Leads Scoring Case Study

Submitted By: Prerak Bhardwaj & Pieyush C Joy
UpGrad Batch Of August, 2020-21

PROBLEM STATEMENT

X Education, an Education company, sells online courses to Professionals. They have abundance of potential leads who visit their website enquiring about the course. But the problem lies in the fact that around 70% of those potential leads don't pan out (i.e., Don't enrol in their course).

The company has hired our team of Data Analysts to pinpoint the traits that a lead shows to ascertain, with a respectable accuracy, if the lead will turn into an actual customer or not. We are required to build a model that gives these leads a score between 0-100. This will help their sales team to focus on only the clients which have a high leads score.

OBJECTIVES

- X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.
- 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.
- We're required to streamline this process by preparing a model that scores each lead.

DATA INFORMATION

- **'Leads.csv'** contains all the information about the leads generated through various sources and their activities.

  This file contains 9240 rows and 37 columns.

  Out of 37 columns, 7 are numeric columns and 30 are non-numeric or categorical columns.

  Current conversion rate of the leads is 39%.

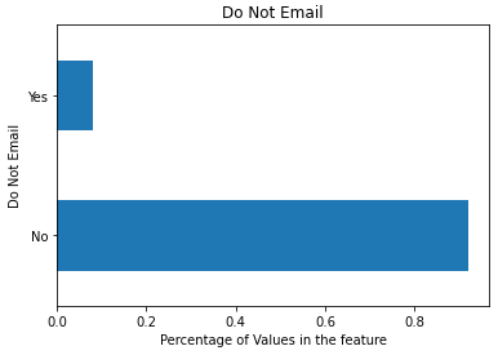
- **'Leads Data Dictionary.csv'** is data dictionary which describes the meaning of the variables present in the "Leads" dataset.

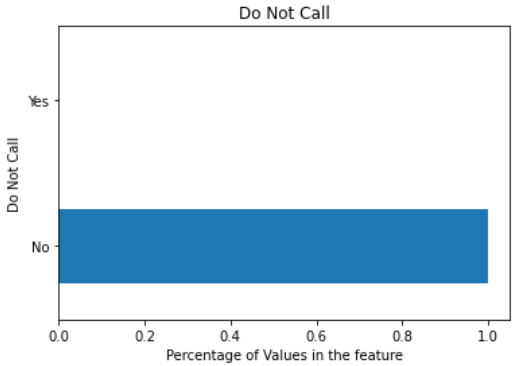
STEPS TO GO ABOUT

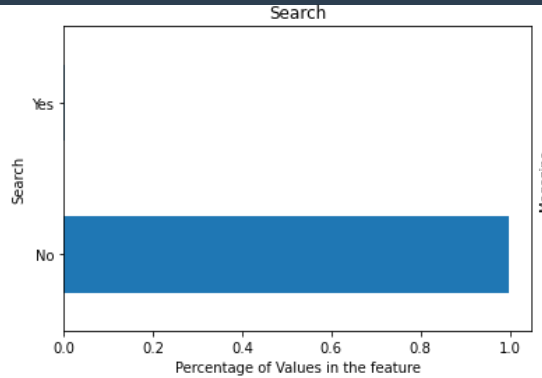
- Imported the necessary data, performed data types checks and inspected the percentage of null values after converting “Select”.
- Data Visualization – Skewness of Categorical features and their removal led to a 40% decline in number of features.
- Dealt with missing values (Imputed numerical features with median due to outliers and categorical ones with “Missing”).
- To reduce the number of dummy columns, grouped niche categorical levels in their respective features as “rare_val”.
- Inspected ‘Class Imbalance’ using a pie-chart, Got rid of ID Columns.
- Did Univariate, Bi-Variate and Multi-Variate analysis on features, created dummies for categorical features.
- Performed train-test split and standardized numerical features. Used RFE to select 25 features to start with.
- Came up with final model “logreg6”, with respectable p-values and non-significant multicollinearity levels.
- Made another Dataframe to store in predictions and values for all possible thresholds.
- Made a Metrics dataframe to evaluate all metrics such as Accuracy, Sensitivity, Specificity, Precision and Recall. Ascertained the threshold using the above metrics. Made an ROC Curve to determine the AUC, which came out to be 96%.
- Evaluated same metrics against test predictions, which performed great… Came up with Score metric in the original dataframe.

