Kaggle - Used Cars

Team: Mihkel Kritšmann, Tarvi Tamm, Kris Porovade

https://github.com/krisporovarde/used-cars

Task 2. Business understanding (1 point)

Our data science project's goal is to predict the price of a used car using its features. This project could be useful for car dealers and used cars listing websites, who want to know what the price of a used car could be. More precisely we try to predict the price of used cars in the USA. We see our project being used for example in finding used-cars listings below their value and flipping them.

Right now there are sources that promise a price estimate for used cars (ex. Edmunds). Our model should offer a more accurate price than competitors in the market right now. The project's constraints right now are location, we look at cars only listed in the US market. There could be a risk of error if used in the EU or other regions. Data is collected from an open-source Kaggle dataset, which has scraped used cars from craigslist. In the future our algorithm can be scaled if more data is collected.

We have people with skills in data analysis and machine learning, and also a basic understanding of automobile terminology. Some terminology we will use to differentiate different topics are *vehicle model* and *learning model.* Model will be the specific vehicle model and the learning model will be the machine learning algorithm.

In a nutshell the goal of this project is to predict the price of a used car using its features and finding out which feature affects the price the most. We consider our project a success if the model gets an accuracy of over 0.9 in test and validation sets.

Task 3. Data understanding (2 points)

The data is taken from kaggle used cars dataset. The dataset consists of 426880 rows and 26 columns. The file type is .csv and we are using a jupyter notebook to analyze and train our learning model. The datatypes of the features are mostly objects, with a few integrals and floats. The prediction will be an integral.
Our data columns: 'id', 'url', 'region', 'region_url', 'price', 'year', 'manufacturer', 'model', 'condition', 'cylinders', 'fuel', 'odometer', 'title_status', 'transmission', 'VIN', 'drive', 'size', 'type', 'paint_color', 'image_url', 'description', 'county', 'state', 'lat', 'long', 'Posting_date'.

Some definitions that might cause confusion:

Year - the year when the car was built
VIN - vehicle identification number

Type - car type (ex. sedan)
Lat - latitude
Long - longitude
Description - sellers own description of the vehicle.

From exploring the data we can identify that the majority of instances have some missing values. After we eliminate instances with missing values we are left with 79195 rows of data. Also we see that the vehicle ‘model’ variable is very unclean. There are duplicates and very many outliers. For this we need to use a string metic (ex. jaro-winkler) to group the same models before we encode it for training.

For training we will only use some of the features we considered important variables for the price prediction. Those are as follows - 'price', 'year', 'manufacturer', 'condition', 'cylinders', 'odometer', 'fuel', 'transmission', 'drive', 'type', 'paint_color', 'size','model'
We left out logistical data, also title_status because it's equal to condition. Lastly we left out ‘description’ because this feature is a long string and will make our learning model overfit.

Task 4. Planning your project (0.5 points)

Our detailed plan consists of these tasks:

- Understanding the Data, (estimated time per team member - 3h)
- Preparing the Data, (estimated time per team member - 3h)
- Choosing a Model, (estimated time per team member - 3h)
- Training the Model, (estimated time per team member - 7h)
- Evaluating the Model, (estimated time per team member - 7h)
- Making predictions, (estimated time per team member - 5h)
- Making a poster for the Poster Session, (estimated time per team member - 2h)

The process of understanding the data is already underway, because we have already made ourselves familiar with the data, with what we are going to move forward with.

Next on the list is preparing the data. We need to clean the data and remove any unwanted data, missing values or rows or columns. We will also need to split the data into training and test sets.

Then comes the main work in which we have to choose an appropriate model to train the data on and hopefully get expected results. Since this step is not so straight-forward, it is also given the largest amount of hours to complete the task.

At last we need to make conclusions and decide whether our model was a success or not. In case it wasn’t, we might have to go back to the previous task.

To finish the group project we need to make a poster to present our results and conclusions in the Poster session.