



Lead Scoring Case Study

VAIBHAV PARAKH

SANKALP SEKSARIA

Problem Objective

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, although X Education gets a lot of leads, its lead conversion rate is very poor which is only 30 %.

To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Goals of the Case Study

There are quite a few goals for this case study.

- 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.
- The CEO has given a ballpark of the target lead conversion rate to be around 80%.

Problem Solving Methods

Data Cleaning and Preparation

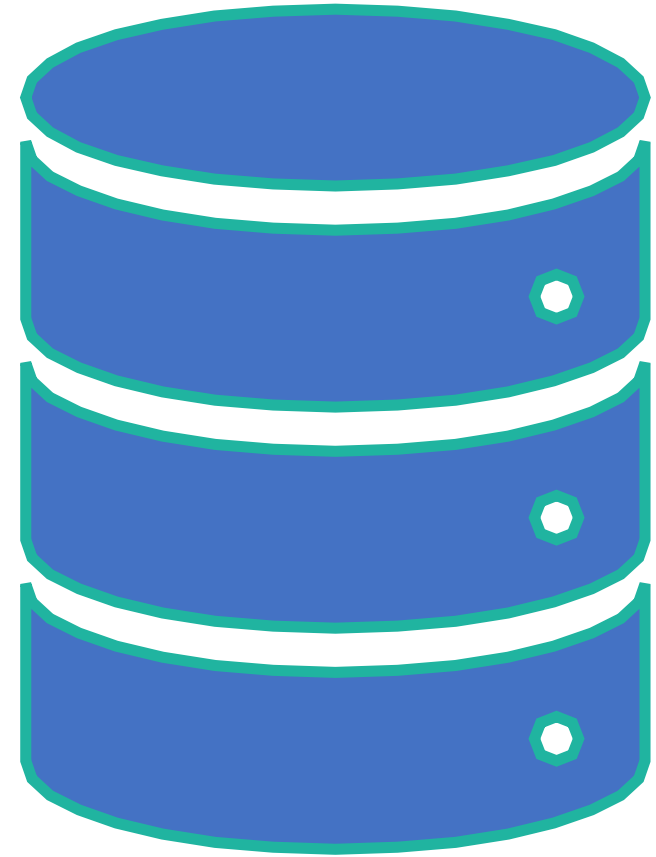
- Identify the data quality and clean based on requirement
- Handle null values based on converted rate without removing data points
- Data Imputation
- Outlier Analysis and treatment

Solve problem

- Variable Processing
- Univariate Analysis (EDA)
- Train Test Split data
- Logistic Regression Model Building

