Linear Regression

Price of Used Cars Prediction

Introduction

Linear Regression Model

Prediction Price of Used Cars Listing

Methodology: Data

MSRP

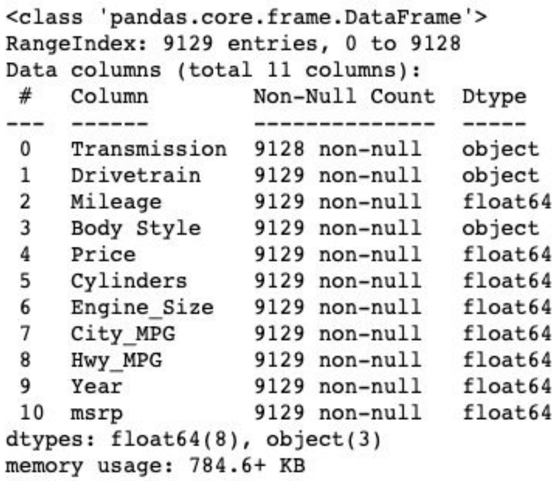
https://www.cars.com/research/

MSRP for latest model of 453 cars

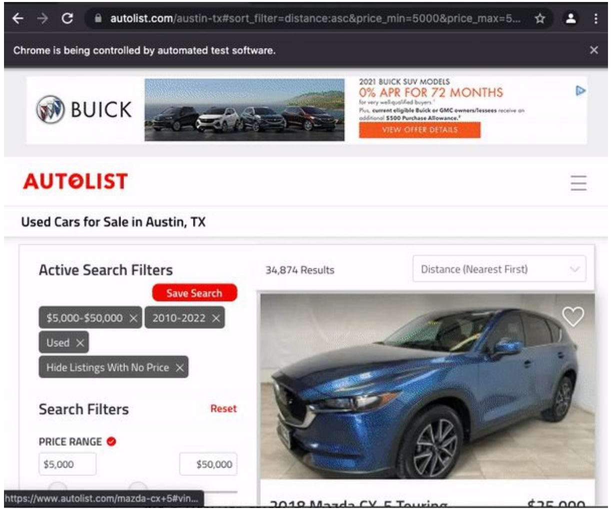
Used Car Listings

https://www.autolist.com/

19 variables for 11809 used cars, including Price.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9129 entries, 0 to 9128
Data columns (total 11 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Transmission   9128 non-null   object
 1   Drivetrain     9129 non-null   object
 2   Mileage        9129 non-null   float64
 3   Body Style     9129 non-null   object
 4   Price          9129 non-null   float64
 5   Cylinders      9129 non-null   float64
 6   Engine_Size    9129 non-null   float64
 7   City_MPG       9129 non-null   float64
 8   Hwy_MPG        9129 non-null   float64
 9   Year           9129 non-null   float64
 10  msrp           9129 non-null   float64
dtypes: float64(8), object(3)
memory usage: 784.6+ KB
```



autolist.com/austin-tx#sort_filter=distance:asc&price_min=5000&price_max=5...

Chrome is being controlled by automated test software.

BUICK
2021 BUICK SUV MODELS
0% APR FOR 72 MONTHS
for very well-qualified buyers.
Plus, current eligible Buick or GMC owners/lessees receive an additional $500 Purchase Allowance.
VIEW OFFER DETAILS
AUTOLIST

Used Cars for Sale in Austin, TX

Active Search Filters

34,874 Results

Distance (Nearest First)

$5,000-$50,000 ✕ 2010-2022 ✕ Used ✕ Hide Listings With No Price ✕

Search Filters

PRICE RANGE

$5,000 $50,000

https://www.autolist.com/mazda-cx+5#vin...

