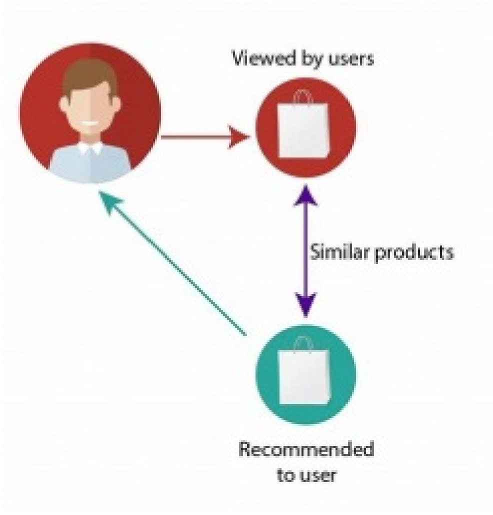
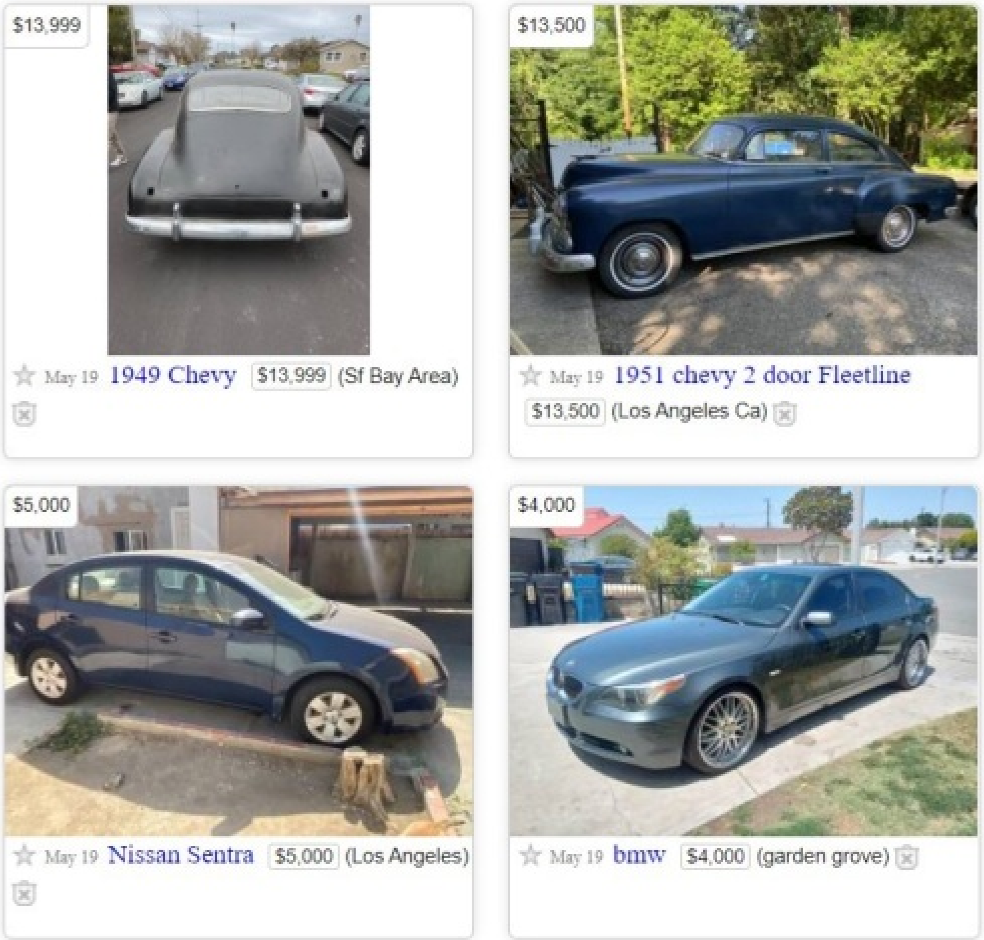
Craigslist’s Used Car Recommendation and Price Prediction Model

Introduction | Presentation structure | Problem summary | Technical approach | Executive summary | Handling missing data and outliers | Content-based recommender

BUILDING A RECOMMENDATION AND PRICE PREDICTION SYSTEM FOR CRAIGSLIST’S USED CARS BUYING AND SELLING PLATFORM




$13,999
☆ May 19 1949 Chevy $13,999 (Sf Bay Area)

$13,500
☆ May 19 1951 chevy 2 door Fleetline $13,500 (Los Angeles Ca)

$5,000
☆ May 19 Nissan Sentra $5,000 (Los Angeles)

$4,000
☆ May 19 bmw $4,000 (garden grove)

BY

Olisa Uzondu

Data Source: https://www.kaggle.com/datasets/austinreese/craigslist-carstrucks-data
References::

Olisa Uzondu

Craigslist’s Used Car Recommendation and Price Prediction Model

BUILDING A RECOMMENDATION AND PRICE PREDICTION SYSTEM FOR CRAIGSLIST’S USED CARS BUYING AND SELLING PLATFORM

CONTENT

1. Problem Summary

2. Technical Approach

3. Executive Summary

4. Supporting Summary 1, 2, 3, 4

5. Conclusion & Recommendations

Data Source: https://www.kaggle.com/datasets/austinreese/craigslist-carstrucks-data
References::

