LEAD SCORING CASE STUDY

Varun Patel
Harsha Belurkar

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 and fill the form if interested with their details. The company also gets leads through past referrals. 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%. We need to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers.

Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. 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.

- Data sourcing and Cleaning
- EDA and Outlier Treatment
- Data Preparation
- Model Building
- Model Evaluation
- Final Inferences

SOLUTION APPROACH

Build a model where in you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance and Logistic Regression fits the purpose

Data Sourcing

- Dataset used : Leads.csv
- We have imported necessary libraries required
- Inspected dataset using Shape, Info, Describe and head functions
- Shape of the data frame : (9240, 37)
- Few columns have "Select" as their values. This gives us the information that either the user haven't selected any option or chose not to select. So these values gives us the same information as the "null" values. Hence, we are converting Select into NULL values
- Total no of columns with missing values > 45% : 7
- Dropped the Columns which are populated less than %, as these might not be helpful in giving the right understanding of the data.
- Retained only the important columns that are found to be relevant for the analysis.

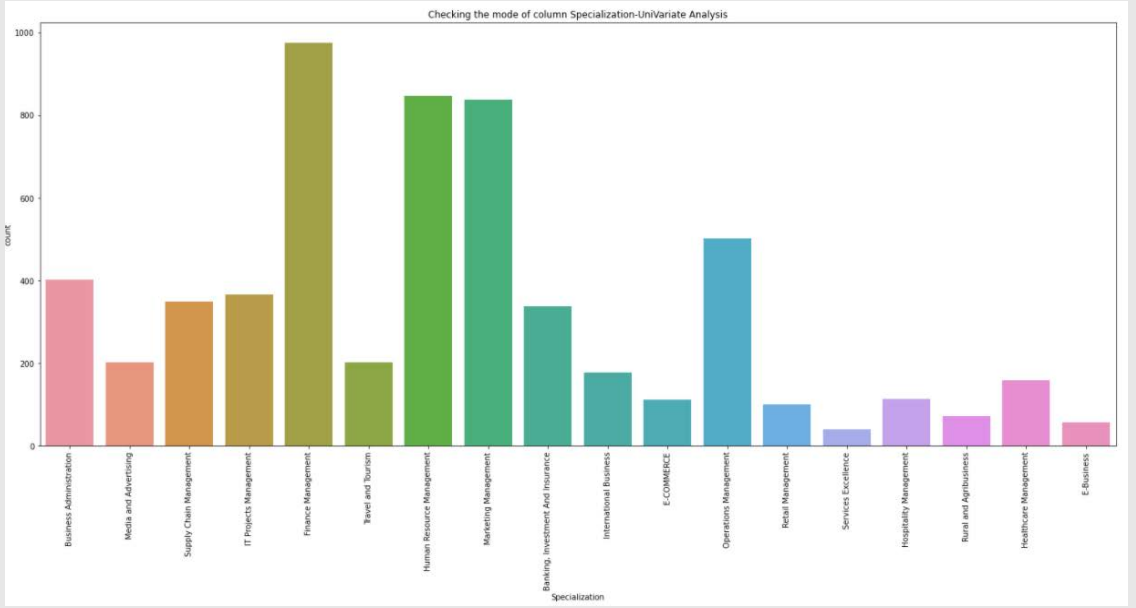
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9240 entries, 0 to 9239
Data columns (total 37 columns):
 #   Column                                         Non-Null Count  Dtype  
---  ------                                         --------------  -----  
 0   Prospect ID                                    9240 non-null   object 
 1   Lead Number                                    9240 non-null   int64  
 2   Lead Origin                                    9240 non-null   object 
 3   Lead Source                                    9204 non-null   object 
 4   Do Not Email                                   9240 non-null   object 
