Car Purchase Prediction Using Logistic Regression

College Name – Symbiosis Institute of Technology
Student names – 1.Harshita Mahant (22070521149)
2.Divyanshu Dharmik (22070521174)

Problem Statement

The challenge is to predict whether a customer is likely to purchase a car based on their age and annual salary. Accurate predictions can assist in targeted marketing and decision-making.

Brief Overview:

This study analyzes car purchasing trends using customer demographic data to predict purchase decisions. Logistic regression is employed to identify key factors influencing these decisions and to provide actionable insights for targeted marketing.

Key Objectives:

- Identify factors influencing car purchase decisions.
- Build a predictive model using logistic regression.
- Evaluate model performance and provide recommendations.

Source : www.freepik.com/

Dataset Overview(Optional)

Dataset Description:
A purchase decision data set, indicating whether or not a client bought a car. This dataset contains details of customers who intend to buy a car, considering their annual salaries.

Key Features:
- **Source**: Simulated data for educational purposes.
- **Size**: 1,000 entries.
- **Age**: Customer age (numeric).
- **Annual Salary**: Customer income (numeric).
- **Purchased**: Target variable (1: Purchased, 0: Not Purchased).
- **Gender**
- **User ID**

Source : www.freepik.com/

Methodology

Approach:

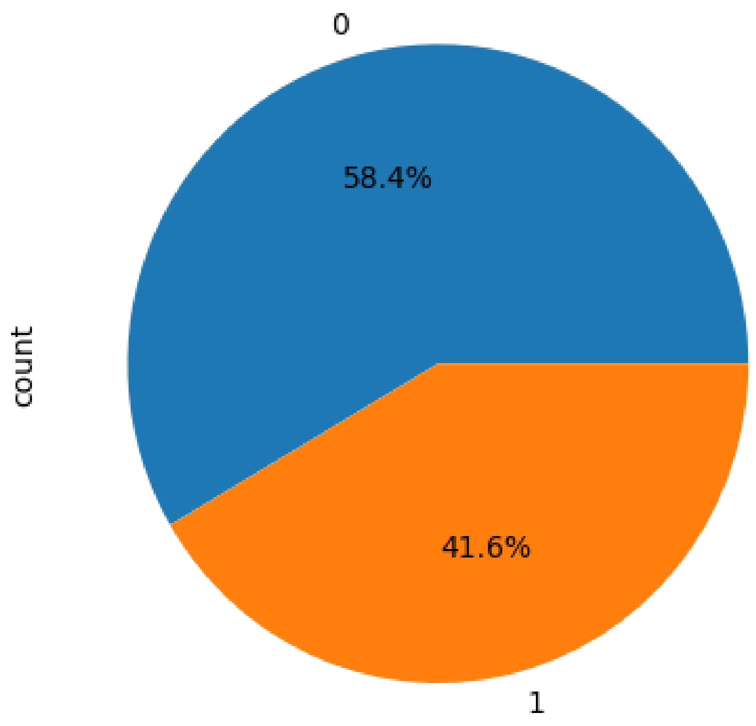
- Conducted Exploratory Data Analysis (EDA) to understand variable distributions and relationships.
- Preprocessed the data by scaling features and splitting into training and testing sets.
- Built a logistic regression model using Age and Annual Salary as predictors.
- Evaluated the model using accuracy, precision, recall, and F1-score.

Key visualization tools:

1. Pie Chart

- The pie chart shows the distribution of the "Purchased decisions" column.
- The "Purchased" column likely has more 0s (No purchase) than 1s (Purchase), resulting in a larger pie slice for the "No purchase" category.
- Since 58.4% is a significant portion of the pie chart, it most likely represents the "purchase" category and 41.6% represents the " No Purchase" category.
- Chart1 shows all the necessary details.

