Predicting Used Car Prices

Linear Regression

Riwa Sabri

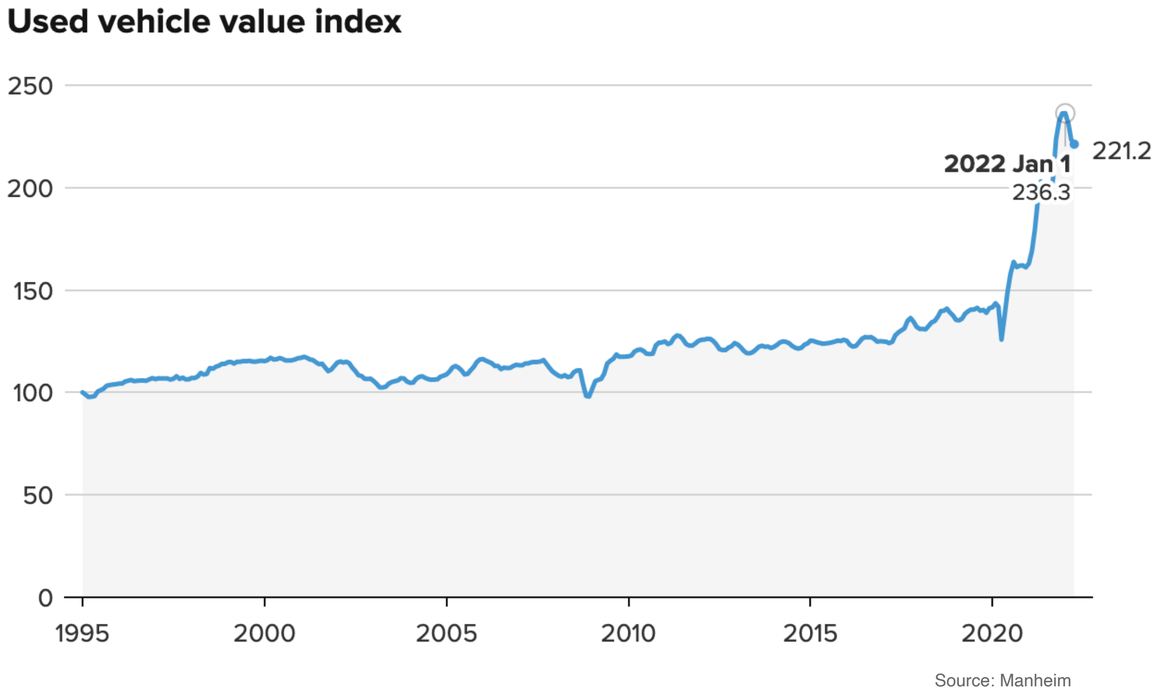
August 2022

Introduction

- The used car market in the US saw a **big surge in demand** and prices in the last year. This is due to the **chip shortage** that reduced supply of new cars and trucks.

- Although this surge has recently started to decelerate, buying a used car is not what it used to be.

- This analysis aims at using a linear regression model to predict used cars prices in a **20 mile radius of ZIP code 11221 in Brooklyn.**



