# LEAD SCORING CASE STUDY

Group Members

Prem Raj

Lalit Kumar

Raghav N

## Problem Statement

An education company named X Education needs assistance in choosing the leads that have the best chance of becoming paying clients. The business wants us to develop a model in which each lead is given a lead score, with higher lead scores indicating a higher likelihood of conversion and lower lead scores indicating a lower likelihood of conversion. The desired lead conversion rate has been estimated by the CEO to be in the range of 80%.

## Goals and Objectives

- There are quite a few goals for this case study.
- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.
- There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

# METHODOLOGY

During the data cleaning and manipulation process, several steps were undertaken:

- **Duplicate Data:** Duplicate data was checked for and handled appropriately to ensure data integrity.
- **Handling NA Values and Missing Values:** NA values and missing values were examined and addressed through suitable methods, such as imputation or removal, to avoid any impact on the analysis.
- **Dropping Columns:** Columns that contained a large number of missing values and were deemed irrelevant for the analysis were dropped from the dataset.
- **Imputation:** If necessary, values were imputed for missing data points using appropriate techniques to ensure the dataset's completeness.
- **Outlier Detection and Handling:** Outliers in the data were identified and managed to avoid their adverse influence on the analysis.

**Exploratory Data Analysis (EDA)** was performed, encompassing the following steps:

- **Univariate Data Analysis:** The count and distribution of variables were examined individually to gain insights into their characteristics.
- **Bivariate Data Analysis:** Correlation coefficients and patterns between variables were analyzed to identify any relationships or dependencies.

# METHODOLOGY

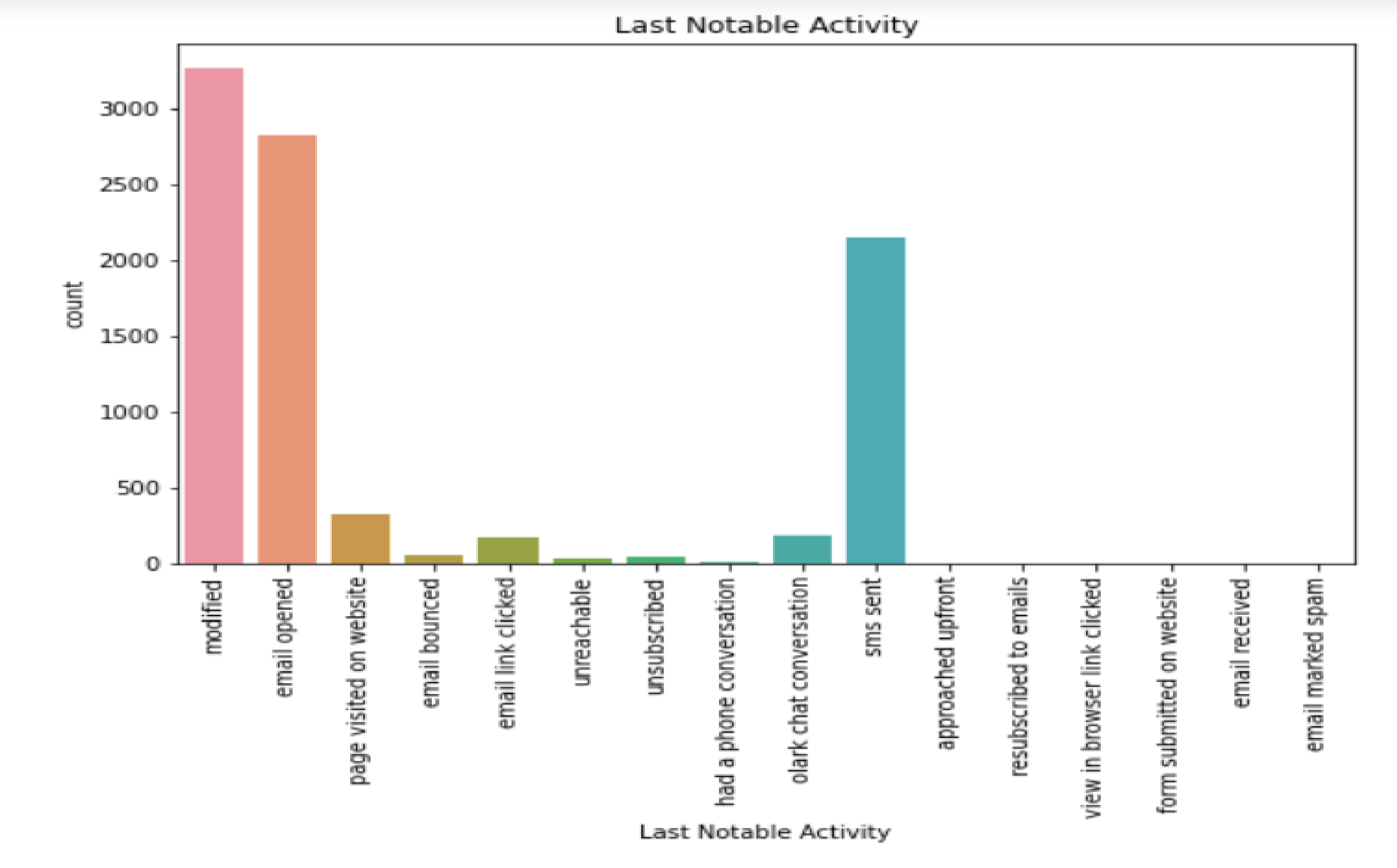
- Feature scaling and encoding techniques were employed to prepare the data for analysis, including the creation of dummy variables as needed.
- A logistic regression model was selected as the classification technique for model building and prediction.
- The model was validated to assess its performance and accuracy in predicting outcomes.
- The findings of the model were presented, highlighting the key insights and patterns discovered during the analysis.
- Conclusions and recommendations were drawn based on the analysis, providing insights and suggestions for further actions or improvements.

