Predicting Used Vehicle Price

Anu

Abstract

The project used a dataset about used vehicles from secondary market in USA. It was aimed at predicting the price of a used vehicle. The work attempted to identify a model with most accuracy by comparing results from popular regression algorithms. Metrics such as r^2 score and mean absolute error were adopted to rank the models. It was inferred that Random Forest regressor gave better results than other techniques. Analysis also showed that vehicle year and odometer reading were the most influencing features.

Motivation

Pricing of used vehicles is not straightforward. It depends on quite a few factors and moreover, doesn't follow any standard depreciation from original price. However, it is important to estimate the price when a vehicle is transferred in secondary market.

By analysing past vehicle quotes, it might be possible to develop a learning model that can predict the price with reasonable accuracy. Such a model if developed would be helpful for vehicle buyers, sellers and dealers to estimate the fair market price of a used vehicle. It would also give a chance for a new vehicle buyer to trade-in the existing old vehicle to the dealer at a fair price. Price of used vehicle is also important for insurance providers so they know value of vehicle being covered.

To summarize, this project is aimed at helping stakeholders of used vehicle market by coming up with a model capable of predicing the price.

Dataset(s)

Used Vehicles in USA

- This research uses a dataset dealing with used vehicles sold and bought in USA
- The dataset has been sourced from kaggle.com
- It has more than 400,000 records and each record has around 25 fields
- Some of the important fields are Manufacturer, Model, Year, Odometer, Condition, Transmission, Location (State) and the Price
- Primary file in the dataset: `Used_Vehicles_USA.csv`

Data Preparation and Cleaning [1 of 3]

Problems identified while reviewing the dataset:

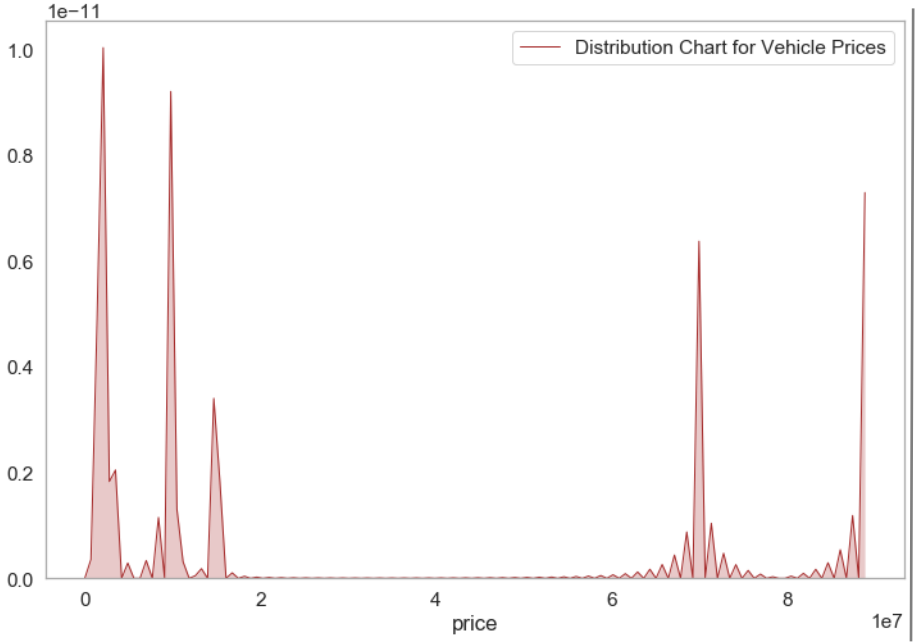
- Missing values (blank or NaNs)
- Invalid data such as zero price
- Outliers in fields like price and year
- Irrelevant fields such as url

