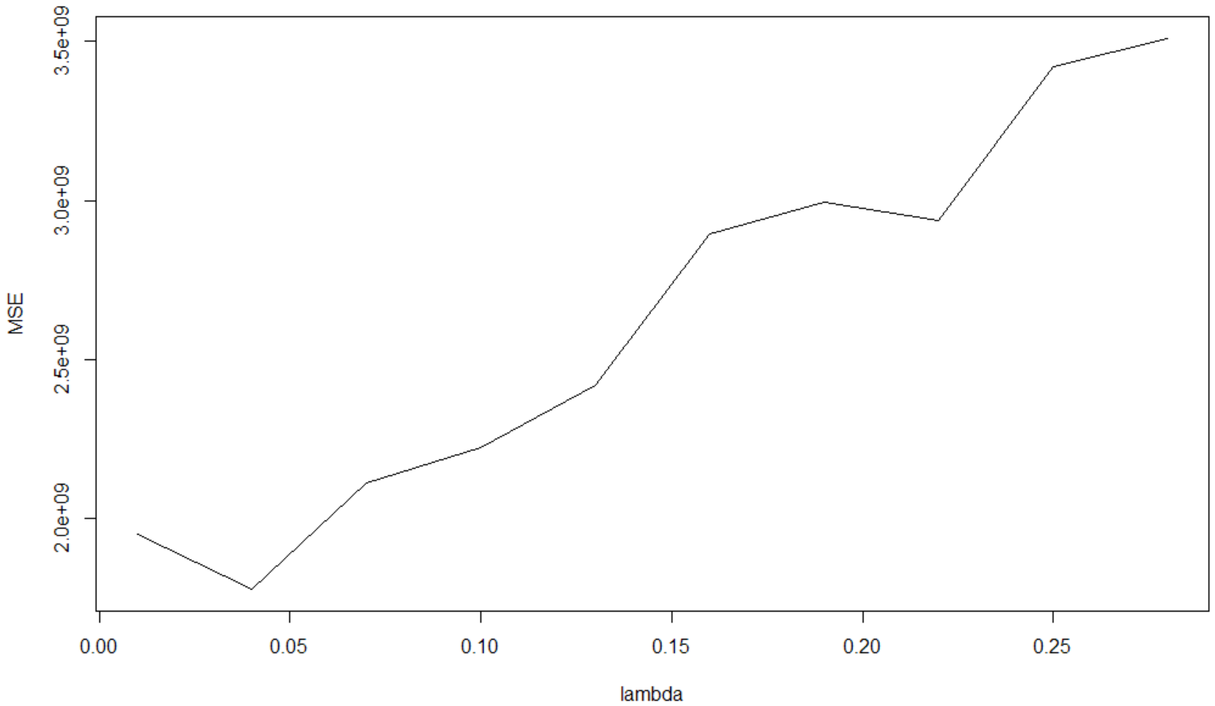
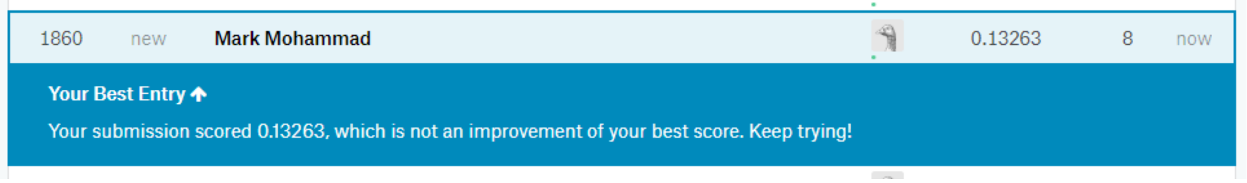
A) The goal of the Housing Prices Project is to predict house prices. The data has 2919 observations, about half with no data for SalePrice that will be held out as the test data. There are 81 variables that contain information on various factors that may affect the sale price of a house in Ames, Iowa. We intend to use cross-validation to experiment with different methods of prediction, namely, boosting and trees.

B) Before doing much work in cleaning the data we did a 4 fold cross-validated gbm to find the best lambda value. We then applied that value (0.07) and ran the model. The kaggle error score on this model of 0.15034 was one of the best we received among all models we tried even though much of the data was missing and we did not attempt any feature engineering. Our next step was to inspect the data for missing values and impute the missing data. Using md.pattern() from the mice package, we found a significant amount of missing data. We did single imputation at first to get a dataset that we could work with to find a reasonable model. The first model was a default boosting model and we received a warning that the variable Utilities had no variation. We removed that variable for all future models. We then did a 4 fold cross validation for lambda with 1500 trees and interaction.depth of 4, and determined that 0.04 was the best value.



. We then ran the model with that lambda value. Based on the variable importance values, we removed a few variables that had 0 importance and ran the model again. Both Kaggle scores were 0.13567, an improvement from the first model that had all the missing data. We realized, however, that we had not verified that all factor variables had been coded properly. We went and changed categorical variables from Integer to Factor and restarted the process. The variable importance showed even more insignificant variables. We took those out and ran the model. The Kaggle score for both models was the same, 0.20335, a decrease from our first models. Disappointed with the reduction in accuracy, we decided to do a multiple imputation on the original data with m=5. We ran a model for each completed dataset and took the average of the values for each observation. The Kaggle score was a slight improvement, 0.19164. This is a plot of the CV error and training error with the line showing the number of trees used in the prediction.

C) Our model that produced the best Kaggle score was a model that we simply took out variables that were considered unimportant in our first models, and also variables that originally had very much missing data. Those variables were Street, PoolQC, PoolArea, MiscFeature, Heating, Alley, FireplaceQu, and Fence. In this model, we did not impute missing values or properly code variables as factors when necessary.

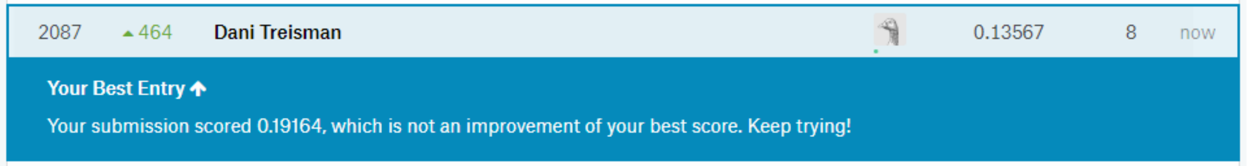


1860 new Mark Mohammad 0.13263 8 now

Your Best Entry ↑

Your submission scored 0.13263, which is not an improvement of your best score. Keep trying!

After correctly coding all variables and fixing missing data with an average of 5 imputed datasets, our final model included all variables except Utilities, which had no variation. This final model utilized boosting with the gbm function. The cross validation error was 960878160. The Kaggle score and rank of our final model…



2087 ▲464 Dani Treisman 0.13567 8 now

Your Best Entry ↑

Your submission scored 0.19164, which is not an improvement of your best score. Keep trying!

D) Our final model did not result in the best score. Even though we achieved a better score with a many missing values and incorrectly coded variables, that is not necessarily correct and we were more comfortable with the lower score and a correct process. If we had more time, we would have created dummy variables for missing entries and also for some variables that had 0 as their square footage where a binary classifier for the existence of that feature would probably give us more information. We also would have liked to try other models such as random forests because it is possible that we were overfitting with our GBM. The main issue we had was the confusion of having a “better” model despite doing the model incorrectly as well as the removal of insignificant variables resulting in the exact same score. We also would have like to cross validate the other parameters since some of them we chose arbitrarily.

.