Machine Learning Prediction of Recessions: *An Imbalanced Classification Approach*

ALIREZA YAZDANI

ALIREZA YAZDANI
is chief data scientist and founder at Calcolo Analytics LLC in Newton, MA.
alirezayazdani21@yahoo.com

KEY FINDINGS

- We examine prediction of US recessions from a machine learning (classification) perspective.
- We employ class imbalance adjustments and carefully analyze its impacts on predictions.
- Ensemble models accurately predict recessions during and preceding major financial crises.

ABSTRACT: *The author examines the problem of predicting recessions from a machine learning perspective and employs a number of machine learning algorithms to predict the likelihood of recession in a given month using historical data from a set of macroeconomic time-series predictors. The author argues that, owing to the low frequency of historical recessions, this problem is better dealt with using an imbalanced classification approach. The author applies measures to compensate for class imbalance and uses various performance metrics to evaluate and compare models. With these measures in place, ensemble machine learning models predict recessions with high accuracy and great reliability. In particular, a random forest model achieves a near perfect true-positive rate within the historical training sample, generalizes extremely well to a test period containing the 2008–2009 financial crisis, and shows elevated recession probabilities during the last few months of 2019, associated with the tightened macroeconomic environment and worsened by the COVID-19 pandemic.*

TOPICS: ***Big data/machine learning, performance measurement, simulations****

Predicting business cycles and recessions is of great importance for investors, businesses, and macroeconomists alike, helping them to foresee financial distress and to seek alternative investment strategies. Important research has been conducted over the past several years on the study and prediction of US recessions, in which great emphasis has been placed historically on identifying financial predictors of recessions. Subsequently, several predictors of recessions have been documented, including the slope of the yield curve (Estrella and Hardouvelis 1991), stock market performance (Estrella and Mishkin 1998), credit market activity (Levanon et al. 2011), nonfarm payrolls (2012), and other employment and interest rate variables (Ng 2014). Traditionally, the dominant modeling approach in the literature has been to estimate the probability of recessions by using probit models, in which the model can admit a binary dependent variable (e.g., recession = 1 and no recession = 0) with continuous (and discrete) predictors. Notable examples of using probit modeling

*All articles are now categorized by topics and subtopics. **View at PM-Research.com.**

methodology for predicting recessions include work by Dueker (1997), Atta-Mensah and Tkacz (1998), Wright (2006), and Liu and Moench (2014).

Technically, a probit model belongs to the family of generalized linear models (GLMs). In essence, GLMs are based on constructing a linear combination of predictors and using an arbitrary transformation, the *link function*, to estimate (the likelihood of) an outcome. Using the identity link function yields the standard linear regression model, whereas in a probit model the inverse Gaussian cumulative distribution link function is used to derive the probability of a binary event (Hastie, Tibshirani, and Friedman 2008). In this sense, probit is a parametric model that, given its global linear functional form, uses constant weight for each predictor across all recessions used in training samples. In addition, a base probit model does not account for possible interactions and conditional relationships among variables, unless they are known and built into the model in advance. There are also well-known issues associated with the model misspecifications and inconsistency of probit regression coefficients when the assumption of normality of the latent variable model residuals is violated (Yatchew and Griliches 1985).

Consequently, it makes sense to critically evaluate the adequacy of probit models in predicting US recessions and to develop alternative modeling strategies less bound by distributional assumptions. In recent years, researchers have argued in favor of using different modeling methods for the aforementioned problem. For example, Sephton (2001) found that nonparametric models such as multiple adaptive regression splines applied to a small set of financial predictors have good predictive power in sample, although power is limited out of sample. He further concluded that using an extended set of predictors consumed by more dynamic (data-driven) models to allow for nonlinearity and variable interactions may offer significant advantages. This observation strongly motivates use of machine learning models for predicting recessions; such models offer greater flexibility toward incorporating a broader selection of predictors and the ability to discover interactions and nonlinear relationships in data.

Recently, machine learning has been used increasingly to find additional predictors and identify business-cycle turning points to predict recessions. A few notable examples involve using neural networks (Qi 2001), learning vector quantization (Giusto and Piger 2017), and support vector machines (SVMs) (James, Abu-Mostafa, and Qiao 2019). Despite recent progress in using machine learning to predict recessions, one important consideration is absent from the literature, and that concerns the low frequency of historical recessions. As reported in many prior works, the number of months of historical US recessions recorded by the National Bureau of Economic Research (NBER) comprises only a small percentage, less than approximately 15%, of all months between 1960 and 2020. In other words, the problem of predicting US recessions using machine learning is an imbalanced classification problem.

Imbalanced classification refers to a situation in which the frequency of one class in the response variable is much lower than another, creating skewed class distributions. Class imbalances may have different severities and result in predictions that are biased toward the majority class and ignore the rare (yet impactful) events in the minority class. Subsequently, standard classification accuracy metrics such as the accuracy rate that are based on measuring the fraction of correctly classified instances in a training sample may become ineffective and misleading (Kuhn and Johnson 2013). This poses a challenge in the training and validation of classification algorithms, and it should be compensated for jointly by making subsampling adjustments and evaluating model performance using metrics robust to class imbalances.

In this article, we address these issues by developing a framework to predict US recessions, in which we (1) use moderately sized predictors that were identified by previous research, (2) compensate for the response variable class imbalance through subsampling adjustments, (3) train and evaluate a suite of machine learning classification algorithms to expand upon those previously reported in the literature, and (4) conduct a detailed analysis of the results by examining a diverse selection of performance metrics and validation strategies to ensure accuracy and reliability of the predictions. To the best of our knowledge, no prior work has addressed these issues in this systematic manner; our article makes a novel contribution to the literature by incorporating some technical delicacies of recession predictions, while generally demonstrating the usefulness of machine learning models in financial applications.

