Lead Scoring Assignment

NIHARIKA DUSA

JAYASHREE SANGLE

JITHU SUBHASH

INTRODUCTION

- 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.

  - 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%.
  - 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.

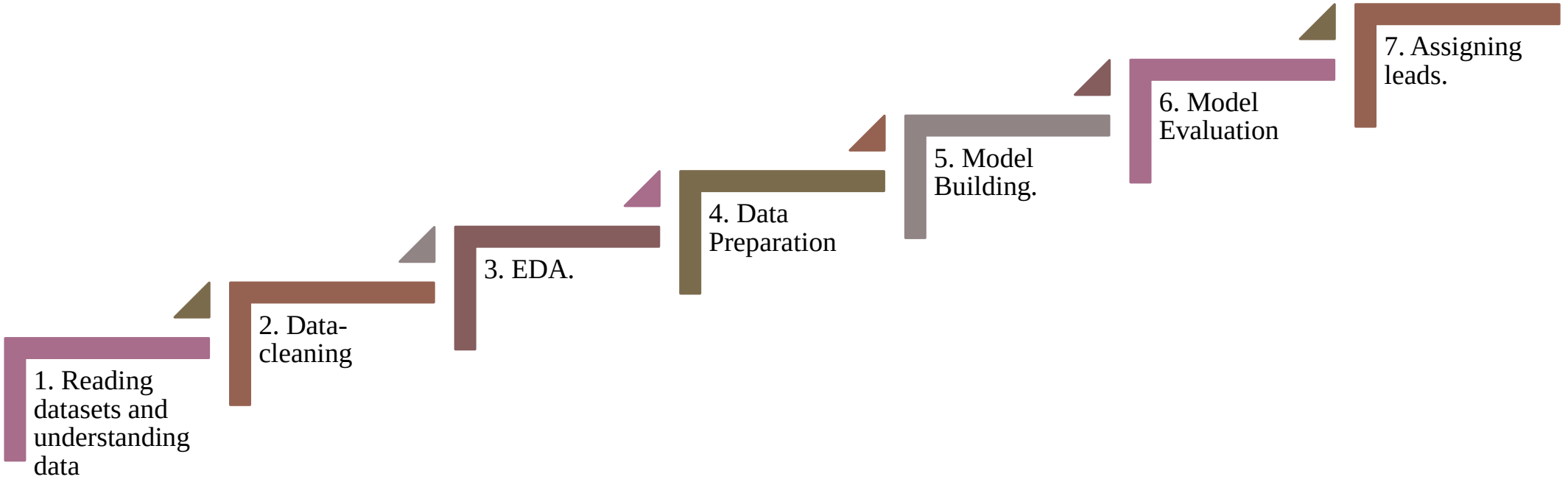
Business Objective

- The company requires you to build a model wherein you need to 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%.

Datasets used:

- Leads.csv

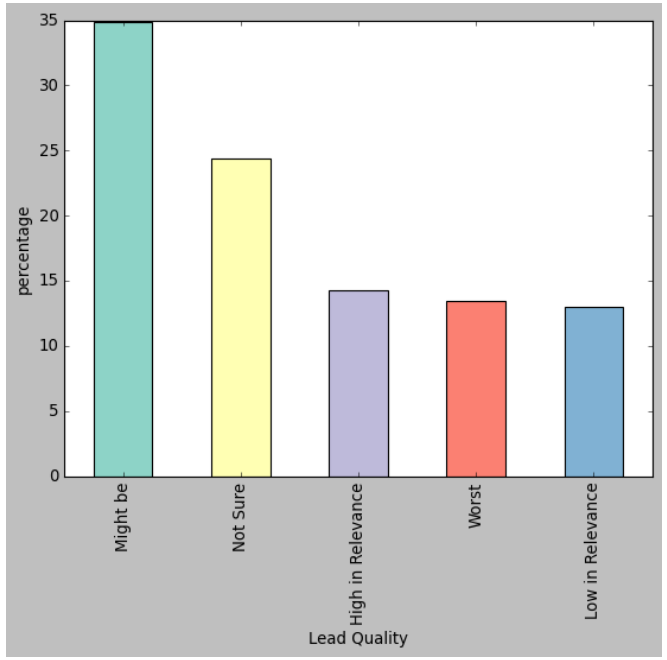
Steps involved

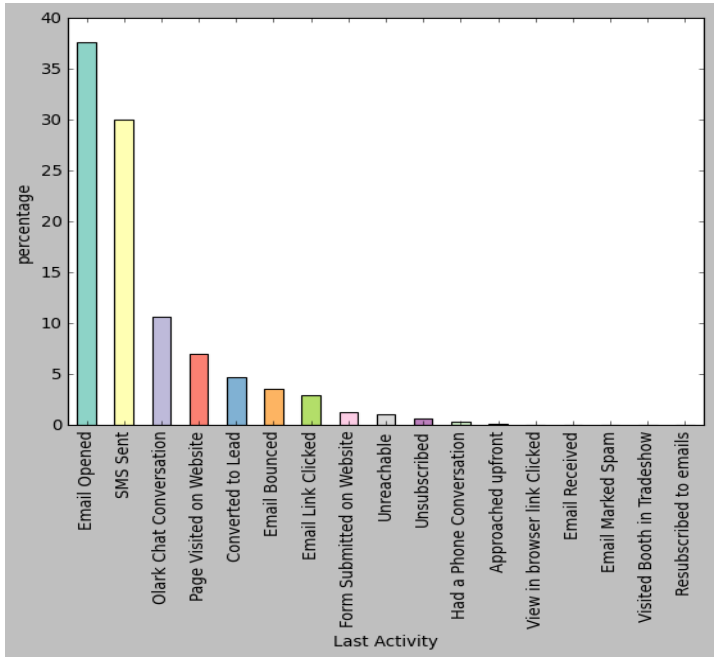


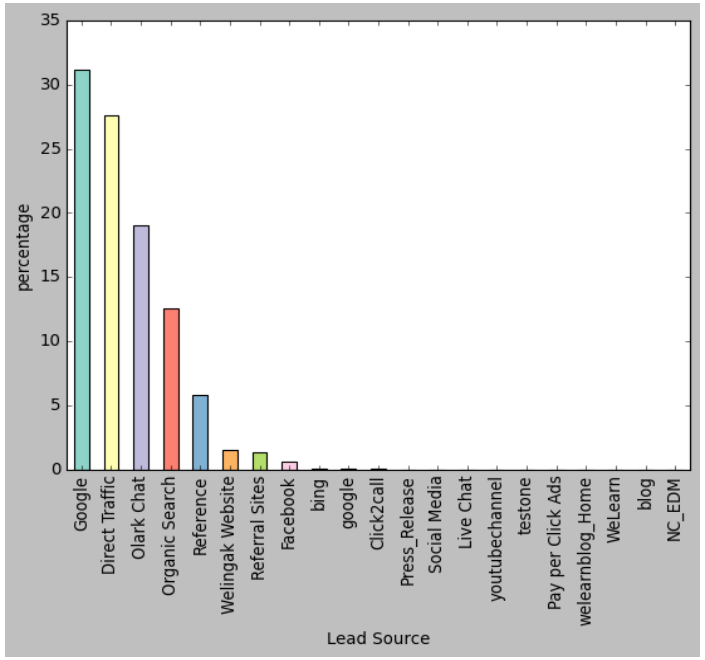
Data Cleaning

Data Cleaning

- There are few null values and ‘Select’ column similar to null values which are taken care.
- Lead Source has spelling error ‘google which as handled



