LEAD SCORING CASE STUDY

Finding the Hot Leads from the Initial Pool of leads by classification model to increase the Target Lead Conversion Rate of X Education

Vishak Nair & Souptik Majumder

26th August2019

STEPS FOLLOWED FOR THE LEAD SCORING

- Understanding the Data
- Data Cleaning drive
- Data Imbalance checks and Outlier Analysis
- Encoding steps to convert into continuous
- Train-Test split
- Scaling to similar range
- Running RFE and selecting the relevant Features
- Backward approach to get the final Logistic model.
- Model Evaluation and final score preparation.

- ➢The stats and the shape of the dataset where well observed initially before any action.
- ➢ Data cleaning was done for all the missing values.
- ➢ Data imbalance, Outliers where checked and taken a call for data preparation.
- ➢ All the categorical variables were converted and train test split done.
- ➢ Standard Scaling was done inorder to bring all the features in comparable scale.
- ➢ Since last dataset contains many features, RFE used to find out the most important features.
- ➢ Features were dropped then using P- value and VIF and final model build.
- ➢ Final model was evaluated based on parameters such as accuracy, sensitivity, ROC curve etc and finalised.

DATASETS USED

- Leads Dataset -> Contains the list of around 9000 past leads with attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc and the information that whether they converted or not.

Problem Statement

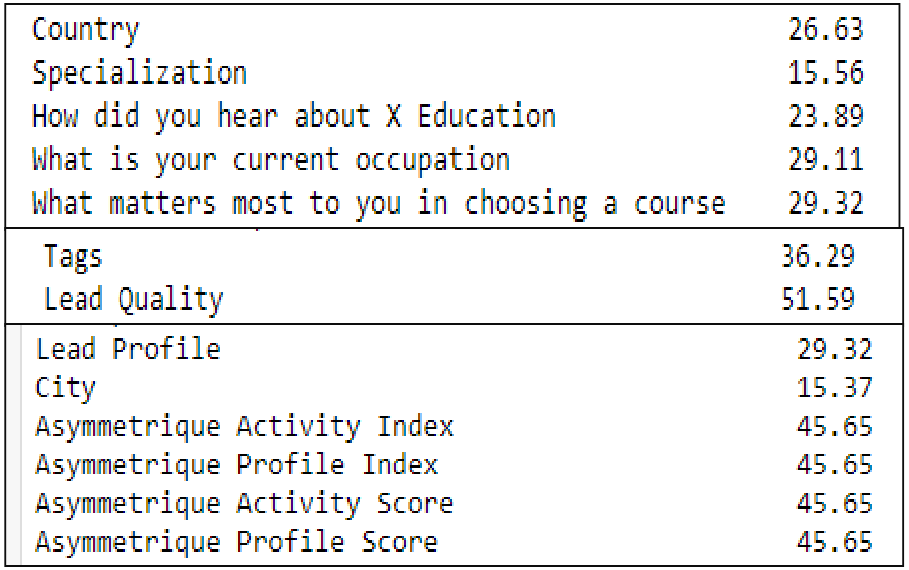
- X Education the online education company sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google. Among the people who land up on the website, people who fill up the form providing their email address or phone number turns out to be a lead. Our aim here is to help find out how to increase the Lead Conversion Rate which is at a low 30% now. Targeting the right leads, is the solution here, for which we will 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. How to target the right leads and how to find out the hot lead, hence marks our problem statement.

The Approach

- We started with **understanding the dataset** and tried to get a look and feel of the data. We observed the initial few records of the dataset; initial statistics to see what each feature has got to offer and the type of each column and most importantly the dimensions of the dataset.
- Started **Data cleaning** drive with Null value treatment and then proceeded with Data Imbalance checks and Outlier Detection and treatment.
- **One Hot encoding** and different approaches to convert the categorical features to continuous for model building.
- **Train-Test split** was done next with random_state of 100 in a ratio of 70% to 30%.
- **Standard Scaling** of the numeric features is other step which we did as part of Data Preparation to get all the features in comparable range.
- **Recursive Feature Elimination** approach to cut short the 126 number of features we had after data preparation.
- From the initial 25 features which RFE suggested, we used a **backward approach** to drop the insignificant features which had either high **p-value** or high **Variance Inflation factor.**
- **Model evaluation** was done by necessary features such as Accuracy, Sensitivity, Specificity, True Positive Rate, False Positive Rate and with the ROC curve.
- **Optimal cut-off** was found out to increase the efficiency of the lead conversion prediction rate.
- **Final dataset** was produced which contains just the **Lead Number and the Lead score** which was found out based on the probability given by the final Logistic Regression Model.

