Progress Report

TEAM: 11
SOPHIE LUEHMANN
GABRIEL SUAREZ

GITHUB REPO: https://github.com/sophieluehmann/ML_project_modeling

Problem Statement:

When someone wishes to sell their car online, it can be very daunting or unclear how to best price their vehicle. Uninformed pricing can cause vehicles to stay on the market for far longer than they should (or even not sell at all), or cause the seller to lose out on money that their vehicle's value could, in fact, yield. For individuals with little knowledge of the market and/or understanding of automobiles, it is likely that this process is far more arduous and time consuming than it needs to be, and we wish to create a machine learning solution to this problem that will allow individuals to receive an optimal price estimate that mitigates the amount of time it takes to sell, and the amount of value lost, based on previously existing used car listings.

Preliminary Data Analysis Results:

Before beginning our data preprocessing, we decided to incorporate new datasets into our project that we identified to be better suited for our purposes as they utilize U.S. listings as opposed to foreign ones (which introduce a litany of issues including: customs clearance, pricing discrepancies, model availability etc.).

In section B we ironed out inconsistencies and irrelevant attributes from our datasets in order to homogenize them and be able to create our master dataset. We removed price outliers, addressing thee following previously mentioned challenge: "There are some unique editions/model years that carry more extravagant values because of factors unaccounted for in the datasets (i.e. first model year a vehicle is produced, a rare options package, restorative work, famous previous owner, etc.)". We also removed year and mileage outliers as well.

In section C we observed the relationships between year and price, at a few different ranges starting from around 1930 going to about 2020, and then refining all the way down to 2002 -

2020. In general, as expected, newer cars cost more. We also observed the sale prices of different makes across all the sets and the master set with boxplots. We looked at the mileage of vehicles vs the price and observed what we again expected which is that the more miles a vehicle has on it the less it is likely to sell for. Lastly, we compared the mileage to the model year of the vehicles and noticed that the older model cars prior to ~1960 tended to have far fewer miles on them than newer vehicles, perhaps because they are collector vehicles.

Algorithm Solutions and Techniques:

Pre-processing and data cleaning:

- Drop irrelevant attributes, remove entries with missing data, encode categorical values, remove outliers, integrate data sets

Exploratory data analysis:

- Descriptive statistics (minimum, maximum, mean, median and quartile values with boxplot), scatter plots, visualize data and observe relationships between price and attributes

Linear Regression

- Makes a prediction by computing a weighted sum of the input features, plus a constant known as the bias term

Polynomial Regression

- A way to fit nonlinear data to a linear model
- Polynomial Regression adds powers of each feature as new features, then trains a linear model on this extended set of features
- When there are multiple features in a dataset, Polynomial Regression finds relationships between features (which is something a plain Linear Regression model cannot do) by adding all combinations of features up to the given degree

Random Forest Regression

- An ensemble of Decision Trees
- Searches for the best feature among a random subset of features when growing trees which results in greater tree diversity
- Trees in different subspaces generalize their classification in complementary ways, and their combined classification can be monotonically improved

Risks and Challenges

- We would like to observe the initial listing price, time until the vehicle is sold, and the final sale price to make our model more informed, but the datasets don't have all of these attributes

- Determining level of granularity to predict (price ranges, vs. a precise price estimate)

- Accuracy of owner's assessment of vehicle's condition could be an issue if too many inflate the condition of their vehicles

- Current market conditions play a role in the price of used cars so fresh data will need to be imported regularly for our estimations to remain accurate

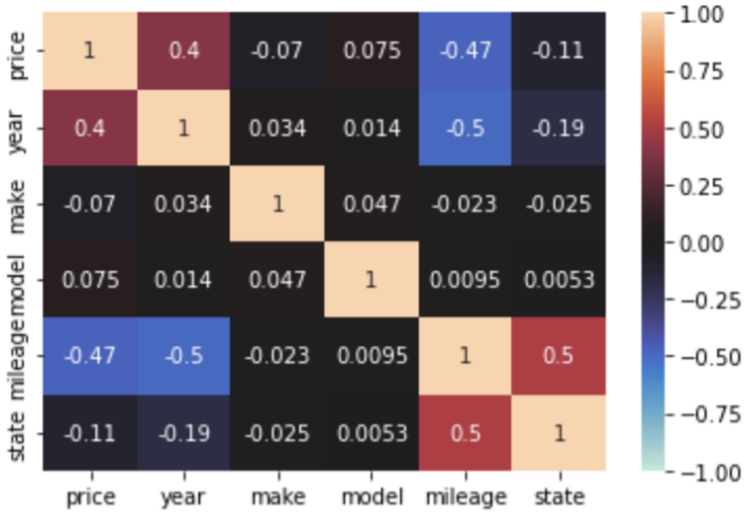
- The more attributes we choose to consider, the smaller our master dataset will become due to many of the rows lacking data. We must find a healthy balance between dataset size and number of attributes utilized in our models.

