XGBoostings are high-quality electric cars price predictors

Zheyan Liu (zl3119)

Contents

Introduction

As the oil price rises, more and more people begin to pay attention to electric cars. First of all, electric cars can save money. No matter where you charge in the U.S., electric cars are cheaper to fuel than gasoline-powered vehicles. Then, electric cars can reduce emissions. Today, the average U.S. electric vehicle emits as much as a gasoline vehicle that gets 73 miles per gallon. As wind and solar power replace coal-fired generation, the emissions performance of electric vehicles will improve. Moreover, electric vehicles offer a better driving experience. Electric engines produce instant torque, which means electric vehicles can narrow the starting line and provide smooth, responsive acceleration and deceleration. Electric vehicles have a lower center of gravity, which improves handling, responsiveness and ride comfort. Rising energy cost would contribute to higher prices of vehicles including second-hand cars; however, electric vehicles, an alternative to traditional motor vehicles, recently plays an increasingly important role in used car market. Basing on that, more families might prefer electric cars. This paper will work on building 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 2021 to May 2021, 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

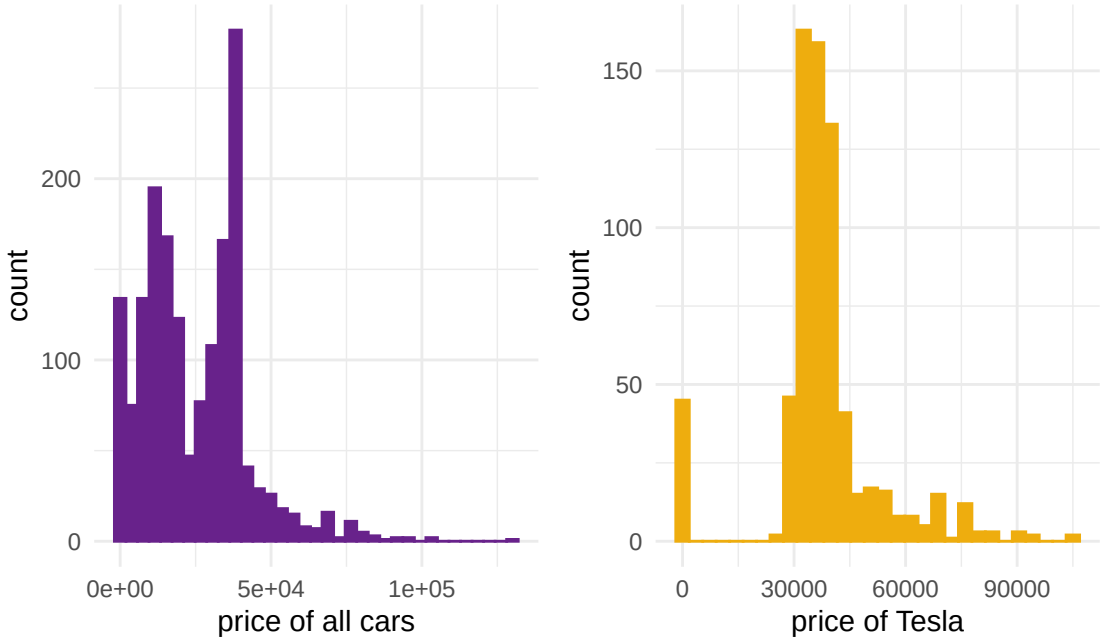
Variable	Type	levels/range	Missing rate	Missing type	Impute method
Price	continuous	0-130000	0	NA	NA
Year	continuous	1901-2022	0.2%	MAR	Median
Odometer	continuous	0-1111111	0.3%	MAR	Median
Lat	continuous	19.64-61.57	1.8%	MAR	Median
Long	continuous	-159.37–70.06	1.8%	MAR	Median
Manufacturer	category	29 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.

