Lead Scoring Case Study using logistic regression

Submitted By –
Kartik Singh

Contents

- Problem statement
- Problem approach
- EDA
- Correlations
- Model Evaluation
- Observations
- Conclusion

Problem Statement

- An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses. They have process of form filling on their website after which the company that individual as a lead.
- 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.
- The typical lead conversion rate at X education is around 30%. Now, this means 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.
- If they successfully identify this set of 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

Business Objective

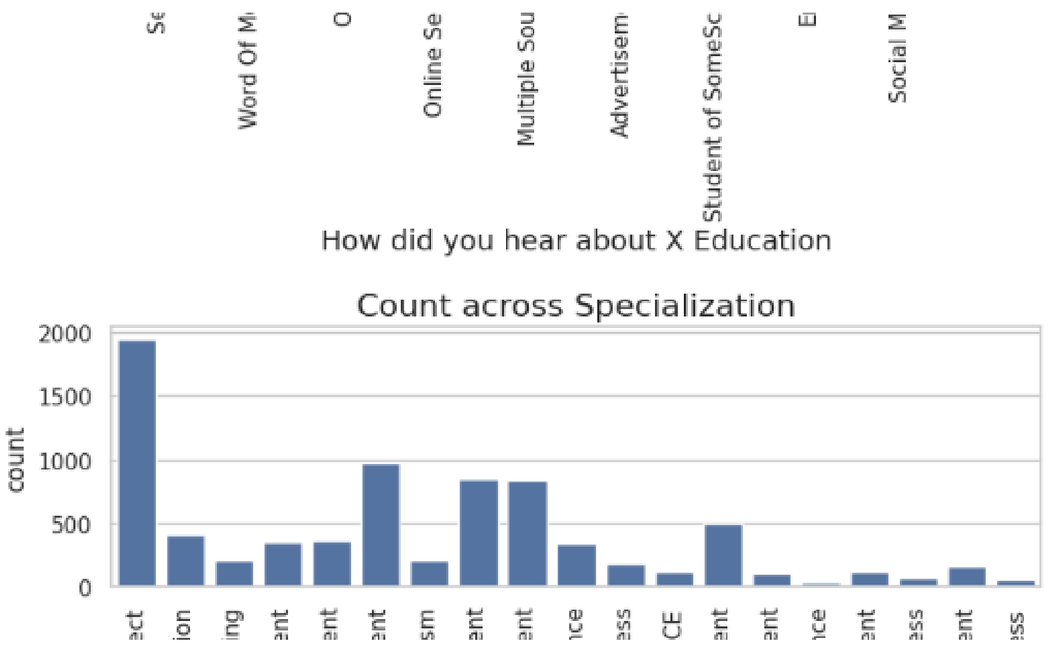
- Lead X wants us to build a model to give every lead a lead score between 0 -100 . So that they can identify the Hot leads and increase their conversion rate as well.
- The CEO want to achieve a lead conversion rate of 80%.
  - They want the model to be able to handle future constraints as well like Peak time actions required, how to utilize full man power and after achieving target what should be the approaches.

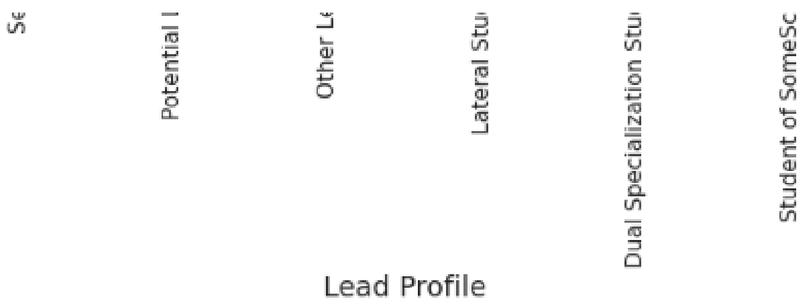
Problem approach

- EDA
- Feature scaling
- Correlations
- Model Building (RFE R-squared VIF and pvalues)
- Model Evaluation
- Making predictions on test set
- Importing the data and inspecting the data frame
- Data preparation
- Dummy variable creation
- Test-Train split

