Used Car Market Analysis

How much is my car worth? Which cars are good investments?

More shoppers today are likely to cite reliability (41% vs. 35% in 2022), finding a vehicle that fits their budget (40% vs. 33% in 2022), and expected costs (26% vs. 21% in 2022) as the most important factors in selecting a vehicle... with 89% saying they'd be willing to switch models and 69% open to switching brands.

— Car Guru Consumer Preferences Survey

Context: It can be difficult to identify a specific price for every vehicles depending on the mileage, trim level, and other varying features

Criteria for Success: Create a regression model for price with an error of $3000 or less per car

Scope: Identify specific models that retain value more or less than the competition

Constraints: Lack of organized vehicle reference data, imbalanced used car listings

Stakeholders: Car Dealers, Buyers, and Analysts

Data Sources: Craigslist; Edmunds, KellyBlueBook for reference car information

Data Wrangling

[400k Used Vehicles Listings Craigslist (Feb 2021) - Kaggle]

Columns

Categorical Data: manufacturer, model, cylinders, transmission, drive, title status, fuel type, size, type, condition, paint color, **VIN**, description, **state, region**

Numerical data: latitude, longitude, **odometer, price**

Data Problem 1: **User Entered Data**

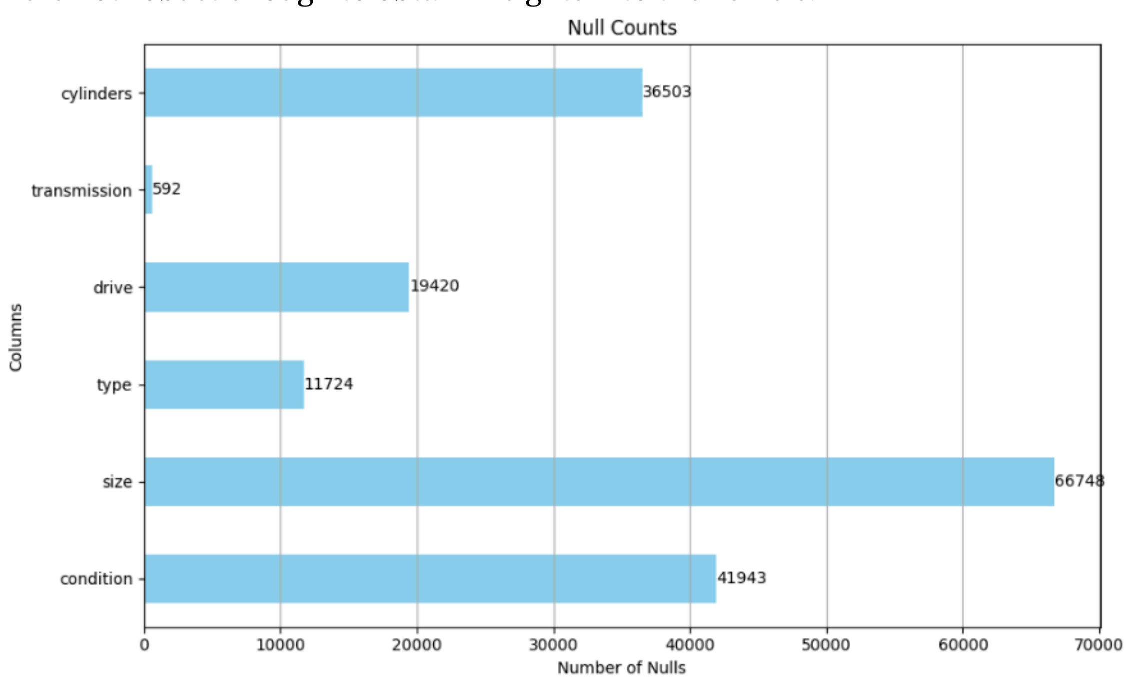
- Incorrect information: ex: 8 cylinder car listed as 6 cylinders
- ‘Condition’ being subjective data
- Unorganized: identical values entered slightly differently ex. Impala vs. Impala LT

Data Problem 2: **No Sale Price**

- Only Listed Price so there is more variance

Data Problem 3: **Duplicate Car Listings**

Other datasets could not address both problems. If a sale price was listed, the other columns were not robust enough to obtain insights into the vehicle.

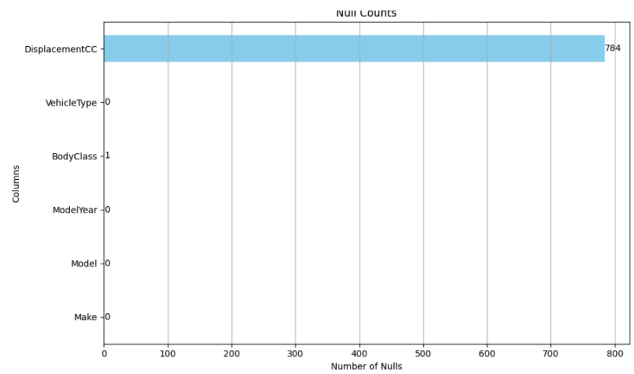


GMC Sierra 1500: 113 Unique Values

```
['sierra 1500 crew cab slt' 'sierra 1500 regular cab'
 'sierra 1500 extended cab slt' 'sierra 1500 limited double'
 'sierra 1500 double cab sle' 'sierra 1500 crew cab sle' 'sierra 1500'
 'sierra 1500 extended cab sle' 'sierra 1500 classic'
 'sierra 1500 regular cab work' 'sierra 1500 double cab'
 'sierra 1500 crew cab' 'sierra 1500 hd crew cab' 'sierra 1500 denali'
 'sierra 1500 at4 automatic' 'sierra 1500 extended cab' 'sierra 1500 base'
 'sierra 1500 4wd crew cab 143' 'sierra 1500 2wd reg cab 119.'
```

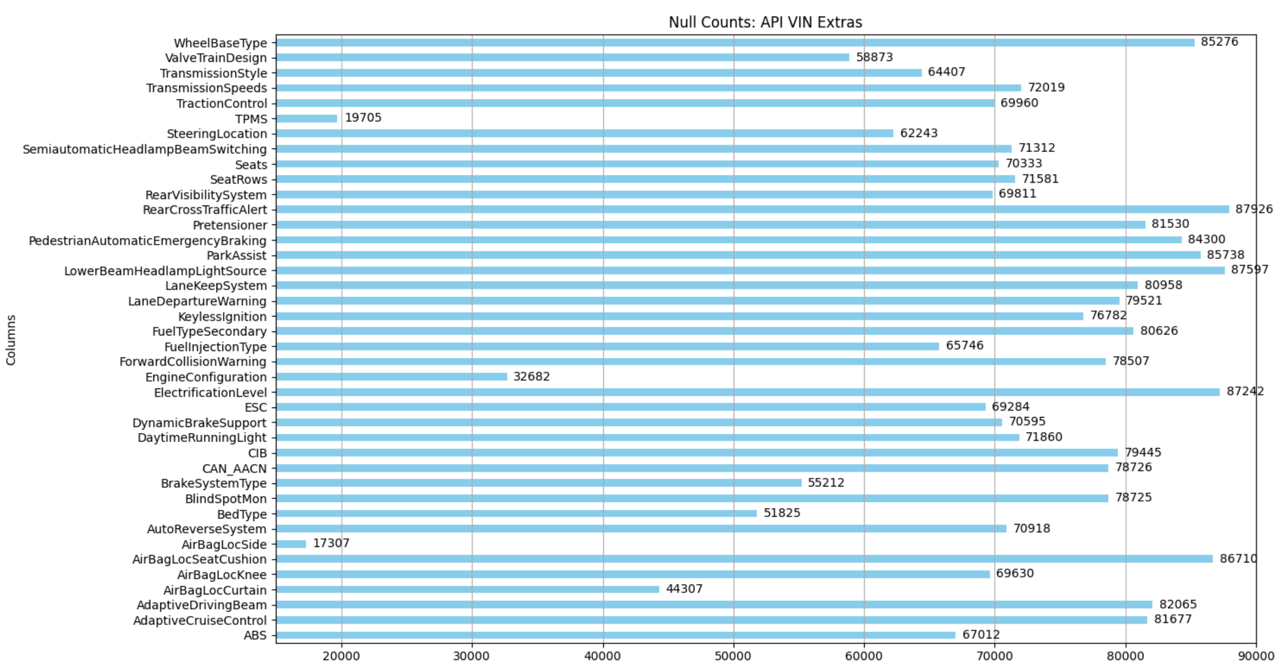
Solution: [VIN Decoder]()

from [NHTSA.gov]()



- **Removes Duplicates,** avoiding skewed predictions
- **Verifies and Organizes Features:**
    - Make
    - Model
    - Year
    - Cylinders
    - Engine Size - 3.0L Engine, 4.6L Engine, etc.
    - Fuel Type - Gasoline, Hybrid, Diesel
    - BodyClass - Sedan, Coupe, Convertible, Pickup, Van, Cargo Van, SUV, MPV
    - VehicleType - Truck, Car
    - Gross Vehicle Weight Rating (GVWR)

Downside: Many Null Values for Extra Features, Series, Trim



Notebook: [Batch VIN Decoding]()

Notebook: [Data Wrangling]()

Imputations

Rows: 93,031

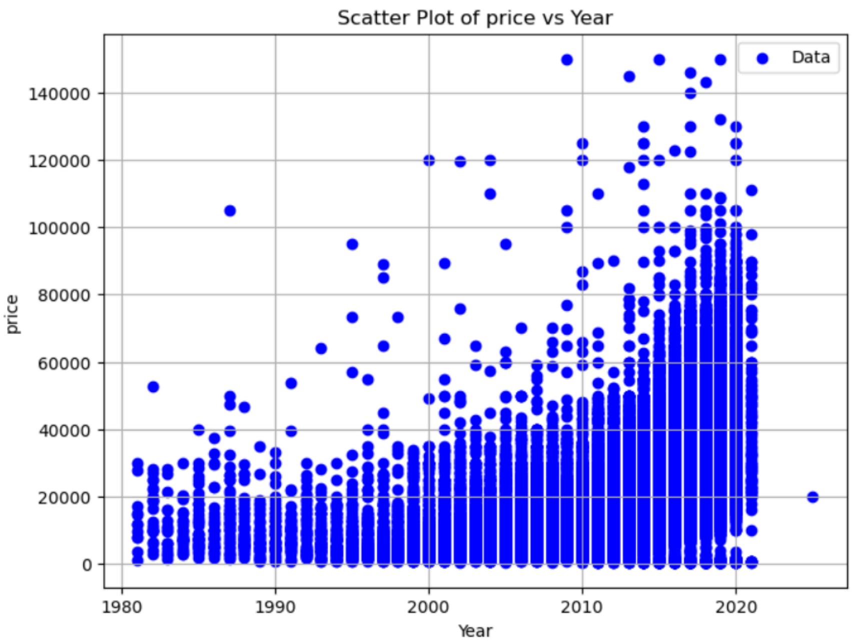
Nulls Values for each Column:

1) **‘Doors’: 15,505**
    - Most were found to be BodyClass ‘pickup’. The number of doors was found to always be 2 for regular cabs, and 4 for extended or mega cabs. Rest: imputed as 4 doors, which is the mode.
2) **DriveType (ex. RWD, FWD, 4WD): 26,070**
    - 5,000 were imputed using the value from craigslist ‘drive column, 15,000 from the missing row’s BodyClass & EngineCylinders ‘DriveType’ mode. Remaining 6,000 did not have EngineCylinders values and were imputed using as specific of information I could find: engine size, make, model, year, trim/series: mode.
3) **EngineCylinders: 8,493**
    - Mode was imputed, according to the vehicle’s Make, Model, Year, and Engine Size.
4) **FuelTypePrimary: 3,149**
    - Referenced Engine Size for Pickup Trucks to find Diesel trucks. Mode Imputation: Gasoline
5) **GVWR: 924**
    - Mode Imputation: 6,000lb or Less
6) **‘DisplacementCC’: 828**
    - **Dropped**
7) **BodyCabType: 37,529**
    - Not Pickup trucks, Encoded as ‘Not Applicable’
8) **EngineHP: 42,709**
    - Mean Imputation / Left as ‘null’ for LGBM, XGBoost, CatBoost
9) **EngineConfiguration: 42,709**
    - Mode Imputation based on Make/Model Combination’s Mode Value

A big challenge for this project has been finding and implementing clean datasets that can increase the feature set for making better predictions. It is a data engineering problem to implement automated cleaning algorithms for the VIN Decoder output as well as the creation of clean datasets that can be referenced to input correct information.

Notebook: EDA

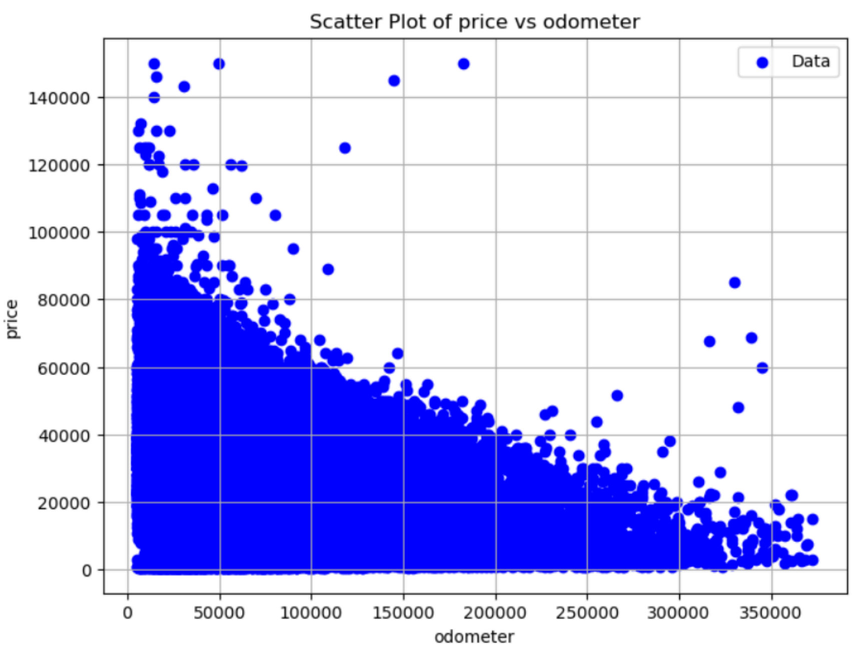
EDA



Mean Price: $19,000

Using 3 standard deviations to remove outliers was ineffective as it only removed prices above $76,000. Price outliers were difficult to spot because refurbished vehicles had high miles but were listed at high price.

Larger vehicles held their value despite being older and higher miles so there was not a consistent ratio of price to odometer or year.



Problem

New Vehicles with Low Miles and Low Price

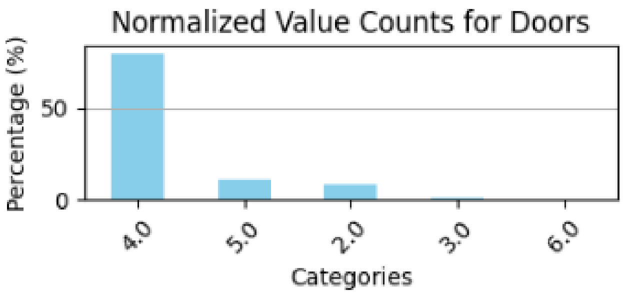
Many 2019-2020 Vehicles are priced between $0 and $20,000.

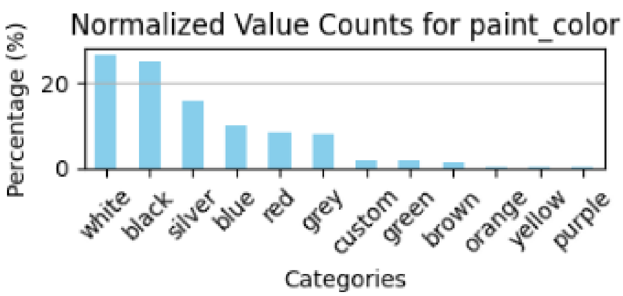
Cars with < 50,000 miles are priced < $20,000.

