Lead Scoring Case Study

Ankita Bondre
Ankush Farmania
Aninda Chakroborty

Problem Statement

- X Education, an online education company, is facing a challenge with its lead conversion process. Despite acquiring a substantial number of leads daily through website visits, form submissions, and referrals, their current lead conversion rate is only around 30%. This poor conversion rate indicates inefficiency in identifying and targeting potential leads effectively.
- In an effort to improve lead conversion rates, X Education seeks a solution to identify and prioritize "Hot Leads" – those most likely to convert into paying customers. The company aims to enhance its lead management system to increase the conversion rate, optimize the sales team's efforts, and ultimately improve overall business profitability.
- How can X Education efficiently identify and prioritize potential "Hot Leads" to enhance its lead conversion rate?

Case Study Goals

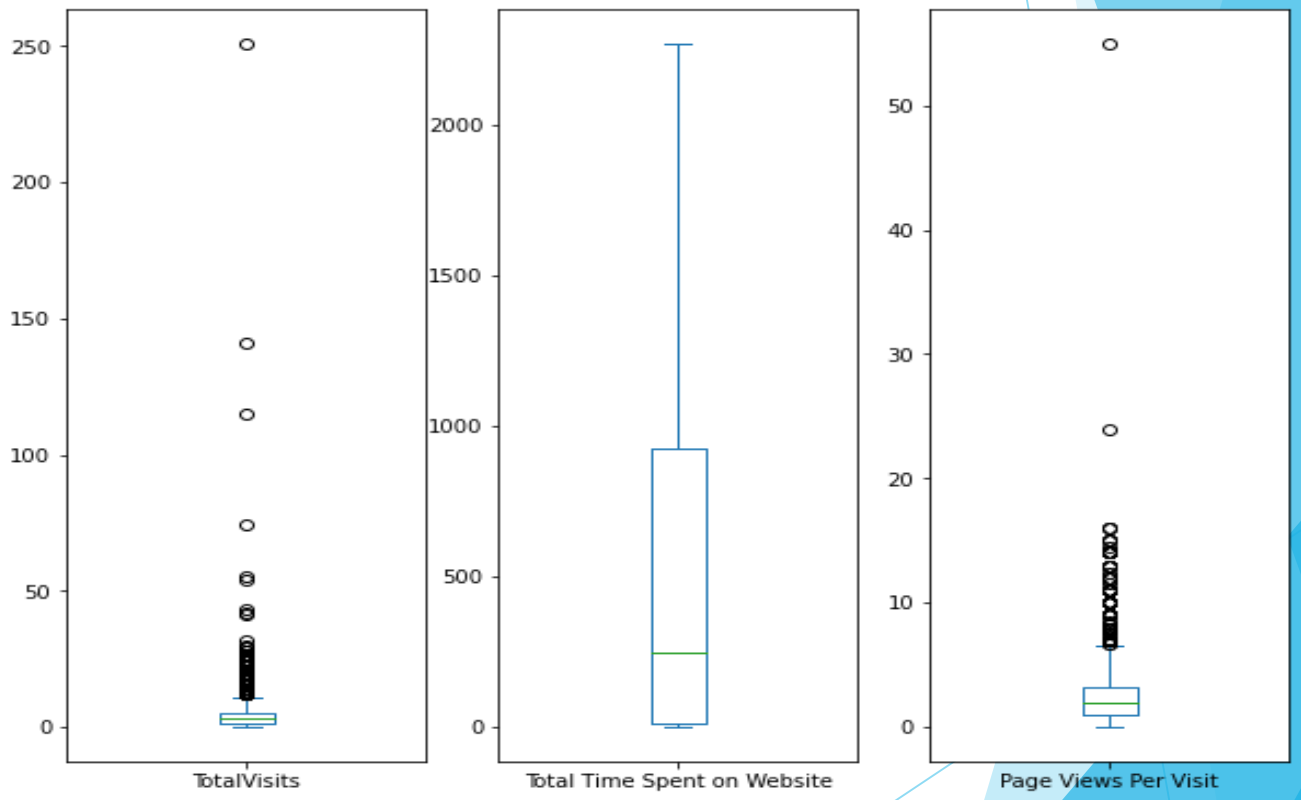
- Develop a logistic regression model to assign lead scores from 0 to 100, enabling efficient targeting of potential leads. Higher scores signify hotter leads with a greater likelihood of conversion, while lower scores represent colder leads less likely to convert.
- Ensure model flexibility to address future changes in the company's requirements by adapting to additional problems and challenges presented by the company.

