MACHINE LEARNING

LEAD SCORE LOGISTIC REGRESSION MODEL

Presented by:
Aayushi Singh and Abhijeet Singh

PROBLEM STATEMENT

X EDUCATION

X Education is an online education platform that sell courses to industry professionals. The company runs ads on various platforms that are linked to ad forms. When a person fills this form he is classified as a lead.

LEAD VS CONVERSION

Now, although X Education gets a lot of leads, its lead conversion rate is very poor - only 30%. To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.

GOAL

The company requires a model wherein we must assign a lead score to each of the leads. Lead score shall determine probability of conversion. The CEO wants the lead conversion rate to be around 80%.

HOW ARE WE ACHIEVING THE GOAL?

BUILD A LOGISTIC REGRESSION MODEL

A logistic regression model will help us prepare a model that can help assign scores to leads.

&

ASSIGN SCORES TO LEADS

A score to be assigned to each lead ranging from 0 to 100. A higher number will indicate higher chances of conversion and vice versa.

STEPS INVOLVED

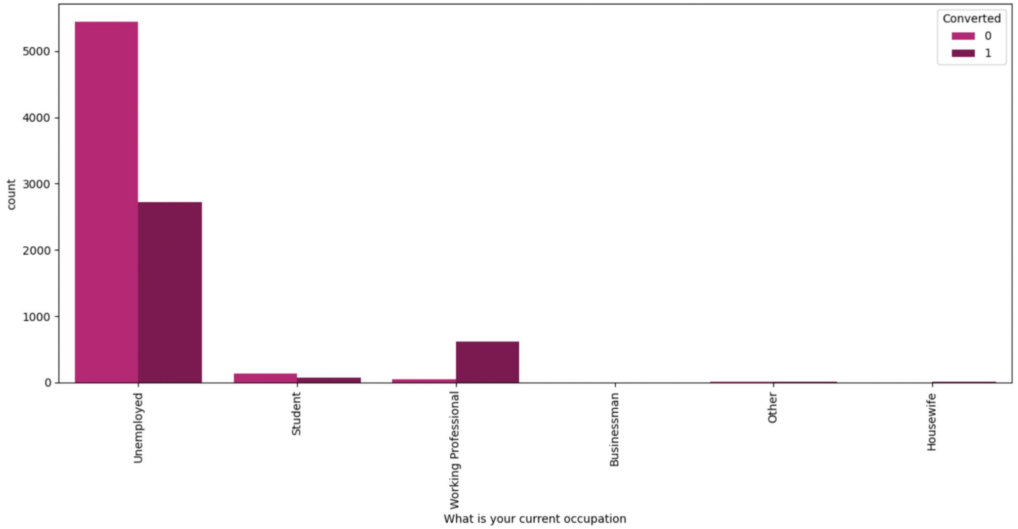
1. Importing **Data** and **Necessary Libraries**.
2. Clean and Prepare the data by **dropping and treating null values** in each column.
3. Perform Exploratory Data Analysis **(EDA)** with data visualisation, to understand and make initial readings on that.
4. Preparing Data for Regression Modelling by **converting binary variables**, **introducing dummy variables**, **splitting data** into train and test sets and **rescaling**.
5. Building the model using **RFE selection** and Manual selection using **GLM**.
6. Assigning **Lead Score** to each lead.
7. **Test the model on Train set**, **make predictions** and calculating **measure metrics** like: Confusion Matrix and Calculating Accuracy, Sensitivity, Specificity, False positive rate, Positive predictive value, Negative predictive value and ROC Curve.
8. **Test the model on Test set**, make predictions and calculating measure metrics like: Confusion Matrix and Calculating Accuracy, Sensitivity, Specificity, Precision and Recall and their Trade off.
9. Generate a list of **'hot leads'** on the basis of lead score.
10. Please find the workings of these steps in the ipython notebook link here.