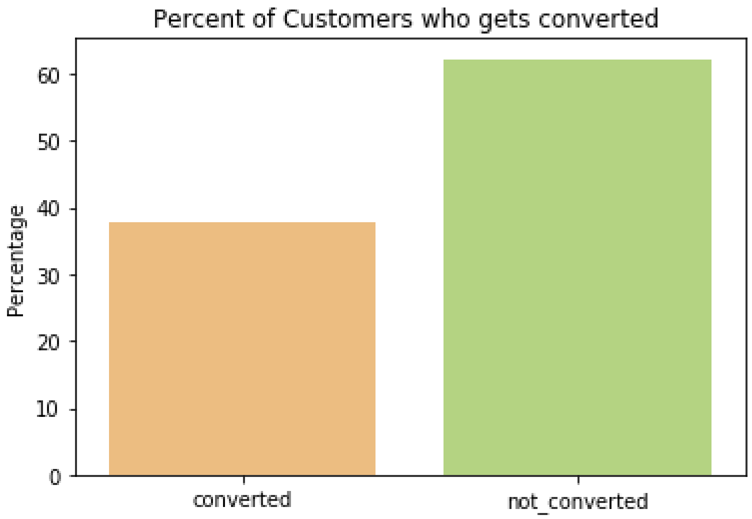
Data Cleaning Drive

- ➢ Some of the features with high number of NULL values are shown here.
- ➢ Each feature was understood properly and the relevance for the final model and decision was taken forth for each of the feature with NULL values.
- ➢ Main Aim here was to keep the initial distribution of data for each feature intact and avoid bringing in Bias.
- ➢ New level of value was introduced wherever necessary, to convert the NULL values to these level.
- ➢ For important features such as Asymmetrique Activity Index and Profile Index for which only 3 levels are there, we have introduced Unknown level to avoid imputing.
- ➢ For features such as Activity Score and Profile Score we have imputed mean to remain distribution same.
- ➢ At the same time, mode was used to impute for some categorical features as well.

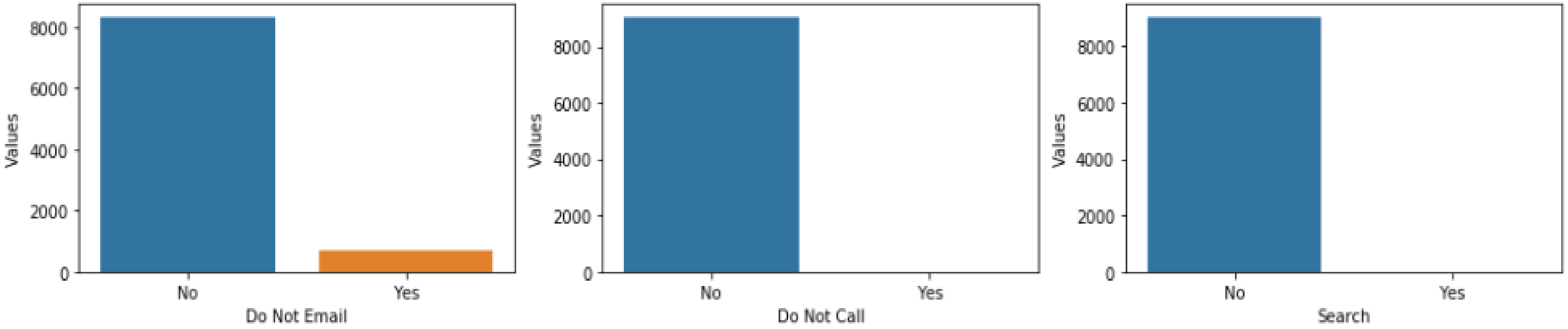


```
Country                                         26.63
Specialization                                  15.56
How did you hear about X Education              23.89
What is your current occupation                 29.11
What matters most to you in choosing a course   29.32
Tags                                            36.29
Lead Quality                                    51.59
Lead Profile                                    29.32
City                                            15.37
Asymmetrique Activity Index                     45.65
Asymmetrique Profile Index                      45.65
Asymmetrique Activity Score                     45.65
Asymmetrique Profile Score                      45.65
```