Methodology

SCRAPING CARS.COM USING BEAUTIFUL SOUP:
USED CARS IN A 20 MI RADIUS FROM NY, 11221

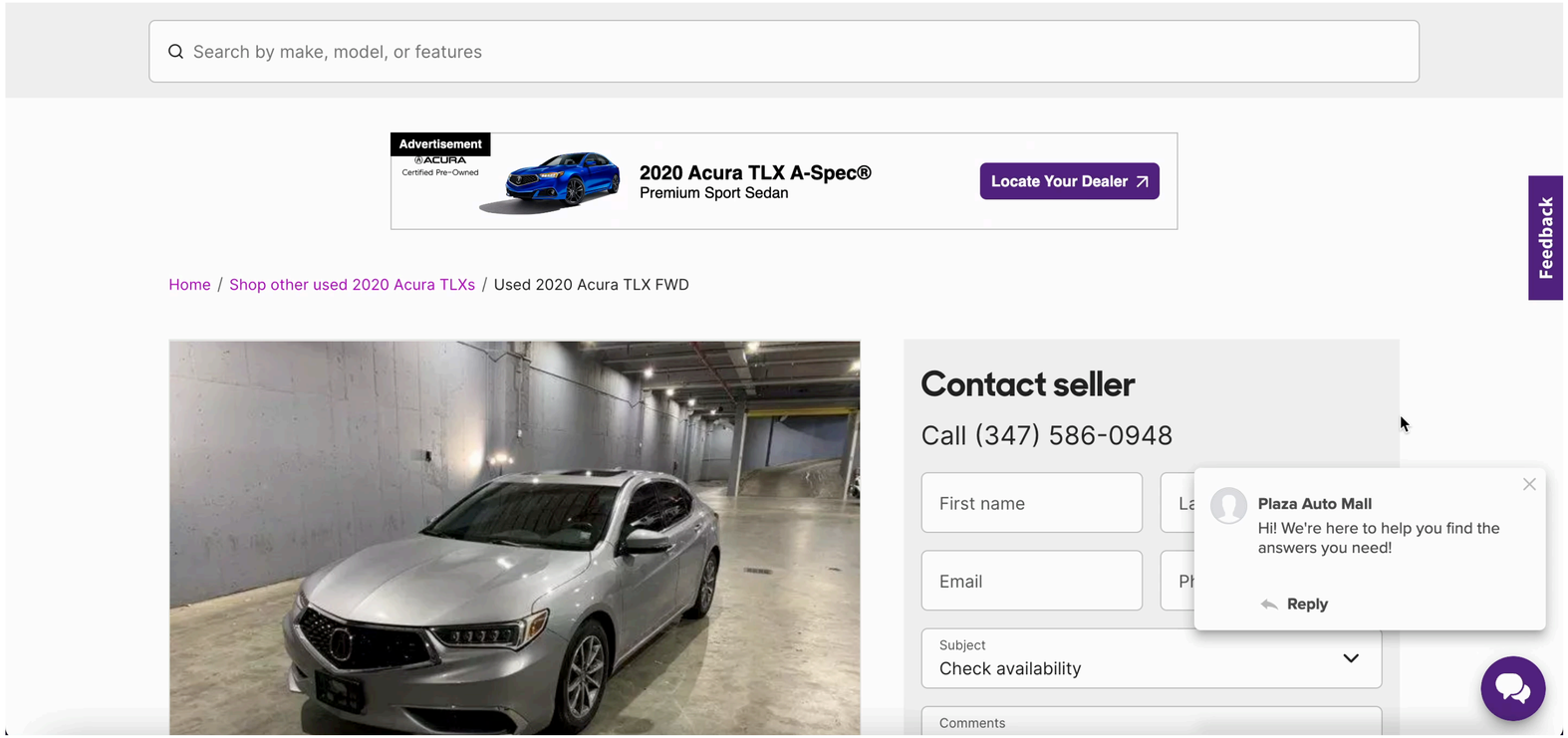
BASELINE MODEL:
ONLY NUMERICAL FEATURES

REGULARIZATION:
LASSO

EDA:
DATA CLEANING AND PREPARATION

IMPROVING ON BASELINE MODEL

A sample row of scraped data after EDA



Year	2020
Make	Acura
Model	TLX
Price	31199.0
Mileage	28762.0
Rating	4.6
Drivetrain	Front-wheel Drive
Average_MPG	28.0
Fuel_Type	Gasoline
Convenience_Count	2
Entertainment_Count	4
Safety_Count	4
Deal_Type	Fair Deal
Engine_Displacement	2.4
Valvetrain_Type	DOHC
Cylinders	I4
Zip_Code	11210
Interior_Color	Ebony
Exterior_Color	Gray
Accidents	None reported
Used_Certified	Used

Baseline model has a validation R^2 of 0.24

The baseline model included numerical features to predict Price.

