Summary Report

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.

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or 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. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

Goals of the Case Study

1. X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a 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%.

2. Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e., is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

Analysis:

We have done the below analysis to generate the lead score and identify the leads which are most likely to be converted into paying customers. We have used the below steps to prepare our analysis:

1. **Loading and cleaning the data:**

    - Imported the lead.csv file and inspected the data frame.
    - Replaced the columns that have 'Select' with null as Select simply means that no value was selected.
    - We then dropped the columns that have more than 45% missing values.
    - Then we imputed the missing values with mode and some of the missing values were imputed with 'Not Specified' so that we do not lose much of the data.
    - We also treated the outliers in the numerical column by removing the top and bottom 1% values.

2. **EDA:**

    - We divided the columns into categorical and continuous columns.
    - Then we did the Univariate analysis of categorical and continuous columns very little insight was obtained from this.
    - To get a clear insight we did a Bivariate analysis of the data and compared each column with the converted value to see how the factors affect the conversion of a lead.
    - This analysis informed us of the factors relating which have high conversion probabilities.
    - During this process, we also dropped some columns that either have unique values or approx. 99% value belonging to one condition, since, it was not going to help us in the analysis process so we decided to drop this.

3. **Dummy Variables Creation:**

- We first mapped the Yes & No data to 0s and 1s.
- We then created the dummy variable for categorical columns and dropped their first value.
- Later, we dropped the original categorical columns.

4. **Splitting dataset into Train and Test Data:**

- We have split the dataset into train and test at 70% and 30% respectively.

5. **Scaling the Numerical Columns:**

- We used MinMaxScaler to scale the numerical columns.

6. **Feature Selection and Model Building:**

- We used RFE was used to select the top 20 relevant variables.
- Later we built the manual feature elimination by eliminating the variable that has a high p-value i.e., whose p-values are greater than 0.05 and eliminating the variable that has a high VIF value i.e., VIF is greater than 5.

7. **Predicting the Target Variable using the Final Model:**

- We got the predicted values on the train set using the Final Model.

8. **Model Evaluation & ROC Curve:**

- Confusion Matrix was made.
- We calculated the Accuracy, Sensitivity, Specificity, Precision and Recall of the train dataset
- We plotted the ROC Curve and the area under the curve came out to be 0.97 which indicates that it is a good model.

9. **Determination of Optimal Cutoff Point:**

- The optimal cutoff point was determined by plotting the Accuracy, Sensitivity and Specificity for various probabilities and the optimal cutoff came out to be 0.3.

10. **Prediction on the Test Data:**

- The prediction on the test set and the Accuracy, Sensitivity and Specificity is calculated.

11. **Final Observation:**

S. No	Particulars	Train Data	Test Data
1	Accuracy	92.02%	92.55%
2	Sensitivity	91.57%	91.39%
3	Specificity	92.29%	93.26%
4	Precision	87.96%	89.09%
5	Recall	91.57%	91.37%

12. Conclusion:

- Leads Sources (such as Google and Direct Traffic) can bring a lot of business to the company as the Lead conversion is very high compared to other sources.
- Leads who spend more time on the website and were last connected by SMS are more likely to get converted.
- Unemployed Leads can be focused as they are more likely to select the course as it is going to help them in getting better career opportunities.
- The next group that can be focused on are the Working Professionals as they are likely to choose the course for better job possibilities.
- People who revert to the email are more likely to opt for the courses.
- This model is able to predict 91% of the converted data accurately and this will help the business in increasing the business revenue as the company will be able to increase their conversions by focusing more on candidates that are more likely to get converted.

Thank You.