Car selling price model evaluation

Tianjian Wang

January 30, 2017

1 Introduction

Marketing is a complicated business, and have many different influencing factors. But there are some kinds of commodities which have less influencing factors. Take cars for example, the price of a car depends on say, model, status, year, and several more factors. For car traders, it is extremely important to have a full knowledge of the market. Since cars have a somewhat simple structure of price, I decided to try and construct a model to predict the price depending on the factors I will describe later.

However, I did not find any research paper on it. Perhaps the topic is a bit too profitable, so that people are reluctant to publish it. The LaTeX template I used for this proposal is from `http://prancer.physics.louisville.edu/classes/308/project_proposal/proposal.tex`.

2 Targets

The target of this project is to construct a regression model that could 'predict' the price of a car given the following factors.

- the nature of the car: vehicle type, model, brand, gearbox, power, fuel type
- the characteristic of the car: year, kilometer, any damage not repaired

The resulting model will be evaluated by the minimum squared error of the model's predicted values it produces versus the real values with the held-out test set.

3 Datasets and Inputs

This dataset was discovered on Kaggle, with over 370,000 used cars scraped from Ebay Kleinanzeigen, the Ebay Classifieds of Germany. The Url of the database:

https://www.kaggle.com/orgesleka/used-cars-database .

4 Way of Solution

It is quite intuitive that the place where the car is in have much influence on the car's price. So I'll first clean the data by calculating the sum of car price in each certain district and divide each car's price by the corresponding sum. I guess that number would be too small so maybe do a log and minus on that, so that the car's price would become more balanced. In this way, we could also easily get the original price later.

For the features, I'll replace all string features with integer classes. Then I'll try to apply some unsupervised methods to try to get more insights of the dataset. I'll try to use most of those supervised methods including basic linear regression, decision trees, random forests, support vector machines, 3NN, 5NN and feed-forward neural network, then see which one is doing the best on the dataset.

5 Benchmark Model and Evaluation Metrics

For the benchmark model of this project, I'll be using decision tree model with all features, improve it as much as I can, and getting its error amount for predicting prices. The error will be precisely computed by taking the difference of the predicted value and the actual price, then squaring the difference, summing up through the test set and finally taking average.

So for the held-out test set P, with p_i as the ith real price in the set, $\hat{p}_i$ as the predicted price in the set, and a total number of k prices, the error e_P would be like this:

$$e_P = \sum (p_i - \hat{p}_i)^2 / k$$

This evaluating procedure will be the same for each and every model I'll be using.

6 Project Design

The project will be combining techniques learned from this nanodegree process. It'll start with data transformation and cleaning. Then I'll use unsupervised methods like PCA and clustering to try to acquire some intuitive insight of the data. Finally, I'll train supervised models on the dataset, and test there performance using cross-validation. Meanwhile, I'll try to use more seaborn to visualize my results.