Features

- Mileage
- Rating
- Average_MPG
- Safety_Count
- Entertainment_Count
- Convenience_Count
- Year
- Engine_Displacement

Model

OLS

Scores

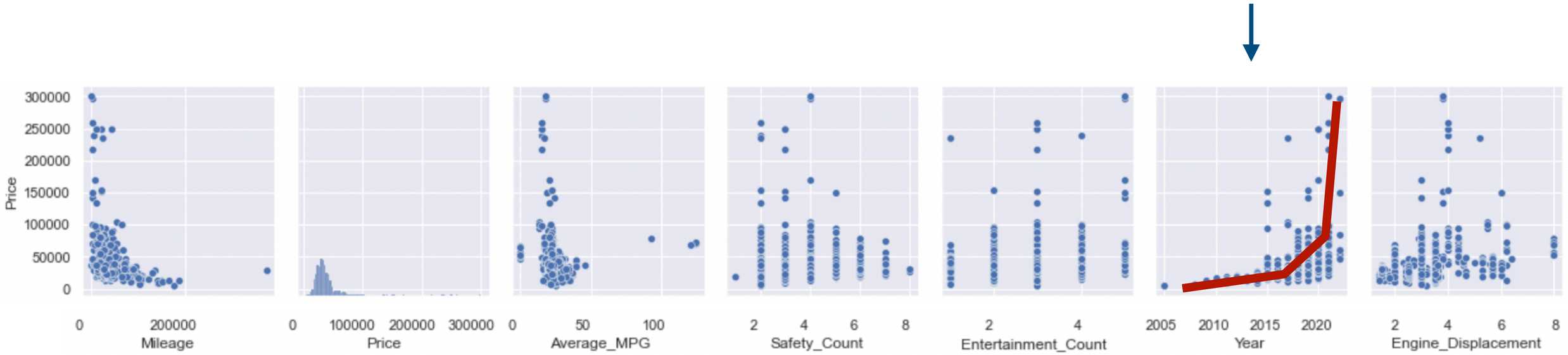
- **Validation R^2:** 0.24
- **MSE:** 11,210.47

Analyzing P-values : Rating and Convenience Count don't pass the null hypothesis test

Rating and Convenience Count have P-values > 0.05, let's drop them.

	coef	std err	t	P>|t|
const	-6.597e+06	1e+06	-6.567	0.000
Mileage	-0.1978	0.033	-5.906	0.000
Rating	345.2725	1005.780	0.343	0.731
Average_MPG	-453.6940	86.778	-5.228	0.000
Safety_Count	-2979.7920	555.213	-5.367	0.000
Entertainment Count	3088.0018	766.810	4.027	0.000
Convenience Count	-14.2496	512.449	-0.028	0.978
Year	3286.4194	497.785	6.602	0.000
Engine_Displacement	7389.8265	649.445	11.379	0.000

Analyzing relationships between features and target : year and price have a polynomial relationship



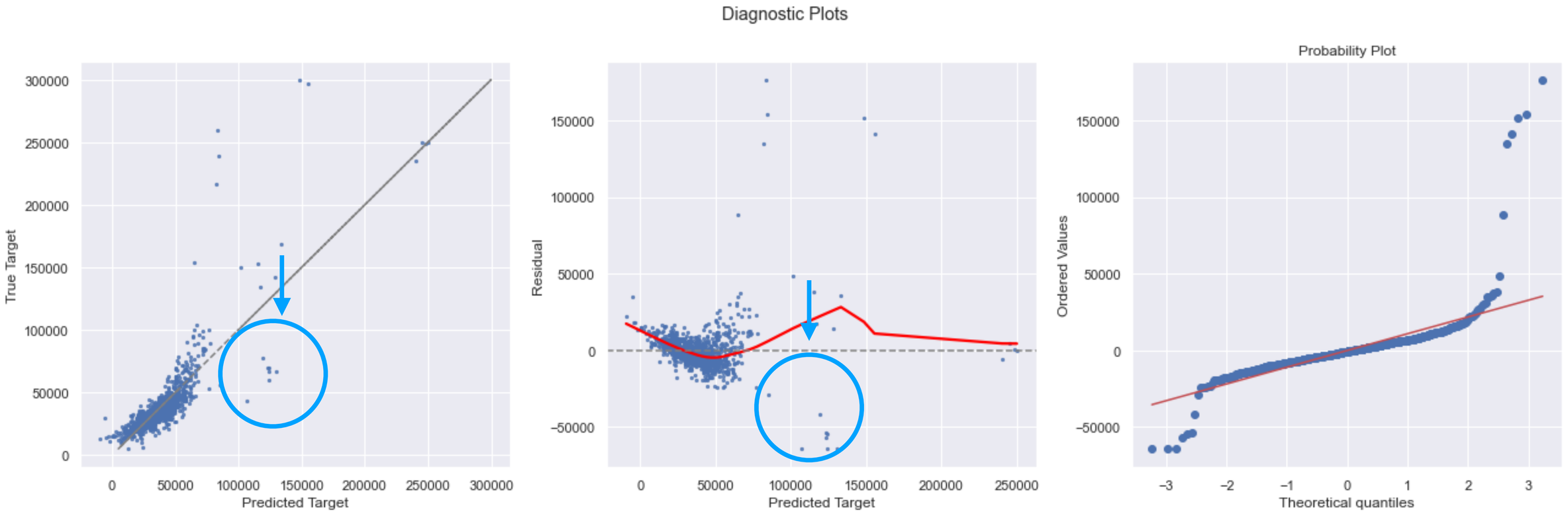
The changes yield a slight improvement in R^2 and MSE