The remainder of this article is as follows. First, we describe our methodology for model development; we introduce data items, transformations, and sampling adjustments for the imbalanced classification approach.

EXHIBIT 1
List of Predictors Used as Features in Machine Learning Models

Series	Acronym	Transformation (lag in months)	Reported by
Federal Funds Rate	FEDFUNDS	DIFF (1M)	Sephton (2001), Ng (2014)
Industrial Production	INDPRO	DIFF_LOG (1M)	Sephton (2001), Chauvet and Piger (2008)
Total Nonfarm Payroll	PAYEMS	DIFF_LOG (1M)	Camacho et al. (2012)
S&P 500 Index	SP500	DIFF_LOG (1M)	Estrella and Mishkin (1998), Qi (2001)
10-Year Treasury Bond	TY10	DIFF (1M)	Estrella and Mishkin (1998)
Unemployment Rate	UNEMPLOY	DIFF_LOG (1M)	Sephton (2001), Ng (2014)
Slope of the Yield Curve	YC	TY10 – FEDFUNDS (1M)	Estrella and Hardouvelis (1991), Estrella and Mishkin (1996), Wright (2006)

We then introduce a number of machine learning classification algorithms used for this analysis and describe certain implementation aspects. We also introduce various performance metrics to evaluate and compare the results of different classification models. We then present a summary of the results from training and test sample performance. Finally, we summarize and conclude.

DATA

The US business cycle expansion and recession dates are obtained from the NBER website.[1] NBER defines a recession as a "significant decline in economic activity spread across the economy, lasting more than a few months, normally visible in real GDP, real income, employment and industrial production." We specify the target variable for our machine learning classification models to be a binary variable, based on NBER recession months, where recession = 1 (or positive) and no recession = 0 (or negative). A list of predictors (features) used for this analysis is given in Exhibit 1. These predictors have been reported by prior research (references in Exhibit 1) as containing information about the macroeconomy and predictors of US recessions.

We retrieved historical time series data from the Federal Reserve Bank of St. Louis website[2] and aggregated them to monthly frequency by sampling from the end-of-month values when the original series are daily or carrying the last observations forward to impute the next two months when the original series are quarterly. We have used appropriate lagging of predictors, in some cases more conservatively than what is reported in the literature, to eliminate the risk of look-ahead bias and publication delays. Moreover, we have detrended and made stationary the predictor time-series data by using transformations such as differencing or differencing of the logarithms of the original series.

The full dataset spans from January 1959 through the end of December 2019 and consists of 732 data points, of which 101 months (circa 14%) are flagged as NBER recessions. This gives a class distribution ratio of approximately 6:1, which may be regarded a case of moderate to severe class imbalance. This situation may be addressed through a number of compensative sampling mechanisms, including (1) down-sampling the majority class to create a roughly equal distribution of the two classes, (2) up-sampling by creating new instances of the minority class through random sampling with replacement, and (3) the synthetic minority oversampling technique, referred to as SMOTE (Chawla et al. 2002), to simultaneously down-sample the majority class and synthesize new data points in the minority class in accordance with a prespecified class distribution ratio.[3]

These subsampling adjustments are generally found to be effective in dealing with negative consequences of imbalanced class distributions, particularly in high-dimensional datasets (Batista, Prati, and Monard 2004; Kuhn and Johnson 2013). Meanwhile,

[1] https://www.nber.org/cycles/cyclesmain.html (accessed May 2020).

[2] https://fred.stlouisfed.org/ (accessed May 2020).

[3] There are other ways to oversample a time series while preserving conditional dependency, using the so-called ADASYN method. See Cao et al. (2011) for theoretical aspects and Dixon, Klabjan, and Wei (2017) for software implementation. We deem this unnecessary here due to using differencing and the reduction of such dependency.

some experimental results suggest that simple down-sampling tends to outperform other methods in most cases involving low-dimensional datasets—see Hulse, Khoshgoftaar, and Napolitano (2007), Wallace et al. (2011), and Blagus and Lusa (2013). Given the relatively small to medium size of our data and other considerations such as the cost of computing and issues associated with introduction of noise through creating synthetic data, we have chosen to use a down-sampling approach for this analysis. In what follows, we run experiments to compare the performance of models using multiple metrics to ensure robustness in model outcome while class imbalance is present.

TRAINING, CROSS-VALIDATION, AND TESTING

We first split full data into two partitions: (1) a training sample from January 1959 to the end of December 2006 and (2) a test sample from January 2007 to the end of December 2019. The training sample consists of 576 months of historical observations (circa 78% of full data), and the test sample consists of 156 months (circa 21% of full data). We hold this test sample out of any model training processes and will only use it to evaluate classification model performance when unseen data outside of the original training sample are encountered. This particular choice of holdout sample challenges models to generalize learnings from earlier history to a previously unseen turbulent period containing recessions associated with the 2008–2009 financial crisis.