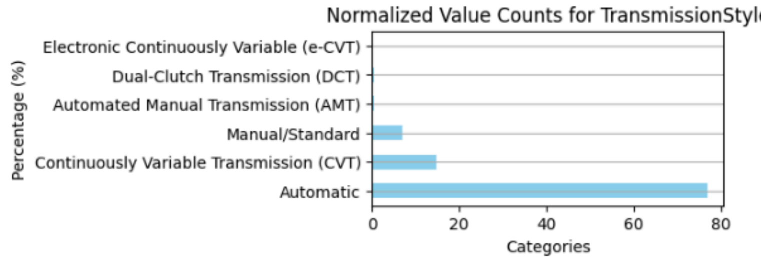
Way too many vehicles at all years and mileages are priced at near $0.

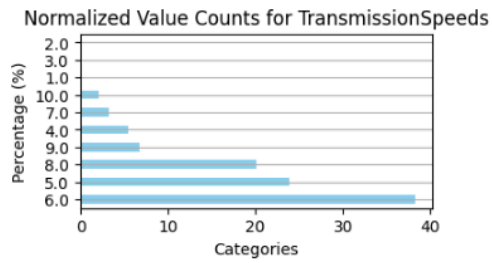
Odometer and Year are highly collinear. Older Vehicles tend to have more miles.

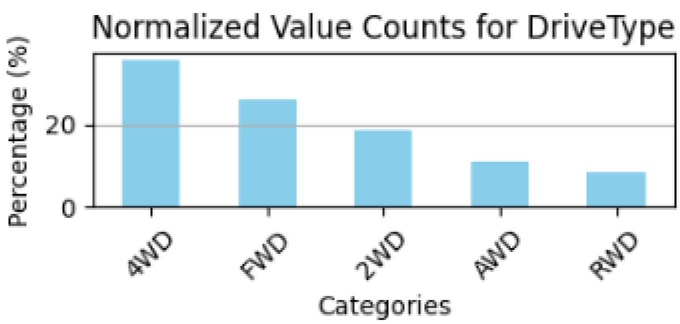
Value Counts

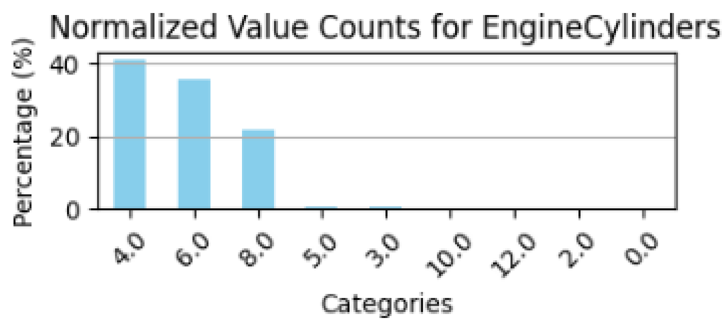




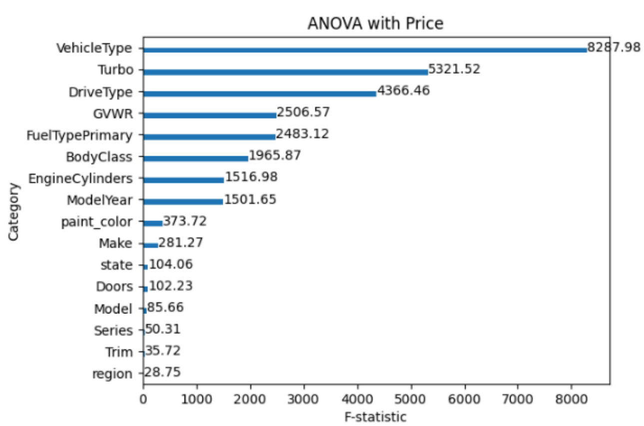








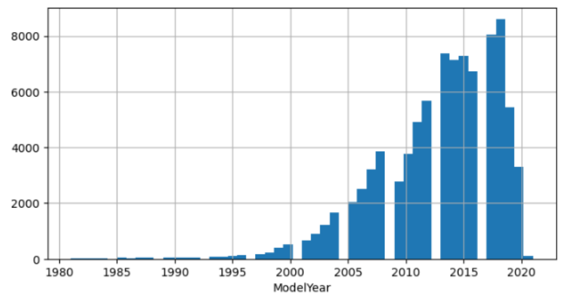
Most Listings are 4 Door, 6 Speed, Automatic, 4Cylinders, White or Black, 4WD.

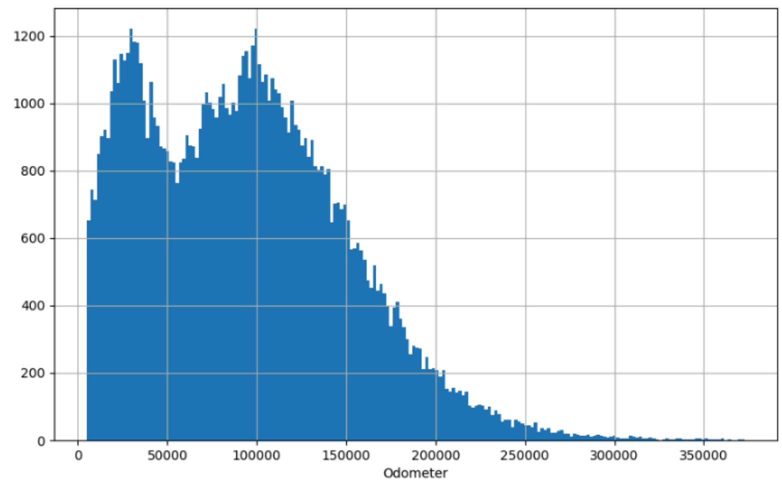


VehicleType, Turbo, DriveType, GVWR, FuelTypePrimary: **highly** correlated with price.

Make, Model, Region, State not so correlated, but still statistically significant.

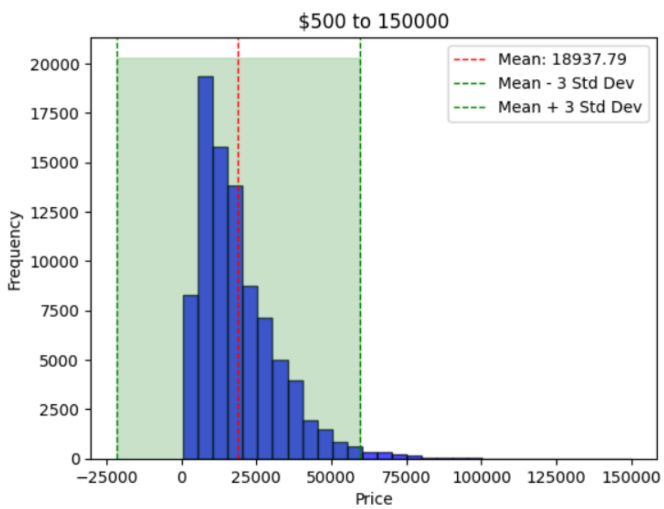
Histograms



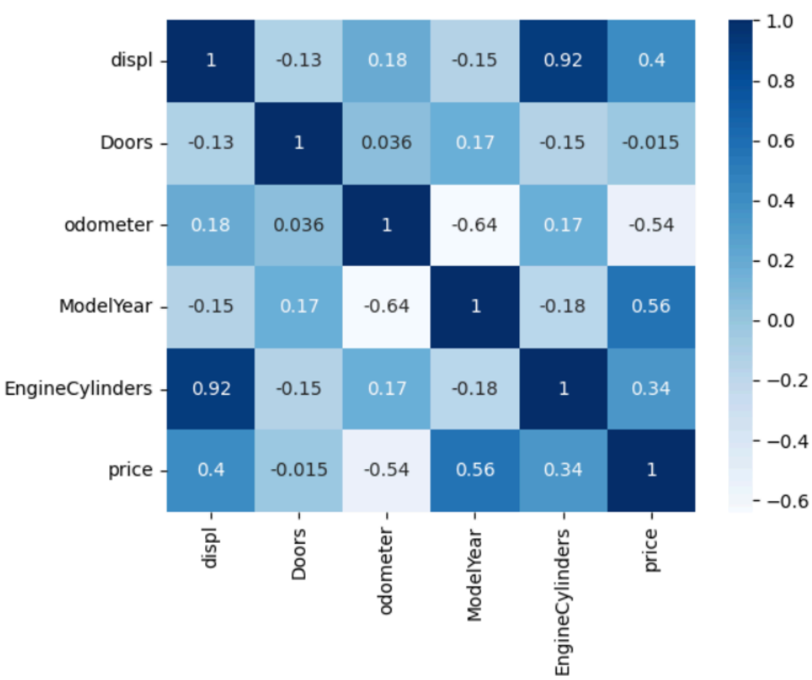


Odometer: New Vehicle hump at 25,000 Miles, Used Vehicle hump at 100,000 miles

Model Year: Some years with very low numbers



Mean Price: A bit high for used vehicles.



:

<u>Price</u>

Year: 0.56 & Odometer: -0.54

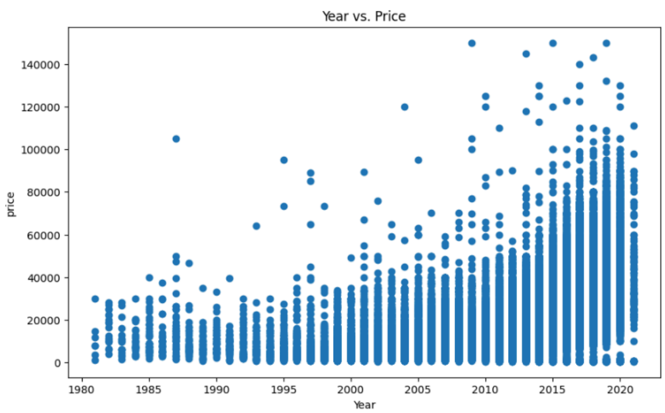
Collinearity: **-0.64**

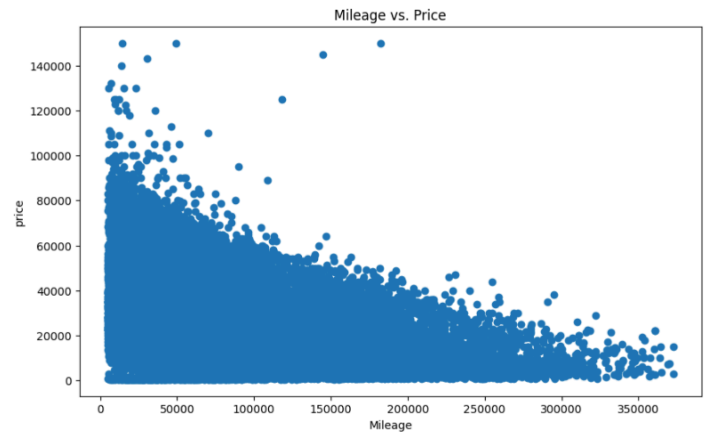
Displ: 0.4 & Cylinders : 0.34

Collinearity: **0.92**

Cylinders should be removed from the model

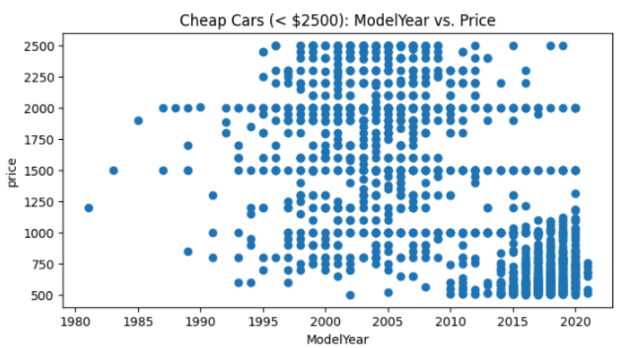
Scatterplots

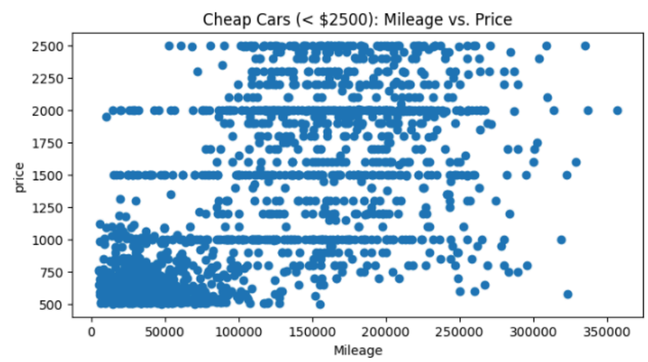




Problem: $500/ 0 Miles / 2020:

Corner Cluster of Low Mileages, New Cars, Low Prices





Problem: No MSRP, Series/Trim, Add-ons

2012 Dodge Challenger

Trim	MSRP
SXT 2dr Coupe (3.6L 6cyl) 305 HP	$ 25,195
R/T 2dr Coupe (5.7L 8cyl) 375 HP	$ 29,995
SRT8 2dr Coupe (6.4L 8cyl) 470 HP	**$ 44,125**

```
{'BasePrice': '83.54%', 'Series': '41.31%', 'Series2': '81.77%', 'Trim': '45.08%', 'Trim2': '97.16%'}
```

```
AdaptiveCruiseControl
{'Optional': 4146, 'Standard': 3567, 'Not Available': 39}
nulls: 0.9143359154851757
```

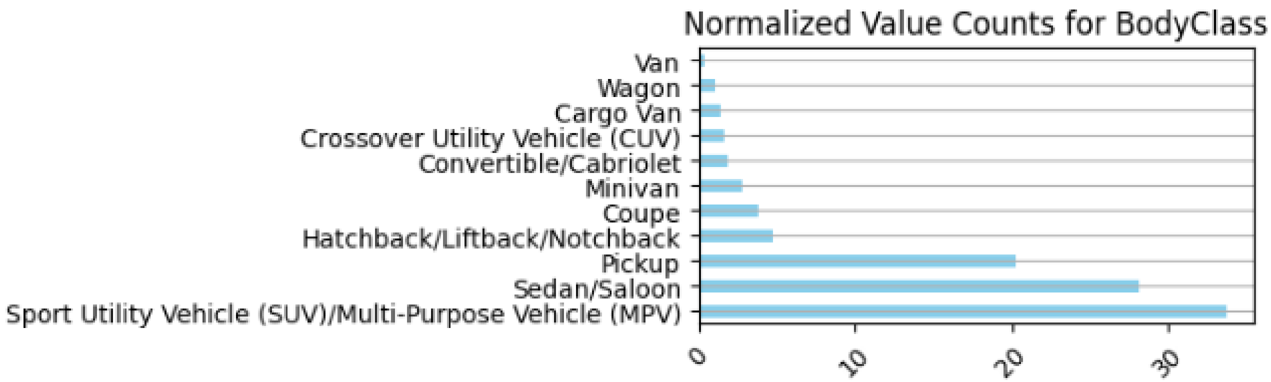
```
ABS
{'Standard': 22507, 'Optional': 18}
nulls: 0.7510857193374073
```

Series and Trim columns are **41 and 45% null.** Despite a smaller increase in engine size and horsepower, SRT8 price rises 300% compared to that of R/T Trim. Model does not have data to account for this.

