Lead Score Case Study

Manidhar Nerella

Rashi Jain

Problem Statement

- An education company named X Education sells online courses to industry professionals
- 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.

Business Objective:

- X education wants to find most promising leads.
- Company have to build a Model which identifies the hot leads.

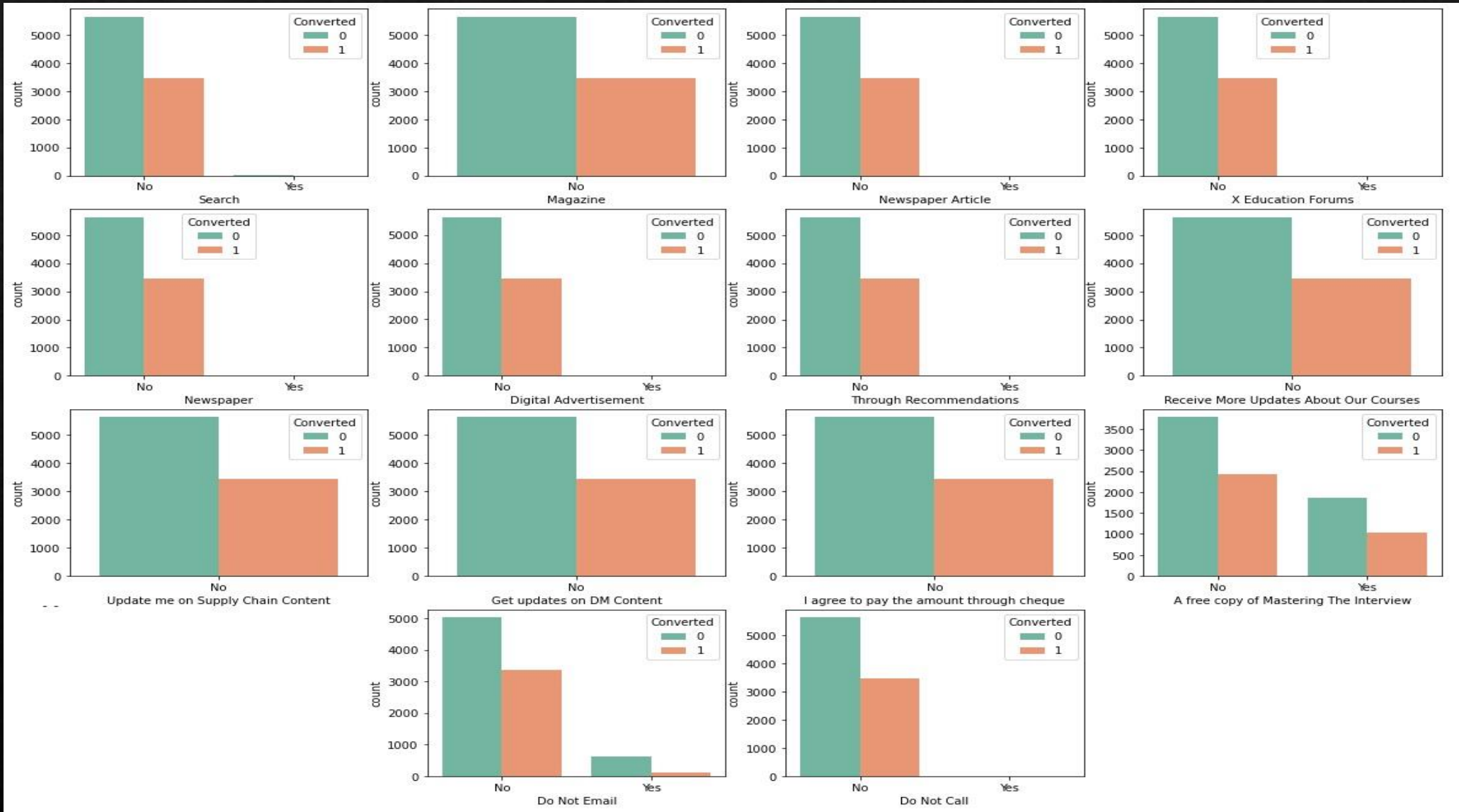
Solution Methodology

- Data cleaning and data manipulation.
  - Handle duplicate data.
  - Handle NaN values and missing values.
  - Imputation of missing values, if necessary.
  - Handle outliers in data.
- Exploratory Data Analysis
- Feature Scaling & Dummy Variables Creation
- Classification technique: Logistic Regression.
- Validation of the model.
- Conclusions and recommendations.

