Covariate Selection for Mortgage Default Analysis Using Survival Models

Dongfang Zhang, Basu Bhandari, Dennis Black

Model Development Department, Comerica Bank, Dallas, Texas, USA
Email: DZhang@comerica.com

How to cite this paper: Zhang, D.F., Bhandari, B. and Black, D. (2021) Covariate Selection for Mortgage Default Analysis Using Survival Models. *Journal of Mathematical Finance*, **11**, 218-233.
https://doi.org/10.4236/jmf.2021.112012

Received: March 10, 2021
Accepted: May 9, 2021
Published: May 12, 2021

Abstract

The mortgage sector plays a pivotal role in the financial services industry, and the U.S. economy in general, with the Federal Reserve, St. Louis, reporting Households and Nonprofit Organizations for One-to-Four-Family Residential Mortgages Liability Level at $10.8T in Q3 2020. It has been in the interest of banks to know which factors are the most influential predicting mortgage default, and the implementation of survival models can utilize data from defaulted obligors as well as non-default obligors who are still making payments as of the sampling period cutoff date. Besides the Cox proportional hazard model and the accelerated failure time model, this paper investigates two machine learning-based models, a random survival forest model, and a Cox proportional hazard neural network model DeepSurv. We compare the accuracy of covariate selection for the Cox model, AFT model, random survival forest model, and DeepSurv model, and this investigation is the first research using machine learning based survival models for mortgage default prediction. The result shows that Random survival forest can achieve the most accurate, and stable, covariate selection, while DeepSurv can achieve the highest accuracy of default prediction, and finally, the covariates selected by the models can be meaningful for mortgage programs throughout the banking industry.

Keywords

Survival Analysis, Mortgage Default, Cox Model, Random Survival Forest, Accelerated Failure Time Model, DeepSurv

1. Introduction

Home building and sales are one of economic engines driving the United States' $21T economy, and the Federal Reserve, St. Louis, reports Households and Nonprofit Organizations for One-to-Four-Family Residential Mortgages, Liability Level, at $10.8T in Q3 2020. Housing foreclosure and mortgage default were major drivers of the 2008 Great recession, however, to the contrary in 2020, existing housing sales are up, generating demand for mortgages even during the Cov-19 crisis and are a bright spot in the US economy as shown below.

Figure 1 ([1]), shows a graphic for existing home sales during the Cov-19 pandemic which shows a strong increase starting in May 2020 in existing home sales totaling 6,690,000 by November 2020 generating revenue for banks but also requiring capital reserves, and consequently, predicting mortgage default will be valuable for a bank's decision on the amount of capital reserve to hold.

Also, two important aspects involved in predicting performance evaluation and prediction interpretation, are respectively: 1) Prediction accuracy, and 2) the rank of covariate importance.

Mortgage default data presents a binary classification problem with an obligor either defaulting or not defaulting, a logistic regression model seems to sufficiently handle this type of classification problem with 1 indicating default and 0 indicating nondefault, however, the classification of 0 as nondefault is incomplete, since the status of this loan is unknown after the end of the sampling period: for example, consider a performing mortgage loan with a loan term of 30 years that does not default or pay-off during the sampling period, then classification of 0 is incomplete, since the status of the loan is unknown from the end of the sampling period to the end of the 30-year loan term.

However, this incomplete data can still give information on default probability, and should not be discarded, instead survival analysis is a modelling methodology that can incorporate this type of incomplete data.

A wide variety of applications for survival analysis abound in economics and

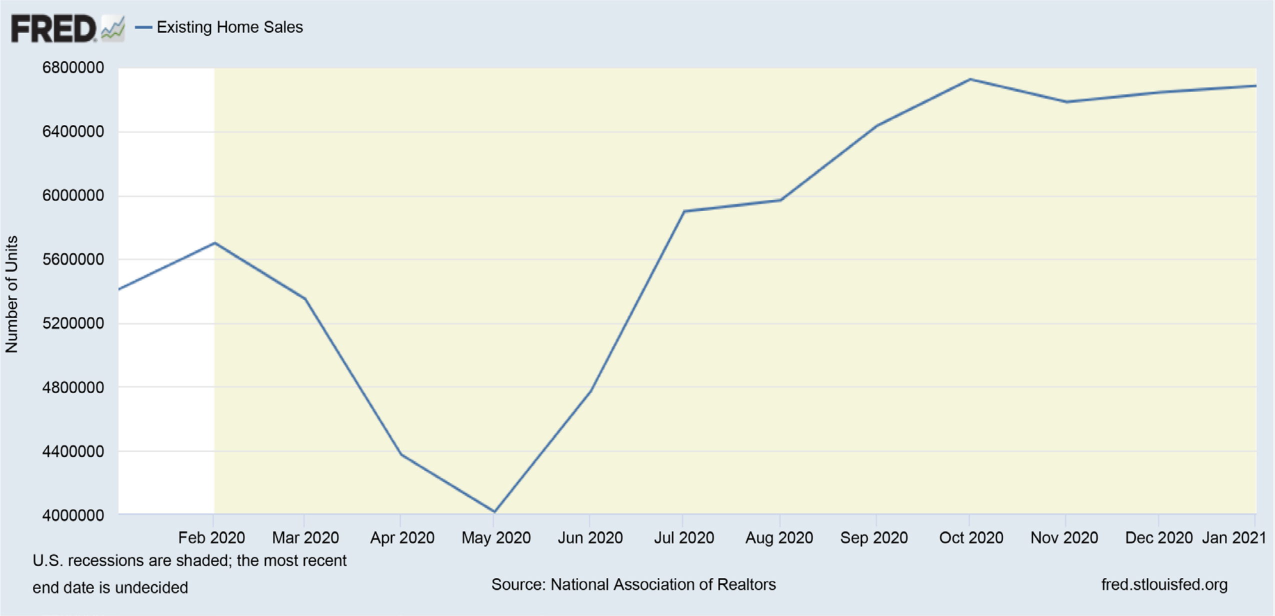

Figure 1. Existing home sales in the united states.

