LEAD SCORING CASE STUDY

PROBLEM STATEMENT:

- An education company named X Education sells online courses to industry professionals.
- The company markets its courses on several websites and search engines like Google.
- When people fill up a form providing their email address or phone number, they are classified to be a lead. The typical lead conversion rate at X education is around 30%.
- Although X Education gets a lot of leads, its lead conversion rate is very poor.
- As a Data scientist we need to identify the 'Hot leads' so that 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 APPROACH

- Steps :
- Data Inspection – Dropping of Unnecessary columns
- Missing Value Treatment and Outlier Analysis
- Univariate and Bi-variate Analysis
- Data Pre-processing – Scaling and Train-Test split
- Modeling and Feature selection using RFE
- Finding Optimal Cutoff using ROC curve
- Sensitivity – specificity view predictions on Test set
- Precision–Recall view predictions on Test set

MISSING VALUE TREATMENT AND OUTLIER ANALYSIS

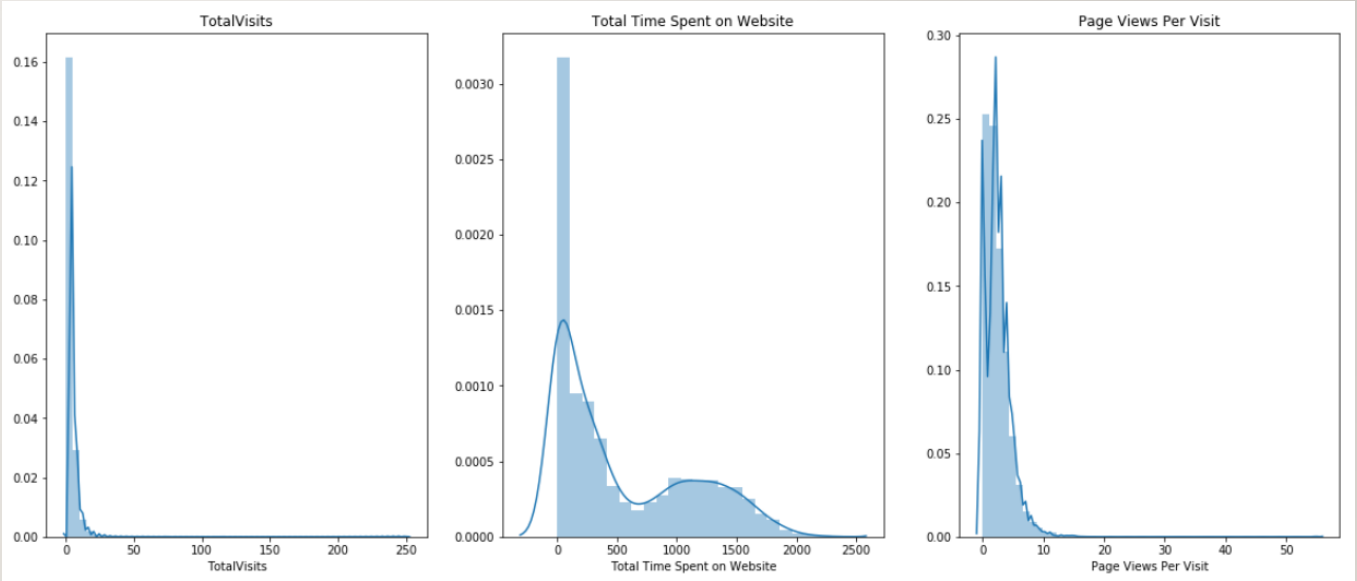
- There are many columns where we have high number of missing values.
- As there are 9000 datapoints in our dataframe, eliminating the columns having greater than 3000 missing values as they are of no use to us.
- For one of the columns its been observed that, around 95% of the data contained in the “Country" feature contains same value, this does not give much insight for the analysis.
- For the customers who have not selected any option by default 'Select' value has been selected for that column. These values are nothing but missing values, hence we need to analyses these value counts of this level 'Select' in all the columns and treat accordingly.
- In some features we have few missing values in the column, its better to drop only the missing value rows instead of dropping column so that we wont loose significant data.

