UC San Diego

Used Car Data Analysis, Visualization and Price Prediction

Team 14: Bing Li, Jie Wu, Atman Patel, Shang Wang

UC San Diego

Contents

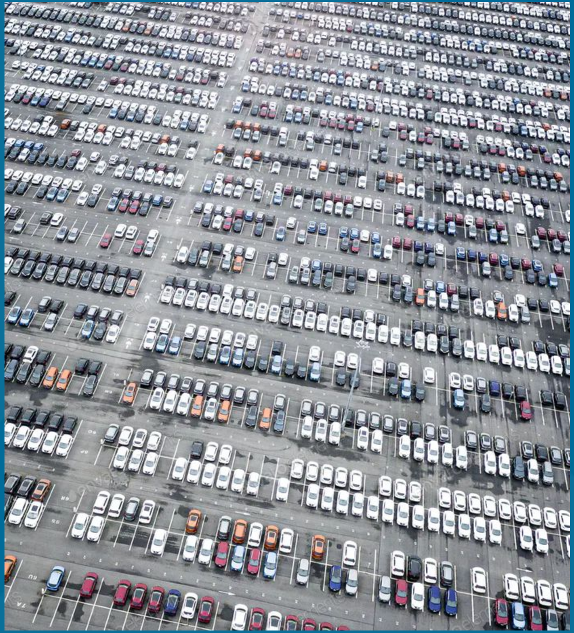
1. Background
2. Methodology
3. Data Cleaning --
4. Data Analysis/Vi
5. Training and pre
   -- Atman Patel, S
6. UI Showcase

UC San Diego

Background

- Motivation: As the number of car owners increase every year, the used car sales would go up. Predicting the resale price could help both buyer and seller negotiate the right amount.

- Goal: To be able to predict the selling price of a used car based on a number of parameters like manufacturer, condition, odometer, car type etc.



Methodology

- Clean the available dataset
- Explore and understand the data to identify relevant/useless features
- Analyze some parameters in detail
- Transform the data into the format a ML model can understand
- Contrast and compare the performance of different regression based models

Dataset

Used car resale dataset

- 400k data points
- Mostly limited to United States
- 25 features

id	url	region	region_url	price
manufacturer	model	condition	cylinders	fuel
title_status	transmission	vin	drive	size
paint_color	image_url	description	county	state
year	odometer	type	lat	long

Data Cleaning

Dropped uncorrelated/redundant columns:

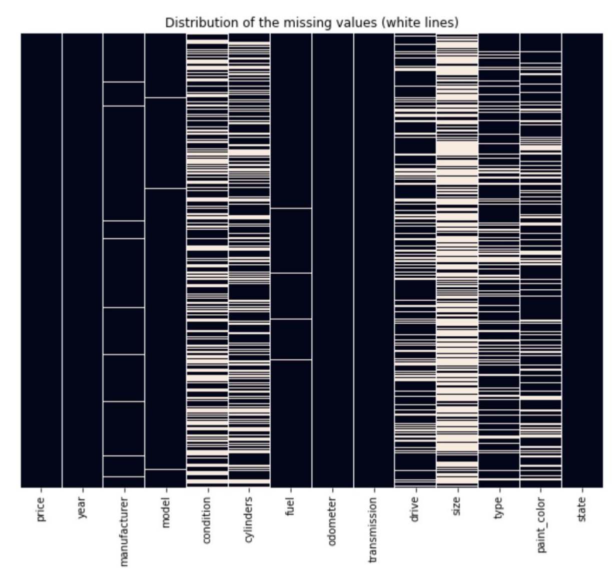
Id, url, region_url, title_status, vin, image_url, description, county, long, lat.

Dealing with missing samples:

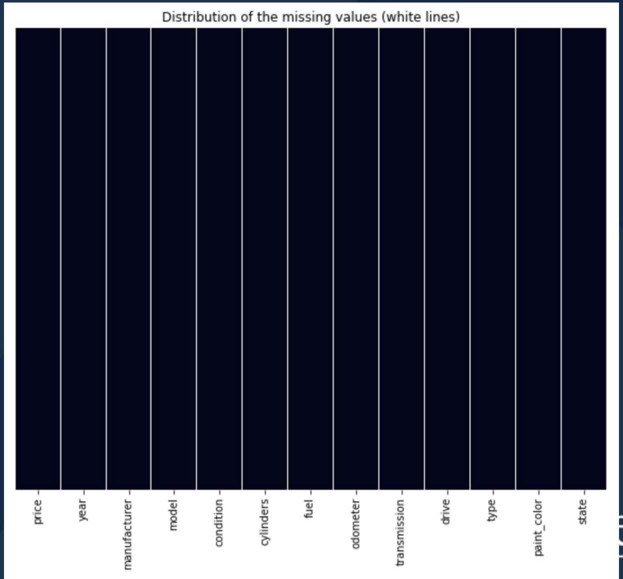
Almost missing all values: Delete the column

