A Machine Learning Based Approach for Financial Risk Assessment of a Small Business Loan Portfolio

Daniel Bereket, *Peng-jen* Chen, *Yash* Tambawala, and *Brent* Wellekens

Abstract. A detailed quantitative assessment of the financial risk of loans issued to small businesses is essential for rating agencies, lenders, investors and other stakeholders to inform investment decision making. In this paper, an approach to assess the risk of a pool of loans using modern machine learning techniques like logistic regression and deep neural networks has been developed. The analysis has been conducted on a dataset of over 150K small business loans originated by the SBA from 1990 to 2014. Two approaches for incorporating macroeconomic features have been explored and the superior discretized-time approach has been used to find the default probability on a 500-loan test dataset. The 1-year and 5-year capital value-at-risk for this test dataset at the 99% and 95% confidence level have been evaluated. Lastly, the impact on different investors has been shown by finding the profit & loss distribution for a junior and senior tranche.

1 Introduction

Small businesses are the backbone of the economy, with 28 million small businesses in the US accounting for nearly 54% of all US sales and providing 55% of all jobs [1]. Despite their vital importance to the US economy, there is little understanding of their credit risk. This is due in large part to most small businesses being privately held companies that are not required to publish public quarterly or yearly financial reports. As a result, data on small business borrowers is sparse and accurately predicting risk is difficult. In light of this, creating advanced models that estimate default probability of loans to small businesses is of great value to loan originators, rating agencies, private investors and other shareholders.

In this paper, modern supervised machine learning techniques were applied to a data set of over 150 thousand small business loans originated by the SBA (Small Business Administration). The goal was to predict the probability of default for a loan using static features available in the SBA data set as well as other external features such as time dependent macro economic covariates. Then, the loss at default and the capital value at risk distributions were estimated for a portfolio of 500 randomly selected loans.

2 Data Exploration

Before delving into the bulk of the data modeling, exploratory data analysis was performed on the loan dataset. The first part of the data exploration section looks at the static loan specific information provided by the SBA. The second part looks at the effect of time-varying, macroeconomic features on loan default rate. Performing this data exploration gives a top-level understanding of which features are presumed to be statistically relevant when building the classification models.

2.1 Static Loan Variables

The data set provided by the SBA initially contained 29 static features for each of the loans originated to the small business borrowers. These features include borrower specific information such as firm location, industry type, term of loan etc. Pivoting the dataset led to some remarkable insights.

Figure 1 illustrates the proportion of loans that defaulted in each state using a heat map. It indicates that there are clear differences in credit risks between states. Michigan, Florida and Georgia, for example, are the riskier states for loans (30% loan defaults) whereas Hawaii, Montana and North Dakota are the safest with less than 10% defaults.

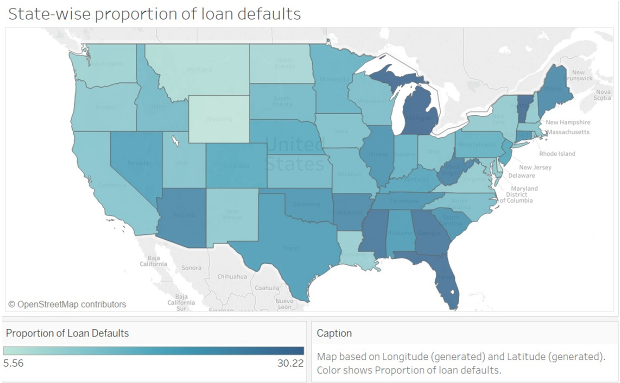


Figure 1: State wise proportion of loan defaults

Figure 2 shows the proportion of loans defaulted in different industries. The proportion of loans defaulted was highest in the Real Estate, Rental and Leasing, and Accommodation and Food Services sectors, and are thus the riskiest industries. Utilities, Transportation and Government sectors had far lower default rates.

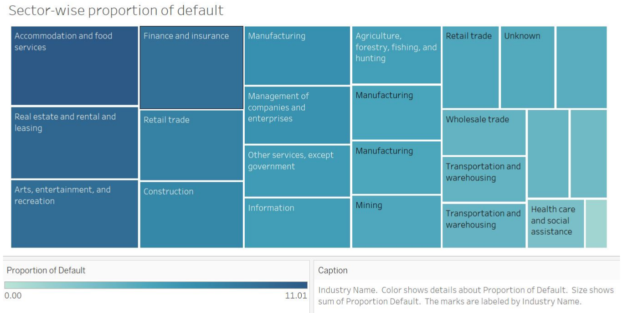


Figure 2: Sector wise proportion of default

Figure 3 highlights that a majority of the defaulted loans had mid low range gross approval amounts.

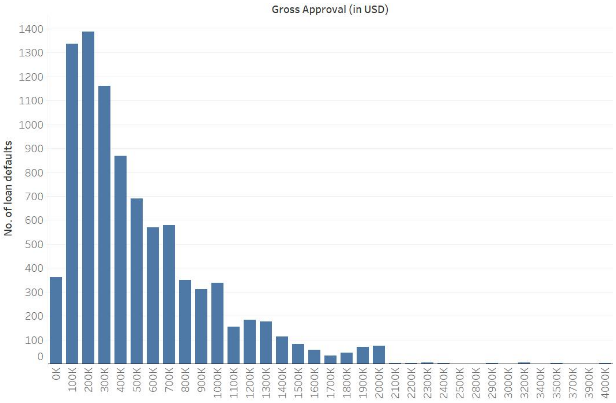


Figure 3: Loans Defaulted vs Gross Approval

Perhaps the most noteworthy insight is the strong time dependence of default rate. There is a strong increase in default rate from around 2007-2008 onwards which coincides with the years of the financial crisis, as shown in Figure 4. This relationship between time and default rate motivated exploration of any potential correlations between time varying features and default.

2.2 Time Varying Features

Exploration of macroeconomic data revealed links between the overall trend of the US economy and the loan default rate. Interest rate, HPI (Housing Price Index), and average S&P 500 return in particular serve as useful predictor variables.