Model Building Steps

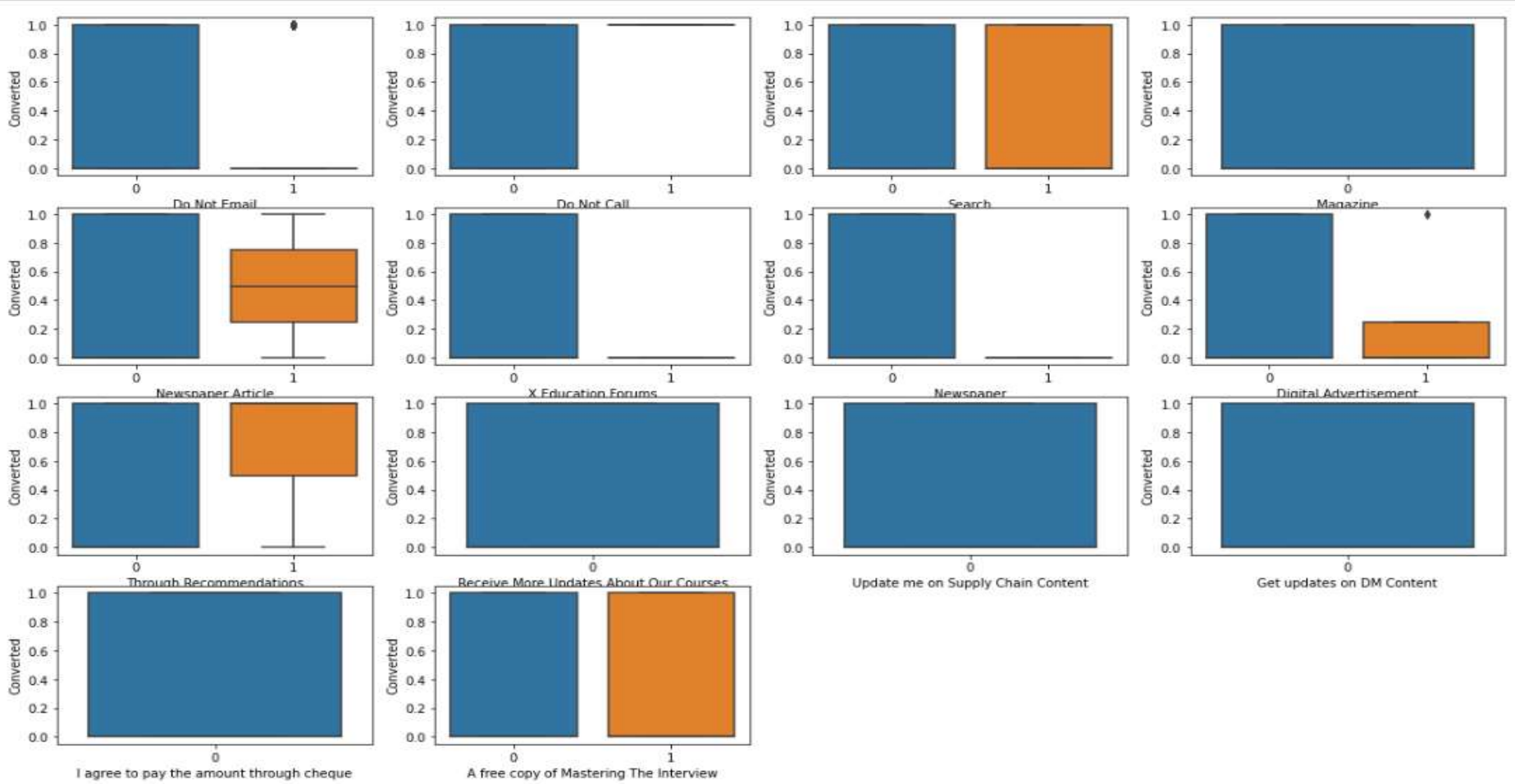
- 1. Import all libraries, read and clean the input data, and carrying out EDA
- 2. Train-Test Split
- 3. Feature Scaling
- 4. Checking Conversion Rate
- 5. Feature Selection using RFE, looking at correlations
- 6. Model Building
- 7. Plotting ROC Curve
- 8. Finding Optimal Cut-off to measure Accuracy metrics of the model
- 9. Making predictions and calculating score on Test data