Modeling and Performance Evaluation

We decided to use the linear regression, polynomial regression and Random Forest regression algorithms as our models. Their performance evaluations are described below:

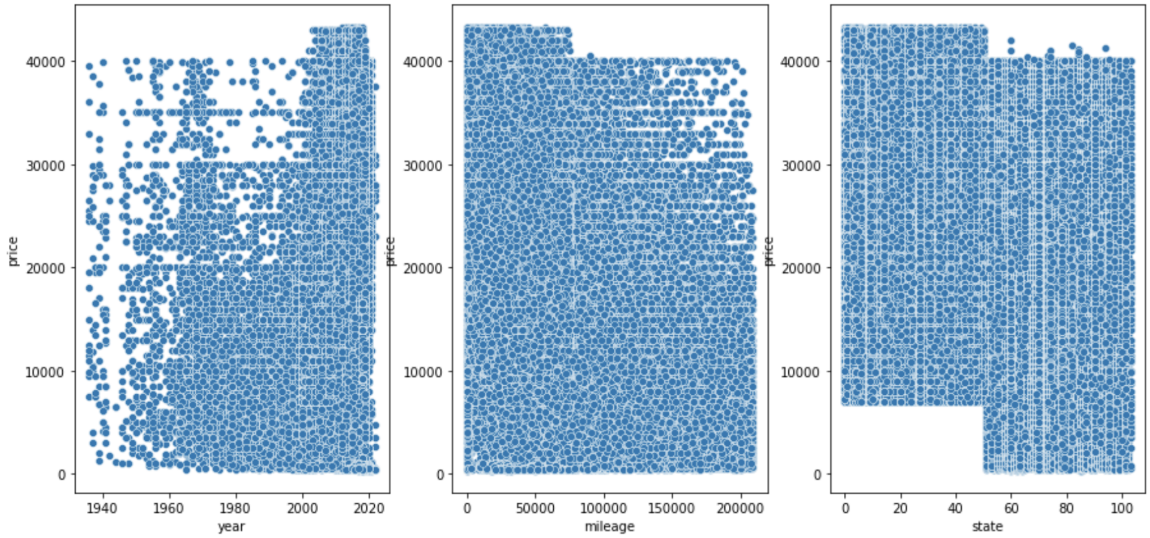
Linear regression

We began by computing a simple regression using only the year and mileage as features, and this gave us our baseline values for both RMSE(7847.19) and R^2 (0.259). Next, we encoded the following three categorical features: (state, make, model), so that we could use them in the regression, in hopes of improving our accuracy. We then used heat map of all the aforementioned features to determine which to use in the next linear regression, and it resulted as follows:

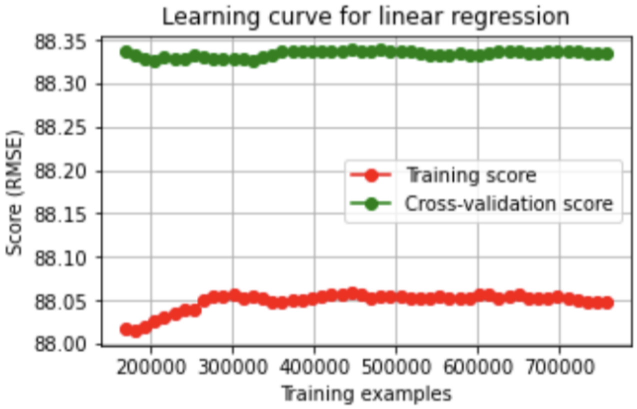


	price	year	make	model	mileage	state
price	1	0.4	-0.07	0.075	-0.47	-0.11
year	0.4	1	0.034	0.014	-0.5	-0.19
make	-0.07	0.034	1	0.047	-0.023	-0.025
model	0.075	0.014	0.047	1	0.0095	0.0053
mileage	-0.47	-0.5	-0.023	0.0095	1	0.5
state	-0.11	-0.19	-0.025	0.0053	0.5	1

After analyzing the resulting heatmap we decided to proceed with year, mileage, and state as our selected attributes for the next regression. Prior to doing so we checked each of their relationships with the price of the vehicles by producing the following three graphs:



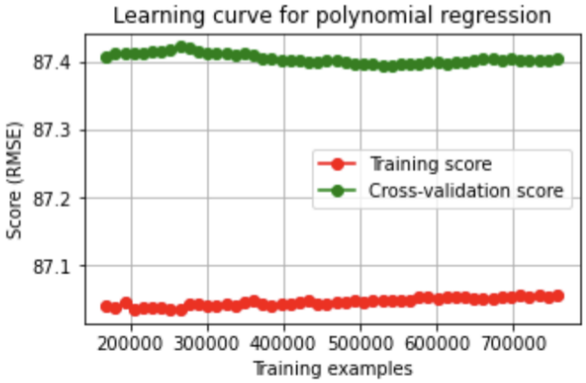
These graphs were pretty disheartening as the distributions were so dense and expansive, causing us to think these attributes alone would likely not be sufficiently informative for our models to be very accurate. Still, we proceeded to use them. We then performed K-cross validation with 10 splits and created a test bed for our models. The average RMSE was around 7757.62 and the average R^2 value was around 0.276. Using the test bed we created our gradient descent linear regression, using a learning rate of 0.1 and an epsilon of 1e-7, and normalized values; it gave us an RMSE of 7772.262 and a R^2 value of around 0.275. The learning curve for this regression appears as follows:



This learning curve is acceptable to us as the difference between the two scores seems to remain within 1.0.