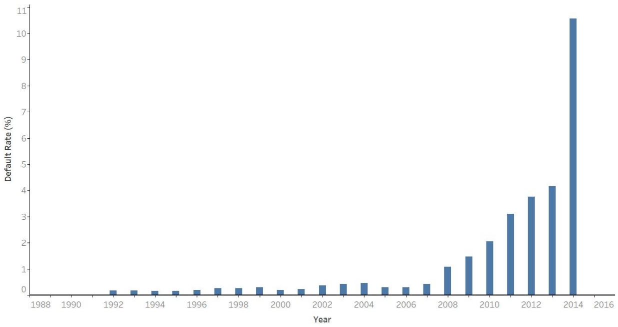


Figure 4: Default rate vs Year of default

Figure 5 shows that default rates are typically higher when interest rates are lower. This may be occurring because interest rates tend to be lower during periods of economic downturn and tend to be higher when the economy is growing.

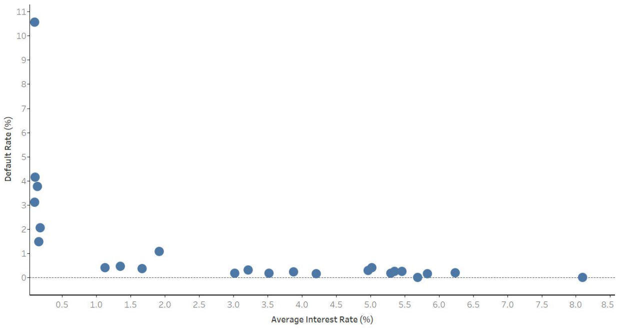


Figure 5: Default rate vs FED Interest rate

Figure 6 shows the default rate versus housing price index plot. It can be observed that high default rates are clustered around average house price indexes between 300 and 350. This could be due to the fact that as observed in section 2.1, the industries most prone to loan defaults have real estate on their balance sheets.

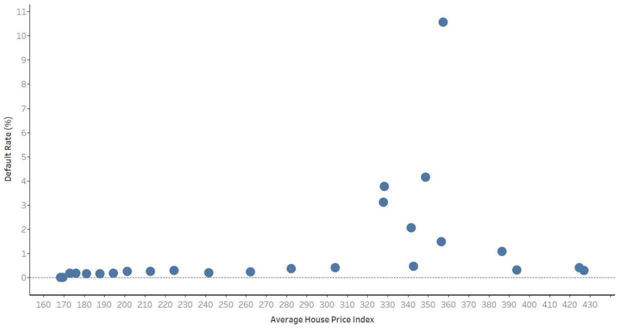


Figure 6: Default rate vs Average housing price index

Figure 7 shows that the default rate is significantly lower when the S&P 500 return is negative. This may seem counterintuitive on the surface but could be due to the fact that the S&P 500 returns are based on yearly averages. Thus there could be a time phase between the burst of an economic bubble and the loan defaults.

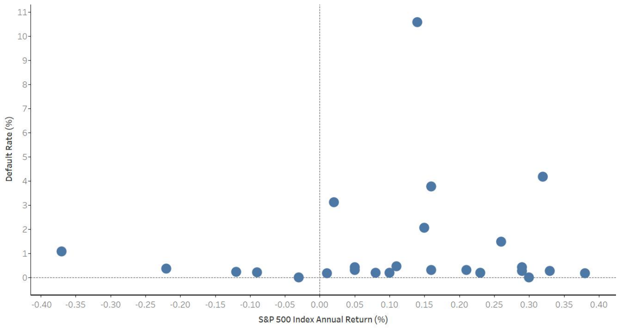


Figure 7: Default rate vs Average S&P 500 return

Now that a top-level overview of the influence of various static and time varying features has been established. The dataset can be cleaned, partitioned and the most statistically significant features can be selected as inputs for the classification models.

3 Feature Engineering

3.1 Data cleaning & sample structuring

The original dataset provided by the SBA contained over 150,000 loans. For the purpose of this investigation, only the loans that were either paid in full or charged off (defaulted) were considered, reducing the number of loans to approximately 54,000.

Macroeconomic variables (HPI, Interest rate, etc.) discussed above were added to the dataset according to the industries the loans were in and the states that they were originated in. For the loans in the regions where the macro economic variables were not available, or the loans did not have NAICS information, the national average was used instead.

Since these macroeconomic features did not stay fixed over time, a discrete-time formulation was adopted (for periods in years) to leverage the predictive power of the time varying features and to address the loan-to-loan correlation due to economic factors. Table 1 illustrates the approach used to discretize each loan into multiple samples for each time-period of the loan for the purpose of prediction.

3.2 Data Partitioning

Due to the fact that most of the defaults occurred in the later years of the dataset as shown in Figure 4, the training, validation and test set were created using the random scrambling technique. This ensured that each data set contained an approximately even amount of defaulted loans on a percentage basis. It also ensured that elements of all time periods, financial crisis in 2007-2008, for example, were included in each data set.

Table 1: Extract of the time discretized data set

	Start	Stop	Interest Rate (%)	Borrower State
Loan 101	1990	1991	8.10	NY
Loan 101	1991	1992	5.69	NY
Loan 101	1992	1993	3.52	NY
Loan 102	1990	1991	8.10	NC
Loan 102	1991	1992	5.69	NC
Loan 102	1992	1993	3.52	NC

Overview of the data partitioning:

Training set = ~ 80% total data
Validation set = ~ 20% total data
Test set = 500 randomly selected loans

The test set consists of loans that have a term of at least 60 months in order to predict losses for both a 1- and 5-year horizon.

3.3 Feature Selection

It is essential to select the appropriate features for prediction in order to avoid over-fitting the training data-set. There are multiple approaches to find the optimum set of features for prediction. In this paper, 2 methods were used, L_1 regularization and the variable importance plot using classification trees, to extract the most statistically significant features.

L_1 regularization (Lasso regularization) on the discrete-time dataset is an effective technique which can be used to select the important features for the model. Figure 8 shows the results of running the L_1 regularization on this dataset and its effect on the cross-validation error. The two vertical lines represent the number of features selected based on min. λ and min. $\lambda + 1se$.

A variable importance plot derived from classification trees is another technique that can be used to find the features that play the most important role in prediction. A feature with a higher value for the mean decrease in gini coefficient has a greater role in partitioning the data into the given classes. Figure 9 shows this plot for the current prediction task. In this analysis, features on the left side of the vertical line (with a mean decrease in gini coefficient below 6) have been considered as ineffective for prediction purposes.