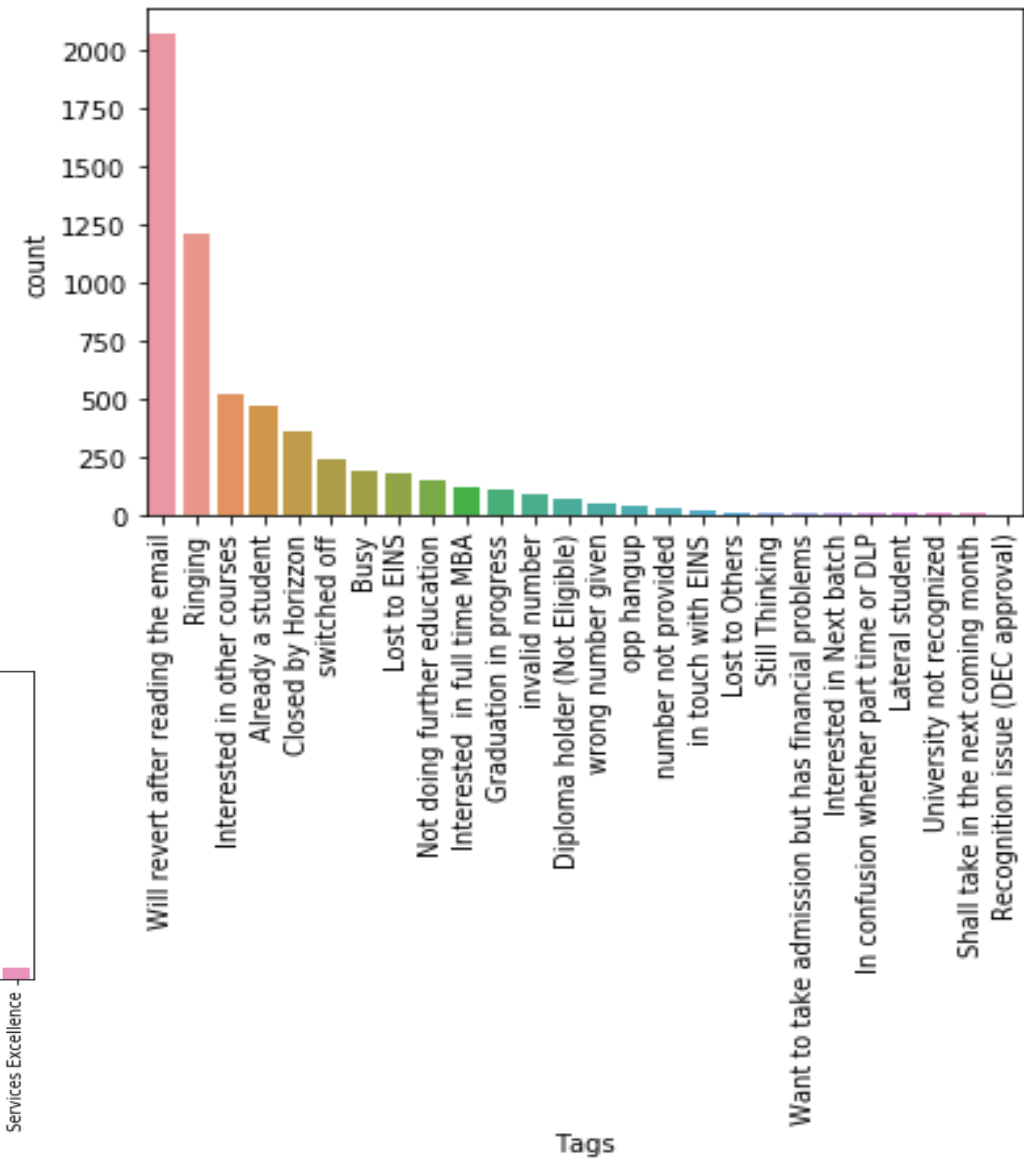
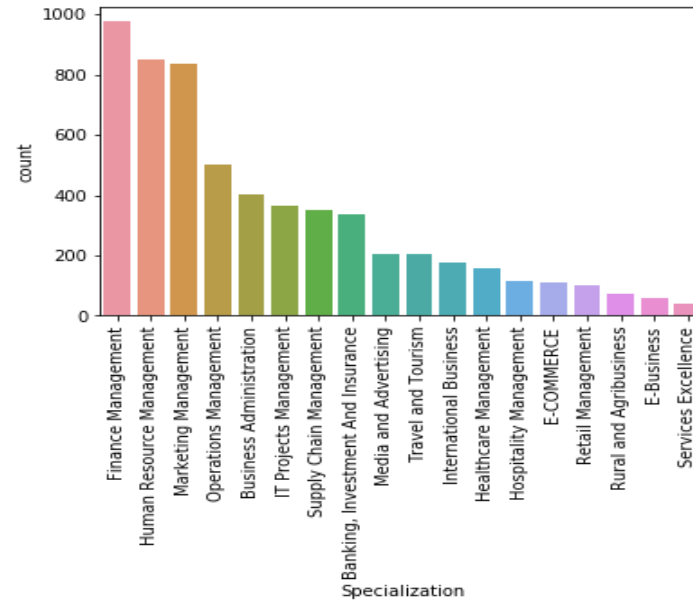
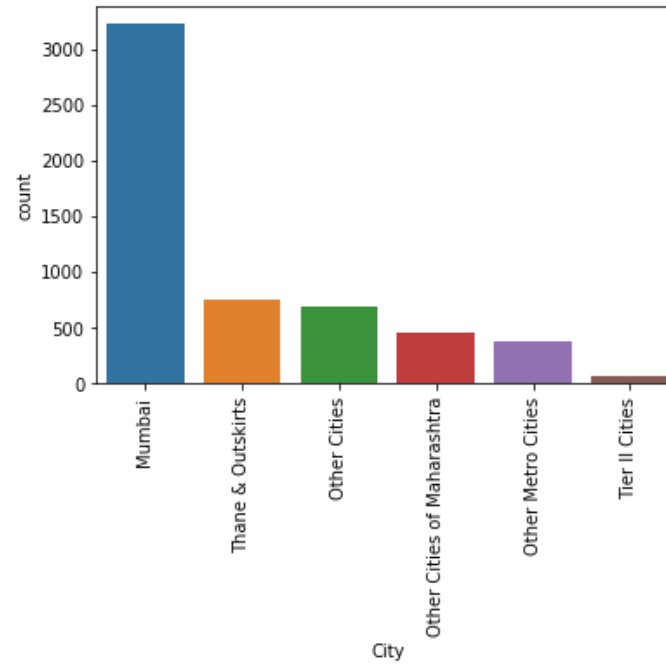
Identify influencing features

- Identify based on Logistic regression model
- Draw Conclusion and recommendations for model.



Data Cleaning

- Checked for duplicated values
- Replaced Select with NaN
- Dropping unnecessary columns with only null values, single unique feature, rating columns.