Polynomial regression

For this regression we used the same X and y values as before, and we scaled the X testing and training values and proceeded to fit the model. We achieved a RMSE for the testing set of about 7772.18 and R^2 score of about 0.276; strikingly similar to our linear regression. The learning curve for this regression appears as follows:



Again, this learning curve is acceptable to us since the difference also remains within 1.0.

Random Forest regression

For our Random Forest regression we normalized and scaled our inputs and achieved a RMSE of approximately 7950.46 (a bit worse than the other two) and a R^2 score of ~0.242 (a bit better than the other two). We were unable to generate a Learning curve for this regression, however.

Plan For Completion

At this stage in the project, we have trained models for the entire master data set. This means that we had to scale down the features our models are trained on because we have only included features that every instance has data for. Specifically, our models are trained to estimate price based on make, model, year, mileage, and state. However, when looking at the heat map, we see that make and model were not significant factors in price estimation so really our models are estimating price only based on mileage, year, and state.

Our plan going forward is to scale back the instances included in the data set we train so that we can include more features to see if there are factors we have not considered that play a significant role in price estimation. The features we will be including in further models are city, title status, region, and potentially certain identifying characters from the vehicles' VINs. We will need to train more models and compare those to the models that we already have, taking into consideration how the reduction of instances in the dataset and expansion of features affects models' performance. Once this is complete, our results will offer a solution to the problem we set out to solve.

To-Do + Time :

- Train models with more features (5-6 hours)
- Analysis of new models + model performance visualization (2 hours)
- Comparison of old models and new models (1 hour)
- Establish results and explain solution (1-2 hours)
- Write final report (1 hour)
- Prepare presentation (1 hour)

Citations:

Gradient Descent function: https://www.kaggle.com/code/hieunt01/used-car-price-predictions-using-linear-regression

Description: The method used in the sklearn pipeline for our linear regression model was found on this webpage.

Used Cars Dataset

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

Description: This data contains information from Craigslist on vehicles for sale. The data is scraped every few months and collects relevant information for car listings.

Instances: 426880
Attributes: 26

* id - entry ID
* url - listing URL
* region - craigslist region
* region_url - region url
* price - entry price

* year - entry year
* manufacturer - manufacturer of vehicle
* condition - condition of vehicle
* cylinder - number of cylinders
* fuel - fuel type
* odometer - miles traveled by vehicle
* title_stats - title status of vehicle
* transmission - transmission of vehicle
* VIN - vehicle identification number
* drive - type of drive
* size - size of vehicle
* type - generic type of vehicle
* image_url - image url
* description - listed description of vehicle
* county - NA
* state - state of listing
* lat - latitude of listing
* long - longitude of listing
* posting_date - date craigslist ad was posted

Used Car Price Predictions

https://www.kaggle.com/datasets/harikrishnareddyb/used-car-price-predictions

Description: This dataset contains information on used cars for sale in the United States. Eight features were assembled for each car sale listing in the dataset.

Instances: 852122
Attributes: 8

* Price - Target Variable.
* Year - Year of the car purchased.
* Mileage - The no.of km's driven by the car.
* City - In which city it was sold.
* State - In which state it was sold.
* Vin - a unique number for a car.
* Make - Manufacturer of the car.
* Model - The model(name) of the car.

US Cars Dataset

https://www.kaggle.com/datasets/doaaalsenani/usa-cers-dataset

Description: This data is scraped from AUCTION EXPORT.com. and includes information on 28 brands of clean and used vehicles for sale in US. Twelve features were assembled for each car in the dataset.

Instances: 2499
Attributes: 12

* Price - The sale price of the vehicle in the ad
* Year - The vehicle registration year
* Brand - The brand of car
* Mode - model of the vehicle
* Color - Color of the vehicle
* State - The state in which the car is being available for purchase
* City - The city in which the car is being available for purchase
* Mileage - miles traveled by vehicle
* Vin - The vehicle identification number is a collection of 17 characters (digits and capital letters)
* Title Status - This feature included binary classification, which are clean title vehicles and salvage insurance
* Lot - A lot number is an identification number assigned to a particular quantity or lot of material from a single manufacturer.For cars, a lot number is combined with a serial number to form the Vehicle Identification Number.
* Condition - Time remaining for sale