Capstone II: Craigslist Used Car Dataset

Biman S. Mondal
June 3, 2025

Problem Statement

The objective of this project is to predict the price of a commonly driven used vehicle in the US. Kelly blue book and other references exist for pricing a single vehicle, but those resources don't offer a way to understand trends. The dataset for this problem is sourced from Kaggle. The used car dataset was developed by scraping primarily Craigslist. Craigslist is a region-specific online marketplace used to sell anything from tools to services to used vehicles. The data was scraped in April and May of 2021 and comprises all the Craigslist regions within the US.

Approach

The project followed the data science methodology. This report comprises the final step in the data science process.

1. Data Wrangling
2. Exploratory Data Analysis (EDA)
3. Pre-processing / Feature Engineering
4. Modeling
5. Documentation

The project utilized *Python* language and *Pandas* dataframes to interrogate the dataset. The plots are created using *Matplotlib* and *Seaborn* libraries. The modeling of the data uses the *sklearn* library. All the code is housed across Jupyter notebooks.

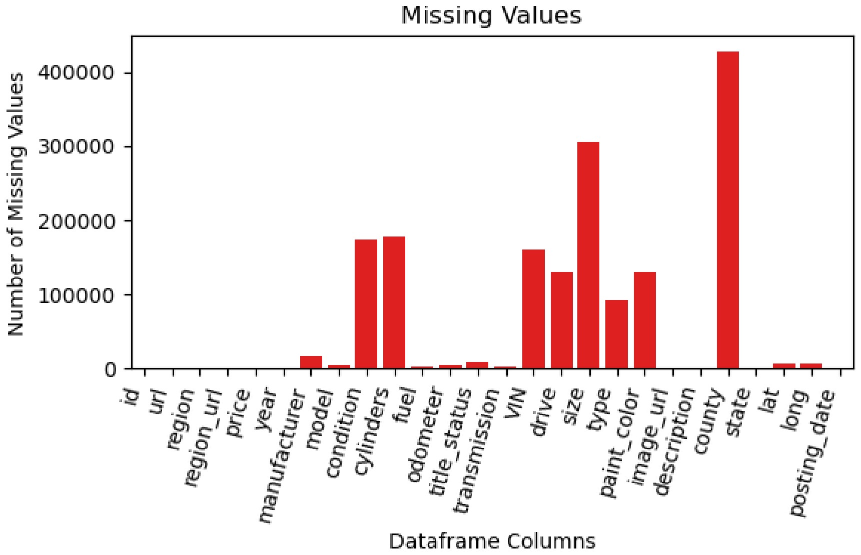
Data Wrangling:

The raw dataset is a .csv file of size 1.4 gB with 426,880 records and 26 fields. The following table is of the columns in the original dataset and their respective data type.

Table 1 - Dataset

#	Column	Non-Null Count	Dtype
0	id	426880 non-null	int64
1	url	426880 non-null	object
2	region	426880 non-null	object
3	region_url	426880 non-null	object
4	price	426880 non-null	int64
5	year	425675 non-null	float64
6	manufacturer	409234 non-null	object
7	model	421603 non-null	object
8	condition	252776 non-null	object
9	cylinders	249202 non-null	object
10	fuel	423867 non-null	object
11	odometer	422480 non-null	float64
12	title_status	418638 non-null	object
13	transmission	424324 non-null	object
14	VIN	265838 non-null	object
15	drive	296313 non-null	object
16	size	120519 non-null	object
17	type	334022 non-null	object
18	paint_color	296677 non-null	object
19	image_url	426812 non-null	object
20	description	426810 non-null	object
21	county	0 non-null	float64
22	state	426880 non-null	object
23	lat	420331 non-null	float64
24	long	420331 non-null	float64
25	posting_date	426812 non-null	object