EXPLORATORY DATA ANALYSIS

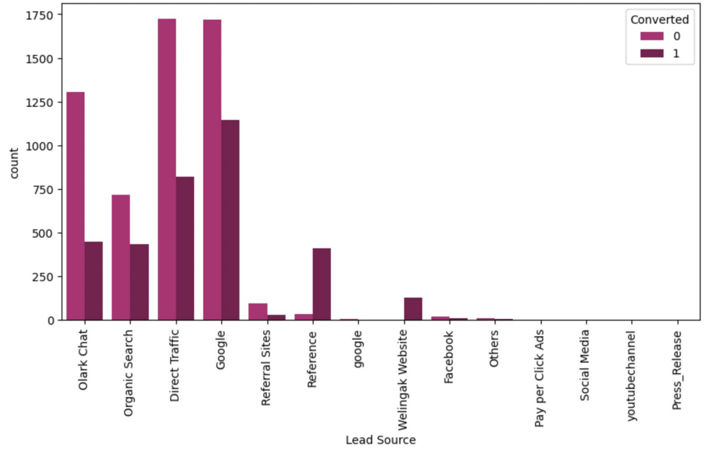
The following slides will contain screens of data visualisation from the Ipython Notebook.

Analysis: Working professionals are more likely to convert to customers i.e.e join course while Unemployed leads and students are also highly interested but approximately 50% only get converted into customers and users.

‘What is your occupation’ column



‘Lead Source’ column

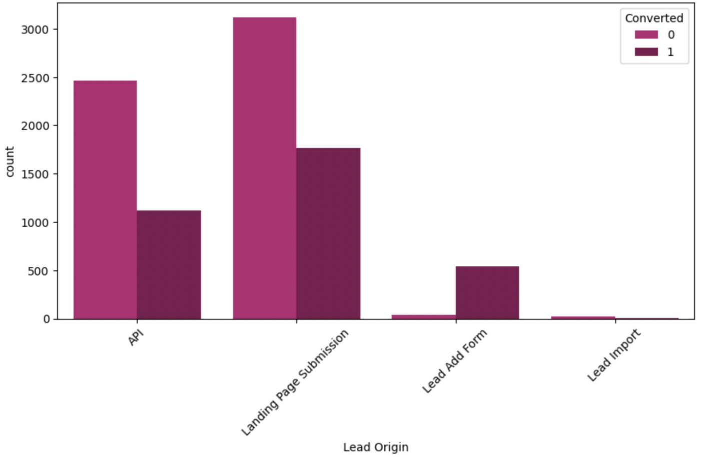


Analysis:

- Google and Direct traffic generates maximum number of leads.
- Conversions through reference leads is vero good.
- Leads through Wellingak website.

Analysis: Leads generation and Conversion ratio is good for API and Landing Page Submission. Conversion rate via Lead ad form is also quite good.

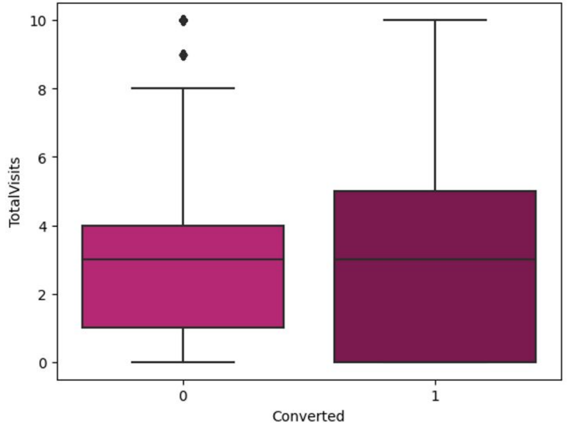
‘Lead Origin’ column



Analysis: Focus on people visiting and spending time on website.

