# Lead scoring case study

Bharat Panera

# Problem Statement

**Low Conversion Rate**: X Education experiences a low lead conversion rate, with only 30% of leads turning into paying customers.

**Efficiency Concerns**: Despite a high volume of daily leads, the conversion process lacks efficiency, resulting in suboptimal outcomes.

**Identifying 'Hot Leads'**: X Education aims to identify 'Hot Leads'—those most likely to convert—to enhance the overall conversion rate.

**Lead Scoring Model**: Develop a lead scoring model to assign scores to leads, prioritizing high-scoring leads for targeted sales efforts.

**Target Conversion Rate**: The CEO sets a target conversion rate of 80%, emphasizing the need for significant improvement in the current conversion rate.

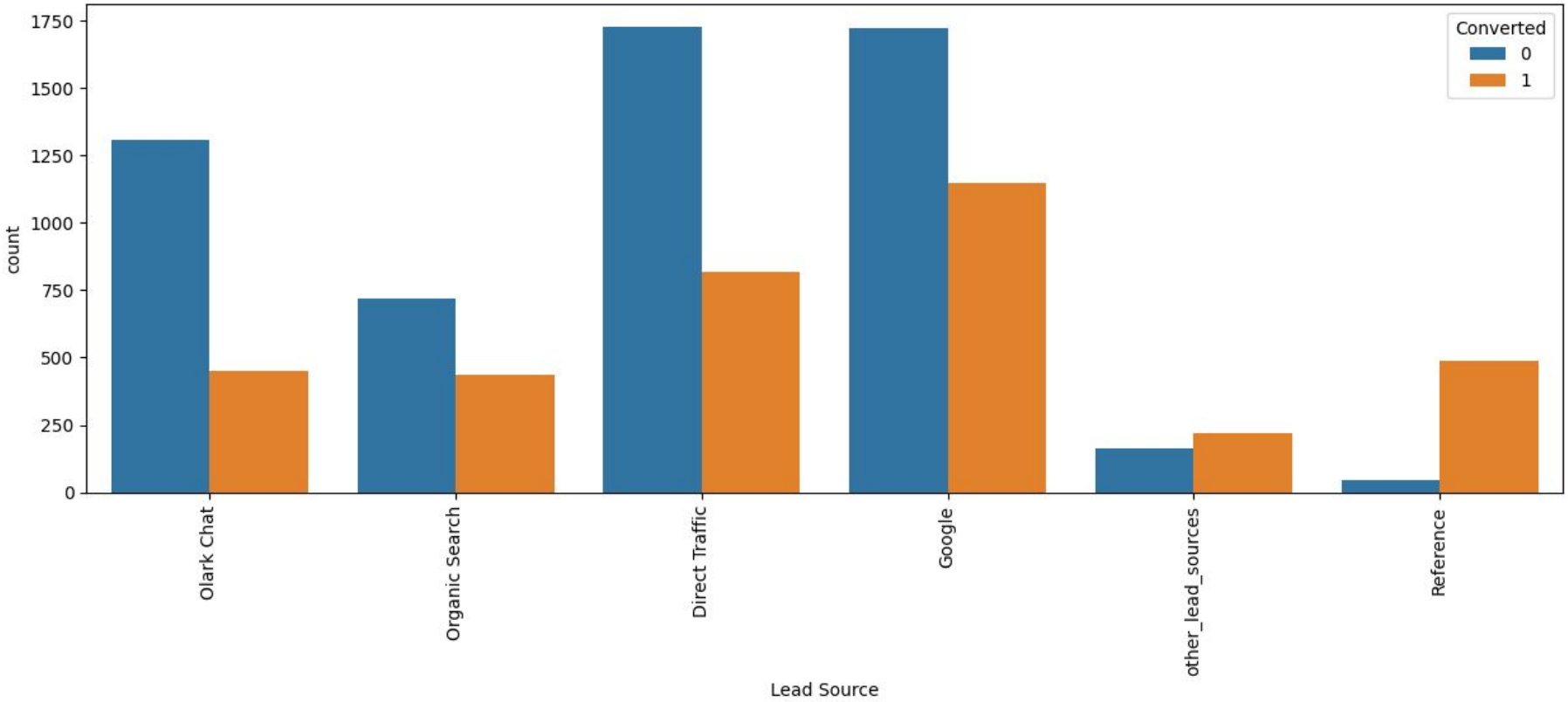
