LEAD SCORING CASE STUDY

BY –S.SOWJANYA

BATCH NO: DS C52

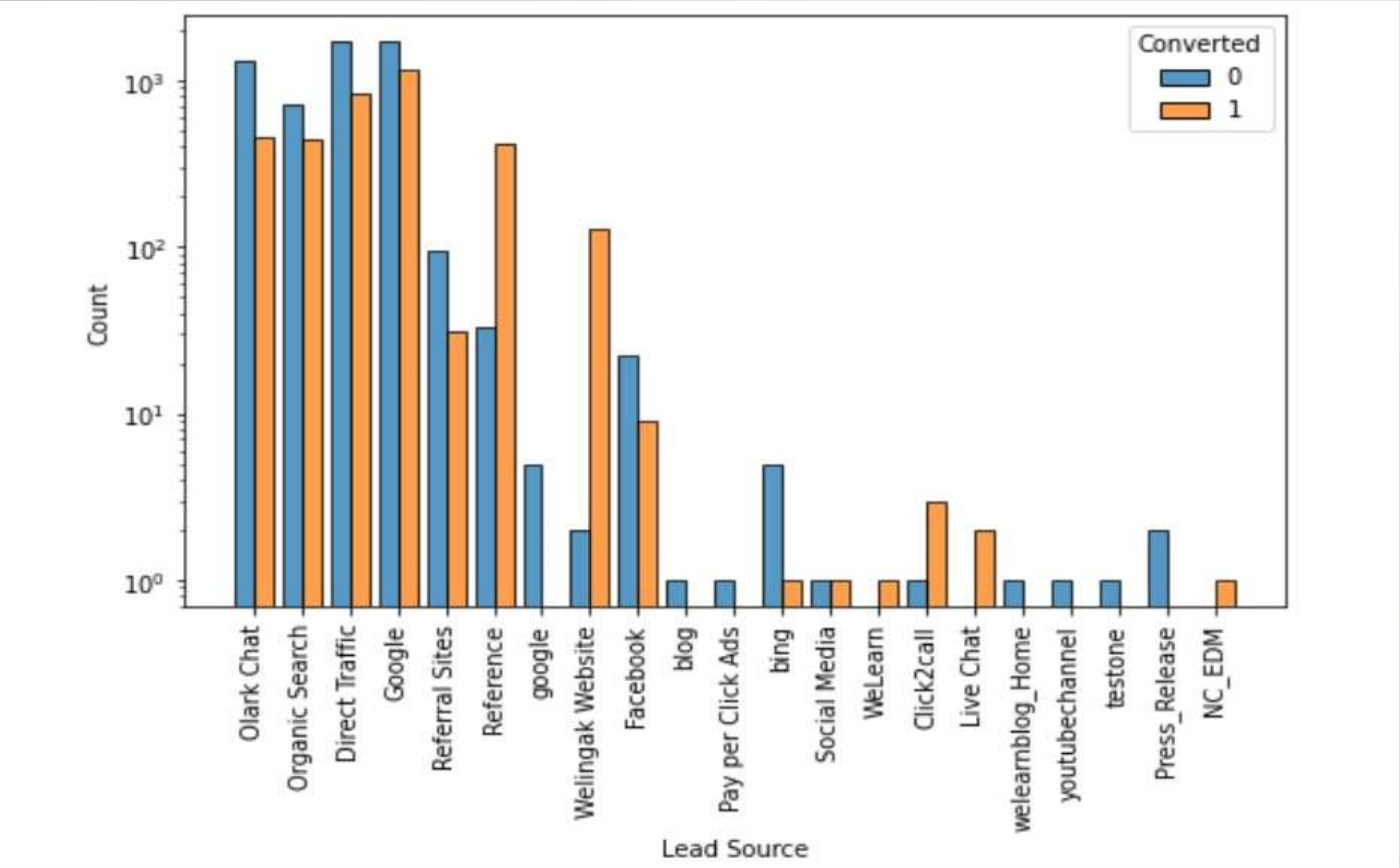
PROBLEM STATEMENT

- An education company named X Education sells online courses to industry professionals.
- The leads are acquired through various mediums and the sales team start making calls, writing emails, etc. to convert the leads.
- The current conversion rate is 30 %. This indicates that most of the leads are not getting converted.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'
- The CEO wants the conversion rate to increase to 80 %

APPROACH FOR ANALYSIS AND MODELLING.