	Baseline Model	Improved Baseline Model	% Change (Rounded)
Validation R^2	0.24	0.26	**+8%**
MSE	11,210.47	11,102.47	**-1%**

We do a lot better by adding categorical variable “Make”!

	Improved Baseline Model	Improved Baseline Model + Make	% Change (Rounded)
Validation R^2	0.26	0.66	**+153%**
MSE	11,210.47	7,029.44	**-37%**

Diagnostic Plots: some cars have a much higher predicted price vs their actual price



While “Make" improves the model fit overall, it’s causing our model to overestimate Luxury cars with high mileage

Name	Mileage	Year	Predicted Price	Actual Price	Residual
2019 Porsche Cayenne Base	26461.0	2019	124063.347306	59900.0	-64163.347306
2015 Porsche Panamera 4	53604.0	2015	106962.221981	42888.0	-64074.221981
2021 Porsche Macan Base	10546.0	2021	130354.001132	66420.0	-63934.001132
2019 Porsche Cayenne Base	22589.0	2019	123707.994107	66988.0	-56719.994107
2019 Porsche Cayenne Base	20132.0	2019	124059.504604	69988.0	-54071.504604
2019 Porsche Cayenne Base	24974.0	2019	123366.784285	69488.0	-53878.784285
2019 Porsche Cayenne S	36859.0	2019	118947.251453	77690.0	-41257.251453
2020 Tesla Model 3 Long Range	13782.0	2020	84953.053190	56400.0	-28553.053190

← For example this 2019 Porsche Cayenne is being valued at $124,063 because it’s luxury, but due to high mileage it’s actual price is $59,900!

After adding a new “high mileage” x “luxury car” feature, new R^2 of 0.7 and improved MSE !

	Improved Baseline Model + Make	Improved Baseline Model + Make +New Feature	% Change (Rounded)
Validation R^2	0.66	0.70	**+6%**
MSE	7,029.44	6,390.17	**-9%**

