# Leads Score Case Study

Prasad Maharana
&
Bibek Kumar

# Problem Statement

- X Education wants to identify whether a particular profile or lead is a hot lead so they can contact only those leads who have a higher chances of converting into sales.
- They want to expand the leads to sales conversion funnel so maximum conversion is obtained from the leads.
- They also want the solution to adapt to changes when the company wants maximum leads to increase the reach or wants maximum conversion to sales from hot leads

# Data Cleaning- Leads Information

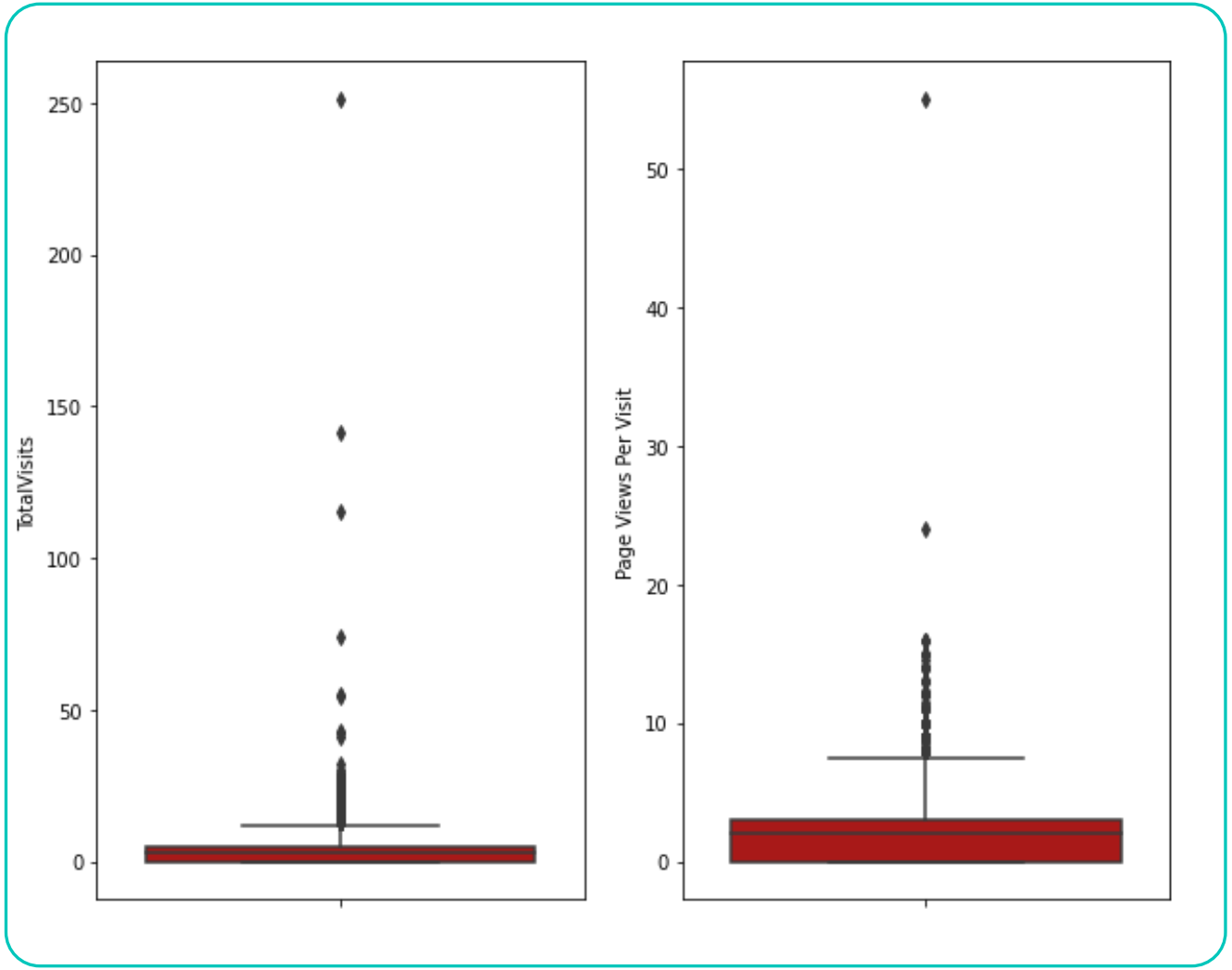
- The initial Dimension was 37 columns and 9240 rows
- There were 7 numerical columns (variables) and 30 categorical columns (variables)
- We dropped redundant columns which did not add value to the data
- We replaced the Select value with Nan or Null value in pandas
- We dropped columns with missing values greater than 30%
- We imputed columns with missing values lesser than 30% with the maximum occurring categorical value or max value
- We reformatted column “Lead source” to avoid data duplication
- We were finally set with 30 columns and 9420 rows