DROPPING OF UNNECESSARY COLUMNS

- From business perspective "City" feature does not impact much whether a customer will convert or not. Hence Dropping “Country” and “City” columns would be opted.
- There are few columns in our dataframe which only one value was majorly present for all the data points. it's best that we drop these columns as they won't help with our analysis.
- We also choose to drop the columns such as ‘Prospect ID’ and ‘Lead Number’ which contains unique values as it will not be useful for us in the analysis.
- After dropping all the missing value columns and unnecessary columns we have 69% of the rows retained for our analysis.

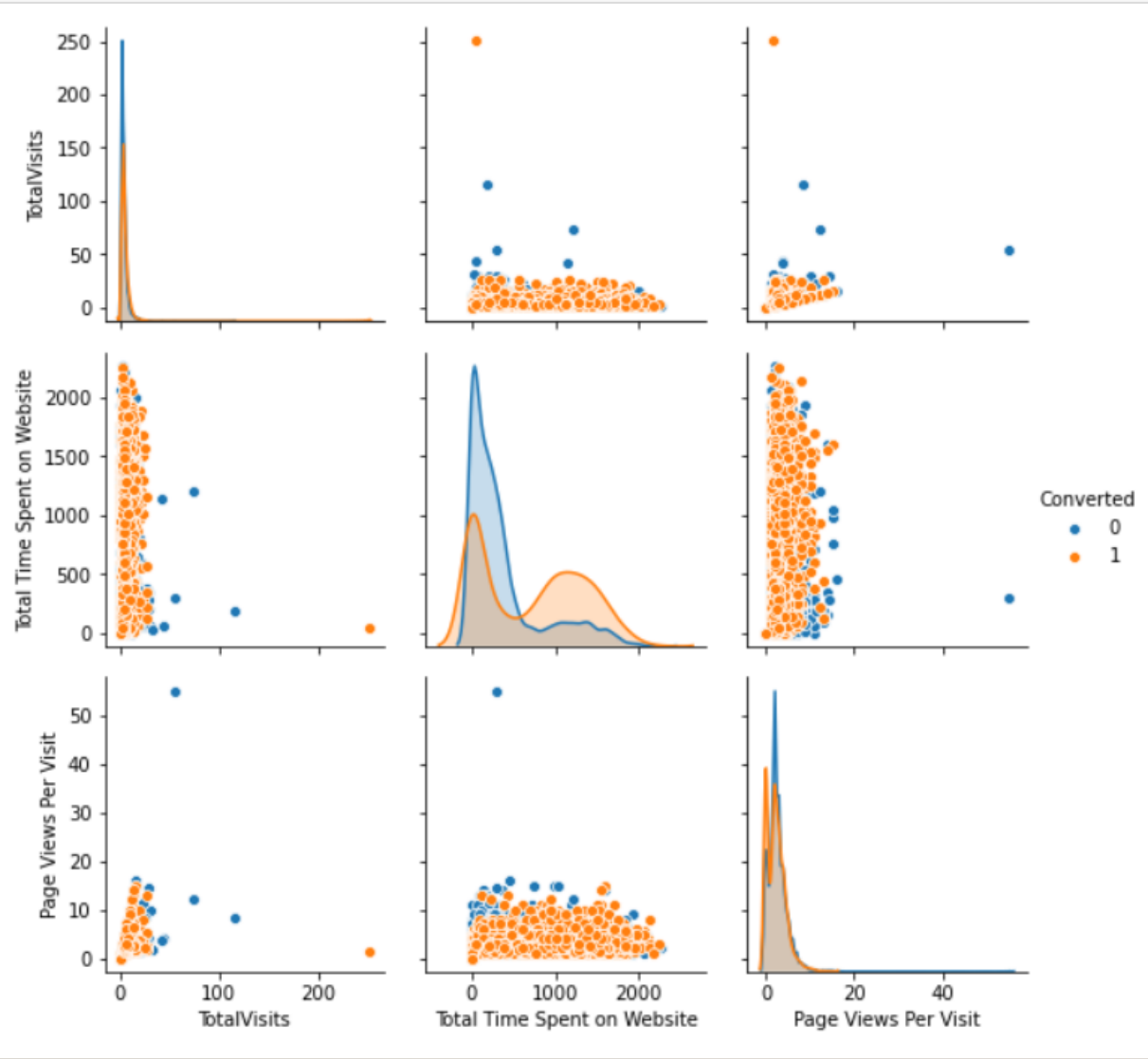
UNIVARIATE ANALYSIS

- Univariate Analysis is done for the continuous variables and it is observed that all the three columns have outliers as the plots are right skewed.



BI-VARIATE ANALYSIS

- From the above chart, we could see that majority of leads are through 'Landing Page Submission' followed by 'API'.
- Conversion rate is around 43% for 'Landing Page Submission' while it is around 50% for 'API'.
- Interesting thing here is that Leads through 'Lead Add From' have around 93% conversion rate.



