Machine Learning Applied to a Custom Financial Data Set

Aaron Belkin-Rosen
abelkinrosen99@gmail.com

May 3, 2023

Introduction: The modern finance industry is abundant with immense amounts of data, and with the power of modern computing and programming tools, every-day traders are able to create and analyze vast data sets. Coupled with a rise in Machine Learning research, powerful tools have become easily accessible for all. Despite these positive developments, these traders are still largely out-gunned by big financial institutions that deploy these technologies at scale. However, the opportunity for individuals to deploy their own advanced algorithms has certainly made things interesting.

In this project, a custom-made data set containing "significant insider buys" from 2012 to 2022 was created following the investing philosophy described by my grandfather in [5]. The data set under investigation was developed by collecting financial metrics commonly used for technical analysis corresponding to the time of each sufficiently large insider transaction. The complete data set has 2095 data points, each containing 102 features. The data set was labeled into 3 categories by the return achieved within a year of the insider buy, and returns greater than 5%, 10%, and less than 5%, were labeled "Good Buy", "Excellent Buy", and "Weak Buy", respectively. This labeling led to a supervised learning approach where the goal was to use ensemble learning techniques to improve classification accuracy above the capabilities of individual classifiers. The simple idea of ensemble learning is to use a committee of classifiers, as opposed to just one, to assign labels to unseen test data. The benefit of this technique is that individual classifiers may vary in their ability to "learn" particular features, and by combining many classifiers, a more comprehensive model may be created. However, the suc-

cess of ensemble learning is highly data set dependent, so many combinations of data dimensionality reduction techniques, classifier types, and parameters were explored to maximize performance. The novelty of this project lies in the combination of these techniques on a data set created following a unique investing philosophy and spanning a significant market duration. To train and evaluate different classifier configurations, an 80/20 split was performed to obtain training and test sets. The ensemble techniques of "bagging" and "stacking" have shown the most empirical promise on tabular financial data sets [4], so these were deployed. Due to the existence of three classes, a classifier scoring significantly higher than 33% classification accuracy, which would be achieved by random guessing, would be sufficient to defend the project's success.

Figure 1: Machine Learning Pipeline.

Methods: In this project, the pipeline shown in Figure 1 was followed. First, min-max and Z-score normalization functions were applied to the raw data set. Min-max normalization transforms the data for each feature to a value between 0 and 1, with the absolute minima and maxima being 0 and 1, respectively, and every other data-point being a decimal. This allows features contrasting in numerical context to be fairly compared. Z-score normalization transforms each feature's data to have a mean of 0 and a standard deviation (std) of 1. This normalized data set was then applied to various feature ranking algorithms including principal component analysis (PCA), minimum redundancy maximum relevance (MRMR), variance ratio (VR), augmented-variance ratio (AVR), and correlation reduction. These feature ranking algorithms were applied to five selected base classifiers and the best performing combinations of feature selection algorithm and base classifier were chosen for further investigation in the ensemble methods.