- Imputed Values with highest count in particular columns
- Segregated all NA values into others as separate entity.
- Highly skewed columns were dropped.



Exploratory Data Analysis

– Numerical Variable

Following are observations

Total visits

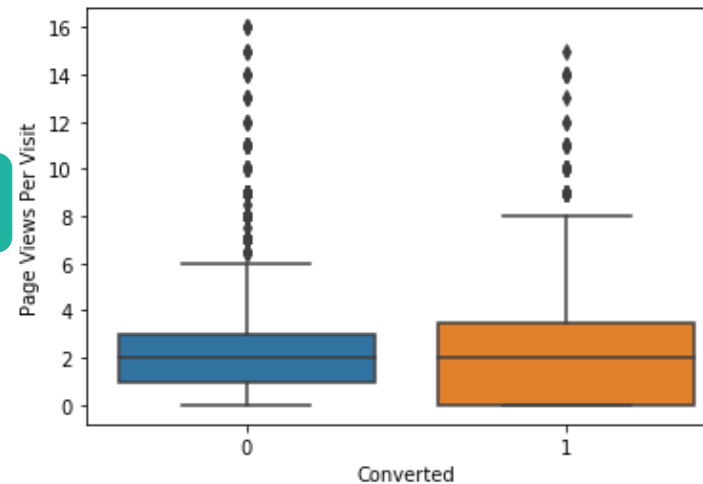
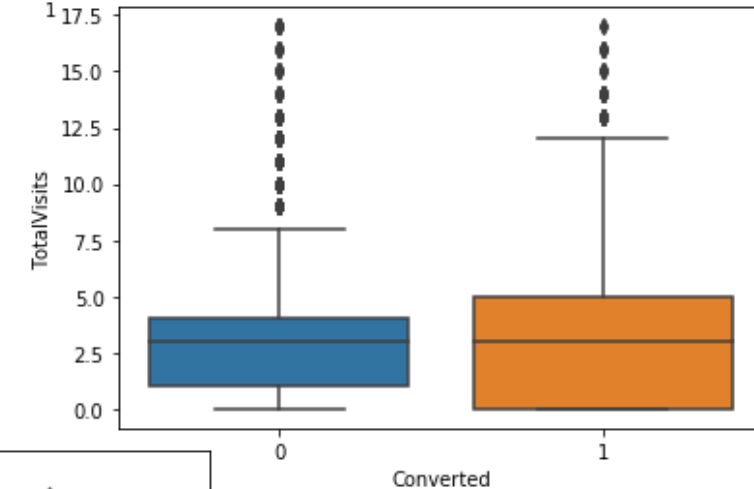
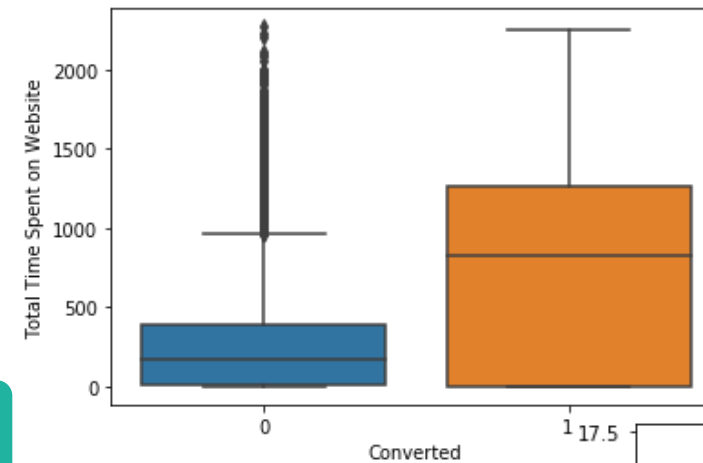
- Median for converted and non converted leads are same.
- Nothing conclusive based on Total Visits.