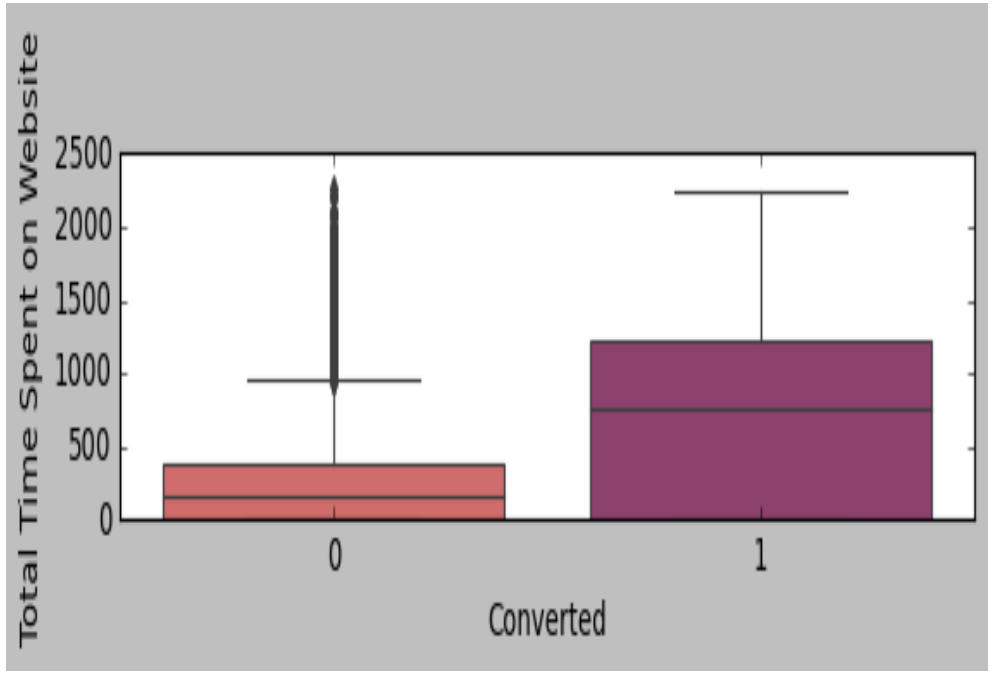


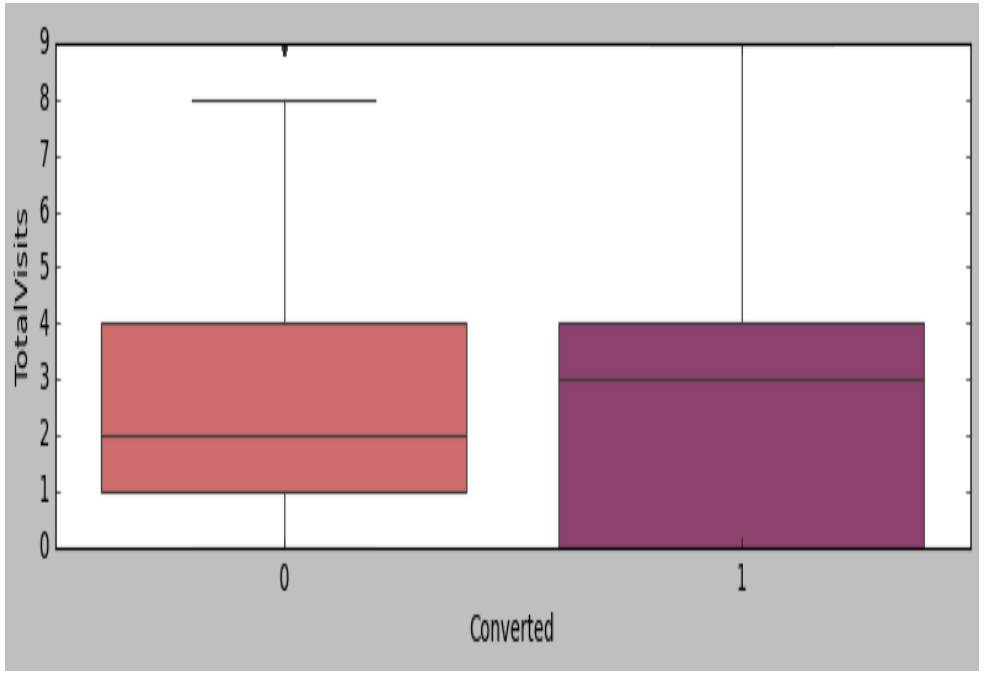


EDA

Numerical variables

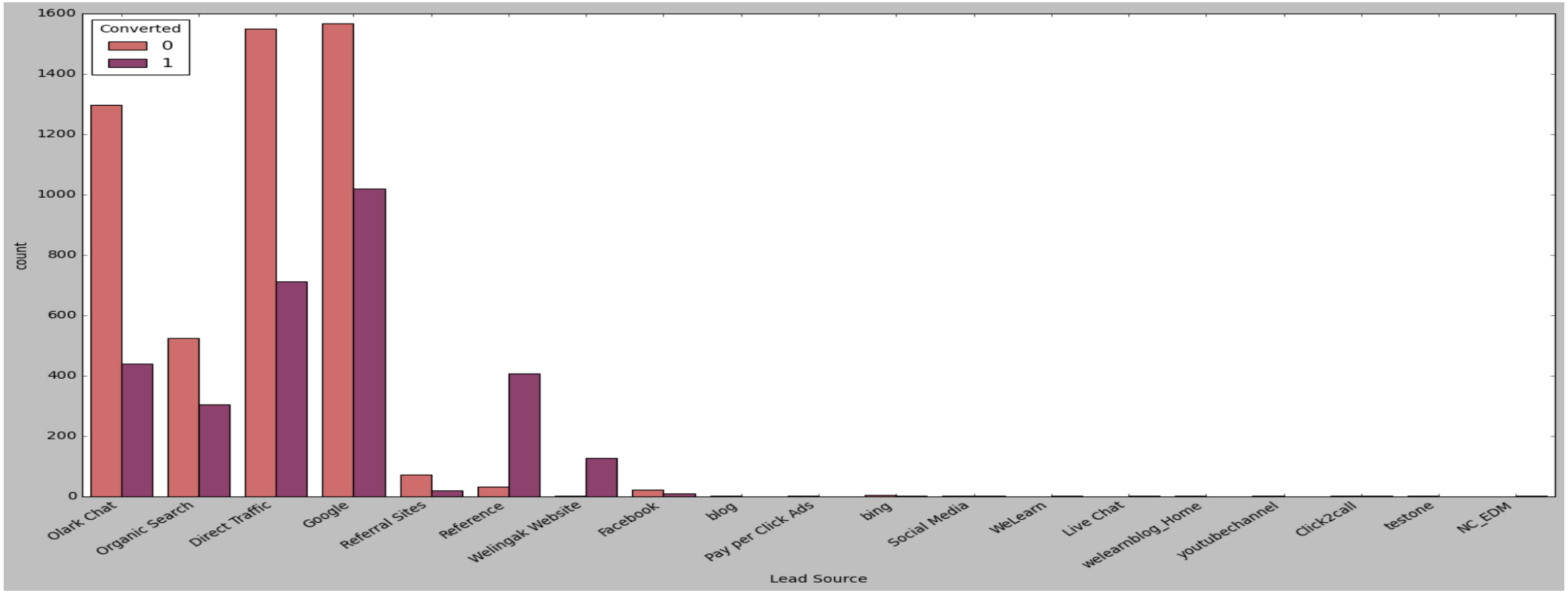
- People who are spending more time on the website are the ones getting converted