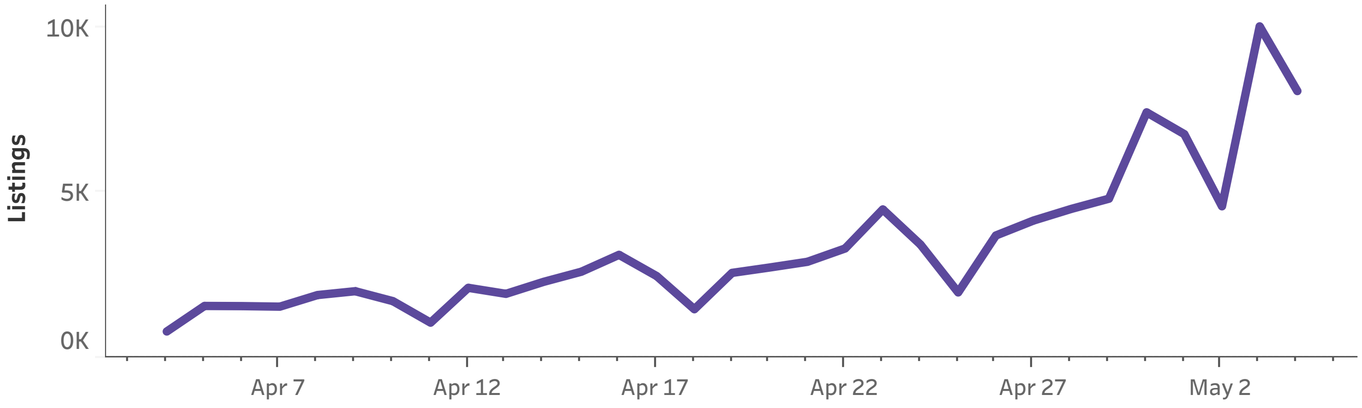
Olisa Uzondu

Craigslist’s Used Car Recommendation and Price Prediction Model

Rising used car listings (~300% between April and May 2021) as a result of **high demand for used cars**, have placed upward **pressure on Craigslist’s ability to increase sales** requiring the inclusion of a **recommendation and a price prediction system.**

The attractiveness of used cars have caused an increase in used car demand	With the current COVID-19 situation, more people are unwilling to rideshare, and so want a personal car for themselves, increasing used car demand
Increase in car demand has facilitated increase in car listings from April to May 2021	With increased demand, more cars are being listed on Craigslist for sale
With used car listings on the rise, more pressure is put on the platform to increase sales	If listings continue to rise at this rate without an improvement to sales strategy, buyers and sellers will switch to craigslist competitors

USED CAR LISTINGS



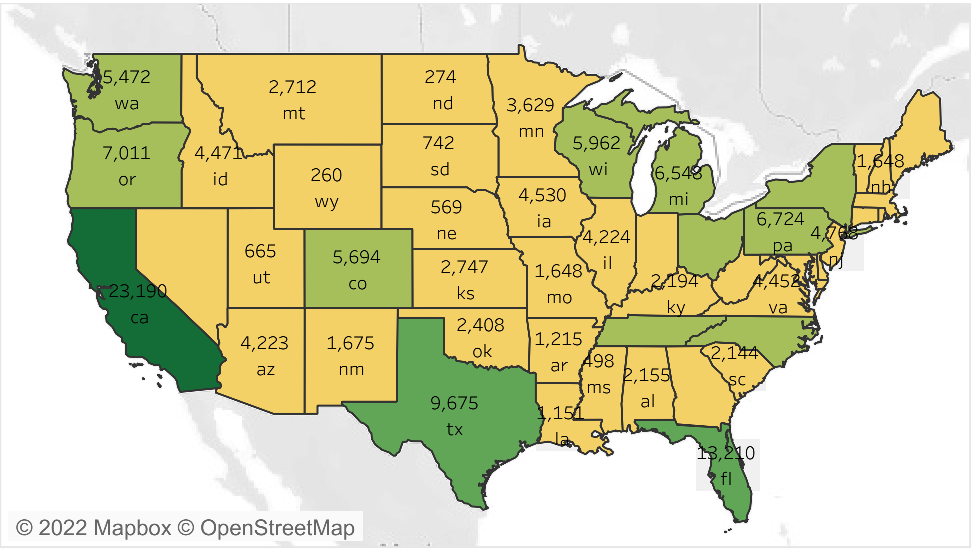
Data Source: https://www.kaggle.com/datasets/austinreese/craigslist-carstrucks-data
References::

Olisa Uzondu

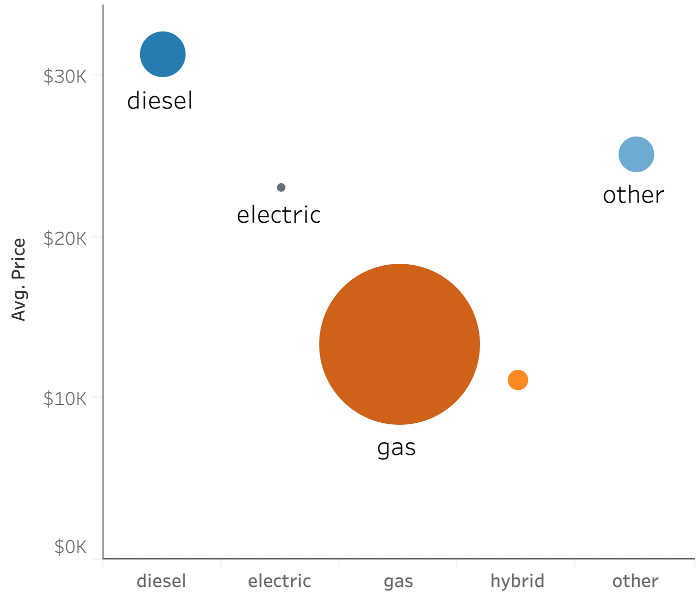
Craigslist’s Used Car Recommendation and Price Prediction Model

Sales can be increased by about 15% through **implementing a recommender engine** to target buyers, **and a price prediction system** that considers factors affecting price in order to ease price estimation headaches over the coming months.

Issues to Explore	Why do we do this?
1. What is the geographical distribution of the car listings?	1. We have to filter out the low listing volume areas from the high listing volume ares. We’d need to focus on hot-spots to boost sales
2. Do these regions have a preference for types of cars listed?	2. To focus efforts, we’d want to target specific types of cars (luxury or non-luuxury)
3. Whether there is a preference or not, what is the major fuel engine type?	3. Fuel prices fluctuate and so it’d be helpful to get an idea of the common fuel engine type of the listed cars to consider in the price model
4. How will age and non-age related factors affect sales?	4. We’d want to explore how these factors affect price to determine what listings to push forward.
5. How should the categorical values be used to establish relationships?	5. For the price prediction model, the categories would need to be transformed for correlation and model building purposes.