EDA – Data Cleaning

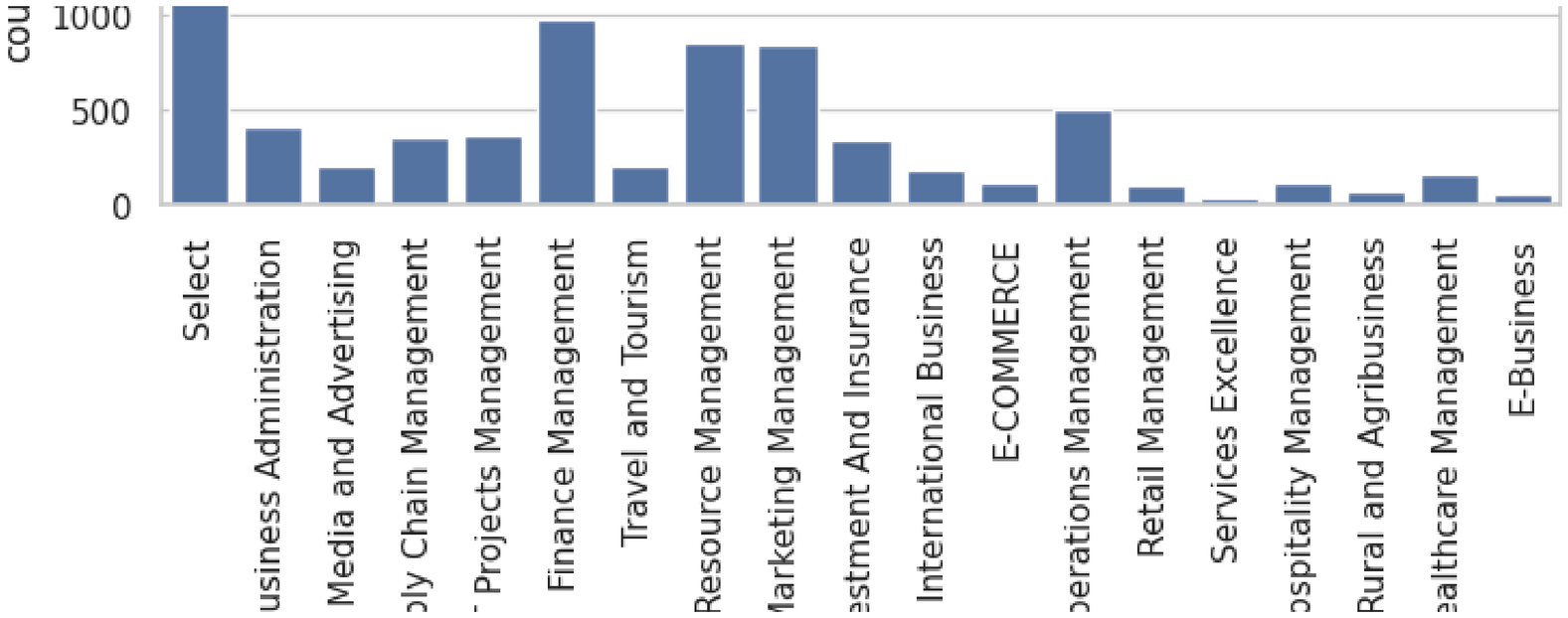
- "In certain columns, there is a category labeled 'Select' that is responsible for managing specific tasks."