BI VARIATE ANALYSIS : NUMERICAL-CATEGORICAL

- Customers who are not converted are spending more time on the website compared with customers who are converted
- From the pairplots above it is evident that we have outliers on the numerical columns
- To handle the skewness, we are implementing **power transformation** which also handles outliers.

DUMMY VARIABLE CREATION:

- Collected all the object variables into a temp variable and created a dummy variables using get_dummies and drop_first parameter
- Concatinated dummy dataframe with original dataframe
- As “Specialization” variable has “select” value, created dummy variable for this and dropped the column for select value. (“Specialization_Select”)
- Added dummy variables for Specialization column to the original dataframe.

DATA PRE-PROCESSING – SCALING AND TRAIN-TEST SPLIT

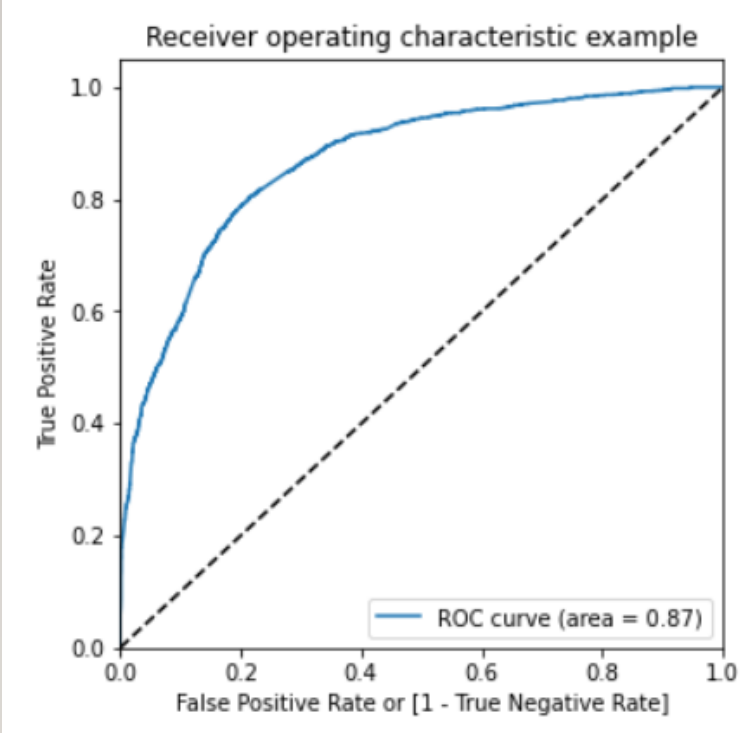
- We have slit the data into Train and Test sets with 70:30 proportions respectively.
- Scaling is done for the train dataset using "fit_transform"
- We have used "MinMaxScaler" for scaling purpose.
- Looked at the Correlation matrix and plotted a heatmap and its been observed that few features have high correlation.

