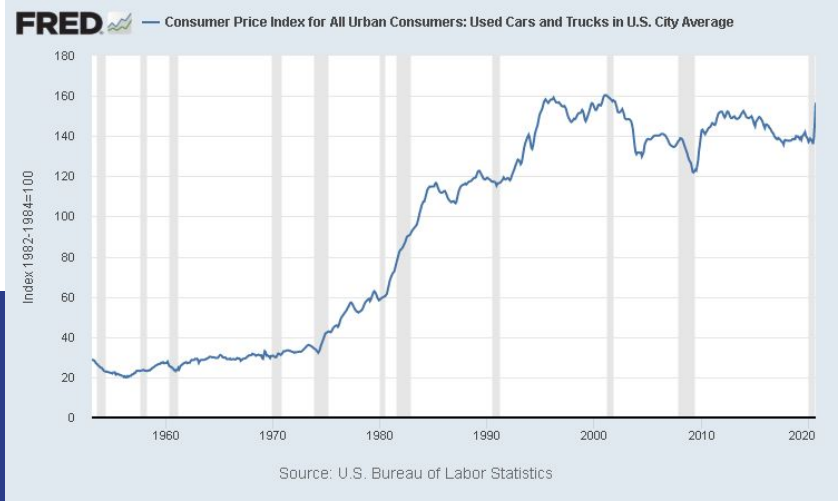
Used Car Price Prediction

Predicting Used Car Resale Price

First Capstone Project
Daniel Kim

Problem Statement & Goal

How can you predict used car resale price?



Data Wrangling

1. Redundancy Removal - Eliminate the same posts with multiple times for more exposure

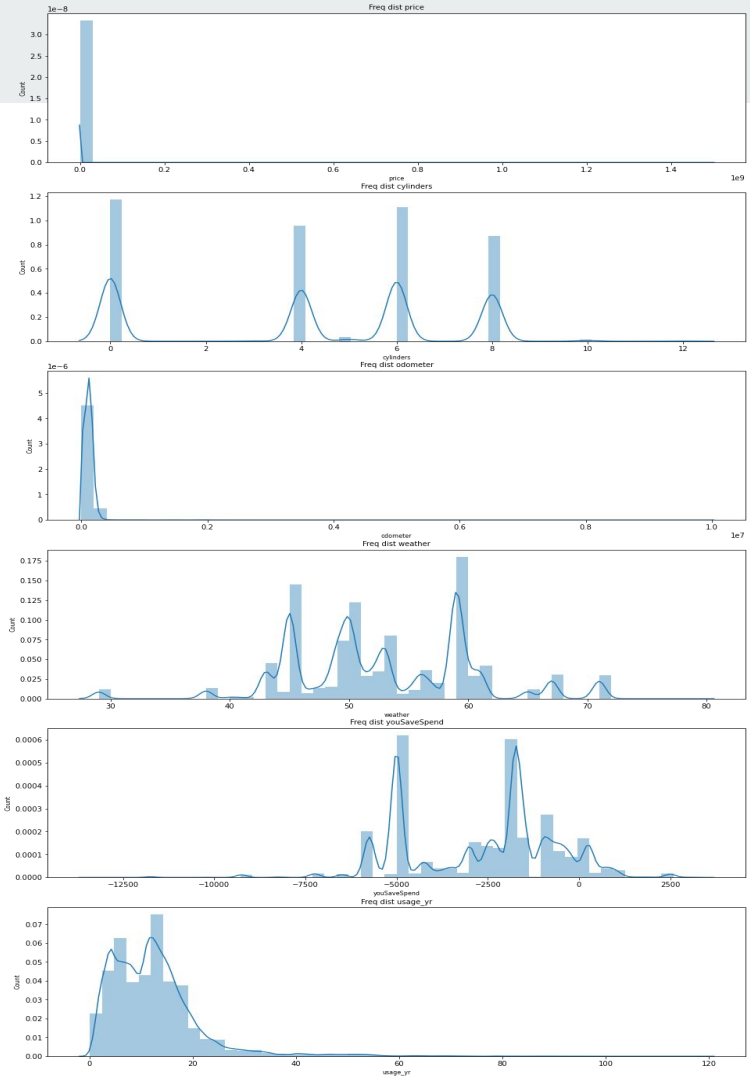
2. Null Values - NaN
    a. Drop rows with unidentifiable values in ‘price’, ‘location’, ‘year’, ‘make’, etc.
    b. Fill in NaN with external data with the following sources:
        i. Reverse Geocoder for exact geo-location
        ii. Merging with external dataset to fill in ‘manufacturer’, ‘cylinder’, ‘drive’, etc by ‘make’.
        iii. Fill in NaN with Majority values

