Car selling price model evaluation

Tianjian Wang

February 1, 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, which is remarkably less than some complicated commodities. 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.

Although it is a rather interesting idea, there has not been many research papers done on this. Perhaps the topic is a bit too profitable, so that people are reluctant to publish it. In a short paper [1], the author explored several supervised methods with data extracted from newspaper advertisements. And in a thesis [2], the author explained in more details the whole process of car value prediction, combined with a lot of fundamental statistical knowledge. These were the only two papers that me and the reviewer are capable of finding. The LaTeX template I used for this proposal is from `http://prancer.physics.louisville.edu/classes/308/project_proposal/proposal.tex`.

2 Problem Statement

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.

There are a total of 20 fields for the dataset, some are discrete features and some are continuous.

The discrete features include:

- dateCrawled : when this ad was first crawled, all field-values are taken from this date
- name : ”name” of the car, including the brand and sometimes the sequence number
- seller : either private or dealer
- offerType : either offer or request
- abtest : only two values ’test’ and ’control’, meaning unclear
- vehicleType : one of the total eight categories
- gearbox : either manual or automatic
- model : the car’s model
- fuelType : one of the total seven fuel categories
- brand: the car’s brand
- notRepairedDamage : if the car has a damage which is not repaired yet, either ’yes’ or ’no’
- postalCode : indicating the region the car is from

The continuous features include:

- price : the price on the ad to sell the car
- yearOfRegistration : at which year the car was first registered
- powerPS : power of the car in PS
- kilometer : how many kilometers the car has driven
- monthOfRegistration : at which month the car was first registered
- dateCreated : the date for which the ad at ebay was created
- nrOfPictures : number of pictures in the ad

- lastSeenOnline : when the crawler saw this ad last online

It's interesting to see that some of the features, although in themselves does not make much sense, could be combined together to produce valuable information. For example, 'lastSeenOnline' and 'dateCreated' could be combined to estimate the time it takes for the car to be sold.

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, such as PCA and clustering. 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 examples, 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. I choose mean squared error function because it's a regression problem, and I want to penalize more heavily the predicted results that have a lot of difference from the label value. So the bigger the difference, the much bigger it will be after the square, which makes this error function ideal.

6 Project Design

The project will be combining techniques learned from this nanodegree process. It'll start with data transformation and cleaning. For example, since the dataset is crawled from the ads, there are many examples with one or more of the fields containing nothing. So first step is to get rid of those. Then like mentioned above, I'll calculate the bias for continuous fields and transform them to get more dense. For discrete fields containing more than 2 possible values, I'll replace them with integer fields, which is a standard method.

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. The tuning of parameters will be done by using exhaustive grid search. Meanwhile, I'll try to use more seaborn to visualize my results.

References

[1] Pudaruth, Sameerchand. Predicting the price of used cars using machine learning techniques. Int. J. Inf. Comput. Technol 4.7 (2014): 753-764.

[2] Lin, Yuchen. Auto Car Sales Prediction: A Statistical Study Using Functional Data Analysis and Time Series. Diss. 2015.