

Financial Risk Analytics Project Report

Denis Bykov, Iryna Kryvda, Scott Murff, David Odgers, Brian Wilcox

1 Introduction

The group was provided with a dataset of approximately 150 thousand SBA loans to work with in analysis. Of those initial loans, about 95 thousand had a final status of canceled or exempt. Not including these loans, the group worked with the remainder of these loans focusing on a variety of tasks. The first task was to systematically evaluate and explore each of the 30 variables provided in the initial dataset. Then we collected and abundance of additional data enabled us to greatly increase the quality of modeling results. The next step was to deal with missing values and perform imputation wherever necessary in order to have a 'complete' dataset to work with. From this point on, the group was set to focus on three kinds of modeling tasks. The first of these came in modeling the Probability of Default over various time spans. Using these calculated probabilities, the next modeling step was to fit a loss given default model. With these two fitted models we are able to predict the loss of a loan given that a simulated default event occurred. The last modeling step of this project came in the form of modeling the Value at Risk and Average Value of Risk (also called shortfall) of a simulated distribution of loans. From these Value at Risk values, we then correlated that to a multi-tranche loan distribution evaluating how such losses would impact a Junior-level and Senior-level investor.

2 Data Preparation

2.1 Data Cleaning and Exploration

The baseline dataset that we were provided had observations on 147,423 SBA 504 program loans that were originated between 1990 and 2014. Of these 72,836 were classified as "EXEMPT" which means they are "exempt from disclosure under FOIA Exemption 4." In addition 1 loan had a missing value for the Loan Status and 19,780 of the loans were canceled. Without sufficient information to be useful for modeling, we delete these three categories of loans from the dataset leaving us with 54,806 loans in our dataset.

Of the 54,806 loans there are 8,982 (16%) defaults and 45,824 (84%) loans that are paid in full. In total this provides a sizable although still relatively small data set from which to model probability of default and loss given default so we need to be judicious about how many covariates we allow into the model. The dataset is somewhat more limited in the number of high quality covariates that we can use to predict. In the remainder of this section we review and explore the baseline covariates. In the next section we describe how we sought to enhance the dataset by merging in time-varying and geographic specific covariates from a wide array of sources.

In total there were 30 variables provided in the initial dataset. We systematically evaluated each variable to explore its range of values, missing-ness, etc. In the table below we list each variable and it's associated treatment in a baseline LGD model based on our evaluation. A few variables of note are the following:

CDC_State is the same as the same as BorrState 94% of the time. Since this variable won't add much new information, and including an extra 50+ dummy variables with a relatively small dataset for fitting that many parameters, we chose to not include this variable.

InitialInterestRate seems like it would be a very useful variable for prediction and has a very intuitive link to probability of default and loss-given default. Unfortunately it is poorly populated with non-missing values only 2% of the time. Furthermore, in the non-missing cases the only unique value is 3.25 which provides no variation that would be useful for prediction. For these reasons we did not include this variable.

NaicsCode is a variable that provides valuable information about which industry the borrowing business is in. The variable as provided contains all 6 digits which is the most specific categorization possible. There are a total of 1,080 industries in the dataset when using the 6 digits code. This is far too many to be of use in fitting a predictive modeling when you consider the relatively small dataset of 55k loans. Instead using the the 6 digits NAICS code we instead use the first 2 digits which segments the data into 26 categories defined by broader industries than that the full 6 digit code. The two most common 2 digit industries are 44 and 72 which correspond to the Retail Trade and Accommodation and Food Services industries respectively. Each of these categories accounts for 8% of the loans. The NAICS code is missing for 31% of the loans. In these cases as discussed in the missing data section below, we create a new category called "MISSING" in order to avoid dropping these loans from the data set.

ProjectState is the the same as BorrState for 99.5% and therefore provide virtually no new information and is not worth including in the model given the relatively small dataset available to fit model parameters.

LoanStatus is a key variable from which we define our response variable for the probability of default model. As noted earlier, of the 54,806 loans there are 8,982 (16%) defaults and 45,824 (84%) loan that did not default.

We use **ThirdPartyDollars**, **GrossApproval**, **GrossChargeOffAmount**, to define our response variable for our loss given default model. We do this by taking the following steps:

1. We interpolate Missing Values of ThirdPartyDollars analyzing the ratio ThirdPartyDollars/GrossApproval when ThirdPartyDollars is not missing. Taking the median ratio we fill missing values of ThirdPartyDollars with MedianRatio*GrossApproval. We believe this is a very intuitive and valid way to interpolate the missing values since the structure of 504 loans would suggest a value of $.5/.4=1.25$. We calculate and use the median value of 1.22. The mean is 1.39 but is skewed slightly higher than the expected

1.25 by a few outliers. Note that without interpolation nearly 70% of loans have a missing value for ThirdPartyDollars. Without this interpolation the dataset would be severely limited in helping build a loss given default model.

2. After the interpolation, we define LGD as follows

$$LGD = \min\left[\frac{ChargeOffAmount}{ThirdPartyDollars + GrossApproval}, 1\right] \quad (1)$$

3. We cap LGD at 1 as shown with the min function above since in general we do not expect chargeoffs to be greater than the entire loan amount at origination.
4. We show a histogram of LGD below which can be seen to be bimodal with peaks at 0 and 0.4. It's interesting to note that the SBA typically guarantees the first 40% of losses for the CDC partner who typically sits in 2nd lien position but is made whole by the SBA in the event of loss. Hence it would appear the SBA is taking large losses while the Third Party Lender is typically protected.

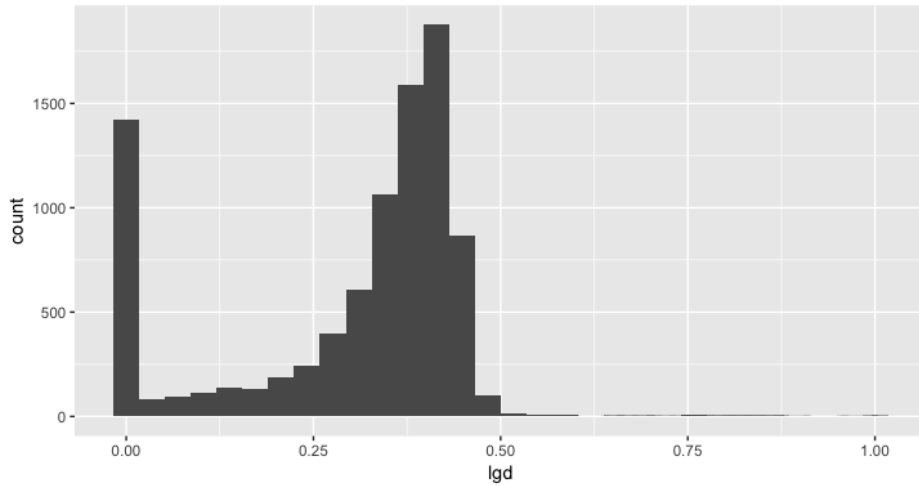


Figure 1: Histogram of LGD

We summarize our treatment of all of the variables provided in the initial dataset in the table below. In the next section we discuss additional data that we merged in to expand the richness of our covariate set.

