HIGH SCHOOL EDITION
Journal of Student Research
Volume 11 Issue 2 (2022)

A Meta-Analysis Evaluating the Performance of Machine Learning Models on Probability of Loan Default

Ely Hahami[1] and Doug Piper[#]

[1]The Lawrenceville School, Lawrenceville, NJ, USA
[#]Advisor

ABSTRACT

There has been a recent increase in the implementation of machine learning algorithms to predict the credit risk of prospective loan applicants. This meta-analysis aims to contribute to the small but growing research on the effects of algorithmic lending. Specifically, we compare the performance of the Logistic Regression (LR) model and Random Forest (RF) model in predicting loan default (PD). Using the area under the receiver operating characteristic curve as a measure of aggregate machine learning model performance, we ultimately find convincing evidence that the RF model is more accurate than the logit model in PD (p-value=0.029, $\alpha = 0.01$). These results have major implications for banks and financial firms as mortgage lending transitions into the FinTech era.

Review of Literature

There is a growing amount of research analyzing machine learning (ML) models on the probability of default (PD). Fuster et. al (2021) build and embed ML techniques in an equilibrium model to analyze both extensive margin (exclusion) and intensive margin (rates) impacts on both PD and disparate impact on minority borrowers. In another study, Zhang (2015) models PD using survival analysis, probability of density curves (pdf), and hazard curves, ultimately finding some strength in using the logit model. Lastly, Zhu et. al (2019) hone in on the RF algorithm and employ the SMOTE method — an oversampling technique that generates synthetic samples from the minority class — to predict PD through real-world user loan data on Lending Club.

Conceptually Understanding Machine Learning in the Context of Lending

Intuitively, mortgage lenders want to make their businesses money — and thus seek to minimize risk when choosing a prospective borrower. The primary focus of determining this credit risk is to predict if a customer will default on his or her mortgage loan in the future — that is, when a borrower fails to pay back a debt according to the initial arrangement. This probability of default is generally estimated through credit bureau data and other borrower characteristics. For a loan origination, a bank generally sets a cut-off threshold and approves a credit to those customers that have the predicted PD less than the predefined threshold. In a general sense, PD is not only important for effective risk and capital management, but also for the pricing of credit assets, bonds, and more sophisticated instruments such as derivatives.

Naturally, prospective borrowers with high default probabilities in the screening stage would be more likely to be denied a loan from a lender. There exists a function $\hat{y}$ — an unbiased point estimate for y that maps borrowers' observable characteristics (such as FICO score or other creditworthiness factors) into some vector x, and consequently

yields the true probability of default. Mathematically, according to Fuster et. al (2021), this relationship can be represented by $p\hat{}(x)= \hat{y}$ for some predictive function $p\hat{}(x)$.

Traditional approaches involve logit models — a binomial regression model — which assume linear values of the vector x and are defined with the following link function:

$\log (g(x)/(1-g(x)) = x'\beta$.

However, one key conceptual difference between this traditional model and more novel machine learning models that are extensively used in classification problems is that these ML models can employ a wider range of functions for $\hat{y}$. Lenders typically use tree-based models and neural networks, which not only minimize statistical loss — namely, the Mean-Squared Error (MSE) — but also assume non-linear values of the vector x.

Intuitively, improvements in statistical technology generally track the true y value more closely. In an example provided by Fuster et. al. (2021), default probabilities with newer technology result in a convex quadratic function for the input x rather than the traditional linear outcome. Consequently, in this meta-analysis, we rely on the machine learning models that track the x vector in a non-linear fashion, and seek to statistically test their accuracy in PD.

Defining and Explaining The Random Forest Model

For this paper, we focus on the Random Forest (RF) model. Based on the use of simple decision trees, RF is a non-linear, non-parametric classification algorithm that bins covariates in the vector of observable lender characteristics, x, to best predict default probability. An RF algorithm works as follows. First, it takes a covariate that is included in the vector x, then searches for a single value that separates this covariate into defaulters and non defaulters. This process is repeated recursively for every "leaf" of the primitive tree. RF also uses K-fold cross-validation — that is, randomly splitting the sample into K subsamples, thus facilitating ideal fit of the data.