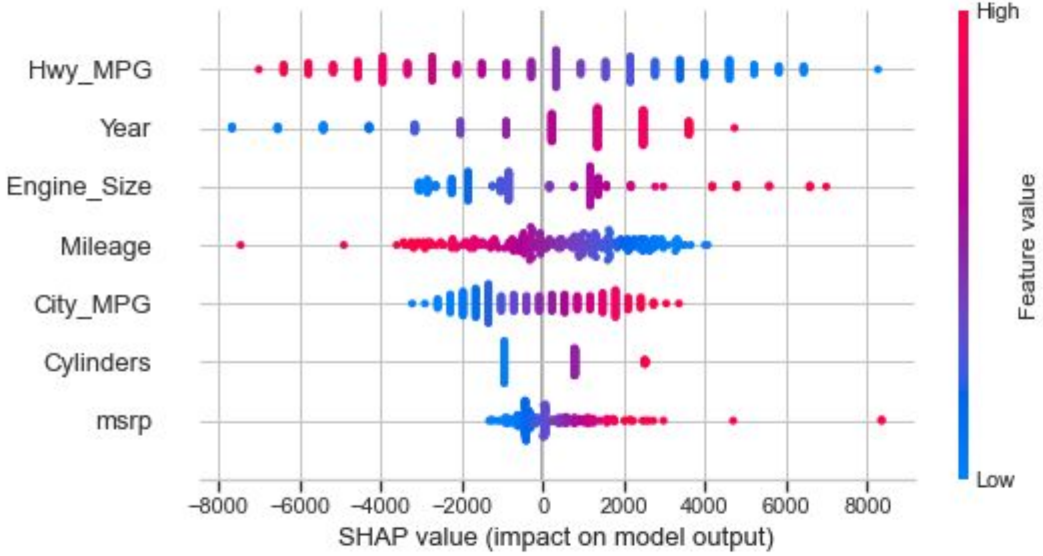
2018 Mazda CX-5 Touring $25,000

Methodology: Libraries & Packages

- Selenium, requests & BeautifulSoup: Scrape and parse HTML
- sklearn.impute : Predictions to impute missing data
- Fuzzywuzzy - Search matches between different sources
- sklearn.linear_model : Create regression models
- yellowbrick.regressor: Drop influential datapoints
- Shap: Visualization of influence of features in model.

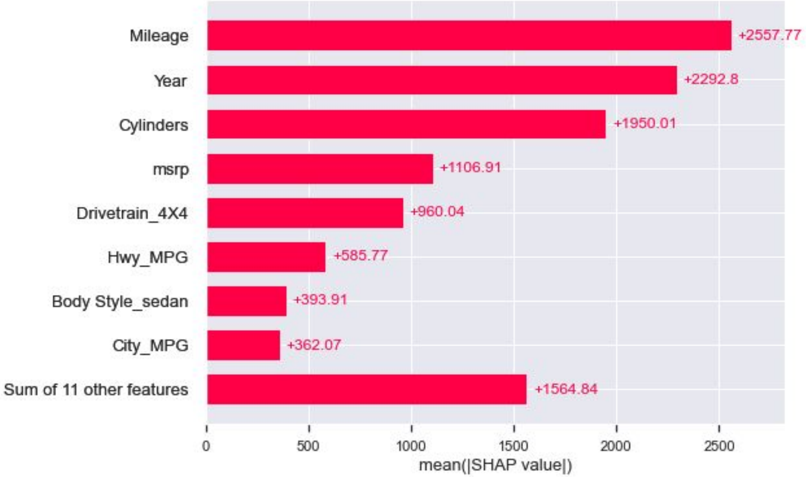