 5   Do Not Call                                    9240 non-null   object 
 6   Converted                                      9240 non-null   int64  
 7   TotalVisits                                    9103 non-null   float64
 8   Total Time Spent on Website                    9240 non-null   int64  
 9   Page Views Per Visit                           9103 non-null   float64
 10  Last Activity                                  9137 non-null   object 
 11  Country                                        6779 non-null   object 
 12  Specialization                                 7802 non-null   object 
 13  How did you hear about X Education             7033 non-null   object 
 14  What is your current occupation                6550 non-null   object 
 15  What matters most to you in choosing a course  6531 non-null   object 
 16  Search                                         9240 non-null   object 
 17  Magazine                                       9240 non-null   object 
 18  Newspaper Article                              9240 non-null   object 
 19  X Education Forums                             9240 non-null   object 
 20  Newspaper                                      9240 non-null   object 
 21  Digital Advertisement                          9240 non-null   object 
 22  Through Recommendations                        9240 non-null   object 
 23  Receive More Updates About Our Courses         9240 non-null   object 
 24  Tags                                           5887 non-null   object 
 25  Lead Quality                                   4473 non-null   object 
 26  Update me on Supply Chain Content              9240 non-null   object 
 27  Get updates on DM Content                      9240 non-null   object 
 28  Lead Profile                                   6531 non-null   object 
 29  City                                           7820 non-null   object 
 30  Asymmetrique Activity Index                    5022 non-null   object 
 31  Asymmetrique Profile Index                     5022 non-null   object 
 32  Asymmetrique Activity Score                    5022 non-null   float64