Now let’s add the remaining categorical features to our model

01 **Model**

02 **Cylinders**

03 **Fuel Type**

04 **Drivetrain**

05 **Accidents**

06 **Used / Certified**

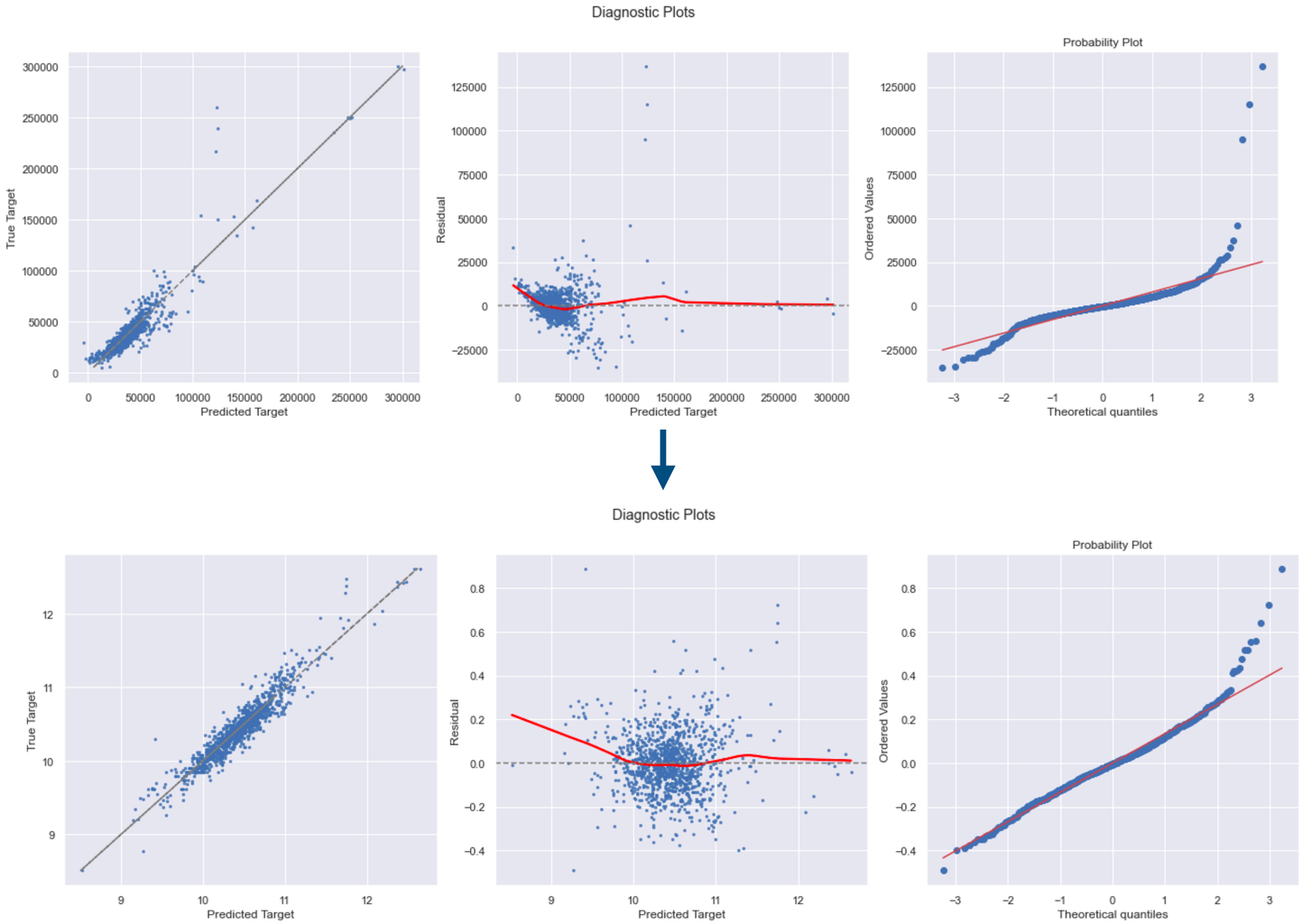
07 **Deal Type**

08 **Valve-train Type**

R^2 went up to 0.79!

	Improved Baseline Model + Make +New Feature	Improved Baseline Model + Make +New Feature+ Added Categoricals	% Change (Rounded)
Validation R^2	0.70	0.79	**+13%**
MSE	6,390.17	5,492.38	**-14%**