other social science disciplines ranging from unemployment analysis to the tendency of a convicted criminal to reoffend (recidivism). [2] examines the unemployment rate calculated from the Current Population Survey and indicates that information on the length of employment is only collected on those individuals unemployed at the time of the survey, and not for an individual's unemployed between surveys. [2] indicates that individuals are unemployed coming into the survey period presenting left censoring, and individuals are employed at the end of the survey period, but become unemployed after the end of the survey, presenting right censoring. [3] examines recidivism from the standpoint of survival analysis and indicates that ad hoc attempts to introduce time varying covariates, without the use of survival analysis introduces unintended consequences. Censoring and time varying covariates need to be formally introduced into the analysis through survival methodology to accommodate incomplete data, and proper likelihood functions need to be developed to examine censoring and time varying covariates.

In this paper, several survival analysis methodologies will be compared in relation to their accuracy of default prediction and accuracy of covariate importance ranking.

Mortgage prediction has been examined by other researchers: [4] used logistic regression on Hongkong residential mortgage data and found that current loan to value ratio (LTV) and unemployment are the two most important factors influencing default. [5] used mortgage data from one financial institution, and employed a Cox proportional hazard model, and the author found that among all the covariates, LTV, and debt-service-coverage ratio had largest impact. [6] compared four classification models, logistic regression, random forest, boosted regression trees, and generalized additive models, and the author found that random forest outperforms other models with prediction accuracy as the metric.

Although these papers discussed the covariate importance based on different models there are several considerations deserving further investigation: First, these papers did not discuss model accuracy, which is a deficiency when discussing covariate importance, and second, most classification models did not utilize the incomplete data, which potentially could be a large fraction of the total dataset.

Survival models incorporate and utilize incomplete data, and several papers have used the famous Cox Proportional Hazard model to examine mortgage default. First, [7] used the Cox model, found a high net property value, and a high house price growth rate decreased mortgage default. Second, [8] researched the effect of mortgage monthly payment paydown on mortgage default using a Cox Proportional Hazards model, and discovered that a high percentage of monthly paydown reduces the risk of mortgage default. Finally, [9] found that the origination loan-to-value ratio and the unemployment rate are important variables predicting mortgage default using the Cox model.

With the popularity of machine learning, several other survival models started being noticed by academicians, such as Random Survival Forest, and deep learning-based survival model. Consequently, this paper will investigate four different survival models, the Cox Proportional Hazard model, Accelerated Failure Time (AFT) model, Random Survival Forest (RSF) model, and the deep learning based DeepSurv model.

To compare the accuracy of the models, training and test data will be constructed and a C-index will serve as the accuracy metric, and the motivating factor for using C-index as the performance metric is because for incomplete data, accuracy could not be defined. This paper will also generate covariate importance ranks for all the models, and construct a model-covariates matrix to discover the best model in terms of covariate importance.

Finally, this paper is arranged as follows: Section 2 will be the introduction of survival theory and models including Cox, AFT, RSF and DeepSurv. Section 3 is building the covariates ranking with Cox model, AFT model, RSF model. Section 4 will be evaluating the effectiveness and accuracy of covariates ranking generated with the three models using DeepSurv model. Section 4 will be the results, discussion and conclusion.

2. Theory

2.1. Censored Data

Censoring arises naturally in time-to-event data when, the starting of an event or ending of the event, are not precisely observed ([10]), and there are various censoring types, for example, right censoring, interval censoring and left censoring. The most common type of censoring is right censoring, where time-to-event is not observed, and as an example consider mortgage default data, where mortgage default is the event of interest in terms of survival analysis. Now assume that for a given dataset, 30% borrower default is observed, and each defaulted observation recorded with an appropriate default date, however, for the remaining 70% of the borrowers in the data, default is not observed, and therefore the observations are recorded as right censored. Although right censored data seems to be a case of missing data, their time-to-event is not actually observed before the end of study, however, these subjects are very valuable because they went a certain amount of time without experiencing an event, and this in itself is informative to the analysis. To use the common statistical models, such as linear regression and logistic regression, for time-to-event data will result in biased estimation and misleading results, since these analyses cannot handle censored data when survival experience is only partially known.

2.2. Survival Function

Survival analysis is a statistical data analytic technique for analyzing time-to-event data, and one fundamental relationship in survival analysis is the survival function. Assume T is a continuous random variable, then the probability of an individual surviving beyond time t can be defined as Equation (1).

$$S(t) = P(T \geq t) = \int_t^{\infty} f(t)\mathrm{d}t = 1 - F(t) \tag{1}$$

where $f(t)$ is the probability density function of an event of interest happens at time *t*. $F(t)$ represents the cumulative probability of an event of interest happened by time *t*. For example, in the case of mortgage default above, the event of interest is mortgage default. $S(t)$ is the probability of mortgage default has not happened until time *t*. Time *t* is not an absolute time stamp, but a time period relative to the start of mortgage borrowing.

The other basic quantity is the hazard function. It is also known as the hazard rate, the instantaneous death rate, or the force of mortality. The hazard function can be expressed as in Equation (2).

$$\lambda(t) = \lim_{dt \to 0} \frac{P(t \leq T \leq t + dt \mid T \geq t)}{dt} = \frac{f(t)}{S(t)} \tag{2}$$

where $P(t \leq T \leq t + \mathrm{d}t \mid T \geq t)$ expresses the conditional probability that the event of interest will happen in time interval *dt* given it did not occur before.

