X Education Lead Scoring

Logistic Regression Assignment

Identifying hot leads to increase lead conversion rates to 80% daily

Created by
Neeta Talekar
Sheetal Atre

1. Overview

- *Project statement and Rationale*
- X Education sells online courses to industry professionals
- Lead scoring is a subtask of their customer relationship management (CRM) to prioritize their sales.
- To increase the lead conversion rate from 30% to 80%

- *Project Objectives*
- To build and evaluate a lead scoring model to assign lead scores from 1-100 points to rank the leads to determine their sales-readiness
- Model must have a Recall score of atleast 80%
- A higher lead score implies that the customer is more likely to engage with the company.
- High priority leads will be passed on to sales and low priority leads may be engaged in lead nurturing campaigns.
- The sales team can focus on the high scoring potential hot leads for follow-up

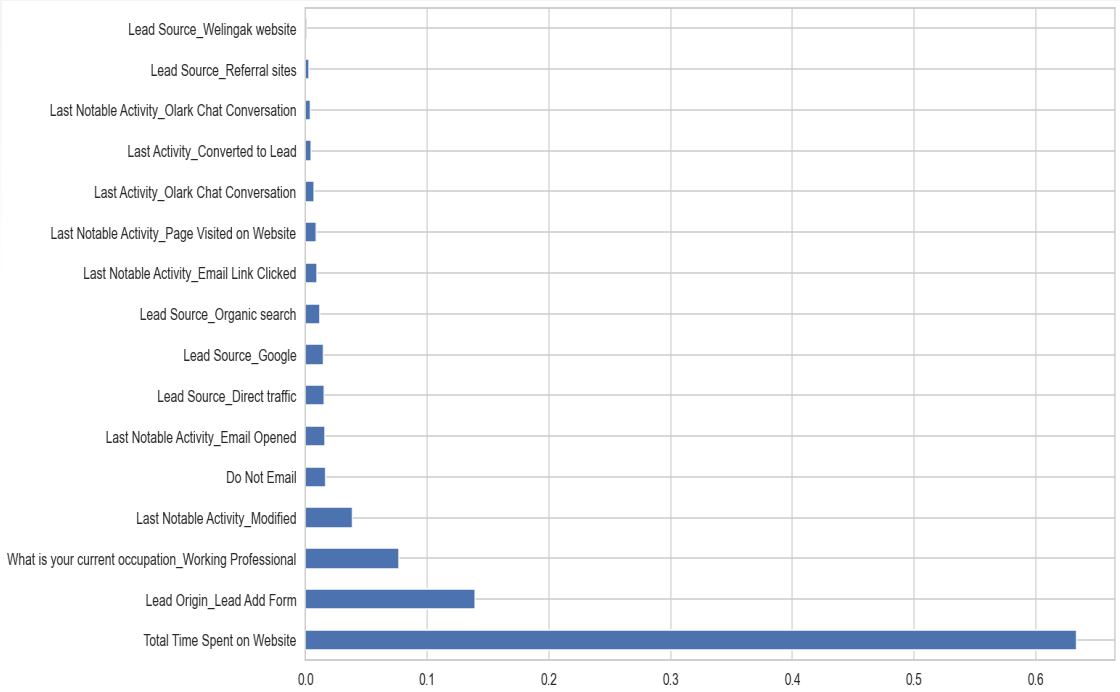
2. Technical Approach

- Step 1: Examine and transform the data to deal with redundant columns, missing values, outliers andcorrelations with the output label
- Step 2 : Identify the positive(Converted) and negative class(Not Converted).
- Step 3: We have 70 variables after cleaning the dataset. Create a variable structure using RFE with Logistic Regression to carefully select the relevant 1/3rd features (= 19 columns) that qualify an ideal lead as it largely influences the quality of the lead scoring output. For RFE implementation first we select all features withvif < less than 5 to remove multicollinearity and then keep only the features with p-value <= 5% to filter out insignificant features.
- Step 4: Build a model using Logistic Regression algorithm which returns a probability value, which we convert to a binary value
- Step 5: In-order to convert the probability to a binary value, we identify a threshold. A value above that threshold is predicted as “Converted"; a value below indicates “Not Converted“ by the model . Since we do not know the ideal threshold value, we evaluate the model for several thresholds from 0.1 to 0.9 and evaluate the model performance using statistical metrics and visualizations
- Step 6 : Calculate and assign the lead scores from the calculated probabilities using the optimal threshold value
- Step 7: Interpret lead score by using statistical methods or visualization tools