An additional consideration when training machine learning models on time-series data concerns the correct use of cross-validation (CV). In a standard k-fold CV, the training data are randomly partitioned into k folds of roughly equal size, and models are trained using data in k – 1 folds while making predictions of samples in the remaining fold. This process is repeated one at a time until all k performance estimates have been made and summarized into a single performance metric (Kuhn and Johnson 2013). For time-series data, however, a random partitioning may not be valid given the requirement that training samples occur prior to the validation or test samples. Alternatively, we employ a CV technique based on the *rolling forecasting origin* (Hyndman and Athanasopoulos 2018), whereby the blocks of training and validation sets move in time to ensure no future information is passed into the model training.[4]

MACHINE LEARNING ALGORITHMS

We employ a number of machine learning classification algorithms for the modeling and prediction of recessions. These machine learning models represent a diverse set of learning mechanisms, and they give us better insight into the overall predictive power and effectiveness of machine learning approach for this problem. These models include

- probit binary classification, used as a benchmark for predicting recessions
- GLMNET (Friedman, Hastie, and Tibshirani 2008) for fitting generalized linear models while using regularization for variable selection
- SVM with radial basis function kernel (Vapnik 1996) to develop nonlinear decision boundaries for class separation problems
- random forest (RF), ensemble learning (Breiman 2001) based on the construction of several decision trees and aggregation (i.e., mode) of the predicted classes
- extreme gradient boosting (XGB), an ensemble method for efficient implementation of gradient boosted trees with formal regularization (Chen and Guestrin 2016)
- neural network (NNET), consisting a single hidden layer feedforward network

We use these algorithms to estimate the probability of a recession month, with historical input features and training labels (recession or no recession), and convert the estimated class probabilities to a binary prediction using a 50% probability threshold.

CLASSIFICATION PERFORMANCE METRICS

Various metrics exist to evaluate and compare the performance of classification models in practice.

[4] We are using windows/blocks with lengths of 24 and 3 months for training and validation, respectively. These lengths are somewhat arbitrary, but a block time-series CV may be improved through data-driven methods, such as a recently proposed modular machine learning approach. See Simonian (2020) for details.

Although some of these metrics are more conventional, they may not always be robust to class imbalances. It is preferable to evaluate and compare models across a number of such metrics to gain greater confidence in the results. For a survey of these metrics and their properties, see Kuhn and Johnson (2013) and Brownlee (2019). We use the following metrics:

- accuracy rate: the proportion of all correctly predicted cases (positive or negative)
- precision: the ratio of true positives over the sum of true positives and true negatives
- sensitivity or recall: the percentage of positive cases correctly predicted
- specificity: the percentage of negative cases correctly predicted
- area under the receiver operating characteristic (ROC) curve (AUC): combining sensitivity and specificity into a single metric, regarded as robust to class imbalances
- F-score: the harmonic mean of precision and recall, defined as $(\frac{2 * Precision * Recall}{Precision + Recall})$
- H-measure (Hand 2009): misclassification-cost-weighted metric adjusting for different classifiers' tendencies that favor sensitivity versus specificity and vice versa
- Kolmogorov–Smirnov (KS): representing a model's ability to distinguish classes

For these metrics, higher scores (capped at 100%) indicate better performance. Moreover, it is possible to optimize a machine learning classification algorithm with respect to a specific penalty function, usually based on an evaluation metric such as those listed. In such a case, the algorithm seeks to optimize (e.g., maximize) that metric during the training process. The selection of one metric versus another may involve a trade-off depending on the problem objective. Here, to optimize the learning process to correctly predict positive classes (i.e., recessions), we choose to maximize the sensitivity metric during model training. It is worth mentioning that, in our dataset, 82 out of 576 training months (circa 14.2%) and 19 out of 156 test months (circa 12.2%) are flagged as recessions. In other words, for our models to be effective, they need to outperform a naïve classification (that always predicts the minority class) with an accuracy rate of 86% on the training sample and 88% on the test sample.

SUMMARY OF RESULTS

With the aforementioned settings, we are now prepared to examine the results of our experiment. First, we take a look at the performance of different models on the training sample, spanning January 1959–December 2006. Exhibit 2 illustrates the training sample ROC curves for different models. The ROC curve is a plot of the true-positive rates (i.e., sensitivity) against the false-positive rates (i.e., 1 – specificity) at various thresholds. It may be used to visualize the performance of a binary classification in terms of sensitivity and specificity trade-offs, rank order the performance of different models, and determine alternate thresholds for class probabilities (Kuhn and Johnson 2013). Except for the RF model, and to some extent XGB, which exhibits a distinct ROC path located at the top (shown in red and dashed dark blue, respectively), the ROC curves of the other models overlap significantly, which makes it hard to interpret and distinguish among model performances using this metric alone.

Subsequently, it is imperative to evaluate and understand model performance using various other metrics to better judge models' predictive power, as well as the similarities and differences among them. These comparisons are presented in Exhibit 3. A quick look at the values in the exhibit suggests that the RF model has achieved the best performance overall, as indicated by various metrics. For example, although the standard accuracy rate or specificity is close across models, RF clearly distinguishes itself by achieving higher scores across most other metrics. In particular, RF achieves a significantly higher H-measure and a maximum possible value of 100% sensitivity, indicating it has correctly predicted all historical recession months in the training sample. Notably, the SVM and XGB models also perform well across different metrics.

