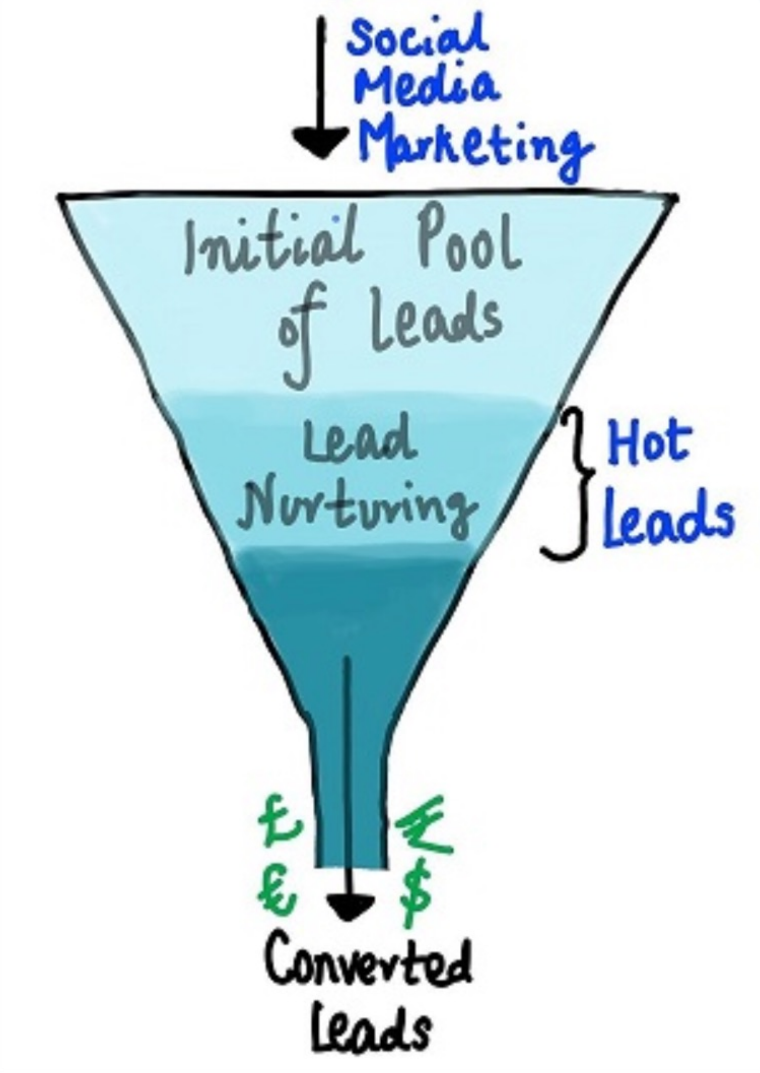
Lead scoring Case study

Tarun Balotia

Problem statement

- X education gets leads from many mediums
- The leads acquired by X education need to be scored to improve the conversion rate of top of the funnel – initial pool of leads
- The model developed should improve the current conversion from 30% to 80%