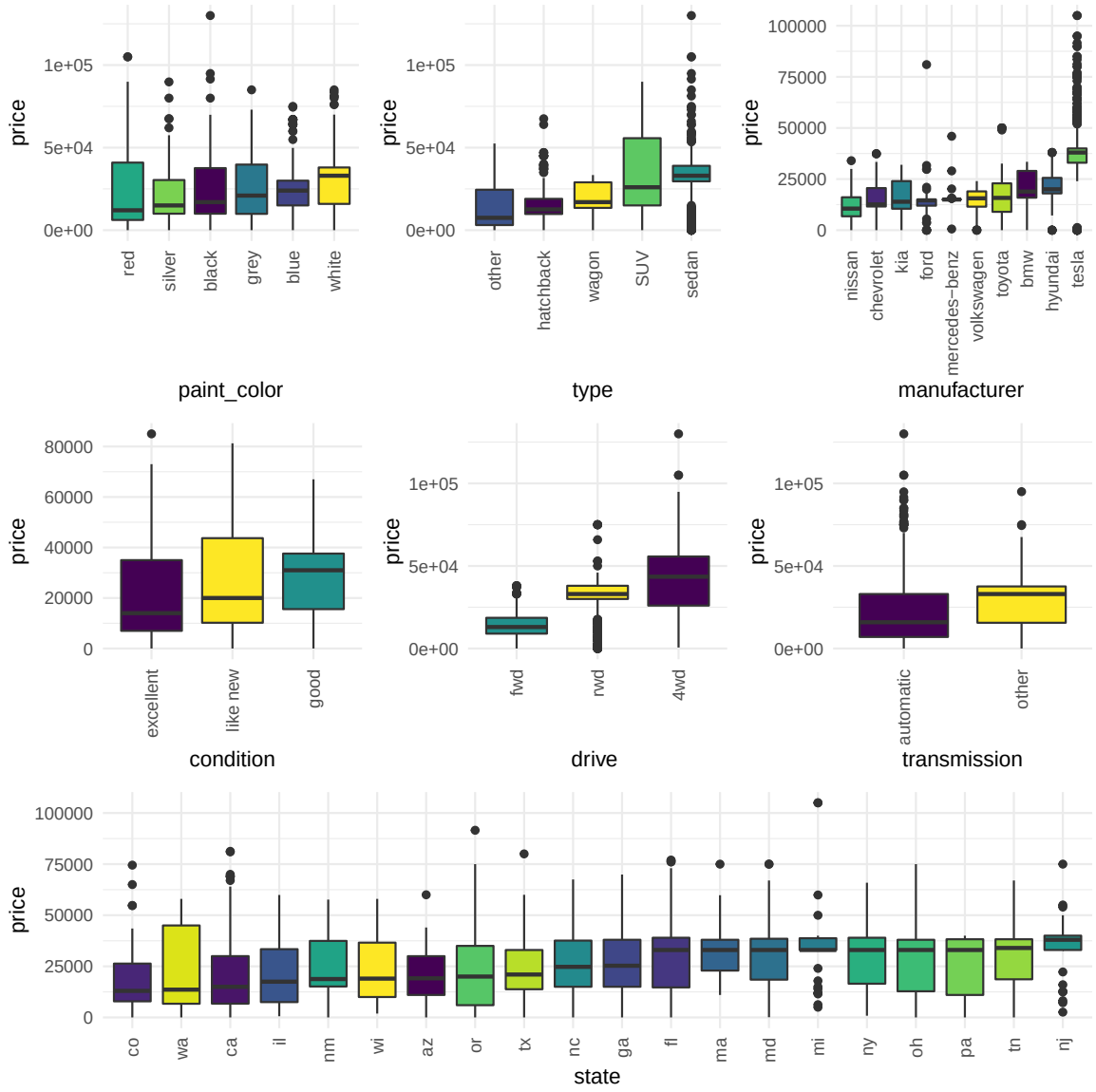
Interesting price distribution



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 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 730.83).

