Lead Scoring Case Study

Submitted By -

Mukesh Chaurasia

Chhavi Kansal

Business Objective

- To help X Education to select the most promising leads(Hot Leads), i.e. the leads that are most likely to convert into paying customers.
- To build a logistic regression model to assign a lead score value between 0 and 100 to each of the leads which can be used by the company to target potential leads.

The objective is thus classified into the following sub-goals:

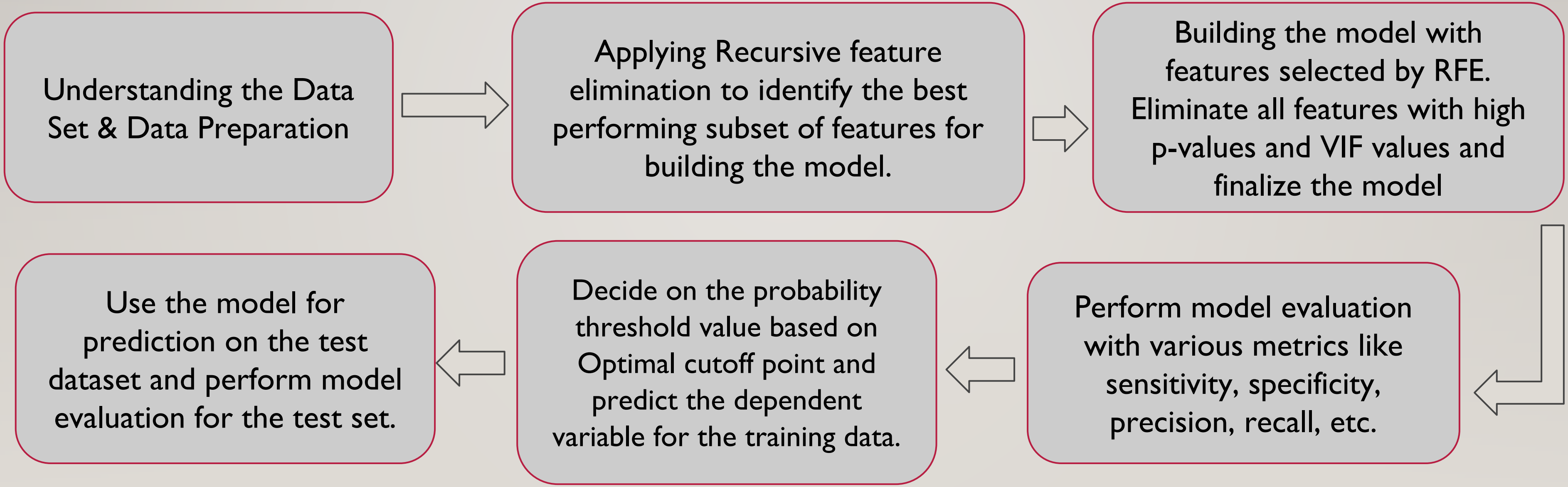
Create a Logistic Regression model to predict the Lead conversion probabilities for each lead.

Decide on a probability threshold value above which a lead will be predicted as converted, whereas not converted if it is below it

Multiply the Lead Conversion probability to arrive at the Lead Score value for each lead.

Problem Solving Methodology

The approach has been to divide the entire case study into various checkpoints to meet each of the sub-goals. The checkpoints are represented in a sequential flow as below:



Understanding the Data Set & Data Preparation

Applying Recursive feature elimination to identify the best performing subset of features for building the model.

Building the model with features selected by RFE. Eliminate all features with high p-values and VIF values and finalize the model

Perform model evaluation with various metrics like sensitivity, specificity, precision, recall, etc.

Decide on the probability threshold value based on Optimal cutoff point and predict the dependent variable for the training data.

Use the model for prediction on the test dataset and perform model evaluation for the test set.

Data Preparation

Replace select values with NAN in columns :
There are 4 columns(Specialization, How did you hear about X Education, Lead Profile and City) which has select values, we replaced select values with NAN values.

Drop the columns with more than 50% missing values:
Columns having more than 50% missing values have no use, so we have dropped all 3 columns ('How did you hear about X Education', 'Lead Profile' and 'Lead Quality')having more than 50% missing values.

Data Preparation contd...

Drop skewed & sales team generated columns
we have dropped the columns with skewed values (means having one value only or one value at very high percentage say more than 80%) and sales team generated columns. Such columns are 'Asymmetrique Activity Index', 'Asymmetrique Profile Index','Asymmetrique Activity Score','Asymmetrique Profile Score','Do Not Email', 'Do Not Call', What matters most to you in choosing a course, Search, Magazine, Newspaper Article, X Education Forums, Newspaper, Digital Advertisement, Through Recommendations, Receive More Updates About Our Courses, Update me on Supply Chain Content, Get updates on DM Content, and I agree to pay the amount through cheque.