# DATA MANIPULATION

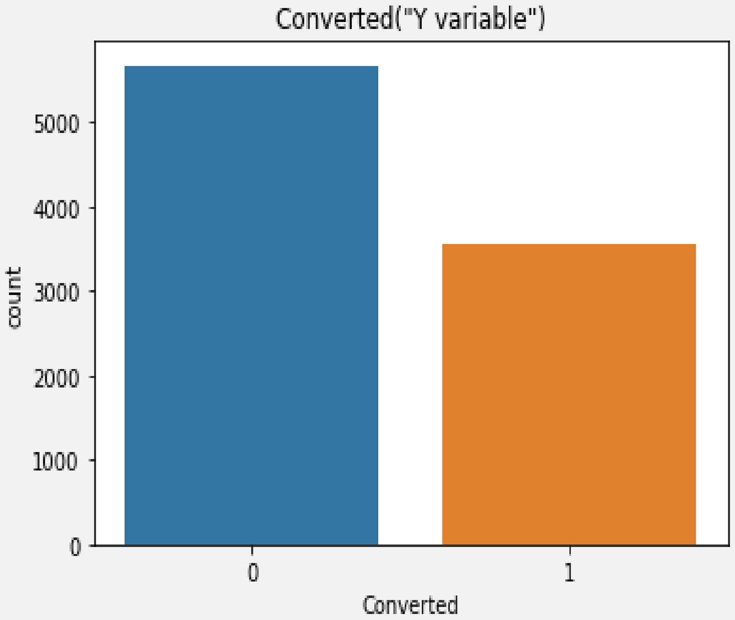
To streamline the analysis, the following steps were taken:

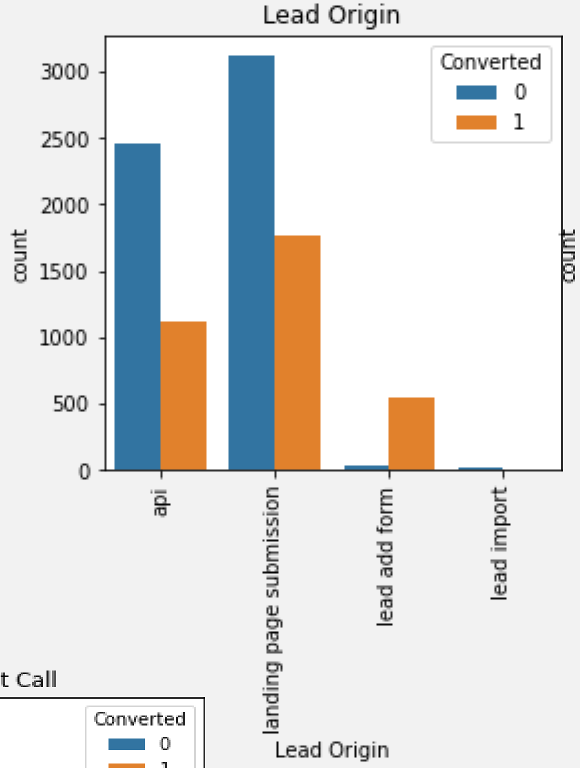
- The dataset consists of 37 rows and 9,240 columns.
- Features such as "Magazine," "Receive More Updates About Our Courses," "Update me on Supply," "Chain Content," "Get updates on DM Content," and "I agree to pay the amount through cheque" were dropped as they had only a single value and provided no significant information.
- The columns "Prospect ID" and "Lead Number" were deemed unnecessary for the analysis and were removed from the dataset.
- Certain object-type variables showed minimal variance in their values. Consequently, features like "Do Not Call," "What matters most to you in choosing course," "Search," "Newspaper Article," "X Education Forums," "Newspaper," and "Digital Advertisement" were dropped.
- Columns with more than 35% missing values, such as "How did you hear about X Education" and "Lead Profile," were dropped from the dataset to ensure data integrity and reliability.
- These steps were implemented to refine the dataset and focus on the most relevant and informative features for the analysis.

