Summary Report: Building a Lead Scoring Model

In this report, we've outline our approach to building a lead scoring model aimed at helping the company target potential leads effectively. The primary goal of this model is to assign lead scores to leads, with higher scores indicating a higher likelihood of conversion, and lower scores suggesting a lower chance of conversion.

Data Cleaning & Exploratory Data Analysis:

Our first step was to read and preprocess the leads data for model development. This data encompassed various features such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. We conducted data cleaning to address missing values and outliers. This process was essential to ensure the model's accuracy and generalizability. Missing Value Treatment -

- For Categorical Variables - A separate missing category has been created to study the impact of missing values on Lead Conversion
- For Numerical Variables - Variables are transformed into Categorical Type, with a seperate bucket for Null Values

Moreover, we created scatter plots for numeric and bar charts for categorical variables to identify trends in the data and use that in our model building process.

Data Splitting and Pre-processing:

For identifying probable clients, we used a logistic regression model. To assess the model's performance, we divided the data into a training set and a testing set. The training set was used to train the logistic regression model, while the testing set was used for evaluation. After all of the encodings and scalings done on the train set, we were left with 205 variables.

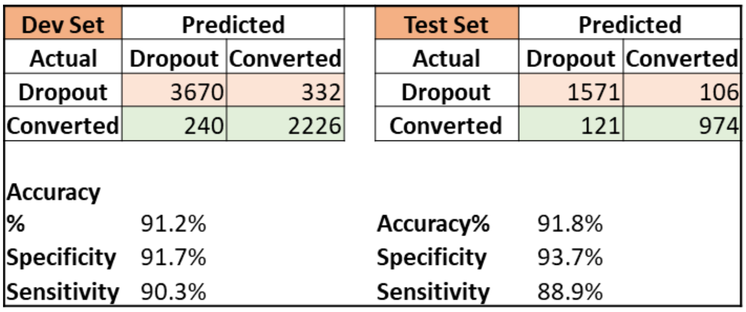
Model Building Process:

We found the 15 top variables in the dataset using RFE, which we then employed in the logistic regression model-building process. We examined the p-values of all the variables after developing our initial model and eliminated those that were not significant. We used the VIF technique to eliminate variables that were linked with other model variables. Following elimination, we retrained the model with the remaining 12 final variables.

Threshold Tuning & Model Evaluation:

One of the critical steps in the lead scoring process is determining an appropriate threshold for lead scores. We used techniques like ROC curve analysis & interaction of accuracy, sensitivity, specificity to choose an optimum threshold. This allowed us to classify leads into "hot" and "cold" categories. Basis this approach, we finalised 0.3 as the ideal cut-off point for our model.

We employed various metrics, including confusion matrix, accuracy%, sensitivity%, specificity%, and ROC-Area under curve, to gauge the model's effectiveness. The results of the model on train & test sets are shared below -



Dev Set	Predicted	
Actual	Dropout	Converted
Dropout	3670	332
Converted	240	2226

Test Set	Predicted	
Actual	Dropout	Converted
Dropout	1571	106
Converted	121	974

Accuracy %	91.2%
Specificity	91.7%
Sensitivity	90.3%

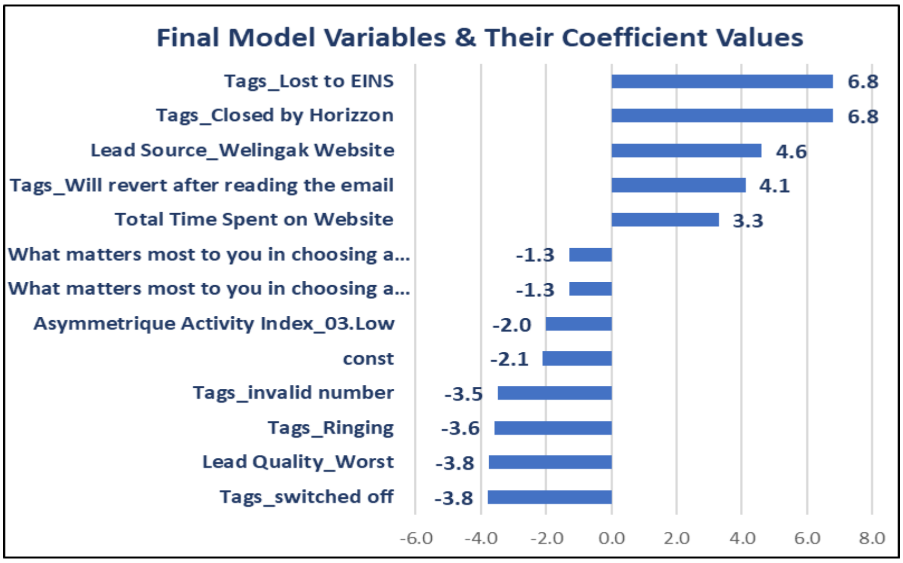
Accuracy%	91.8%
Specificity	93.7%
Sensitivity	88.9%

Scoring Leads and Business Interpretation:

The trained logistic regression model can be applied to the lead data to generate lead scores. The probability values can be converted in 100 digit score, each ranging from 0 to 100. These scores will provide a clear indication of the conversion potential for each lead. Leads with higher scores will be classified as "hot" and can become top priorities for the sales team, while leads with lower scores will be considered "cold".

Key Findings:

- X Education can transform the model's likelihood to lead scores and should target possible leads with scores more than 30 or probabilities greater than 0.3.
- A few variables improve the likelihood of lead conversion based on the Model Outcome. As a result, the business should prioritise consumers with favourable variables, i.e. variables with positive coefficient values in the model.



In conclusion, our approach to building a lead scoring model involved a systematic process of data cleaning, EDA, pre-processing, modelling, threshold tuning, & model evaluation. In addition, by assigning lead scores, we enable the company to focus its efforts on leads with the highest conversion potential, which can significantly impact business outcomes.