Lead Scoring Case Study

- Abhishek Sharma
- Moni Kumari
- Shuchi Agrawal

CASE STUDY DESCRIPTION

- 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.
- When these people fill up a form providing their email address or phone number, they are classified to be a lead which will be then passed to Sales team to start making calls or send emails to convert these leads. The typical lead conversion rate at X education is around 30%.
- Now, although X Education gets a lot of leads, its lead conversion rate is very poor. 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.

PROBLEM STATEMENT

- The company requires to build a model wherein a lead score is assigned 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.
- This case study focuses on building 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.
- Identification of such leads which can possible be converted is the focus of the case study.

APPROACH

To improve the lead conversion rate to be around 80%, Logistic Regression model is created to identify the important variables and derive insights on how to improve the lead conversion count. Below Steps are performed in the case study for the outcome :

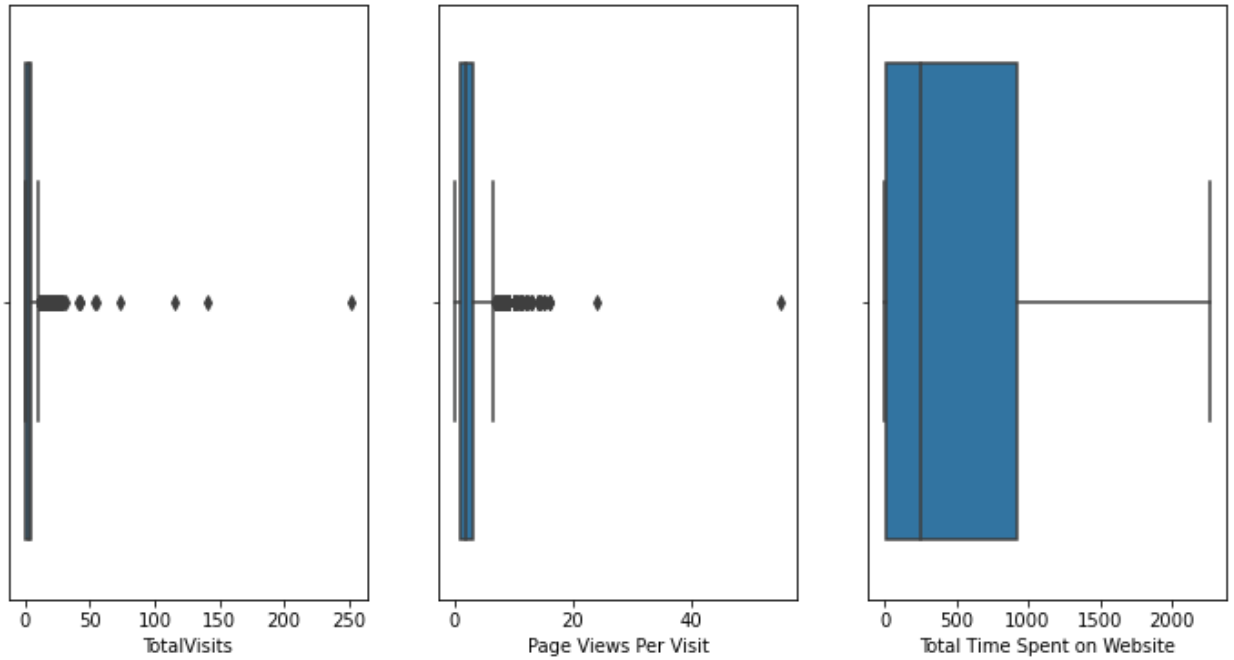
- Importing and understanding the data
- Data Cleaning - This step includes Missing Values Check , Handling them and Data Imputation.
- Exploratory Data Analysis -
- Data Preparation for Modelling – This step includes creating dummies for categorical variables, Train-Test Data Split and Scaling.
- Modelling – This step includes Feature Selection, Recursive Model Building to find the optimal model,
- Model Evaluation - using Performance Metrics & Building ROC Curve, Finding Optimal Cut-Off point
- Predictions on Test data using Final Model
- Final Evaluation using Performance Metrics on Test Data
- Calculating Lead Score