Variable Number	Usage in Expanded Model	Usage in Baseline Model	Variable Name
1		Drop	Program
2		Drop	BorrName
3		Drop	BorrStreet
4		Drop	BorrCity
5		Dummies	BorrState
6	Use as a merge key	Drop	BorrZip
7		Drop	CDC_Name
8		Drop	CDC_Street
9		Drop	CDC_City
10		Drop. 94% the same as BorrState	CDC_State
11	Use as a merge key	Drop	CDC_Zip
12		Drop	ThirdPartyLender_Name
13		Drop	ThirdPartyLender_City
14		Dummies	ThirdPartyLender_State
15		Use to Derive Depdent Variable	ThirdPartyDollars
16		Use to Derive Depdent Variable	GrossApproval
17	Use as a merge key	Drop	ApprovalDate
18	Descriptive but won't be used in model	Drop	ApprovalFiscalYear
19		Dummies	DeliveryMethod
20		Dummies	subpgmdesc
21		Drop. Only has two values: NA, 3.25	InitialInterestRate
22		Dummies	TermInMonths
23		Drop. Create Dummies using first 2 digits	NaicsCode
24		Descriptive but won't be used in model	NaicsDescription
25		Drop	ProjectCounty
26		Drop. 99.5% the same as BorrState	ProjectState
27		Dummies	BusinessType
28		Drop in LGD since all loan will have CHGOFF	LoanStatus
29	Use in a hazard model.	Drop	ChargeOffDate
30		Use to Derive Depdent Variable	GrossChargeOffAmount

Figure 2: Baseline Variable Treatment

2.2 Merging in Other Data

Our initial hypotheses around the nature of the loan defaults suggested that macroeconomic indicators would prove helpful in classification. Moreover, we expected to see a lot of variance in these economic indicators caused by location. Our aim was to find highly granular data that went down to the county level and was reported monthly. As a result, we used mostly public government databases that spanned the majority of 1990 to 2014, the timespan of the SBA loans. Out of those datasets we got a diverse set of indicators ranging from agricultural spending to the number of patents filed. On top of the public data, we also merged in private data curated by Zillow.

One of the data sets came from the USDA, which provided information on agricultural research development. It recorded yearly spending in both the private and public sectors covering 1990 through 2009. The covariates included nominal spending values and inflation adjusted numbers. For classification the data was merged based on the approval year. For hazard modeling, the yearly value was used for all months within that year. When exploring the agricultural research data, an interesting trend appeared: agricultural spending in the private sector significantly declined a year before an economic recession.

Another dataset we included looked at types of jobs in the US at different snapshots in time. The data was collected decades apart leading to many complications with filling the missing data. It was recorded at the county level and looked at total number of people employed, the number of people in creative jobs, and the number of artists. We merged it in by county name and used a constant value across all years within each decade. The lack of continuous data across the SBA loan years resulted in these covariates not being useful for classification or hazard modeling.

A major hypothesis was that intellectual property would be a good indicator of business health. In other words, areas with more patents filed are less likely to default. We merged in patent data that looked at the yearly aggregate of patents filed per state. To relate each state to one another we created a covariate looking at each state's fraction of patents filed that year. The data was merged in by state and approval year, however it only spanned 2002-2014. Nonetheless, these covariates didn't seem to have any effect on our classification model and were omitted from the hazard model.

We also hypothesized that affluent areas would be less likely to experience SBA defaults. To find a proxy for the affluence of an area we used farmer's markets. The covariates were aggregates of the number of regular and organic farmer's markets within a zipcode. Since the farmer's market data was only available after 2014, we assumed that those farmer's markets were established much earlier and therefore valid for most of the SBA data. As a result we merged in the data by zipcode alone. Nonetheless, these covariates didn't seem to have any effect on our classification model and were omitted from the hazard model.

As general economic indicators we relied on the S&P 500, Consumer Confidence Index (CCI), and Business Confidence Index (BCI). The S&P 500 tracks the 500 largest equities in the US stock market and therefore displays US economic conditions rather well. For the S&P 500 we used the daily high, low, open, close, adjusted close, volume, and returns. The S&P 500 data was merged in by the approval date for classification, whereas the first value of a month was used for the hazard model. The CCI is the result of a national survey of consumers on market conditions that's updated monthly. In addition, the BCI complements these indicators by surveying enterprises on production, current market events, and their future outlooks. Both the BCI and CCI were merged in by the approval month for classification and by the month in the hazard model. All of these datasets spanned the full length of the SBA data (1990-2014).

Another important economic indicator was the unemployment rate. We combined several datasets to include the unemployment rate at the county level year by year, as well as the rate across the entire US month by month. Data was merged in by county and approval year as well as the approval month for hazard modeling. The unemployment data spanned 1990 through 2014.

In order to gauge the health of different US counties, we looked at educational spending. We extracted the county and state level finances of public school districts as reported by the US Census Bureau. The data was available from 1992 onwards in yearly increments for both state and county. Since there were some-

times multiple school districts per county, the data was averaged within each county. Both the county and state level data included covariates such as revenue from different sources and spending by category. For classification, the data was merged by project county by year, and project state by year. Similarly, for the hazard model a yearly constant was used.

For gauging local conditions, income tax data was quite helpful. This data was provided from the IRS and compiled over 2004-2014. It included the number of exemptions, number of returns, adjusted gross income, salaries, capital gains, and other fields commonly found on personal income tax forms. This data was merged in at the zipcode level by year for classification. For hazard modeling the yearly value was held constant across every month within that year.

Lastly, we merged in a private dataset from Zillow containing real estate values. Zillow provided several datasets including their seasonally adjusted home value index, median home prices per square foot, percent of homes increasing in value, and percent of homes decreasing in value. The Zillow data mapped several regions within a county, leading us to average the data across counties to achieve one value per county. The data spanned 1996 through 2014. For classification we merged by approval month and county, and for the hazard model we used county and current month.

2.3 Missing Value Handling

For the baseline set of variables discussed in the data preparation and cleaning section above we generally took the approach of creating a new level of categorical variable called "MISSING." This enabled us to avoid dropping rows when a particular variable was missing and also allowed us to assess whether the category "MISSING" of a particular covariate was predictive.

ThirdPartyDollars is the only numeric variable in the baseline set that had missing values. We described in the data preparation and cleaning session how we dealt with these missing values in order to define a robust measure of LGD.

Merging in external data created another opportunity to deal with missing values. As described in the previous section, not all datasets that we merged in spanned the entire time range of the loan sample, 1990 through 2014. In these cases, missing values were introduced for some of the variables that we merged in if the loan origination date fell outside the timespan of the dataset we were merging in. In these cases we interpolated values using the following methods:


1. Mean imputation.
2. Filling in 0s when missing data didn't pertain to business of loan.
3. Mean imputation based on historical averages for state level data.
4. Filling in the next adjacent value if we used country-wide data with many discontinuities.

3 Probability of Default Modeling

The group took several approaches to default probability modeling. Notably, there were two major strategies in trying to calculate default probabilities. One of these involved implementing a variety of logit model and deep learning approaches. The other involved hazard modeling. We discuss each in the following two subsections.

