Lead Scoring Case Study

Submitted by: Manojkumar Padmanaban
Manjunath
Manoj

Objective

- Tackle the problem of improving the lead conversion rate for education company named X Education sells online courses to industry professionals.
- In order to address the problem of low lead conversion rate for X Education, we proposed building a logistic regression model to assign lead scores to each lead.
- The lead score ranges from 0 to 100, with higher scores representing a higher likelihood of conversion (hot data) and lower scores representing a lower likelihood of conversion (cold leads).
- This approach helps the sales team prioritize data with a higher probability of conversion, ultimately increasing the overall conversion rate.

Data Preparation and Missing Values treatment

- Data cleaning: Handling missing values, removing duplicates, and fixing inconsistencies.
- Analyzed features and columns importance
- Checked the values and removed unnecessary columns based on use case understanding
- Data transformation: Encoding categorical variables, normalizing or scaling numerical variables.
- Investigated the dataset for missing values.
- Handled missing values by filling with mean/median/mode, using interpolation, or removing instances (based on the proportion of missing values and their impact on the analysis)
- Encoding Categorical Variables Explored the distribution of numerical variables.
- Applied normalization or standardization methods to scale numerical variables, ensuring they have comparable ranges and do not disproportionately impact the model.

Exploratory Data Analysis (EDA)

- Perform EDA to identify trends, patterns, and relationships between the variables in the dataset.
- Understand which features might be important for predicting lead conversion.

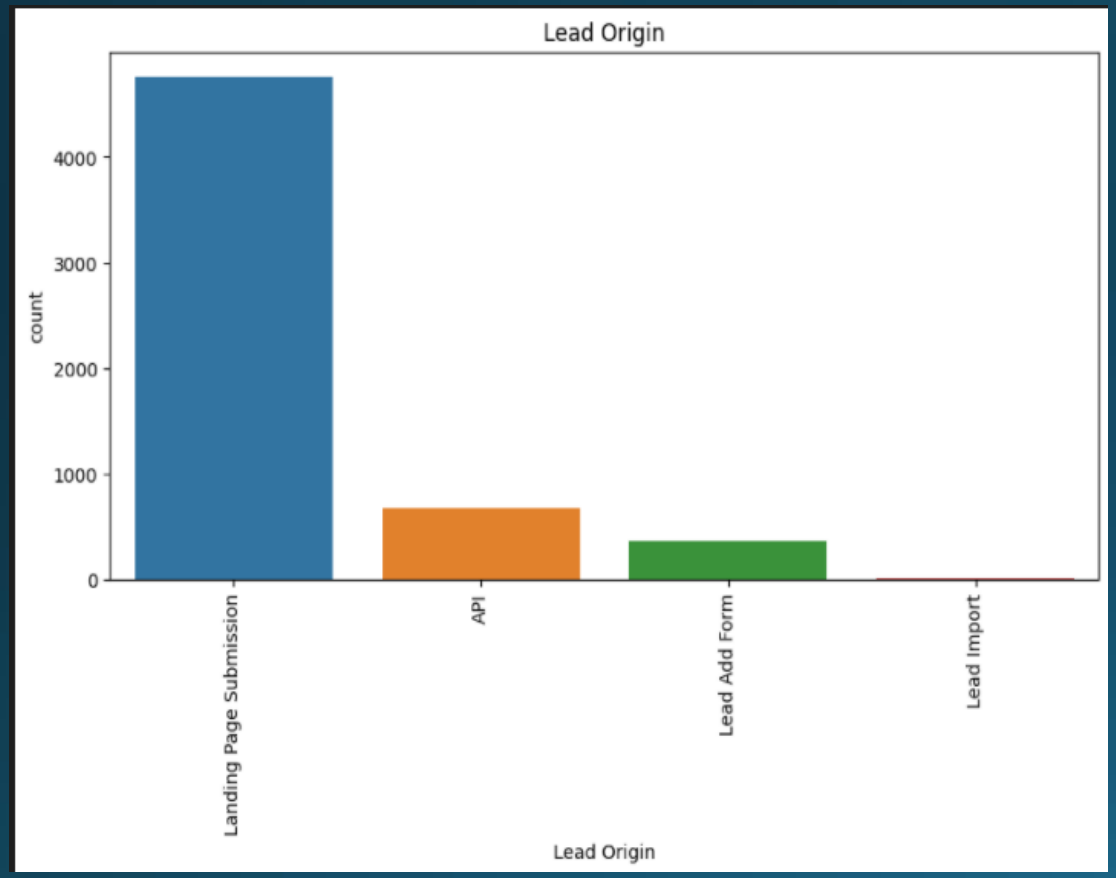
- Univariate analysis for categorical variables