3. Insights to focus on Important features impacting Lead Score

Sales team should focus on leads who:

- Spend more time on the website
- Have submitted the form showing interest
- Working professionals and those users who leads generated from Referral sites
- Connect directly/Google to the Lead Source
- Users who have modified any information, opened/sent emails, visited website should be focused on
- Users who use Clark Char conversation, Welingak website should also be focused on



4. Confusion Matrix

- Definitions:
  - "Converted" is a positive class.
  - "Not Converted" is a negative class.
- If model predicts Converted, we send a mail which costs us INR 5
- If the lead actually Converts, we will have a profit of INR 100
- In marketing and lead scoring models it is more important to be able to detect the positive than the negative class
- Recall must be 80%.Means If the model is predicting 100 hot leads, 80% should be converted.

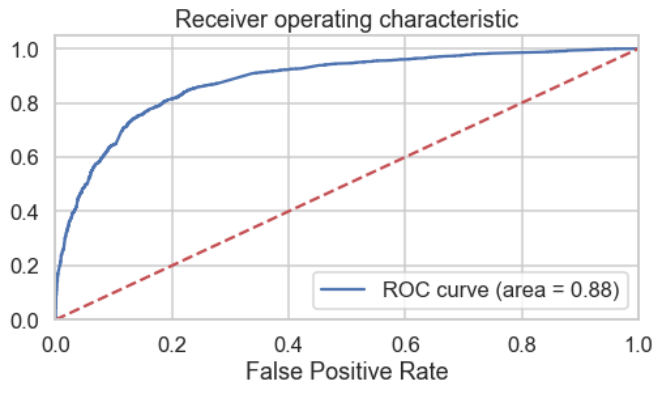
True Positive (TP): model correctly predicts the positive class • Reality: Converted • Model predicted: "Converted and send mail" • Outcome:Cost = 5, Profit=100 • Total cost = 100-5=95	False Positive (FP): incorrectly predicts the positive class • Reality: Not Converted • Model predicted: "Converted and send mail" • Outcome: Cost = 5, Profit=0 • Total cost = 0-5=-5
False Negative (FN): incorrectly predicts the negative class • Reality:Converted • Model predicted: "Not Converted and don't send mail" • Outcome:Cost=0,Profit=100 • Total cost = 100-0=100	True Negative (TN): model correctly predicts the negative class • Reality: Not Converted • Shepherd said: "Not Converted, and don't send mail" • Outcome:Cost=0, Profit=0 • Total cost = 0-0=0

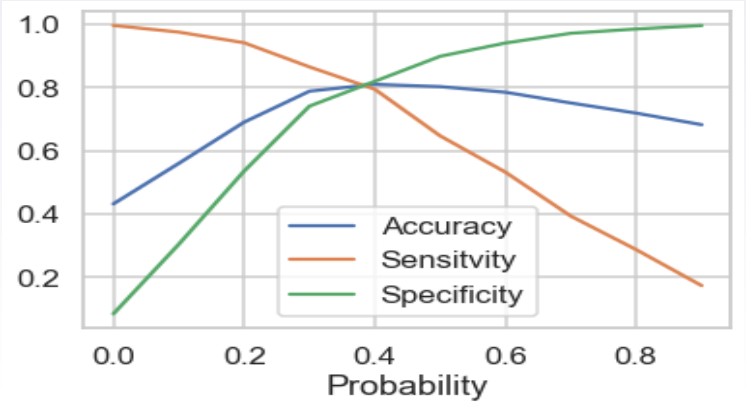
5. Model Evaluation Metrics

- Accuracy
  - Accuracy means fraction of predictions our model got right. For eg Accuracy = 0.81, means 81% of total 100 predictions is right. This is a class-imbalanced data set, as there is a significant disparity between the number of positive and negative labels