If a given feature was proven to be effective in both of these methods, it was considered to be statistically significant and included in the subsequent prediction models. This selection method tries to ensure that a good balance between variance and bias is achieved. The selected

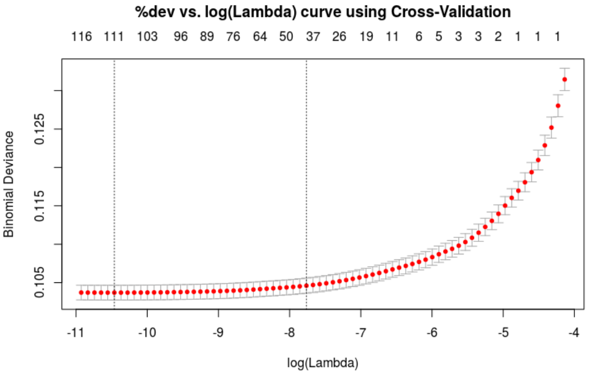


Figure 8: L_1 regularization

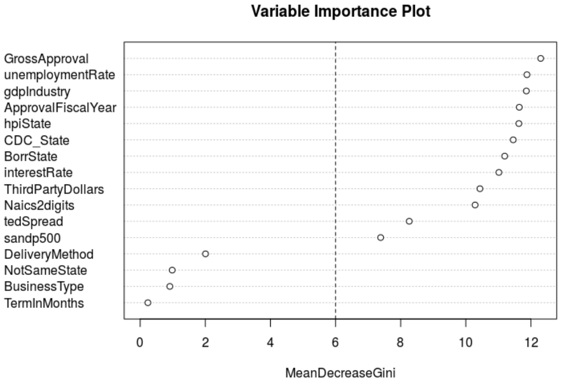


Figure 9: Variable importance plot

feature variables are shown in Table 2. A full overview of which variables were selected for each method can be found in the appendix.

Table 2: Selected static and time varying covariates

Static	Time Varying
Approval fiscal year	Industry GDP
Borrower state	HPI per state
Gross loan approval	Federal interest rate
NAICS first 2 digits	Average S&P 500 return
Not same state	Ted spread
Third party dollars	Unemployment rate by state

4 Model Selection

In order to predict the probability of default of the loans in the SBA dataset, two different models, logistic regression and deep learning neural nets with multiple layer sizes, were used. The prediction was done on the loan data set that was discretized for 1-year time periods as described in section 3. The performance of these discretized-time loan data set models was compared to a baseline model using the static-time loan data set.

4.1 Static-time models for baseline setup

For the purpose of evaluating the models, AUC (area under the curve) was used as the main performance metric. First, an AUC based on the validation set for the logistic regression model without macro economic variables was constructed. The AUC of this model was used as a baseline. Then, to investigate the effect of the macro economic variables, a logistic regression model was constructed that included these variables at the time the loans were initiated and an AUC was evaluated that could be used a second baseline. The feature selection procedure discussed in section 3.3 was also performed for this model. The AUC of the first baseline model was 0.828, and the AUC of the second baseline model that included the macro economic variables was 0.847. Thus, it is clear that macro economic variables improved the model performance.

4.2 Discrete-time dataset for logistic regression and deep learning models

During the lifetime of a loan, the loan default event is dependent on macroeconomic factors that vary with time. Thus, it was decided to include these features into the prediction model to improve the model's predictive performance. There are many candidate supervised machine learning models that can be used for this purpose.
Two Model configurations were considered:

1. Simple logistic regression on the discrete time dataset using select features

2. Multiple deep learning models with different number of layers and neuron structure on the discrete time dataset using select features

For the discrete-time formulation, the state of a loan is defined in a given period as U^n_{t-1}, which is the state of the n^{th} loan at time t [2].In the dataset, there are only two possible states for a loan at time t. They were defined as:

$$U^n_t = 0 \ (Not\ in\ Default) \quad U^n_t = 1\,(Default)$$

In both of these discrete-time formulations, the conditional probability of transitioning to default at t was computed. The conditional probability of default for loan i at time t was defined as p^t_i. The probability of default for time $t \leq i$ was defined as $P^n_{def,i}$ and is formulated according to the equation:

$$P^n_{def,i} = 1 - P^n_{surv,i} = 1 - \prod_{i=1}^{i}(1 - p^n_i)$$

Using these computed probability of default values ($P^n_{def,i}$), the test performance and the risk capital were computed.

The discrete time formulation is computationally less-expensive vs. the static time models mentioned above. Another advantage of the discrete time formulation is that it captures loan-to-loan correlation due to common geographic and macroeconomic features.

4.2.1 Model Framework

The model computes the conditional probability of default defined as:

$$P\left[U^n_t = 1 \,\middle|\, F_{t-1}\right] = h_\theta(1, X^n_{t-1})$$

Here, F_{t-1} is defined as all information available at time $t-1$. h_θ is the probability transition function and θ is the parameter to be estimated. In this case, h_θ refers to the neural network architecture. Note that logistic regression is a special case of a neural network architecture in which there are no layers. X^n_{t-1} are the static and time-varying features of loan n at time $t-1$.[2]

In the neural network function, a linear combination of basis functions are repeatedly passed into a simple non-linear link function (soft-max function) to produce highly non-linear functions. Typically, neural networks tend to have a low-bias and high variance which can be corrected through various methods. In this case, regularization and dropout were used to address over-fitting. The parameters used for the deep learning model can be seen in table 3. As stated before, a neural network with zero hidden layers is a logistic regression model.

Table 3: Neural Network Model Parameters

Network Parameter	Attribute
Number of Hidden Layers	1,3,5
Number of Hidden Units	50,150,250
Activation function	Rectified Linear Units with dropout

4.2.2 Computational Implementation

The modeling approach is computationally intensive as the size of the dataset runs into 700K+ samples and the architecture of the deep neural network model is complex. The time to train the deep learning model would be long if typical personal computing resources were used. Hence, virtual CPU resources were utilized (Amazon EC2 Instances) over the cloud from Amazon Web Services. In order to configure an vCPU instance on Amazon AWS, an Amazon EC2 Instance with RStudio Server Machine Image was configured [3]. The "m4.2xlarge" EC2 instance was used, which has 32GB of memory 8 vCPUs.

There are several packages available in R to implement deep learning models. In this case, h2o.ai, (an open source deep learning platform) was used. The h2o deep learning server was configured to run on all 8 vCPUs in parallel. An input dropout ratio of 0.2 was used, and rectified linear units were used as the activation function. The h2o deep learning package provides an easy to manage memory when dealing with big data and it is quickly becoming the most popular deep learning implementation package in the R environment.

