Aaron Lee
009085596
CMPE-256, Summer 2019

Individual Project: Vehicle Recommendation System Report

I. Motivation

Purchasing a brand-new vehicle is a huge investment for people because it is expensive and typically a long-term purchase. Selecting the correct vehicle can be a long and gruesome process with so many different options on the market. Vehicles range from electric to internal combustion engines, roadsters to minivans, and simple self-operated to vehicles with driver's assistance nowadays. There are countless options that people can choose from, so creating a recommendation system for people interested in making a big purchase can make it whole lot easier.

II. Approach

The topic for this project is to develop a recommendation system for people looking into purchasing a new vehicle. This recommendation system will take into consideration the inputs listed below, which are all optional, for what the buyer is looking for, and output a maximum of 10 vehicles that fit into their requirements.

Inputs:

1. Budget (Maximum Price)
2. Car Style (sedan, coupe, etc)
3. Brand
4. Most important Aspects (Choose all that apply):
    a. Comfort
    b. Driving
    c. Interior
    d. Technology
    e. Utility

III. Data

a. Data Collection

For data collection, I was originally going to pull data from edmunds.com using REST API's. However, I found out that they have discontinued allowing the general public use their APIs in 2017, http://edmunds.mashery.com/.

I resorted to collecting data by using Python's *selenium* library and I began to automate the process of going through the web to extract data from their webpages and store it into a csv file. By providing the xpath of the Edmunds HTML page on where all the possible elements are, I'm able to read the data for each feature and store the values into a python dictionary.

For the data of vehicles, I limited it to brand new vehicles ranging from 2019 – 2020 depending on the manufacturer and model. From Edmunds.com, I was able to extract data for 380 different vehicles and able to get overall ratings for the followings features and put it into a python dictionary.

1. Edmund Overall Rating
2. Driving Performance
3. Comfort on the road
4. Interior ergonomics
5. Technology

6. Cargo and Utility

The category section was also found by first creating a list of all possible vehicle types. Using Python's *selenium,* a loop was created to go though the list of vehicle categories and find the term frequency on each of the vehicle's link. If one category had the most, the vehicle would then be categorized as that.

Some problems that I ran into during the data collection process from this website was that the vehicles did not have a standardized template across all 380 vehicles. There were three distinct formats I recognized while scraping through the data. One would have the values listed in a table, the second is listed under a drop-down box, and then the last format would not have any ratings listed under the features. I handled this is the python script in nested try-catch blocks and if there were no ratings listed, I assigned each feature as 'N/A' and added that vehicle into a separate list, for record purposes. There was a lot of trial and error because of this, but after running fixing these problems on a small sample subset of ten vehicle, I was able to collect all the data of the new vehicles listed on their website.

With the raw data, I saved the data into a csv file called 'edmund_ratings.csv'. The data extraction process can be found in the Jupyter Notebook file called '*data_extraction_from_edmunds.ipynb'*.

	Manufacturer	Vehicle	year	category	MSRP(Min)	MSRP(Max)	comfort	driving	interior	overall rating	tech	utility	
0	acura	2019 Acura ILX Sedan	2019	sedan	$25,900	$31,550	NaN	NaN	NaN	NaN	NaN	NaN	https://www.edmunds.com/acura
1	acura	2019 Acura MDX Hybrid	2019	suv	$44,300	$60,050	8.0	8.0	7.5	8.0	7.5	9.0	https://www.edmunds.com/acura/r
2	acura	2019 Acura MDX SUV	2019	suv	$44,300	$60,050	8.0	8.0	7.5	8.0	7.5	9.0	https://www.edmunds.com/acura/r
3	acura	2019 Acura NSX Coupe	2019	coupe	$157,500	$157,500	6.9	8.5	7.5	6.9	5.5	4.0	https://www.edmunds.com/acura/
4	acura	2019 Acura RLX Hybrid	2019	sedan	$54,900	$61,900	NaN	NaN	NaN	NaN	NaN	NaN	https://www.edmunds.com/acura
5	acura	2019 Acura RLX Sedan	2019	sedan	$54,900	$61,900	NaN	NaN	NaN	NaN	NaN	NaN	https://www.edmunds.com/acura

Figure 1: Head of the 6 vehicles in Raw Data Set with Missing Values

b. Data Processing

To prepare and clean this data, the dataset that had missing values needed to be removed. Using the pandas library, the 'pandas.DataFrame.dropna' function was used to remove missing values. These missing values came from vehicles that did not have any ratings given to it on the website.

