

Proceeding Paper

Macroeconomic Adverse Selection in Machine Learning Models of Credit Risk †

Joseph L. Breeden *[‡]  and Yevgeniya Leonova †

Deep Future Analytics LLC, Santa Fe, NM 87505, USA; leonova@deepfutureanalytics.com

* Correspondence: breeden@deepfutureanalytics.com

† Presented at the 9th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 12–14 July 2023.

‡ These authors contributed equally to this work.

Abstract: Macroeconomic adverse selection is computed as a time series of forecast residuals via the vintage origination model for an industry dataset of auto loans. The adverse selection time series are computed separately as model residuals using logistic regression, neural networks, and stochastic gradient boosted trees to predict defaults in the first 24 months of a loan. Panel data versions of these models with lifecycle and environment inputs from a segmented Age-Period-Cohort analysis were also estimated. The estimates show that panel data methods make better use of available data to provide faster estimates of adverse selection risk in recent vintages and incorporate defaults at any age of the loan. The nonlinear modeling advantages of neural networks and stochastic gradient boosted trees did not significantly alter the estimates of adverse selection. Overall, all methods confirmed that macroeconomic adverse selection was dramatically higher in 2021 and 2022 for US auto loan originations.

Keywords: adverse selection; credit scoring; survival models; neural networks; stochastic gradient boosted trees



Citation: Breeden, J.; Leonova, Y. Macroeconomic Adverse Selection in Machine Learning Models of Credit Risk. *Eng. Proc.* **2023**, *39*, 95. <https://doi.org/10.3390/engproc2023039095>

Academic Editors: Ignacio Rojas, Hector Pomares, Luis Javier Herrera, Fernando Rojas and Olga Valenzuela

Published: 24 July 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The COVID-19 pandemic brought rapid, dramatic swings in economic conditions and consumer behavior. Monitoring of credit quality has shown deterioration in many loan categories. In auto lending, credit quality deterioration appears to have begun in the second quarter of 2021 and extended at least through the end of 2022. When normalized for changes in the credit quality of borrowers using logistic regression models, the residual credit risk appears to correlate to the rapid rise in new and used car prices and the rise in auto loan interest rates. This suggests a period of macroeconomic adverse selection similar to what was observed between 2006 and 2009.

Since the 2009 mortgage crisis, the lending industry has widely adopted new methods from machine learning and artificial intelligence in lending. Adoption for credit risk assessment and underwriting has been slower than other industries because of regulatory demands, but research and experimentation are extensive [1] and deployment will continue to grow. Given that machine learning has greater flexibility for finding nonlinear patterns, some proponents have suggested that such methods may be able to incorporate the structure that is showing up as macroeconomic adverse selection in regression-based models.

The current research analyzes auto loan data for originations from 2002 through 2022. Origination scores predicting the likelihood of being 60+ days past due (DPD) are estimated using logistic regression, discrete time survival models, stochastic gradient boosted trees (SGBT), SGBT with Age-Period-Cohort (APC) inputs, neural networks, and neural networks with APC inputs. For each model, time series of the residual errors are estimated by origination (vintage) month and compared across models. This research is the first to compare the estimation of adverse selection time series by vintage across regression

and machine learning methods. Although some lenders quantify adverse selection across product types on internal data, our study is the first to publish a history of macroeconomic adverse selection for auto loans on a broad industry dataset.

This study finds clear advantages to panel data methods with APC inputs over traditional scoring methods for rapidly estimating adverse selection within a portfolio. This advantage is both in better use of the available data and having a stronger baseline versus age of the loan and calendar date against which to compare.

Section 2 provides an overview of the literature. Section 3.1 describes the available data. Section 3.2 provides brief descriptions of the modeling techniques used. Section 4 provides the results.

2. Literature Review

Adverse selection broadly means that the credit quality of borrowers is not what was expected when the loans were written according to the loan-origination model employed. The earliest work on adverse selection [2] described how this can happen when lenders compete on loan pricing and terms. Given a choice, borrowers will apply for the loan with better terms first. The borrowers rejected by the bank with better pricing then apply to the lender with less attractive pricing. The bank with higher pricing may have expected higher yields, but by attracting only the riskier borrowers, the losses could be higher and yields actually fall.

Adverse selection through competitive pressure can be described as microeconomic adverse selection. Sometimes trends of adverse selection are apparent across the industry. Dubbed macroeconomic adverse selection [3], this has been found to correlate to changes in the cost of the goods being purchased and changes in the cost of borrowing [4]. The theory is that depending upon macroeconomic conditions, the pool of borrowers may shift in ways not observable from the data typically available on the borrowers.