Data Preparation contd…

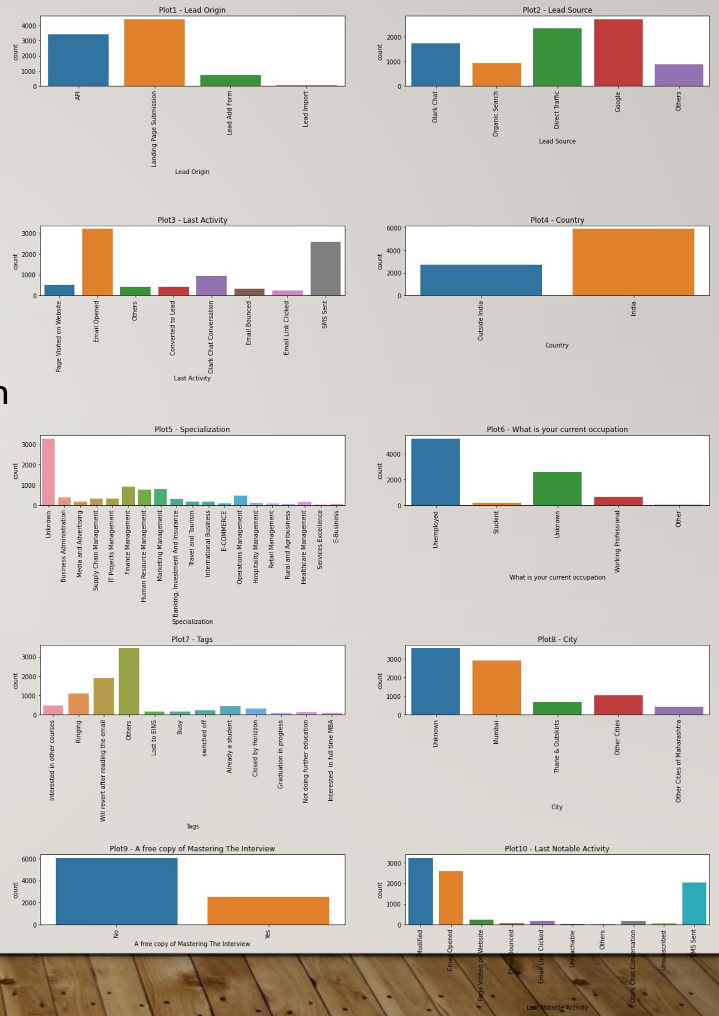
Impute the missing values:
The columns ‘TotalVisits’ and ‘Page Views Per Visit’ are continuous variables with outliers. Hence the null values for these columns were imputed with the column median values.

Merge the low percentage labels into separate category:
There are column labels having very less values, we have merged all such labels which has less than 1% values.

Outlier treatment:
The outliers present in the columns 'TotalVisits‘ & 'Page Views Per Visit‘ were finally removed based on interquartile Outlier range analysis.

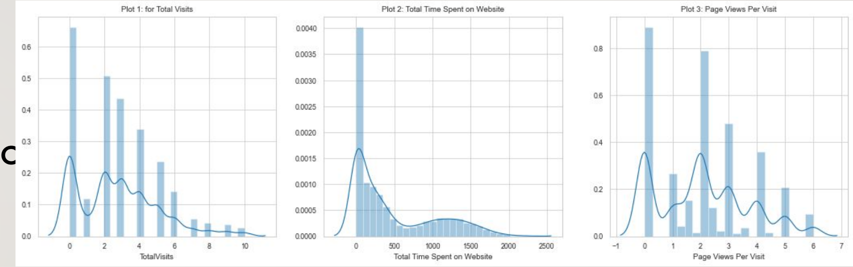
EDA (Univariate analysis Categorical variables)

- Major leads are generated through landing page submission or through API integration
- Google is the most preferred lead generation source. Direct traffic and olark chat are also contributed significantly in lead generation.
- Email open and SMS sent are main last activity while email link clicked is least.
- 70% leads are from India rest are from different countries other than India.
- Finance management is the most preferred specialization while services excellence is the least
- Most of the leads has occupation as unemployed. Few are the student too.
- Revert after reading the email are most preferred tag while graduation in progress is least preferred.
- Most of the leads are from Mumbai and nearby cities only.
- 70% leads are not opted A free copy of Mastering The Interview
- Modified, email opened and sms sent are most last notable activity.



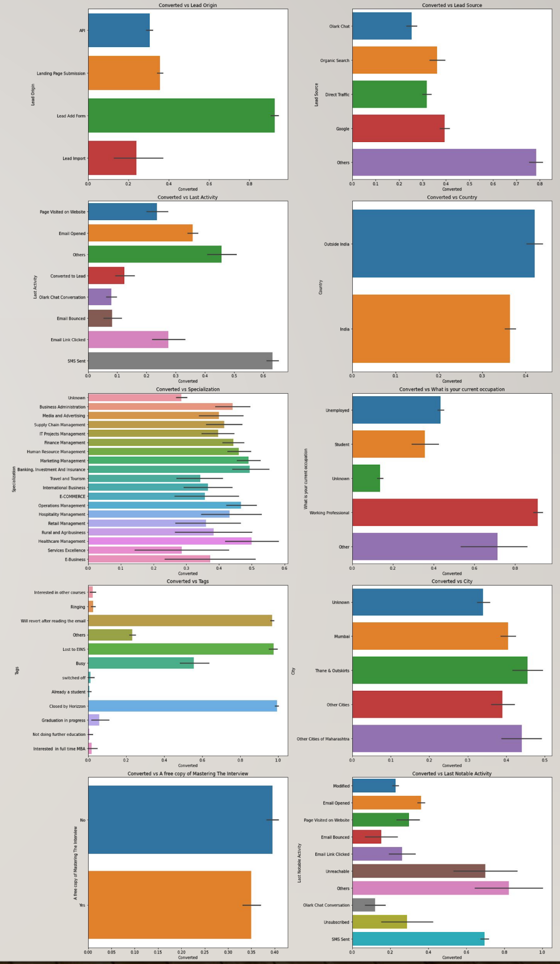
EDA (Univariate analysis Continuous variables)

- Many leads has visited 2 and 3 times on website. Few have not visited website too.
- Most of the leads has average time less than 500 millisecond while few leads has avg time between 1000 to 1500 milliseconds too.
- Many leads has visited 2 page before getting captured while few have not visited any page.



