Lead Scoring Case Study

Submitted by Abhinav Phirani, Abhishek Sharma

EPGPDS - C53

Problem Statement

- X Education - an online edtech company selling courses is currently running at ~30% lead conversion rate. It currently uses several websites and google for bringing people to websites and then these leads are converted through watching videos, filling form, calls, SMS etc. Company also uses referrals. Company wants to identify more leads that have higher conversion rate, calling them 'Hot Leads'.

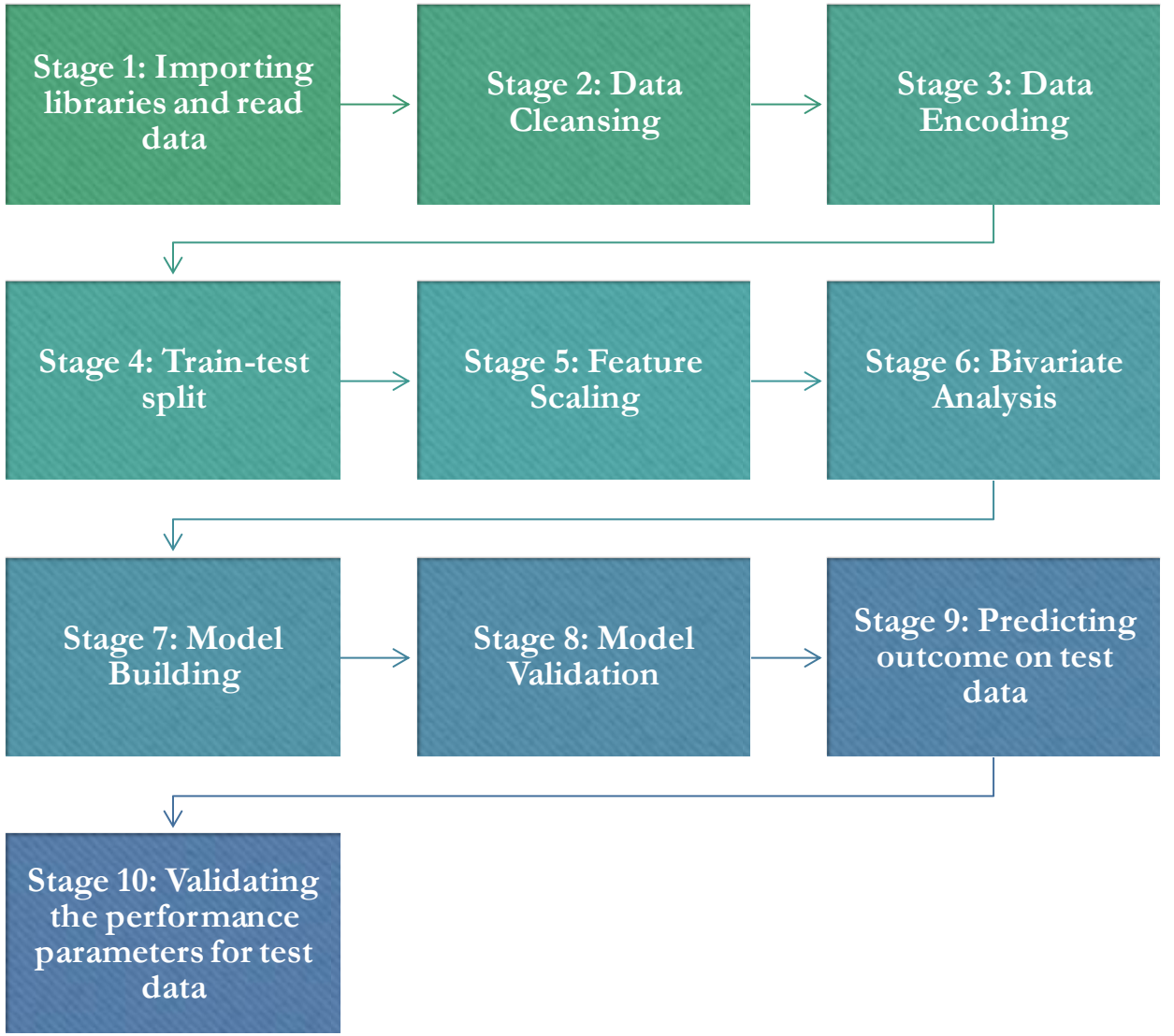
Lead Conversion Funnel

Data Description

- ~ 9000 data points on historical prospects and their conversion status
- Key attributes impacting conversion like Lead Source, Total Time Spent on Website, Total Visits, Last Activity, Do not Email, Do not Call, Current Occupation status, Lead Origin, Country etc.
- Target Variable : Converted – 0/1

```
Data columns (total 37 columns):
Prospect ID                                    9240 non-null object
Lead Number                                    9240 non-null int64
Lead Origin                                    9240 non-null object
Lead Source                                    9204 non-null object
Do Not Email                                   9240 non-null object
Do Not Call                                    9240 non-null object
Converted                                      9240 non-null int64
TotalVisits                                    9103 non-null float64
Total Time Spent on Website                    9240 non-null int64
Page Views Per Visit                           9103 non-null float64
Last Activity                                  9137 non-null object
Country                                        6779 non-null object
Specialization                                 7802 non-null object
How did you hear about X Education             7033 non-null object
What is your current occupation                6550 non-null object
What matters most to you in choosing a course  6531 non-null object
Search                                         9240 non-null object
Magazine                                       9240 non-null object
Newspaper Article                              9240 non-null object
X Education Forums                             9240 non-null object
Newspaper                                      9240 non-null object
Digital Advertisement                          9240 non-null object
Through Recommendations                        9240 non-null object
Receive More Updates About Our Courses         9240 non-null object
Tags                                           5887 non-null object
Lead Quality                                   4473 non-null object
Update me on Supply Chain Content              9240 non-null object
Get updates on DM Content                      9240 non-null object
Lead Profile                                   6531 non-null object
City                                           7820 non-null object
Asymmetrique Activity Index                    5022 non-null object
Asymmetrique Profile Index                     5022 non-null object
Asymmetrique Activity Score                    5022 non-null float64
Asymmetrique Profile Score                     5022 non-null float64
I agree to pay the amount through cheque       9240 non-null object
A free copy of Mastering The Interview         9240 non-null object
Last Notable Activity                          9240 non-null object
```

