Lead Scoring Case Study

Presented by,

Lohith Raj M

Kondapalli Anil Kumar

Problem Statement

- An education company named X Education, which sells online courses to industry professionals, wants to identify hot leads from initial pool of leads.
- The sales team would then focus on the hot leads, and would try and convert all the hot leads through regular e-mail and phone call conversations.
- For the identification of the hot leads the company needs a model which assigns lead score to each lead based on the probability of a lead getting converted.
- The sales team has been given a target of 80% conversion rate from the current rate of around 30%. The model needs to assist sales team in achieving this.
- For this purpose a logistic regression model is built.

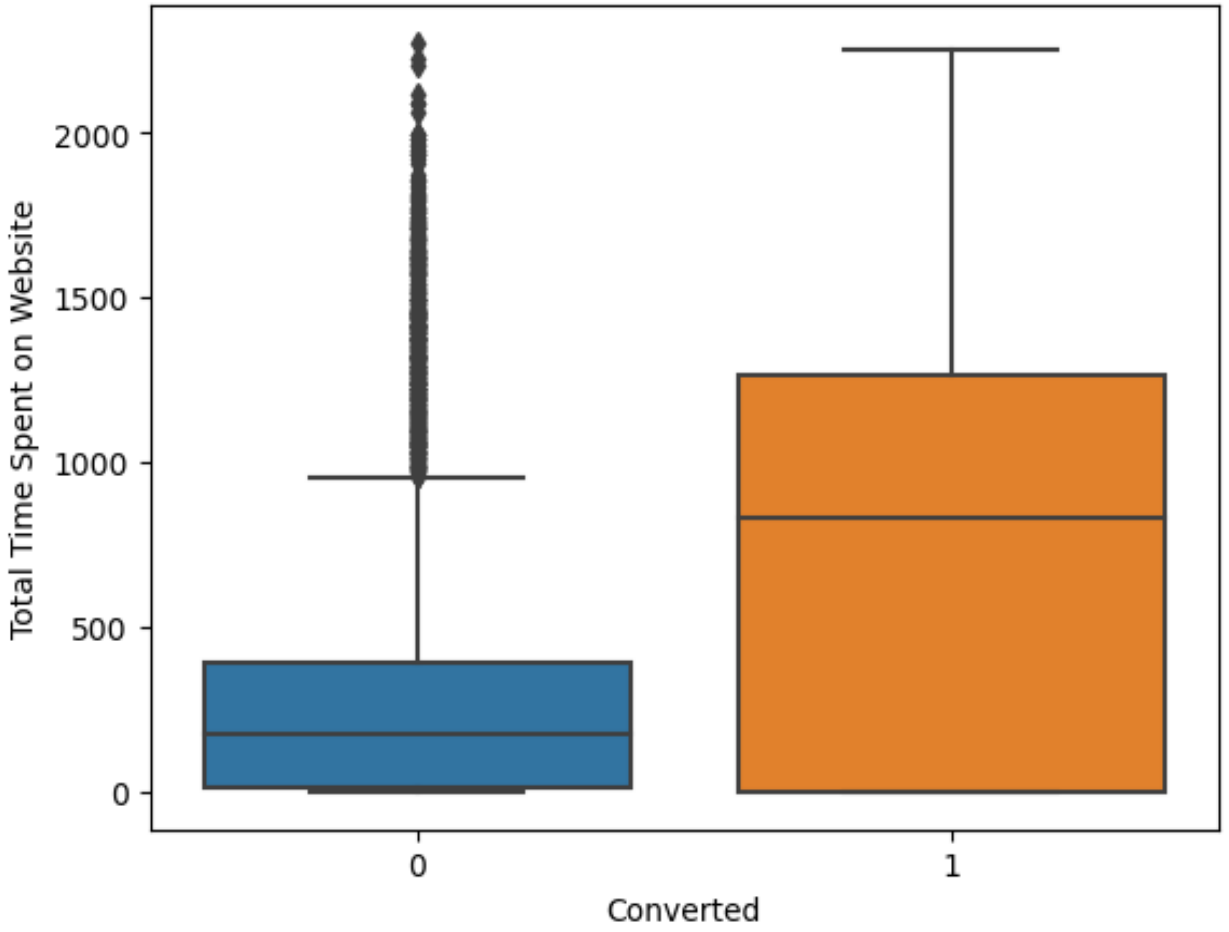
Approach of the Study

1. Reading & Understanding the Data - checking basic info & stats of the data set
2. Data Cleaning - handling outliers and missing values
3. EDA - univariate and bivariate analysis, & heat map for numerical features
4. Data Preparation - splitting the data into train & test sets, & scaling the features
5. Model Building - logistic model with eventual 10 parameters is built
6. Checking the Model Metrics - accuracy, sensitivity and specificity is checked
7. Finding the Optimal Cut-off value - optimal value of probability is calculated
8. Making predictions on the test set using the Logistic Model
9. Conclusion - Results are highlighted