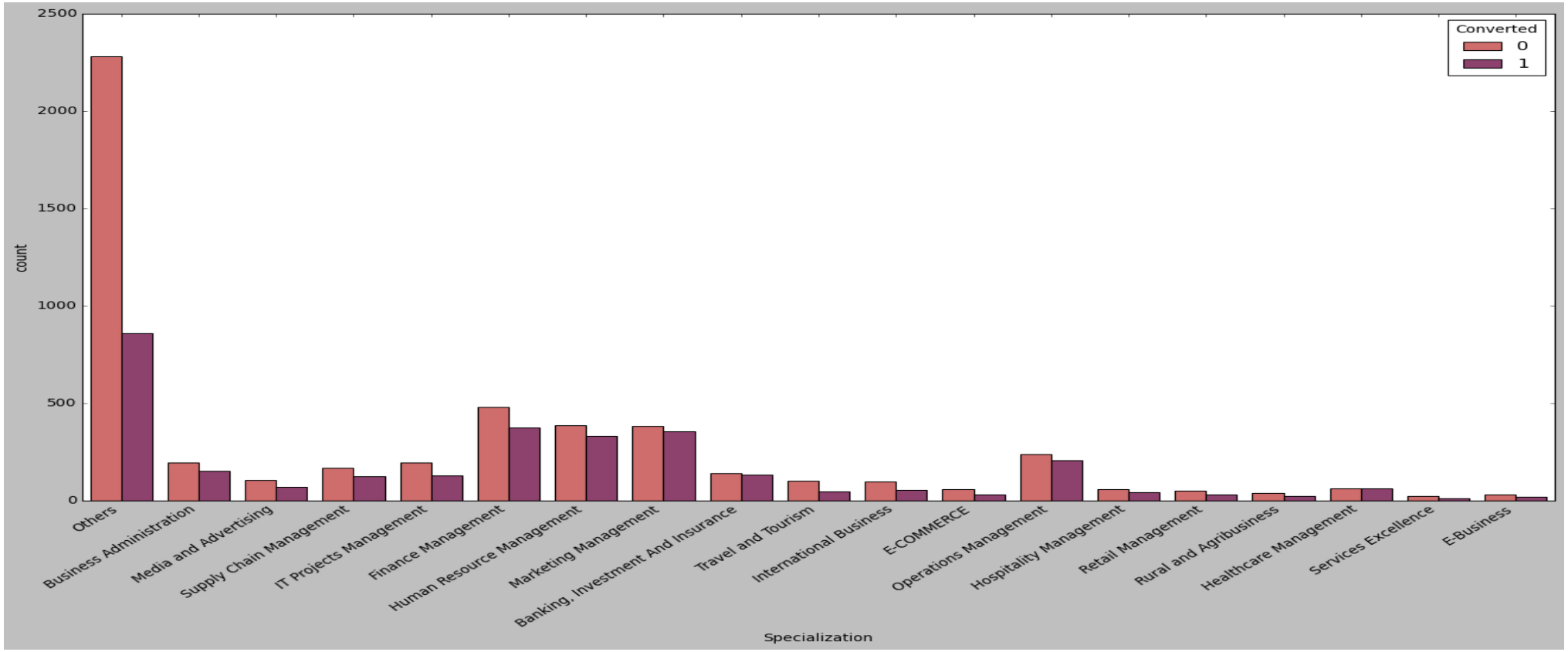
Lead Source

- Direct traffic and Google generate more leads.
- Reference and Welingak website leads have more conversion rate