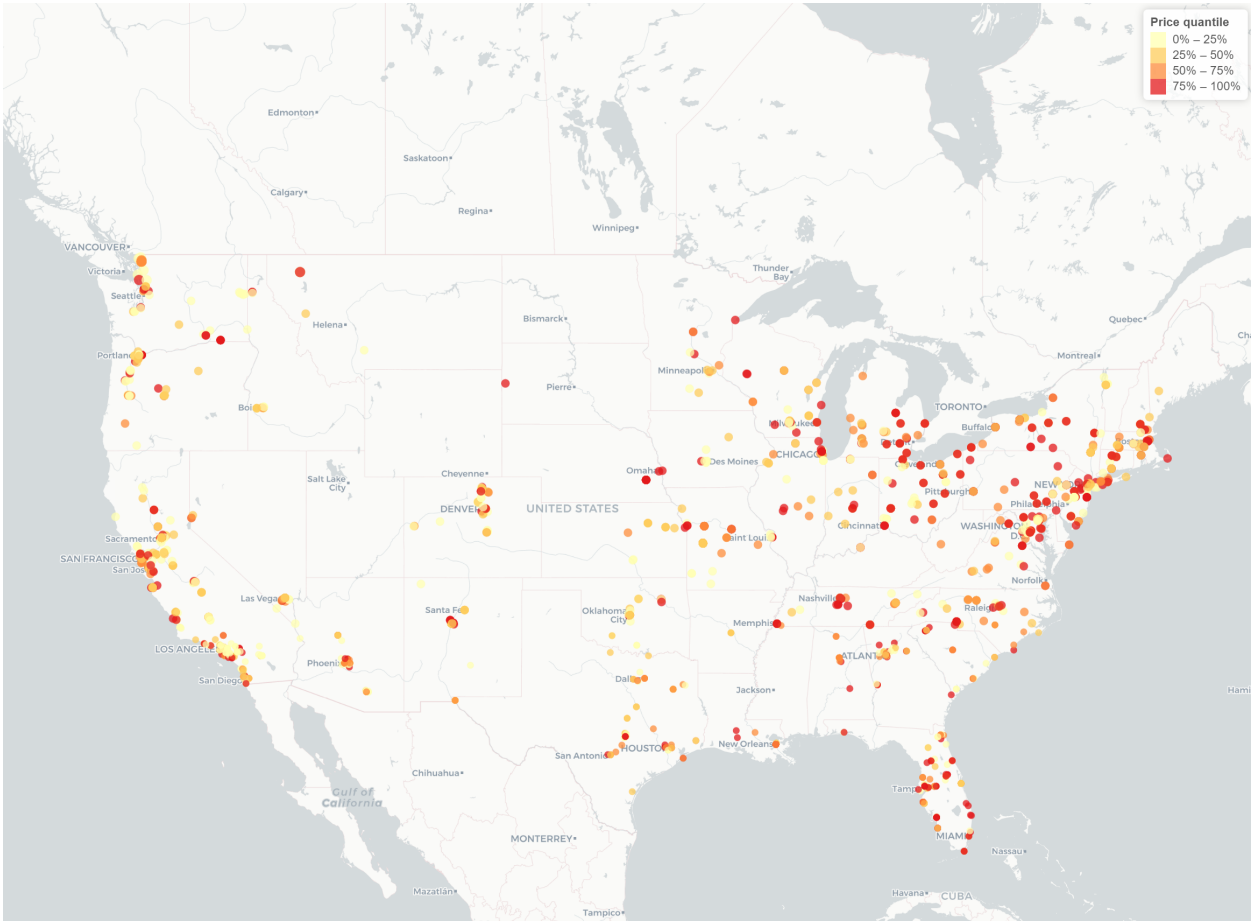
Price vs category variables



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

We used Lasso, Regression Tree and Gradient Boosting Tree to predict the price. We 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, We selected Regression Tree because it is easy to interpret and it captures the interaction between variables. Finally, We 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 403.4287935 with 300 steps, the best-tune λ is 153.9748531.