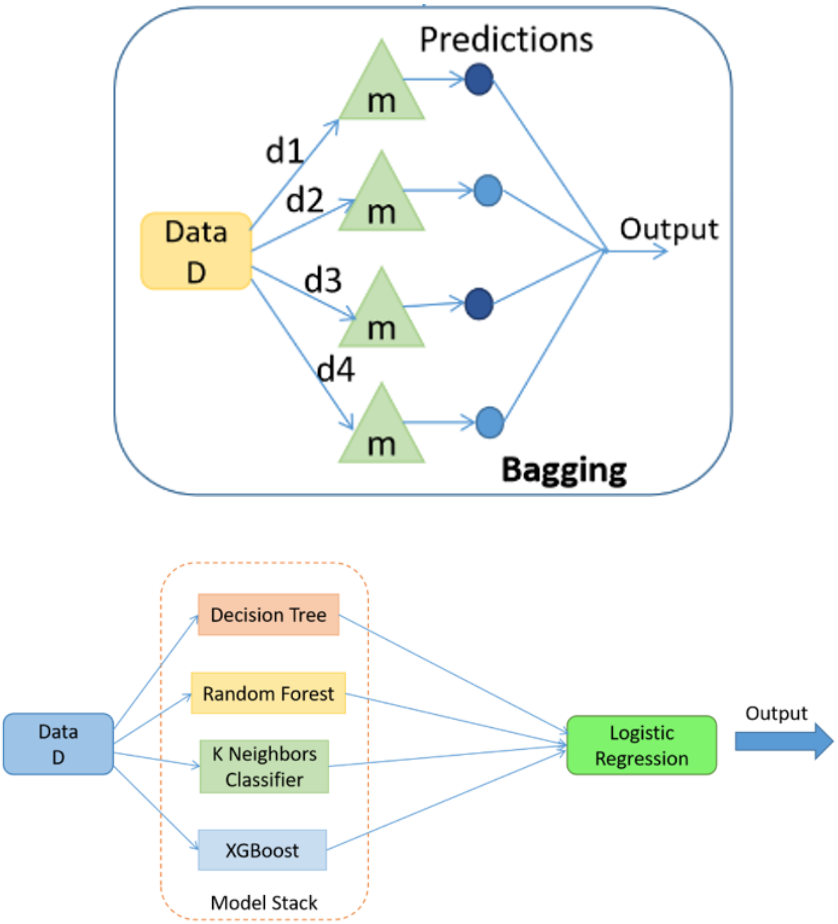

Figure 2: Bagging (top) and Stacking (botton) methods [2].

These base classifiers consisted of gradient-boosted decision trees (XGBoost and CatBoost [3]), geometric decision boundary classifiers (K-Nearest Neighbors and Support Vector Machines), and a neural network (MLP). Lastly, the ensemble methods of bagging and boosting were deployed, parameters were empirically optimized, and selected model configurations were reported in the results.

The bagging and boosting techniques are shown in Figure 2. In bagging, a feature selection filter is applied to the data set, a training/test split is performed, and an ensemble of "N" homogeneous classifiers is trained using the training set. After training, a test data set is fed through each classifier and the ensemble selects the mode of the model predictions as the final classification. Each classifier is trained by randomly sampling, with replacement, 80% of the training set's data points and 20 random features. Additionally, N was selected to be 40. These parameters were empirically selected to maximize performance and combinations of type of classifier and feature selection filter were explored in experimentation. In stacking, a "stack" of base models is trained using a training set and the training data is fed through each classifier in the stack to generate a feature space of predictions. This feature space and its corresponding labels are used to train a level 1 classifier, which maps predictions from the stack to expected outputs. The output of the level 1 classifier is the final prediction [1]. The parameters that were optimized for were the weight of each type of classifier in the ensemble, the number of features in the reduced data set, and the type of level 1 classifier used for the final classification. The parameters that were selected were weighting the gradient-boosted decision trees 1.2:1 with respect to the other classifiers, and selecting the top 25 features in each filter. In experimentation, the type of feature selection filter and level 1 classifier were explored.

No Filter Mean Cross Val. (MCV)	XGBoost	CatBoost	KNN	SVM	MLP
MCV	0.50	0.51	0.47	0.48	0.49

Filter Mean Cross Val.	XGBoost	CatBoost	KNN	SVM	MLP
Correlation	0.47	0.52	0.46	0.46	0.47
PCA	0.49	0.49	0.48	0.48	0.46
MRMR	0.52	0.53	0.51	0.49	0.46
VR	0.50	0.48	0.51	0.48	0.48
AVR	0.50	0.50	0.49	0.47	0.49

Figure 3: Mean-Cross Validation of each classifier on raw data set (top) and reduced data set (bottom) by feature selection algorithm.

Results: The result shown in Figure 3 explores the effect of feature selection algorithm on classification accuracy using 5-Fold Mean Cross Validation (MCV) as the evaluation criterion. The MCV is computed by re-sampling the training set numerous times to train a model on each fold and taking the average accuracy. These results serve as the justification for the classifier/filter combinations selected for further investigation in the bagging and stacking ensemble techniques. Figures 4 and 5 show the results of bagging and stacking on min-max and Z-score normalized data, respectively. These figures show the average accuracy and false positive rates for each classifier/filter combination used in bagging and stacking. When trusting a model to evaluate the strength of a stock purchase, its false positive rate is just as important as the overall model accuracy. The false positive rate is reported as the ratio of predicted strong buys that turned out to be weak to the total number of predicted strong buys. The importance of this metric is illustrated by a confusion matrix in Figure 6, which shows that despite having very similar overall accuracy scores, two models can perform quite differently in false-positive rate. Since these results are based on the mean of 10 iterations of bagging and stacking, standard error of the mean (SEM) was calculated and reported using an error bar.

