LEAD SCORING CASE STUDY SUMMARY

SUBMITTED BY: GURPREET KAUR ||| SHIVAM SHARMA ||| (DSC43 BATCH)

Problem Statement

1. **X Education** sells online courses to industry professionals. X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers. **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. **Achieve lead conversion rate to be around more than 80%.**
3. There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well.
4. 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**.
5. Finding the Top three variables in your Model which **contribute** most towards the **Probability of a Lead Getting Converted**

PROCESS FOLLOWED

- **Step 1: Reading and Understanding the Data**

- **Step 2: Data Cleaning, Manipulation (Missing Values Treatment) & Outlier Analysis with Visualization**

    1. Check for Duplicates in Dataset and Specific ID Columns
    2. Handling the 'Select' Level that is present in many of the Categorical Variables
    3. Dropping Columns with more than 45% Missing Values
    4. Categorical Attributes Analysis & Treatment (Imputation)
    5. Imbalanced Attributes Analysis & Treatment
    6. Numerical Attributes Analysis & Treatment
    7. Outlier Analysis & Treatment
    8. Checking Skewness
    9. Dropping Columns
    10. Data Imbalance Analysis
    11. Percentage of Rows Retained

- **Step 3: Exploratory Data Analysis (EDA)**
    1. Visualising the Data on Categorical & Numerical
    2. Univariate & Bivariate Analysis
    3. Analysis Using Correlation Matrix & Heat-Map
- **Step 4: Data Preparation for Modelling**
    1. Dummy Variable Creation
    2. Splitting the Data into Training and Testing Sets
    3. Feature Scaling
- **Step 5: Building a Logistic Regression Model By Mixed Approach**
    1. Feature Selection By RFE
    2. Model Building
    3. Assessing the model with StatsModels (P-Value, VIF)
- **Step 6: Creating Prediction & Statistical Analysis on the Train dataset**
- **Step 7: Making Predictions on Test Dataset Using the Final Model**
- **Step 8: Model Evaluation & Performance**
    1. Plotting the ROC Curve ('Receiver Operating Characteristic' Curve)
    2. Finding Optimal Cut-off Point
    3. Precision Recall Trade-Off
    4. Making predictions on the Test Set
    5. Statistical Analysis Of Final Model (Confusion Matrix, Accuracy, Sensitivity, Specificity, Recall, Precision, FPR, PPV, NPV, F1 Score, Classification Report)
    6. Calculating Cross Validation Score
- **Step 9: Lead Score Calculation**
- **Step 10: Hot Leads Determination**
- **Step 11: Feature Importance Determination**
- **Step 12: Business Insights (Top 3 Parameters)**
- **Step 13: Final Report with Results & Recommendations**

CONCLUSION & RECOMMENDATIONS

Our Final model has following characteristics:

- All variables have **p-value < 0.05**, showing **Significant Features Contributing Towards Lead Conversion**.
- All the features have very **low VIF values**, meaning, there is **hardly any Multicollinearity** among the features. This is also evident from the heat map.
- The **ROC curve** has a value of **0.95**, which proves it is very **Good Predictive Model!**
- The overall **Accuracy of Around 88% at a probability threshold of 0.34** on the test dataset is also very acceptable.

- **For Train Dataset**
  - **Accuracy: 88.52%**
  - **Sensitivity/Recall: 88.17%**
  - **Specificity**: **88.74%**
  - **False postive rate** - predicting the lead conversion when the lead does not convert: **0.11**
  - **Precision/Positive predictive value**: **82.85%**
  - **Negative predictive value**: **92.40%**
  - **ROC**: **0.95**
  - **F1 Score**: **0.85**

- **For Test Dataset**
  - **Accuracy : 88.15%**
  - **Sensitivity/Recall : 86.31%**
  - **Specificity: 89.33%**
  - **False postive rate -** predicting the lead conversion when the lead does not convert**: 0.10**
  - **Precision/Positive predictive value: 83.87%**
  - **Negative predictive value: 91.03%**
  - **ROC : 0.95**

- The optimal threshold for the model is 0.34 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.
- Eleven features were selected as the most significant in predicting the conversion.

- **Features with Positive Coefficient Values**
  - Tags_Lost to EINS
  - Tags_Closed by Horizzon
  - Tags_Will revert after reading the email
  - Tags_Busy
  - Lead Origin_Lead Add Form
  - What is your current occupation_Working Professional
  - Last Notable Activity_SMS Sent
  - Total Time Spent on Website

- **Features with Negative Coefficient Values**
  - Last Activity_Email Bounced
  - Last Notable Activity_Modified
  - Last Notable Activity_Olark Chat Conversation

The Model seems to predict the Lead Conversion Rate - Probability More than 80% & we should be able to give the CEO confidence in making good Turn Over Business based on this Model- Potential Leads!