Lead Scoring Case Study

Logistic Regression Modeling

Team
Nistha Kumar & Gaurav Rana

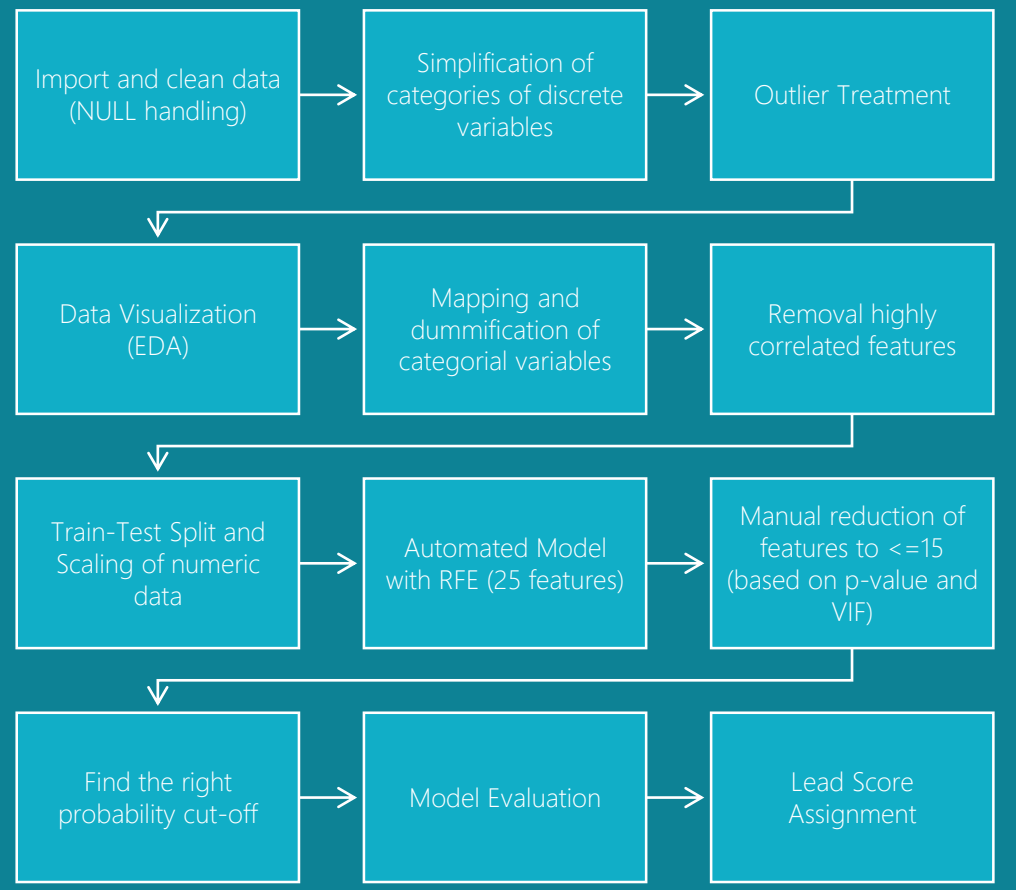
Table of Content

Problem Statement & Solution Approach

Problem Statement

- X Education, that sells online courses, currently has a lead conversion rate of 30%
- The company wants to identify "Hot Leads" – the most potential leads, so that sales team focuses on them (rather than contacting everyone)
- Target "lead conversion rate" > ~80%
- Data set available – 9000 leads and its data
  - User Input variables
  - Score variables
  - Activity variables
- Objective is to create a logistic regression model that can assign a "Lead Score" (0-100) to each lead that sales can use (higher score means lead is hot)