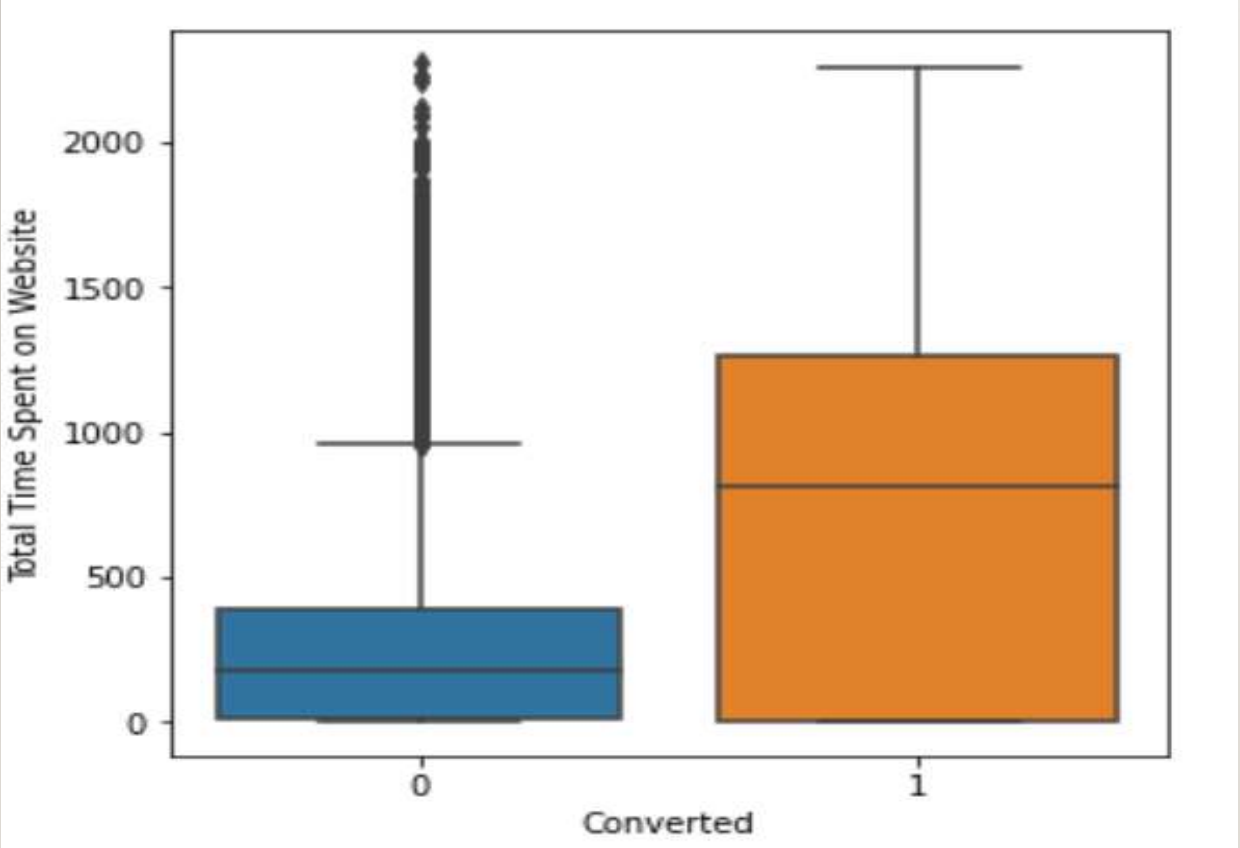
- Cleaning and Understanding the data
- Exploratory Data Analysis for finding out most useful variable for conversion
- Preparing the data for model building
- Build the logistic Regression model
- Test the model on train and test dataset
- Evaluate the model with different measures and matrices
- Interpret the model and its parameters