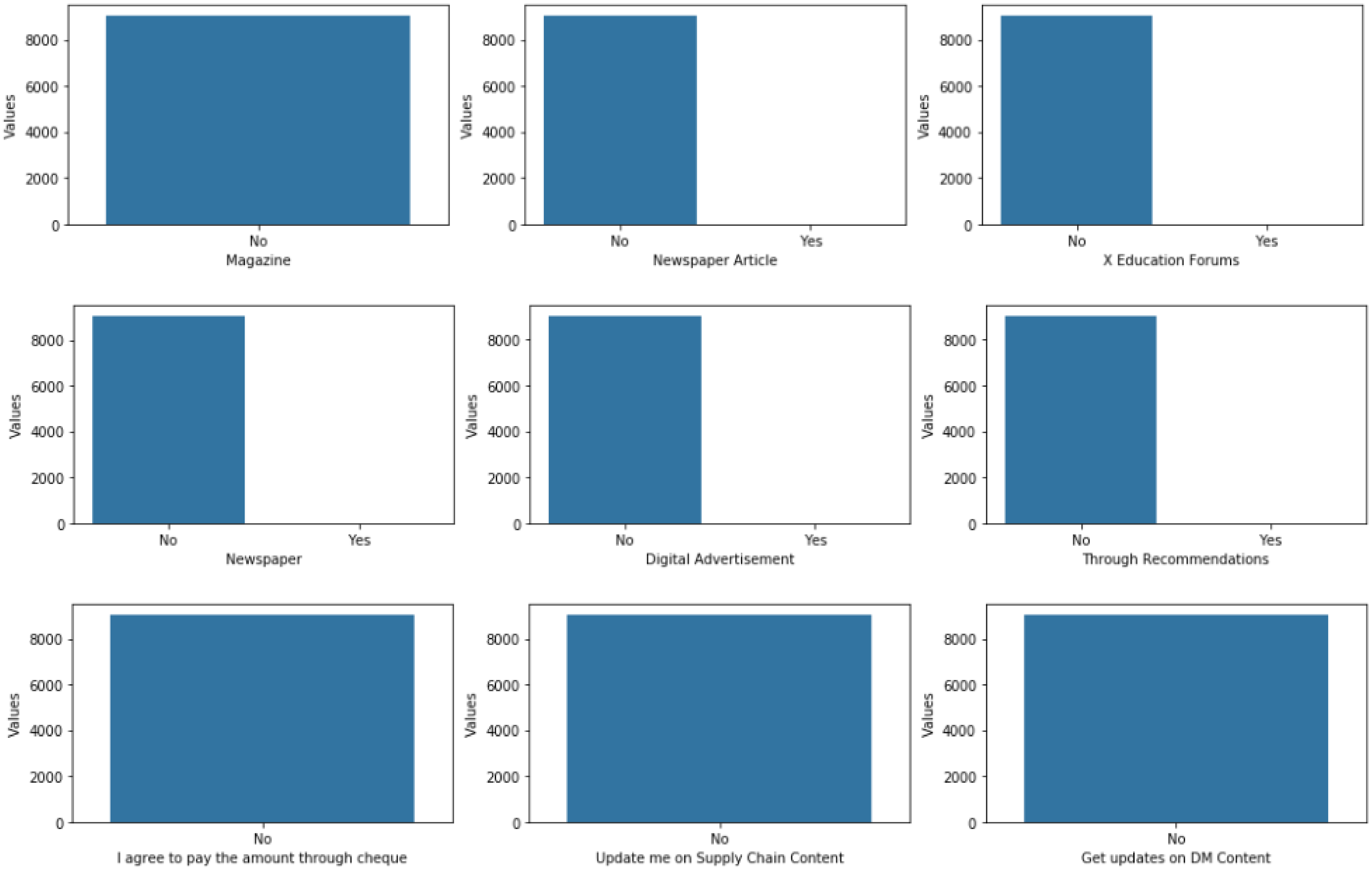
Data Imbalance



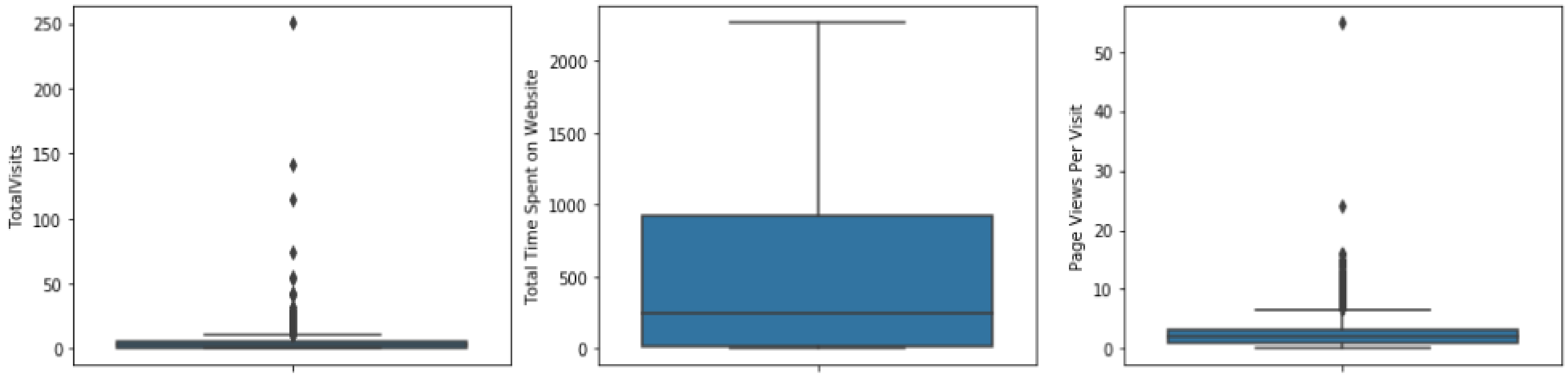
- ➢ We can see a little bit of Data Imbalance between these 2 categories of converted and Not converted. We can see out of total percentage around 30%+ ~ nearly 40% of the leads are only getting converted which supports the initial claim that `The typical lead conversion rate at X education is around 30%.`
- ➢ The below graphs explain the clear Data imbalance for many Binary features such as ‘Do Not Email', 'Do Not Call', 'Search', 'Magazine', 'Newspaper Article‘ etc hence, all these columns won't add any relevant information and give us variance for model building and hence dropped.



