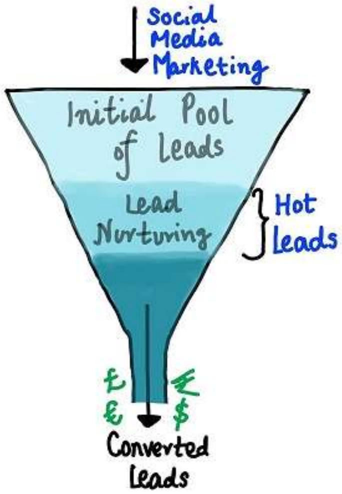
LEAD SCORING CASE STUDY

PROBLEM STATEMENT

An education company named X Education sells online courses to industry professionals. The company markets its courses via several platforms, when people fill up a form providing their email address or phone number, they are classified to be a lead. The current lead conversion rate at X education is around 30%. A typical lead conversion process can be represented using the following funnel:



The company requires a model wherein lead score is assigned to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

ANALYSIS PROCEDURE

- 1. Load and inspect the dataset

- 2. Null values handling

- 3. Analysis of numerical columns

- 4. Analyzing categorical type columns

- 5. Train-test split

- 6. Scaling the features

- 7. Feature selection using RFE

- 8. Model building

- 9. Finding the optimal cut-off point

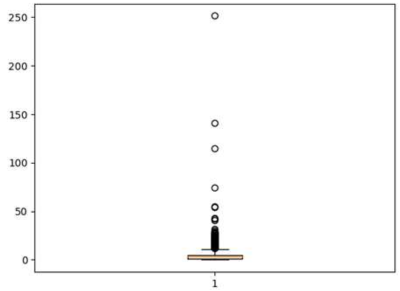
- 10. Training data predictions on final model

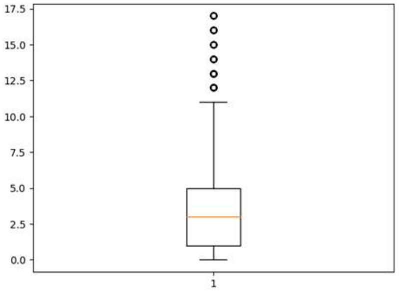
- 11. ROC curve

- 12. Making Predictions on test dataset

ANALYSIS OF NUMERICAL COLUMNS

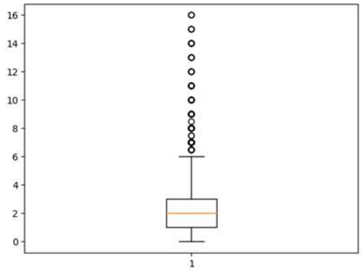
Column- ‘TotalVisits’: There were outliers in the column, which were dealt with by cleaning the column and maintaining data for +- 3 standard deviations which accounts for about 99.7% of data.

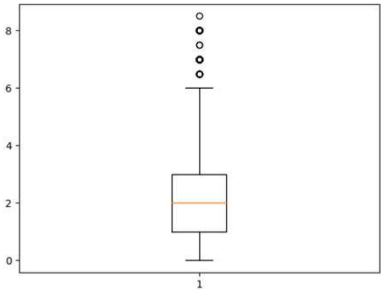




Column- ‘Total Time Spent on Website’: There were no outliers in this column, data was consistent.

Column- ‘Page Views Per Visit’: There were outliers in the column, which were dealt with identically as above.





ANALYSIS OF CATEGORICAL COLUMNS

A lot of categorical columns did not make sense from a model building perspective either due to the data being highly skewed or the purpose of the data in the column not making business sense. However, no column was dropped due to the latter because of lack of business explanation.

Business insights from this section of analysis:

- Column-' **Do Not Call**': Highly skewed data showed that about 99% of leads opted not to be called.
- Columns-'**Search**','**Magazine**','**Newspaper Article**','**X Education Forms**','**Newspaper**','**Digital Advertisement**': Highly skewed data showed that majority of the leads hadn't seen any ad of the company on any means. This is a key thing that must be highlighted with advertisement team.
- Column-' **Through Recommendations**': Majority of leads did not come in via a recommendation.
- Column-' **Receive More Updates About Our Courses**': Highly skewed data, indicating majority of the leads do not want updates about company's courses
- Column-' **Update me on Supply Chain Content**': Highly skewed data, indicating majority of the leads do not want updates about company's supply chain content.
- Column-' **Get updates on DM Content**': Highly skewed data, indicating majority of the leads do not want updates about company's DM content.