EDA- LEAD SOURCEVS CONVERTED



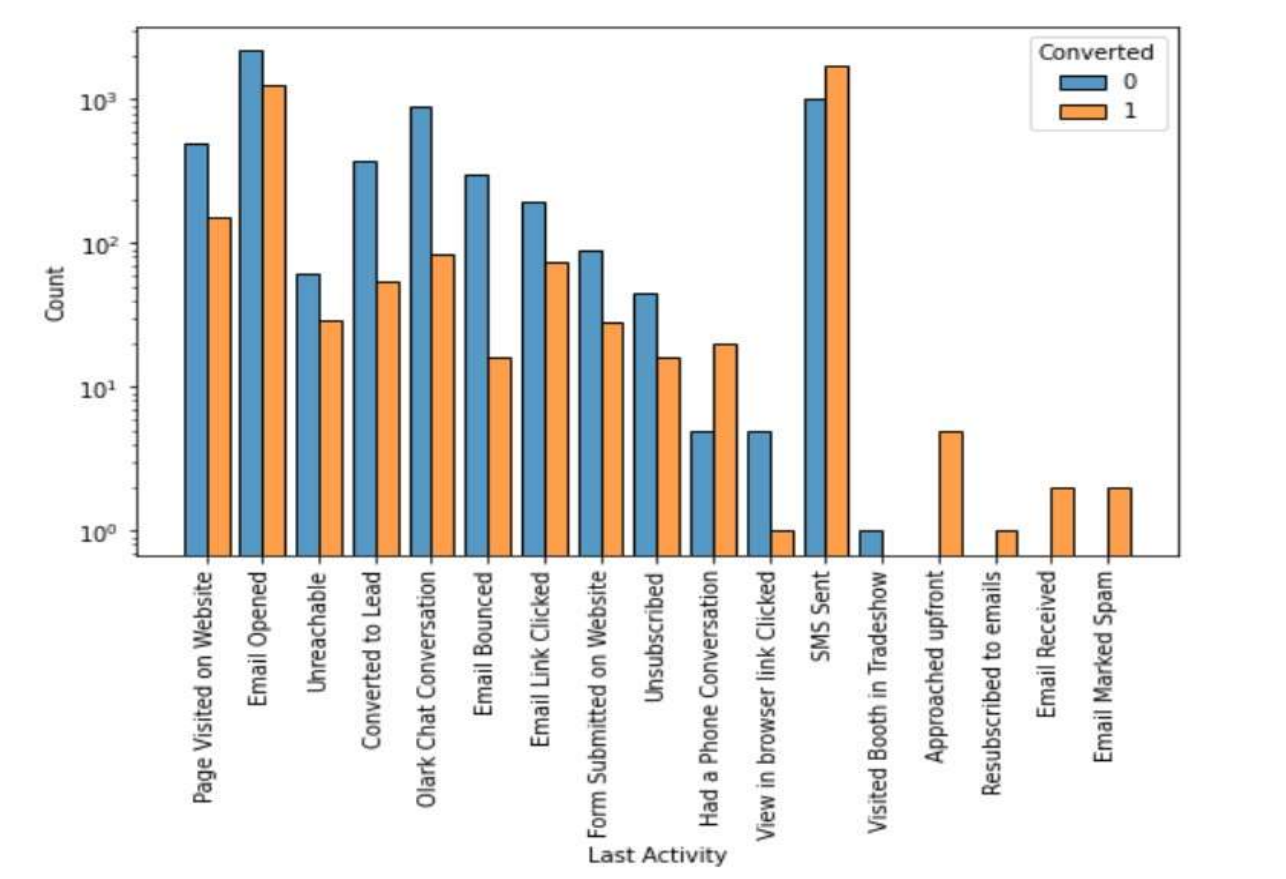
- In lead source, categories such as "Wellingak website" and "reference" have high higher conversion numbers.
- Also the absolute value of count is also considerable.
- "Click2call" and "live chat" also have higher conversion numbers, however the count value is very less.