3.1 Logistic and Deep Learning Modeling

In these methods, the group looked to model default probabilities of the approximately 55,000 loans. The dataset for this challenge included all of the initial data that the group was provided with as well as the various datasets the group took the time to collect during this process. These datasets included county-level income, unemployment rates, educations funding, Zillow housing data, etc. Additionally, the group decided to implement one-hot encoding to account for various kinds of categorical data such as sector codes, borrowing/lending state, etc. In doing this, the group had a wide variety of data to work with and trying to draw simultaneous correlations between these risk factors and charge-offs proved to be a complex procedure. Notably, several of the data columns had features with large variations between values so the group decided to take the log of these values as a way to reduce the variation of information as well as normalize the data. In prepping the dataset for training, the group used range normalization to keep values bounded between 0 and 1. Upon cleaning the model for predictions, the group was finally ready to implement models on the data. A simple logistic regression model using this dataset proved to have some success but was ultimately incapable of drawing significant correlations between risk factors and default. It became evident that a nonlinear model would be needed to predict default probabilities accurately.



#	Color
0	Red
1	Green
2	Blue
3	Red
4	Blue

#	Red	Green	Blue
0	1	0	0
1	0	1	0
2	0	0	1
3	1	0	0
4	0	0	1

Figure 3: One-hot Encoding Example

Therefore, the group turned to implementing several diamond-shaped deep learning schemes in Keras as an attempt to obtain better default probability results.[1] Notably, the group implemented several different diamond networks to predict if loans were likely to default or not. This shape was picked mainly because it had proven to have successful results in other classification problems that included a high order of categorical and sparse data (which is exactly the dataset that the group is working with). This served to be a focal point for the group as it became evident that several of these categorical values had high correlations with the outputs of the loans. Other architecture details included the groups selection of layers, activation functions, optimizers, and dropout. The group decided to add and remove layers not at the end of the network but rather in the middle of the

network. This was done to maintain learning symmetry within the model so that there would not be an equal number of nodes for layers equidistant to the middle layer. The except to this, of course, is the logistic layer which is added to the end of the diamond output to produce a binary output for each loan. While the number of layers differed in results for each of the neural networks, the other hyperparameters of the model had little to no variation throughout testing (with the exception of tuning them and retraining the models across different variations). Choosing activation functions turned out to be a relatively straightforward process, although it did not seem this way initially. Using a deep model (6+ layers) was going to be difficult if the group decided to go with a Rectified Linear Unit (ReLU) activation function. This is due to the problem commonly known in the deep learning community as the vanishing gradient issue. To counter this, several other activation functions were considered such as Leaky ReLU and Shifted ReLU but ultimately the group decided on implementing an Exponential Linear Unit (ELU) which served to not only have better classification results in other datasets but also deals with the vanishing gradient problem and is proven to have faster run times with more complex models [2].

$$f(x) = \begin{cases} x & x \geq 0 \\ \alpha(e^x - 1) & x \leq 0 \end{cases} \quad (2)$$

$$f'(x) = \begin{cases} 1 & x \geq 0 \\ f(x) + \alpha & x \leq 0 \end{cases} \quad (3)$$

Additionally, this activation function has the benefit of handling inputs of the data that may be log-based well by providing a lower bound on their gradient. In choosing which Optimizer to use, the group again went about the route of evaluating the data that they were working with. They determined that due to the sparsity of the data, it was necessary to select an adaptive optimization algorithm that can handle the sparse cases of the data. Therefore, the group turned to using the Adaptive Gradient (AdaGrad) methodology which keeps track of past gradients in respect to each risk factor [3].

$$\beta_{t+1} = \beta_t + \frac{\eta}{\sqrt{G_t + \epsilon}} g_t \quad (4)$$

The benefit of this is that the algorithm makes smaller updates for features that are included more frequently while making larger updates for features that are rarer within the dataset. Choosing dropout proved to be fairly straightforward as the group varied dropout levels from 0 to 0.5 using the most dropout at the middle (most dense) level of the diamond and decreasing dropout for each layer farther apart from the middle. The loss function used is simply binary cross-entropy which is a standard amongst binary classification problems. Additionally, the group included a streaming metric for receiver operating curve area under the curve (ROC-AUC) in order to keep track if the algorithm was improving each iteration in terms. Some results are included in this report.

Model	Train Accuracy	Test Accuracy	Train AUC	Test AUC
Logistic Model	0.8680	0.8686	0.8455	0.8555
10 Layer Network	0.9209	0.9043	0.9384	0.9309
12 Layer Network	0.9302	0.8721	0.9420	0.8936
Best 1 Year Results	0.9999	0.9999	0.9999	0.9999
Best 5 Year Results	0.9473	0.9325	0.9347	0.9275

Table 1: Accuracy and ROC-AUC Results

Upon evaluating the models, the group found promising results from the 8 and 10 layer models used to evaluate probability of default. Overall, these provided the best test set accuracies and ROC-AUC values of the models implemented. The worst performing of the models in terms of overall performance was the logistic model simply because it is unable to capture the same nonlinearities present in the model that the group was working with. The best training set model turned out to be the 12 layer model because it captured the highest degree of nonlinearity but overall it failed to generalize well to the test set.

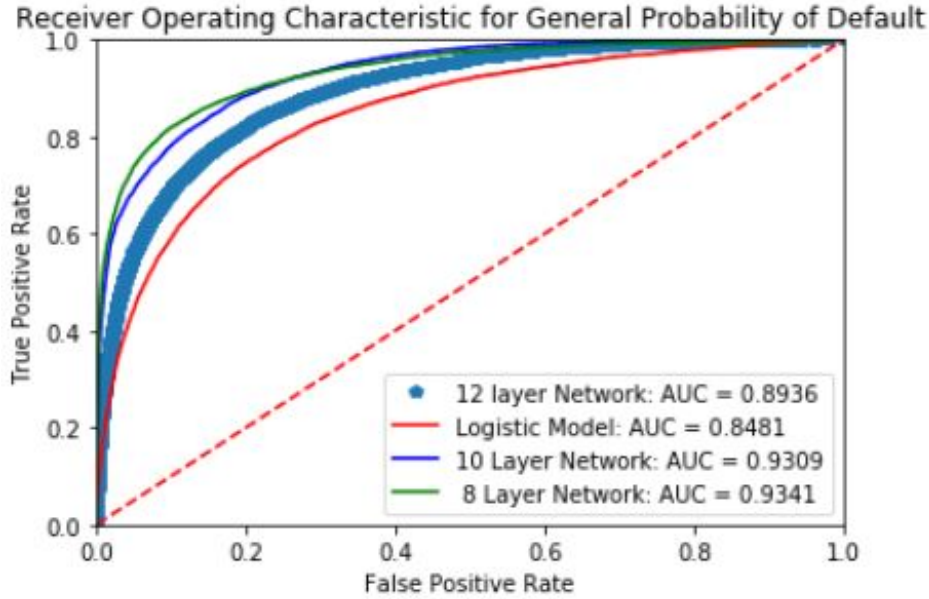


Figure 4: ROC-AUC

Some of the difficulties encountered within this modeling were overfitting (in models with more layers) as well as large misadjustment around local optima in the objective for several iterations until finally breaking out of the local optima and improving the overall accuracy/ROC-AUC. Additionally, for the 1 year default probabilities the dataset had only 2 of 55,000 results default within one year. This proved to be difficult in accurately predicting outcomes of loans in that timespan. As the deep learning model struggled to break out of local optima in this case.

