**MACHINE LEARNING PROJECT ON CAR PURCHASING DATA**

**Data Description:**

This dataset tells us about people who might buy cars. We know their name, email, country, gender, age, salary, debt, net worth, and if they bought a car. This information helps us understand who's more likely to buy cars so that we can focus our marketing.

The dataset provides the following attributes for each potential customer:

**Customer Information:**

* **Customer Name:** The full name of the customer.
* **Customer e-mail:** The customer's primary email address, used for communication and identification.

**Personal Information:**

* **Country:** The country where the customer resides. This information is used to understand regional preferences and market trends.
* **Gender:** The customer's gender identity. This data can be used to analyse gender-based differences in purchasing behaviour.
* **Age:** The age of the customer. Age is often a significant factor in determining car preferences and purchasing power.

**Financial Information:**

* **Annual Salary:** The annual income of the customer. This is a crucial factor in assessing the customer's affordability and purchasing power.
* **Credit Card Debt:** The total amount of outstanding credit card debt. High levels of debt can indicate financial strain and may influence a customer's ability to obtain financing for a car purchase.
* **Net Worth:** The customer's total assets minus their total liabilities. Net worth provides a comprehensive view of the customer's financial health and can be used to assess their overall purchasing power.
* **Car Purchase Amount:** Target Value

**Problem Statement:**

Problem Definition: Given a dataset containing information about potential car buyers (customer name, email, country, gender, age, annual salary, credit card debt, net worth), predict whether or not a customer will purchase a car.

**Target Variable:** The target variable is "Car Purchase Amount." However, since we are interested in predicting whether a customer will purchase a car, we can transform this into a binary classification problem. We can create a new binary variable indicating whether a car purchase was made (1) or not (0).

**Machine Learning Task:** The machine learning task is a binary classification problem. The goal is to build a model that can accurately predict whether a potential customer will purchase a car based on the provided features.

**All the preprocessing steps:**

**1. Importing Necessary Libraries:**

Import essential Python libraries like pandas, NumPy, matplotlib, seaborn, and sklearn for data manipulation, visualization, and machine learning.

**2. Loading the Data:**

Load your car price dataset into a Pandas Data Frame

**3. Knowing about the data:**

Get a basic understanding of the dataset, including its shape, data types, and summary statistics.

**4. Data Preprocessing and EDA:**

Handle missing values, outliers, and inconsistencies in the data.

Explore the data through exploratory data analysis (EDA) to identify patterns, relationships, and potential issues.

**5. Finding Correlation Between Independent Variables:**

Identify any strong correlations between independent variables to avoid multicollinearity issues.

**6. Checking Outliers:**

Detect and handle outliers that might skew the model's performance.

**7. Features Selection:**

Selecting appropriate columns to be fitted in the model

**8. Doing Some Visualization in Data:**

Create visualizations (e.g., histograms, scatter plots, box plots) to understand the distribution of features and their relationships with the target variable (car price).

**9. Using Train-Test Split in Data for Modelling:**

Divide the dataset into training and testing sets to evaluate the model's performance on unseen data.

**10. Fitting the Random Forest Model in Data for Prediction of Car Price:**

Create a Random Forest model and train it on the training data.

**11. Using Some Statistics Like R2 and Accuracy of Prediction Data:**

Evaluate the model's performance using metrics like R-squared, Random Forest. These metrics help assess the model's accuracy and generalization ability.

**What model(s) applied and why:**

We used the model Random Forest in my car price prediction dataset. There are some reasons to use the Random Forest model in this dataset.

Car prices are often influenced by various factors, such as Annual Salary, Credit Card Debt, and Net Worth. Random forest model fits in hierarchy way to get the best Results. While this assumption might not be perfectly accurate in all cases, it can provide a reasonable approximation.

**Result(evaluation) analysis:**

I used Random Forest for statical measure I used R-square for accuracy.

**Random Forest CV R²:** 0.9427800840300996

**Random Forest Test R²:** 0.9526113402846064

**Google Colab notebook link:**

[**https://colab.research.google.com/drive/1oYTXs-TwBzFk6SothIvwGHc2Ku9EDrOH?usp=sharing**](https://colab.research.google.com/drive/1oYTXs-TwBzFk6SothIvwGHc2Ku9EDrOH?usp=sharing)

---------------------------------------------------------------- x ------------------------------------------------------------------

**from:** Nikita Kumari CDS 6th Sem  **To:** Priya Sharma Mam

Parul CDS 5th Sem