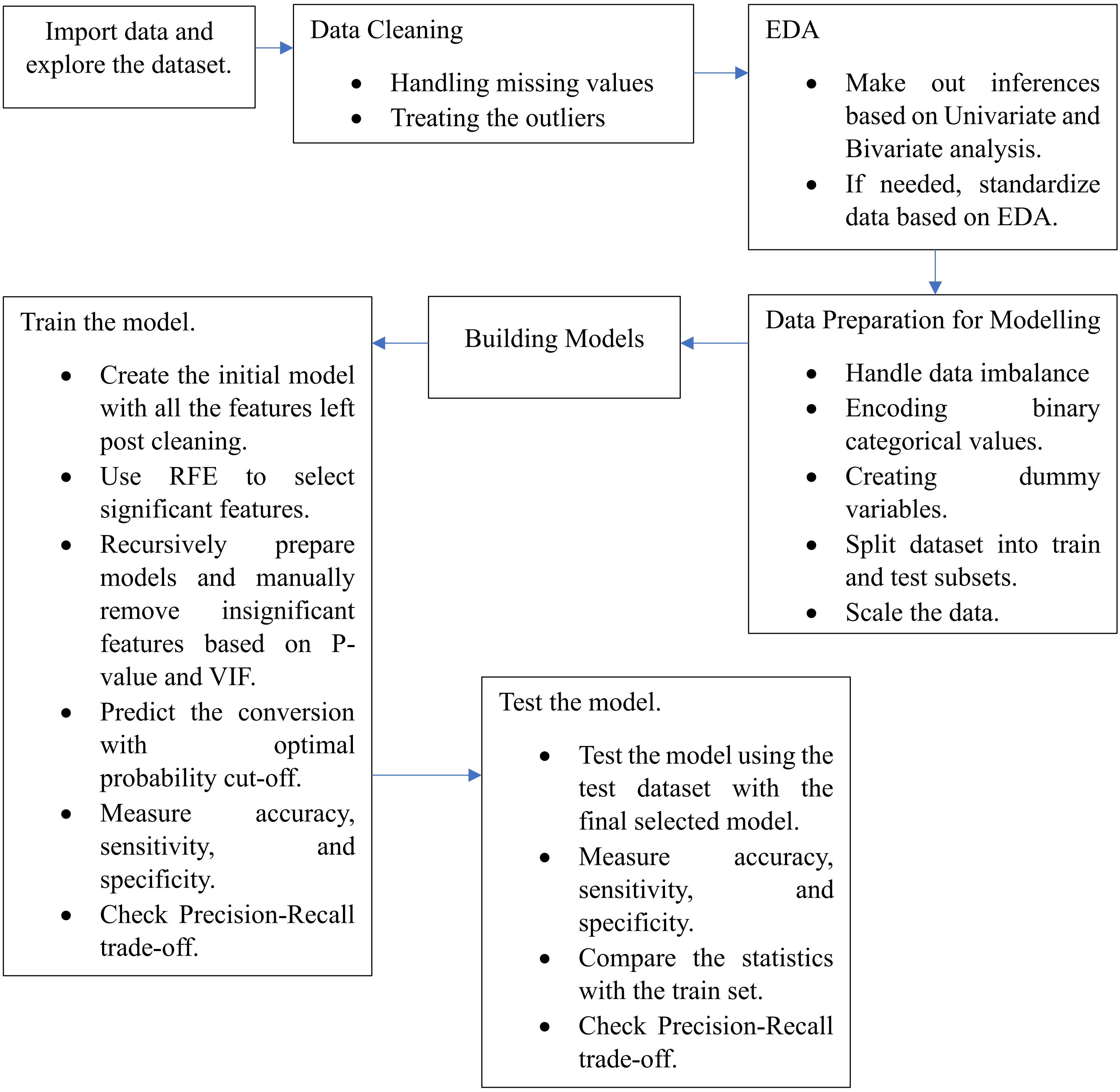
Summary

Prepared By – Vishesh Divya, Vivek Chauhan and Aryan Jain

Problem Statement and Business Objective –

X Education is an education company that sells online courses to industry professionals. Many interested individuals land on their website and check out the courses. The company has a process of form filling and then the individual is a lead. Typically, the current lead conversion rate is around 30%. X Education wants to select the most promising leads that can be potential customers, known as "Hot Leads". The aim is to build a logistic regression model to increase the lead conversion rate.

Methodology



Import data and explore the dataset.

Data Cleaning

- Handling missing values
- Treating the outliers

EDA

- Make out inferences based on Univariate and Bivariate analysis.
- If needed, standardize data based on EDA.

Data Preparation for Modelling

- Handle data imbalance
- Encoding binary categorical values.
- Creating dummy variables.
- Split dataset into train and test subsets.
- Scale the data.

Building Models

Train the model.

- Create the initial model with all the features left post cleaning.
- Use RFE to select significant features.
- Recursively prepare models and manually remove insignificant features based on P-value and VIF.
- Predict the conversion with optimal probability cut-off.
- Measure accuracy, sensitivity, and specificity.
- Check Precision-Recall trade-off.

Test the model.

- Test the model using the test dataset with the final selected model.
- Measure accuracy, sensitivity, and specificity.
- Compare the statistics with the train set.
- Check Precision-Recall trade-off.

Step 1 – Importing and exploring the dataset

This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. which may or may not be useful in ultimately deciding whether a lead will be converted or not. The target variable, in this case, is the column 'Converted' which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn't converted.

Step 2 – Data Cleaning

1. We checked and dropped variables that had high percentage of null values. Then, we imputed the missing values for certain variables.
2. Created new categories for similar values or values having low frequency to improve the readability of categorical variables.
3. For Numerical variables, we checked and treated the outliers.