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

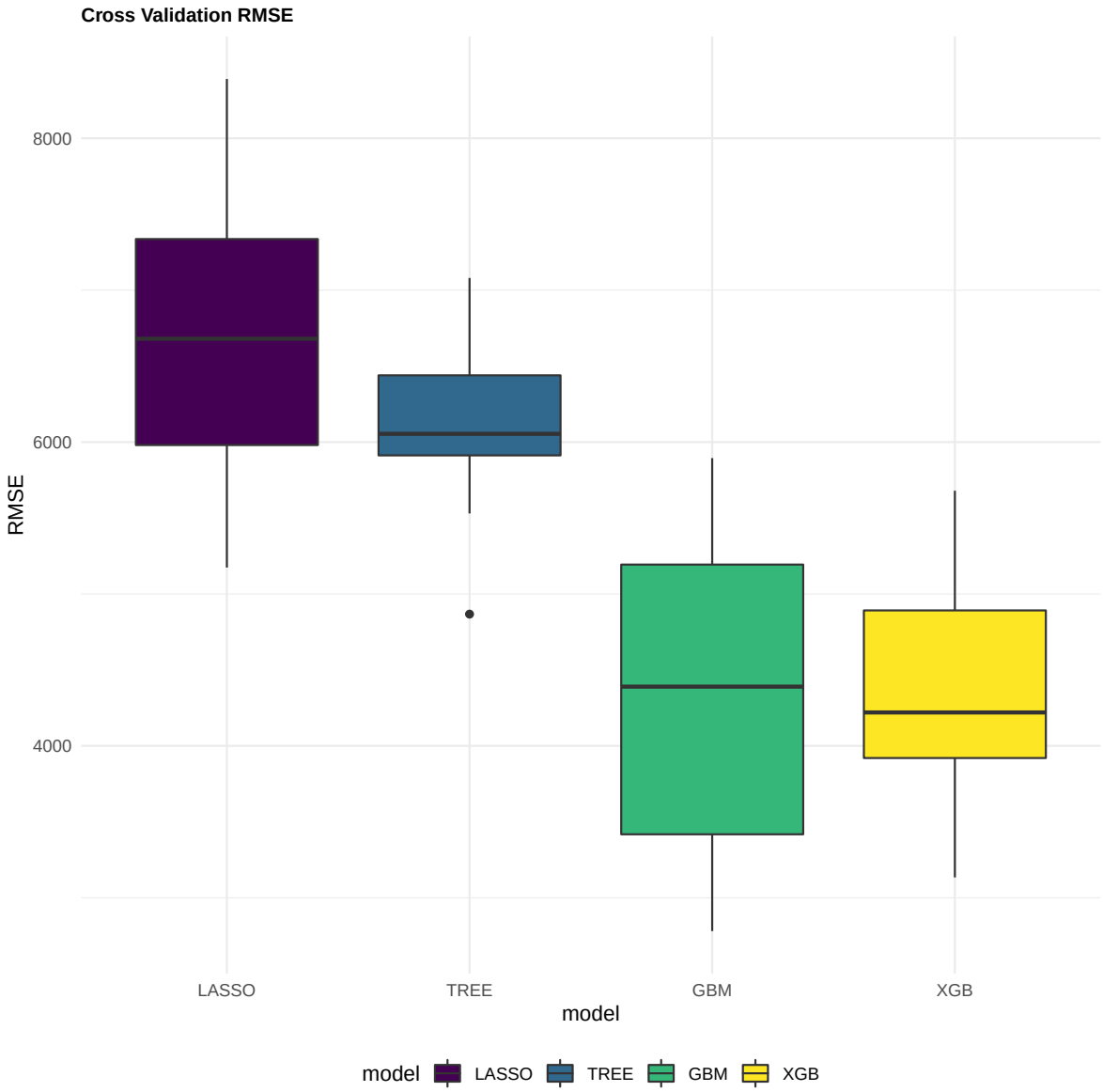
XGBoost

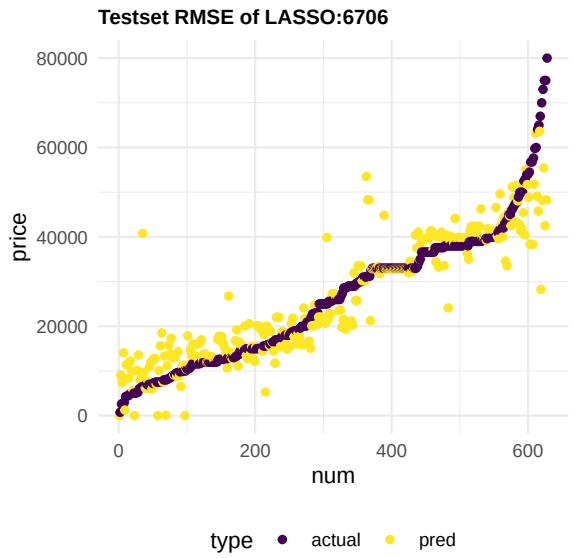
nrounds is the number of decision trees in the final model – 500, 1000, 1500 were selected as options; *eta* is analogous to learning rate in the model, makin the model more robust by shrinking weights – 0.01 and 0.05 were seleced; *max_depth* is the max depth of a tree where higher depth will allow model to learn relations very specific to a particular sample – 2, 4, 6 were selected as options; *colsample_bytree* denotes the fraction of columns to be randomly samples for each tree – the selection is from 0.5 to 0.9 with 0.08 as one step; *subsample* denotes the fraction of observations to be randomly samples for each tree – 0.5 and 1 were selected as options; *gamma* specifies the minimum loss reduction required to make a split – 0 and 50 were selected as options; *min_child_weight* is the minimum sum of instance weight needed in a child – 0 and 20 were selected as options.