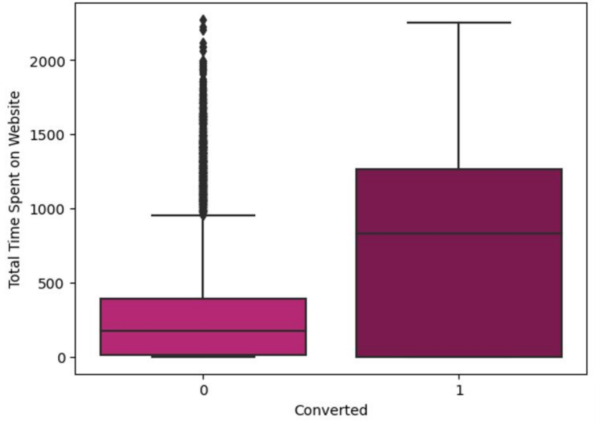
Focus should be made on the people who visit the website as the volume of conversion of people visiting the website is high.

‘Total Visits’ (Website) column



Also, it is clearly visible that the people who spend more time on website has a high conversion rate.

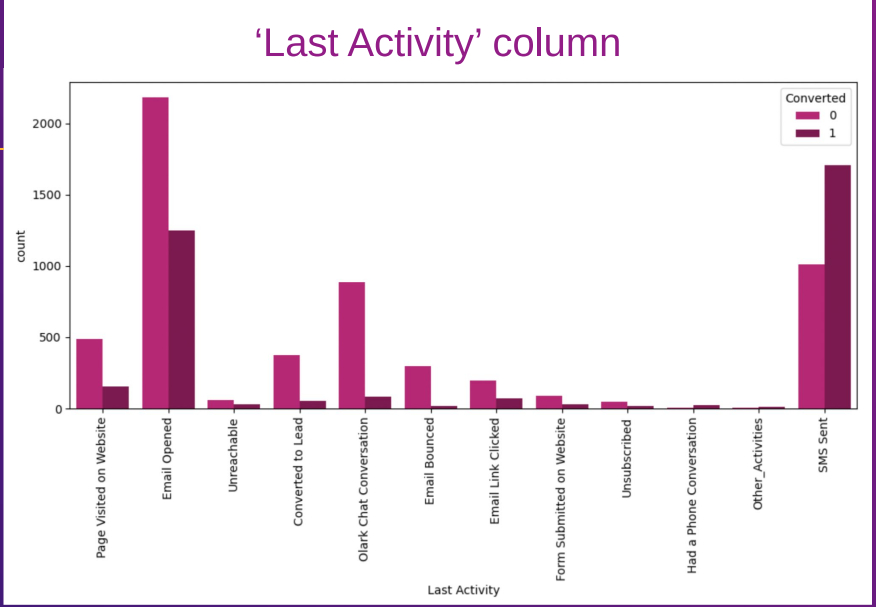
‘Total Time Spent on Website’ column



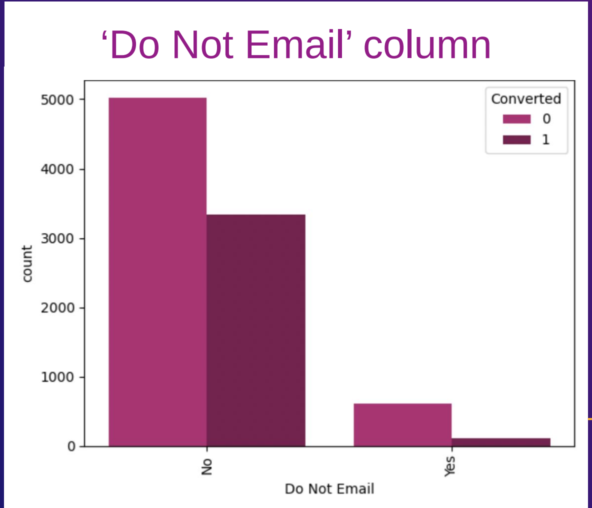
Analysis: Last Activity Matters…
High rate of conversion for