ASSUMPTIONS

- As suggested by business category ‘SELECT’ in was replaced with null.
- Considering a threshold of 40% all the columns with missing value above threshold were dropped. as applying any imputation on such huge missing values can impact the overall analysis of case study which is not recommended.
- For few Category columns, null values has been replaced with a new Category as “Others” to segregate the data.
- For few Category columns, merged the category in Others category which has low volume of records.
- For Numerical columns, null values has been imputed with IQR*1.5 of the variable for those where mean and median are same but max value is way out of range.
- Dropped few unnecessary columns where data was heavily skewed to not impact the overall model building.

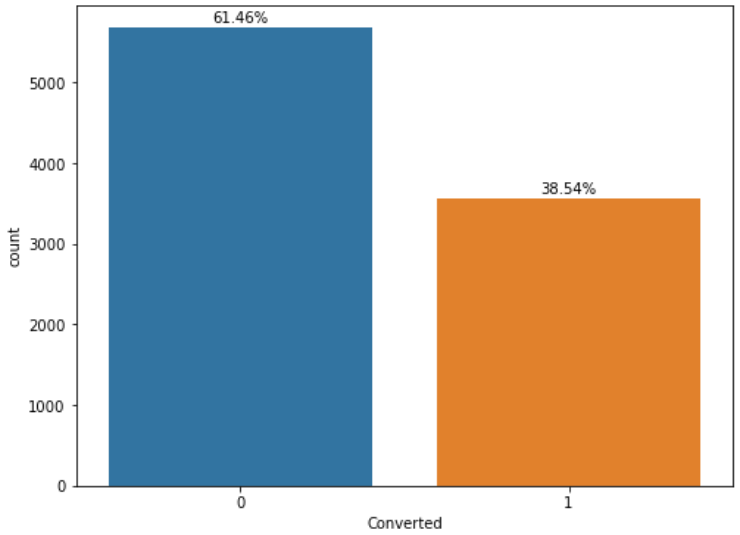
OUTLIERS TREATMENT

- Observed Outliers for Numerical columns using BoxPlot.
- To handle outliers for features TotalVisit and Page View Per Visit considered the top range of data till 99 percentile to be able to build proper model.



UNIVARIATE ANALYSIS

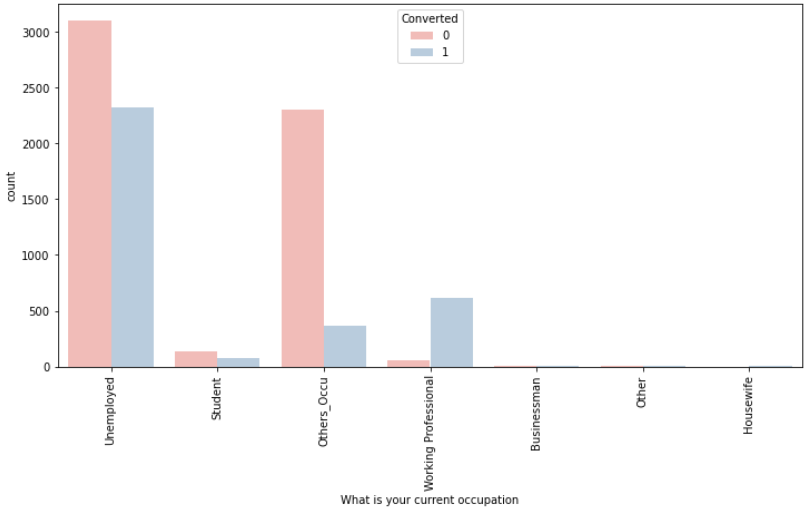
- From the available data we analyzed that conversion rate is 38.54% .

