Report

(Random\_state = 9047)

Step 1: clean the data

I imputed the missing data with either the median or mode of the variables they belong to, and transformed the string variables I may use to numeric variables. Then I found there were some extreme prices, I dropped these outliers and deleted the lines where the price values were 0. Then I chose proper X variables and standardized them.

Step 2: Pick independent variables

I tried different combinations of different variables that may be the factors influencing the car price as the variables of decision tree regressor model to compare their prediction performance, found the combination of variables “year”, “Model” and “odometer” have relatively higher predicting accuracy.

Step 3: split the data

I split the dataset into training dataset and testing dataset according to the ratio of 8:2.

Step 4: Choose best model

I respectively fitted six models: K-Neighbors Regressor, Decision Tree Regressor, Random Forest Regressor, Bagging Regressor, Adaboost Regressor, XGBoost to the training data, predicted the car price using them with testing data, measure their predicting accuracy rate. For the KNN model, I iterated to fit the model with different K values to the data and calculated corresponding MSLE values, then I plotted the graph with these MSLE values to find the K value leading to the least MSLE and used it as the KNN model’s parameter. For other models, I repeated to try different parameter values to try to make these models have higher predicting accuracy rate. The output of “training data R2 Score” reflect if the models are overfitting, the output of “testing data R2 Score” represents the model’s’ predicting accuracy.

It turned out that the decision tree regressor with the accuracy rate of 0.7495 had the highest predicting accuracy rate among these models, so I chose it as my model.

Reference:

1. Abhash Panwar. (2021) used-car-price-prediction. Available from: https://github.com/abhashpanwar/used-car-price-prediction
2. Srivathsan G. (2020 Predicting the price of used cars. Available from: https://towardsdatascience.com/predicting-the-price-of-used-cars-891d13faf3fc