The best-performing model in this experiment (i.e., RF) is a nonlinear ensemble model. The top panel of Exhibit 4 illustrates the three largest contributors to the RF model predictions, known as a *variable importance* plot. The bottom panel is a one-dimensional partial dependency plot (Friedman 2001) depicting the nonlinear marginal contribution by the top predictor in the RF model. Although we do not attempt a formal variable selection in this work given the vast literature on this topic, it is worth noting that the slope of the yield curve (YC) and total nonfarm payrolls (PAYEMS) are

Exhibit 2

Comparison of Model ROCs (training sample)

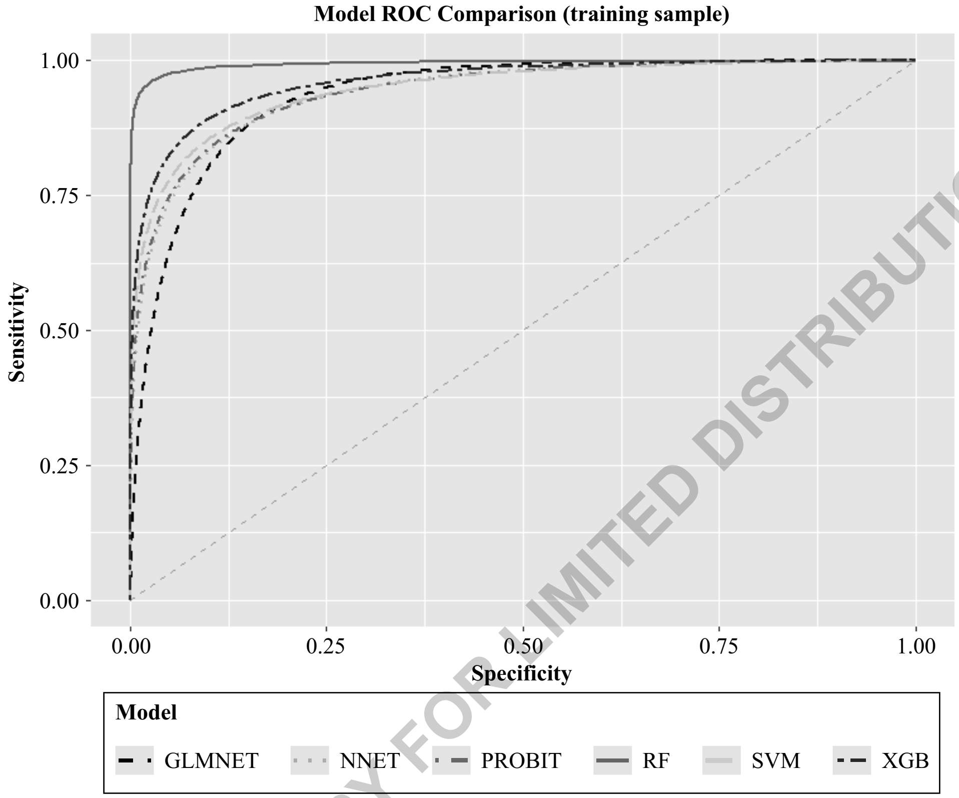


Exhibit 3

Performance of Different Classification Models (training sample)

Model	Accuracy	AUC	F-Score	H-Measure	KS	Precision	Sensitivity	Specificity
PROBIT	87%	89%	53%	64%	79%	53%	93%	86%
GLMNET	88%	89%	54%	64%	79%	54%	91%	87%
SVM	90%	90%	59%	67%	80%	59%	90%	90%
RF	91%	95%	61%	78%	89%	61%	100%	89%
XGB	89%	90%	58%	66%	79%	58%	90%	89%
NNET	86%	89%	52%	62%	78%	51%	93%	85%

the top two predictors across nearly all models in the training sample. The Industrial Production Index and unemployment rate are among the second most important predictors picked by the models.

Studying the performance of models only on training data may not be adequate for evaluation and is usually not recommended for model selection owing to the risk of overfitting. We now evaluate model performance using the test sample, spanning January 2007 to December 2019, for a more honest assessment of how these models may perform with previously unseen data. Exhibit 5 illustrates the test sample ROC curves for different models. Here, the level of similarities and overlap in the trajectories of the ROC curves makes it very difficult to know which model performed best on the test sample, although it is not hard to see that the NNET model (gray ROC curve) underperformed in this period.