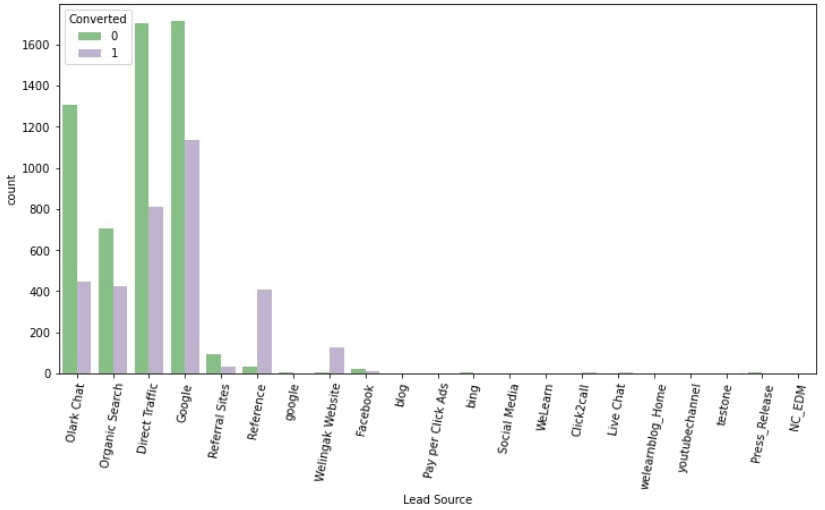


BIVARIATE ANALYSIS

From the below plots it can be infer that

- Working Professionals have the higher conversion
- Unemployed have the highest count in the lead category and additional focus can be given to them in conversion
- References as the leadsource has the highest conversion and the top two count of leads are from Direct Traffic or Google

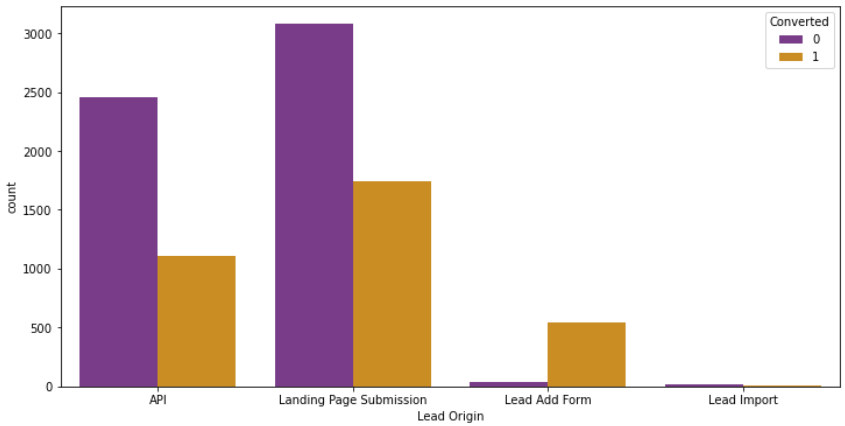


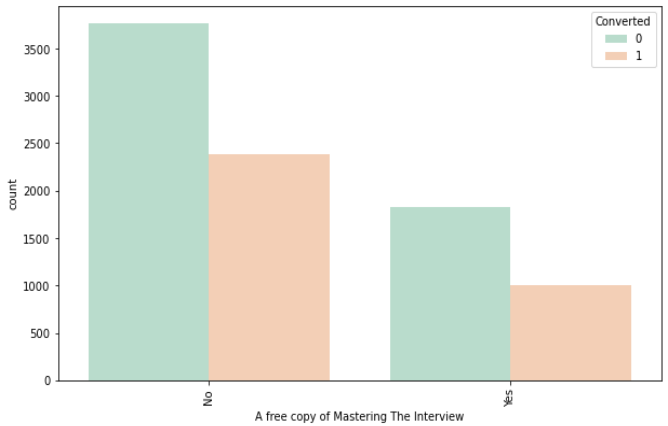


BIVARIATE ANALYSIS

From above plots it can be infer that:

- Lead Origin as Landing Page Submission has the highest count of leads along with most conversions
- People who said No for Free copy of Mastering the interview are highest in the conversion

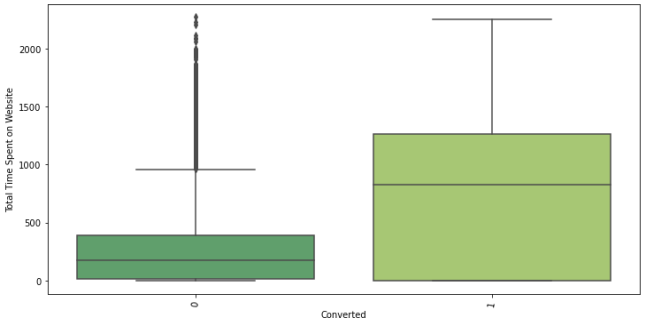


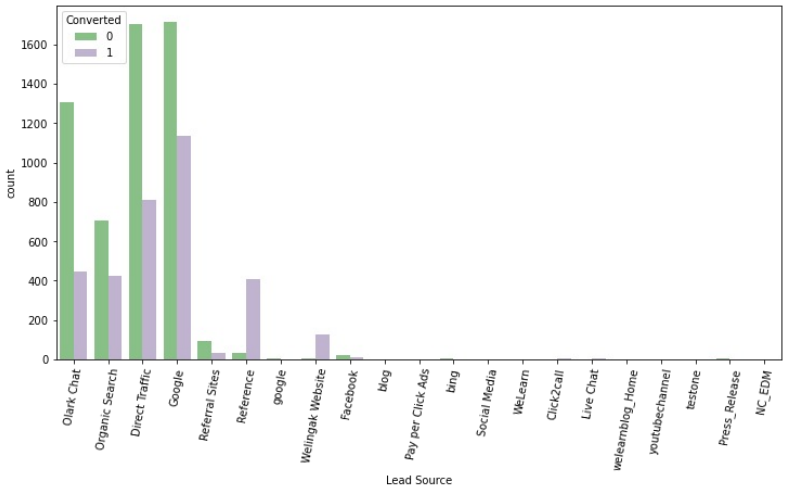


BIVARIATE ANALYSIS

From above plots it can be infer that

- we can see that leads spending more time on website are majorly converted irrespective of Specialization.
- Lead Source as Wellingak Website, ClarkChat, Referral Sites & Organic Search are the ones who have most of them converted amongst the other lead source.

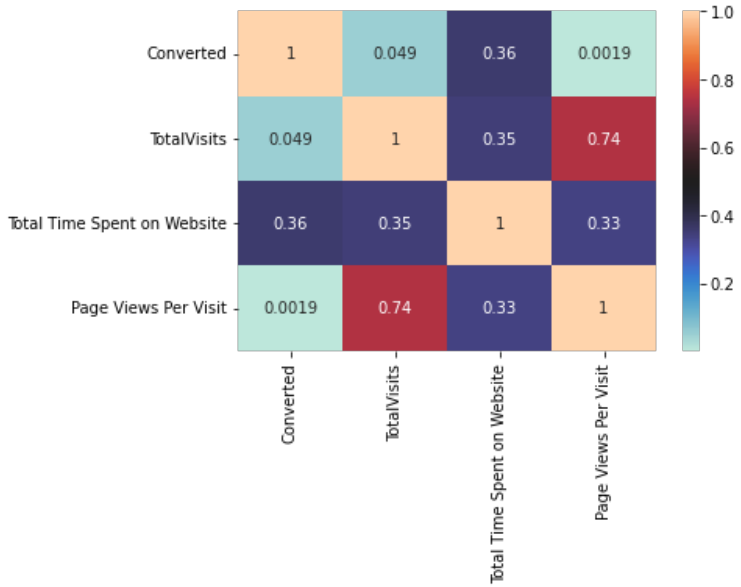




CORRELATION MATRIX

From above Correlation Matrix, we can see that

- Converted has very high positive correlation with Total Time Spent on Website and lowest correlation with relationship with Page View Per Visit.