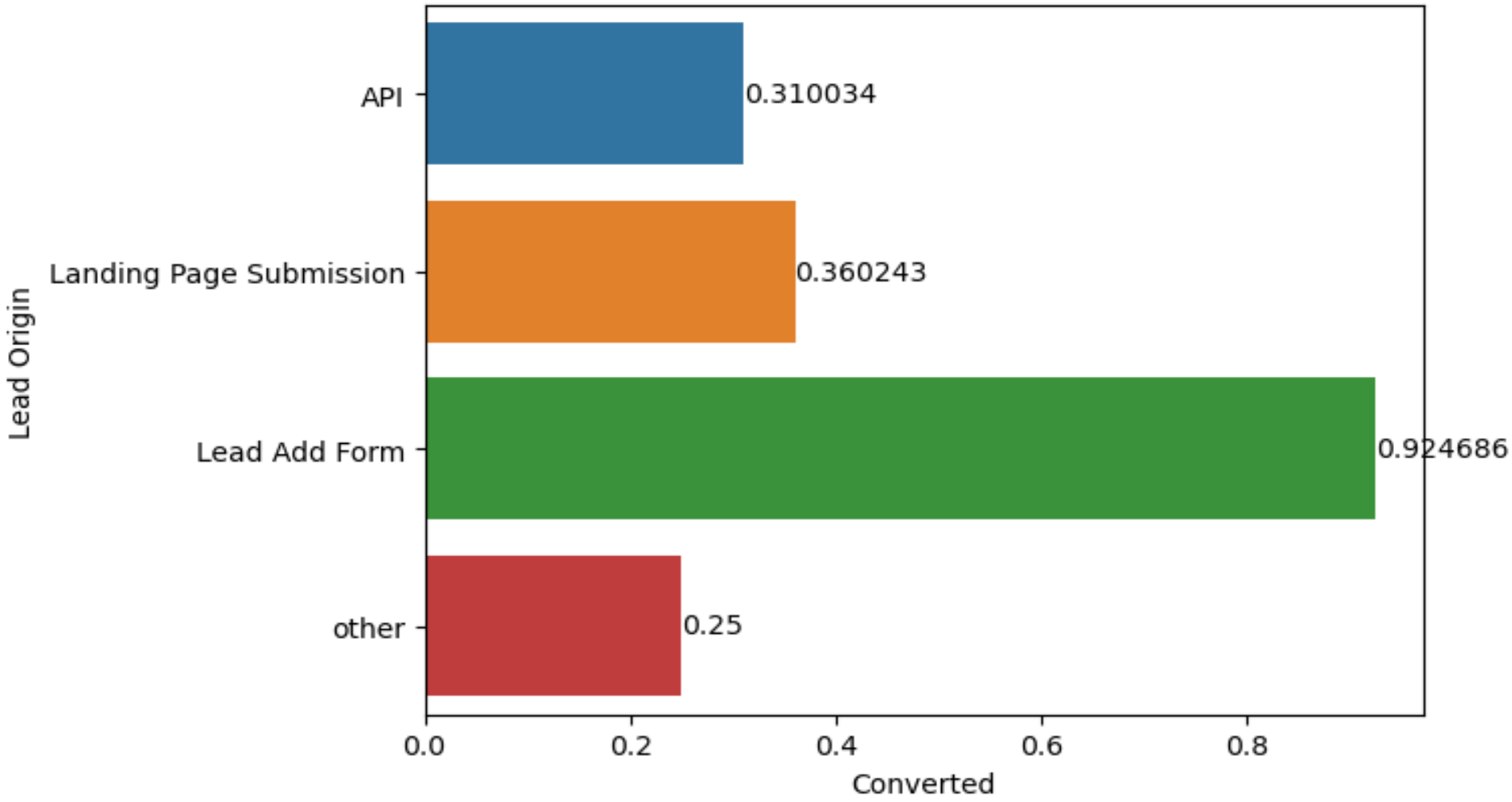
EDA

a) Numerical Features

- The Box plot of “Total Time Spent on Website” vs “Converted” is plotted.
- For the leads that are converted time spent on the website is clearly high, with median value close to 1000 units of time spent on the website.