Solution Approach



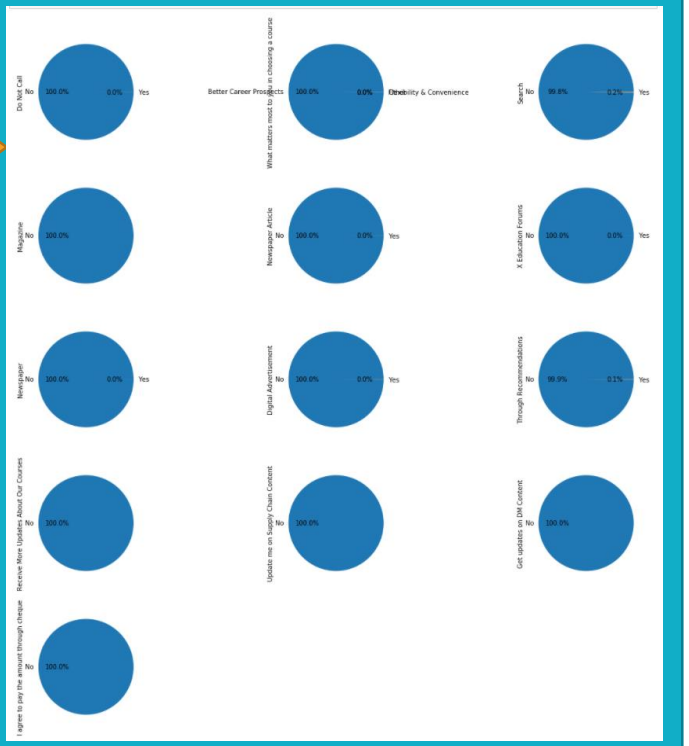
Data Preparation & Understanding(1/3)
(NULL handling & Simplification of Categorical Variables)

Treatment of Columns with > 40% NULL Values
- "Select" was treated as Null (no selection in UI)
- All columns with > 40% NULLs were dropped
- Such Columns deleted (with null %age)
  - How did you hear about X Education (78%)
  - Lead Profile (74%)
  - Lead Quality (51%)
  - Asymmetrique Profile Score (46%)
  - Asymmetrique Activity Score (46%)
  - Asymmetrique Profile Index (46%)
  - Asymmetrique Activity Index (46%)

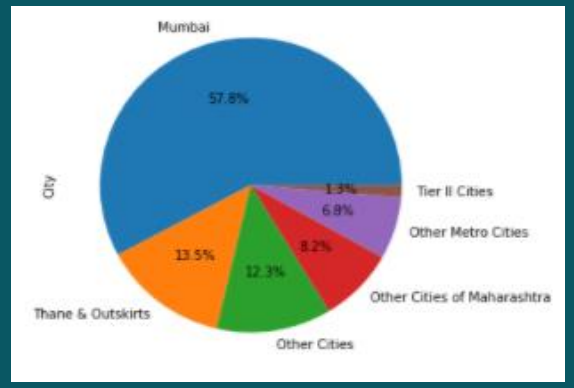
Treatment of Rows with > 6 NULL Values
- 20 (out of 9000+) rows with more than 6 columns as Nulls were deleted

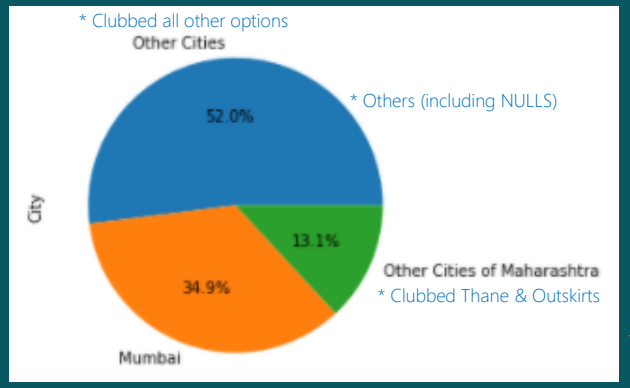
Score variables deleted
- Score variables are updated by the Sales team when in contact with the customer, hence cannot be used to predict a lead score - therefore they were deleted
- "Tags" was one such column (in addition to previously deleted columns)

Dropping of categorical columns that were heavily skewed
- 13+ categorical columns had more than 99.9% percent values dominated by 1 category – which adds no value to a classification model – hence dropped



Simplification of "Large number" of categories into broader categories
- Applied to 7 categorical variables (Lead Origin, Lead Source, Last Activity, Specialization, Current occupation, City, Last Notable Activity)
- Example:



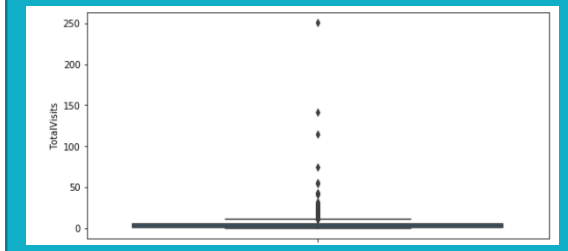


Data Preparation & Understanding(2/3)
(Outlier Treatment, Correlation Check)

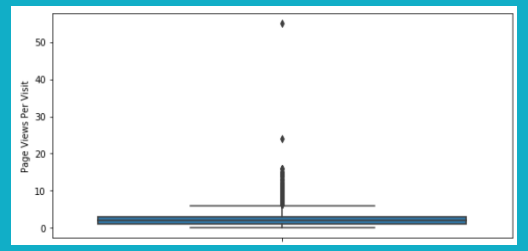
Outliers in Numerical Columns

- We had upper range outliers in "Total Visits" & "Page Views Per Visit" – 95 percentile capping was done to treat the same

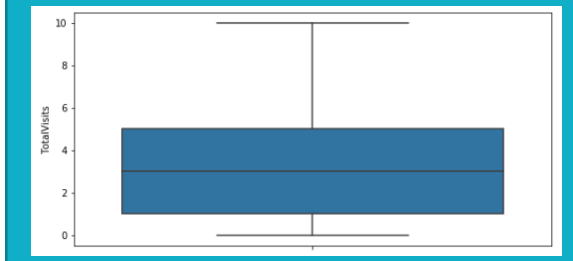
Total Visits (Before)

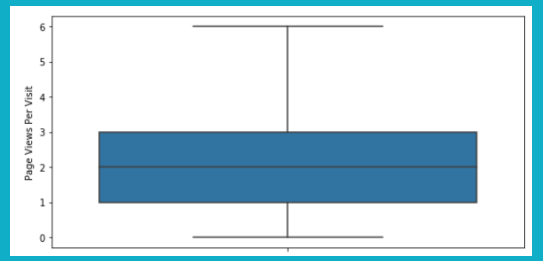


Page Views/Visit (Before)



