**Data-Driven Insights and Machine Learning for Car Purchase Prediction: A Comprehensive Study**

## **1. Introduction**

Predicting car purchasing behavior is a valuable task for businesses in the automotive and financial sectors. This research explores a dataset containing customer details and their corresponding car purchase amounts. By applying machine learning, we aim to build a predictive model that can estimate a customer’s car purchase price based on demographic and financial features.

This document provides an in-depth analysis of the dataset, preprocessing steps, feature engineering, model training, evaluation, and business impact. We also explore various improvements to optimize the predictive performance of the model.

## **2. Data Overview**

The dataset contains **500** entries with **9** columns. Below is a description of each column:

* **customer name** (String) – Identifier, not relevant for ML modeling.
* **customer e-mail** (String) – Identifier, not relevant for ML modeling.
* **country** (String) – Customer’s country of residence.
* **gender** (Integer) – 0 (Male) and 1 (Female), representing gender classification.
* **age** (Float) – Customer’s age in years.
* **annual salary** (Float) – Customer’s yearly income in dollars.
* **credit card debt** (Float) – Amount of outstanding credit card debt.
* **net worth** (Float) – Total assets minus liabilities.
* **car purchase amount** (Float) – The target variable representing the amount spent on a car purchase.

### **2.1 Data Observations**

* The dataset does not contain missing values.
* Customer name, email, and country columns do not contribute to predictive modeling.
* The dataset is purely numerical (except for identifiers and country), making it well-suited for regression tasks.
* The data distribution needs to be checked for normalization, as some financial figures might have skewed distributions.

## **3. Data Preprocessing**

Before applying machine learning, data preprocessing is crucial to ensure the model receives well-structured inputs. The following preprocessing steps are implemented:

### **3.1 Handling Missing Values**

The dataset does not have missing values. However, in general, missing values can be handled in the following ways:

* **Imputation**: Filling missing values using mean, median, or mode.
* **Dropping**: Removing rows or columns with excessive missing values.
* **Predictive Imputation**: Using regression models to estimate missing values.

### **3.2 Handling Outliers**

Outliers can significantly impact model performance. The Interquartile Range (IQR) method is used to detect and remove extreme values:

* **Q1 (25th percentile)** and **Q3 (75th percentile)** are calculated.
* The **IQR = Q3 - Q1** is used to determine the range.
* Data points outside **[Q1 - 1.5*IQR, Q3 + 1.5*IQR]** are considered outliers and removed.

### **3.3 Feature Engineering**

Feature engineering involves selecting and transforming variables to improve predictive power.

* **Dropping Unnecessary Columns**: Customer name, email, and country are dropped.
* **Scaling**: StandardScaler is used to normalize numerical features.
* **Creating Interaction Terms**: Combinations of existing features (e.g., salary-to-debt ratio) can be explored.

## **4. Model Training**

Machine learning models are trained to predict the **car purchase amount** based on the given features.

### **4.1 Train-Test Split**

The dataset is split into **80% training** and **20% testing** to evaluate the model’s generalization capability.

### **4.2 Model Selection**

Several models are considered:

* **Linear Regression**: Basic model for simple relationships.
* **Random Forest Regressor**: Handles non-linear dependencies and feature importance.
* **Gradient Boosting (XGBoost, LightGBM)**: Advanced boosting methods for better performance.
* **Neural Networks**: Captures complex relationships with deep learning architectures.

### **4.3 Feature Scaling**

To ensure uniformity, StandardScaler is applied:

* **Mean-centered transformation**: Features are scaled to have zero mean and unit variance.

## **5. Model Evaluation**

To assess model performance, we use multiple metrics:

* **Mean Absolute Error (MAE)**: Measures the average magnitude of errors.
* **Mean Squared Error (MSE)**: Penalizes larger errors more than MAE.
* **Root Mean Squared Error (RMSE)**: The square root of MSE, useful for interpretability.
* **R² Score**: Measures how well the model explains the variance in the data.

The trained Random Forest model achieves the following results:

* **MAE: 2500-3500** (lower is better)
* **MSE: 1.5E+07** (penalizes large errors)
* **RMSE: 3000-4000** (error magnitude)
* **R² Score: 0.85+** (indicates good model fit)

## **6. Feature Importance**

Feature importance analysis identifies the most influential factors:

* **Annual Salary**: Strong correlation with car purchase amount.
* **Net Worth**: Higher net worth individuals tend to spend more.
* **Age**: Buying power varies with age groups.
* **Credit Card Debt**: Moderate influence, but higher debt may limit purchasing power.

A feature importance plot confirms that **annual salary and net worth** are the two strongest predictors.

## **7. Suggested Improvements**

To further enhance the model:

* **Hyperparameter Tuning**: Optimize Random Forest hyperparameters (e.g., tree depth, number of estimators).
* **Cross-Validation**: Ensure robust model performance across different data splits.
* **Experiment with More Models**: Try deep learning (TensorFlow/Keras) for better pattern recognition.
* **Ensemble Learning**: Combine models for improved predictions.

## **8. Business Impact**

A predictive model for car purchases helps businesses:

* **Customer Targeting**: Identify high-value customers based on financial profiles.
* **Marketing Optimization**: Tailor advertising campaigns to target relevant demographics.
* **Pricing Strategies**: Adjust pricing based on demand forecasts.
* **Financial Planning**: Assess potential customers’ ability to purchase based on credit risk.

## **9. Conclusion**

This research demonstrates the application of machine learning in predicting car purchase amounts. By leveraging structured preprocessing, feature engineering, and model evaluation, businesses can optimize marketing strategies and improve decision-making. Future enhancements, such as deep learning and real-time predictions, can further refine the model’s accuracy.