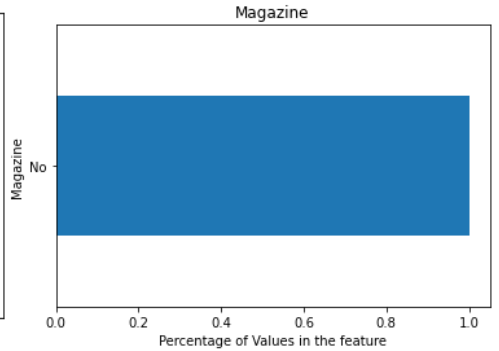
DATA SKEWNESS – CATEGORICAL FEATURES & Y/N VARIABLES

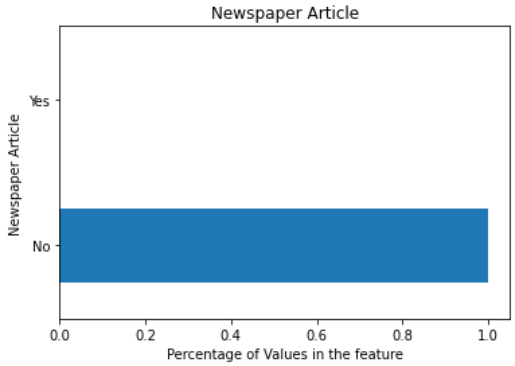
THE SKEWNESS TOWARDS 1 VALUE IS CLEARLY VISIBLE IN ALL THE BELOW CHARTS

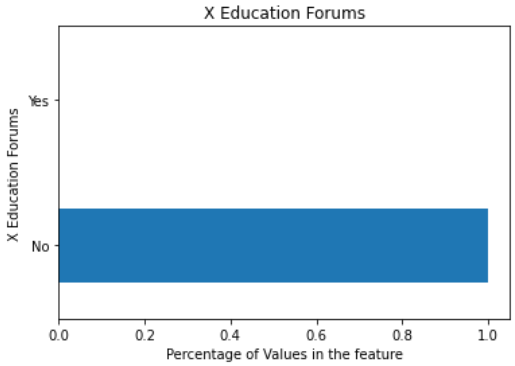


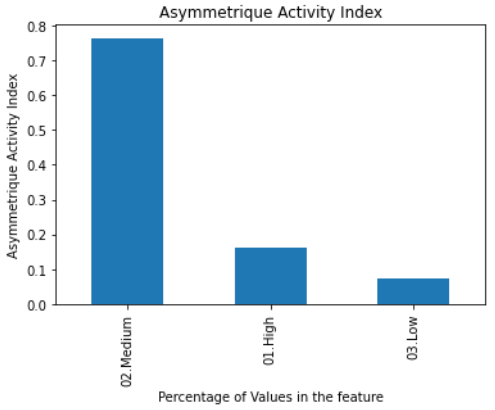