Value Counts: VIN Decoder

70 Unique Makes (Chevrolet, Ford, ..)
878 Unique Models (F-150, Impala,..)

75% of BodyClass Values are: Sedan, Pickup, SUV



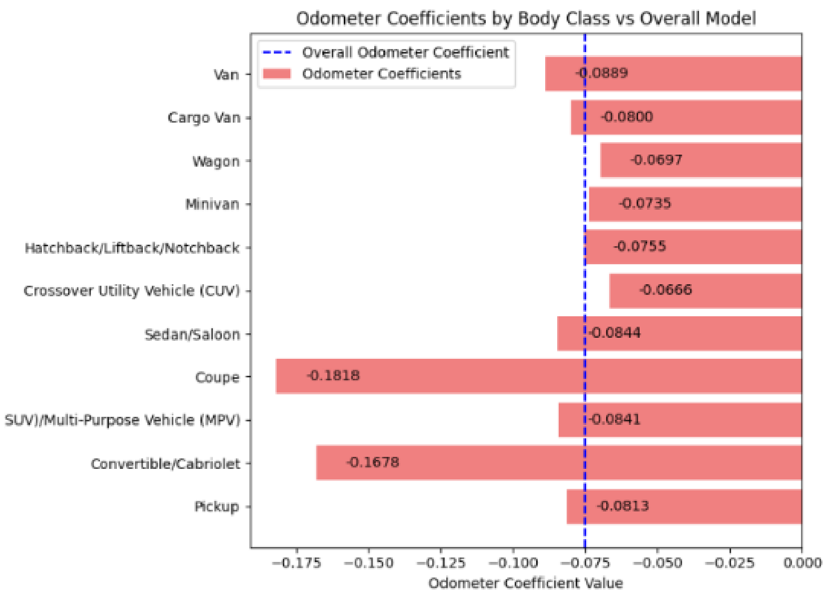
Weight Ratings (GVWR) are binned with overlap:

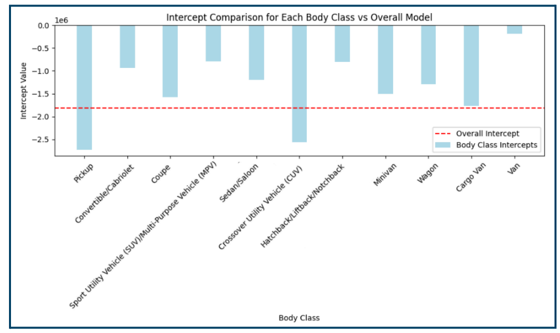
```
Class 1: 6,000 lb or less (2,722 kg or less)              32276
Class 2E: 6,001 - 7,000 lb (2,722 - 3,175 kg)             14567
Class 1D: 5,001 - 6,000 lb (2,268 - 2,722 kg)             13316
Class 1C: 4,001 - 5,000 lb (1,814 - 2,268 kg)             13182
Class 2F: 7,001 - 8,000 lb (3,175 - 3,629 kg)              7022
Class 2H: 9,001 - 10,000 lb (4,082 - 4,536 kg)             3093
Class 1B: 3,001 - 4,000 lb (1,360 - 1,814 kg)              2354
Class 2G: 8,001 - 9,000 lb (3,629 - 4,082 kg)              1709
Class 3: 10,001 - 14,000 lb (4,536 - 6,350 kg)             1628
Class 2: 6,001 - 10,000 lb (2,722 - 4,536 kg)               401
Class 8: 33,001 lb and above (14,969 kg and above)           18
Class 1A: 3,000 lb or less (1,360 kg or less)                11
Class 4: 14,001 - 16,000 lb (6,350 - 7,258 kg)                6
Class 7: 26,001 - 33,000 lb (11,794 - 14,969 kg)              2
Class 6: 19,501 - 26,000 lb (8,845 - 11,794 kg)               1
```

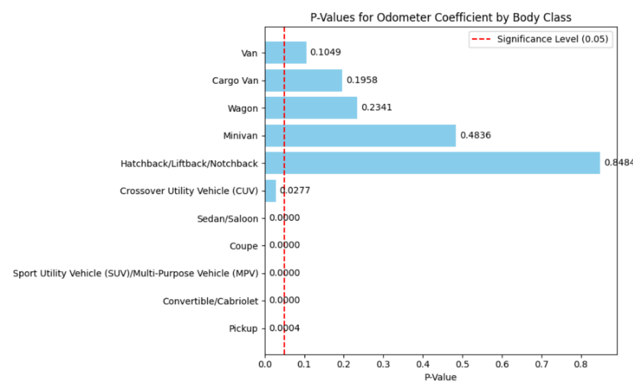
Most imputed at **<6,000lb**
More than half are in a specific range of 1,000lbs.

BodyClass Analysis

Depreciation Rates







Linear Regression on **price** with: **Year, Mileage**

All Vehicles compared to each individual BodyClass.

Keeping ModelYear as a factor ensured a fair comparison of odometer.

P-Values statistically significant: **Pickup, Convertible, SUV, Coupe, Sedan, CUV**

Faster Depreciation Rate:	**Slower Depreciation Rate:**
Pickup: 8.4%	Crossover 11.2%
SUV: 12.1%	
Sedan: 12.5%	
Convertible: 123.7%	
Coupe: 142.4%	

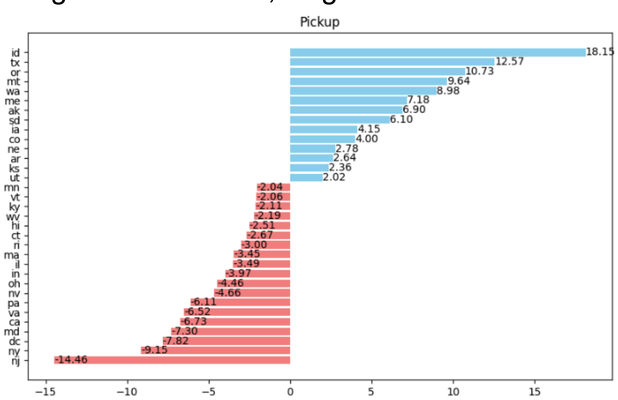
`'State' & 'BodyClass'` Chi-Square Test:

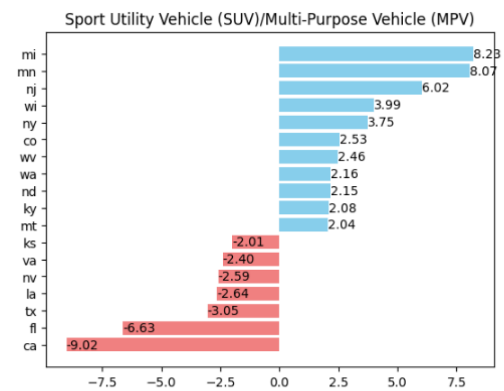
Chi-Square Statistic: 4550.524864966057
P-Value: 0.0
Degrees of Freedom: 500 (11 BodyClass, 51 States)
Critical Value: 554.5 (0.05 Significance Value)

Reject the null hypothesis: There is a significant association between State and BodyClass.

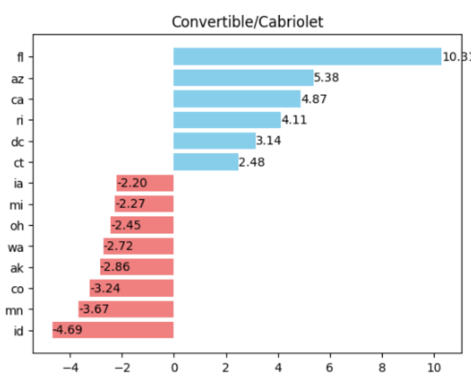
2) Create Contingency Table
3) Create Residuals for each Cell in Contingency Table

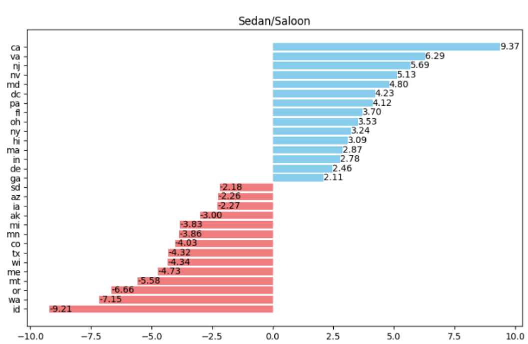
A residual greater than 2 or less than -2 suggests a significant deviation from independence. Positive = Higher Association, Negative = Lower Association



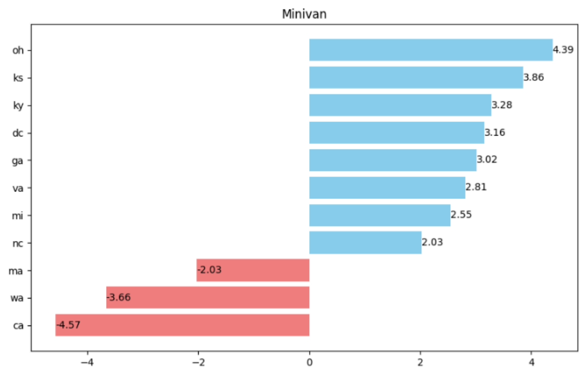


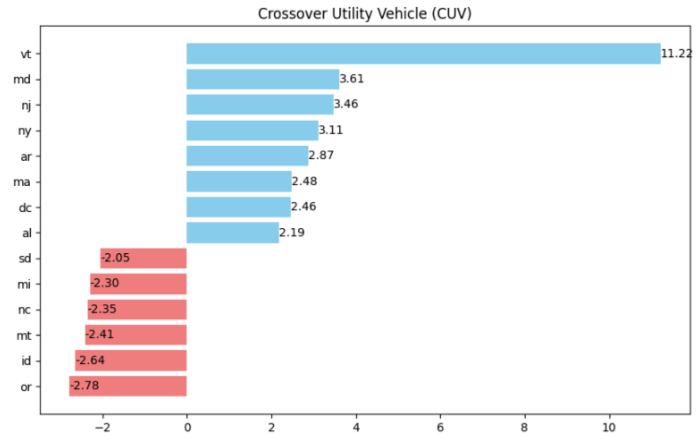
Pickup Trucks are the most variable in demand: Lower demand in Urban Areas: NJ, NY, DC
Increased demand in Midwest: Idaho, Oregon, Washington. Idaho only wants pickup trucks

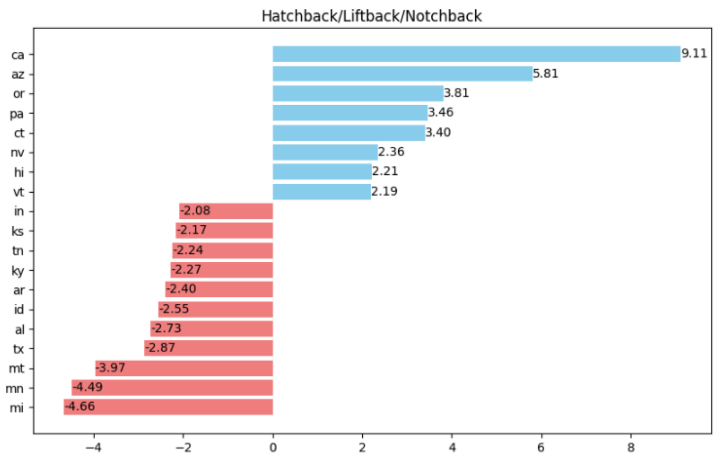




Florida: #1 Preference for Convertibles by a lot, Preference for Coupes, dislike of SUV & Wagon





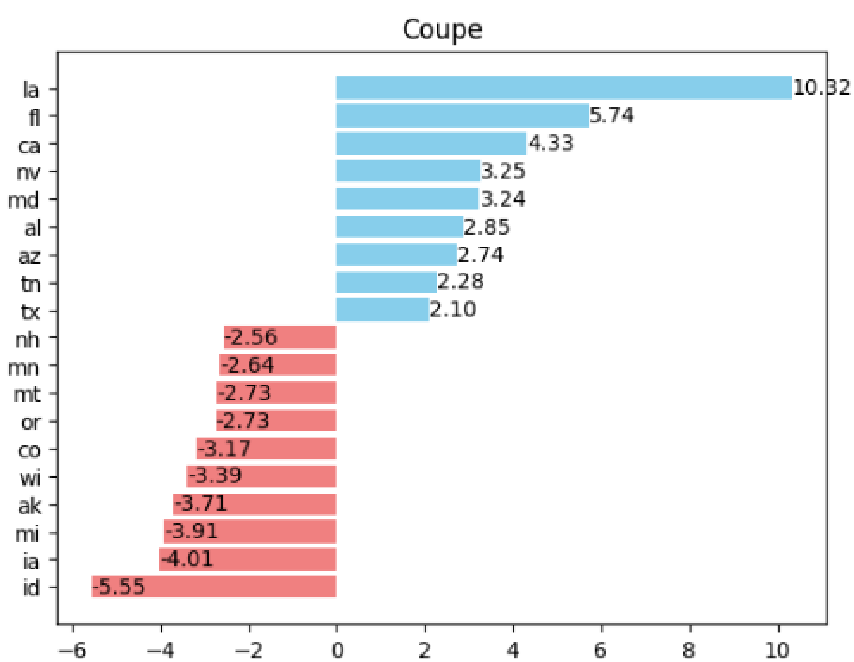


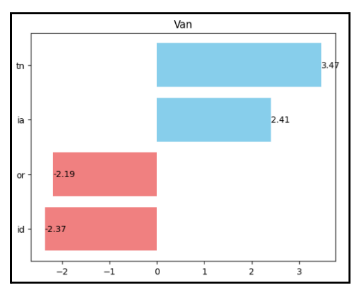
California: Preference for Hatchbacks & Sedan, Dislike of Pickup Trucks, SUV, Minivans

