



CASE STUDY ON LEAD SCORING

By,

Manushankar PM

Sumit Batra

Vedant Dubey

TABLE OF CONTENT

- Introduction of X education company
- Problem statement
- Objective of case study
- Goals of the case study
- Steps in building the model
- Model evaluation
- Problem solution

INTRODUCTION TO X EDUCATION COMPANY

- An education company named X Education sells online courses to industry professionals.
- Many professionals who are interested in the courses land on their website and browse for courses.
- The company markets its courses on several websites and search engines like Google.
- Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos.
- When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals.
- Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not.

PROBLEM STATEMENT

- The typical lead conversion rate at X education is around 30%.
- Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'.
- the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

OBJECTIVE OF THE CASE STUDY

- Select the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- Build a model wherein one can assign a lead score 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%.

GOALS OF THE CASE STUDY

- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads.
- A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.
- Predict the model and test the model so that the model would be able to adjust to if the company's requirement changes in the future

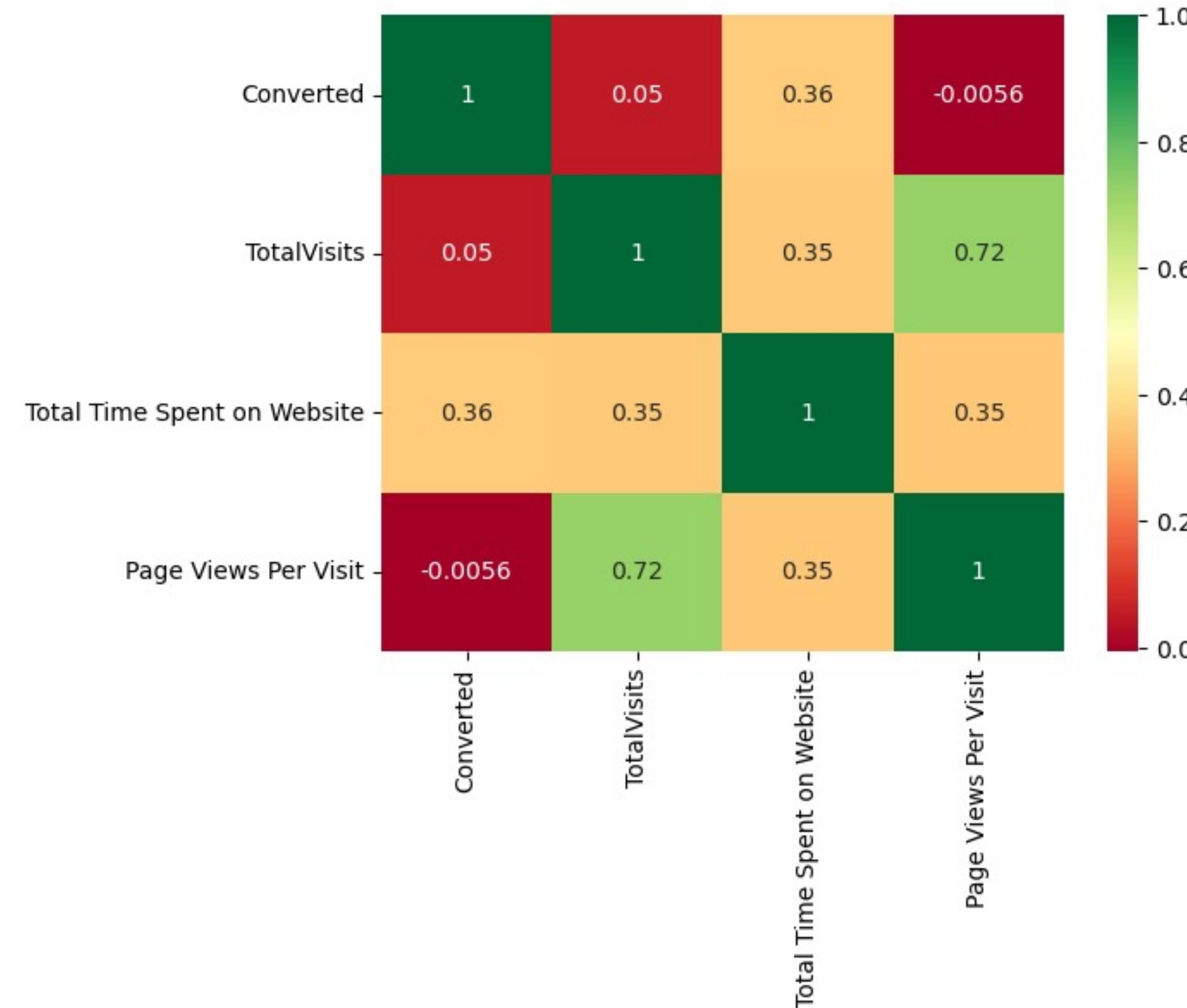
STEPS IN BUILDING THE MODEL

- Data cleaning
- Eda
- Data preparation
- Model building

DATA CLEANING

- The target variable, in this case, is the column 'Converted' which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn't converted.
- A level called 'Select' for categorical variables which is same as null values, is handled by deleting it after the split is done.
- Handling Missing values
 - Deleting the columns with over 30% null values
 - Deleting the columns which has only single values
 - Deleting the columns which has no use for modelling like lead Number, Project ID
 - Converting few of the null values to unknown
 - Deleting the rows for which the missing values are less than 2%.

