Lead Scoring Case Study

Bijit Sarkar

Abhishek Rout

Abhirup Mukherjee

Problem statement

- **“Build a logistic regression model to assign a lead score to each of the leads which can be used by the company to filter out the leads that are most likely to convert into paying customers.”**

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.

Business relevance

- The model will increase the company revenue by improving the lead conversion rate
- Current lead conversion rate is 30%. The company expects that by deploying the model, the conversion rate would improve to 80%

Assumptions

- Columns more that 40% missing values were dropped
- Columns having very low or no variance were dropped
- ‘Select’ value throughout the data is treated as ‘null’ value
- The missing values in the categorical columns were replaced by the column’s respective mode (i.e. most occurring value in the column)
- The missing values in the continuous columns were replaced by the column’s respective median (i.e. The value below which 50% of a particular column’s datapoints reside)
- Highly correlated dummy variables were dropped

Datasets

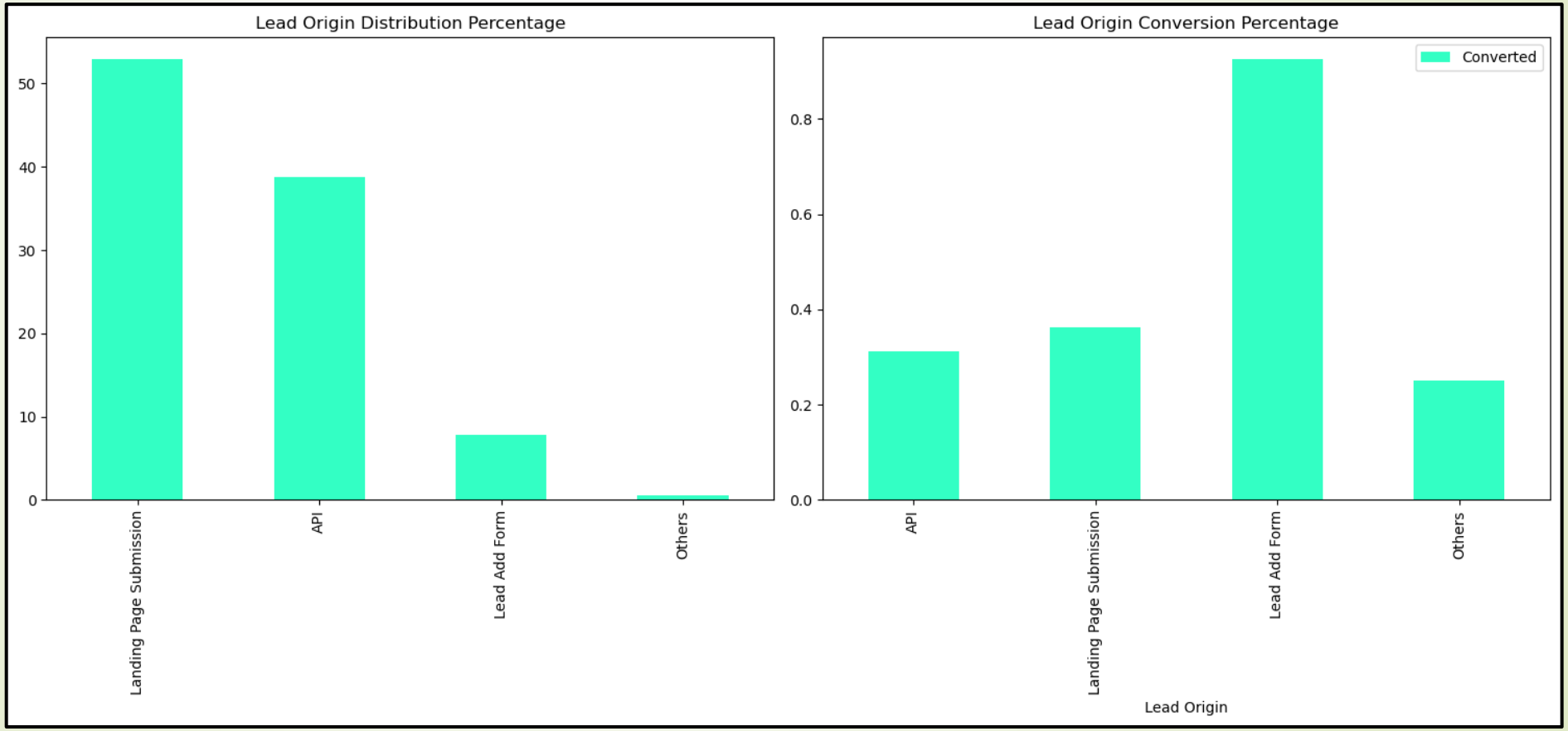
- ***Leads.csv*** (Lead scoring document)
- ***Leads_Data_Dictionary.xlsx*** (Data dictionary. Tells us about the details of individual columns in the given datasets)