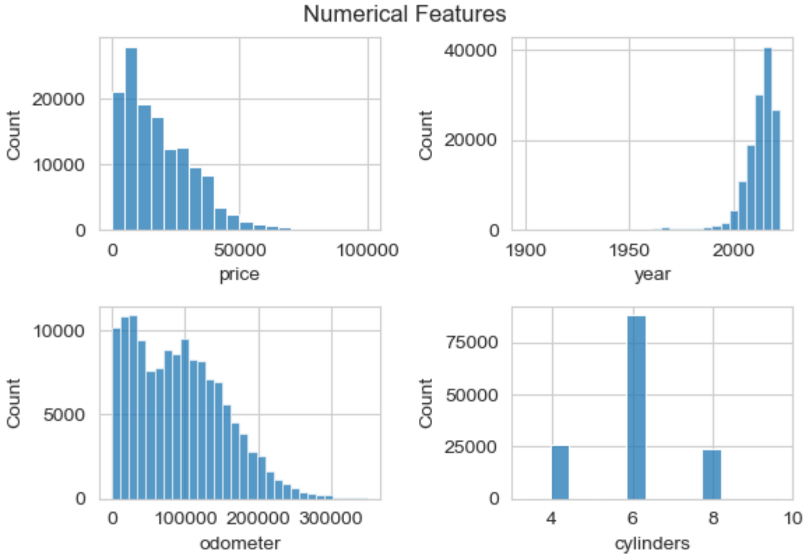
Note the highlighted rows in the table were eventually dropped after considerable data exploration but initially this was not the plan. The following figure is a plot of the missing data which shows the dataset includes a considerable amount of null data.



Initially, there was a motivation to determine price variance with respect to region. Note that the Craigslist *‘region’* and *‘county’* are not necessarily the same. A custom function was created using an API of the complete US (county-city-state dataset) to fill in the missing ‘county’ column. In hindsight, this was an unnecessary step, and the approach was abandoned.

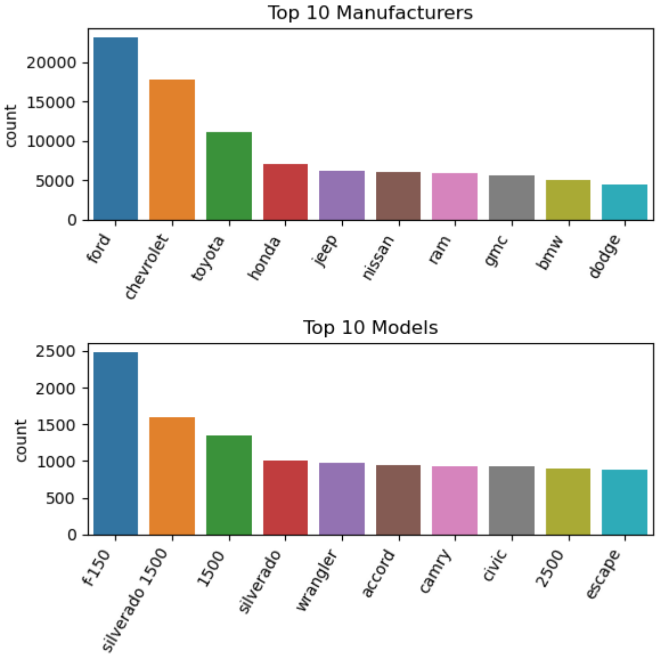
EDA:

There are 4 major numerical data. The ‘price’ is the target, ‘year’, ‘odometer’, and ‘cylinders’ are numerical data. Originally ‘cylinders’ were of object type (example: “4 cylinders”) and had to be text processed and converted to numerical. The following is the histograms of the numerical features of the dataset.