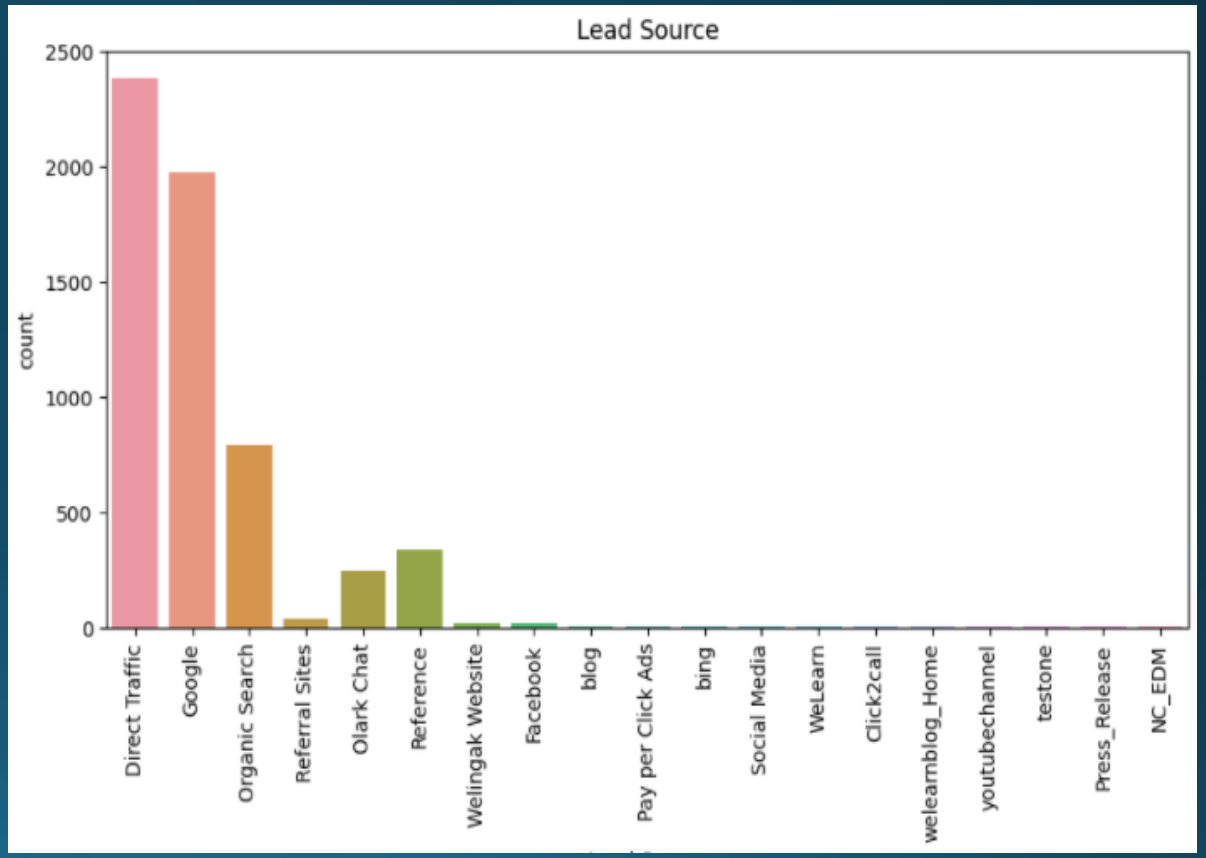
- Univariate analysis for Continuous variables