EDA Correlation



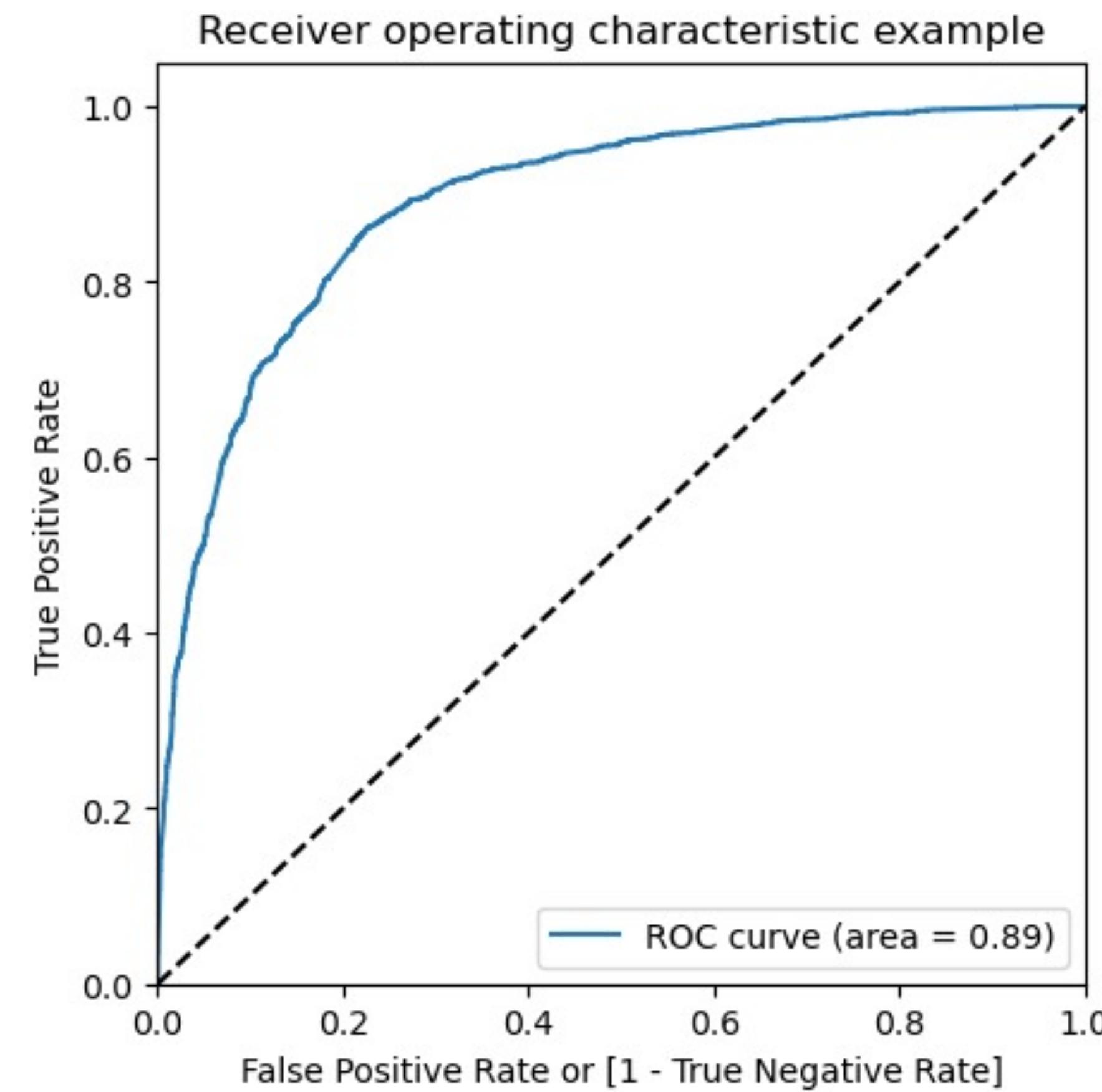
Observations:

1. There is not much correlation between converted and visited.
2. There is a decent correlation between time spent on website and converted.