Data Imbalance

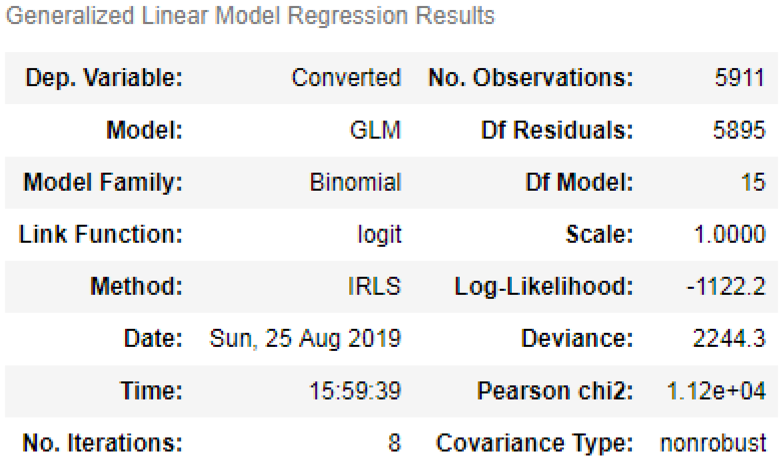


Outlier Analysis



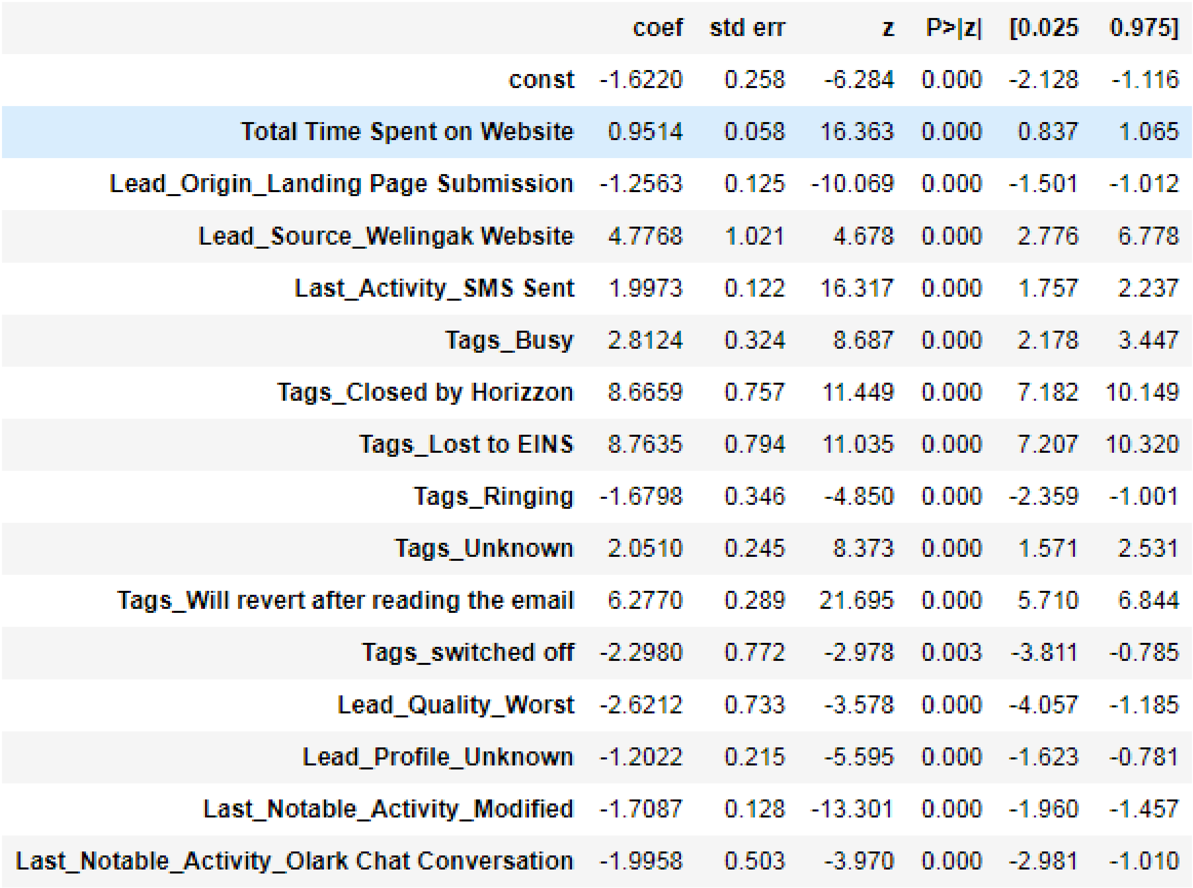
➢ We can clearly spot outliers in the features such as `TotalVisits and Page Views Per Visit`. Total time spent on Website doesn't have any outliers. We can see that there are as many as 250 visits recorded for total visits by possible leads. As high as this number of visits to a website seems to be not like a correct capture and hence we can remove these outliers. Similarly for the page views per visit, as many as 20+ page views in a single visit seems to be not correct. We can remove these as well.

Final Model and the Features



Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	5911
Model:	GLM	Df Residuals:	5895
Model Family:	Binomial	Df Model:	15
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1122.2
Date:	Sun, 25 Aug 2019	Deviance:	2244.3
Time:	15:59:39	Pearson chi2:	1.12e+04
No. Iterations:	8	Covariance Type:	nonrobust

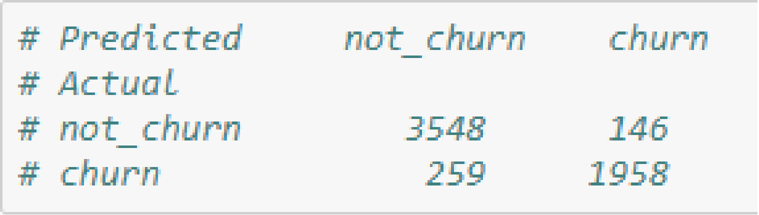