- Bivariate analysis for categorical variables vs target variable

- Bivariate analysis for continuous variables vs target variable

Univariate analysis Observations

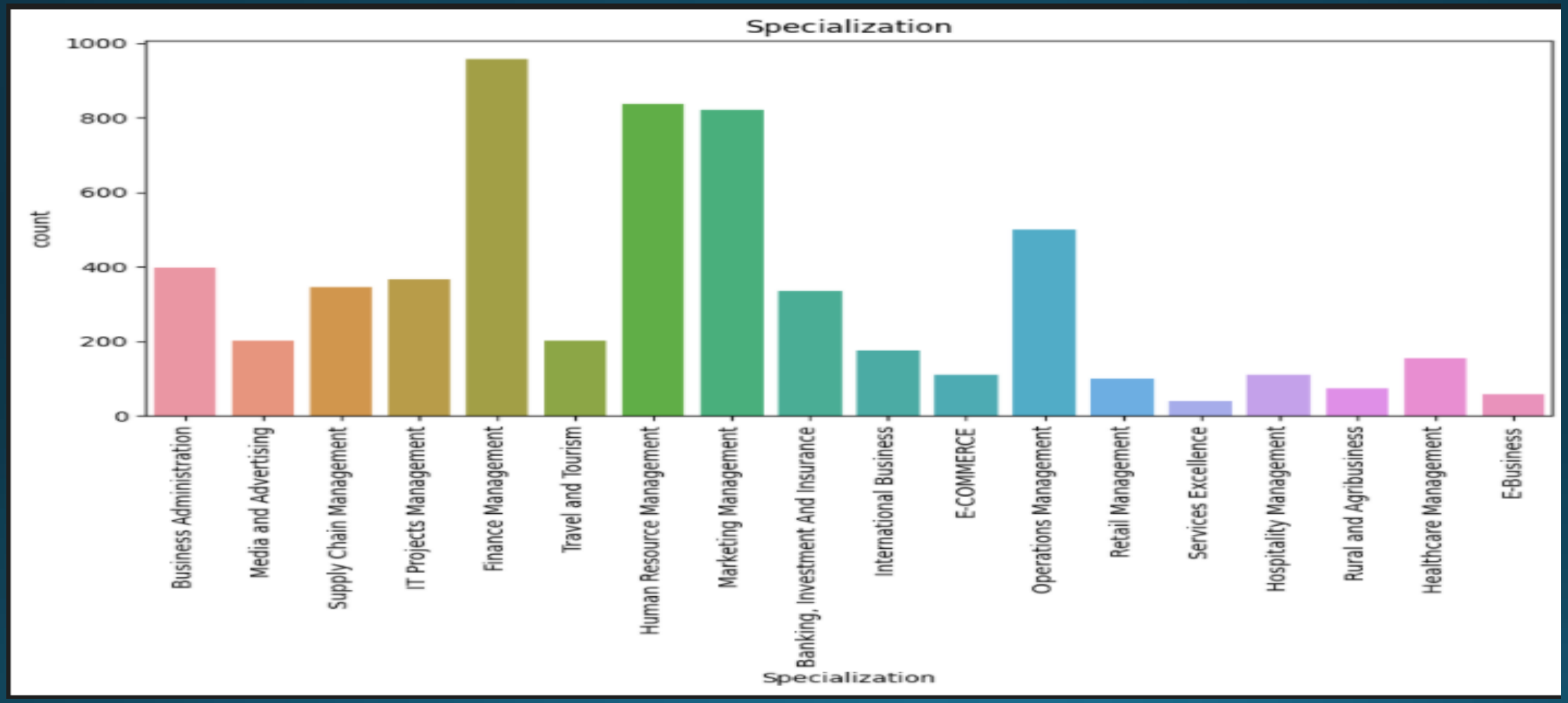




Most of the Leads originate from the Landing page submission

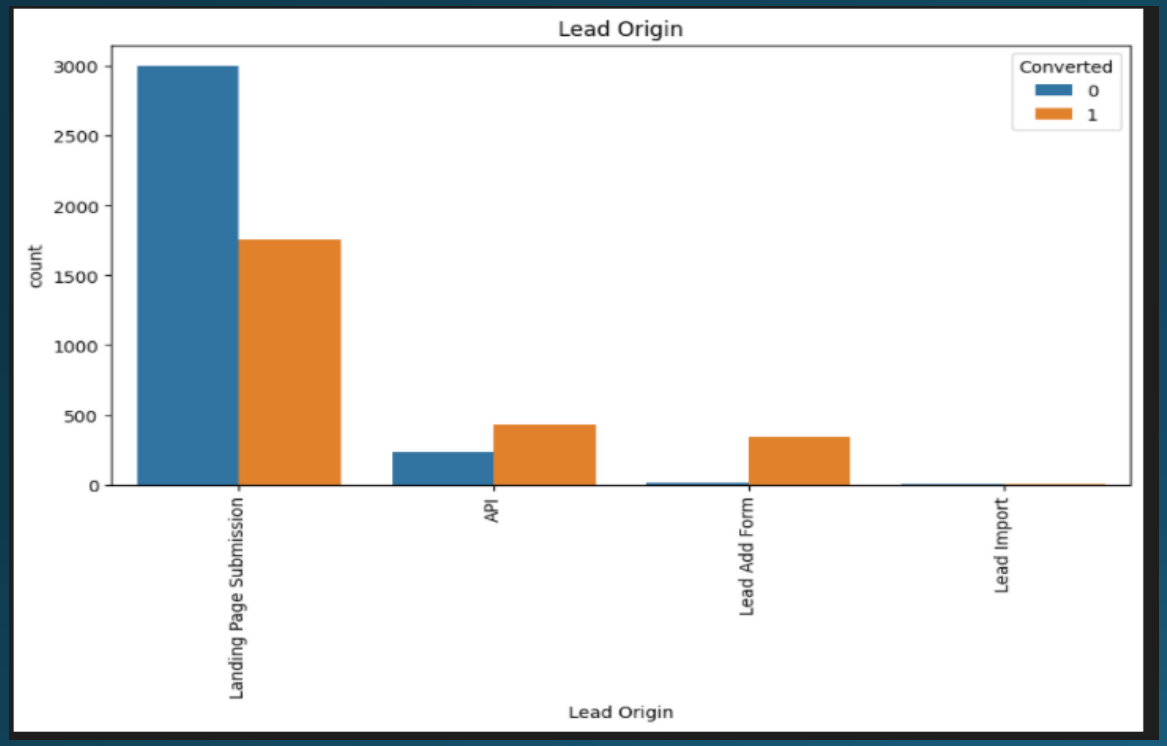
Direct traffic , google search and organic search are the most popular sources of Leads