# Data Transformation

- We replaced categorical columns with Yes/No values with 1/0 binary values
- We created dummy variables for the 8 columns which had categorical values but not the Yes/No columns and drop the first columns for each dummy variable(To avoid the dummy variable trap) and joined it to the main dataset
- Dropped the original columns which were converted into dummy variables
- We then performed an outlier check on the result

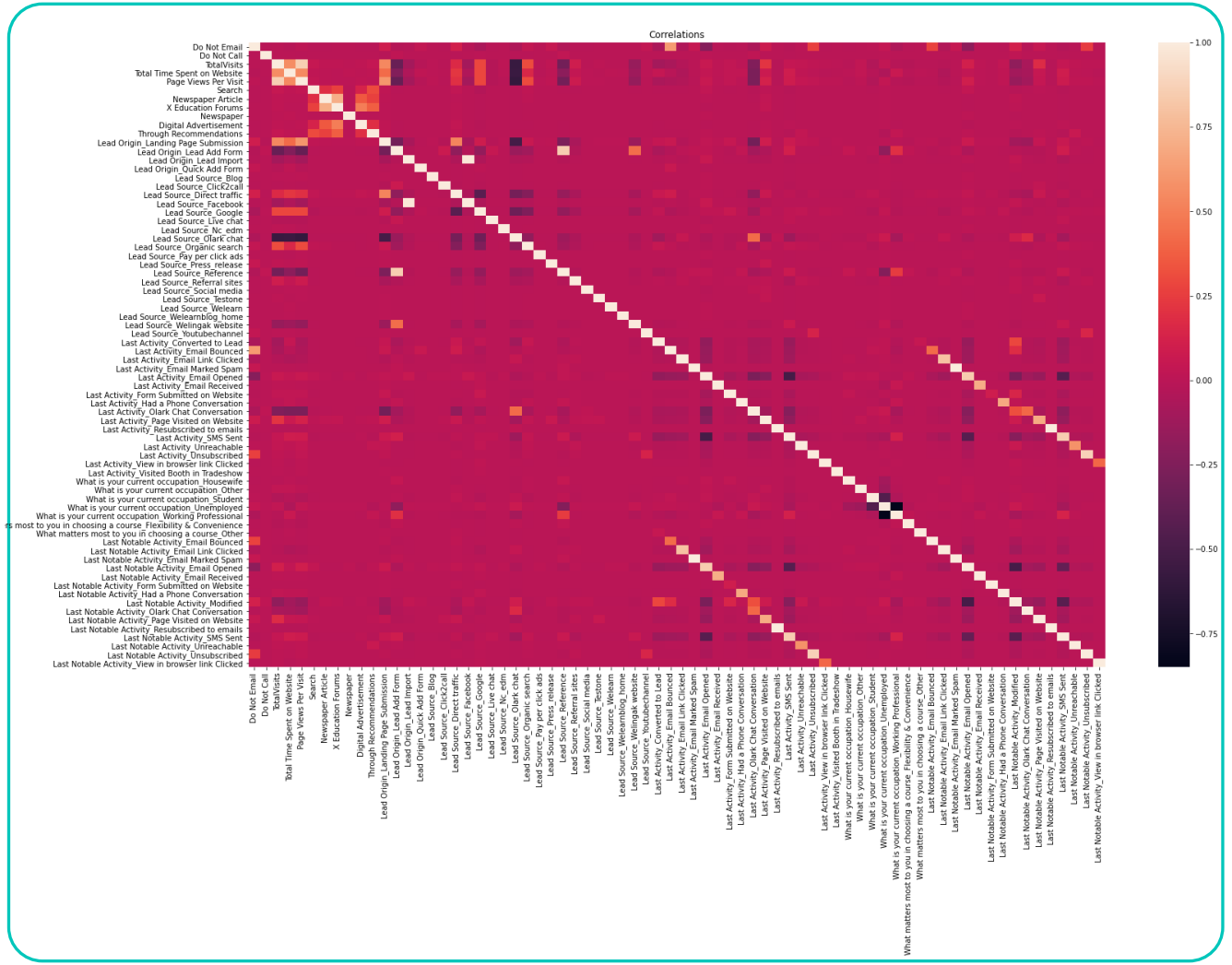
# Outlier Detection

- The Blox plot for TotalVisits and Page Views Per Visit shows presence of outliers
- The outliers can be either dealt with Binning or by using a scaling
- We decided to not perform outlier treatment and normalize using a min max scaler in the later stage



