Analyzing the Risk of Mortgage Default

Grace Deng

Statistics Honors Thesis - Fall 2016

.

Adviser: David Aldous

Abstract

This paper analyzes the risk of mortgage default and prepay for single-family, 30-year fixed rate mortgages using a variety of machine learning and survival analysis methods. Predictions are made for homeowner choices to continue payment, default, or prepay using both parametric and non-parametric models. These models include Binary Logit, Multinomial Logit, K-Nearest Neighbors, K-fold Cross Validation, and Random Forest. Replications of each model, with various combination of parameters, were performed in order to identify the best model; the Random Forest model with 150 trees and 4 entries yielded the highest accuracy at 93%. Difference in survival time (months of mortgage payment until termination) are then compared for owner-occupied homes versus investment homes. Investors are found to pay mortgages longer than primary homeowners. Meanwhile, primary homeowners are less likely to default and more likely to prepay than their investment counterparts, suggesting the presence of an endowment effect. Survival curves are also plotted for six Cox Proportional Hazard models after checking relevant assumptions.

Keywords: Default, Prepay, Machine Learning, Survival Analysis, Non-Parametric Methods

[1]

[1]**Acknowledgements:** I would like to thank my adviser, David Aldous, for his invaluable support and guidance.

1 Introduction

One of the major causes of the 2007 financial crisis was due to an overflow of subprime mortgages and resulting defaults. Before non-government agencies took over the mortgage-backed securities market, most mortgages adhered to the underwriting rules by agencies such as Fannie Mae and Freddie Mac. Non-government agencies later loosened the underwriting guidelines, resulting in a bubble in the market that was sensitive to changes in the economy and consumer expectations. The goal of this project is to explore single-family loan performance data published by Fannie Mae and analyze the risk of default and prepay using a range of statistical techniques.

Traditional econometric models tend to focus on finding significance or causality for specific predictors that adhere to a theoretical framework. Results are presented mainly in terms of statistically significant predictors or R^2; the models used were mainly regressions, multinomial logit, or Cox Proportional Hazard model instead of machine learning methods. Furthermore, few analyze their models in terms of out-of-sample accuracy and distinguished between training and testing datasets.

The analysis of this paper, on the other hand, will begin with simple logit models, extend to multi-class multinomial logit, and move on to non-parametric machine learning models such as K-Nearest Neighbors, K-fold Cross Validation, and Random Forest that may fare better with potential non-linearity in the predictors. The corresponding accuracy rates and confusion matrices will then be analyzed. Since primary and investment homeowners likely have very different mindsets, median survival time of these two demographics will be compared and survival curves of corresponding Cox Proportional Hazard models will be examined.

Contemporaneous economic variables, loan-specific variables, and borrower demographics will all be used as predictors in order to form a more holistic picture of the causes of mortgage decisions. Variables traditionally used to capture market conditions include local unemployment, which not only serves as a proxy for the income level but also captures some of the consumer expectations regarding housing prices. An underutilized covariate in previous research, local rent, will be included in order to measure the mobility between ownership and rental housing. Because the dataset used is very extensive, it will also be possible to measure geographic variation through state indicators for all 50 states. Loan variables consist of equity (measured by contemporaneous loan-to-value ratio), interest rate, loan age, origination year, and loan purpose. Borrower characteristics, such as whether a borrower is a first time home-buyer, the primary resident in the house, and is creditworthy (FICO scores) will be used to capture individual variation that affects mortgage decisions. Different types of loans will be compared to test for the presence of an "Endowment Effect", where borrowers may be less willing to default on residential property versus an investment property, holding other factors such as interest rate and principal constant.

The Fannie Mae website contains 15 years of data (2000-2015) on the acquisition and performances of 30-year fixed rate mortgage loans, divided by quarters. Data for each quarter is split into two files: Acquisitions and Performance, containing 24 and 29 variables respectively. The Fannie Mae data will be the main dataset, supplemented by data on other economic variables from FRED and Bureau of Labor Statistics.