The work by Breeden in 2011 [3] observed this through three credit cycles between 1990 and 2006 for mortgages. Breeden and Canals-Cerda in 2016 [4] performed a detailed analysis of credit quality before and through the 2009 mortgage crisis to adjust for all loan-application information and found that half of the credit quality deterioration could not be explained by poor underwriting. Instead, it appeared to correlate to the cost of homes and mortgage interest rates. This effect can be observed in all product categories. Calem, Canon, and Nakamura (2011) [5] related adverse selection in home equity lines of credit to county-level unemployment and consumer confidence.

3. Materials and Methods

3.1. Data

Auto loan performance data from 26 lenders was modeled in order to assess adverse selection trends. This dataset included 1,244,651 loans originated between January 2005 and December 2022. Typical origination variables were used, including Bureau Score, Loan-To-Value (LTV) ratio, Debt-To-Income (DTI) ratio, channel (direct or indirect), collateral type (new or used), term, state, and anonymized lender ID.

Behavioral variables such as delinquency that are included in a typical behavior score are a post-origination attempt to adapt to the difference between origination expectation and post-origination reality. For that reason, including delinquency in our models would dilute the measure of adverse selection. However, some information is available immediately after the loans are originated that would not be included in a traditional origination score. Primary among these are the offered interest rate on the loan (APR), and the balance of the loan. The annual percentage rate (APR) offered by the lender may incorporate information not made available for creation of the origination scores, such as adjustments for specific dealers, form of employment, broader relationship with the lender, etc. Therefore, we created a measure $\Delta APR(i, v, L)$ that computed the difference between the APR on a specific loan within a specific vintage $APR(i, v, L)$ and the average APR for that vintage by

lender $\overline{APR}(v, L)$ where v is the vintage and L is the lender in order to incorporate some of this missing knowledge.

$$\Delta APR(i, v, L) = APR(i, v, L) - \overline{APR}(v, L) \tag{1}$$

To make measures of adverse selection useful to portfolio managers, the estimation needs to occur as early as possible in the life of a vintage. Therefore, default D has been defined as the date on which an account first becomes ≥ 60 days past due (DPD). When logistic regression (LR), neural networks (NN), or stochastic gradient boosted trees (SGBT) were used to create a traditional origination credit score, the outcome period for default was the first 24 months of the life of the loan. A total of 2428 such defaults exist within the dataset, a 24-month default rate of 0.2%.

When using models that are normalized to a lifecycle or hazard function, defaults at any loan age are relevant. The total number of defaults through the entire life of the loans was 55,435, a 4.45% lifetime default rate. The panel data approach thus incorporates defaults from vintages that have not yet reached 24 months old and considers additional defaults later in the lifetime of the loans, which will be important to interpreting the results obtained later.

3.2. Algorithms

Several modeling techniques were compared in order to determine the robustness of estimations of macroeconomic adverse selection as driven by external forces rather than simply estimation noise.

3.2.1. Logistic Regression

Logistic regression is the traditional method for creating origination scores. For a thorough introduction to credit scoring, see Thomas, Crook, and Edelman [6] or Anderson [7]. The cumulative probability of default for the first 24 months of the loan is predicted as

$$\text{logit}(D_i) \sim \sum_{j=1}^n c_j s_{ij} + c_0 \tag{2}$$

where the c_j are the n estimated coefficients for the scoring factors s_{ij} for account i . The set of scoring factors s are chosen to optimize the Akaike Information Criterion (AIC). Some of the variables are binned to capture nonlinearities.

After the model has been created, the residual error by vintage is estimated by creating a second regression with the original model $M(c, s_i)$ as a fixed input.

$$\text{logit}(D_i) \sim M(c, s_i) + \sum_v g_v \delta_v \tag{3}$$

where δ_v is a delta function for vintage date v and g_v is the corresponding coefficient. Fixed effects by vintage (dummy variables) could have been included in the original regression, Equation (2). In some situations, this can shift some of the explanatory power of the scoring factors to the vintage fixed effect. Since the goal here is to obtain maximum explanatory power from the scoring factors and use the vintage effects only to measure the residuals, a two-step process was employed.

3.2.2. Age-Period-Cohort Models

Age-Period-Cohort (APC) models [8–10] for vintage analysis explain the risk of default at each observation period as a combination of functions of the age a of the loan, the calendar date t , and the vintage date v . These functions can be spline approximations, non-parametric, or other forms, but are generally not tied to specific scoring factors.

