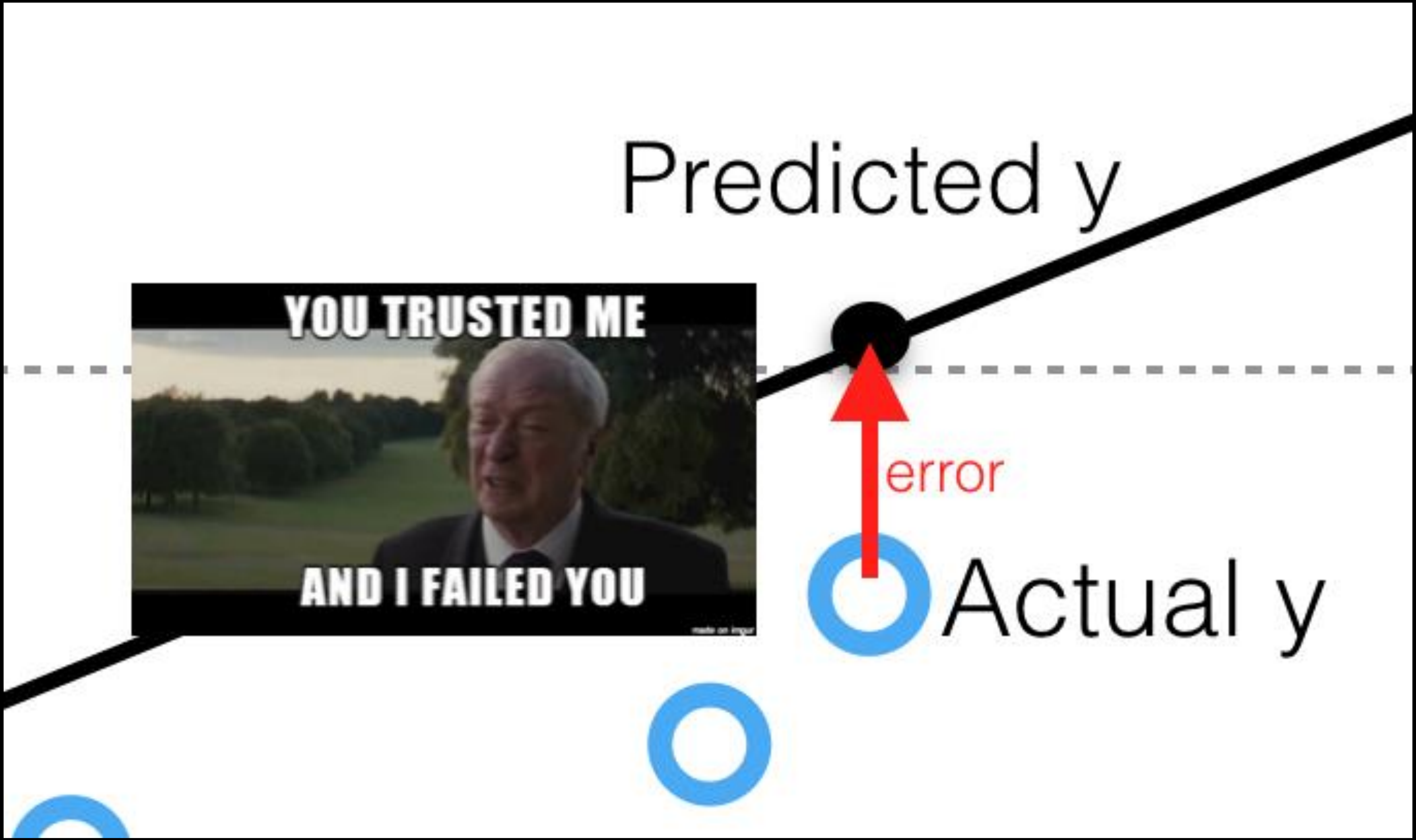
REGRESSION MODEL

CAR PRICE PREDICTION WEB APP

Steven L Truong

Friday, 16/04/2021


Predicted y
YOU TRUSTED ME
AND I FAILED YOU
error
Actual y

INTRODUCTION

❖**MOTIVATION**:

- Buy and sell used cars is always a big decision.
- Create the tools to predict the as closest car’s price as possible.

INTRODUCTION

❖ **MOTIVATION**:

- Buy and sell used cars is always a big decision.
- Create the tools to predict the as closest car’s price as possible.

❖ **OBJECTIVES**:

- Build predictive models using data gathered from the internet.
- Conclude the best model ready for production.

INTRODUCTION

❖ **MOTIVATION**:

- Buy and sell used cars is always a big decision.
- Create the tools to predict the as closest car's price as possible.

❖ **OBJECTIVES**:

- Build predictive models using data gathered from the internet.
- Conclude the best model ready for production.

❖ **GOALS**:

- Write the web app and deploy the model to the cloud.

METHODOLOGY

- **Data source:**
  - Scrape from cars.com

METHODOLOGY

- **Data source:**
  - Scrape from cars.com
- **Tools:**
  - BeautifulSoup, Numpy and Pandas
  - Matplotlib and Seaborn
  - Scikit-learn and XGBoost
  - Streamlit and Heroku

METHODOLOGY

Data Scraping and Preparation

- **Use BeaufulSoup to scrape data from cars.com**
- **Clean the data to be ready for EDA**

METHODOLOGY

Data Scraping and Preparation

- Use BeaufulSoup to scrape data from cars.com
- Clean the data to be ready for EDA

Exploratory Analysis

- Exploratory Data Analysis
- Look at the features’ correlations for insights before modeling.

METHODOLOGY

Data Scraping and Preparation

- Use BeaufulSoup to scrape data from cars.com
- Clean the data to be ready for EDA

Exploratory Analysis

- Exploratory Data Analysis
- Look at the features' correlations for insights before modeling.

Modeling

- Build baseline models.
- Cross validation and choose the final model.