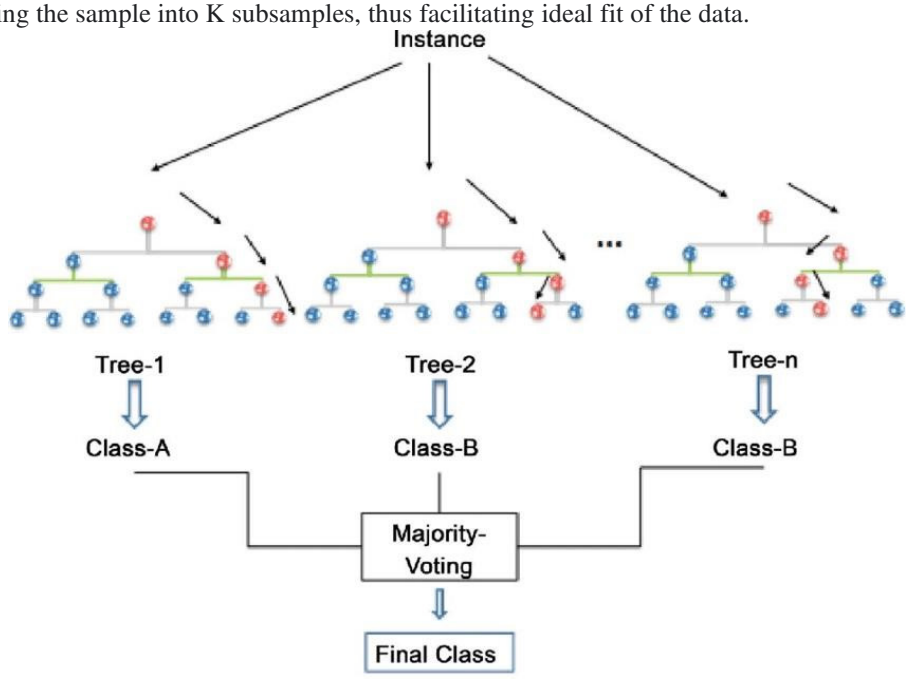


Figure 1: Demonstration of the random forest methodology.

Introducing ROC and AUC in the Context of Confusion Matrix Theory

Confusion Matrix Theory gives us insight not only into the types of errors being made by a classifier. The performance criterion of the appropriate machine learning classifier in PD is largely based on ROC and AUC. A receiver operating

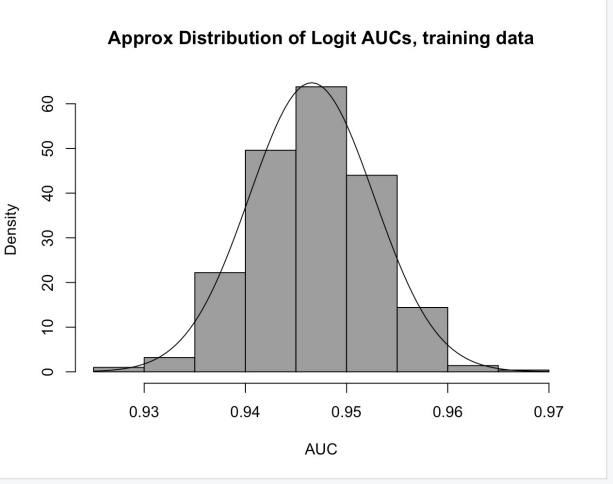
characteristic (ROC) curve is a graph showing the performance of a classification model at all classification thresholds. This plots two parameters: specificity, or false positive rate, and sensitivity, or true positive rate. We define True Positive Rate (TPR) as (TP)/(TP+FN), where FN denotes a false negative, and denote False Positive Rate (FPR) as FP/(FP+TN). AUC, which will be of primary focus, measures the entire two-dimensional area underneath the entire ROC curve (ie. integral calculus) from (0,0) to (1,1). Consequently, the ideal ROC curve hugs the upper left corner of the chart and has an AUC close to 1. Mathematically, if the points on the ROC curve are defined as (x_1, y_1), (x_2, y_2), (x_3, y_3), AUC can be estimated as:

$$\tfrac{1}{2} \Sigma (x_{i+1} - x_i)*(y_i+y_{i+1}).$$

Overall, AUC is desirable because it is classification-threshold-invariant and scale-invariant.