Exploratory Data Analysis (EDA)

It is observed that both "TotalVisits" and "Page Views Per Visit" contain outliers, as evident from the box plots.
We shall drop these outliers as these may disturb our analysis and model building.



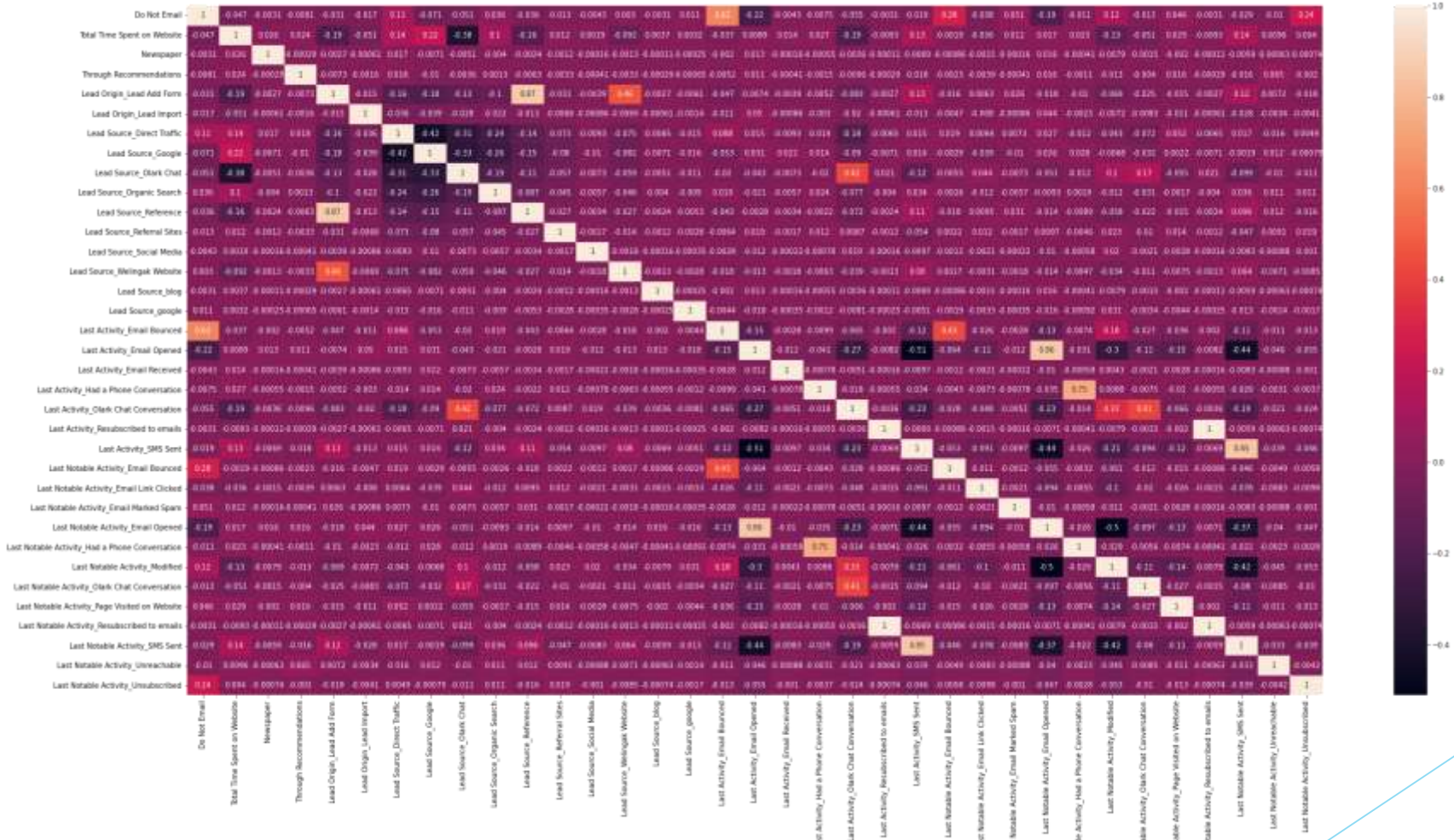
Studying/Visualizing Target variable (Converted) with other Categorical Columns