Lead Conversion Process - Demonstrated as a funnel

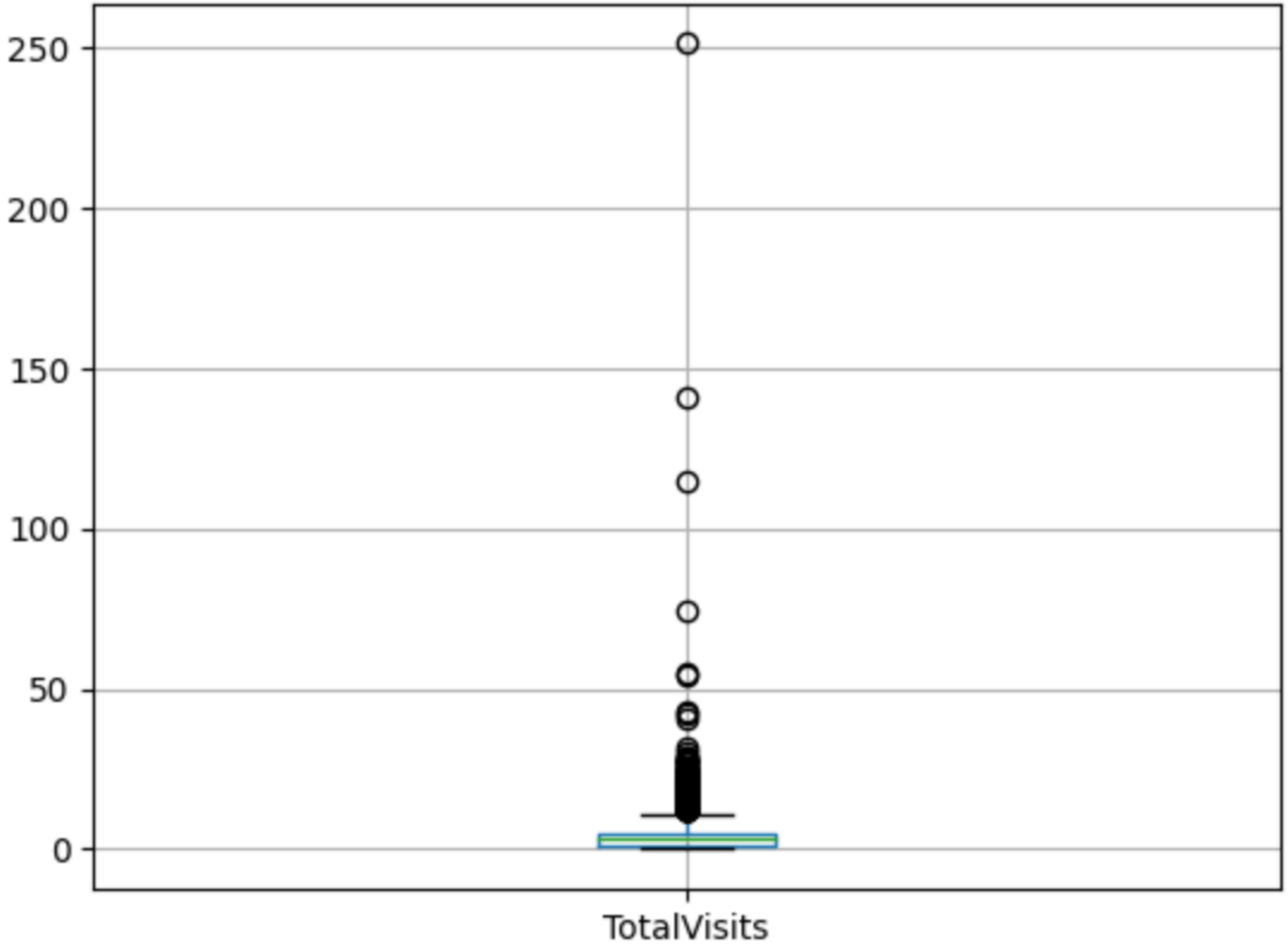
Approach

- We did an EDA on the provided data, cleaned the data to remove nulls and imputed the data
- Used Recursive feature selection for eliminating features
- Build a logistic regression model after removing high VIF and high p-values features
- Find the optimal probability threshold for classifying as converted or not

Modelling steps

- Null value handling
- Numerical and categorical columns analysis
- Test, Train data split
- Feature scaling
- RFE(recursive feature elimination)
- Model building
- Optimal threshold
- Predict on test data and get model metrics