Because $a = t - v$, a model-specification error exists if no constraints are imposed. In applications to credit risk analysis, the following representation is common.

$$D \sim b_0 + b_1a + F'(a) + b_2v + G'(v) + H'(t) \quad (4)$$

where b_0 is the intercept, b_1 and b_2 are the linear coefficients for a and v , and $F'(a)$, $G'(v)$, and $H'(t)$ are the nonlinear functions that have zero mean and no linear component. For explanation, these are usually combined as $F(a) = b_0 + b_1a + F'(a)$, $G(v) = b_2v + G'(v)$, $H(t) = H'(t)$ where $F(a)$ is called the lifecycle measuring the timing of losses through the life of the loan, $G(v)$ is the vintage function measuring credit risk by vintage, and $H(t)$ is the environment function measuring the net impact from the environment (primarily economic conditions). The primary advantage of APC models is the ability to separate these effects, so the credit risk function captures the full amount of credit quality variation, but cleaned of impacts from the macroeconomic environment and normalized for differences in the age of the loans. The credit risk function does not adjust for loan-level changes in underwriting, so it is not a perfect measure of adverse selection. However, if the analysis is segmented by key measures such as bureau score and term, a net residual credit risk function can be extracted that can serve as an adverse selection measure.

3.2.3. Discrete Time Survival Models

Discrete time survival models [11,12] are a form of panel regression where each account is observed each month to predict default/no default. The regression equation can include nonparametric lifecycle and environment functions as in APC models and scoring factors as in logistic regression. Cox proportional hazards [13] models are the original continuous time formulation of this, where the APC-style lifecycle is a discrete time version of the Cox PH hazard function.

Previous work has shown that survival models that are estimated via a partial likelihood estimation as with Cox PH or a logistic regression estimation of the full panel model have instabilities in the context of credit risk modeling [14]. The instability occurs, in part, because the model-specification error of the APC model appears as colinearity between the scoring factors, lifecycle, and environmental factors of the survival model.

Breeden [15] proposed a solution to this where an initial APC decomposition is performed as described above and the lifecycle and environment functions are taken as fixed inputs to a second panel logistic regression estimation.

$$D \sim F(a) + H(t) + \sum_{j=1}^n c_j s_{ij} + c_0 \quad (5)$$

This two-step process resolves any colinearities between scoring factors, lifecycle, and environment while retaining maximum explanatory power for the scoring factors. In practice, this has proven to create scores that are more stable through changes in the environment while retaining account-level predictive accuracy.

Adverse selection is measured in a final step as described in Equation (3) except as a panel logistic regression.

3.2.4. Artificial Neural Network

Using artificial neural networks (NN) for credit risk forecasting has been the subject of numerous publications [16–18]. The problem design is similar to creating a logistic regression credit score, but with the network allowing for nonlinearity and interaction effects that would need to be discovered manually and encoded into the inputs of a regression model.

The available training data for auto loan defaults is not particularly complex compared to many applications of neural networks, and thus is not a showcase for the nonlinear wonders of machine learning. However, it is sufficient to address the question of whether

adverse selection is a model-specific error or due to a hidden variable that is not discoverable by any algorithm using only traditional data.

The neural network architecture was correspondingly simple. The network had an input layer, five fully connected layers with softplus activation functions, and a sigmoid output node. Softplus is less efficient and some argue less interpretable than ReLU activation functions, but it had better convergence performance in this context. The target was the same binary indicator of default within 24 months as used in the logistic regression model with a binary cross-entropy loss function.

Neural networks such as this do not function well when defaults comprise only 0.2% of the training data. Previous research has shown that at least a 4:1 or 3:1 ratio is needed for proper network estimation [19,20]. In this case, all default accounts were included and four times as many non-default accounts were randomly sampled from the dataset. The resulting network predictions need to be balanced back in order to match the overall default probability of the original training dataset.

3.2.5. NN + APC

Within the domain of credit risk modeling, having data from 2005 through 2022 is considered a significant amount of history. Compared to economic cycles, it is not. One problem with neural networks or any scoring technique with a wide (24 month) outcome period is that fragments of an economic cycle get confused with scoring attributes. The primary theoretical advantage of discrete time survival models over logistic regression is creating a distinction between environmental trends and credit risk trends that are explainable from scoring factors.

Analogous to the discrete time survival models, the lifecycle and environment from the APC models can be provided as inputs to the neural network with the data arranged as a panel of repeated observations for the accounts until default or payoff [21]. The network architecture is arranged such that the APC inputs $O(a, t) = F(a) + H(t)$ in units of log-odds of default are passed to the final node as an offset without modification. The neural network is used only as a replacement for the credit risk component, effectively modeling the account-level residuals around the long-term trends of lifecycle and environment.