Total Time spent on website

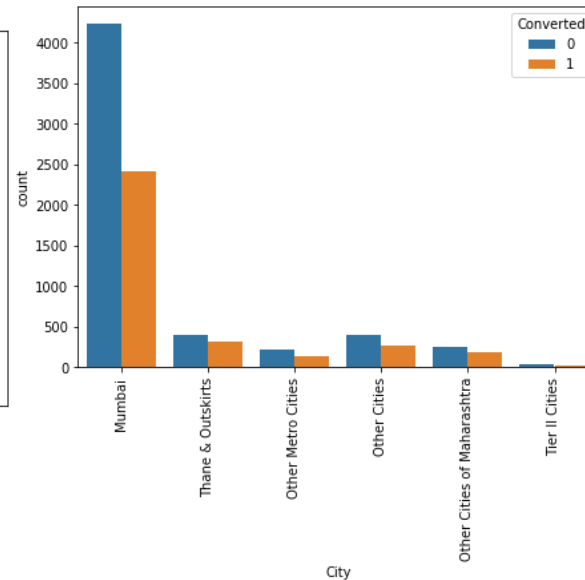
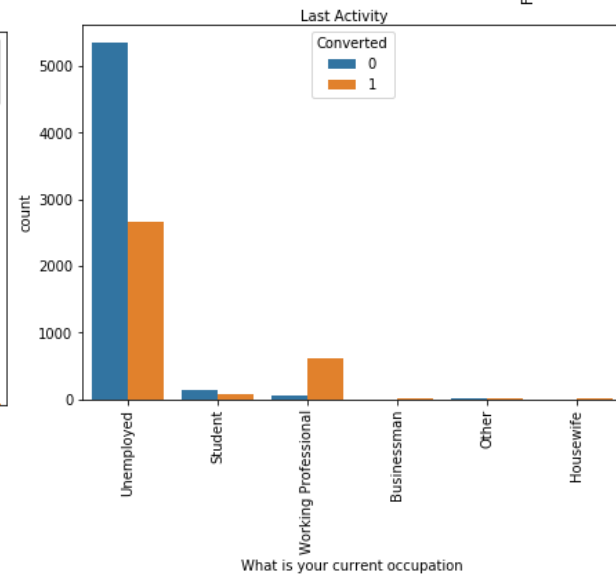
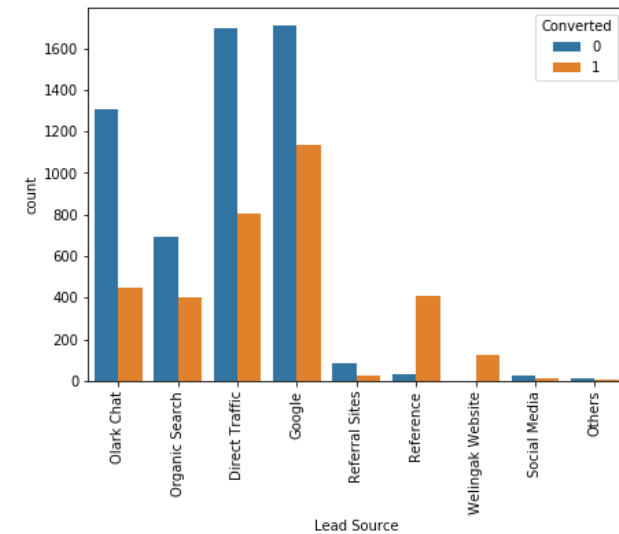
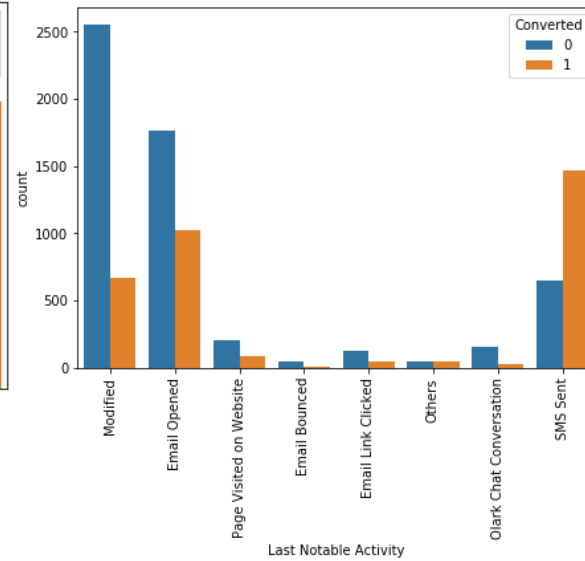
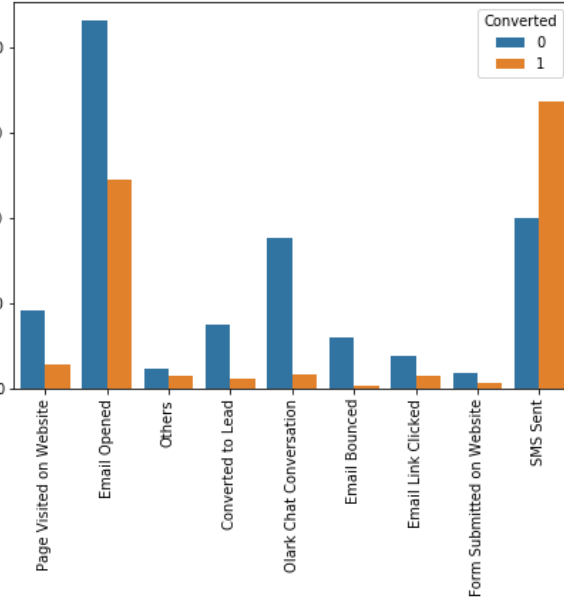
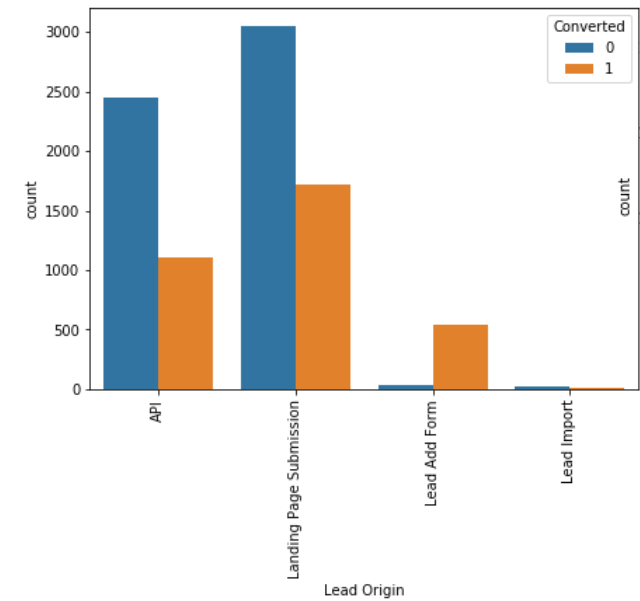
- Leads spending more time on the website are more likely to be converted
- Website should be made more engaging to make leads spend more time.