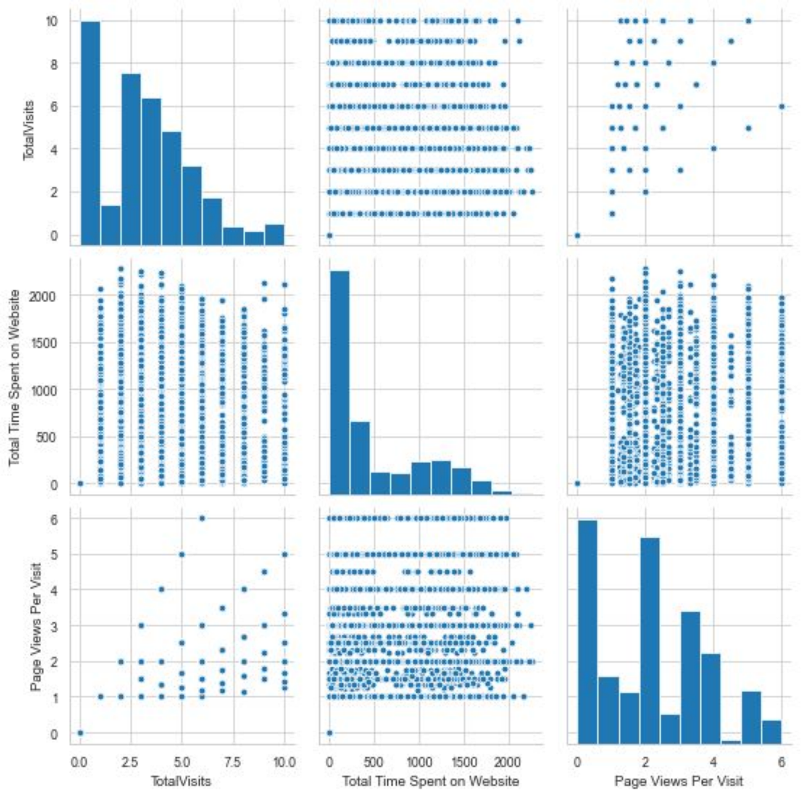
EDA (Bivariate analysis Categorical variables)

- Leads generated through `lead ad form` are converted most while leads generated through `lead import` are converted least.
- Leads generated through `Google` and `Organic` source are converted most.
- Leads as last activity SMS sent are converted most
- There are significant conversion in leads from `outside India`.
- `Marketing management`, `banking investment and insurance` and `health care management` leads are better converted.
- `Working professional` leads are most converted.
- Will revert after reading the email and lost to EINS are the most preferred tags for lead conversion
- Leads from `Thane and outskirts` are most converted
- Leads opted `A free copy of Mastering The Interview` converted significantly
- Leads with `Last notable activity` as SMS sent are converted most.



EDA (Bivariate analysis Continuous variables)

- There is no correlation between total visits and total time spent on website
- Total visits and page views per visit has positive correlation. As total visits increases page visits also increases.



Data Preparation for Modeling

Binary Encoding:

All skewed binary variables(Yes/No) are already dropped. Only ‘A free copy of Mastering The Interview' is converted into 1/0.

Dummy Encoding:

Dummy variables are created for all categorical columns. (Country, Lead Origin, Lead Source, Tags, What is your current occupation, Specialization, City, Last Activity and Last Notable Activity)

Test-train split:

The original dataframe was split into train and test dataset. The train dataset was used to train the model and test dataset was used to evaluate the model.

Data Preparation for Modeling Contd...

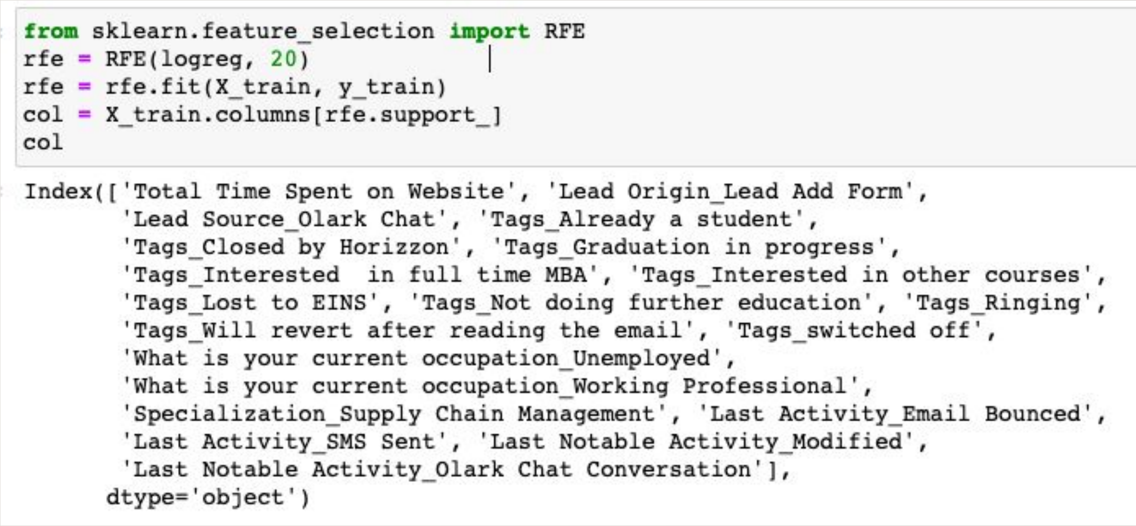
Feature Scaling:

Scaling helps in interpretation. It is important to have all variables(specially categorical ones which has values 0 and 1) on the same scale for the model to be easily interpretable.