For proper estimation, the dataset still requires balancing. The input offset needs to be adjusted with an additive constant for any change in default probabilities due to rebalancing. The revised offset, O' , is

$$O' = O(a, t) + \left(\log\left(\frac{\bar{p}}{1 - \bar{p}}\right) - \bar{O} \right) \quad (6)$$

When the network produces forecasts, the original offset $O(a, t)$ is used without the rebalancing adjustment factor. As with the plain NN, the final dataset for model estimation under-sampled the loans that never default in order to achieve a 4:1 ratio with loans that eventually will default. Model training was performed on 80% of this balanced dataset and cross-validation on 20% to determine the stopping point.

The part of the network dedicated to processing the origination factors can have the same architecture as that used without the APC inputs. However, providing APC inputs often allows for a simpler network architecture. The target variable for the network was default that occurs at any point in the life of the loan, as done in the DTSM, allowing the larger panel dataset to be modeled.

3.2.6. Stochastic Gradient Boosted Trees

Decision trees are as old as credit risk modeling [22,23]. The multidimensional space described by the scoring attributes is split with hyperplanes to separate good from bad accounts. A slightly more sophisticated version fits a regression model within each terminal node of the tree, as in CART [24]. Stochastic gradient boosted trees [25,26] are essentially an ensemble modeling approach where each new regression tree is weighted to explain the

data points that were less explainable by the preceding set of regression trees. Trees are added until no significant improvement is obtained on a test set.

Tree-based methods do not suffer from multicollinearity problems as regression does, so additional inputs can be provided without destabilizing the model. Therefore, the SGB Tree was provided with all of the inputs given to the logistic regression and neural network models as well as factor variables for state and lender. These additional inputs might allow the algorithm to better handle outliers. The target variable is again whether an account defaults within the first 24 months, as used in the logistic regression and neural network origination scoring models. Unlike the neural network approach, no balancing of default and non-default data is required for model convergence.

In most credit scoring competitions, SGBT has been a winning approach. Recent research by Grinsztajn, Oyallon, and Varoquaux [27] suggests that tree-based models will perform better than neural networks for tabular data structures where neighboring input factors may have no ordering or continuity. Neural networks have been found to excel in sound and image processing applications where the inputs are neighboring pixels in an image or sequential points in the time sampling.

For the current work, the goal is not to declare a winner, but rather simply to compare the residuals of these methods versus vintage origination date. For consistency of comparison, vintage date is again excluded from the inputs and adverse selection is quantified via a final logistic regression as in Equation (3) where $M(s_i)$ is the full ensemble of trees applied to forecasting account i and held as a fixed input when measuring the vintage residuals.

3.2.7. SGBT + APC

Some implementations of stochastic gradient boosted trees allow for the same kind of fixed inputs as logistic regression and the NN+APC algorithm above. Again using $O(a, t) = F(a) + H(t)$ as a fixed input allows us to create an SGBT credit risk panel model that is centered around the long-term trends of lifecycle and environment. As observed with NN+APC, the resulting hybrid model can be both simpler and more robust out-of-sample as compared to the stand-alone SGBT model.

The inputs to the credit scoring SGBT model were the same as for the DTSM using the full panel dataset, where defaults occurring at any age are included. This is the same dataset used for NN+APC models. Because of the volume of data, the model was estimated on a 5% random sample of loans, including the full history for each loan. During training, 80% of the 5% sample was used for training and 20% for cross-validation to determine the stopping point. Model residuals by vintage were estimated by applying the models to the full dataset.

4. Results

Nine separate models were estimated using seven different techniques, including the APC decomposition. For regression models, measures of LTV and term were binned to allow for nonlinearities in their relationship to default. collateral type and channel categorical variables are measured relative to their reference levels, which are indicated with a 0 estimate. Not all lenders reported DTI, so a separate flag for DTI missing was included and DTI missing was interacted with DTI to capture the correlation to default.

All of these independent variables are available at loan origination, which is the traditional design of an origination score. For purposes of measuring adverse selection, we are concerned with the loans that are actually booked. Therefore, we can additionally incorporate information available just after origination. We call this a “post-origination score”. The most useful factor was found to be $\Delta APR(i, v)$ as defined in Equation (1). Adding $\Delta APR(i, v)$ to the model improved the in-sample fit, lowering AIC from 23,800 to 23,314, and actually improved the significance of the channel and collateral type coefficients. The rest of the models were estimated post-origination so that the adverse selection measure removes as much structure as possible from the independent variables.

To confirm that the models were estimated properly, receiver operating characteristic (ROC) curves were estimated for each model. Because of the different datasets for the models with 24-month outcome periods and models with APC inputs using panel data, the Gini coefficients are unlikely to be directly comparable. Further, the NN and SGBT models were estimated on samples and applied to the full dataset, so most of those test results are out-of-sample. Regardless of the many test differences, the results in Table 1 show that all models are working as expected. Logistic regression performs normally given that it uses only the first 24 months of performance data and therefore is missing a majority of the defaults that occur later. APC has the lowest Gini coefficient of the panel methods, because it has no loan-level information and makes no attempt to be a scoring model. The DTSM, NN, and APC models all perform comparably given the uncertainty in estimating the test statistics and different handling of the data. For example, SGBT + APC Post-Origin Score has the most sampling disadvantages of the models, yet performed comparably.