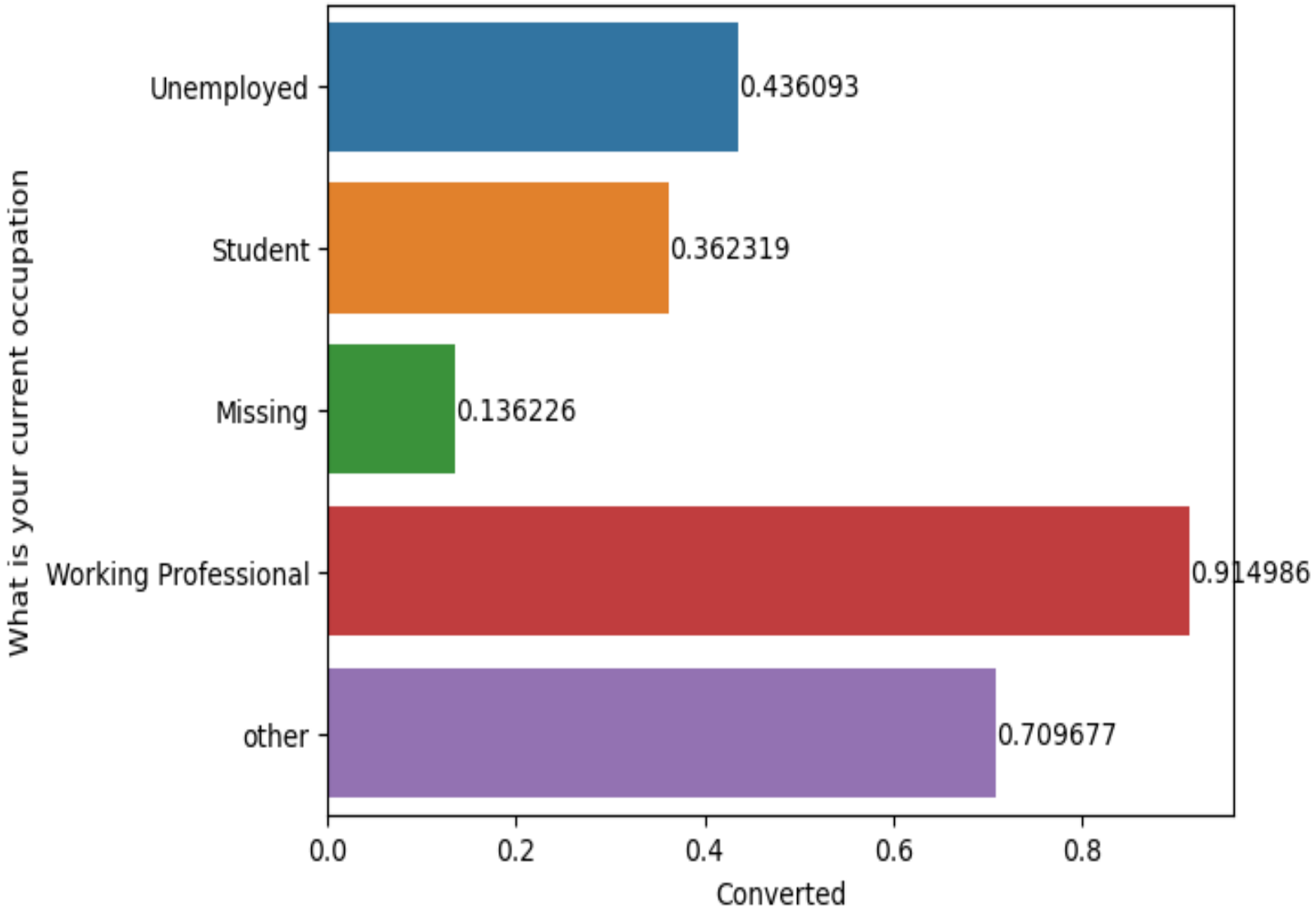


b) Categorical Features



- Conversion rate of the data set is 38.45%
- Conversion rate of the class “Lead Orgin_Lead Add Form” is close to 92%.

- Conversion rate of the class “What is your current occupation_Working Professional” is over 90%.
- Conversion rate of the class “What is your current occupation_Unemployed” is over 43.6%.