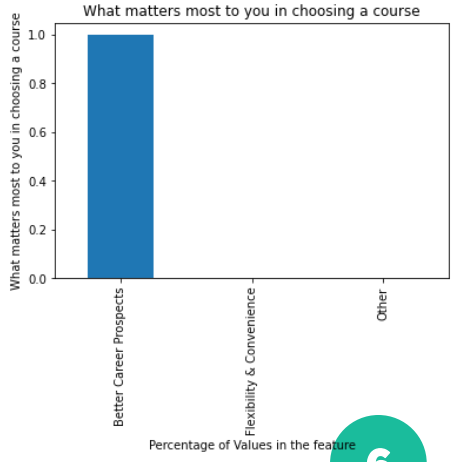




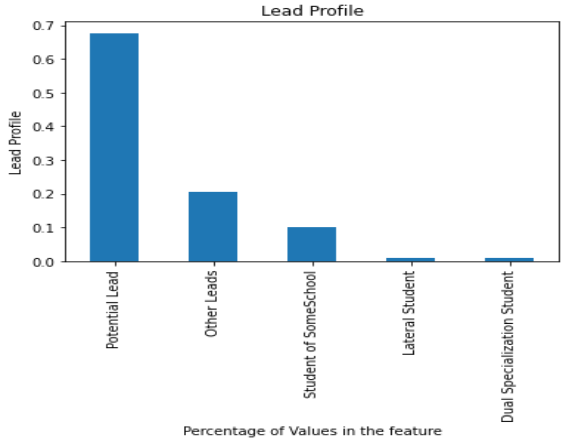


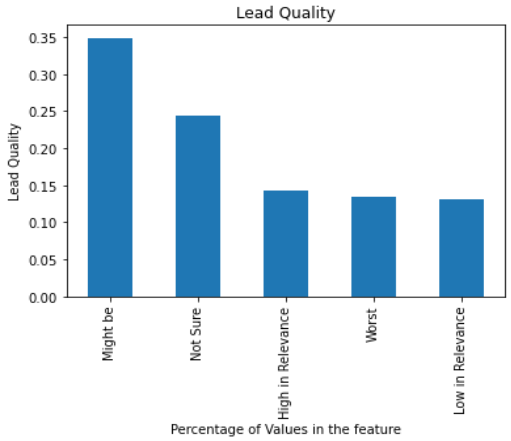


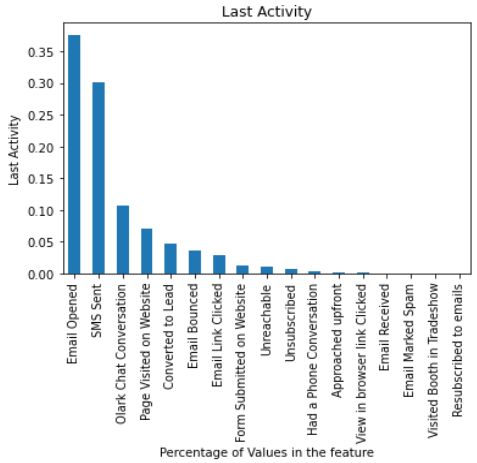


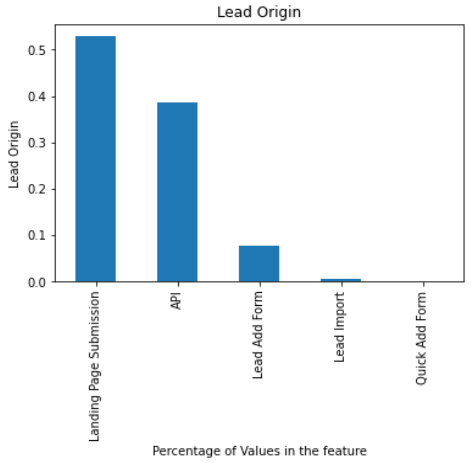


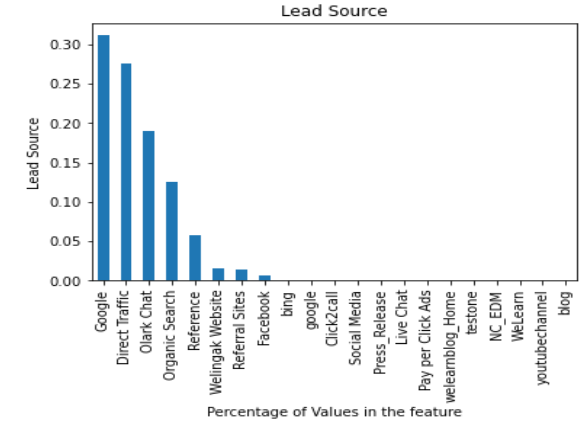
DATA VISUALISATION- CATEGORICAL FEATURES

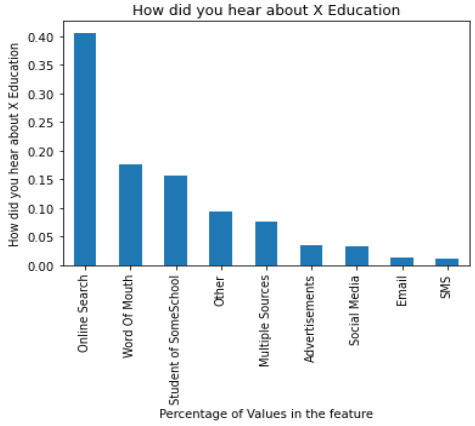