Idaho: Preference for Pickup Trucks,
Dislike of almost everything else

Louisiana: Extremely Favorable towards Coupes

Vermont: Extremely Favorable towards CUV

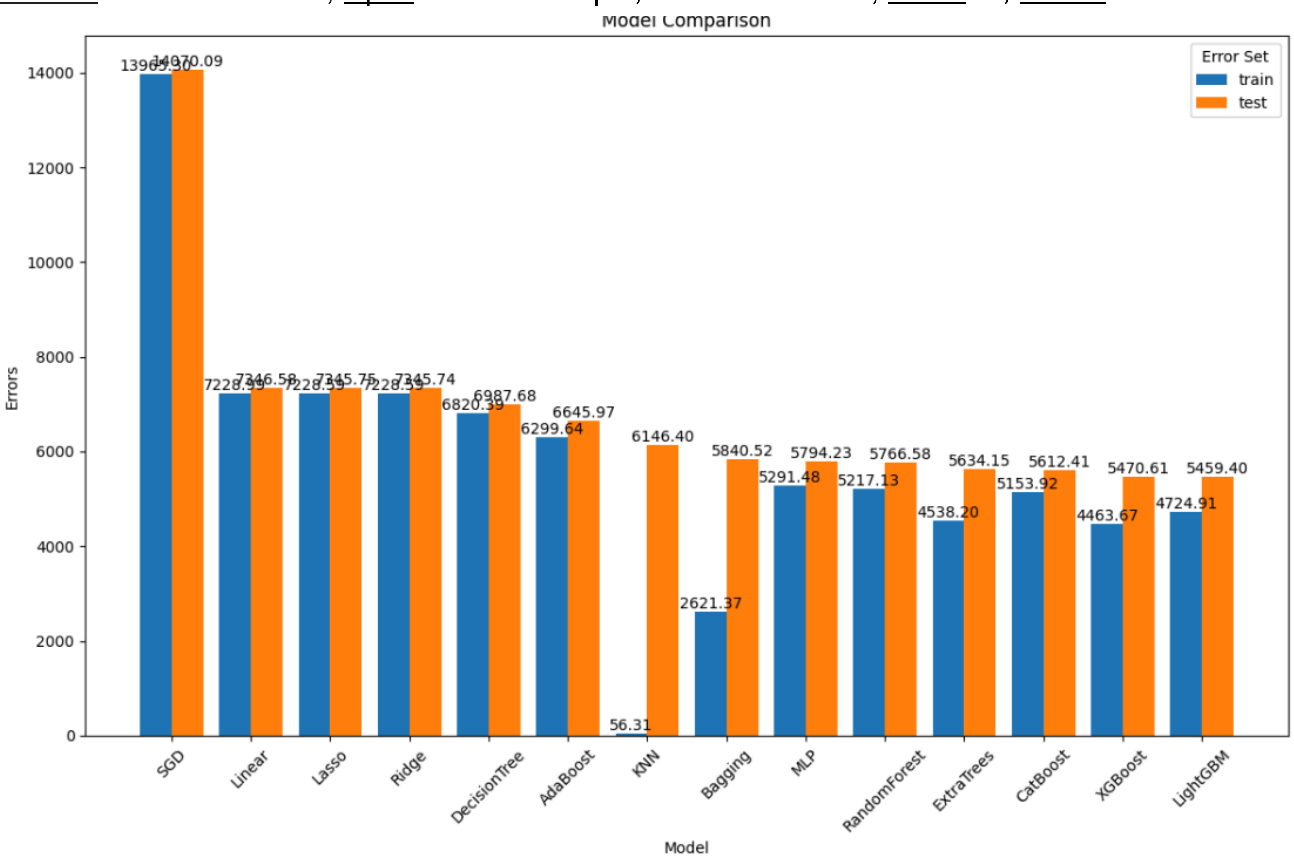




Modeling

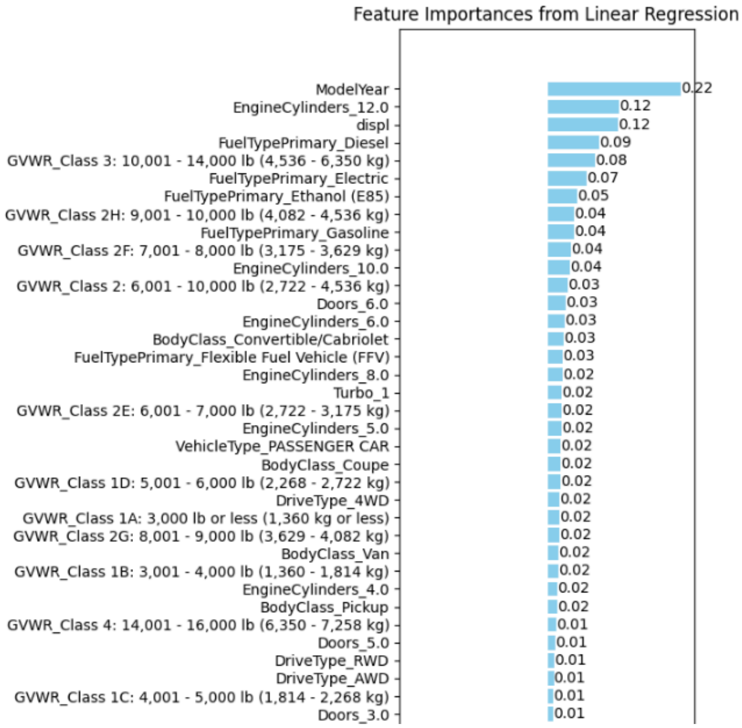
Features Used:
Categories: 'Turbo', 'BodyClass', 'FuelTypePrimary', 'EngineCylinders', 'VehicleType', 'DriveType','GVWR', 'Doors', 'paint_color'
Numerical Data: 'odometer', 'ModelYear', 'displ'

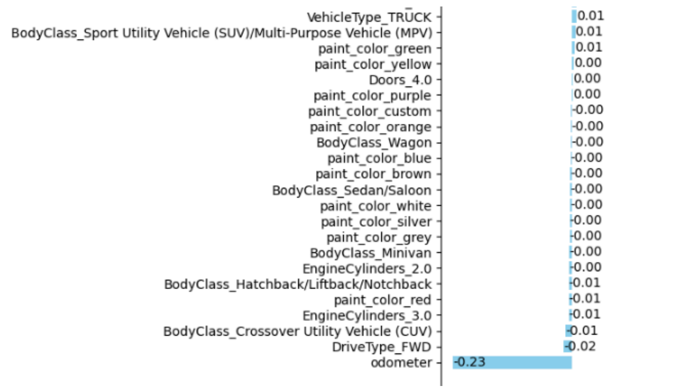
Scaler: StandardScaler, Split: Train Test Split, StandardScaler, Folds: 3, Metric: RMSE



Test RMSE: $5459.40

Feature Importances



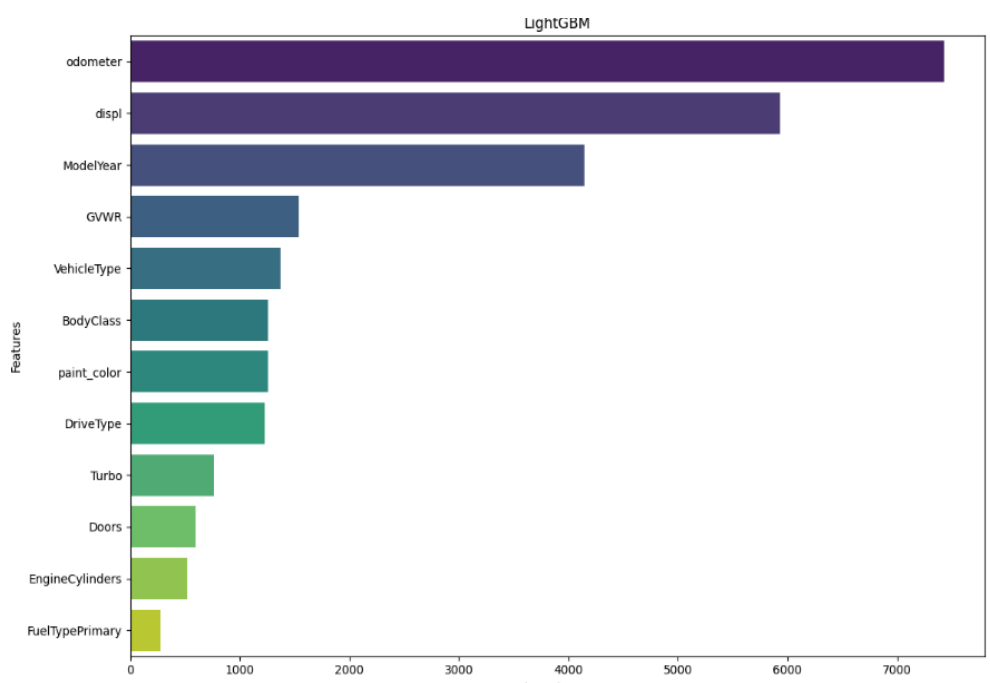


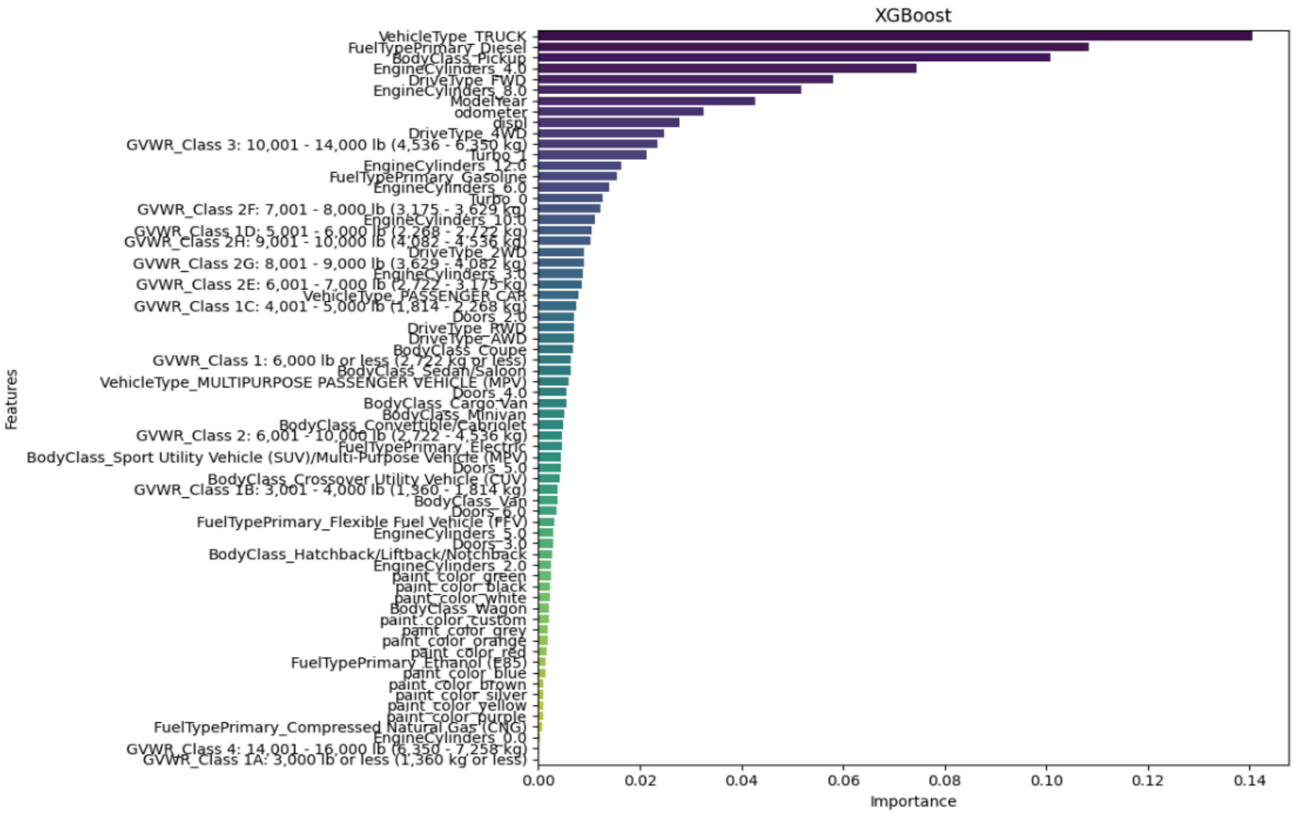
(Linear Coefficients)

Green cars are slightly higher priced.
Red cars are slightly lower priced.

Odometer is the most important feature. **FuelTypePrimary** found to be a more important feature for a linear model compared to Gradient Boosted Trees.

Odometer, displ (enginesize/displacement) and ModelYear proved to be the top Features throughout the entire Modeling process.





Diesel Pickup trucks are more highly in demand and higher priced than the competition.

Model Selection: Gradient Boosted Trees

XGBoost

- Dummy Variables
- Level-Wise Tree Building
- Imputes Missing Values Automatically
- Slower, More Memory

CatBoost

- Avoids Overfitting while handling data with many discrete values (such as Series, Trim)
- Learns Relationships between these categorical variables using their relative mean

LightGBM

- Fast Training, Prone to Overfitting
- Bins Continuous Variables
- Can automatically bundle features that are rarely non-zero at the same time, thus reducing dimensionality and speeding up computations.

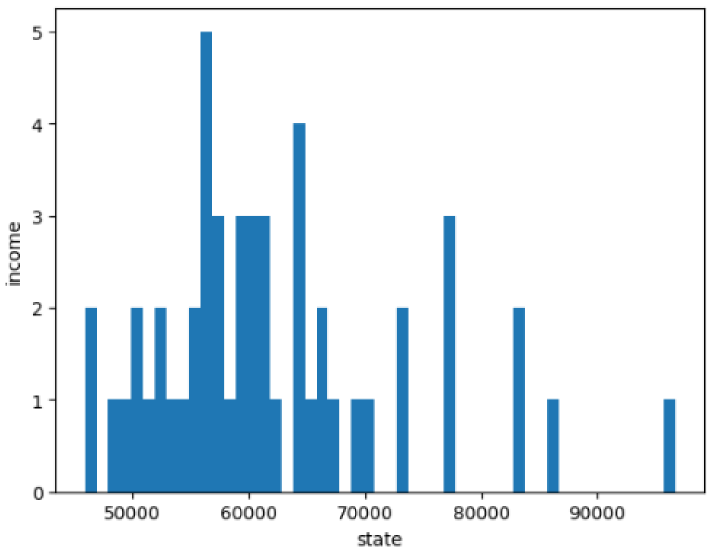
New Features

Local Economic Factors

How important is the specific Make/Model of the vehicle as opposed to competing Makes and Models that share the same BodyClass, Engine Size, EngineCylinders, and other features?

Does the specific state/region affect the local market for vehicles?

2021 Median state Income



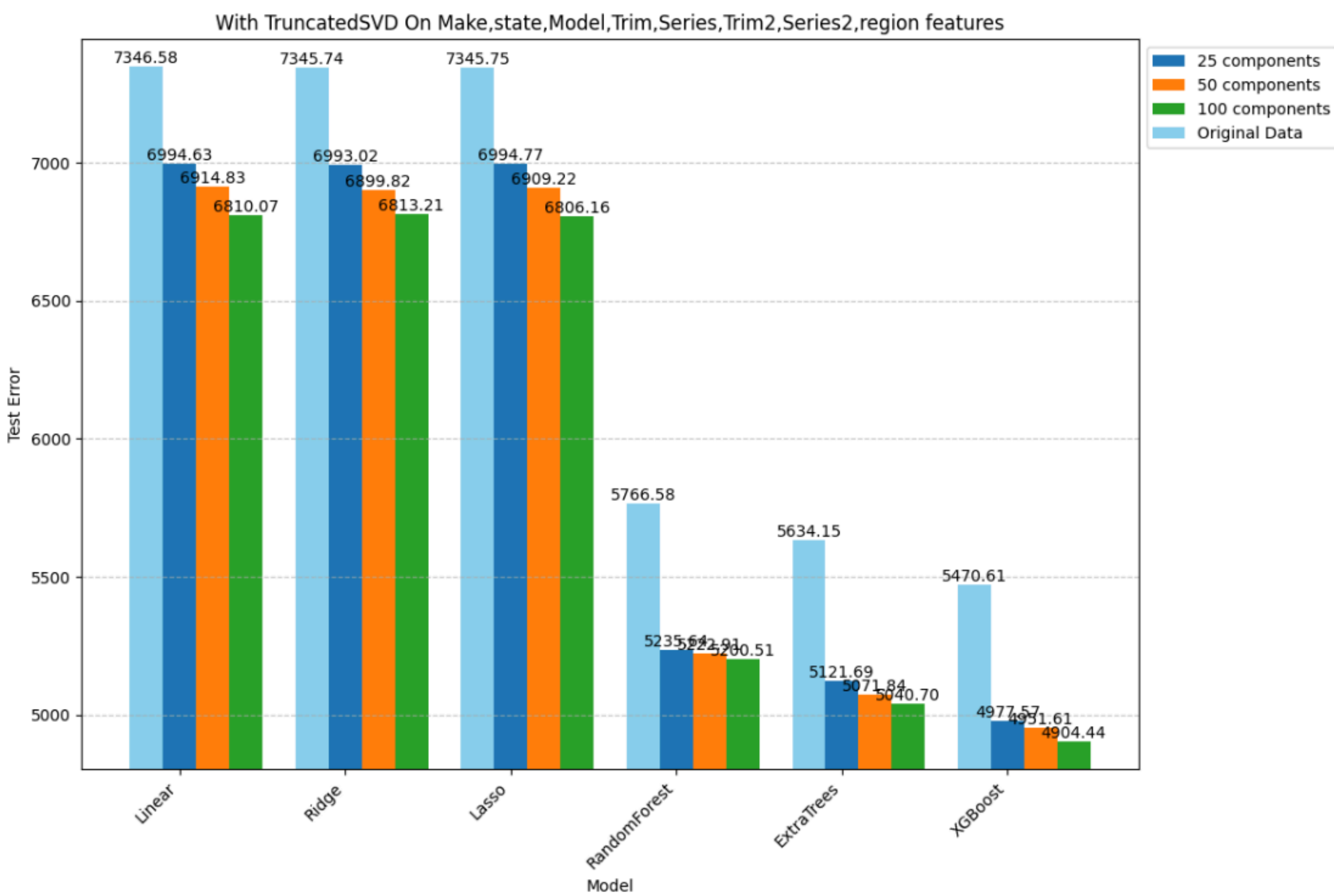
	feature	unique_vals
4	Series	1748
3	Trim	1737
2	Model	815
7	region	399
6	Series2	184
5	Trim2	76
0	Make	57
1	state	51

Make, Model, State, Region, Series, Trim Analysis

These columns introduced **5,067** new features. TruncatedSVD 'squashed' the variance of these columns into < 100 components, and the results were compared to the results without these features, seen above.