DATA CLEANING

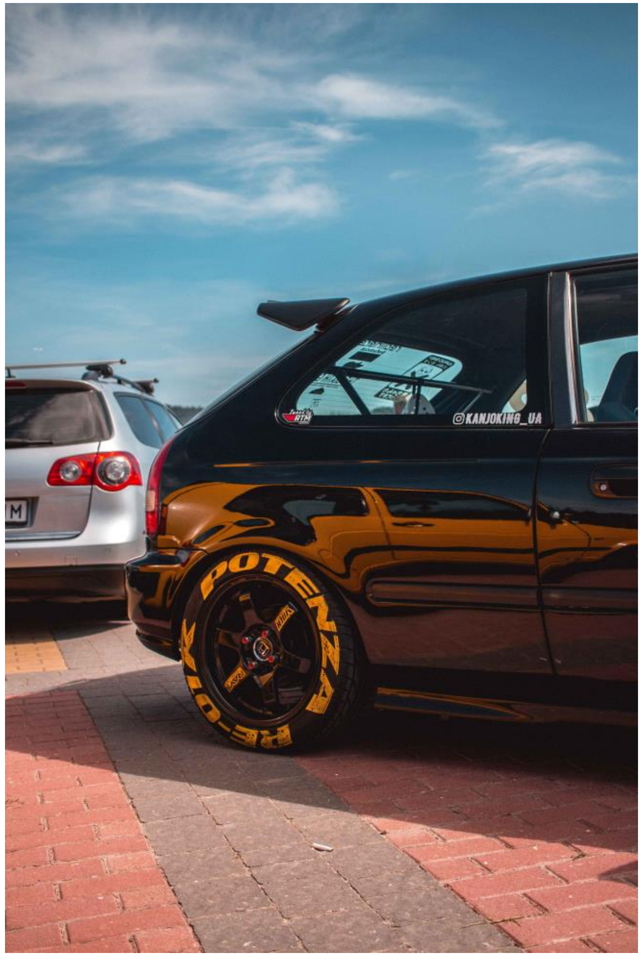
DATA CLEANING

DATA CLEANING

Or we could say “cars cleaning”

DATA CLEANING

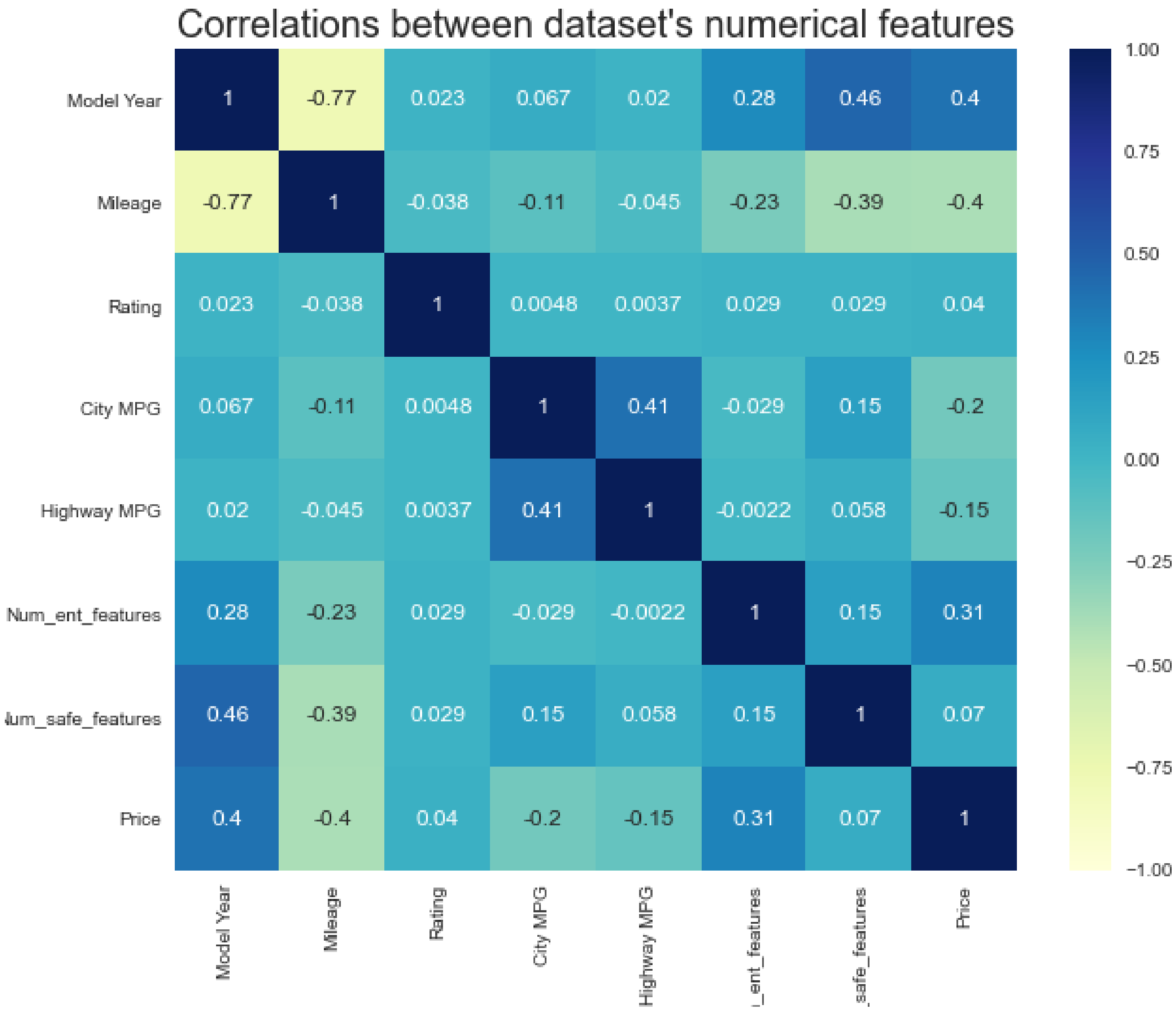
Or we could say “cars cleaning”