Missing several values: Delete those rows

Before

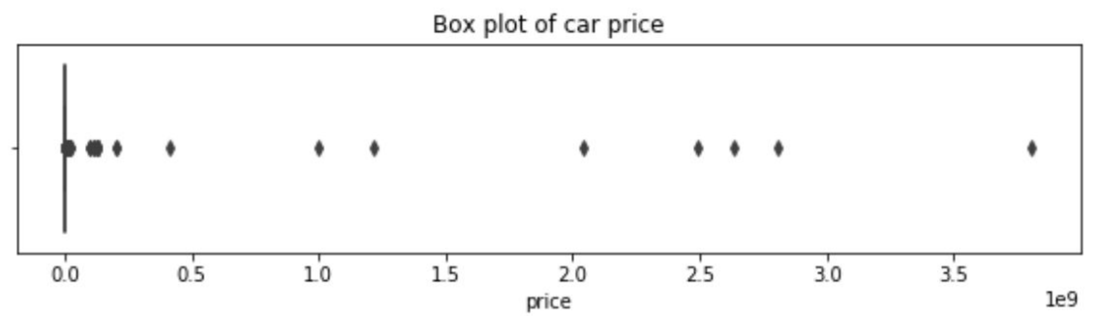


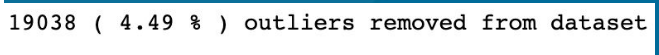
After



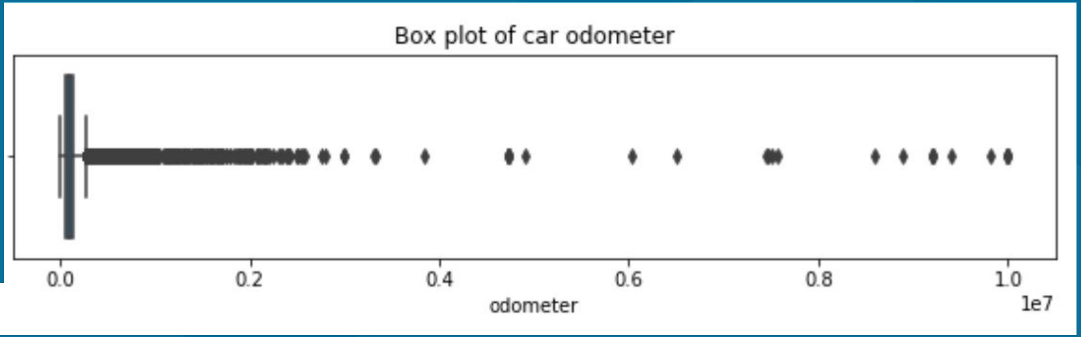
UC San Diego

Data Cleaning - Remove outliers (for car price and odometer)



