Sale Price on Usage, Equipment Type, Configuration

Chul, Morgan, Walker

Goals

- Predict Sale Price based on features of the data
- Use Linear Models to Minimize RMSLE
- Score Models with Testing set

- Features we wanted to work with:
  - Year Made, Machine Hours Current Meter, fiModel, Product Size, Drive System, Ride Control, Stick, Turbocharged

Git Work Flow

One team member forked linear-regression Repo

Others cloned Repo locally and navigated to separate branch

Worked on separate branches on different features

git add .

git commit -m ‘message’

git push origin <select branch>

Merged branches with Master and combined features

The Data

Zipped CSV File

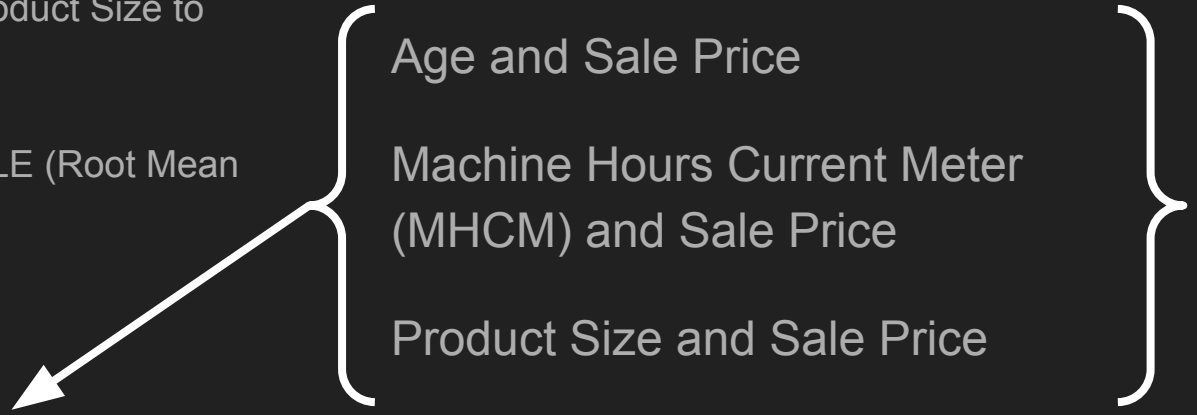
- 401125 rows and 54 columns
- Pandas, Numpy
- LOTS of NaN ‘s

What features can we even include?

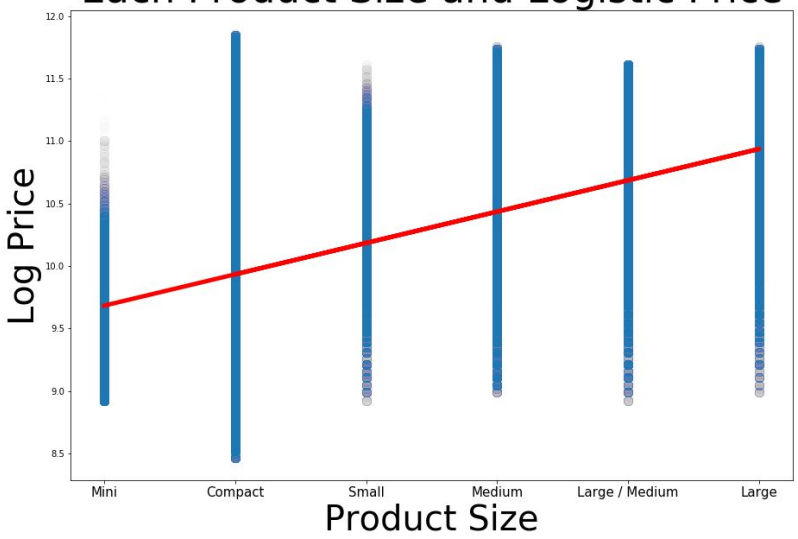
- Imputing Missing values
- Looking at certain years