Data Preparation

DATA PREPARATION STEPS –

- Converted binary variables (Yes/No) to 0 and 1 for model building.
- Created Dummy variables for all category columns using pd.get_dummies

TRAIN-TEST SPLIT

- Split the data into train and test data frame in 70-30% ratio using sklearn library.

FEATURE SCALING

Used Standard Scaler to standardize the numerical column values so that they can be analysed at same scale. If we don't have comparable scales, then some of the coefficients as obtained by fitting the model might be very large or very small as compared to the other coefficients which makes model evaluation very difficult.

MODEL BUILDING

- Build 1st Logistic Regression training model using all features.
- To build best fit model, we used Recursive Feature Elimination technique to get the top 15 features to build out next model
- For each model build, we have checked for p-value should be less than 0.05
- To remove Multicollinearity, calculated Variance Inflation Factor(VIF) to check if feature variables are not correlated with each other.
- Dropped the features which have high p-value and highly correlated one by one and recursively build the model to get optimal model.

Dep. Variable:	Converted	**No. Observations:**	6293
Model:	GLM	**Df Residuals:**	6277
Model Family:	Binomial	**Df Model:**	15
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1296.8
Date:	Mon, 19 Jun 2023	**Deviance:**	2593.6
Time:	00:19:12	**Pearson chi2:**	9.14e+03
No. Iterations:	23		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-3.0544	0.185	-16.468	0.000	-3.418	-2.691
Lead Source_Welingak Website	3.4441	1.022	3.371	0.001	1.442	5.446
What is your current occupation_Others_Occu	-2.4257	0.138	-17.632	0.000	-2.695	-2.156
Tags_Already a student	-2.7170	1.019	-2.667	0.008	-4.713	-0.721
Tags_Busy	2.0086	0.288	6.970	0.000	1.444	2.573
Tags_Closed by Horizzon	7.5840	0.735	10.323	0.000	6.144	9.024
Tags_Lost to EINS	7.9650	0.634	12.558	0.000	6.722	9.208
Tags_Not doing further education	-1.7417	1.034	-1.684	0.092	-3.769	0.286
Tags_Others_Tags	2.9676	0.215	13.828	0.000	2.547	3.388
Tags_Ringing	-2.1533	0.294	-7.322	0.000	-2.730	-1.577
Tags_Will revert after reading the email	5.8992	0.243	24.232	0.000	5.422	6.376
Tags_invalid number	-23.0011	1.59e+04	-0.001	0.999	-3.13e+04	3.12e+04
Tags_number not provided	-23.1623	3.14e+04	-0.001	0.999	-6.16e+04	6.15e+04
Tags_switched off	-3.3336	0.746	-4.467	0.000	-4.796	-1.871
Tags_wrong number given	-23.1261	1.98e+04	-0.001	0.999	-3.89e+04	3.89e+04
Last Notable Activity_SMS Sent	2.7938	0.135	20.748	0.000	2.530	3.058