95% percentile upper range capping

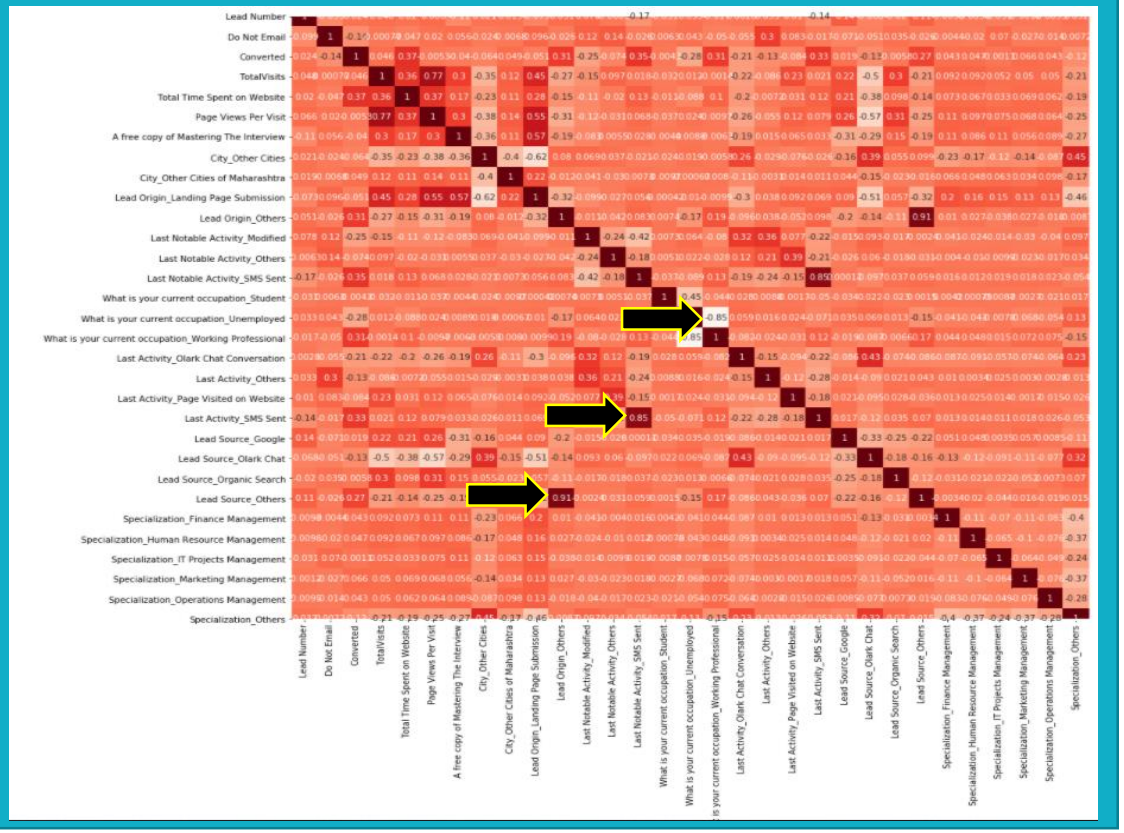




Dummification was done for all retained categorical variables, deleting the first category for each

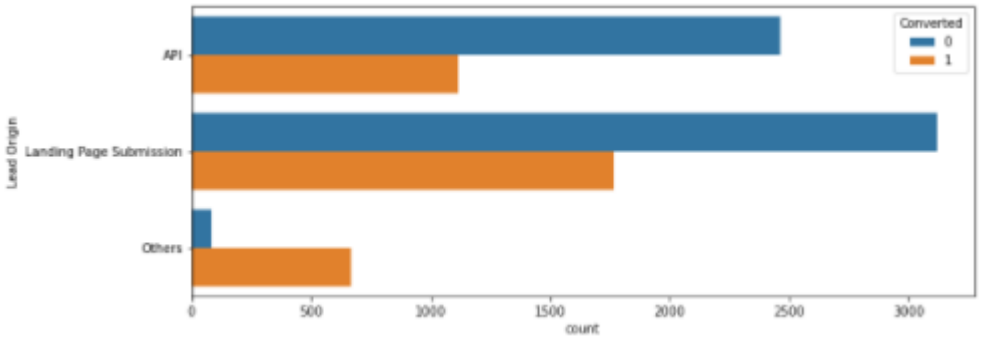
Collinearity Check between columns

- Deleted "Lead Source_Others","Last Notable Activity_SMS Sent","What is your current occupation_Unemployed" because they were highly correlated to other features

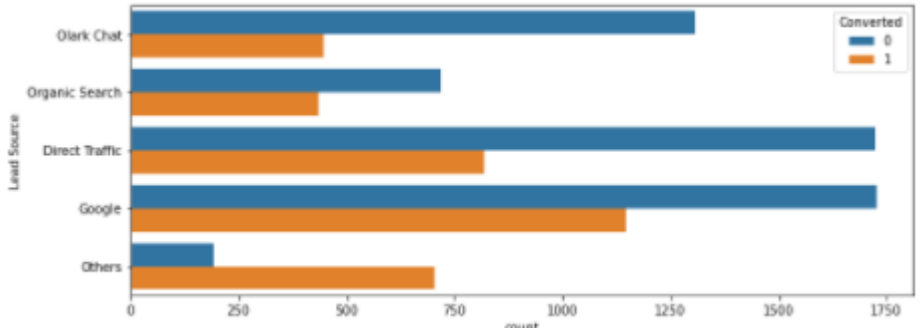


Data Preparation & Understanding(3/3)