End to end model development process



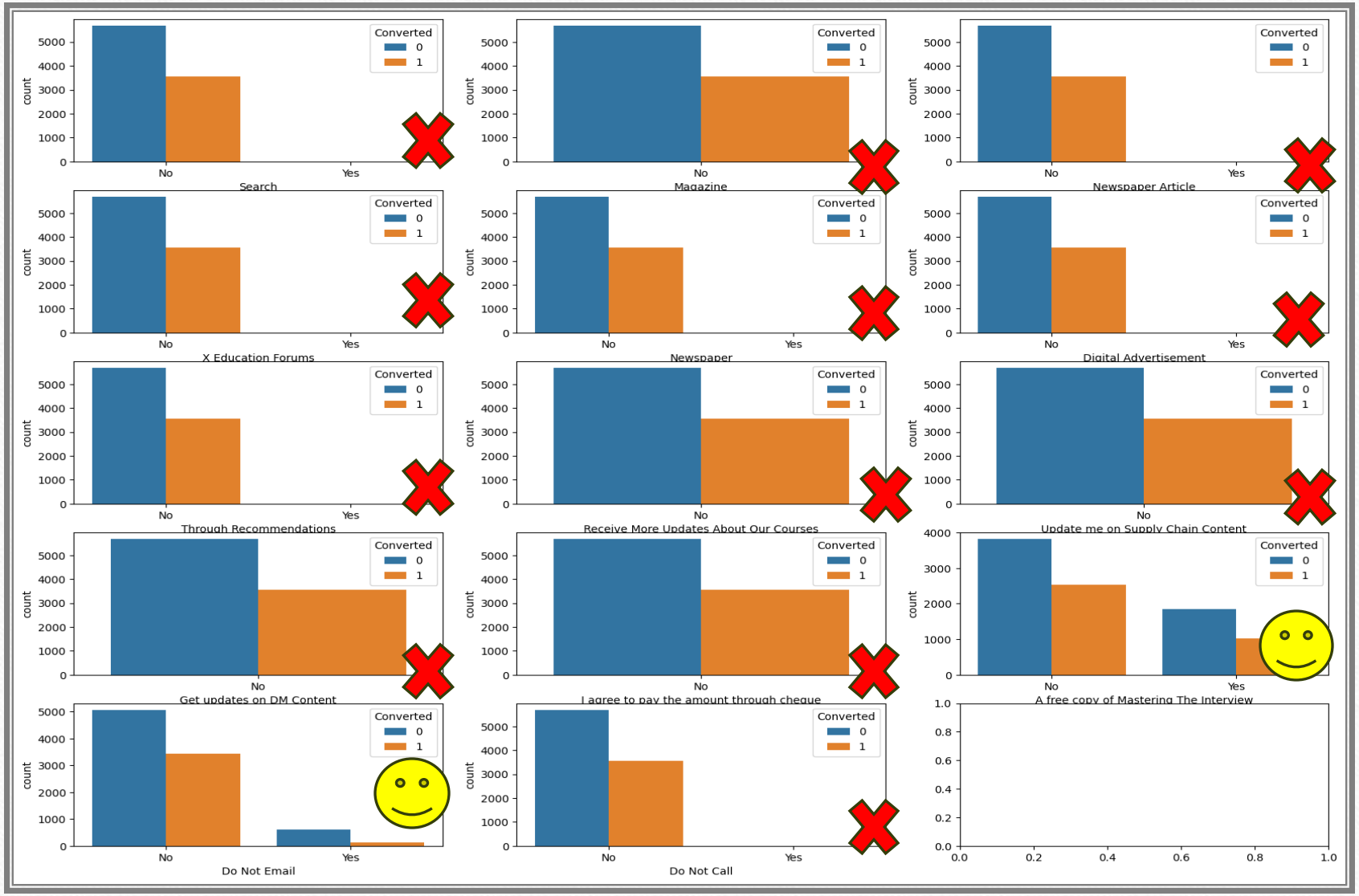
Data Cleansing

- Removal of unique values 'Prospect ID' and 'Lead Number'.
- Drop the columns with > 30% null values

```
(leads_df.isnull().sum()/len(leads_df))*100
```

```
Lead Origin                                    0.000000
Lead Source                                    0.389610
Do Not Email                                   0.000000
Do Not Call                                    0.000000
Converted                                      0.000000
TotalVisits                                    1.482684
Total Time Spent on Website                    0.000000
Page Views Per Visit                           1.482684
Last Activity                                  1.114719
Country                                       26.634199
Specialization                                36.580087
How did you hear about X Education            78.463203
What is your current occupation               29.112554
What matters most to you in choosing a course 29.318182
Search                                         0.000000
Magazine                                       0.000000
Newspaper Article                              0.000000
X Education Forums                             0.000000
Newspaper                                      0.000000
Digital Advertisement                          0.000000
Through Recommendations                        0.000000
Receive More Updates About Our Courses         0.000000
Tags                                          36.287879
Lead Quality                                  51.590909
Update me on Supply Chain Content              0.000000
Get updates on DM Content                      0.000000
Lead Profile                                  74.188312
City                                          39.707792
Asymmetrique Activity Index                   45.649351
Asymmetrique Profile Index                    45.649351
Asymmetrique Activity Score                   45.649351
Asymmetrique Profile Score                    45.649351
I agree to pay the amount through cheque       0.000000
A free copy of Mastering The Interview         0.000000
Last Notable Activity                          0.000000
dtype: float64
```

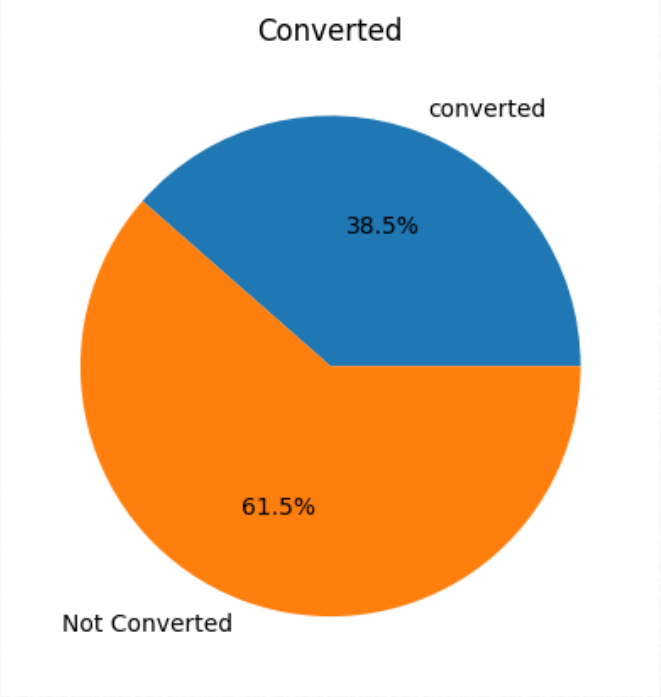
Remove the imbalanced variables, categorical data