Answers:

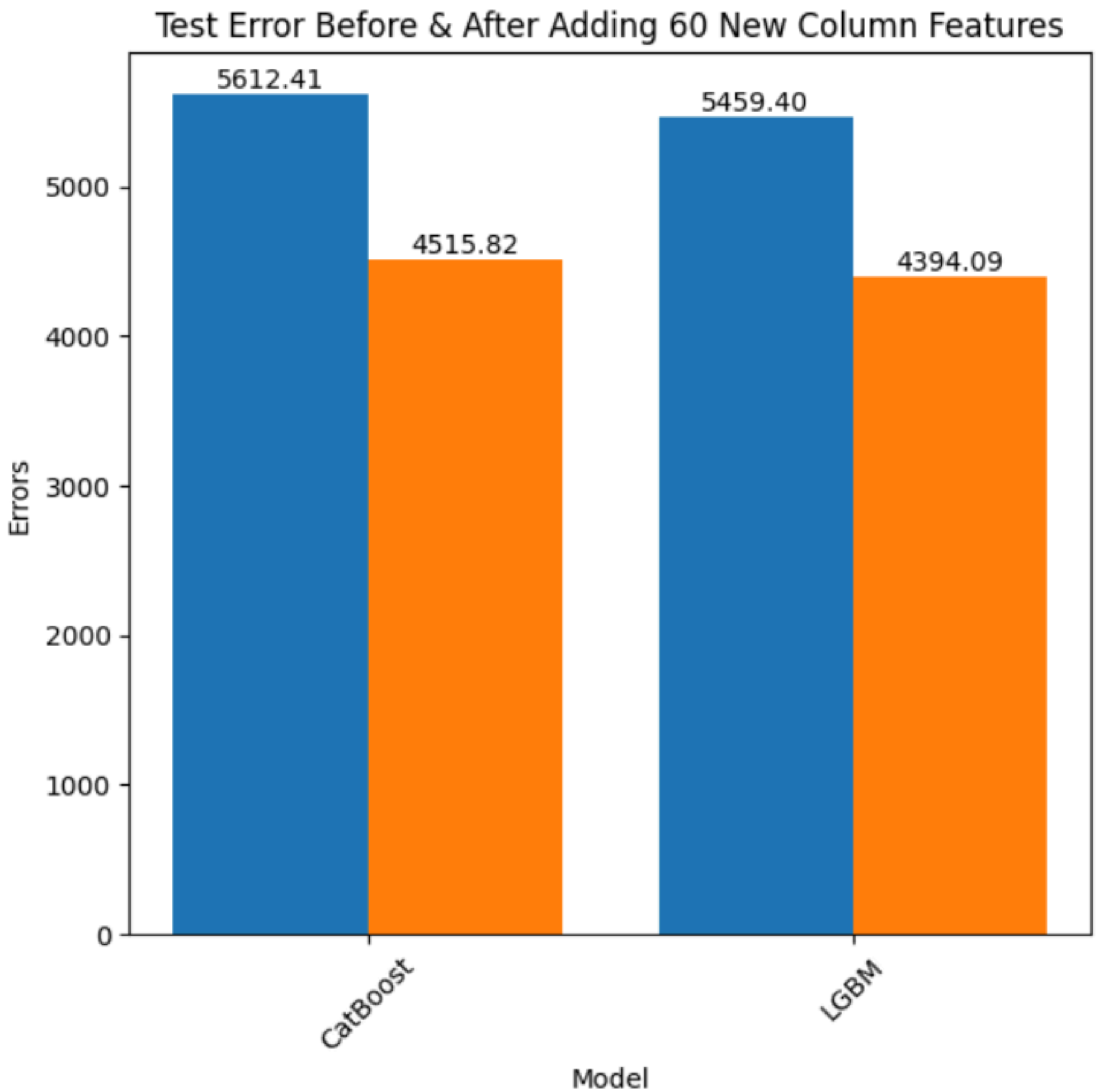
1. Do region & state play a role in vehicle prices?
2. Are certain Makes/Models more/less valuable than the competition within BodyClass?



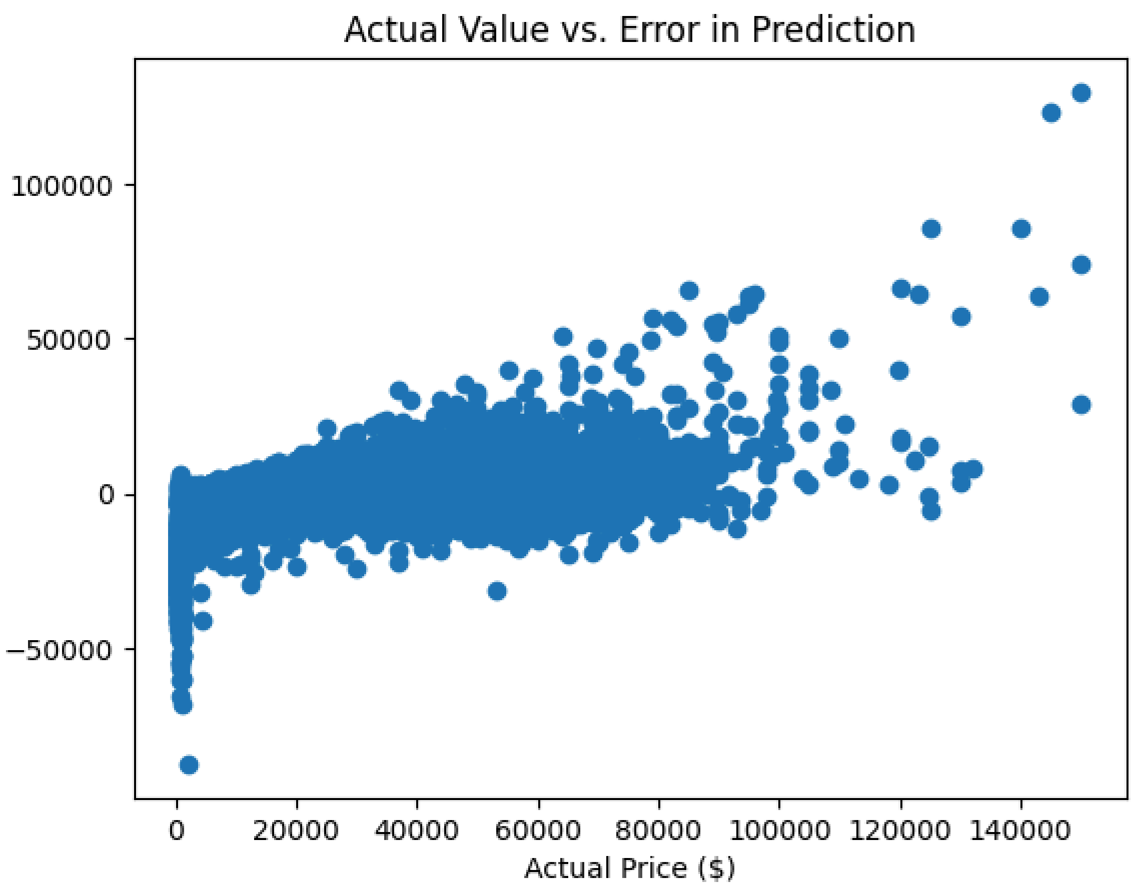
RMSE decreases about 7.5% when squashing variance down to 100 components.

Final Models: CatBoost & LightGBM

- These two models do not require dummy variable encoding.
- CatBoost finds the relationship between categorical variables using integer encoding, which has proven upsides.
- LightGBM performs feature selection within the training process. These are naturally a good fit for the dataset which has a large number of sparse, inconsistent categorical variables.



Residual Analysis



While it makes sense for errors to increase as the price increases. Vehicles priced below **$5000 should not be estimated to be $50,000** or more. The deviation from the larger trend can be seen in the bottom left of the graph as the dots fall in a straight line.

The dataset came with a 'description' column that confirmed the presence of:

1. 'Down Payment' in 'price' Listings

These are generally newer vehicles that are higher-end, and the dealer has listed just a portion of the total price in the 'price' column. Exploring the 'description' column of high percentage errors can confirm there are hundreds of these listings, with errors in prediction sometimes above 4000% and listed prices generally below $3000. These listings are best **removed** from the dataset

2. Significantly Damaged Vehicle Listings

These vehicles show very similar errors as the 'Down Payment' listings. Typically listed for $5000 or less. The description will contain words such as: 'sold as-is', 'doesn't run', or 'for parts'. These listings are best **removed** from the dataset

Other Large Errors:

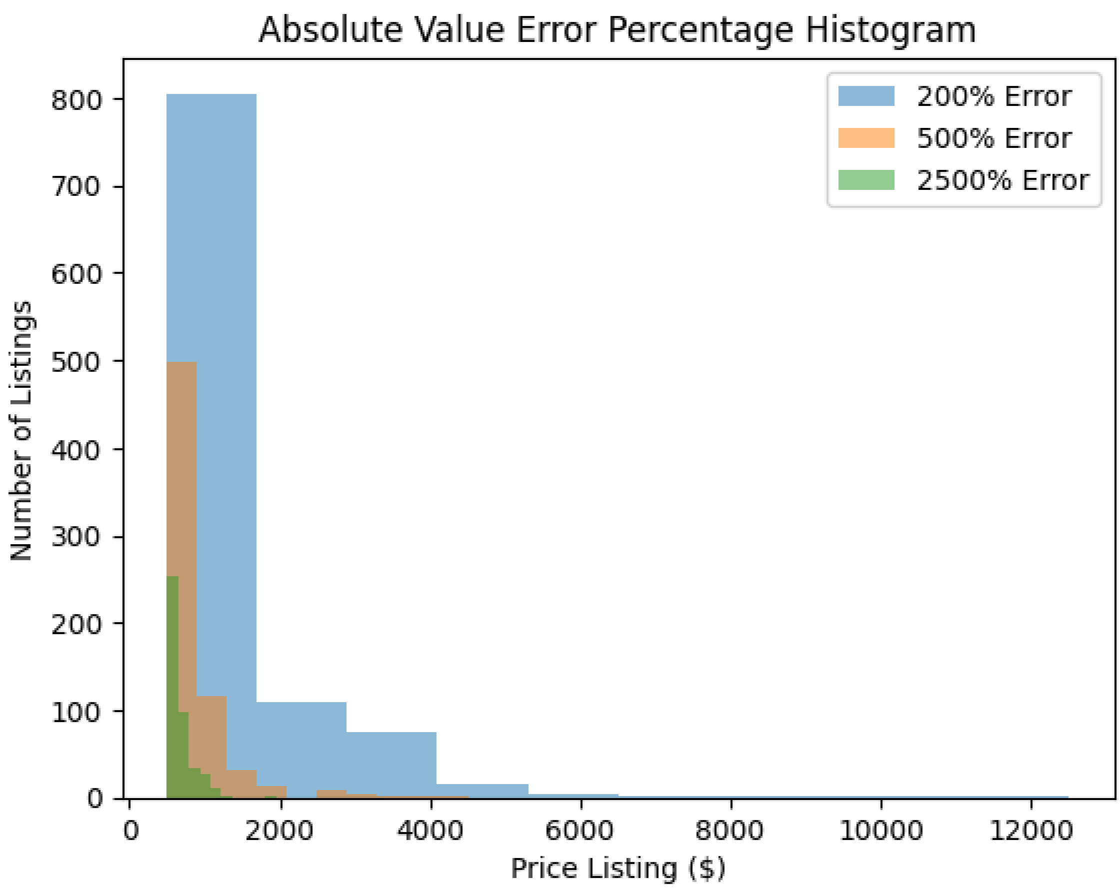
3. Significantly Modified Vehicles

These are vehicles where the car owner put significant investment into upgrading various aspects of the car appearance. The error percentage was not quite as high, with the list price being just a tad above average. Frequently, listings were sports cars and convertibles.

4. Cargo Vans

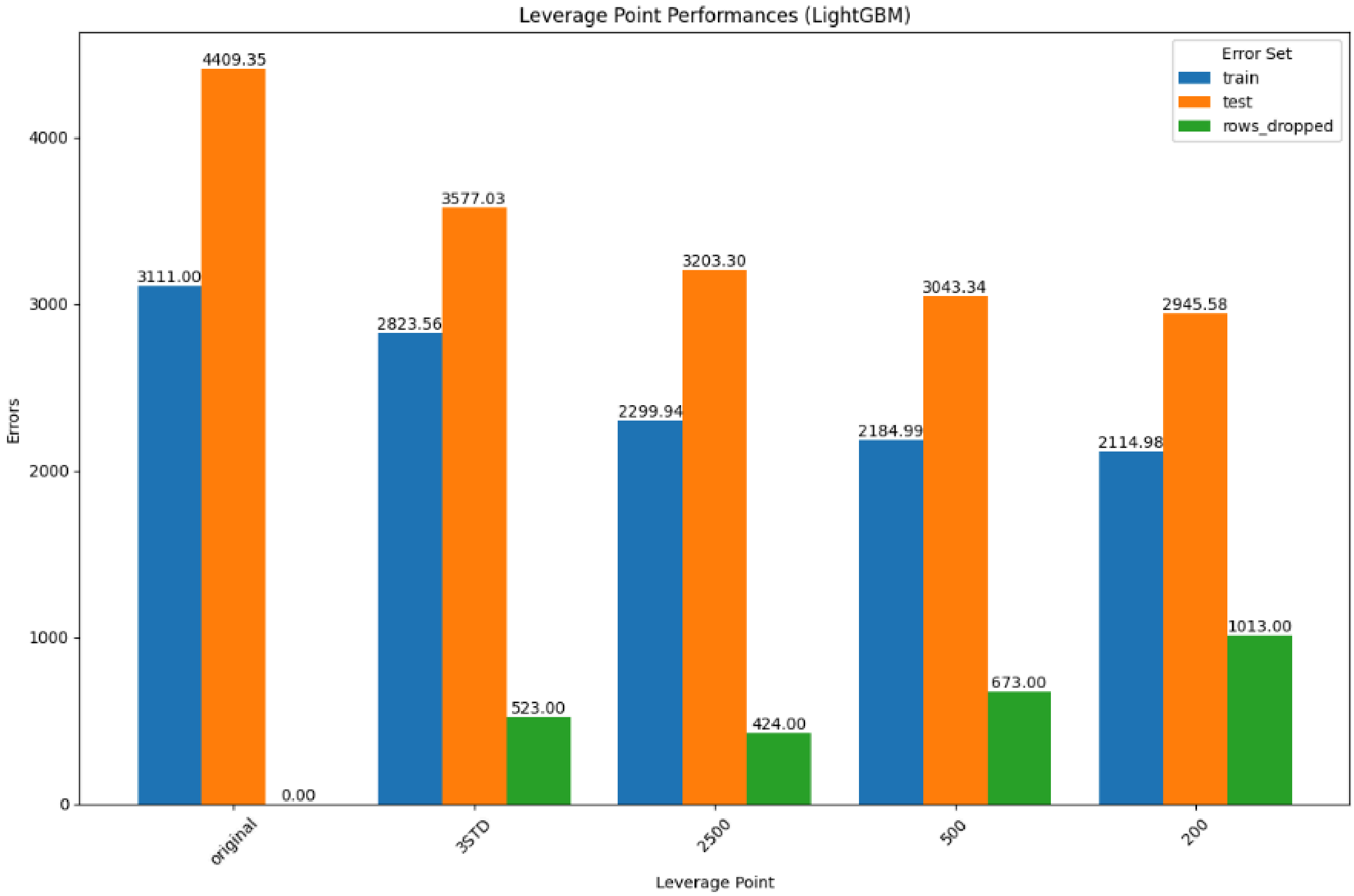
These listings showed high errors as they were a minority of the dataset. It may be that there is a high variance to the depreciation on liveable vehicles such as RVs and Vans, either holding their value despite high miles, or wearing down and requiring significant upkeep investments to maintain the vehicle. Future explorations of this data should involve SMOTE on Vans and Cargo Vans to train on balanced classes.

```
Number of Listings with |200| % prediction error or more: 1013
Number of Listings with |500| % prediction error or more: 673
Number of Listings with |2500| % prediction error or more: 424
```