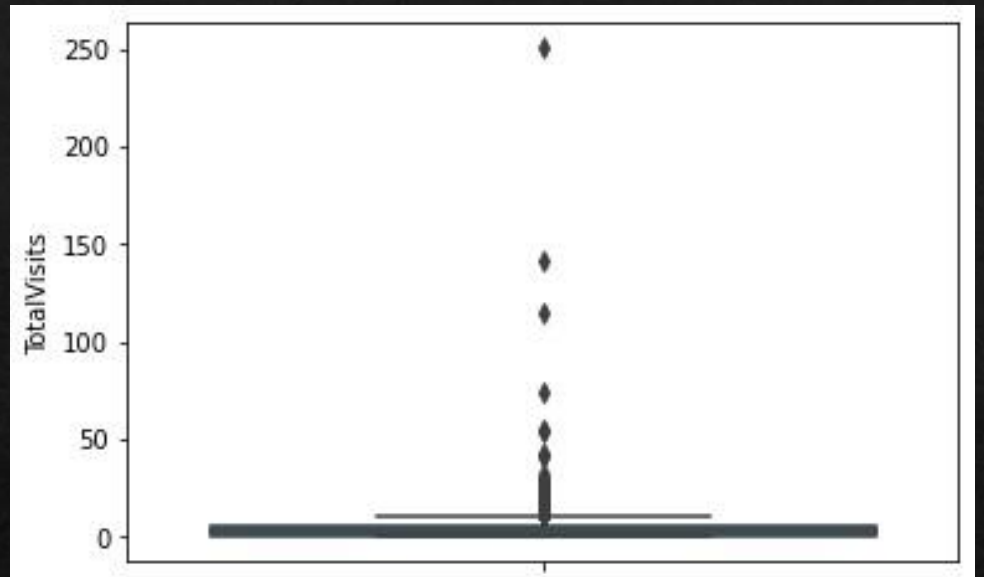
Data Cleaning

- Total Number of Rows =37 and Number of Columns =9240
- we have removing those columns where the percentage of missing data is more than 35%.
- Impute NA for other columns as “Not Provided” as removing all those values will affect model building.

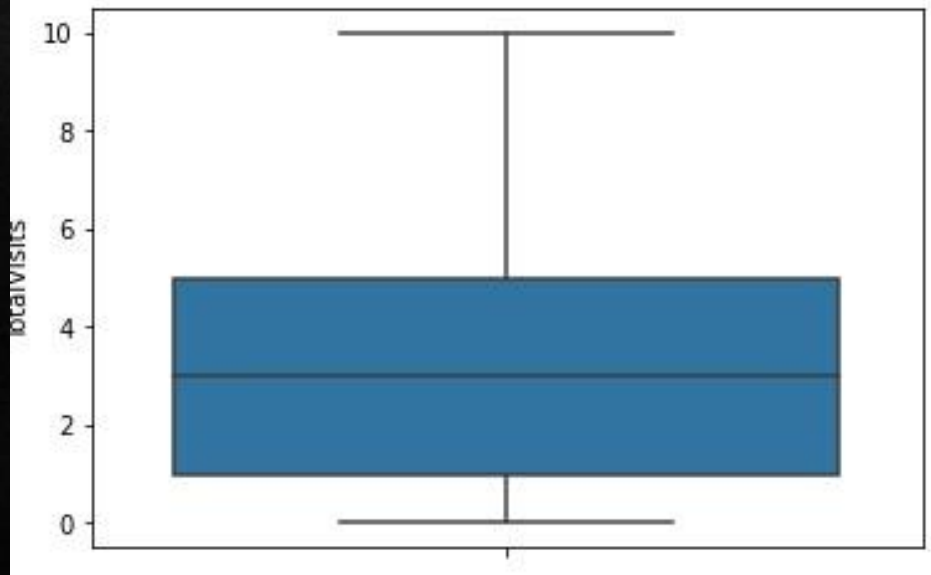
EDA (Categorical Variable)

