Predicting Customer Purchase Behavior

Project Overview

Objective

The goal of this project is to develop a machine learning model that predicts the likelihood of a customer making a purchase based on their demographic features like age, gender, and annual income. The insights generated will assist businesses in targeting the right customer segments, optimizing marketing strategies, and ultimately improving sales conversion rates.

Problem Statement

Businesses often struggle to efficiently target potential customers. The challenge is to identify key demographic factors influencing purchase behavior to avoid ineffective marketing investments. By predicting purchase likelihood, businesses can allocate resources more effectively and enhance customer engagement.

Data Summary

Source

The dataset was sourced from Kaggle and contains information on:

- **Demographics:** Age, gender, and annual income.
- **Behavioral Attributes:** Time spent on the website and purchase status.

Data Wrangling

- The dataset initially had duplicates representing customers who did not make purchases, which were removed to avoid skewing the analysis.
- The dataset had no missing values, outliers, or inappropriate data types.

For detailed data wrangling processes, refer to the Data Wrangling Notebook

Exploratory Data Analysis (EDA)

Key insights from EDA include:

- **Time on Site:** Customers who made a purchase typically spent around 35 minutes on the website, compared to 20 minutes for those who did not purchase.
- **Demographic Patterns:** Clear relationships were observed between demographic features and purchase behavior.

Visualizations revealed important trends that informed the selection of appropriate machine learning models.

Preprocessing and Feature Engineering

- **Numerical Features:** Scaled using `MinMaxScaler`.
- **Categorical Features:** Encoded using `OneHotEncoder`.

The preprocessing pipeline ensures all features are appropriately transformed before feeding them into machine learning models.

Model Selection and Evaluation

Models Implemented

1. **Logistic Regression**
2. **Random Forest Classifier**
3. **Gradient Boosting Classifier**

Best Model: Random Forest Classifier

- **Accuracy:** 93% on the test set.
- **Cross-Validation Score:** 91%.
- **Feature Importance:** Annual income and age were identified as the key drivers of purchase decisions.

Evaluation Metrics

- **Accuracy:** Measures overall prediction correctness.
- **Classification Report:** Provides precision, recall, and F1-score for each class (purchasers and non-purchasers).

Key Findings

- **High Accuracy:** The Random Forest model effectively predicted purchase behavior with 93% accuracy.
- **Actionable Insights:**
    - **Marketing Focus:** Targeting specific income brackets and age groups can increase purchase likelihood.
    - **Site Engagement:** Encouraging more time on the website may boost conversions.

Future Work

1. **Incorporate Behavioral Features:** Add metrics like browsing behavior and purchase history for richer insights.
2. **Advanced Models:** Explore customer conversion through discounts.
3. **Scalability:** Test the model on larger, more diverse datasets to ensure generalizability.
4. **Cost-Benefit Analysis:** Optimize models for business goals like revenue maximization.

Conclusion

This project demonstrates the power of machine learning in understanding and predicting customer behavior. By leveraging demographic insights, businesses can make data-driven decisions to enhance marketing efficiency and customer experience.

For detailed code and documentation, visit the GitHub Repository