Spam Detection

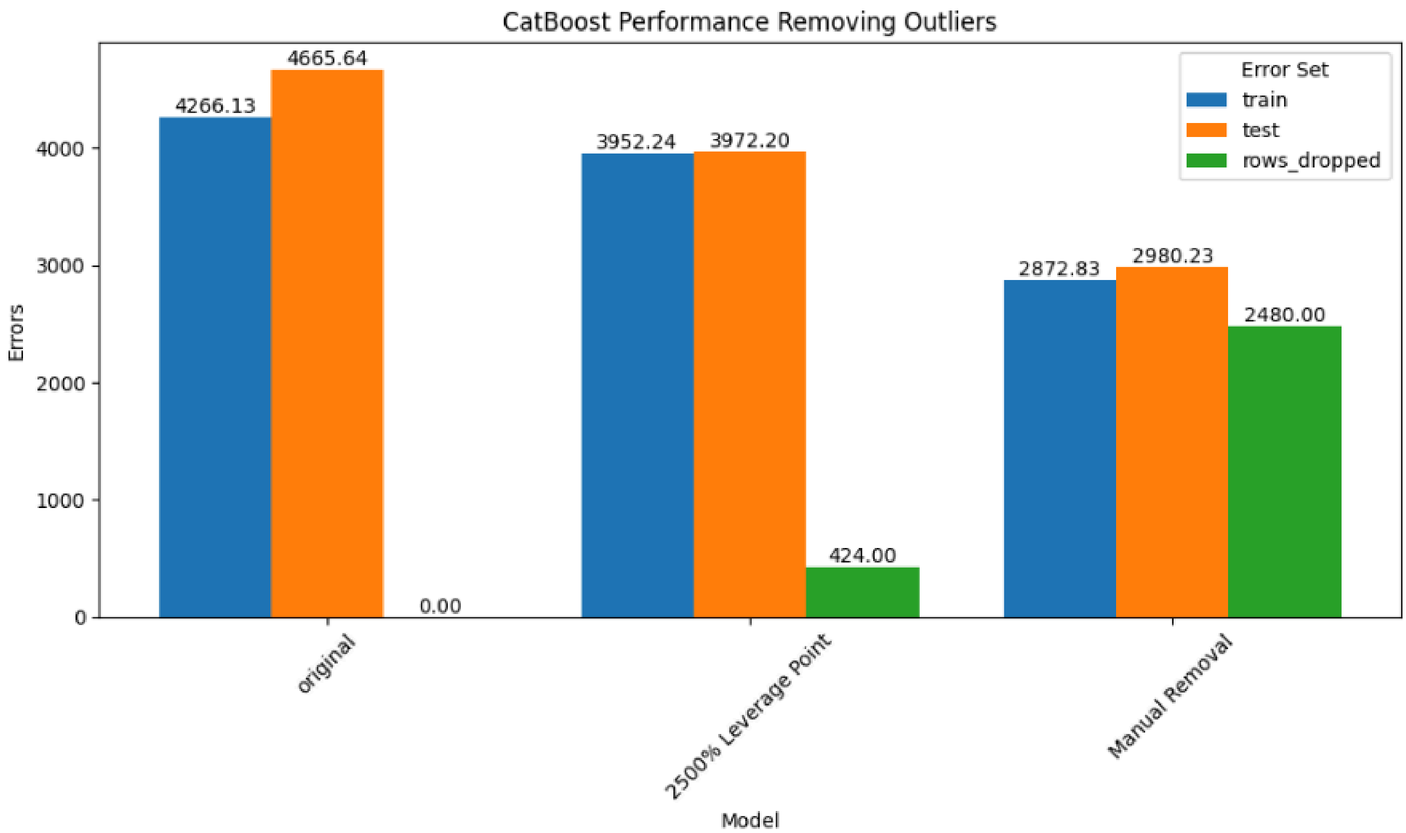
3 Standard Deviations from Mean Absolute Percentage Error
vs.
200%, 500%, 2500% or greater Absolute Percentage Error



Simply using a >2500% Percentage error threshold **retained 99 more rows** of data, while **improving error reduction by over $300**

Manual Outlier Identification:

By running numerous tests, similar to the above, and inspecting 'description' columns for indications of cars that did not run or cars that were listed as down payment or cars with heavy modifications, **2480** rows were removed.



This **reduced $0.50 in errors per row**.
Notebook: [Modeling]

XGBoost:

Top 10 Overall Feature	**Importance**
TractionControl: **Standard**	0.1523
RearVisibilitySystem: **Standard**	0.0674
SemiautomaticHeadlampBeamSwitching: **Standard**	0.0522
BodyCabType: **Crew/SuperCrew/CrewMax**	0.0500
ESC: **Standard**	0.0401
BodyClass: **Pickup**	0.0251
FuelTypePrimary: **Diesel**	0.0257
EngineCylinders	0.0186
VehicleType: **Truck**	0.0120
DayTimeRunningLight: **Standard**	0.0120

Top 5 Makes	**Importance**
Porsche	0.003383
Jeep	0.002809
Land Rover	0.002344
Mercedes-Benz	0.001840
RAM	0.001834

Top 5 Models	**Importance**
Porsche 911	0.005904
Jeep Wrangler	0.004085
Acura TLX	0.003151
GMC Sierra HD	0.002492
Cadillac XT6	0.002194

Depreciation Curve Experiment Procedure:

Train model without state, region, state_income.

Based on feature importance of a specific Model/Series/Trim, identify most similar vehicles to each unique vehicle (drop duplicates, subset on: control columns plus series/trim).

‘Control’ columns: displ, Turbo, VehicleType, GVWR, BodyCabType, EngineCylinders, BodyClass, & ModelYear

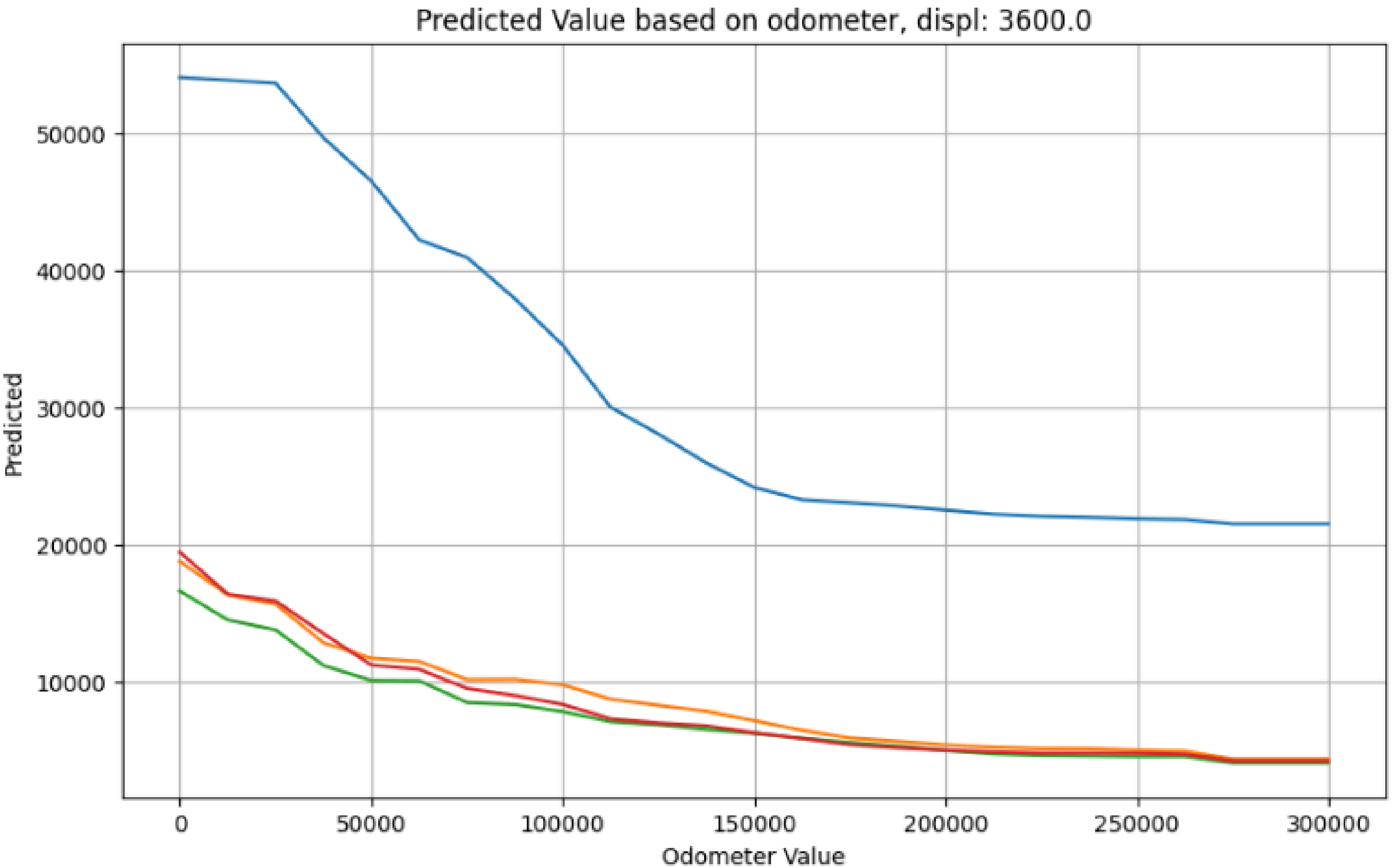
Variable: Odometer: 0 to 300,000 miles, new row every 12,500 miles