- 187,168 raw data points were scraped.
- Clean data set has 122,351 rows and 18 columns

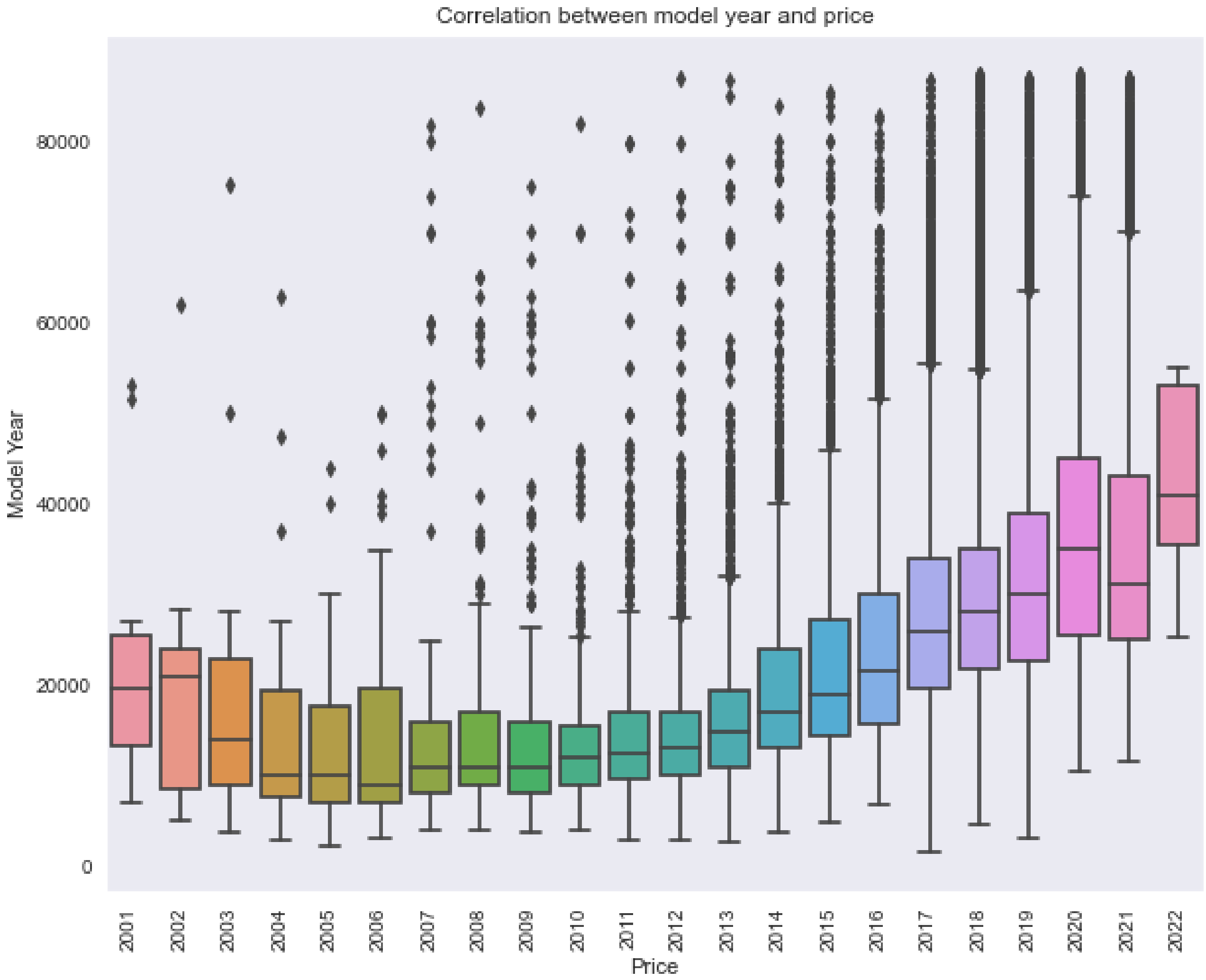
EDA

- Price is positively correlated with Model Year and negatively correlated with Mileage.
- Slightly positively correlated with num_ent_features.
- Not so much for the rest of the features.



EDA

- Price is positively correlated with Model Year.
- There are outliers all over the place.
- Generally speaking, the newer the more expensive car.



MODELS

Pre features engineered.

MODELS

Pre features engineered.

- Linear Regression Model:
  - R^2 for test set: 0.290

- Polynomial Regression Model:
  - R^2 for test set: 0.474

MODELS

Pre features engineered.

- Linear Regression Model:
  - R^2 for test set: 0.290
- Polynomial Regression Model:
  - R^2 for test set: 0.474
- Random Forest Regressor:
  - R^2 for test set: 0.577
- Gradient Boosted Regressor:
  - R^2 for test set: 0.605
- Extreme Gradient Boosting (XGBoost):
  - R^2 for test set: 0.729

MODELS

Pre features engineered.

- Linear Regression Model:
  - R^2 for test set: 0.290
- Polynomial Regression Model:
  - R^2 for test set: 0.474

In general, they all underfit

- Random Forest Regressor:
  - R^2 for test set: 0.577
- Gradient Boosted Regressor:
  - R^2 for test set: 0.605
- Extreme Gradient Boosting (XGBoost):
  - R^2 for test set: 0.729

MODELS

Work with categorical features!

MODELS

Work with categorical features!

Intuitively, car's brand (make) determines the product's price, so let's work on that.

MODELS

Work with categorical features!

- Linear Regression Model:
  - R^2 for test set: 0.505
- Polynomial Regression Model:
  - R^2 for test set: 0.695
- Extreme Gradient Boosting (XGBoost):
  - R^2 for test set: 0.870

MODELS

Work with categorical features!