Numerical Features

Numerical Features’ SHAP Values

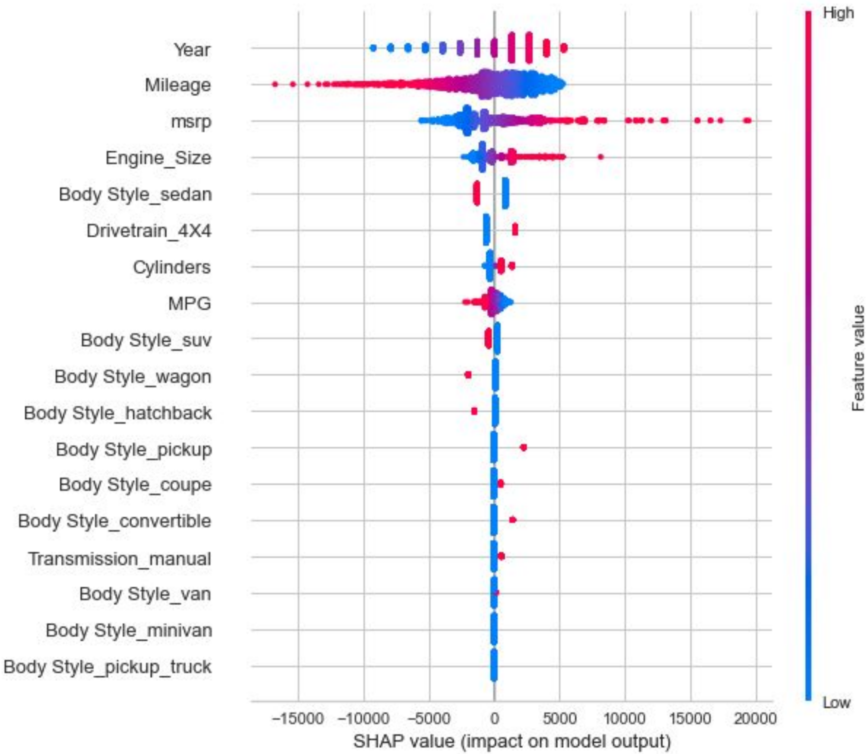


All Features

Mean SHAP value of Features



SHAP Values of all Features

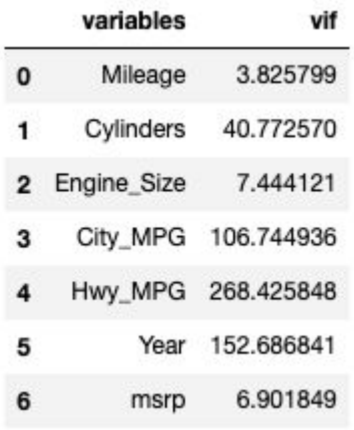


Correlated Features & Regularization

Impute missing values

Imperfect Collinearity

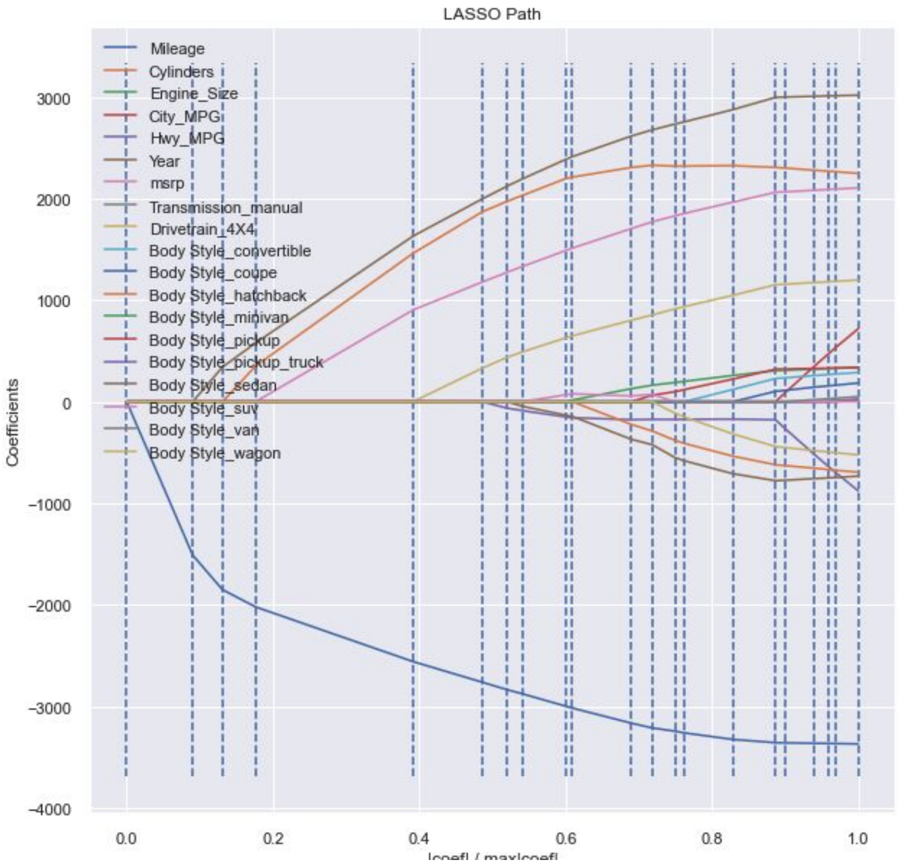
Regularization



	variables	vif
0	Mileage	3.825799
1	Cylinders	40.772570
2	Engine_Size	7.444121
