Japan Used Cars Price Prediction Project

---

1. Introduction

Overview: This project aims to predict the prices of used cars listed on tc-v.com, Japan’s largest online used car marketplace. Accurate price predictions can help both buyers and sellers make informed decisions.

Data Source: The dataset consists of various features about the cars, including brand, model, year, mileage, and other relevant attributes, scraped from tc-v.com.

Problem Statement

Cars' data was scraped from tc-v.com and it included Information about Japan's largest online used car marketplace. Ten features were assembled for each car in the dataset

• This dataset includes 10 features:

Feature, Type, Description

1. Price - The sale price of the vehicle in the ad

2. Mark - The brand of car

3. Model - Model of the vehicle

4. Year - The vehicle registration year

5. Mileage - miles traveled by vehicle

6. Engine capacity - The measurement of the total volume of the cylinders in the engine

7. Transmission - The type of gearbox used by the car

8. Drive - Wheel drive(2wd, 4wd and awd)

9. Hand drive - Left-hand traffic (LHT) and right-hand traffic (RHT)

10. Fuel - The type of fuel used by the car(gasoline, diesel, hybrid, LPG, and CNG)

Predict the price of an unknown car.

2. Understanding the Dataset

Features Description:

1. **Price (Integer):** The sale price of the vehicle.
2. **Mark (String):** The brand of the car (e.g., Toyota, Nissan).
3. **Model (String):** The model of the car (e.g., Corolla, Civic).
4. **Year (Integer):** The vehicle's registration year.
5. **Mileage (Integer):** The distance traveled by the vehicle (in kilometers).
6. **Engine Capacity (Integer):** The total volume of the engine's cylinders (in cc).
7. **Transmission (String):** Type of gearbox (e.g., manual, automatic).
8. **Drive (String):** Wheel drive (e.g., 2WD, 4WD).
9. **Hand Drive (String):** Left-hand traffic (LHT) or right-hand traffic (RHT).
10. **Fuel (String):** Fuel type (e.g., gasoline, diesel, hybrid).

Data Cleaning:

- Handled missing values by filling in averages for numerical fields and mode for categorical fields.
- Corrected data types where necessary.
- df.head() - returns the first 5 rows of the Data Frame. This is useful for quickly inspecting the structure and contents of the Data Frame.
- df.reindex() - method is a powerful tool for modifying the index of a DataFrame, allowing for reordering and handling missing values effectively.
- df.drop() - for removing unwanted rows or columns from a DataFrame, allowing for effective data manipulation and cleaning.
- df.shape - attribute is a simple and effective way to get the dimensions of a DataFrame, providing valuable information about the size and structure of the data.
- df.columns - attribute is a straightforward and effective way to access the column labels of a DataFrame, providing essential information for data analysis and manipulation.
- df.describe() - method for generating descriptive statistics of a DataFrame, providing valuable insights into the data's structure and characteristics.
- df.info() - method is a concise and informative way to get a summary of a DataFrame, providing essential information about its structure, data types, and memory usage.
- df.isnull().sum() - is used to check for missing (null) values in a DataFrame and to count the number of null values in each column.
- df.duplicated().sum() - is used to identify and count the number of duplicate rows in a DataFrame.

3. Data Pre-processing

Handling Missing Values:

- Imputed missing values in the ‘price’ and ‘mileage’ columns using the mean values of similar cars.

Encoding Categorical Variables:

- Applied One-Hot Encoding to categorical features like ‘Mark’, ‘Model’, and ‘Fuel’.

Feature Scaling:

- Used Min-Max Scaling to normalize features such as mileage and engine capacity to a 0-1 range.

4. Exploratory Data Analysis (EDA)

Feature Analysis:

- Used histograms to visualize the distribution of prices, years, and mileage.
- Box plots helped identify outliers in features like mileage and engine capacity.

Univariate Analysis:

- The term univariate analysis refers to the analysis of one variable prefix “uni” means “one.” Univariate analysis aims to understand the distribution of values for a single variable.

Multivariate Analysis:

- Multivariate analysis (MVA) is a statistical technique used to analyze and interpret data involving multiple variables. It involves the simultaneous observation and analysis of more than one statistical outcome variable at a time.

Variance Inflation Factor (VIF):

- The Variance Inflation Factor (VIF) is a crucial tool in regression analysis for identifying multicollinearity among independent variables, allowing researchers to make informed decisions about model specification and variable selection to ensure reliable statistical inference.

Correlation Analysis:

- A correlation matrix was generated to examine relationships between features. Strong correlations were found between the car's year and price.

Outlier Detection:

- Outliers in mileage and price were identified and either removed or treated to avoid skewing the model.

5. Model Selection

Choosing the Right Model:

- Several regression models were tested, including Linear Regression, Lasso, Ridge, Random Forest, K Nearest Neighbours, and Decision Tree. Based on R2 scores, performed best.

Model Training:

- The training process involved splitting the data into training and testing sets (80/20 split). The model was trained on the training set using the algorithm.

6. Model Evaluation

Metrics Used:

- Evaluated the model using Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE).

Results Interpretation:

- The final model achieved a Random Forest of indicating a reasonable prediction accuracy.

7. Model Deployment

Exporting the Model:

- The trained model was exported as a pickle file (model.pkl) for deployment.

8. Future Work

Deployment Strategy:

- A Flask web application was created to allow users to input car features and get price predictions in real-time.

Model Improvements:

- Incorporating more features like car color and previous ownership history could improve the model’s accuracy.

Feature Engineering:

- Creating new features such as price depreciation rates based on the year could enhance predictions.

Scaling the Model:

- The model could be scaled to handle a larger dataset by using cloud-based solutions.

9. Conclusion

Summary

- The project successfully built a model that predicts the prices of used cars with good accuracy. The deployment of this model can be beneficial to the automotive marketplace in Japan.

Business Impact

- This model can help users make more informed buying decisions and help sellers price their vehicles competitively.

10. References

- tc-v.com: Data source for the used car information.
- Python Libraries: Used libraries such as Pandas, NumPy, Matplotlib, Scikit-learn, and Pickle