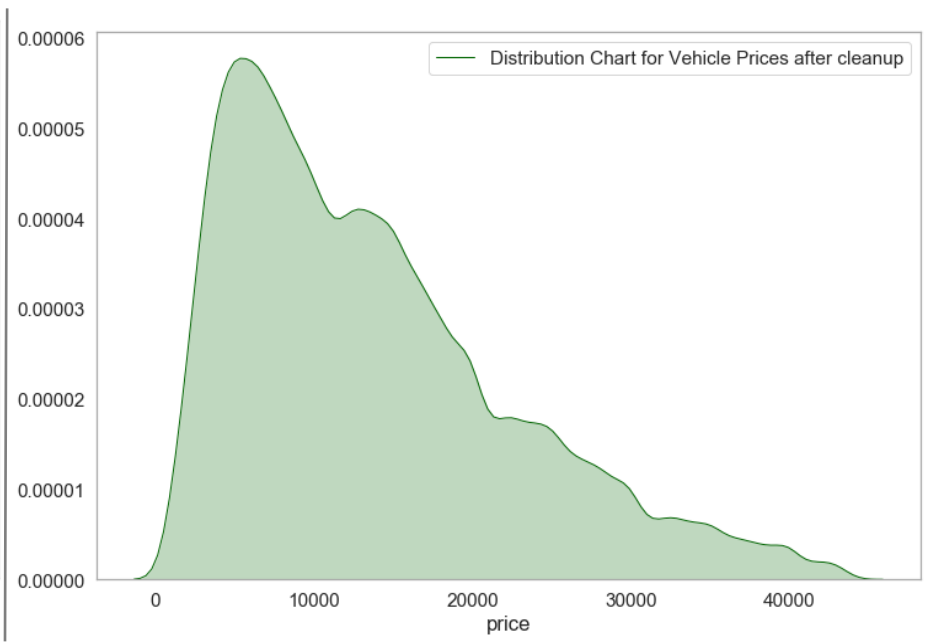
In order to rectify these problems and make the dataset usable for learning, following actions were performed:

- Dropped irrelevant fields, especially those that have significant missing values
- Drop zero prices
- Analyzed the distribution of fields to identify outliers and cut them out

Data Preparation and Cleaning [2 of 3]

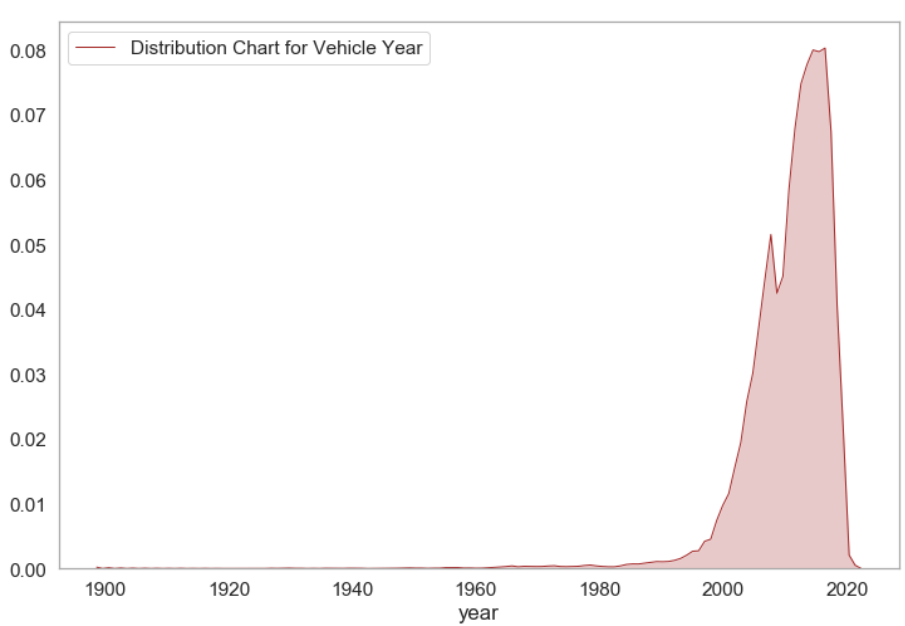
Price: Following charts show distribution before and after cleaning. It can be noted that the outliers (<2%) are removed which makes the distribution meaningful to learn.