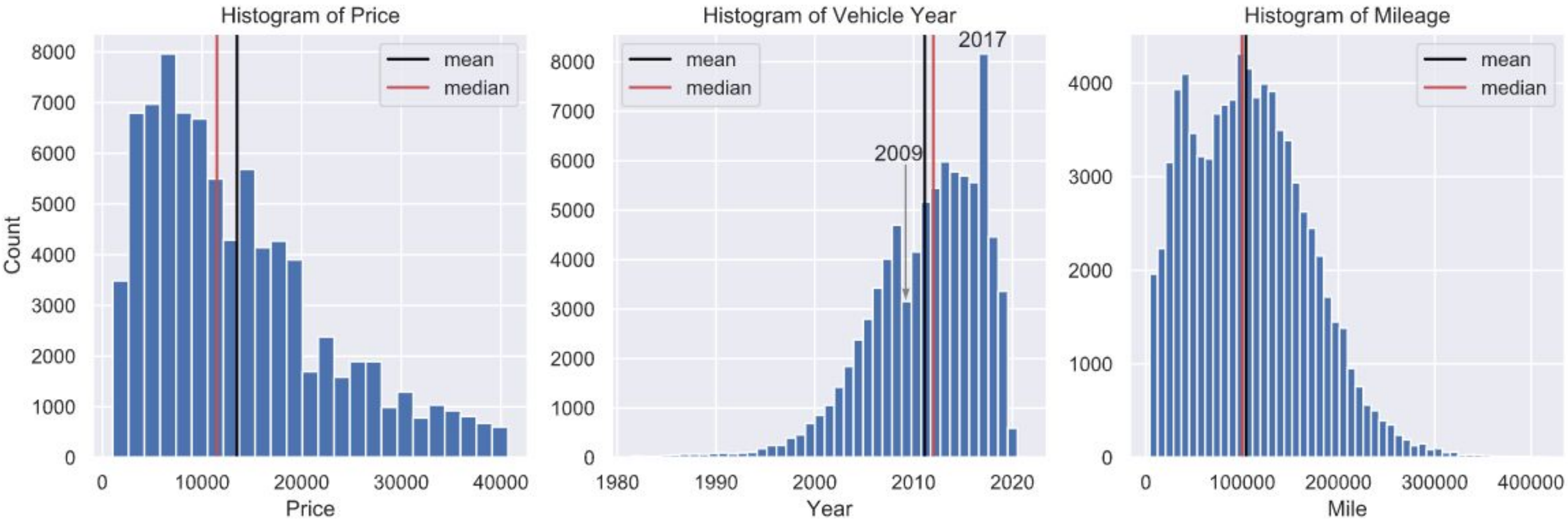


```
19038 ( 4.49 % ) outliers removed from dataset
```



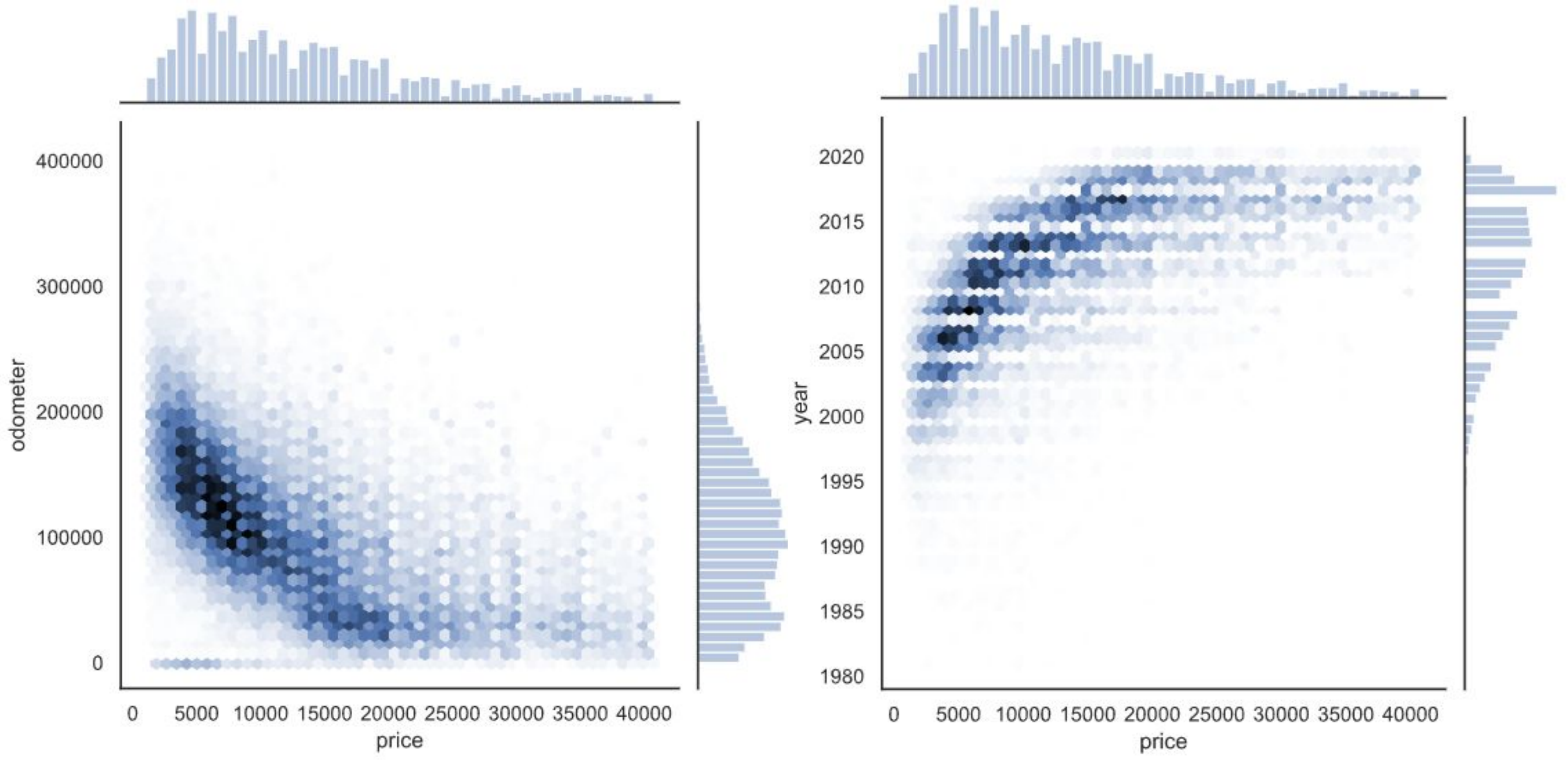
```
148365 ( 36.65 % ) outliers removed from dataset
```