Studying/Visualizing Target variable (Converted) with other Categorical Columns

- As seen in the figure, categorical columns "Magazine", "Receive More Updates About Our Courses", "Update me on Supply Chain Content", "Get updates on DM Content", "I agree to pay the amount through cheque" are being removed from the dataset because they do not provide meaningful or informative variation with respect to the target variable (i.e. Converted), and including them in the analysis would likely not contribute to predictive modeling or understanding lead behavior.
- Other variables shows some significant variation with the target variable (i.e. Converted)

Looking At Correlations

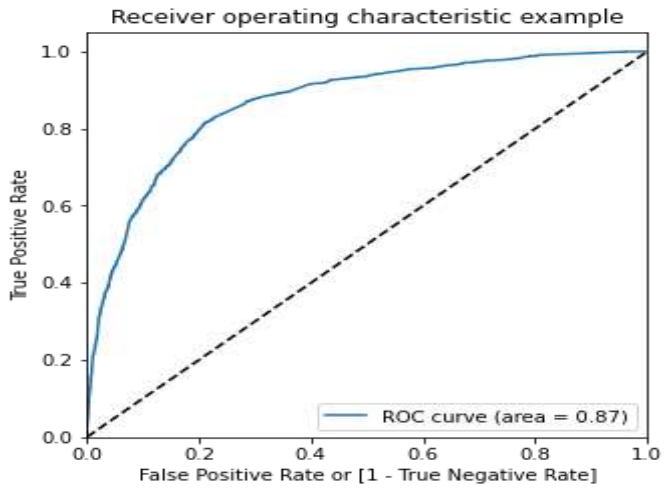


Summary from Correlations

- Excluded the dummy variables like "*Lead Origin_Lead Add Form*", "*Last Activity_Email Opened*", "*Last Activity_Had a Phone Conversation*", "*Last Activity_Resubscribed to emails*", "*Last Activity_SMS Sent*", and "*Last Notable Activity_Email Opened*" due to their strong correlation, exceeding 50%, with other variables. Having them in our model may lead to overfitting of the model.

