2021 Craigslist used electric cars price Prediction

Zheyan Liu (zl3119)

Contents

Introduction

With the soaring oil price, more and more people are considering buying an electric car to save money. We would like to build models to help predict the used electric car price so that customers can use this model to determine whether the deal is reasonable.

We used data from Craigslist, which is the world's largest collection of used vehicles for sale. The original data contains price of used car from Apr 2011 to May 2011, it contains 426880 observations and 18 variables. Since we are only interested in cars fueled by electricity, data is reduced to 1698 observations.

Research Questions

- Find appropriate way to handle missing values.
- Conduct exploratory data analysis to find interesting facts about the data.
- Build and compare machine learning models to find the best one for the prediction task.

Data preparation

We drop variables that has missing rate higher than 35%. For variables with relatively high missing rate (>2%), we analyze the missing pattern, whether they are MAR or MNAR. For MNAR categorical variables, missingness is treat as an attribute *NAN_cat*. For MAR categorical variables, missing values are imputed with mode. After selection and imputation, the final variables for model are as follow

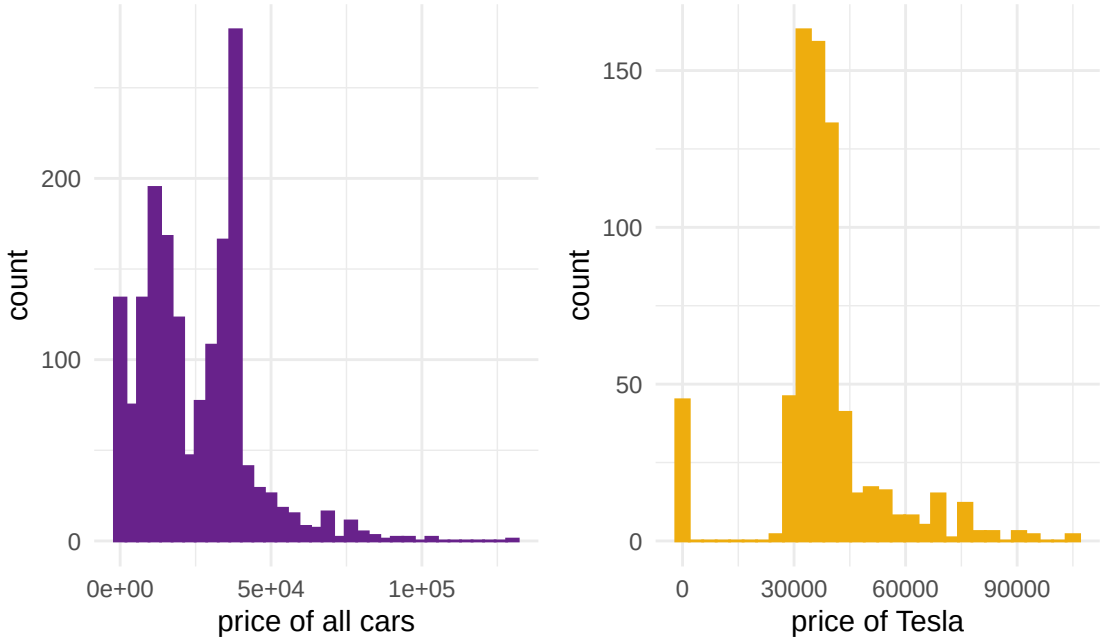
Variable	Type	levels/range	Missing rate	Missing type	Impute method
Price	continuous	0-130000	0	NA	NA
Year	continuous	1901-2022	0.2%	MAR	Median
Odometer	continuous	0-1111111	0.3%	MAR	Median
Lat	continuous	19.64-61.57	1.8%	MAR	Median
Long	continuous	-159.37–70.06	1.8%	MAR	Median
Manufacturer	category	29 levels	4.8%	MNAR	NA as attribute
Condition	category	6 levels	31.6%	MNAR	NA as attribute
Title_status	category	6 levels	1.5%	MNAR	NA as attribute
Transmission	category	3 levels	1.1%	MAR	Mode
Drive	category	3 levels	19.9%	MAR	Mode
Type	category	11 levels	10.1%	MAR	Mode
State	category	49 levels	0	NA	NA
Paint_color	category	11 levels	26.4%	MAR	Mode

Manufacturer, *Title_status* and *Condition* are considered MNAR because the records with missing values clearly has lower price compared to other category.

Exploratory data analysis

We discovered some interesting facts through visualization. Note that all the exploratory data analysis are based on the raw data without imputation.