- Linear Regression Model:
  - R^2 for test set: 0.505
- Polynomial Regression Model:
  - R^2 for test set: 0.695
- Extreme Gradient Boosting (XGBoost):
  - R^2 for test set: 0.870

We have better results, can we improve our performance?

MODELS

Work with categorical features!

Let’s dummify the entire dataset!

MODELS

L1/L2 Regularization

K-Fold Cross-Validation!

- Linear Regression Model:
  - R^2 for test set: 0.869
  - RMSE = 4787.90
- Lasso Model:
  - R^2 for test set: 0.866
- Extreme Gradient Boosting (XGBoost):
  - R^2 for test set: 0.920
  - RMSE = 3749.5
- Ridge Model:
  - R^2 for test set: 0.865

PREDICTION

2017 Chevrolet Camaro 2SS

44,953 Mileage, Gasoline engine

City MPG 16 – Highway MPG 25

RWD – Engine 6.2L V8 – 8 speed Manual

Linear Regression Model predicts

$35,235

Extreme Gradient Boosting (XGBoost) predicts

$35,089

PREDICTION

2017 Chevrolet Camaro 2SS

44,953 Mileage, Gasoline engine

City MPG 16 – Highway MPG 25

RWD – Engine 6.2L V8 – 8 speed Manual

Linear Regression Model predicts

$35,235

Extreme Gradient Boosting (XGBoost) predicts

$35,893

True value

$38,395

PREDICTION

2018 INFINITY Q60 3.0t LUXE

18, 719 Mileage, Gasoline engine

City MPG 19 – Highway MPG 27

AWD – Engine 3.0 V6 – 7 speed Automatic

Linear Regression Model predicts

$37,604

Extreme Gradient Boosting (XGBoost) predicts

$35,689

PREDICTION

2018 INFINITY Q60 3.0t LUXE

18, 719 Mileage, Gasoline engine

City MPG 19 – Highway MPG 27

AWD – Engine 3.0 V6 – 7 speed Automatic

Linear Regression Model predicts

$37,604

Extreme Gradient Boosting (XGBoost) predicts

$35,689

True value

$32,500

CONCLUSION

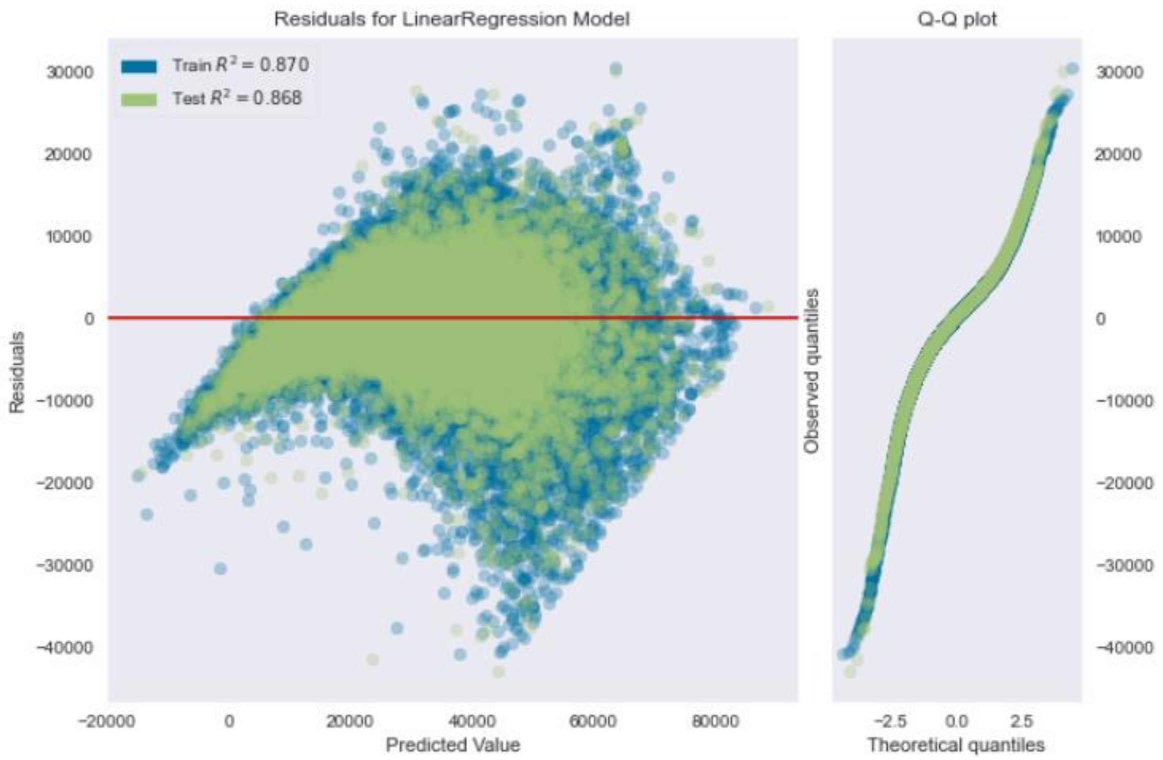
L1/L2 Regularization

K-Fold Cross-Validation!

- Linear Regression Model:
  - R^2 for test set: 0.869
  - RMSE = 4787.90
- Lasso Model:
  - R^2 for test set: 0.866
- Extreme Gradient Boosting (XGBoost):
  - R^2 for test set: 0.920
  - RMSE = 3749.5
- Ridge Model:
  - R^2 for test set: 0.865