Remaining Attributes after dropping not required variables

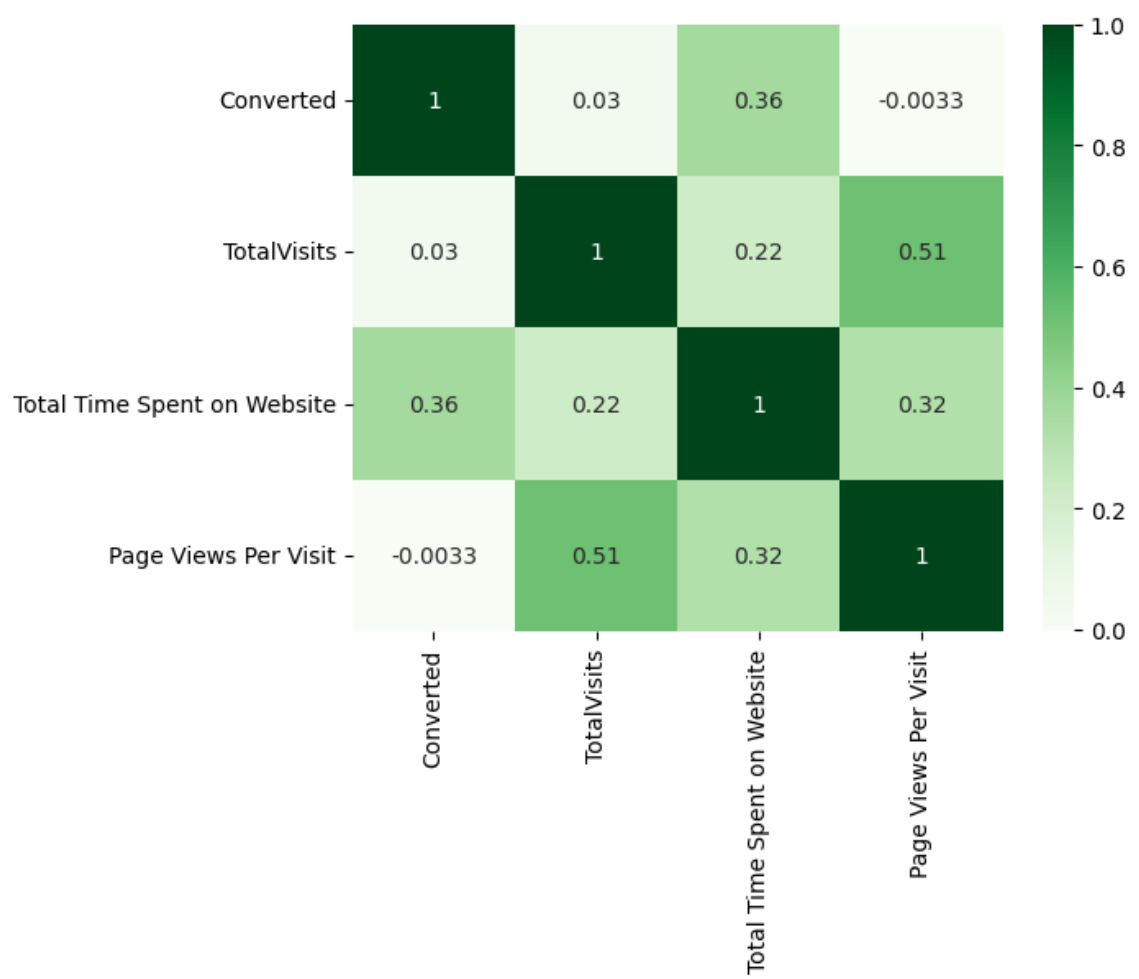
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9240 entries, 0 to 9239
Data columns (total 9 columns):
 #   Column                                  Non-Null Count  Dtype  
---  ------                                  --------------  -----  
 0   Lead Origin                             9240 non-null   object 
 1   Lead Source                             9240 non-null   object 
 2   Do Not Email                            9240 non-null   object 