P-Value & VIF of Model Parameters

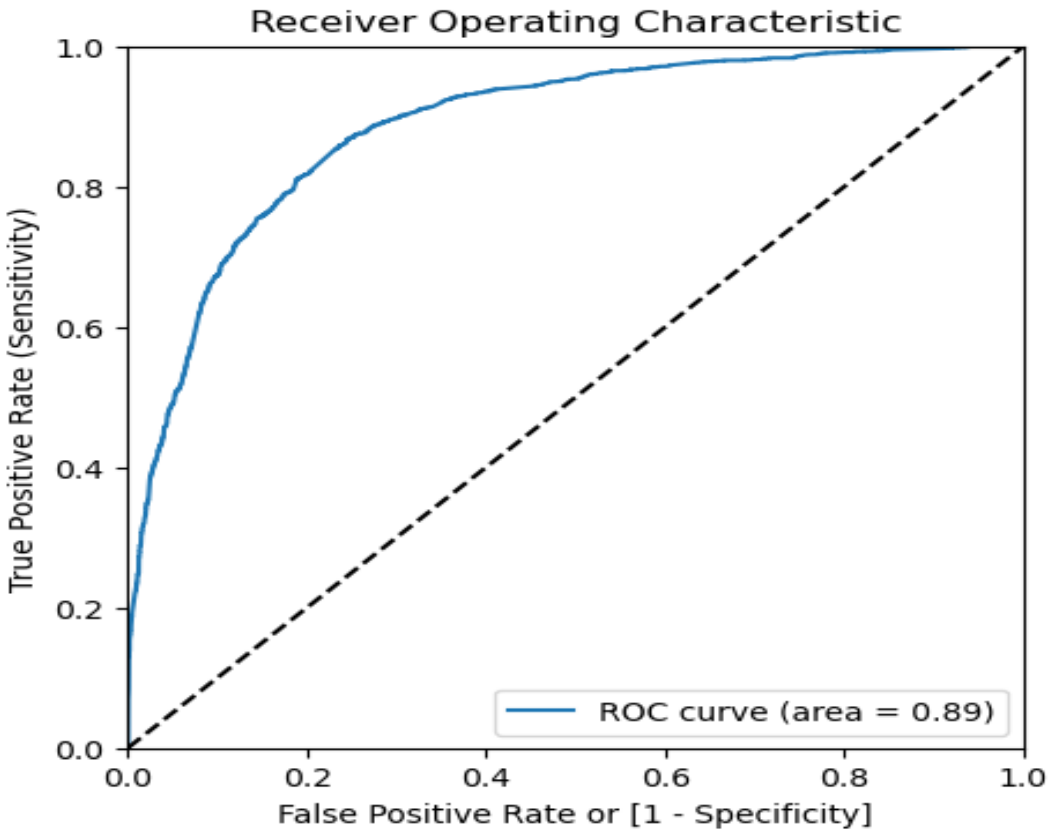
	P>|z|
const	0.000
Total Time Spent on Website	0.000
Lead Origin_Lead Add Form	0.000
Lead Source_Olark Chat	0.000
Lead Source_Welingak Website	0.011
Last Activity_Email Bounced	0.000
Last Activity_Olark Chat Conversation	0.000
Last Activity_SMS Sent	0.000
What is your current occupation_Unemployed	0.000
What is your current occupation_Working Professional	0.000
Last Notable Activity_Modified	0.000

p-value of all the features is less than 0.05, which indicates all the features are statistically significant