 33  Asymmetrique Profile Score                     5022 non-null   float64
 34  I agree to pay the amount through cheque       9240 non-null   object 
 35  A free copy of Mastering The Interview         9240 non-null   object 
 36  Last Notable Activity                          9240 non-null   object 
dtypes: float64(4), int64(3), object(30)
memory usage: 2.6+ MB
```

Data Cleaning

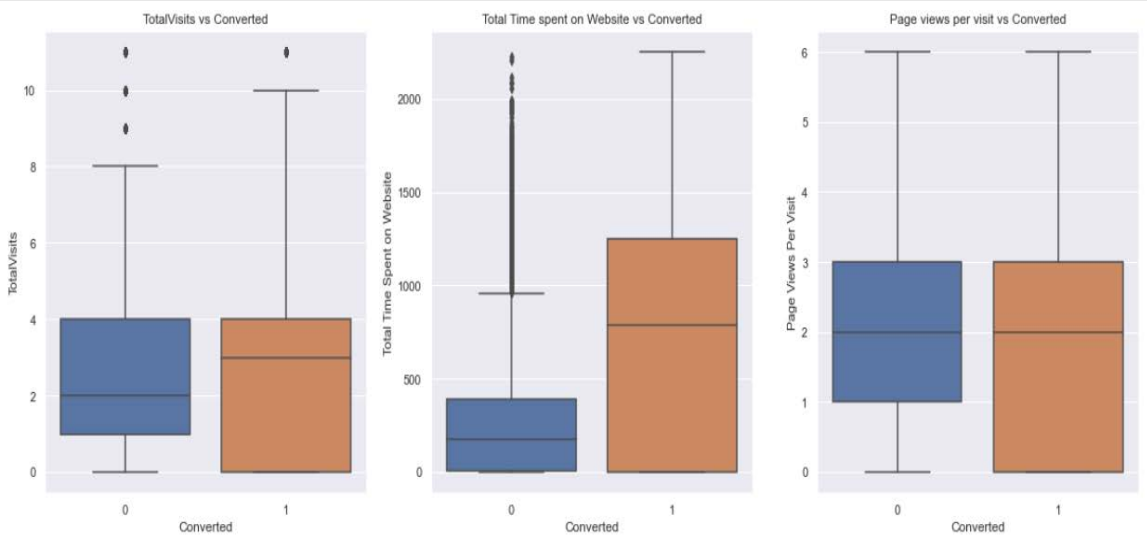
- Dropping the score variable as it is generated by sales team and variables having only 1 unique value as it does not give much insights for our model
- Binary Mapping : converting the values Yes/No to O and 1
- Data Imputation : for columns <45% missing values
- -- City, What matters most to you in choosing a course, What is your current occupation and Country : Imputed with Mode
- -- Specialization : As the specialization is distributed across multiple values ,we can't impute it with mode and therefore we can put the Nan values as Unknown category
- -- Last 4 columns < 2 % missing values and are hence dropped
- Merging Categorical values : Identifying the categorical columns with many categories and merging two or more categories of a column into a category named "Others"

	Total_missing_values	Percent_missing_values
City	3669.0	39.707792
Specialization	3380.0	36.580087
What matters most to you in choosing a course	2709.0	29.318182
What is your current occupation	2690.0	29.112554
Country	2461.0	26.634199
TotalVisits	137.0	1.482684
Page Views Per Visit	137.0	1.482684
Last Activity	103.0	1.114719
Lead Source	36.0	0.389610



EDA and Outlier Treatment

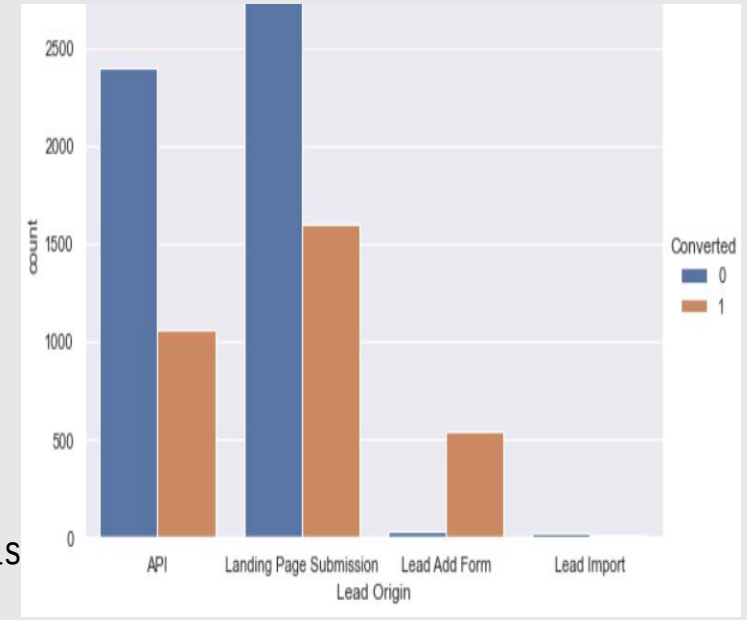
- We found that all numerical columns – Total Visits , Total time spent on website and page views per visit had outliers
- The outlier count was around 4% or less and choose the drop them
- Out of the 3 numerical variables TotalVisits and Page_Views_Per_visit doesnt show much effect on conversion rate, but Total_Time_Spent_on_Website seems to have an impact on conversion rate with more time spend leading to likelier conversion.

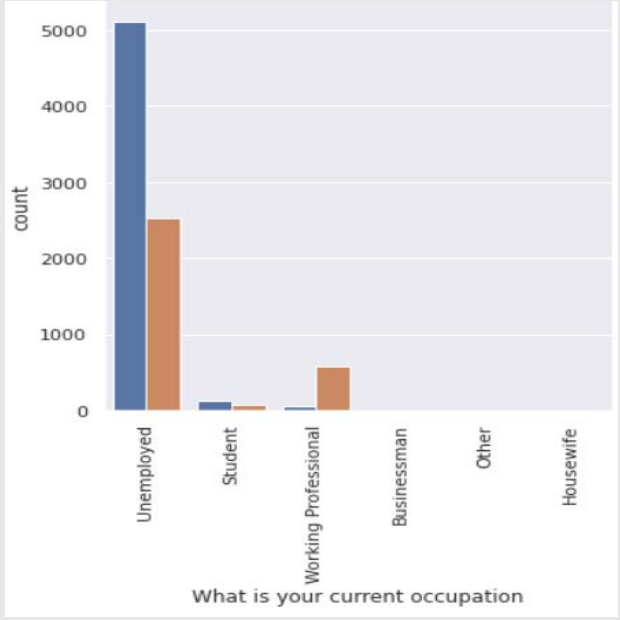


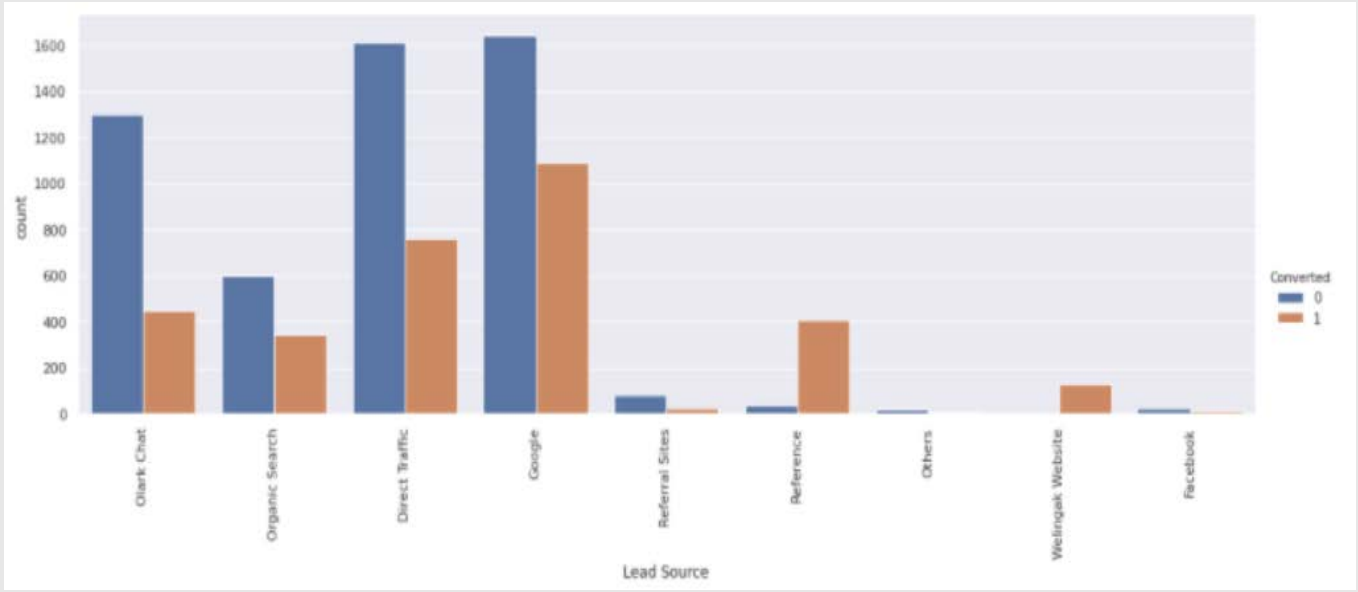

