Summary Report - Lead Scoring Case Study

Group – Krishnan Vybhava Srinivasan and Nithya V

Background - X Education requires you to build a model to assign a lead score to each of the leads in a manner where the target lead conversion rate would be around 80%

Process:

Data Cleaning

- The data had 37 Columns and 9240 rows.
- **Missing Values -** Dropped columns with significant missing values.
- **Imputing Values/ Outlier Treatment/ EDA**
  - We performed multivariate and univariate analysis of on columns:
    - **Imputation of Values for Missing items**
    - **Outlier Treatment** - outliers where adjustment for values beyond the range of 2% - 95%
    - **Visualization -** we dropped **14 columns** and categorical columns which had infrequent values were changed to generic values such as Others, etc.,
- **Data Preparation**
  - Categorical columns was converted to Dummy variables

- **Model Building Preparation**
  - Removed the Response Variable 'Converted' and copied into a different data-frame
  - **Split 70% of data for training and balance for test**
  - With Standard Scaler technique standardized the continuous columns

- **Model Building**
  - Per Summary statistics a lot of variables were insignificant and need feature elimination. We decided to use RFE method and then move to manual for feature elimination.
  - We used the Generalized Linear Models method we try to fit a logit curve to a binomial data. We started with output as 20 features
    - The logistic regression curve gives you the probabilities of conversion
    - From 20 Features identified we dropped 5 features based on p values and VIFs
    - **Remaining 15 features** statistics all p values are below 0.05 and the VIF values are below 3.5.

- **Model Evaluation**
  - Assumed items with probability greater than 0.5. Derived confusion metrics and results are:
    - Accuracy – 92%
    - Sensitivity – 85%
    - Specificity – 96%
    - Area under the Roc curve -96%

Summary Report - Lead Scoring Case Study

Group – Krishnan Vybhava Srinivasan and Nithya V

    - We need to evaluate optimal cut off point.
    - **Plotted a curve with accuracy, sensitivity, and specificity and based on which concluded the optimal cut-off was around 0.3**
    - **Based on this got decent metrics :**
        - **Accuracy – 92%**
        - **Sensitivity – 86%**
        - **Specificity – 96%**
    - Computed precision and Recall, which as 93% and 86%, which looked good.
    - Precision and recall tradeoff chart also indicated cut off as 0.3 as being ideal
- **Predictions on Test Set**
    - **We perform the predictions on the test set and metrics based on that are:**
        - **Accuracy Score – 91%**
        - **Sensitivity – 84%**
        - **Specificity – 96%**
- **Given the metrics of train and test data is similar it can be concluded the model is robust**

Conclusion

Any lead with probability score greater than 30% is a **Hot Lead** and others are **cold lead**.

- Key Variables which Reduces Conversion Probability:
    - Do Not Email
    - Last Activity – Olark Chat Conversation
    - Lead Quality – Not Sure and Worst
    - Tags –Ringing, Switched off,
    - Last Noteable Activity – Email Link Clicked and Modified

- Key Variables Increase Conversion Probability
    - Occupation – Working Professional
    - Tags – Lost to EINS, Closed by Horizzon, Busy, Will revert after reading email
    - Last Noteable Activity – SMS Sent, Wellngak Website