2 Data

The final dataset used in this paper is compiled from multiple sources. The Fannie Mae mortgage dataset contains many individual loan variables such as observation period, original unpaid balance, original interest rate, FICO scores, etc. as well as some limited information on borrower characteristics (i.e. number of borrowers, Debt-to-Income ratio) Only loans that are single-family homes (no condos), and either owner-occupied or investment home (no secondary) are used, all of which originated from 2006 or later. All loans are 30-year Fixed Rate Mortgages (FRM), meaning that the interest rate and monthly mortgage payment are "locked in" for the full amortization period. Other data sources include the FRED (Federal Reserve Economic Data), the US Bureau of Labor Statistics, and the Department of Housing and Urban Development websites. When possible, variables such as local unemployment and rent are matched by observation period (dated by months) and Metropolitian Statistical Area (MSA). Because the raw data is very extensive, a randomized subset of approximately 5.5 million observations is chosen for the analysis. Of these observations, 2345347 are in the base class "Paying" (42.5%), 366585 loans have defaulted (6.6%), 2801014 loans have prepaid (50.8%).

Five non-categorical variables are used as predictors: rent-to-mortgage ratio (Rent Ratio), contemporaneous loan-to-value ratio (TLTV), unemployment rate, loan age in months (Age), and the original 30-year interest rate. Histograms for these variables can be found in the Appendix; the plots for TLTV and Interest Rate are fairly symmetric and normally distributed, while the plots for Unemployment Rate and Loan Age are slightly right skewed. The Rent Ratio is very left skewed, with some extreme outliers suggesting that for these loans, the monthly mortgage payment was much lower than the median rent. There are several explanations for this phenomenon; one possibility is that the homeowner began with a large down payment (ie. had a lot of cash on hand), so the mortgage balance was relative small compared to equity and resulted in a small monthly payment. Outliers (5.3%) for Rent Ratio are identified using Tukey's method (see Appendix). Categorical variables used as predictors include the First Time Home-buyer Indicator (Yes, No, or Unknown), the Loan Purpose indicator (Primary, Refinance, or Cash-out Refinance), the FICO score, Year of Origination for the loan, Occupation indicator (Owner-Occupied or Investment home), and State indicators for all 50 states.

3 Simple Logit Default

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-5.0399	0.1523	-33.09	0.0000
rentratio	-0.1233	0.0053	-23.22	0.0000
first_time_indicatorU	0.3401	0.2039	1.67	0.0953
first_time_indicatorY	0.0705	0.0183	3.85	0.0001
loan_purposeP	-0.6062	0.0146	-41.58	0.0000
loan_purposeR	-0.1597	0.0132	-12.07	0.0000
occupPrimary	0.1708	0.0198	8.62	0.0000
tltv	0.0457	0.0005	98.99	0.0000
fico_b580-619	-0.0127	0.0606	-0.21	0.8338
fico_b620-639	0.0309	0.0602	0.51	0.6076
fico_b640-659	-0.0086	0.0593	-0.15	0.8842
fico_b660-679	-0.1170	0.0589	-1.99	0.0470
fico_b680-699	-0.2550	0.0587	-4.34	0.0000
fico_b700-719	-0.3722	0.0588	-6.33	0.0000
fico_b720-739	-0.5554	0.0590	-9.41	0.0000
fico_b740-759	-0.7427	0.0591	-12.56	0.0000
fico_b760-850	-1.1730	0.0583	-20.13	0.0000
uemp_rate	0.6330	0.0042	149.36	0.0000
origyear_indicator2007	-0.7956	0.0163	-48.66	0.0000
origyear_indicator2008	-1.3720	0.0173	-79.48	0.0000
origyear_indicator2009	-2.3175	0.0238	-97.38	0.0000
origyear_indicator2010	-3.6648	0.0294	-124.52	0.0000
origyear_indicator2011	-4.8852	0.0340	-143.69	0.0000
origyear_indicator2012	-6.6857	0.0437	-153.06	0.0000
underwater1	-0.1957	0.0206	-9.48	0.0000
age	-0.0743	0.0003	-273.82	0.0000
orig_rate	0.3674	0.0124	29.58	0.0000
stateAL	0.2663	0.1155	2.31	0.0211
stateAR	0.5474	0.1177	4.65	0.0000
stateAZ	-0.1640	0.1060	-1.55	0.1218
stateCA	-1.2463	0.1040	-11.99	0.0000
stateCO	0.5172	0.1110	4.66	0.0000
stateCT	-0.1753	0.1079	-1.63	0.1041
stateDC	0.9964	0.1284	7.76	0.0000
stateDE	0.0082	0.1270	0.06	0.9487
stateFL	-0.4325	0.1040	-4.16	0.0000
stateGA	-0.4994	0.1058	-4.72	0.0000
stateHI	1.5524	0.1242	12.50	0.0000
stateIA	1.4691	0.1233	11.91	0.0000
...	...	...	...	...
stateWY	0.7973	0.1897	4.20	0.0000