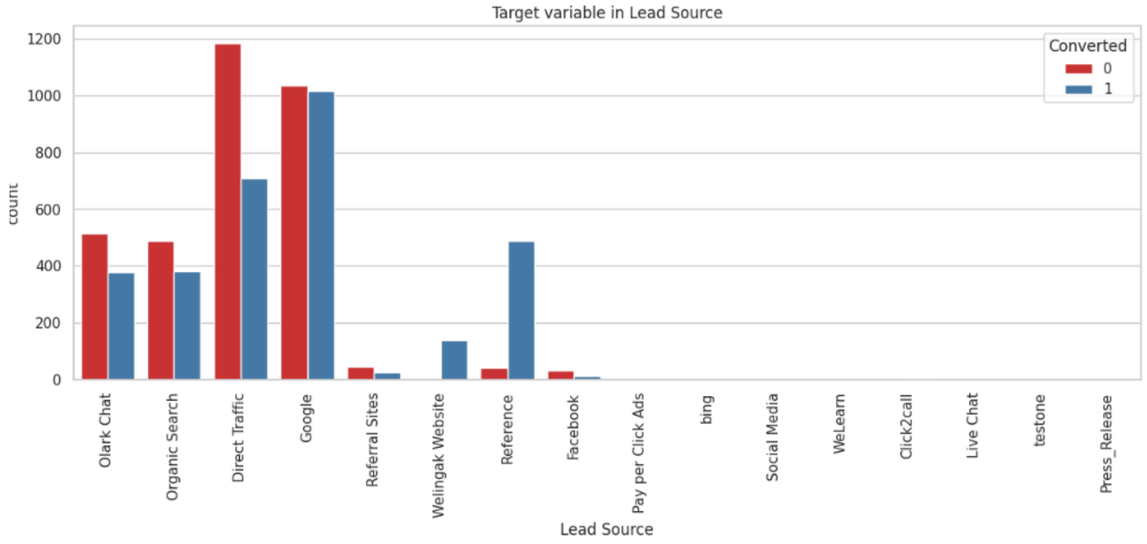
Specialization

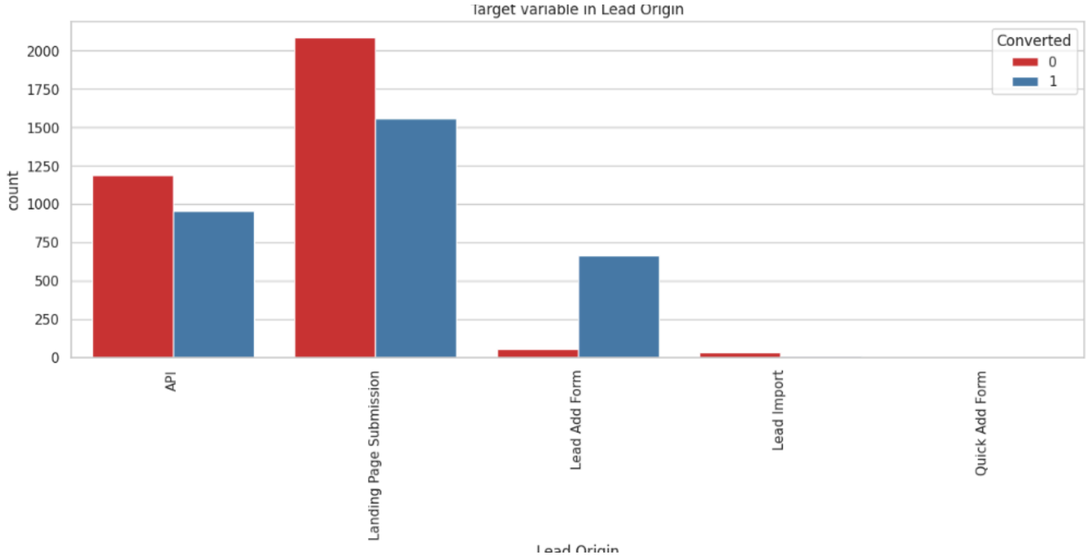
- Leads specializing in HR, Finance, and Marketing Management have a higher likelihood of conversion.