# Solution Approach

1. **Data Understanding**
    - Initial exploration to comprehend the dataset's structure and variables.
2. **Exploratory Data Analysis (EDA)**
    - Handle missing values.
    - Analyze categorical and numerical columns.
3. **Preparing Data for Modeling**
    - Create dummy variables.
    - Split data into train/test sets.
    - Scaling Features.
4. **Model Building and Training**
    - Train logistic regression model.
5. **Testing and Validation**
    - Test model on test data.
    - Validate model performance.

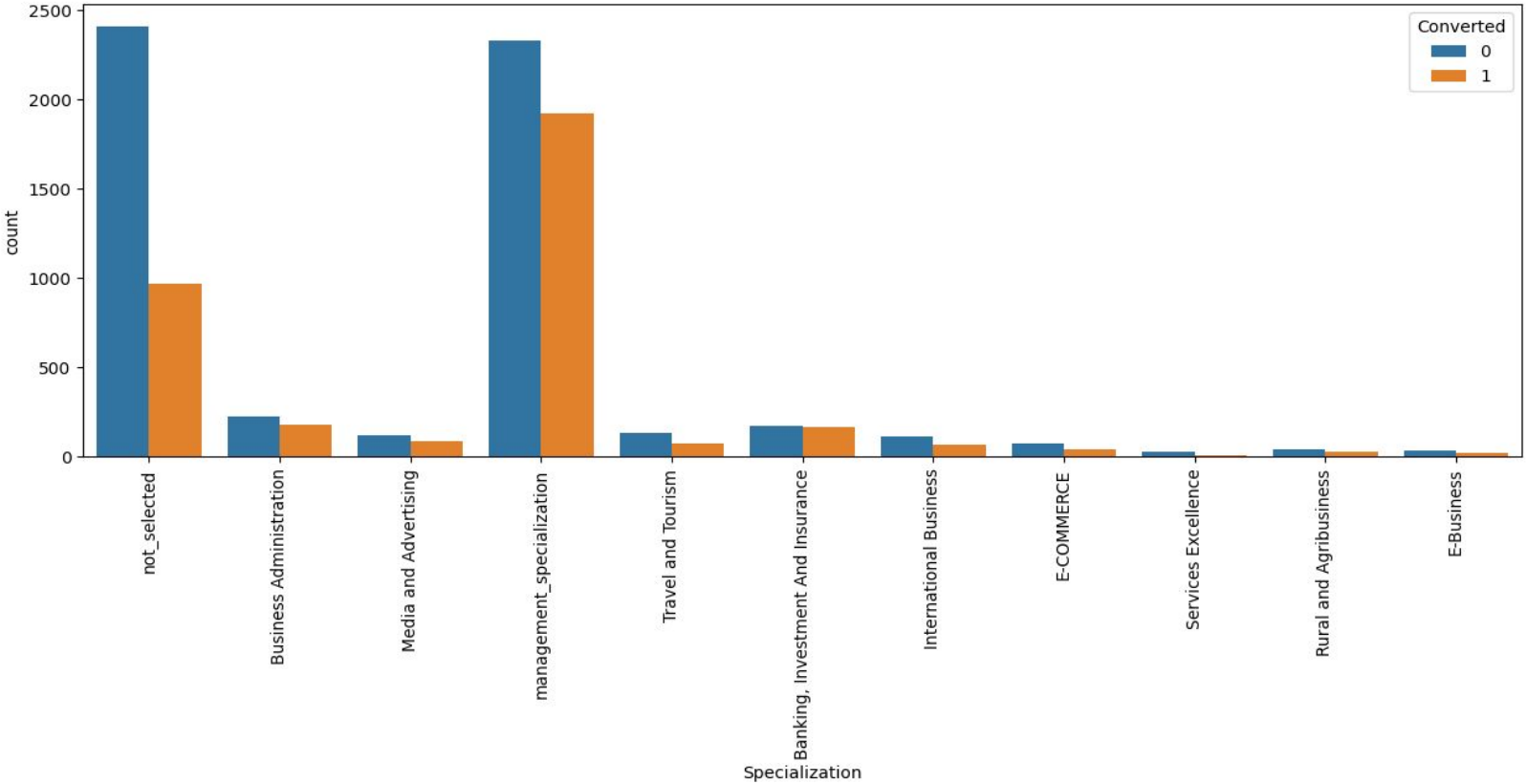
# Exploratory Data Analysis (EDA)