3.2 Perry-Cox Hazard Modeling

To study the effects of different covariates on loan defaults we turned towards the Cox Proportional-Hazards model. In order to capture the time-varying effects we expanded the original loan dataset by months, creating a dataset with over eight million observations. Those eight million observations were merged with our additional covariates based on current months rather than approval dates. The specific variables and methods used for merging the data into the hazard model are described in section 2.2.

$$\lambda_i(t) = \lambda(t|X_i) = \lambda_0(t)\exp(\beta_1 X_{1i} + \dots + \beta_n X_{ni}) \quad (5)$$

The baseline hazard function, Equation 5, shows that if the hazard ratio of a coefficient is greater than one, that coefficient increases the likelihood of default, and vice versa for hazard ratios less than one. The higher the hazard ratio, the less likely a loan is to survive as time goes on.

We ran a multivariate Cox regression analysis to discover the effects of each covariate on loan defaults, and see those effects relative to every other covariate. For our hazard model we only focused on the additional covariates, omitting the original SBA dataset as well as datasets that had no effect in our classification models (such as patent filings and farmer’s markets). Additionally, we separated income tax data from the rest of the covariates when running the regression because it had significantly fewer observations and many more missing variables.

Although Table 2 doesn’t include income tax data, we find that the largest regression coefficients were tied to macroeconomic variables as well as the Zillow data. The biggest changes in hazard ratios result from changes in basic metrics such as the CCI, unemployment rate, and S&P 500 daily returns. Overall, spending data on education or agriculture had almost no effect on the hazard ratio.

Factoring in the results of Table 3, we can’t draw conclusions about all the large coefficients from Table 2. Public spending on agricultural research was statistically significant and had a small coefficient, which indicates it has a limited effect on the survival rate. Out of all the S&P 500 variables, the high or max values were the most statistically significant. The positive beta coefficient on `sp_High` suggests that a monthly peak in the stock market is more likely to lead to loan defaults. While not an intuitive conclusion, it could be capturing a reverse trend where the SBA loan defaults are a precursor to market corrections, and would therefore correlate with market peaks. CCI was also statistically significant and had a relatively large beta coefficient. With a hazard ratio greater than one and a positive coefficient, a higher CCI correlates with more SBA loan defaults. Again this is not initially intuitive, but a recent analysis by MarketWatch shows that the stock market has performed better following a low CCI value [7]. While both county and US unemployment rates had a similar impact on hazard rates, only the county level rate was statistically significant. The relationship between unemployment and hazard rates shows that a higher unemployment rate correlates with more SBA loan defaults. Out of all the Zillow datasets, only their custom ZHVI index was found to be statistically significant. The negative coefficient

Table 2: Coefficients from Multivariate Cox Regression

	coef	exp(coef)	se(coef)
ag_Public_nom	5.762×10^{-9}	1	2.493×10^{-9}
ag_Private_nom	-3.058×10^{-9}	1	2.180×10^{-9}
ag_Public_dollar	-6.764×10^{-9}	1	2.100×10^{-9}
ag_Private_dollar	2.088×10^{-9}	1	1.839×10^{-9}
sp_Open	-0.03145	0.969	0.01979
sp_High	0.0261	1.026	0.01036
sp_Low	-0.01601	0.9841	0.009776
sp_Close	0.01951	1.02	0.02038
sp_Volume	-9.737×10^{-11}	1	1.296×10^{-10}
sp>Returns	-27.25	1.467×10^{-12}	20.02
CCI	0.4213	1.524	0.1093
BCI	-0.09449	0.9098	0.08223
unemp_rate_US	0.08731	1.091	0.2027
unemp_rate_pty	0.07521	1.078	0.02383
zill_ZHVI	-0.004556	0.9955	0.0009909
zill_SizeRank_zhvi	1.270×10^{-4}	1	5.517×10^{-5}
zill_PctDecrease	1.459×10^{-4}	1	0.01183
zill_SizeRank_dec	-6.377×10^{-5}	0.9999	6.188×10^{-5}
zill_PctIncrease	-1.072×10^{-3}	0.9989	0.01093
ed_pty_ENROLL	9.542×10^{-6}	1	5.968×10^{-6}
ed_pty_TOTALREV	4.201×10^{-6}	1	2.171×10^{-6}
ed_pty_TFEDREV	2.487×10^{-6}	1	4.886×10^{-6}
ed_pty_TSTREV	-2.266×10^{-7}	1	1.264×10^{-6}
ed_pty_TOTALEXP	-1.023×10^{-5}	1	5.203×10^{-6}
ed_pty_TCURINST	3.648×10^{-6}	1	4.881×10^{-6}
ed_pty_TCURSSVC	7.078×10^{-6}	1	6.093×10^{-6}
ed_pty_TCURONON	-8.524×10^{-7}	1	9.834×10^{-6}
ed_pty_TCAPOUT	7.535×10^{-6}	1	6.231×10^{-6}
ed_st_ENROLL	4.454×10^{-7}	1	1.715×10^{-7}
ed_st_TOTALREVENUE	-7.464×10^{-10}	1	1.077×10^{-7}
ed_st_FEDERALREVENUE	5.322×10^{-7}	1	2.541×10^{-7}
ed_st_STATE_REVENUE	-9.033×10^{-8}	1	4.446×10^{-8}
ed_st_TOTAL_EXPENDITURE	-3.569×10^{-8}	1	1.621×10^{-7}
ed_st_INSTRUCTION_EXPENDITURE	2.826×10^{-8}	1	1.572×10^{-7}
ed_st_SUPPORT_SERVICES_EXPENDITURE	1.213×10^{-7}	1	2.274×10^{-7}
ed_st_OTHER_EXPENDITURE	-1.964×10^{-7}	1	6.819×10^{-7}
ed_st_CAPITAL_OUTLAY_EXPENDITURE	-3.972×10^{-7}	1	2.114×10^{-7}