Combining Equation (1) and Equation (2), Equation (3) can also be derived, which shows that survival and hazard function provide equivalent information.

$$S(t) = \exp\left(-\int_0^t \lambda(x)\mathrm{d}x\right) \tag{3}$$

where $\int_0^t \lambda(x)$ is called cumulative hazard, which every model uses to calculate the survival function, *S*(*t*).

2.3. Concordance Index

In time-to-event data, because some outcomes are unknown, it will not be possible to use accuracy or area under curve (AUC) to evaluate the performance of a model, however, [11] proposed a rank-based method to judge the prediction capability of survival models. Every survival model generates a risk score for each subject, and usually, the risk score is the median survival time of a subject, and then all possible appropriate subject pairs will be evaluated as principles shown in [11]. For example, if both subjects of the pair are not censored, and the median survival time of A is larger than that of B, and the time to event of A is larger than that of B, [11] calls this a concordant pair. [11] also explained how to handle concordance when only one subject is censored, or two subjects are censored. Finally, the ratio of concordance counts to the counts of all valid pairs is the concordance index (C-index) where the range of the C-index varies between 0.5 to 1, and when the C-index is 0.5, the model is no better than a random guess, and when the C-index is 1, the model can predict perfectly. [12] found that the C-index is the weighted average of time specific AUC, which explains its popularity as the metric of choice for evaluating survival models.

3. Results and Discussion

3.1. Data Introduction and Exploration

The dataset used in this paper has 50,000 U.S mortgage borrowers (obligors), and is the dataset used by [13], which can be downloaded from their book's website, and at the end of the tracking of each mortgage borrower, some borrowers were recorded as defaulted or finished payment. In this data, 30% of obligors defaulted, only 17% of the obligors were recorded as continuously paying, and rest of obligors finished their loan term. Each mortgage is associated with an origination time, record time, indicator of default or indicator of payoff, and other covariates associated with the mortgage. Names and explanations of all the 15 covariates are in **Table 1**, and these 15 covariates can be grouped as macroeconomic variables (gdp, uer, hpi, interest_rate), loan related variables (LTV, LTV_orig, FICO_orig, investor_orig, balance, balance_orig, hpi_orig, Interest_rate_orig), and property related variables (Retype_co_orig, Retype_PU_orig, Retype_SF_orig).

The four macroeconomic covariates are grouped, and their paired correlations calculated, as shown in **Figure 2**. **Figure 2** shows the covariate hpi has a strong negative correlation with the covariate uer, which from the economic perspective is supported, as the economy gets stronger more people are employed, decreasing the unemployment rate, and similarly a more active economy increases disposable income which implies more homes are purchased, increasing the hpi leading to a negative correlation between uer and hpi. The covariate gdp has a fairly strong positive correlation with hpi, where a sustained increase in economic activity increases gdp, and similarly increases home sales, which in turn increases hpi leading to a positive correlation between gdp and hpi. Likewise, a

Table 1. Description of covariates in mortgage data set.

Covariate name	Description
Balance	Outstanding balance at observation time
LTV	Loan-to-value ratio at observation time
interest_rate	Interest rate at observation time
hpi	House price index at observation time
gdp	Gross domestic product (GDP) growth at observation time, in %
uer	Unemployment rate at observation time, in %
REtype_CO_orig	Real estate type condominium = 1, otherwise = 0
REtype_PU_orig	Real estate type planned urban development = 1, otherwise = 0
REtype_SF_orig	Single family home = 1, otherwise = 0
investor_orig	Investor borrower = 1, otherwise = 0
balance_orig	Outstanding balance at origination time
FICO_orig	FICO score at origination time, in %*
LTV_orig	Loan-to-value ratio at origination time, in %
Interest_Rate_orig	Interest rate at origination time, in %
hpi_orig	House price index at origination time, base year = 100

*FICO score is a credit score created by Fair Isaac Corporation.

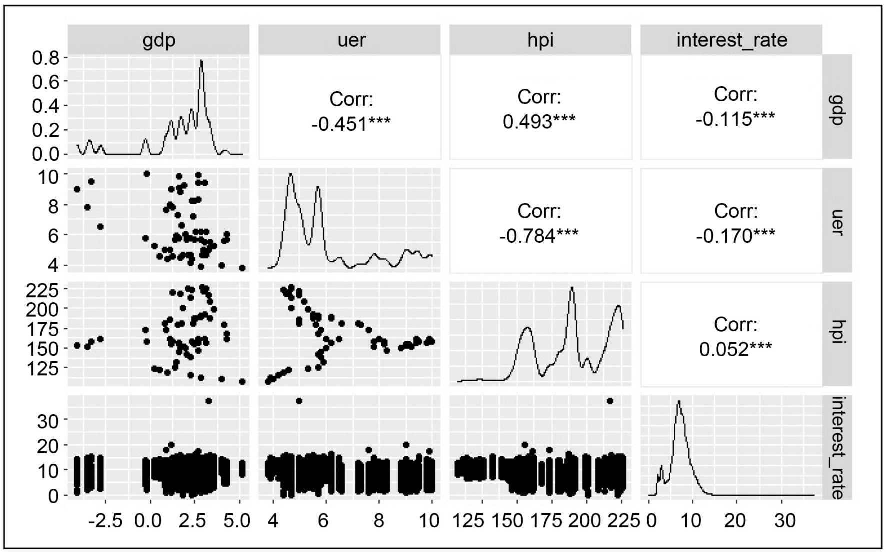

Figure 2. Correlations between macroeconomic covariates.

sustained increase in economic activity increases employment, which in turn decreases the unemployment rate, uer, and presents a fairly strong negative correlation between gdp and uer in Figure 2.

Before fitting survival models, the data set is preprocessed by performing two steps. First, it is moving the origination date of each mortgage to 0, and the second, is keeping only the last record of each mortgage, and computing the time from origination date to the last observation. The default indicator variable takes on two values, the value 1 if the mortgage has defaulted during the sampling window, and 0 if the observation has not defaulted, that is, survived and is censored, and finally, left censoring is avoided by assuming all loans start from the first observation.

