Lead Scoring Case Study

Prathyusha A, Kolahal Anil Kumar
and Manish Gehani

Business Problem

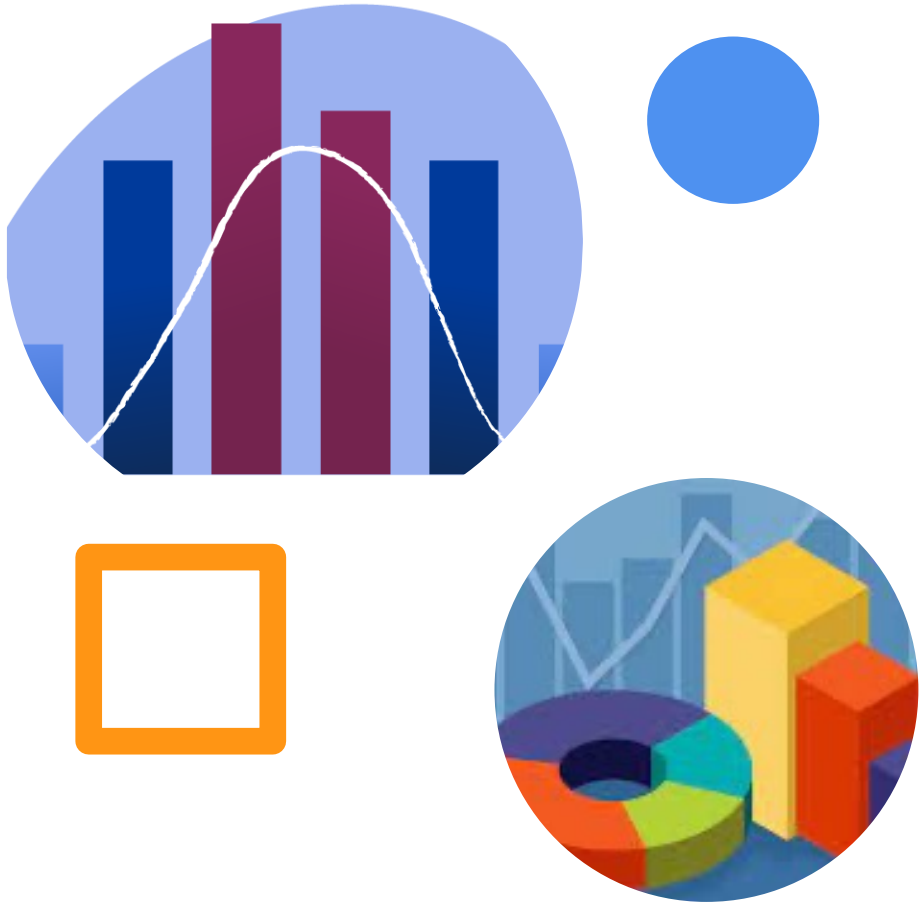
•X Education wants to improve their lead conversion rate by identifying 'hot leads'

•The company aims to develop a lead scoring model to predict the conversion chance of each lead

•The model will be based on historical data and features of the leads, such as their source and profile

•The goal is to provide the sales team with a prioritized list of leads most likely to convert into paying customers

•This will enable the sales team to focus their efforts on the 'hot leads' and improve the lead conversion rate.