	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.6220	0.258	-6.284	0.000	-2.128	-1.116
Total Time Spent on Website	0.9514	0.058	16.363	0.000	0.837	1.065
Lead_Origin_Landing Page Submission	-1.2563	0.125	-10.069	0.000	-1.501	-1.012
Lead_Source_Welingak Website	4.7768	1.021	4.678	0.000	2.776	6.778
Last_Activity_SMS Sent	1.9973	0.122	16.317	0.000	1.757	2.237
Tags_Busy	2.8124	0.324	8.687	0.000	2.178	3.447
Tags_Closed by Horizzon	8.6659	0.757	11.449	0.000	7.182	10.149
Tags_Lost to EINS	8.7635	0.794	11.035	0.000	7.207	10.320
Tags_Ringing	-1.6798	0.346	-4.850	0.000	-2.359	-1.001
Tags_Unknown	2.0510	0.245	8.373	0.000	1.571	2.531
Tags_Will revert after reading the email	6.2770	0.289	21.695	0.000	5.710	6.844
Tags_switched off	-2.2980	0.772	-2.978	0.003	-3.811	-0.785
Lead_Quality_Worst	-2.6212	0.733	-3.578	0.000	-4.057	-1.185
Lead_Profile_Unknown	-1.2022	0.215	-5.595	0.000	-1.623	-0.781
Last_Notable_Activity_Modified	-1.7087	0.128	-13.301	0.000	-1.960	-1.457
Last_Notable_Activity_Olark Chat Conversation	-1.9958	0.503	-3.970	0.000	-2.981	-1.010