Before fitting any survival model, it is standard practice to generate Kaplan-Meier survival curves to explore the impact of univariate data on survival, and since most of the covariates in the mortgage data set are continuous, dummy variables are generated as follows: For any continuous covariate, if the value is larger than the mean of the covariate, it is labeled as 1, otherwise it is labeled as 0. Figure 3 is the Kaplan-Meier survival curves of all 15 covariates.

In each survival curve, if the two curves are overlapping, it signifies the value of that covariate does not matter to the survival time of the mortgage, and if the two curves separate, it indicates the covariate impacts survival time. Figure 3 shows that a few covariates have no impact on survival time, such as investor_orig, REtype_CO_orig, REtype_PU_orig, and REtype_SF_orig, however, for many of the univariates there is separation between the two curves, indicating an impact on survival time, for example, FICO_orig, interest_rate, uer, interest_rate_orig, balance, hpi, balance_orig, hpi_orig, LTV, gdp, LTV_orig. Note, the first graph examines FICO score at origination, and for FICO scores at origination above the mean, there is less risk, represented by the higher turquoise survival curve, versus, the lower FICO scores at origination represented by the

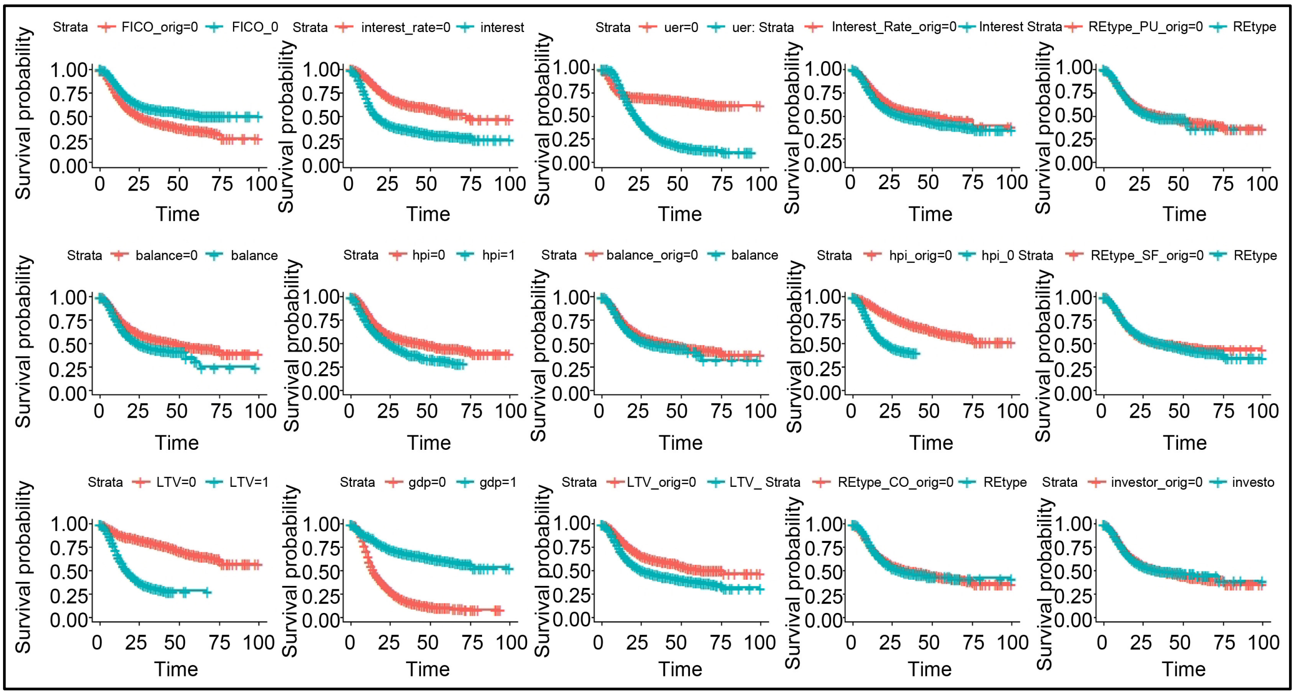

Figure 3. Kaplan-Meier survival curves of all covariates.

lower red survival curve indicating shorter survival time, and the other univariates that show separation follows a similar logic. Although the Kaplan-Meier curve is visually straightforward, there is a drawback, it does not detect correlations.

3.2. Cox proportional Hazard Model

In D.R. Cox's famous 1972 paper [14], Regression Models and Life-Tables, in the Journal of the Royal Statistical Society, Cox proposed a model for handling time-to-event data, explaining the effects of multi-covariates with continuous or categorical covariates, and this model, the Cox model, is expressed in Equation (4).

$$\lambda(t \mid X_i) = \lambda_0(t)\exp(X_i \cdot \beta) \tag{4}$$

where $X_i = \{X_{i1}, X_{i2}, \cdots, X_{in}\}$ are the values of covariates of object *i*. The Cox model attempts to find the effect of covariates on the hazard rate, $\lambda(t)$, by multiplying the base hazard rate, which changes with time, and an exponentiated linear combination of covariates. The above model implies the effect of the covariates on the hazard rate does not change over time, and the Cox model is called a proportional hazards model since the ratio of the hazard rate of one object, X_i over that of another object, X_j is a constant.

