Kitchen Sink Gradient Boosting

Just take all the variables and throw them into a gradient boosting machine

Approaches

- Psuedocontinuous dependent variable using inverse gaussian transformation of Dep_Var
- 3 Sets of trees with different parameters, use averaged results
- All trees are just stubs (max depth of 2 or 3)
- Use only 35% of features in each tree
- Slow learning rates (0.005 or less), more than100 boosting rounds
- Validate using last 4 quarters in training set

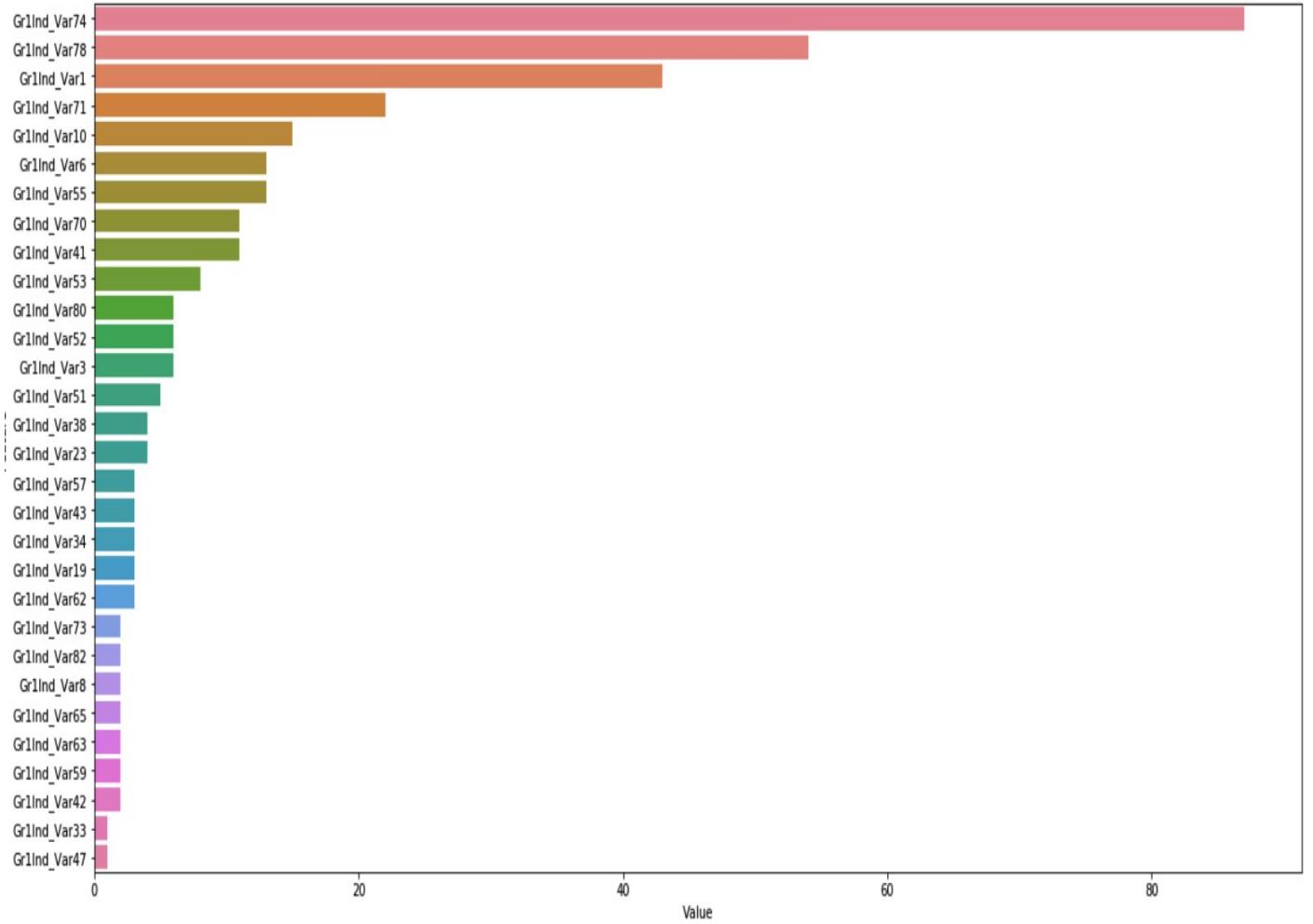
Details

- Used LightGBM
- min_data_in_leaf = 500
- 1 of 3 models used bagging, using 78% of data, resampled every 5 boosting rounds
- Model weights are 0.39, 0.39, and 0.22
- Psuedocontinuous predictions combined, then inverse transformed and rounded to integer
- Did not use individual stock ID's. (Didn't help.) Time-related variables had no effect.

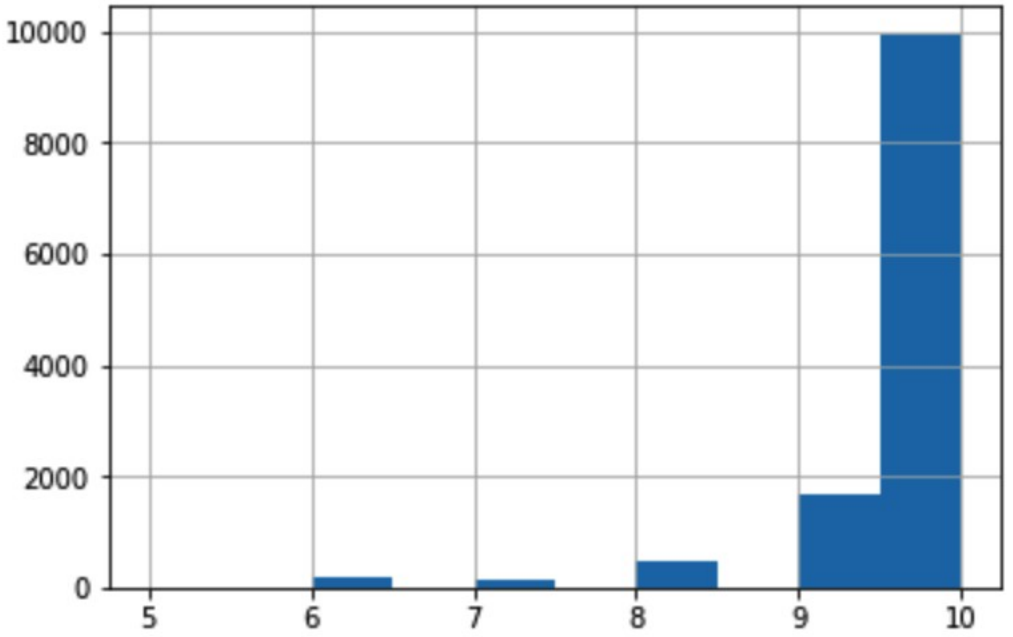
Results

- On validation, Spearman r=0.109 (but hyperparamters were probably overfit, test performance may be weaker)
- Most features showed low importance, but eliminating unimportant features did not help validation results
- Most important: Gr1Ind_Var78, Gr1Ind_Var74, and Gr1Ind_Var1
- Gr2 variables generally not important, but most important of them was Gr2Ind_Var12
- Most predictions were bin 10, all others were lower

Most Important Variables



Distribution of Predicitons



Code is available.