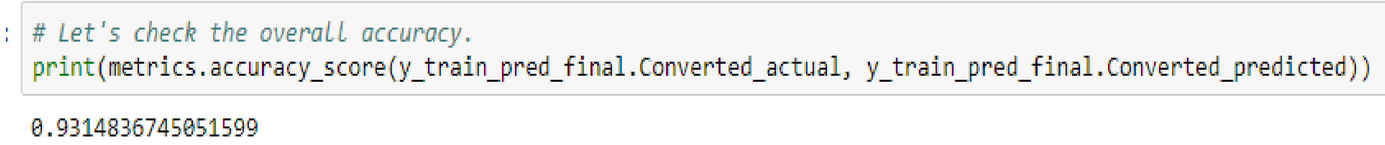
- ➢ We can see the Model parameters and the model features of the final model we selected.
- ➢ Here some of the features are encoded features which was created as part of One Hot Encoding process such as 'Tags_' features.
- ➢ We can see that the p- values for these features are less than 0.05 hence making all these features significant.
- ➢ Since all these features here are scaled, the coefficient of each feature here shows the contribution towards the final marking of probability for each lead whether the lead will get converted or not.

Model Performance on TEST set

Confusion Matrix and other parameters

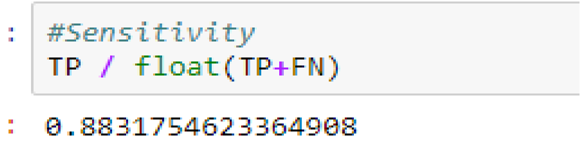


```
# Predicted      not_churn     churn
# Actual
# not_churn          3548        146
# churn               259       1958
```



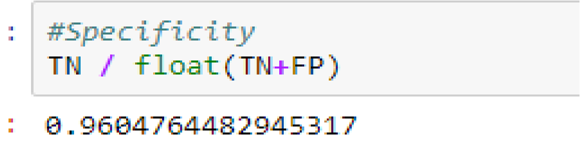
```
# Let's check the overall accuracy.
print(metrics.accuracy_score(y_train_pred_final.Converted_actual, y_train_pred_final.Converted_predicted))
```

```
0.9314836745051599
```



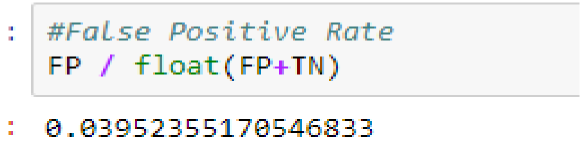
```
#Sensitivity
TP / float(TP+FN)
```

```
0.8831754623364908
```



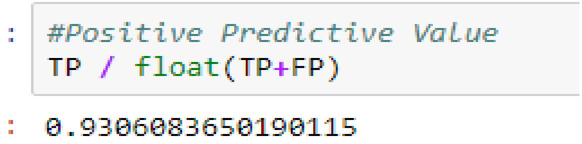
```
#Specificity
TN / float(TN+FP)
```

```
0.9604764482945317
```



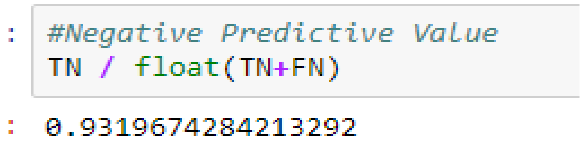
```
#False Positive Rate
FP / float(FP+TN)
```

```
0.03952355170546833
```



```
#Positive Predictive Value
TP / float(TP+FP)
```

```
0.9306083650190115
```



```
#Negative Predictive Value
TN / float(TN+FN)
```

```
0.9319674284213292
```

➢ Sensitivity of 88%,Specificity of 96%, FPR of only around 4% and Positive Predictive Value of around 93% is what our model gives in the test set which are really good values

➢**Accuracy** determines the overall predicted accuracy of the model which is 93% now.