Model Evaluation

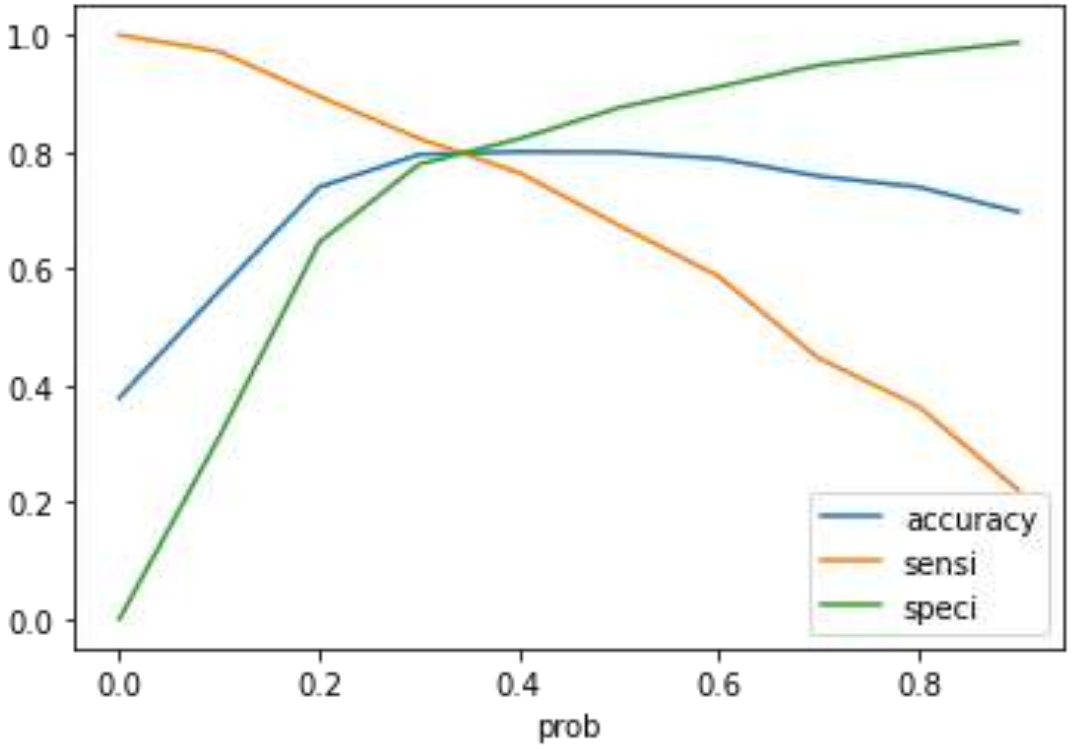
	Features	VIF
3	Lead Source_Olark Chat	1.64
6	Last Activity_Olark Chat Conversation	1.54
9	Last Notable Activity_Modified	1.40
1	Total Time Spent on Website	1.26
10	Last Notable Activity_SMS Sent	1.17
4	Lead Source_Reference	1.11
0	Do Not Email	1.10
5	Lead Source_Welingak Website	1.04
7	Last Notable Activity_Email Link Clicked	1.02
2	Lead Origin_Lead Import	1.00
8	Last Notable Activity_Had a Phone Conversation	1.00
11	Last Notable Activity_Unreachable	1.00

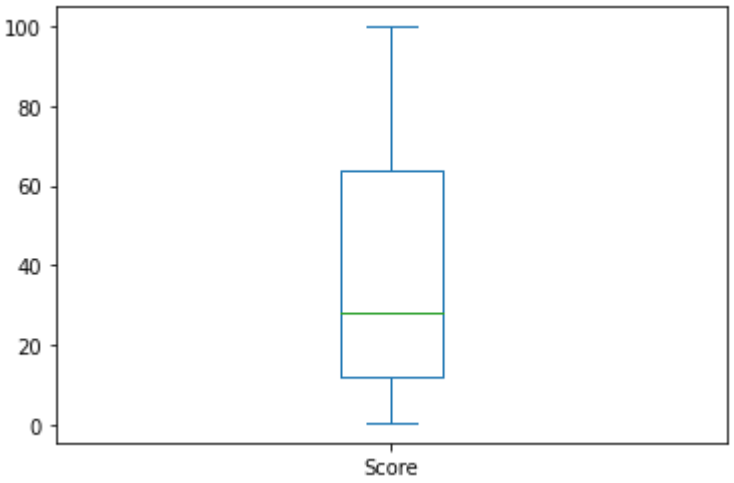


- In our final model we could see all the variables have a very low VIF value (i.e.<5) as per the industrial norms, is good for a model.
- The ROC Curve value, which stands at 0.87, is close to 1 and the curve approaching to top left corner of the garph indicating a strong predictive model performance.