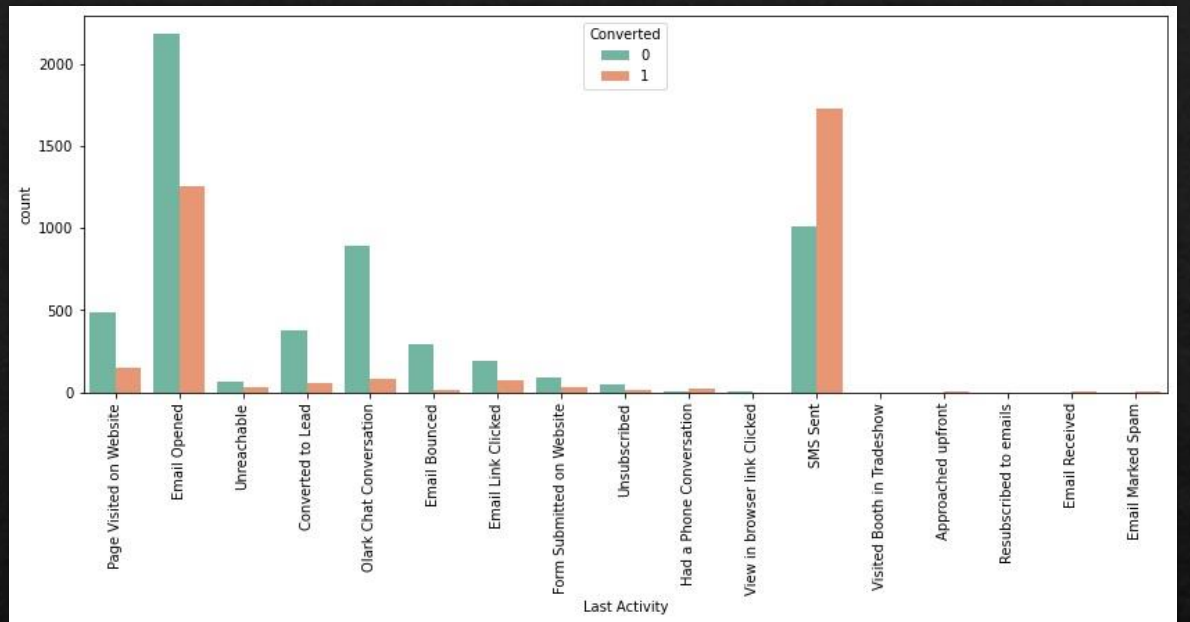


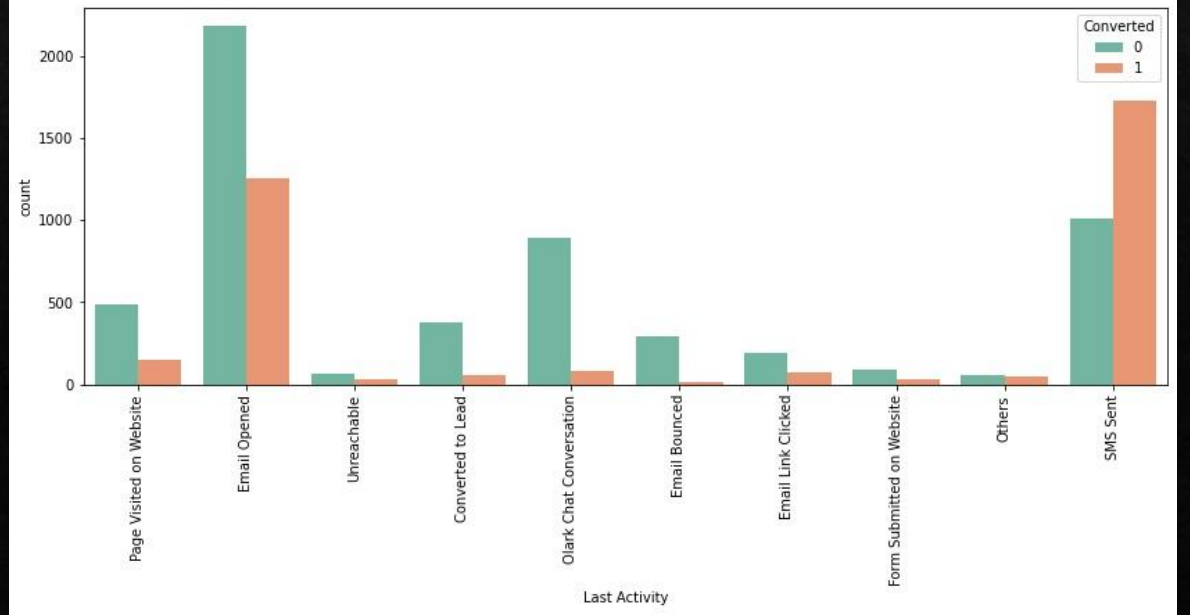
EDA (Numeric Variable)