3	City_MPG	106.744936
4	Hwy_MPG	268.425848
5	Year	152.686841
6	msrp	6.901849

Variables Dropped

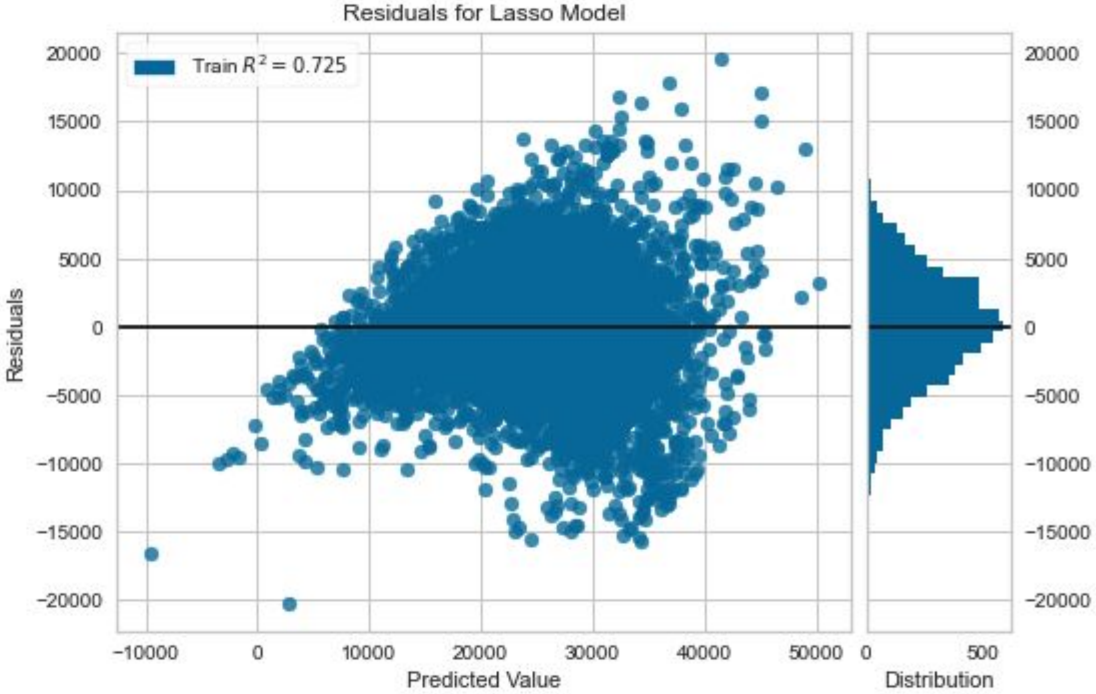
- City MPG
- Manual Transmission
- Body Style - Minivan
- Body Style - Van
- Body Style - SUV
- Body Style - Pickup Truck
- Body Style Hatchback



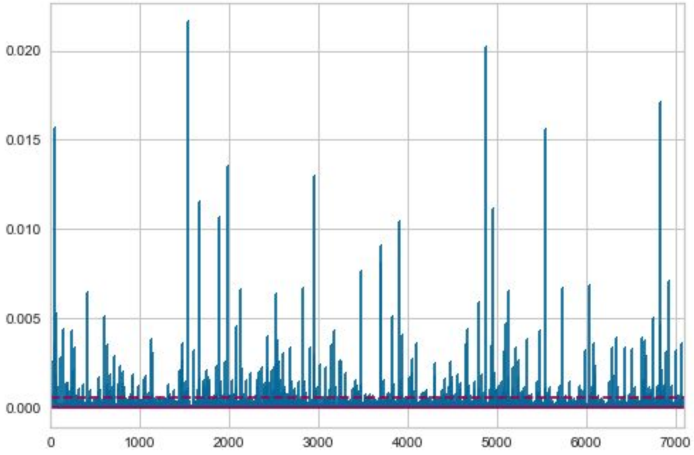
Regularization Results

Model	Mean Squared Error	Mean Absolute Error	R^2 Score	Parameters
Ridge (L1)	1.837455e^07	3335.628	0.725271	Alpha: 12.9