‘Standardisation’ was used to scale the data for modelling. It basically brings all of the data into a standard normal distribution with mean at zero and standard deviation one.

Feature Selection using RFE

Recursive feature elimination is an optimization technique for finding the best performing subset of features. It is based on the idea of repeatedly constructing a model and choosing either the best (based on coefficients), setting the feature aside and then repeating the process with the rest of the features. This process is applied until all the features in the dataset are exhausted. Features are then ranked according to when they were eliminated.



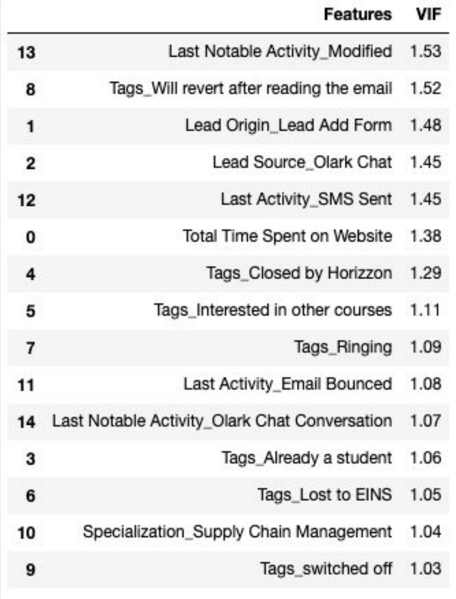
```
from sklearn.feature_selection import RFE
rfe = RFE(logreg, 20)
rfe = rfe.fit(X_train, y_train)
col = X_train.columns[rfe.support_]
col
```

```
Index(['Total Time Spent on Website', 'Lead Origin_Lead Add Form',
       'Lead Source_Olark Chat', 'Tags_Already a student',
       'Tags_Closed by Horizzon', 'Tags_Graduation in progress',
       'Tags_Interested  in full time MBA', 'Tags_Interested in other courses',
       'Tags_Lost to EINS', 'Tags_Not doing further education', 'Tags_Ringing',
       'Tags_Will revert after reading the email', 'Tags_switched off',
       'What is your current occupation_Unemployed',
       'What is your current occupation_Working Professional',
       'Specialization_Supply Chain Management', 'Last Activity_Email Bounced',
       'Last Activity_SMS Sent', 'Last Notable Activity_Modified',
       'Last Notable Activity_Olark Chat Conversation'],
      dtype='object')
```

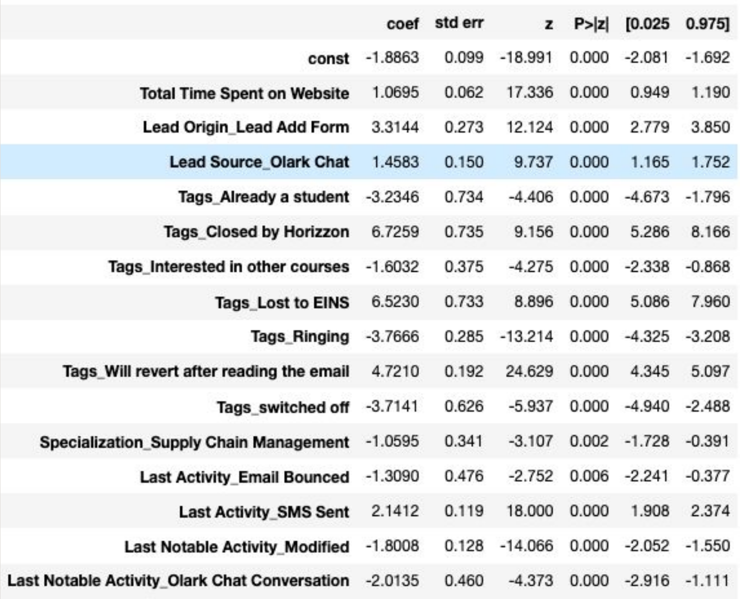
Running RFE with the output number of the variable equal to 20.

Model Building

- Generalized Linear Models from StatsModels is used to build the Logistic Regression model.
- The model is built initially with the 20 variables selected by RFE.
- Unwanted features are dropped serially after checking p values (< 0.5) and VIF (< 3) and model is built multiple times.
- The final model with 15 features, passes both the significance test and the multicollinearity test.