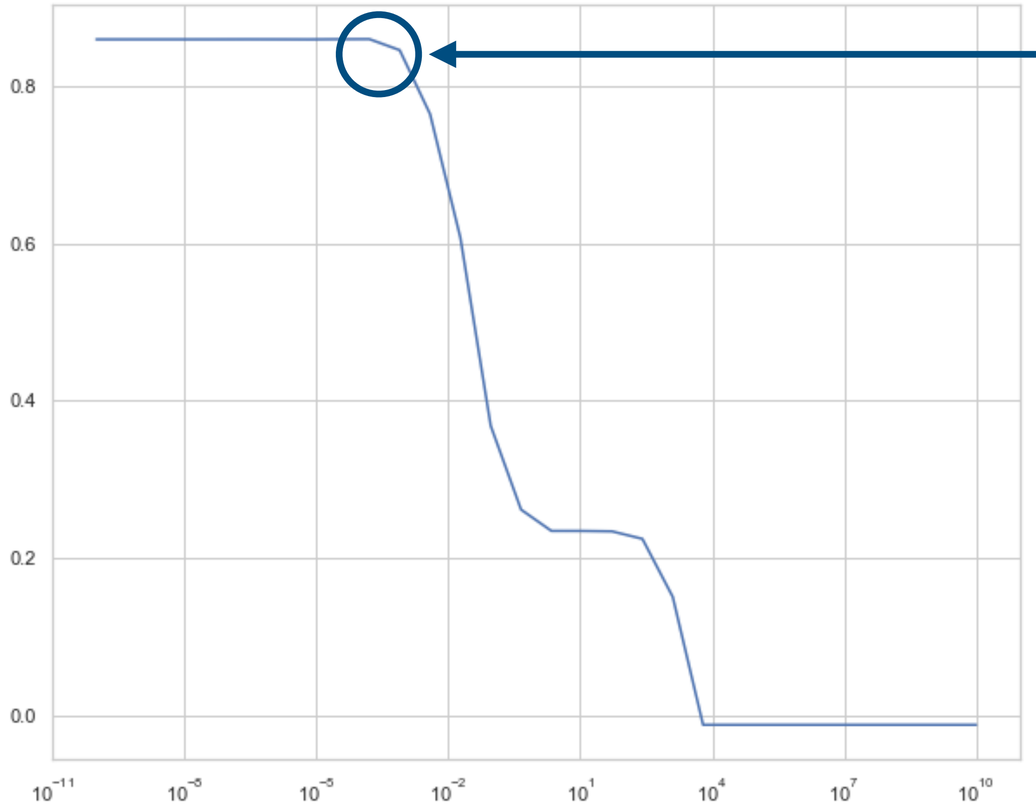
Distribution of residuals suggest that a log transformation can improve performance



Log Transformation of Target leads to a new R^2 of 0.85!

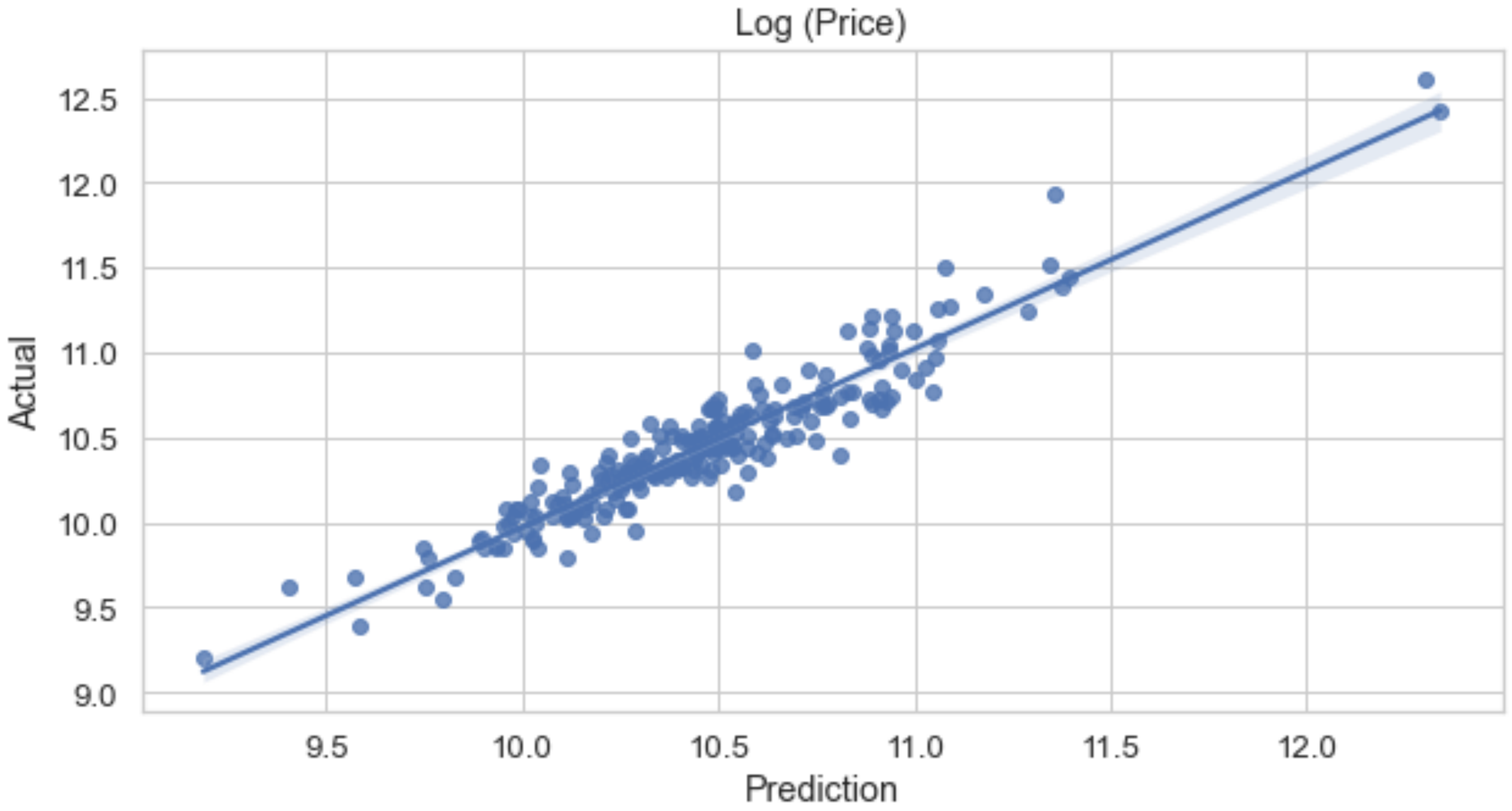
	Improved Baseline Model + Make +New Feature+ Added Categoricals	Improved Baseline Model + Make +New Feature+ Added Categoricals+Log Transformation	% Change (Rounded)
Validation R^2	0.79	0.85	**+8%**
MSE	5,492.38	0.11	**N/A**