Table 1: Simple Logit - Default vs. Paying

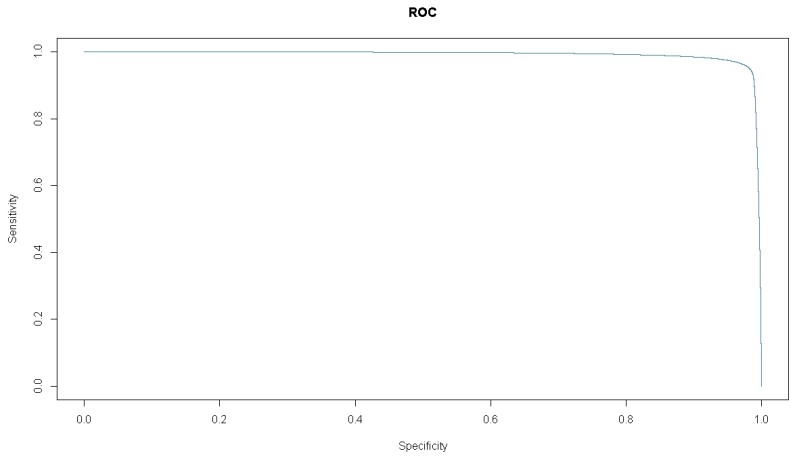


Figure 1: ROC Curve - Default Logit

		True Class	
		Paying	Default
Predicted	Paying	463773	2755
	Default	11451	64408
	Overall Accuracy:	**0.9738**	
	Sensitivity:	**0.9590**	
	Specificity:	**0.9759**	

Table 2: Confusion Matrix - Default Logit

In the Default Logit, all loans that prepaid are dropped to avoid bias in the results. 80% of the data is randomly chosen as the training dataset, and the rest as the testing dataset. From Table 1, it can be seen that apart from an occasional FICO score category or state indicator, all predictors are statistically significant at the 1% level. As FICO scores increase, the corresponding coefficient becomes increasingly negative, which makes sense since borrowers are less likely to default if they have greater credibility. Origination Year coefficients follows a similar pattern, which could be due to the fact that during the Great Recession, interest rates and housing prices all continued to decline and people who applied for mortgages in the later years received better deals. In addition, TLTV, Unemployment Rate, and Interest Rate all have positive coefficients, suggesting that increases in the loan-to-value ratio (less equity), unemployment rate (risk of income loss and economic recession), or mortgage interest (higher monthly payments) all contribute to default. Finally, geographic location does play an important role, where state indicators for California, Florida, and D.C. etc. show less risk of default compared to states such as Arizona, Hawaii, and Wyoming.

The ROC plot in Figure 1 shows the performance of the simple logit as the cutoff threshold for fitted values is varied. The optimal cutoff, which maximizes the sensitivity (true positive rate) and specificity (true negative rate) is 0.111. Overall accuracy of the model is 97.38%, and both sensitivity and specificity are fairly high as well, meaning that the model is equally good at identifying loans that will default and loans that will continue to pay.

4 Multinomial Logit

		True Class		
		Paying	Default	Prepay
Predicted	Paying	465098	1871	23646
	Default	31	24895	11615
	Prepay	4133	46791	524510
Overall Accuracy: 0.477				

Table 3: Confusion Matrix - Multinomial Logit

Mortgage termination can also be due to prepay, such as refinance or the homeowner deciding to sell the property and cash out on capital gains. Since prepay is a competing risk for mortgage default, it is worthwhile to use multi-class classification, such as Multinomial Logit (MNL) in order to analyze the full dataset. Using the neural net package in R, coefficients were estimated for the MNL using the training dataset (converged after 10 iterations). The model is then used to predict the relative probabilities of paying, default, and prepay for each loan in the testing dataset; the final class prediction is based on the highest probability. The final accuracy is 47.7% after comparing the predicted class to the true class, only somewhat better than a blind guess 33.33%.