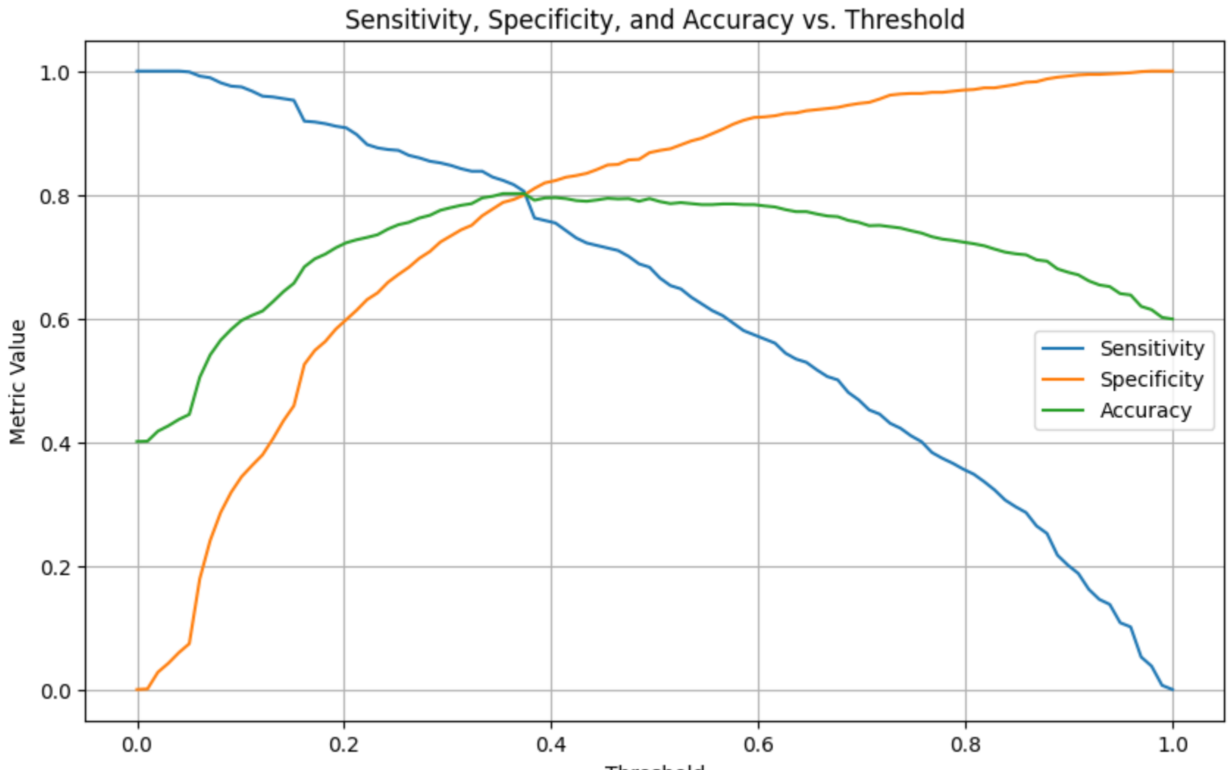
Numerical columns Analysis



Categorical columns Analysis

- Some categorical columns were dropped as follows:
  - Lead quality , lead profile
  - Prospect ID, lead number
  - What matters most to you in choosing a course
  - What is your current occupation
  - Country

Model Metrics

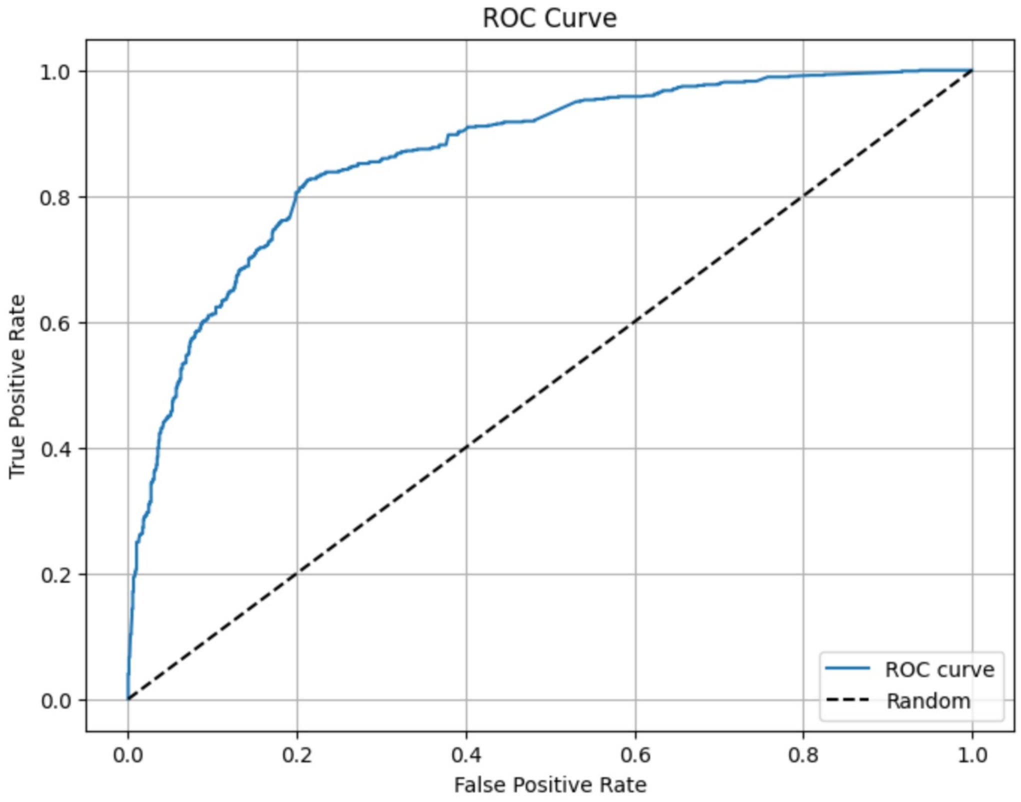


- For finding the optimal threshold that would lead to prediction of converted or not, we plot the sensitivity, specificity and accuracy plot to balance all metrics
- This number comes to approx 0.35

Logistic regression Model Building

- MinMax Scaling done for numerical columns
- Used RFE to arrive at 20 features
  - RFE -It is a feature selection technique used to select the most important features from a given dataset. The main idea behind RFE is to recursively remove less important features and build the model again until the desired number of features is reached or until the model's performance stops improving.
- Remove all features with VIF>3 and p-value>0.05

Model Metrics



- Area under the ROC curve = 0.87

Test data metrics

- Accuracy: 0.80
- Sensitivity: 0.83

Recommendations for lead outreach

- The top variables that point to high converting leads are
  - Total time spent on website
  - Lead origin
  - Lead source
  - Last activity
- The above features must be noted, and leads must be prioritized based on these
- In case of increased workforce, lower threshold leads can be picked up in favour of more customers opting for course
- In times of reducing cost/lower outbound calls, only the high scoring leads(with higher probability threshold) should be called.