Exploratory Data Analysis - Histogram

After Data Normalization, all other numeric features showed no normal distribution except for ‘usage_yr’, which has right-skewness.

This can be interpreted that most used car seller would like to sell their car within around 20 years in that a used car with under 20 years of usage cannot be considered as ‘Classic Car’ which has sufficient historical interest to be collectible and worth preserving or restoring rather than scrapping.

https://en.wikipedia.org/wiki/Classic_car

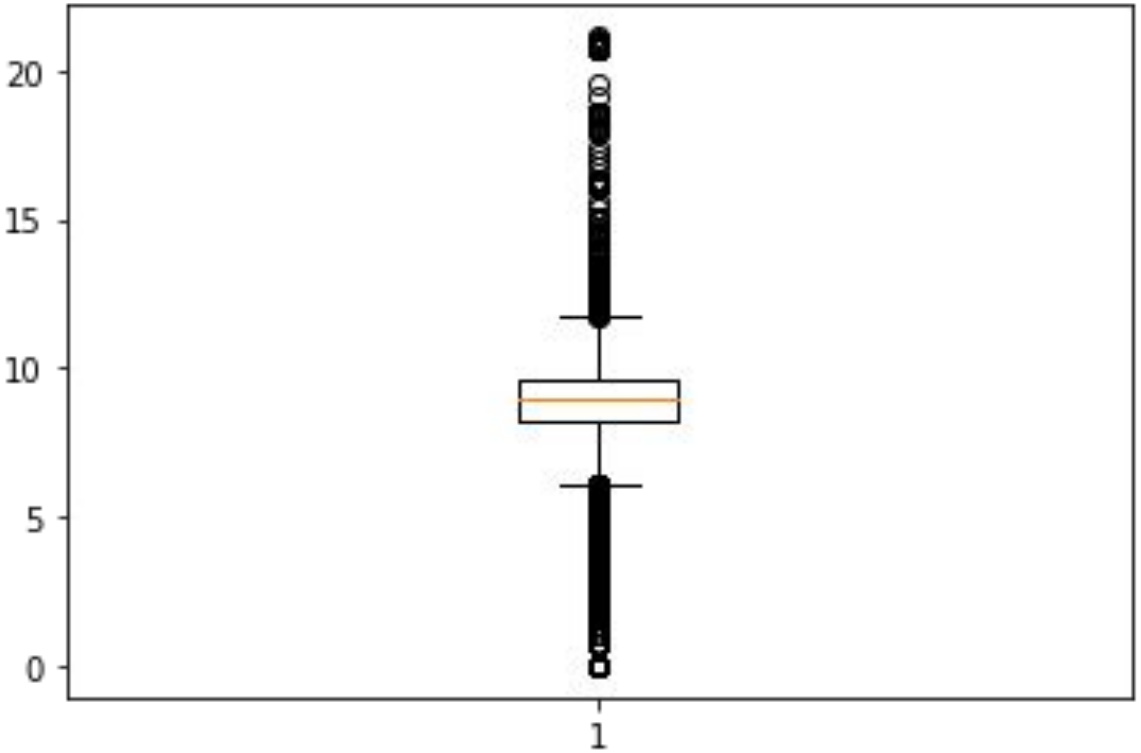


Exploratory Data Analysis - Heatmap

As you see, numerical features have not distinctive correlation to ‘price’.