# Data Preparation

- We begin with Splitting of the independent (Features) variables and the dependent (Target) variables
- We Checked the correlation between the Feature variables and dropped features which had a high correlation with multiple other features
- It was difficult to detect multicollinearity with so many variables so we dropped the ones that were visible and then dropped the others in the model tuning section



# Train Test Split

- We started with splitting the data into Train set and test set
- We scaled the features using a min max scaler to normalize the numerical variables
- We checked for class imbalance and found the target class to be around 39% (Lead conversion rate)
- We then continued to build the model (Modelling)

# Modelling

- We used the stats model api to create a logistic regression model with all the features and then checked the summary for the model
- There were around 70 variables so we used RFE to select the best 20 variables
- We carry forwarded only those variables that rfe supported and created another model basis the new columns
- After checking the model summary, we started tuning the model
- We dropped high p-value features one by one and rebuild the model after dropping variables.
- We checked the Variance inflation Factor at each rebuild to ensure no multicollinearity

## Features

| | Features | VIF |
|---|---|---|
| 0 | const | 7.77 |
| 14 | Last Notable Activity_Modified | 1.99 |
| 9 | Last Activity_Email Bounced | 1.76 |
| 1 | Do Not Email | 1.75 |
| 10 | Last Activity_Olark Chat Conversation | 1.69 |
| 13 | Last Notable Activity_Email Opened | 1.61 |
| 4 | Page Views Per Visit | 1.54 |
| 5 | Lead Origin_Lead Add Form | 1.47 |
| 2 | TotalVisits | 1.40 |
| 15 | Last Notable Activity_Olark Chat Conversation | 1.36 |
| 7 | Lead Source_Welingak website | 1.22 |
| 8 | Last Activity_Converted to Lead | 1.20 |
| 16 | Last Notable Activity_Page Visited on Website | 1.19 |
| 3 | Total Time Spent on Website | 1.17 |
| 6 | Lead Source_Direct traffic | 1.10 |
| 11 | What is your current occupation_Working Profes... | 1.09 |
| 12 | Last Notable Activity_Email Link Clicked | 1.06 |

## Model Summary

Generalized Linear Model Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | Converted | No. Observations: | 6468 |
| Model: | GLM | Df Residuals: | 6451 |
| Model Family: | Binomial | Df Model: | 16 |
| Link Function: | logit | Scale: | 1.0000 |
| Method: | IRLS | Log-Likelihood: | -2704.1 |
| Date: | Tue, 07 Sep 2021 | Deviance: | 5408.2 |
| Time: | 14:52:18 | Pearson chi2: | 7.06e+03 |
| No. Iterations: | 7 | | |
| Covariance Type: | nonrobust | | |

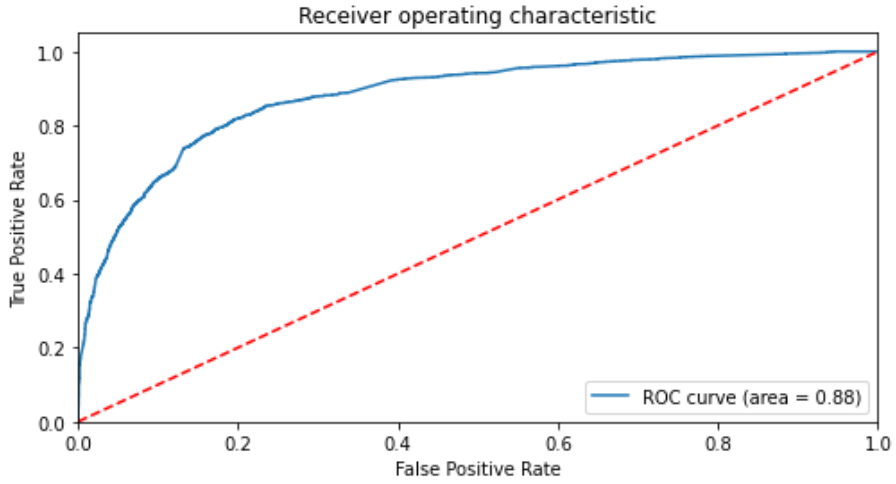
| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | -0.1250 | 0.086 | -1.448 | 0.148 | -0.294 | 0.044 |
| Do Not Email | -1.2932 | 0.196 | -6.614 | 0.000 | -1.676 | -0.910 |
| TotalVisits | 7.6160 | 2.227 | 3.421 | 0.001 | 3.252 | 11.980 |
| Total Time Spent on Website | 4.1736 | 0.153 | 27.210 | 0.000 | 3.873 | 4.474 |
| Page Views Per Visit | -8.2288 | 1.178 | -6.984 | 0.000 | -10.538 | -5.920 |
| Lead Origin_Lead Add Form | 3.0347 | 0.191 | 15.864 | 0.000 | 2.660 | 3.410 |
| Lead Source_Direct traffic | -0.5162 | 0.079 | -6.553 | 0.000 | -0.671 | -0.362 |
| Lead Source_Welingak website | 1.9384 | 0.743 | 2.608 | 0.009 | 0.481 | 3.395 |
| Last Activity_Converted to Lead | -1.2225 | 0.223 | -5.471 | 0.000 | -1.660 | -0.785 |
| Last Activity_Email Bounced | -1.1717 | 0.346 | -3.390 | 0.001 | -1.849 | -0.494 |
| Last Activity_Olark Chat Conversation | -1.0466 | 0.192 | -5.439 | 0.000 | -1.424 | -0.669 |
| What is your current occupation_Working Professional | 2.8064 | 0.189 | 14.870 | 0.000 | 2.436 | 3.176 |
| Last Notable Activity_Email Link Clicked | -1.7978 | 0.272 | -6.609 | 0.000 | -2.331 | -1.265 |
| Last Notable Activity_Email Opened | -1.3344 | 0.086 | -15.429 | 0.000 | -1.504 | -1.165 |
| Last Notable Activity_Modified | -1.7129 | 0.099 | -17.352 | 0.000 | -1.906 | -1.519 |
| Last Notable Activity_Olark Chat Conversation | -1.4171 | 0.373 | -3.795 | 0.000 | -2.149 | -0.685 |
| Last Notable Activity_Page Visited on Website | -1.8468 | 0.201 | -9.198 | 0.000 | -2.240 | -1.453 |