Create predictions for each vehicle at each odometer reading

Plot results

Porsche 911

2005



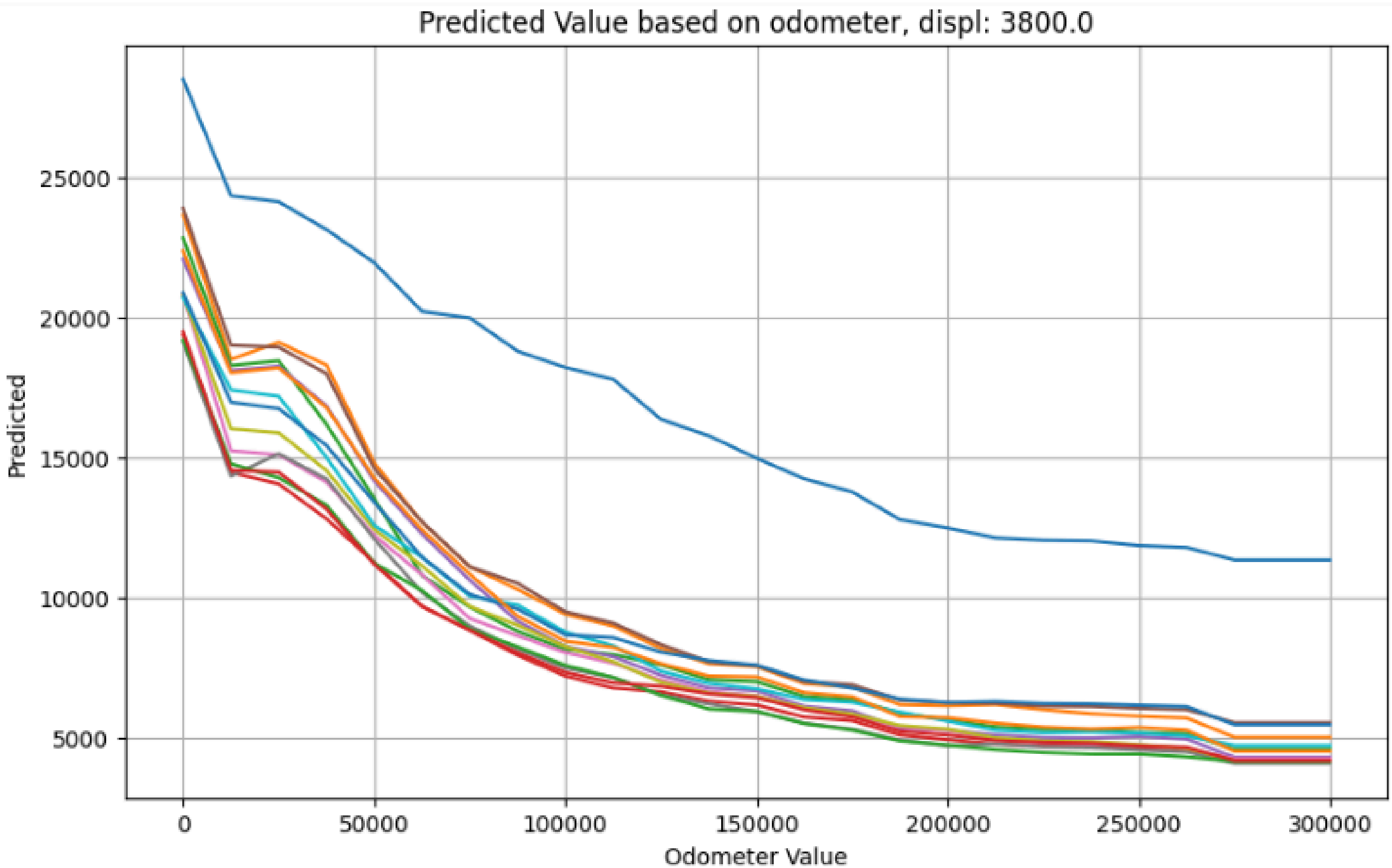
MSRP

Porsche 911 - $83,400

Infiniti G35 - $30, 700

Jeep Wrangler

2007 - V6



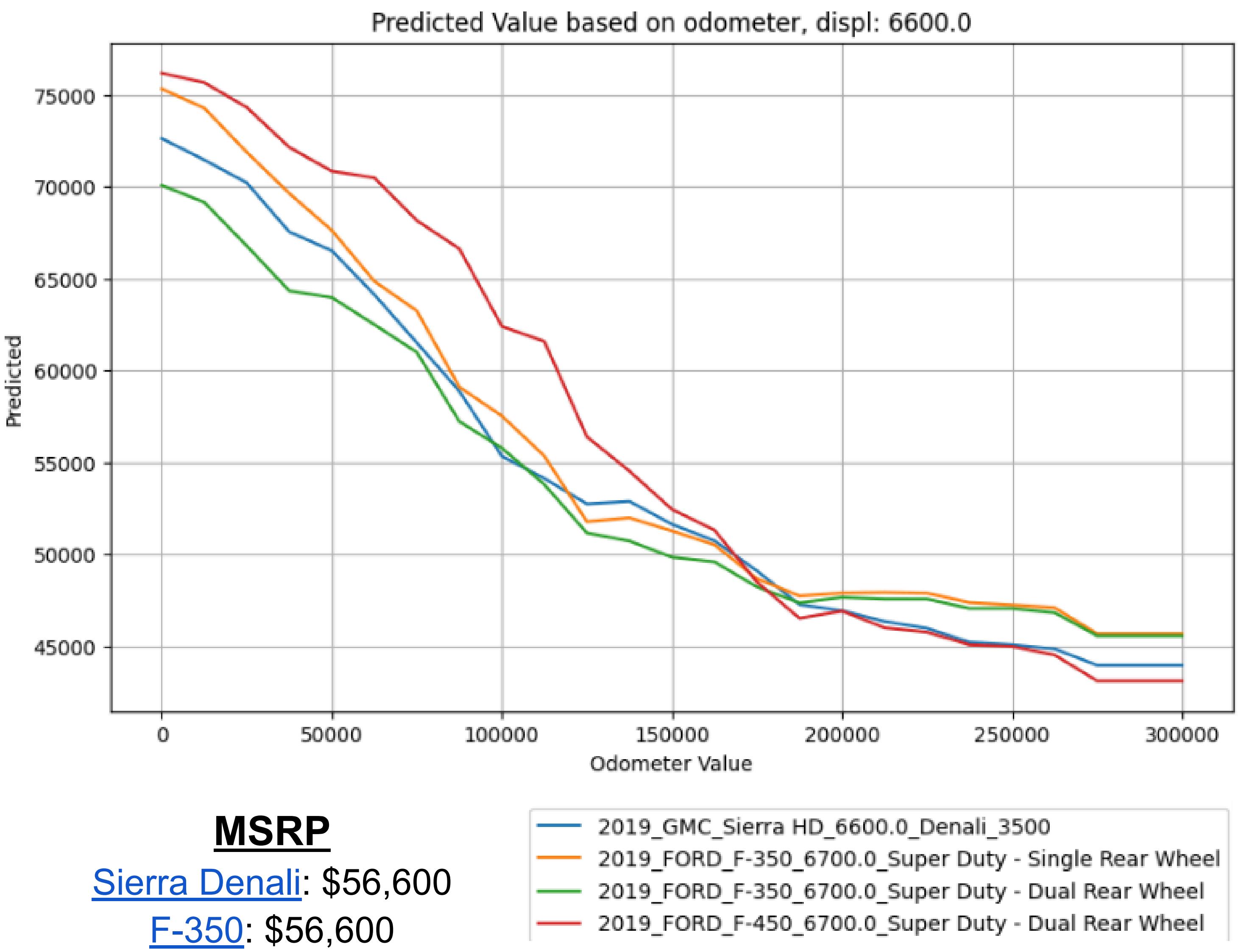
MSRP

Wrangler (Unlimited X): $22,530

Pathfinder: $25,000 - $31,000

GMC Sierra HD

2019 - 6.6L Turbo Diesel



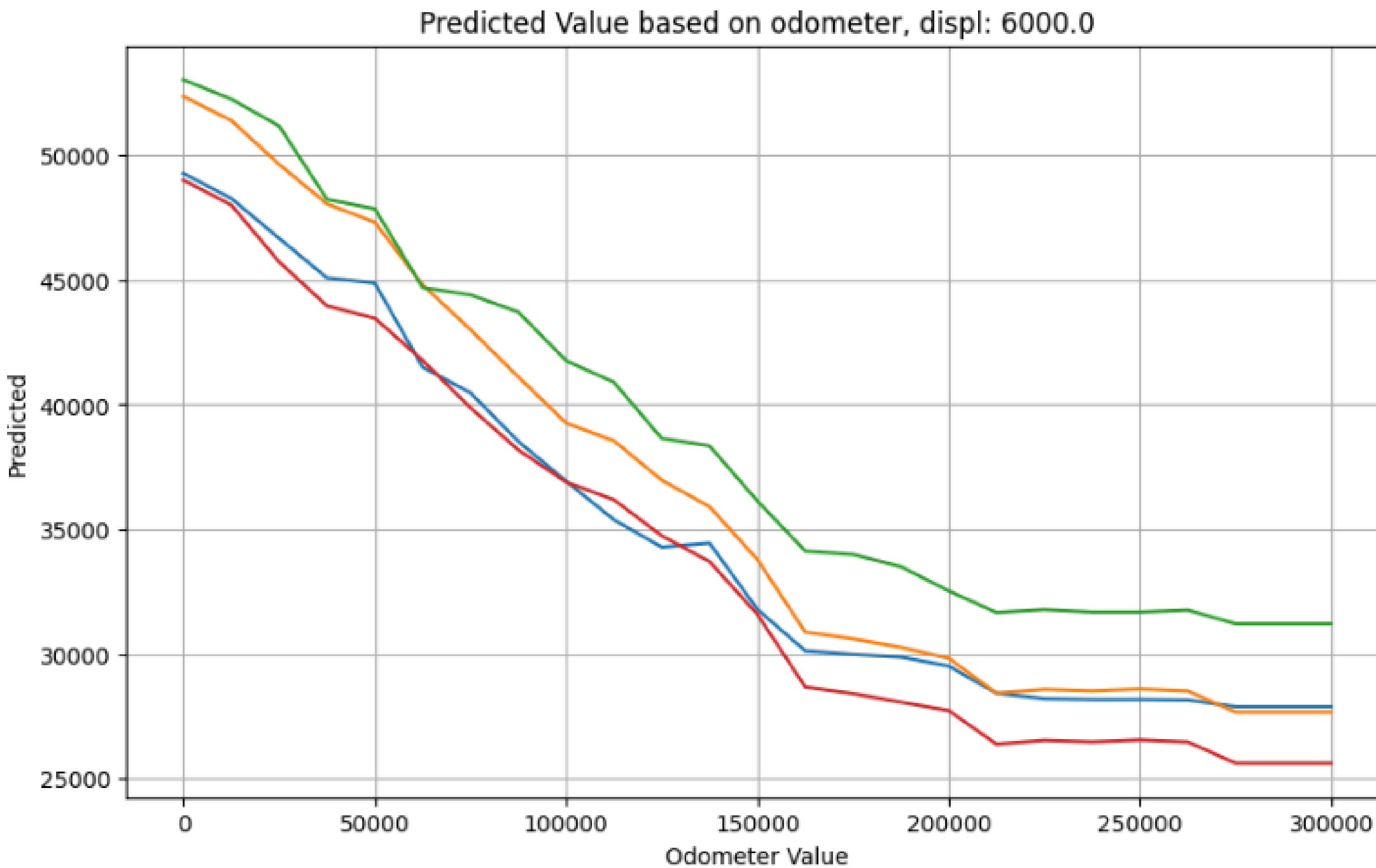
MSRP

Sierra Denali: $56,600

F-350: $56,600

F-450: $56,600

2019 - 6.0L Diesel

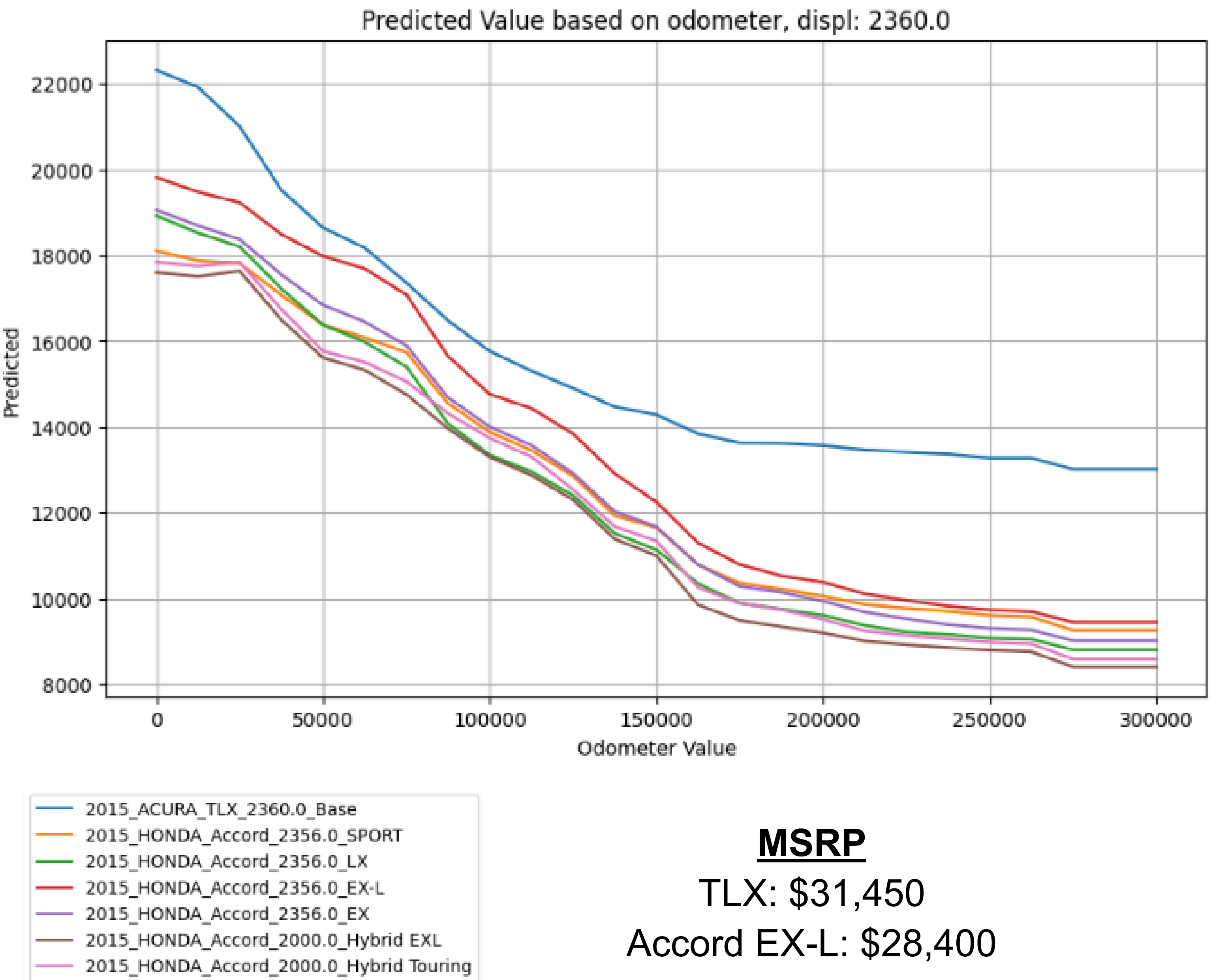


MSRP

Ram 2500 Tradesman: $39,850
Sierra 2500 Fleet: $40,000
Ram 2500 BigHorn: $42,100
Ford F-250 SuperDuty: $43,000