Business Objective

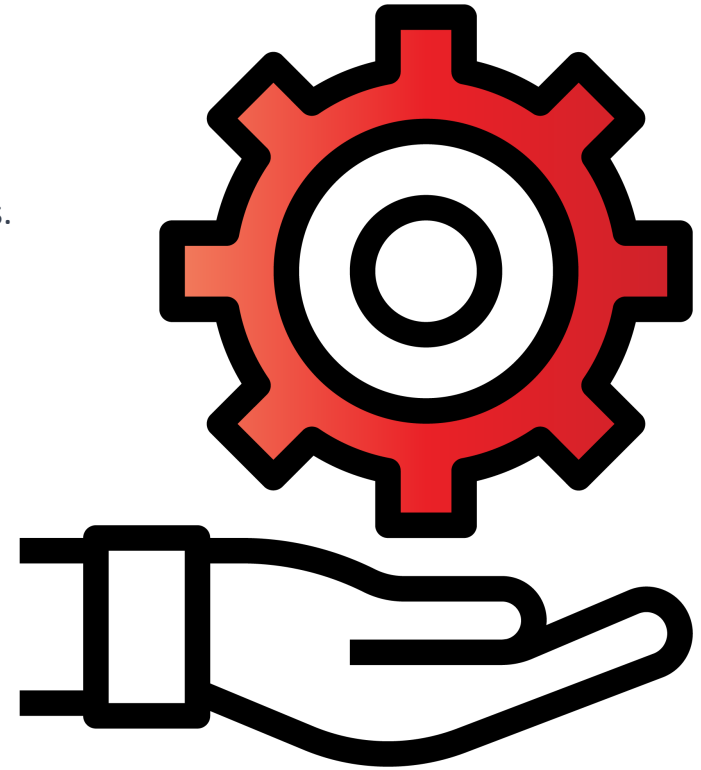
- The Business Objective Is To Build A Logistic Regression Model To Identify The Hot/Potential Leads And Achieve The Lead Conversion Rate To 80%.”

Understanding Dataset

- The objective is to predict the conversion chance of leads and develop a lead scoring model to improve the lead conversion rate.
- The dataset contains around 9000 data points with various attributes like Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc.
- The target variable is 'Converted' with 1 indicating converted and 0 indicating not converted.
- The data dictionary provided in the zip folder provides more information about the dataset.
- The categorical variables have a level called 'Select' which needs to be handled since it is as good as a null value.

Methodology

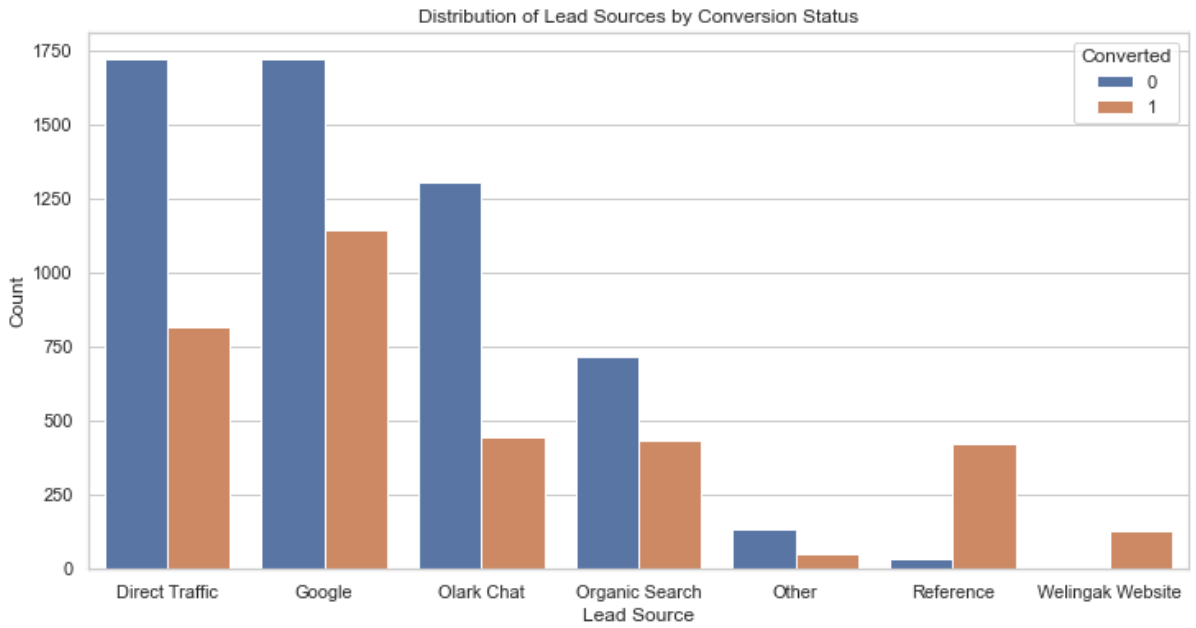
- Data cleaning and manipulation:
  - Identify and handle duplicate data.
  - Identify and handle missing values and NA values.
  - Remove columns with a large number of missing values or that are not useful for analysis.
  - Impute missing values if necessary.
  - Identify and handle outliers in data.
- Exploratory Data Analysis (EDA):
  - Analyze univariate data
  - Analyze bivariate data
- Feature Scaling & Encoding
- Modeling: Use logistic regression for model building and prediction.
- Validate the model.
- Present the model.
- Draw conclusions and provide recommendations.