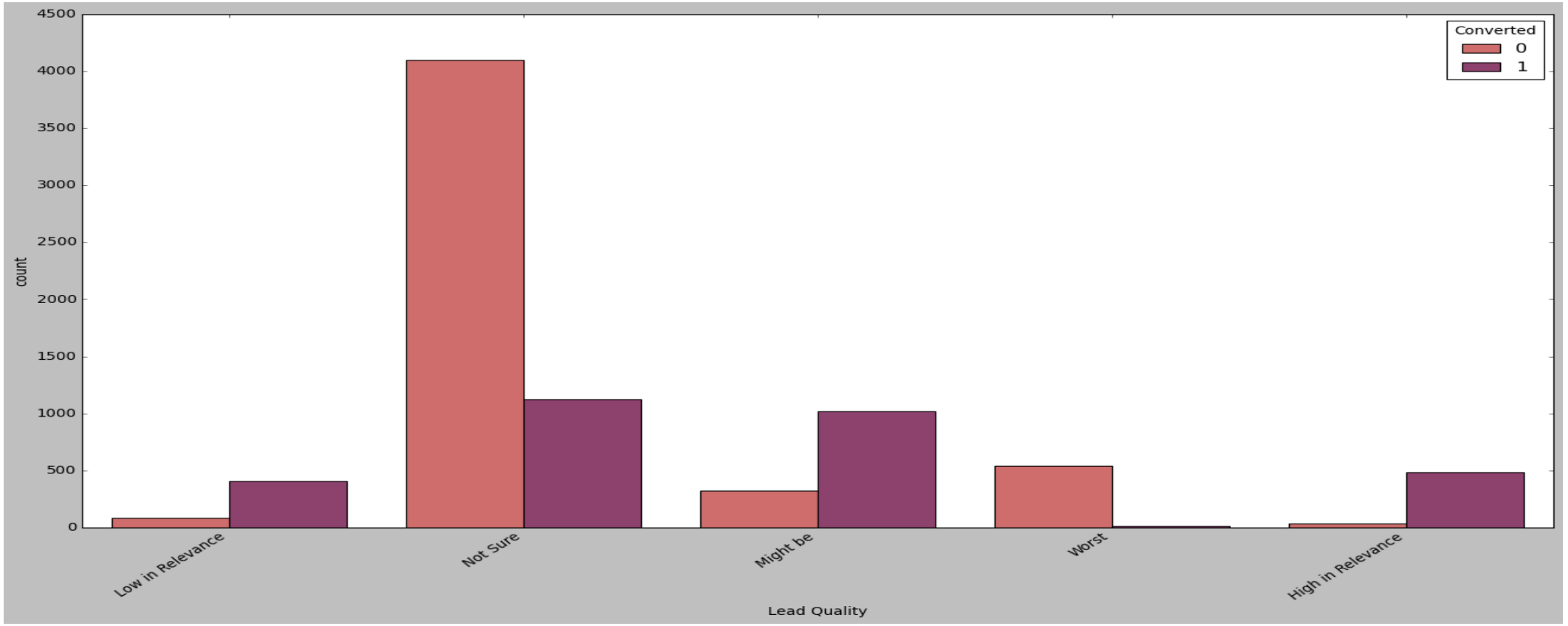
Specialization

- All specializations seem to have same conversion rates



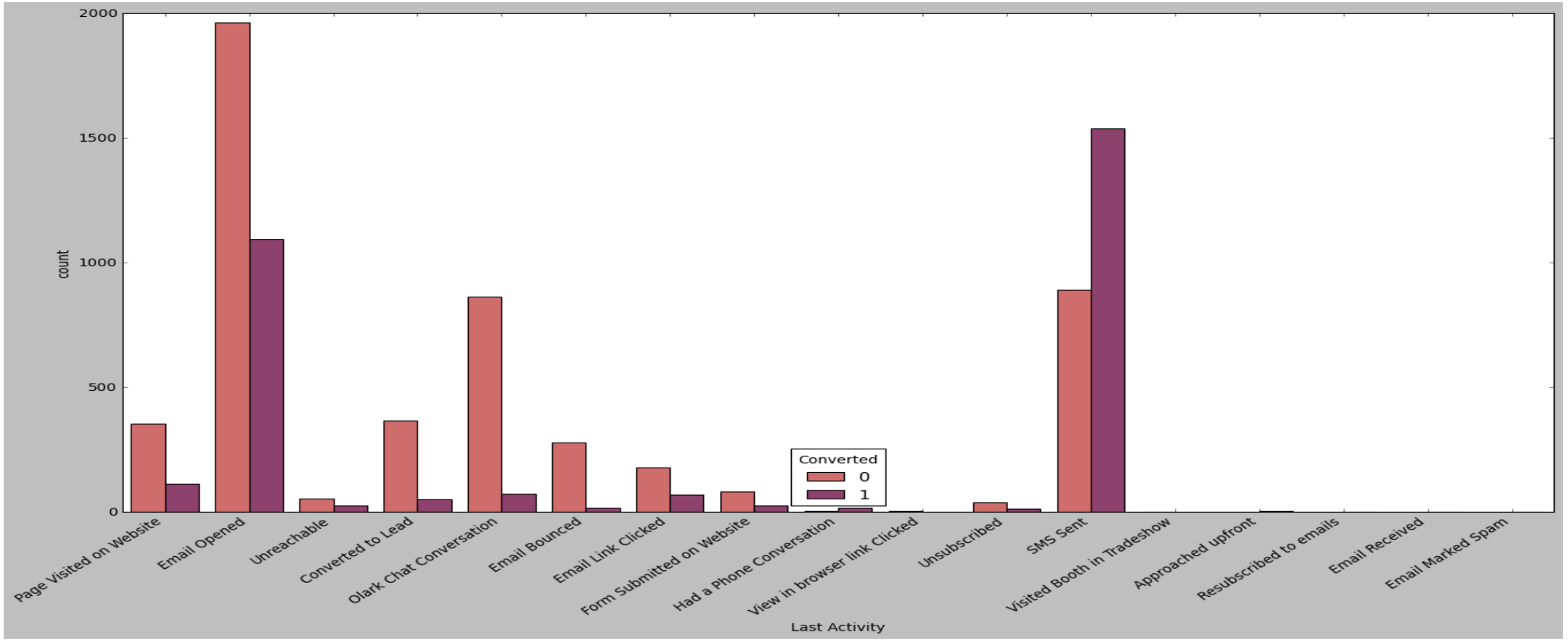
Lead Quality

- “Might be’ has more conversion rates and “worst” has less conversion rate



Last activity

- Email Opened and SMS sent has more leads converted



Model Evaluation

Final Model

- All P-values are less than 0.05 or equal to 0.

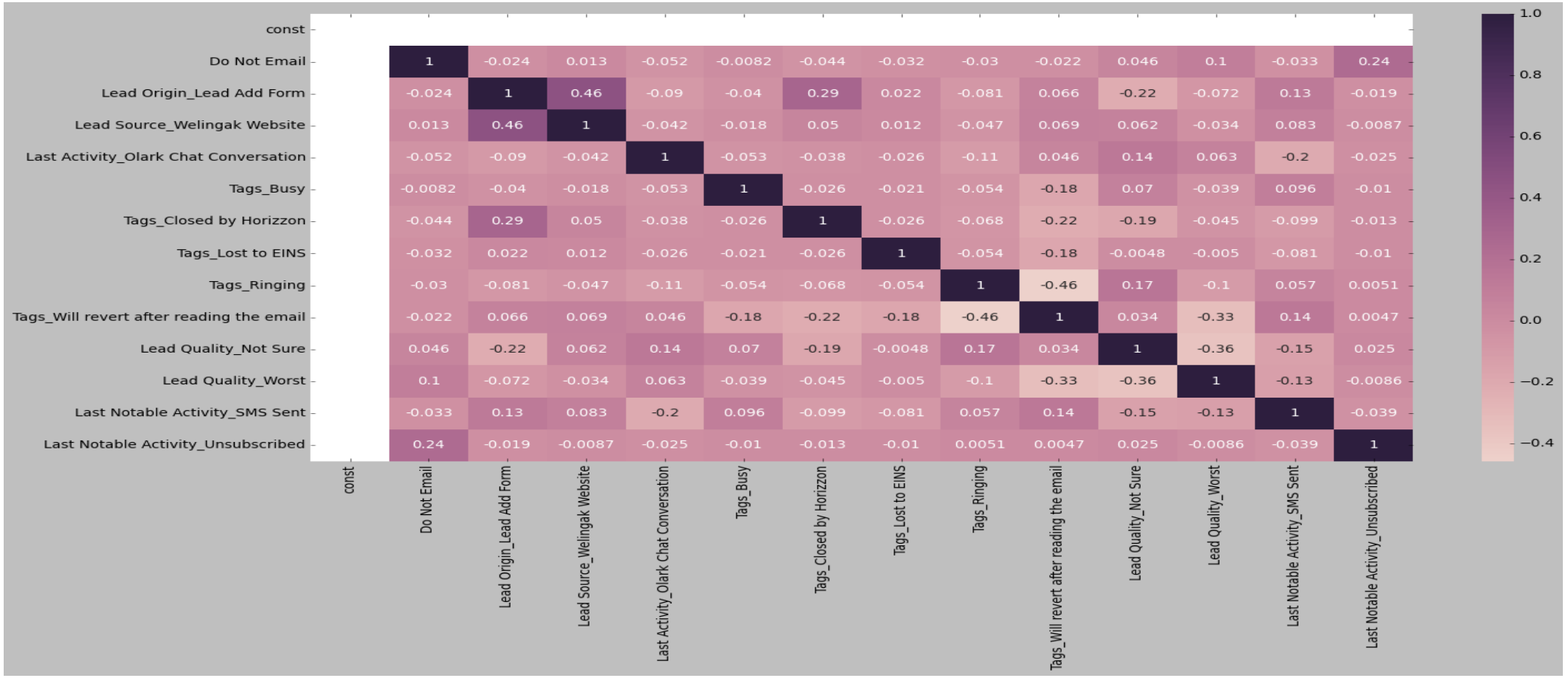
Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	5686
Model:	GLM	Df Residuals:	5672
Model Family:	Binomial	Df Model:	13
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1415.7
Date:	Tue, 18 Jul 2023	Deviance:	2831.5
Time:	11:02:40	Pearson chi2:	2.13e+04
No. Iterations:	8	Pseudo R-squ. (CS):	0.5641
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.3271	0.228	-10.193	0.000	-2.775	-1.880
Do Not Email	-1.5367	0.242	-6.358	0.000	-2.010	-1.063
Lead Origin_Lead Add Form	1.6044	0.448	3.580	0.000	0.726	2.483
Lead Source_Welingak Website	2.6365	0.865	3.049	0.002	0.942	4.331
Last Activity_Olark Chat Conversation	-1.4260	0.193	-7.382	0.000	-1.805	-1.047
Tags_Busy	4.5403	0.337	13.482	0.000	3.880	5.200
Tags_Closed by Horizzon	8.0716	0.786	10.271	0.000	6.531	9.612
Tags_Lost to EINS	8.9701	0.576	15.568	0.000	7.841	10.099
Tags_Ringing	-0.7998	0.338	-2.366	0.018	-1.462	-0.137
Tags_Will revert after reading the email	4.5231	0.248	18.236	0.000	4.037	5.009
Lead Quality_Not Sure	-3.4610	0.137	-25.293	0.000	-3.729	-3.193
Lead Quality_Worst	-3.3073	0.844	-3.918	0.000	-4.962	-1.653
Last Notable Activity_SMS Sent	2.3984	0.121	19.805	0.000	2.161	2.636
Last Notable Activity_Unsubscribed	2.0639	0.616	3.352	0.001	0.857	3.271