 3   Converted                               9240 non-null   int64  
 4   TotalVisits                             9103 non-null   float64
 5   Total Time Spent on Website             9240 non-null   int64  
 6   Page Views Per Visit                    9103 non-null   float64
 7   What is your current occupation         9240 non-null   object 
 8   A free copy of Mastering The Interview  9240 non-null   object 
dtypes: float64(2), int64(2), object(5)
memory usage: 649.8+ KB
```

Numerical data analysis

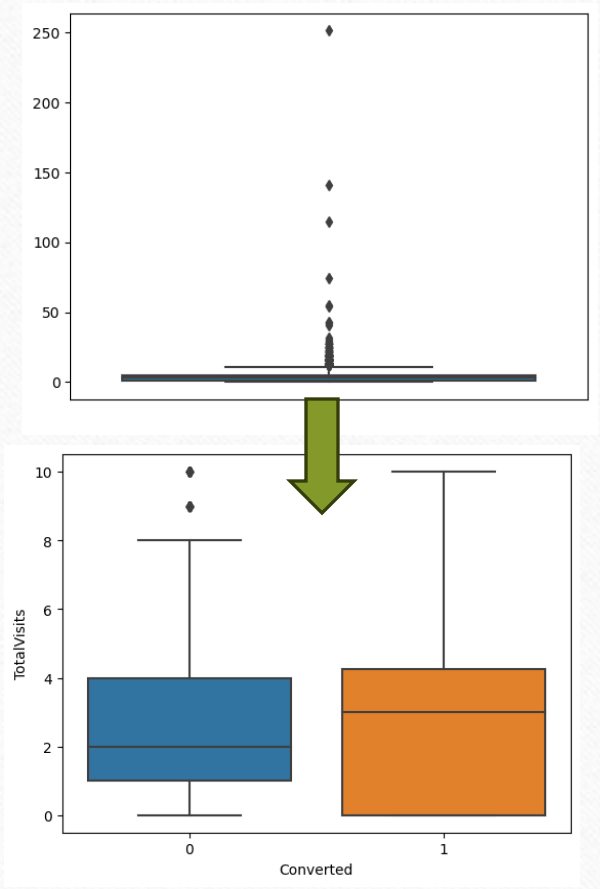