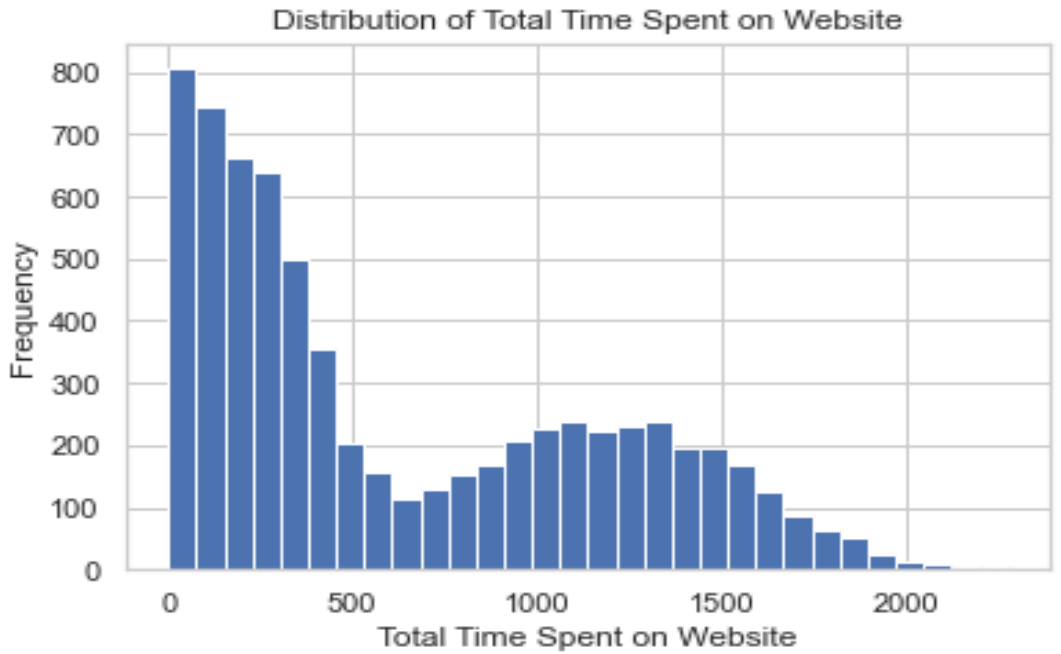


Data Manipulation

- The dataset has 37 rows and 9240 columns.
- Features with single values like "Magazine", "Receive More Updates About Our Courses", "Update me on Supply Chain Content", "Get updates on DM Content", and "I agree to pay the amount through cheque" were dropped.
- "Prospect ID" and "Lead Number" were removed as they are not necessary for analysis.
- Features with low variance were dropped based on value counts for some object type variables, including "Do Not Call", "Search", "Newspaper Article", "X Education Forums", "Newspaper", and "Digital Advertisement".
- Columns with more than 35% missing values, such as "How did you hear about X Education" and "Lead Profile", were dropped.

