LEAD SCORING CASE STUDY

Akanksha Dwivedi

PROBLEM STATEMENT

- 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 know most promising leads.
- For that they want to build a Model which identifies the hot leads.
- Deployment of the model for the future use.

SOLUTION METHODOLOGY

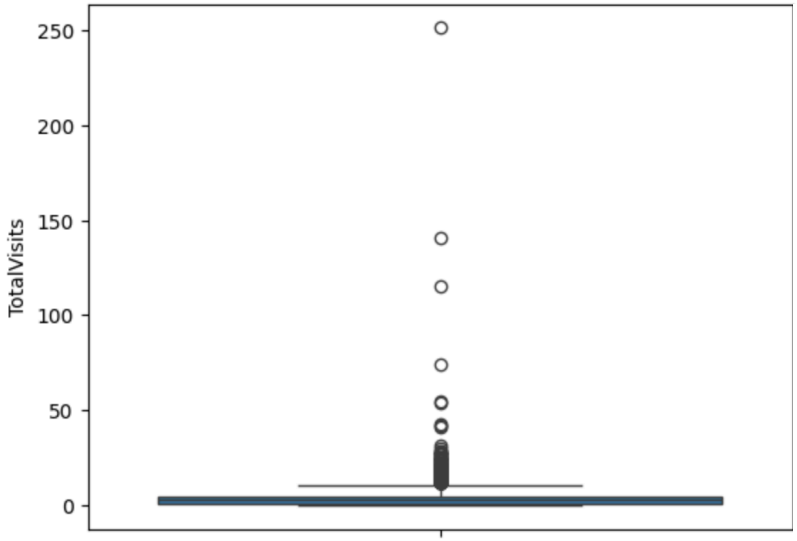
- Import Python Packages
- Reading and Understanding the dataset
- Data Cleaning
- Exploratory Data Analysis
- Data Preparation
- Train-Test Split
- Feature Scaling
- Model Building
- Model Evaluation
- Making Predictions on test dataset
- Model Parameters
- Conclusion
- Recommendations

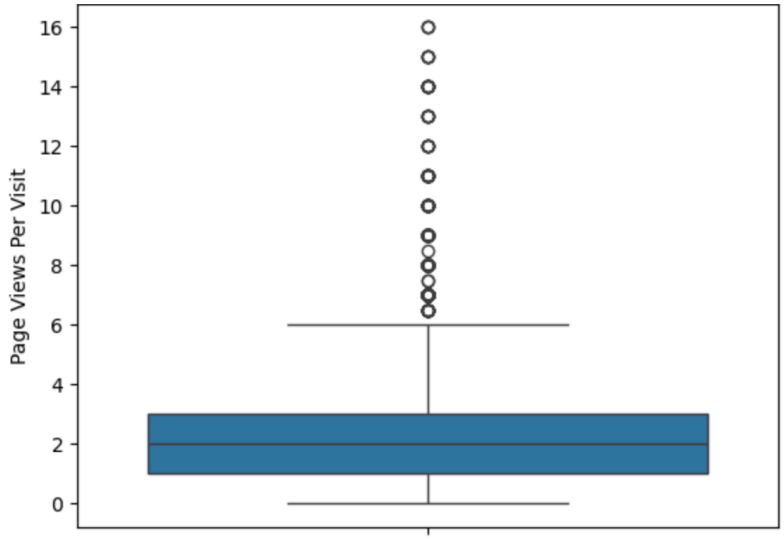
DATA CLEANING

- Delete redundant columns i.e 'Prospect ID' and 'Lead Number'
- Delete the columns with unique values
- Treatment for ‘Select’ values
  - Many of the categorical variables have a level called 'Select' which needs to be handled. The columns having ‘Select’ were imputed with NaN
- Treatment for null values
  - Delete the columns where null value percentage is > 40%
  - Impute the null values for the columns where null value percentage > 15% with mode
  - Treat columns where null value percentage ~1%

DATA CLEANING

- Outlier Treatment
  - There were outliers for column 'TotalVisits' and 'Page Views Per Visit’
  - The outliers are present only in the upper range so the outliers were treated by capping the upper range to 99%