Table 3: Statistical Significance of Multivariate Cox Regression

	z	Pr(> z)	
ag_Public_nom	2.312	0.020797	*
ag_Private_nom	-1.403	0.160763	
ag_Public_dollar	-3.221	0.001276	**
ag_Private_dollar	1.136	0.256055	
sp_Open	-1.589	0.111988	
sp_High	2.519	0.011777	*
sp_Low	-1.638	0.101387	
sp_Close	0.957	0.33834	
sp_Volume	-0.751	0.4526	
sp>Returns	-1.361	0.173578	
CCI	3.854	0.000116	***
BCI	-1.149	0.250549	
unemp_rate_US	0.431	0.666656	
unemp_rate_cty	3.157	0.001596	**
zill_ZHVI	-4.598	0.00000427	***
zill_SizeRank_zhvi	2.303	0.021306	*
zill_PctDecrease	0.012	0.990156	
zill_SizeRank_dec	-1.031	0.302683	
zill_PctIncrease	-0.098	0.921856	
ed_cty_ENROLL	1.599	0.109857	
ed_cty_TOTALREV	1.935	0.052936	.
ed_cty_TFEDREV	0.509	0.610688	
ed_cty_TSTREV	-0.179	0.857753	
ed_cty_TOTALEX	-1.966	0.049317	*
ed_cty_TCURINST	0.747	0.454809	
ed_cty_TCURSSVC	1.162	0.245412	
ed_cty_TCURONON	-0.087	0.930925	
ed_cty_TCAPOUT	1.209	0.22653	
ed_st_ENROLL	2.597	0.009408	**
ed_st_TOTAL_REVENUE	-0.007	0.99447	
ed_st_FEDERAL_REVENUE	2.095	0.036202	*
ed_st_STATE_REVENUE	-2.032	0.042194	*
ed_st_TOTAL_EXPENDITURE	-0.22	0.825683	
ed_st_INSTRUCTION_EXPENDITURE	0.18	0.857359	
ed_st_SUPPORT_SERVICES_EXPENDITURE	0.534	0.593592	
ed_st_OTHER_EXPENDITURE	-0.288	0.773368	
ed_st_CAPITAL_OUTLAY_EXPENDITURE	-1.879	0.0603	.
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1			

on zill_ZHVI shows that higher home values correlate with a decrease in loan defaults. Lastly, educational spending was statistically significant but the tiny coefficients suggest the effects were minimal relative to the other covariates.

Running a separate multivariate Cox regression on just income tax data, we found that three factors generally stood out as statistically significant: adjusted gross income, taxable interest, and business net income amount. Although our income tax data was broken into the number of exemptions and the amounts of different reported values, only the amounts were statistically significant. Moreover, not all income levels were equal in their significance. Mostly income levels 2 and 3 (\$25k-\$50k and \$50k-\$75k respectively) registered as statistically significant, representing the middle class. None of the coefficients were very large, however, so we turned to predicted survival curves to gauge the effect of these variables on the hazard rate.

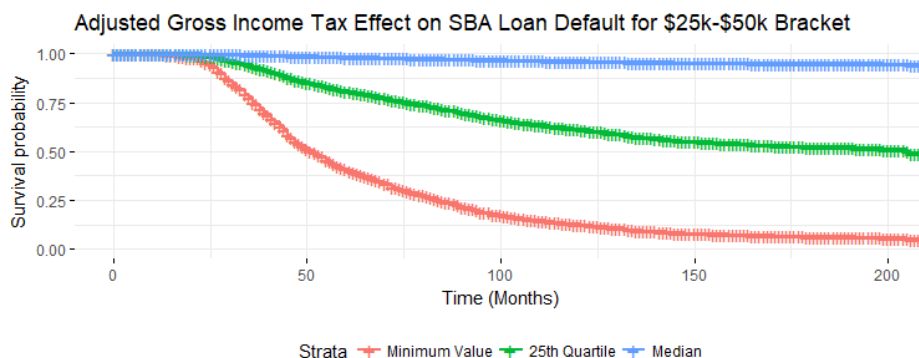


Figure 5: Effect of AGI using predicted survival fit

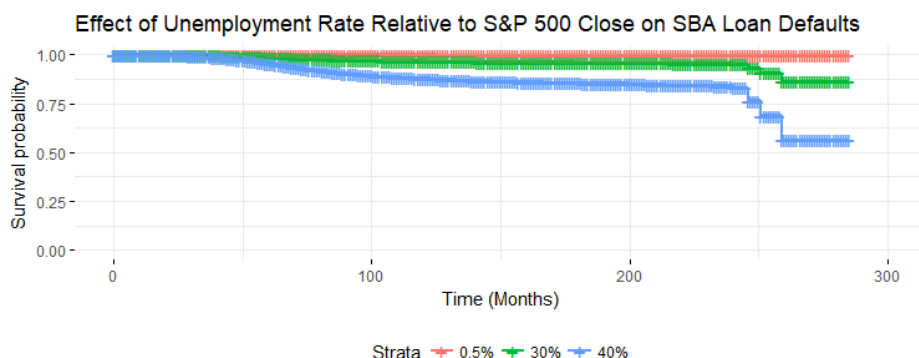


Figure 6: Effect of unemployment rates using predicted survival fit

By holding other covariates constant around their mean, we're able to fit a predicted survival curve that shows the relative impact of a single covariate. In figure 5 we looked at all income tax variables for the \$25k-\$50k tax bracket. We fitted three curves at three different quartiles of AGI data while holding the remaining variables constant. A clear pattern emerged, the lower the AGI value in a certain county, the higher likelihood of default in that area. In addition,

the survival probability drops off drastically for extreme values of AGI (i.e. the smallest AGIs). As you approach the median AGI value there is almost no drop in survival over the entire loan data timespan.

We also fit a prediction curve to the county level unemployment data while holding the S&P 500 monthly close constant in figure 6. Several difficulties in matching county level data led to discontinuities, but the overall trend is quite interesting. In general, we find that very high unemployment rates (maximum recorded in our dataset was 40.5) lead to significant dropoffs in loan survival. Nonetheless, even a high unemployment rate of 30 only leads to about a 20% dropoff in loan survival. So while unemployment rates do directly correlate with loan defaults, only extreme values seem to really tip the scales.

It's also worth noting there were many difficulties in the hazard modeling. Firstly, by expanding the data temporally by months, we ended up with only 8,000 defaults for eight million observations. This led to large discontinuities in the data and several covariates only had values for non-default observations (or perfect classification). In hindsight, doing a temporal expansion by loan years rather than months would have been better. Additionally, our focus on county level data created a lot of noise. Merging by county was difficult and created many missing values. In fact, the coefficients and hazard ratios are so small in Table 2 because of the limited amount of data (Cox regression model ended up using 2.3 million complete observations with only 482 default events). Nonetheless, we believe the statistically significant variables and their trends (positive or negative beta coefficients) are valid.

4 Loss Given Default Modeling

Refer to the data cleaning and exploration section to review how we defined the response variable for loss given default modeling. It follows from the way that we have defined LGD that it is bounded between 0 and 1, taking on continuous values in that range. Because of the special nature of this range, we follow Papke and Wooldrige (1996) in modeling this loss rate using fractional response regression, a type of Generalized Linear Model.

The model is specified as follows:

$$E(y_i|\mathbf{x}_i) = G(\mathbf{x}_i\boldsymbol{\beta}), i = 1 \dots, N \quad (6)$$

where G is a function mapping its inputs in the range $(0, 1)$. Specifically, we choose the function G to be the logistic function:

$$G(z) = \frac{\exp(z)}{1 + \exp(z)} \quad (7)$$

The likelihood function takes the following form:

$$l_i = y_i \log[G(\mathbf{x}_i|\beta)] + (1 - y_i) \log[1 - G(\mathbf{x}_i|\beta)] \quad (8)$$

In this setup of the likelihood function, the number of loans drops out since it does not depend on the parameters and we call it a quasi-maximum likelihood estimator (QLME) [5]

We implement the model in R using the GLM function with a quasibinomial('logit') link function.