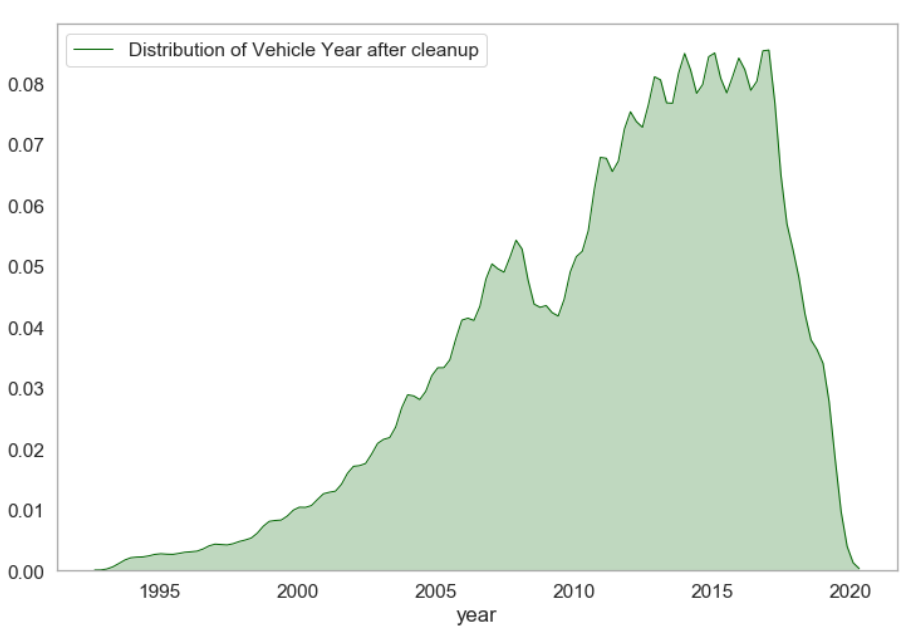




Data Preparation and Cleaning [3 of 3]

Year: The charts show distribution of year before and after cleaning. It can be noted here too that after the cleanup, distribution is less biased.





Research Question(s)

- This project focuses on prediction of price of a used vehicle in USA
- Given a set of vehicle features, the model that is built will predict the price
- More precisely, following questions are expected to be answered by this work:

1. Can we predict market price of a used vehicle using past buy/sell data?
2. Which learning model among popular ones would predict with most accuracy?
3. Which vehicle feature has most influence on price in used vehicle market?

The research would be supported by useful visualization and conclusion.