4.3 Model Comparison

In order to compare the different modeling approaches, the out-of-sample performance of the model was computed on a portfolio of 500 randomly selected loans described in section 3.2. The receiver-operating characteristic curves were used for this purpose. The area under the curves (AUC) were calculated and used as the metric for out-of-sample comparison. The AUC is a superior metric to mean squared error for comparing the predictive power of the models in this investigation, as it is assumed than an investor cares more about accurately predicting which loans will default (true positive rate) versus identifying certain loans as defaulting even though they survive (true negative rate). First, the AUC of the deep neural network models was computed on a validation set and the best model was then selected. The AUC obtained from the discrete-time formulation was compared with the AUC from the static-time models to find potential improvements in model performance over the baseline. After the selection of the best model, the AUC was computed on the test set containing 500 loans.Table 4 shows an overview of the aforementioned models. The ROC curve for the first three logistic regression models is shown in figure 10.

The logistic regression with time varying macro economic features exhibits the best validation AUC and is more computationally efficient than the deep learning model. The validation AUCs can be compared for the various deep neural network architectures according to figure 11. The logistic regression, which is a special case of the neural network, has the highest AUC. Thus it was chosen as the prime model to predict the probability of default for the test set portfolio consisting of 500 randomly selected loans. There might be multiple reasons behind the better performance of the logistic regression. One possible reason might be that the deep neural network might be over-fitting the validation set. This effect can be controlled by using an ensemble model for the deep learning models to control for the higher variance. Also, in cases where there are many more loans in the dataset than the existing dataset, it might be possible that the deep

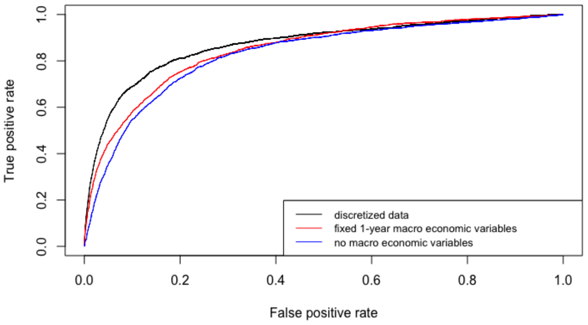


Figure 10: Validation Set ROC curve of the logistic regression models

Table 4: Validation AUC of the Baselines and the Models

Model	Dataset	Validation AUC
Logistic regression w/o macroeconomic features	Static-time	0.828
Logistic regression w. macroeconomic features	Static-time	0.847
Logistic regression w. macroeconomic features	Discrete-time varying	0.871
Deep Neural Network (3 layers, 250 units) w. macroeconomic features	Discrete-time varying	0.858

learning model is better than the logistic regression.

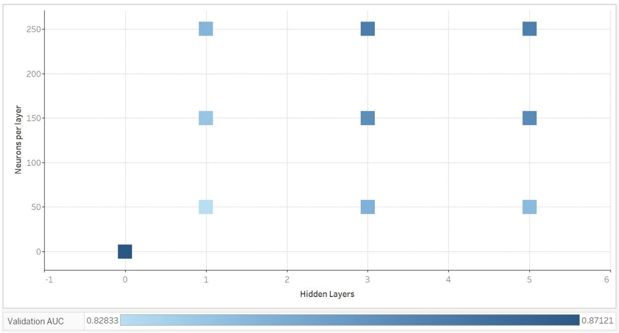


Figure 11: Validation AUC grid for no. of layers vs. neurons per layers

4.4 Out-of-sample test performance

The out-of-sample test performance was found on a dataset containing 500 randomly selected loans for a 5-year horizon. In reality, the macroeconomic features in future periods are unknown in the modeling process. This would directly impact the test performance of loan default prediction models. Thus, to assess the impact of macroeconomic features, two different datasets were used to compute the test set. In the first test dataset, the macroeconomic features observed at the start of every loan were fixed. In the second test dataset, the actual macroeconomic features known beforehand were used. An alternative method to assess the impact of the macroeconomic features would be to use a time-series forecast to predict these features in future periods. Because it is difficult to predict these macroeconomic features beyond a 1-3 year horizon, the former approach was used.

Table 5 shows the AUCs on the test set for these two cases. The test set AUC for the case when the actual macroeconomic features are used is much higher than the one where fixed macroeconomic features are used for future predictions. These results demonstrate the importance of using accurate estimates of macroeconomic features for predicting the probability of default. In practical settings, the macroeconomic features could be obtained using state-of-the-art auto-regression techniques, Winter-Holt's forecast or a simple heuristic forecast based on expert opinion.

Table 5: Test set AUC for Best Model

Test Data Set	AUC
500 loans w. fixed macroeconomic features	0.869
500 loans w. actual macroeconomic features	0.901

5 Loss Given Default

Understanding how much money is lost on defaulting loans provides a more complete understanding of the risk associated with each loan. This is used by investors to quantify the losses that occur when a default happens. In order to estimate this, the ratio between the Charge-Off amount and the Gross Approval was evaluated and subsequently multiplied by the Gross Approval to obtain the Charge-Off amount. In order to do so, the datasets were filtered to only include defaulting loans, and the following three models were considered for estimating loss:

1. Linear Regression
2. Bagging (Decision Trees)
3. Polynomial Splines

All of these models utilize the macro-economic features discussed earlier. Additionally, model 1 served as a benchmark for the other two models. All the models were evaluated using the MSE on the validation set, as this was a regression problem rather than a classification problem.

Table 6: Model MSE Comparison

Model	MSE
Linear Regression	0.090
Bagging	0.031
Natural Regression Splines	0.071

Table 6 shows that bagging displays a significantly better MSE than the other two models, and was thus selected for use in subsequent portions of the experiment.

6 Portfolio Analytics

Investors are interested in analyzing and estimating the risk associated with their investments. In this case, the investment consists of a portfolio of 500 randomly selected loans originated by the SBA. In order to determine what the potential downside of this investment could be, a loss at default distribution was generated using Monte Carlo simulation. This loss distribution was subsequently used to determine the 99% and 95% Value at Risk parameters and the loss distribution for different tranche classes. These are important parameters to know for the investor as it gives an indication of how much capital should be retained in order to cover monetary losses.

6.1 Monte Carlo Simulation

The Monte Carlo simulation is comprised of three steps that were performed on each of the 10000 iterations:

1. Generate a random uniform variable V
2. For each loan n, if $P(U_i^n = 1) \geq V$ then the loan defaults (i is the time horizon)
3. Estimate the loss per loan that defaults and sum over all defaulted loans in the portfolio

Parametric and non-parametric approaches were taken to determine the charge-off amounts for each loan. The parametric approach takes the feature values for the specific loan as input to the bagging model previously developed and uses the outputted ratio to estimate the Gross Charge-Off Amount. The non-parametric approach uses the recorded Gross Charge-off amounts to generate the loss distribution.