Approach for end to end analysis

- Formulate the problem statement
- Import and understand the data
- Preprocessing data (Drop unwanted columns/EDA/missing value/outliers detection/convert categorical variables/get dummy etc.)
- Train- test split
- Feature scaling
- Model building
- Model evaluation

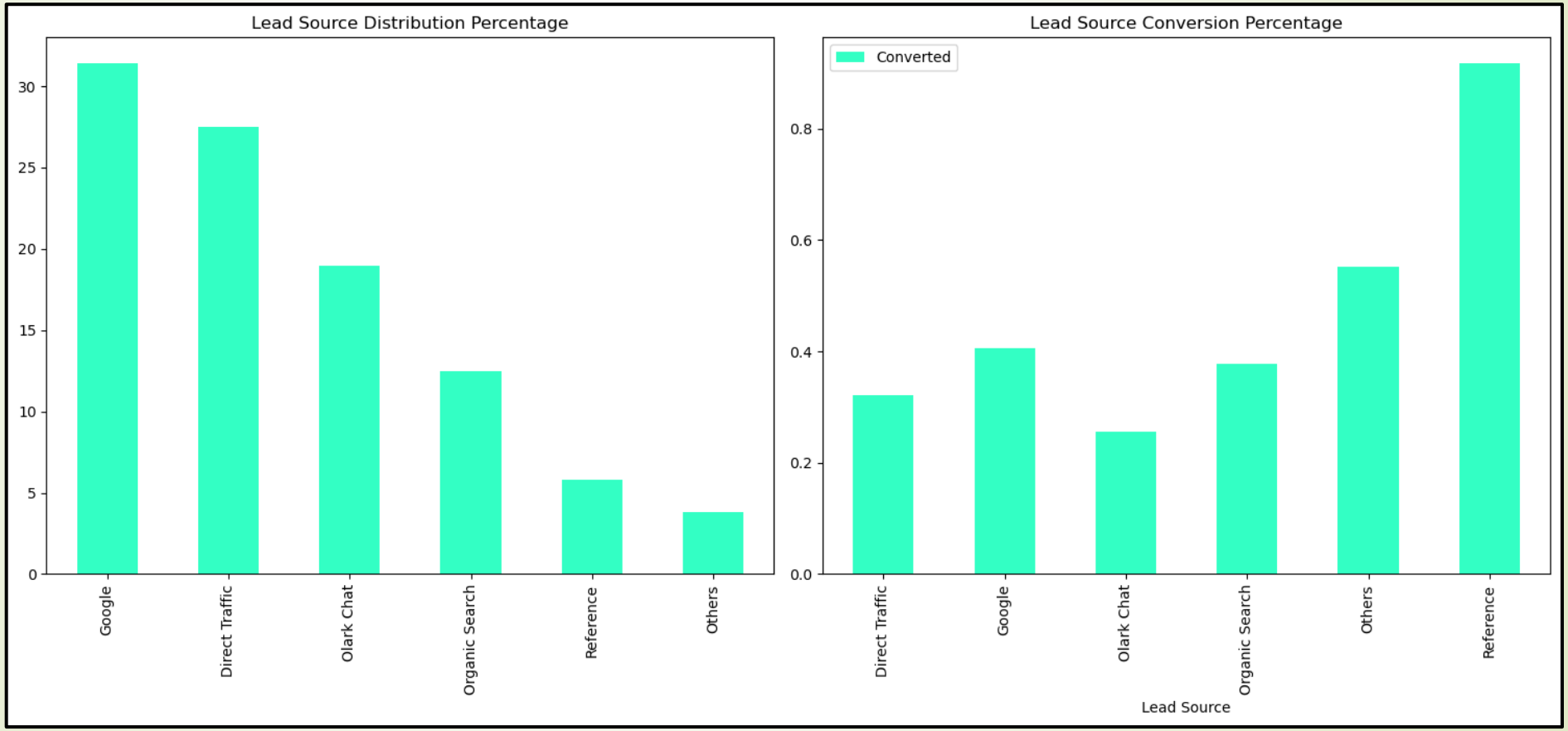
EDA insights

- Maximum distribution of lead origin belongs to ***Landing page submission***. But the maximum lead origin distribution belongs to ***Lead add form***