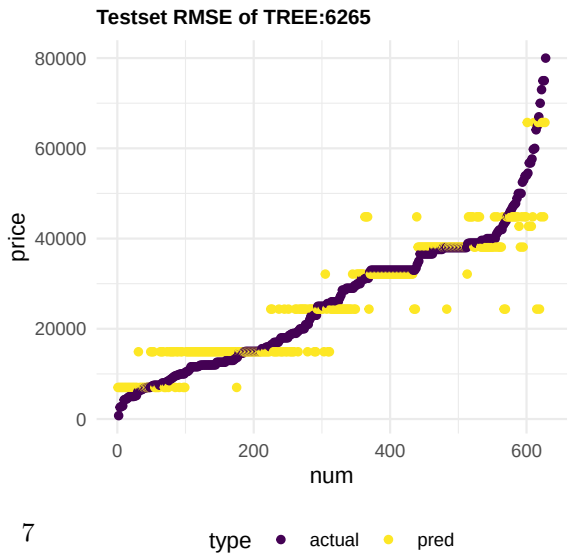
The optimal *nrounds* is 1000, optimal *max_depth* is 4, optimal *eta* is 0.05, optimal *gamma* is 0, optimal *colsample_bytree* is 0.6, optimal *min_child_weight* is 2, and the optimal *subsample* is 0.9.

RMSE for XGBoost is 4528.881043.