L1-regularized generalized linear model (LASSO regression) was introduced by Tibshirani ([15]), as shown in Equation (5), in which β is the coefficient vector and λ is the regularization parameter.

$$\hat{\beta}(\lambda) = \arg\min_{\beta}\left[-\log\{X;\beta\} + \lambda|\beta|\right] \tag{5}$$

Lasso regression has the quality of shrinking and selecting covariates, and Tibshirani ([16]) shows that LASSO regression can select the best set of covariates, compared with other covariates selection methods, and the covariate selection of the Cox model can be incorporated into the LASSO regression, as illustrated in [17] [18]. This algorithm is included in glmnet package of R, which is used in this paper, and as Tibshirani ([18]) indicates, all covariates should be standardized for the purpose of covariate selection, otherwise the coefficients cannot be compared.

Figure 4 shows how the LASSO Cox model works, when λ increases, all the coefficients shrink, and coefficients that are not statistically significant are suppressed to zero excluding those covariates from the final model. Depending on how many variables are selected, an appropriate λ can be chosen, and in this paper, five covariates are selected with λ set at 0.05. The position of λ is shown as the vertical red line in **Figure 4**, which corresponds to the value of $\log(\lambda)$ as −3.

Figure 4 also shows the behavior of the LASSO modeling framework with the log of lambda on the x-axis, coefficient values on the y-axis, number of parameters on the top horizontal axis, colored lines in the graph representing the number of variables in the model, and finally, this plot corresponds to Tibshirani's [15] graphic on page 273. The log of lambda at −3, exponentiated, obtains 0.0497871 which corresponds to the selection of $\lambda = 0.05$, yielding 5 variables as the appropriate model, and the turquoise (half oval) line shows the 5 variables are non-zero, emphasized by the red horizontal line.

The five covariates are listed in **Table 2**, from left to right, according to their importance in the Cox model, and covariate importance is ranked by the absolute values of their coefficients, which can be explained with the survival function of the Cox model, as shown in Equation (6).

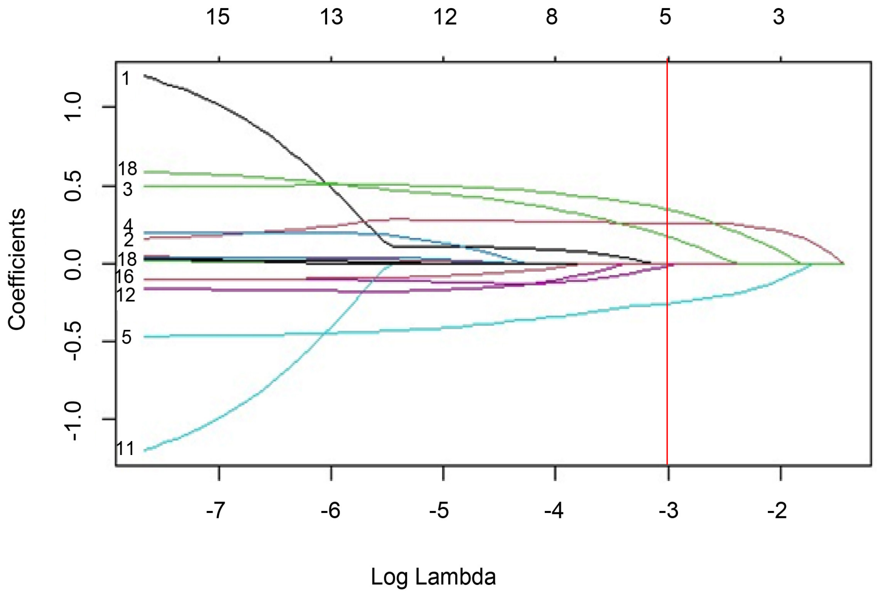

Figure 4. Variation of coefficients of Cox model with λ.

Table 2. The outcome from LASSO Cox model.

Variable	interest_rate	LTV	gdp	hpi_orig	FICO_orig
Coefficient	0.35	0.26	−0.25	0.17	−0.01

$$S(t \mid X_i) = \left[S_0(t)\right]^{\exp(\beta \cdot X_i)} \tag{6}$$

From Equation (6) it can be concluded that when a coefficient is 0, the covariate has no impact on the survival function and, when the coefficient is larger than 0, it will reduce survival time, and when the coefficient is negative, it increases survival time. This can explain why the coefficient of gdp and FICO_orig are smaller than 0, *i.e.*, since a higher gdp growth rate and larger FICO scores have a positive impact on survival time, the coefficients are less than 0, which in turn, positively affects survival time. The coefficient values for interest rate, LTV, hpi_orig are positive, since the higher values for those risk drivers indicate the possibility of a shorter survival time, *i.e.*, interest rate is a measure of default risk, the higher the interest rate, the higher the risk of defaults, and for LTV the larger the loan in relation to the value of the property the higher the risk, and finally for hpi_orig, the higher the house price at origination the higher the mortgage payment, and the more difficult for the obligor to make larger payments over the business cycle. Now, given that all the covariates are standardized to the same magnitude, the absolute value of the coefficient reflects the extent survival time can be reduced, and the survival function altered.

3.3. Accelerated Failure Time Model

Like the Cox model, the accelerated failure time (AFT) model is also a linear model, and the L1-regularized, Lasso penalty, AFT model will be employed with this data to choose the five most significant covariates that drive failure time. There are several parametric AFT models, and the Weibull AFT model is the most popular since it has characteristics of both a proportional hazard model and an accelerated failure time model. Equation (7) shows the Weibull AFT model.

$$\log(T_i) = X_i^{\mathrm{T}}\beta + \sigma\varepsilon_i \tag{7}$$

In Equation (7), ε_i is an i.i.d. random variable that satisfies the log-Weibull distribution and σ is a scale parameter, and since Weibull AFT is a parametric AFT model, the expected survival time can be derived as Equation (8), which can give a clear indication how covariates impact survival time ([19]).

$$E(T) = \exp\left(X^{\mathrm{T}}\beta\right)\Gamma(\sigma+1) \tag{8}$$

The Lifelines python package will be used in this section, and similar to the Cox model, all the covariates are standardized before applying the AFT model, and from Equation (8), we can find that when one covariate's coefficient is 0, the covariate does not have an impact on survival time. When the covariate's coefficient is larger than 0, it has positive impact on survival time, and therefore, the coefficients from the AFT model usually are opposite of that from the Cox model, as shown in **Table 3**.