Page View Per Visit

- No concrete reading from the page view per visit graphs.



Exploratory Data Analysis – Univariate Analysis



Distribution of Lead Origin

Landing page submission is comparatively high than the rest of the categories, lead form has high certainty in lead conversion.

Distribution of Lead Source

Google is the best lead source among all categories in the lead source, Direct traffic, Olark Chat and organic search are some of the best entities in lead source.

Distribution of occupation

Working professionals and Unemployed going for the course have high chances of joining it

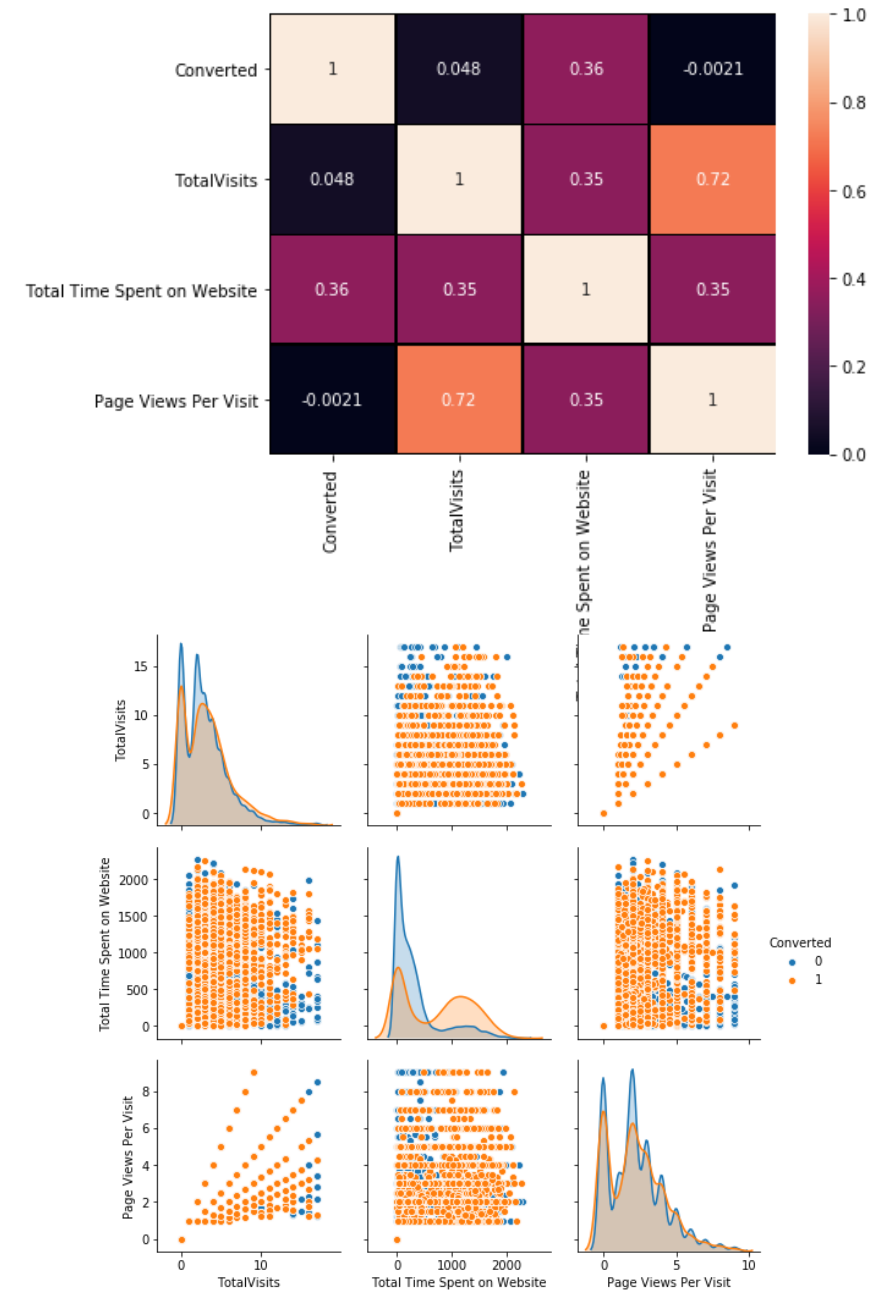
Distribution of the city

Most leads are from Mumbai city with maximum conversion count from the same

Exploratory Data Analysis

– Bi variate Analysis

- From pair plot we can observe clearly that our dataset has highly skewed values with lot of random peaks.
- With heat map we can inference Total visits and Page view per visit has high correlation than other features.
- Total Visits and converted has very low correlation.
- Total Visits and Total time spent on website has a reasonable correlation.
- There is positive correlation between total time spent on website and conversion.
- There is almost no correlation in Page views per visit and total visits with conversion.



Model Building

- For Model building we need to scale and split data into train and test dataset.
- We will be using Logistic Regression for building the model.
- Variable selection done through RFE(recursive feature elimination) and further we remove features with high p value and VIF value.
- Analyzing various parameters for train dataset Specificity, Sensitivity, Accuracy, Precision and Recall for train data.
- Plot the ROC Curve which shows trade off between sensitivity and specificity.

```
col = X_train.columns[rfe.support_]
```

```
col
```

```
Index(['Total Time Spent on Website', 'LO_Lead Add Form', 'LS_Olark Chat',  
      'LS_Welingak Website', 'LA_SMS Sent', 'Tags_Already a student',  
      'Tags_Busy', 'Tags_Closed by Horizzon', 'Tags_Lost to EINS',  
      'Tags_Not Specified', 'Tags_Ringing',  
      'Tags_Will revert after reading the email', 'Tags_switched off',  
      'LNA_Modified', 'LNA_Olark Chat Conversation'],  
      dtype='object')
```

```
X_train.columns[~rfe.support_]
```

```
Index(['Do Not Email', 'TotalVisits', 'Page Views Per Visit',  
      'A free copy of Mastering The Interview', 'LO_Landing Page Submission',  
      'LO_Lead Import', 'LS_Direct Traffic', 'LS_Google', 'LS_Organic Search',  
      'LS_Reference', 'LS_Referral Sites', 'LS_Social Media',  
      'LA_Converted to Lead', 'LA_Email Bounced', 'LA_Email Link Clicked',  
      'LA_Email Opened', 'LA_Form Submitted on Website',  
      'LA_Olark Chat Conversation', 'LA_Page Visited on Website',  
      'S_Banking, Investment And Insurance', 'S_Business Administration',  
      'S_E-Business', 'S_E-COMMERCE', 'S_International Business',  
      'S_Management Specialization', 'S_Media and Advertising',  
      'S_Rural and Agribusiness', 'S_Services Excellence',  
      'S_Travel and Tourism', 'CO_Businessman', 'CO_Housewife', 'CO_Student',  
      'CO_Unemployed', 'CO_Working Professional',  
      'Tags_Graduation in progress', 'Tags_Interested in full time MBA',  
      'Tags_Interested in other courses', 'Tags_Not doing further education',  
      'City_Mumbai', 'City_Other Cities', 'City_Other Cities of Maharashtra',  
      'City_Other Metro Cities', 'City_Thane & Outskirts',  
      'LNA_Email Bounced', 'LNA_Email Link Clicked', 'LNA_Email Opened',  
      'LNA_Page Visited on Website', 'LNA_SMS Sent'],  
      dtype='object')
```


Logistic Regression Model

Using RFE and Manual feature elimination for features having P-value more than 0.05 and VIF more than 5. We reached a final model with P - Value less than 0.051 and VIF less than 5.

Generalized Linear Model Regression Results			
Dep. Variable:	Converted	No. Observations:	6246
Model:	GLM	Df Residuals:	6230
Model Family:	Binomial	Df Model:	15
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1236.1
Date:	Fri, 04 Dec 2020	Deviance:	2472.3
Time:	21:51:25	Pearson chi2:	8.86e+03
No. Iterations:	8		
Covariance Type:	nonrobust		