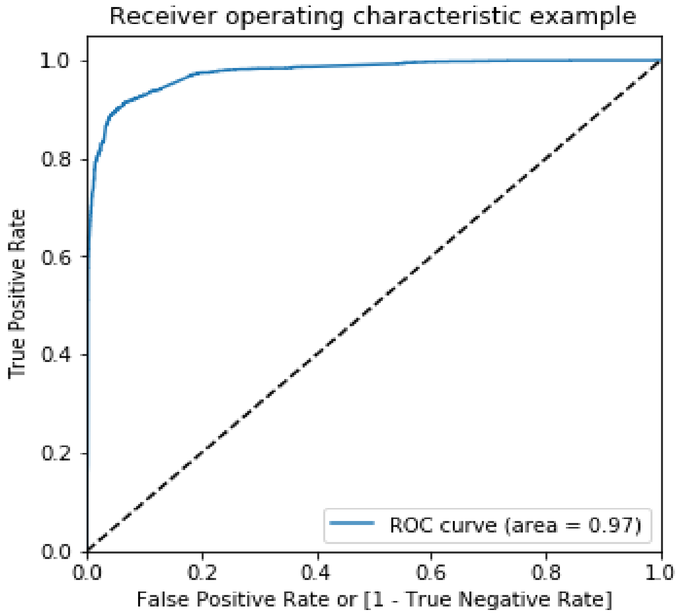
➢ **Sensitivity** (also called the true positive rate, or the recall) measures the proportion of actual positives which are correctly identified as such. The value of 88% shows a very good model.

➢**Specificity** (also called the true negative rate) measures the proportion of negatives which are correctly identified as such. We are getting a good percentage of specificity with our model.

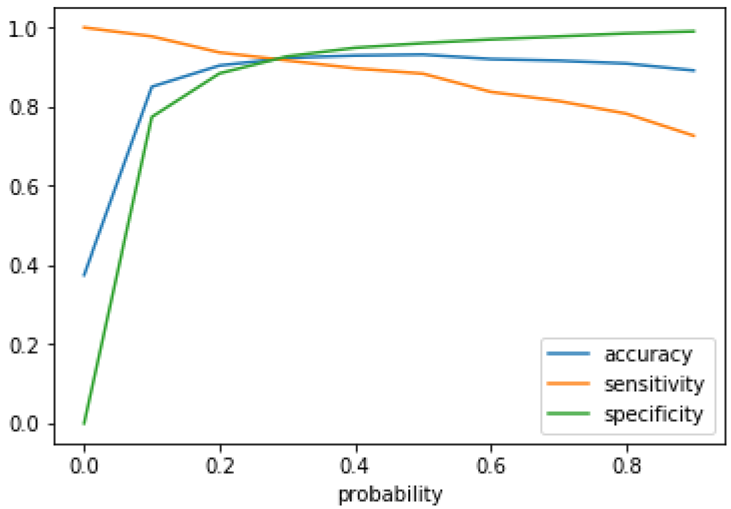
Model Performance continued.

ROC Curve

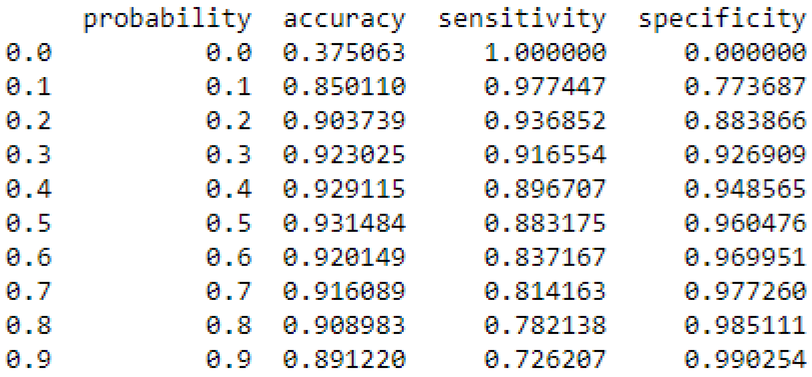
- ROC or Receiver Operating Characteristic Curve is a plot that shows the diagnostic ability of binary classifiers. We know that the classifiers that give curves closer to top-left corner indicate a better performance. Also for good models, the AUC or Area Under Curve would be higher. In our ROC plot above, we can see we are getting good AUC and a good trade off between the Sensitivity and Specificity.