- People who opened the mail
- Had Olark conversation/ SMS conversation

‘Last Activity’ column



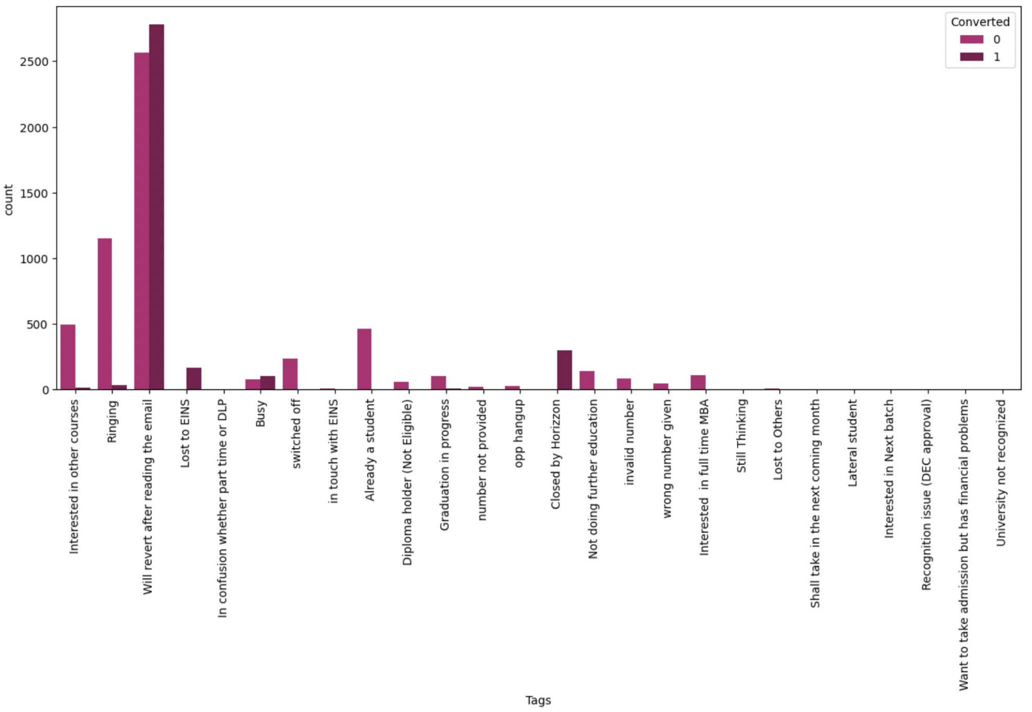
‘Do Not Email’ column



Analysis: There are high conversions via E-mailers. Especially for leads who opt for emailers.