© 2022 Mapbox © OpenStreetMap



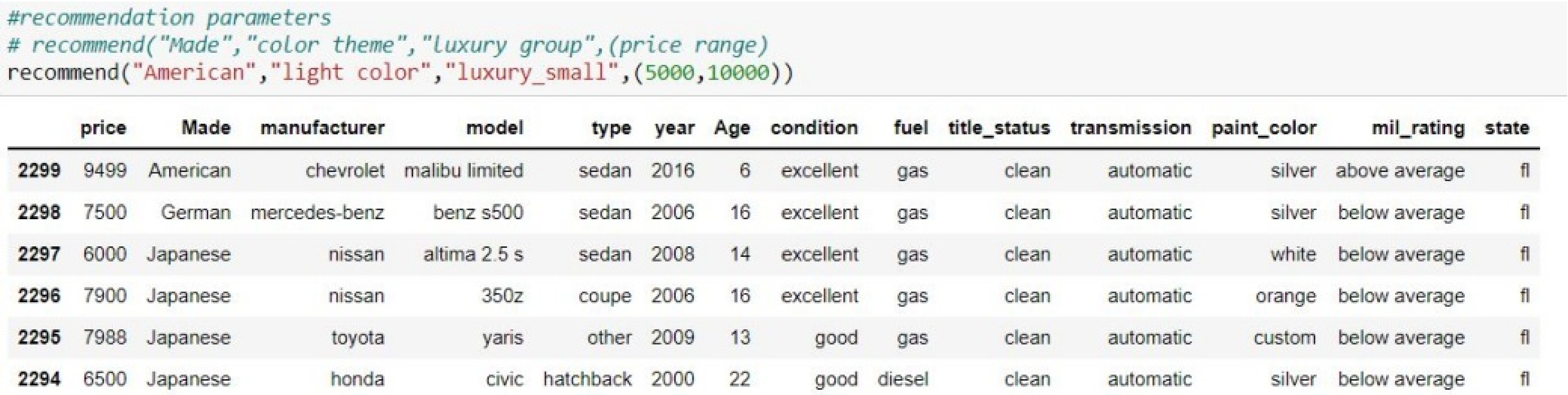
Data Source: https://www.kaggle.com/datasets/austinreese/craigslist-carstrucks-data
References::

Olisa Uzondu

Craigslist’s Used Car Recommendation and Price Prediction Model

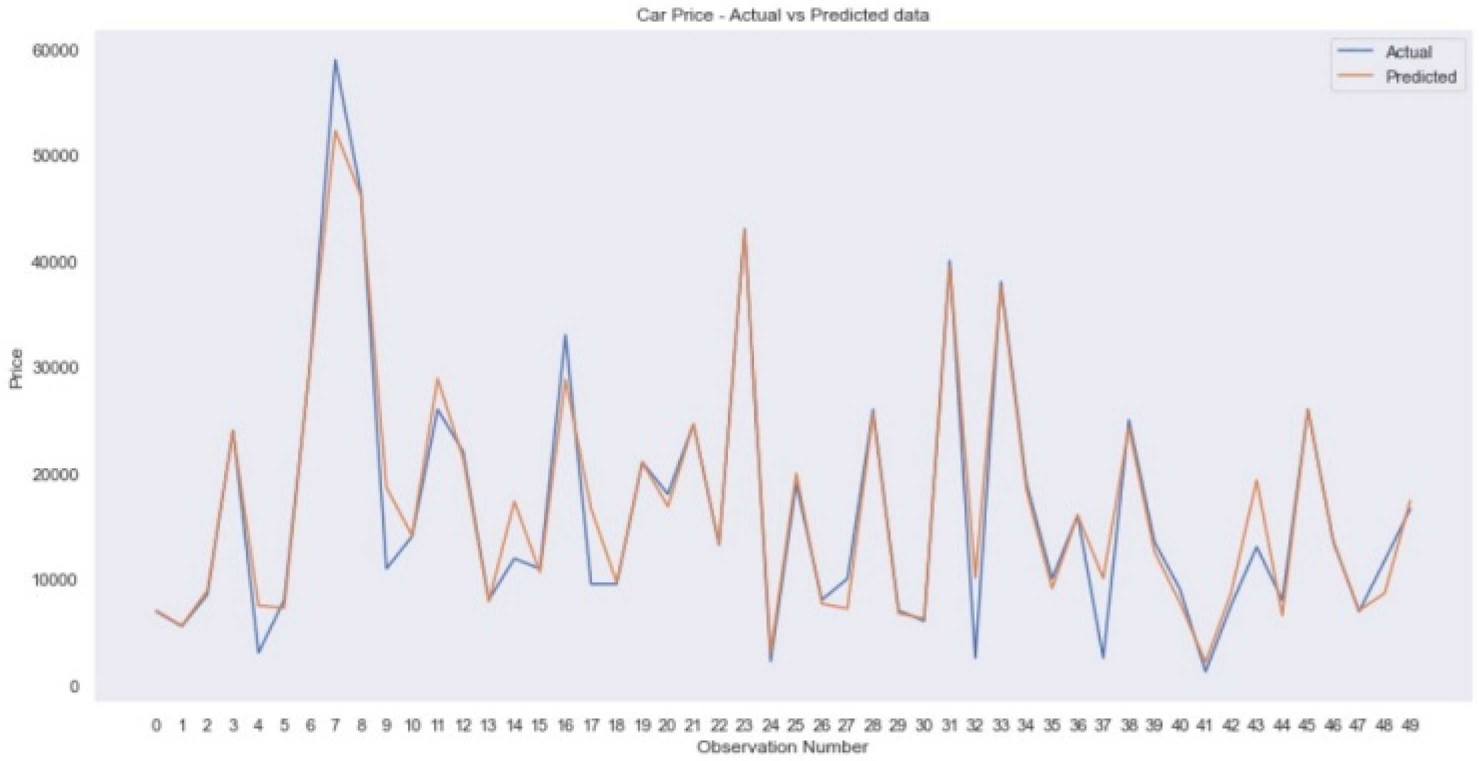
Presentation structure | Problem summary | Technical approach | Executive summary | Handling missing data and outliers | Content-based recommender | Correlation matrix