Initially we chose to fit a baseline model using the following the variables which enter the models as dummies: BorrState, ThirdPartyLender_State, DeliveryMethod, TermInMonths, NaicsCode, BusinessType, and Total Loan amount. Under this specification, with a 75% training set and 25% validation set we obtain the results shown in the first row of the table below:

Model Specification	Training RMSE	Validation RMSE
Baseline	0.133156	0.1361161
Add Unemployment	0.1331238	0.136214
Add 5 Year PD	0.1331244	0.1362158

Figure 7: LGD Model Results

After fitting the baseline model we added the unemployment rate to the model to determine whether this would provide an incremental improvement in prediction accuracy. As shown in the second row of the table, surprisingly the unemployment rate does not give any lift to the prediction accuracy.

Our next effort at improving the model consisted of adding the 5 year PD from the neural network model described earlier as a driver. The literature has shown that for at least public firms PD can be an important driver of LGD and we wanted to test whether the same was true in our sample of small business loans. [6] As shown in the third row of the table above we don't find that PD provides any substantial lift for the prediction accuracy of LGD in our sample of small business loans.

In the end we use the full specified model containing both unemployment and 5 year PD as drivers in addition to the baseline set of covariates. The model performance between the various specifications is essentially the same and we choose the fully specified model for theoretical reasons since we anticipate an economic link between the unemployment, PD and LGD.

We now can use our fitted LGD model to predict the loss at default for all 54,806 loans in the dataset. In addition we can take the product of the LGD estimate and the PD estimate from the models above to generate an estimate of total loss. We do this in the next section in order to calculate VaR estimates and losses in various tranches of a hypothetical security backed by 500 random loans.

5 VaR and Tranche Modeling

In order to make the right investment decisions investors have to model and evaluate the risks associated with their portfolio. In the case of our project we analyzed a portfolio of 500 randomly selected loans from the top half of riskiest loans in our pool 55k SBA 504 loans. We determined this was the best method of choosing the loans for this project because we were able to get the right risk profile to spread risk across our tranches.

5.1 Portfolio Simulation

To perform the risk analysis we generated a loss distribution by applying Monte Carlo simulation. From this loss distribution we were able to calculate a continuous Value at Risk and Average Value at Risk spanning the 95% and 99% levels and determine the loss distribution for senior and junior tranches. Value at Risk is an important metric that helps investment professionals to estimate how much capital they should retain in order to remain solvent with a chosen probability. VaR has become one of the most popular techniques in risk management due to its practical relevance and intuitiveness.

We selected our random sample of 500 loans from half the full set of 54,806, which were the top 27,403 most risky loans, using simple random sampling without replacement.

Once we selected the 500 loans we performed the following steps for each loan 5,000 times:

1. Generate a random uniform variable V
2. If $P(U = 1) \geq V$ then the loan defaults where $P(U = 1)$ is estimated by our PD model at both 1 and 5 year default horizons
3. Estimate the loss per loan that defaults by using the LGD model estimate for that loan
4. Estimate the sum of the loss over all defaulted loans in the portfolio

After performing 5,000 simulations we obtained the plot of loss distribution for 1-year and 5-year time horizons. Figure 8 shows the 1 and 5-year loss distribution respectively.

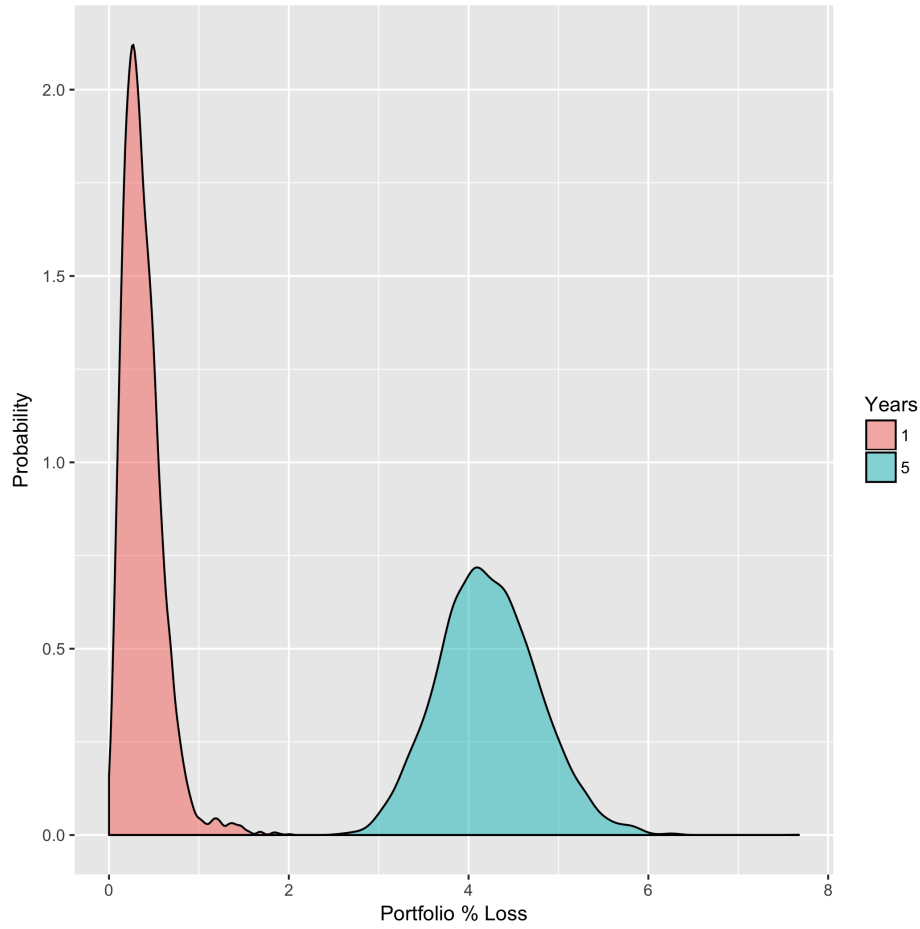


Figure 8: Simulated Percent Portfolio Loss Distribution for 1 and 5 Year Periods

5.2 Value at Risk Distribution

Banks have to assess the risks before making investment decisions. They also need to estimate the worst-case scenario in order to ensure that they have available funds to absorb the losses. VaR helps to determine the amount of funds that banks need to retain to cover the losses and Average VaR provides a more conservative measure analyzing the tail of the loss distribution. We calculated both VaR and Average VaR (Expected Shortfall) between 90% and 100% at steps of .0001 in order to plot a nearly continuous range of VaR and AVaR estimates.

Figure 9 show the 1-year Expected Shortfall distribution for 90% to 100% confidence levels. Figure 10 show the 5-year Expected Shortfall for 90% to 100% confidence levels.

As expected the AVaR is always greater than or equal to VaR and provides a more conservative risk measure by looking at the distribution of losses given that

a rare event has occurred.