							Features	VIF
	coef	std err	z	P> z	[0.025	0.975]		
const	-3.5312	0.222	-15.925	0.000	-3.966	-3.097	1 LO_Lead Add Form	1.77
Total Time Spent on Website	1.0779	0.061	17.634	0.000	0.958	1.198	9 Tags_Not Specified	1.69
LO_Lead Add Form	1.7357	0.421	4.127	0.000	0.911	2.560	11 Tags_Will revert after reading the email	1.66
LS_Olark Chat	1.3026	0.145	8.998	0.000	1.019	1.586	4 LA_SMS Sent	1.65
LS_Welingak Website	3.5853	0.849	4.223	0.000	1.921	5.249	2 LS_Olark Chat	1.64
LA_SMS Sent	1.9855	0.117	17.005	0.000	1.757	2.214	13 LNA_Modified	1.49
Tags_Already a student	-1.2401	0.635	-1.952	0.051	-2.485	0.005	0 Total Time Spent on Website	1.46
Tags_Busy	2.6859	0.307	8.737	0.000	2.083	3.288	3 LS_Welingak Website	1.32
Tags_Closed by Horizzon	8.5865	0.765	11.220	0.000	7.087	10.086	7 Tags_Closed by Horizzon	1.21
Tags_Lost to EINS	7.4777	0.572	13.072	0.000	6.357	8.599	10 Tags_Ringing	1.12
Tags_Not Specified	1.9071	0.219	8.697	0.000	1.477	2.337	5 Tags_Already a student	1.08
Tags_Ringing	-1.6067	0.316	-5.078	0.000	-2.227	-0.987	14 LNA_Olark Chat Conversation	1.08
Tags_Will revert after reading the email	6.4811	0.277	23.406	0.000	5.938	7.024	8 Tags_Lost to EINS	1.07
Tags_switched off	-2.2534	0.769	-2.929	0.003	-3.761	-0.746	6 Tags_Busy	1.05
LNA_Modified	-1.7770	0.127	-13.969	0.000	-2.026	-1.528	12 Tags_switched off	1.03
LNA_Olark Chat Conversation	-1.8176	0.422	-4.303	0.000	-2.646	-0.990		

Final Logistic Regression Model

Our final model has P-values tending to 0 and VIF values less than 2 which suggests that the model is good to go and can be used to make predictions on the test data.

Generalized Linear Model Regression Results

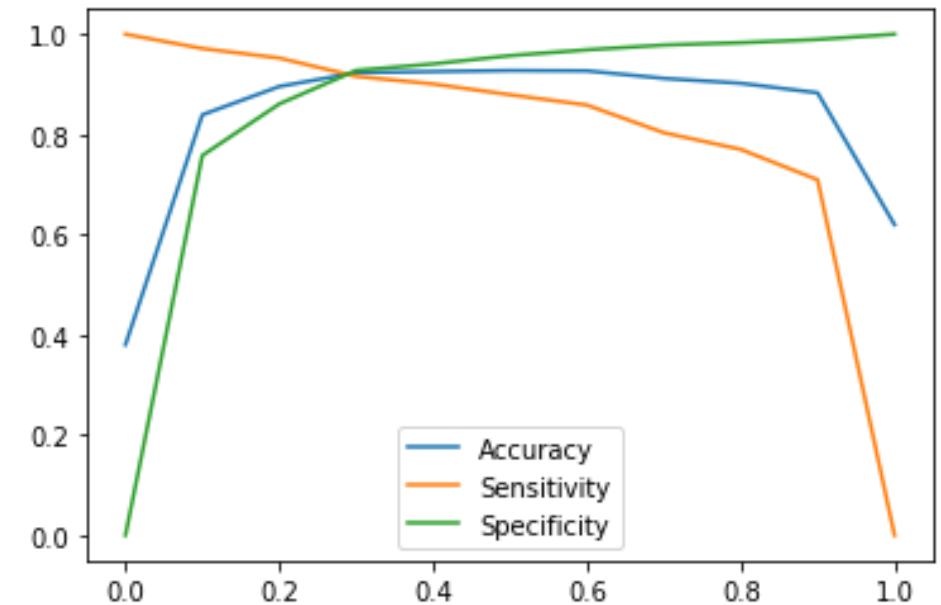
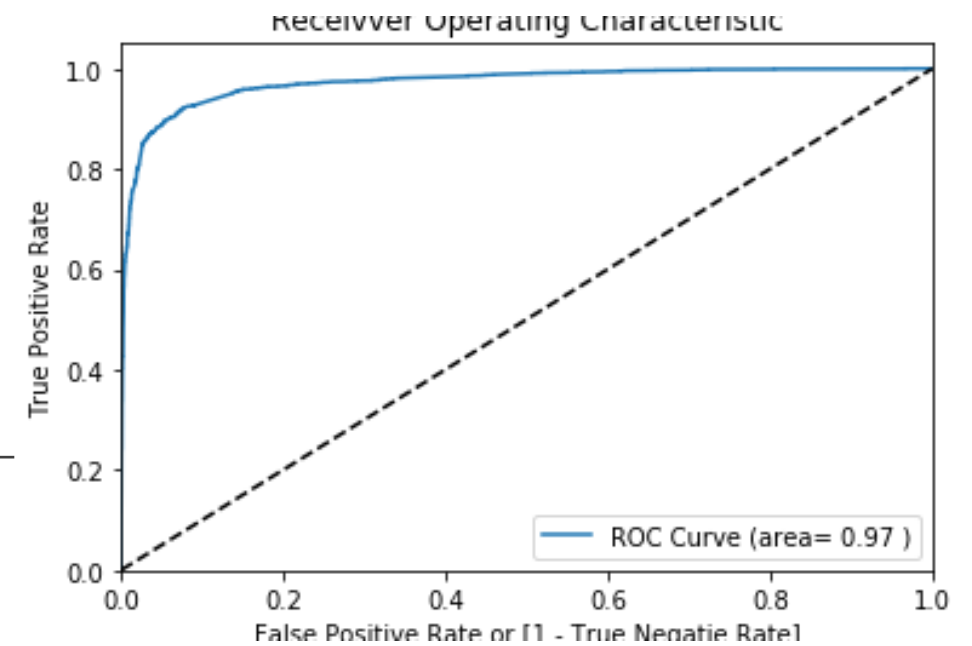
Dep. Variable:	Converted	No. Observations:	6246
Model:	GLM	Df Residuals:	6231
Model Family:	Binomial	Df Model:	14
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1238.6
Date:	Sun, 06 Dec 2020	Deviance:	2477.3
Time:	05:54:27	Pearson chi2:	8.87e+03
No. Iterations:	8		
Covariance Type:	nonrobust		

