SUMMARY

PROBLEM STATEMENT:

Industry professionals can purchase online courses from X Education, a company that provides education. Many experts who are interested in the courses visit their website on any given day and search for courses.

On numerous websites and search engines like Google, the firm advertises its courses. When these folks arrive, they could peruse the courses, complete a form for the course, or watch some videos on the internet. These people are categorized as leads when they fill out a form with their phone number or email address. Additionally, the business receives leads from earlier recommendations. Once these leads are obtained, sales team members begin calling, sending emails, etc.

STEPS FOLLOWED:

1. Reading and comprehending data:

Here, in an effort to understand how the data appeared and felt, we noted the following:

- The number of columns and rows
- Types of data for each column
- Checking how the data appears in the first few rows and the distribution of the data
- Verifying any duplicates that may exist.

2. Data Cleaning:

Here, we looked for errors in the dataset and corrected any incorrect column names.

- Looking for null values and impute them using the proper techniques.
- For categorical columns, mode imputation was employed.
- If there is no skewness in the data, we impute numerical columns using the mean
- If there is skewness, we impute numerical columns using the median.

3. Data Visualization and Outliers Treatment:

- To determine which categorical columns, make the most sense, we did univariate analysis on the columns.
- We then deleted any columns whose variance is close to zero.
- We used bivariate analysis to examine how categorical columns varied in relation to the Converted column.
- We used box plots to do univariate analysis on numerical columns to determine whether the data contained any outliers.

- To determine how the leads are related to these columns, we used bivariate analysis on numerical columns with the Converted column.
- To handle the outliers in the data set, we employed the IQR approach.
- In order to determine the columns that are correlated, we also plotted the correlation matrix in this phase.

4. Feature Scaling:

At this point, there were no outliers and our data was quite clean. We are aware that logistic regression eliminates input parameters are given as numbers. As a result, we changed all of the category columns to numbers.

- Binary mapping was used to convert the two-level "Yes" and "No" columns to numbers.
- Columns with more than two levels were transformed utilising to create dummy columns using **pd.get dummies command.**

5. Model Building:

We have created a model and removed attributes using the recursive feature elimination technique on the qualities that are still there. The model accuracy is used by RFE to determine which characteristics (and combination of factors) most significantly influence forecasting the desired quality. In this phase, we used statistics to stabilize the model, where we verified that the p-values were less in the library 0.05 and vif values to be less than 5. Variance. Multicollinearity is treated using the inflation factor (vif). After building the stable model, we predicted probabilities on the train set and created a new column predicted with 1 if probability is greater than .5 else 0. We calculated the confusion matrix on this predicted column to the actual converted column.

6. Model Evaluation

7. Conclusion:

After trying out saveral models, our final model has following characteristics:

All p-values are very close to zero. VIFs for all features are very low. There is hardly any multicollinearity present. The overall testing accuracy of 90.78% at a probability threshold of 0.05 is also very good.

Dataset Accuracy Sensitivity Specificity False Positive Rate Positive Predictive Value Negative Predictive Value AUC Train 0.9111 0.8573 0.9449 0.0550 0.9070 0.9135 0.9488 Test 0.9078 0.8412 0.9457 0.0542 0.8984 0.9126 0.9388

The optimal threshold for the model is 0.20 which is calculated based on tradeoff between sensitivity,specificity and accuracy. According to business needs, this threshold can be changed to increase or decrease a specific metric.

High sensitivity ensures that most of the leads who are likely to convert are correctly predicted, while high specificity ensures that most of the leads who are not likely to convert are correctly predicted.

Twelve features were selected as the most significant in predicting the conversion:

Features having positive impact on conversion probability in decreasing order of impact: Features with Positive Coefficient Values Tags_Lost to EINS Tags_Closed by Horizzon Tags_Will revert after reading the email Tags_Busy Lead Source_Welingak Website Last Notable Activity_SMS Sent Lead Origin_Lead Add Form

Features having negative impact on conversion probability in decreasing order of impact: Features with Negative Coefficient Values Lead Quality_Worst Lead Quality_Not Sure Tags_switched off Tags_Ringing Do Not Email