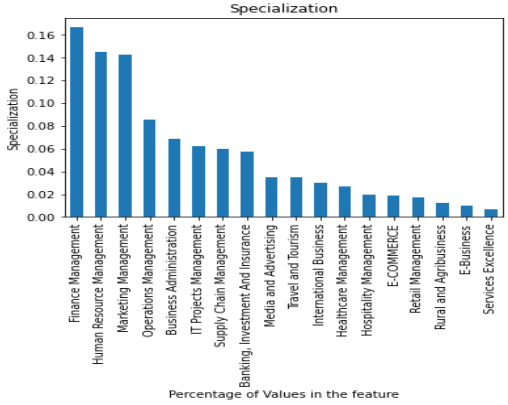


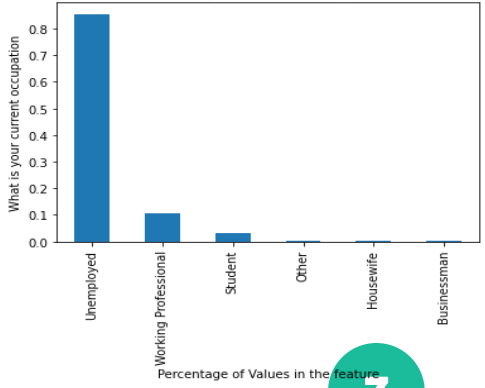




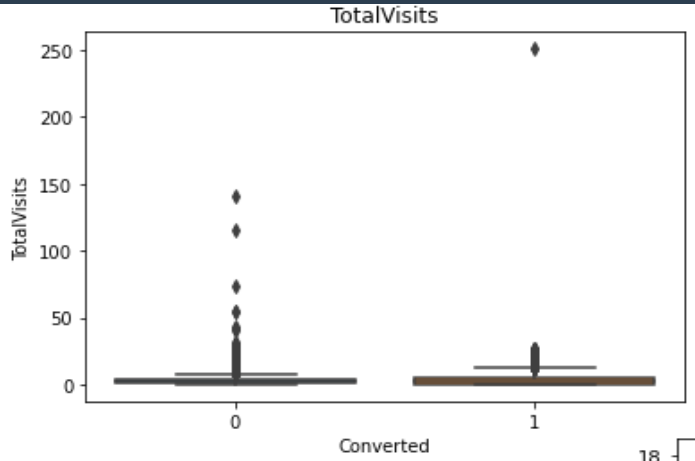


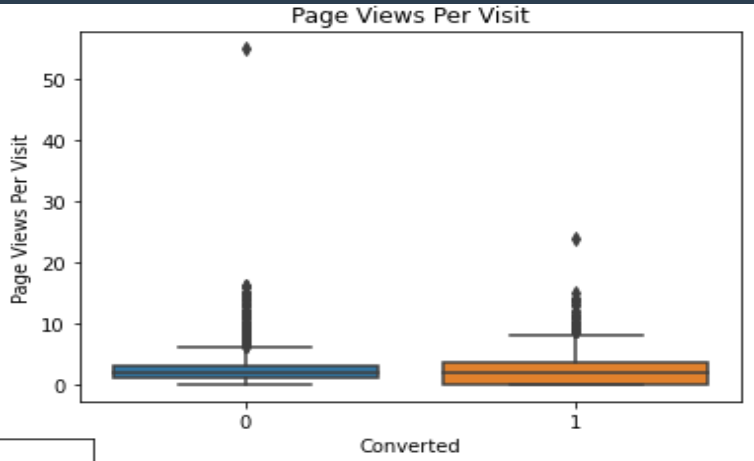


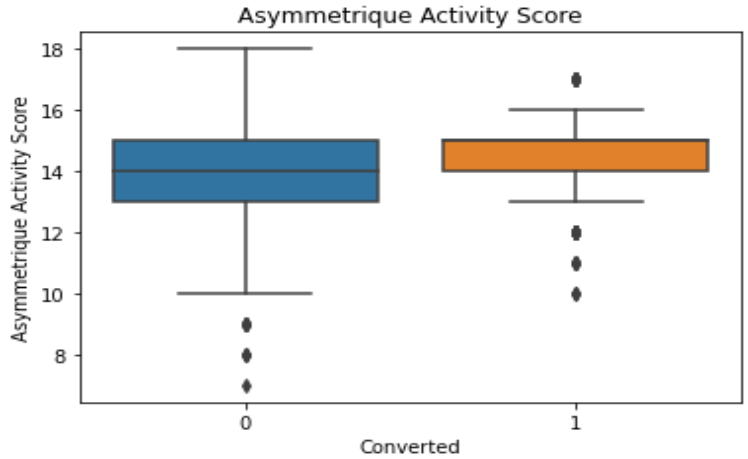




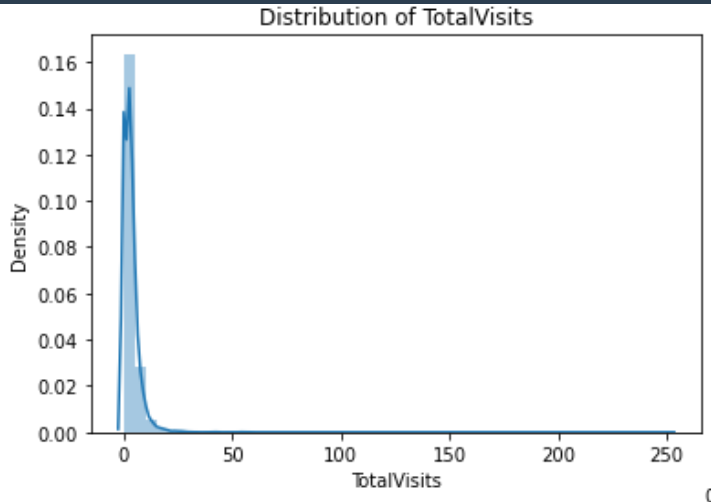
DATA VISUALISATION – OUTLIERS IN NUMERICAL FEATURES



