2021

Satish Agrawal, SANJAY JARAS

BELlevue university

**Abstract**

On an average, there are 1.88 vehicles per USA household. Owning a car is becoming more and more common worldwide. The current market size of used car vehicles is around $89 billion. The buyer confidence will likely increase over the next five years as the economy recovers from the coronavirus pandemic. It is projected that the revenue for used car dealers in the USA will amount to approximately $123.3 billion by the year 2024[1]. In the US alone, there are 123,905 businesses involved in the used car market. Due to the increased price of new cars and customers with a lack of funds, used car sales are on a global increase. Developing countries adopted the lease culture instead of owning a new car due to affordability. Therefore, the rise in used car sales is increasing exponentially. There is abundance of used car sales data out and available. This project aims at using the multiple attributes of the used car sales and train a model that can predict a good price for a used car. Using machine learning algorithms like linear regression, K-Neighbors Regressor, Random Forest Regressor, etc., we aim to build a model that will help in reliably predicting used car price.

**Intro/background of the problem**

The price of the new vehicle is determined by the manufacturer. The manufacturer considers variety of factors including government taxes, used raw materials cost, labors involved, intended profit margin per unit and many others to come up with the Manufacturer Suggested Retail Price (MSRP). So, buyers of new cars are a bit more confident about the price of new cars; which is not always true with the price of an old car. The used car buying is a very complex process, as an average buyer might not think of all the variables affecting or involved in the price. Car sellers seldom take advantage of such a scenario by listing unfair prices owing to the demand. On the seller's part also it's quite hard to estimate the used car price manually. Generally, experienced sellers can think of some parameters like mileage on the vehicle, condition of the car, fuel type, and vehicle age, etc. However, for experienced sellers also, it is hard to consider all parameters while estimating used vehicle price. So there is a necessity for a used car price prediction system to reliably determine the fair price of the car using various vehicle parameters. There are existing models in the market that estimates the used vehicle price; we are not sure about their accuracy and biases.

This project aims to train models for the data set chosen. The data set is one from Craigslist used car listings. We are expecting it to be generalized to the rest of the world and other listing portals. The data will be collected from data sources mentioned in a data source section. Later after data cleaning and feature extraction, we will train different models using this cleaned dataset. We will compare all trained machine learning models for accuracy while training, cross-validations, and final testing. We will pick a model that is more efficient and accurate for deployment.

**Methods**

We chose a data set from Kaggle @ <https://www.kaggle.com/austinreese/craigslist-carstrucks-data>. This data set has more than 450k records and 25 attributes for each record. We are training a supervised learning model, price being the target variable. We have split our data into training set and test to avoid test data getting used in imputation and training. We are dropping few columns from the get-go thinking those may not correlate as much (or at all) to contribute to the price of the used car. We dropped unique listing id, image URL, listing URL, region URL, VIN and description. It is apparent that these columns do not have direct correlation with the target variable. As part of data cleansing process, we looked at the distribution of the listing year column. We dropped extreme values like anything less than 1995 and more than 2020. We also removed the outliers from the odometer readings and the final price. As part of feature engineering we have added a new column for the age of the car at the time of listing. The value for the age in month derived from the model year and the date of listing. For the model year month we considered September of model year minus one as most of the car manufacturers launch their next year car models in August/September of current year. We found that distribution of the age for the data set is right skewed distribution with single peak.

We also noticed significant amount of missing values in few of the columns. We implemented iterative imputation using many of the estimators to reduce the mean square error. We chose Bayesian ridge, decision tree regressor, extra trees regressor, K neighbor regressor, and lasso regression and ran through the data set for columns with missing values using the negative mean square error being the scoring metric. It came out to be the Bayesian ridge the best imputer for the data set.

There are many categorical attributes in the data set which are key to our modeling project for price prediction, for example, manufacturer, model, cylinders (number of cylinders), title status, transmission (auto versus manual), fuel (type of fuel used) and few others. We used one hot encoding to convert categorical columns into numerical columns to use with model training. We are also planning to use label encoding approach and compare the models. We are using skikit learn pipelines for all the steps so that we can reuse all these steps for test data.

**Results**

We split the data set into train and test set using 80-20 split. We want to make sure that we have some unseen data set aside for the model the see how it performs. At this time of writing of this analysis, we have successfully trained a linear regression model using the data. We are seeing the R2 score of 87.2639%.

**Discussion/conclusion – Next Steps**

The initial results at this point of time on the project seems to have acceptable accuracy to be able to predict the price for a used car given other attributes. We have spent enough time in handling missing values and outliers. Most of the attributes are now close to a normal distribution which is helping the predictability of the model. Currently we imputed all data once however this approach seems to be wrong. So we will be splitting train and test data set before imputing so that same imputaion pipeline can be easily applied onto new test data in future. We are working towards tuning hyper-parameters of the Linear regression model and come up with a better performing model. We also intend to train a few other models and to hyper-parameter tuning and choose the best model for the problem. We are planning to use GridSearchCV and do an exhaustive training of multiple algorithms with multiple hyper-parameters.

**Acknowledgments**

The authors of this project have referred to Kaggle for the data set and some of the notebooks available in the Kaggle. There is significant number of solutions available in the Internet for the problems. We have referred to some of those available solutions and have designed our own solution. We have also referred to multiple websites including towards data science articles from various authors, geeks for geeks, machine learning website, mediam.com and many others for the basic machine learning and statistics concepts and practical examples.

**References**

* Aurélien Géron – June 2019 - Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow – O’Reilly Publication
* Panwar Abhash - Used Car Price Prediction using Machine Learning - https://towardsdatascience.com/used-car-price-prediction-using-machine-learning-e3be02d977b2
* Used car dataset - <https://www.kaggle.com/austinreese/craigslist-carstrucks-data>
* Enes Gokce - Predicting Used Car Prices with Machine Learning Techniques - https://towardsdatascience.com/predicting-used-car-prices-with-machine-learning-techniques-8a9d8313952