Exploratory Data Analysis - Outliers

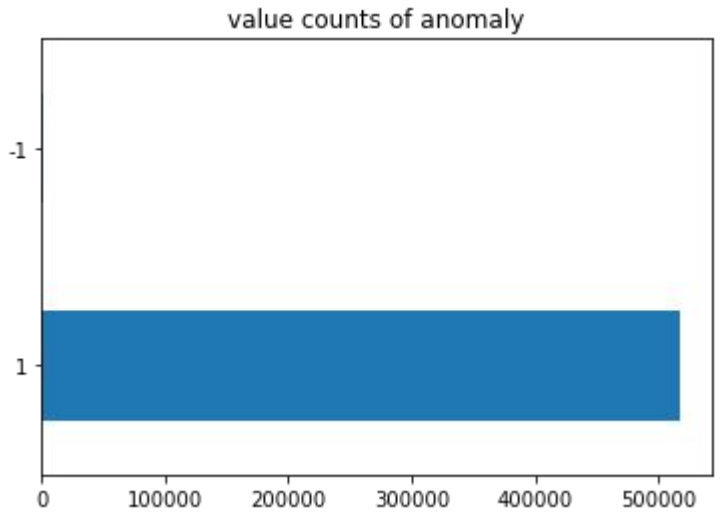
As target, ‘price’ has too many outliers even though ‘price’ was processed with ‘log’



Machine Learning - Removing Outliers

With Isolation Forest algorithm, 0.04% of outliers (-1) were identified as 2,078 rows.

The 2,078 outliers were removed.



Machine Learning - Feature Importance

Among 3 different algorithms, ‘odometer’ and ‘usage_yr’ have key importance to determine used car prices.

Algorithms	Decision Tree Regressor	Random Forest Regressor	Gradient Regressor
Feature Importance %	**‘Odometer’ : 46.8%** ‘Weather’ : 9.5% **‘Usage_yr’ : 15.5%** ‘Vclass_sport utility vehicle - 4wd’ : 5.3% ‘VClass_standard pickup trucks 4wd’ : 9.8%	‘Cylinders’ : 5.1% **‘Odometer’ : 33.6%** ‘Weather’ : 8.6% **‘Usage_yr’ : 16.9%**	**‘Odometer’ : 16.6%** **‘Usage_yr’ : 18.7%** **'division_Middle Atlantic' : 18.3%** ‘Manufacturer_cadillac’ : 6.2% **‘manufacturer_chevrolet’ : 14.6%**

Machine Learning - Prediction Accuracy

Random Forest Regressor showed the best score (only positive as 0.005) in spite of an enormous amount of Run Time (853.5 minutes).

Gradient Boost Regressor showed disappointing results: the lowest score (-0.31) in consideration with a lot of Run Time taken.

Algorithms	Decision Tree Regressor	**Random Forest Regressor**	Gradient Boost Regressor
Accuracy (R-Squared)	-0.000182501274920 03276	**0.0051472583318550 31**	-0.308376677905855 55
Randomized Search CV Hyper Parameter Run Time	2.5 seconds	**853.5 minutes**	112.4 minutes
Best Parameters by Randomized Search CV	'splitter': 'random', 'min_samples_split': 15, 'min_samples_leaf': 10, 'max_features': 'log2', 'max_depth': 5, 'criterion': 'friedman_mse'	**'n_estimators': 120, 'min_samples_split': 15, 'min_samples_leaf': 5, 'max_features': 'sqrt', 'max_depth': 15, 'criterion': 'mse'**	'n_estimators': 300, 'min_samples_split': 100, 'min_samples_leaf': 5, 'max_features': 'sqrt', 'max_depth': 25, 'loss': 'ls', 'criterion': 'friedman_mse'

Recommendation - For More Accurate Price

1. Use Random Forest Regressor
2. Use parameters as below:
    a. ‘n_estimators': 120
    b. 'min_samples_split': 15
    c. 'min_samples_leaf': 5
    d. 'max_features': 'sqrt'
    e. 'max_depth': 15
    f. 'criterion': 'mse
3. Cut Down ‘usage_yr’ by 20 years.
4. Limit such conditions as location, weather, type, etc.