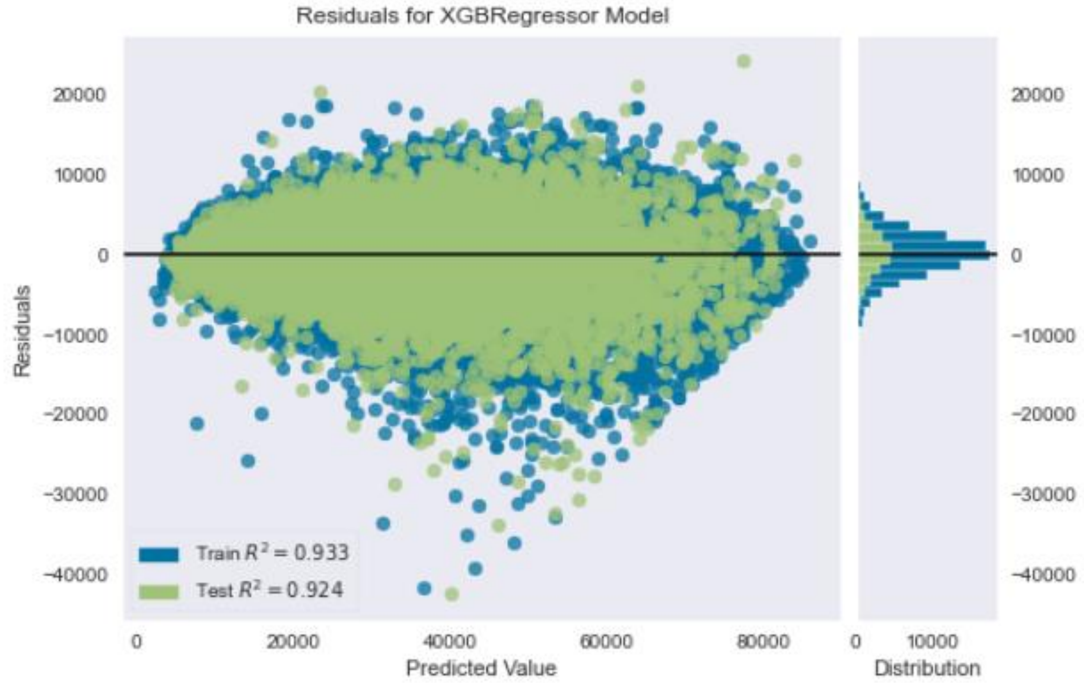
RESIDUALS

- Linear Regression Model:
  - R^2 for train set: 0.871
  - R^2 for validation set: 869
  - RMSE = 4787.90



RESIDUALS

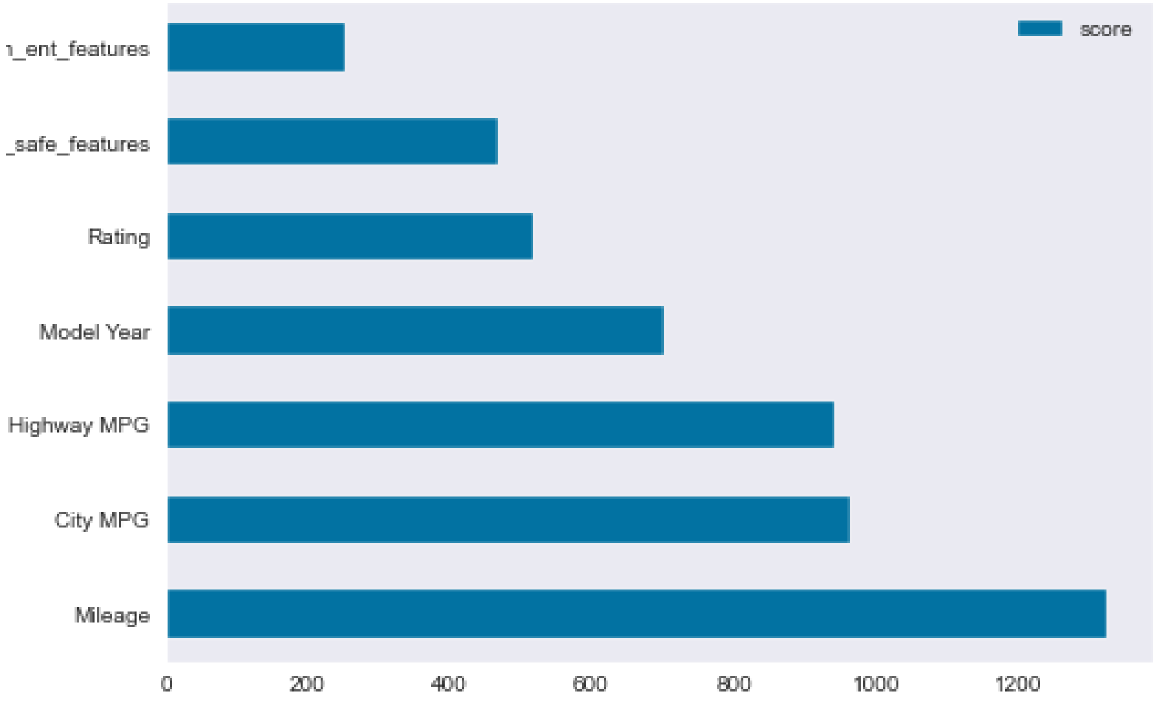
- Extreme Gradient Boosting (XGBoost):
  - R^2 for train set: 0.932
  - R^2 for validation set: 0.920
  - RMSE = 3749.5



ResidualsPlot(ax=<AxesSubplot:title={'center':'Residuals for XGBRegressor Mo

FEATURE IMPORTANCE

- Extreme Gradient Boosting (XGBoost):
  - R^2 for train set: 0.932
  - R^2 for validation set: 0.920
  - RMSE = 3749.5



RECAP

RECAP

Linear Model

0.290 → 1 cat feature → 0.505

0.505 → L1/L2/5-Fold CV → 0.869

0.505 → All cat features → 0.869

RECAP

Linear Model: 0.290 → 1 cat feature → 0.505 → L1/L2/5-Fold CV, All cat features → 0.869

XGBoost: 0.729 → 1 cat feature → 0.870 → All cat features → 0.920

WHAT'S NEXT?

Orignal question?