```
# Defining outlier treatment function
def outlier_treatment(datacolumn):
 sorted(datacolumn)
 Q1,Q3 = np.percentile(datacolumn , [25,75])
 IQR = Q3 - Q1
 lower_range = Q1 - (1.5 * IQR)
 upper_range = Q3 + (1.5 * IQR)
 return lower_range,upper_range

#Calculating IQR
lowerbound,upperbound = outlier_treatment(df_cleaned.TotalVisits)
print(lowerbound,upperbound)
```

EDA

- We would need to analyze the distribution of various values of all the above considered fields against the "Converted" value .
- Intent is to check how the conversion column is dependent on the variable
- Bivariate Analysis Inferences for top variables :
- Lead Origin : API and Landing page submission has high non converted count whereas lead ad form has high converted value
- Lead Source : All the Categories have high non converted count except references and welingak website have high converted count
- What is your current occupation : The conversion rate is very high for working professionals, but count is less. So this course should be advertised more to working professionals to increase the count.
- Last Notable Activity : The conversion rate is high for 'SMS Sent'.







Data Preparation

Created Dummy variables for categorical variables :

- 'Lead Origin', 'Lead Source', 'Specialization', 'What is your current occupation', 'City', 'Last Notable Activity' and dropped the original variables after concatenation

