Car Price Prediction Using Machine Learning

Name: khorn kimkhimmengkheang
Date: Tue 22 Apr

1/ Abstract

This project focuses on predicting the selling price of used cars using machine learning. It uses a dataset that includes various car details like brand, year, fuel type, and kilometers driven. The model helps both buyers and sellers by giving them a fair price estimate based on past data. This way, users can make better decisions when buying or selling used cars.

2/ Problem Statement

People often don't know how much a used car should cost. Car prices can change a lot depending on how old the car is, what type of fuel it uses, how far it has been driven, and other features. Because of this, people sometimes pay too much or sell their cars for too little. This project builds a model that can predict car prices using past data.

3/ Objectives

- Understand the dataset and clean it.
- Analyze the data through graphs and visuals.
- Build a model to predict car prices.
- Save the cleaned data for future use.

4/ Tools and Technologies Used

- **Programming Language:** Python
- **Libraries:** pandas, numpy, matplotlib, seaborn, scikit-learn
- **Environment:** Google Colab / Jupyter Notebook

5/ Dataset Description

- The dataset is taken from a CSV file(kaggle): `CAR DETAILS FROM CAR DEKHO.csv`
- Important columns: `name`, `year`, `selling_price`, `km_driven`, `fuel`, `seller_type`, `transmission`, `owner`
- The dataset was checked for duplicates and missing values.

6/ Data Preprocessing

- Loaded the dataset using pandas.
- Displayed the first 30 rows to understand the structure.
- Checked the column names and data types.
- Removed duplicate rows to avoid repeated data.
- Dropped rows with missing values in important columns like `selling_price` and `km_driven`.

- Removed outliers by keeping only the cars below the 99th percentile in `selling_price` and `km_driven`.
- Removed extra white spaces from column names.

7/ Exploratory Data Analysis (EDA)

- **Histogram:** Showed the distribution of car prices using a blue-colored bar chart.
- **Box Plot:** Helped to check for outliers in the `km_driven` column.
- **Scatter Plot:** Plotted `selling_price` against `km_driven` to understand their relationship.

8/ Model Building

- In this stage, the cleaned dataset can be used to train machine learning models such as Linear Regression or Decision Tree.
- The data would be split into training and testing sets.
- The model would be evaluated based on accuracy or error metrics (e.g., R2 score).

9/ Results

- The dataset was cleaned and visualized successfully.
- Outliers and duplicates were removed.
- Data is ready for model training and prediction.
- Cleaned data was saved as `cleaned_car_dataset.csv`

10/ Conclusion

By using this model, people can get an idea of how much a used car is worth based on real data. This can help buyers avoid paying too much and help sellers set fair prices. The model will be useful in making smarter car deals.