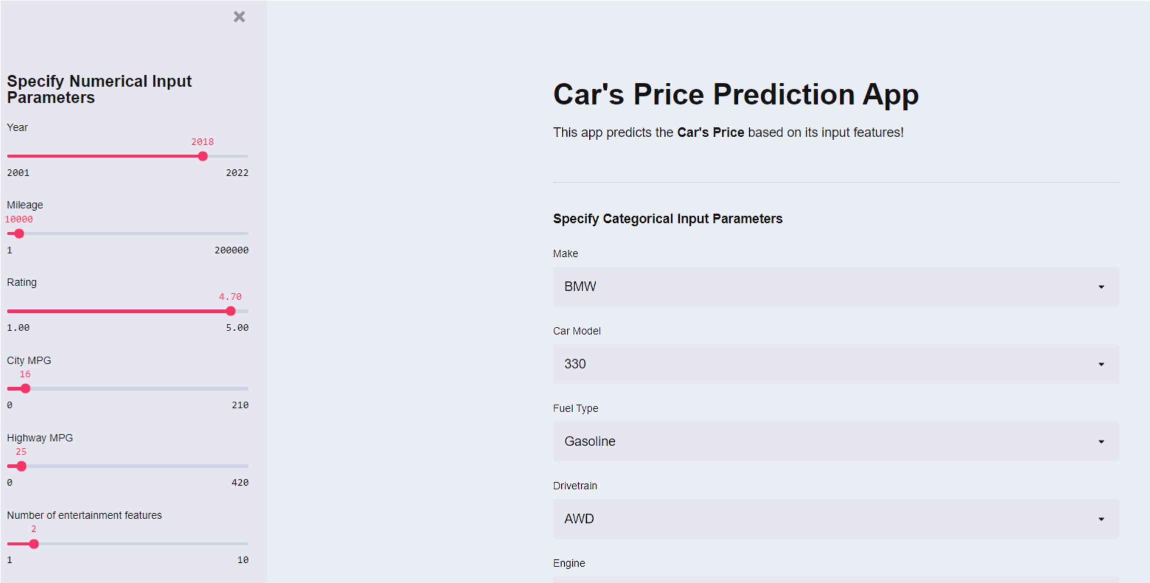
BUILD THE INTERACTIVE WEB APP AND DEPLOY IT TO THE CLOUD!

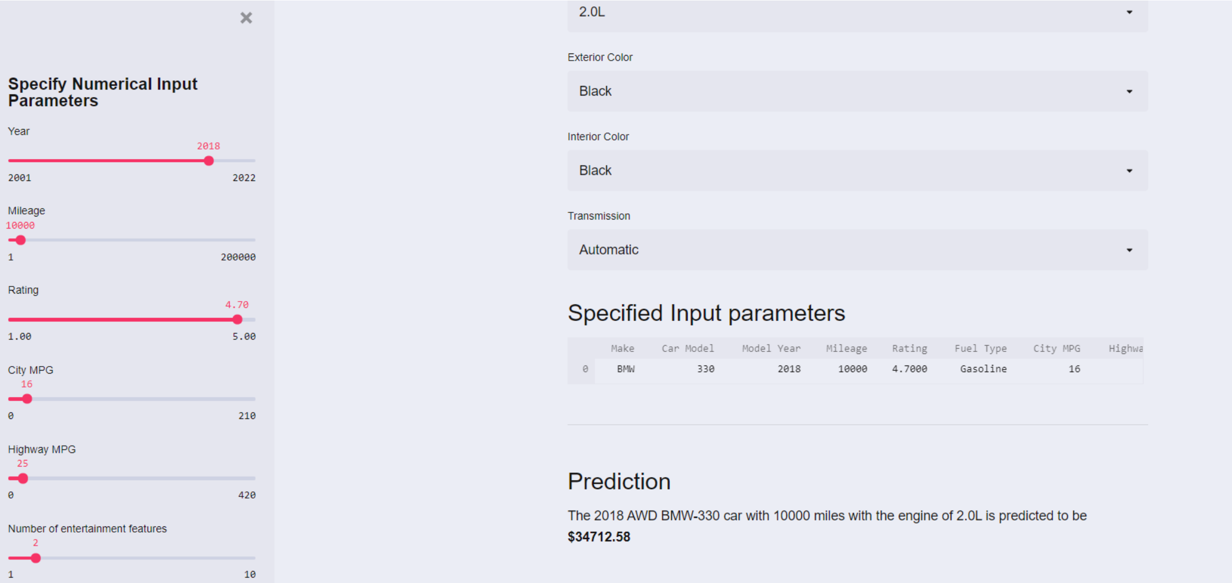
BUILD THE INTERACTIVE WEB APP AND DEPLOY IT TO THE CLOUD!

https://car-predictor-regression.herokuapp.com/

BUILD THE INTERACTIVE WEB APP AND DEPLOY IT TO THE CLOUD!