Methods

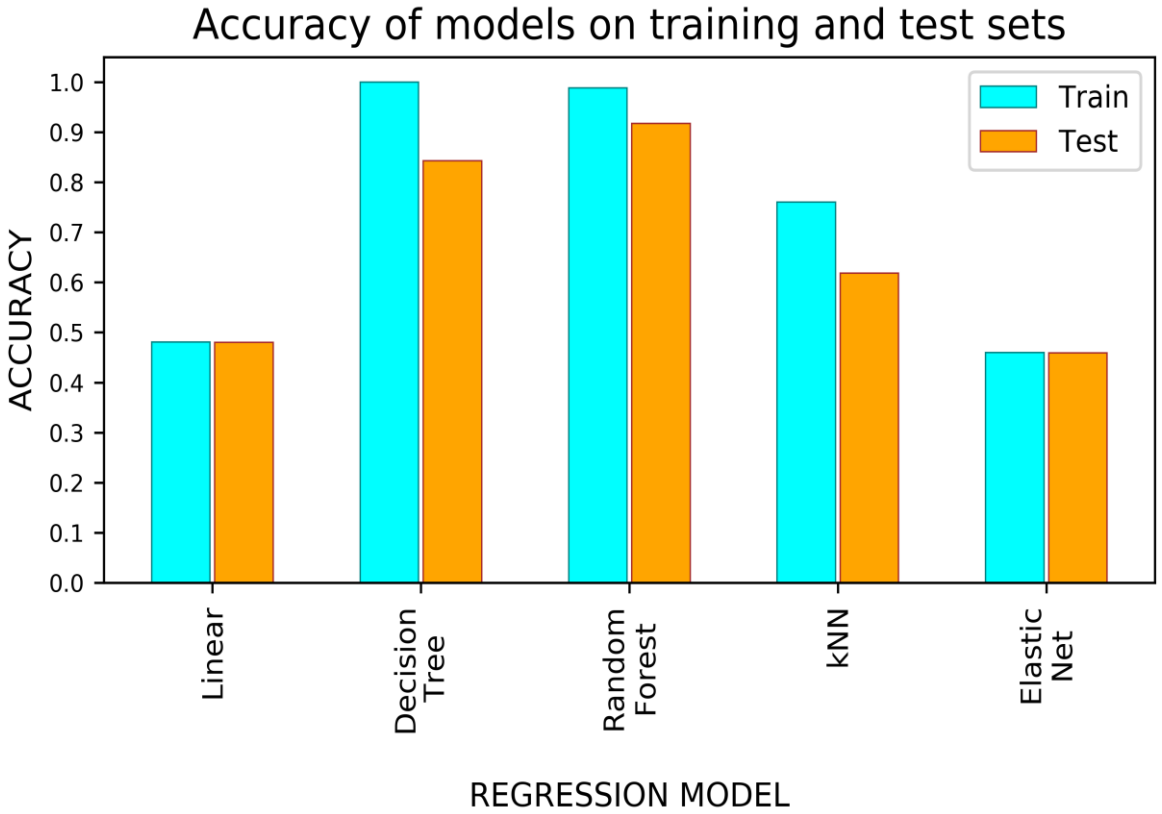
- The dataset contains the target variable price and those factors that would influence it. Hence this research would employ supervised learning methods.
- As the target variable is continuous, regression methods were chosen to come up with learning models.
- Regression algorithms used and compared are Linear Regression, Decision Tree, Elastic-Net, Random Forest and k-Nearest Neighbours.
- Metrics used to compare the results would be accuracy, mean absolute error and root mean squared error.
- The correlation between price and the influencing factors (features) has been visualized using a bar chart.

Findings [1 of 4]

- All the regression algorithms used in this research were able to predict price but with varying accuracy levels and error values
- The linear models gave the least accuracy and turned out to be most erroneous, which revealed that the relationship between price and features may not be linear
- Though couple of models were able to predict the price with reasonable accuracy, it may not be still within acceptable limits from a specific buyer or seller
- Correlation analysis between price and the features came with a large range, indicating that we can differentiate most influencing features from the least
- Specific findings are visualized and elaborated under slides to follow

Findings [2 of 4]

- The bar chart compares the accuracy of models on both training and test sets
- Random Forest regressor has the highest accuracy on test set (r^2 score)
- Decision Tree has most difference in accuracy between training and test – it could reveal overfitting
- Linear models seem unusable as prediction is quite weak

