Summary Report - Lead Score Case Study

The X Education company requires us to build a model wherein we need to assign a lead score to each of the leads between 0 and 1 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.

So, we build a logistic regression model to assign a lead score between 0 and 1 to each of the leads which can be used by the company to target potential leads.

The following are the steps that we have used:

1. Reading and Understanding of the Dataset
    - The Data provided to us by the company has the 9240 data points as rows and 37 attributes as columns. Out of this 37 attributes, 30 attributes are Categorical and the rest 7 variables are numerical.
    - When we checked the data info, then we found that there were missing data points /Null Values for various attributes in the dataset which we need to handle before making the model.

2. Data Cleaning
    - Once we know the percentage of missing values was available in each columns, we dropped those variables that has missing count of more than 40%.
    - For the Columns in which missing count was less than 40%, we handled these values by replacing with their respective mode values (for categorical variables) and median value (for numerical value). For some categorical columns where the missing values were over 25% of their data, we made a new category such as Not Available/Not Specified etc. so that we don't lose any data in our analysis.
    - After treating the missing values in the dataset, we also handled the outliers in the numerical columns by dropping them.
    - We also found the few categorical variables such as country of customer, Magazine, etc. were having imbalance when we were checking the data imbalance in the data frame. We decided to drop them as they don't provide any useful info due to singularity in nature of the data in our model building.
    - Also at end, we dropped the unnecessary columns such as prospect id, lead number as they hold values in our analysis.

3. EDA – Exploratory Data Analysis
    - After cleaning the data, we started our exploratory data analysis (EDA) process with the remaining attributes. We have done Univariate Analysis of each attributes and Bivariate Analysis of attributes with the target variable "Converted" to check whether is their trend in data.
    - During Our EDA process, we inferred lot of insights out of the data.

4. Data Preparation for Model Building
    - For Preparing the Data for model building, we converted the all categorical variables into the dummy variables and dropping the Not Available/Not Specified of each attribute.

5. Feature Scaling using Standard Scalar
    - In preparing the data for model building, for the numerical variables, we scaled them with Standard Scalar from Scikit Learn.

6. Train-Test Split of the data
    - After preparing the data for Model, then we created two data frame as “X” and “y” where X has all the features from the original data frame except target variable data in it and y has only the target variable data in it.
    - Now, we split the data in 70-30 ratio as train and test dataset respectively.

7. Feature Selections with Recursive Feature Elimination (RFE)
    - Here, we run the feature selection algorithm using Recursive Elimination (RFE) method to select only 20 features for model.

8. Model Building
    - Once we used recursive feature elimination to remove most of features and left with only 20 features. We initialize the Logistic Algorithm and fit & transform our model with train data set.
    - Later, the variables were removed manually from the model depending on their p-value and VIF values (The variables with VIF < 5 and p-value < 0.05 were only kept).
    - At the final model, we left with 16 variables in the model where all the p-values of all variables is less p-values (less than 0.05) and low VIF values (<5.0).

9. Model Evaluation
    - We found the accuracy of our model based on the train data by comparing the actual y train and predicated y values.
    - For evaluating our model, The Confusion Matrix was also made. With the help of confusion matrix, we calculated the different metrics such as specificity, sensitivity, false positive rate, positive predictive rate and negative predictive rate for our model.
    - Later on, we also found the optimum cut off value (using ROC curve and plot between accuracy, sensitivity and specificity) and we got an optimum cut off as 0.35.
    - With optimal cut off value as 0.35, we calculated the accuracy of our model and other metrics as well shown in the final observation.

10. Making Prediction on the Test data
    - Here, we fitted test data for making prediction and checking the various metrics.
    - We calculated the accuracy of our model and other metrics as well based on the test data shown in the final observation.

11. Final Observation
    - Overall the Linear regression model that we build has the following metric values.

S No	Metrics	Value from Train Data	Value from Test Data
1	Accuracy	*90.96 %*	*90.86 %*
2	Sensitivity	*89.55 %*	*90.17 %*
3	Specificity	*91.83 %*	*91.32 %*
4	False Positive Rate	*8.17 %*	*8.68 %*
5	Positive Predictive Rate	*87.07 %*	*87.24 %*
6	Negative Predictive Rate	*93.46 %*	*93.39 %*

We recommend the following insights that we observed during our analysis and also from model, where the X education company should be more aggressive towards to convert the potential leads into the customers.

1. When Lead identifies by the “Lead Add Form” as the Origin.
2. When Lead Source are “Google”, “Direct Traffic”, “Welingak website” and “Cloak Chat”
3. The Potential buyers like to visit the website frequently and spend time on the website.
4. When the last activity performed by customers is “SMS Sent” and “Email Clicked/Open”.
5. When the Status of the Leads is “Will revert after reading the email” and “Closed by horizon”.
6. When the leads are of “Working Professional” in the “Management Specialization” or in the “Banking, investment and Insurance” field as their industry domain.
7. For any Student leads, if “SMS Sent” performed by student as their last notable activity are the potential buyers of online course offered by X educations.