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

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 errorness in the dataset, it contains some prices equal or very close to 0. We remove the 7% lowest price records in the dataset (7% quantile on price is quantile(df$price, 0.07)).

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.

Building model and tuning parameters

Use cross validation to select the best parameter or parameter combination for each model.

Lasso

The parameter λ controls the L1 Regularization, the bigger the λ, the fewer variables in the model. Set the candidate values of λ to be from 0.1353353 to 0.0067379 with 300 steps, the best-tune λ is 148.4131591.

Regression Tree

The parameter *max tree depth* determines how many splits/leaves the tree can get. A lower *max tree depth* may result in underfitting while a higher *max tree depth* can lead to overfitting. Set the candidate values of parameter to be 1 to 10 with step of 1, the best-tune *max tree depth* is 8

Gradient Boosting Regression Tree

There are several parameters in the GBM model. *num of trees* controls the number of estimators/base-trees in the ensemble model.*interaction depth* is similar to *max tree depth* in the Regression Tree, it determines the highest level of variable interactions allowed while training the model. *shrinkage* is considered as the learning rate. It is used for reducing, the impact of each additional fitted base-tree. For *num of trees* and *interaction depth*, small value may cost underfitting and the bigger one can result in overfitting while *shrinkage* does just the opposite.

Set the range for *num of trees* to be 200 to 500 with step of 100, the range for *interaction depth* to be 2 to 7 with step of 1 and *shrinkage* to be 0.05 or 0.1. The best-tune *num of trees* is 500, *interaction depth* 7 is and *shrinkage* is 0.1

Model performance

The cross validation median RMSE LASSO(6681) > TREE(6054) » GBM(4274). The reason is that most variables do not have linear relationship with price. Additionally, Lasso model does not allow interaction between variables while Tree-based model considers that.

What is more, I plotted the scatterplot of actual and predicted value. Note that Lasso can predict the price to be negative values, which is not reasonable in practice. I use a simple function $price_{pred} = max(0, price_{pred})$ to correct that. On the test set, it is clear that these model do not fit well on extreme values (very high or very low price). Also, the prediction of regression tree looks like a step function.

Important variables

The plot shows the most important 15 variables for each model. For the Lasso model, the variable importance is the coefficients. For all three models, variable year and odometer are among the top three most important variables. From the coefficients of Lasso, the car price is negatively correlated with the odometer and positively related with manufactor year. In addition, they have some shared important variables such as if manufactored by tesla, whether the car type is SUV and whether the car is rear-Wheel drive.

Model Limitations

- Lasso model does not have good prediction performance in the test set and cross validation
- Regression tree model is explainable but the predictionm performance can be improved
- GBM model fits data well but the ensemble method is like "black boxes", we can not clearly explain how it works

Conclusions

Some conlusions from the data preparation step, analysis step and modeling step:

- I determined the missing pattern for missing variables and used mode imputation for categorical variables, median imputation for continuous variable
- Some interesting facts about data are discovered such as large number of Tesla cars causing the price histogram to have two vertices. Another fact is the interesting distribution of electric car sales
- The GBM model has the best prediction performance, it has RMSE 4274 at cross validation and RMSE 4729 at the test set while the median car price in the dataset is 2.599×10^4
- If interpretability is also considered, we should select Regression Tree model because it has acceptable prediction error with RMSE 6054 at cross validation and it has good interpretability

Appendix

- All the data, code and documents are in the github, check out the repository **here**
- The report is within three pages excluding plots and tables, check out the excluded version **here**