Dep. Variable:	Converted	**No. Observations:**	6293
Model:	GLM	**Df Residuals:**	6278
Model Family:	Binomial	**Df Model:**	14
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1306.3
Date:	Mon, 19 Jun 2023	**Deviance:**	2612.7
Time:	00:19:12	**Pearson chi2:**	8.89e+03
No. Iterations:	23		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-3.3056	0.189	-17.468	0.000	-3.677	-2.935
Lead Source_Welingak Website	3.4422	1.021	3.370	0.001	1.440	5.444
What is your current occupation_Others_Occu	-2.4104	0.137	-17.650	0.000	-2.678	-2.143
Tags_Already a student	-2.4658	1.019	-2.419	0.016	-4.464	-0.468
Tags_Busy	2.2945	0.286	8.033	0.000	1.735	2.854
Tags_Closed by Horizzon	7.8353	0.736	10.651	0.000	6.393	9.277
Tags_Lost to EINS	8.2043	0.635	12.910	0.000	6.959	9.450
Tags_Not doing further education	-1.4731	1.033	-1.426	0.154	-3.498	0.552
Tags_Others_Tags	3.2278	0.217	14.898	0.000	2.803	3.652
Tags_Ringing	-1.8314	0.289	-6.334	0.000	-2.398	-1.265
Tags_Will revert after reading the email	6.1554	0.246	25.044	0.000	5.674	6.637
Tags_number not provided	-22.8453	3.15e+04	-0.001	0.999	-6.18e+04	6.18e+04
Tags_switched off	-3.0083	0.744	-4.043	0.000	-4.467	-1.550
Tags_wrong number given	-22.8098	1.99e+04	-0.001	0.999	-3.91e+04	3.9e+04
Last Notable Activity_SMS Sent	2.7118	0.131	20.721	0.000	2.455	2.968

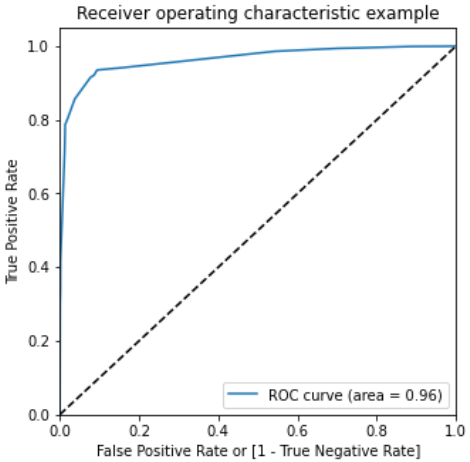
Generalized Linear Model Regression Results

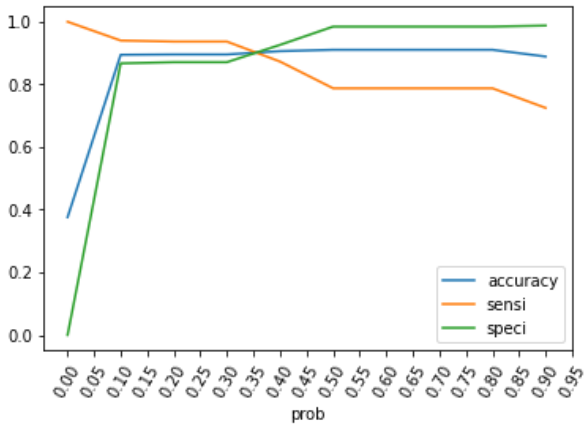
Dep. Variable:	Converted	**No. Observations:**	6293
Model:	GLM	**Df Residuals:**	6279
Model Family:	Binomial	**Df Model:**	13
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1308.8
Date:	Mon, 19 Jun 2023	**Deviance:**	2617.5
Time:	00:19:12	**Pearson chi2:**	8.85e+03
No. Iterations:	22		
Covariance Type:	nonrobust		