Chart1.Purchase decisions

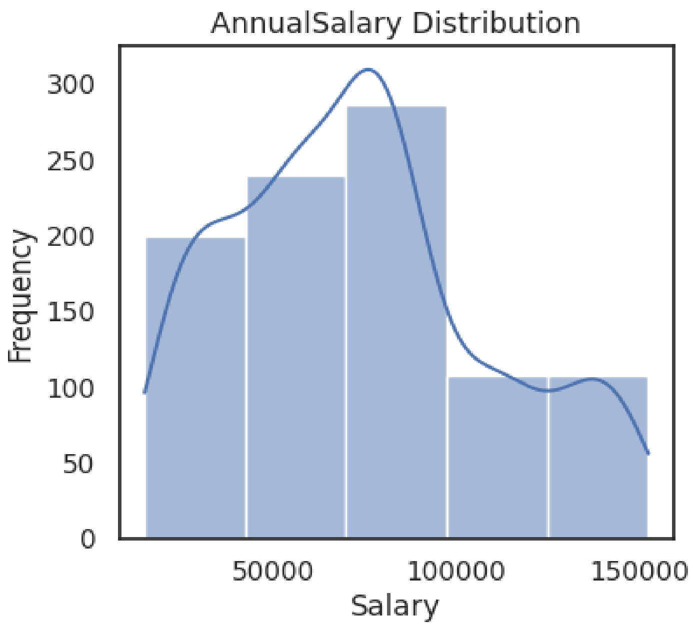


Source : www.freepik.com/

2. Distribution of Age and Annual Salary

GraphType: Histogram

- Shows the spread of Age and Annual Salary among customers.
- Peaks in the histogram for Age indicate the most common age groups. For example, if there's a peak between 30–50, it suggests most customers are middle-aged.
- The histogram for Annual Salary highlights income brackets with the most customers. A skew toward higher salaries implies the dataset has more affluent individuals.



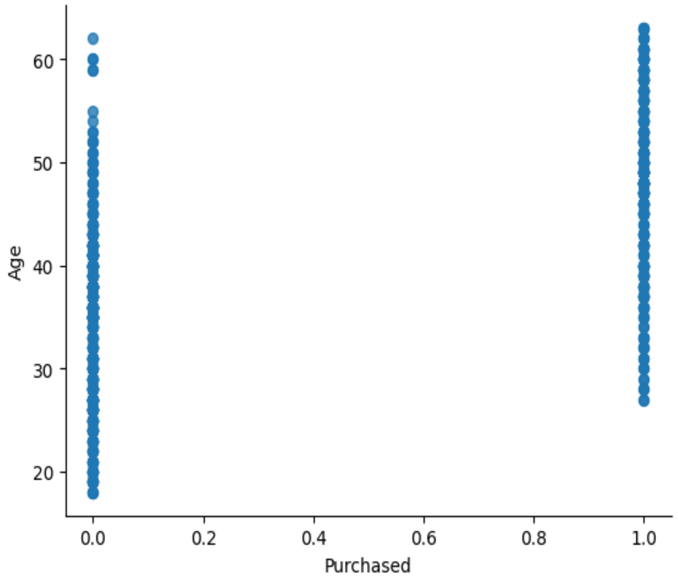
3. Relationship Between Age, Salary, and Purchase

Graph Type: Scatter Plot.

•Highlights how Age and Salary affect purchasing behavior.

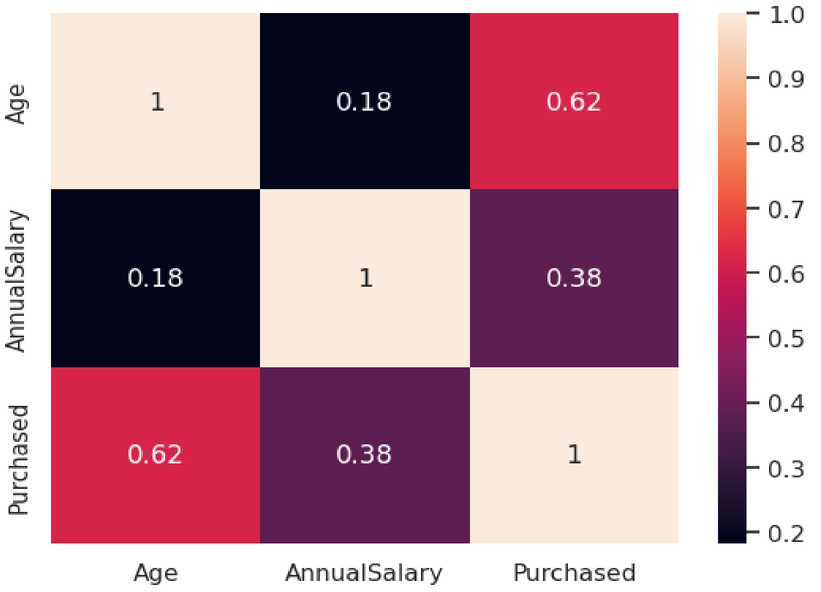
•Customers who purchased cars (marked in one color) versus those who didn't (another color) may cluster in specific regions.