Data Analysis - Histograms



Data Analysis - Continuous Variables

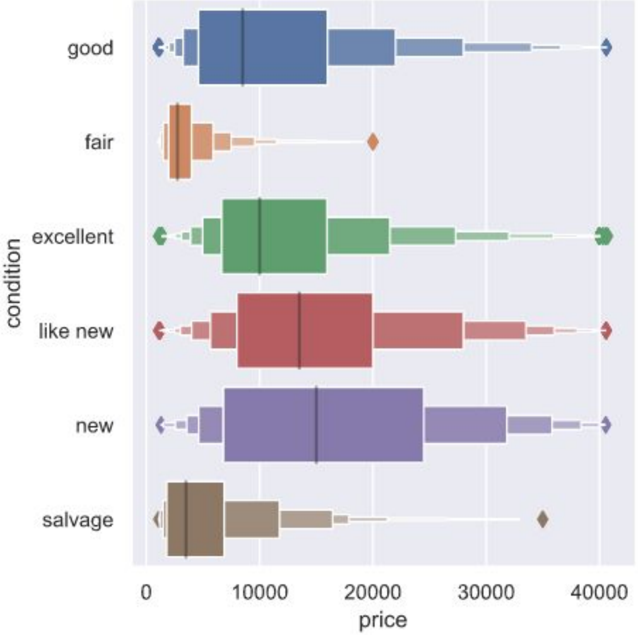
Price vs. Mileage and Year



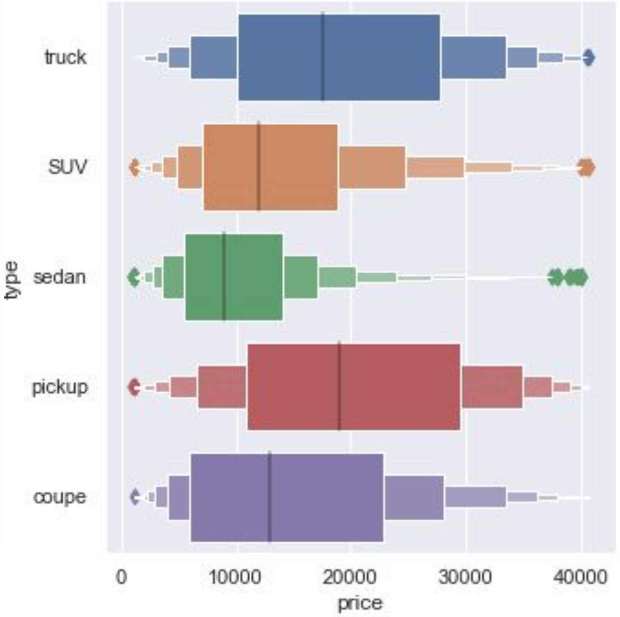
UC San Diego

Data Analysis - Discrete Variables

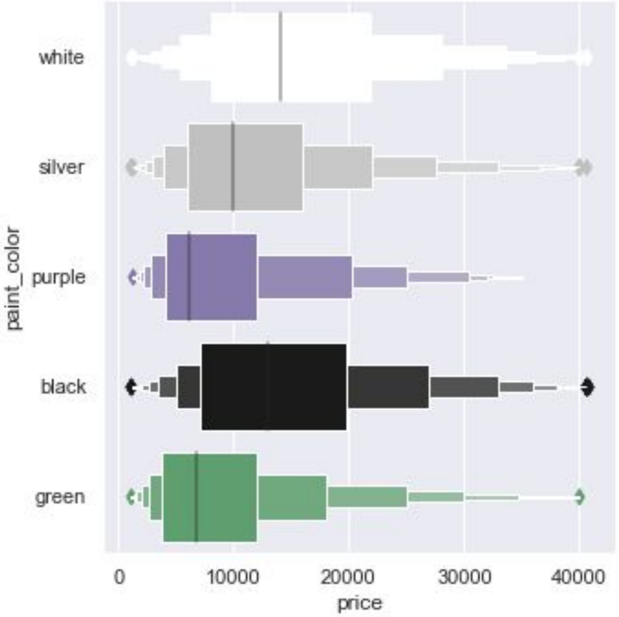
Price vs. Condition, Type and Color



Price vs Condition

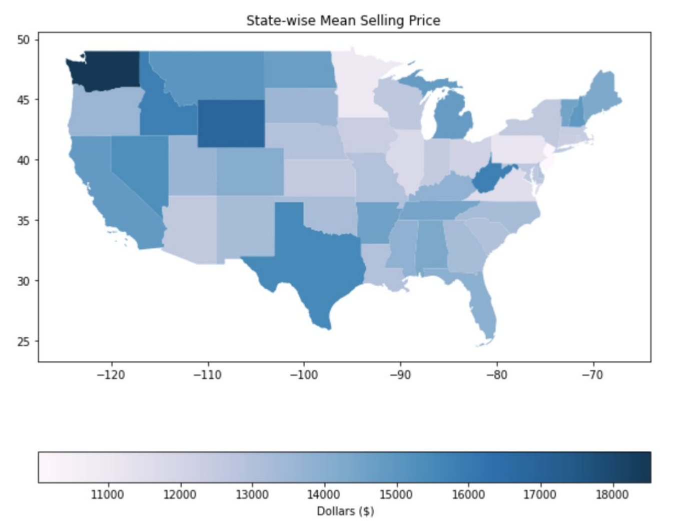


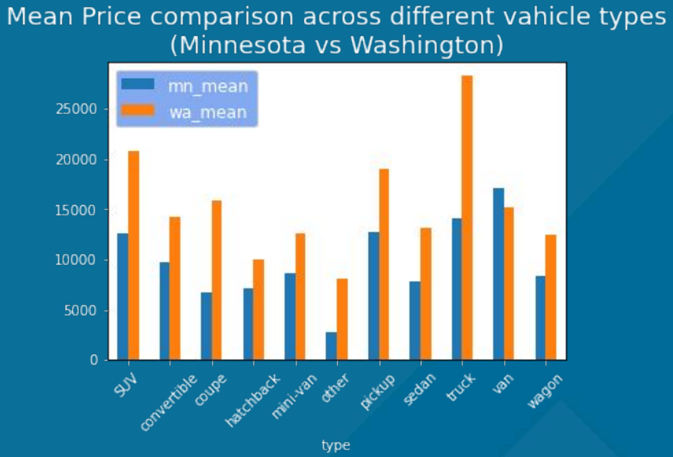