Another step that needed to be done was to remove the commas and dollar sign to the columns in MSRP(Min) and MSRP(Max). To do this, the python library for regular expressions was used. By iterating over each row in the MSRP column, I would substitute all characters that are not 0 through 9 and replace it with an empty space.

These steps can be found in the Jupyter notebook 'data_processing_and_knowledge-based-modeling.'

This dataset was then saved to a csv called 'post_process_edmund_rating.csv'.

	Manufacturer	Vehicle	year	category	MSRP(Min)	MSRP(Max)	comfort	driving	interior	overall rating	tech	utility	link
1	acura	2019 Acura MDX Hybrid	2019	suv	44300	60050	8.0	8.0	7.5	8.0	7.5	9.0	https://www.edmunds.com/acura/mdx/2019/
2	acura	2019 Acura MDX SUV	2019	suv	44300	60050	8.0	8.0	7.5	8.0	7.5	9.0	https://www.edmunds.com/acura/mdx/2019/
3	acura	2019 Acura NSX Coupe	2019	coupe	157500	157500	6.9	8.5	7.5	6.9	5.5	4.0	https://www.edmunds.com/acura/nsx/2019/
6	acura	2020 Acura RDX SUV	2020	suv	37600	47700	8.0	7.5	7.5	7.9	8.5	8.5	https://www.edmunds.com/acura/rdx/2020/
7	acura	2020 Acura TLX Sedan	2020	sedan	33000	48950	8.0	7.5	8.0	7.5	7.0	7.0	https://www.edmunds.com/acura/tlx/2020/
15	audi	2019 Audi A4 Sedan	2019	sedan	37400	50800	8.2	8.5	8.5	8.2	9.0	8.0	https://www.edmunds.com/audi/a4/2019/

Figure 2: Example of 6 vehicles after removing Missing Values and removing '$' and ','

IV. Implementation

A knowledge-based model will be used to implement this Vehicle Recommendation System.

This system will calculate calculate a utility value, then check and sort the vehicles based off the manufacturer, category, and/or price, and display the top 10 vehicles that fit those requirements.

For Step One, a utility score is calculated based off of the Edmund ratings. With this score, we will multiply the score by a weight and calculate the Utility that will fit the user's requirement.

A web application using Django Web Framework was used. The simple User Interface can be seen below in Figure 3 with multiple user inputs.