		coef	std err	z	P> z	[0.025	0.975]		Features	VIF
	const	-3.7490	0.208	-18.040	0.000	-4.156	-3.342	1	LO_Lead Add Form	1.77
	Total Time Spent on Website	1.0712	0.061	17.622	0.000	0.952	1.190	10	Tags_Will revert after reading the email	1.66
	LO_Lead Add Form	1.7479	0.424	4.118	0.000	0.916	2.580	4	LA_SMS Sent	1.65
	LS_Olark Chat	1.2885	0.144	8.927	0.000	1.006	1.571	8	Tags_Not Specified	1.65
	LS_Welingak Website	3.5634	0.851	4.187	0.000	1.895	5.231	2	LS_Olark Chat	1.62
	LA_SMS Sent	2.0096	0.117	17.214	0.000	1.781	2.238	0	Total Time Spent on Website	1.46
	Tags_Busy	2.8922	0.299	9.664	0.000	2.306	3.479	12	LNA_Modified	1.42
	Tags_Closed by Horizon	8.7931	0.762	11.538	0.000	7.299	10.287	3	LS_Welingak Website	1.32
	Tags_Lost to EINS	7.6899	0.567	13.554	0.000	6.578	8.802	6	Tags_Closed by Horizon	1.21
	Tags_Not Specified	2.1195	0.206	10.277	0.000	1.715	2.524	9	Tags_Ringing	1.12
	Tags_Ringing	-1.4025	0.308	-4.548	0.000	-2.007	-0.798	7	Tags_Lost to EINS	1.07
	Tags_Will revert after reading the email	6.6925	0.267	25.053	0.000	6.169	7.216	13	LNA_Olark Chat Conversation	1.07
	Tags_switched off	-2.0495	0.766	-2.676	0.007	-3.550	-0.549	5	Tags_Busy	1.05
	LNA_Modified	-1.7700	0.127	-13.908	0.000	-2.019	-1.521	11	Tags_switched off	1.03
	LNA_Olark Chat Conversation	-1.8100	0.421	-4.295	0.000	-2.636	-0.984			

Plotting ROC Curve

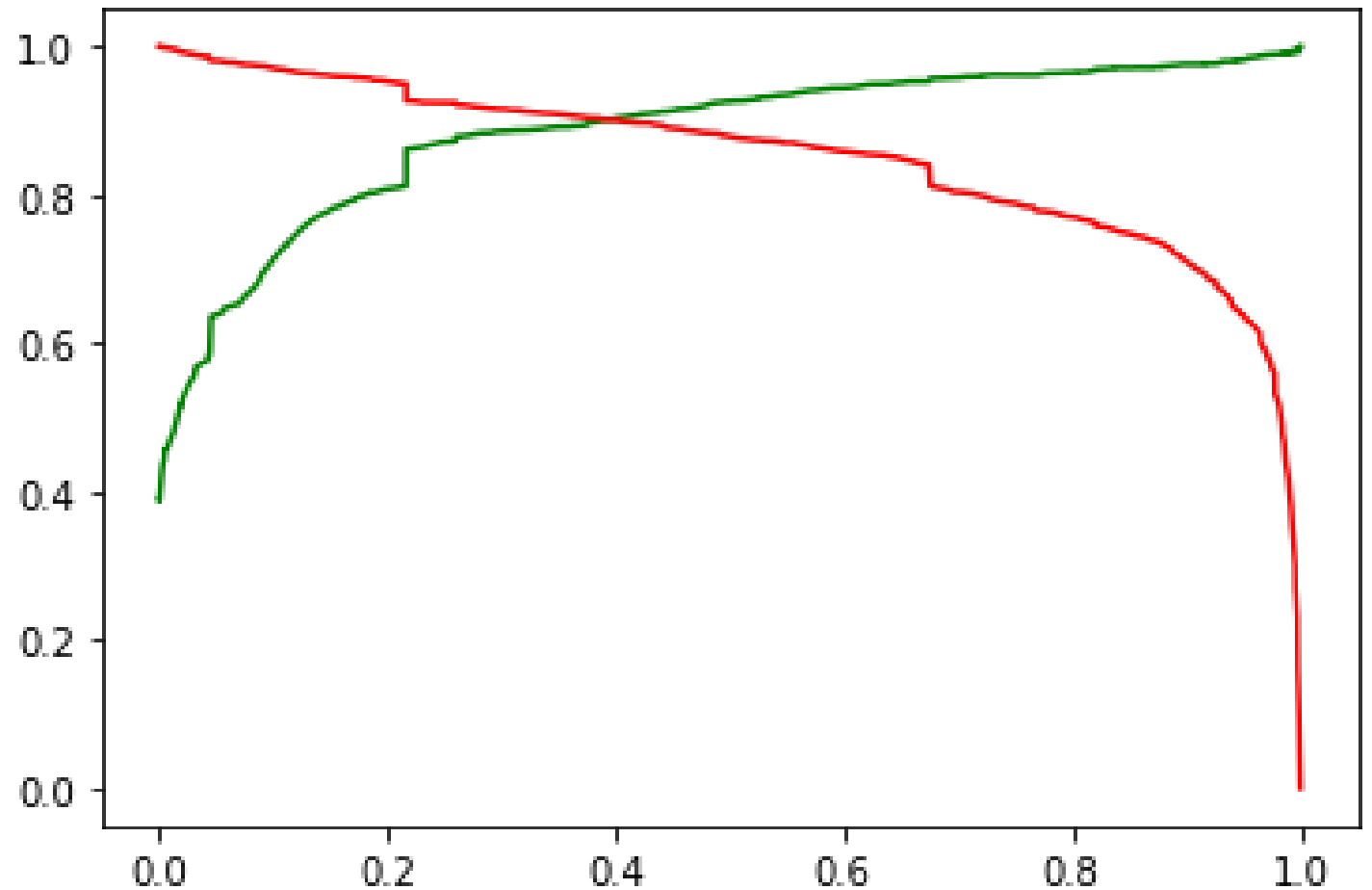
ROC curve demonstrates several things

- It shows trade off between sensitivity and specificity.
- The closer the curve follows left hand border and then the top border of the ROC space, this proves better accuracy of the test.
- The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.



