What's it worth?

Predicting the price of second hand cars

Radhika Mardikar, Hil Alcee, Ricollis Jones, Tres Pimentel

Motivation

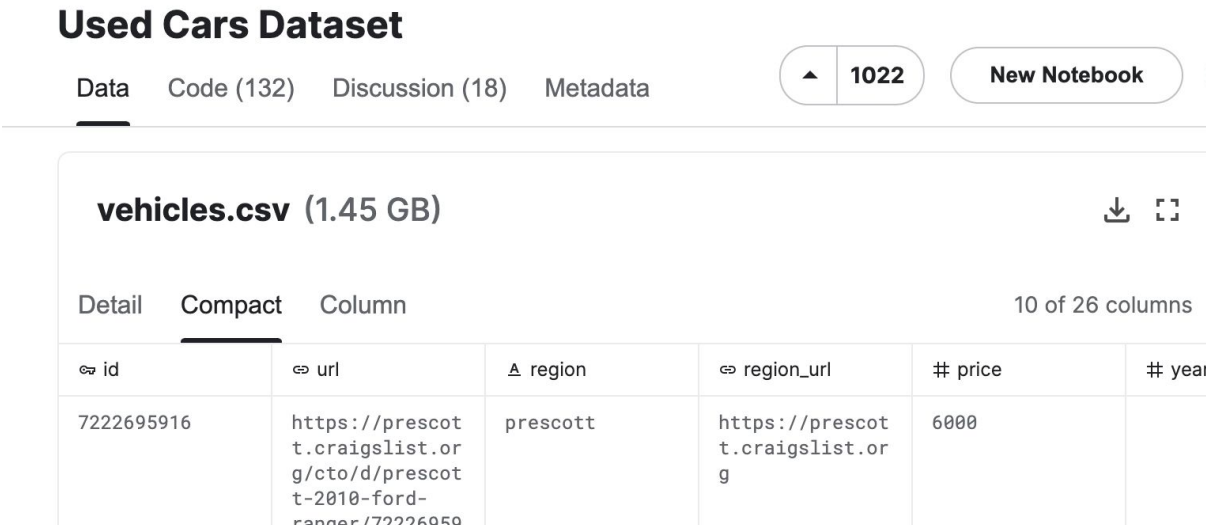
- Huge market for 2nd hand cars
- Great variability in price
- Predict the price of 2nd hand cars

Goal

- Try to find something interesting
- Tell users how much (exactly) their car is worth
- Give users a ballpark estimate of car worth

Dataset

- Kaggle (Craigslist)
- 500k examples
- Lots of features!

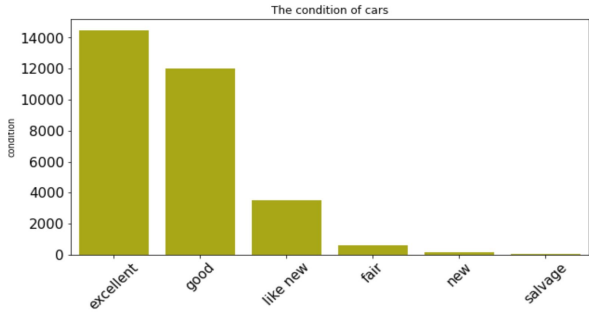


EDA

- Chose top 17 Manufacturers to keep within scope of project and timeline
- Dropped where user entry was not included or NaN values to aid with removing skewness from data.
- Outliers skewed certain numerical categories greatly (Price, Odometer, Year)
    - Set thresholds to remove outliers ($50,000 max, Greater than $1 min, 180,000 Miles max, 1960 year minimum)
    - Incorrect entry possible as there were prices in the trillions.
- Dropped any unnecessary columns that didn't pertain to scope of project or model building (I.E. VIN #, Lat, Long, image URL,etc.)
- Majority of our sample of used cars were within the Excellent-Good condition range as car condition plays a role in the price of a used car's resale value.

```
print("The maximum and minimum car prices")
print(df['price'].max())
print(df['price'].min())
df['price'].nsmallest(n=10000)
```

```
The maximum and minimum car prices
3736928711
0
```

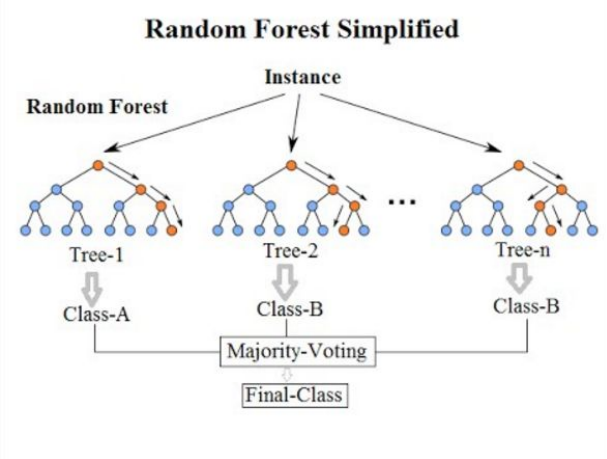


Model Development

- Target variable - Price
- 80/10/10 split
- Linear Regression model
- Polynomial regression model
- Decision Tree
- Random Forest Regressor
- Random Forest Classifier

Results

- Random Forest Regressor with n_estimators = 18, max_depth = 11, criterion = ‘mse’
  - R2 = .868
  - About 87% percent of the variance in price can be predicted from our model
- Random Forest Classifier with 4 price bins, n_estmators = 12, max_depth = 11, criterion = ‘entropy’
  - Accuracy = .686
  - Bin size = $12,500
    - Accuracy vs utility?



Further work on this dataset

- Model: user input field with a lot of variability
  - Inclusion in model gave us very low accuracy, curse of dimensionality, low # of observations for each unique model type
- Manufacturers
  - Using full list of manufacturers gave us poor accuracy