Optimal cut-off and Accuracy Metrics of the model

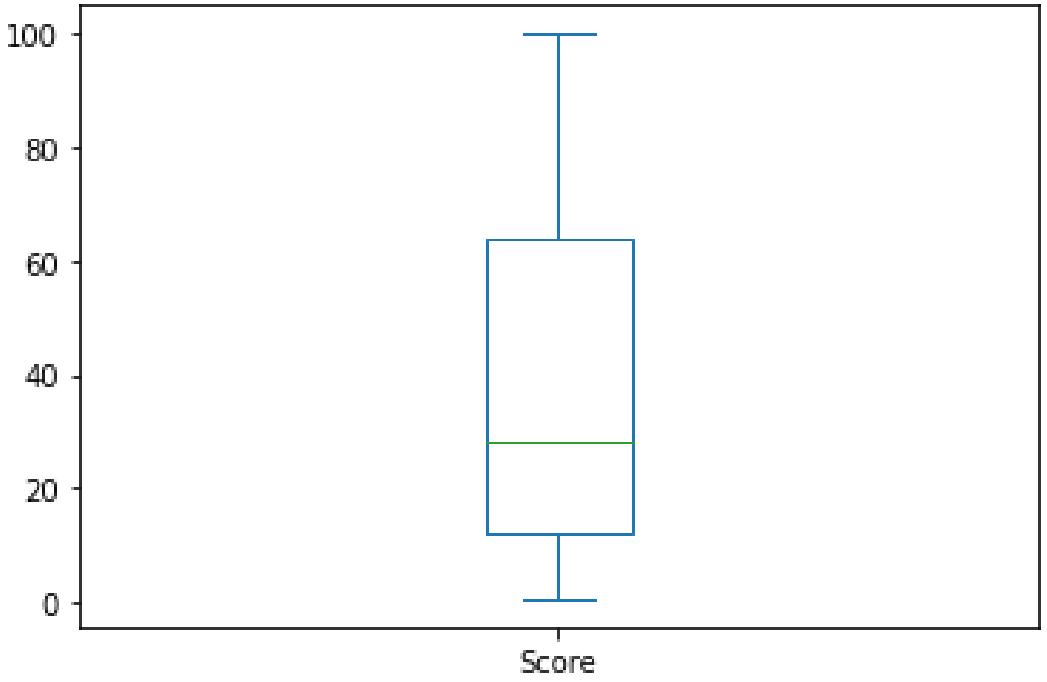
- In the plot having ***accuracy***, ***sensitivity*** and ***specificity*** plotted for various probabilities, we could see they coincide approximately at probability of 0.3, which shall be our optimal cut-off to proceed with.
- On applying the optimal cut-off, our result metrics on train dataset looks like:
  Accuracy: 79.52%
  Sensitivity: 82.27%
  Specificity: 77.85%
  Precision Score: 69.36%
  Recall Score: 82.27%
- After converting the probabilities to the score, we can see the score ranges between 0 to 100 in train data, (as discussed in the problem statement)





Performance on Test Data

- On getting the predictions on test data and applying the optimal cut-off, our result metrics on test dataset looks like:
  - Accuracy: 77.17%
  - Sensitivity: 82.21%
  - Specificity: 74.11%
  - Precision Score: 65.89%
  - Recall Score: 82.21%
- After converting the probabilities to the score, we can see the score ranges between 0 to 100 in test data, (as discussed in the problem statement)



Conclusion

- **We could see our model has an accuracy of 79% and 77% on train and test data respectively. With 82% of sensitivity on both test and train data (which show's model's good ability to predict conversion of lead, as sensitivity is the measure of power to predict true positive which was asked by CEO to get it around 80%).**
- **A total of 12 features are able to define the maximum variance in the conversion rate. In these 12 features, dummy variables from "*Lead Source*" and *"Lead Notable Activity"* have greater weightage. Other features like "*Do Not Email*" and *"Total Time Spent on Website"* are significant enough in scoring the leads.**
- **Also the score assigned to the Leads fall under the range i.e. 0 to 100, in both train and test dataset**

Thank You!!!!