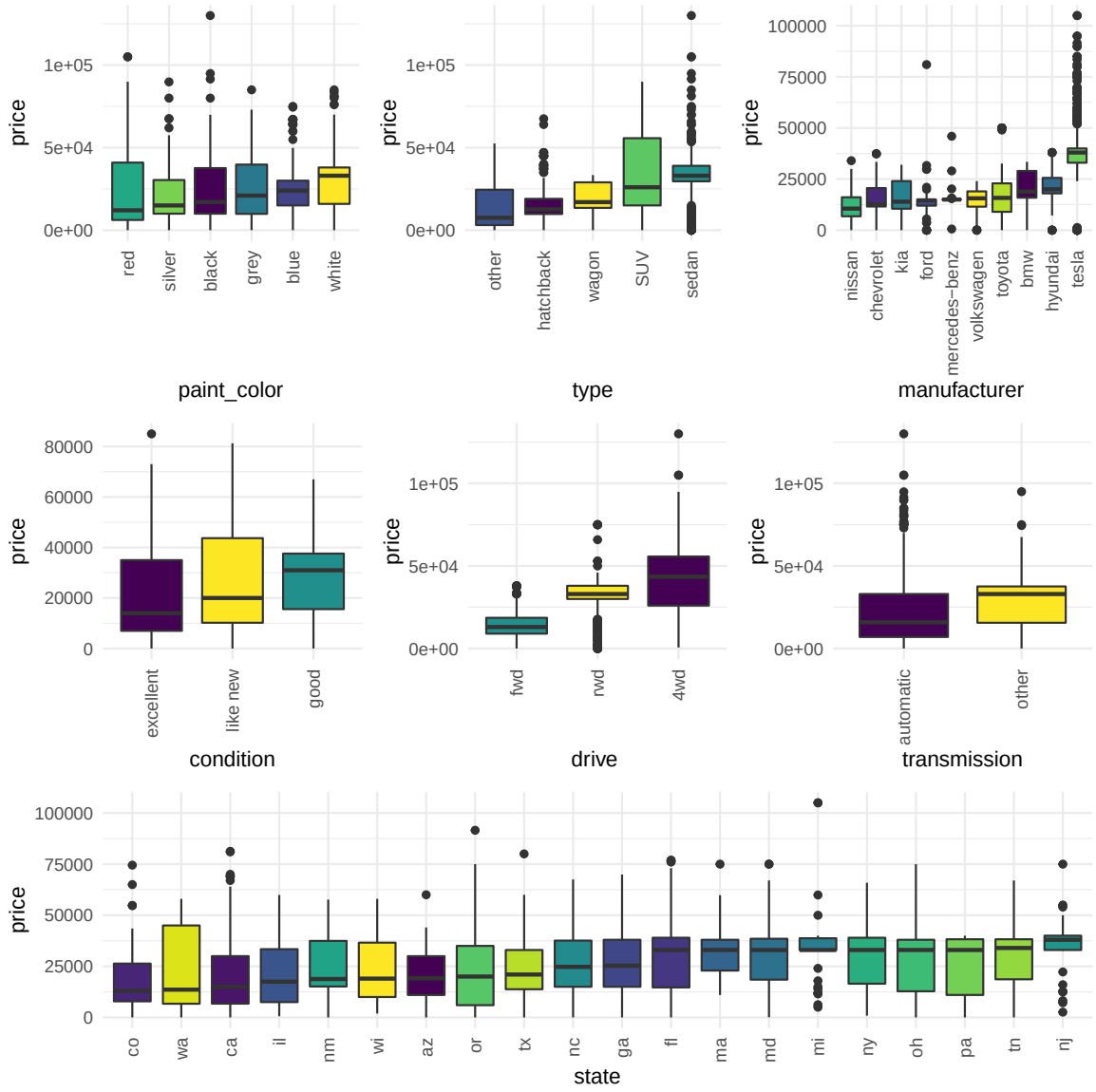
Interesting price distribution



From the first price histogram, there is clearly two vertices of the price, and the price is relatively skewed to the left. The reason is that 689 out of 1698 observations in the dataset is manufacturered by Tesla, and Tesla has a higher price than most of other brands.

In addition, there are some outliers in the dataset, it contains some prices equal or very close to 0. Therefore, we will remove the 2% records with low price in the training data.