The covariates selected with the AFT model are consistent with the Cox model, and the signs of the coefficients are opposite of the Cox model, which confirms the theoretical analysis.

3.4. Random Survival Forest

The random survival forest (RSF) model derives from the Random forest model of Breiman ([20]), and contrasted to the Cox model and the AFT model, which are parametric and continuous models, a random survival forest model is a non-parametric, discrete model. It has the advantage that it does not depend on any distributional assumptions, and the drawback is it is hard to explain the quantitative effect of different covariates, although it can still generate the rank order of covariate importance.

Like Random forests, RSF models also produce hundreds of decision trees based on some splitting rule, and the most commonly used splitting rule is the log-rank statistic. For each tree, a subset of the covariates is selected randomly based on the square root of p, where p is the number of covariates, then recursively a covariate is chosen, and its splitting value determined, so that the left node and the right node of the tree has the maximum difference of the log-rank statistics ([21]). The log-rank statistic measures how large is the difference in hazard rates of two groups, and in the case of RSF, it measures the difference in hazard rates between left node and right node in the current split.

The covariate ranking in RSF is similar with that of Random forest. It calculates the drop of prediction accuracy on the test data excluding the selected covariate, and since RSF is an ensemble algorithm, there are efficient ways to implement this process ([22]). This paper uses the random ForestSRC package of R, and **Figure 5** shows the covariate importance from RSF. Unlike the Cox and AFT models, coefficients have quantitative meaning on how they impact survival time, covariate ranking from RSF does not ([23]).

Note that gdp, LTV, uer, interest_rate and hpi_orig are the top five covariates as in the other rankings above.

3.5. DeepSurv

DeepSurv presents as a deep learning algorithm based on a Cox model ([24]), keeping the structure of the neural network algorithm, and retaining the Cox proportional hazards paradigm as a one layer response output with appropriate updated parameters. The effects of covariates are modeled with a multiple layer neural network to capture the interactions between covariates, and like other

Table 3. Outcome from Weibull AFT model.

Variable	interest_rate	gdp	hpi_orig	LTV	FICO_orig
Coefficient	−0.29	0.23	−0.20	−0.18	0.04

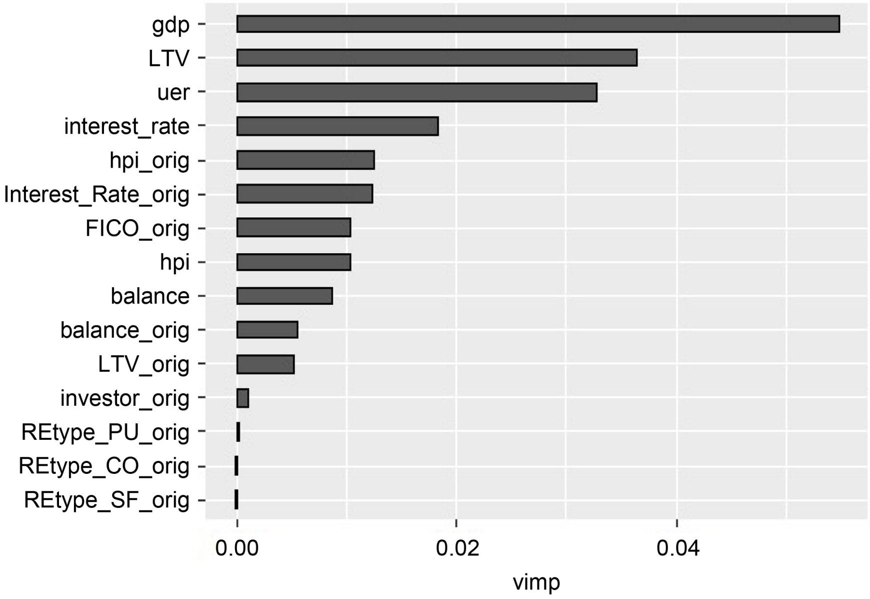

Figure 5. Covariate rank from RSF model.

deep learning neural network algorithms, DeepSurv also uses alternate fully connected layers and drop out layers to avoid overfitting. DeepSurv, also uses a scaled Exponential Linear Unit (SELU) as the activation function with a hazard function output, and finally, the loss function is the average negative log partial likelihood with regularization ([24]). Katzman, briefly describes the algorithm below.

DeepSurv is a multi-layer perceptron similar to the Faraggi-Simon network. However, we allow a deep architecture (i.e., more than one hidden layer) and apply modern techniques such as weight decay regularization, Rectified Linear Units (ReLU) … Batch Normalization … dropout … stochastic gradient descent with Nesterov momentum … gradient clipping … and learning rate.

Scheduling … The output of the network is a single node, which estimates the risk function $\hat{h}_\theta(x)$ *parameterized by the weights of the network* [20].

As seen above, Deepsurv is a highly flexible model facilitated, in part, by modifying the basic gradient descent algorithm into a more adaptable method, and also, as noted, allowing for more neural network layers, introducing more parameters, within the hidden layer framework ([24]), giving more tractability to the neural network environment, and in this paper, the implementation of Deepsurv is accomplished using the python Pycox package ([25]).

The default structure of the DeepSurv neural network will be employed, which is composed of two hidden layers each with thirty-two nodes, a ReLU activation function, a batch norm, and 10% drop out, and finally, the data is split as 80% training data and 20% test data. Training data is used to fit the model, and test data is used for model evaluation by applying the C-index metric to determine the best model fit, and finally, all the models were given a random state, so the results are repeatable.