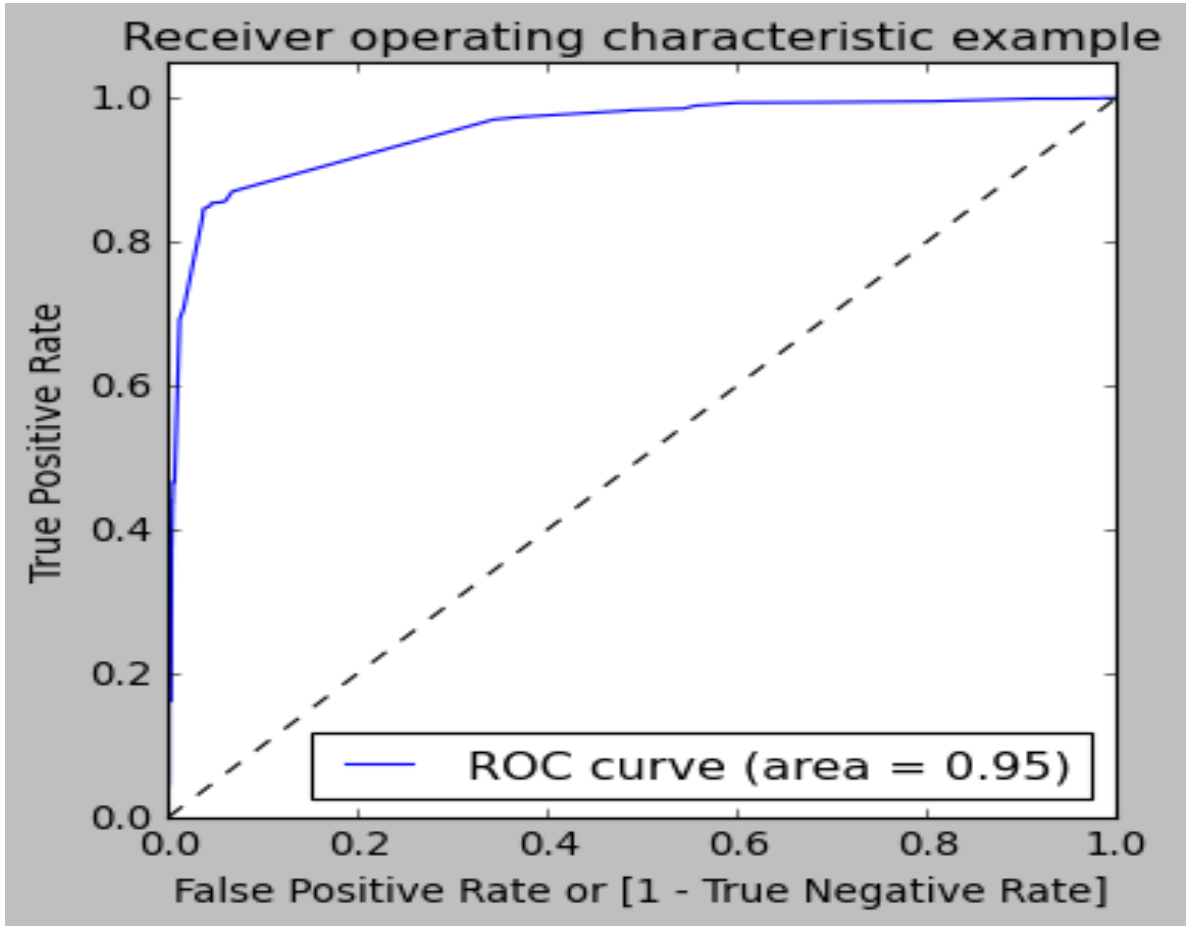
Correlations

- We can say there is very less multicollinearity or no correlations between variables.