Table 1. Gini coefficients for the models tested.

Method	Traditional	+ APC
LR/DTSM Orig	0.664	0.853
LR/DTSM Post-Origin	0.703	0.853
NN	0.839	0.822
SGBT	0.868	0.836
APC		0.773

With confirmation that the models are performing properly, the following analysis compares the model residuals by vintage. Figure 1 computes the change in adverse selection by vintage comparing origination to post-origination logistic regression scores. The figure shows that including APR information in the score does refine our understanding of adverse selection in 2016–2017, where the origination score residuals are overestimated and 2018–2019 where the origination score residuals were underestimated. The scale of ± 0.1 in units of change in log-odds is roughly equivalent to a $\pm 10\%$ change in credit risk. This is not large compared to the underlying measures of adverse selection shown in subsequent graphs, but not immaterial. In general, we conclude that post-origination models provide some advantage when measuring adverse selection as a way to incorporate underwriting policy changes that might not otherwise be captured in the models.

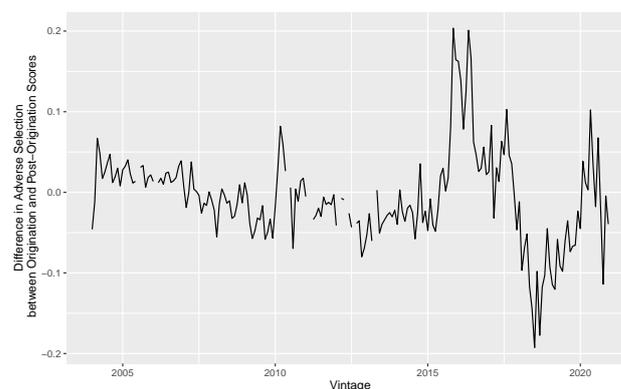


Figure 1. The difference between macroeconomic adverse selection measures for a logistic regression score using information available at origination versus that which is available post-origination but excluding behavioral data.

The next step was to compare adverse selection as measured for logistic regression, neural networks, and stochastic gradient boosted trees, Figure 2. These models are post-origination credit scores using default in the first 24 months as the target variable. Tests were run using other outcome periods, but they were less effective. Extending the outcome

period to 36 months captures significantly more loan defaults, but it delays the measurement of adverse selection by that same three years. In order to have business value, the waiting time for estimating adverse selection must be as short as possible. At the other extreme, we could have looked at the first 12 months in the life of a loan to assess residual credit risk. Some lenders even focus on first payment default as an early warning indicator. Although potentially useful, we would run out of data with which to construct models. The trade-offs are challenging.

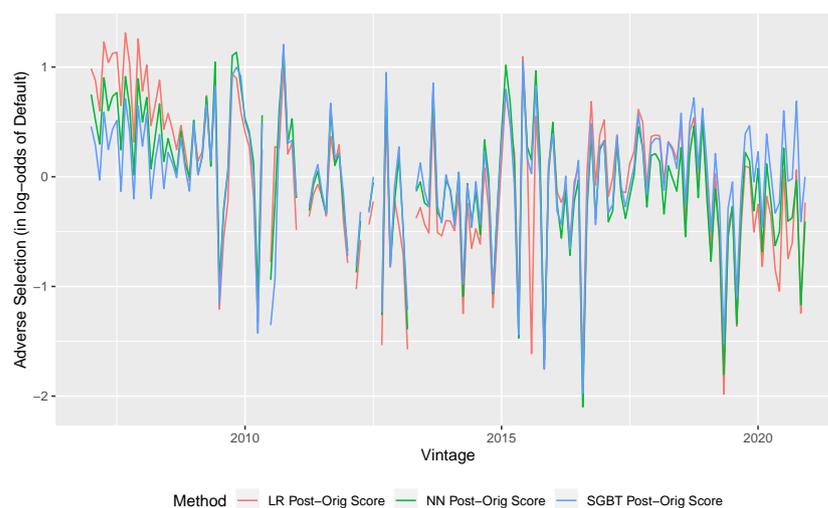


Figure 2. A comparison of macroeconomic adverse selection measures from logistic regression, neural networks, and stochastic gradient boosted trees using defaults within the first 24 months of a loan as the target variable.