Through the utilization of **data cleansing and preprocessing techniques**, and **scikit-learn (sklearn) machine learning library**, a **content-based recommender system** able to recommend the **top 6 most similar vehicles** to a viewed listing was created, and a **price prediction model with 87.8% accuracy.**



```
#recommendation parameters
# recommend("Made","color theme","luxury group",(price range)
recommend("American","light color","luxury_small",(5000,10000))
```

	price	Made	manufacturer	model	type	year	Age	condition	fuel	title_status	transmission	paint_color	mil_rating	state
2299	9499	American	chevrolet	malibu limited	sedan	2016	6	excellent	gas	clean	automatic	silver	above average	fl
2298	7500	German	mercedes-benz	benz s500	sedan	2006	16	excellent	gas	clean	automatic	silver	below average	fl
2297	6000	Japanese	nissan	altima 2.5 s	sedan	2008	14	excellent	gas	clean	automatic	white	below average	fl
2296	7900	Japanese	nissan	350z	coupe	2006	16	excellent	gas	clean	automatic	orange	below average	fl
2295	7988	Japanese	toyota	yaris	other	2009	13	good	gas	clean	automatic	custom	below average	fl
2294	6500	Japanese	honda	civic	hatchback	2000	22	good	diesel	clean	automatic	silver	below average	fl

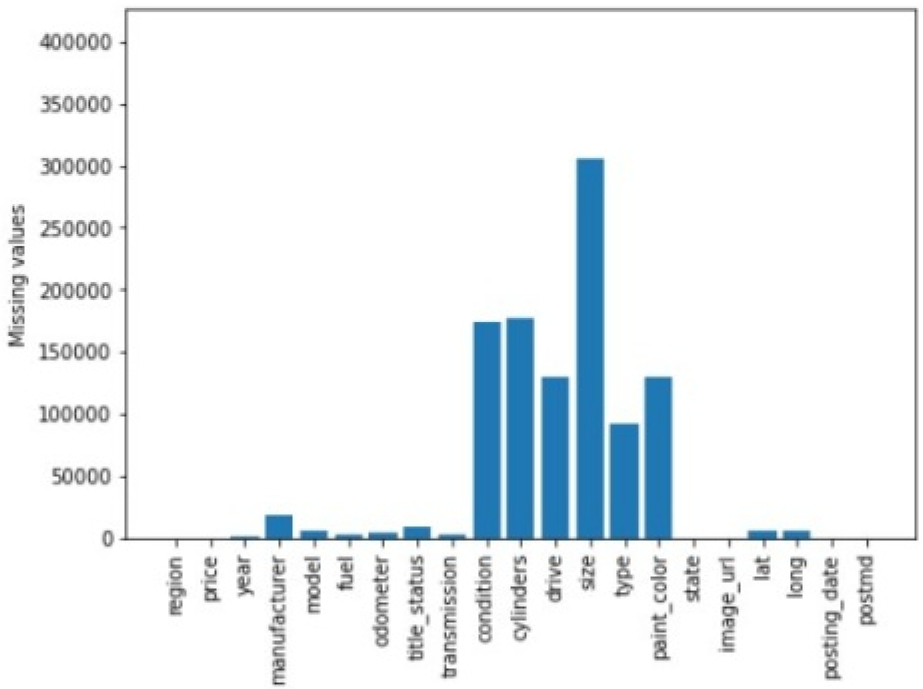


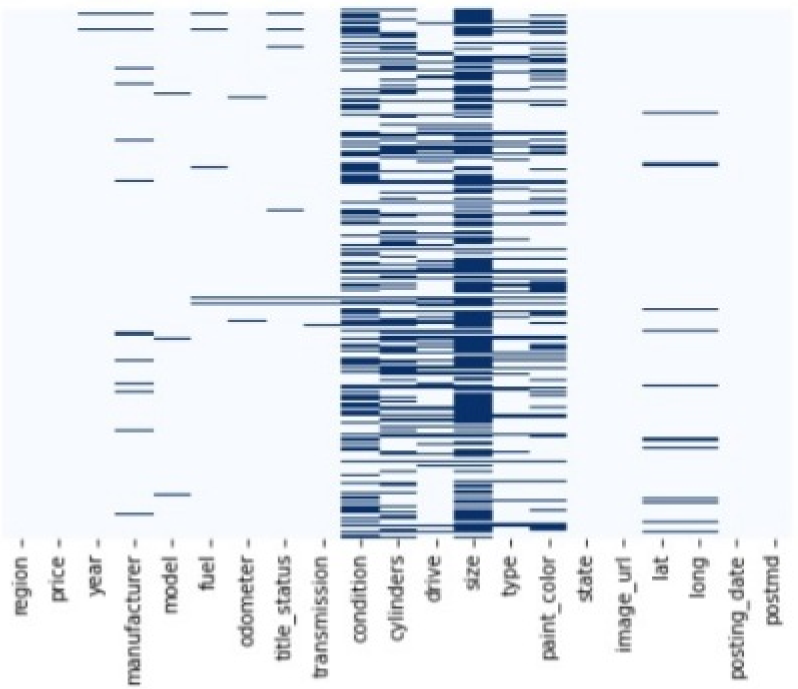
Data Source: https://www.kaggle.com/datasets/austinreese/craigslist-carstrucks-data
References::