After Handling outliiers







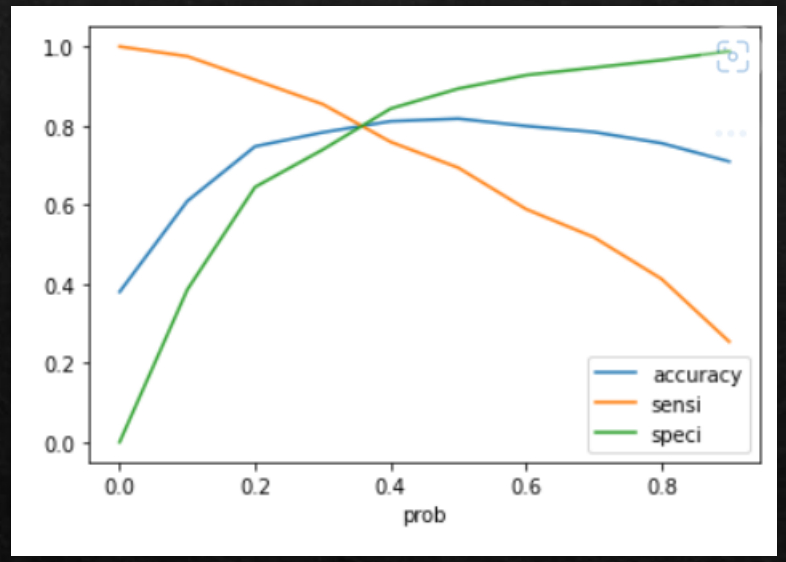
Data Conversion

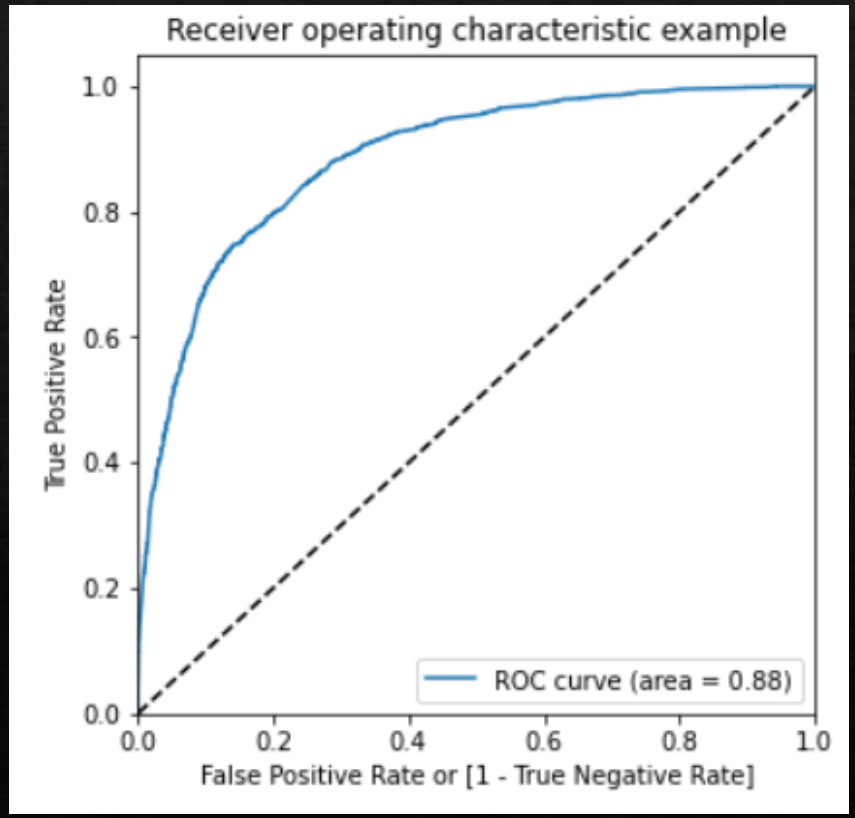
- In this step we transform the data into such a way that it is useful for future analysis. In the given case where there were only Yes or No in a particular column, we have replaced it with 1's and 0's. Whereas when there are more than 2 options we have added dummy columns to them for further analysis.
- Final number of columns: 51

Model Building

- Splitting the Data into Training and Testing Sets. We have chosen 70:30 ratio.
- Use RFE for Feature Selection
- Building Model by removing the variable whose p- value is greater than 0.05 and vif value is greater than 5
- Make predictions on test data set
- Overall accuracy 81.67%

ROC Curve





Optimal Cutoff

Probability where we get balanced sensitivity and specificity that is 0.3

Final Prediction

- The final prediction percentage in the given case is around 85% which is more than the 80% level desired by the CEO. Hence the model is working properly and is time to test the model with the test data.
- Following are the final results after the cut off point has been fixed at 0.3
  - Accuracy - 78.29%
  - Sensitivity - 85.41%
  - Specificity - 73.94%
- Finally, we test the data using the test data and following are the results derived after using the test data
  - Accuracy - 79.38%
  - Sensitivity - 86.08%
  - Specificity - 75.25%

Conclusion

- While we have checked both Sensitivity-Specificity as well as Precision and Recall Metrics, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction.
- Accuracy, Sensitivity and Specificity values of test set are around 79%, 86% and 75% which are approximately closer to the respective values calculated using trained set.
- Also the lead score calculated in the trained set of data shows the conversion rate on the final predicted model is more than 80%.
- Hence overall this model seems to be good.

Recommendations

- Following variable affect the the buyer the most in descending order and can be used for further analysis:

1. The total time spend on the Website.
2. Total number of visits.
3. When the lead source was:
4. When the last activity was:
5. When the lead origin is Lead add format.
6. When their current occupation is as a working professional.