Predicting Used Car Price

Liyang Du (ld477), Mi Zhou (mz558)

Abstract

A sophisticated used car value estimator is a catalyst for the healthy development of used car market. Though numerous studies of vehicle price prediction based on a variety of machine learning algorithms have been well developed, few focuses on the comparison of different algorithms in pricing estimation. In this project we collect roughly 550,313 used car dealing records for empirical analysis on a thorough comparison of two algorithms: multiple linear regression and Adaboost. Current results show that though Adaboost has a lower coefficient of determination (adjusted R^2), it shows great advantage in the model including massive features compared with multiple linear regression. This indicates that Adaboost could be a better algorithm when handling models with a large number of features, yet multiple linear regression shows obvious advantage when dealing with models with well selected features.

1 Dataset Description

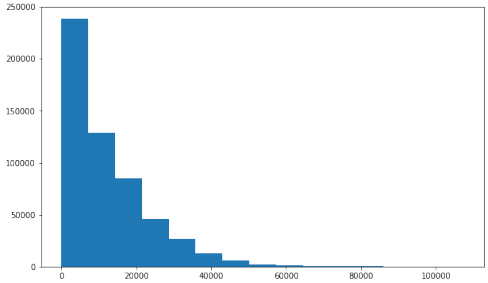
1.1 Data Characteristics

The dataset was scraped from Craigslist. It has 550313 rows and 22 features on used car sales, containing most all relevant information on car sales including price, condition, manufacturer, latitude/longitude, and 16 other categories. The output that we wish to predict is the price of the used car. Out of the 22 features, there are 3 url type features (url, city url, image url), 5 numeric type featuers (price, year, odometer, latitude, longitude), 13 categorical type features (city, manufacturer, make, condition, cylinders, fuel, title status, transmission, VIN, drive, size, type, paint color), and 1 text features (desc - meaning description). We hope that such a large dataset will lend itself to the creation of a state-of-the-art pricing model.

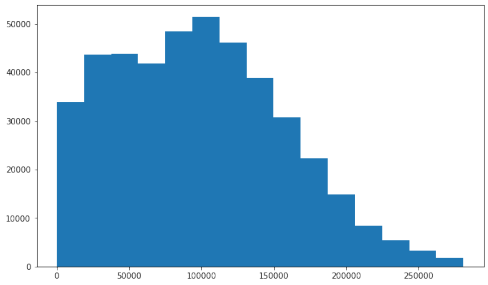
Out of the 22 features, there are 17 features having missing values except for city, price, url, city url, image url. The features that have missing ratio higher than 20% are condition, cylinders, odometer, VIN, drive, size, type, paint color. The count of missing values are: year: 1487, manufacturer: 26915, make: 9677, condition: 250074, cylinders: 218997, fuel: 4741, odometer: 110800, title status: 4024, transmission: 4055, VIN: 239238, drive: 165838, size: 366256, type: 159179, paint color: 180021, latitude: 11790, longitude: 11790.

1.2 Data Visualization

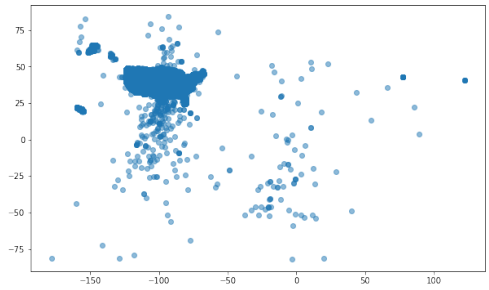
We visualize some featurues about our data. First we look at the price histogram. The price is extremely heavy-tailed: some cars have extremely high prices. So we plot the histogram of the price of cars with the lowest 99.9% price. Also note that there are around 600 cars that have extremely price above 99.9% of the cars.



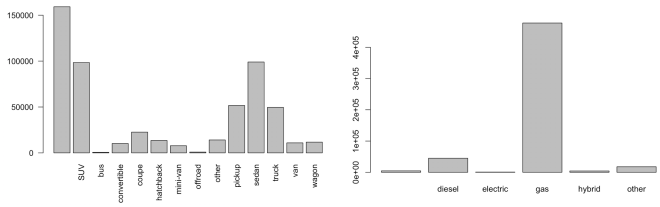
Then we look at the odometer, which we consider is an important feature in determining the price of the used cars. After dropping NA values, the odometer is also extremely heavy-tailed and we only plot the histogram of the lowest 99% odometers.



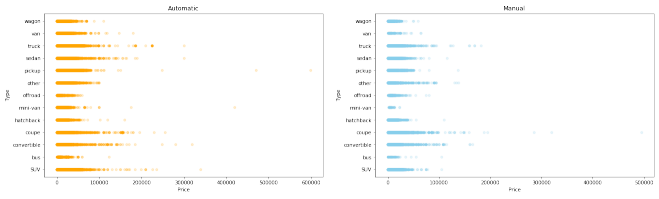
Another interesting thing in this dataset is that it provides the longitude and latitude of every trade. Since most price predictions do not consider using these features, let us have a look of the geographical property of the dataset.



Then let us look at some categorical features. Since car types and fuels might be a good impact on the price, we look at the fequency plot of the types and fuels. Note that the most left column stands for NA values.



