A REGRESSION STORY: BUILDING A PRICE PREDICTIVE MODEL FOR CARS

Neshat J. Heravi

GitHub: Neshat-JH

10/13/2021

Introduction

- The cars prices have fluctuated in the past year or so due to the pandemic aftermath. The incentive is to create a predictive model based on recent data so that buyers have a more realistic information about car prices.

Objective

- The objective of this project is to build regression predictive models for car prices based on their features. The model with best evaluation metrics and highest accuracy will be selected for use.

Goal

- The goal is to help Pacific Northwest car buyers to have a more realistic idea of the prices amid recent price fluctuations.

Data

- Data is scraped from cars.com website using BeautifulSoup Python library.
- Data includes Seattle and Portland cities and areas within the 30 miles radius.

Seller's info

Elliott Bay Auto Brokers

4.7 ★★★★½ (305 reviews)

13001 Aurora Ave N Seattle, WA 98133

Features

Convenience	Navigation System Power Liftgate Remote Start
Entertainment	Apple CarPlay/Android Auto Bluetooth HomeLink Premium Sound System
Exterior	Alloy Wheels Sunroof/Moonroof
Safety	Backup Camera Brake Assist Stability Control
Seating	Leather Seats Memory Seat Third Row Seating

Used

Save

2018 Mercedes-Benz GLE 350 Base

35,219 mi.

$38,999

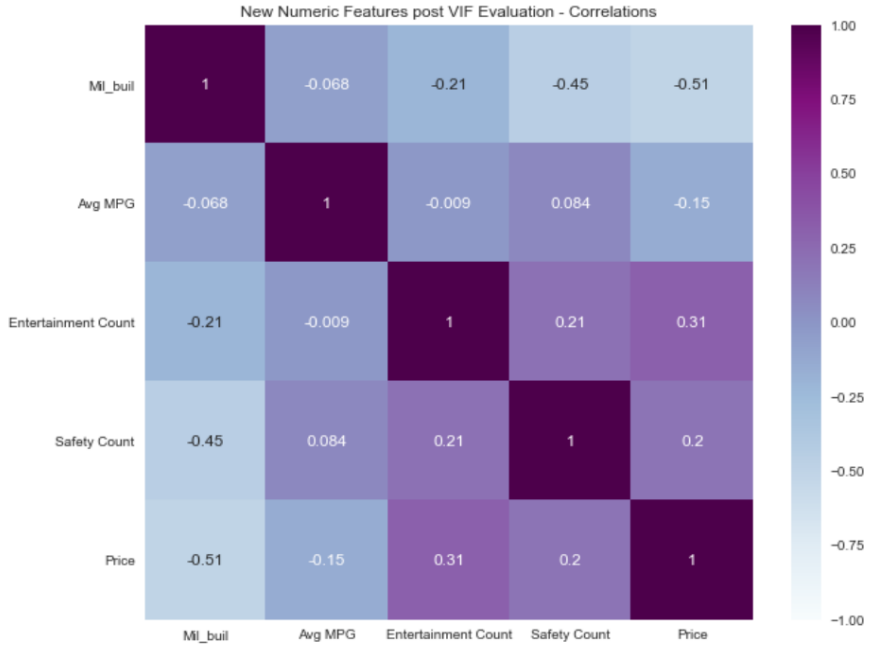
Good Deal | Home Delivery | Virtual Appointments

Basics

Exterior color	Polar White
Interior color	Black
Drivetrain	Rear-wheel Drive
MPG	0–23
Fuel type	Gasoline
Transmission	Automatic
Engine	Premium Unleaded V-6 3.5 L/213
VIN	4JGDA5JBXJB070148
Stock #	070148
Mileage	35,219 mi.
Vehicle history	CARFAX Report

Methodology

- Data cleaning:
  - *Disintegrating some columns: Built Year, Make and Model from Name*
  - *Detecting and handling missing data and outliers*
- Exploratory Data Analysis (EDA):
  - *Looking into feature correlations, VIF, pairplots etc.*
  - *Understanding the transformations needed to make data suitable for regression model*
- Building Models:
  - *Creating different regression models.*
  - *Using cross-validations, and using evaluation metrics to select the final model*



Regression Models: Evaluation Metrics

- **Base Linear Regression (numeric features only):**
  - R^2 Training set: 0.35
  - R^2 Test set: 0.37
  - MAE: 8,940.20
  - RMSE: 11,545.34