Price vs Type

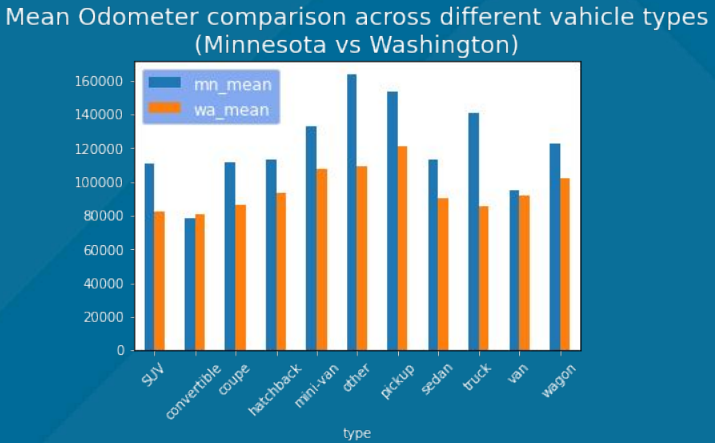


Price vs Color

State wise differences


State-wise Mean Selling Price


Mean Price comparison across different vahicle types (Minnesota vs Washington)


Mean Odometer comparison across different vahicle types (Minnesota vs Washington)

Price Prediction

As we saw, the resale price is affected by a lot of factors.