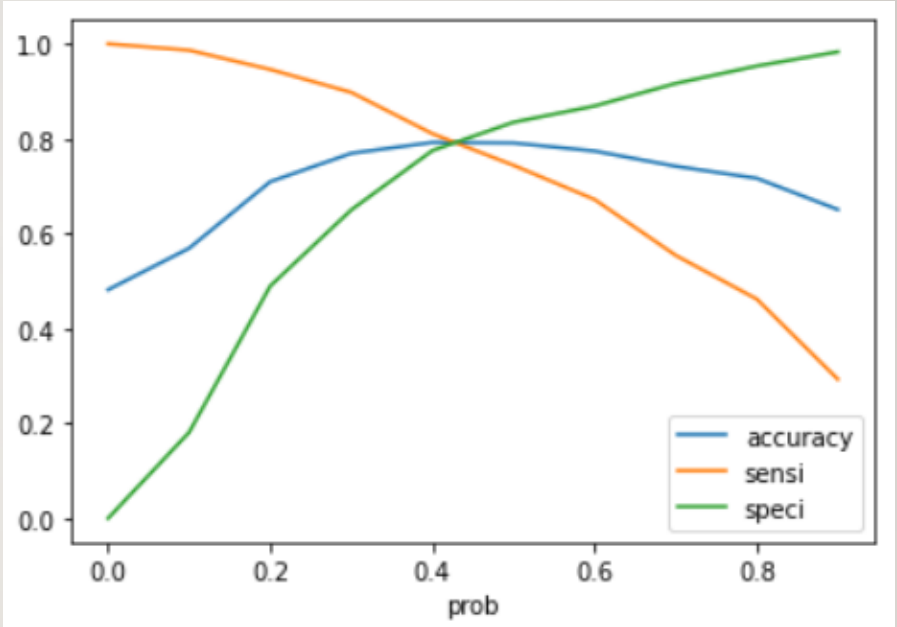
MODEL BUILDING

- As we have more than 100 features, manual selection of features is time consuming process. We have used RFE to select top 20 features and proceeded with manual changes.
- After looking at the model summary we noticed that there are features with high p-values.
- After checking the P-values and VIF values we removed 'Lead Source_Reference' feature.
- Later we iterated through same steps to achieve at the final set of features.

MODEL EVALUATION & FINDING OPTIMAL CUT-OFF:

- We predicted the probability for each observation using the final model.
- Choosing the arbitrary cut-off of 0.5 we have predicted the conversion and calculated the accuracy. (0.79)
- Plotted ROC curve and its been observed that area under the curve is 0.87.
- Plotted accuracy, sensitivity and specificity for cut-off ranging from 0.0 to 0.09.
- After looking at the plots its noticed that the cut-off is 0.42





ROC CURVE AND ACCURACY, SENSITIVITY, SPECIFICITY

PREDICTIONS ON TEST SET: SENSITIVITY AND SPECIFICITY VIEW

- We have predicted target variable (Converted) with optimal cut-off of 0.42 above which the lead is considered as "Hot Lead".
- Below are the metrics for test set predictions
- From the 'Sensitivity' and 'Specificity' view :
- we have got good accuracy(78%)
- Sensitivity : 79%
- Specificity : 78%

Confusion matrics

Actual/ Predicted	0	1
0	773	223
1	193	723

PREDICTIONS ON TEST SET: PRECISION AND RECALL VIEW

- We have predicted target variable (Converted) with optimal cut-off of 0.44 above which the lead is considered as "Hot Lead".
- Below are the metrics for test set predictions
- From the 'Precision' and 'Recall' view :
- we have got good accuracy(78%)
- Precision: 78%
- Recall: 77%

Confusion matrics

Actual/ Predicted	0	1
0	793	203
1	211	705

SUMMARY

- Overall we have achieved a good metrics (accuracy) from both specificity-sensitivity and precision-recall views for test set predictions.
- We can go-ahead with this model for future predictions to identify the 'Hot Leads'
- The top three variable which contribute most towards the probability of a lead getting converted are:
- **Total Visits**
- **Total Time Spent on Website**
- **Lead Origin_Lead Add Form**