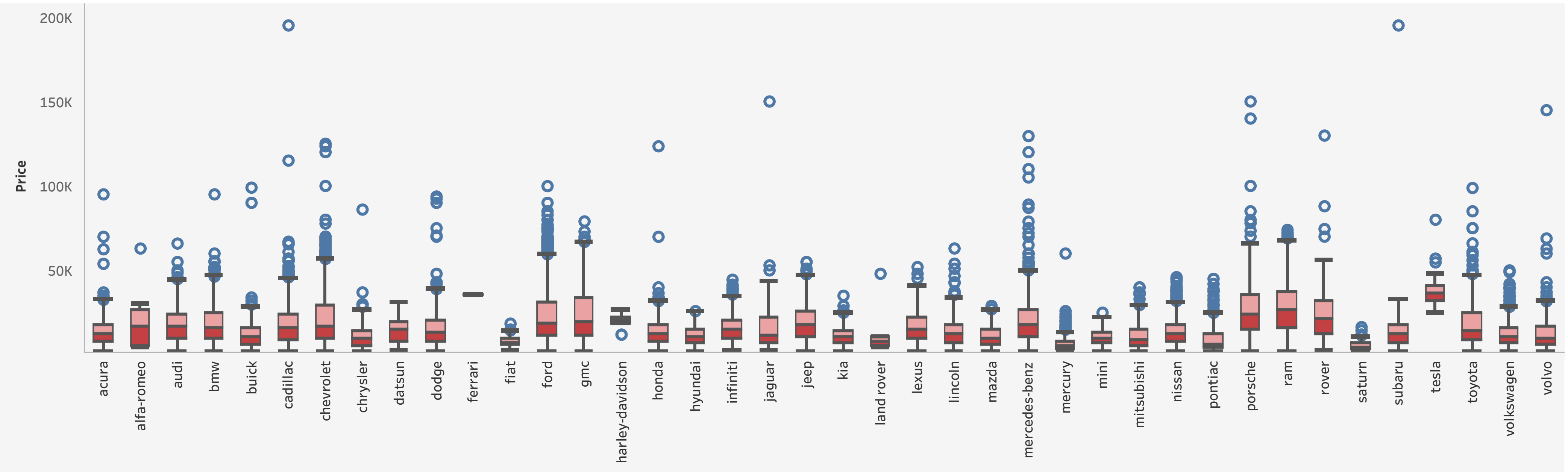
Olisa Uzondu

Craigslist’s Used Car Recommendation and Price Prediction Model

Examining the dataset revealed a **high number of missing entries** that involved **dropping several unnecessary columns** and multiple rows of missing data , **replacing some missing data** with their column counterparts **using numpy’s ”random.choice()” function**, and filtering out price and distance traveled **outliers** to have a cleaned dataset for proper analysis.





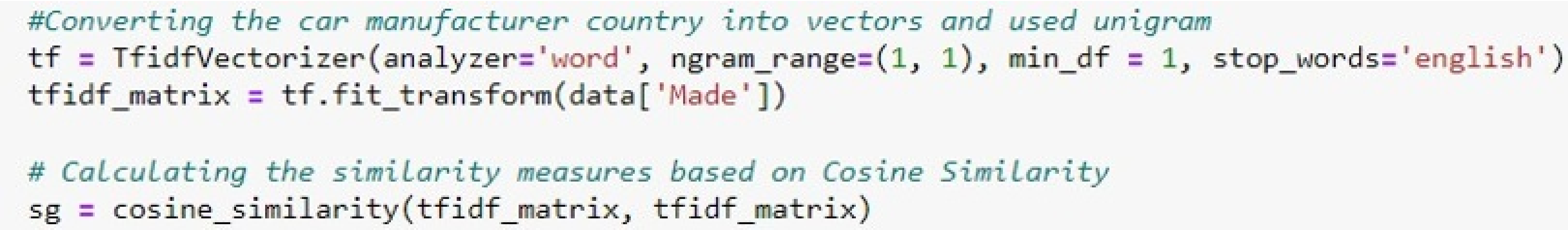


Data Source: https://www.kaggle.com/datasets/austinreese/craigslist-carstrucks-data
References::

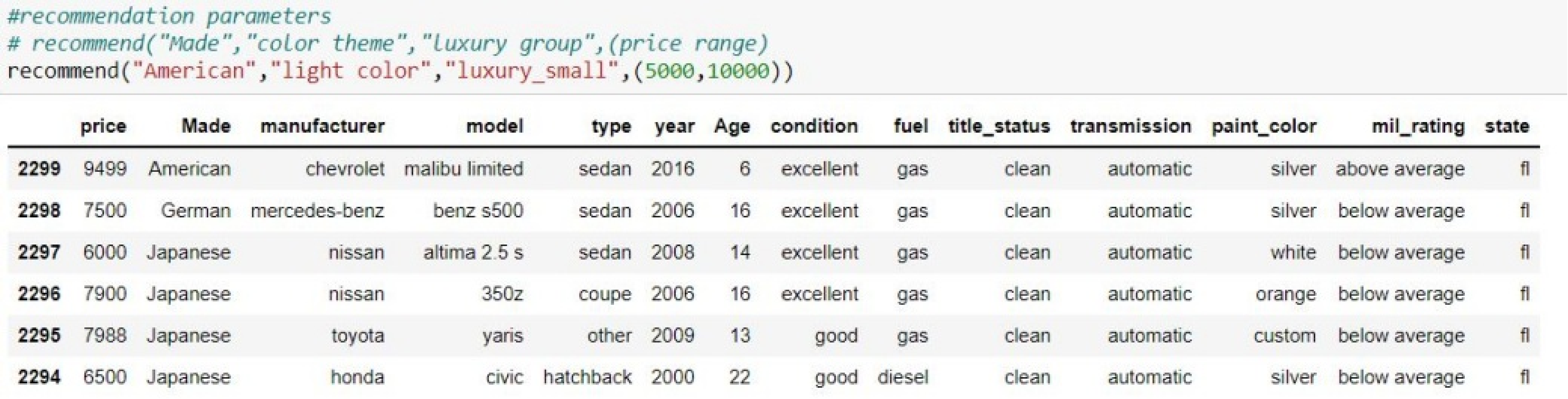
Olisa Uzondu

Craigslist’s Used Car Recommendation and Price Prediction Model

Further analysis of the dataset **indicated several categories with numerous distinct variables** and so were **classified into larger clusters** to simplify the recommendation model and improve its processing speed and output variety which **sklearn’s cosine similarity metric** was used.



```
#Converting the car manufacturer country into vectors and used unigram
tf = TfidfVectorizer(analyzer='word', ngram_range=(1, 1), min_df = 1, stop_words='english')
tfidf_matrix = tf.fit_transform(data['Made'])

# Calculating the similarity measures based on Cosine Similarity
sg = cosine_similarity(tfidf_matrix, tfidf_matrix)
```



```
#recommendation parameters
# recommend("Made","color theme","luxury group",(price range)
recommend("American","light color","luxury_small",(5000,10000))
```