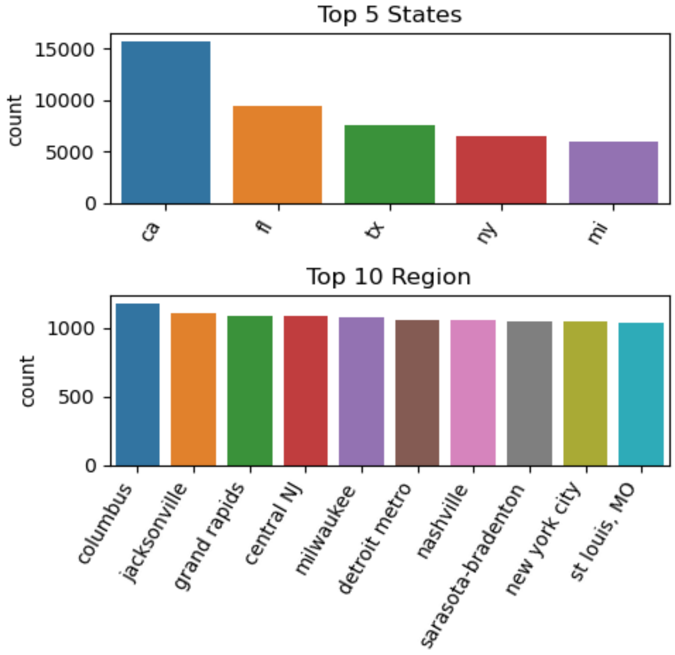


The numerical data shows that for an average vehicle: the price below $50,000, was made in the 2000’s, the mileage is around 100k, and the engine is most likely 6-cylinders. There is a considerable amount of information to parse from the ‘description’ feature however for brevity this was not pursued. For the ‘price’, all rows associated with null values were dropped. The considerable number of na values were replaced with the median of each feature.

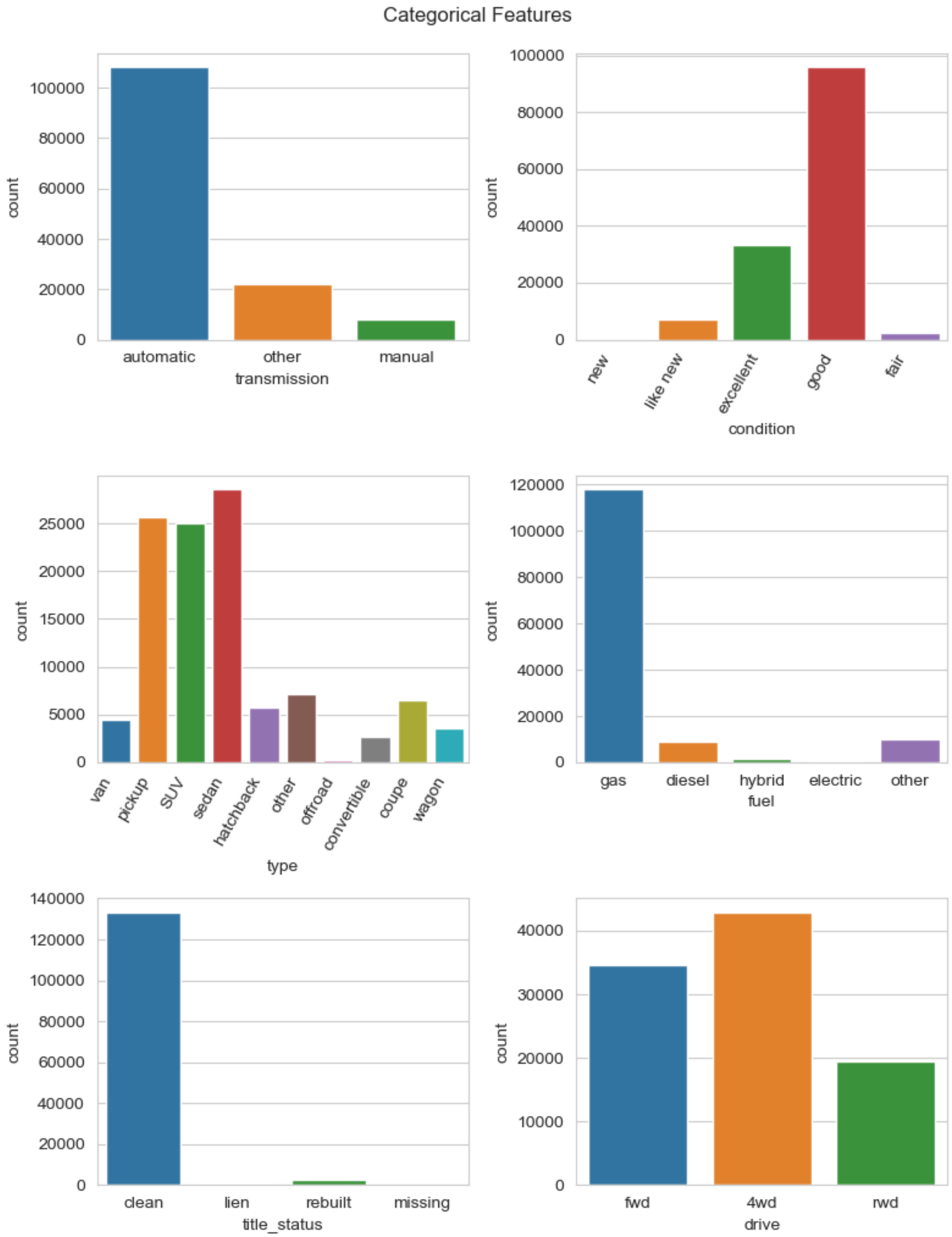
Inspecting the model and manufacturers, the data shows that the dataset is dominated by pickup trucks primarily by Ford F-150 and Chevy Silverado. The following plot is a list of the top 10 manufacturers and models in the dataset.