EDA

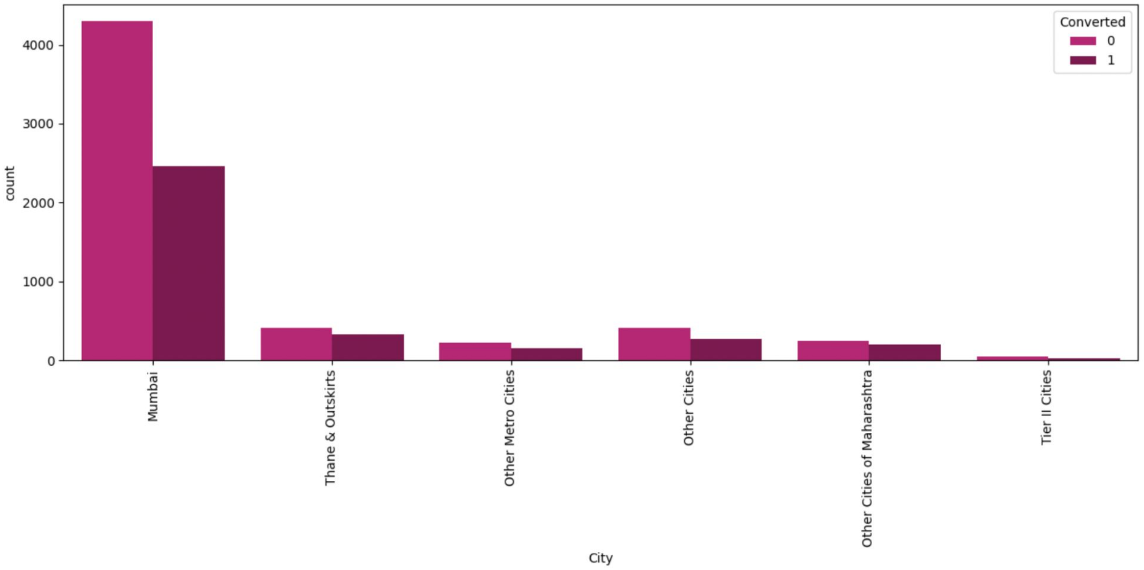
‘Tags’ column



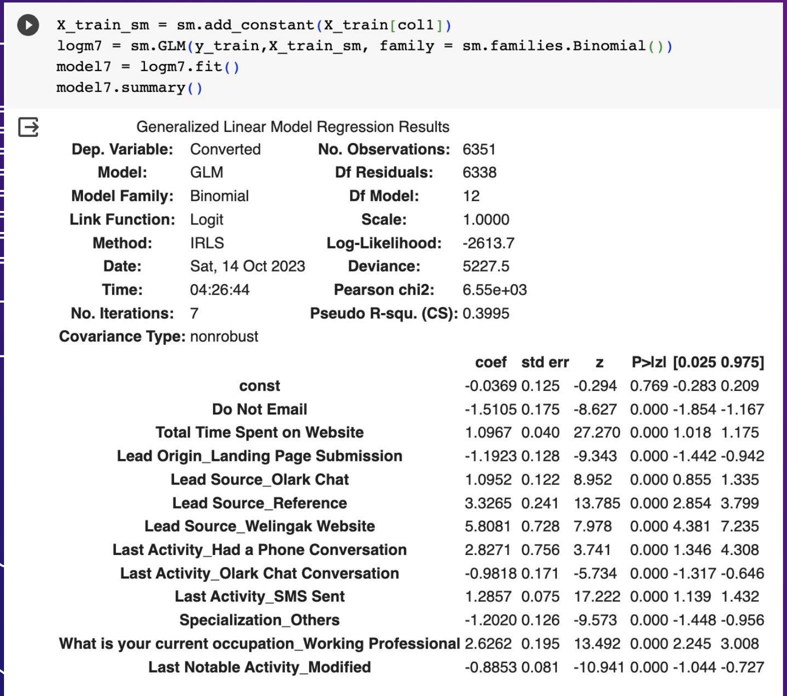
Analysis:

- Leads with the tag ‘Will revert after reading the email’ have extremely high conversion rates.
- Horizonn has closed all leads directed to them - direct conversions. There are no non-converted leads by Horizonn.
- Lost to EINS also has a very high conversion rate

‘City’ column



Analysis: Most leads generated and converted are from Mumbai.



