**Used Cars Price Prediction**

Team Name: Hello World

Team Members: Xiwen Hao 1518122

**Introduction**

**Background**

The used cars market is becoming more and more prosperous, with more and more transactions. The characteristics of the used cars market itself are that buyers and sellers are in an asymmetric information structure. The traditional pricing method is based on asset evaluation to evaluate the used cars price, which is greatly affected by subjective factors and not accurate enough and has not yet formed a set of recognized and reliable used cars price evaluation system. Therefore, a more scientific and accurate valuation model is urgently needed.

**Importance**

The importance is finding a suitable model based on the dataset that can effectively predict the used car’s price, which could help dealers to standardize the pricing of used cars and help customers to choose the appropriate used car according to their budget. Through scientific evaluation and prediction, the entire used car buying and selling market will be fairer and more stable.

**Goal**

Our goal is to study and analyze the impact of individual factors on the price of used cars and summarize the relationship between them. Besides, we will preprocess the dataset and execute different models to find the most suitable model with the highest accuracy rate to effectively predict the price of used cars.

**Methodology**

To achieve the goal of our project and find a good model with a higher accuracy rate, we are going to preprocess our raw data and use feature engineering to improve the performance of our models. We divide those tasks into the following steps:

1. **Drop irrelevant columns:** we will manually analyze each column and choose those unnecessary columns, then drop them.
2. **Handle null values:** we will drop all rows that contain null values. Even though we drop all those rows, we still have enough data for us to analyze.
3. **Handle outliers:** we will visualize the distribution of some features and find some data that are unusual in the real world, then drop them.
4. **Data visualization:** we will visualize the relationship between each feature and the price of used cars. Then we will analyze the impact of those features on the price of used cars.
5. **Handle categorical features:** since machine learning algorithms require numerical data and we have some categorical features that are not numerical. We use encode those features into numerical features.
6. **Select appropriate models:** normally, we use regression models when we want to predict a continuous set of values for a given independent variable. We choose the following models to help us with the analysis:
   1. Linear Regression
   2. Lasso Regression
   3. Polynomial Linear Regression
   4. Decision Tree
   5. Random Forest
7. **Conclusion:** compare and analyze the result of those models and find the best model with the highest accuracy rate.

**Description of Dataset**

Our dataset is Used Cars Dataset. The dataset used in this project was downloaded from Kaggle. It was uploaded by Austin Reese. And this data is scraped every few months, so it’s pretty new and worth to analyze it.

The dataset contains 26 columns. It contains all the most relevant information about car sales provided by Craigslist. Those columns are [id, url, region, region\_url, price, year, manufacturer, model, condition, cylinders, fuel, odometer, title\_status, transmission, VIN, drive, size, type, paint\_color, image\_url, description, state, lat, long, posting\_data].

The description of those columns is as follow:

id: Entry ID

url: Listing URL

region: Craigslist region

region\_url: Region URL

price: The price of the used car

year: The production date of the car

Manufacturer: The manufacturer of the car

Model: The model of the car

Condition: The condition of the car

Cylinders: The number of cylinders

Fuel: The fuel type of the car

Odometer: How many miles traveled by the car so far

Title\_status: Title status of the car

Transmission: The transmission of the car

VIN: Vehicle identification number

Drive: Type of drive

Size: Size of vehicle

Type: Generic type of vehicle

paint\_color: The color of vehicle

image\_url: Image URL

description: Listed description of vehicle

state: The state of the car

lat: Latitude of listing

long: Longitude of listing

posting\_date: The date of posting

The dataset contains 458213 rows and the size of the csv file of the dataset is 1.34GB.

download link: [Used Cars Dataset | Kaggle](https://www.kaggle.com/austinreese/craigslist-carstrucks-data)

**Conclusion**

Before predicting the price of used cars with five models, we also analyzed the impact of each feature on the price of used cars and came to the following conclusions:

* Cars with prices ranging from 3000 to 10000 are more popular
* Cars that traveled 70000 to 150000 miles are more popular
* Diesel cars are more expensive than other cars
* Cars equipped with all-wheel drives tend to be more expensive
* Top three states with the most transaction are CA, FL and NY
* The average car price in most states is between 10000 to 15000
* the average price of more popular manufacturers is between 10000 to 15000

Then we start to train our data with five different models. After preprocessing the raw data by dropping all missing values and handling outliers. Then the data was split into 8:2 train to test ratio. Then we construct five models to train our data. And the MSE score, RMSE score, and R2 score of each model are shown in the table:

Table: Results of Models

|  |  |  |  |
| --- | --- | --- | --- |
|  | MSE | RMSE | R2 |
| Linear Regression | 29920631.4776 | 5469.9755 | 0.4914 |
| Lasso Regression | 29920631.5221 | 5469.9755 | 0.4914 |
| Polynomial Linear Regression | 17195299.5673 | 4146.7215 | 0.7077 |
| Decision Tree | 16538524.6618 | 4066.7584 | 0.7189 |
| Random Forest | 8436659.0076 | 2904.5927 | 0.8566 |

From the results obtained from these five models, the Random Forest Regression model produced the best results. Hence it is the regression model selected for predicting the price of used cars.