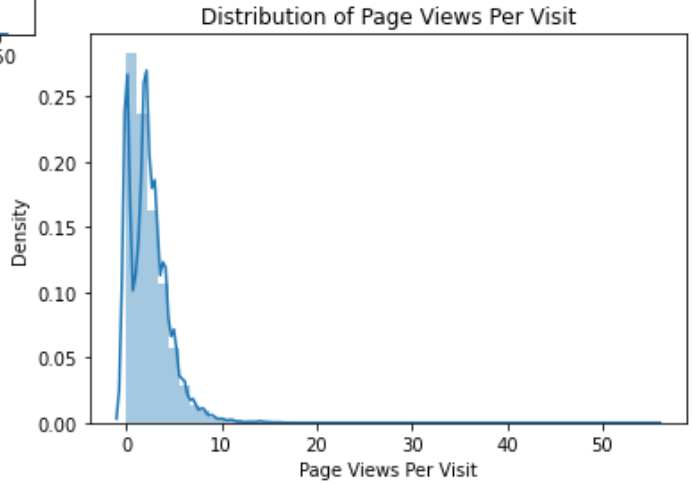


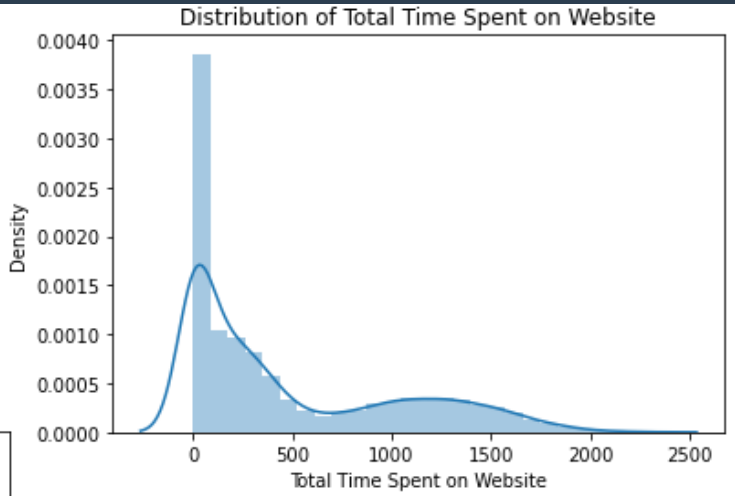


DATA VISUALISATION – DISTRIBUTION OF NUMERIC FEATURES

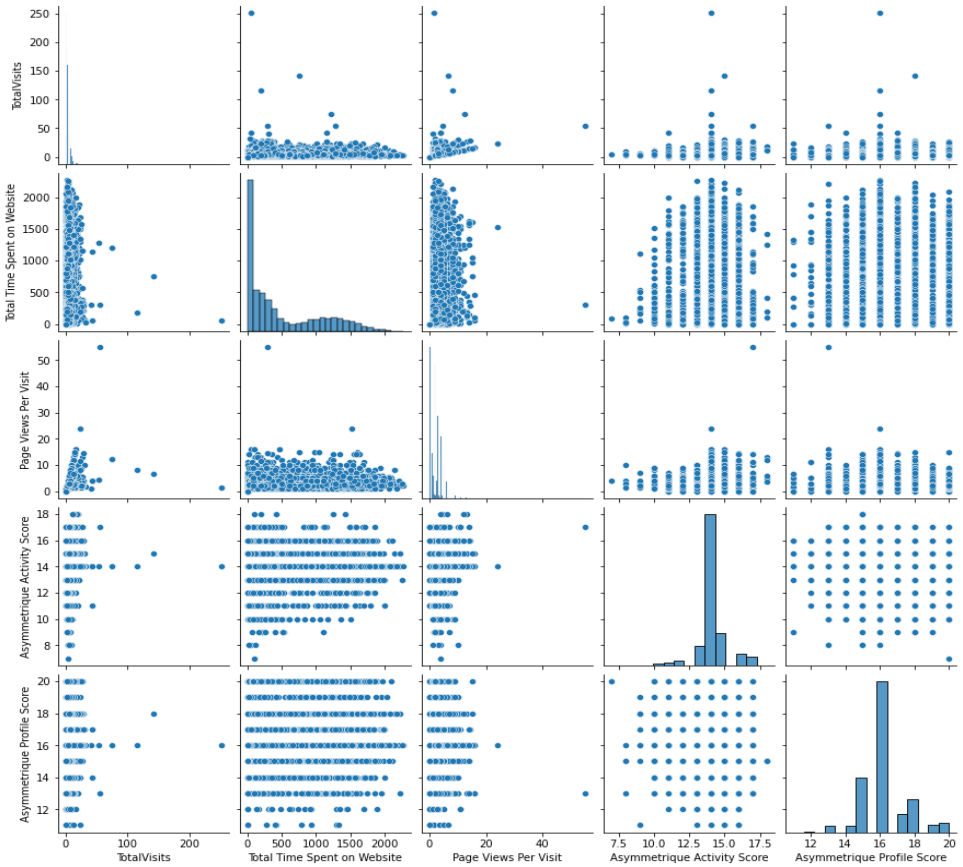


THESE CHARTS CLEARLY DEPICT THE DISTRIBUTION OF NUMERIC FEATURES, WHICH IS CLEARLY NOT NORMAL, BUT THEIR CORRELATION ISN'T MUCH