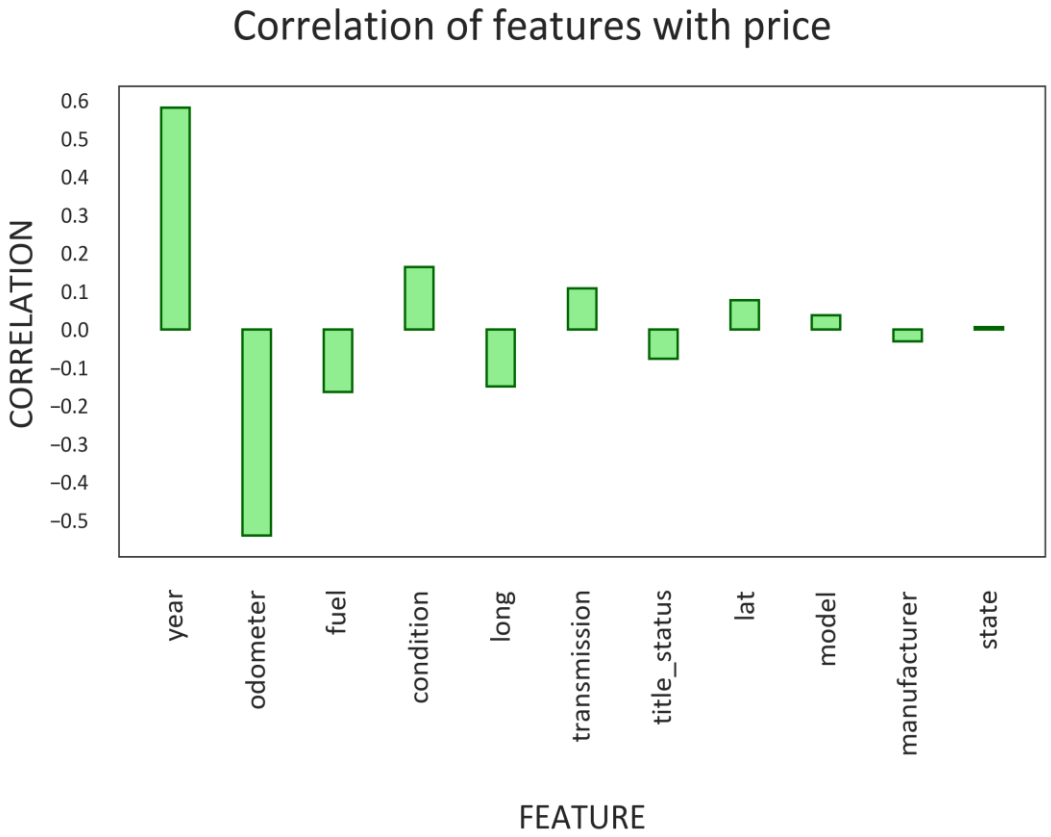


Findings [3 of 4]

- This bar chart compares the error metrics among models on test set
- Random Forest regressor continues to top the list with least errors in prediction
- Moreover, it shows less difference between MAE and RMSE hence low variance in deviation from expected
- Again, linear models seem to have large error hence may not be usable

Findings [4 of 4]

- This chart illustrates how features are correlated with price
- With the dataset used, vehicle **year** seems to have most influence over price, followed by **odometer**
- Year has positive correlation with price and Odometer has a negative correlation



Limitations

- Since the dataset is based on used vehicle market in USA, the models developed may not be fit enough to predict in other regions.
- Data doesn’t provide visibility to the actual transaction being made between buyer and seller, hence price field reflects the asking price.
- The price agreed between two private parties is typically different than when a dealer is involved for the same vehicle. The dataset doesn’t specify the category.

Conclusions

1. *Can we predict market price of a used vehicle using past buy/sell data?*

Prediction with reasonable accuracy seems possible from the models adopted but that accuracy level may still be unacceptable from the eyes of specific buyer or seller.

2. *Which learning model among popular ones would predict with most accuracy?*

In terms of both accuracy and root mean squared error, **Random Forest** regressor was found to top the list in this experiment with the dataset. It was followed by Decision Tree regressor.

3. *Which vehicle feature has most influence on price in used vehicle market?*

Vehicle **year** showed maximum correlation with price among other features followed by odometer reading.

Future work

- As the data was collected over a time period and across locations, it is quite possible that the price is influenced by the demand and supply of used vehicles. I think demand/supply is an important influencer for price, those details can be acquired and used in future research

Acknowledgements

- The data for this research is obtained from kaggle.com

- No feedback was received on the work

References

- https://www.kaggle.com/datasets
- https://pandas.pydata.org/pandas-docs/stable/reference/index.html
- https://scikit-learn.org/stable/supervised_learning.html
- https://matplotlib.org/contents.html
- https://seaborn.pydata.org/api.html