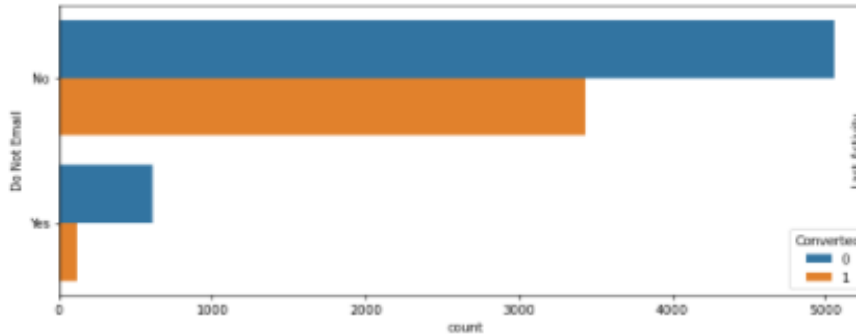
(Data Visualizations – key insights)



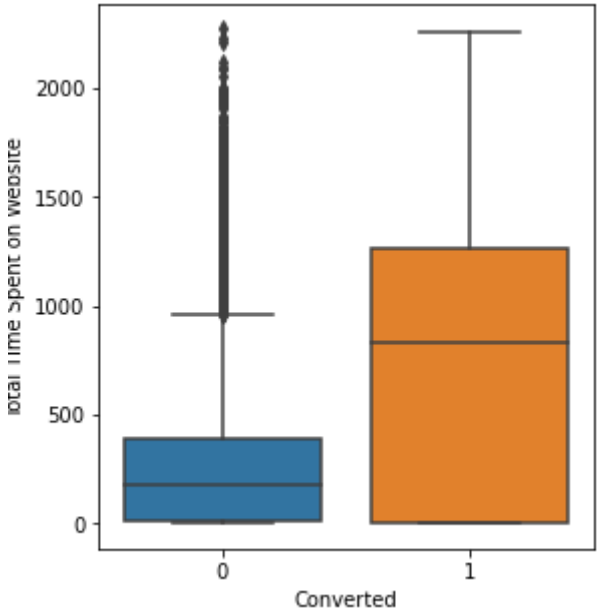
Lead Origin: To improve overall lead conversion rate, focus is needed more on improving lead conversion of API and Landing Page Submission origin



Lead Source: To improve overall lead conversion rate, focus is needed more on improving lead conversion of olark chat, organic search, direct traffic, and google leads.(Google and Direct traffic contributing to maximum number of leads)