Precision and recall trade off

As per Precision-Recall trade off, the cut off is around 0.4.



Lead Score for varying cut off probability

Here we will examine the projected lead score for different cut off probability to estimate the lead. So, this will be a useful template to change the cut off based on business needs.

```
# Adding Lead Score Column as per buisness requirement
```

```
y_test_pred_final['Lead Score'] = y_test_pred_final['Conversion_prob']*100  
y_test_pred_final.head()
```

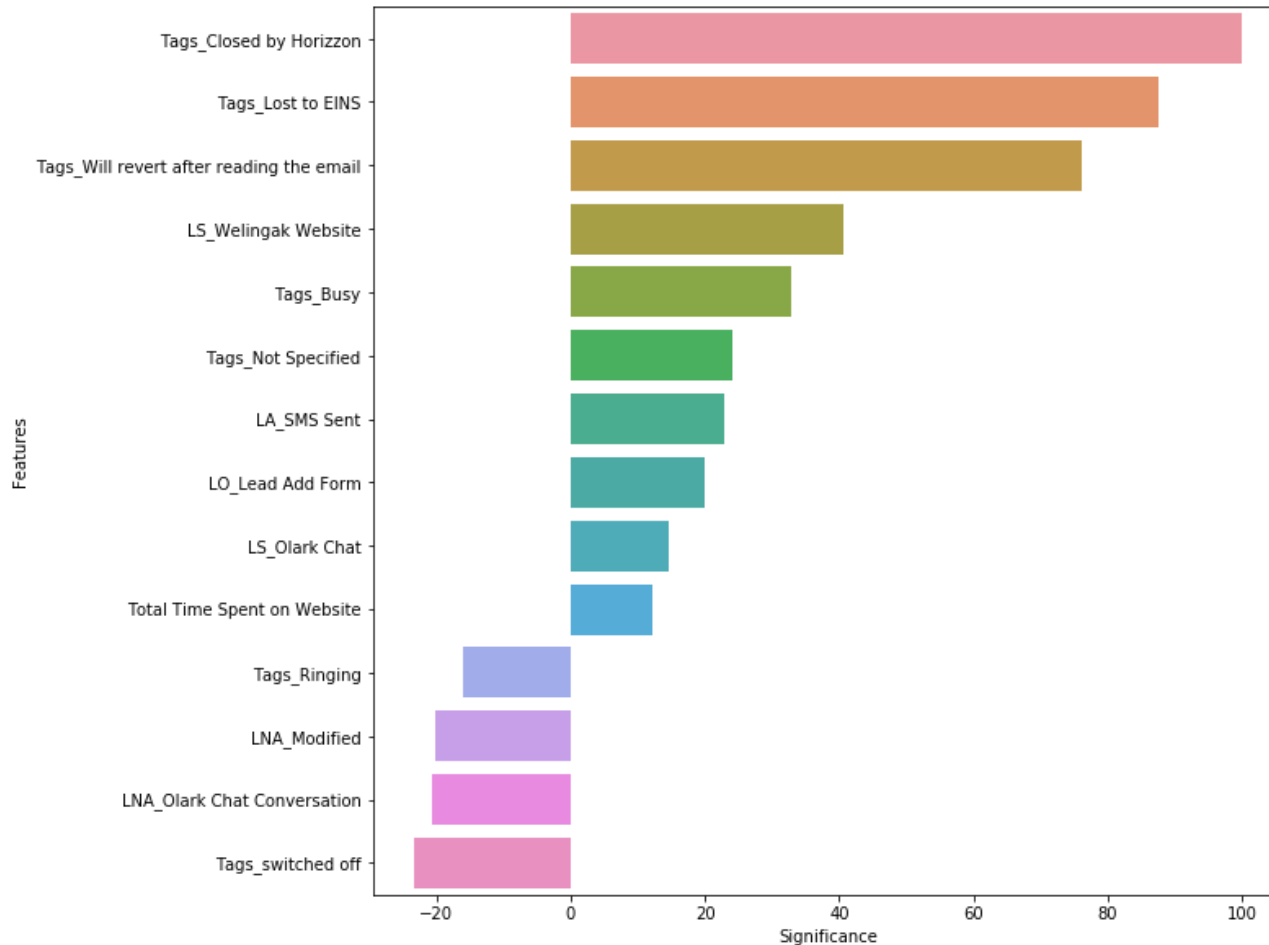
	StudentID	Converted	Conversion_prob	final_predicted	Lead Score
0	7625	0	0.178961	0	17.896082
1	5207	1	0.976689	1	97.668887
2	2390	1	0.996859	1	99.685870
3	4362	0	0.025840	0	2.583958
4	1023	0	0.016779	0	1.677865

```
# Let's also sort the leads by lead score to enable the team to close hot leads as per the lead score
```

```
y_test_pred_final.sort_values(by = 'Lead Score', ascending = False, inplace = True)  
y_test_pred_final.head()
```

	StudentID	Converted	Conversion_prob	final_predicted	Lead Score
779	4062	1	0.999974	1	99.997423
326	3339	1	0.999936	1	99.993584
2582	8103	1	0.999918	1	99.991847
218	6028	1	0.999860	1	99.986017
1936	2354	1	0.999807	1	99.980666

Relative Feature Importance and Recommendations



Top 3 variables that contribute the most towards the probability of a lead getting converted are:

- Tags_Closed by Horizzon
- Tags_lost to EINS
- Tags_will revert after reading the email

X education company needs to focus on following key aspects to improve overall conversion rate:

- Focus on the top 3 tags which are very positive for business.
- Focus on working professional who have high conversion rate.
- Increase user engagement on wellingak website since it helps higher conversion.
- Improving lead add form also improves lead conversion with high certainty