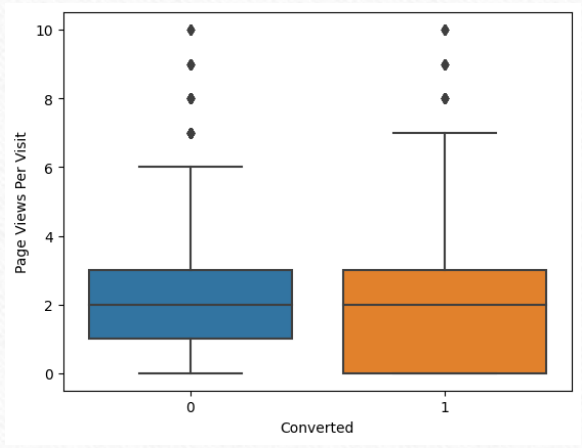
Y

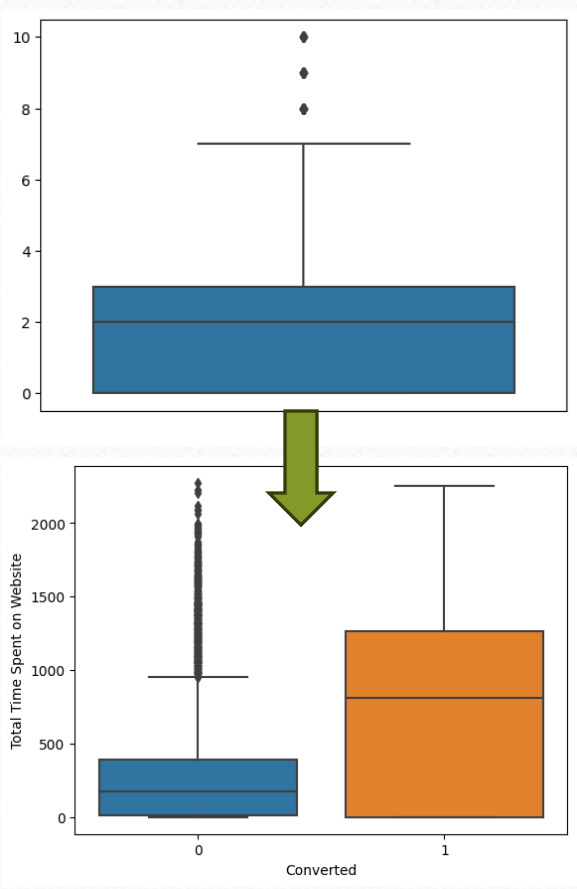
Xs



Check for Outliers and remove them







Data Preparation

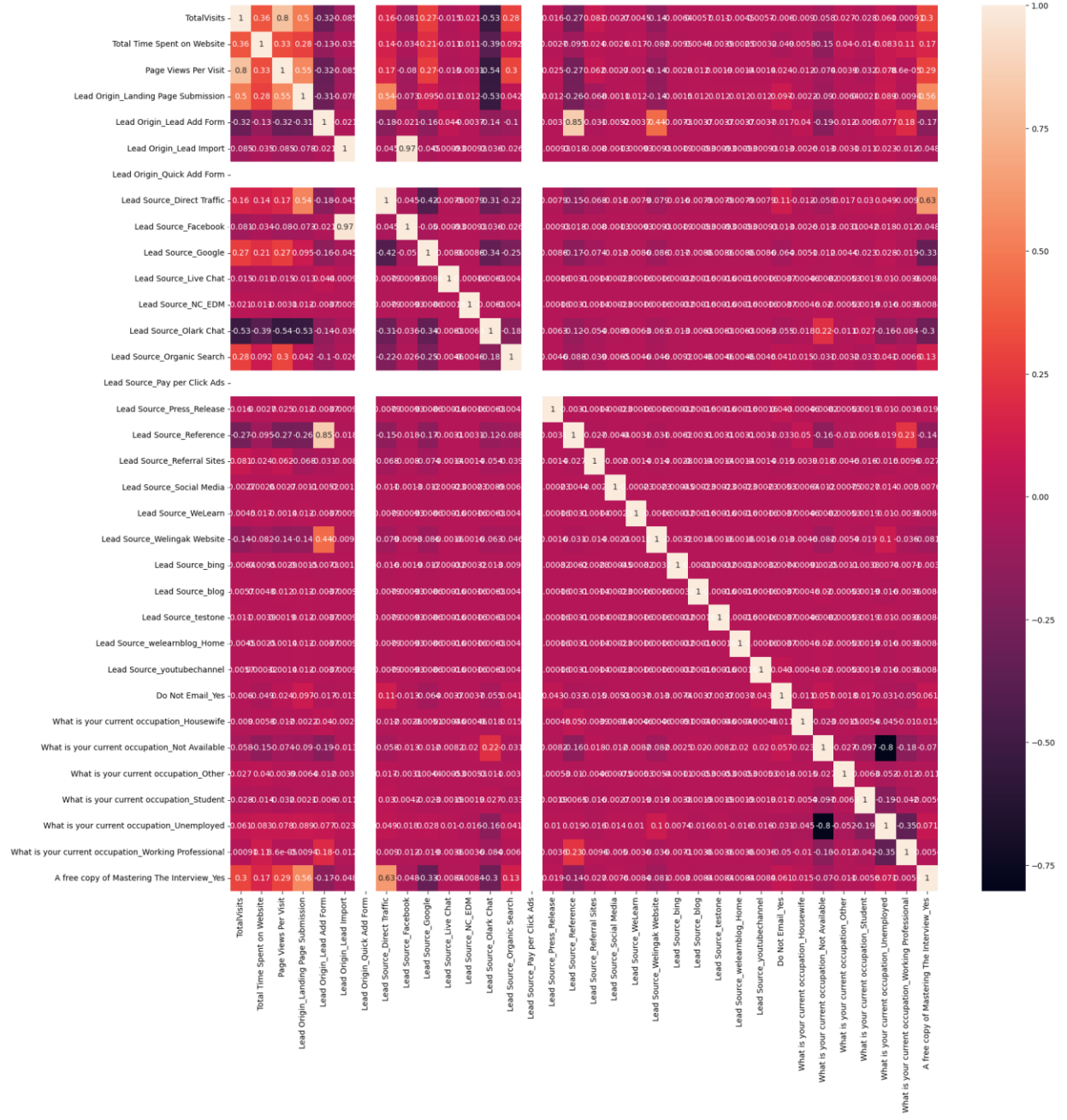
- Create Dummy variables
- Encode Categorical variables to 0/1 and remove main columns

- **Test – Train split : 70% training data, 30% test data**

- **Feature Scaling: - Standard Scaler to scale all variables**

Correlation Matrix

Identify highly correlated parameters (>0.8) and remove them



Model Building and Recursive Feature Elimination