EDA- TOTAL TIME SPENT ON WEBSITE VS CONVERTED



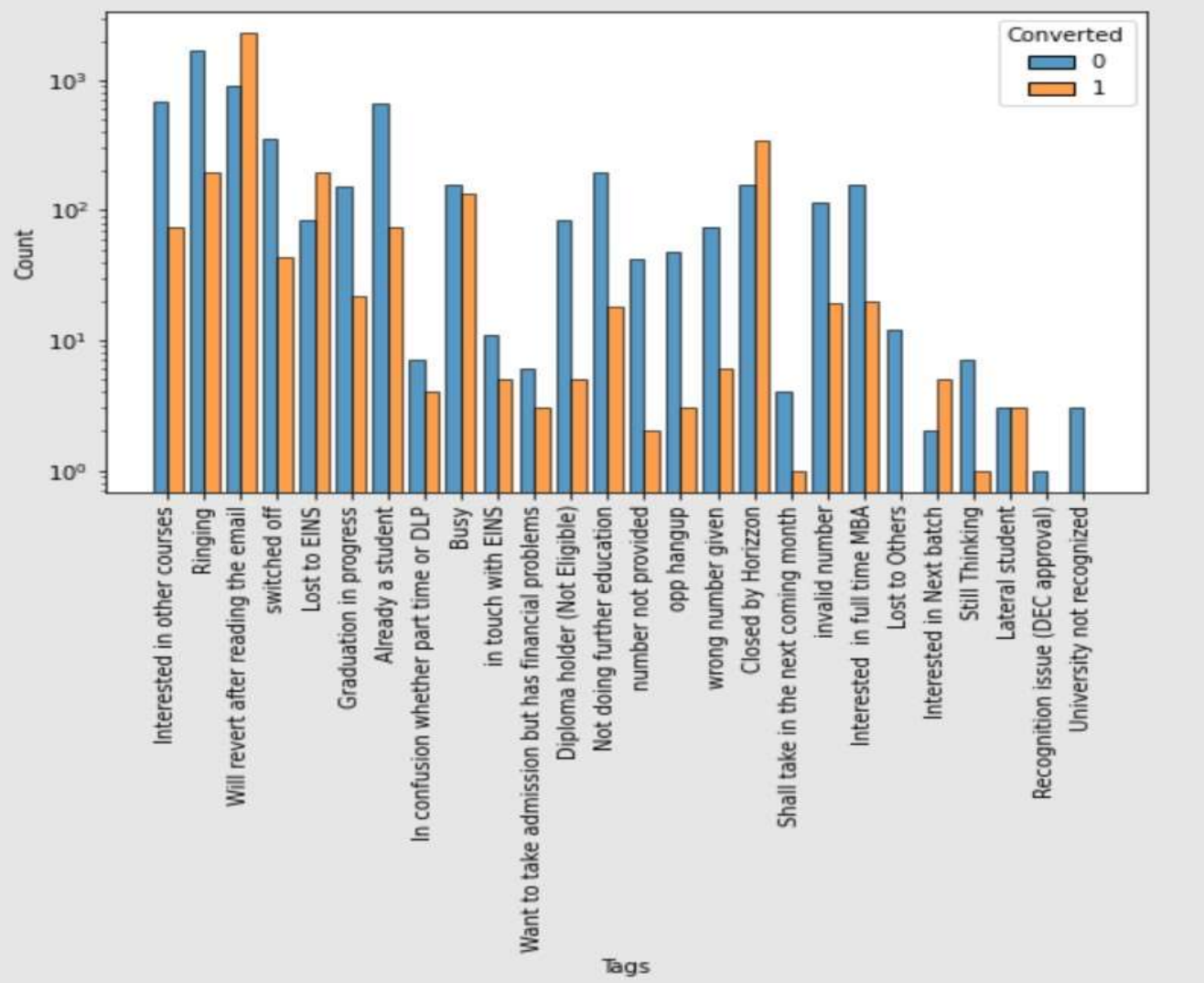
- The box plot show that leads which spent large amount of time on website can be converted.
- However there are outliers of non-converted portion as well.
- So the range of time spent need to be more carefully inspected.