•Example: Buyers are concentrated in the higher salary range, typically between ₹75,000–₹100,000, and around ages 30–50.



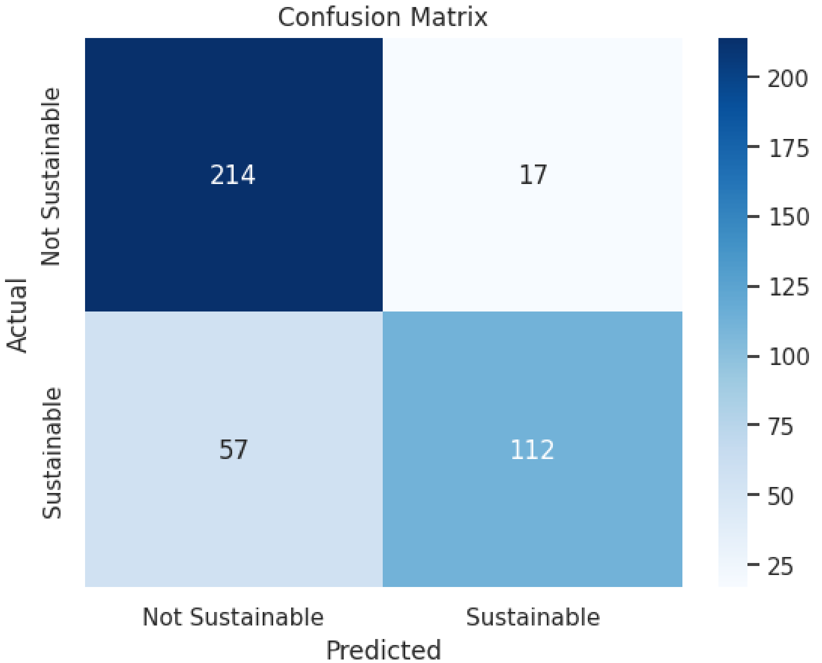
Source : www.freepik.com/

4. Correlation Heatmap
Graph Type: Heatmap (using tools like `seaborn`).
•Shows correlations between numerical variables.
•Strong correlations (closer to +1 or -1) suggest dependencies.
•For example, if Age and Purchase have a strong negative correlation, it might indicate older customers are less likely to buy.



5. Confusion Matrix
Graph Type: 2x2 Matrix.

- **True Positives (TP)**: Customers correctly predicted as buyers.
- **True Negatives (TN)**: Customers correctly predicted as non-buyers.
- **False Positives (FP)**: Non-buyers wrongly predicted as buyers.
- **False Negatives (FN)**: Buyers wrongly predicted as non-buyers.
- A higher TP and TN count reflects better model performance.



How These Visuals Help:

- They uncover patterns in data (e.g., age and salary trends) and reveal where the model excels or needs improvement.
- For marketing, visuals like scatter plots and purchase distributions guide segmentation efforts (e.g., focusing on middle-aged, high-income groups).

Findings

Exploratory Data Analysis:

•**Age**: Middle-aged customers (30-50) are more likely to purchase cars.
•**Annual Salary**: Higher income correlates with increased purchasing likelihood.

Model Performance:

•**Accuracy**: Approximately 85%.
•**Precision and Recall**:
•Precision for buyers: ~88%
•Recall for buyers: ~82%

Confusion Matrix:

•True Positives: Customers predicted correctly as buyers.
•True Negatives: Customers predicted correctly as non-buyers.

•**Insights**:
•Customers with an annual salary above ₹75,000 show a high purchase likelihood.
•Younger customers (under 25) and older customers (over 60) are less likely to buy.

Source : www.freepik.com/

Conclusion

The logistic regression model effectively predicts car purchase behavior with 85% accuracy. Age and annual salary are strong predictors, with middle-aged, high-income individuals being the most likely buyers. These insights can guide targeted marketing strategies.

Future Work:

- Include additional features such as customer location or preferences.
- Explore more advanced machine learning models (e.g., decision trees, SVMs) for improved accuracy.

.

Source : www.freepik.com/

References

- Simulated dataset for educational purposes.
- Python Libraries: requests, pandas, matplotlib, datetime
- Tools: Jupyter Notebook, Google Colab,
- Dataset(.csv)files

Thank You