Second-hand Car Discount Prediction Midterm Report

Xiaopeng Yuan

Wentao Zhang

Yu Zhang

Mingxiang Zhao

**1.Motivation:**

In this project, we are trying to give an object enough second-hand car price to the customer. Most of the price now is given by the car agency themselves and it is influenced by many different variables. There are not open and transparent rules on how to get and calculated the car’s price. And we also know it’s impossible for anybody to make this kind restrictive rules for such a huge and mature market, even the president. So instead of making new rules for the market, we choosing to learn from the market and trying to get some rules from it.

We scrape the latest data from car reviews website and process the reviews using NLP to find out people's’ attitudes towards different brands. It is very interesting, and it’s useful because 80 percent of business-relevant information originates in unstructured form. Therefore, we can dig out hidden information and find some patterns for the brand itself and we will consider using it as one of the variables in our price predicting models.

**2.Background**

Some recent paper already mentions this area, like using machine learning to predict the car price, here we just pick one of them as an example, in the paper ***Predicting the Price of Used Cars using Machine Learning Techniques*** the link of this paper was shown below.(<https://www.ripublication.com/irph/ijict_spl/ijictv4n7spl_17.pdf>) We can find the author just pick up the brand, cinder, volume, year, mileage/km price as variables, but he/she ignore the car itself. Sometimes the car price is also influenced by the car’s body style and the rarity. For example, a 6 Cyl engine Mustang is much easier to get a lower price than an 8 Cyl Mustang because there are too much second hand 6 Cyl engine Mustang in the market. Thus, in our personal opinion, we thought the variable in this paper is not enough. And it also does not consider about if there’s quality problem news being published for this brand how will it influence the car price. Therefore, in our model, we are trying to consider more variables and make the prediction model more accurate.

**3.Approach:**

Linear regression, sentiment analysis, classification.

The main objective of our work is to predict the price of second-hand cars. For instance, when the user provides year, brand and type of an SUV, we can predict the price of it.

**4.Data sources**

Description of data sources:

We gathered data from autos.nj.com. This is a website that you can buy new and used cars in New Jersey. You can find cars here with specific features as you want, such as brand, price, color, mode and so on.

Descriptive statistics of the data: The information we get for each car is as follow: brand, price, engine, transmission, drivetrain, mileage, and location.

Explorative analyses:

There are so many factors can influence the price and some factors even cannot be quantified. So it’s scarcely possible to get a precise function. So we decide to limit the scope of the automobile market. In order to reduce the influence of customers’ psychology for the new model, we decide to focus on used automobile specifically study on a popular model such as SUV. Our goal is to find a method to give a prediction to the price of used SUV in New Jersey. We are not going to provide an accurate price for each kind of SUV, we will offer is a reasonable range instead.

Metric:

PR/per hour: Number of positive reviews per hour on twitter

Features: Price, engine, brand, mileage, drivetrain, location, PR/per hour

**5.Methodology:**

We firstly get data from autos.nj.com and then get data from the Twitter on the basis of data we get from autos.nj.com.

What we get from the website are some key factors such as name, brand, price, engine, and so on. We will format the data and then go to the second part. We write a crawler to get the data we need from the website and codes will be attached to the proposal.

In the second part, we select the data we need from the result. To be specific, we will select some mainstream cars to analyze the variation trend of the cars’ prices. If someone wants to buy one of or a series of these, they may find what they want depends on their budget. In terms of the factors that may influence the price of the cars, we take the year, mileage, certification status, a condition into consideration. What’s more, we come up with an interesting idea, which is searching the comments about a car in twitter and implement sentiment analysis on the comments. (Input of this method is the brand-related tweets, output is the content of the tweets) After that, we calculate the ratio of positive and negative sentiment, this will be a new factor which may influence the estimated price of a car. For instance, when the rate of negative sentiment has reached over 50%, we may deduct the estimated price of a car to 5%, with the rate goes up, the estimated price will be deducted accordingly. When it reaches 70%, we push out a warning or notice that the car is notorious.

In a nutshell, we will utilize machine learning (linear regression) and text mining to implement this project.

**6.Challenges and open questions:**

One challenge is that the data we need from twitter is much less than we expected. For instance, when we search “Honda”, which, in my perspective, is one of the most popular car brands in the world, we just get less than 1 tweet per minute. We felt overwhelmed about this result. After that, we search BMW, which is one of the most prestigious brands in the world. The result is 94 tweets/hour. We felt frustrated again. Finally, we decided to use the time for space(data). We continue to scrape data by using tweepy for a long time. Then we get the data we want.

The problems we met in gathering data from autos.nj.com is that there was too much invalid information, so we need to eliminate them totally. We also needed to set a rational interval in our code so that we won’t be banned by website and at the same time the time cost was acceptable

A question is that how to make use of the data from car forums. We have viewed the car forums you provided for us, but we think it’s tough to use them. Because we saw many meaningless or useless posts. Also, there are some posts related to more than one brand or type of car which may lead to ambiguity.

Although we focus on used SUV market in New Jersey, there are still many interference factors from external such as the preference of engine number or output of a special model. We still need to use universal data for the whole used automobile market and apply to our project appropriately.