The 1-year and 5-year losses generated from a parametric and non-parametric approach were then used to generate: loss distributions, 95% and 99% VaR distributions, and tranche loss distributions for each respective time horizon.

6.2 Loss Distributions

After the 10000 simulations were performed, plots of the parametric and non-parametric loss distributions were generated for both 1-year and 5-year time horizons. In both instances, loss amounts of 0 were removed from the loss distributions only if no default events occurred during the iteration. This was done to condition the distributions on scenarios where a default occurs.

Figure 12 shows the 1-year loss distributions from the simulation using parametric and non-parametric approaches. The non-parametric distribution only contains Charge-off values of 0, while the parametric distribution had a strong peak at 0 and was strongly skewed right. (Explain the standard normal for non-parametric)

Figure 13 shows the 5-year loss distributions. The parametric curve is extremely similar to the non-parametric curve. Additionally, both are right-biased, although less so than for the 1-year loss distributions.

Both time horizons show a clear bias towards 0 losses. This can be explained by the fact that most of the 500 loans did not default, and among the loans that did default, several did not default within the first five years or had a 0 charge off amount. (If needed, use filler statement for VaR and Tranches).

6.3 Value at Risk Distribution

When underwriting loans, banks want to ensure that in the event of defaults they have enough funds saved to cover their losses. Value-at-risk (VaR) is a metric that is used to determine how much money should be retained. For a given significance level α, the VaR is the $100 - \alpha$ percentile within the distribution of losses.

Using the previous loss values, a bootstrapping procedure comprised of the following steps was iterated 10000 times:

1. Bootstrap the loss values
2. Calculate the Value-at-Risk at the 95% and 99% significance level
3. Generate 95% and 99% confidence intervals for the Value-at-Risk

4. Calculate an Average Value-at-Risk for both significance levels

Figure 14 shows the parametric and non-parametric 1-year VaR distributions for a 95% significance level. The 95% confidence interval for the parametric distribution is [0.157, 0.157] and the Average VaR (AVaR) is 0.157. The 95% confidence interval for the non-parametric distribution is [0, 0] with an AVaR of 0.

Figure 15 shows the parametric and non-parametric 1-year VaR distributions for a 99% significance level. The parametric distribution has a 99% confidence interval of [0.157, 0.157] with an AVaR of 0.157. The non-parametric distribution has a 99% confidence interval of [0, 0] with an AVaR of 0.

Figure 16 shows the parametric and non-parametric 5-year VaR distributions for a 95% significance level. The 95% confidence interval for the parametric distribution is [0.15743, 0.15753] and the AVaR is 0.15749. The 95% confidence interval for the non-parametric distribution is [0.1276, 0.129] and the AVaR is 0.1287.

Figure 17 shows the parametric and non-parametric 5-year VaR distributions for a 99% significance level. The parametric distribution has a 99% confidence interval of [0.1574, 0.1574] and the AVaR is 0.1574. The non-parametric distribution has a 99% confidence interval of [0.1344, 0.1344] and an AVaR is 0.1344.

The parametric and non-parametric distributions provide different insights into how much money needs to be saved. The parametric distribution provides higher estimates for the VaR than the non-parametric distribution. If one were to assume the worst-case, then they would look at the statistics from the parametric distribution and withhold approximately 16% of the total loan pool.

However, this procedure provides unreliable results. As mentioned before, the 500 loans are biased towards not defaulting and having 0 charge-off. As a result, most of the loss values are close to each other and close to 0, with a small number of values significantly larger. As a result, when the quantile is being estimated, there are not enough values to provide a precise percentile, and the maximum is selected instead as a rough approximation. In order to provide more robust VaR estimates, a larger set of loans needs to be examined such that the ratio of defaulting loans to non-defaulting loans is more equitable.

Table 7: AVAR/Confidence Intervals

Par/Non-par	α	horizon	AVaR	CI-Low	CI-High
Par	95	1	0.157	0.157	0.157
Non-par	95	1	0	0	0
Par	99	1	0.157	0.157	0.157
Non-par	99	1	0	0	0
Par	95	5	0.15749	0.15743	0.15753
Non-par	95	5	0.1287	0.1276	0.239
Par	99	5	0.1574	0.1574	0.1574
Non-par	99	5	0.1344	0.1344	0.1344

6.4 Securitization and Tranche Loss Distributions

Oftentimes, banks seek to profit further from these loans by securitizing them. When this is done, several different tranches, or separate securities, are created, where more senior tranches are prioritized within the pool of funds from the loans. Thus, the price of the different tranches are affected separately by losses. More junior tranches are hit by losses to the funds first, and once they are wiped out, the following tranches are hit in order of seniority. In this section, the loss distributions of two tranches are considered: a junior tranche encompassing funds in the [5%, 15%] quantiles of total funds, and a senior tranche encompassing funds in the [15%, 100%] quantiles of total funds.

The bootstrapped losses can be used to measure the loss distribution of both tranches. These losses can also be measured with respect to the amount of money in each of the tranches. If the loss amount surpasses 5% of the original loan funds, then that would correspond to a loss in the junior tranche. Likewise, if the loss amount surpasses 15% of the original loan funds, then that would correspond to a loss in the senior tranche.

The 1-year parametric plots show that roughly all of the losses will not hit the senior tranche, and roughly 90% will not hit the junior tranche. Within the remaining 10% of losses impacting the junior tranche, the loss percentages to the tranche are roughly evenly distributed. The 1-year non-parametric plots show that none of the losses impact the junior or senior tranche.

Both the parametric and non-parametric 5-year tranche plots show roughly all of the losses not impacting the senior tranche and 50% not affecting the junior tranche. Both plots also show the spread of loss percentages among the losses impacting the junior tranche to be roughly uniform.

The parametric 5-year tranche distribution closely resembled the 5-year non-parametric distribution. This confirms that parametric loss estimators are effective in measuring the actual loss distribution of securities derived from securitizing the test set portfolio. Additionally, the plots indicate that a 1-year time horizon is much less risky for investors than a 5-year time horizon. Within a 1-year time horizon, there is only a 10% risk of the junior tranche incurring losses and a 1% risk of the senior tranche incurring (minimal) losses, and within a 5-year time horizon, the risk percentages shift to 50% and 1%, respectively. Therefore, if an investor seeks to minimize their risk of loss, they should hold shares in the junior tranche until a year after approval, and should hold shares in the senior tranche for a long-term period.