Model performance







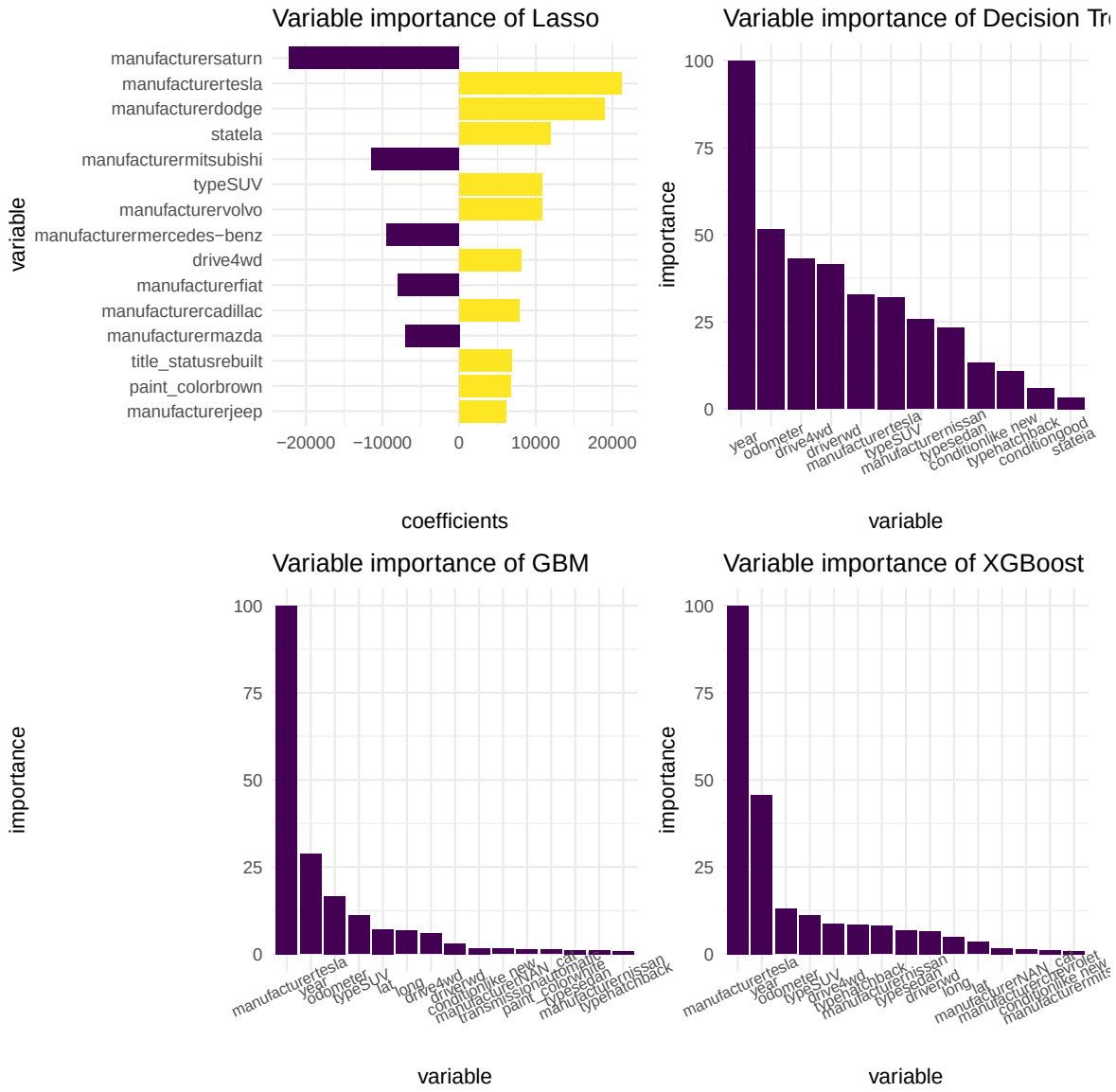
The full training performances are shown below:

TRAIN RESULTS	MAE (median)	RMSE (median)	RSQUARED (median)
LASSO	4251.711	6680.215	0.8031168
TREE	3982.900	6054.306	0.8325880
GBM	2400.654	4390.029	**0.9146599**
XGB	**2102.903**	**4219.979**	0.9136081

The cross validation median RMSE LASSO(6680) > TREE(6054) » GBM(4390) > XGB(4219.979). So, the XGB should be seleceted as final model. 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, we 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. we used 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.