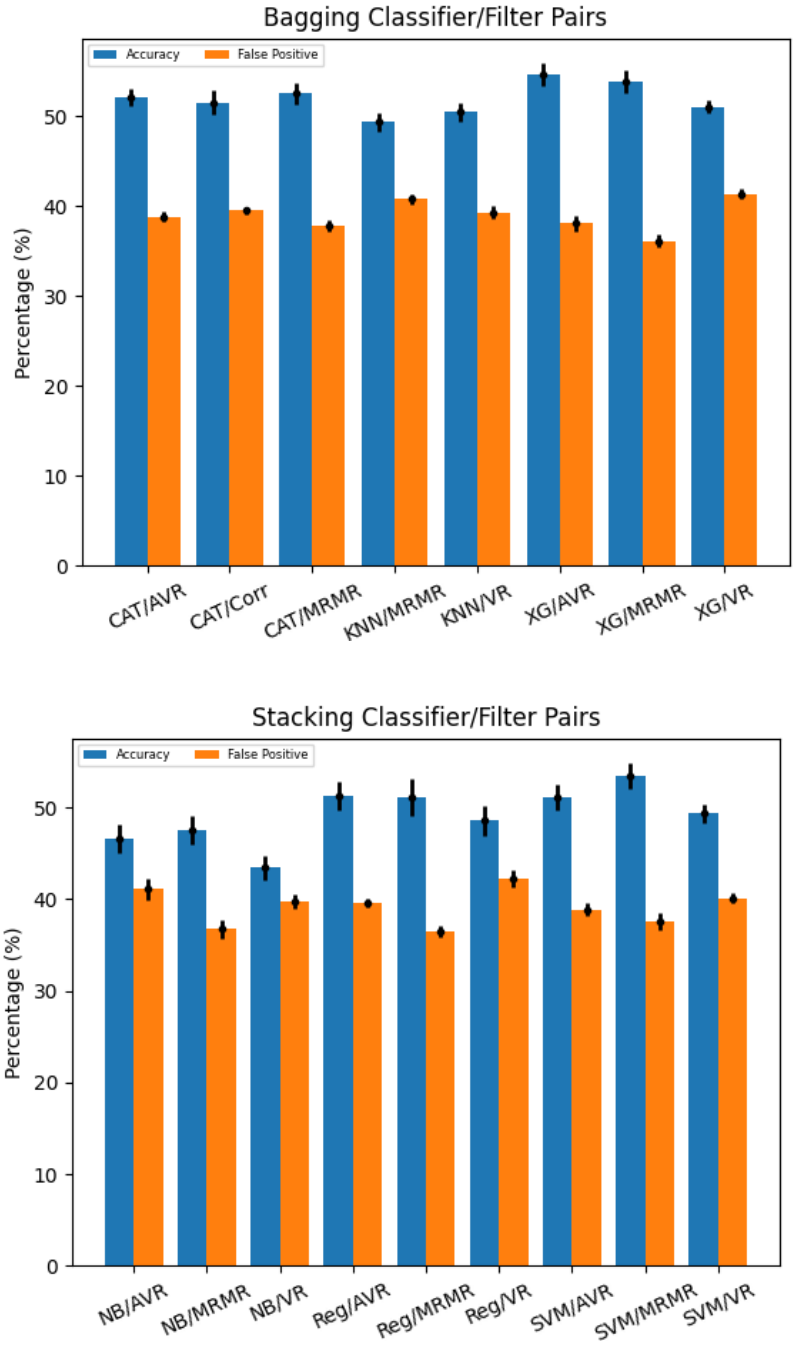

Figure 4: Bagging and Stacking results using min-max normalization.