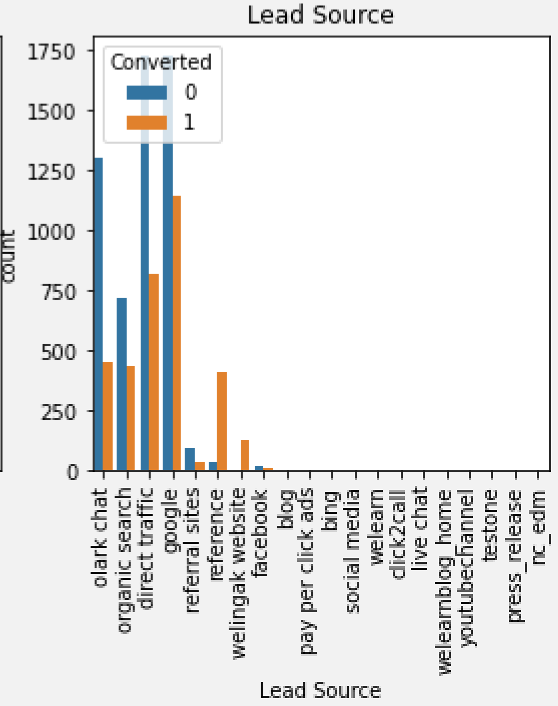
# EXPLORATORY DATA ANALYSIS (EDA)