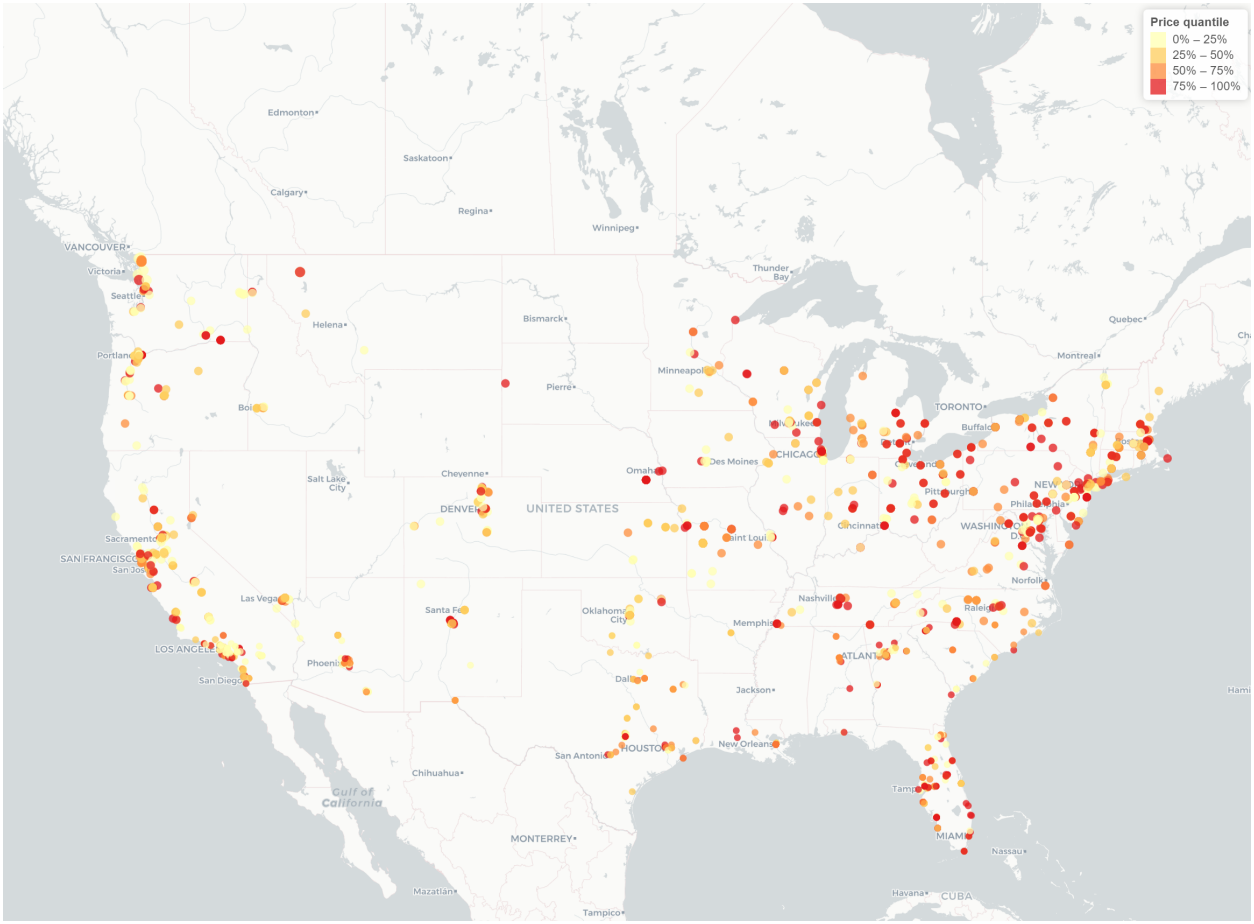
Price vs category variables



Despite some common knowledge, here is some interesting factors from these boxplot:

- White cars has the highest price while red ones has the lowest median price. However, prices of red cars are scattered.
- 4wd cars has the highest price and the reason behind this ican be car type. 123 out of 166 4wd cars are SUV and Sedan.
- Electric cars New Jersy has the highest median price while California is one of the states with the lowest median price.

Price map



Most car sales takes place near the Coast or the Great Lakes Region. In addition, the car price in the East Coast is clearly higher than that in the West Coast

Models

I used Lasso, Regression Tree and Gradient Boosting Tree to predict the price. I used Lasso because there is a considerable number of variables (239 including dummy variables) in the training data. And L1 regularization can help reduce dimension and avoid multicollinearity. In addition, I selected Regression Tree because it is easy to interpret and it captures the interaction between variables. Finally, I adopted the ensemble model GBM to better utilize the good property of tree-based models, a single tree can have high bias while boosting methods fits the residual of last round to gradually reduce bias.

Model preparation

Conduct model preparation with exact following steps

- Impute the data and divide the data into train set and test set. Test set takes up 20%.
- Using MinMaxScaler to scale all continuous variable in range [0, 1] so that the Lasso coefficients are comparable.
- Remove the 2% records with low price in the training set.