Step 3 – EDA

1. Performed univariate analysis for categorical and numerical variables to get valuable insights on their effect on the target variable.

Step 4 – Data Preparation

1. Encoded columns with binary values, No and Yes to 0 and 1 respectively.
2. Created dummy variables for the remaining categorical variables with multiple values and, dropped the original variables post dummy variable creation.
3. Split-up the dataset into train and test subsets using 70-30 split.
4. Used the Standard scaler to scale the numerical features.

Step 5 – Model Building

1. Used RFE to select the top 20 features.
2. Used the statistics to recursively run the model and manually eliminated features based on P-values and VIFs and dropped the insignificant ones.
3. Finally, we arrived at the 13 most significant features with P-values<0.05 and VIF<5.

Step 6 – Prediction on Train Data

1. Initially, we kept the probability threshold as 0.5 and ran our model.
2. Based on the assumption, we calculated confusion matrix, accuracy, sensitivity, specificity, and other metrics to check the reliability of our model.
3. Plotted the ROC curve and the area under the curve came out to be 0.88 which is pretty good.

4. Plotted accuracy-sensitivity-specificity graph to find an optimal cut-off point which came out to be ~0.35 and calculated confusion matrix, accuracy, sensitivity, specificity.
5. Checked the Precision-Recall trade-off and the balance point came out to be ~0.4, based on this, our metrics came out to be - accuracy 80.55%, sensitivity 80.62% and specificity 80.50%.

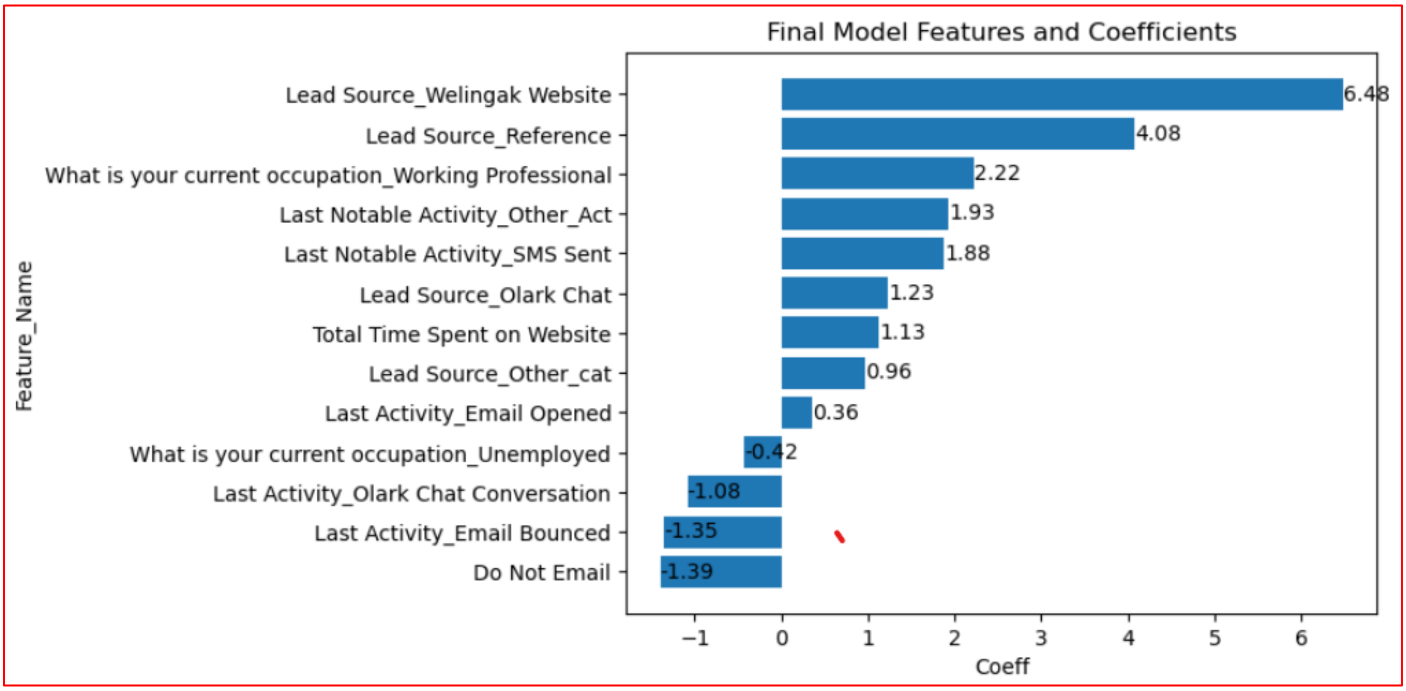
Step 7 – Prediction on Test Data

1. Implemented the learning on the test data and calculated all the metrics. Following are the overall stats -
   - Accuracy – 80.63%
   - Specificity – 81.26%
   - Sensitivity – 79.58%
   - Precision – 72.08%
   - Recall – 79.58%

Conclusion

We have some recommendations for the Marketing and Sales team –

1. Leads from Weligak website have a higher chance of getting converted.
2. Team should keep an active track of references provided by the existing customers.
3. They should design campaigns or some strategies specifically for Working Professionals.
4. People spending more time on the website are more likely to get on-boarded.


Final Model Features and Coefficients
Feature_Name
Lead Source_Welingak Website
Lead Source_Reference
What is your current occupation_Working Professional
Last Notable Activity_Other_Act
Last Notable Activity_SMS Sent
Lead Source_Olark Chat
Total Time Spent on Website
Lead Source_Other_cat
Last Activity_Email Opened
What is your current occupation_Unemployed
Last Activity_Olark Chat Conversation
Last Activity_Email Bounced
Do Not Email
Coeff