EXHIBIT 4
Predictor Contributions for RF

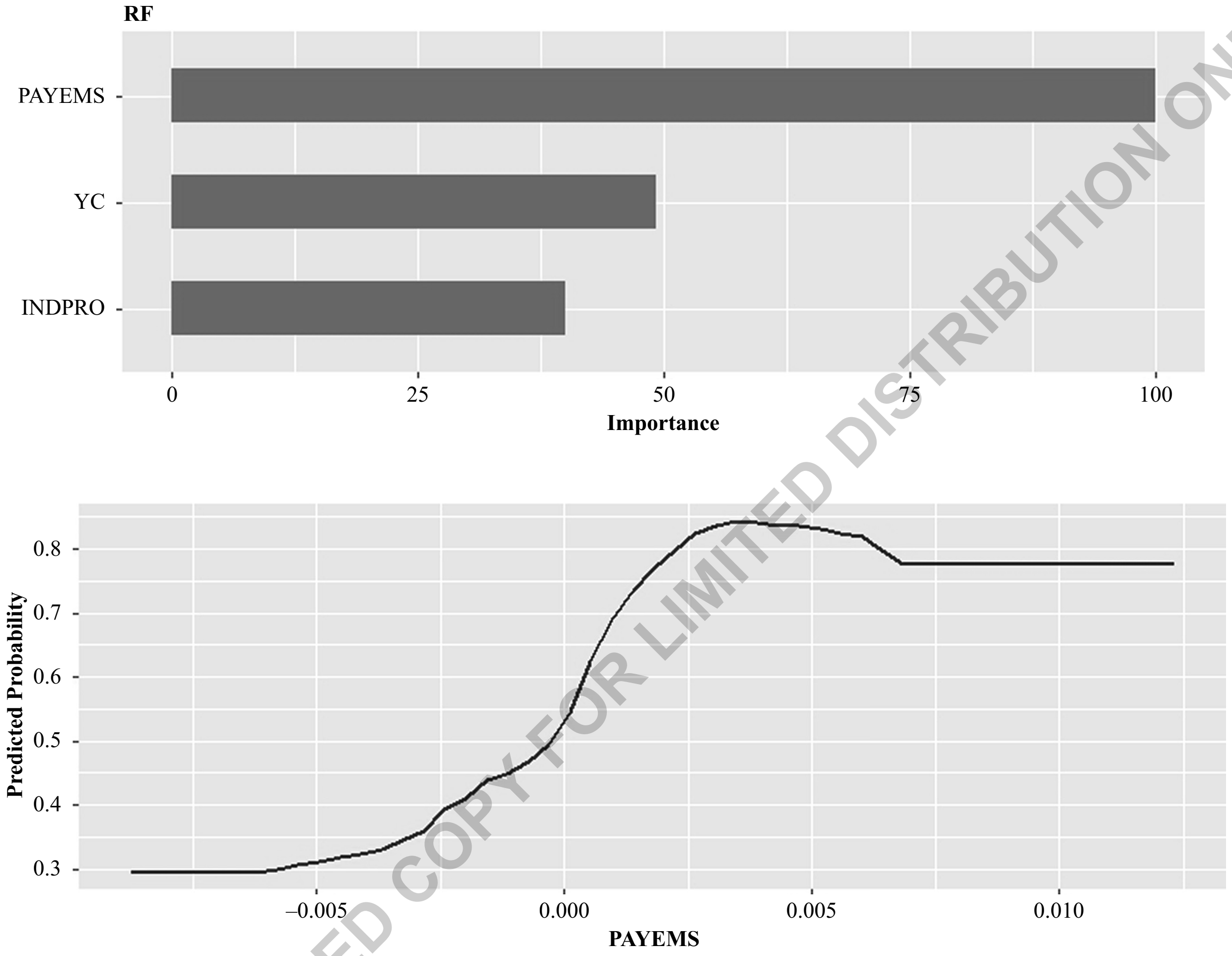


Exhibit 6 presents various metrics for evaluating model performance in the test sample. The two ensemble models of RF and XGB achieve the best performance overall, with RF coming in at the top once again. We may attribute this observation to the well-recognized property of ensemble models achieving lower variance (Hastie, Tibshirani, and Friedman 2008), resulting in learning well from training data and generalizing well to unseen test data.

A notable observation is a perfect sensitivity score by the GLMNET model and its overall decent performance across both the training and test periods. This is most likely an effect of the regularization processes employed by the algorithm, whereby a few strong predictors (e.g., unemployment rate, the slope of the yield curve, and total nonfarm payroll) with persistent signals are chosen for making predictions, albeit in a linear fashion. On the other hand, a rather simplistic NNET model seems to underperform in sample and outside of it. This indicates the need for developing a more complex architecture design (i.e., deeper network with larger size) and carrying out more extensive calibration (e.g., hyperparameter tuning) to achieve better performance that may be subject to a much higher cost of computing.

Next, we combine model predictions for training and test data (based on the same models trained only once on the training sample) and calculate performance metrics on a full sample, the summary of which is shown in Exhibit 7. Acknowledging the fact that some of these values may be influenced by the large sample size in our training data, they still show similar patterns, which clearly shows the advantage and predictive power of the

EXHIBIT 5

Comparison of Model ROCs (test sample)

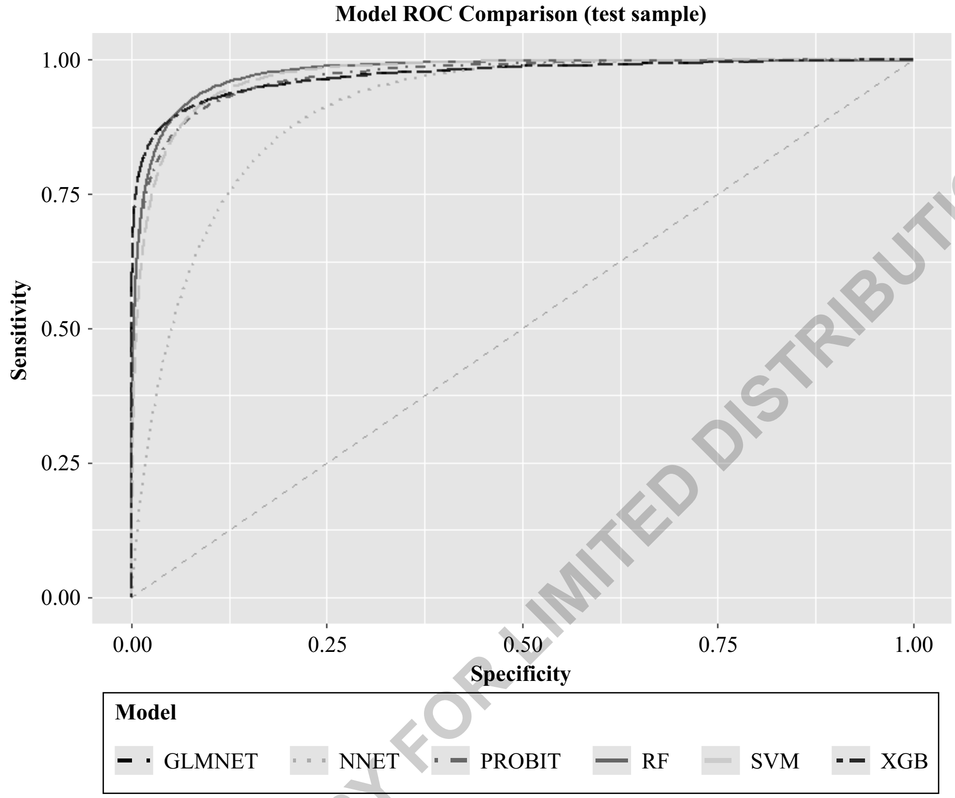


EXHIBIT 6

Performance of Different Classification Models (test sample)