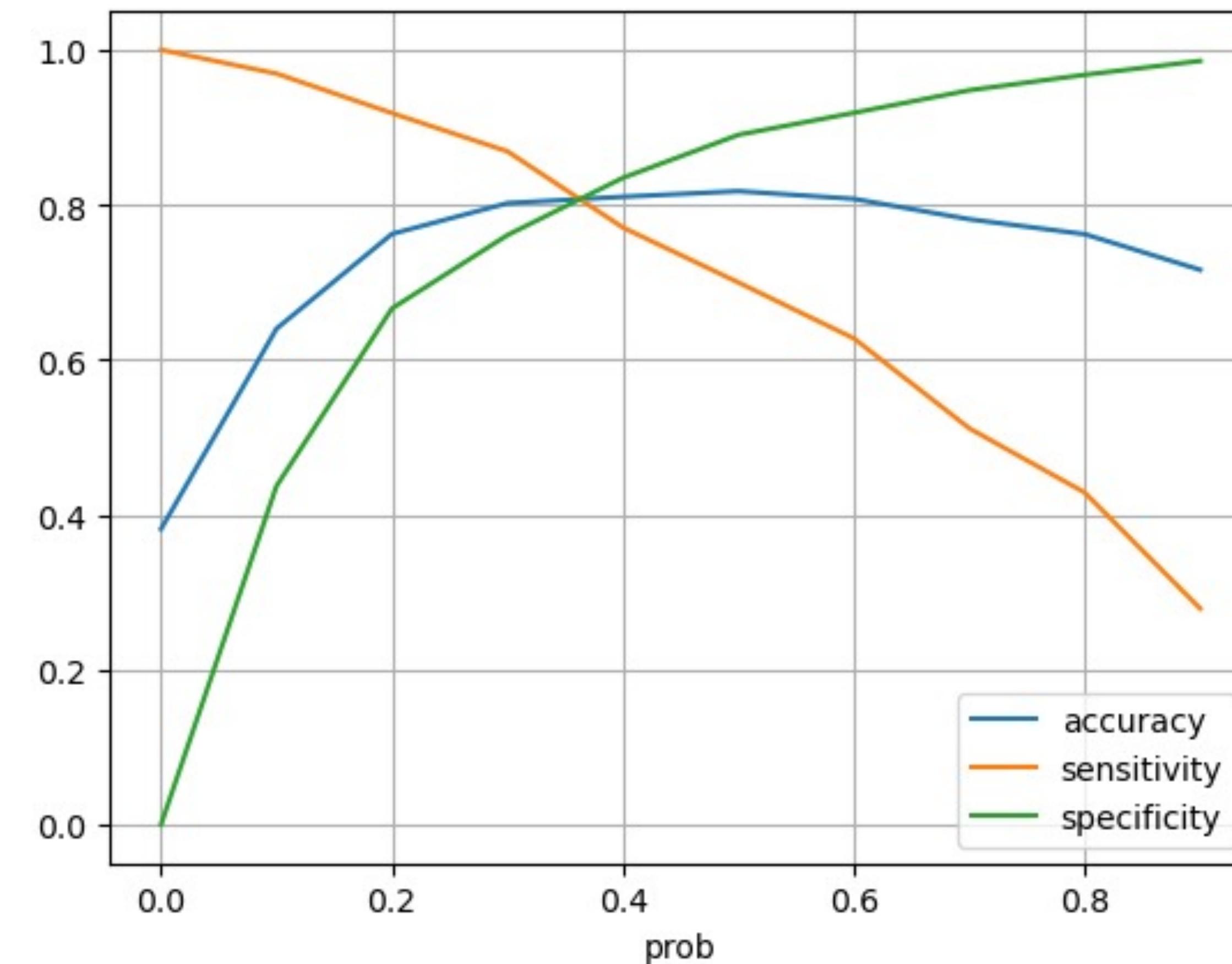
MODEL BUILDING

- Logistic Regression Model for binary classification, as it showed promise in handling the target variable, "Converted".
- RFE was preformed to select the most important columns, the number of variables used for model building is 20
- Manual Feature Reduction process was used to build models by dropping variables with p value greater than 0.05.
- The after multiple Iterations model nine was good with less p-value and VIF less than 5.

- The area under the ROC curve is 0.89 which represents a good fit.



- At around 0.38, all the Three metrics i.e. Accuracy, Sensitivity and Specificity have optimal values. Therefore took 0.38 as a cutoff probability.
- The three metrics for training dataset and test dataset are very close to each other which represents the model is a good fit.



PROBLEM SOLUTION

Based on the Final Model suggestion can be provided to the company as:

- I. Identify the Hot leads based on the high probability of conversion and spend efforts on them.
2. Send out automated Emails, SMS and Chatbots to the Leads.
3. Notify the Leads with the latest offers and updates of the company.
4. Improve the quality of the Interaction with the leads by increasing more sales team.