5 K-Nearest Neighbors

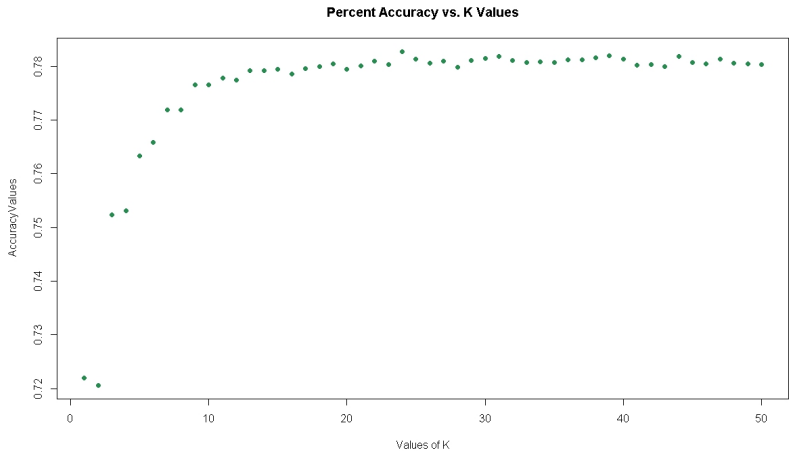


Figure 2: Accuracy vs. K values - KNN

		True Class		
		Paying	Default	Prepay
Predicted	Paying	7264	89	1894
	Default	5	300	138
	Prepay	1244	1001	8065
Overall Accuracy: 0.781				

Table 4: Confusion Matrix - K-Nearest Neighbors, K*=24

Due to the low performance of the MNL, which may be the result of non-linearity in the predictors, non-parametric methods are then explored. A randomized 100,000 subset is chosen for K-Nearest Neighbors (KNN) model, since using the full dataset would be too computationally expensive. 50 KNN models are performed, where K varied from 1 to 50, in order to find the optimal K with highest accuracy. In Figure 2, it can be seen that accuracy starts off at about 72% before plateauing off at around 78%, with the best K equal to 24. Table 4 gives the confusion matrix for the model with K=24, where overall accuracy is 78.1%. This is a significant improvement in multi-class prediction compared to the MNL.

6 K-Fold Cross Validation

The goal of cross validation is to limit overfitting and observe how results will generalize when using an out-of-sample independent dataset. K-fold Cross Validation is not exhaustive (compared to Leave-One-Out-Cross Validation), but it is a good approximation and should perform well with the large number of observations in the dataset. Similar to KNN, 50 K-fold Cross Validation models are performed and each model's accuracy is plotted against corresponding K values (Figure 2). The best performing model is K=23, and the resulting confusion matrix shows an accuracy of 78.5%, slightly better than KNN results.

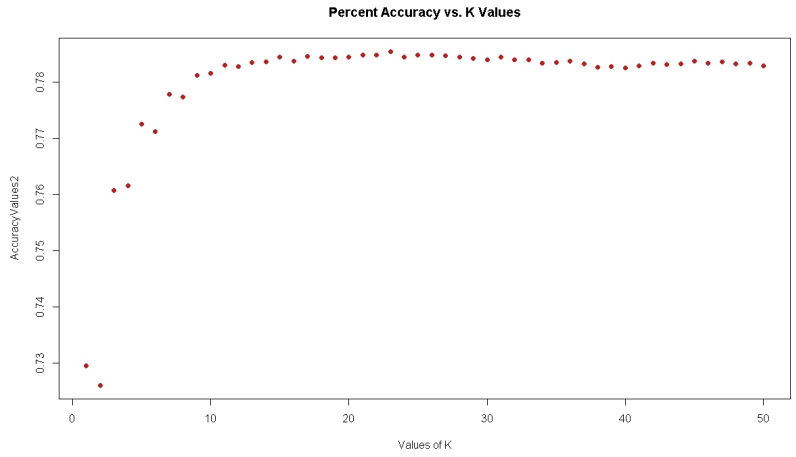


Figure 3: Accuracy vs. K values - K-fold Cross Validation

		True Class		
		Paying	Default	Prepay
Predicted	Paying	29074	356	7370
	Default	16	1076	619
	Prepay	4921	3891	32677
Overall Accuracy: 0.785				

Table 5: Confusion Matrix - K-fold Cross Validation, K*=23

7 Random Forest

150 random forest models are performed, based on combinations of the two parameters that can be altered in the random forest model: the number of trees and the number of randomly selected entries. Research has shown that random forest could yield better results for large datasets given the same number of predictors [1]. The table below shows the resulting accuracy on the testing dataset, and is rounded to the nearest 0.01. The highest accuracy achieved is 0.92965 (see bolded in Table 6), which resulted from the combination of 80 trees X 4 random predictors and the combination of 150 trees X 4 random predictors at each node.