- **Linear Regression (post log transformation of target):**
  - R^2 Training set: 0.45
  - R^2 Validation set: 0.47
  - R^2 Test set: 0.45

- **Polynomial Regression:**
  - R^2 Training set: 0.55
  - R^2 Validation set: 0.55
  - R^2 Test set: 0.54

- **Lasso Regression:**
  - R^2 Training set: 0.87
  - R^2 Validation set: 0.85

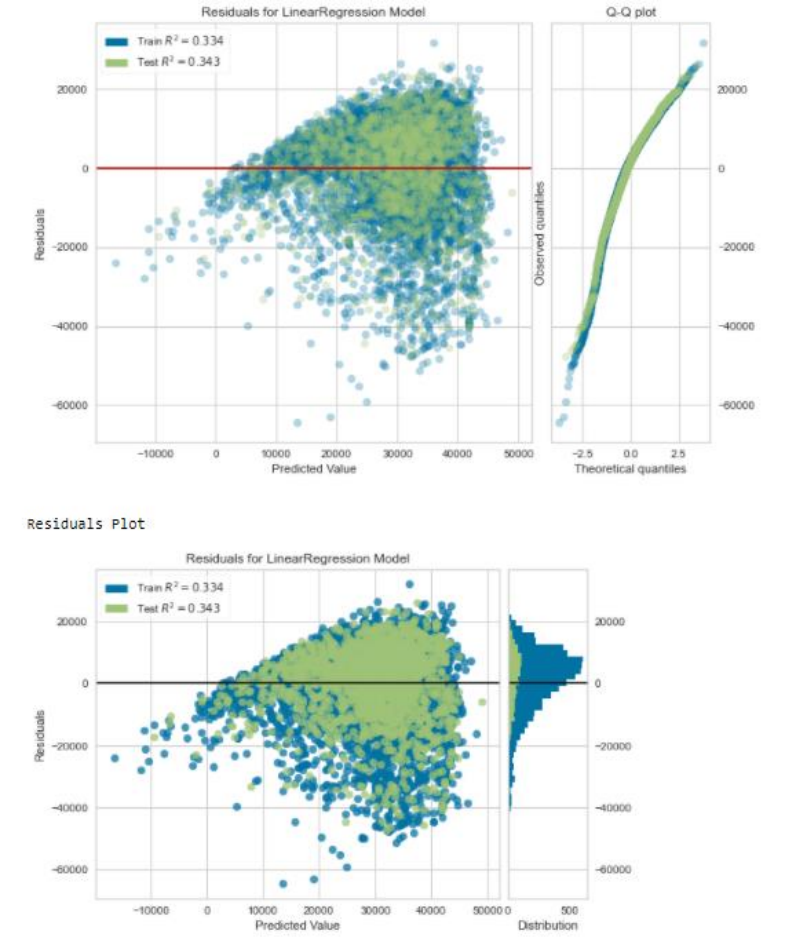
- **Ridge Regression:**
  - R^2 Training set: 0.88
  - R^2 Validation set: 0.85