1. **Handling Null Values**:
    - Removed columns with null values exceeding 40% to ensure data integrity and model performance.
2. **Categorical Column Analysis**:
    - Conducted detailed analysis of each categorical column:
        - Examined distribution and importance of each category.
        - Identified columns deemed unimportant and removed them from further analysis.
3. **Numerical Column Analysis**:
    - Investigated each numerical column:
        - Conducted outlier treatment to ensure data consistency and model robustness.
        - Assessed correlation between numerical columns and the target variable "converted" to identify potential predictors.

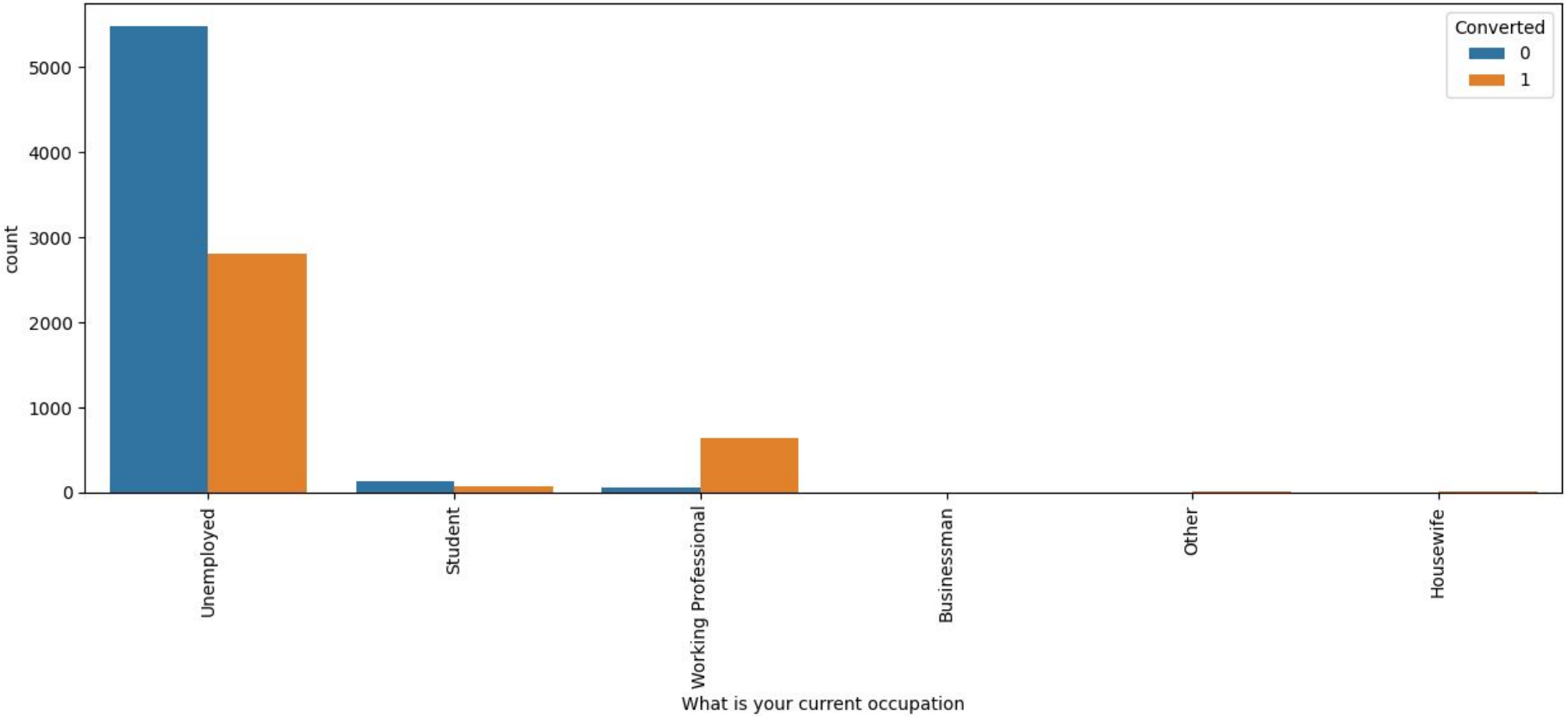
# Correlation of Categorical Variables - “Lead Source”