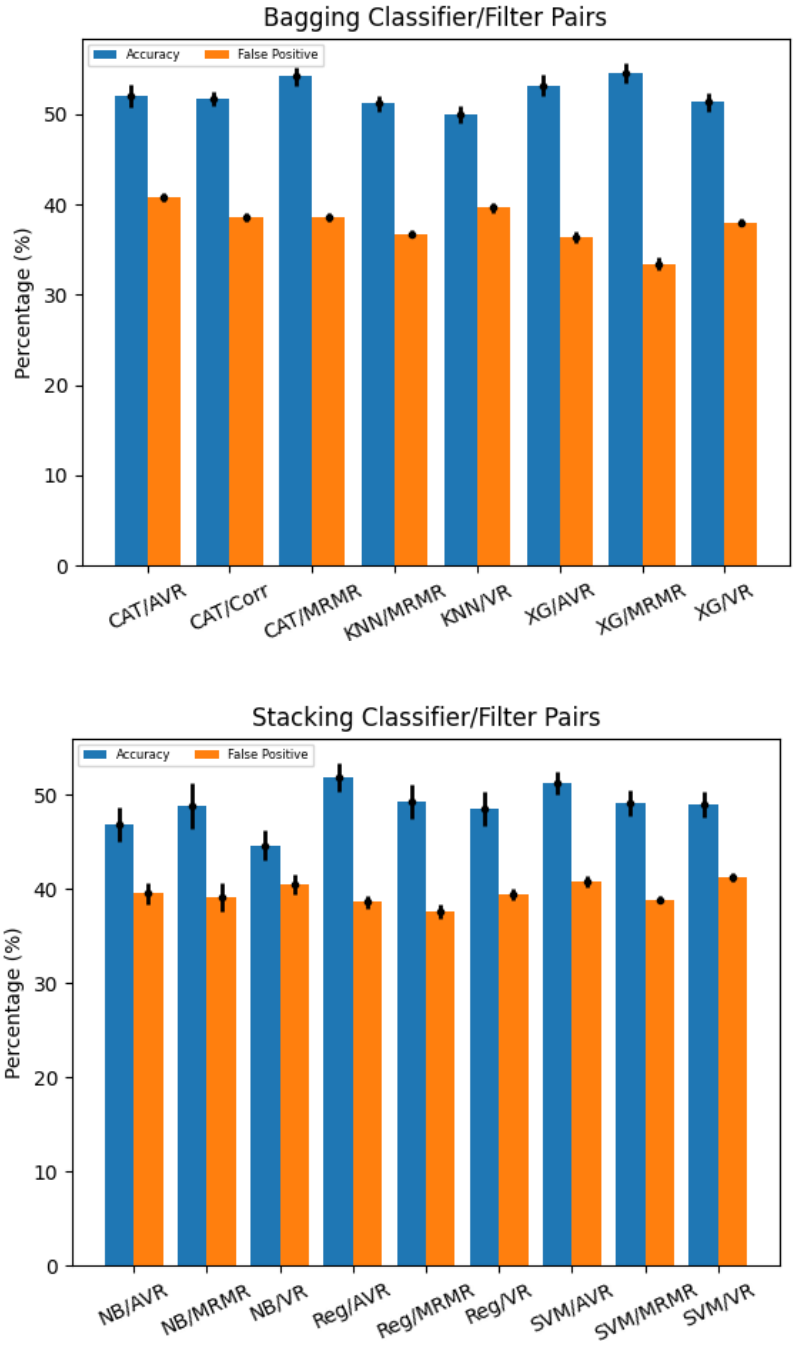

Figure 5: Bagging and Stacking results using Z-score normalization.