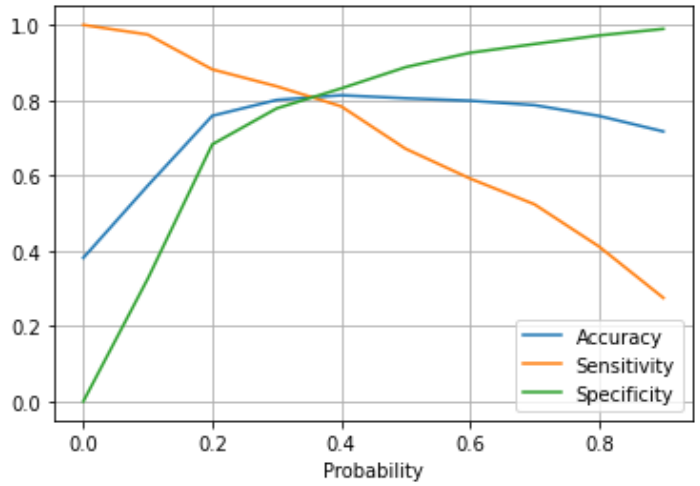
# Final Model with Features

- We concluded that the model has no multicollinearity and did not contain any features with p-value greater than the acceptance range of 5%
- Note : Const is a constant field which is not a part of the final set of features but a step of the model building and should be ignored in features list

# Prediction and Evaluation

- We tested the model by predicting the train set target variable (y_train)
- We plotted the ROC curve basis the model's metrics on the train set and found the area under ROC to be 0.88 which means it was a good model
- We continued to find the optimal cut off for defining which range would define the "Leads Converted (1)" and "Leads not converted (0)"
- For that we created a range of values from 0.0 to 0.9 and calculated the accuracy, sensitivity and specificity for each of the cut-off values and then plotted the three metrics against the probability range.
- The point where all the three metrics met was the optimal cut-off and we found it to be around 0.35



# Prediction and Evaluation

- We set the cut-off for the probability for classification to 0.35 and computed the confusion matrix for the final result set
- The precision was 72% i.e. Ability to identify 72% as leads and recall was 81% i.e. ability to correctly identify as leads was 81%
- This concluded that model was good and we proceeded to predict the test data
- On the test data, we found the precision to be 74% and recall to be 80% which was in range of the test set.

# Leads score

- To assign a leads score, we used the probability column from the prediction with x100 multiplier. The score range was 0-100%.
- This score signifies that each lead with a higher score meant a hot lead and each lead with a lower score was a cold lead.
- The model could easily adapt to the two conditions of the sales team.
- If they want maximum reach, they can contact leads with scores in range of 40-100%
- If they want maximum conversion but lesser calls, they can contact leads with scores in range of 80-100%

# Conclusion

- The Accuracy, Precision and Recall score we got from test set in aceptable range.
- We have high recall score than precision score which we were exactly looking for.
- In business terms, this model has an ability to adjust with the company's requirements in coming future.
- This concludes that the model is in stable state.

Important features responsible for good conversion rate or the ones' which contributes more towards the probability of a lead getting converted are :

- Last Notable Activity_Had a Phone Conversation
- Total Visits
- Total Time Spend
- What is your current occupation_Working Professional