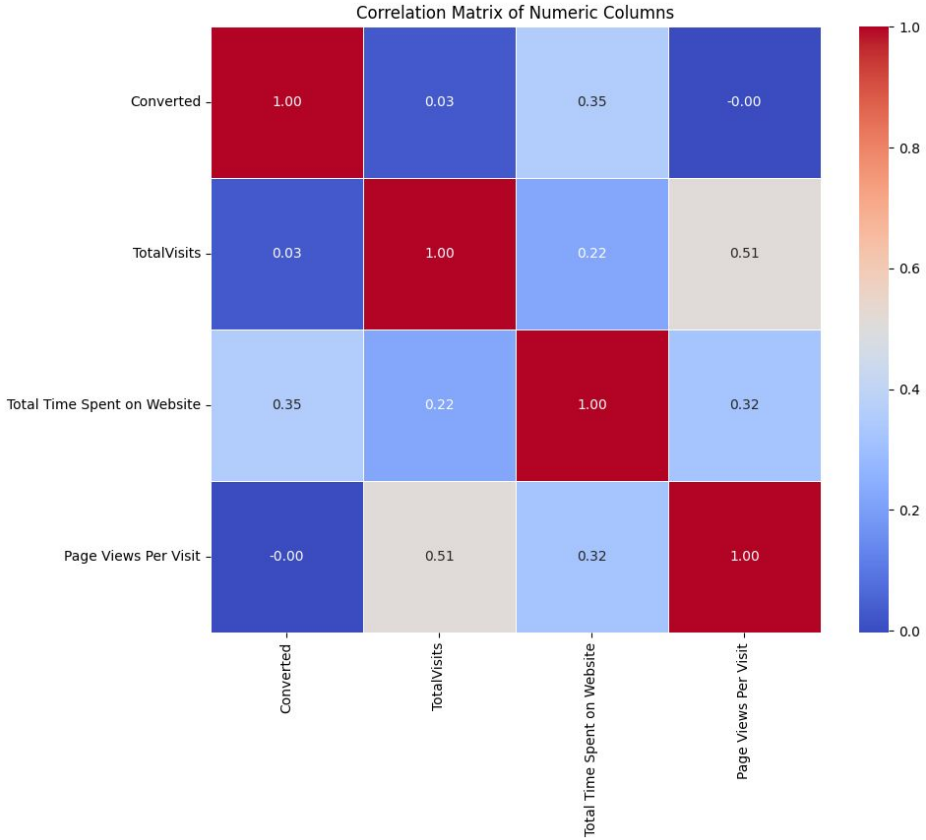
# Correlation of Categorical Variables - “Specialization”