The following are a plot of the top regions and states which the dataset is composed of. Intuitively the largest states by population have the largest number of listings. Note that regions are not necessarily the proper city area.



The other categorical features of importance are plotted in a seaborn count plot below. The ‘type’ feature has many categories, several were combined. Note that Seaborn does not quantify *na* values and the data still required imputing.

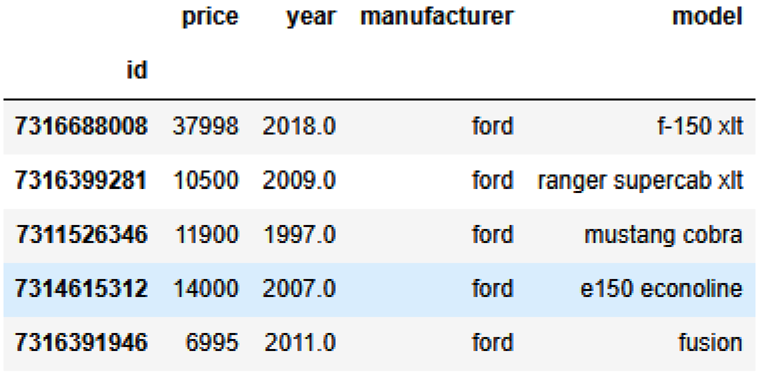


Feature Engineering:

The previous sections mentioned that the dataset has significant missing and null data. All the null values in the target feature ‘price’, ‘year’, model, and manufacturer were removed in order to perform modeling. The odometer null values and *na* values were estimated with a custom function, taking the ‘year’ and ‘condition’ of the vehicle and calculating the median of all vehicles for that specific model. Any ‘na’ values in the ‘model’ and ‘manufacturer’ were automatically dropped since parsing the ‘description’ feature is not pursued.

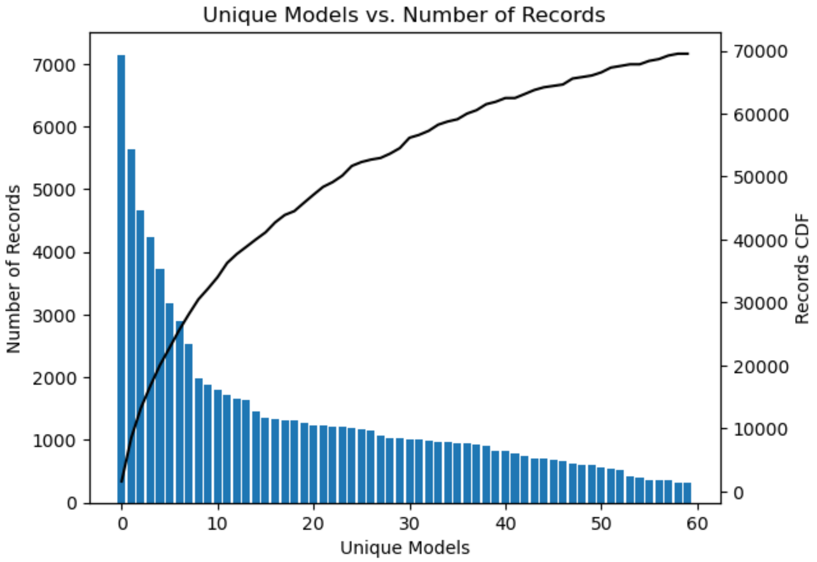
For the ‘type’ feature, trucks and pickup trucks were combined, ‘van’, and ‘minivan’ were combined, and ‘coupe’, ‘convertible’, and ‘wagon’ were combined as ‘other’. The records associated with the ‘bus’ category were removed.

