customer is Olark Chat Conversation or Email Bounced, Lead Origin is Landing Page Submission, Specialization is Not Specified by the customer and Customer is Unemployed, negatively impact the lead conversion rate.