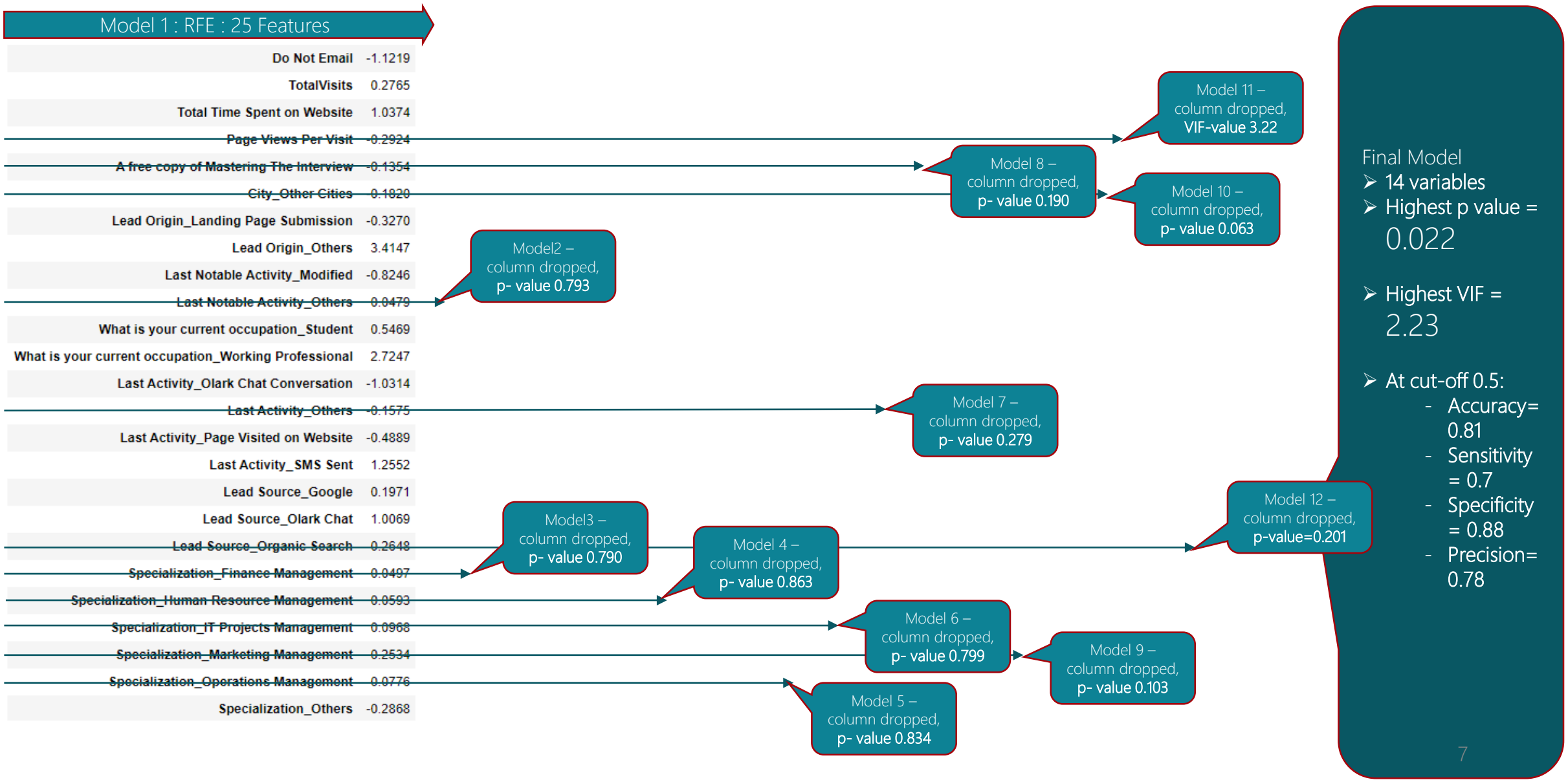


Do Not Email: To improve overall lead conversion rate, focus is needed more on customers entering the value as "No". they have better conversion rate.

