Lead Score Case Study Summary

Problem Summary:

- X Education is an Education company as a company would like to increase its lead conversion rate higher than its current level which is 30% by filtering out its leads to find those with high potential.

Business Goal:

To find factors which would suffice the company's objective of identifying potential leads and their sources to increase their lead conversion, thereby adding to their overall revenue.

Requirement guideline:

A logistic regression model with more than 80% conversion rate to assign was required to be built that would assign a lead score between 0 and 100 to each of the leads which would be used by the company to target potential leads.

Approach:

As it was a classification problem, logistic regression was used.

Below are the steps that were taken:

1. Data Understanding:

Here we tried to get the look and feel of the data, we observed the following things:

- Data types and Number of rows and columns
- Checking the first few rows how the data looks
- Checking for duplicates, if any.

2. Data Cleaning:

The dataset was checked for discrepancies :

- Columns with select as a variable entry were converted to NaN and those with above 40% null values were dropped.
- Categorical data from columns containing entries with minimal counts were grouped into others.

3. Data Visualization and Outliers Treatment:

- Univariate analysis and bi-variate analysis were performed on both categorical and numerical columns.
- A heat map was plotted to identify the correlation between numerical columns.

4. Data Preparation for model building and feature scaling :

- Created dummy variables for categorical columns.
- Numerical variables were scaled using Standard Scalar.

5. Feature selection:

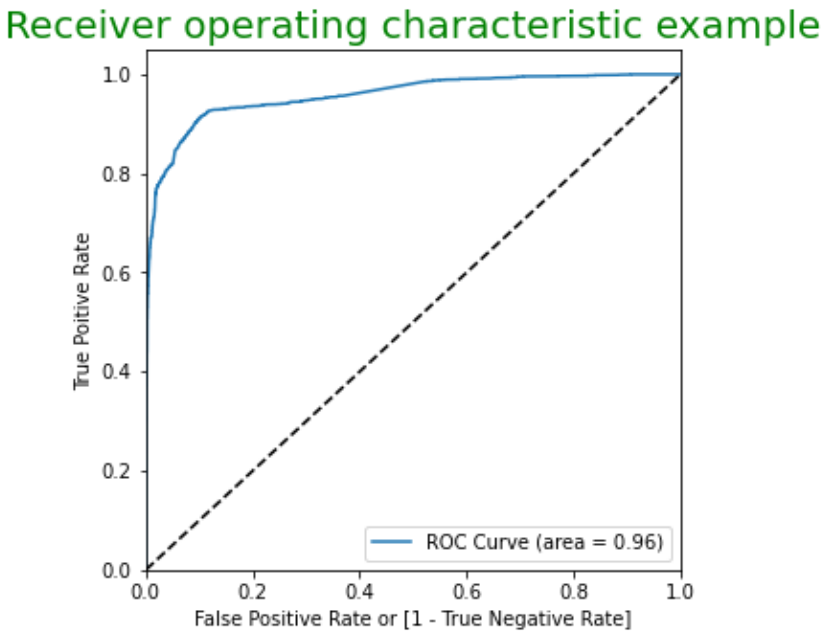
- Recursive Feature Elimination technique(RFE) was used for feature selection to select the top 20 variables that would affect the model with the remaining features dropped.

- A correlation matrix of the final features was then plotted to check how they affect the target variable.

6. Model building:

A model was then built using the top 20 features and those that had high p-values or Variance Inflation Factors(VIF) were then iteratively removed until a stable model was found.

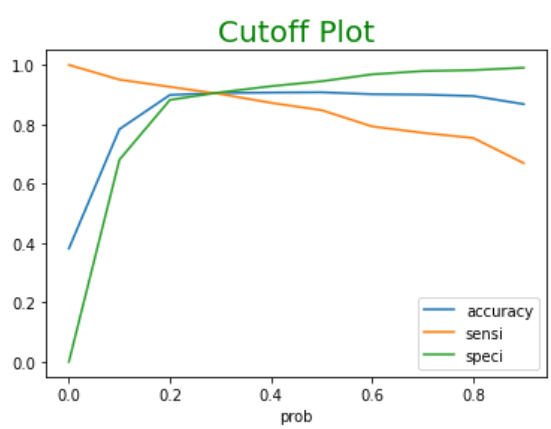
Then the train set was used to predict probabilities and create a new column predicted with 1 if the probability is greater than .5 else 0. The confusion matrix was then calculated on this predicted column to the actual converted column. The roc curve was plotted to find the area under the curve to understand the performance of the model.



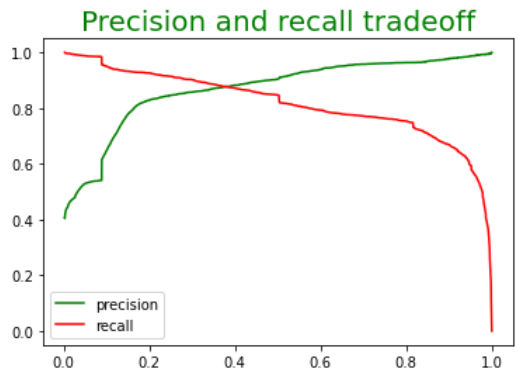
The area of the ROC curve indicates the performance of the model. The closer to 1 the area under the curve, the better the model.

7. Model Evaluation on Train Set:

With probabilities from 0.0 to 0.9, the 3 metrics -accuracy, sensitivity and specificity were calculated, To make predictions on the train dataset, an optimum cutoff of 0.3 was found from the intersection of sensitivity, specificity and accuracy as shown in the below figure:



To make predictions on the test dataset, the optimum cutoff was considered as obtained from the Precision recall graph of the train dataset as shown below figure:



We can observe that 0.37 is the tradeoff between Precision and Recall. However,a threshold was chosen at 0.30 so as to favour recall and to consider any Prospect Lead with Conversion Probability higher than 30 % to be a hot Lead.

8. Predictions on Test Set:

After finalizing the optimum cutoff and calculating the metrics on the train set, the data was predicted on the test set. Below are the observations:

Train Data:

Accuracy: 90.55%
Sensitivity: 90.19%
Specificity: 90.78%

Precision: 90.48%
Recall: 84.75%

Test Data:

Accuracy: 90.8%
Sensitivity: 92%
Specificity: 90%

Precision: 86.06%
Recall: 91.51%

9. Conclusion

- As the final model is derived, it can be observed that the probability of conversion will highly be influenced by the leads closed by Horizzon.
- Welingak website seems to be a source for hot leads as it has a strong positive impact on the model.
- Students or potential customers who chose to revert after reading the email were more likely to be converted as leads.
- Leads whose phones were generally switched off or ringing were less likely to be converted as leads.
- It was also anticipated that potential leads who are already students or interested in other courses had a lesser chance of becoming strong leads.
- In general, the Tags column seems to be an important column as most of the variables from the model are converted to dummy variables with high coefficient magnitudes.
- Using this model, the sales team of X education can easily filter models using the lead score for identifying and converting high-potential leads.