7 Conclusion

The logistic regression model with macro economic variables using the discrete time formulation proved to be the best model. It outperformed the baseline logistic regression model and deep learning models that were considered.

Incorporating time varying elements such as macro economic variables into the probability of default classification models improves their predictive accuracy. This is likely due to the fact that the macro economic variables that were chosen (FED interest rate, average HPI, unemployment rate, average return of S&P 500 etc) capture the overall state of the economy. The financial health of small businesses is directly influenced by this. The key drawback of using time varying elements to predict probability of default across future time horizons is that they themselves need to be forecast as the future is unknown. This means that the overall predictive power of the logistic regression model using discrete time formulation is proportional to the accuracy of the time varying feature regression models.

The parametric approach to determine loss at default distributions, capital value of risks and tranche loss distributions laid out in the portfolio analytics section of the report is a sophisticated, more granular, alternative to the non parametric (less computationally expensive) approach often used in the financial industry by rating agencies.

8 References

1. "Small Business Trends | The U.S. Small Business Administration | SBA.Gov". Sba.gov. N.p., 2017. Web. 24 Mar. 2017.

2. Sirignano, Justin, Apaar Sadhwani, and Kay Giesecke. "Deep Learning For Mortgage Risk". (2016): n. pag. Web. 25 Mar. 2017.

3. Aslett, Louis. "Rstudio Server Amazon Machine Image (AMI) - Louis Aslett". Louisaslett.com. N.p., 2017. Web. 26 Mar. 2017.

9 External Feature Data Sources

"Annual Unemployment Rates By State | Iowa Community Indicators Program". Icip.iastate.edu. N.p., 2017. Web. 24 Mar. 2017.

"BEA : Gross-Domestic-Product-(GDP)-By-Industry Data". Bea.gov. N.p., 2017. Web. 24 Mar. 2017.

"Download Data - Effective Federal Funds Rate (DFF) | FRED | St. Louis Fed". Fred.stlouisfed.org. N.p., 2017. Web. 24 Mar. 2017.

"Download Data - TED Spread (TEDRATE) | FRED | St. Louis Fed". Fred.stlouisfed.org. N.p., 2017. Web. 24 Mar. 2017.

"House Price Index Datasets | Federal Housing Finance Agency". Fhfa.gov. N.p., 2017. Web. 24 Mar. 2017.

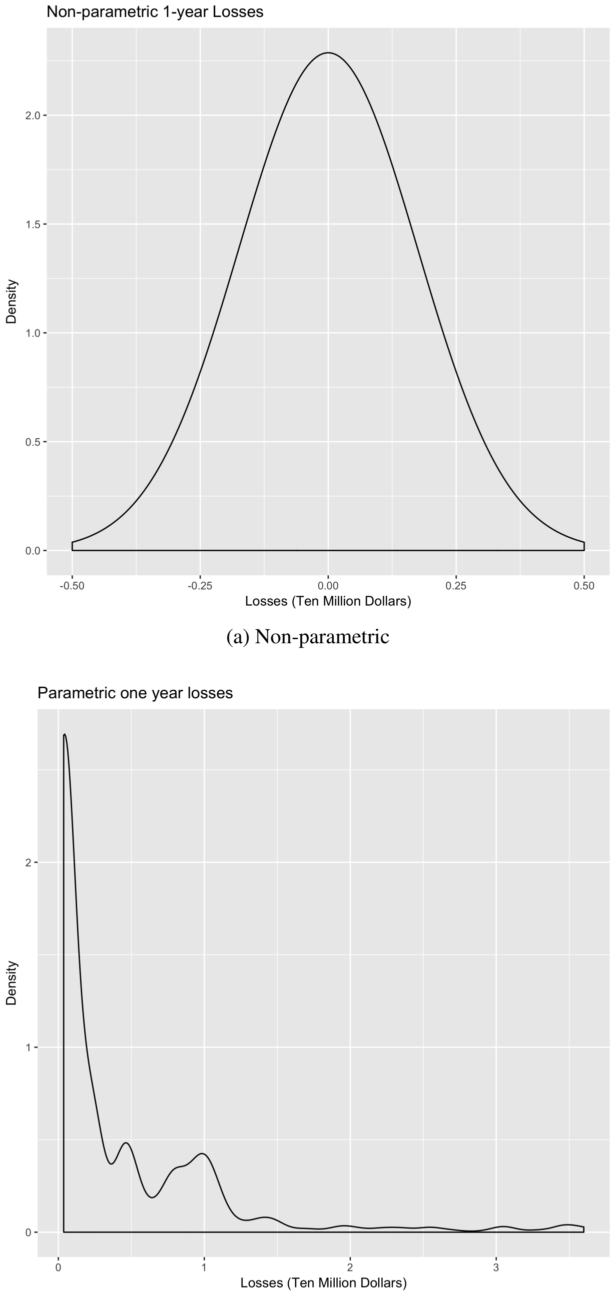
"S&P 500 Annual Total Return (Yearly)". Ycharts.com. N.p., 2017. Web. 24 Mar. 2017.

10 Appendix

Table 8: Overview of feature variables extracted at each step in covariate selection method.

Feature Variable	**L_1 Regularization**	**Variable Importance**
Approval fiscal year	✓	✓
Borrower state	✓	✓
*CDC state	✓	✓
*Not same state	✓	
Business type		✓
Delivery method	✓	
Industry GDP	✓	✓
Gross loan approval	✓	✓
HPI per state	✓	✓
Federal interest rate	✓	✓
NAICS 2 digit	✓	✓
Average S&P return	✓	✓
Term in months	✓	
TED spread	✓	✓
Third party dollars	✓	✓
Unemployment rate	✓	✓