Now that we’re identified the important ones, we need to transform these features:

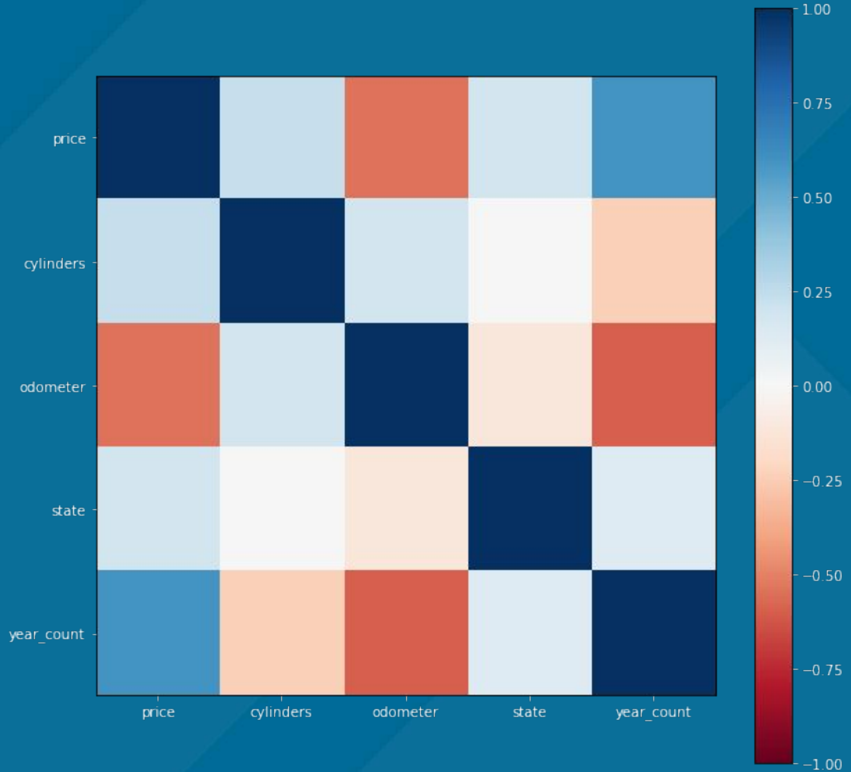
- Use as it is (real number values)
- One hot encoding
- Bucketing
- Drop/Remove

UC San Diego

Price Prediction

Correlations between

- Price
- Year
- Manufacturer
- Cylinders
- Odometer



Price Prediction

Models:

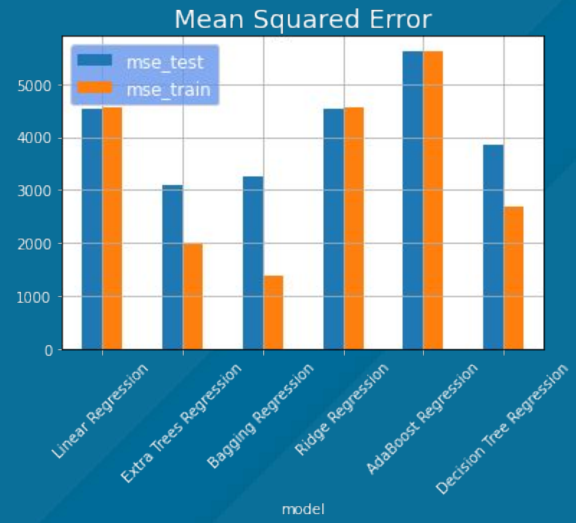
1. Linear Methods - Linear Regression, Ridge Regression
2. Ensemble Methods - Decision Tree Regression, Bagging Regression, Extra Trees Regression, AdaBoost Regression