Modeling

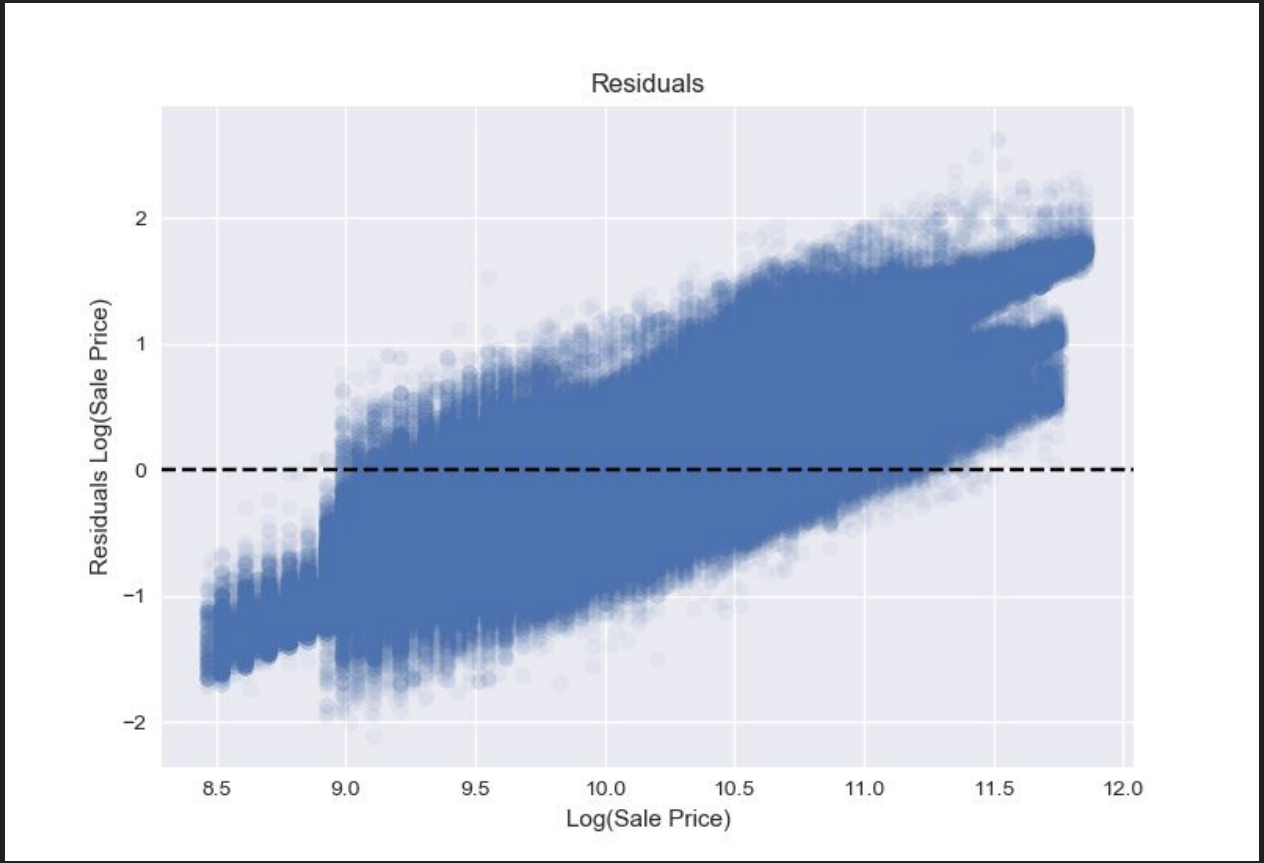
- OLS Linear Regression
  - Age, Machine Hours, Product Size to predict Sale Price
  - Summary of MSRE
  - Scoring Method of RMSLE (Root Mean Squared Error)



Final model using Age, MHCM, Product Size to predict Sale Price


Each Product Size and Logistic Price
Log Price
12.0
11.5
11.0
10.5
10.0
9.5
9.0
8.5
Mini
Compact
Small
Medium
Large / Medium
Large
Product Size


Usage Price Model
Log(Sale Price)
12.0
11.5
11.0
10.5
10.0
9.5
9.0
8.5
0
5000
10000
15000
20000
25000
30000
Machine Hours


Residuals
Residuals Log(Sale Price)
2
1
0
−1
−2
8.5
9.0
9.5
10.0
10.5
11.0
11.5
12.0
Log(Sale Price)

Performance

OLS Linear Regression Model, R squared = 0.35

Scoring on RMSLE and modeling on RMSE