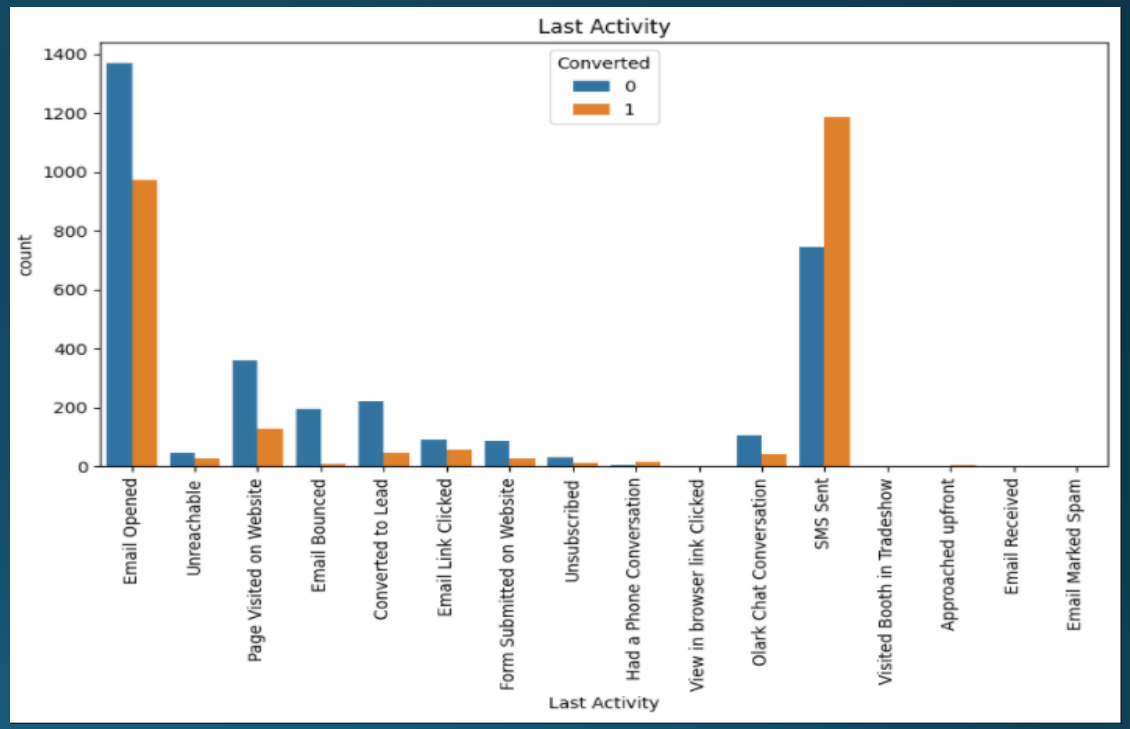
Univariate analysis Observations



Finance HR and Marketing are the most popular specialization that people are interested