- **Gradient Boosted Regressor:**
  - R^2 Test set: 0.80

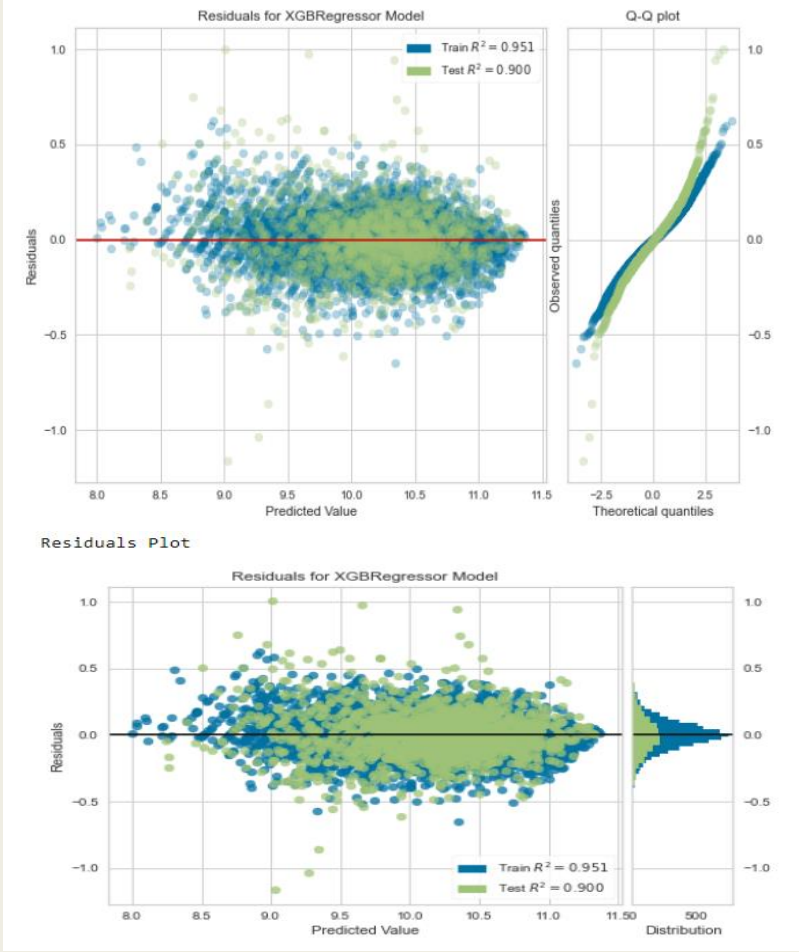
- **XGBoost Regressor:**
  - R^2 Training set: 0.95
  - R^2 Validation set: 0.95
  - R^2 Test set: 0.90
  - MAE: 3,323.50
  - RMSE: 4,682.70