- Precision, Recall/Sensitivity/TPR and Specificity/TNR - to account for errors and to identify the optimal threshold probability to turn a probability model into a classification model
  - Precision = What proportion of positive identifications was actually correct? Our model has a precision of 0.73 means, when it predicts a lead is Converted, it is correct 73% of the time. Percentage of leads predicted as Converted that were correctly classified
  - Sensitivity/Recall = What proportion of actual positives was identified correctly? Our model has a recall of 0.79 means, it correctly identifies 79% of total Converted leads. Percentage of leads actually Converted that were correctly classified
  - Specificity = What proportion of unsuccessful conversions that are correctly classified? Our model has specificity of 0.82 means, leads predicted as Non Converted from all the leads, but who have actually Converted

6. Model Evaluation – Threshold for default training model





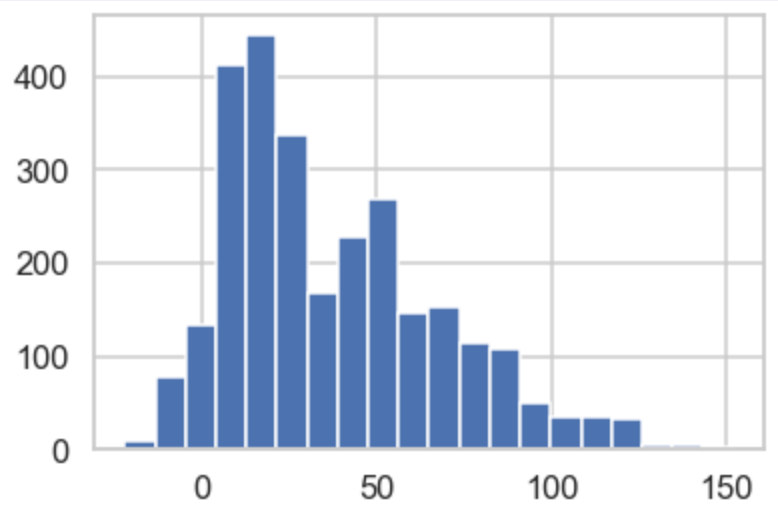
Optimal Threshold Probability cut-off

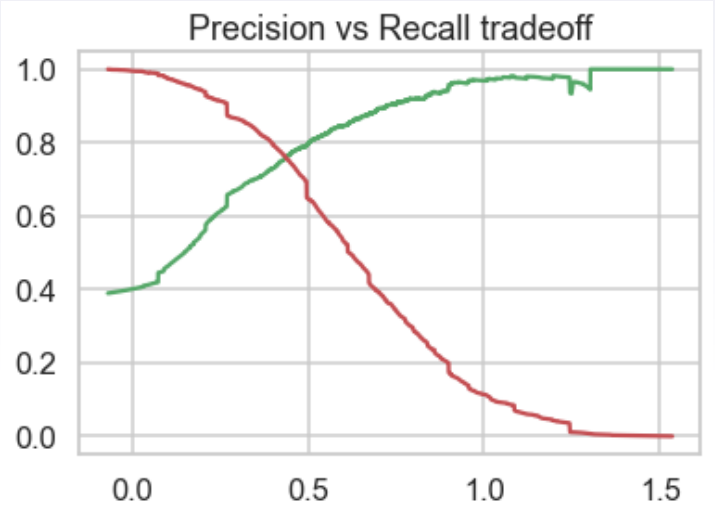
Optimal threshold probability is the convergent point of accuracy, sensitivity and specificity which is 0.4. All probabilities above this value will be considered as "Converted" in our model. All probabilities generated by our model below this value will be marked "Not Converted".

Area Under Curve(AUC) of Receiver Operating Characteristic (ROC) curve - to find optimal cut-off point :

ROC curve plots TPR against FPR across different probability threshold. As threshold probability decreases TPR increases and TNR decreases. This trade-off can be visualised using a ROC curve, allowing us to pick an optimal threshold.
Bigger the AUC(0.88) indicates a better model, where customers with higher predicted lead scores are most likely to correspond to the actual purchase, whereas customers with poor quality have little incentive to buy the company's products or services.

7. Interpreting Lead Score Values





- Higher lead scores indicates most likely the user will get Converted.
- Precision Vs Recall (called F1 score)
  - We have a 0.40 precision-recall trade-off value. Means any Conversion Probability above 40% can be considered as Hot leads. The users having lead score above 40 is a good lead to follow up.

THANKS!