WITH TARGET VARIABLE. THEY WON'T BE IN THE SELECTED FEATURES.





DATA VISUALISATION- BIVARIATE FOR NUMERICAL

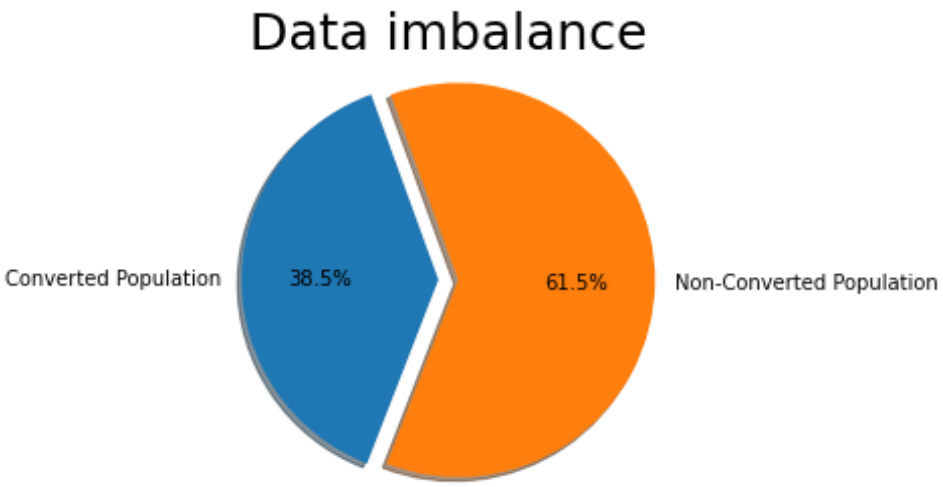


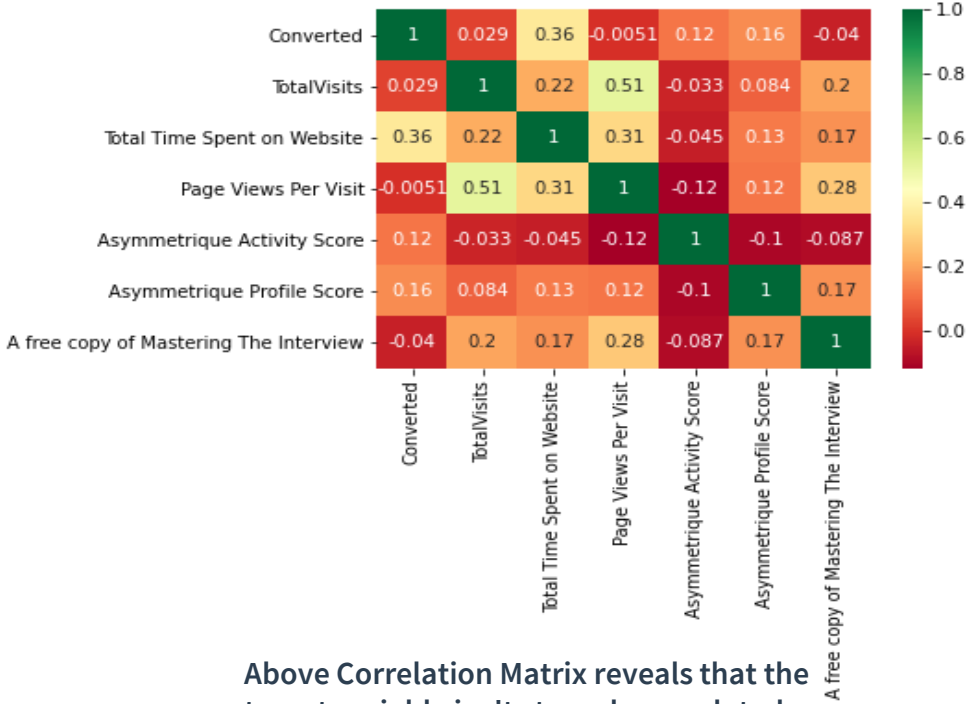
These scatter plots make the relationships between numerical features and discrete numerical features abundantly clear..