```
[ ]  # importing necessary library
     import statsmodels.api as sm
```

```
[ ]  from sklearn.linear_model import LogisticRegression
     logreg = LogisticRegression()

     from sklearn.feature_selection import RFE
     # targeting 15 features in output
     rfe = RFE(estimator=logreg, n_features_to_select=15)
     rfe = rfe.fit(X_train, y_train)
```

Fit the model and check P-value for each feature

Iteration 1: Remove variable with highest P-value and fit the model again and repeat till P-value is <0.05 for all variables

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6220
Model:	GLM	**Df Residuals:**	6204
Model Family:	Binomial	**Df Model:**	15
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2820.4
Date:	Sun, 20 Aug 2023	**Deviance:**	5640.8
Time:	15:26:41	**Pearson chi2:**	7.00e+03
No. Iterations:	21	**Pseudo R-squ. (CS):**	0.3431
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.1226	0.095	-11.876	0.000	-1.308	-0.937
Total Time Spent on Website	1.0566	0.039	27.232	0.000	0.981	1.133
Lead Source_Direct Traffic	-1.1103	0.108	-10.321	0.000	-1.321	-0.899
Lead Source_Facebook	-0.7939	0.469	-1.694	0.090	-1.712	0.125
Lead Source_Google	-0.7738	0.103	-7.510	0.000	-0.976	-0.572
Lead Source_Live Chat	23.2806	4.82e+04	0.000	1.000	-9.44e+04	9.45e+04
Lead Source_Organic Search	-0.9288	0.128	-7.246	0.000	-1.180	-0.678
Lead Source_Reference	2.4979	0.213	11.729	0.000	2.080	2.915
Lead Source_Referral Sites	-1.4702	0.353	-4.159	0.000	-2.163	-0.777
Lead Source_Welingak Website	4.5881	0.721	6.364	0.000	3.175	6.001
Do Not Email_Yes	-1.1453	0.159	-7.198	0.000	-1.457	-0.833
What is your current occupation_Housewife	23.7109	1.47e+04	0.002	0.999	-2.88e+04	2.88e+04
What is your current occupation_Other	0.8514	0.655	1.301	0.193	-0.432	2.135
What is your current occupation_Student	1.1984	0.217	5.517	0.000	0.773	1.624
What is your current occupation_Unemployed	1.3285	0.084	15.862	0.000	1.164	1.493
What is your current occupation_Working Professional	3.8867	0.201	19.358	0.000	3.493	4.280

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6220
Model:	GLM	**Df Residuals:**	6208
Model Family:	Binomial	**Df Model:**	11
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2833.6
Date:	Sun, 20 Aug 2023	**Deviance:**	5667.1
Time:	15:27:20	**Pearson chi2:**	7.05e+03
No. Iterations:	7	**Pseudo R-squ. (CS):**	0.3403
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.1128	0.093	-11.940	0.000	-1.295	-0.930
Total Time Spent on Website	1.0556	0.039	27.279	0.000	0.980	1.131
Lead Source_Direct Traffic	-1.0781	0.106	-10.124	0.000	-1.287	-0.869
Lead Source_Google	-0.7424	0.102	-7.285	0.000	-0.942	-0.543
Lead Source_Organic Search	-0.8896	0.127	-7.000	0.000	-1.139	-0.641
Lead Source_Reference	2.5634	0.212	12.069	0.000	2.147	2.980
Lead Source_Referral Sites	-1.4463	0.353	-4.101	0.000	-2.137	-0.755
Lead Source_Welingak Website	4.6199	0.721	6.409	0.000	3.207	6.033
Do Not Email_Yes	-1.1513	0.159	-7.243	0.000	-1.463	-0.840
What is your current occupation_Student	1.1577	0.217	5.335	0.000	0.732	1.583
What is your current occupation_Unemployed	1.2869	0.082	15.616	0.000	1.125	1.448
What is your current occupation_Working Professional	3.8436	0.200	19.200	0.000	3.451	4.236

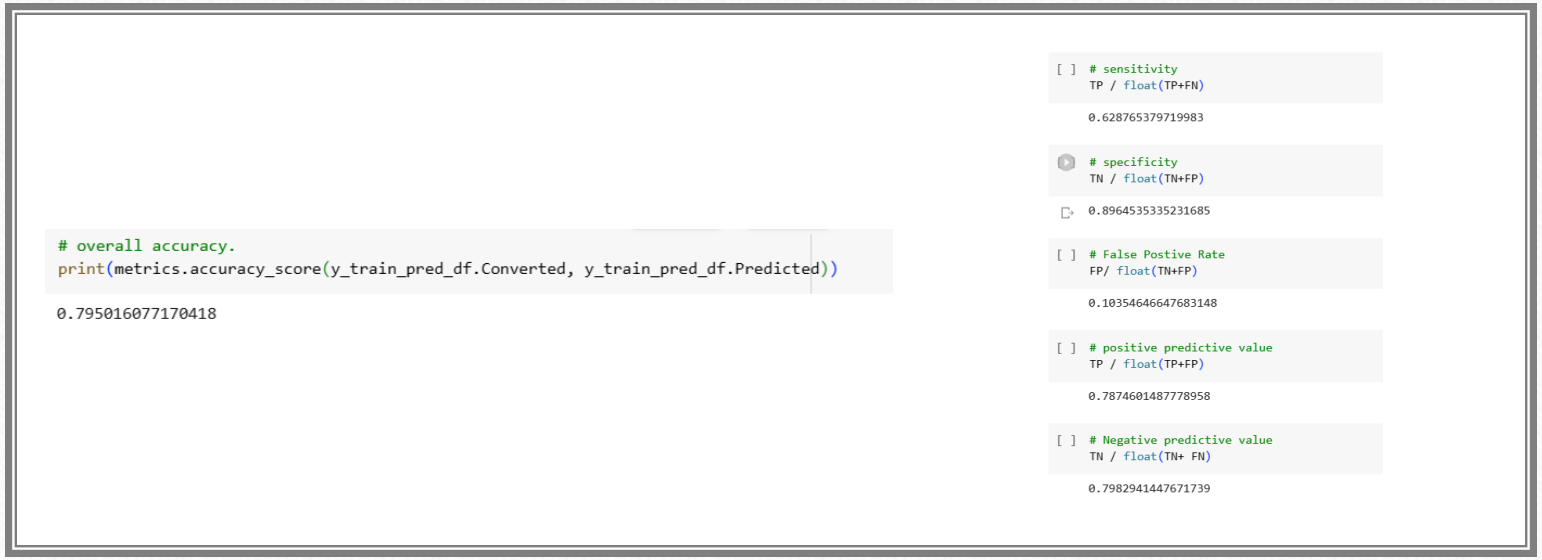
	Features	VIF
0	Total Time Spent on Website	1.10
1	Lead Source_Direct Traffic	1.73
2	Lead Source_Google	1.72
3	Lead Source_Organic Search	1.30
4	Lead Source_Reference	1.28
5	Lead Source_Referral Sites	1.02
6	Lead Source_Welingak Website	1.07
7	Do Not Email_Yes	1.11
8	What is your current occupation_Student	1.05
9	What is your current occupation_Unemployed	2.56
10	What is your current occupation_Working Profes...	1.29

Iteration 5

VIF

P-value <0.05

VIF <5