```
# Creating dummies
dummy = pd.get_dummies(df_cleaned[['Lead Origin','Lead Source','Specialization','What is your current occupation','City'
                                   'Last Notable Activity']], drop_first=True)
dummy.head()
```

	Lead Origin_Landing Page Submission	Lead Origin_Lead Add Form	Lead Origin_Lead Import	Lead Source_Facebook	Lead Source_Google	Lead Source_Olark Chat	Lead Source_Organic Search	Lead Source_Others	Lead Source_Reference	S
0	0	0	0	0	0	1	0	0	0	
1	0	0	0	0	0	0	1	0	0	
2	1	0	0	0	0	0	0	0	0	
3	1	0	0	0	0	0	0	0	0	
4	1	0	0	0	1	0	0	0	0	

Test Train Split has been divided in ratio 70 : 30 :

- Converted is response variable and remaining are feature variables

Feature Scaling :

- We have used Standard Scaler to standardize the independent features present in the data

The Present Lead conversion rate is at 37.66 %

Model Building : Feature selection using RFE

- Among all columns the top 20 columns useful for logistic regression
- Here is the snippet of code used and columns

```
In [69]: logreg = LogisticRegression()
         from sklearn.feature_selection import RFE
         rfe = RFE(logreg, 20)             # running RFE with 20 variables as output
         rfe = rfe.fit(X_train, y_train)