https://car-predictor-regression.herokuapp.com/





THANK YOU

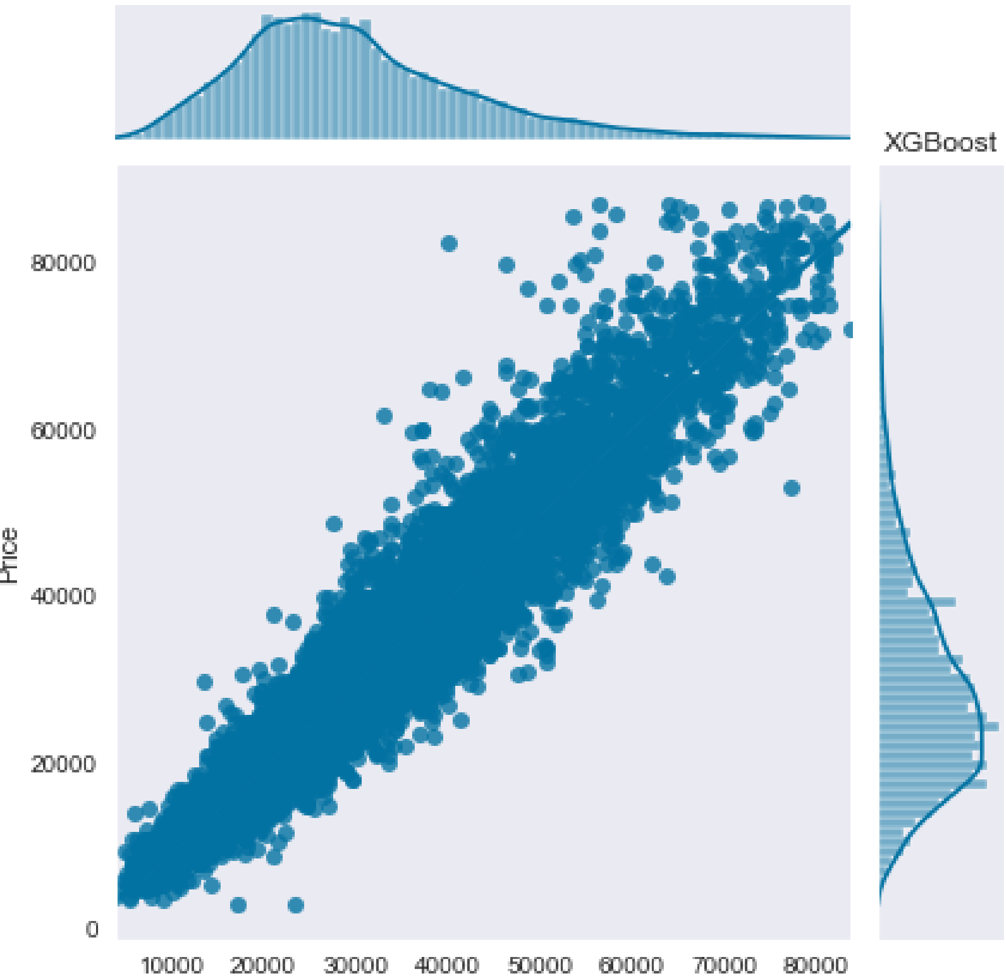
STEVEN L TRUONG

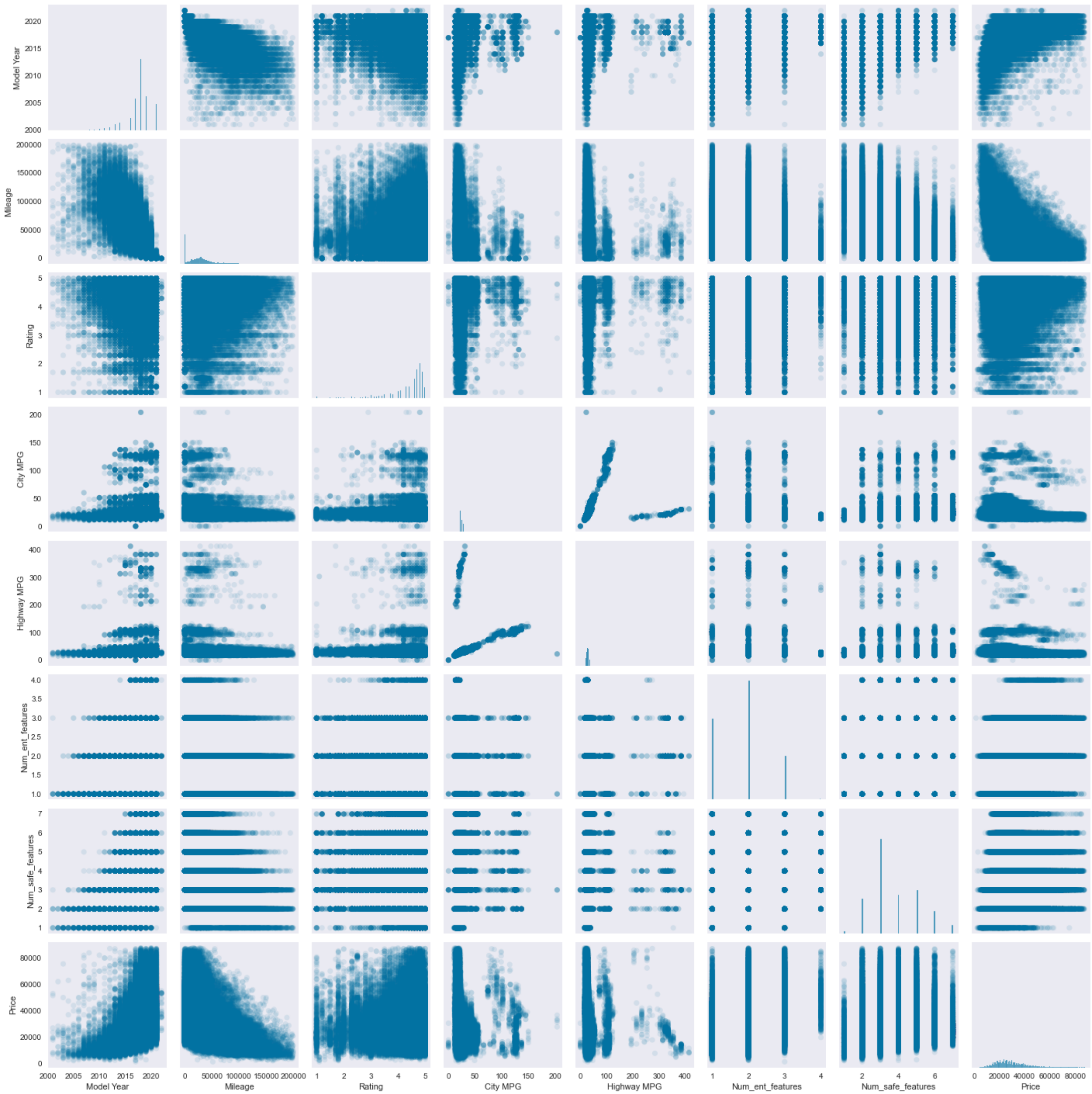
https://github.com/luongtruong77

tqluong77@gmail.com

QUESTIONS?