Most notable when comparing these measures of adverse selection is the overall similarity. This suggests that adverse selection is more of an attribute of the loans than of the models. Although the NN and SGBT models can capture much more nonlinearity and interaction between variables, the modeling technique does not fully explain the structure within the data. The model residuals (adverse selection) for NN and SGBT are closer to the through-the-cycle average during 2020, 2014, and 2007, but they are not flattened entirely. Notably, the periods of high risk from 2008 through 2009 are still present and are consistent with prior mortgage studies of heightened adverse selection during that period. Better loan quality from 2011 through 2014 is also consistent with the prior mortgage results, so although the estimates are volatile by monthly vintage, they are broadly consistent with expectations.

One solution to the trade-off between quicker response and more defaults is the use of a survival modeling approach where defaults at any age are compared to a baseline expectation from a hazard function or lifecycle. This kind of analysis could be implemented in many ways. Beginning with an Age-Period-Cohort decomposition provides a complete measure of credit risk by vintage but without explanation, Figure 3. That APC decomposition uses lifecycles segmented by bureau score and term, so it is adjusted for dominant scoring factors, but not LTV, DTI, channel, or collateral type.

Taking lifecycle and environment as fixed inputs to a panel logistic regression (DTSM) allows us to further adjust for lender shifts in origination volume by LTV, DTI, channel, or collateral type. The adverse selection measured for the DTSM is overlaid in Figure 3. The comparison of APC vintage function to DTSM adverse selection shows that they are very similar. A small divergence occurs between 2012 and 2016, but recent measures are very well aligned. This suggests that the segmented APC analysis is a quick, computationally efficient way to capture most of the adverse selection problem, although there can be situations where an account-level score brings further refinement. Those advantages might

be more acute when measured for a single lender where the volume by loan attributes can swing more rapidly.

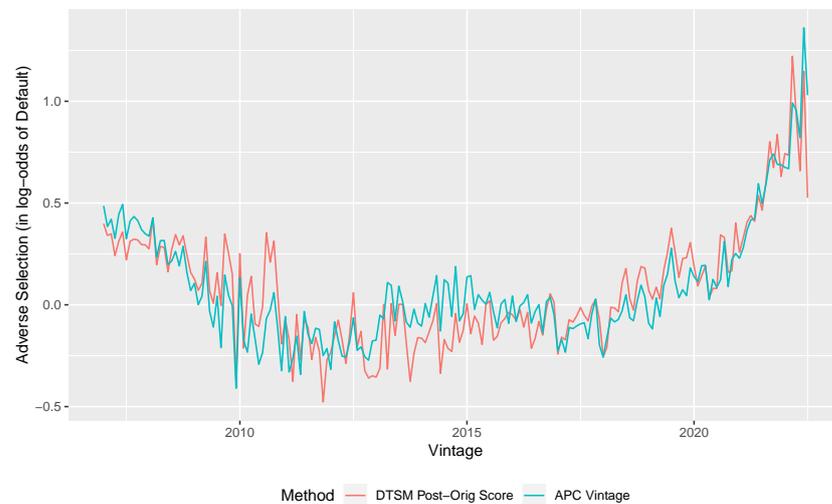


Figure 3. A comparison of the credit risk estimate by vintage from an Age-Period-Cohort analysis and the residuals by vintage from a discrete time survival model with the same APC lifecycles and environment as fixed inputs.

Figure 4 compares the three methods of panel estimation with APC inputs. From a data perspective, the comparison is still not fair. Even though large servers were used for the analysis, the full panel dataset (performance data for every month of every loan) is 32,028,587 rows of data. That far exceeds what could be processed using standard libraries for stochastic gradient boosting (gbm) and neural networks (keras) in R in reasonable time. Therefore, the NN and SGBT models downsampled the non-default loans so that the model training sets were only 342,895 rows of observations. Conversely, the APC vintage decomposition uses all of the data in vintage aggregate form and the DTSM used all observations within the panel data. Regardless of sampling and algorithm used, the vintage-aggregate residuals for each model are remarkably similar.