	price	Made	manufacturer	model	type	year	Age	condition	fuel	title_status	transmission	paint_color	mil_rating	state
2299	9499	American	chevrolet	malibu limited	sedan	2016	6	excellent	gas	clean	automatic	silver	above average	fl
2298	7500	German	mercedes-benz	benz s500	sedan	2006	16	excellent	gas	clean	automatic	silver	below average	fl
2297	6000	Japanese	nissan	altima 2.5 s	sedan	2008	14	excellent	gas	clean	automatic	white	below average	fl
2296	7900	Japanese	nissan	350z	coupe	2006	16	excellent	gas	clean	automatic	orange	below average	fl
2295	7988	Japanese	toyota	yaris	other	2009	13	good	gas	clean	automatic	custom	below average	fl
2294	6500	Japanese	honda	civic	hatchback	2000	22	good	diesel	clean	automatic	silver	below average	fl

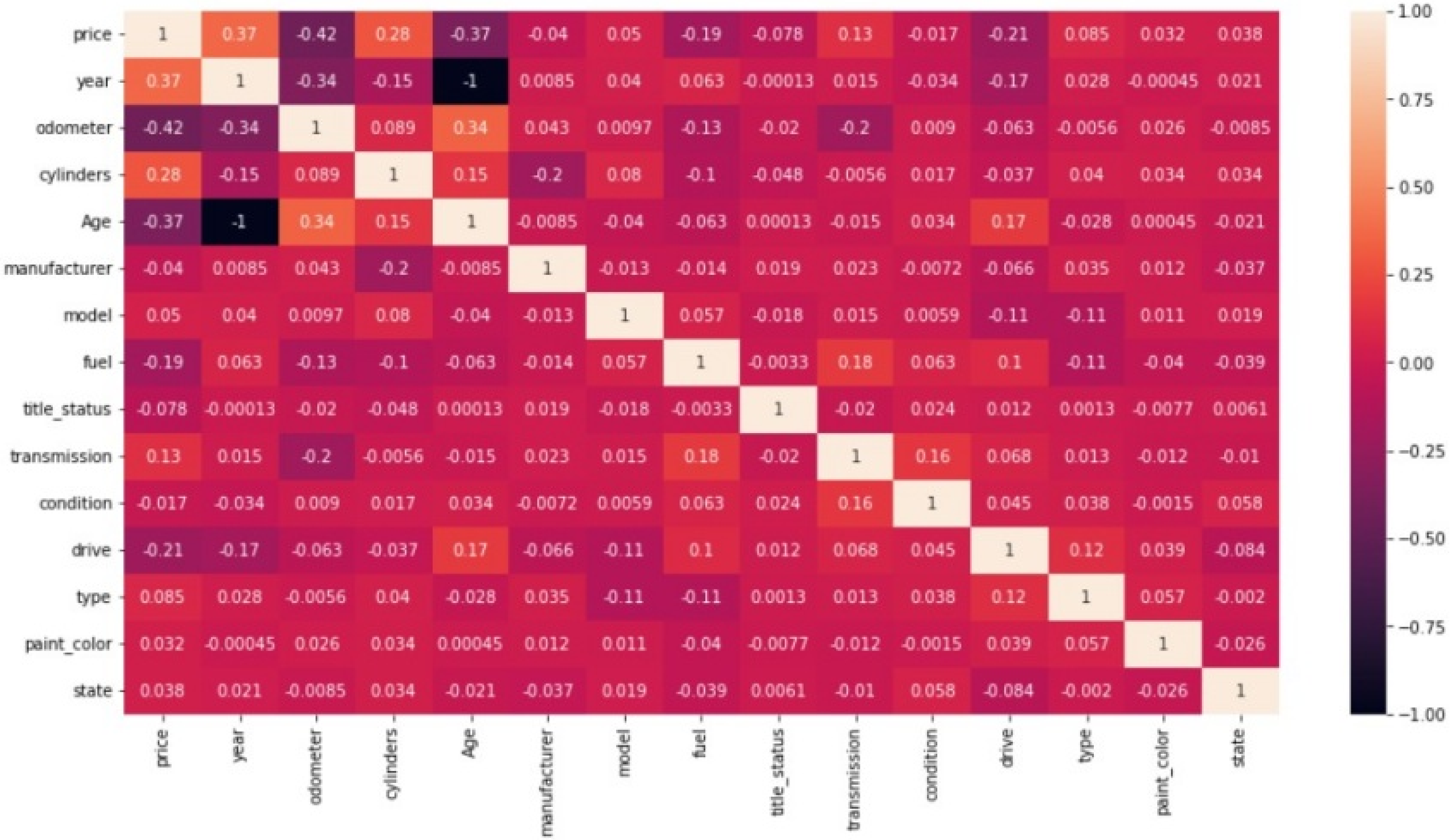
Data Source: https://www.kaggle.com/datasets/austinreese/craigslist-carstrucks-data
References:: https://www.kdnuggets.com/2020/07/building-content-based-book-recommendation-engine.html
*For python script see my jupyter notebook at https://1drv.ms/u/s!AnhfmE8xuYYA5ToSdheDFlzPl1mh?e=rIzaL1

Olisa Uzondu

Craigslist’s Used Car Recommendation and Price Prediction Model

Using **sklearns’s OrdinalEncoder** to transform the categorical features, there seemed to be **little to no correlation** to price except for age, distance traveled, and number of cylinders.

This low level strength of correlation tells us that not one variable has high influence in price determination. And so, most of the variables will be used in the price prediction model.