# Correlation of Categorical variables - “What is your current occupation”



# Numerical Column Analysis



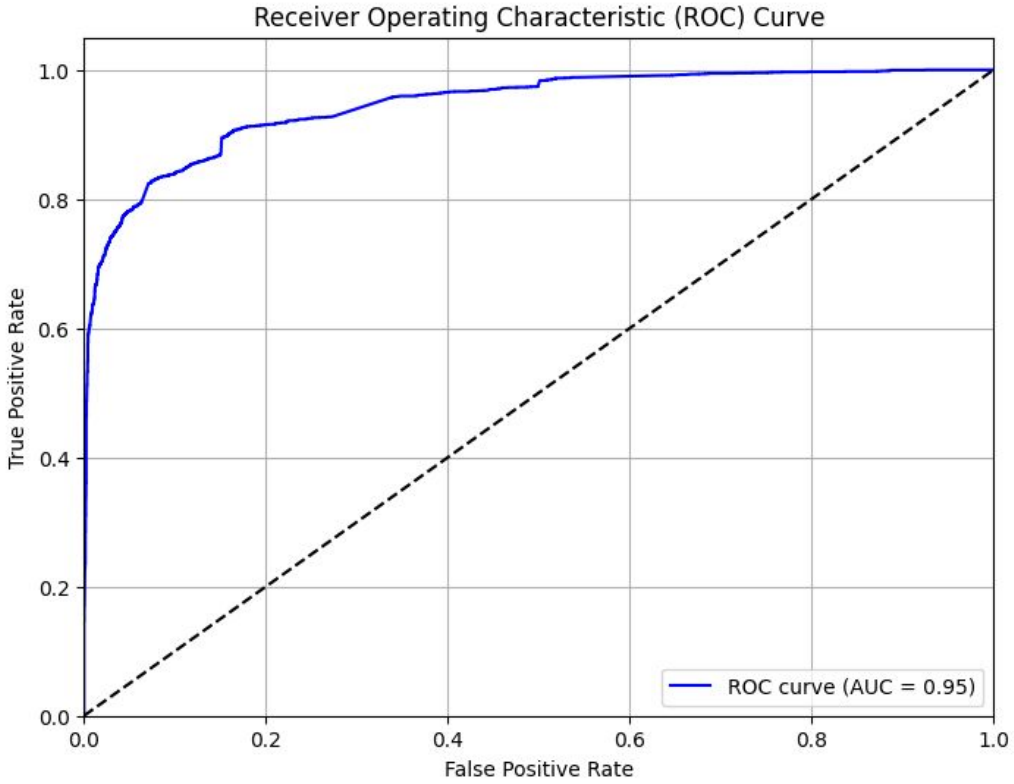
# Preparing Data for Modeling

1. **Creation of Dummy Variables**:
    - Converted categorical variables into numerical format using dummy encoding.
    - Ensured compatibility with machine learning algorithms by representing categorical data as binary indicators.
2. **Splitting Data into Train and Test Datasets**:
    - Divided the dataset into training (70%) and testing (30%) sets to assess model performance.
    - Maintained data separation to prevent overfitting and ensure unbiased evaluation of the model.
3. **Scaling the Features with StandardScaler**:
    - Applied StandardScaler() for feature scaling to standardize feature magnitudes.
    - Facilitated convergence and stability of the logistic regression model during training.

# Model Building and Training

1. **Feature Selection with Recursive Feature Elimination (RFE)**:
    - Utilized Recursive Feature Elimination (RFE) to select the best 15 features for model training.
    - Identified the most relevant features contributing to predictive performance.
2. **Model Creation with Statsmodel Library**:
    - Constructed initial logistic regression models using the Statsmodel library.
3. **Iterative Model Refinement**:
    - Iteratively refined the models by evaluating feature importance based on statistical significance.
    - Retained features with p-values less than 0.05 to ensure statistical significance.
    - Implemented Variance Inflation Factor (VIF) analysis to address multicollinearity, keeping VIF values below 5.
    - Iteratively created and evaluated a total of four logistic regression models to optimize predictive accuracy.
4. **Model Evaluation**:
    - Assessed the performance of each model using accuracy metrics.
    - Achieved a high accuracy rate of 88.66% through iterative refinement and feature selection.

# ROC Curve


Receiver Operating Characteristic (ROC) Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (AUC = 0.95)

# Precision-Recall Curve