EDA- LAST ACTIVITYVS CONVERTED



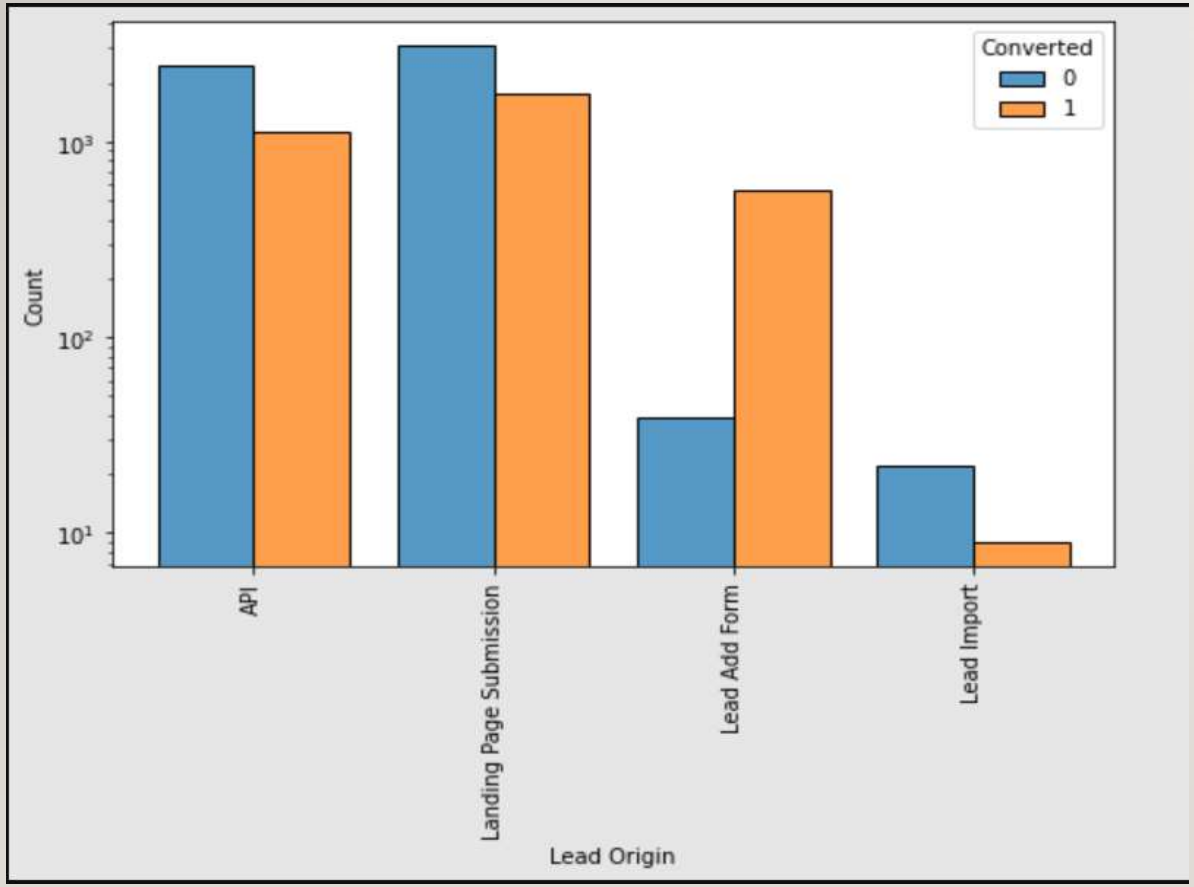
- The bar graph shows classes from "last activity" which were performed by customer
- We can see high conversion for categories such as "SMS sent" and "Had a phone conversation".
- There are categories such as "Approached upfront", "Email received", "Email marked spam", which also have good conversion rate, but the count of data points is very less.