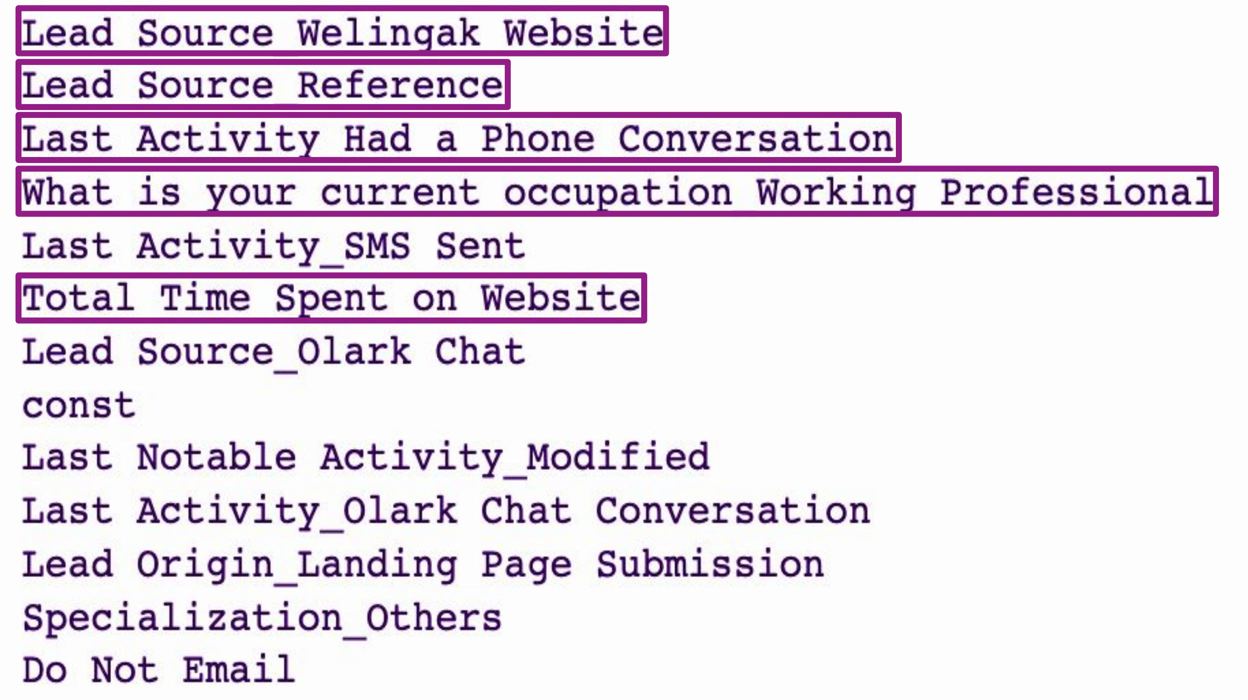
```
X_train_sm = sm.add_constant(X_train[col1])
logm7 = sm.GLM(y_train,X_train_sm, family = sm.families.Binomial())
model7 = logm7.fit()
model7.summary()
```

```
                Generalized Linear Model Regression Results
   Dep. Variable:  Converted          No. Observations:  6351
          Model:  GLM                  Df Residuals:     6338
   Model Family:  Binomial                 Df Model:     12
  Link Function:  Logit                       Scale:     1.0000
        Method:  IRLS                 Log-Likelihood:   -2613.7
          Date:  Sat, 14 Oct 2023           Deviance:    5227.5
          Time:  04:26:44               Pearson chi2:    6.55e+03
  No. Iterations:  7              Pseudo R-squ. (CS): 0.3995
Covariance Type: nonrobust
                                                     coef   std err   z      P>|z|  [0.025  0.975]
                       const                       -0.0369  0.125   -0.294   0.769  -0.283  0.209
                   Do Not Email                    -1.5105  0.175   -8.627   0.000  -1.854  -1.167
             Total Time Spent on Website            1.0967  0.040   27.270   0.000  1.018   1.175
         Lead Origin_Landing Page Submission       -1.1923  0.128   -9.343   0.000  -1.442  -0.942
               Lead Source_Olark Chat               1.0952  0.122   8.952    0.000  0.855   1.335
               Lead Source_Reference                3.3265  0.241   13.785   0.000  2.854   3.799
            Lead Source_Welingak Website            5.8081  0.728   7.978    0.000  4.381   7.235
         Last Activity_Had a Phone Conversation     2.8271  0.756   3.741    0.000  1.346   4.308
          Last Activity_Olark Chat Conversation    -0.9818  0.171   -5.734   0.000  -1.317  -0.646
                Last Activity_SMS Sent              1.2857  0.075   17.222   0.000  1.139   1.432
                Specialization_Others              -1.2020  0.126   -9.573   0.000  -1.448  -0.956
What is your current occupation_Working Professional 2.6262 0.195   13.492   0.000  2.245   3.008
             Last Notable Activity_Modified        -0.8853  0.081   -10.941  0.000  -1.044  -0.727
```

OUR FINAL MODEL

Our final Logistic Regression model is Model #7 with P Values for all variables as 0.000 and all VIF values below 2.5. There are a total of 12 variables in this model.