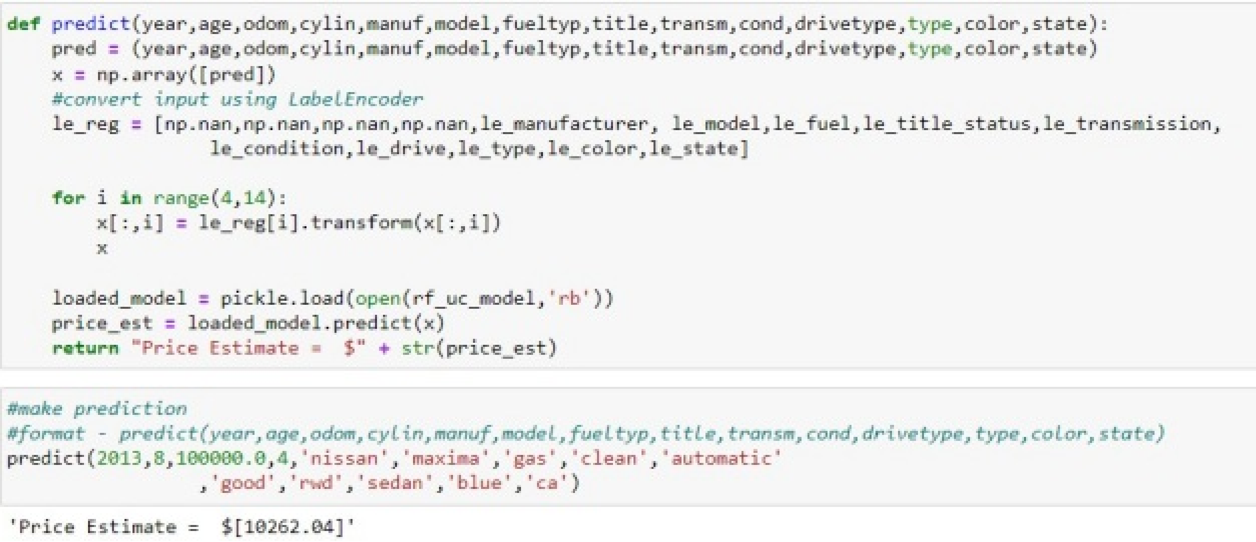
	price	year	odometer	cylinders	Age	manufacturer	model	fuel	title_status	transmission	condition	drive	type	paint_color	state
price	1	0.37	-0.42	0.28	-0.37	-0.04	0.05	-0.19	-0.078	0.13	-0.017	-0.21	0.085	0.032	0.038
year	0.37	1	-0.34	-0.15	-1	0.0085	0.04	0.063	-0.00013	0.015	-0.034	-0.17	0.028	-0.00045	0.021
odometer	-0.42	-0.34	1	0.089	0.34	0.043	0.0097	-0.13	-0.02	-0.2	0.009	-0.063	-0.0056	0.026	-0.0085
cylinders	0.28	-0.15	0.089	1	0.15	-0.2	0.08	-0.1	-0.048	-0.0056	0.017	-0.037	0.04	0.034	0.034
Age	-0.37	-1	0.34	0.15	1	-0.0085	-0.04	-0.063	0.00013	-0.015	0.034	0.17	-0.028	0.00045	-0.021
manufacturer	-0.04	0.0085	0.043	-0.2	-0.0085	1	-0.013	-0.014	0.019	0.023	-0.0072	-0.066	0.035	0.012	-0.037
model	0.05	0.04	0.0097	0.08	-0.04	-0.013	1	0.057	-0.018	0.015	0.0059	-0.11	-0.11	0.011	0.019
fuel	-0.19	0.063	-0.13	-0.1	-0.063	-0.014	0.057	1	-0.0033	0.18	0.063	0.1	-0.11	-0.04	-0.039
title_status	-0.078	-0.00013	-0.02	-0.048	0.00013	0.019	-0.018	-0.0033	1	-0.02	0.024	0.012	0.0013	-0.0077	0.0061
transmission	0.13	0.015	-0.2	-0.0056	-0.015	0.023	0.015	0.18	-0.02	1	0.16	0.068	0.013	-0.012	-0.01
condition	-0.017	-0.034	0.009	0.017	0.034	-0.0072	0.0059	0.063	0.024	0.16	1	0.045	0.038	-0.0015	0.058
drive	-0.21	-0.17	-0.063	-0.037	0.17	-0.066	-0.11	0.1	0.012	0.068	0.045	1	0.12	0.039	-0.084
type	0.085	0.028	-0.0056	0.04	-0.028	0.035	-0.11	-0.11	0.0013	0.013	0.038	0.12	1	0.057	-0.002
paint_color	0.032	-0.00045	0.026	0.034	0.00045	0.012	0.011	-0.04	-0.0077	-0.012	-0.0015	0.039	0.057	1	-0.026
state	0.038	0.021	-0.0085	0.034	-0.021	-0.037	0.019	-0.039	0.0061	-0.01	0.058	-0.084	-0.002	-0.026	1

Data Source: https://www.kaggle.com/datasets/austinreese/craigslist-carstrucks-data
References:: https://www.analyticsvidhya.com/blog/2021/07/car-price-prediction-machine-learning-vs-deep-learning/
*For python script see my jupyter notebook at https://1drv.ms/u/s!AnhfmE8xuYYA5ToSdheDFlzPl1mh?e=rIzaL1

Olisa Uzondu

Craigslist’s Used Car Recommendation and Price Prediction Model

Using **sklearn’s LabelEncoder**, target variables were encoded and fed into three (3) (LinearRegression, XGBoostRegressor, RandomForestRegressor) machine learning models, out of which the **RandomForest model** performed the best with an **87.8% accuracy**

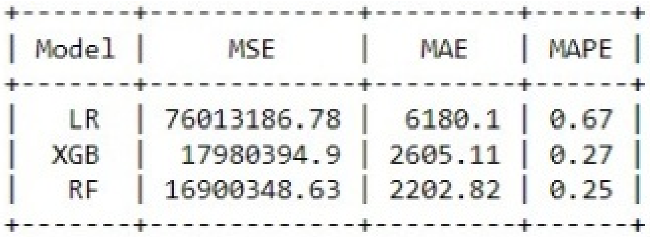


```
def predict(year,age,odom,cylin,manuf,model,fueltyp,title,transm,cond,drivetype,type,color,state):
    pred = (year,age,odom,cylin,manuf,model,fueltyp,title,transm,cond,drivetype,type,color,state)
    x = np.array([pred])
    #convert input using LabelEncoder
    le_reg = [np.nan,np.nan,np.nan,np.nan,le_manufacturer, le_model,le_fuel,le_title_status,le_transmission,
                   le_condition,le_drive,le_type,le_color,le_state]

    for i in range(4,14):
        x[:,i] = le_reg[i].transform(x[:,i])
        x

    loaded_model = pickle.load(open(rf_uc_model,'rb'))
    price_est = loaded_model.predict(x)
    return "Price Estimate =  $" + str(price_est)
```