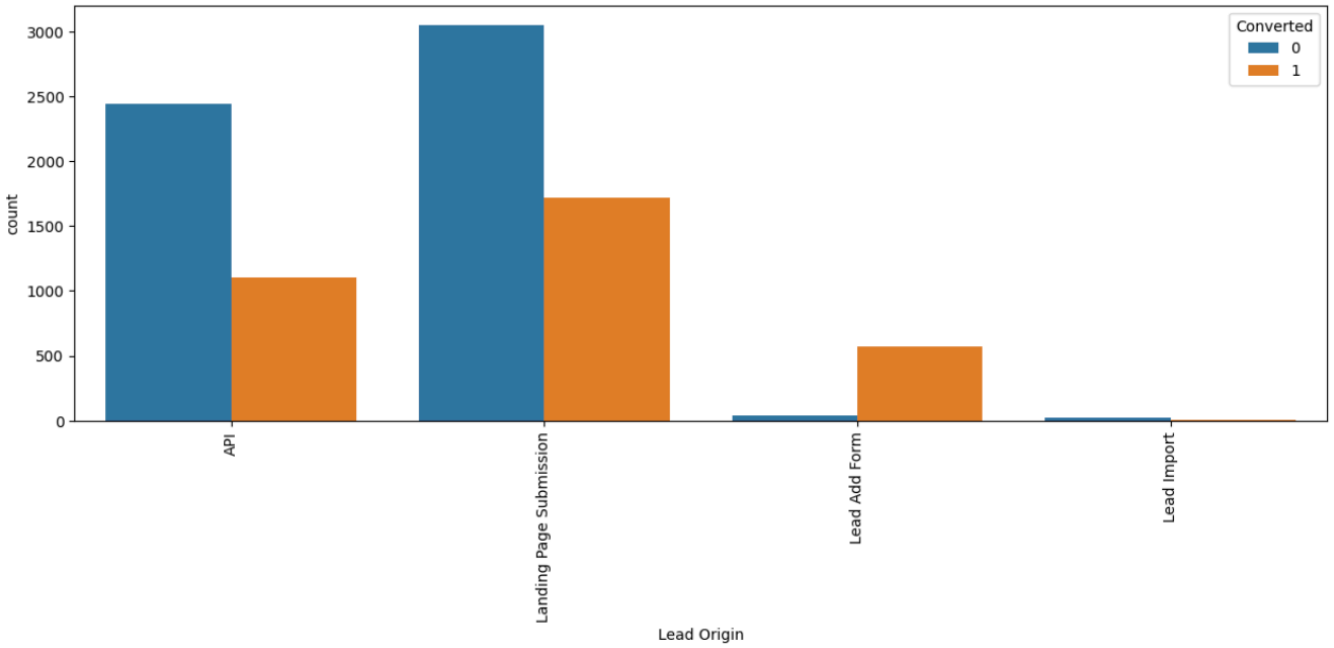




UNIVARIATE ANALYSIS

LEAD ORIGIN

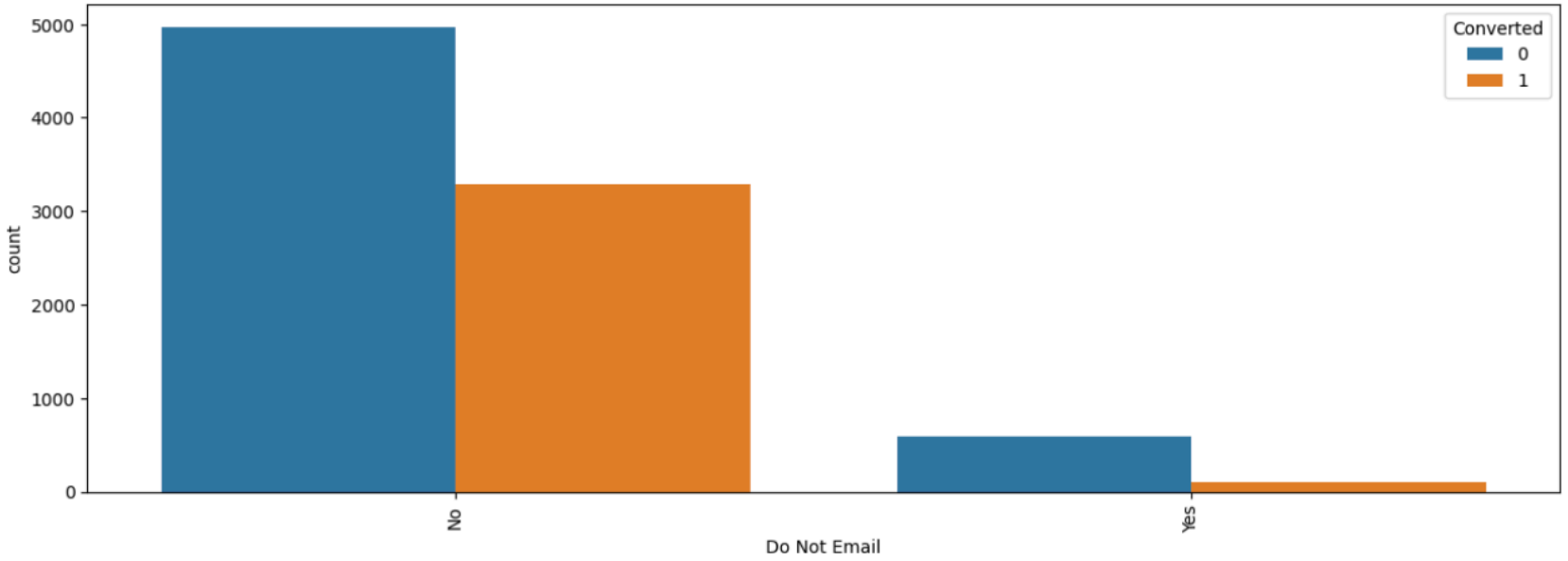
- 'API' and 'Landing Page Submission' have both high number or leads and conversion rate.
- 'Lead Add Form' has a very high conversion rate as compared to the number of leads they have.
- 'Leads Import' have very few leads