Also, these plots make it clear that among our 5 chosen numerical features, 2 are discrete features while other 3 are actual numerical features.

These Discrete Numeric Features have a sort of order to them.

CLASS IMBALANCE & CORRELATION HEATMAP





Above Pie-Chart shows that our population is divided in almost 3:2 ratio of Non-Conversion V/s Conversion, which is not ideal but decent enough to work with

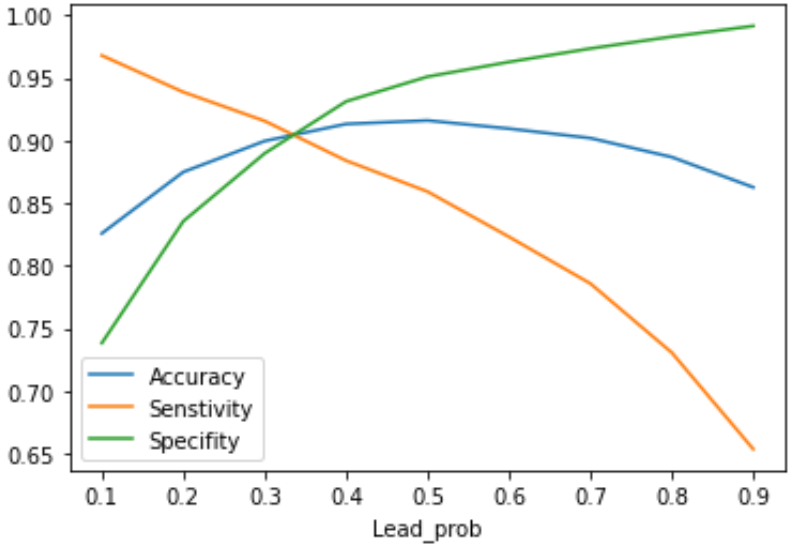
Above Correlation Matrix reveals that the target variable isn't strongly correlated with any of the numerical variables, hence much of our model's strength must come from categorical features

METRICS DATAFRAME

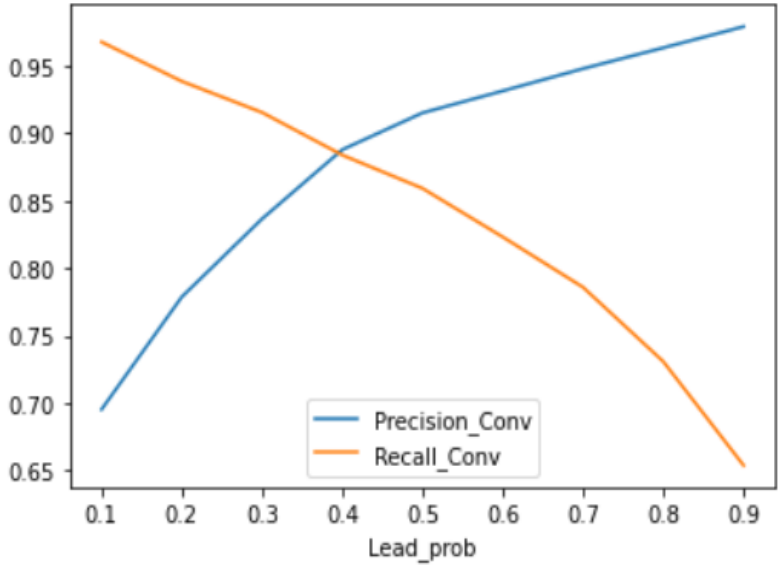
	Lead_prob	Accuracy	Senstivity	Specifity	Precision_Conv	Precision_NoConv	Recall_Conv	Recall_NoConv
0.1	0.1	0.825912	0.967964	0.738381	0.695108	0.973962	0.967964	0.738381
0.2	0.2	0.874923	0.938767	0.835582	0.778675	0.956795	0.938767	0.835582
0.3	0.3	0.899660	0.915653	0.889805	0.836606	0.944813	0.915653	0.889805
0.4	0.4	0.913265	0.884023	0.931284	0.887984	0.928732	0.884023	0.931284
0.5	0.5	0.916048	0.859286	0.951024	0.915335	0.916446	0.859286	0.951024
0.6	0.6	0.909555	0.823195	0.962769	0.931620	0.898345	0.823195	0.962769
0.7	0.7	0.901979	0.785888	0.973513	0.948141	0.880651	0.785888	0.973513
0.8	0.8	0.886827	0.730738	0.983008	0.963636	0.855589	0.730738	0.983008
0.9	0.9	0.862709	0.653690	0.991504	0.979344	0.822895	0.653690	0.991504

THE ABOVE DATAFRAME SHOWS DIFFERENT PROBABILITY THRESHOLDS AS AGAINST THEIR RESPECTIVE EVALUATION METRICS.
OUR CHOSEN THRESHOLD “0.4”, PERFORMS SPLENDIDLY IN ALL OF THEM.

