Data Science Challenge

Tuo Sun & Yidan Wang & Xilai Nian

Dataset Description:

For more detailed description, please find the pdf file on Google Drive.

1. **Car_data.json** (use Python crawler to fetch used car data from cars.com)
   The json file contains all used-car information on website like ID code, vehicle basic information, title, price, odometer, all features, and seller info. (each car has its unique ID)
2. **Key_pos.csv**
   This csv provides the key for zip code API. (https://zippopotam.us/)
   It has *ID* and *Position* two attributes extracted from cars_data.json. Use Position zip code to obtain detailed location information (latitude, longitude, place and state name) of the used-car on sale from the API.
3. **Key_vin.csv**
   This csv provides the key for NHTSA VIN API. (https://vpic.nhtsa.dot.gov/api/)
   If there is a need for more detailed car information, the *ID* and *VIN* attributes extracted from cars_data.json give access to obtain vehicle VIN information. The data obtained from the API may have missing values, which needs to be properly pre-processed.

Data Challenges:

1. **Data cleaning**
   The datasets have some useless features and some of them may have missing values. Therefore, it is necessary to pre-process the data like extracting useful features for analysis, dealing with missing values to make the dataset neat and friendly.
2. **Data transformation**
   In car_data.json, there is a “title” attribute which saves the information of cars make, model and license time. The “title” should be processed to multi useful attributes for better analysis.
   In addition, the json format is just for saving raw data. After data process, it’s necessary to convert the data to DataFrame and save as other better format like csv.
3. **Data analysis: Predict the used-car price**
   Once given some features of the used-car, the model can give a predict value. This data analysis requires building a prediction model. Choose suitable features from the dataset, find and train a predictive model that may have a good performance on the dataset. Then, show the effect of the model.