We might also look at some joint distribution of price and some features. This way we can have an intuition of how one feature is correlated with price. So we look at the joint distribution of the price and car types under 3 different transmissions (automatic, manual, other).



1.3 Data Cleaning

In this part we illustrate our approach in data cleaning and preprocessing. The first step we adopt is dropping the 3 url type features (url, city url, image url) because in our models the url features are hard to use. Also we drop the text feature (desc) because it is also hard to use in the model. Now we have 18 features: 1 price to be predicted by the other 17 features.

The second step is dealing with missing values. If we have features that have very high missing ratios (e.g. 90%), then we should drop these features because of the high missing ratios. But in this dataset the highest missing ratio is around 60%, so we rather consider filling missing values. One possible way is filling with 0 (numeric) and NA (categorical). But we think this way it introduces more bias than filling with most frequent values or mean values. So, we fill missing with feature mean (numeric) or most frequent value (categorical).

1.4 Modelling Logistics

When fitting different models, we may encounter over-fitting and under-fitting. We also want to test the effectiveness of the models. So we illustrate the whole steps that we will take.

After choosing a model (linear regression, Adaboost, etc), we conduct the train/test split, making 80% of the data into training data and the remaining 20% test data. We further split the training data into 5 folds to conduct 5-fold cross-validation.

Then we train the model on the training data using 5-fold cross-validation, to get the parameters and hyper-parameters: The final model we choose will have the lowest average error among the 5 folds. Finally we test the model on the test data to see the predicting performance of the model.

Then we can compare the training error and test error. If both errors are low, then we are good. If both errors are high, then we are facing under-fitting. We should consider increasing the model complexity and using smaller penalty parameters. If the training error is low and test error is high, then we are facing over-fitting. We should consider decreasing the model complexity, using fewer features and using larger penalty parameters. Then when we finally achieve the relatively-best model, the test error and adjusted R^2 is the measure of the effectiveness of the model we develop.

2 Model Analysis

2.1 Model Selection

2.1.1 Multiple Linear Regression

We first applied multiple linear regression to predict the used car price as we assumed price is linearly dependent on various explanatory features such as millage, car type, maker and etc. The objective function is de-

fined as:

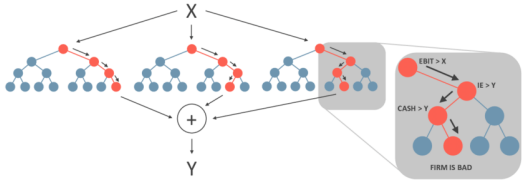
$$p_i = \sum_{j=0}^{d} x_{ij}\beta_j + e_i, \quad i = 0, 1, 2, ..., n \tag{1}$$

$$E(e_i|\{x_n\}) = 0 \quad and \quad E(e_i^2|\{x_n\}) = \sigma^2 \tag{2}$$

where p_i is the price whose expected value depends on the covariates x_i and e_i represents the error terms and is assumed to be independent between observations. σ is unknown and usually the covariate x_0 is a constant 1 and β_0 is the intercept. In the multiple linear regression model above, the parameters β_i and the variance σ^2 are to be estimated.

2.1.2 Adaboost

We also applied Adaboost for the used car price prediction. We used random forests as weak learners and convert the classification model into the equivalent regression model. The accuracy of weak learners is better than random guessing for arbitrary training data and able to handle weighted training samples. Given these two properties of weak learners, AdaBoost provides a framework to combine these weak learners to obtain a final regression model whose accuracy is significantly higher than the accuracy of any single weak learner. In each iteration, AdaBoost attempts to improve upon its errors for particular examples in the training set by minimizing the errors for those in the previous model.



Though individual training might create over-fittings, they are overcome by averaging out the predictions of individual trees with a goal to reduce the variance and ensure consistency.

2.2 Feature Selection

At the start of our experiment, we chose 15 features from the total 21 features because 4 features are of images which we are not currently able to utilize and another two features are the longitude and latitude of the dealing locations which we consider are entirely irrelevant features to the used car price. When testing the performance of our multiple linear regression model, we also calculate the p-value for every single feature and pick out those features whose p-values ≤ 0.05 as the most relevant features and feature with p-values $\in (0.05, 0.1]$ as relevant features.

3 Experiment Results

To test the feasibility of both models, we calculate the adjusted R^2 to measure our model performance. We found that multiple linear regression predicts with an adjusted R^2 0.5921, and a root MSE 8144.379 using all features excluding latitudes and longitude, which indicates that the multiple linear regression model explains 59.21% of the total variance on the validation set. On the other hand, Adaboost predicts with a root MSE 11099.49.

4 Next Steps

1. To further investigate the feasibility of machine learning techniques on used car pricing, we will look to apply more advanced techniques such as LSTM, fuzzy logic and genetic algorithms.
2. To further prevent our models from over-fitting, we are planning to apply more sophisticated regularization techniques and thus we require to normalize some of our features since they are no longer invariant under scaling.
3. In order to ensure the generality of our model, we are going to test our models' functionality with various independent datasets, i.e. we will extend the testing data with eBay and OLX used cars data sets and validate our proposed models.
4. To further improve the measurement of model accuracy, We will design and test more metrics and loss functions and seek out the ones that are the most suitable for specific learning algorithms.
5. We intend to develop a fully automatic, interactive system that contains a repository of used car information, enabling users to know the prices of similar cars via a recommendation engine.