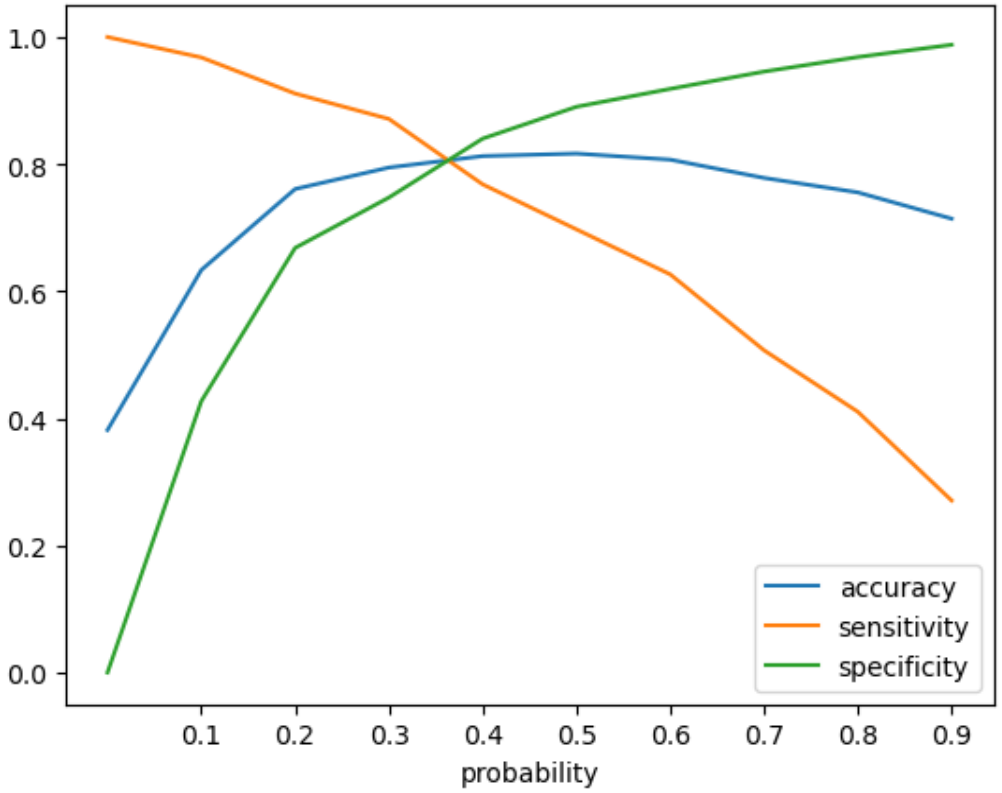
Features	VIF
What is your current occupation_Unemployed	1.68
Lead Source_Olark Chat	1.62
Last Notable Activity_Modified	1.60
Last Activity_Olark Chat Conversation	1.57
Lead Origin_Lead Add Form	1.54
Last Activity_SMS Sent	1.44
Lead Source_Welingak Website	1.29
Total Time Spent on Website	1.25
What is your current occupation_Working Profes...	1.22
Last Activity_Email Bounced	1.10

VIF of all the features is less than 5 indicates the presence of acceptable level of multicollinearity in the model

ROC Curve & Optimal Cut-Off Value



The AUC of the model is 0.89, this indicates the model is working satisfactorily & not just guessing randomly.



Probability at the Intersection of the curves of accuracy, sensitivity and specificity is optimal cut-off value, 0.35.

Interpreting parameters of the logistic model

- "Lead Origin_Lead Add Form" has 35.61 times the odds of the other classes of "Lead Origin", in getting the lead converted.
- "Last Activity_Email Bounced" is associated with 0.82 or 82% (1 - 0.18) reduction in getting the lead converted.
- An increase in 1 standard deviation in "Total Time Spent on Website" is associated with 2.98 times increase in the odds of a lead getting converted.

	Coefficient	odds
Lead Origin_Lead Add Form	3.572736	35.61
What is your current occupation_Working Professional	3.531187	34.16
Lead Source_Welingak Website	2.608567	13.58
Lead Source_Olark Chat	1.264354	3.54
Last Activity_SMS Sent	1.146114	3.15
Total Time Spent on Website	1.090863	2.98
What is your current occupation_Unemployed	1.018824	2.77
Last Notable Activity_Modified	-0.753641	0.47
Last Activity_Olark Chat Conversation	-0.892922	0.41
Last Activity_Email Bounced	-1.723753	0.18
const	-1.988132	0.14

Metrics of the Train & Test sets

- The difference between the metrics of the train set and test set is in acceptable level.
- This indicates there is no overfitting on the train set, and model can be generalised.

train set (Logistic regression model)

- accuracy: 80.84%
- sensitivity: 81.55%
- specificity: 80.41%

test set

- accuracy: 79.98%
- sensitivity: 80.04%
- specificity: 79.94%

Hot Leads

Top 5 Hot leads based on the lead score of the test set

	Lead No.	Lead Score
2662	627106	99.89
1601	627462	99.81
1093	591536	99.78
484	658648	99.75
33	625862	99.68

Bottom 5 Hot leads based on the lead score of the test set

	Lead No.	Lead Score
1450	654594	35.19
2202	583595	35.09
2149	614845	35.09
2568	602582	35.06
541	597873	35.04

Conclusion

- The sales team has the target of achieving 80% conversion rate from the current rate of around 30%, in order to achieve this objective Lead score assigned by the model to each lead can be used.
- The leads with scores greater than say 80, don’t need too much prompting, otherwise it could have a negative effect. Regular e-mail form of communication and making calls only when necessary, may prove sufficient.
- The leads near the cut-off, 35 as per the model, need careful monitoring through regular phone calls or else they could be lost. The company needs to strategize at what stages they need to make phone calls and what will be the necessary frequency of the phone calls to convert the leads.