UNIVARIATE ANALYSIS

DO NOT EMAIL

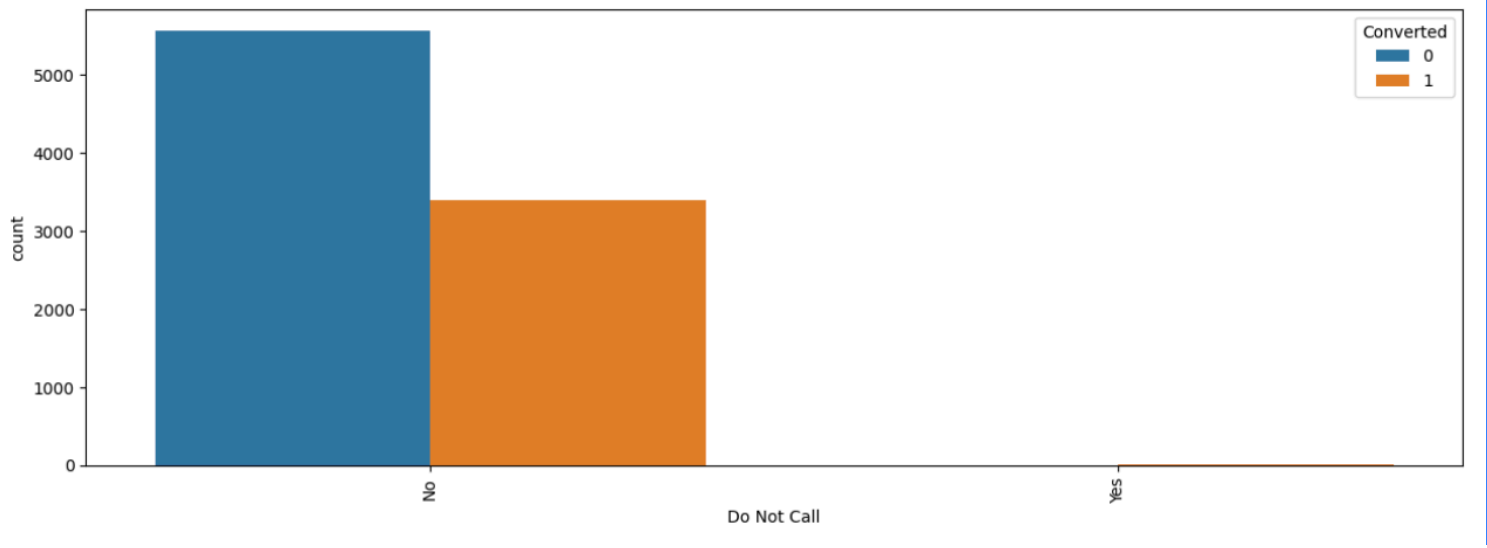
- 92% of the people don’t want to be emailed about the course.



UNIVARIATE ANALYSIS

DO NOT CALL

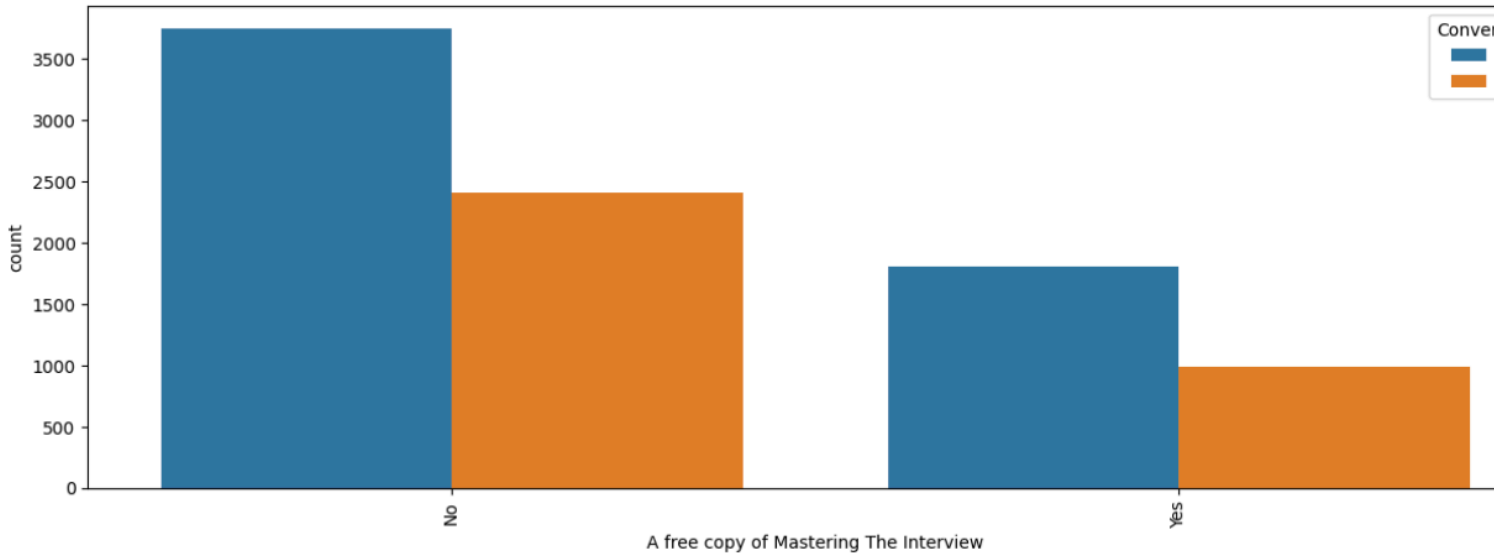
- We can see from the value count data that 'No' is having more than 99.9% of the data. We can safely drop this column, as this will not add much to the analysis.