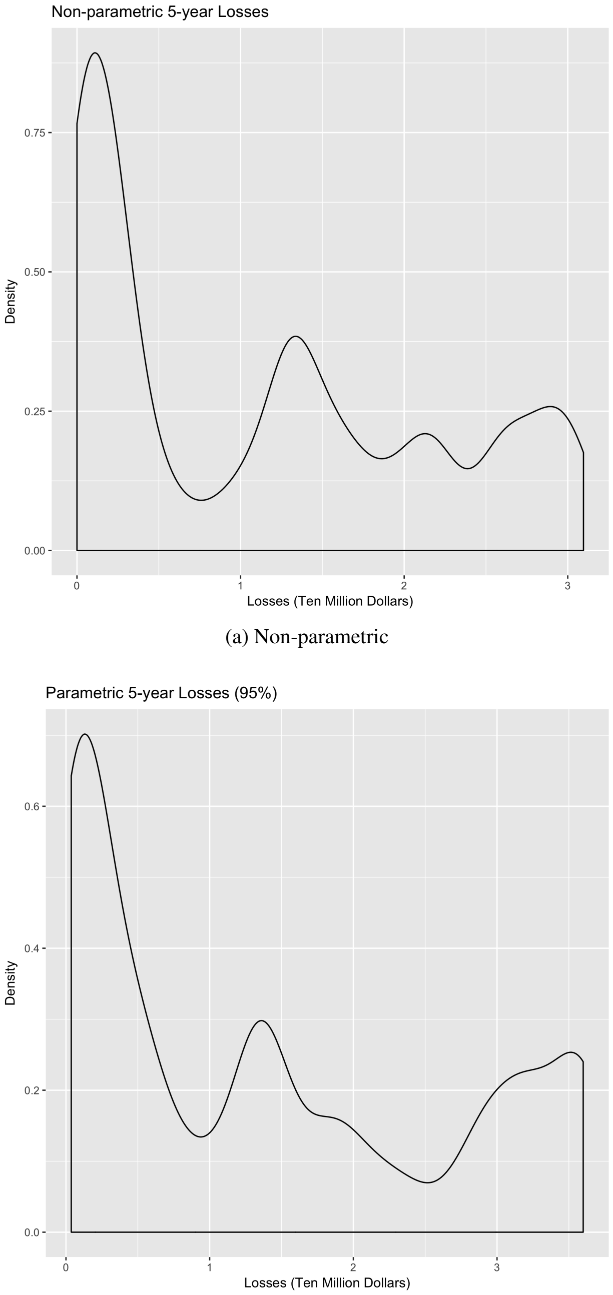
*While Not Same State did not appear to be significant in the variance importance plot , it was decided to include it rather than CDC State because including Borrower State and Not Same State is the same as including Borrower State and CDC State, and so Not Same State would be a more efficient choice.