	Features	VIF
13	Last Notable Activity_Modified	1.53
8	Tags_Will revert after reading the email	1.52
1	Lead Origin_Lead Add Form	1.48
2	Lead Source_Olark Chat	1.45
12	Last Activity_SMS Sent	1.45
0	Total Time Spent on Website	1.38
4	Tags_Closed by Horizzon	1.29
5	Tags_Interested in other courses	1.11
7	Tags_Ringing	1.09
11	Last Activity_Email Bounced	1.08
14	Last Notable Activity_Olark Chat Conversation	1.07
3	Tags_Already a student	1.06
6	Tags_Lost to EINS	1.05
10	Specialization_Supply Chain Management	1.04
9	Tags_switched off	1.03

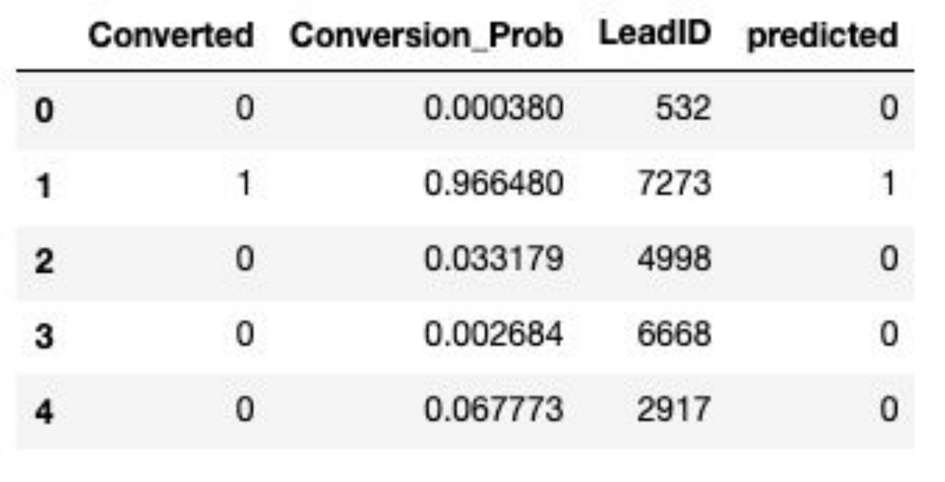


	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.8863	0.099	-18.991	0.000	-2.081	-1.692
Total Time Spent on Website	1.0695	0.062	17.336	0.000	0.949	1.190
Lead Origin_Lead Add Form	3.3144	0.273	12.124	0.000	2.779	3.850
Lead Source_Olark Chat	1.4583	0.150	9.737	0.000	1.165	1.752
Tags_Already a student	-3.2346	0.734	-4.406	0.000	-4.673	-1.796
Tags_Closed by Horizzon	6.7259	0.735	9.156	0.000	5.286	8.166
Tags_Interested in other courses	-1.6032	0.375	-4.275	0.000	-2.338	-0.868
Tags_Lost to EINS	6.5230	0.733	8.896	0.000	5.086	7.960
Tags_Ringing	-3.7666	0.285	-13.214	0.000	-4.325	-3.208
Tags_Will revert after reading the email	4.7210	0.192	24.629	0.000	4.345	5.097
Tags_switched off	-3.7141	0.626	-5.937	0.000	-4.940	-2.488
Specialization_Supply Chain Management	-1.0595	0.341	-3.107	0.002	-1.728	-0.391
Last Activity_Email Bounced	-1.3090	0.476	-2.752	0.006	-2.241	-0.377
Last Activity_SMS Sent	2.1412	0.119	18.000	0.000	1.908	2.374
Last Notable Activity_Modified	-1.8008	0.128	-14.066	0.000	-2.052	-1.550
Last Notable Activity_Olark Chat Conversation	-2.0135	0.460	-4.373	0.000	-2.916	-1.111

Predicting the conversion Probability on Train dataset

- Creating a dataframe with the actual Converted flag and the predicted probabilities.

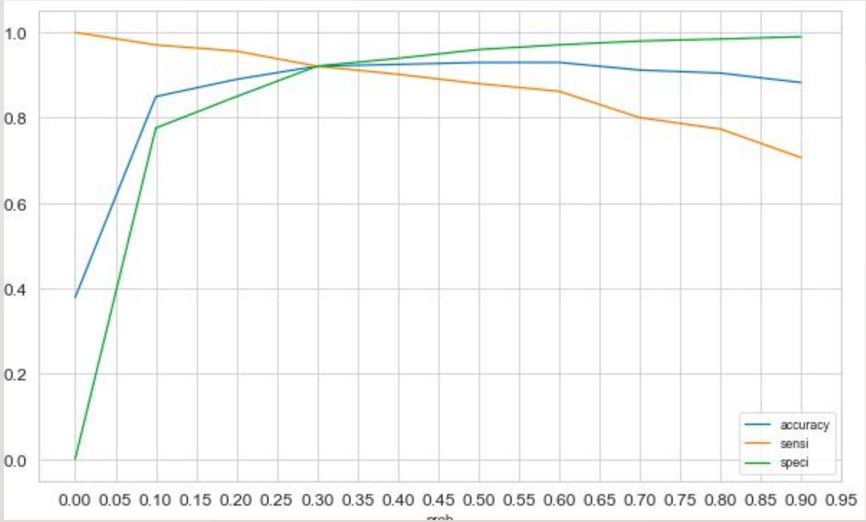
- Creating new column 'predicted' with 1 if Conversion_Prob > 0.5 else 0 Showing top 5 records of the dataframe in the picture on the right.