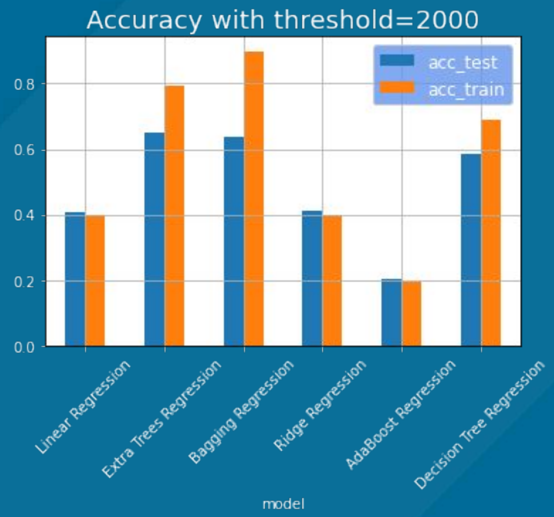
Evaluation Metrics:

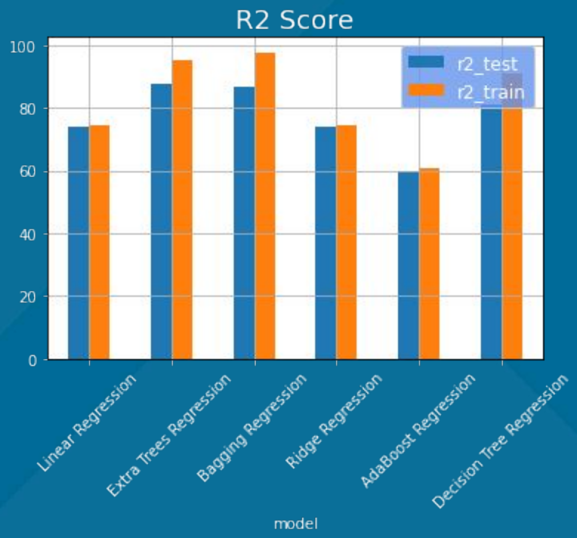
1. Accuracy - Correct price (ground truth) within Predicted price +- threshold
2. R2 Score

UC San Diego

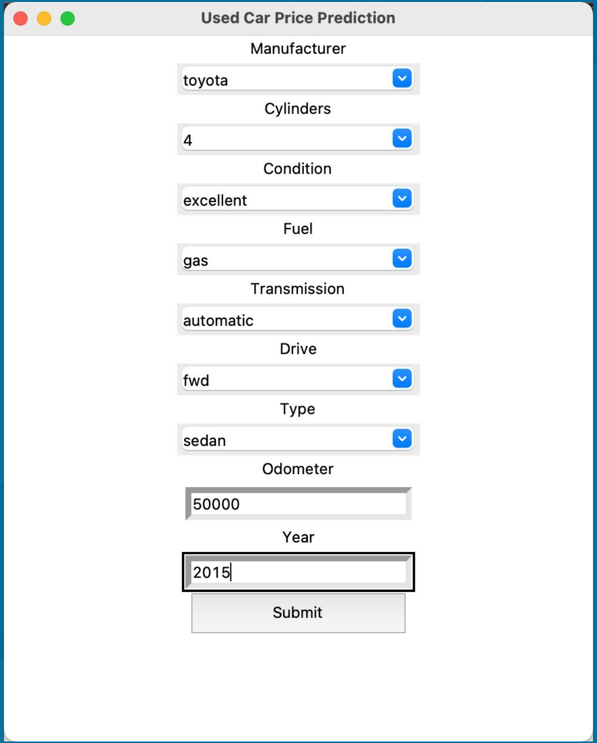
Price Prediction - Model Comparison







Price Prediction UI



Used Car Price Prediction

Manufacturer: toyota

Cylinders: 4

Condition: excellent

Fuel: gas

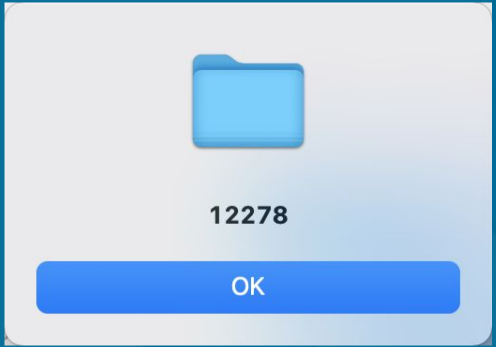
Transmission: automatic

Drive: fwd

Type: sedan

Odometer: 50000

Year: 2015



12278

Future work

1. Try Neural Network.
2. Use larger dataset to get more accurate predicted price.
3. Try to use vehicle model information as well.
4. Use pygal to plot interactive figures for visualization based on this dataset.

Thank You!

Questions?