TEAM - 592817
Hemant Modi
Immidi Shreyas

Project Report: Car Purchase Prediction

1. Introduction

1.1 Project Overview

The project centers on developing a robust machine learning (ML) solution designed to predict car purchases using customer data analysis. By amalgamating essential demographic and income-related features, the model aims to forecast the likelihood of a customer making a car purchase.

1.2 Purpose

The primary objective of this endeavor is to create a reliable predictive system leveraging machine learning algorithms. This system assists in estimating the probability of a potential customer's inclination towards purchasing a car. The predictive insights garnered from this model will offer invaluable guidance to automotive dealerships in targeting potential buyers effectively.

2. Literature Survey

2.1 Existing Problem

The current landscape in automotive sales and marketing faces challenges in accurately predicting customer behavior and purchase patterns. Traditional approaches often lack precision in estimating a potential buyer's inclination towards purchasing a car based on demographic and income-related factors.

2.2 References

The literature review draws insights from seminal research articles, industry reports, and studies concerning predictive analytics in automotive sales.

2.3 Problem Statement Definition

The core problem revolves around effectively leveraging customer data to predict car purchases accurately. The project aims to address this by formulating a machine learning solution that utilizes demographic and income-related features to forecast the likelihood of a customer making a car purchase.

3. Ideation & Proposed Solution

3.1 Empathy Map Canvas

Discuss insights gained from understanding potential car buyers, their needs, desires, and challenges through an empathy map.
Link for the Map - Empathy Map Mural
3.2 Ideation & Brainstorming
Detail the process of generating ideas and brainstorming solutions for predicting car purchases effectively.
Link for the Map - Brainstorm Map

4. Requirement Analysis

4.1 Functional Requirements
Data Collection:
- Demographic Information: Gather comprehensive data including age, gender, and historical purchase patterns.
- Income Data: Acquire accurate annual income details of potential customers.
- Previous Purchase History: Capture past car purchase records for predictive modeling.

Preprocessing:
- Data Cleaning: Ensure data integrity by handling missing values and outliers.
- Feature Engineering: Derive relevant features from collected data for enhanced model performance.
- Normalization/Scaling: Normalize or scale features for better convergence in machine learning algorithms.

Model Development:
- Algorithm Selection: Choose appropriate machine learning algorithms for predictive modeling.
- Hyperparameter Tuning: Optimize algorithm parameters for improved accuracy.
- Model Training: Train the model using historical data to predict car purchase likelihood.

4.2 Non-Functional Requirements
Performance:
- Accuracy: Achieve a high accuracy rate in predicting car purchases to ensure reliable insights.
- Scalability: Design the system to handle an increasing volume of customer data effectively.
- Response Time: Ensure quick response times for predictions within the application.
User Interface:
- Ease of Use: Develop a user-friendly interface for easy input of customer data.

- Accessibility: Ensure the application is accessible across various devices and platforms.

Security:
- Data Privacy: Implement stringent data privacy measures to protect customer information.
- Model Confidentiality: Safeguard the integrity of the predictive model from unauthorized access or modifications.

Compliance:
- Regulatory Compliance: Ensure compliance with data protection regulations and industry standards.

5. Project Design

5.1 Data Flow Diagrams & User Stories

DFDs visually represent how data moves within a system. For a car purchase prediction AI/ML model, processes include data collection, preprocessing, model training, and prediction. Data flows between user input, processed data, training data, and prediction results, while external entities like the user interface interact with the system.
User Stories are concise descriptions of features from the user's perspective. In the context of the AI/ML model, users want to input preferences for analysis, desire clear prediction results.User Stories articulate user needs and guide the development process