Optimal Cut off point



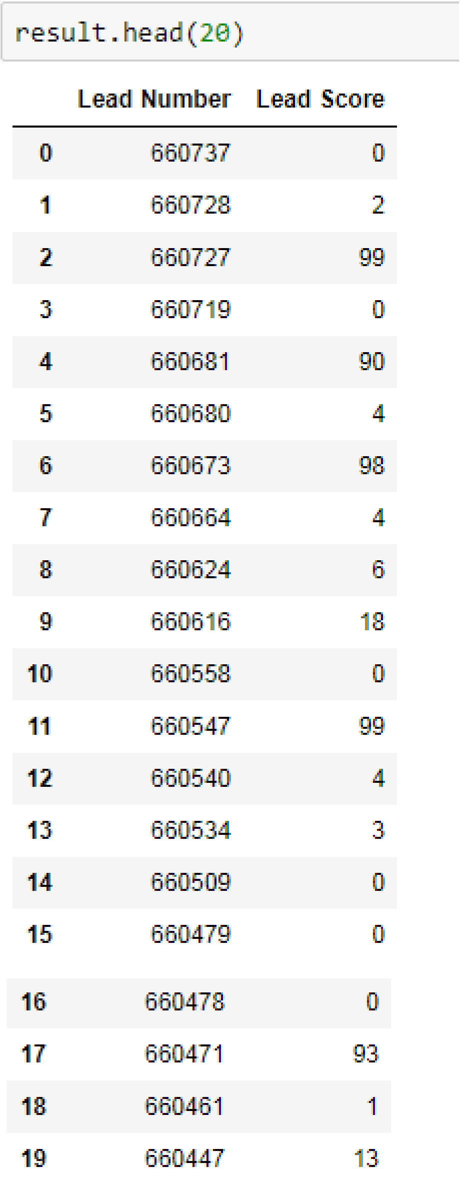
- From the plot between the Accuracy, Sensitivity and Specificity we can see, probability around 0.3 would be a good cut-off. For cut-off probability of 0.3, we getting accuracy of 92.3%,specificity of 92.7% and more importantly sensitivity of 91.7%.



```
     probability  accuracy  sensitivity  specificity
0.0          0.0  0.375063     1.000000     0.000000
0.1          0.1  0.850110     0.977447     0.773687
0.2          0.2  0.903739     0.936852     0.883866
0.3          0.3  0.923025     0.916554     0.926909
0.4          0.4  0.929115     0.896707     0.948565
0.5          0.5  0.931484     0.883175     0.960476
0.6          0.6  0.920149     0.837167     0.969951
0.7          0.7  0.916089     0.814163     0.977260
0.8          0.8  0.908983     0.782138     0.985111
0.9          0.9  0.891220     0.726207     0.990254
```

Test set, Final Dataset and Recommendations

- ➢ We are getting a accuracy of 92% in our test set. Sensitivity of 92%,Specificity of 91% is also got which indicates the model performs not only in the train dataset, but also in the test dataset which indicates the model performance would be as good in different cases.
- ➢ In the side shows some sample leads and the lead score assigned with the help of our model from our final dataset.
- ➢ The Lead Score here is a value which we assigned to each lead in range 0-100, that shows the lead is a hot lead or not. The higher the score means higher the chance, lead will be converted.
- ➢ For example in the example shown in the side, we can see leads such as 660727,660681 and some more having values in range of 90s indicating they are hot leads.
- ➢ Depending on the quarter performance, whether they met the lead conversion rate for the quarter and need a less rate for the rest of the month, or an aggressive drive to chip in more leads can be done based on our final model.
- ➢ The Lead score will indicate the potential leads to target on increasing efficiency.
- ➢ The final model features will help in selecting the right marketing strategy and where to concentrate more as part of budget saving and efficient utilisation of pool amount for lead conversion.



```
result.head(20)
```

```
    Lead Number  Lead Score
0        660737           0
1        660728           2
2        660727          99
3        660719           0
4        660681          90
5        660680           4
6        660673          98
7        660664           4
8        660624           6
9        660616          18
10       660558           0
11       660547          99
12       660540           4
13       660534           3
14       660509           0
15       660479           0
16       660478           0
17       660471          93
18       660461           1
19       660447          13
```

THANK YOU