Unlike the Cox model, which can identify the coefficients of covariates, DeepSurv is a black box model, and consequently is not the optimal choice for coefficient explanation or selection, however, as with neural networks in general, DeepSurv is an excellent prediction model. Kim ([26]), and Zhu ([27]) claim DeepSurv can achieve higher prediction accuracy than other survival models, and this is confirmed with results in **Table 4**, where all the four models were fit with training data and predicted using the test data on all covariates. **Table 4** shows that DeepSurv indeed can achieve much higher accuracy than other models on the mortgage data, where DeepSurv obtains a 16.15% higher percentage change in the C-index than the next highest C-index score attained by RSF.

Next, DeepSurv is used as a tool to compare and evaluate, the performance of covariate ranking obtained from the other models, and the covariate ranking will be evaluated at 5 levels first, the top covariate, then the top 2 covariates, continuing until finally, the top 5 covariates are evaluated. **Table 5** shows the results. Note the Cox model, and the AFT model C-index levels off from the top 4 to top 5 covariates, and the RSF model C-index increases from the top 4 to the top 5 covariates by just 2.8%. **Table 5** shows RSF can pick better covariates at every level, since its C-index outperforms the other two models on all the levels. Cox and the AFT model perform similarly on each level because they have four covariates overlapping for the top five covariates.

Figure 6 is the C-index on test data using the top N covariates of the RSF model, and this figure shows that the top five covariates from the RSF model essentially achieve maximum C-index accuracy, as marked by the horizontal red line in the graphic. Notice the C-index varies very little from 5 covariates to 14 covariates, and as with scree plots and elbow plots, the top 5 covariates are chosen to be the most parsimonious model covariates that have the highest C-index.

Finally, in Section 3.2 and Section 3.3, the choice of a 5-covariate model with a selection of $\lambda = 0.05$, is supported by the conclusions gleaned from **Table 5** and **Figure 6**, and the preceding paragraphs.

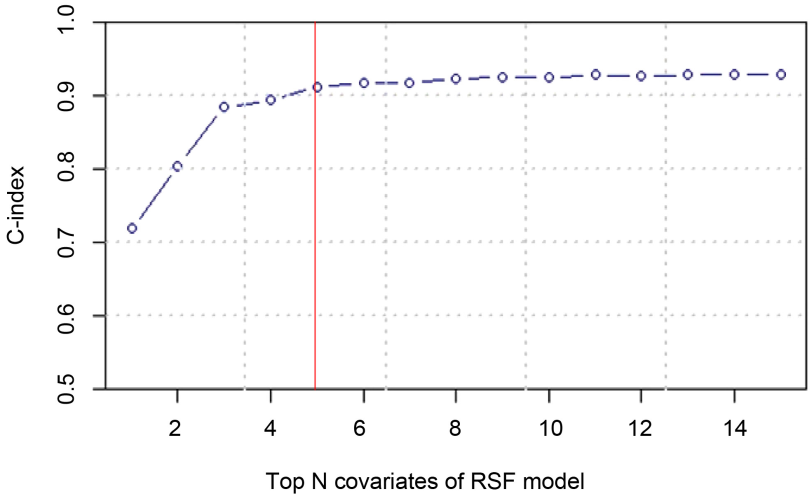

Figure 6. C-index with top N covariates of RSF using DeepSurv model.

Table 4. C-index with different survival models.

Model	Cox	AFT	RSF	DeepSurv
C-index	0.798	0.789	0.799	0.928

Table 5. C-index with covariate ranking of different models.

model	Top 1	Top 2	Top 3	Top 4	Top 5
Cox	0.688	0.768	0.844	0.859	0.865
AFT	0.687	0.776	0.837	0.867	0.865
RSF	0.724	0.805	0.881	0.890	0.915

4. Conclusions

Determining the probability of mortgage default is a critical part of a bank's risk assessment profile affecting originations, relationship management, and loss reserves, consequently, determining the best modeling algorithm is also critical to a bank's overall financial strength. Public mortgage data with 15 covariates, and a binary variable indicating default or nondefault were procured, organized, and analyzed to determine the covariate selection and ranking capability of several widely used and studied survival models. The aim was not to search all the variations of survival models, but to demonstrate the capability of survival models to enhance the understanding of mortgage default through the selection of a judicious set of covariates that explain default and enhance senior managements understanding of an obligor's potential for default. Results from a Kaplan-Meier analysis and Cox Proportional Lasso regression show that interest_rate, LTV, gdp, hpi_orig, and FICO_orig are highly effective explanatory variables to determine mortgage default.

Further analysis shows that DeepSurv can achieve far better prediction accuracy than the other models in this study, and using the C-index as the measure of goodness-of-fit for the Cox, AFT, and RSF models, the RSF model achieves the best goodness-of-fit ranking. Among all the 15 covariates, the RSF model picked 5 covariates which can successfully predict mortgage default, and finally, the chosen top 2 covariates are gdp growth rate and the loan to value ratio, and this result is consistent with findings from the literature ([4] [5] [6]).

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] National Association of Realtors (n.d.) Existing Home Sales [EXHOSLUSM495S]. Federal Reserve Economic Data, Federal Reserve Bank of St. Louis, St. Louis. https://fred.stlouisfed.org/series/EXHOSLUSM495S

[2] Kiefer, N.M. (1988) Economic Duration Data and Hazard Functions. *Journal of*

Economic Literature, **26**, 646-679.