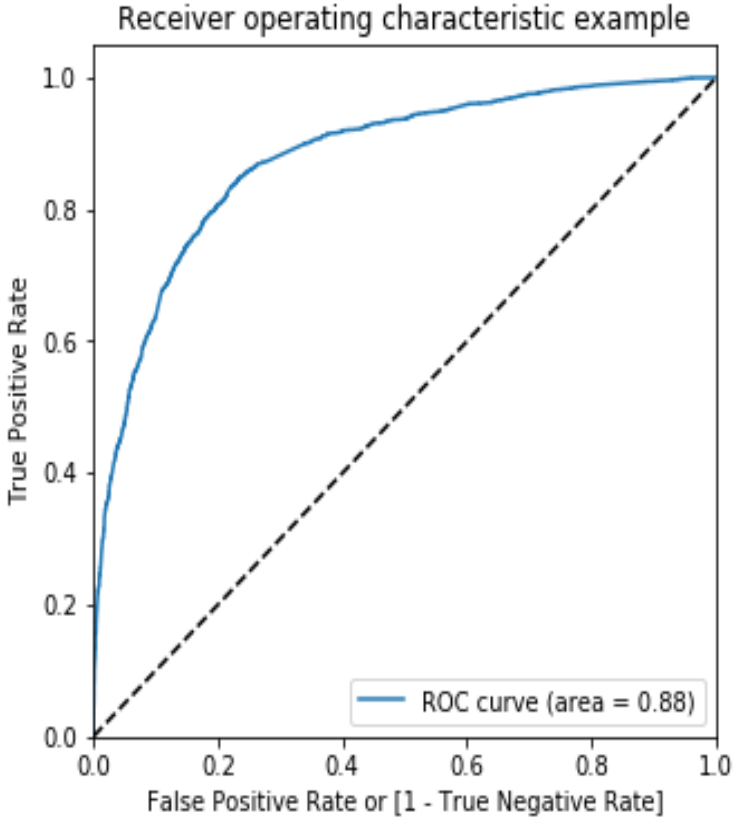


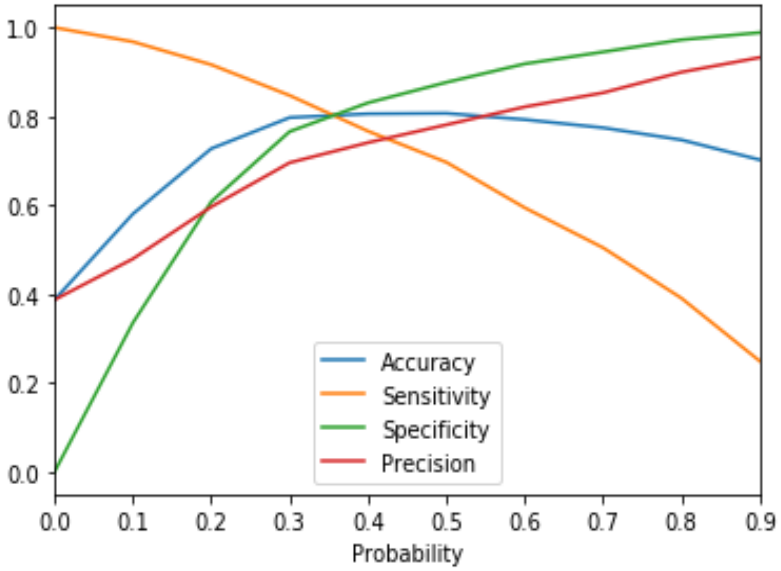
Lead spending more time on the website are more likely to be converted. Hence website should be made more engaging to increase conversions

Data Modeling



Model Evaluation (AUC, Metrics – Train/Test)





Cut-off selected = 0.3

Summary of evaluation

- **Train Data (As per Cut-off 0.3):**
  - Accuracy= 0.797
  - Sensitivity= 0.847
  - Specificity= 0.766
  - Precision= 0.696
- **Test Data (As per Cut-off 0.3):**
  - Accuracy= 0.801
  - Sensitivity= 0.854
  - Specificity= 0.769
  - Precision= 0.694

