Mid-Quarter Project Progress Report

Group Member: Paul Ling, Jinzhuang Li, Jintong Li, Alex Ma, Junkai Xiang

Data Description:

The dataset we choose contains information regarding the basic information about the used car. For the column, it includes brand, models, mileage, price, model_year, fuel_type, engine, transmission, ext_col, int_col, accident, clean title, and price. Take a closer look into details of the data, for instance, in the model column, major car brands like Ford, Lexus, and Audi are shown as attributes. And for the color column, it comprises exact features, like Black, Blue, Moonlight Cloud, etc as attributes. Consequently, the dataset should be considered as exhaustive which contains all the essential traits about a specific used car in the market. The dataset also contains 4009 data points in total, for which is a suitable size for training a model for the price evaluation given required characteristics.

EDA:

For the Exploratory Data Analysis, we utilize various sorts of methods. Firstly, we check cardinality for all categorical variables. Some columns contain a large number of attributes. For instance, the “model” column contains 1896 variables, the “engine” column has 1146 variables. We also conduct the missing values check among the columns “fuel_type”, “accident” and “clean_title”. For the “fuel_type” column, we transfer the data points with “missing”. For the “accident” column, the “no reported” variables are replaced with “no”, the “At least 1 accident or damage reported” is transferred to “yes”. This process not only solves the problem of the miss of specific values, but also normalizes the data into the form that we can easily train with. Furthermore, the similar strategy is also implemented for the column “clean_titile”. The feature “accident or damage reported” is substituted with “no”. This is for the sake of making the whole column with only variables “yes” or “no”, for which is increasingly clear and practical in the process of training the model.

In analyzing the dataset, we used box plots on the car data, which came up with many outliers both in mileage and price. To reduce these outliers, we used log and square root transformations. Log seemed to work better on mileage and the square root seemed better for price. Still, a significant number of outliers existed, thus indicating that more steps should be taken. Moving forward, we will decide whether these outliers will have an impact on our analysis to a significant extent or whether we are going to keep them or remove them, considering the trade-off between the completeness of the data set and the reliability of the analysis.

Additionally, there is a huge imbalance problem in our data: threefold higher numbers of

vehicles without any accident compared to those with accident history, and vehicles with a clean title are more than 15 times those that are not. This might make the model biased during the prediction; therefore, we will look at performing either oversampling or undersampling techniques for further steps to keep the insights balanced and accurate.

Feature Selection:

For our used car price prediction model, we prioritized features with a direct impact on vehicle pricing, specifically Year, Mileage, Brand, and Accident History. Year and Mileage were selected as essential indicators of a car's age, condition, and potential resale value—newer cars generally command higher prices due to lower depreciation, while lower mileage suggests reduced wear and is often associated with higher value.We chose Brand over Model to capture brand-based value, as each manufacturer's overall reputation and reliability influence market prices. Model was excluded to keep the model generalizable, because there are over 1,800 unique models in this dataset, which would most likely lead to overfitting. Additionally, Accident History was selected, given that vehicles with past accident records typically see reduced resale prices. In contrast, Fuel Type, Engine, Transmission, and Interior & Exterior Color were deprioritized due to their limited impact on resale price for used cars.

In the second half of this quarter, we plan to reassess our selected features with quantitative methods such as correlation analysis or feature importance scores to evaluate the individual contribution of each feature.

Early Plan for Model Development

1. Model Choices:

Linear Regression: This model will help us find a direct relationship between features like mileage, year, and price.

Polynomial Regression: If data patterns seem nonlinear (e.g., price changes with mileage)polynomial regression will capture those relationships.

Neural Network: To handle complex patterns, we'll use a neural network. This model can analyze multiple factors at once for better predictions..

2. Steps to Build Models:

Data Splitting: We'll divide the data into three parts: training(70%), validation (15%), andLesting(15%) to ensure fair performance testing.

Data Preparation: Mileage and price will be standardized (normalized), and categorical data(like car brand) will be converted into numeric form so the models can use them effectively.

Feature Transformations: Based on our EDA, we'll apply transformations like log-scaling on mileage to minimize the effect of outliers.

3.Tuning Models:

Linear and Polynomial Models: We'll test different polynomial degrees to pick the one that best fits our data.

Neural Network: We'll test different setups for the neural network, adjusting layers and neuroncounts for accuracy.

4.Evaluation:

Metrics: We'll measure model accuracy with Root Mean Squared Error (RMSE) and Mean Absolute Error(MAE)and compare them..

Comparison: We'll create simple charts to show which model predicts car prices best and identify any over- or under-estimalion.

5.Implementation Timeline:

Starting November 4, we'll spend three weeks on building and testing models, leaving time for improvements and preparing our final report.

Literature Review:

It is indispensable for us to learn from the experiences from the past research and take valuable traits from them. From the paper Bukvić, L., et al. (2022) [1], they use the classical process of treating the data from Croatian used cars market, by first doing EDA to get rid of the outliers. Next, the researchers did carefully feature selection. They didn't choose those columns with high cardinality. Hence, they mainly measure the relationship of mileage, brand, and years with price. The models primarily utilized by the researchers are matlab linear regression and random forest, which is a model that allows using multiple columns to generate a single model.

Since we have already reached the Artificial Neural Network in the quarter, we should also consider the usage of ANN into our training. Based on Zhou, X., et al. (2020) [2], BP (Back-Propagation) neural network is an ideal choice for training the model, given that our project has multiple columns. Furthermore, BPNN also doesn't rely on empirical formulas and can create rules based on the existing data to elaborate the patterns of data.

Difficulties Encountered:

The value we are trying to predict, the response variable "price" still has a large amount of outliers even after using log-transformation. In the future, we have to decide how to handle these outliers carefully, either keep them and train a model that is more robust to noise, or delete these corresponding observations. Meanwhile, we observed imbalance data in all categorical variables. Oversampling, undersampling, or other methods should be done to balance the dataset. And the last difficulty we are facing right now is we kept the variable "brand" despite its high cardinality feature, because it is considered as one of the most important features of car price in reality. The balance between the computation complexity and model accuracy is a matter that we should evaluate further. .

Team member contribution:

- Jintong Li: EDA and feature selection codes, writing of difficulties encountered.
- Paul Ling: Feature selection, project management
- Junkai Xiang: The writing of Data description, EDA, and Literature Review.
- Jinzhuang Li: EDA, project starting codes, project management
- Alex Ma: The writing of Early Plan for Model Development

Reference

1. Bukvić L, Pašagić Škrinjar J, Fratrović T, Abramović B. Price Prediction and Classification of Used-Vehicles Using Supervised Machine Learning. *Sustainability*. 2022; 14(24):17034. https://doi.org/10.3390/su142417034
2. Zhou, X. (2020). The usage of artificial intelligence in the commodity house price evaluation model. *Journal of Ambient Intelligence and Humanized Computing*. https://doi.org/10.1007/s12652-019-01616-4