UNIVARIATE ANALYSIS

A FREE COPY OF MASTERING THE INTERVIEW

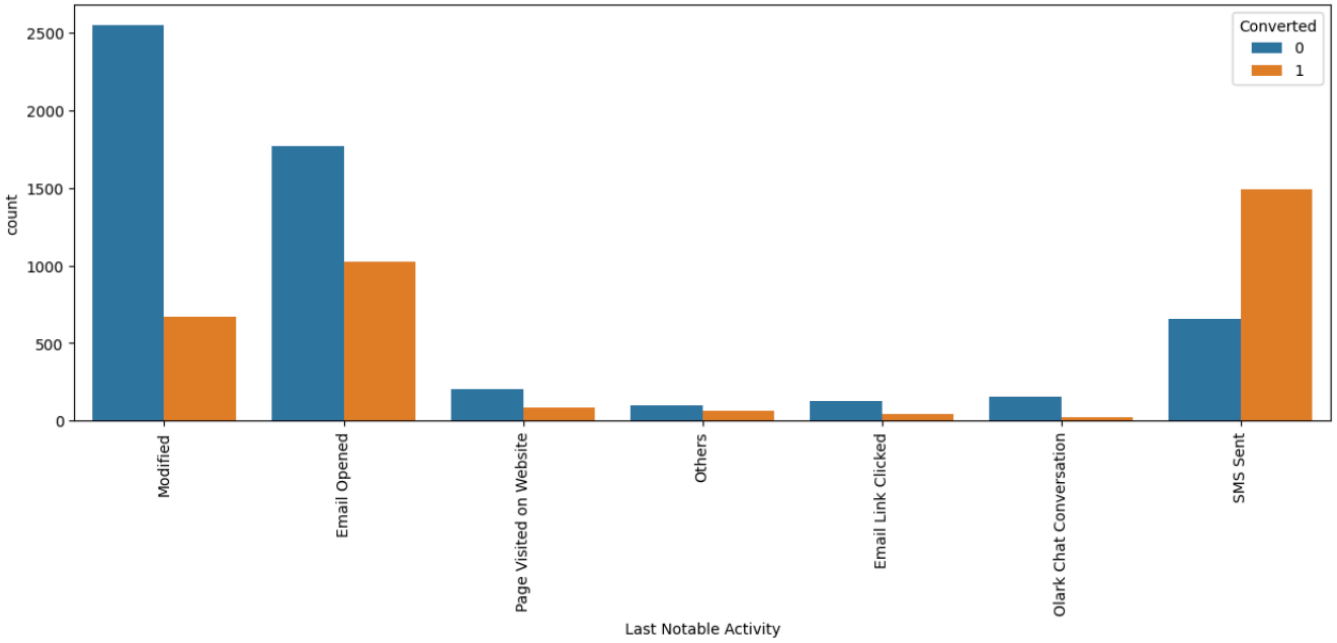
- The number of leads who do not ask for the free copy are high. This group can be focused for conversion.



UNIVARIATE ANALYSIS

LAST NOTABLE ACTIVITY

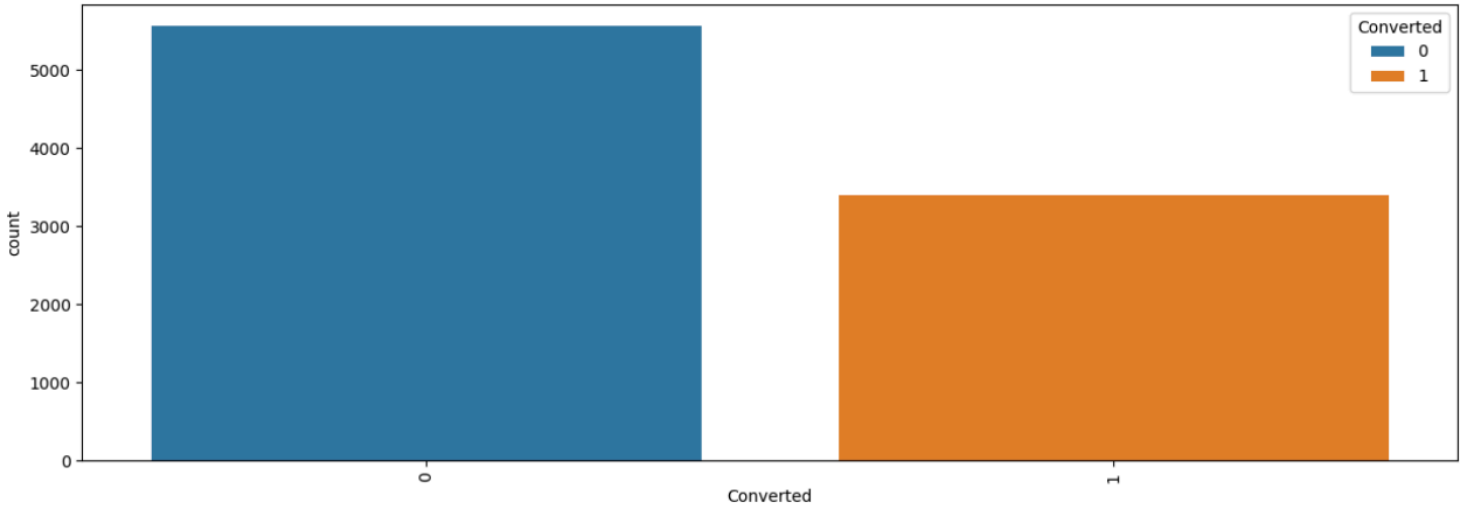
- ‘Modified' and 'Email Opened' have high number of leads. This section can be targeted to increase the conversion rate.
- SMS sent have high conversion rate.