PLOT CHARTS VISUALIZATION – EVALUATION METRICS



The above plot chart shows the SENSITIVITY SPECIFICITY TRADEOFF, that occurs because we can't have the highest of both… We're supposed to find the perfect balance as per our needs. This charts specifies "0.3" being that golden mark

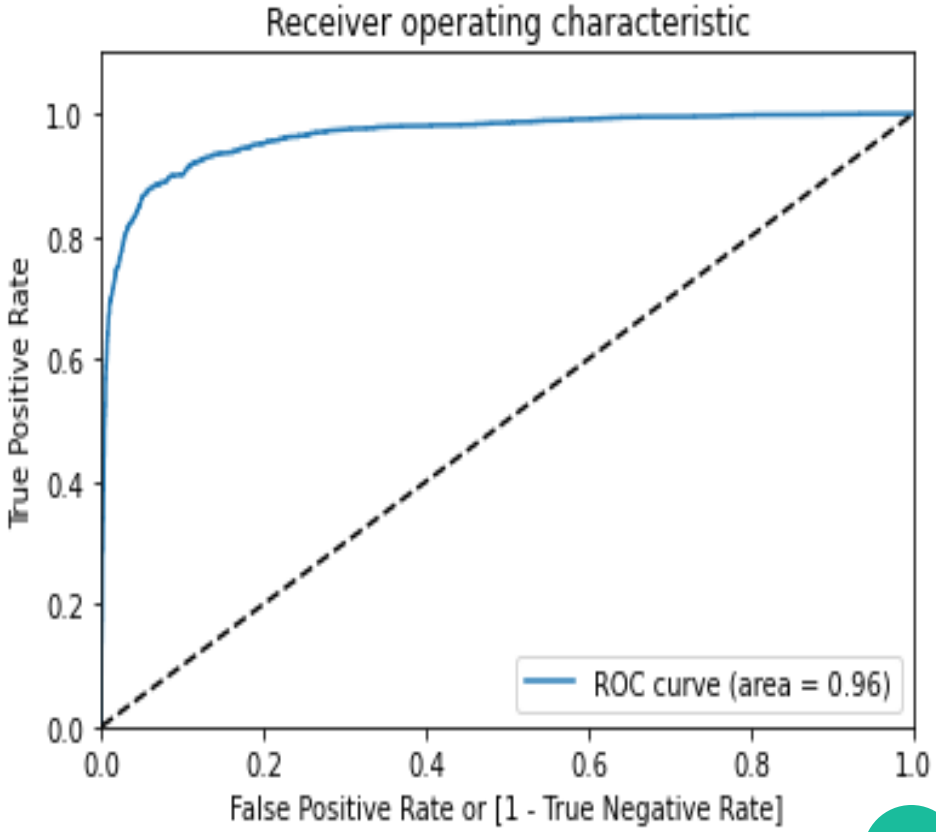


The above chart shows the PRECISION of class "1" and RECALL of class "1", and depicts a similar trade-off, which proves, "0.4" being the optimal threshold…

ROC CURVE – AREA UNDER THE CURVE

RECEIVER OPERATING CHARACTERISTIC CURVE

ROC, by determining the AUC - Area Under The Curve, judges the goodness of the model, by plotting the True Positivity Rate against False Positive Rate.
Since, our ROC is pretty close to the upper left part of the chart, it means the model is pretty well built.
The AUC in our case is 96%, which is just spectacular…



CONCLUSION: -

- From our final model, it is clear that numerical features don’t play as important role in lead’s conversion as does the categorical features like Tags, Lead Profile, Last Activity, Occupation of the candidate etc.
- The Score metric that is derived from all this, by multiplying probability of each conversion by 100 and rounding up to 0 decimal point. This metric should prove to be a boon for the Sales Team and they can finally focus on selling instead of judging whom to call.
- Candidates must be somehow made to abide by certain highly positive indicators like Lead add forms, Tags of reverting after E-Mail, these can certainly increase their chances of converting into actual customers.

	Score	Converted
Score	1.000000	0.851643
Converted	0.851643	1.000000

THIS SHOWS THAT OUR DERIVED METRIC, SCORE IS WORKING AS EXPECTED, IT IS STRONGLY CORRELATED WITH TARGET VARIABLE, CONVERTED, IN THE POSITIVE SENSE