(a) Non-parametric

(b) Parametric

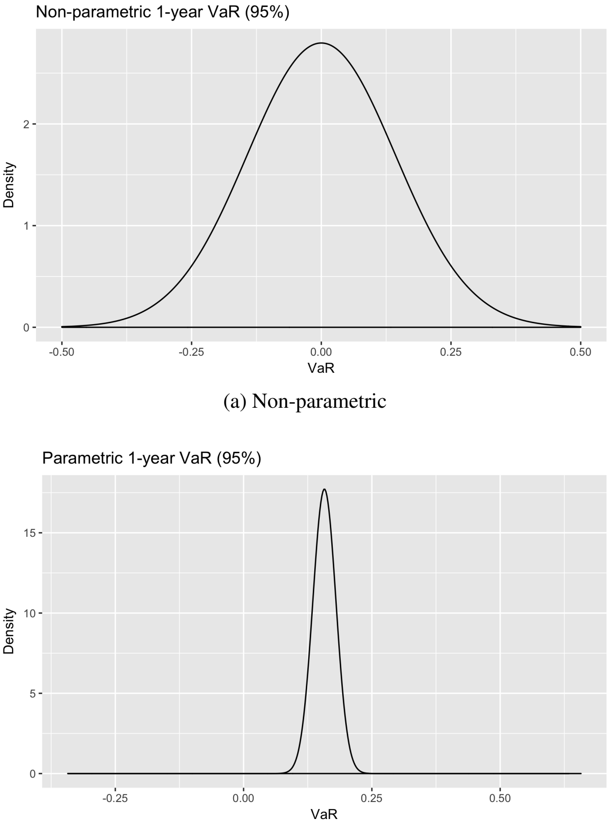
Figure 12: 1-year loss distributions



(a) Non-parametric

(b) Parametric

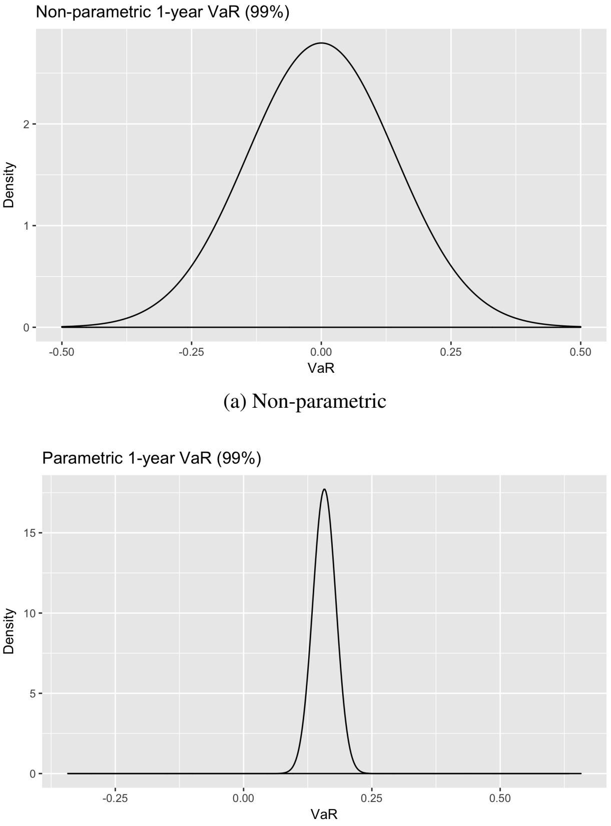
Figure 13: 5-year loss distributions



(a) Non-parametric

(b) Parametric

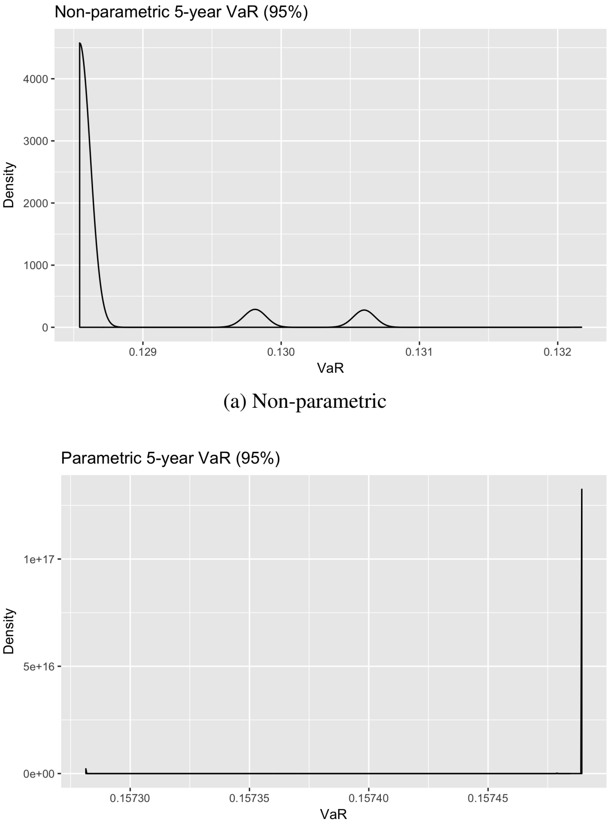
Figure 14: 1-year VaR distributions (95% Level)



(a) Non-parametric

(b) Parametric

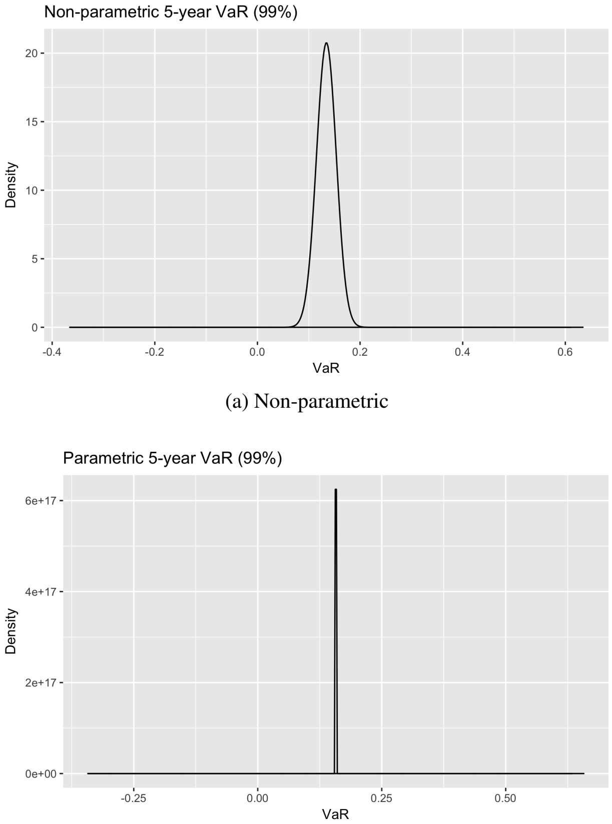
Figure 15: 1-year VaR distributions (99% Level)



(a) Non-parametric

(b) Parametric

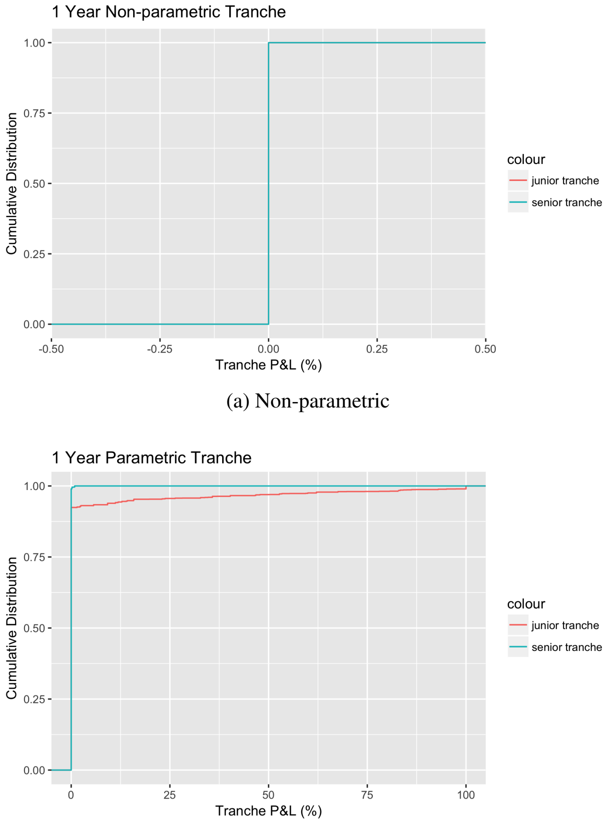
Figure 16: 5-year VaR distributions (95% Level)



(a) Non-parametric

(b) Parametric

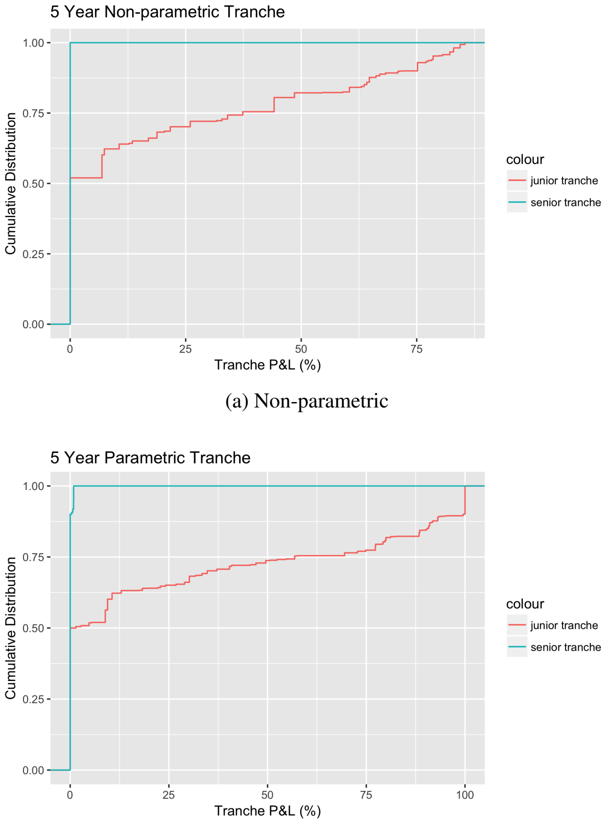
Figure 17: 5-year VaR distributions (99% Level)



(a) Non-parametric

(b) Parametric

Figure 18: 1-year Tranche distributions



(a) Non-parametric

(b) Parametric

Figure 19: 5-year Tranche distributions