Regression Models: Residuals Comparison

Residuals for Linear Regression

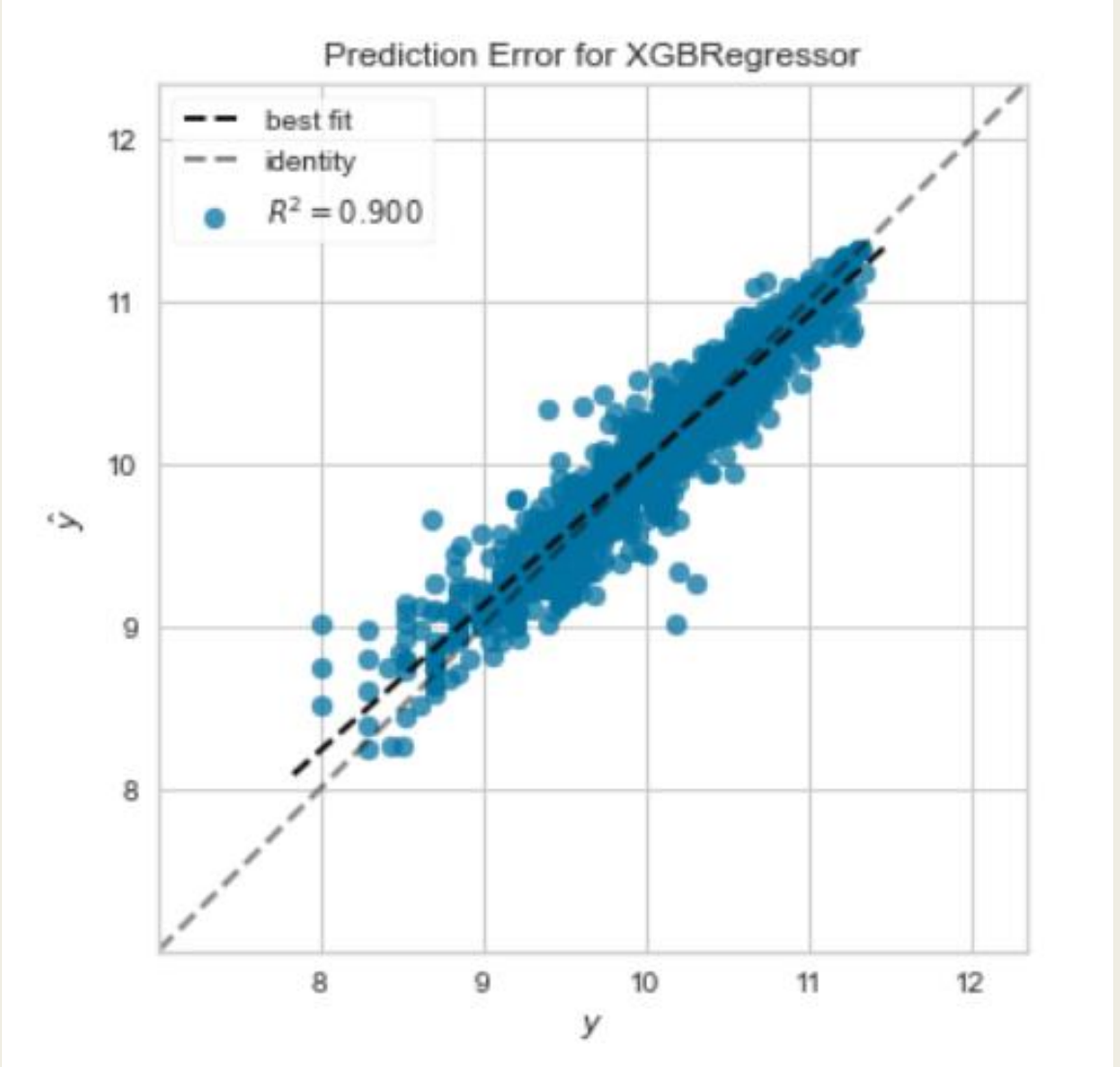


Residuals for XGBoost Regressor



Conclusion

- The objective of this project was to built and select the most accurate regression model to predict cars' prices. Thus, based on the evaluation metrics the **XGBoost Regressor** is the most accurate model with R^2 score of **90%** on the test set.