Majority of the feature engineering effort dealt with wrangling the ‘model’ feature. With the null records removed, there are still 13,707 unique models in the dataset. Instead of choosing from a drop down prescribed menu, Craigslist users are able to input the manufacturer and model manually. This leads to numerous incorrect inputs and omissions. For the model section this results in thousands of model categories. For example, see below the table of Ford models with a f-150 and a f-150 xlt.

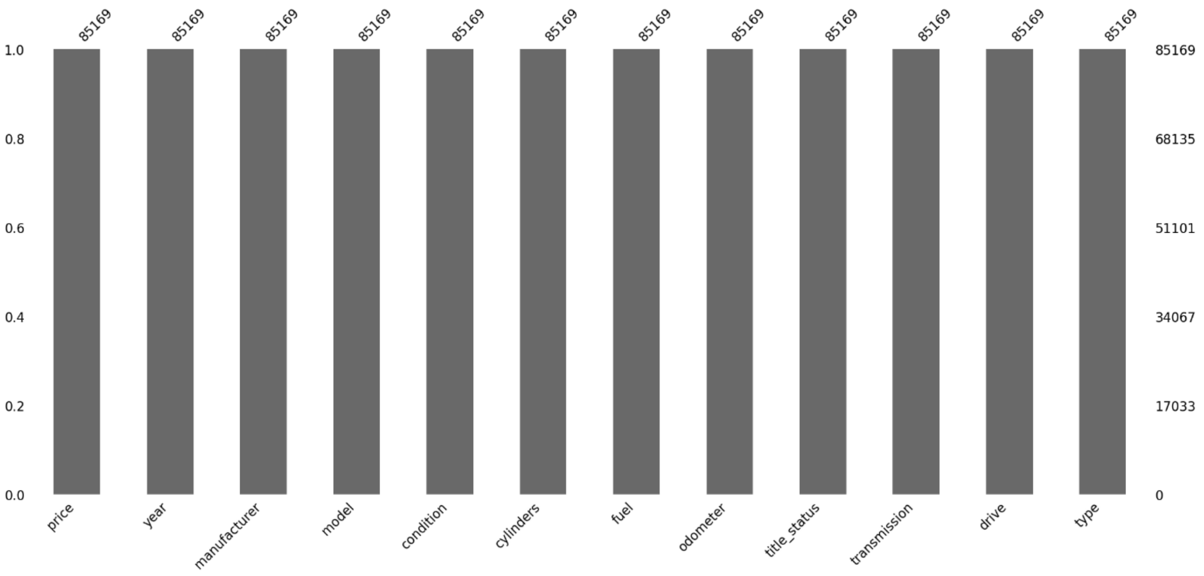


	price	year	manufacturer	model
id				
7316688008	37998	2018.0	ford	f-150 xlt
7316399281	10500	2009.0	ford	ranger supercab xlt
7311526346	11900	1997.0	ford	mustang cobra
7314615312	14000	2007.0	ford	e150 econoline
7316391946	6995	2011.0	ford	fusion

Transforming the dataset with OHE on this current field would expand the dataset to an unmanageable size. The ‘model’ feature was consolidated with the most common car models using regular expressions. For example, records with models: ‘F-150’, ‘150’, and ‘ford f-150 xlt’ were all standardized to ‘f-150’. Even after consolidating to top 60 model names there was still nearly 5.5k unique models.