Lasso (L2)	1.837435e^07	3335.626	0.725274	Alpha: 2.0
Elastic Net (L1 & L2)	1.83735e^07	3335.624	0.725274	Alpha:1.97 L1: 1

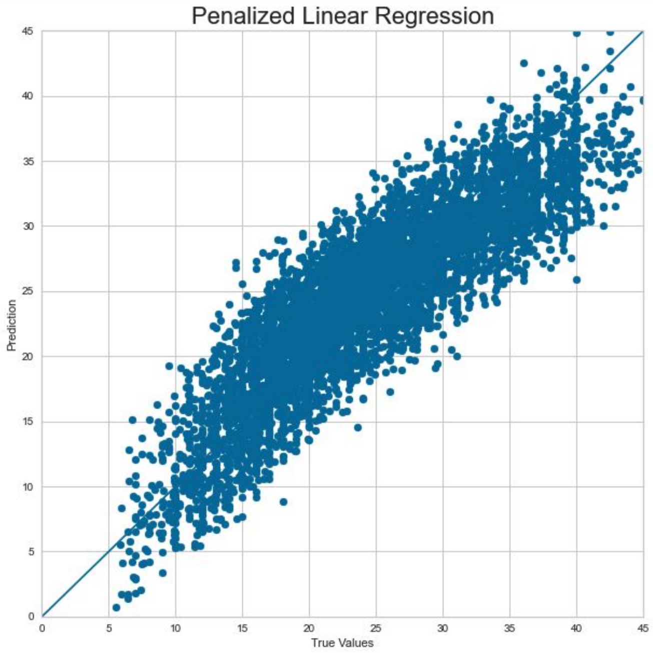
Residual Analysis

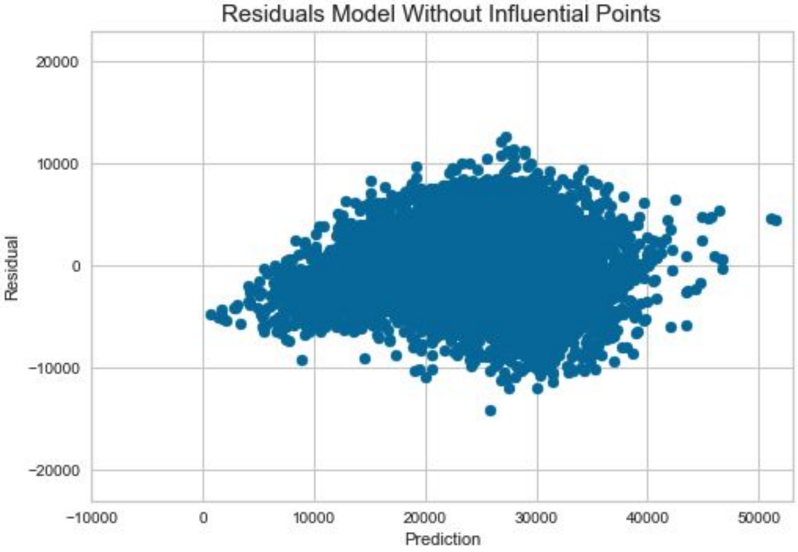


399 Influential Points



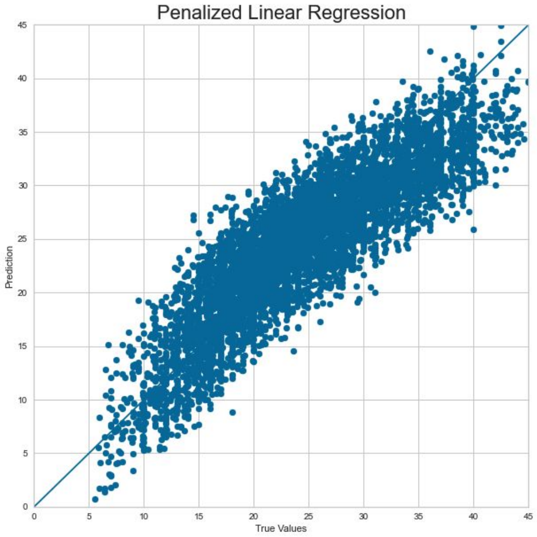
Regression without Influential Points

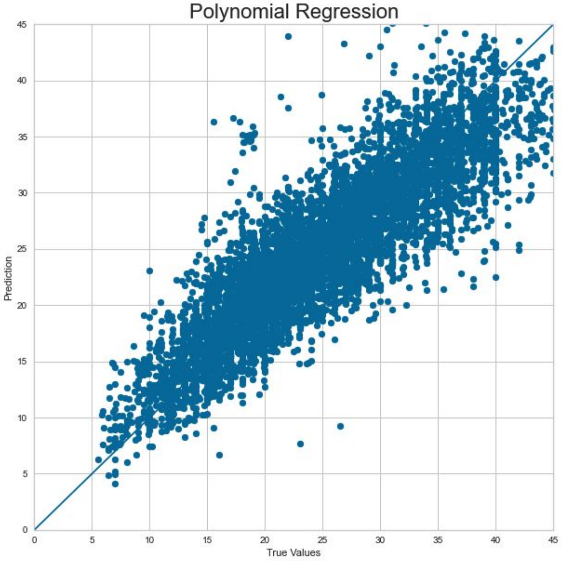




Linear Model	Mean Absolute Error	Root Mean Squared Error	R^2 Score	Parameters