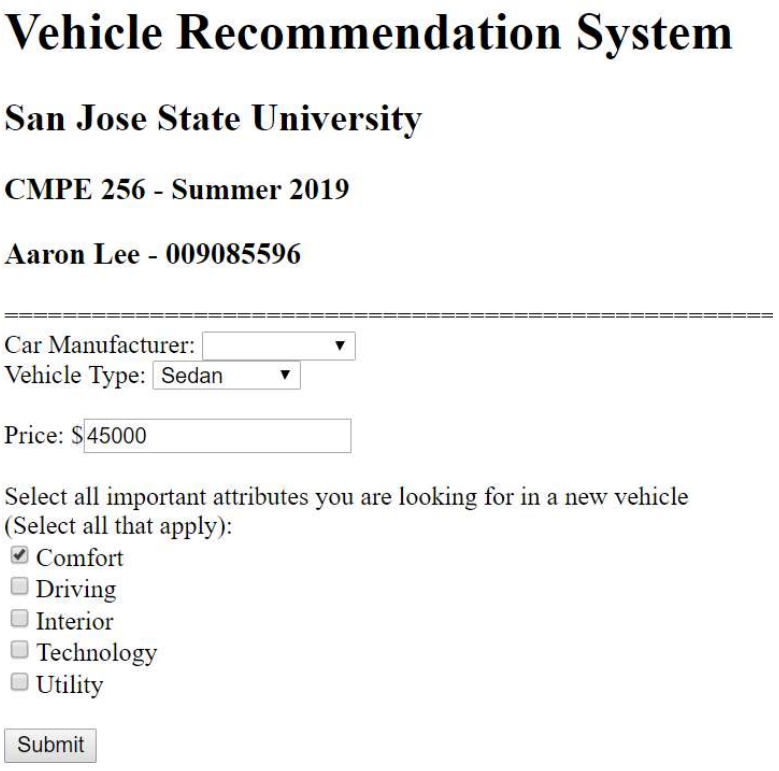

Figure 3: User Interface for Recommendation System

The output of the system will depend on what is submitted on the form. If zero selection is made for the features of comfort, driving, interior, technology, or utility, and submit was clicked, then the average will be taken to calculate the 'Calculated Utility' rating.

	Manufacturer	Vehicle	year	category	MSRP(Min)	MSRP(Max)	comfort	driving	interior	overall rating	tech	utility	link
1	acura	2019 Acura MDX Hybrid	2019	suv	$44,300	$60,050	8.0	8.0	7.5	8.0	7.5	9.0	https://www.edmunds.com/acura/mdx/2019/

Figure 4: Data for Sample Calculation

For example, the utility for the data shown in Figure 4 is:

$$Calculated\ Utility = 8.0 + 8.0 + 7.5 + 7.5 + 9.0 / 5 = 8$$

If a user does select one of the options, then it will be calculated as shown (using the inputs from Figure 3):

$$n = \#\ of\ items\ selected = 2$$
$$weight = 1/n = ½ = 0.5$$
$$Calculated\ Utility = weight * (Comfort) + weight * (Interior)$$
$$Calculated\ Utilty = 4 + 3.725 = 7.725$$

This calculated utility, with and without user selection, will be calculated for all vehicles and a column will be inserted into the pandas data frame.

After calculating the utility for each vehicle, the list will be sorted. By using the example in Figure 3, since the vehicle type and price were inputted, then the data will be sorted based off those requirements as shown in Figure 5.

Recommendations

	Calculated Utility	Manufacturer	Vehicle	year	category	MSRP(Min)	MSRP(Max)	comfort	driving	interior	overall rating	tech	utility	link
89	8.4	honda	2018 Honda Civic Type R Touring	2018	sedan	18940	26800	8.4	8.5	8.0	8.4	8.0	8.5	https://www.edmunds.com/honda/civic/2018/
88	8.4	honda	2018 Honda Civic Si w/Summer Tires	2018	sedan	18940	26800	8.4	8.5	8.0	8.4	8.0	8.5	https://www.edmunds.com/honda/civic/2018/
91	8.3	honda	2019 Honda Accord Sedan	2019	sedan	23720	35950	8.3	8.5	8.0	8.3	8.0	9.0	https://www.edmunds.com/honda/accord/2019/
152	8.2	mazda	2019 Mazda 6 Sedan	2019	sedan	23800	35100	8.2	8.5	8.5	8.2	7.5	8.0	https://www.edmunds.com/mazda/6/2019/
97	8.1	honda	2019 Honda Insight Sedan	2019	sedan	22930	28190	8.1	7.5	8.5	8.1	8.0	8.0	https://www.edmunds.com/honda/insight/2019/
129	8.1	kia	2018 Kia Forte Hatchback	2018	sedan	16800	21700	8.1	8.5	7.5	8.1	7.5	8.5	https://www.edmunds.com/kia/forte/2018/
141	8.1	lexus	2019 Lexus ES 350 Sedan	2019	sedan	39750	44285	8.1	8.0	8.0	8.1	8.0	8.0	https://www.edmunds.com/lexus/es-350/2019/
37	7.9	buick	2019 Buick Regal Sportback Hatchback	2019	sedan	25070	39070	7.9	7.5	7.5	7.9	8.0	9.0	https://www.edmunds.com/buick/regal-sportback/2019/
36	7.9	buick	2019 Buick Regal Sportback GS	2019	sedan	25070	39070	7.9	7.5	7.5	7.9	8.0	9.0	https://www.edmunds.com/buick/regal-sportback/2019/
166	7.8	toyota	2019 Toyota Camry Hybrid Sedan	2019	sedan	28400	32975	7.8	7.5	8.5	7.8	6.5	8.5	https://www.edmunds.com/toyota/camry-hybrid/2019/

Figure 5: Recommendations for Sedan with Best Comfort and Under $45,000

If no selections of the drop-down list were made, then all the vehicle manufacturers and styles will be taken into consideration. This is shown in Figure 6.

