AMS 598 - Project3 Report

Weiwei Tao

October 26, 2021

1 Problem Description

In this project, we are working on predicting sale price of bulldozers sold at auctions. Two datasets were given: training.csv with 52 features and 300,000 records and test.csv contains 101,125 records. The goal is to use given features to predict sale price which is provided for each record in the training dataset.

2 Sale Price prediction

My analysis includes following steps:

1. Deal with missing values: impute missing MachineHoursCurrentMeter with 0; remove columns with large amount of missing (> 80%) and impute columns that contain less than 80% of missing with 'Unknown'.
2. Feature generation: from sale date, I generated sales year, month and day. The minumum value for year made is 1000, which is obviously an errorness. I replaced those values that are less than 10 percentile with 1966 (10th percentile of year made).
3. Feature wrangling: remove tokens, spaces and numbers for state and fiProductClassDesc.; combine levels with low frequency of ProductGroup and state to reduce total number of levels so as to avoid overfitting in modeling.
4. String indexing for categorical variables: before modeling, it is necessary to convert columns of strings to numbers through string indexing.
5. Data modeling and prediction: the training dataset is splitted into a training dataset and a validation dataset. The validation dataset is used to parameter selection. I utilized random forest for price prediction because random forest is able to prevent overfitting through bagging and feature selection and it is not sensitive to outliers. I have also performed parameter tunning (maxDepth and numTrees) to find out the best model for our data.

3 Prediction results

Random forest regressor with maxDepth of 20 and numTrees of 20 yield a pretty low root mean squared error of 10896, which is used for price prediction in test dataset.

Random forest model also returns feature importance score for each features as shown in Fig. 1. From the feature importance figure, we can conclude that yearMade, produce size, product class and saleyear are top key features related to our prediction.

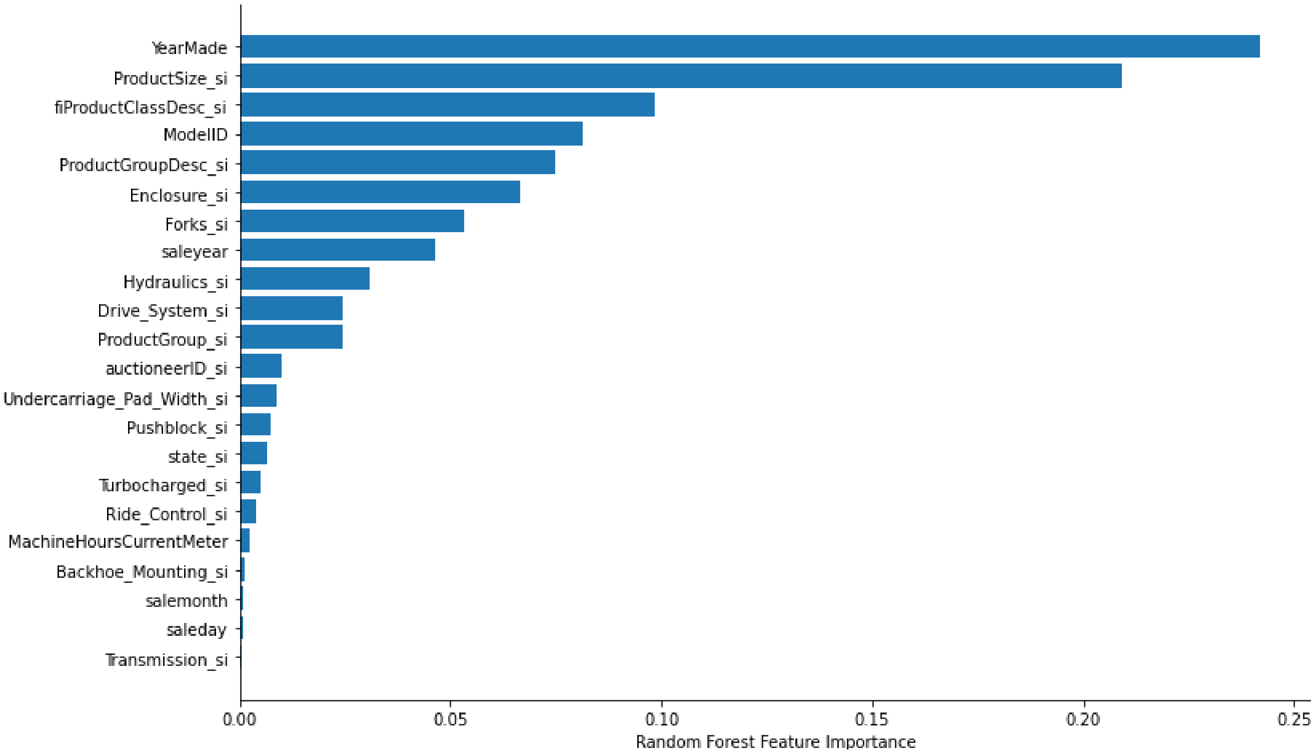

Figure 1: Feature importance from Random Forest Regressor.