	Converted	Conversion_Prob	LeadID	predicted
0	0	0.000380	532	0
1	1	0.966480	7273	1
2	0	0.033179	4998	0
3	0	0.002684	6668	0
4	0	0.067773	2917	0

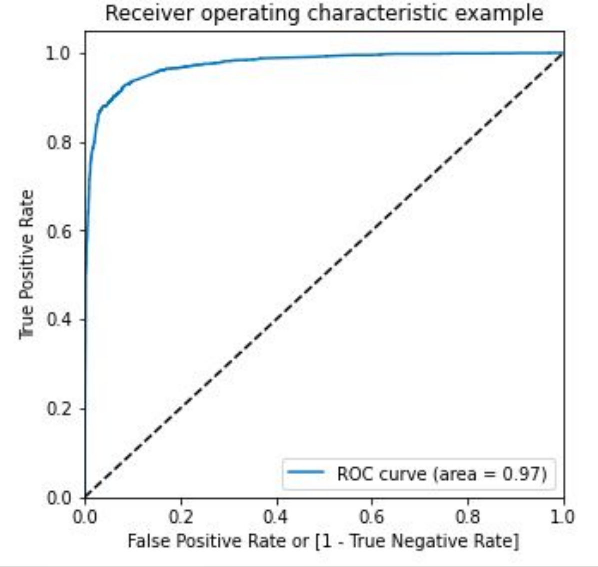
Finding optimum probability threshold

- The accuracy sensitivity and specificity was calculated for various values of probability threshold and plotted in the graph to the right.
- From the curve above, 0.30 is found to be the optimum point for cutoff probability.
- At this threshold value, all the 3 metrics - accuracy sensitivity and specificity was found to be well around 90% which is a well acceptable value.



Plotting the ROC curve and calculating AUC

- It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
- By determining the Area under the curve (AUC) of the ROC curve, the goodness of the model is determined. Since the ROC curve is more towards the upper-left corner of the graph, it means that the model is very good. The larger the AUC, the better will is the model.
- The value of AUC for our model is 0.9737



Evaluating the model on train dataset

Confusion Matrix

# Predicted	Not Converted	Converted
# Actual		
Not Converted	3593	151
Converted	274	2009

Probability Threshold
0.30

Accuracy
TP +TN/ (TP+TN+FN+FP)
0.9208

Sensitivity
TP / (TP+FN)
0.9207

Specificity
TN / (TN+FP)
0.9209

False Positive Rate
FP/ (TN+FP)
.0791

Positive Predictive Value
TP / (TP+FP)
0.8765

Negative Predictive Value
TN / (TN+ FN)
0.9501

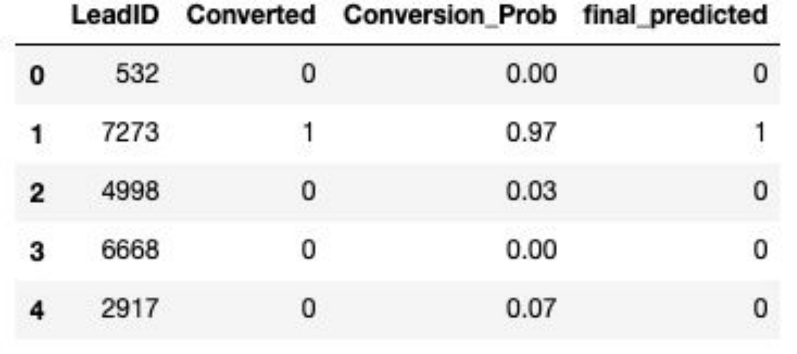
Precision
TP / TP + FP
0.8765

Recall
TP / TP + FN
0.9207

F1 score = 2× (Precision*Recall)/(Precision+ Recall)
0.8980

Making Prediction on test dataset

- The final model on the train dataset is used to make predictions for the test dataset
- The train data set was scaled using the scaler.transform function that was used to scale the train dataset.
- The Predicted probabilities were added to the leads in the test dataframe.
- Using the probability threshold value of 0.36, the leads from the test dataset were predicted if they will convert or not.



	LeadID	Converted	Conversion_Prob	final_predicted
0	532	0	0.00	0
1	7273	1	0.97	1
2	4998	0	0.03	0
3	6668	0	0.00	0
4	2917	0	0.07	0

Evaluating the model on Test dataset

Accuracy
TP +TN/ (TP+TN+FN+FP)

0.9052

Sensitivity
TP / (TP+FN)

0.8988

Specificity
TN / (TN+FP)

0.9092

False Positive Rate
FP/ (TN+FP)

.0907

Positive Predictive Value
TP / (TP+FP)

0.8637

Negative Predictive Value
TN / (TN+ FN)

0.9335

Precision
TP / TP + FP

0.8636

Recall
TP / TP + FN

0.8988

F1 score = 2× (Precision*Recall)/(Precision+ Recall)

0.8809

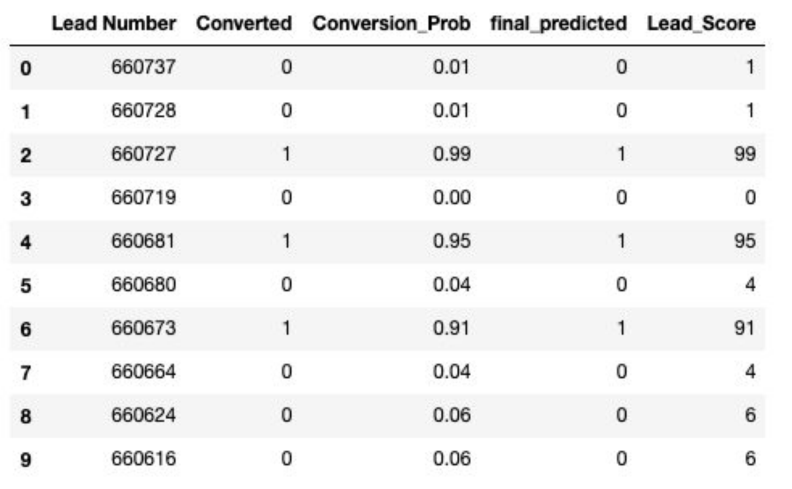
Lead Score calculation on complete data

Lead Score is calculated for all the leads in the original dataframe.