	# Predictors										
	1	2	3	4	5	6	7	8	9	10	
	10	0.79	0.91	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92
	20	0.84	0.92	0.93	0.93	0.93	0.93	0.93	0.93	0.92	0.92
	30	0.84	0.92	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.92
	40	0.84	0.92	0.93	0.93	0.93	0.93	0.93	0.93	0.92	0.93
	50	0.85	0.92	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93
	60	0.86	0.92	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93
#Trees	70	0.86	0.92	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93
	80	0.86	0.92	0.93	***0.93**	0.93	0.93	0.93	0.93	0.93	0.93
	90	0.85	0.92	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93
	100	0.86	0.92	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93
	110	0.87	0.92	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93
	120	0.85	0.92	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93
	130	0.85	0.92	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93
	140	0.86	0.92	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93
	150	0.86	0.92	0.93	***0.93**	0.93	0.93	0.93	0.93	0.93	0.93

Table 6: Accuracy Matrix Rounded to Nearest 1% - #Trees X #Predictors

*Denotes a combination that yielded the highest accuracy, 0.92965.

		True Class		
		Paying	Default	Prepay
Predicted	Paying	8487	25	346
	Default	0	492	184
	Prepay	36	826	9604
Overall Accuracy: 0.9297				

Table 7: Confusion Matrix - Random Forest, 150 Trees X 4 Predictors

So far, the random forest model produced the highest accuracy at about 93%. From the confusion matrix above, it seems that the model has a harder time distinguishing between Default and Prepay classes. This suggests that there are subtle differences between the two types of homeowners, perhaps not captured by the predictors, especially since the number of loan observations is much greater than the number of predictors (N >> p). However, the model is extremely good at predicting the true Paying class, with a true positive rate of 99.6% (only 36 out of 8523 cases misclassified). From an institutional lender perspective, this model would be very useful in identifying borrowers that will pay back the loan and generate profit in the long-run; although prepay homeowners will pay back the full principal, the lender earns more interest when the borrower pays over the full amortization period.

8 Survival Analysis

Survival analysis refers to methodology for analyzing data where the variable of interest is the time until a designated event happens, such as cancer remission, failure of a machine, etc. In this paper's context, the event would be mortgage termination, and the survival rate would be calculated by the proportion of loans that are still continuing payment after the cutoff observation period. Survival curves are not commonly used in predictive modeling, but it would be interesting to look at graphical representations of mortgage termination. The survival time for a loan can be defined as the age of the loan, in months, when a default or prepay occurs; a uniform cutoff period means there will be right censoring in the data.

8.1 Survival Rates and Time

	Default	Paying	Prepay	Total	Median Survival Time
Primary	6132	38277	47471	91880	31 Months
	0.07	0.42	0.52	100%	
Investment	581.00	4247.00	3292.00	8120	34 Months
	0.07	0.52	0.41	100%	

Table 8: Summary Statistic - Primary vs. Investment Home Loans

	N	Observed	Expected	$\frac{(O-E)^2}{E}$	$\frac{(O-E)^2}{V}$
Primary	91880	53603	52163	39.8	440
Investment	8120	3873	5313	390.4	440
$\tilde{\chi}^2 = 440$ on 1 degrees of freedom, p-value = 0					