Recommendations

	Calculated Utility	Manufacturer	Vehicle	year	category	MSRP(Min)	MSRP(Max)	comfort	driving	interior	overall rating	tech	utility	link
139	9.0	kia	2020 Kia Telluride SUV	2020	suv	31690	43490	9.0	8.0	8.5	8.4	8.0	8.0	https://www.edmunds.com/kia/telluride/2020/
183	8.6	volkswagen	2019 Volkswagen Golf R Hatchback	2019	hatchback	40395	41495	8.6	9.0	9.0	8.6	8.5	8.0	https://www.edmunds.com/volkswagen/golf-r/2019/
27	8.6	bmw	2019 BMW M5 Sedan	2019	sedan	102700	110000	8.6	9.0	8.5	8.6	8.5	8.0	https://www.edmunds.com/bmw/m5/2019/
12	8.5	audi	2019 Audi S4 Sedan	2019	sedan	50200	57800	8.5	9.0	8.5	8.5	8.5	7.5	https://www.edmunds.com/audi/s4/2019/
32	8.4	bmw	2019 BMW i3 Hatchback	2019	hatchback	44450	51500	8.4	9.0	8.0	8.4	8.0	8.0	https://www.edmunds.com/bmw/i3/2019/
181	8.4	volkswagen	2019 Volkswagen Golf GTI Hatchback	2019	hatchback	27595	37095	8.4	7.5	8.5	8.4	8.5	8.5	https://www.edmunds.com/volkswagen/golf-gti/2019/
89	8.4	honda	2018 Honda Civic Type R Touring	2018	sedan	18940	26800	8.4	8.5	8.0	8.4	8.0	8.5	https://www.edmunds.com/honda/civic/2018/
88	8.4	honda	2018 Honda Civic Si w/Summer Tires	2018	sedan	18940	26800	8.4	8.5	8.0	8.4	8.0	8.5	https://www.edmunds.com/honda/civic/2018/
101	8.4	honda	2019 Honda Ridgeline Crew Cab	2019	truck	29990	43420	8.4	8.5	9.0	8.4	8.0	8.5	https://www.edmunds.com/honda/ridgeline/2019/
91	8.3	honda	2019 Honda Accord Sedan	2019	sedan	23720	35950	8.3	8.5	8.0	8.3	8.0	9.0	https://www.edmunds.com/honda/accord/2019/

Figure 6: Recommendations for Vehicles with Highest Comfort Ratings

V. Solution Evaluation

After completing this recommendation system, I found that it can be a useful tool in sorting through the vehicles and generating a recommendation based on users' explicit requirements. One of the drawbacks to this would be that from Edmunds selection, there are vehicles missing that do not have a rating. These vehicles were listed with the n/a values when collecting the data. There are also some vehicle manufacturers that were not listed, such as Tesla.

However, if there was a larger dataset with consistent feature ratings, then I think this could be a much better tool by being able to create a hybrid recommendation system. There were other sites such as Kelly Bluebook and Consumer Reports, but their features and attributes were not the same as the features Edmund uses (and not free).

VI. Impact to Problem

For purchasing a new vehicle, this can be a helpful tool in getting high-level ratings and details all in one table. It gives users an idea of what to expect and how it may compare to other vehicles that are within their price range. This recommendation system can make them aware of other options (not all) that are out there and that they may want to consider for their big long-term purchase.

VII. References

[1] Babu Thomas, L., & Vaidhehi, V. (2018). The Design of Web Based Car Recommendation System using Hybrid Recommender Algorithm. *International Journal of Engineering & Technology, 7*(3.4), 192-196. doi:http://dx.doi.org/10.14419/ijet.v7i3.4.16772