Project Midterm Report: Predicting Used Car Price

Yihao Zhu, Xinci Liu, Lirou Luo, Liming Zhao

11/09/2020

1 Introduction

The current pandemic is altering the used-car industry: transactions are shifted away from the dealership lots and largely digitized. It becomes essential for prospect buyers to know how to best utilize the information provided online to judge the value of their prospect investments. Buyers should understand how each feature is factored into the pricing of a used car, and whether their targeting vehicles possess features that would enable them to carry values into the foreseeing future. These days, there are a plenty of car-dealing website that provide details on different types of car models that were listed, yet most data features seem complex and uncorrelated to the general public.

In this project, our goal is to analyze the data we scrapped from https://www.truecar.com/used-cars-for-sale/listings to identify the primary factors influencing the listing price of used cars in the U.S., with the hope of using this information to benefit customers. This website provides general information on about 1 million used cars that were listed, such as brand, model, mileage, price, year of manufacturing, location, ownership and accident records. We will use these features to predict on the listing price of used cars, and these predictions will be validated with the actual prices listed.

It should be noted that it is not practical to analyze all brands and all models, because encoding brands and models will give rise to numerous and redundant features. Therefore, we intend to build models only on several typical and popular cars. Another issue is that we only have access to nearly ten thousand samples at one time so that we do not have adequate data to analyze a particular car model. Therefore, our strategy is to first scrap ten thousand used car samples spreading all kinds of models, from which we are enabled to investigate the popularity of all models. Then, we plan to concentrate on those popular models and scrap those individually. That is, we will have nearly ten thousand samples for each model, which reasonably and greatly enlarges our data sets. In this midterm report, we illustrate the data set processing and featurization, which is the same as subsequent new data sets for individual models.

2 Data Distribution & Processing

The latest available data set on used-cars we could find was on cars listed online in 2017. Since cars are expected to depreciate overtime, we would like to have data on cars listed this year in order to make valid price predictions. Therefore we decided to use the data extraction tool *Bazhuayu* to scrap information on used cars currently listed on the website *Truecar*. We originally obtained 9,122 lines of records and 11 columns on car features. To tidy up the data, we used python to split the columns such as '*additional_description*' that contains more than one feature, and also applied feature transformation such as one-hot encoding for features with set values (see Table 1). We then went through the data and filtered out records that have missing values or were 'N/A'. In this way, we obtained a cleaner data set with information on 7,954 used cars and 15 columns containing one feature each, which is ready to be used to visualize the data distribution. We used bar graph or histogram to plot each non-Boolean feature against count. As seen in Fig.1 below, we are able to identify the most common model type of used cars listed on Truecar, which includes *Toyota*'s 'Camry' and 'Corolla' models, *Nissan*'s 'Sentra' and 'Altima', *Honda*'s 'Civic' etc.. Since there are 368 unique car model types in our data set, we decided that it would be more appropriate to select and only look at the top twenty most common car model types. We also looked at the price

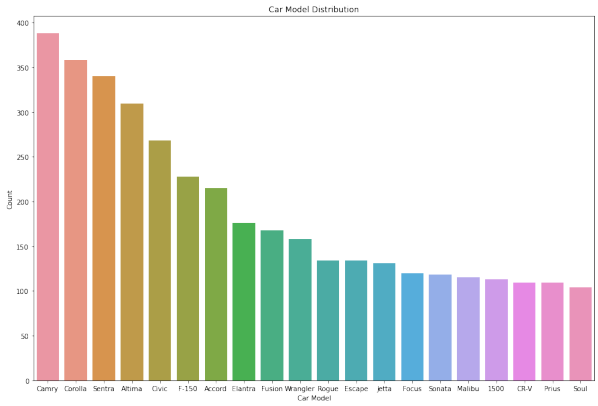

Figure 1: Car Model Type Distribution

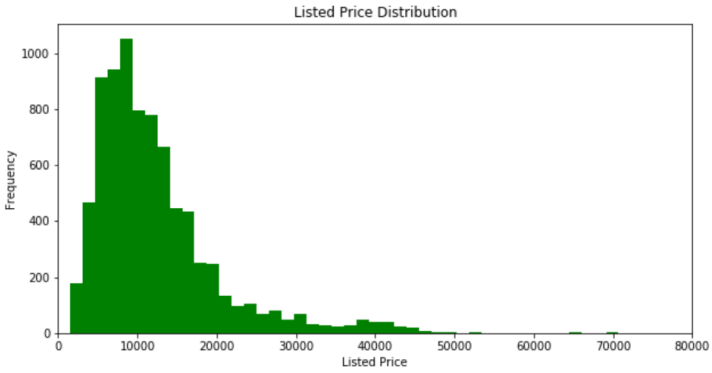

Figure 2: Listed Price Distribution

distribution (Fig.2), and found that most cars are listed for prices ranging from $1,500 to $ 80,000, apart from an outlier of price $158,778. The ipynb file in our repository also contains distributions of Mileage, interior/exterior color and demographic information. With these observations, we decided to scrap another set of data from the *Truecar* website, this time setting conditions on the desired car model types, at the same time making sure that the samples cover a broad enough price range to reflect the population.