EDA - TAGS VS CONVERTED



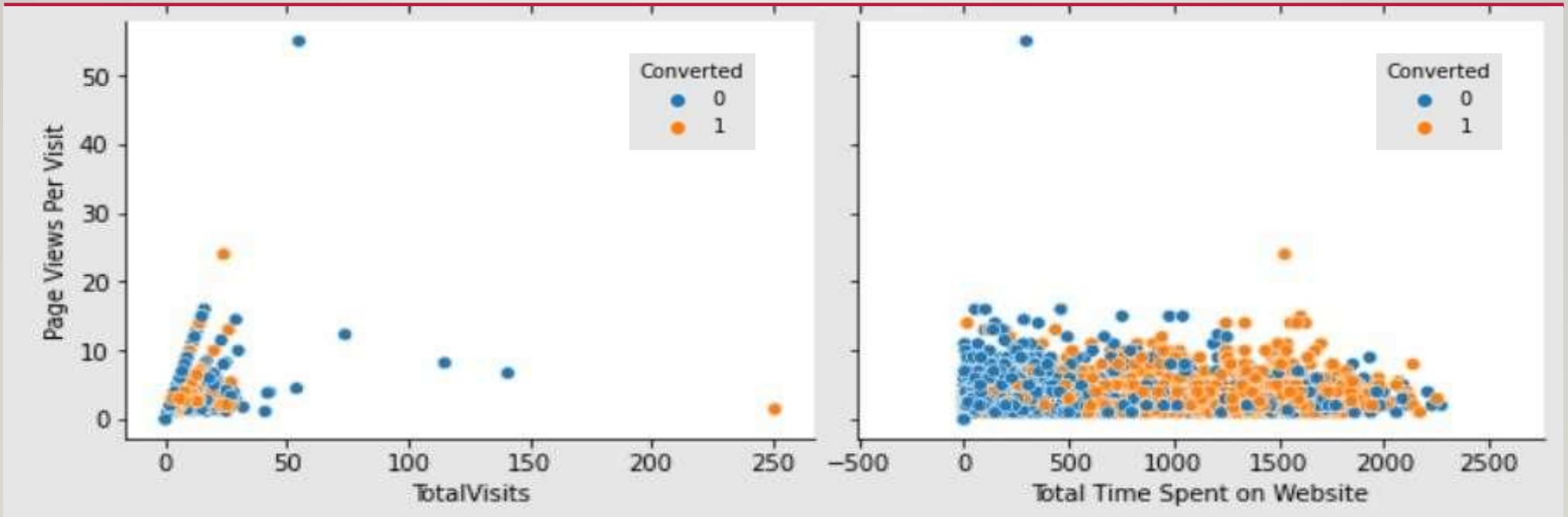
- The Tags categories such as "Will revert after reading the email", "Lost to EINS", "closed by horizzon" , and "busy" have high conversion rate.
- The count of this categories are also substantially high.

EDA – LEAD ORIGINVS CONVERTED



- The chart represents the origin of the lead.
- The leads which are generated through lead add form have much higher conversion rate as compared to other sources.
- The sales team can circulate more such forms to generate the leads.

EDA – PAIR-PLOT FOR PAGEVIEWED PERVISITVS TOTALVISITS, TOTALTIME SPENT ONWEBSITE,



- There is linear relationship between page views per visit and total visits. Hence we might not consider one of this features as significant.
- The total time spent on website vs page views per visit have not linear relationship.

