Attempting to Predict the Usability of Kaggle Datasets

Marc Marquez
University of Florida
Gainesville, Florida

ABSTRACT

A dataset of Kaggle datasets is taken and used to predict the Usability of said dataset with the predictors log of file size, filetype CSV, filetype Other, Medal type, Months since 2015, and day of the week of upload. The multiple linear regression approach fails due to the limited range of values for Usability coupled with a possible nonlinear relationship, leading to a nonparametric approach to be taken instead. Several tree-based models were trained and tested, and the one with the smallest mean squared error was selected. This turns out to be a boosted tree model of 2000 trees, a shrinkage parameter of 0.01, and an interaction depth of 4, which in turn yielded a mean squared error of 5.1983. This is rather large for an estimate of a parameter that ranges from 0 to 10, suggesting that none of these variables can fully explain Usability (i.e. some missing variables may do better), or that some of the predictors or Usability itself are subject to some form of bias (e.g. sampling).

INTRODUCTION

Kaggle is a website known for hosting data science competitions, teaching courses on data science, and offering publicly available datasets. As of April 20th, 2024, there are 317,983 datasets in Kaggle, with hundreds more being added to the site daily. However, not all datasets are created equal; many contain well-updated, well-documented, and well-explained variables, while others are hardly updated and riddled with missing, awkwardly formatted, or incorrect

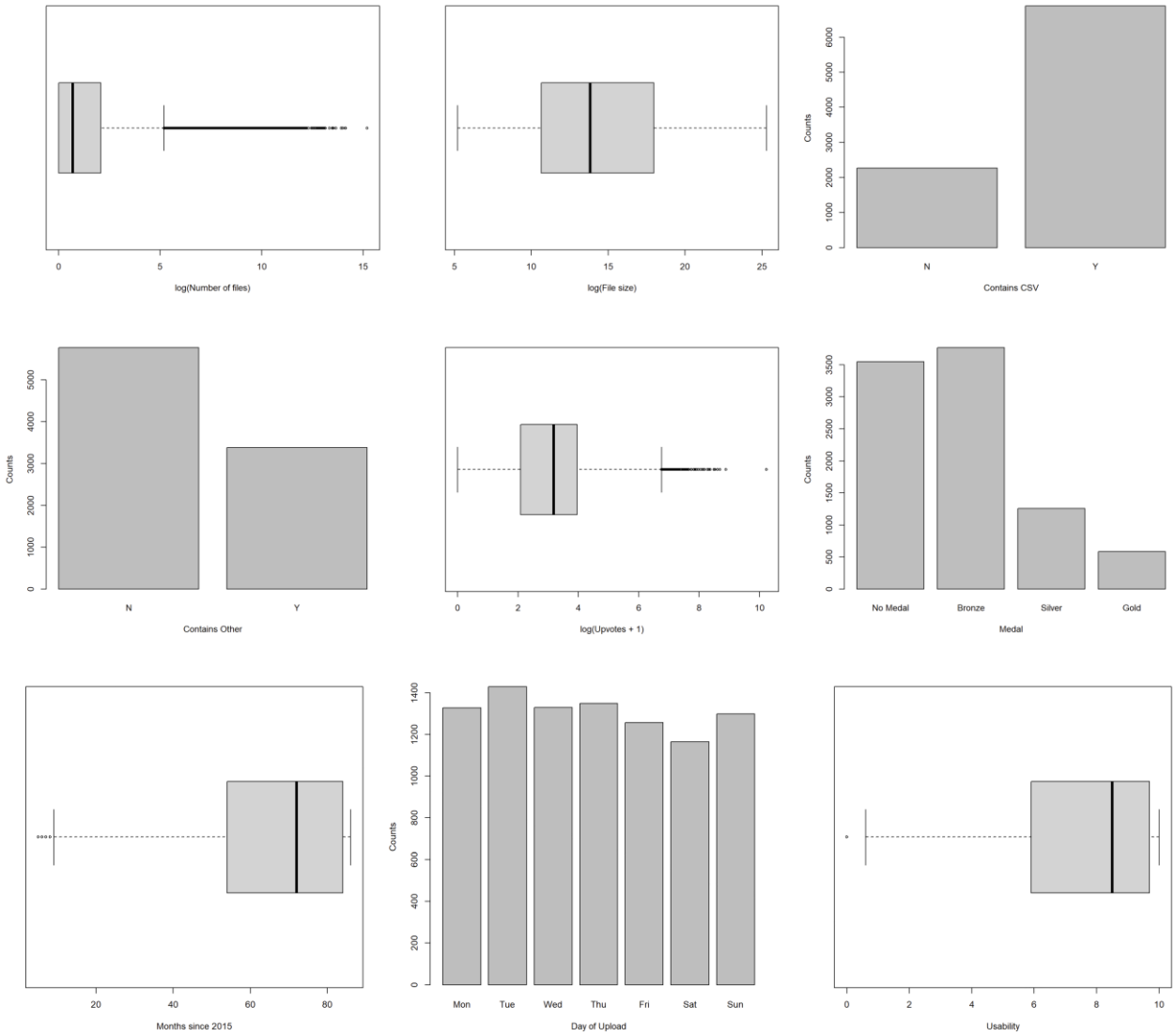
values for confusing variables with no explanation. The Usability feature was created to allow users to differentiate between one and the other. According to Kaggle, a dataset's Usability score is based off its completeness (whether it includes subtitles, tags, an overall description, and a cover image), credibility (proper source inclusion, whether it is public, and update frequency), and compatibility (licensing, file formatting, file descriptions, and column/variable descriptions). However, there are also other measures of a dataset's usefulness, such as its number of Upvotes, and whether the dataset has a certain medal. Theoretically, these other measures should line up with Usability, i.e., a dataset with high Usability should have many upvotes and a gold medal. The purpose of this study, then, is to check if said measures, coupled with other aspects of a dataset, can predict a dataset's usability.