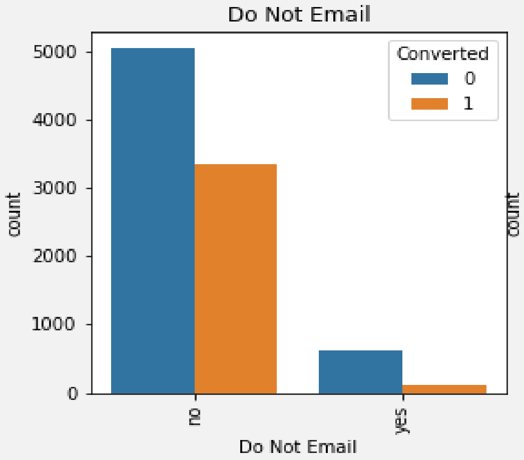


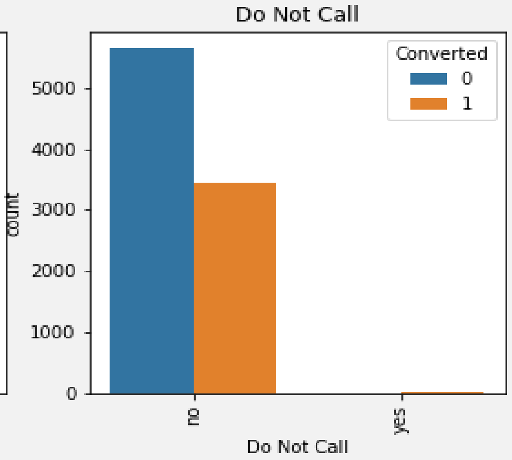
# CATEGORICAL VARIABLES TO CONVERTED



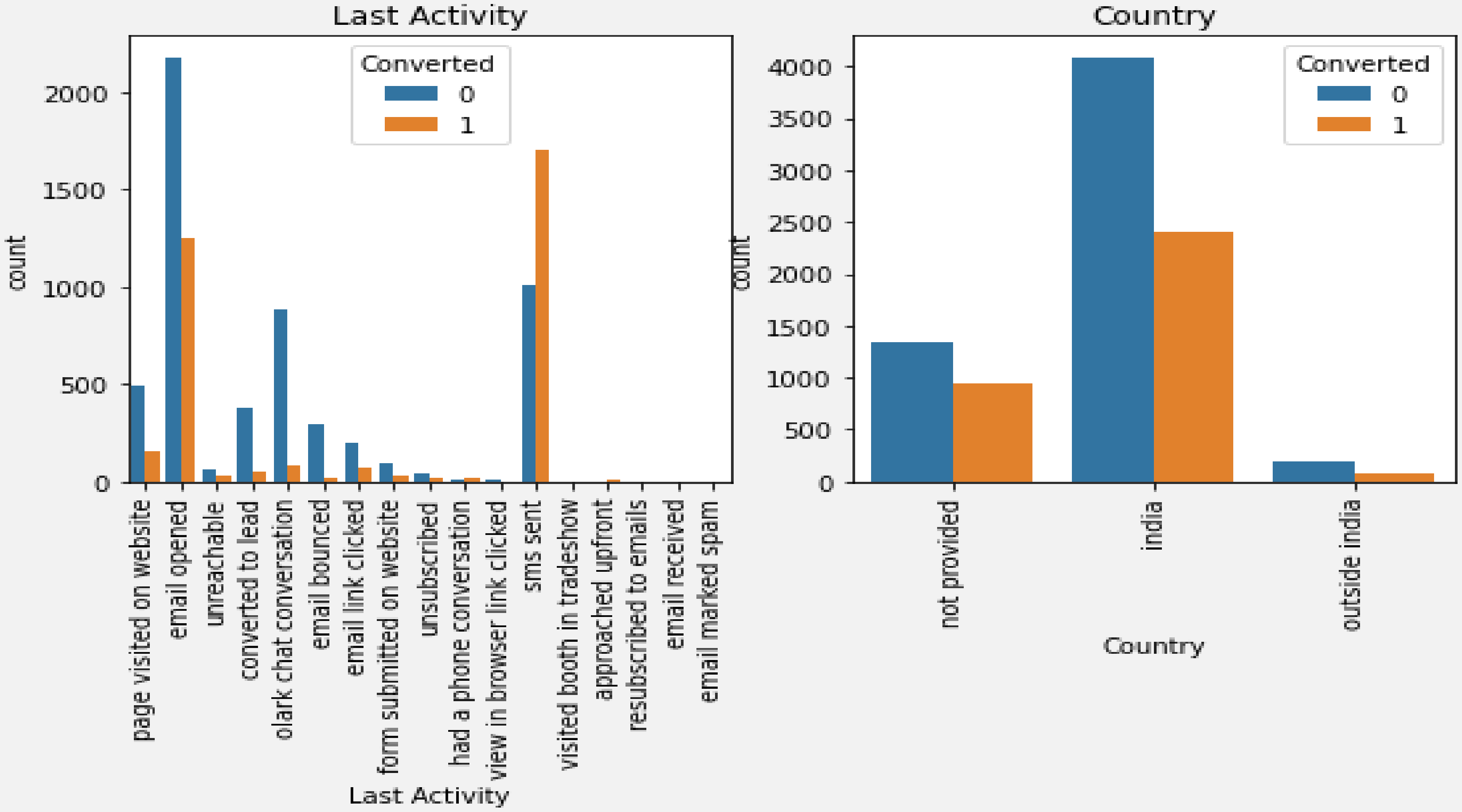




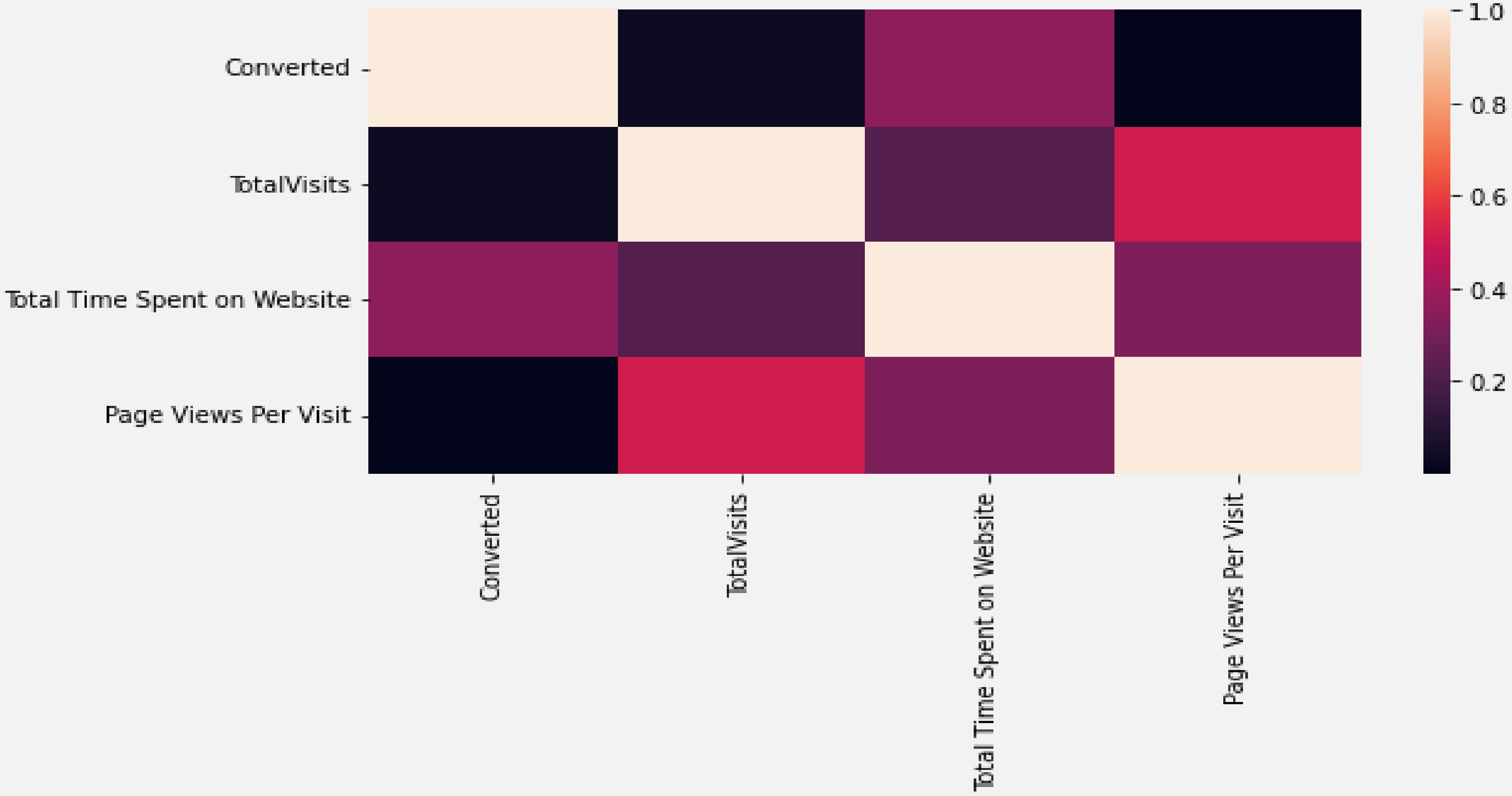




# CATEGORICAL VARIABLES TO CONVERTED



# CORRELATION AMONG VARIABLES



# MODEL BUILDING

To prepare the data for regression analysis, the following steps were followed:

- Train-Test Split:
- The dataset was divided into training and testing sets using a 70:30 ratio, respectively.
- RFE was applied to select the most relevant variables for the regression model. The output of RFE included 15 variables.
- The model was constructed by removing variables with a p-value greater than 0.05 and a VIF (variance inflation factor) value greater than 5. This step aimed to improve the model's accuracy and eliminate insignificant predictors.
- The trained model was used to make predictions on the test dataset.
- The overall accuracy of the model was determined to be 81%, indicating a satisfactory level of prediction performance.

By following these steps, the dataset was effectively split, relevant features were selected, and a regression model was built and evaluated to provide accurate predictions.

# CONCLUSION

Based on the analysis conducted, it has been identified that certain variables have a significant impact on potential buyers and their likelihood to change their minds and purchase courses from X Education. These variables, in descending order of importance, are:

**Total time spent on the Website,**

**Total number of visits,**

Lead source:

- **Google & Direct traffic,**
- **Organic search & Welingak website**

**Last activity:** The last activity the potential buyers engaged in before making a decision is also significant. The following activities were found to be influential:

- **SMS & Olark chat conversation**

**Lead origin**: When the lead originates from the Lead add format, it has a positive impact on the potential buyer's decision-making process.

**Current occupation**: Potential buyers who are working professionals are more likely to change their minds and make a purchase.

.