Bivariate analysis variables vs target variable

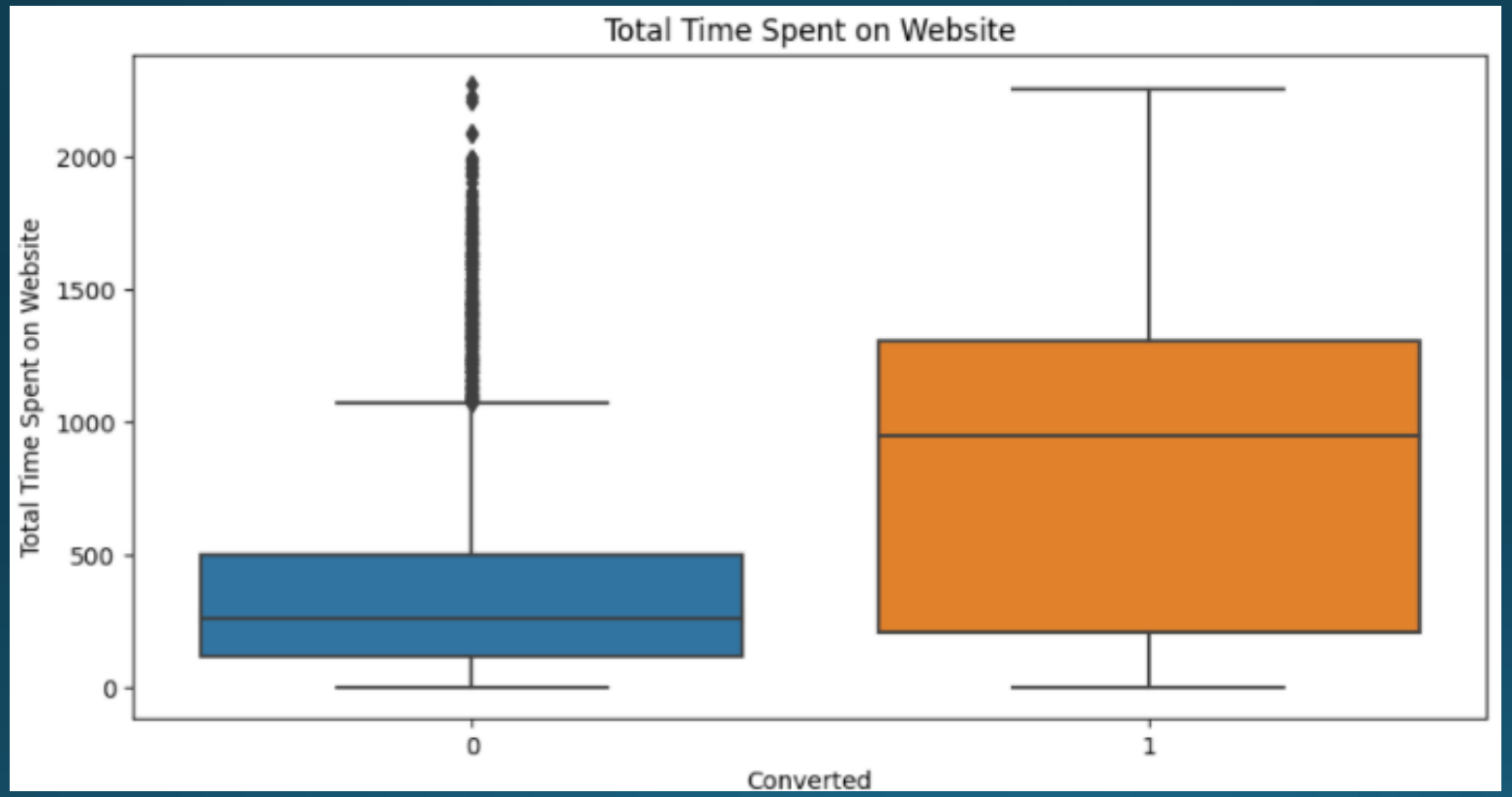




Api and Lead Add Form are very effective as the conversion rate is very high