METHODS

Data

The data was taken from a Kaggle dataset containing 9159 entries of datasets and includes the name, author, number of files, file size, types of files included (CSV, JSON, SQLITE, and/or other), number of upvotes, type of medal (none, bronze, silver, gold), date of upload, day of upload (Monday, Tuesday, etc.), time of upload, and the response variable, Usability (on a scale from 0 to 10, where 10 is the best). Of these, time, name, and author were excluded, as the former does not specify a time zone (rendering it meaningless), and the latter are useless. File type was expanded into four separate variables, corresponding to whether the file does or does not contain a CSV, JSON, SQLITE, or other type of file. Date of upload was changed to months since 2015 (Jan. 2016 is 1, Feb. 2016 is 2, etc.). Finally, file size contained data in bytes, kilobytes, and megabytes, and was entirely converted to the former for simplicity.

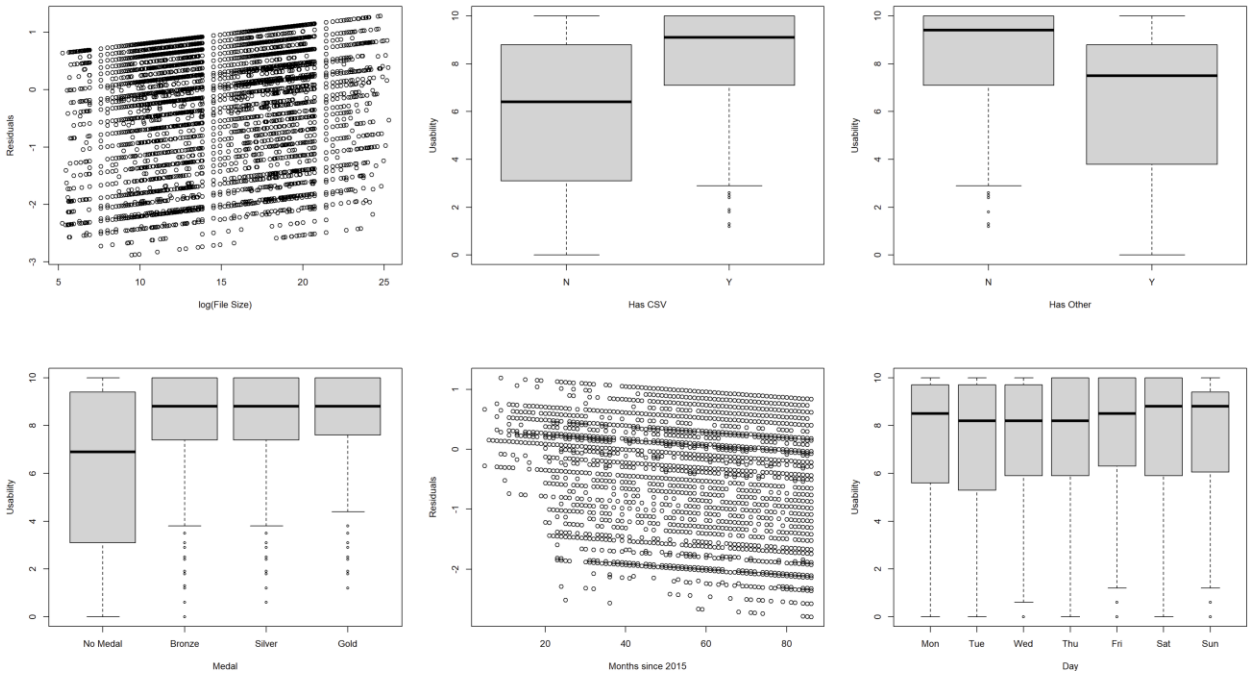
Distributions of Variables



Of the current 10 predictors, several are dropped due to poor sampling. There are far too few entries containing JSON and SQLITE files (332 and 32, respectively), hence they are now incorporated into the Other filetype. All five quantitative predictors and the response Usability contained outliers; the amount in months since 2015 and Usability are minimal (13 and 87, respectively), but number of files (1879), file size (1822), and number of upvotes (988) had far too many and were transformed via the natural log to reduce them. The transformed file size had

no outliers, but the transformed number of files had 1451, and was also dropped. The transformed number of upvotes did have far fewer outliers (109); however, they appear far more extreme. In addition, according to Kaggle, dataset medals are awarded based on the number of upvotes from non-novice users (not including the creator). Thus, datasets with more upvotes will tend to have medals, suggesting some correlation. For these reasons, the number of Upvotes is also dropped. The remaining quantitative variables, Months since 2015 and file size, are not correlated which each other (correlation coefficient of 0.04925).