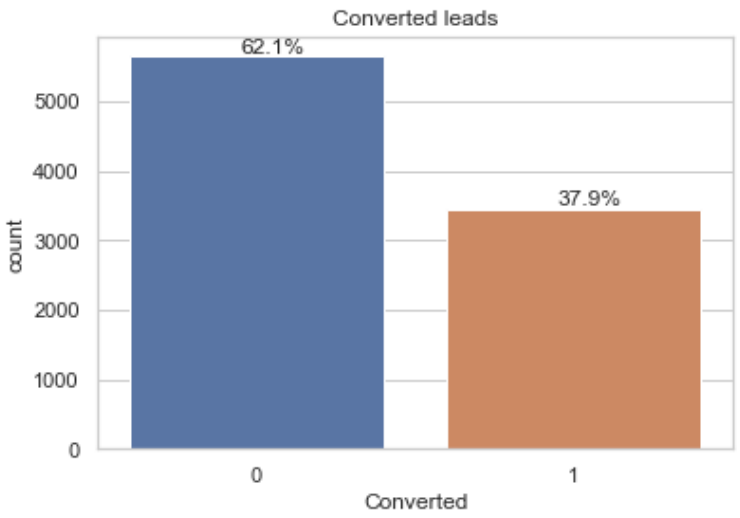
EDA



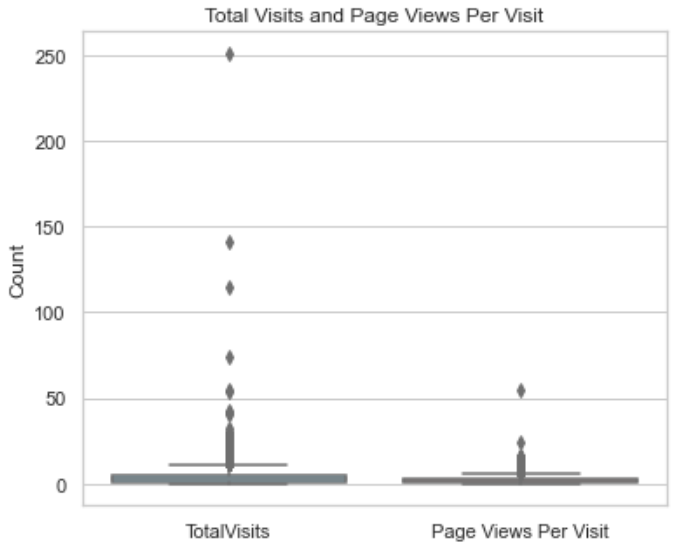


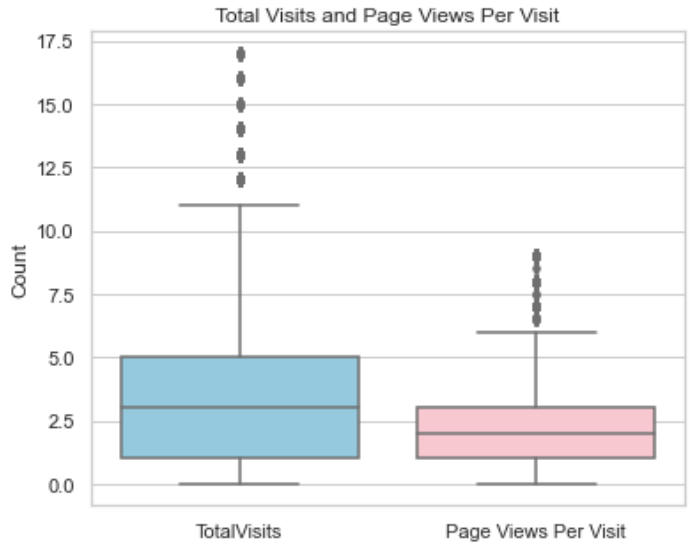
- References and Welingak Website have the highest conversion
- Olark Chat and Direct Traffic have a high number of leads but poor conversion
- Focus should be on increasing the conversion on the aforementioned websites
- Also, leads to Welingak Website should be increased
- The average time spent on the website is ~8 minutes
- It is not surprising that higher the time spent, better is the conversion. If an individual spends over 10 minutes, he is more that 50% likely to join