MODEL BUILDING

The most important part of the feature variables selection to build the model was done using the Recursive Feature Elimination technique. RFE is a technique that recursively eliminates less important features from a model to improve its performance.

Finalized a list of 25 features and started building the logistic regression model. Once the first model was built, features were eliminated manually to get to the final iteration of the model.

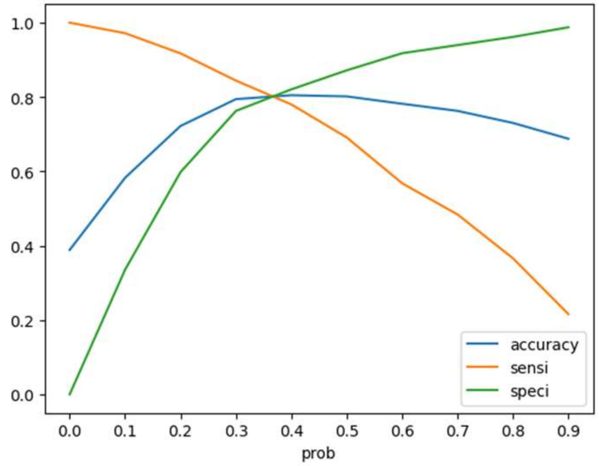
Features were eliminated based on p-value and VIF score in below sequence:

1. High p-value & high VIF
2. Low p-value & high VIF
3. High p-value & low VIF

The model was built to compute the probability of conversion of each lead.

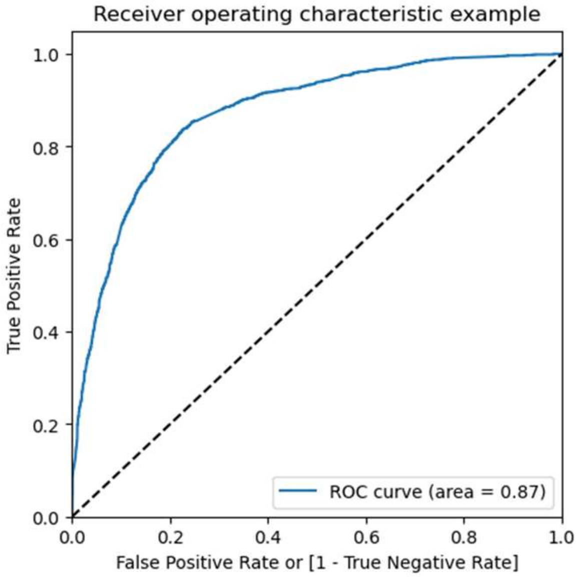
OPTIMAL CUT-OFF & ROC CURVE

As per problem statement the sensitivity of the model is paramount and the target is for it to be around 80%. A plot was made between the accuracy, sensitivity and specificity of the model using which the final cut-off of 0.35 was decided.



Predictions were then made using the optimal cut-off on the training dataset.

Once predictions were made the ROC curve for the model was drawn to verify the model performance. The Receiver Operating Characteristic (ROC) curve is a graphical representation that illustrates the performance of a binary classification model by plotting the true positive rate vs. false positive rate, helping to evaluate the model's discrimination threshold and overall Predictive power.



TESTING THE MODEL

The model performance was tested on the test dataset.

1. The numeric columns of the test dataset were scaled.
2. The feature variables in the final model were filtered in from the set of feature variables of test set.
3. Predictions were made on the test set, i.e. probability of conversion of a lead was calculated. Followed by categorizing the leads based on cut-off probability decided previously.
4. Finally the accuracy, sensitivity and specificity were calculated which came out to be 78.86%, 78.67%, and 78.97% respectively.

Hence, it can be concluded upon final testing of the model that the required target of around 80% model sensitivity has been achieved.