A Comparison of RF and Logit in PD

Below gives the approximately normal distribution of AUCs using the logit model for 1000 simulated runs, coded in R. According to Zhang (2015), The mean AUC of the training data — the initial dataset you use to teach a machine learning application to perform criteria — was 0.9466 with a standard deviation of 0.0061. The mean AUC for the testing data (sometimes called validation data) was 0.8795 with a standard deviation of 0.0388. These AUCs were based on 20,918 individual observations from loans originated in 2004, with information about these loans randomly collected monthly from January 2005 to May 2010.



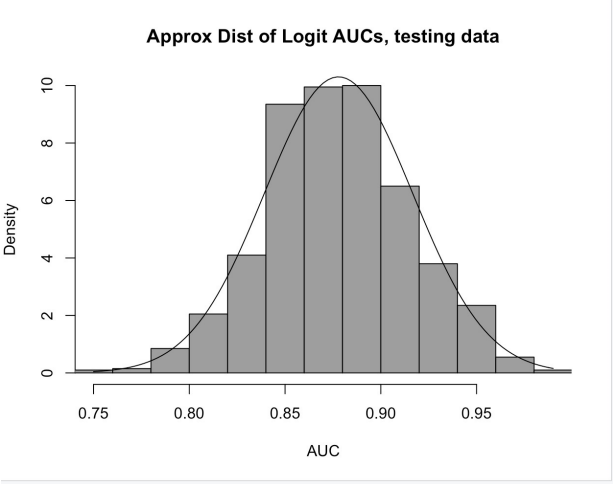


Figure 2: Histogram with inserted normal distribution curve for Logit Model training data AUCs.

Figure 3: Histogram with inserted normal distribution curve for Logit Model testing data AUCs.

We combine the training and testing ML data simulations to get $\bar{x}$, the sample mean AUC of the logit model, and S_x the sample standard deviation of logit AUCs. It is important to note that we add the variances of the training and testing data set, and then take the square root, in order to get a combined sample standard deviation. Performing these calculations, we obtain $\bar{x}_{logit} = 0.91305$ and $S_{logirt} = 0.019638$. Below is the combined histogram:

$$E(X+Y)=E(X) + E(Y)$$
$$\sigma(x+y) = \sqrt{(\sigma^2x + \sigma^2y)}$$

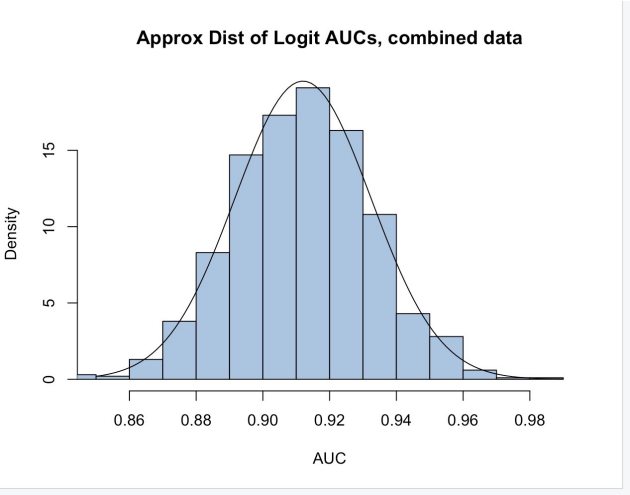


Figure 4: Histogram with inserted normal distribution curve for Logit Model, combined data AUCs.

According to Zhu et. al (2015), the AUC in the Random Forest Regression — which contains >115,000 observations from the Lending Club for the first quarter of 2019 — has a sample mean ($\bar{x}$) of 0.983 and we calculate the sample standard deviation (σ) to be 0.02. We thus have two representative samples — one from a logit model, and one from an RF model.

