Assignment Summary Report

Problem Statement:

X Education is an online eduction provider to industry professional. They market their products through various digital media and capture the leads(people who click ads/ watch videos/ enter contact info) and have sales team try and convert those leads to actual customers. The conversion rate though is not very great and hence, X education has appointed us to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires us to build a model wherein we need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%. They want to solve this problem using data and analytics.

Approach taken:

To approach this problem we are given a dataset with >9000 data points and we need to utilize them to work on a target variable 'Converted' so that we increase the overall lead conversion ratio. We need to identify the potential 'Hot Leads' so that sales can specifically focus on those leads and try and convert almost all of them. We chose to do the following steps to go forward with this case study:

1. Read and understand the data
2. Clean the data - Data Preparation (This includes handling nulls and treating outliers and making the dataset to a most useful form so that we can drive best results)
3. Performing univariate and bivariate analysis to get a glimpse and summary deductions from the dataset
4. Creating Dummies for categorical Variables - Data Preparation
5. Performing train-test split and scaling - Data Preparation
6. Modelling - RFE, GLM, optimal cutoffs
7. Check the validity of the model by using cunfusion matrix and metrics such as Accuracy, Sensitivity, Speficity, Recall and Precision
8. Test the data on the test set to check how the model is performing
9. Drive insights on the outcome and suggest recommendations based on optimal threshold

Steps:

1. **Data Cleaning:** There were several columns with >30% null values so we had to drop these columns, also apart from nulls there were several columns which had Select as

one of values so we had to consider those as well as nulls to clean/ drop / impute our columns

2. **Data Preparation:** Check for duplicates in columns like Prospect ID, remove columns which had just singular values as these values don't really impact the analysis, convert the Yes/No Flags to 1/0 values so that they can be directly used in modelling. Next steps would be to create dummy variables of categorical variables and create the test train sets. Post which we would need to scale the numerical variables so that all are standardized.
3. **Modelling:** We use GLM modelling and logistic regression to classify the leads to Converted or not. Post which we use RFE to contain the variables to 20 and then use manual fine tuning to remove low significance and multi collinear variables. Post this modelling, we predict the values for the train data sets and set a threshold to convert the probability to predicted 1/0 flags
4. **Accuracy and Metrics:** Accuracy came out to be 81% and then Specificity and Senstivity were also calculated. Sensitivity came out to be nearly 80% which is quite high, recall value also came out to be 80%.
5. **Testing the model on test dataset:** This step is done to compare the metric values of accuracy, sensitivity, recall etc of the test dataset and train data set. In our case both came out to be same, which means the model is really good to be used in production

Results and Recommendations:

Based on the the above result we see Lead Origin : Lead Add Form(In a positive fashion), Occupation: Unemployed (In a negative fashion) and Occupation: Unknown / Student (Negetively) impact the model the most and we used focus on the leads which have positive coefficients and leave those who have negative cofficients

Sales team should primarily focus on people who are:

• Spending a lot of time on the website • Leads are adding form • Leeds from Olark Chat • Leads where the last activity was sending SMS

Apart from these they should focus less on the leads with the following characteristics:

• Leads who do not email • Leads who already have been converted before • Leads who leave after Olark Conversation, do not proceed forward • Leads who are Students, Unemployed or their occupation is unknown • Leads who are not very active in terms of Activity of platforms