In [70]: rfe.support_

Out[70]: array([ True, False,  True,  True,  True,  True, False, False,  True,
                False, False,  True, False,  True, False, False, False, False,
                False,  True, False, False, False, False, False, False, False,
                False, False, False, False,  True,  True, False,  True,  True,
                 True, False, False, False, False, False, False, False,  True,
                 True,  True, False, False,  True,  True,  True])
```

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.1224	0.669	0.183	0.855	-1.189	1.434
Do Not Email	-1.4452	0.196	-7.362	0.000	-1.830	-1.060
Total Time Spent on Website	1.0550	0.041	25.520	0.000	0.974	1.136
Lead Origin_Landing Page Submission	-1.0118	0.132	-7.658	0.000	-1.271	-0.753
Lead Origin_Lead Add Form	1.6034	0.917	1.748	0.080	-0.194	3.401
Lead Origin_Lead Import	0.8657	0.571	1.517	0.129	-0.253	1.984
Lead Source_Olark Chat	0.9369	0.123	7.624	0.000	0.696	1.178
Lead Source_Reference	1.9094	0.939	2.034	0.042	0.069	3.750
Lead Source_Welingak Website	5.1460	1.366	3.767	0.000	2.469	7.823
Specialization_Hospitality Management	-0.8737	0.341	-2.560	0.010	-1.543	-0.205
Specialization_Unknown	-1.1683	0.125	-9.309	0.000	-1.414	-0.922
What is your current occupation_Housewife	23.5581	2.21e+04	0.001	0.999	-4.33e+04	4.33e+04
What is your current occupation_Student	-0.1594	0.699	-0.228	0.820	-1.530	1.211
What is your current occupation_Unemployed	-0.4261	0.663	-0.643	0.520	-1.725	0.873
What is your current occupation_Working Professional	2.1757	0.690	3.152	0.002	0.823	3.529
Last Notable Activity_Had a Phone Conversation	22.9679	2.31e+04	0.001	0.999	-4.52e+04	4.52e+04
Last Notable Activity_Modified	-0.5928	0.086	-6.868	0.000	-0.762	-0.424
Last Notable Activity_Olark Chat Conversation	-1.1803	0.326	-3.619	0.000	-1.819	-0.541
Last Notable Activity_SMS Sent	1.4894	0.089	16.679	0.000	1.314	1.664
Last Notable Activity_Unreachable	1.8687	0.612	3.054	0.002	0.670	3.068
Last Notable Activity_Unsubscribed	1.2155	0.529	2.299	0.021	0.179	2.252

Model Building : Logistic regression model

- Logistic regression model is built using GLM() function under stats model library
- Using manual elimination, we have dropped the Columns till we attained significant P value near to zero and VIF below 2.5

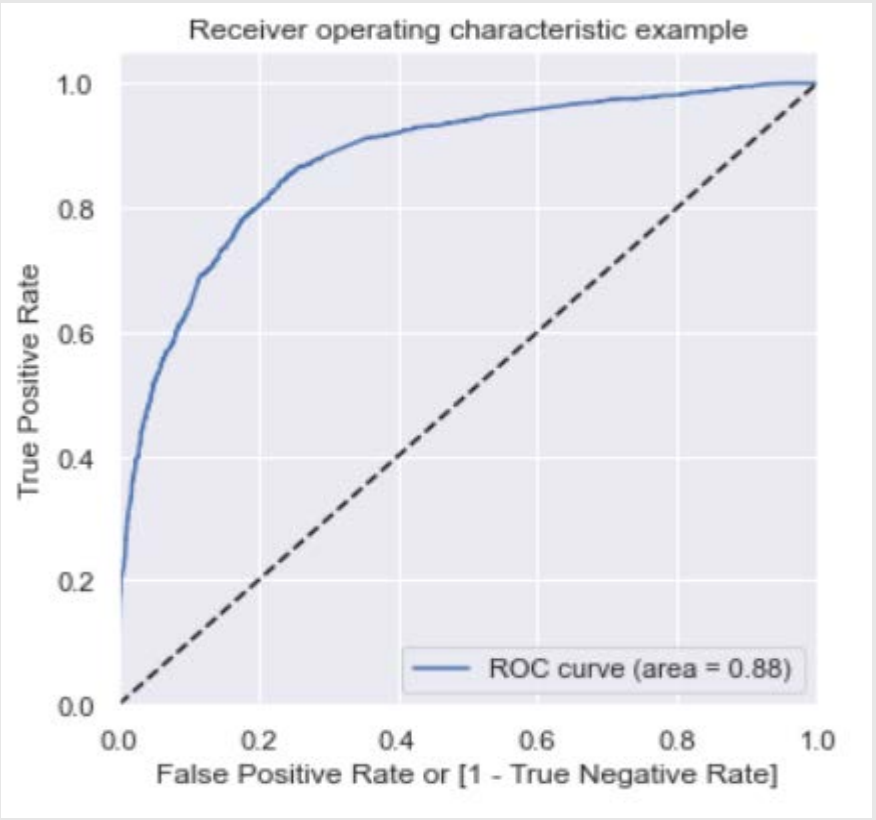
	Features	VIF
7	Specialization_Unknown	2.27
3	Lead Source_Olark Chat	2.00
9	Last Notable Activity_Modified	1.94
2	Lead Origin_Landing Page Submission	1.80
11	Last Notable Activity_SMS Sent	1.61
1	Total Time Spent on Website	1.30
4	Lead Source_Reference	1.23
0	Do Not Email	1.20
8	What is your current occupation_Working Profes...	1.19
5	Lead Source_Welingak Website	1.09