link: - Data Flow Doc

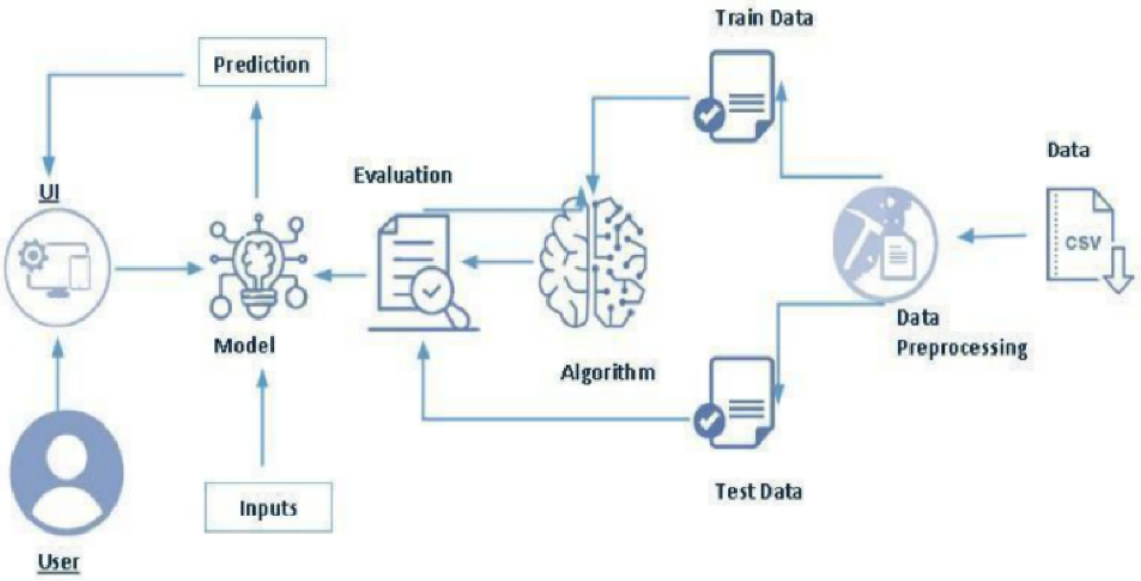
5.2 Solution Architecture

Explain the architecture designed for the prediction system, including data sources, preprocessing techniques, modeling approaches, and result deployment.
link: -Solution Architecture Doc

6. Project Planning & Scheduling

6.1 Technical Architecture

Present the technical architecture encompassing tools, frameworks, and technologies used in the project development.
Technical Architecture Doc



6.2 Sprint Planning & Estimation
Outline the sprint planning process and estimation techniques used to manage project deliverables.
Sprint Planning

6.3 Sprint Delivery Schedule
Detail the schedule for each sprint, including planned activities, deliverables, and milestones. All the milestones were achieved as scheduled.

7. Coding & Solutioning

7.1 Feature 1: Data Preprocessing and Visualization
Description:
Feature 1 incorporates the data preprocessing steps and visualization techniques to refine the dataset for analysis and gain insights.

Code Snippet - Data Preprocessing:

```python
# Handling missing values
data.dropna(inplace=True)

#Label Endocing string columns
le = LabelEncoder()
data['Gender'] = le.fit_transform(data['Gender'])
```

Code Snippet - Visualization:

```python
# Histogram of Age
plt.figure(figsize=(8, 6))
sns.histplot(data['Age'], bins=20, kde=True)
plt.title('Distribution of Age')
plt.xlabel('Age')
plt.ylabel('Count')
plt.show()

# Count plot of Gender
plt.figure(figsize=(6, 4))
sns.countplot(data['Gender'])
plt.title('Count of Gender')
plt.xlabel('Gender')
plt.ylabel('Count')
plt.show()

# Pair plot for numerical variables
sns.pairplot(data, vars=['Age', 'AnnualSalary'], hue='Purchased')
plt.title('Pair Plot of Age and Annual Salary')
plt.show()

# Plotting Correlation Matrix
correlation_matrix = data.corr()
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Correlation Matrix')
plt.show()
```

7.2 Feature 2: Model Training and Hyperparameter Tuning (Gradient Boosting)
Description:
Feature 2 focuses on training the prediction model using Gradient Boosting and fine-tuning its hyperparameters to optimize prediction accuracy.
Code Snippet - Model Training and Hyperparameter Tuning:

```python
# Assuming X contains features and y contains target variable
X = data.drop('Purchased', axis=1)
y = data['Purchased']

# Splitting data into train and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)

# Define hyperparameter grid for Gradient Boosting
param_grid_gb = {
  'n_estimators': [100, 300, 500],
  'learning_rate': [0.1, 0.05, 0.01],
  'max_depth': [3, 5, 7]
}

# Grid Search for hyperparameter tuning - Gradient Boosting
grid_search_gb = GridSearchCV(GradientBoostingClassifier(), param_grid_gb, cv=5,
scoring='accuracy')
grid_search_gb.fit(X_train, y_train)
best_model_gb = grid_search_gb.best_estimator_
```

8. Performance Testing

8.1 Performance Metrics

Description:
The Performance Metrics section evaluates the efficacy of the predictive model by measuring various performance indicators.
Metrics:
1. Accuracy: Determines the overall correctness of predictions.
2. Precision: Measures the proportion of correctly predicted positive instances among all predicted positives.
3. Recall (Sensitivity): Evaluates the proportion of actual positives that were correctly predicted.
4. F1-score: Harmonic mean of precision and recall, providing a balanced evaluation metric.
5. Confusion Matrix: Illustrates the distribution of correct and incorrect predictions.