Model

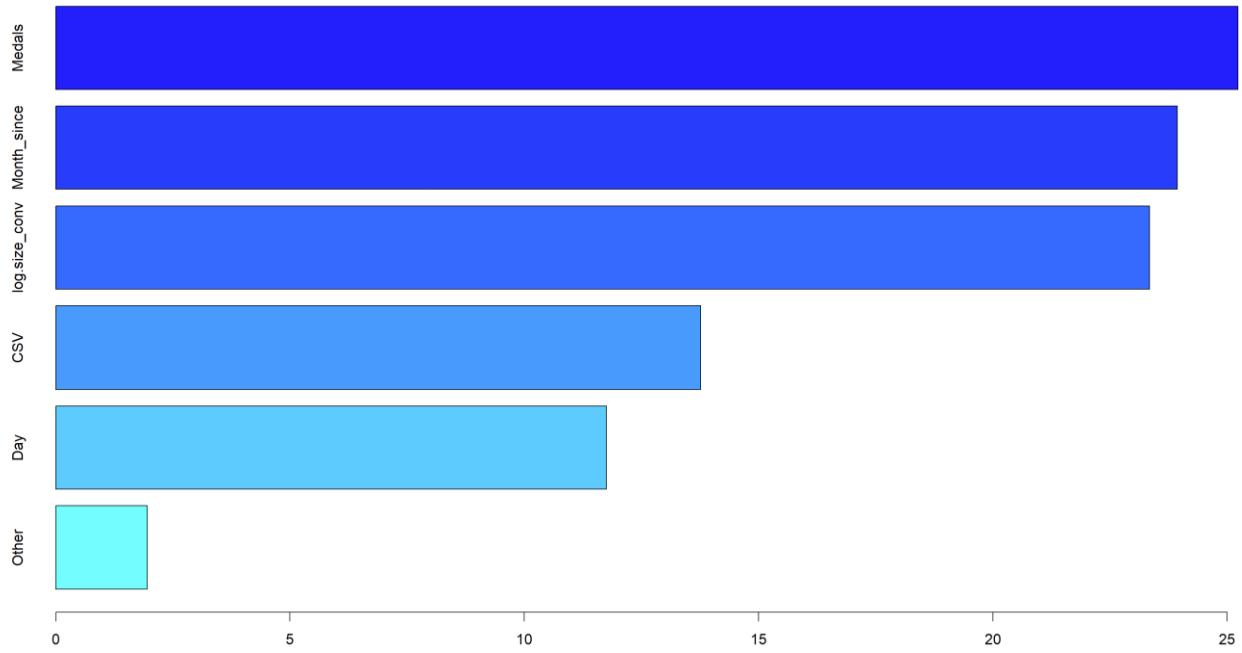


Initially, a multiple linear regression model is used, with quantitative predictors log of file size and months since, and categorical predictors filetype CSV, filetype Other, Medal type, and Day. However, initial modelling of Usability with one factor each yielded residuals with linear trends for the quantitative predictors and wildly heteroskedastic errors for the categorical predictors (using Levene's test yielded a p-value of $3.857 * 10^{-52}$ for filetype CSV, $2.1201 *$