Acura TLX

2015 - V4

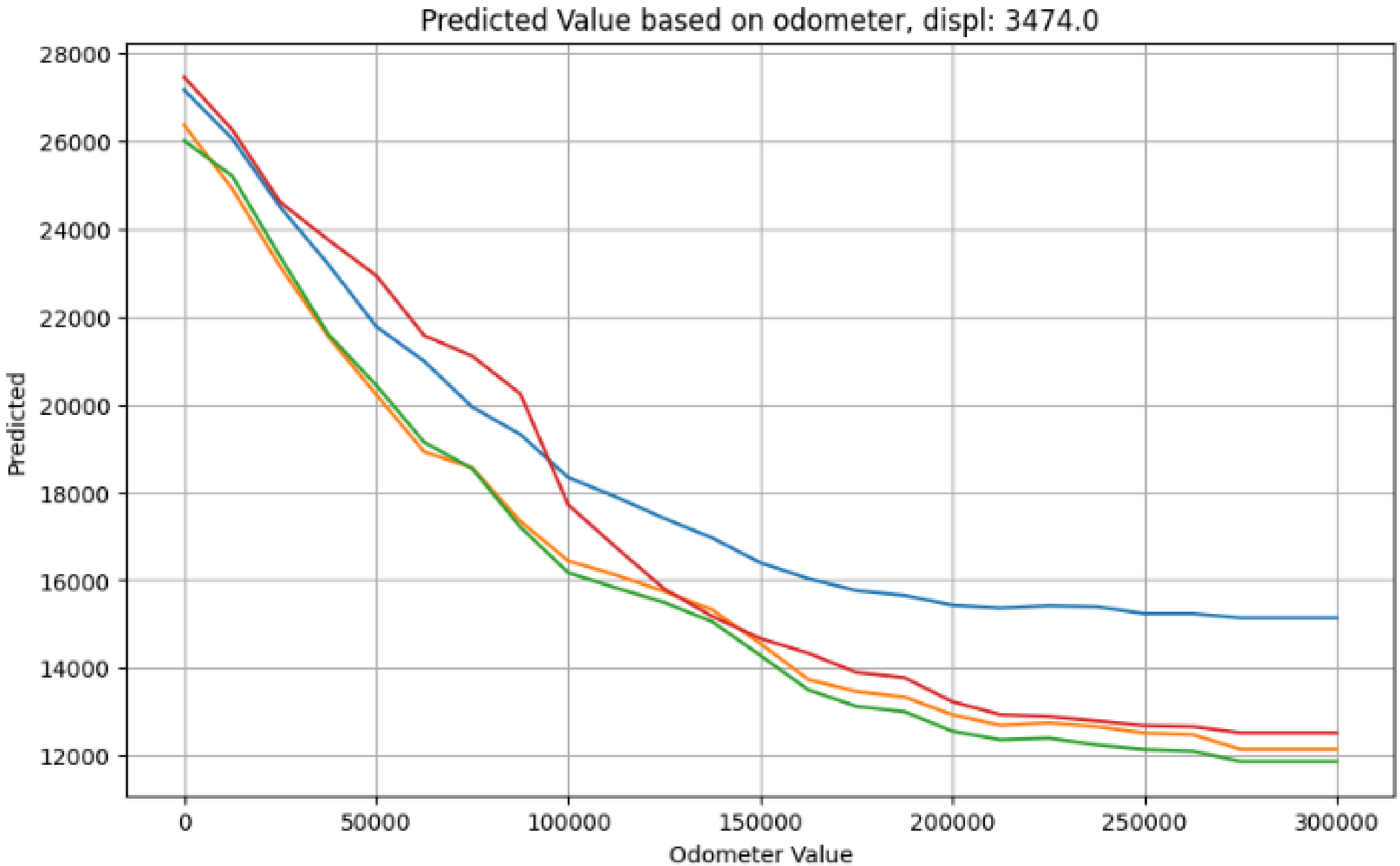


MSRP

TLX: $31,450

Accord EX-L: $28,400

2015 - V6

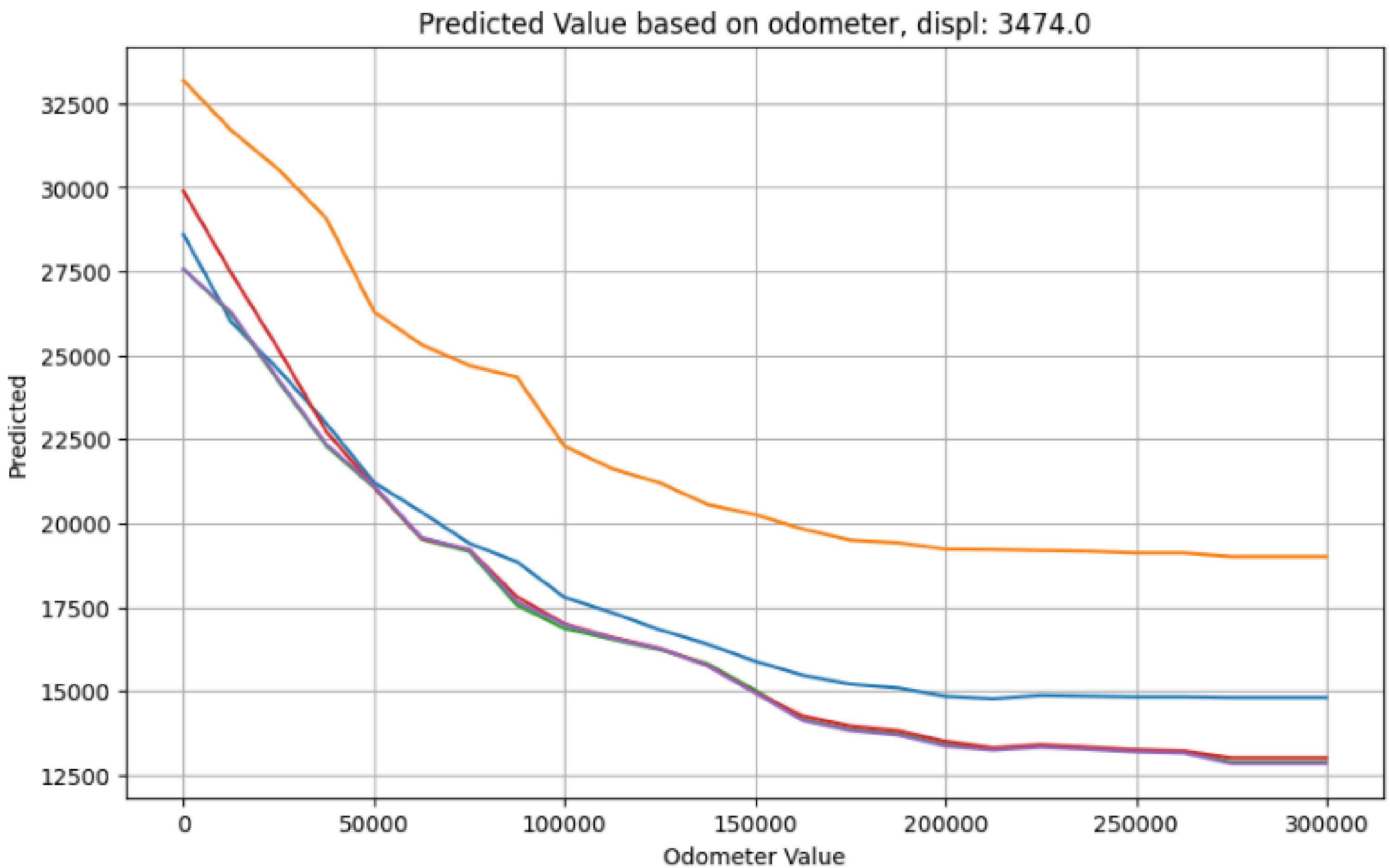


<u>MSRP</u>

RLX: $48,450

TLX: **$35,320**

2016 - V6

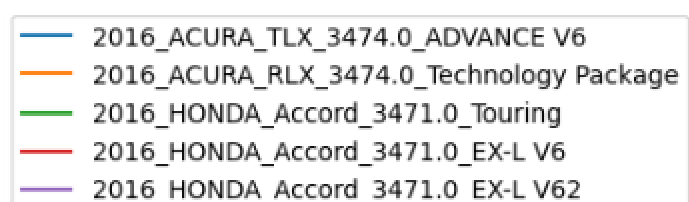


MSRP:

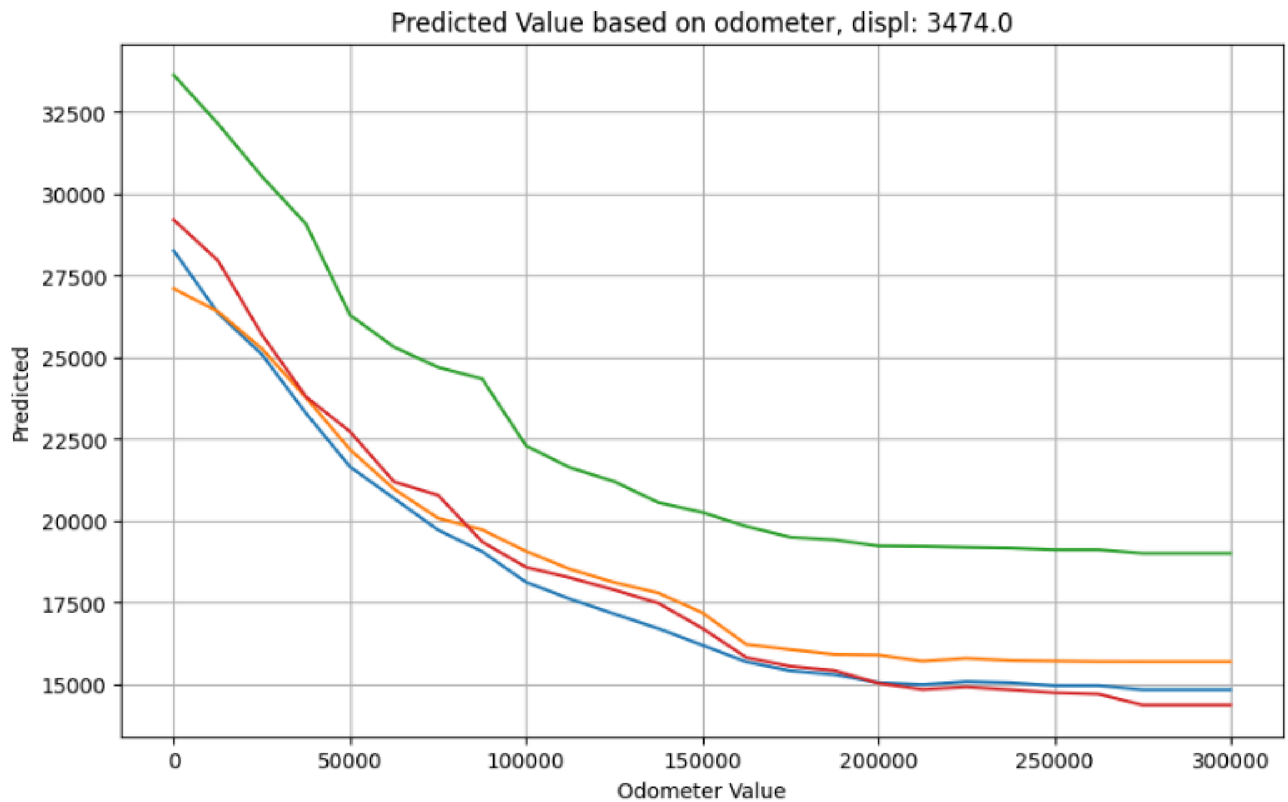
TLX Advance: $42,600

RLX Tech: $54,450

Accord EX-L: $30,740



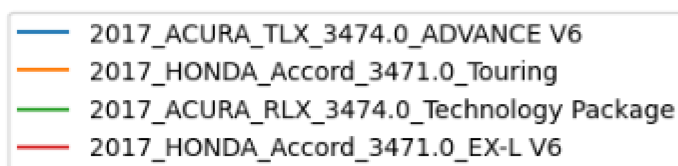
2017 - V6



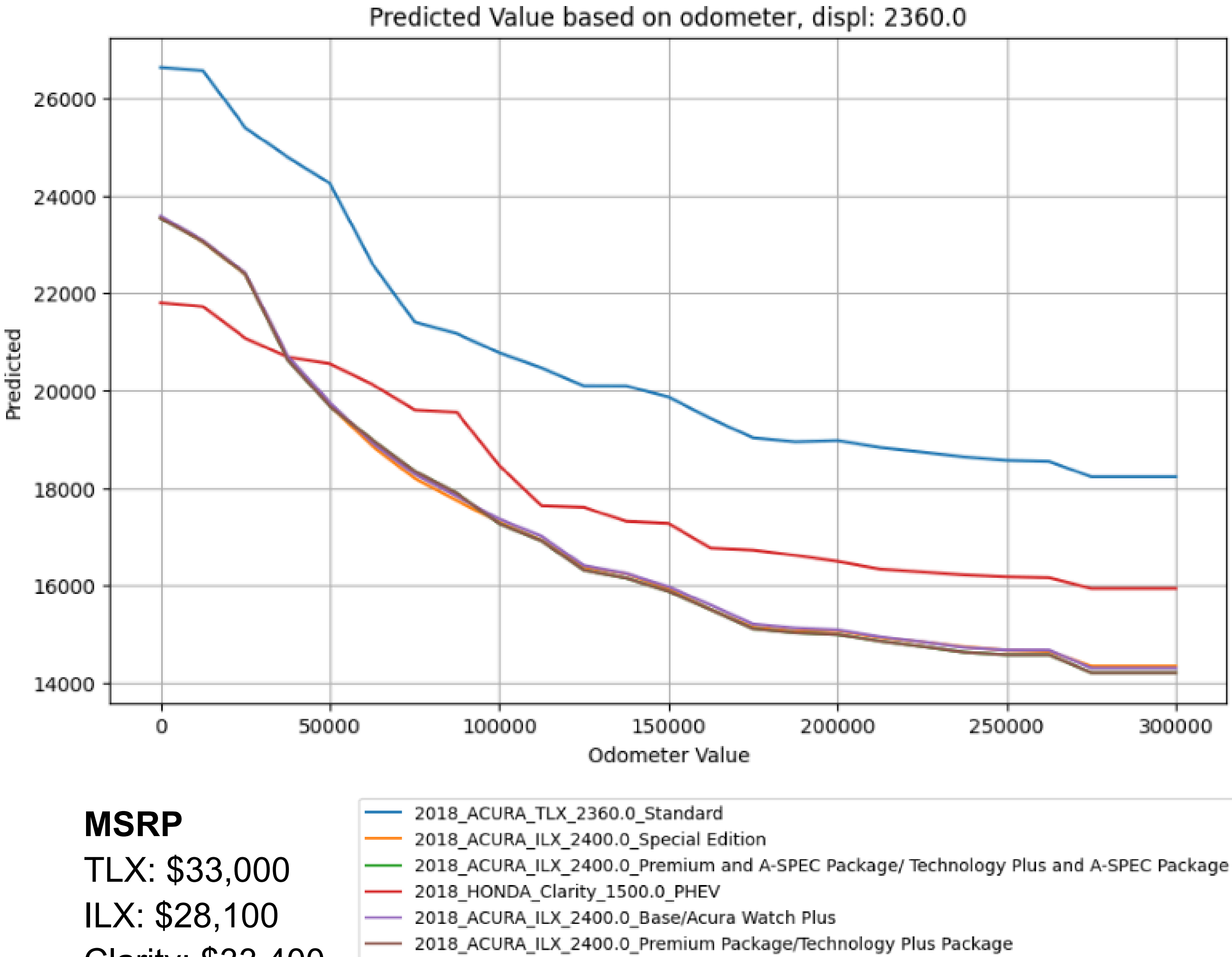
MSRP:

[TLX Advance]: $42,700

[RLX Tech]: $54,450



2018 - V4



MSRP

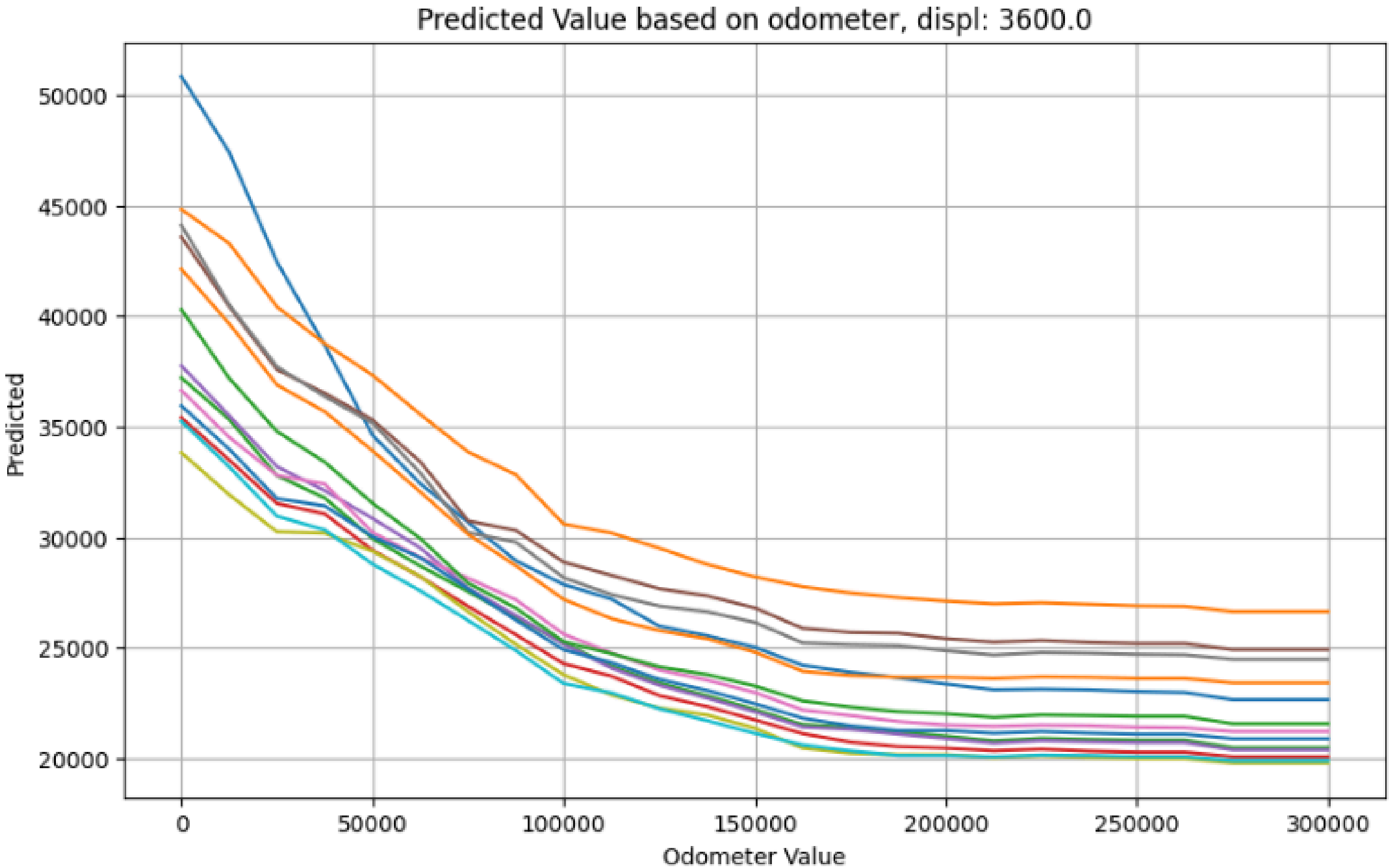
TLX: $33,000

ILX: $28,100

Clarity: $33,400

Cadillac XT6

2020



MSRP

XT6 FWD - $52,695

Enclave Avenir - $53,800

Future Work

1. **Data Cleaning**: Series/Trim can be used to obtain new features by…

2. **Data Scraping:** Edmunds, KBB
    - MSRP
    - Fuel Economy
    - Horsepower & Torque
    - Towing Capacity
    - Safety/Luxury Features

3. **NHTSA API:** Crash Ratings, Recalls, Investigations, Complaints

4. **CI/CD Pipeline:** Spam Detection, Automated Data Manipulation: Cleaning up VIN Output, Imputing Missing Values

5. **Time Series Analysis:** Identify global and local trends in demand/prices