Formula for Lead Score calculation is:

Lead Score = 100 * Conversion Probability

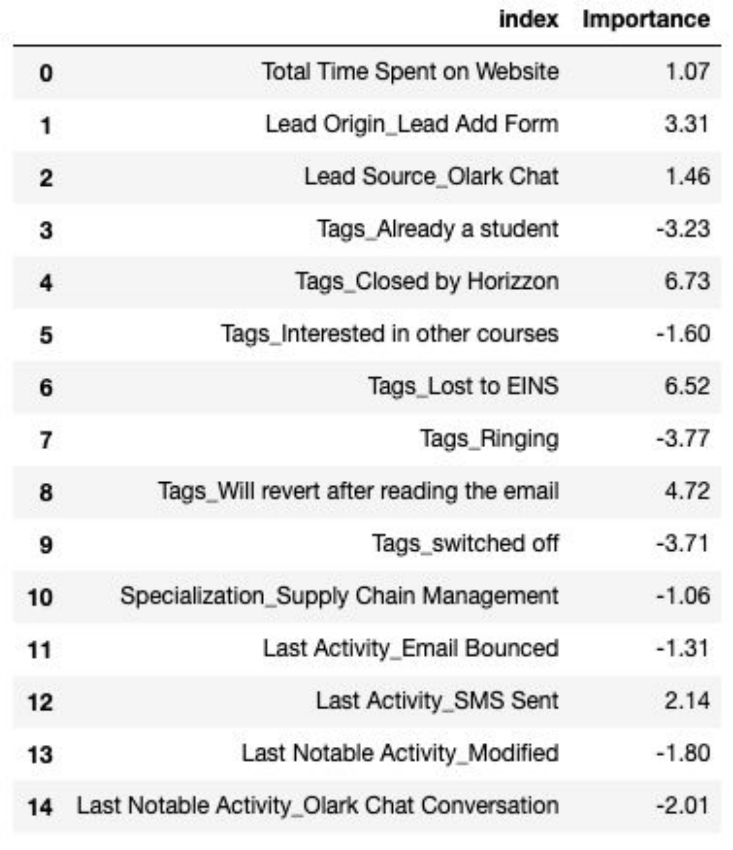
- The train and test dataset is concatenated to get the entire list of leads available.
- The Conversion Probability is multiplied by 100 to obtain the Lead Score for each lead.
- Higher the lead score, higher is the probability of a lead getting converted and vice versa,



	Lead Number	Converted	Conversion_Prob	final_predicted	Lead_Score
0	660737	0	0.01	0	1
1	660728	0	0.01	0	1
2	660727	1	0.99	1	99
3	660719	0	0.00	0	0
4	660681	1	0.95	1	95
5	660680	0	0.04	0	4
6	660673	1	0.91	1	91
7	660664	0	0.04	0	4
8	660624	0	0.06	0	6
9	660616	0	0.06	0	6