EDA: Categorical data



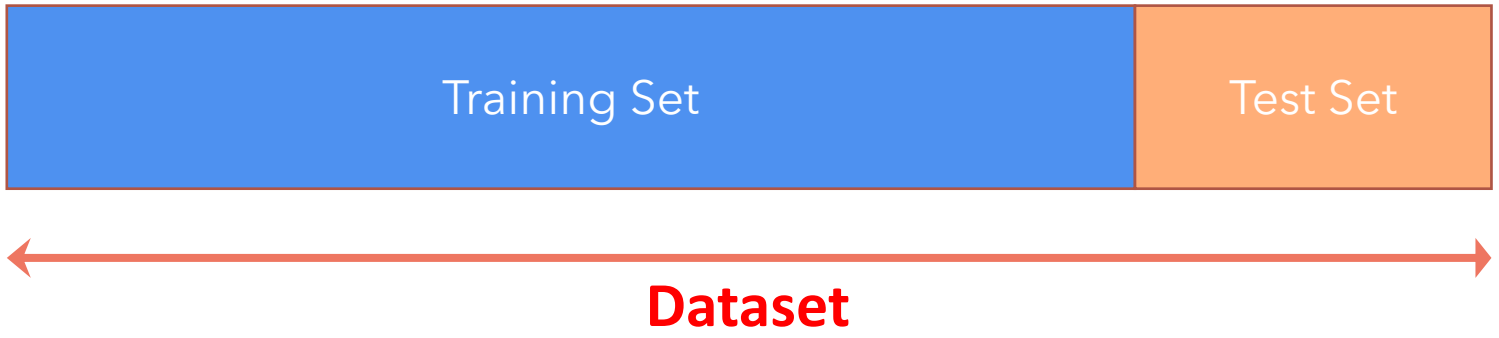
Almost two-third leads were not converted

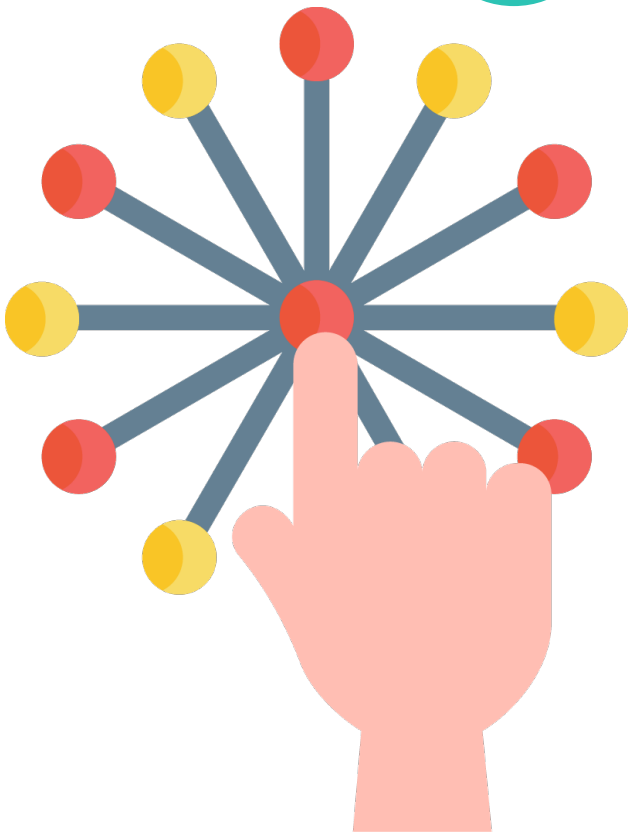




Treating Outliers

Model Building





- Use RFE for Feature Selection
- Running RFE with 15 variables as output
- Building Model by removing the variable whose p- value is greater than 0.05 and vif value is greater than 5
- Predictions on test data set