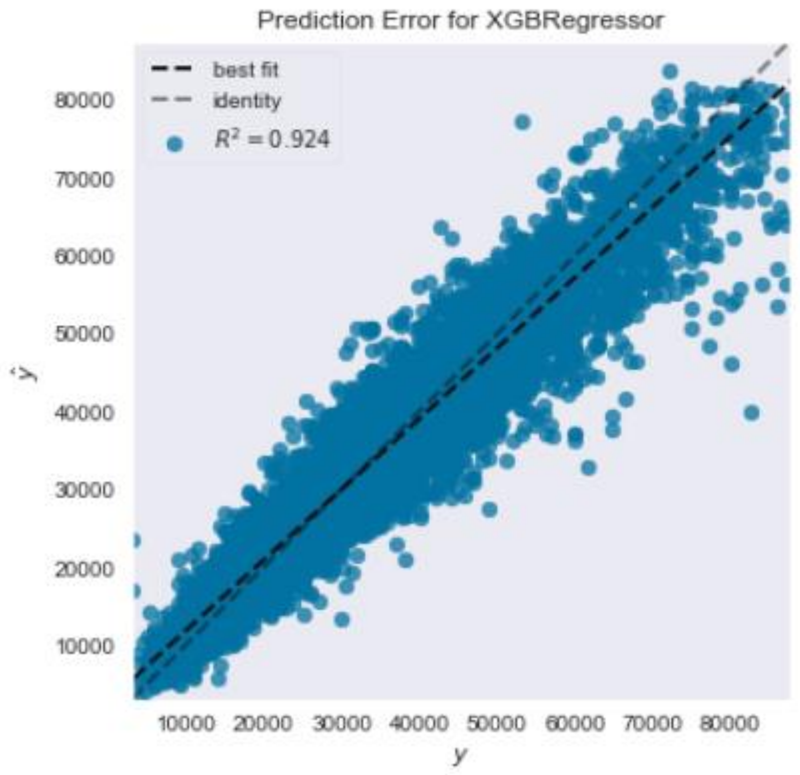
APPENDIX

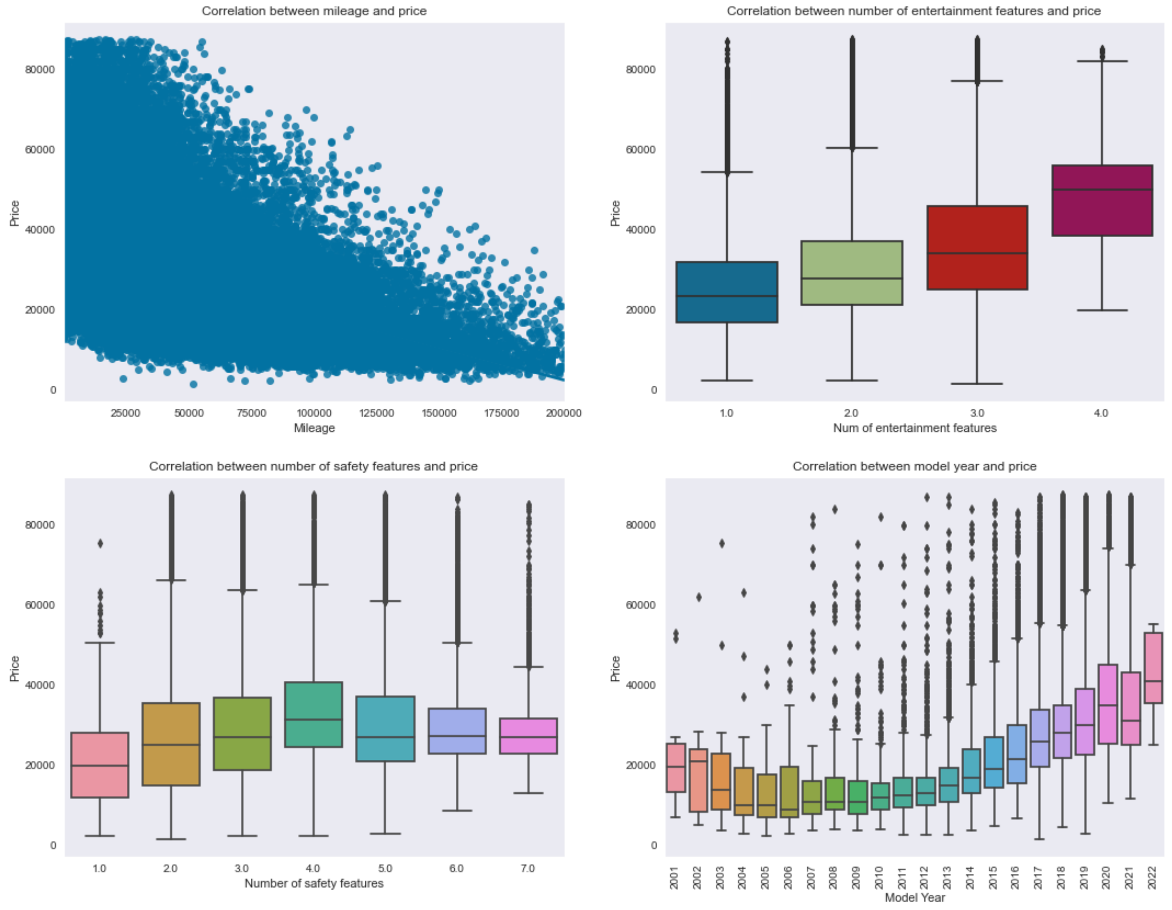

XGBoost
Price
80000
60000
40000
20000
0
10000
20000
30000
40000
50000
60000
70000
80000


Model Year
Mileage
Rating
City MPG
Highway MPG
Num_ent_features
Num_safe_features
Price

APPENDIX

Prediction Error Plot





Documentation on how to use streamlit to build the interactive app : https://streamlit.io/