```
# overall accuracy.
print(metrics.accuracy_score(y_train_pred_df.Converted, y_train_pred_df.Predicted))
```

```
0.795016077170418
```

```
[ ] # sensitivity
    TP / float(TP+FN)
```

```
0.628765379719983
```

```
# specificity
TN / float(TN+FP)
```

```
0.8964535335231685
```

```
[ ] # False Postive Rate
    FP/ float(TN+FP)
```

```
0.10354646647683148
```

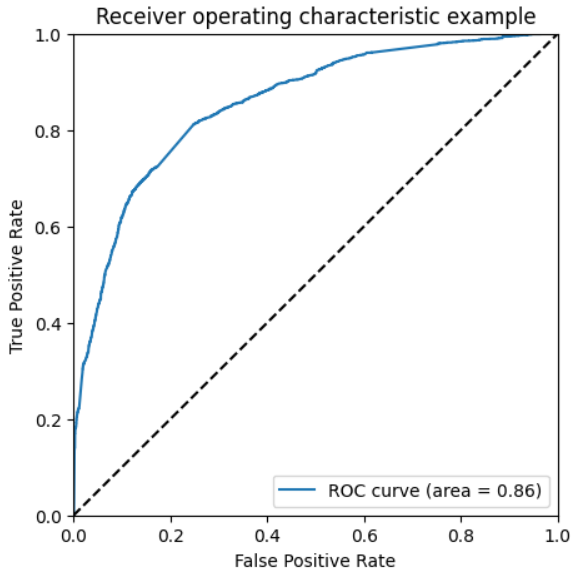
```
[ ] # positive predictive value
    TP / float(TP+FP)
```

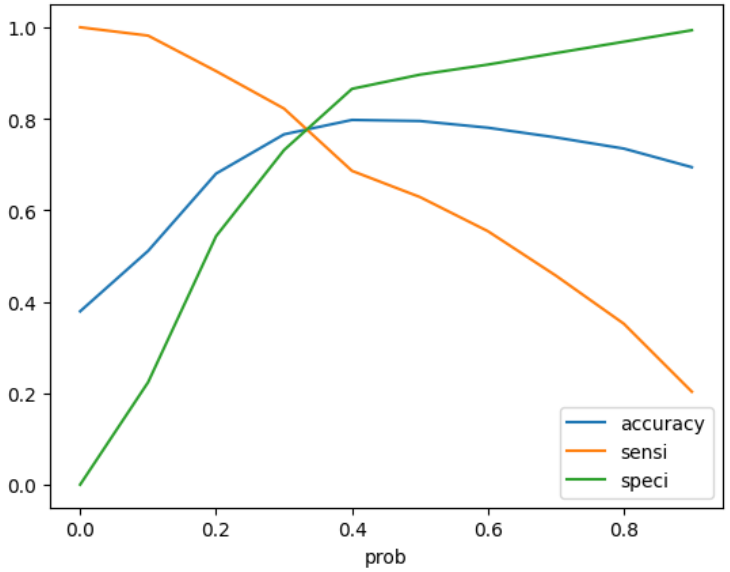
```
0.7874601487778958
```