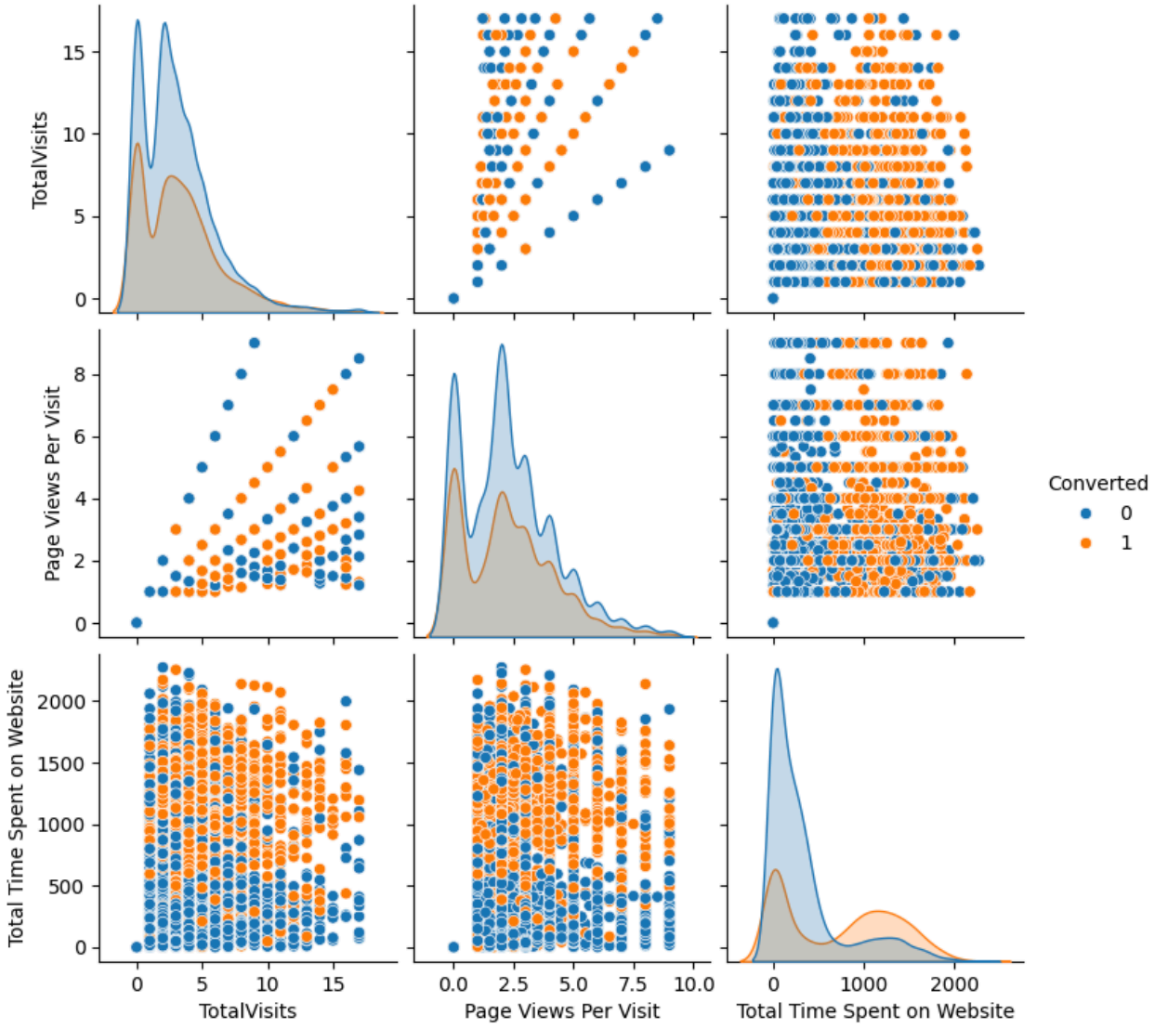
EXPLORATORY DATA ANALYSIS

DATA IMBALANCE RATIO

- From value count and count plot we can see that data is properly balanced.
- Imbalance ratio: 61.08%
- **Conversion rate is 37.92%**, meaning only 37.92% of th people have converted to leads.



BIVARIATE ANALYSIS



CORRELATION MATRIX

- 'TotalVisits' and 'Page Views Per Visit' have a high correlation index of 0.72
- 'Total Time Spent on Website' has a correlation index of 0.35 with target variable 'Converted'
- All other numerical columns have low correlation with target variable 'Converted'

DATA PREPARATION

- Converting Binary Variables to 0/1 for 'Do Not Email' and 'A free copy of Mastering The Interview' features
- Created Dummy features for categorical variables

FEATURE SCALING

- Scaled the numerical features using StandardScaler
- Insights after scaling the features:
  - 'Lead Source_Facebook' and 'Lead Origin_Lead Import' have a high correlation of 0.98
  - 'Lead Origin_Lead Add Form' and 'Lead Source_Referance' have a high correlation of 0.85
  - 'TotalVisits' and 'Page Views Per Visit' have a correlation of 0.72
  - 'Lead Origin_Lead Add Form' ,'Lead Source_Welingak Website', 'Last Activity_SMS Sent' and 'What is your current Occupation_Working Professionals' have positive correlation with the target variable 'Converted'