Table 1: Feature Summary from the Original Data Set

feature	**type**	**encoding**	**feature**	**type**	**encoding**
make[a]	categorical	NA	state	categorical	NA
model[a]	categorical	NA	color	categorical	NA
year[b]	numerical	NA	value[c]	categorical	NA
mileage[b]	numerical	NA	value description	string	NA
trim level[b]	categorical	ordinal number	URL	string	NA
number of accidents[b]	numerical	NA	picture	graphical	NA
number of owners[b]	numerical	NA	discount[d]	numerical	NA
1 usage[b]	categorical	one-hot encoding			

[a]: For the prediction of a specific car model, *make* and *model* are not useful features.
[b]: The feature is used in the prediction.
[c]: The website evaluated the list price of the cars, e.g., excellent price and fair price, which should not be included in the prediction.
[d]: Discount varies dealer-by-dealer. We only intend to predict the listed price excluding the discount.

3 Feature Selection & Preliminary Model Fitting

Due to the reasons mentioned at the end of the introduction section, preliminary analysis and model fitting were only done on a specific car model. For the purpose of demonstration, and due to the limited amount of data we currently have, we chose to analyze *Toyota Camry* which is the car model that most frequently appeared in our data set. The features we selected are 'mileage','year','number of accidents','number of owners','usage', and our target is 'price'. We didn't select 'trim level', 'color' and 'state' at this time for the following reasons. Feature 'trim level' is extremely messy across different car models, and within the same car model trim levels are represented by different notations every year. Therefore, it is better for us to standardize the trim level for a car model after we acquire the data for that specific model and after that do we take trim level into consideration. As for 'color' and 'state', we don't think they have a significant impact on determining the price, so right now we decided not to include them for a preliminary analysis. However, they may be included for further parameter tuning in the future.

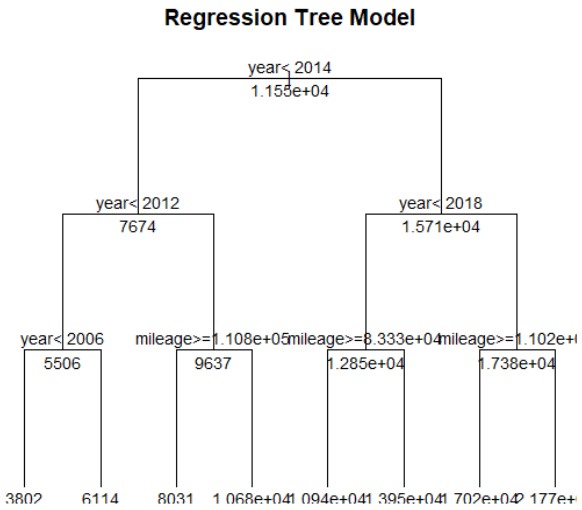

Figure 3: Regression Tree Model

```
> glm_cv5_model
Generalized Linear Model

387 samples
  6 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 309, 308, 310, 311, 310
Resampling results:

  RMSE      Rsquared   MAE
  1768.775  0.8776229  1315.633
```

Figure 4: Result from 5-fold Cross-Validation

The code for preliminary model fitting is written in R due to R having some built-in methods that display some useful statistical values for analysis. (The detailed code is available on our GitHub repository.) For the first part, we employed the old-school way of randomly splitting the data into 80% being in training set and 20% being in test set. Then, we trained a regular linear regression model and a regression tree model using the features mentioned above. We calculated the Root Mean Squared Error(RMSE) as well as the Mean Absolute Error(MAE) for both the training set and test set. The training RMSE for linear regression model is \$1718.641 and the test RMSE is \$1796.956. The training MAE for linear regression model is \$1244.901 and the test MAE is \$1460.87. These results show that the error between training and test set are quite small, which indicates that our linear regression model is not overfitting the data for *Toyota Camry*. As for the regression tree model, the training RMSE is \$1595.082 and the test RMSE is \$1627.595. The training MAE is \$1166.537 and the test MAE is \$1307.465. These numbers also show that our regression tree is not overfitting the data. Another point worth mentioning is that we also plotted our regression tree model and from the plot we discovered that for *Toyota Camry* the most important features are year and mileage. However, feature importance could change after we obtained more data for each car model, and we expect it to vary a lot among different car models. The last thing we did was running a 5-fold cross validation on our Toyota Camry data set using a regular linear regression model. The cross validation showed that the average RMSE is \$1768.775, the average MAE is \$1315.633, and most importantly, the R-squared value is 0.8776229, which means that the five features we selected for our preliminary analysis account for roughly 87.8% of the total variance within our data set. This means that our features have great influence on Toyota Camry's listed price.

4 Future Steps

1. Determine the models to be studied. The metric is that those models are popular and their MSRPs (manufacturer's suggested retail price) are spread across a wide range.

2. Scrap more data from car-dealer website on selected car models (ones that most commonly seen listed on car-dealer websites).

3. After obtaining more new data, clean up the trim level columns using ordinal encoding.

4. Calculate the deprecation rate defined as the deprecation divided by its MSRP, which is the target in the subsequent regressions. It potentially helps to get rid of the impact from different MSRPs.