DATA CONSIDERATION FOR MODEL BUILDING

	Var1	Var2	Correlation
907	Lead Source_Facebook	Lead Origin_Lead Import	0.967632
1706	Lead Source_Reference	Lead Origin_Lead Add Form	0.849425
6564	What is your current occupation_Working Profes...	What is your current occupation_Unemployed	0.843135
2900	Last Activity_Email Bounced	Do Not Email	0.616344
804	Lead Source_Direct Traffic	A free copy of Mastering The Interview	0.594715

- The Pearson correlation table states the top five highly correlated variables.
- These variable can prove redundant while creating the model.

MODEL PARAMETERS AND EVALUATION METRICS

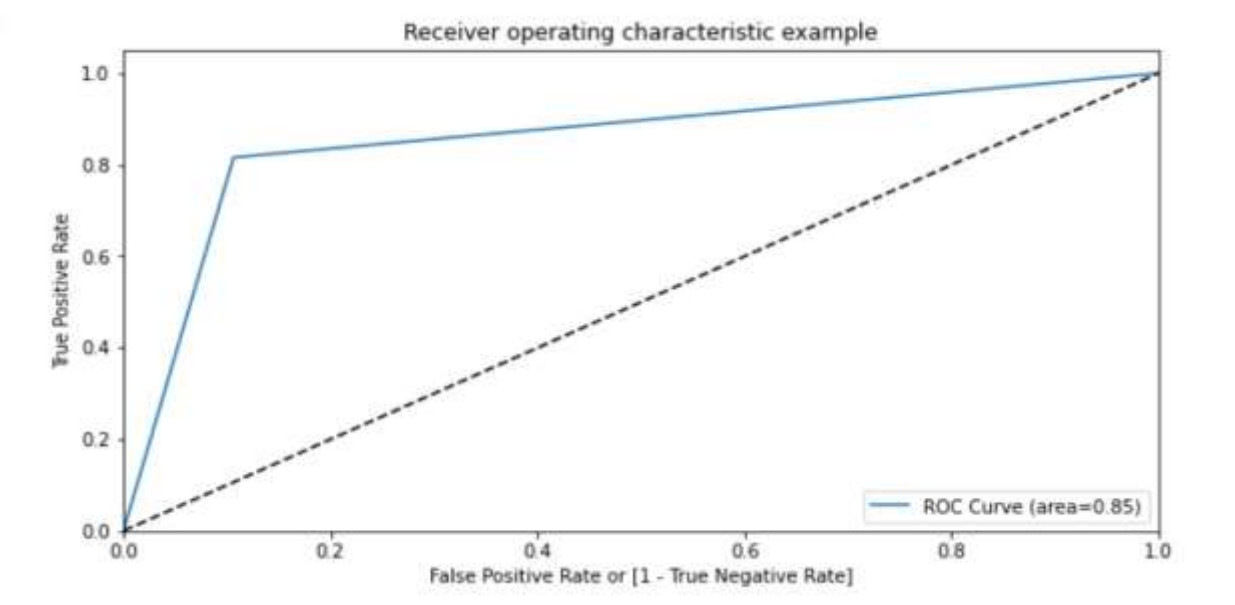
Feature	Importance
const	-2.790608
Do Not Email	-1.555664
Total Time Spent on Website	1.082816
Lead Origin_Lead Add Form	3.237448
Lead Source_Facebook	1.084463
Lead Source_Olark Chat	0.979384
Lead Source_Welingak Website	3.003525
Last Activity_Converted to Lead	-1.093272
Last Activity_Olark Chat Conversation	-1.633684
Last Activity_SMS Sent	1.428152
Last Activity_Unsubscribed	1.459553
Tags_Busy	1.599004
Tags_Closed by Horizzon	3.003928
Tags_Interested in Next batch	2.433024
Tags_Lost to EINS	3.547735
Tags_Will revert after reading the email	2.995554
Lead Quality_Worst	-0.965822