Model	Accuracy	AUC	F-Score	H-Measure	KS	Precision	Sensitivity	Specificity
PROBIT	91%	90%	71%	70%	81%	59%	89%	91%
GLMNET	91%	95%	73%	79%	90%	58%	100%	90%
SVM	93%	91%	76%	74%	83%	65%	89%	93%
RF	94%	94%	80%	82%	89%	69%	95%	94%
XGB	94%	94%	78%	80%	88%	67%	95%	93%
NNET	92%	84%	68%	58%	68%	64%	74%	94%

ensemble methods. Here, the RF model is a distinct best performer, with XGB coming second, followed closely by SVM and others.

As we observed, nonlinear machine learning models seem to outperform the benchmark probit model. In addition, the general performance similarity among some models when measured by the basic accuracy metrics highlights the importance of evaluating and comparing different models using various alternative methods, including those more robust to class imbalances and misclassification cost. In particular, this includes metrics such as the F-score and H-measure that seem to better distinguish models across all three of the training, test, and full sample datasets.

Finally, we may have a look at the historical NBER recessions and compare those with the results predicted by the RF model (Exhibit 8). Historical recessions are shown as gray bars, and the RF estimated probability of

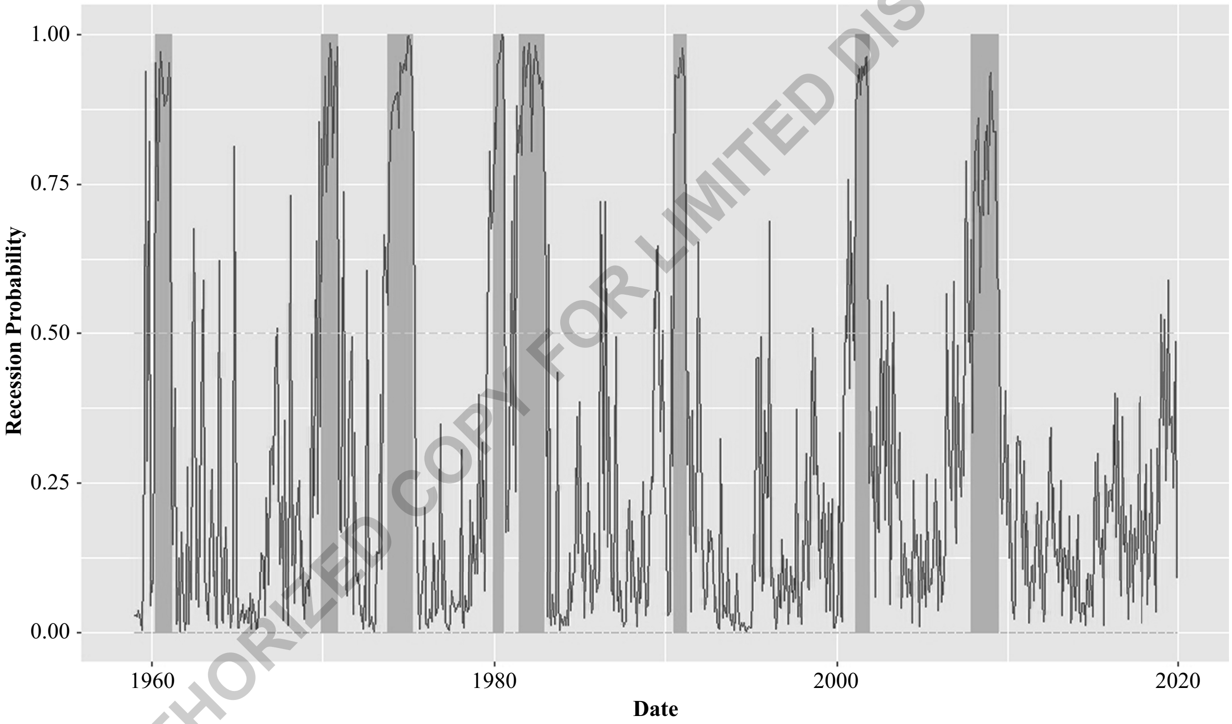
Exhibit 7

Performance of Different Classification Models (full sample)

Model	Accuracy	AUC	F-Score	H-Measure	KS	Precision	Sensitivity	Specificity
PROBIT	88%	90%	68%	65%	79%	54%	92%	87%
GLMNET	88%	90%	69%	67%	81%	55%	93%	88%
SVM	90%	90%	72%	69%	81%	60%	90%	90%
RF	92%	95%	76%	79%	89%	62%	99%	90%
XGB	90%	91%	72%	69%	81%	59%	91%	90%
NNET	88%	88%	66%	62%	76%	53%	89%	87%

Exhibit 8

Historical NBER Recessions (gray bars) and RF Predicted Probability (red line)



a given month being in recession is shown by a continuous red line. A predicted probability greater than 0.5 will be flagged as a month in recession by the RF model. There is a near perfect match between NBER historical recessions and the probabilities above 0.5 predicted by RF. This corresponds to the very high RF sensitivity score observed in Exhibits 3, 6, and 7.