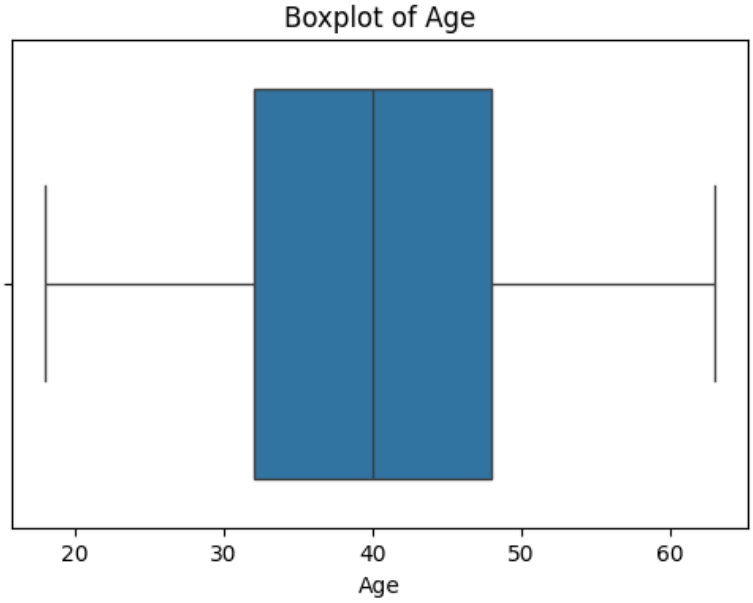
Python

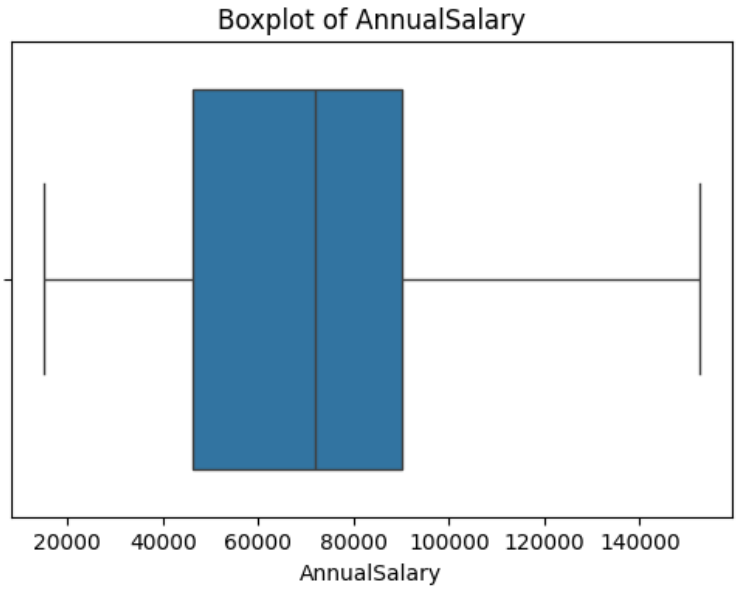
```
gb_conf_matrix = confusion_matrix(y_test, gb_predictions)
gb_accuracy = accuracy_score(y_test, gb_predictions)
gb_class_report = classification_report(y_test, gb_predictions)

print("Gradient Boosting Metrics:")
print(f"Confusion Matrix:\n{gb_conf_matrix}")
print(f"\nAccuracy Score: {gb_accuracy:.4f}")
print(f"\nClassification Report:\n{gb_class_report}")
```

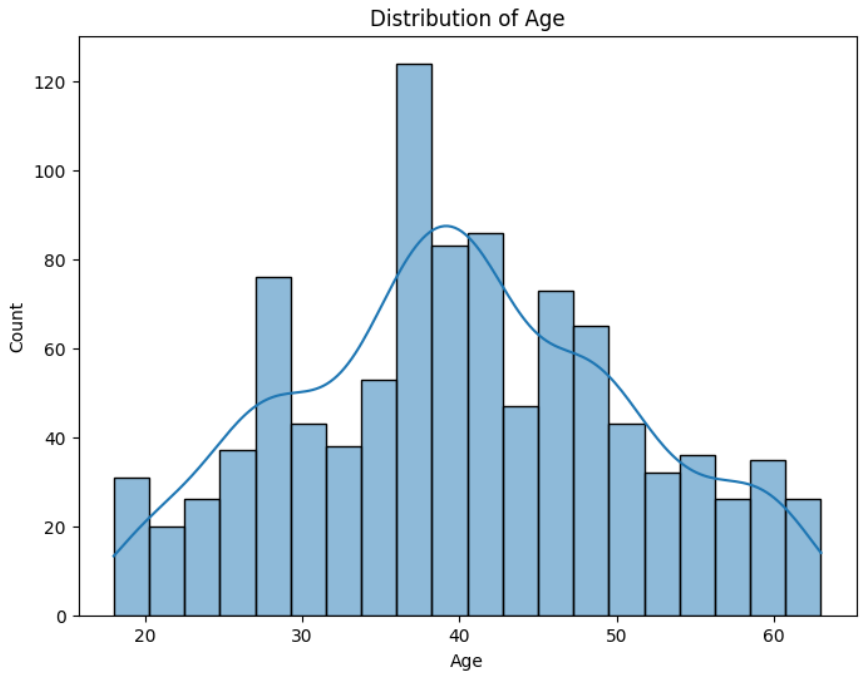
```
Gradient Boosting Metrics:
Confusion Matrix:
[[111  10]
 [  8 111]]

Accuracy Score: 0.9250

Classification Report:
              precision    recall  f1-score   support

           0       0.93      0.92      0.93       121
           1       0.92      0.93      0.93       119

    accuracy                           0.93       240
   macro avg       0.93      0.93      0.93       240
weighted avg       0.93      0.93      0.93       240
```