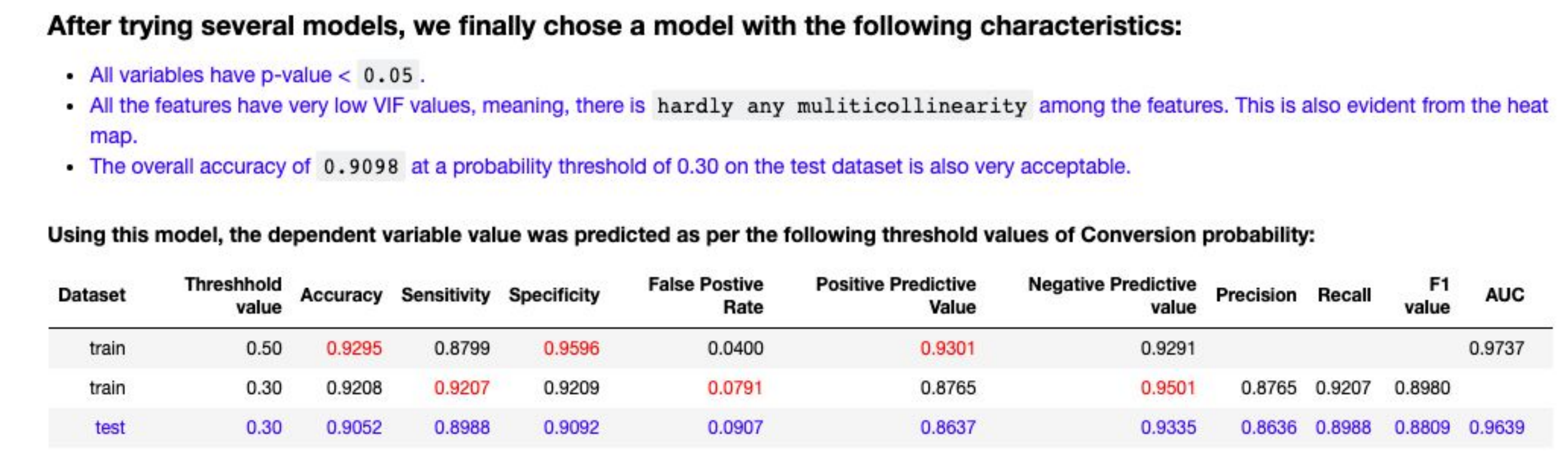
Determining Important Features

- 15 features have been used by our model to successfully predict if a lead will get converted or not.
- The Coefficient (beta) values for each of these features from the model parameters are used to determine the order of importance of these features.
- Features with high positive beta values are the ones that contribute most towards the probability of a lead getting converted.
- Similarly, features with high negative beta values contribute the least.



	index	Importance
0	Total Time Spent on Website	1.07
1	Lead Origin_Lead Add Form	3.31
2	Lead Source_Olark Chat	1.46
3	Tags_Already a student	-3.23
4	Tags_Closed by Horizzon	6.73
5	Tags_Interested in other courses	-1.60
6	Tags_Lost to EINS	6.52
7	Tags_Ringing	-3.77
8	Tags_Will revert after reading the email	4.72
9	Tags_switched off	-3.71
10	Specialization_Supply Chain Management	-1.06
11	Last Activity_Email Bounced	-1.31
12	Last Activity_SMS Sent	2.14
13	Last Notable Activity_Modified	-1.80
14	Last Notable Activity_Olark Chat Conversation	-2.01

Inference



After trying several models, we finally chose a model with the following characteristics:

- All variables have p-value < `0.05`.
- All the features have very low VIF values, meaning, there is `hardly any mulilticollinearity` among the features. This is also evident from the heat map.
- The overall accuracy of `0.9098` at a probability threshold of 0.30 on the test dataset is also very acceptable.

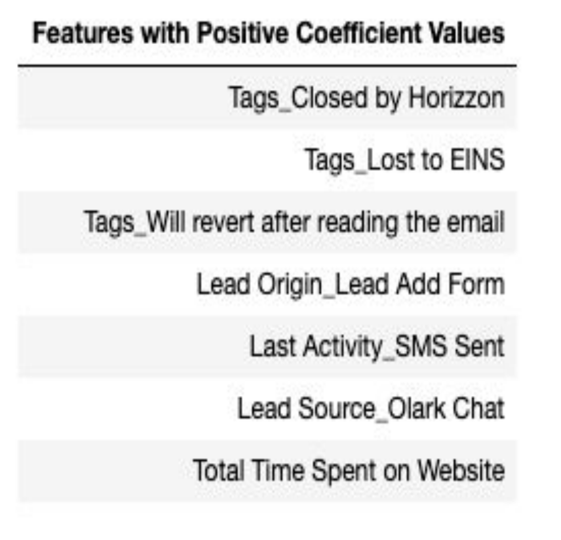
Using this model, the dependent variable value was predicted as per the following threshold values of Conversion probability:

Dataset	Threshhold value	Accuracy	Sensitivity	Specificity	False Postive Rate	Positive Predictive Value	Negative Predictive value	Precision	Recall	F1 value	AUC
train	0.50	0.9295	0.8799	0.9596	0.0400	0.9301	0.9291				0.9737
train	0.30	0.9208	0.9207	0.9209	0.0791	0.8765	0.9501	0.8765	0.9207	0.8980	
test	0.30	0.9052	0.8988	0.9092	0.0907	0.8637	0.9335	0.8636	0.8988	0.8809	0.9639

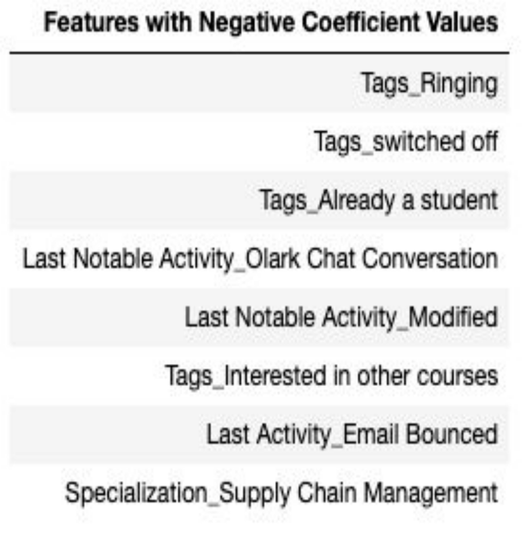
Inference Contd...

Based on our model, some features are identified which contribute most to a Lead getting converted successfully.

- The conversion probability of a lead increases with increase in values of the following features in descending order
- The conversion probability of a lead increases with decrease in values of the following features in descending order



Features with Positive Coefficient Values
Tags_Closed by Horizzon
Tags_Lost to EINS
Tags_Will revert after reading the email
Lead Origin_Lead Add Form
Last Activity_SMS Sent
Lead Source_Olark Chat
Total Time Spent on Website



Features with Negative Coefficient Values
Tags_Ringing
Tags_switched off
Tags_Already a student
Last Notable Activity_Olark Chat Conversation
Last Notable Activity_Modified
Tags_Interested in other courses
Last Activity_Email Bounced
Specialization_Supply Chain Management

Thank You