[3] Allison, P.D. (1982) Discrete-Time Methods for the Analysis of Event Histories. *Sociological Methodology*, **13**, 61-98. https://doi.org/10.2307/270718

[4] Wong, J., Fung, L., Fong, T. and Sze, A. (2004) Residential Mortgage Default Risk and the Loan-to-Value Ratio. *Hong Kong Monetary Authority Quarterly Bulletin*, **4**, 35-45. https://doi.org/10.2139/ssrn.1331270

[5] Ciochetti, B.A, Deng, Y., Lee, G., Shilling, J.D. and Yao, R. (2003) A Proportional Hazards Model of Commercial Mortgage Default with Originator Bias. *The Journal of Real Estate Finance and Economics*, **27**, 5-23. https://doi.org/10.1023/A:1023694912018

[6] Fitzpatrick, T. and Mues, C. (2016) An Empirical Comparison of Classification Algorithms for Mortgage Default Prediction: Evidence from a Distressed Mortgage Market. *European Journal of Operational Research*, **249**, 427-439. https://doi.org/10.1016/j.ejor.2015.09.014

[7] Bajari, P., Chu, C.S. and Park, M. (2008) An Empirical Model of Subprime Mortgage Default from 2000 to 2007. Technical Report, National Bureau of Economic Research, Cambridge. https://doi.org/10.3386/w14625

[8] Zhu, J., Janowiak, J., Ji, L., Karamon, K. and McManus, D. (2015) The Effect of Mortgage Payment Reduction on Default: Evidence from the Home Affordable Refinance Program. *Real Estate Economics*, **43**, 1035-1054. https://doi.org/10.1111/1540-6229.12104

[9] Deng, Y., Quigley, J.M., Van Order, R. and Freddie, M. (1996) Mortgage Default and Low Downpayment Loans: the Costs of Public Subsidy. *Regional Science and urban Economic*, **26**, 267-288. https://doi.org/10.1016/0166-0462(95)02116-7

[10] Moore, D.F. (2016) Applied Survival Analysis Using R. Springer, Cham, 3. https://doi.org/10.1007/978-3-319-31245-3

[11] Harrell, F.E., Califf, R.M., Pryor, D.B., Lee, K.L. and Rosati, R.A. (1982) Evaluating the Yield of Medical Tests. *JAMA*, **247**, 2543-2546. https://doi.org/10.1001/jama.1982.03320430047030

[12] Heagerty, P.J. and Zheng, Y. (2005) Survival Model Predictive Accuracy and Roc curves. *Biometrics*, **61**, 92-105. https://doi.org/10.1111/j.0006-341X.2005.030814.x

[13] Baesens, B., Roesch, D. and Scheule, H. (2016) Credit Risk Analytics: Measurement Techniques, Applications, and Examples in SAS. John Wiley & Sons, Hoboken. http://www.creditriskanalytics.net/

[14] Cox, D.R. (1972) Regression Models and Life-Tables. *Journal of the Royal Statistical Society*: *Series B* (*Methodological*), **34**, 187-220. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x

[15] Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*: *Series B* (*Methodological*), **58**, 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[16] Hastie, T., Tibshirani, R. and Tibshirani, R.J. (2017) Extended Comparisons of Best Subset Selection, Forward Stepwise Selection, and the Lasso. https://arxiv.org/abs/1707.08692

[17] Park, M.Y. and Hastie, T. (2007) L1-Regularization Path Algorithm for Generalized Linear Models. *Journal of the Royal Statistical Society*: *Series B* (*Statistical Methodology*), **69**, 659-677. https://doi.org/10.1111/j.1467-9868.2007.00607.x

[18] Tibshirani, R. (1997) The Lasso Method for Variable Selection in the Cox Model. *Statistics in Medicine*, **16**, 385-395.

https://doi.org/10.1002/(SICI)1097-0258(19970228)16:4%3C385::AID-SIM380%3E3.0.CO;2-3

[19] Liu, E. (2018) Using Weibull Accelerated Failure Time Regression Model to Predict Survival Time and Life Expectancy. https://doi.org/10.1101/362186 https://www.biorxiv.org/content/10.1101/362186v2.full.pdf

[20] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32. https://doi.org/10.1023/A:1010933404324

[21] Nasejje, J.B., Mwambi, H., Dheda, K. and Lesosky, M. (2017) A Comparison of the Conditional Inference Survival Forest Model to Random Survival Forests Based on a Simulation Study as Well as on Two Applications with Time-to-Event Data. *BMC Medical Research Methodology*, **17**, Article No. 115. https://doi.org/10.1186/s12874-017-0383-8

[22] Ishwaran, H., Kogalur, U.B., Chen, X. and Minn, A.J. (2011) Random Survival Forests for High-Dimensional Data. *Statistical Analysis and Data Mining*: *The ASA Data Science Journal*, **4**, 115-132. https://doi.org/10.1002/sam.10103

[23] Ishwaran, H., Kogalur, U.B., Blackstone, E.H. and Lauer, M.S. (2008) Random Survival Forests. *The Annals of Applied Statistics*, **2**, 841-860. https://doi.org/10.1214/08-AOAS169

[24] Katzman, J.L., Shaham, U., Cloninger, A., Bates, J., Jiang, T. and Kluger, Y. (2018) Deepsurv: Personalized Treatment Recommender System Using a Cox Proportional Hazards Deep Neural Network. *BMC Medical Research Methodology*, **18**, Article No. 24. https://doi.org/10.1186/s12874-018-0482-1

[25] https://github.com/havakv/pycox

[26] Kim, D.W., Lee, S., Kwon, S., Nam, W., Cha, I.-H. and Kim, H.J. (2019) Deep Learning-Based Survival Prediction of Oral Cancer Patients. *Scientific Reports*, **9**, Article No. 6994. https://doi.org/10.1038/s41598-019-43372-7

[27] Zhu, X., Yao, J. and Huang, J. (2016) Deep Convolutional Neural Network for Survival Analysis with Pathological Images. 2016 *IEEE International Conference on Bioinformatics and Biomedicine*, Shenzhen, 15-18 December 2016, 544-547. https://doi.org/10.1109/BIBM.2016.7822579