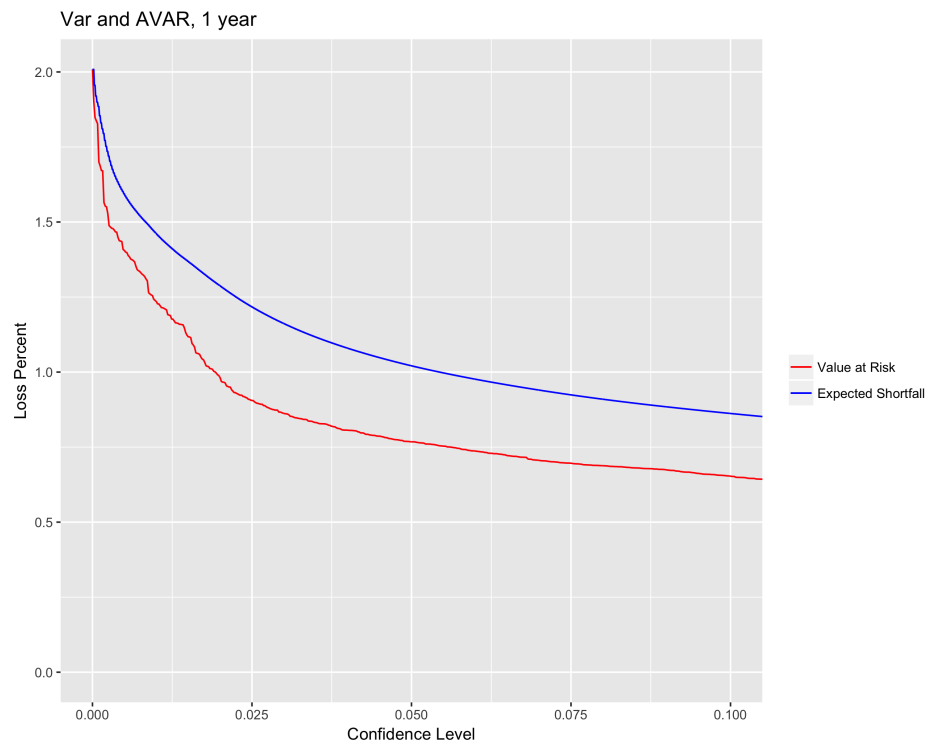


Figure 9: Value at Risk after 1 year

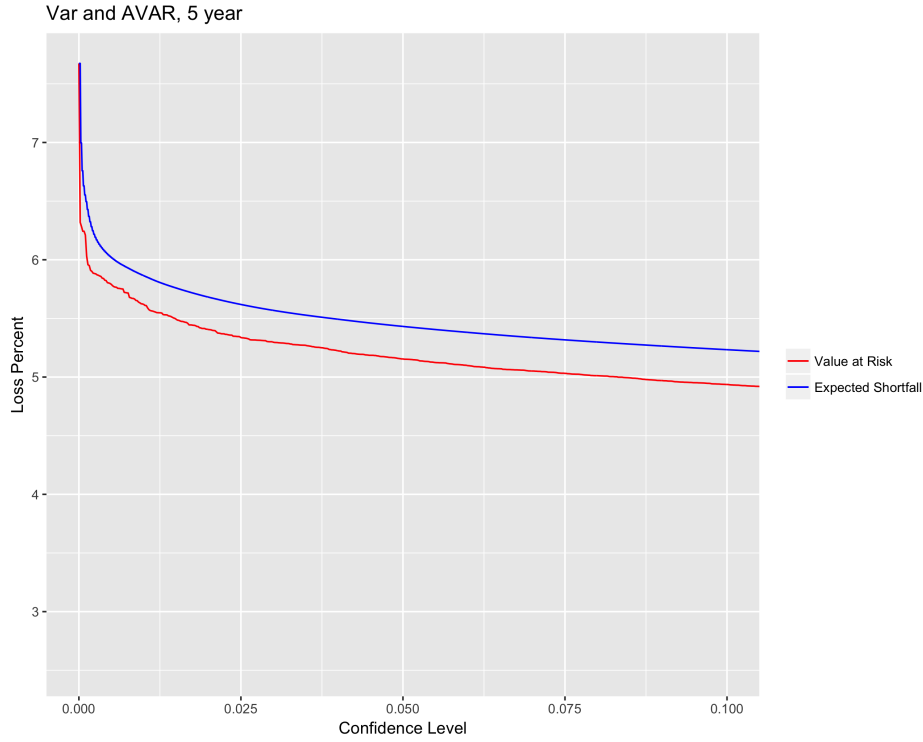


Figure 10: Value at Risk after 5 years

5.3 Securitization and Tranche Loss Distribution

Banks can use securitization to increase profits from selling the loans. Securitization is a process of creating financial instruments by combining other financial assets and marketing different ties of repackaged instruments to investors.

After securitization, several different tranches are created and prioritized within the pool of the funds. More junior tranches are hit by losses to the fund first. After that senior tranches are hit with respect to their seniority.

We calculated the distribution of two tranches - a junior tranche incorporating funds in the [5%, 15%] quantiles of total funds, and a senior tranche incorporating funds in the [15%, 100%] quantiles of total funds. We used the losses calculated previously to measure the loss distribution of our tranches. The case when the loss amount surpasses 5% of the original loan funds would correspond to a loss in the junior tranche. The case when the loss amount surpasses 15% of the original loan funds would correspond to a loss in the senior tranche.

We generated the plots of 1-year and 5-year senior and junior tranches (Figures 11 and 12) to compare how different tranches will be hit by losses. The plots demonstrate us that most of the losses will not hit senior tranche. After analyzing the plots we can conclude that 5 year horizon is more risky for investors than 1-

year horizon.