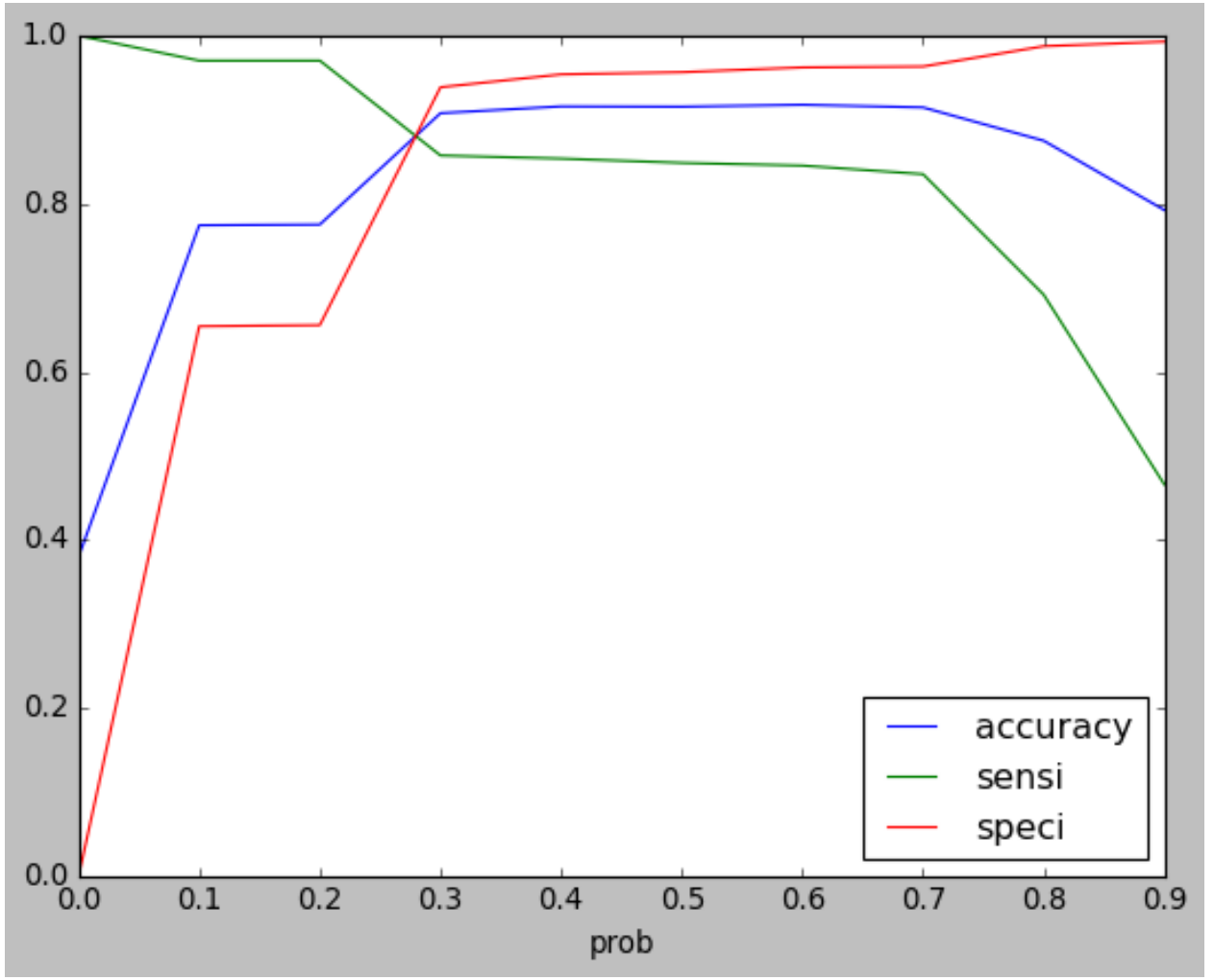
ROC Curve

- Area under the curve is 0.95



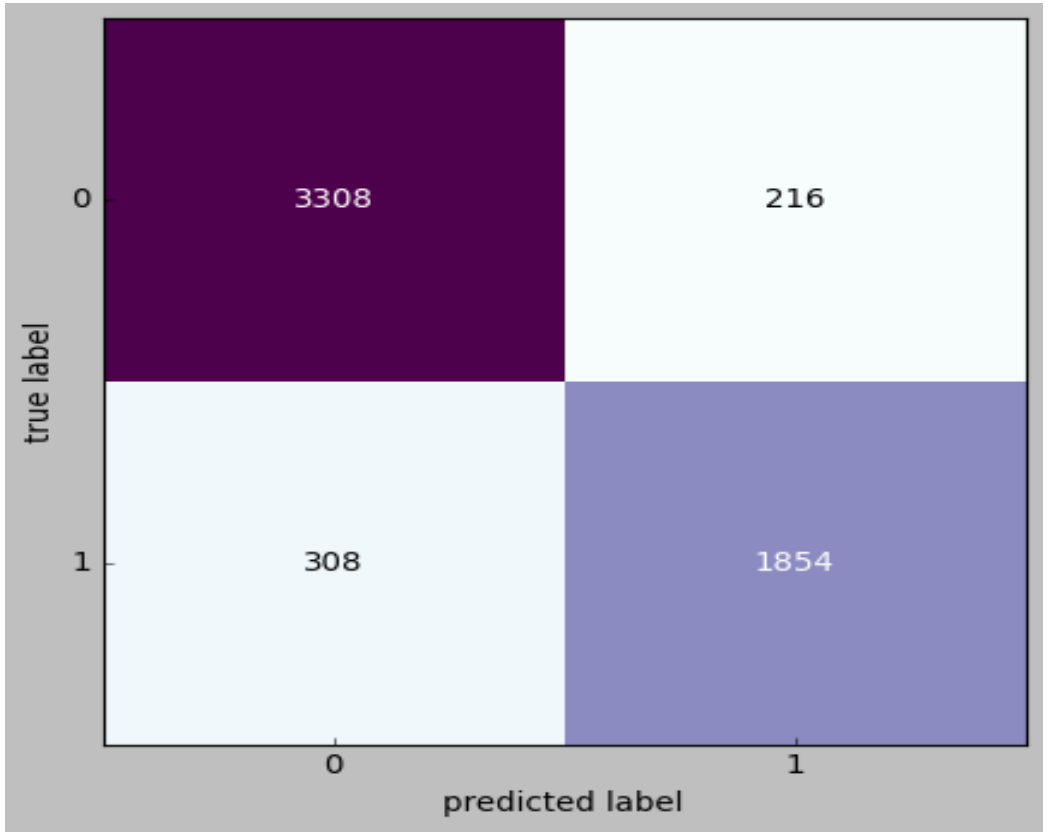
Optimal point cut-off

- 0.25 is the optimal point cutoff probability
- Graph shows sensitivity,specificity and accuracy

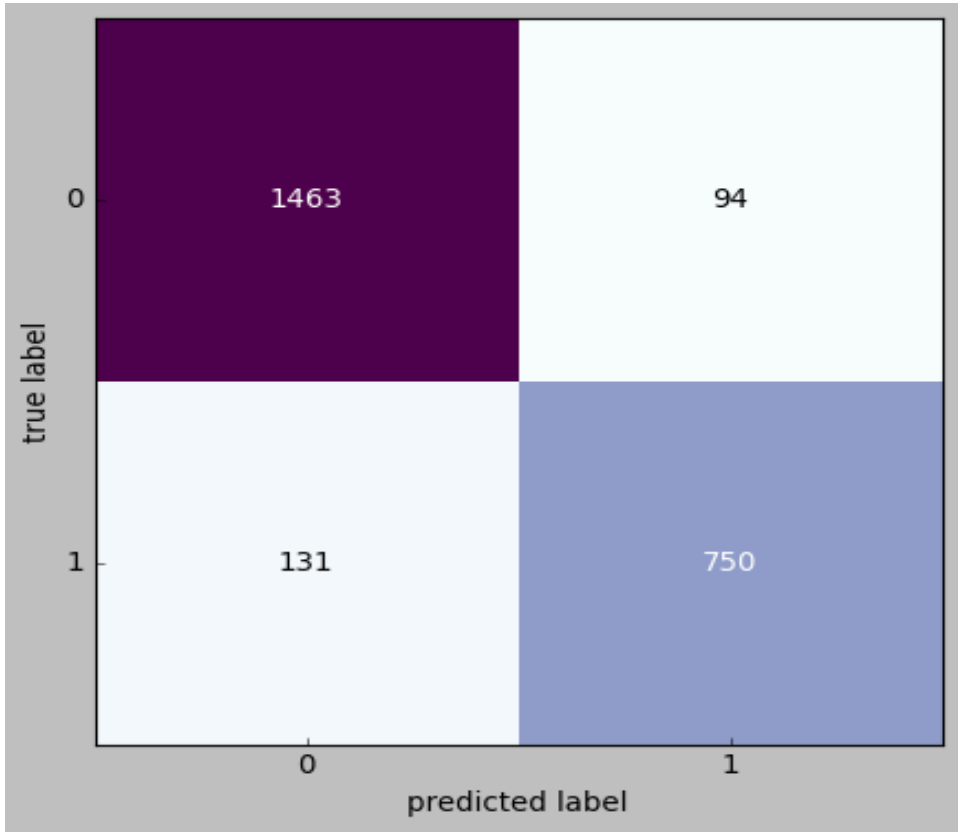


Confusion Matrix

- Confusion matrix for train data



- Confusion matrix for test data



Observations

➢Train data and Test data metrics

➢Train Data:
•Accuracy:91.5%
•Sensitivity:84.8%
•Specificity:95.6%

➢Test Data:
•Accuracy:90.7%
•Sensitivity:85.1%
•Specificity:93.9%

➢List of Final Features

Tags_Lost to EINS	8.970135
Tags_Closed by Horizzon	8.071577
Tags_Busy	4.540301
Tags_Will revert after reading the email	4.523072
Lead Source_Welingak Website	2.636515
Last Notable Activity_SMS Sent	2.398414
Last Notable Activity_Unsubscribed	2.06391
Lead Origin_Lead Add Form	1.604442
Tags_Ringing	-0.79979
Last Activity_Olark Chat Conversation	-1.426
Do Not Email	-1.53665
const	-2.3271
Lead Quality_Worst	-3.30733
Lead Quality_Not Sure	-3.46103

Recommendations

- The company should focus on leads with high conversion probabilities (lead score >= 85) as these have a higher chance of conversion.
- Leads from Google and references, as well as those who have opened emails and received SMS, have higher conversion rates. The company can prioritize these leads in their marketing and sales efforts.
- Leads with 'Will revert after reading the email' and 'Not Sure' in the 'Tags' column have higher conversion rates, indicating they are potential targets for follow-up.
- The city 'Mumbai' seems to have a higher conversion rate, so the company can focus more on leads from Mumbai.

- **The company should continue to monitor and refine the model periodically as the business and lead characteristics may change over time.**

THANK YOU