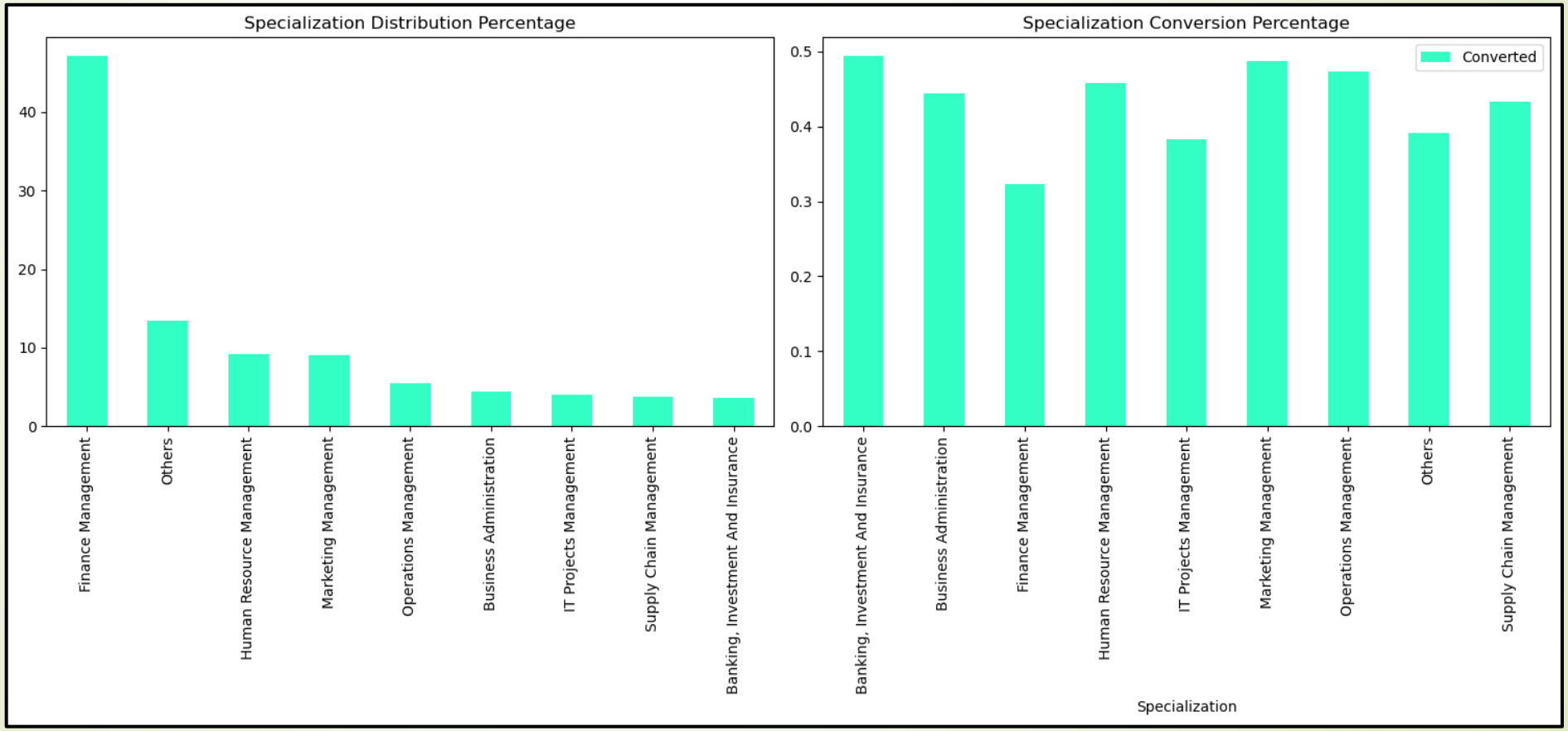
EDA insights

- Maximum distribution of lead source belongs to ***google***. But the maximum converted leads have came through ***references***.