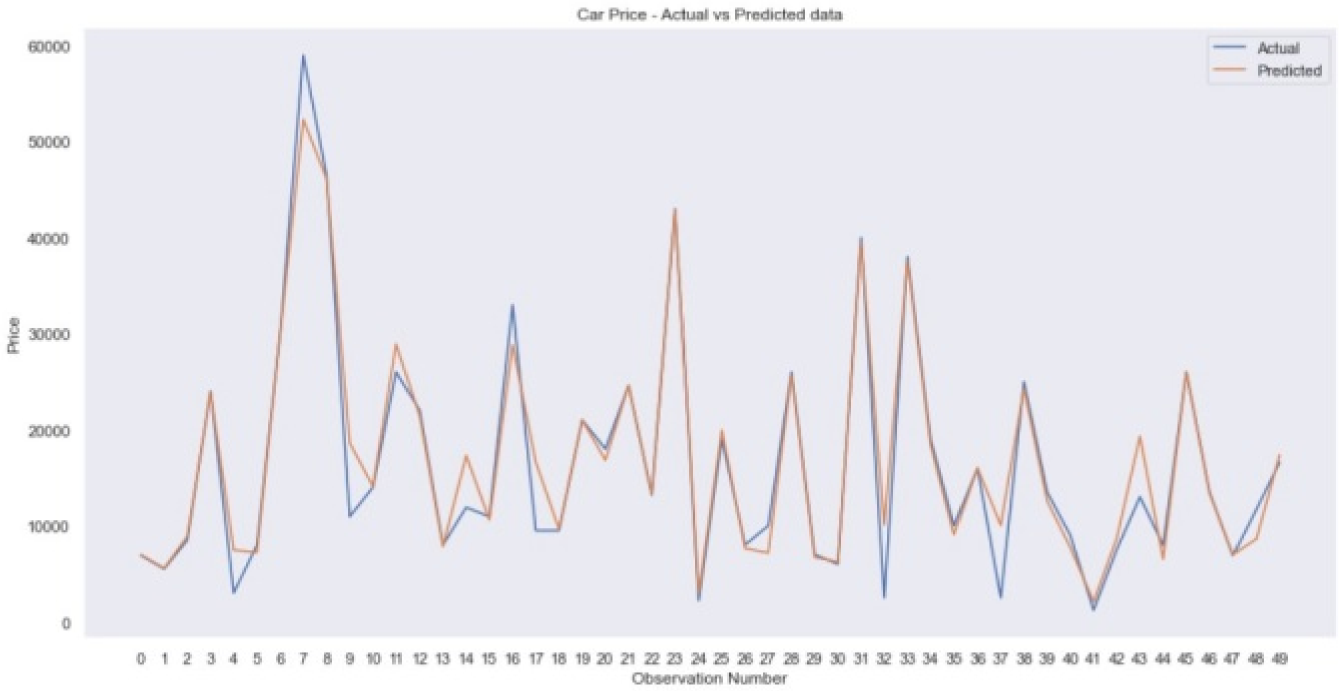
```
#make prediction
#format - predict(year,age,odom,cylin,manuf,model,fueltyp,title,transm,cond,drivetype,type,color,state)
predict(2013,8,100000.0,4,'nissan','maxima','gas','clean','automatic'
                 ,'good','rwd','sedan','blue','ca')
```

```
'Price Estimate =  $[10262.04]'
```

Model Evaluation Table



```
+-------+-------------+---------+------+
| Model |     MSE     |   MAE   | MAPE |
+-------+-------------+---------+------+
|   LR  | 76013186.78 |  6180.1 | 0.67 |
|  XGB  |  17980394.9 | 2605.11 | 0.27 |
|   RF  | 16900348.63 | 2202.82 | 0.25 |
+-------+-------------+---------+------+
```



Data Source: https://www.kaggle.com/datasets/austinreese/craigslist-carstrucks-data
References:: https://www.analyticsvidhya.com/blog/2021/07/car-price-prediction-machine-learning-vs-deep-learning/
*For python script see my jupyter notebook at https://1drv.ms/u/s!AnhfmE8xuYYA5ToSdheDFlzPl1mh?e=rIzaL1

Olisa Uzondu

Craigslist’s Used Car Recommendation and Price Prediction Model

BUILDING A RECOMMENDATION AND PRICE PREDICTION SYSTEM FOR CRAIGSLIST’S USED CARS BUYING AND SELLING PLATFORM

CONCLUSION

A recommender system and a price prediction model for the used car dataset was successfully created after extensive data cleaning and preprocessing.

The recommender system is able to print out the 6 most similar cars to what the user sets as a guiding parameter much like a search engine filter. ALternatively, the top 6 similar listings to the listing currently being viewed are displayed .

The price prediction model was achieved using a Random Forest Regressor machine learning model with an 87.79% accuracy.

RECOMMENDATION

Price prediction model

- Use a cleaner data preprocessing techniques to prepare dataset for machine learning to improve prediction results.
- Develop an interactive user interface where users can estimate the price of a car based on various features
- Examine price disparity between regions, e.g. east and west coast, and segregate the price prediction models accordingly
- Try this model with other car datasets to see how it performs with other datasets

Recommender system

- Explore other methods of building a content based recommender system to incorporate location, seasonal weather condition, vehicle main use case, etc
- implement an input function where users input their choices to specification prompts

Data Source: https://www.kaggle.com/datasets/austinreese/craigslist-carstrucks-data
References::

Olisa Uzondu