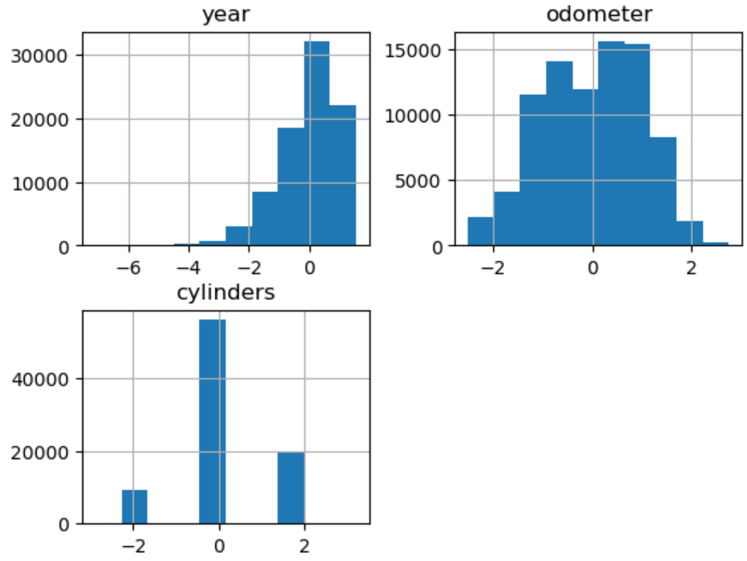
The figure above plots the top 60 models and shows the number of records associated with each model. The top model, Ford F-150, has over 7k records. There is an exponential decline with the top models and after 10 models, the decline is constant i.e. the majority of the dataset comes from the top models. For brevity, the top 60 consolidated models were kept in the dataset for the modeling step and the rest discarded. The finalized clean dataset has 85k records (reduced from over 440k+ records) with 12 fields and is devoid of *na* values.



Modeling:

To model the dataset, the categorical fields in the datasets must be transformed into numerical via one hot encoding (OHE). There are 19 manufacturers and 60 car models in the dataset after the feature engineering processing. The OHE transforms the dataset to have 99 purely numerical fields.

The dataset is split using '*train_test_split*' function from *'sklearn'* into 80% train and 20% test. This split ratio was picked based on experience and was not modified to determine its effect. A linear regression model resulted in a R^2 of 0.701 and mean absolute percent error (MAPE) of 2.99%. The numerical features 'year', 'odometer', and 'cylinders' are of various magnitudes. The data was scaled using '*PowerTransformer'* in sklearn's library using the *'yeo-johnson'* method. The results of the scaling are shown below.



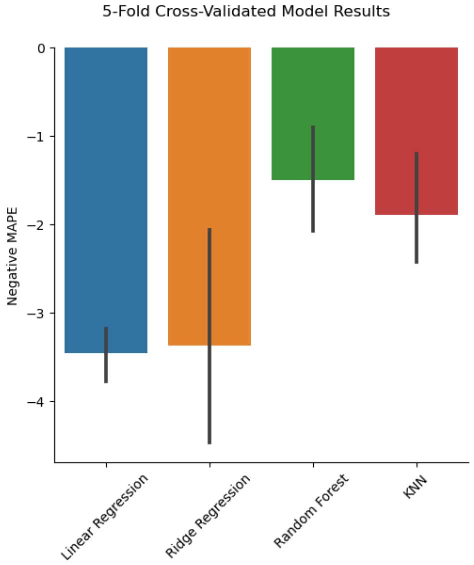
However, running Ridge regression with the scaled data did not improve the results significantly as expected. Modifying the alpha had no appreciable difference in the prediction rate either. The ridge regression yielded a R^2 of 0.711 and a MAPE of 3.25%.

Next a random forest regression analysis via '*RandomForestRegressor'* was performed with only setting the *n_jobs =-1* for parallel processing. The RF model yielded a R^2 of 0.949 and mean absolute percent error of 0.43%. The last model attempted was k-Nearest Neighbor using '*KneighborsRegressor*'. Initially the k was set to 3 neighbors, but a 'GridSearchCV' showed that k=1 was optimum for the dataset. The KNN single result had a R^2 of 0.764 and and mean absolute percent error of 0.66%.

Cross validation prevents overfitting the model and ensures the model is generalizable to unseen data. A five-fold cross validation was performed on all 4 models discussed above with the train dataset. With the train

Results

The following plot is the 5-fold cross validation results with negative MAPE as the scoring. The results of the modeling show that the random forest regression model performs the best to predict the price with an average MAPE of 1.5%. The runner up was the KNN regression model with an average MAPE score of 1.88% error on the price.



One caveat of the model is the final negotiated price of the vehicle is unknown. On Craigslist, the price is set by the user, however the purchaser has an option to negotiate price. The dataset does not include the final sale price and therefore, the price prediction has unknown errors from this fact.

Post modeling, there is no specific recommendation to provide for this dataset. This is a complex dataset and given the scope of the project, a lot of data needed to be omitted to efficiently model the data. Overall, the top 60 models from the dataset are modeled with 4 different regression models with the best predicting price with only a 1.5% error.