- The parameters and the importance of each significant parameter is generated through logistic regression.
- The model gave accuracy of 0.864 on training data set and 0.865 on test data set. In both the cases the

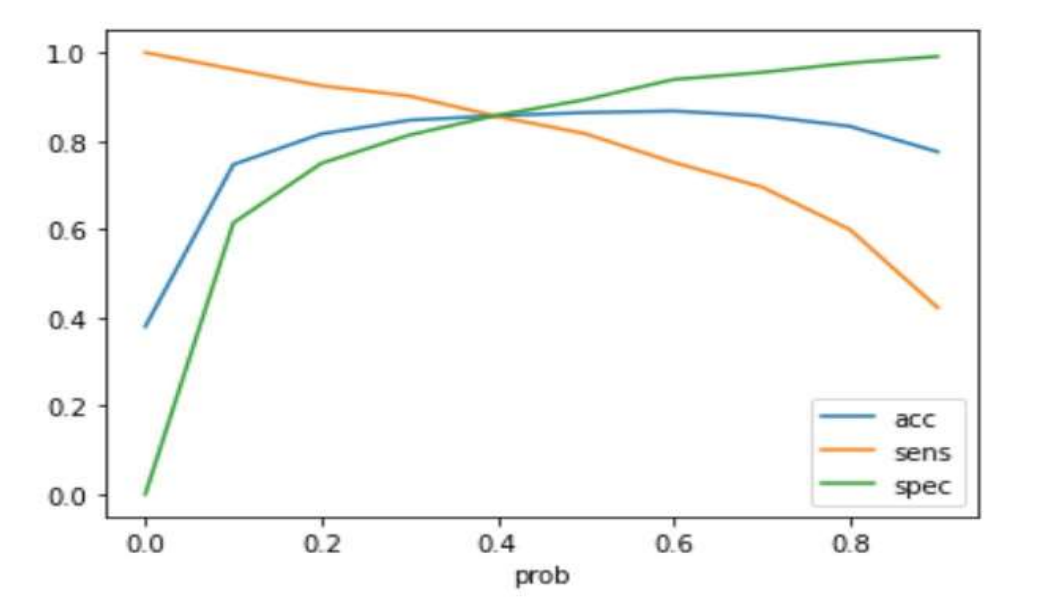
```
Sensitivity     = 0.831207065750736
Specificity     = 0.8860981308411215
Recall          = 0.831207065750736
Precision       = 0.8128598848368522
```

below

ROC CURVE AND OPTIMUM CUT-OFF



- The Receiver Operating Characteristics curve has 0.85 area under the curve.
- This indicates that the model is a good fit.



- For optimum cut-off curve, we discovered that the optimum cut-off is 0.5

RECOMMENDATIONS

Based on Exploratory Data Analysis and Linear Regression, following are the recommendation to improve the conversion rate

- A lead which is spending more than average amount of time on website can be a hot lead. The sales team can look out for more such leads.
- The leads sourced from Facebook, Olark Chat, and Welingak Website are highly convertible leads. The sales team can focus on improving number of leads from these sources.
- Sending SMS can have a bigger impact on the leads.
- The leads which are tagged as 'closed by horizon', 'Tags_Busy', 'interested in next batch', 'lost to EINS', and 'will revert after reading the email' can also prove hot leads. Prompt actions from sales team on such tagged leads can improve the flow of customers.

CONCLUSIONS

- The final model has Sensitivity of 0.831 on the test data set, this means the model is able to predict 83% customers out of all the converted customers, (Positive conversion) correctly.
- It has Precision of 0.81, this means 81% of predicted hot leads are True Hot Leads.
- The accuracy on training data set and test data set is almost similar, proving that the model is stable.
- We can go ahead with the final model and use it for improving the conversion rate of the leads for X Educations.