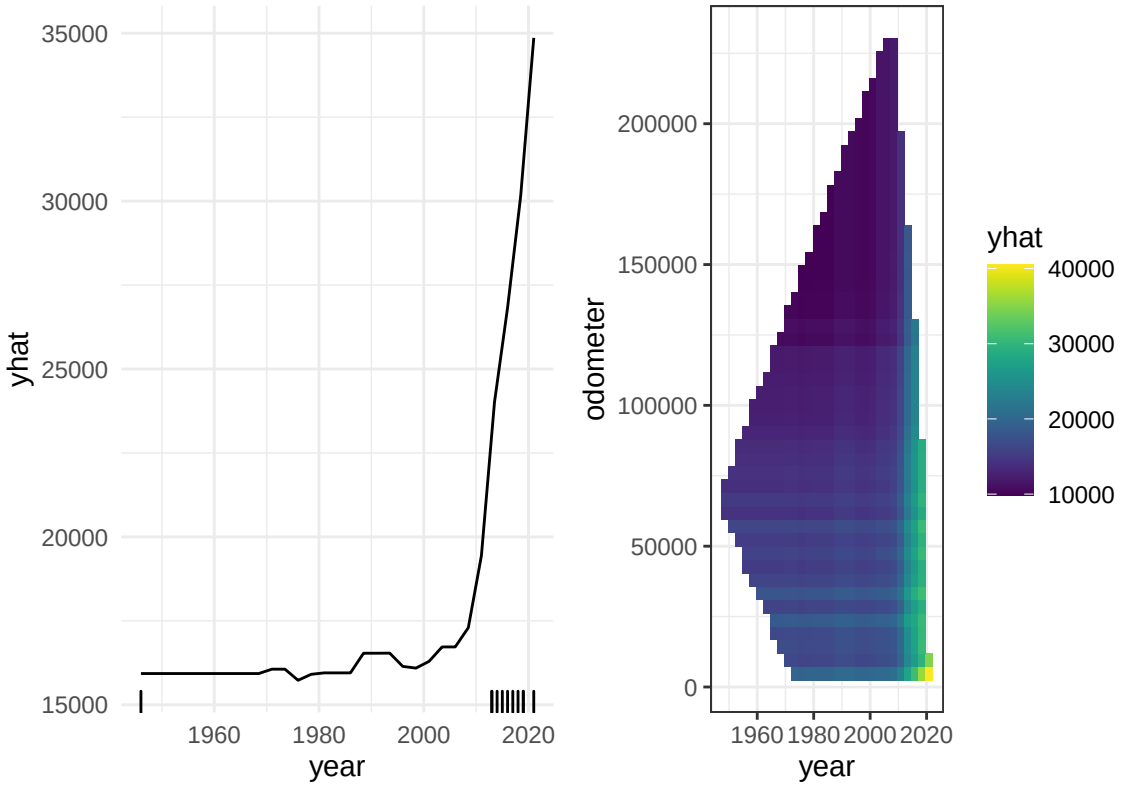
Model Limitations

- 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 and XGB models fit data well but the ensemble method is like "black boxes", we can not clearly explain how it works

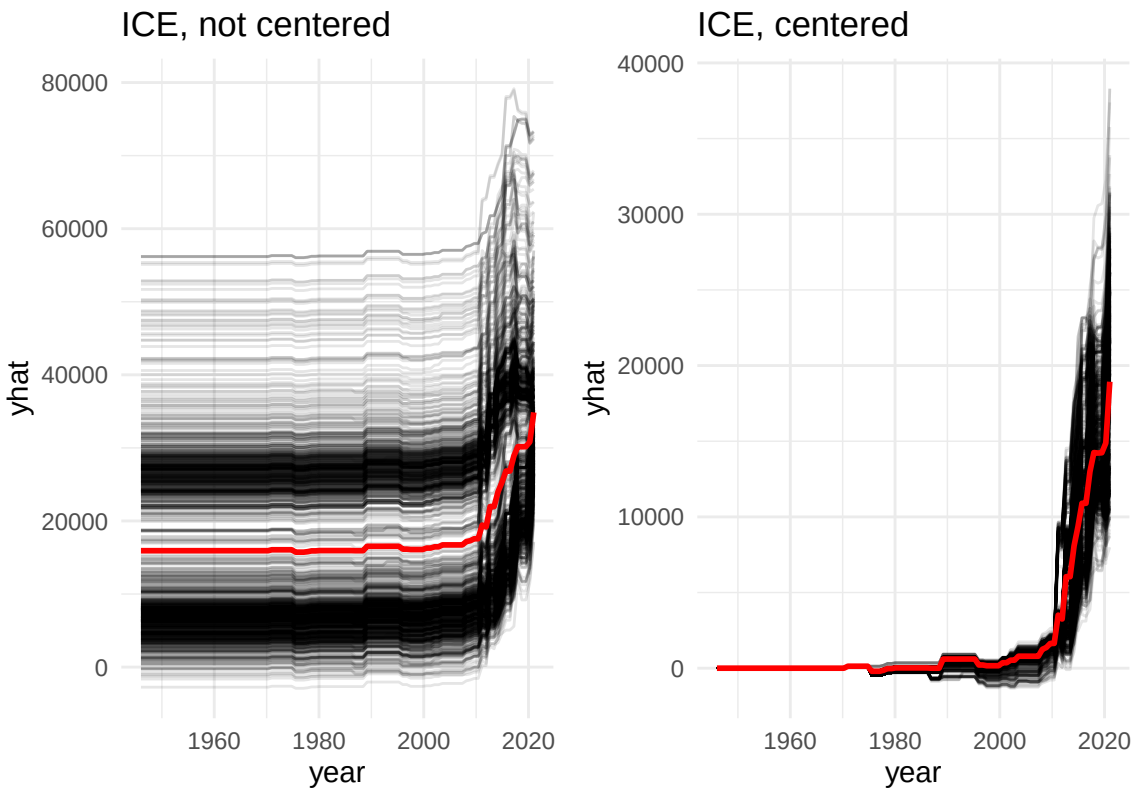
Black box of XGBoost

The best model in cross validation and test set is XGBoost. However, XGBoost is a blackbox model. Therefore, We will use partial dependence plots, individual conditional expectation curves and the lime packages to try to give prediction explanations.