10	Last Notable Activity_Olark Chat Conversation	1.09
13	Last Notable Activity_Unsubscribed	1.08
6	Specialization_Hospitality Management	1.02
12	Last Notable Activity_Unreachable	1.01

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.2186	0.130	-1.678	0.093	-0.474	0.037
Do Not Email	-1.4530	0.196	-7.405	0.000	-1.838	-1.068
Total Time Spent on Website	1.0487	0.041	25.597	0.000	0.968	1.129
Lead Origin_Landing Page Submission	-1.0615	0.130	-8.155	0.000	-1.317	-0.806
Lead Source_Olark Chat	0.9090	0.122	7.481	0.000	0.671	1.147
Lead Source_Reference	3.4845	0.246	14.153	0.000	3.002	3.967
Lead Source_Welingak Website	6.7171	1.018	6.598	0.000	4.722	8.712
Specialization_Hospitality Management	-0.8907	0.343	-2.600	0.009	-1.562	-0.219
Specialization_Unknown	-1.2137	0.125	-9.729	0.000	-1.458	-0.969
What is your current occupation_Working Professional	2.5886	0.201	12.862	0.000	2.194	2.983
Last Notable Activity_Modified	-0.6068	0.086	-7.063	0.000	-0.775	-0.438
Last Notable Activity_Olark Chat Conversation	-1.1934	0.326	-3.663	0.000	-1.832	-0.555
Last Notable Activity_SMS Sent	1.4625	0.089	16.467	0.000	1.288	1.637
Last Notable Activity_Unreachable	1.8390	0.610	3.015	0.003	0.644	3.034
Last Notable Activity_Unsubscribed	1.1892	0.528	2.252	0.024	0.154	2.224

Model Evaluation

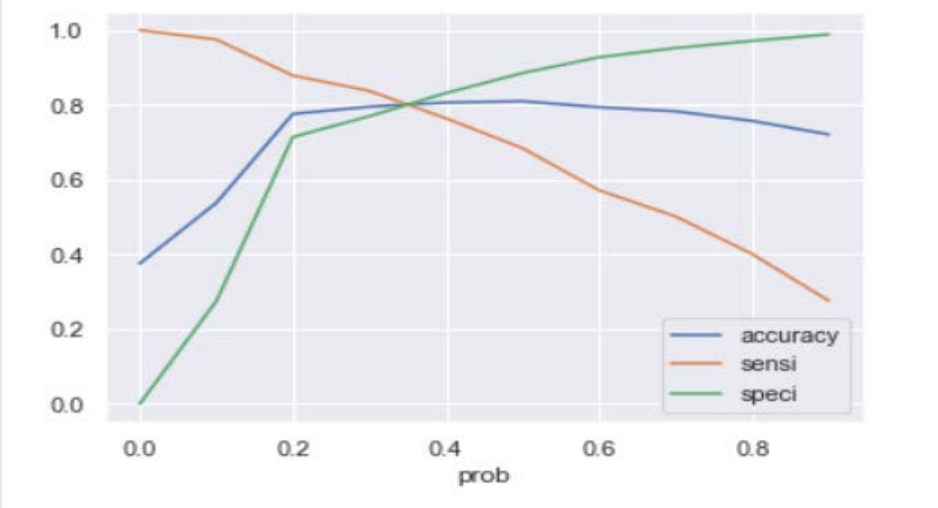
- In general we assume that lead is converted(1) if conversion probability is greater than 0.5
- Creating new column 'predicted' with 1 if Conversion Probability > 0.5 else 0
- Computing ROC curve and Optimal cutoff point as our assumption may or may not be right
- ROC curve : It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test. The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.
- ROC curve of this model is 0.88

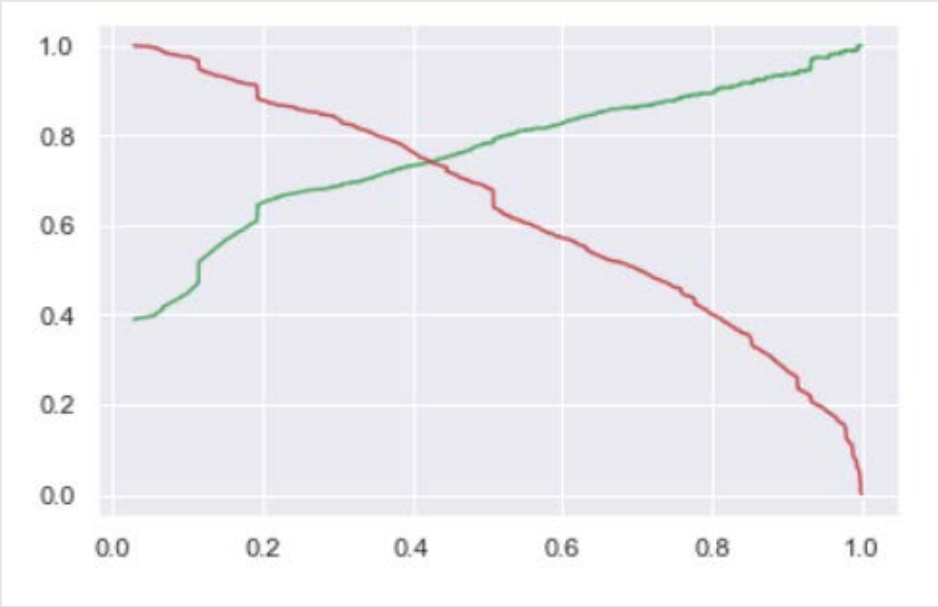
	Converted	Converted_Prob	Prospect ID	predicted
0	0	0.067705	6227	0
1	0	0.127978	6322	0
2	0	0.059202	3644	0
3	0	0.086698	3011	0
4	1	0.914758	8267	1



Model Evaluation

- Finding Optimal cutoff point : Optimal cutoff probability is that probability where we get balanced sensitivity and specificity
- From the curve above, 0.35 is the optimum point to take it as a cutoff probability.
- We want to predict the Lead score between 0 to 100. So multiplying the probability with 100
- We have predicted the lead score for both train and test data using the same optimal point
- The second graph here represents the precision and recall trade off
- Using confusion matrix we have calculated the various evaluation metrics





ModelEvaluation

Train data set

- Accuracy: 80%
- *Sensitivity/Recall*: 80%
- *Specificity*: 80%
- Precision: 70%

Test data set

- Accuracy: 82%
- *Sensitivity/Recall*: 82%:
- *Specificity*: 82%
- Precision: 70%

Train data set results:

	Converted	Converted_Prob	Prospect ID	final_predicted	Lead_Score
0	0	0.067705	6227	0	7
1	0	0.127978	6322	0	13
2	0	0.059202	3644	0	6
3	0	0.086698	3011	0	9
4	1	0.914758	8267	1	91

Test data set results:

	Prospect ID	Converted	Converted_Prob	final_predicted	Lead_Score
0	9058	0	0.102429	0	10
1	5002	0	0.498294	1	50
2	3472	0	0.131460	0	13
3	6698	1	0.988450	1	99
4	6387	0	0.178725	0	18

Final Inferences

The above logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads

The business can calculate the Lead score of each candidate using the optimal cut off and if the lead score is greater than 35 (as the cut off probability is 0.35) then that candidate can be considered as a Hot Lead and the sales team can concentrate more on that person to convert him into a potential paying customer.

To increase the lead score of the candidate the business can target the Lead source feature and try to increase the count of categories Welingak Website and Reference as it has more potential leads. Moreover, the business can target on working professionals as they can turn into possible prospect lead. The business needs to target less on leads who have selected the 'Do not email'.

If there is business decides to the increase the budget for sales team then it can reduce the probability cut off from 0.35 i.e., can target customers with lead score less than 35 as this change would lead to an increase in the targeted customers, and it might result in high lead conversion rate in turn improving the profits of the business.

And if the business wants to concentrate on any other vertical and intends to reduce the sales teamwork then it can choose to increase the cut off from 0.35 to a higher one. This strategy can be applied when the business has reached its target and is willing to accept less lead conversion rate. During the application of this strategy, make sure that you target at least those with lead source of welingak and reference categories and working professionals as these have very high prospects of getting hot leads and potential paying customers.

To conclude the conversion rate of the leads is 80% when the lead with lead score 38 and above is targeted.


THANK
YOU
GRACIAS
ARIGATO
SHUKURIA
BOLZÏN
MERCI
BÏYAN
SHUKRIA
TINGKI
DANKSCHEEN
SPASSIBO
NUHUN
SNACHALHUYA
CHALTU
TASHAKKUR ATU
YAQHANYELAY
YUSPAGARÀTAM
WABEEJA
MAITEKA
HUI
SUKSAMA
EKHMET
GUI
HATUR
UNALCHÉESH
SPASIBO
DENKAUJA
NEHACHALHIYA
MIERSI
DHANYAVAAD
ANIHA
ATTO
MAAKE
GRAZIE
MEHRBANI
PALDIES
LAH
KOMAPSUMNIDA
SANCO
MERASTAWHY
GAEJTHO
GOZAIMASHITA
EFCHARISTO
AGUYJE
FAKAAUE
TAVTAPUCH
MEDAWAGSE
BAIKA
JUSPAXAR
EKOJU
SIKOMO
MAKÉTAJ
MIMMONCHAR