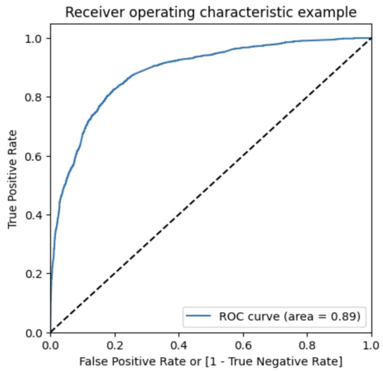
THE FINAL VARIABLES V/S VARIABLES IMPACTING CONVERSION RATES:



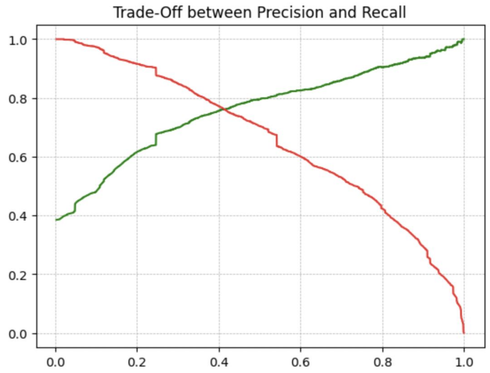
```
Lead Source Welingak Website
Lead Source Reference
Last Activity Had a Phone Conversation
What is your current occupation Working Professional
Last Activity_SMS Sent
Total Time Spent on Website
Lead Source_Olark Chat
const
Last Notable Activity_Modified
Last Activity_Olark Chat Conversation
Lead Origin_Landing Page Submission
Specialization_Others
Do Not Email
```

ROC Curve area = 90%

ROC Curve	Precision - Recall Trade-Off



Reading: ROC Curve area is almost 90%, this signifies our model is good model.



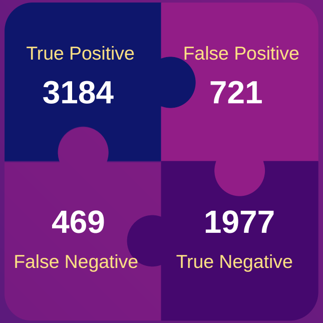
The Trade-Off between Precision and Recall is 0.42.

The Trade-Off between Precision and Recall is 0.42. Therefore we can safely say that a probability of 42% is good enough to be considered as a hot lead.

MEASUREMENT METRICS

TRAIN SET

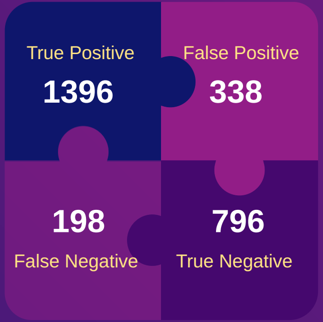
Accuracy: **81.26%**
Specificity: **80.26%**
Sensitivity: **81.54%**
Precision - 79.55% ; Recall - 70.44%



True Positive 3184	False Positive 721
469 False Negative	1977 True Negative

TEST SET

Accuracy: **80.5%**
Specificity: **80.5%**
Sensitivity: **80.5%**



True Positive 1396	False Positive 338
198 False Negative	796 True Negative

HAS 80% CONVERSION (THE CEO'S BALLPARK) BEEN ACHIEVED?

On the basis of lead_score assignment, the leads that earlier did not have conversion status as per available history, could also be guided properly and appropriate actions can now be taken on the basis of the assigned lead scores.
Sensitivity is a Metric used to show that the conversion is achieved on the data.

Sensitivity of the data on test data is 80.5%

CONCLUSIONS AND SUGGESTIONS:

- Leads with **Lead score above 42** can be classified as a Hot Lead.
- A list of IDs of **1,024 Hot Leads** has been generated.
- All calling, communication and marketing actions must be directed towards **hot leads**.
  Further bucketing of scores can be done in order to further augment audience and divide bandwidth accordingly. For instance,
  - 42-50: 4th Priority Hot Leads
  - 51-70: 3rd Priority Hot Leads
  - 71-85: 2nd Priority Hot Leads
  - 85-100: 1st priority Hot Leads
- On the basis of model, efforts **must** be dedicated towards leads:
  - That are Working Professionals
  - Coming from Welingak Website
  - Coming via References
  - Last activity was conversation over call, SMS, Email and Olark Chat.

THANK YOU !