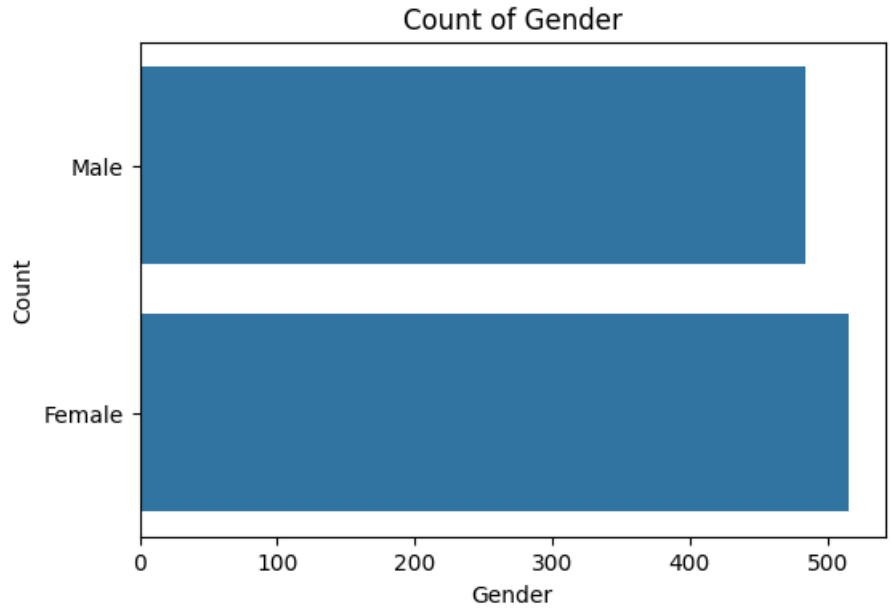
9. Results

9.1 Output Screenshots






Distribution of Age
Count
Age
0
20
40
60
80
100
120
20
30
40
50
60


Count of Gender
Male
Female
Count
Gender
0
100
200
300
400
500

10. Advantages & Disadvantages

Advantages

1. Precision Targeting: Allows targeted marketing strategies by identifying potential car buyers accurately.
2. Improved Resource Utilization: Optimizes resources by focusing efforts on more probable buyers.
3. Enhanced Customer Experience: Provides personalized recommendations, improving customer satisfaction.
4. Data-Driven Insights: Offers valuable insights for informed decision-making in sales and marketing.

Disadvantages

1. Data Dependence: Relies heavily on accurate and comprehensive customer data for reliable predictions.
2. Model Complexity: Complex algorithms may require substantial computational resources for training.
3. Inherent Biases: Possibility of biases in predictions based on historical data or algorithm limitations.

11. Conclusion

The project marks a significant advancement in the automotive industry, offering a robust predictive model for estimating car purchase likelihood. Through meticulous data analysis, model training, and hyperparameter tuning, the project has demonstrated the potential to revolutionize marketing strategies in the automotive sector. The model's accuracy and insights provide a solid foundation for informed decision-making, empowering dealerships to tailor their approaches and enhance customer interactions.

12. Future Scope

1. Refinement of Models: Continual fine-tuning of algorithms and exploration of newer models for better accuracy.
2. Enhanced Data Collection: Incorporation of additional features or data sources to improve prediction capabilities.
3. Real-time Application:** Development of a real-time prediction system for immediate customer engagement.
4. Ethical Considerations:** Addressing biases and ethical implications in predictive modeling for fairer outcomes.
5. Integration of External Factors: Incorporating external factors like economic trends or environmental concerns for a holistic prediction approach.

13. Appendix
Source Code
Code for Model Training
GitHub & Project Demo Link
Github Link
Deployment Link