Lead source and lead origin

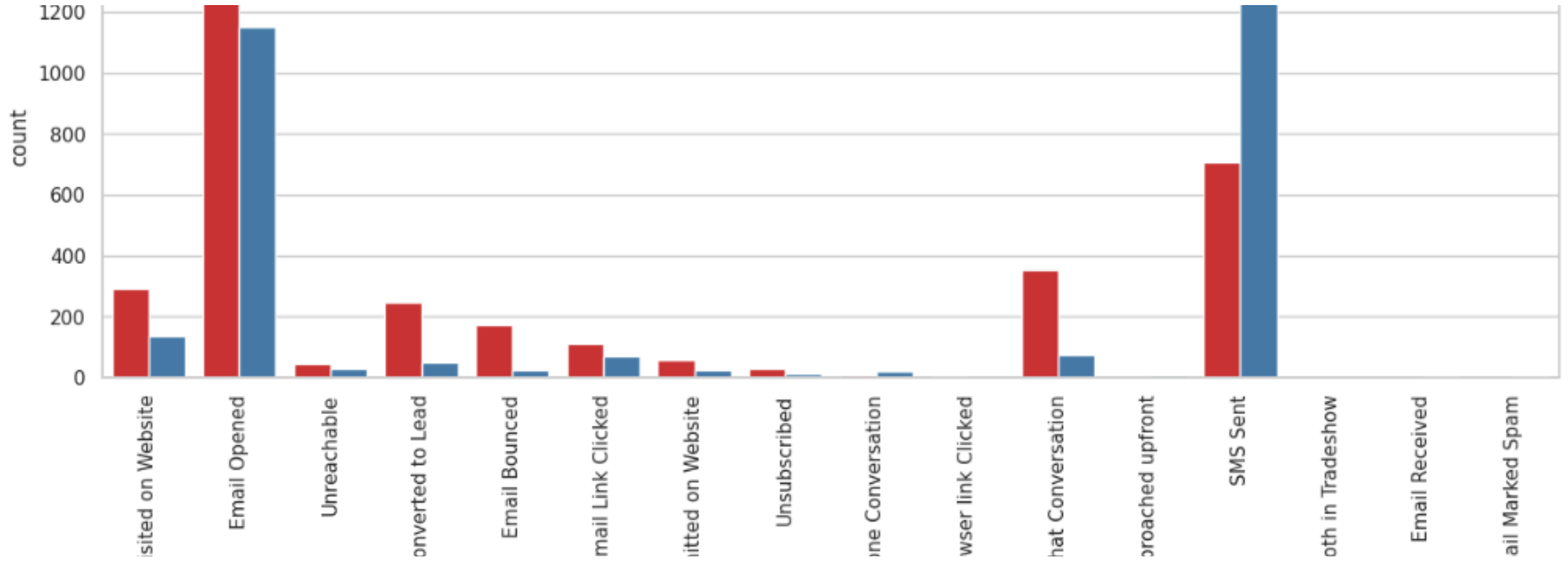
- Leads obtained through Google and direct traffic have a higher likelihood of conversion. Meanwhile, the majority of leads originate from submissions.