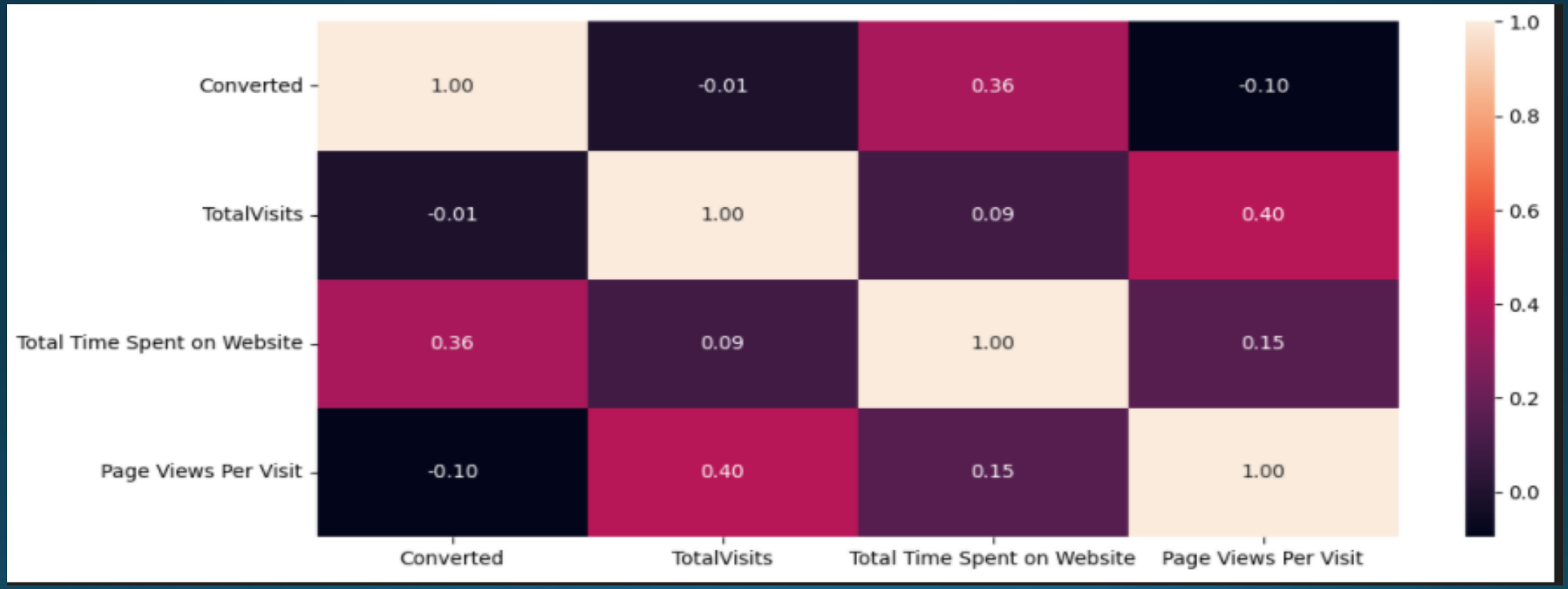
SMS and email are the predominant mode for reaching candidates in that
SMS seems to be very effective as the conversion is high **61 %**