In other words, each NBER recession month prior to 2007 was correctly predicted by RF, which corresponds to a sensitivity score of 100%. Post-2007, all but one NBER recession month were also correctly predicted by RF, corresponding to a sensitivity score of 95%. Examining the only false-negative prediction by RF in this period reveals that this instance corresponds to December 2007, which is the first month flagged as the NBER recession after a relatively long period of expansion preceding the 2008–2009 financial crisis—particularly challenging for models to predict.

On the other hand, RF shows some months as having elevated probability levels greater than 0.5 during training data that are not categorized as NBER recessions. This accounts for some false positives committed by the model and a slightly lower specificity score (i.e., 90% on full sample, 89% on training, and 94% on test sample). A fascinating observation is that RF predicts elevated probabilities during the last few months of 2019, most likely in response to the tightened financial and macroeconomic environment associated with the heightened trade wars and COVID-19 pandemic.

SUMMARY AND CONCLUSION

The problem of predicting US recessions has been studied extensively over the past several years, and a number of modeling methodologies and strong predictors of recessions have been documented. More recently, a few papers have emerged in the literature based on addressing this problem using machine learning algorithms. Building on past research, we examine this problem and argue that, due to the rarity of historical recessions, this is better addressed from an imbalanced classification standpoint. We examine a number of subsampling methods to adjust for the skewed class distributions and develop a machine learning framework for predicting recessions using different algorithms. We carefully look at model performances inside and outside of a training sample and use several classification accuracy metrics with varied properties and discrimination powers to better evaluate these models and gain greater confidence in their findings. Our observations reveal that ensemble machine learning methods exhibit superior predictive power in learning from history and making good predictions both within and outside of a training period. These models, in particular, draw upon the ability of nonlinear information processing and make highly accurate predictions of the recession months associated with the previously unseen financial crisis of 2008–2009. Moreover, these models estimate an elevated probability of recessions during the last few months of 2019, associated with a tightened macroeconomic environment and the COVID-19 pandemic. In addition, the studied machine learning models further confirm the importance of the top contributing variables identified by past research in predicting US recessions. In conclusion, we argue that using nonlinear machine learning models helps in better discovering various types of relationships in constantly changing financial data and enables deployment of flexible data-driven predictive modeling strategies.

REFERENCES

Atta-Mensah, J., and G. Tkacz. "Predicting Canadian Recessions Using Financial Variables: A Probit Approach." Working paper no. 98-5, Bank of Canada, 1998.

Batista, G., R. Prati, and M. Monard. 2004. "A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data." *ACM SIGKDD Explorations Newsletter* 6 (1): 20–29.

Blagus, R., and L. Lusa. 2013. "SMOTE for High-Dimensional Class-Imbalanced Data." *BMC Bioinformatics* 14: 106.

Breiman, L. 2001. "Random Forests." *Machine Learning.* 45 (1): 5–32.

Brownlee, J. "A Gentle Introduction to Imbalanced Classification." 2019, https://machinelearningmastery.com/what-is-imbalanced-classification/.

Camacho, M., G. Perez-Quiros, and P. Poncela. 2012. Markov-switching dynamic factor models in real time. CEPR Working Paper No. 8866.

Cao, H., X. Li, D. Yew-Kwong, and S.K. Ng. "SPO: Structure Preserving Oversampling for Imbalanced Time Series Classification." In *2011 IEEE 11th International Conference on Data Mining*, pp. 1008–1013. Washington, DC: IEEE, 2011.

Chauvet, M., and J. Piger. 2008. "A Comparison of the Real Time Performance of Business Cycle Dating Methods." *Journal of Business and Economic Statistics* 26: 42–49.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. "SMOTE: Synthetic Minority Over-Sampling Technique." *Journal of Artificial Intelligence Research* 16: 341–378.

Chen, T., and C. Guestrin. "XGBoost: A Scalable Tree Boosting System." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794. New York: Association for Computing Machinery, 2016.

Dixon, M., D. Klabjan, and L. Wei. 2017. "OSTSC: Over Sampling for Time Series Classification in R." *arXiv* 1711.09545.

Dueker, M. J. 1997. "Strengthening the Case for the Yield Curve as a Predictor of U.S. Recessions." *St. Louis Fed Review* (Mar): 41–51.

Estrella, A., and G. A. Hardouvelis. 1991. "The Term Structure as a Predictor of Real Economic Activity." *The Journal of Finance* 46 (2): 555–576.

Estrella, A., and F. S. Mishkin. 1996. "The Yield Curve as a Predictor of U.S. Recessions." *Current Issues in Economics and Finance* 2 (7): 1–6.

——. 1998. "Predicting U.S. Recessions: Financial Variables as Leading Indicators." *The Review of Economics and Statistics* 80 (1): 45–61.

Friedman, J., T. Hastie, and R. Tibshirani. 2008. "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software* 33, no. 1 (Feb): 1–22.

Friedman, J. H. 2001. "Greedy Function Approximation: A Gradient Boosting Machine." *Annals of Statistics* 29 (5): 1189–1232.

Giusto, A., and J. Piger. 2017. "Identifying Business Cycle Turning Points in Real Time with Vector Quantization." *International Journal of Forecasting* 33 (1): 174–184.