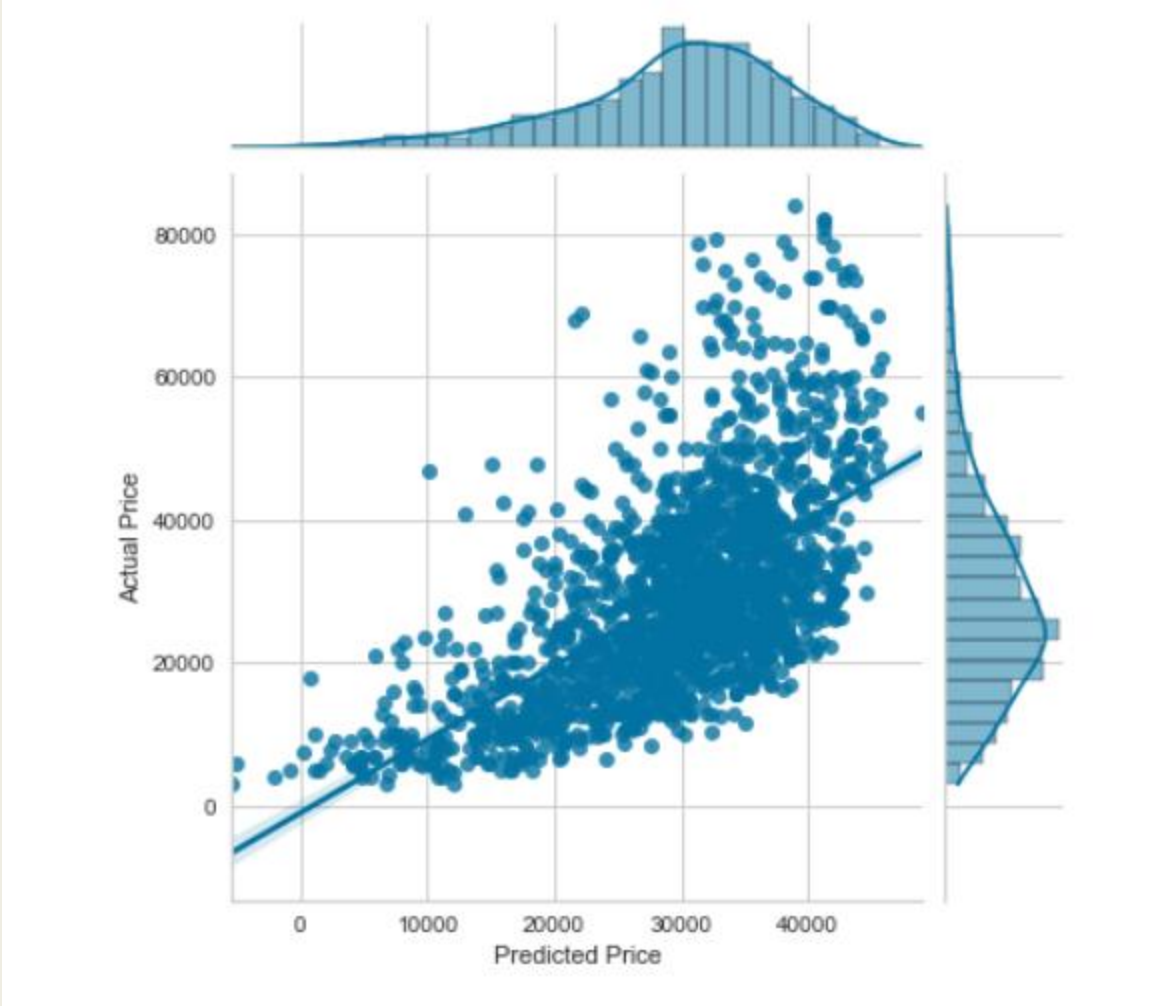


Future Works

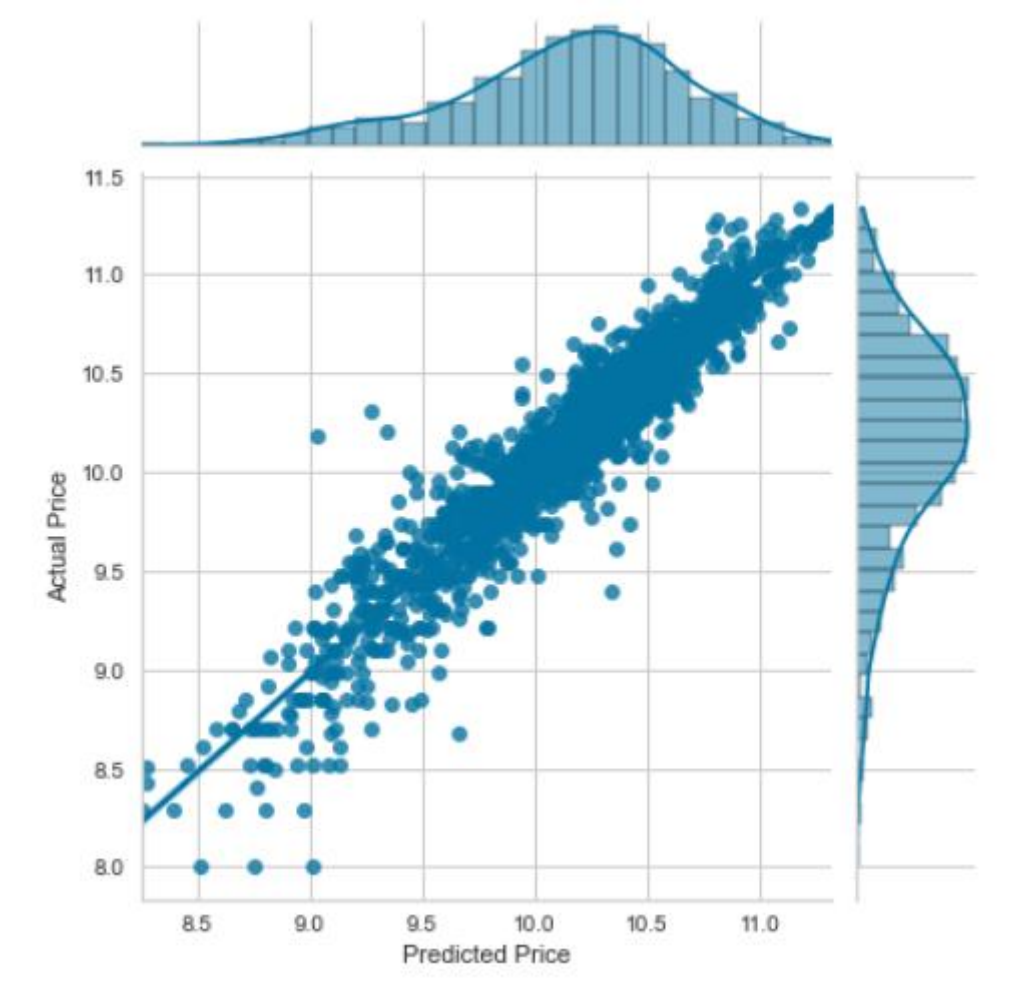
- More thorough data cleaning is suggested; specially handling the outliers will result in more accurate models
- A more detailed hyper-parameter tuning can increase the accuracy of Gradient Boosted and XGBoost regressors. (using GridSearch)
- Data from multiple sources can increase the reliability and also generalization of the models.

Appendix

Jointplot of Base Model: Linear Regression



Jointplot of Final Model: XGBoost Regressor



Appendix: Correlation of Numeric Features