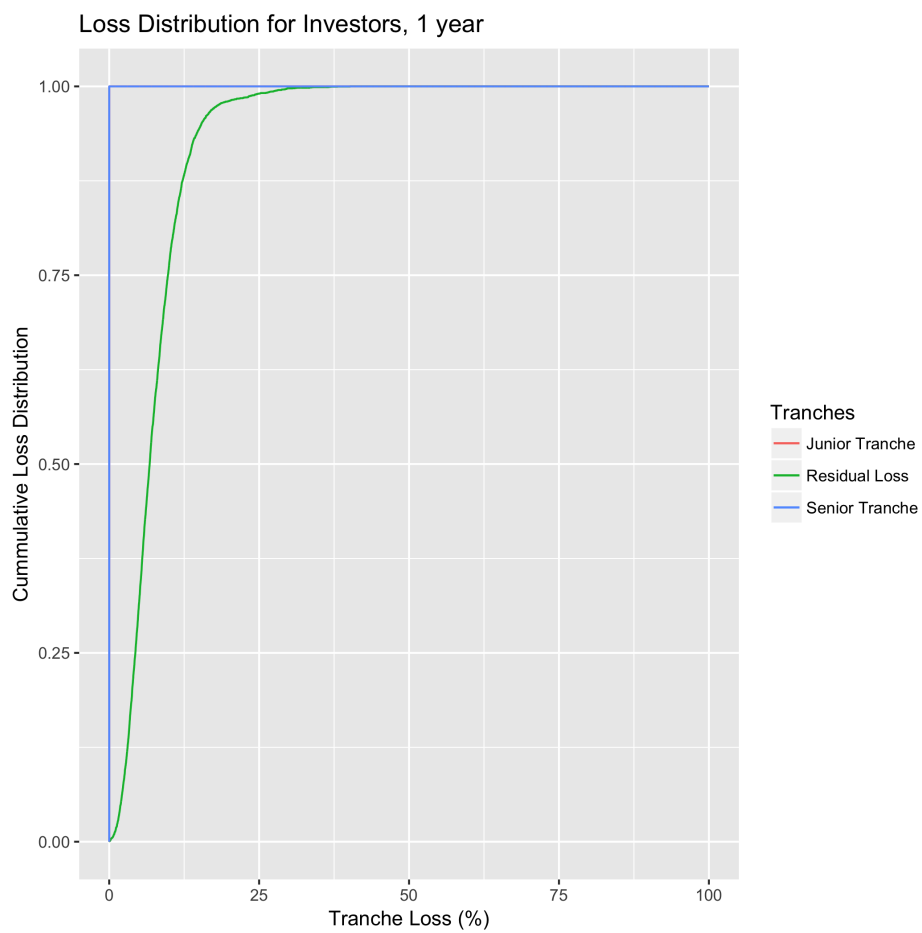


Figure 11: Tranche Loss Distribution after 1 year

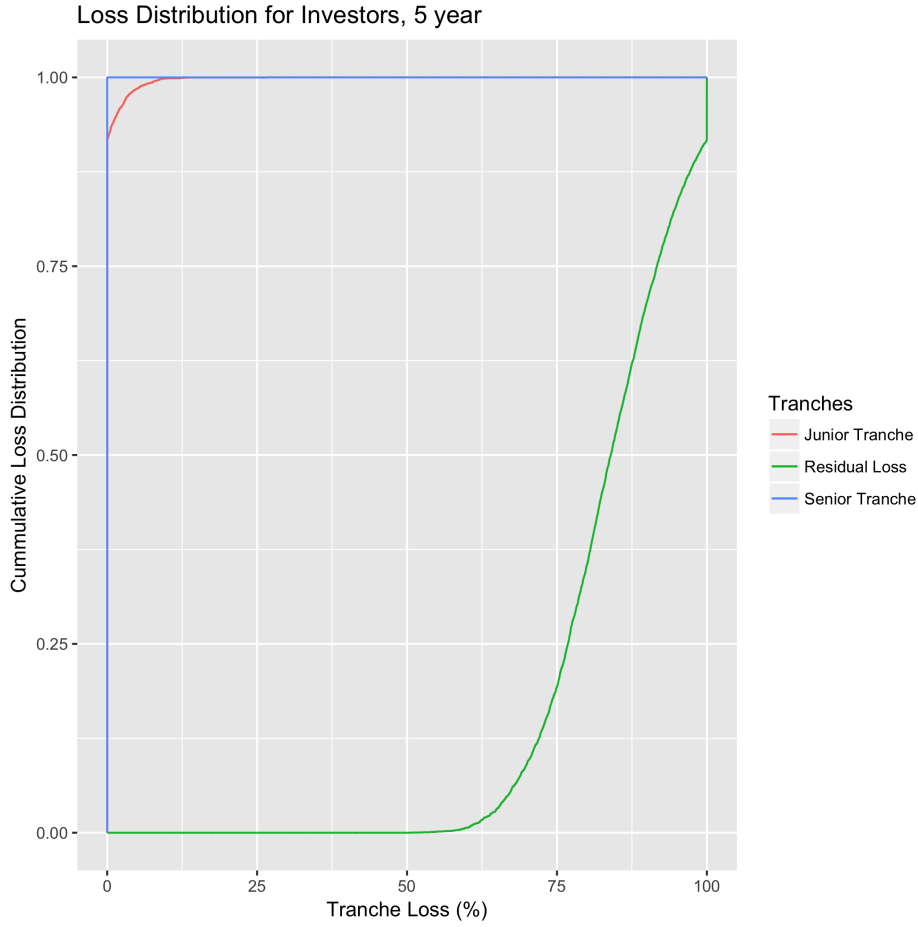


Figure 12: Tranche Loss Distribution after 5 years

6 Conclusion

In this project we were able to go through all of the steps of deriving an estimate for losses on a portfolio of 500 randomly selected SBA 504 loans. We began by undertaking a thorough analysis of the baseline dataset and researched the particulars of the 504 loan program. Next we merged in economic and other data to greatly enlarge our set of available predictors. We chose a strategy to deal with missing values and then fit models for PD and LGD. Using these two models we were able to estimate the total loss on 500 randomly selected loans and calculate VaR and the loss distributions of senior and junior investor classes. The project gave us new insight into the risk profile of these investor classes as well as the risks borne by banks, the Small Business Administration, and hypothetical investors.

7 Contributions

Brian focused on the deep learning modeling used for predicting the probability of defaults for the loans. Brian also served as project manager helping to steer the direction of the project and ensure everyone had a sufficient task to do on a weekly basis.

Scott focused on data preparation and cleaning and loss given default modeling. Scott researched the SBA 504 loan program to understand the typical loan structure which was useful in helping figure out how to interpolate the missing values for the ThirdPartyDollars which was key to having a good measure of LGD.

Denis focused on finding and merging additional datasets for classification and hazard modeling. Denis modified the additional datasets in order to efficiently match them with SBA values by geography and time. He also worked on the hazard model.

Iryna focused on preparing Portfolio Simulations, Value at Risk measures, and Tranche loss modeling techniques.

David focused on preparing Portfolio Simulations, Value at Risk measures, and Tranche loss modeling techniques.

References

- [1] Zhang et. al; aXiv:1601.02376 *Deep Learning over Multi-field Categorical Data*
- [2] Shelhamer et. al; arXiv:1511.07289 *Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)*
- [3] Duchi et. al; *Adaptive Subgradient Methods for Online Learning and Stochastic Optimization*
- [4] Leslie E. Papke, Jeffrey M. Wooldridge; Journal of Applied Econometrics, 1996, vol. 11, issue 6, 619-32 *Econometric methods for fractional response variables with an application to 401(k) plan participation rates*
- [5] Oberhofer, Harald and Pfaffermayr, Michael; Contemporary Economics, ISSN 2084-0845, Vizja Press & IT, Warsaw, Vol. 6, Iss. 3, pp. 56-64 *Fractional Response Models - A Replication Exercise of Papke and Wooldridge (1996)*
- [6] Jon Frye; Federal Reserve Bank of Chicago, <https://www.chicagofed.org/media/others/people/research-resources/frye-jon/frye-lgd-as-a-function-of-the-default-rate-091013-pdf.pdf> *Loss given default as a function of the default rate*
- [7] Mark Hulbert; MarketWatch, 30 Mar. 2017, www.marketwatch.com/story/why-soaring-consumer-confidence-should-worry-investors-2017-03-30. *Why Soaring Consumer Confidence Should Worry Investors*