Checking the Conditions of a 2-Sample t Significance Test

Given our $\bar{x}_{logit}$ and $\bar{x}_{rf}$, we seek to determine if there is a difference in true average AUC between the two machine learning models. We thus use a 2-sample significance test with the t test statistic (based on the t distribution), and set a conservative significance level of $\alpha = 0.01$. In terms of the randomness conditions, we observe that, according to Zhang (2015), the data that accounts for the individual observations were taken *randomly* every year for each mortgage loan. The 2019 mortgage lending club data is also random. There is no need to test the 10% rule — the notion that sample sizes should be no more than 10% of the population — because we are *not* sampling without replacement from a finite population. Moreover, in the Zhang (2015) data, we need not check the Central Limit Theorem — which states that the distribution of sample means approximates a normal distribution as the sample size gets larger regardless of the population's distribution — because the sampling distribution is already normally distributed. While the 2019 lending club data has n=4, there are no glaring outliers or patterns that yield a non-normal sampling distribution. Still, we must proceed with caution because we are not sure this sampling distribution normality exists. We do know, however, that there is independence between and within groups. Thus, taking into account these condition checks, we consider the following null and alternative hypotheses:

H_0: $\mu_{RF} = \mu_{logit}$

H_a: $\mu_{RF} > \mu_{logit}$

Results

We input $\bar{x}_{logit}$, $\bar{x}_{RF}$, S_{logit}, S_{RF}, n_{logit}, and n_{RF} and use a calculation software to determine the following results:

2-SampTTest

$\mu_{RF} > \mu_{logit}$

p-value = 0.0029410505
degrees of freedom (df) = 3.023182546

Recall the p-value denotes the probability of getting a statistic as extreme or more extreme than the one from our sample, given the null hypothesis is true. In other words, it is a type of conditional probability. Proceeding with caution because of the lack of assurance of a normally distributed lender club sampling distribution, we can interpret our results: since our p-value (~0.0029) is less than our conservative significance level ($\alpha = 0.01$), we reject the null hypothesis H_0. There is convincing evidence that the random forest algorithm's true average ROC AUC exceeds that of the logit model. However, it is also important to note that while AUC does provide a relatively thorough aggregate measure of performance, it is not only the criteria measure. Alternatives and complements of ML performance include the Gini Index, an internal tree node, and Brier Score, among others. It is also important to note that while the Logit model estimates probability given the vector of characteristics x, the RF model provides a binary classification given a set of covariates x.

Conclusion and Discussion of Future Research

In this meta-analysis, we have statistically compared the RF and logit models and have found that the RF model performs better in terms of PD. These findings imply that major banks and mortgage lending firms perhaps should institute the RF model over the traditional logit model if they seek a more accurate predictor in this category. However, it is important to note that future research is direly needed. At the academic level, it is important to research additional ML algorithms — such as Gaussian Naïve Bayes (NB), K-nearest neighbors classification (KNN), and Classification and Regression Tree (CART) — and test their accuracy in PD. These accuracies should then be compared to both the Logit and RF models, and these findings perhaps could be generalized to other potential credit markets.

References

Agrawal, A., J. Gans & Goldfarb, 2018, Prediction machines: the simple economics of artificial intelligence (Boston, MA: Harvard BusinessBusiness Review Press).

Chen, W., & Samuelson, F. W. (2014). The average receiver operating characteristic curve in multireader multicase imaging studies. The British Journal of Radiology, 87(1040), 20140016. https://doi.org/10.1259/bjr.20140016

Classification: Roc curve and auc | machine learning crash course. (n.d.). Google Developers. Retrieved March 21, 2022, from https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

Fuster, A., Goldsmith, Pinkham, P., Ramadorai, T., & Walther, A. (2022). Predictably unequal? The effects of machine learning on credit markets. The Journal of Finance, 77(1), 5–47. https://doi.org/10.1111/jofi.13090

Lee, M. S. A., & Floridi, L. (2021). Algorithmic fairness in mortgage lending: From absolute conditions to relational trade-offs. Minds and Machines, 31(1), 165–191. https://doi.org/10.1007/s11023-020-09529-4

Steil, J. P., Albright, L., Rugh, J. S., & Massey, D. S. (2018). The social structure of mortgage discrimination. Housing Studies, 33(5), 759–776. https://doi.org/10.1080/02673037.2017.1390076

Zhang, Q. (2015). Modeling the probability of mortgage default via logistic regression and survival analysis. Open Access Master's Theses. https://doi.org/10.23860/thesis-zhang-qingfen-2015

Zhu, L., Qiu, D., Ergu, D., Ying, C., & Liu, K. (2019). A study on predicting loan default based on the random forest algorithm. Procedia Computer Science, 162, 503–513. https://doi.org/10.1016/j.procs.2019.12.017