Regularization with Lasso using GridSearch



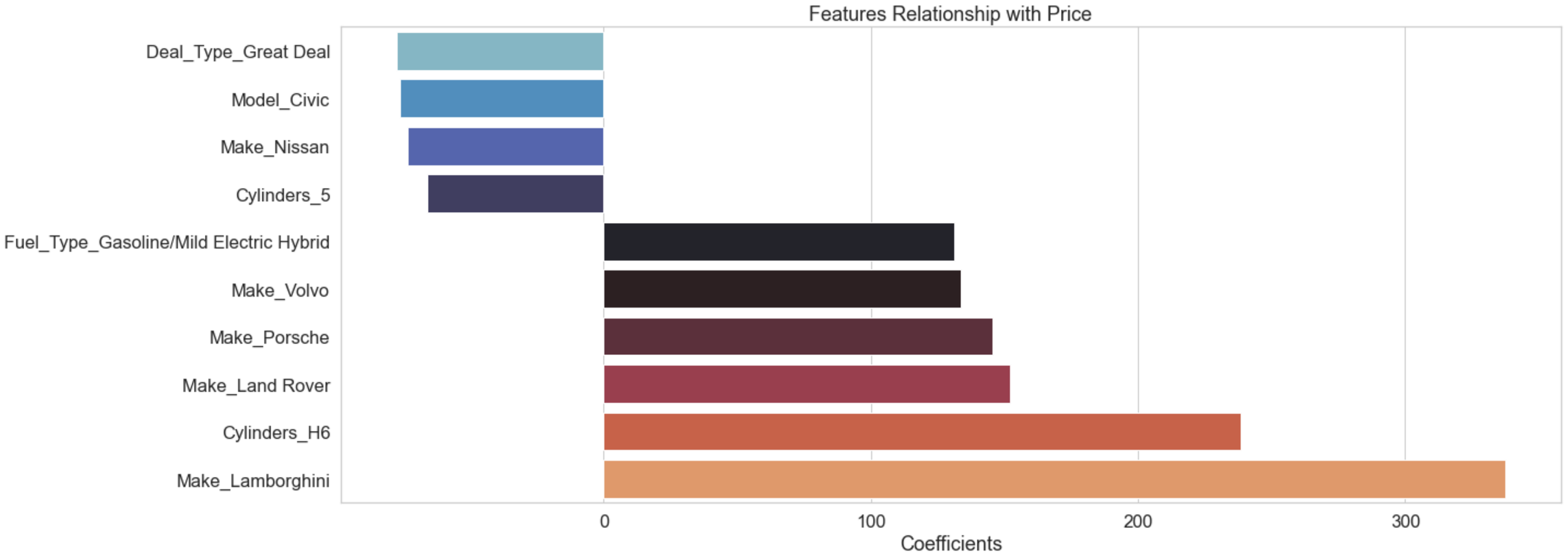
Optimal Alpha

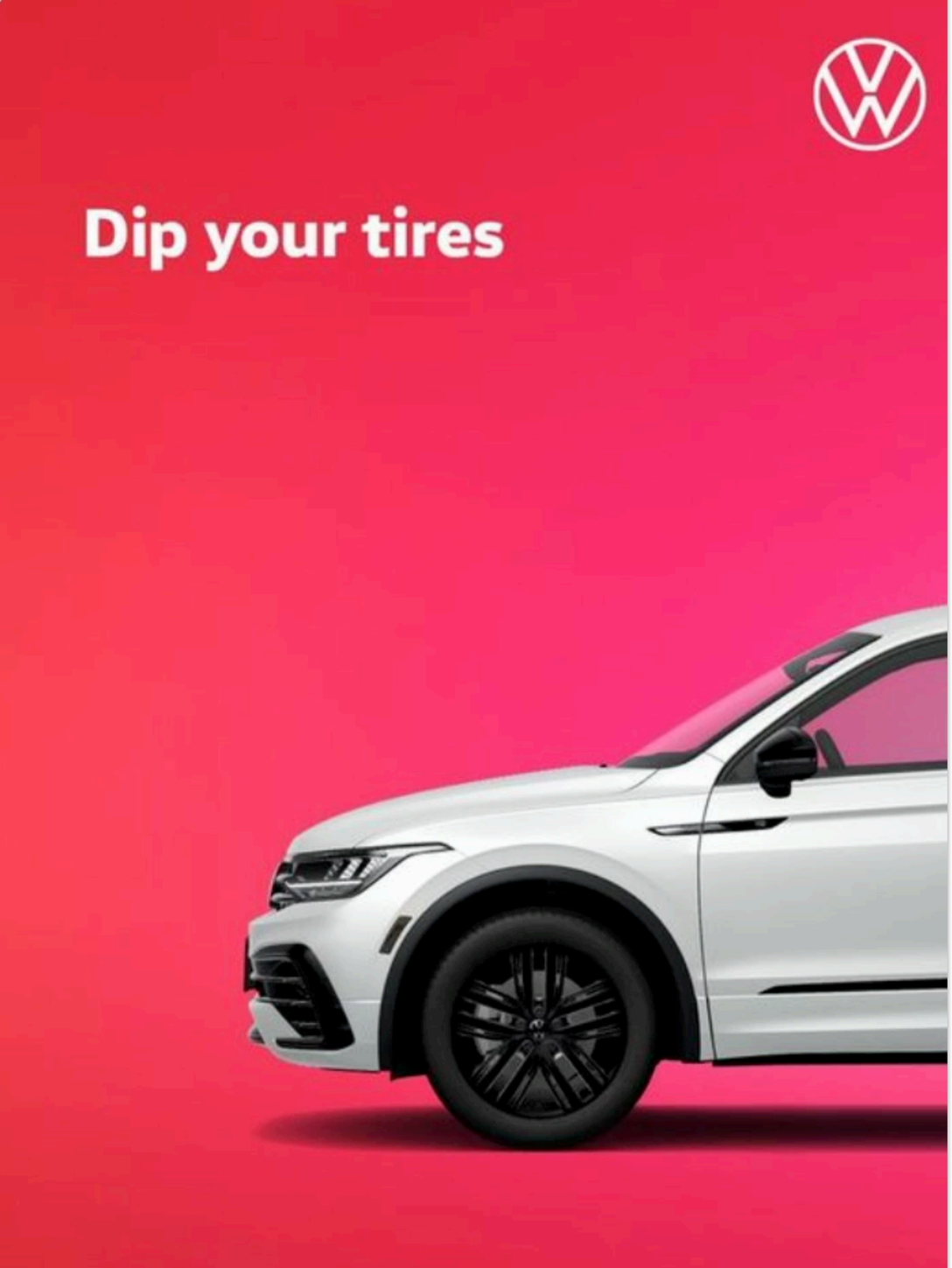
Final Model using Lasso to address multicollinearity and feature prioritization



R^2 Test : 0.89

Final Model Coefficients





Thank you