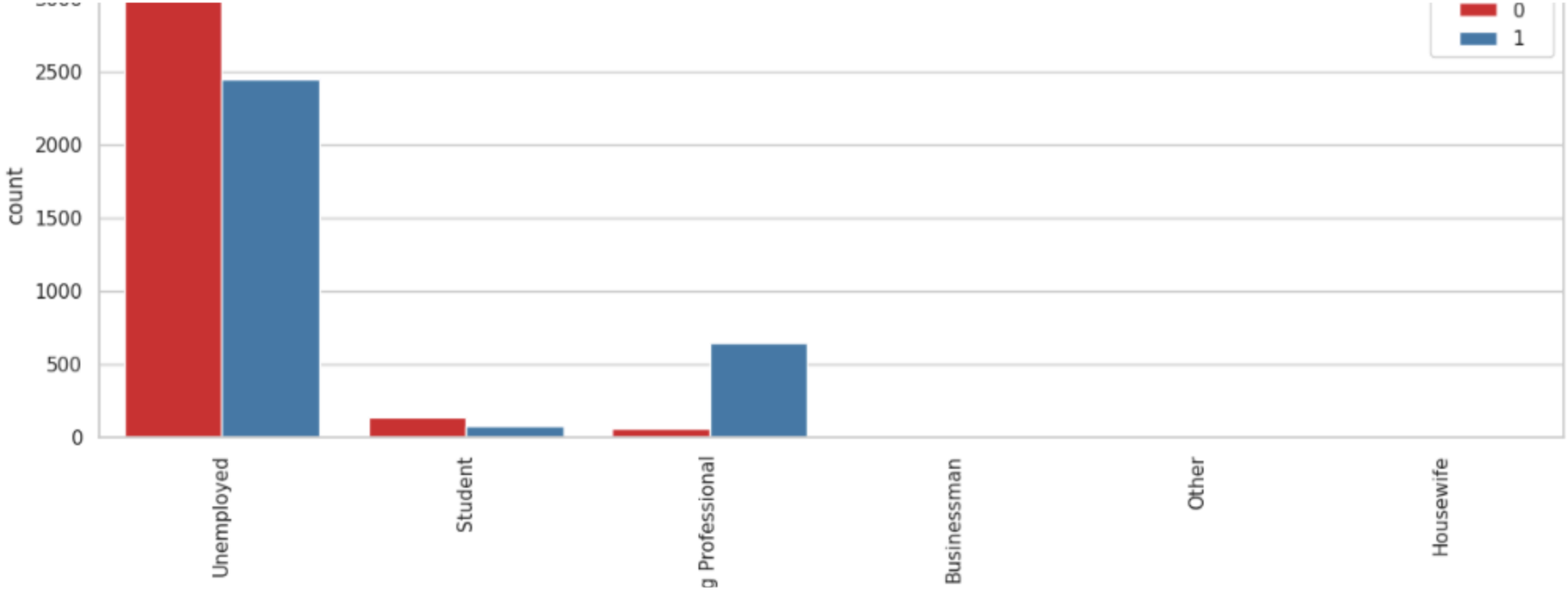
Last lead activity

- Leads that open emails have a high probability of converting, and similarly, sending SMS messages also proves beneficial for conversions.