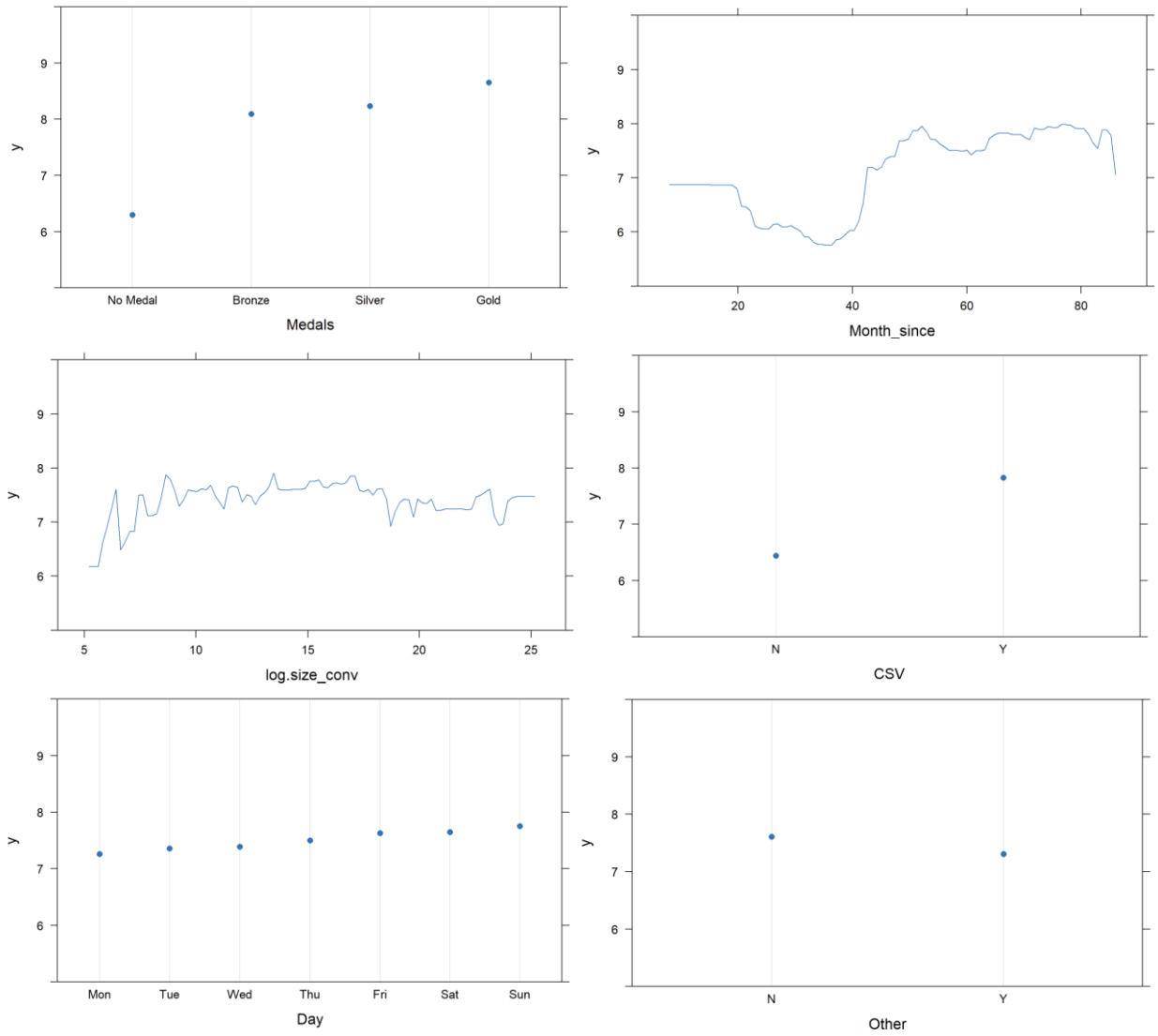
10^{-36} for filetype Other, $1.5526 * 10^{-241}$ for Medal type, and 0.002702 for Day. This is likely due to Usability being limited to values between 0 and 10, inflating the residuals corresponding to fitted values outside this range. This suggests that the relationship between Usability and these predictors is nonlinear, but any attempts to transform the data still yielded similar results. Hence, a nonparametric approach was taken.

Several tree-based models were trained on a random selection of 3000 entries and tested on the other 6159. The models fit were a regular decision tree with cross-validated nodes, a bagged tree made from 200 trees, a random forest made from 200 trees, four separate boosted trees made from 2000 trees and a shrinkage parameter of 0.01, but each with a different interaction depth (1, 2, 3, 4), and a Bayesian Additive Regression tree. The mean squared error between the predicted and actual testing values were taken for each one, and the one with the smallest was selected. This ends up being the boosted tree with an interaction depth of 4.

RESULTS



Based off the relative influence plot, medal type, months since 2015, and the log of file size appear to be the most influential variables in the model with relative influences of 25.2352%, 23.9388%, and 23.3470%, respectively. Below them are filetype CSV, with a relative influence of 13.7616%, and day of the week of upload, with 11.7600%. Filetype other is far below with 1.9575%.



Based on the partial dependence plots, datasets with no medal will tend to have a lower Usability than those with a medal, and those that contain a CSV file will tend to have higher Usability. It is worth noting, however, that despite having the lowest mean square error of the

tree models, it is still an error of 5.1983, which is rather large for predicting a value that lies between 0 and 10.

CONCLUSION

Overall, the variables given by the dataset appear to be poor predictors of Usability, despite their appearances. This may suggest that that there are missing variables that may better explain Usability or that some of the predictors or Usability itself are subject to some form of bias (ironically, this dataset has a usability of 10 while only having 57 upvotes and no medal). Despite the guidelines for acquiring a high Usability being visible, Kaggle never released an exact formula to calculate it. Perhaps a dataset with less skewed variables, or more entries overall may yield a better predictive model; however, given the size of this dataset (9159) accounts for nearly 3% of the entirety of Kaggle datasets, acquiring any more data may incite further bias.