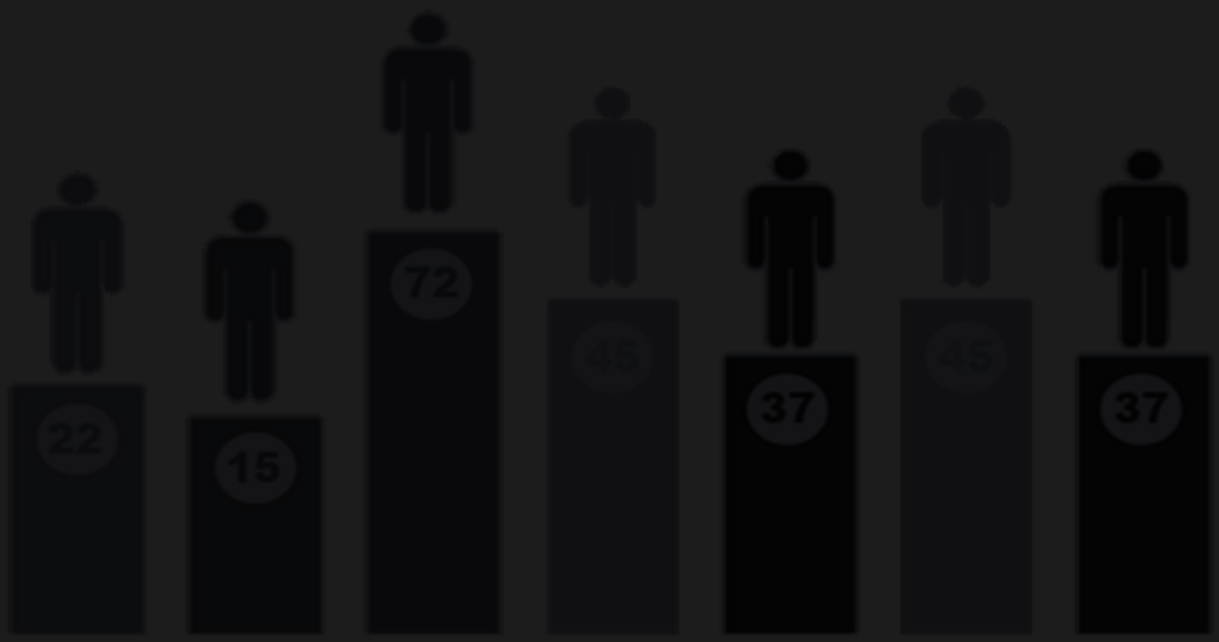
Lead Score Case Study

Summary Report

Abhishek Bhattacharya
Jithin Prakash K



LEAD SCORE CASE STUDY

SUMMARY REPORT

The Lead Score Case study is carried out by creating logistic model to predict if a given lead would convert or not. Below steps are taken to build the model:

1. Reading and Understanding the data.

   The data from the csv file is read to a dataframe using pandas library in Python. The head, shape, info, and description of dataset is analyzed to understand the strength, quality and data content along with the missing and unique values.

2. Data Preprocessing and Visualization

   - Firstly, each column with unique values are analyzed and features with low variance, (i.e. a particular value appearing most frequently compared to others) are dropped after visualizing the data.
   - For categorical features, count plot is used to check frequency. Features with category, "Select" are considered as unavailable or missing and they are changed to nan values. The rarely occurring categories (i.e. frequency less than 1%), are combined as "others". *'Country'* and *'What matters most to you in choosing a course'* columns showed low variance, and are dropped. Further, the features with high missing values are dropped.
   - Imputation is carried out for features with lower missing values. Mode and median are used to impute categorical features and numeric features respectively. Box plot is used to analyze outliers. Outlier data is removed from dataset. Pair plot and heatmap are used to find the correlation and the distribution of the numerical data.

3. Data Preparation

   The categorical features are one hot encoded by dropping "others" or the last category. The correlation is checked for the dummy features and the highly correlated feature-categories are dropped.

4. Model Building steps

   - Test Train Split: The resultant dataframe is split into test and train data using 70%-30% split.
   - Feature scaling: Standard-scaler is used to fit and transform the train dataset.
   - Initial Model: Initial model is built using stats-models GLM with all the features in the dataset, which returned most features with higher p-value.
   - RFE: is used to eliminate such features using sklearn logistic-Regression, which reduced the number of features to 15.
   - Model refining: VIF and p-values are checked to remove the highly correlated, collinear or insignificant variables to return final model with low VIF and low p-values.

5. Evaluation
    - The model is used to obtain the predicted probability values, which is used to obtain converted leads with base cut-off of 0.5. The confusion matrix, accuracy score, specificity, sensitivity, precision, recall, f-score (accuracy measures) etc. are calculated. Precision-recall and ROC curves are plotted to understand the area under the curve (AUC =0.88) depicting a good predictive model.
    - Optimal cutoff value (OCV) is obtained by plotting the data for accuracy, specificity and sensitivity together to return the optimal point as 0.33. The OCV is used to recalculate all the accuracy measures. Accuracy obtained is 80%

6. Predicting Test Data

    The test data is scaled, using standard-scaler. The prediction is made using the final model and the conversion probability is obtained. The OCV (0.33) is used to obtain the model predicted converted leads. All the accuracy measures are calculated for test data. Accuracy is obtained is 79%.

7. Conclusion

The obtained formula:

$$logit(p) = log\left(\frac{p}{1-p}\right)$$
$$= (4.49\ x\ LeadSource\_Welingak\ Website) + (2.99\ x\ LeadSource\_Reference)$$
$$+ (2.71\ x\ CurrentOccupation\_Working\ Professional) - (1.33\ x\ LeadSource\_Direct\ Traffic)$$
$$- (1.32\ x\ LastActivity\_Email\ Bounced) + (1.24\ x\ LastActivity\_SMS\ Sent)$$
$$- (1.16\ x\ LeadSource\_Organic\ Search) + (1.09\ x\ TotalTimeSpent)$$
$$- (1.03\ x\ LastActivity\_Olark\ Chat\ Conversation) - (1.01\ x\ LeadSource\_Referral\ Sites)$$
$$- (0.88\ x\ LastNotableActivity\_Modified) - (0.88\ x\ LeadSource\_Google) - (0.06\ x\ Intercept)$$

The features those strongly impact lead conversion are:

- Lead Source - Welingak Website, Reference, Direct Traffic (Negative), Organic Search (Negative), Referral Sites (Negative), Google (Negative)
- Current Occupation - Working Professional
- Last Activity - Email Bounced (Negative), SMS Sent, Olark Chat Conversation (Negative)
- Total Time Spent
- Last Notable Activity - Modified (Negative)