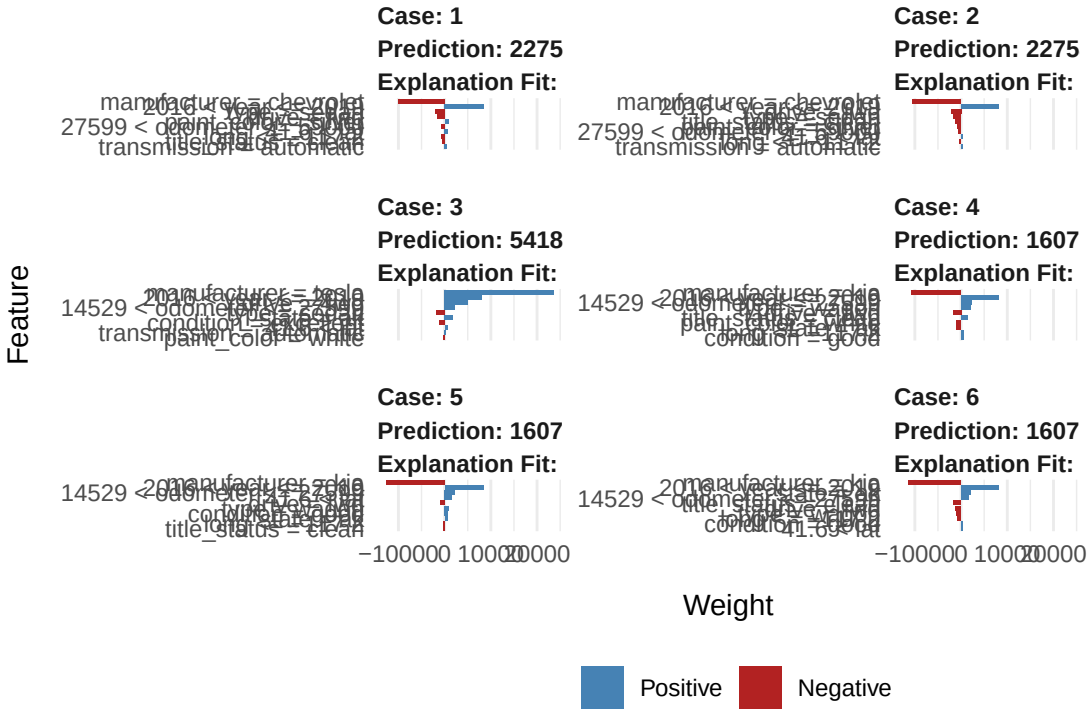
Partial dependence plots We have identified some relevant variables in the variable importance part, important ones are year and odometer. The next step is to attempt to understand how the response variable changes based on these variables. We wull use partial dependence plots to plot the change in the average predicted value as specified feature(s) vary over their marginal distri- bution.



Individual conditional expectation curves ICE curves are an extension of partial dependence plots plots but, rather than plot the average marginal effect on the response variable, we plot the change in the predicted response variable for each observation as we vary each predictor variable.



Lime Once an explainer has been created using the lime() function, it can be used to explain the result of the model on new observations. The explain() function takes new observation along with the explainer and returns a data.frame with prediction explanations, one observation per row. The function plot_features() creates a compact visual representation of the explanations for each case and label combination in an explanation. Each extracted feature is shown with its weight, thus giving the importance of the feature in the label prediction.



Conclusions

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

- We 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 XGB model has been selected by the best training performance, it has RMSE 4219.979 at training set and RMSE 4684 at the test set while the median car price in the dataset is 2.599×10^4.
- Interpreting black-box XGB is done by partial dependence plots, individual conditional expectation curves and local interpretable model-agnostic explainations.
- If more rigorous 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**