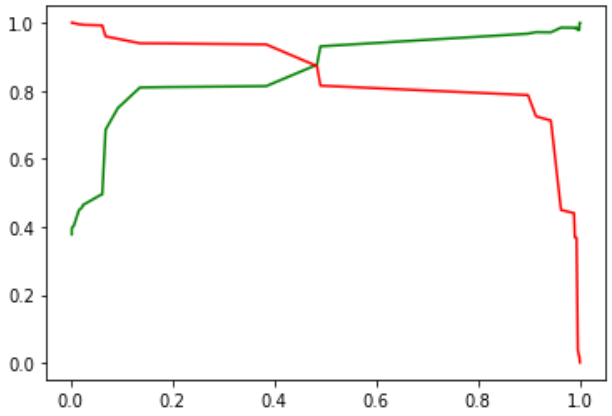
	coef	std err	z	P>|z|	[0.025	0.975]
const	-3.3706	0.190	-17.728	0.000	-3.743	-2.998
Lead Source_Welingak Website	3.4417	1.021	3.370	0.001	1.440	5.443
What is your current occupation_Others_Occu	-2.4069	0.136	-17.654	0.000	-2.674	-2.140
Tags_Already a student	-2.4009	1.019	-2.355	0.019	-4.399	-0.403
Tags_Busy	2.3678	0.285	8.308	0.000	1.809	2.926
Tags_Closed by Horizzon	7.9003	0.736	10.736	0.000	6.458	9.343
Tags_Lost to EINS	8.2665	0.636	13.002	0.000	7.020	9.513
Tags_Not doing further education	-1.4041	1.033	-1.359	0.174	-3.429	0.620
Tags_Others_Tags	3.2950	0.217	15.177	0.000	2.869	3.720
Tags_Ringing	-1.7496	0.288	-6.077	0.000	-2.314	-1.185
Tags_Will revert after reading the email	6.2216	0.246	25.259	0.000	5.739	6.704
Tags_switched off	-2.9258	0.744	-3.935	0.000	-4.383	-1.468
Tags_wrong number given	-21.7296	1.21e+04	-0.002	0.999	-2.37e+04	2.37e+04
Last Notable Activity_SMS Sent	2.6922	0.130	20.694	0.000	2.437	2.947

MODEL EVALUATION – TRAIN DATA

- After getting optimal model, evaluated performance metrics score Accuracy, Recall, Precision.
- ROC curve plotted that 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.
- Calculated Optimal cutoff point between sensitivity & specificity. From below plot, we have received 0.38 as the optimal cut-off point.
- Also checked Precision and Recall trade-off as this will help us to identify the predicted CONVERTED is actual CONVERTED
- The Precision and Recall tradeoff came out to be 0.5, we have considered that as our cut-off probability on test data.







MODEL EVALUATION – TEST DATA

- After running the final optimal model on test dataset we observed that Evaluation matrix has similar values for both training and test dataset:
- Training Dataset:
  - - Accuracy: 91.03%
  - - Sensitivity: 78.75%
  - - Specificity: 98.4%

- Test Dataset
  - Accuracy – 91.69%
  - Sensitivity - 80.81%
  - Specificity – 98.4%

LEAD SCORE PREDICTION

- The final_predicted column shows the conversion probability of prospective lead.
- Lead Score above 39 have a high tendency of converting to a Hot Lead category.

	Converted	Converted_Prob	ProspectID	final_predicted	Lead_Score
0	1	0.993615	1467	1	99
1	0	0.062122	108	0	6
2	1	0.993615	7858	1	99
3	1	0.383896	5220	1	38
4	0	0.062122	3871	0	6

Training data

	ProspectID	Converted	Convert_Prob	final_predicted	Lead_Score
8692	8692	0	0.490458	0	49
6126	6126	1	0.993615	1	99
5198	5198	1	0.062122	0	6
4979	4979	1	0.490458	0	49
9225	9225	0	0.092822	0	9

Test data

RECOMMENDATIONS

- Leads who were Tagged as Closed by Horizzon, Lost to EINS were the max who got converted to permanent customer.
- Prospect spending more time on website have high changes of becoming Hot Leads therefore Sales team can provide more focus on reaching out to those.
- Lead Score with Welingak websites and referral are the ones who have the highest amount of conversions therefore additional marketing can be done
- on the websites and sales team can sent the course details and promotional offers to existing users to get more Hot Leads
- Leads contacted via email/sms has higher chances of conversion.
- Unemployed/Working Professionals as Occupation category can generate more leads by reaching out to them and providing information about the courses available.

CONCLUSION

- The final model shows 92.2% accuracy with Recall as 85.7% and Precision as 96.2%
- The optimal cut-off was selected based on Precision and Recall trade off score.
- The model also worked fine on test dataset with Recall as 87.4% and Precision as 96%
- Overall the model looks good and is able to identify the correct leads which has high chances of conversion using Lead Score prediction