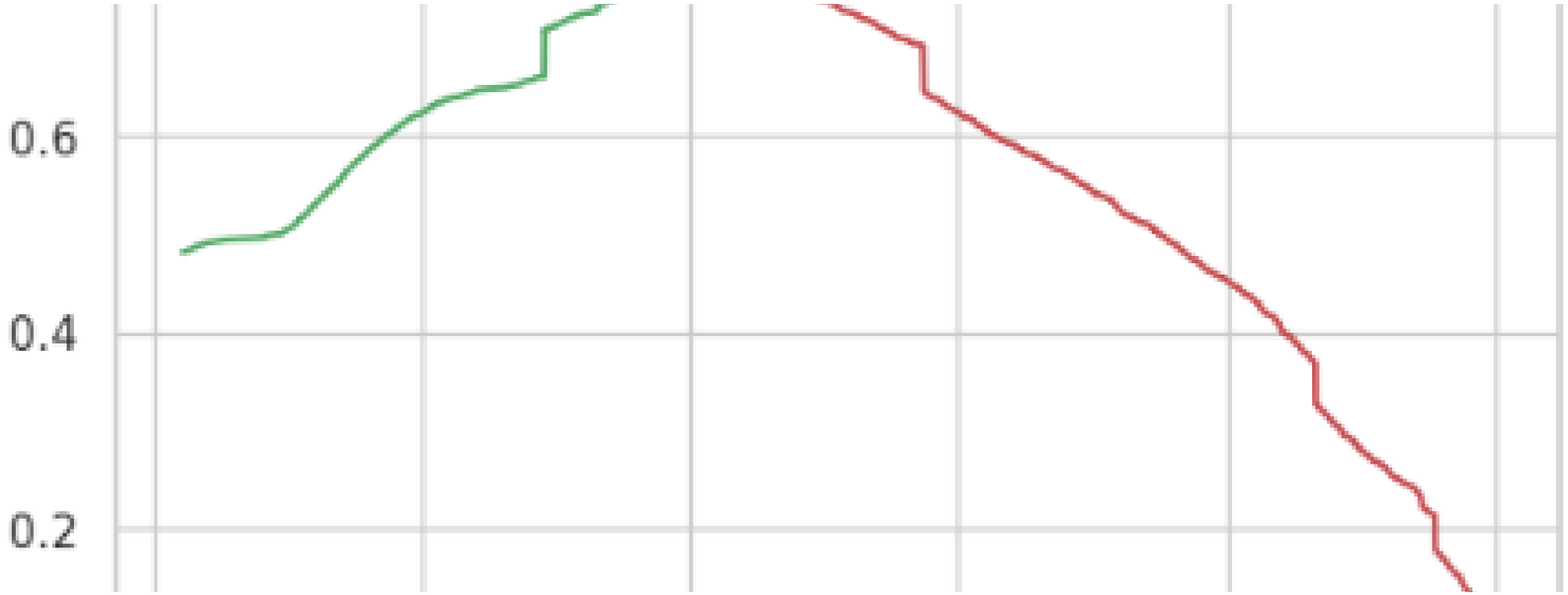


Occupation

- Leads which are unemployed are most interested to join the course than others

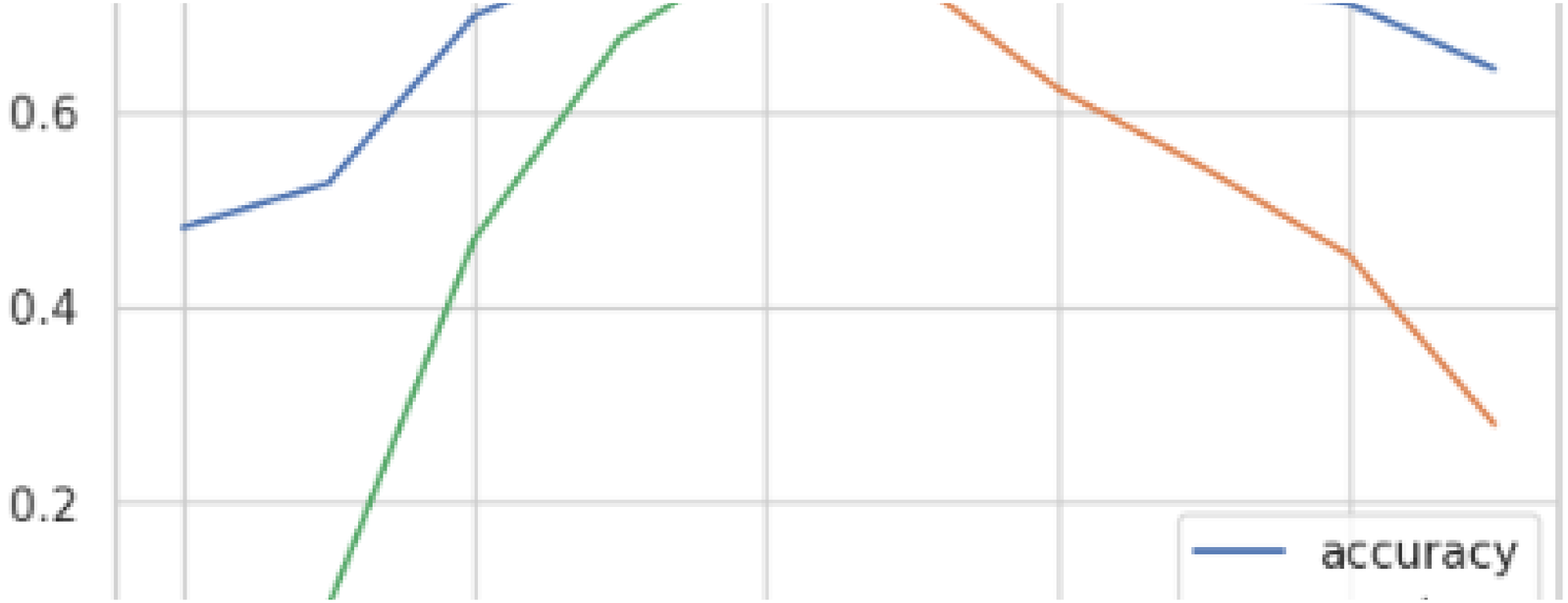


Model evaluation



ROC curve

- "Given that 0.42 represents the tradeoff between Precision and Recall, we can confidently decide to focus on any prospect lead with a conversion probability greater than this threshold."



Observations

- Train Data:

1. Accuracy – 80%
2. Sensitivity – 77%
3. Specificity – 80%

- Test Data:

4. Accuracy – 80%
5. Sensitivity – 77%
6. Specificity – 80%

- Final Features list:

- Lead Source_Olark Chat
- Specialization_Others
- Lead Origin_Lead Add Form Lead Source_Welingak Website Total Time Spent on Website Lead Origin_Landing Page

Conclusion

- We see that the conversion rate is 30-35% (close to average) for API and Landing page submission. But very low for Lead Add form and Lead import. Therefore we can intervene that we need to focus more on the leads originated from API and Landing page submission.
- We see max number of leads are generated by google / direct
- traffic. Max conversion ratio is by reference and welingak website.
- Leads who spent more time on website, more likely to convert.

THANK YOU