ROC Curve

•The ROC Curve is a graphical representation of the model's ability to distinguish between classes

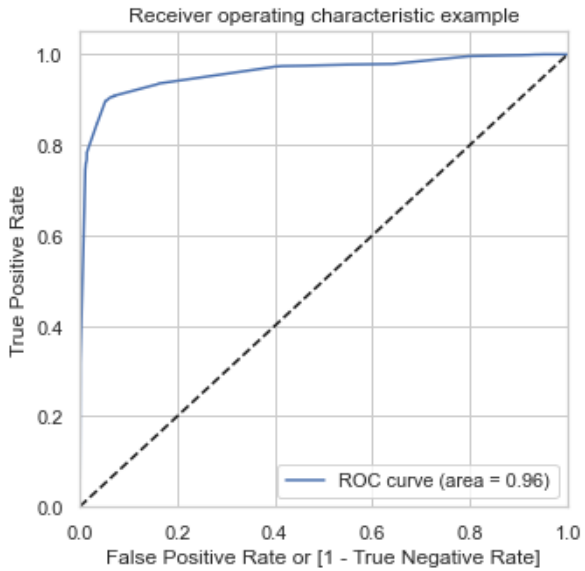
•The AUC (Area Under Curve) measures how accurately the model distinguishes between 1's and 0's

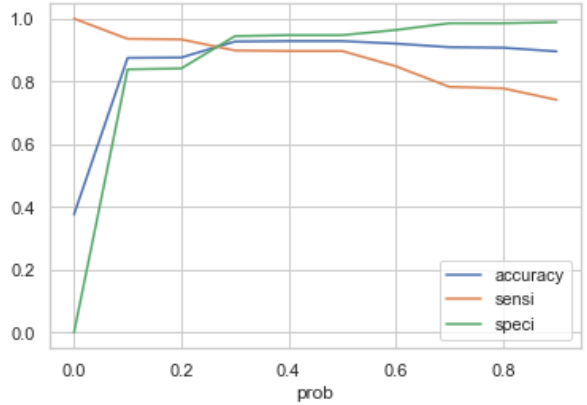
•The AUC for this model is approximately 0.96, indicating that the model accurately distinguishes between 1's and 0's about 96% of the time

•A model with an AUC of 0.96 is considered to be very stable

•The optimal cut-off point for the model is found to be at 0.28, which means that the sensitivity and specificity of the model are balanced at this point

•Any conversion probability greater than 28% is predicted as a lead using the X-Train with probability = 0.28





Assessing the Model's Performance

Train Set

- Accuracy: 93%
- Sensitivity: 90%
- Specificity: 94%

Test Set

- Accuracy: 93%
- Sensitivity: 90%
- Specificity: 95%

- After the model building process, the sensitivity value was found to be higher than the required 80%.
- When evaluating the model on the test set, the model evaluation parameters remained the same, indicating that the model is highly stable.

Conclusion

1.References and Welingak Website have the highest conversion ratio – leads need to be increased

2.Olark Chat and Direct Traffic have a high number of leads but poor conversion – follow-up is required in this area as there is a potential for conversion

3.The average time spent on the website is ~8 minutes which is slightly shy of the 10 minute mark beyond which leads have more than a 50% probability of being converted; perhaps the website needs to be a bit more engaging and containing more information

4.SMS' work better than other forms of communication / involvement - should look to target customers using this means

5.Target working professionals as they seem to be opting for it more than the others

6.Individuals who were not reachable were quite a big chunk so it might help if they are

7.approached via other means viz. SMS, email, etc. or an alternate contact number is sought

Thank you