Hand, D. J. 2009. "Measuring Classifier Performance: A Coherent Alternative to the Area under the ROC Curve." *Machine Learning* 77: 103–123.

Hastie, T., R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning, Data Mining, Inference, and Prediction*, 2nd ed. New York: Springer, 2008.

Hulse, J. V., T. M. Khoshgoftaar, and A. Napolitano. "Experimental Perspectives on Learning from Imbalanced Data." In *Proceedings of the 24th International Conference on Machine Learning*, pp. 935–942. Corvallis, Oregon: Oregon State University, 2007.

Hyndman, R. J., and G. Athanasopoulos. *Forecasting: Principles and Practice*, 2nd ed. Melbourne, Australia: OTexts, 2018.

James, A., Y. S. Abu-Mostafa, and X. Qiao. 2019. "Machine Learning for Recession Prediction and Dynamic Asset Allocation." *The Journal of Financial Data Science* 1 (3): 41–56.

Kuhn, M., and K. Johnson. *Applied Predictive Modeling*. New York: Springer, 2013.

Levanon, G., J.C. Manini, A. Ozyildirim, B. Schaitkin, and J. Tanchua. "Using a Leading Credit Index to Predict Turning Points in the U.S. Business Cycle." Working paper 11-05, The Conference Board, Economics Program, 2011.

Liu, W., and E. Moench. "What Predicts U.S. Recessions?" Staff report no. 691, Federal Reserve Bank of New York, 2014.

Ng, S. 2014. "Viewpoint: Boosting Recessions." *Canadian Journal of Economics* 47 (1): 1–34.

Qi, M. 2001. "Predicting US Recessions with Leading Indicators via Neural Network Models." *International Journal of Forecasting* 17: 383–401.

Sephton, P. 2001. "Forecasting Recessions: Can We Do Better on MARS?" *Federal Reserve Bank of St. Louis* 83 (2): 39–50.

Simonian, J. 2020. "Modular Machine Learning for Model Validation: An Application to the Fundamental Law of Active Management." *The Journal of Financial Data Science* 2 (2): 41–50.

Vapnik, V. *The Nature of Statistical Learnings Theory*. New York: Springer, 1996.

Wallace, B., K. Small, C. Brodley, and T. Trikalinos. "Class Imbalance, Redux." In *2011 IEEE 11th International Conference on Data Mining (ICDM)*, pp. 754–763. Washington, DC: IEEE, 2011.

Wright, J. H. "The Yield Curve and Predicting Recessions." Discussion paper, Board of Governors of the Federal Reserve System, 2006.

Yatchew, A., and Z. Griliches. 1985. "Specification Error in Probit Models." *The Review of Economics and Statistics* 67 (1): 134–139.

To order reprints of this article, please contact David Rowe at d.rowe@pageantmedia.com or 646-891-2157.

ADDITIONAL READING

Machine Learning for Recession Prediction and Dynamic Asset Allocation

Alexander James, Yaser S. Abu-Mostafa, and Xiao Qiao
The Journal of Financial Data Science
https://jfds.pm-research.com/content/1/3/41

ABSTRACT: *The authors introduce a novel application of support vector machines (SVM), an important machine learning algorithm,*

to determine the beginning and end of recessions in real time. Nowcasting, forecasting a condition in the present time because the full information will not be available until later, is key for recessions, which are only determined months after the fact. The authors show that SVM has excellent predictive performance for this task, capturing all six recessions from 1973 to 2018 and providing the signal with minimal delay. The authors take advantage of the timeliness of SVM signals to test dynamic asset allocation between stocks and bonds. A dynamic risk budgeting approach using SVM outputs appears superior to an equal-risk contribution portfolio, improving the average returns by 85 bps per annum without increased tail risk.

Modular Machine Learning for Model Validation: *An Application to the Fundamental Law of Active Management*

Joseph Simonian

The Journal of Financial Data Science

https://jfds.pm-research.com/content/2/2/41

ABSTRACT: *The author introduces a modular machine learning framework for model validation in which the output from one procedure serves as the input to another procedure within a single validation framework. A defining feature of the described methodology is the use of both traditional econometrics and data science. The author uses an econometric model in the first module to classify data in an economically intuitive way. Proceeding modules apply data science techniques to evaluate the predictive characteristics of the model components. The author applies his framework to the fundamental law of active management, a well-known formal characterization of portfolio managers' alpha generation process. In contrast to standard applications of the law, in which it has been used to evaluate a manager's existing active management process, the author recasts the law within his framework as a means to test investment signals for potential use, individually or collectively, in a manager's investment process. To illustrate how this application works, the author provides an example using the well-known Fama–French factors as test signals.*

Derivation of a Dynamic Market Risk Signal Using Kernel PCA and Machine Learning

Alireza Yazdani

The Journal of Financial Data Science

https://jfds.pm-research.com/content/2/3/73

ABSTRACT: *Kernel principal component analysis (PCA) is an extension of the conventional PCA method that employs a kernel transformation whereby hidden patterns in possibly multidimensional data may be detected and extracted more explicitly. In this article, the author applies the method of kernel PCA to a currency prediction case study and derives an aggregate market signal. It is observed that this signal has desirable information-compression properties and may be used as a predictive risk indicator in the return prediction models. Used alongside common drivers of exchange rates, a kernel PCA signal enhances in-sample and out-of-sample risk-adjusted performance across a range of machine learning strategies. In particular, the author observes that a kernel PCA signal remains robust and predictive during volatile market conditions. The kernel PCA signal may be used as a machine learning feature to inform and support data-driven risk management strategies.*