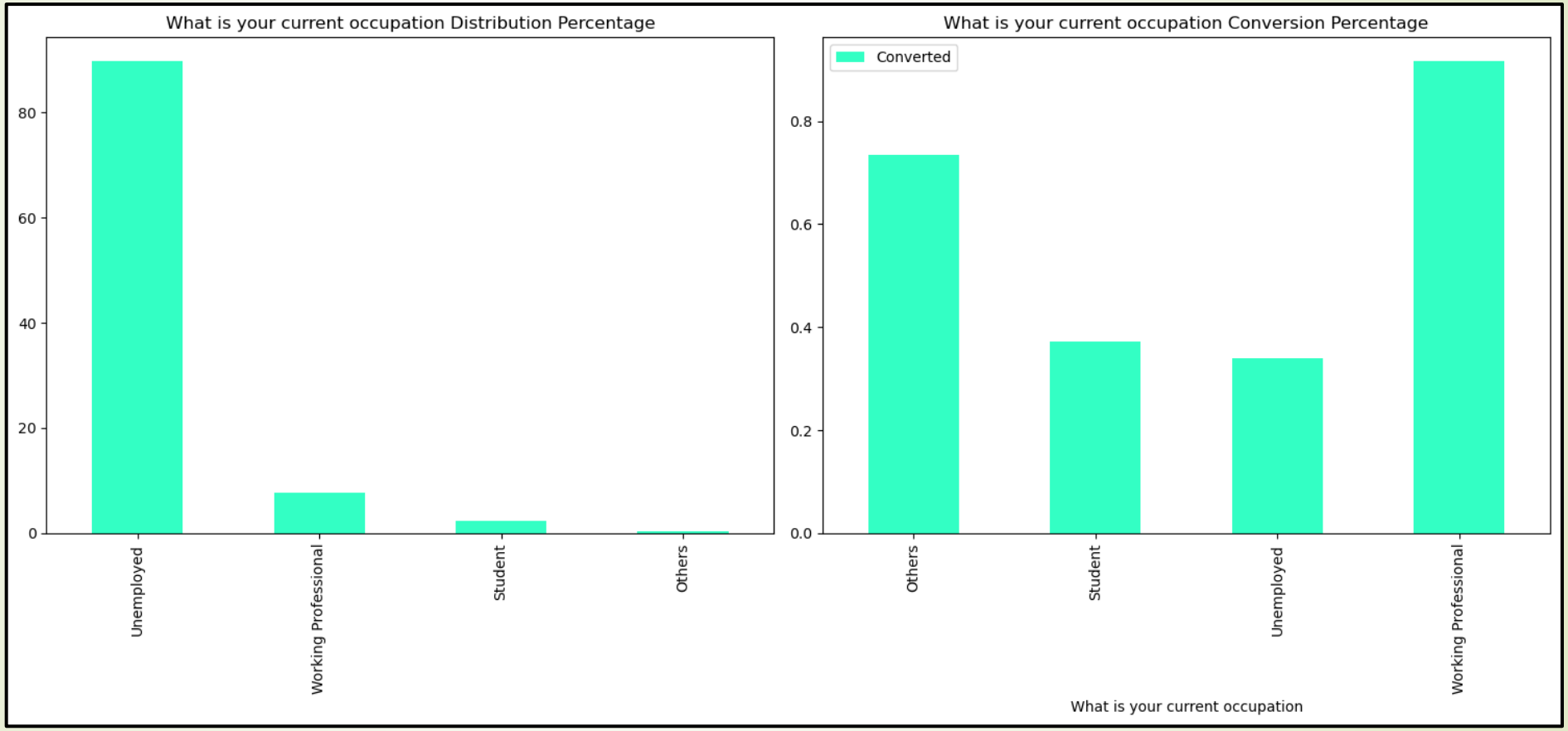
EDA insights

- Very few leads belong from the ***banking, investment and insurance*** domain. But this particular domain has the highest lead conversion rate



EDA insights

- Most of the converted leads are working professionals



EDA insights

- Page Views Per Visit & TotalVisits are correlated. But we considered to keep them.

Feature scaling

- As the numeric variables have outliers, we selected ***StandardScaler*** for feature scaling as StandardScaler is affected less by the presence of outliers.