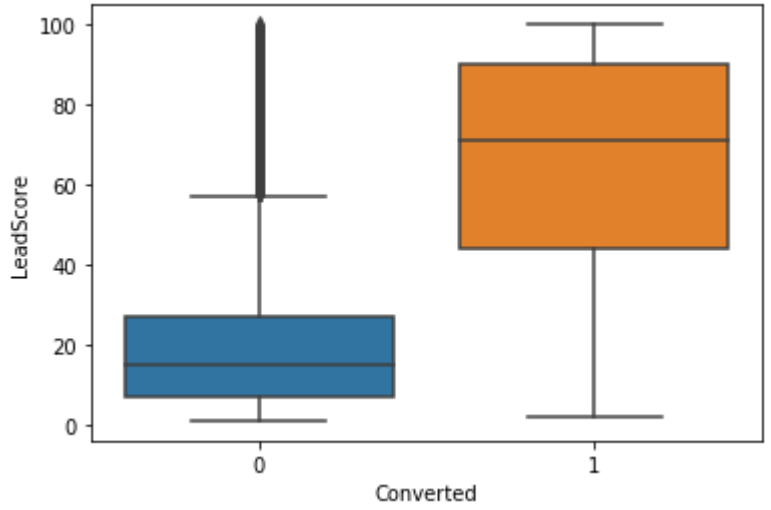
Lead Scoring

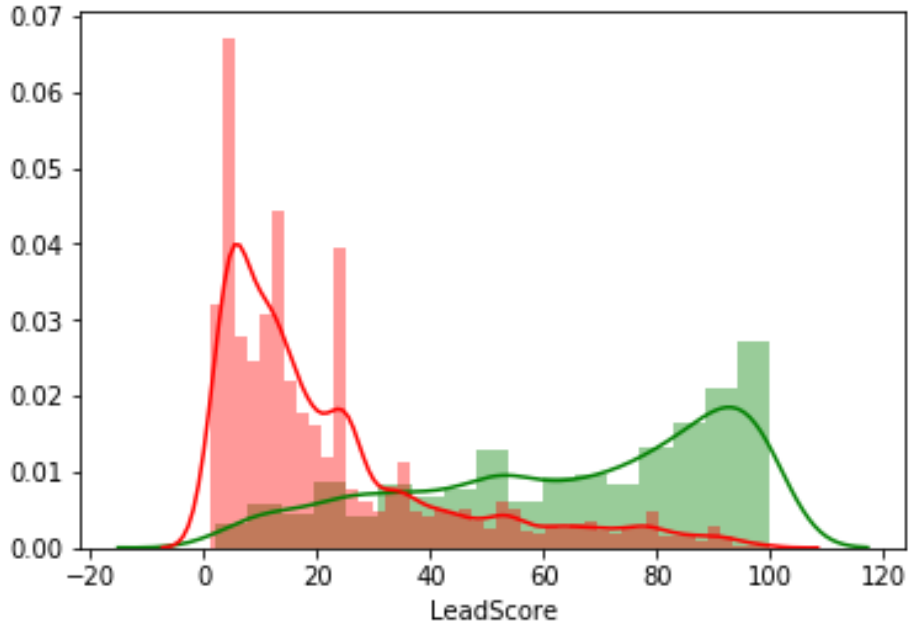
Lead Scores assigned from 1 to 100 (probability * 100)

As can be seen, "converted" leads have got clearly higher scores, and "non-converted" have got lesser scores

Distribution of Lead Scores for Converted (Green) and Non-Converted (Red) seems to clearly demarcate the "hot leads"

	Converted	Lead Score
6784	1	100
6647	1	100
8055	1	100
2665	1	100
8052	1	100
...	...	...
8614	0	1
5643	0	1
5523	0	1
5910	0	1
8947	0	1





Model Summary

- As per the final model there are 14 fields that significantly contribute towards the probability of a lead getting converted
- Out of these 14, following are the three top variables:
  - Wherever the "Lead Origin" is "Others" which means either of (Lead Add Form, Lead Import, Quick Add Form) then it increases the "log of odds" of getting converted by 3.6 times
  - If the lead is a "Working professional" it increases the "log of odds" of getting converted by 2.7 times
  - If the last activity performed by the customer is "SMS Sent" then it increases the "log of odds" of getting converted by 1.26 times

Do Not Email	-1.1763
TotalVisits	0.1594
Total Time Spent on Website	1.0311
Lead Origin_Landing Page Submission	-0.4113
Lead Origin_Others	3.6057
Last Notable Activity_Modified	-0.8604
What is your current occupation_Student	0.5524
What is your current occupation_Working Professional	2.7175
Last Activity_Olark Chat Conversation	-0.9595
Last Activity_Page Visited on Website	-0.3573
Last Activity_SMS Sent	1.2559
Lead Source_Google	0.1831
Lead Source_Olark Chat	1.1586
Specialization_Others	-0.4002

- Based on the model created (Explained above) and its metrics explained on the right, we seem to have arrived at a good model to predict whether the lead would convert or not
- Based on the probability of conversion, provided by the model , a Lead Score has been provided (1 to 100) – higher the lead score, higher the chances of conversion – which should be referred by the Sales team to prioritize the phone calls

Confusion Matrix

Predicted → Actual ↓ Test Data 0.3 cutoff	Not converted	Converted
Not Converted	1318 (TN)	396 (FP)
Converted	154 (FN)	898 (TP)

Metrics

Metric (On test data – 0.3 cutoff)	%age
Accuracy	0.801
Sensitivity	0.854
Specificity	0.769
Precision	0.694

Thank You