Bivariate analysis variables vs target variable



More time Spent leads to more conversion explaining the interest of the candidate.
Maintaining a good website is also critical for good initial experience by the users

Correlation analysis



There is no significant correlation between the feature columns

Feature Selection

- Initial approach: Recursive Feature Elimination with Cross-Validation (RFECV) using Logistic Regression as the estimator (73 optimal features identified).
- Further reduction: Select Best with chi-squared test to select the top 20 features.
  - Total Time Spent on Website
  - Page Views Per Visit
  - Lead Origin_Landing Page Submission
  - Lead Origin_Lead Add Form
  - Lead Source_Direct Traffic
  - Lead Source_Olark Chat
  - Lead Source_Reference
  - Lead Source_Welingak Website
  - Do Not Email_Yes
  - Last Activity_Converted to Lead
  - Last Activity_Email Bounced
  - Last Activity_Form Submitted on Website
  - Last Activity_Olark Chat Conversation
  - Last Activity_Page Visited on Website
  - Last Activity_SMS Sent
  - A free copy of Mastering The Interview_Yes
  - Last Notable Activity_Email Bounced
  - Last Notable Activity_Modified
  - Last Notable Activity_SMS Senta

Model Training and Evaluation

- Data split: Training and testing sets.
- Feature scaling: StandardScaler.
- Model: Logistic Regression with GridSearchCV for hyper parameter tuning.
- Performance metrics: Accuracy, precision, recall, F1-score, and ROC AUC score.
- The best model is give by the features selected by second method using Chi2 selection

Key Results

- Model performance: Reasonable performance with reduced feature set.
- Accuracy: 0.779
- Precision: 0.770
- Recall: 0.710
- F1-score: 0.739
- ROC AUC score: 0.772
- Model performs reasonably well in distinguishing between the two classes.

Precision-Recall Curve and AP Score

- Calculated predicted probabilities and computed precision-recall curve.
- Plotted the curve and calculated the Average Precision (AP) score.
- AP score: 0.83 (83% average precision throughout the range of recall values).

Key Observations & Suggestions

- **Maintaining a very good website** is very important as most leads originate from Landing page submission
- **Search index optimization** can improve the footfall as Direct traffic , google search and organic search are the most common sources
- **Finance , HR and Marketing** are the most popular specialization that people are interested.so we can focus on improving these courses.
- Api and Lead Add Form are very effective as the conversion rate is very high
- SMS seems to be very effective as the conversion is high 61 %
- More time Spent leads to more conversion explaining the interest of the candidate.

Thank You