Building the model

- Started with the pre-processed data
- Most effective predictor variables were evaluated using RFE (We set the number of predictors = 30)
- Built the model using **GLM** (Generalized Linear Model) from **Statsmodel** library
- Checked the summary of the model and analysed different p – values
- VIF (Variance Inflation Factor) was evaluated for the selected predictor variables
- The columns having high p –value & high VIF were dropped
- Built the model again and re-checked p-value & VIF
- Repeated these steps until all the p-values are almost equal to 0 and all the VIF's are less than 5

Insights from the model

The following predictors have a positive weightage on the target variable:

- 1. Last Notable Activity
- 2. What is your current occupation
- 3. Total Time Spent on Website
- 4. Last Activity
- 5. Lead Source

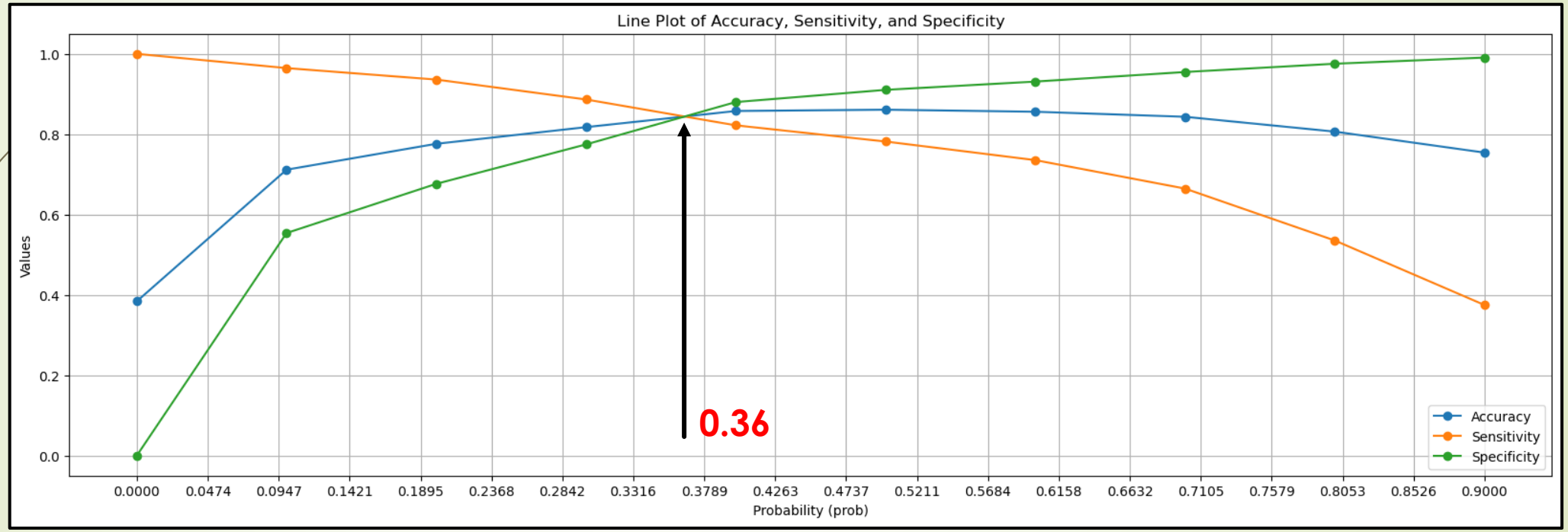
Insights from the model

The following predictors have a negative weightage on the target variable:

- 1. Lead Origin
- 2. Ringing
- 3. Interested in other courses
- 4. Already a student

Model evaluation: Select cut-off

From the model, few performance metrics were calculated (**Accuracy**, **sensitivity** & **specificity**) and plotted against the probability of a lead to convert. From the graph, it's evident that the optimal cut-off should be **0.36**



RESULTS

Model performance

Following are the different performance parameters calculated by applying test data on the model:

- Accuracy score: **87.01%**
- Sensitivity: **87.09%**
- Specificity: **86.96%**
- Positive predictive value: **80.93%**

CONCLUSIONS & SUGGESTIONS

- Following are top 3 categorical/ dummy variables which contribute the most towards the probability of a lead getting converted:
  - Last notable activity/ SMS sent
  - Current Occupation
  - Last activity/ Email opened
- Phone call priority: Focus on **'hot leads'** predicted as 1 by the model
- Segmentation by Lead scoring: Segment customers by Conversion Probability score, **target highest scores first**
- Personalized communication: Train interns for effective personalized communication, highlight benefits of X Education, tailor solutions

- Follow-up strategy: Structured follow-up plan, increase frequency for positive responses
- Feedback loop: Regular intern feedback for refining calling strategy and adjustments
- Offer discounts: Target low-scoring leads with special discounts
- Selective Calling: Prioritize positive attributes, avoid negatives
- Marketing Focus: Emphasize content marketing, targeted emails, and nurturing for lead engagement

THANK YOU