With Influential Points	1.837435e^07	3335.62	0.725	Alpha: 2.0 (L1)
Without Influential Points	1.357837e^07	2952.71	0.769	Alpha: 2.0 (L1)

Polynomial Regression



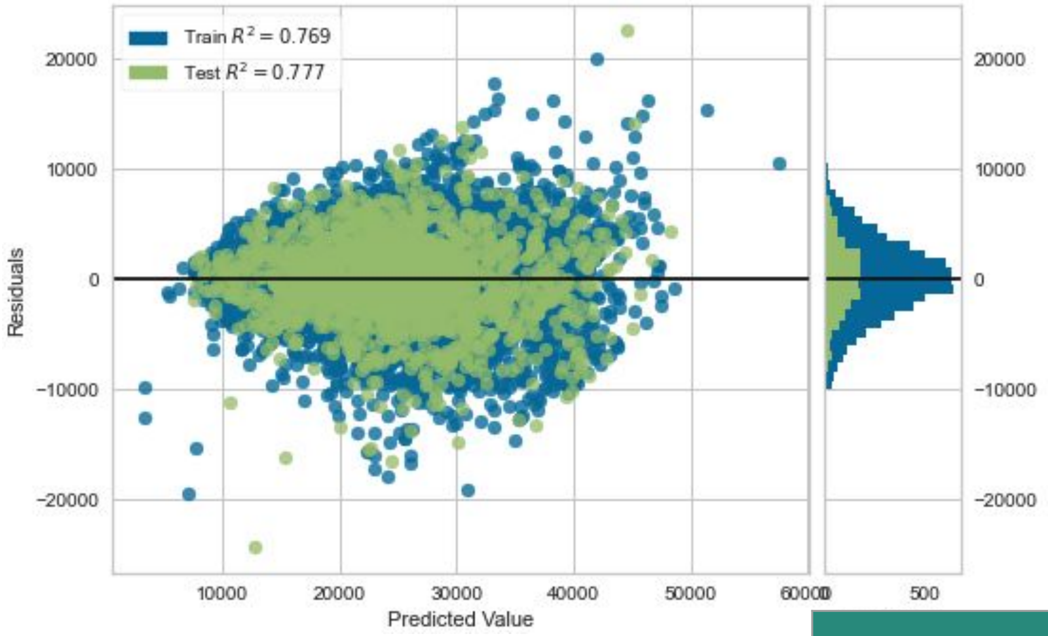


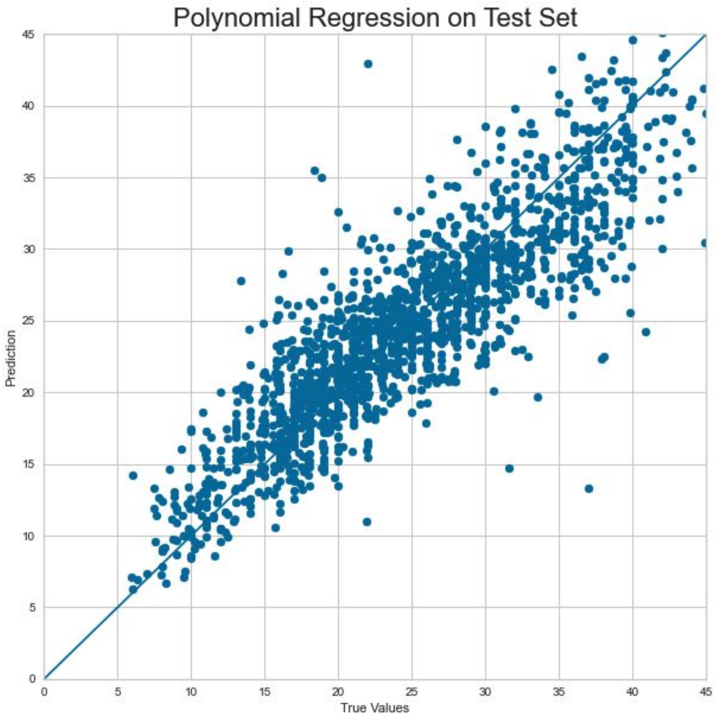
Linear Model	Mean Absolute Error	Root Mean Squared Error	R^2 Score	Parameters
Penalized Linear Model	1.357e^07	2953	0.769	Lambda 2.0 (L1)
Polynomial Model	1.201e^07	2714	0.796	

Out of Sample Performance

Linear Model	Mean Squared Error	Root Mean Squared Error	R^2 Score
Polynomial Model	1.627e^07	3016	0.770

Results: Unseen Data





Linear Model	Mean Squared Error	Root Mean Squared Error	R^2 Score
Polynomial Model	1.627e^07	3016	0.770

Conclusions

- Low Bias

- High Variance

- RMSE: ~$3.000

Future Work

Regularization in Polynomial Regression

Need features indicating level of car (basic/premium)

MSRP for individual car.

Remove influential datapoints from SHAP Summary Plot