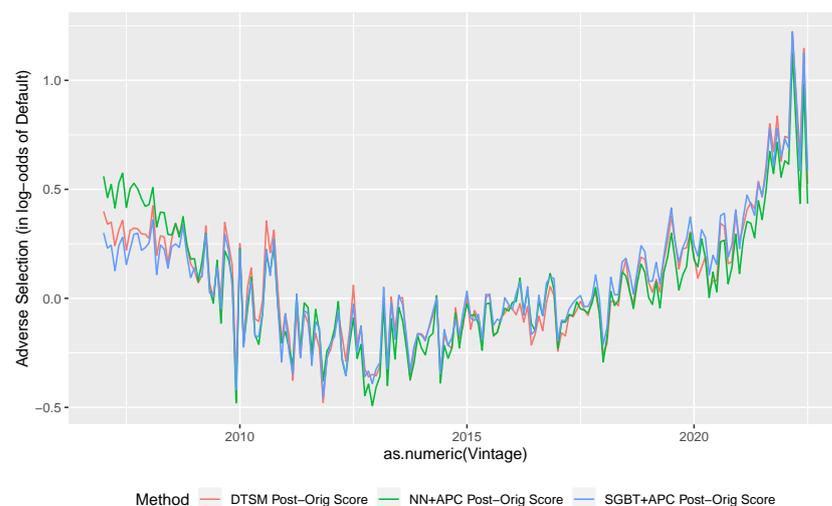


Figure 4. A comparison of the macroeconomic adverse selection estimates from the best DTSM, NN, and SGBT models.

The biggest difference in estimating adverse selection by vintage is seen to be the difference between traditional scoring data and panel data. Comparing Figure 2 to Figure 4

makes clear that scoring methods are significantly more volatile in their adverse selection estimates, simply due to the smaller number of defaults available for modeling, only 4.4% of the total number of lifetime defaults observed. Moreover, because of the 24-month lag in estimation, the scoring approach provides no indication of recent trends.

The recent trends in adverse selection are quite important. All of the panel approaches with APC inputs show that adverse selection has been dramatically higher since February 2021. Within the auto lending industry, this is assumed to be caused by the jump first in the cost of new and used vehicles and later by the increase in the cost of borrowing. Those pressures are hypothesized to have pushed the “value shoppers” out of the market, leaving the less flexible or financially savvy buyers. This is the same dynamic observed leading up to the 2009 recession.

In the second half of 2022, those selection pressures began to ease with the cost of vehicles coming down and auto loan APRs decreasing by the end of the year. The dramatic drops seen at the end of these trends for the most recent months are based upon only a few months of observations and have correspondingly large uncertainties.

These results provide compelling evidence that a panel approach with APC inputs is superior to measuring adverse selection from a traditional scoring approach, but it leaves open which modeling technique is best.

5. Conclusions

The concept of macroeconomic adverse selection became clear during the period 2006 through 2009 when poor quality loans were originated beyond what lenders could expect from their usual scoring inputs. The conditions of rapidly rising home prices and rising interest rates appear to have created an unappealing environment for financially cautious borrowers. The macroeconomic conditions in 2021–2022 resemble this prior period, but with even more extreme rates of change. This led us to suspect that adverse selection would again occur.

This study was undertaken in part to confirm this intuition about the presence of macroeconomic adverse selection in recent auto originations, which was shown here. In addition, the analysis demonstrated that models which create scores estimated relative to lifecycle and environment measures from APC or survival analysis can more rapidly and accurately identify emerging periods of adverse selection. This is valuable from a business perspective so that measuring adverse selection becomes actionable intelligence rather than a retrospective curiosity.

Contrary to some suggestions, machine learning models cannot explain adverse selection as missing nonlinear structure. Rather, the adverse selection measured from neural network and stochastic gradient boosted tree models, even with APC inputs, had residual credit risk by vintage that was statistically unchanged relative to discrete time survival models. This confirms that residual credit risk by vintage should not be viewed as model error but rather as a real indicator of macroeconomic adverse selection. Ideally, sociodemographic data not currently available for model development might quantify the presence or absence of value shoppers, but they would have to be data that are not restricted due to discrimination risks or privacy concerns.

Author Contributions: Conceptualization, J.L.B. and Y.L.; methodology, J.L.B. and Y.L.; software, J.L.B. and Y.L., validation, J.L.B. and Y.L.; writing, J.L.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable

Data Availability Statement: The data used in this study is proprietary to the contributing institutions and not publicly available.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Breeden, J. A survey of machine learning in credit risk. *J. Credit. Risk* **2021**, *17*, 3. [[CrossRef](#)]
2. Wilson, C. Adverse selection. In *Allocation, Information and Markets*; Palgrave Macmillan: London, UK, 1989; pp. 31–34.
3. Breeden, J.L. Macroeconomic adverse selection: How consumer demand drives credit quality. In Proceedings of the Credit Scoring and Credit Control XII Conference, Edinburgh, UK, 30 August–1 September 2011.
4. Breeden, J.L.; Canals-Cerdá, J.J. Consumer risk appetite, the credit cycle, and the housing bubble. *J. Credit. Risk* **2018**, *14*, 1–30. [[CrossRef](#)]
5. Calem, P.S.; Cannon, M.; Nakamura, L.I. *Credit Cycle and Adverse Selection Effects in Consumer Credit Markets-Evidence from the Heloc Market*; FRB of Philadelphia: Philadelphia, PA, USA, 2011; Working Paper No. 11–13.
6. Thomas, L.; Crook, J.; Edelman, D. *Credit Scoring and Its Applications*; SIAM: Singapore, 2017.
7. Anderson, R. *Credit Intelligence & Modelling: Many Paths through the Forest*; Oxford University Press: Oxford, UK, 2019.
8. Fu, W. *A Practical Guide to Age-Period-Cohort Analysis: The Identification Problem and Beyond*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2018.
9. Holford, T.R. The estimation of age, period and cohort effects for vital rates. *Biometrics* **1983**, *39*, 311–324. [[CrossRef](#)] [[PubMed](#)]
10. Mason, W.M.; Fienberg, S. *Cohort Analysis in Social Research: Beyond the Identification Problem*; Springer: Berlin/Heidelberg, Germany, 1985.
11. De Leonardis, D.; Rocci, R. Assessing the default risk by means of a discrete-time survival analysis approach. *Appl. Stoch. Model. Bus. Ind.* **2008**, *24*, 291–306. [[CrossRef](#)]
12. Stepanova, M.; Thomas, L. Survival analysis methods for personal loan data. *Oper. Res.* **2002**, *50*, 277–289. [[CrossRef](#)]
13. Cox, D.R. Regression models and life-tables. *J. R. Stat. Soc. Ser.* **1972**, *34*, 187–220. [[CrossRef](#)]
14. Breeden, J.L.; Leonova, E.; Bellotti, A. *Instabilities Using Cox ph for Forecasting or Stress Testing Loan Portfolios*; Researchgate: Berlin, Germany, 2019.
15. Breeden, J.L. Incorporating lifecycle and environment in loan-level forecasts and stress tests. *Eur. J. Oper. Res.* **2016**, *255*, 649–658. [[CrossRef](#)]
16. Angelini, E.; Di Tollo, G.; Roli, A. A neural network approach for credit risk evaluation. *Q. Rev. Econ. Financ.* **2008**, *48*, 733–755. [[CrossRef](#)]
17. Desai, V.S.; Crook, J.N.; Overstreet, G.A., Jr. A comparison of neural networks and linear scoring models in the credit union environment. *Eur. J. Oper. Res.* **1996**, *95*, 24–37. [[CrossRef](#)]
18. Khashman, A. Neural networks for credit risk evaluation: Investigation of different neural models and learning schemes. *Expert Syst. Appl.* **2010**, *37*, 6233–6239. [[CrossRef](#)]
19. Laurikkala, J. Improving identification of difficult small classes by balancing class distribution. In Proceedings of the Artificial Intelligence in Medicine: 8th Conference on Artificial Intelligence in Medicine in Europe, AIME 2001, Cascais, Portugal, 1–4 July 2001; Springer: Berlin/Heidelberg, Germany, 2001; pp. 63–66.
20. Sundarkumar, G.G.; Ravi, V. A novel hybrid undersampling method for mining unbalanced datasets in banking and insurance. *Eng. Appl. Artif. Intell.* **2015**, *37*, 368–377. [[CrossRef](#)]
21. Breeden, J.L.; Leonova, E. When big data isn't enough: Solving the long-range forecasting problem in supervised learning. In Proceedings of the 2019 International Conference on Modeling, Simulation, Optimization and Numerical Techniques (SMONT 2019), Shenzhen, China, 27–28 February 2019; Atlantis Press: Amsterdam, The Netherlands, 2019; pp. 229–232.
22. Quinlan, J.R. Induction of decision trees. *Mach. Learn.* **1986**, *1*, 81–106. [[CrossRef](#)]
23. Ali, K.; Pazzani, M. Error reduction through learning multiple descriptions. *Mach. Learn.* **1996**, *24*, 172–202. [[CrossRef](#)]
24. Breiman, L.; Friedman, J.; Stone, C.J.; Olshen, R.A. *Classification and Regression Trees*; CRC Press: Boca Raton, FL, USA, 1984.
25. Bastos, J. *Credit Scoring with Boosted Decision Trees*; Technical Report MPRA Paper No. 8034; CEMAPRE, School of Economics and Management (ISEG), Technical University of Lisbon: Lisbon, Portugal, 2007.
26. Chang, Y.-C.; Chang, K.-H.; Wu, G.-J. Application of extreme gradient boosting trees in the construction of credit risk assessment models for financial institutions. *Appl. Soft Comput.* **2018**, *73*, 914–920. [[CrossRef](#)]
27. Grinsztajn, L.; Edouard Oyallon, E.; Varoquaux, G. Why do tree-based models still outperform deep learning on tabular data? *arXiv* **2022**, arXiv:2207.08815.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.