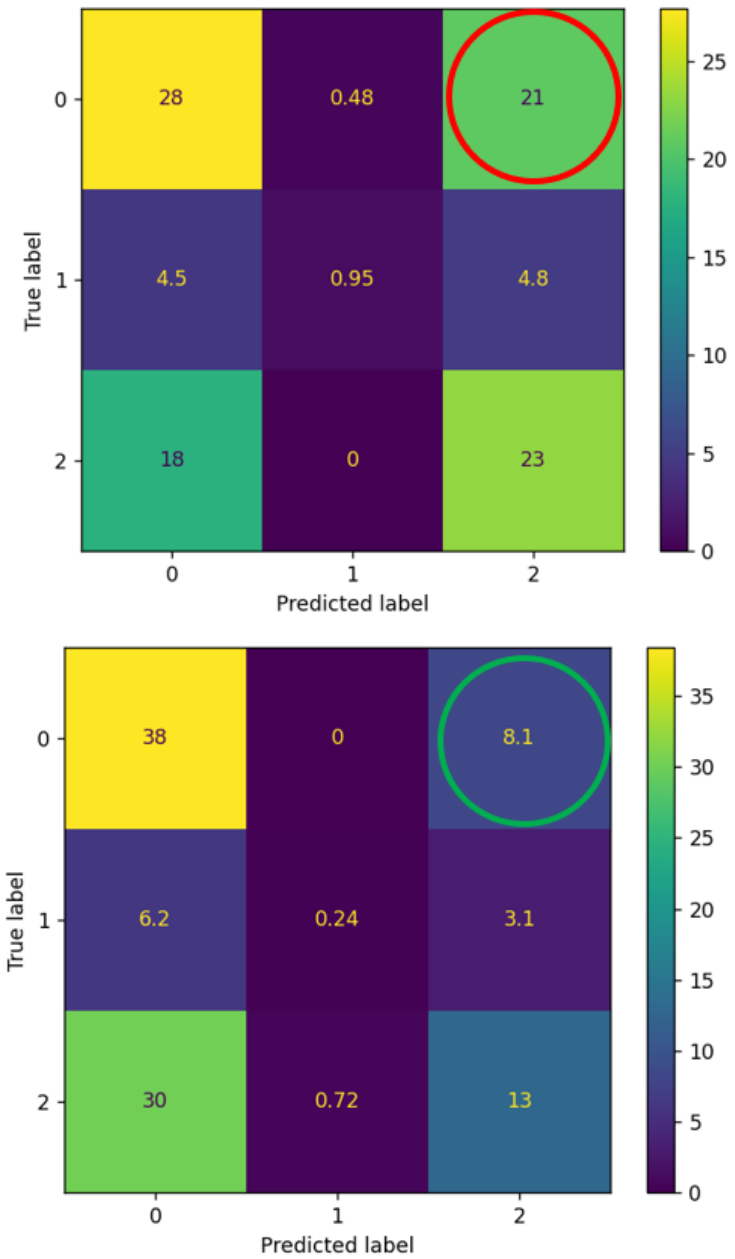

Figure 6: Confusion matrix with each square normalized to percentage of total test set with labels 0, 1, and 2 representing weak, good, and excellent buys, respectively.

Discussion: The results shown by Figures 4 and 5 show that bagging clearly outperformed stacking in both min-max and z-score normalized data sets. The best classifier/filter combination for min-max and z-score normalized data sets in bagging were XGBoost/AVR (54.68%±1.24%) and XGBoost/MRMR (54.56%±1.07%), respectively. Despite having a slightly worse overall classification accuracy, XGBoost/MRMR greatly outperformed XGBoost/AVR in false positive rate, scoring 33.42%±0.67% versus 38.11%±0.84%.

The best level 1 classifier and filter combination for min-max and z-score normalized data sets in stacking were SVM/MRMR (53.46%±1.38%) and Regression/AVR (51.91%±1.47%), respectively. Since the false positive rates for these configurations were 37.59%±1.00% and 38.67%±0.71%, bagging also clearly outperformed stacking in this metric as well. The confusion matrix for z-score normalized XGBoost/MRMR, the most promising classifier/filter combination, is shown in Figure 7.

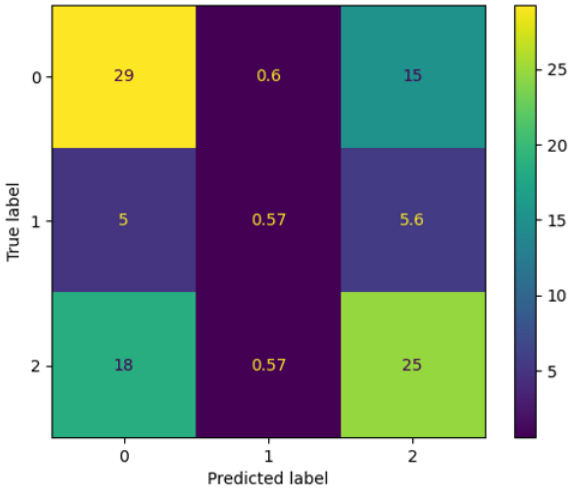

Figure 7: Most promising classifier configuration (Z-score normalized, bagging XGBoost/MRMR).

Conclusion: This work has explored the amalgamation of normalization, feature selection filters, and ensemble learning techniques on a custom financial data set containing significant insider buys from 2012 to 2022. Despite some promising results, additional model configurations should be explored to further maximize classification accuracy and minimize false positive rate.

References

[1] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[2] *Ensemble Stacking for Machine Learning and Deep Learning*. https://www.analyticsvidhya.com/blog/2021/08/ensemble-stacking-for-machine-learning-and-deep-learning/.

[3] John T Hancock and Taghi M. Khoshgoftaar. "CatBoost for big data: an interdisciplinary review". In: *Journal of Big Data* 7.94 (2020). DOI: https://doi.org/10.1186/s40537-020-00369-8.

[4] Isaac Kofi Niti, Benjamin Asubam Weyori, and Adebayo Felix Adekoya. "A comprehensive evaluation of ensemble learning for stock-market prediction". In: *Journal of Big Data* 7.20 (2020). DOI: https://doi.org/10.1186/s40537-020-00299-5.

[5] Gerald Harris Rosen. *A New Science of Stock Market Investing: How to Predict Stock Price Movements Consistently and Profitably*. Harper Collins, 1990. ISBN: 9780887303937.