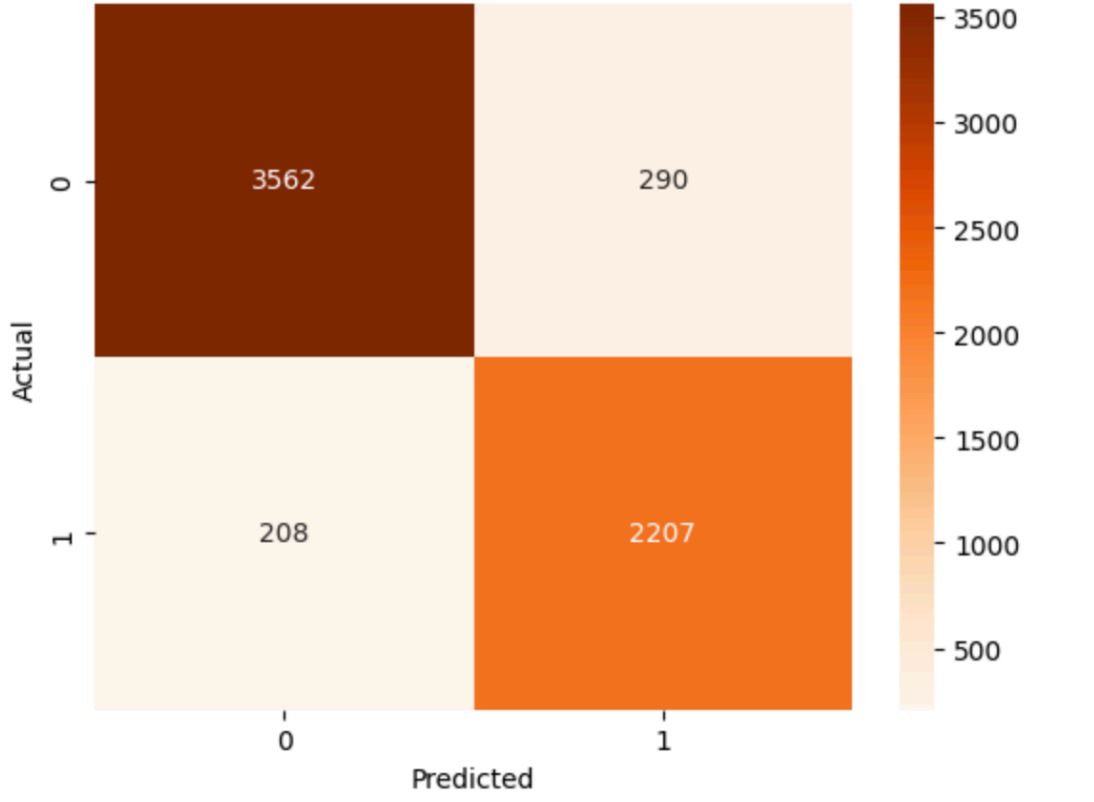
MODEL BUILDING

- Splitting the data into training and test dataset
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio
- Used Recursive Feature Elimination(RFE) to reduce features from 49 to 15
- Manual Feature Elimination process was used to build models by dropping variables with p-value > 0.05 and VIF > 5
- Total 2 models were built before reaching final Model 3 which was stable with (p-values < 0.05). No sign of multicollinearity with VIF < 5
- The final model had 13 variables, we used it for making predictions on the train and test set
- Overall Accuracy is ~91%

CONFUSION MATRIX

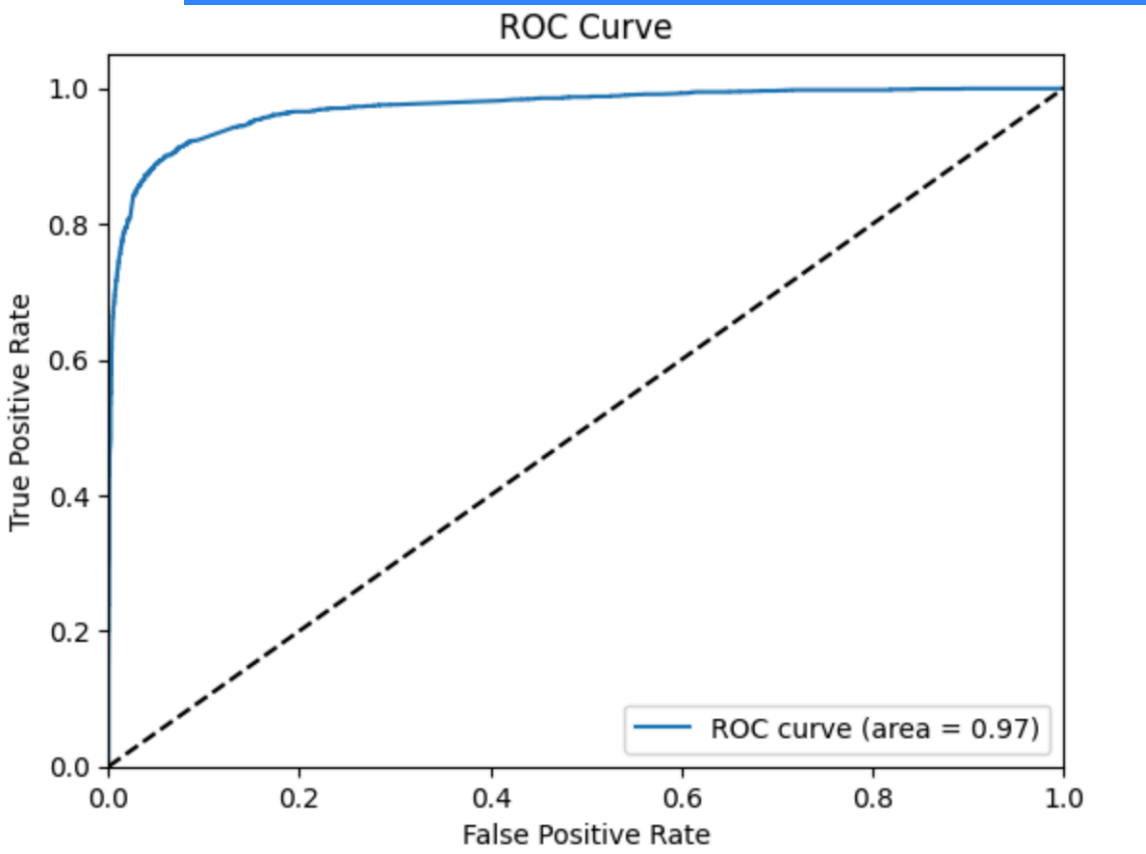
Metrics for the Train Data are below:

- Accuracy : 91.99%
- Sensitivity : 91.35%
- Specificity : 92.39%
- Precision : 88.28%
- Recall : 91.35%



ROC CURVE

- ROC curve shows the tradeoff between sensitivity and specificity.
- The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
- We are getting a good value of 0.97 indicating a good predictive model.

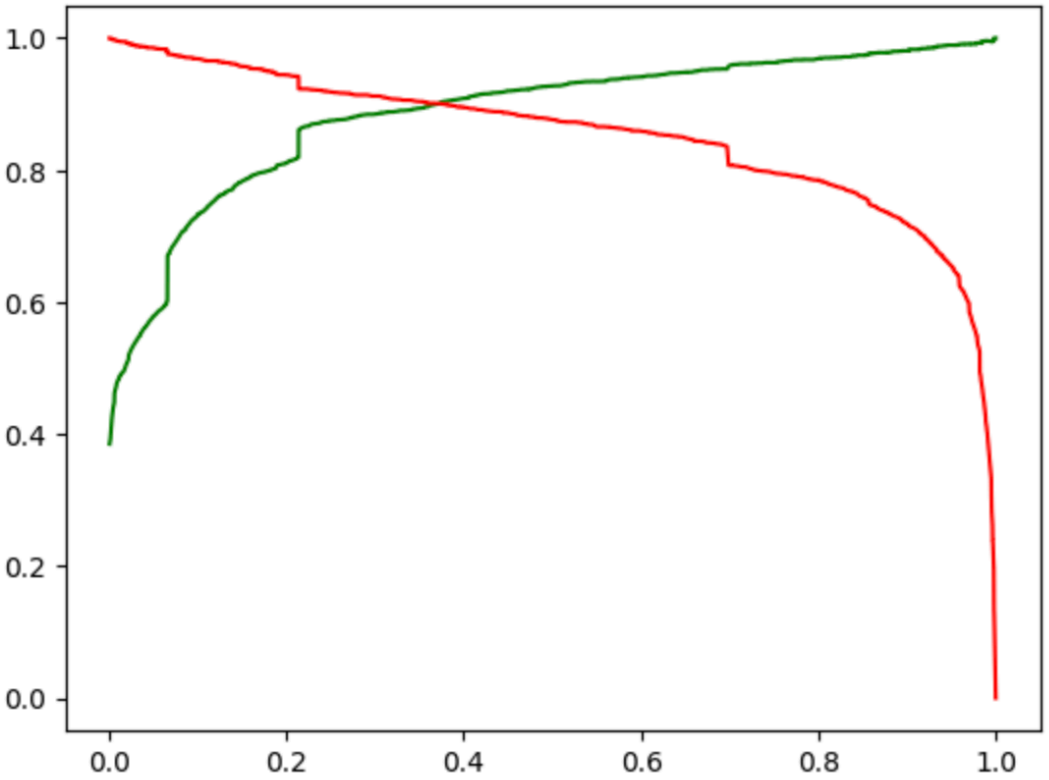


PRECISION RECALL TRADEOFF CURVE

From the Precision - Recall trade off curve cut off point(0.385) we see that:

- True Positive numbers have decreased and True Negative numbers have increased
- Sensitivity/Recall has decreased.

Thus, we cannot use Precision-Recall trade-off method. Thus we will use 0.286 as optimal cutoff point.



MAKING PREDICTIONS ON TEST DATASET

- Before predicting on the test set, we need to standardize the test set and need to have exact same columns present in our final train dataset
- After doing the above step, we started predicting the test set, and the new prediction values were saved in a new data frame.
- After this we did model evaluation i.e. finding the accuracy, precision, and recall.
- Evaluation metrics for train and test are very close to around 91%
- The sensitivity value for test data is 92.86% while for train data is also 91.35%.
- The accuracy value is 92.70%. This indicates that the model is performing well for the test data set also
- Lead score is created on test dataset to identify hot leads – high the lead score higher the chance of conversion, low the lead score lower the chance of getting converted.