```
[ ] # Negative predictive value
    TN / float(TN+ FN)
```

```
0.7982941447671739
```

Predict Y-train and Check performance





Draw ROC curve

ROC 0.86 is close to 1 hence good fit

optimal probability cut-off

Optimal prob cut-off at 0.3

```
[ ]  y_train_pred_df['Lead_Score'] = y_train_pred_df.Model_prob.map( lambda x: round(x*100))

     y_train_pred_df[['Converted','Model_prob','final_Predicted','Lead_Score']].head()
```

	Converted	Model_prob	final_Predicted	Lead_Score
7806	0	0.237803	0	24
926	0	0.328122	1	33
3565	0	0.666980	1	67
7572	0	0.115852	0	12
4363	0	0.146377	0	15

Calculate Lead score and predict accuracy based on Model probability calculation

Prediction Accuracy of 82.223%

```
checking_test_df = y_test_df.loc[y_test_df['Converted']==1,['Final_Pred
checking_test_df['Final_Predicted'].value_counts()
```

```
1    871
0    168
Name: Final_Predicted, dtype: int64
```

```
871/(871+168)
```

```
0.8383060635226179
```

```
#sensitivity of our logistic regression model
TP / float(TP+FN)
```

```
0.8383060635226179
```

```
#specificity
TN / float(TN+FP)
```

```
0.7180589680589681
```

Test the same model on test data at cut-off prob of 0.3 and calculate performance

Conclusion

Our final logistic regression model has following attributes and respective coefficients

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6220
Model:	GLM	**Df Residuals:**	6208
Model Family:	Binomial	**Df Model:**	11
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2833.6
Date:	Sun, 20 Aug 2023	**Deviance:**	5667.1
Time:	15:27:20	**Pearson chi2:**	7.05e+03
No. Iterations:	7	**Pseudo R-squ. (CS):**	0.3403
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.1128	0.093	-11.940	0.000	-1.295	-0.930
Total Time Spent on Website	1.0556	0.039	27.279	0.000	0.980	1.131
Lead Source_Direct Traffic	-1.0781	0.106	-10.124	0.000	-1.287	-0.869
Lead Source_Google	-0.7424	0.102	-7.285	0.000	-0.942	-0.543
Lead Source_Organic Search	-0.8896	0.127	-7.000	0.000	-1.139	-0.641
Lead Source_Reference	2.5634	0.212	12.069	0.000	2.147	2.980
Lead Source_Referral Sites	-1.4463	0.353	-4.101	0.000	-2.137	-0.755
Lead Source_Welingak Website	4.6199	0.721	6.409	0.000	3.207	6.033
Do Not Email_Yes	-1.1513	0.159	-7.243	0.000	-1.463	-0.840
What is your current occupation_Student	1.1577	0.217	5.335	0.000	0.732	1.583
What is your current occupation_Unemployed	1.2869	0.082	15.616	0.000	1.125	1.448
What is your current occupation_Working Professional	3.8436	0.200	19.200	0.000	3.451	4.236

Subjective Questions -1

Which are the top three variables in your model which contribute most towards the probability of a lead getting converted?

Solution: Based on the coefficient values, the top three variables that contribute most are:

- **Lead Source**
- **What is your current occupation**
- **Do Not Email**

What are the top 3 categorical/dummy variables in the model which should be focused the most on in order to increase the probability of lead conversion?

Solution: Again, based on the coefficient values, the top three categorical/dummy variables contributing most towards lead conversion are the same:

- **Lead Source - Welingak Website**
- **What is your current occupation - Working Professional**
- **Lead Source - Reference**

Subjective Questions -2

X Education has a period of 2 months every year during which they hire some interns. The sales team, in particular, has around 10 interns allotted to them. So during this phase, they wish to make the lead conversion more aggressive. So they want almost all of the potential leads (i.e. the customers who have been predicted as 1 by the model) to be converted and hence, want to make phone calls to as much of such people as possible. Suggest a good strategy they should employ at this stage.

Solution: The final prediction is calculated based on optimal cut off value of 0.3. To maximize the chances of conversion, the company should reach to all prospects which have a conversion possibility, i.e. value of 1 and probability of 0.3.

Similarly, at times, the company reaches its target for a quarter before the deadline. During this time, the company wants the sales team to focus on some new work as well. So during this time, the company's aim is to not make phone calls unless it's extremely necessary, i.e. they want to minimize the rate of useless phone calls. Suggest a strategy they should employ at this stage.

Solution: Post meeting the target if we are able to run through all the calls at optimal cut-off probability of 0.3, we need to maximize the positives, even if the total number of calls go down. Therefore, we need high sensitivity, even if conversion probability is low. We can slowly start moving from 0.3 probability to 0.2 and 0.1 and start calling all possible prospects with value equal to 1.

Thank you