MODEL PARAMETERS

- The final logistic regression model has **13 features**
- **Equation of line:**
  - Converted = -0.765079 - 0.367659 X Lead Origin_Landing Page Submission + 0.403844 X Lead Origin_Lead Add Form + 0.274769 X Lead Source_Olark Chat + 0.468541 X Lead Source_Welingak Website + 0.547402 X Last Activity_SMS Sent + 1.291309 X Tags_Closed by Horizzon + 0.840598 X Tags_Lost to EINS - 1.010874 X Tags_Others - 1.170860 X Tags_Ringing + 1.851406 X Tags_Will revert after reading the email + 0.626698 X Last Notable Activity_Email Opened + 0.978982 X Last Notable Activity_SMS Sent

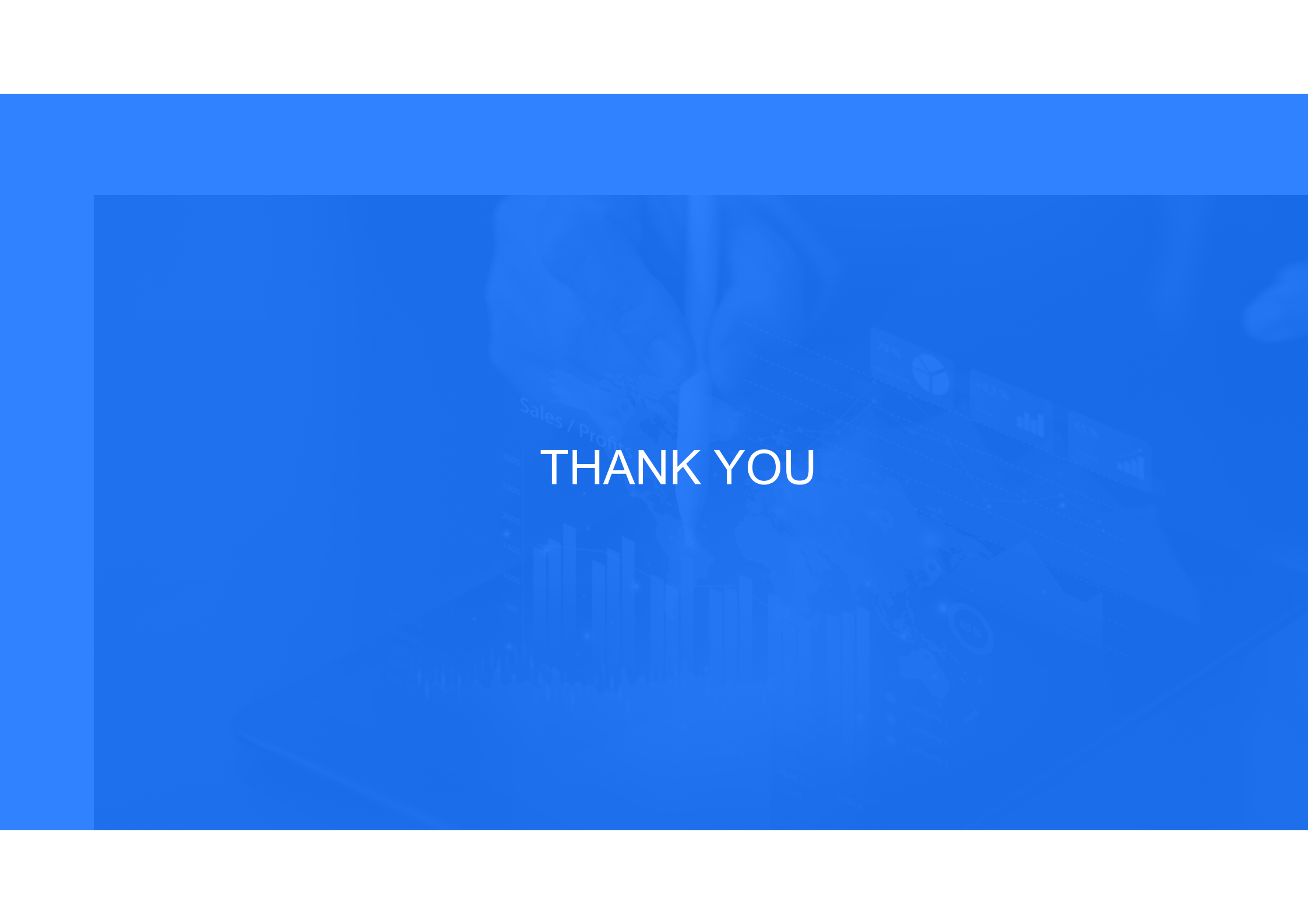
CONCLUSION

- The model achieved a sensitivity of 91.35% in the train set and 92.86% in the test set, using a cut-off value of 0.286
- The model also achieved an accuracy of ~91%
- The Optimal cutoff probability point is 0.286. Converted probability greater than 0.286 will be predicted as Converted lead and probability smaller than 0.286 will be predicted as not Converted lead
- **Top three features** that contribute positively to predict hot leads are:
  - Tags_Will revert after reading the email
  - Total Time Spent on Website
  - Last Notable Activity_SMS Sent

RECOMMENDATIONS

To improve the potential lead conversion rate X-Education should focus on the top important features:

- **Tags_Will revert after reading the email:** As the leads with tags, will revert after reading the email is high, so the company should focus more on email marketing
- **Total Time Spent on Website:** Leads spending more time on the website can be our potential lead
- **Last Notable Activity_SMS Sent** Lead whose last activity is sms sent can be potential lead for the company
- **Tags_Closed by Horizzon:** Tags closed by Horizzon have a good conversion rate
- Focus on features with **positive coefficients**
- **Working professionals** to be targeted as they have a high conversion rate
- Develop strategies to attract high-quality leads from **performing lead sources**


THANK YOU