Table 9: Log Rank Test - Primary vs. Investment Home Loans

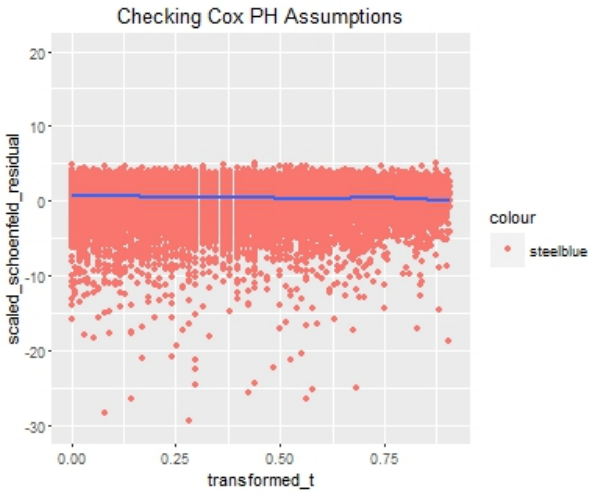
It can be seen from the Median Survival Time column (Table 8) that loans for Investment homes tend to pay longer than Owner Occupied (Primary) homes. The difference in survival time could be explained by the income levels of investment and primary homeowners; assuming that people do not buy investment homes before their primary residences, the people who can afford the down payment for an investment home and qualify for a second loan would necessarily earn more than first time home buyers. To formally test for difference in survival distributions, a log-rank test will be used, since it is non-parametric and more appropriate with survival data that tends to be right skewed and censored. In Table 9, the results of the log-rank test is shown. A $\tilde{\chi}^2$ value of 440 and p-value of 0 signifies that the difference in survival time between primary and investment homes are statistically significant at the 1% level. Finally, observing the base rates for different mortgage choices in Table 8, it can be concluded that although primary and investment homes both have the same default rates, primary homeowners tend to prefer prepay over default, suggesting that there may be an endowment effect where residents become attached to their homes.

8.2 Cox Proportional Hazard Model

The Cox Proportional Hazard model (Cox PH) is a semi-parametric model that does not require a specific hazard function. The dataset is first split between Primary and Investment homes, then further divided into two subsets, one that included only Paying or Default loans (Default Only), and one that included only Paying or Prepay loans (Prepay Only). Six Cox PH models are then fitted onto the datasets: Primary homes, Investment homes, Primary-Default Only, Primary-Prepay Only, Investment-Default Only, and Investment-Prepay Only. All models have statistically significant p-values (0) for the Likelihood Ratio Test, the Wald Test, and Log Rank Test. The resulting survival curves, as well as a plot of scaled Schoenfeld residuals to test the Proportional Hazards assumption of Cox regressions, are shown below.

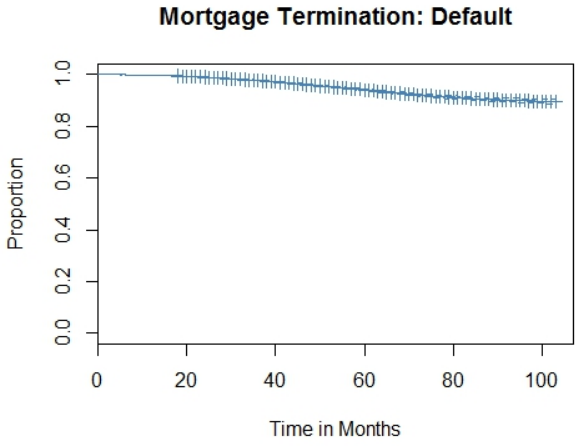
Ideally, the Schoenfeld residual plot would be centered about 0, which all but the model for Default-Only Owner-Occupied homes satisfy. Due to the large amount of data, the survival curves appear more smooth than typical "staircase" survival plots. Since default accounts for a small portion of mortage termination, the survival curves for Default Only models are relatively flat for both primary and investment homes. On the other hand, it is easy to see the prepay accounts for most of the mortgage termination, and it is interesting to note that the corresponding survival curves, for all homes, show that the entire population "dies" at about 100 months. In other words, a 30-year Fixed Rate Mortgage is not expected to last more than 8.33 years, only about 28% of the full amortization period.

(a) Survival Curve

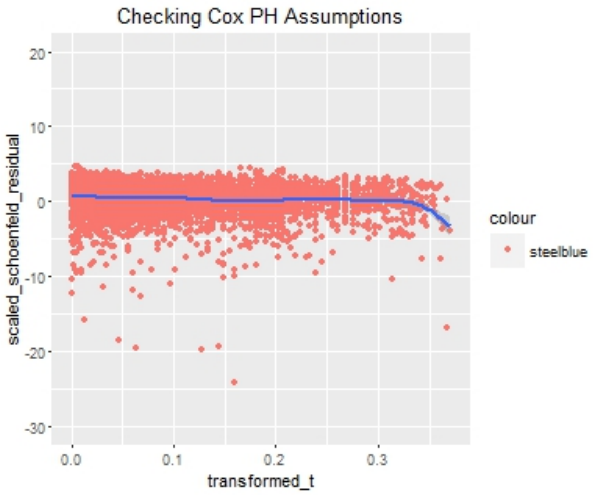


(b) Checking Assumptions

Figure 4: Cox PH Model for Owner-Occupied Homes - Default and Prepay

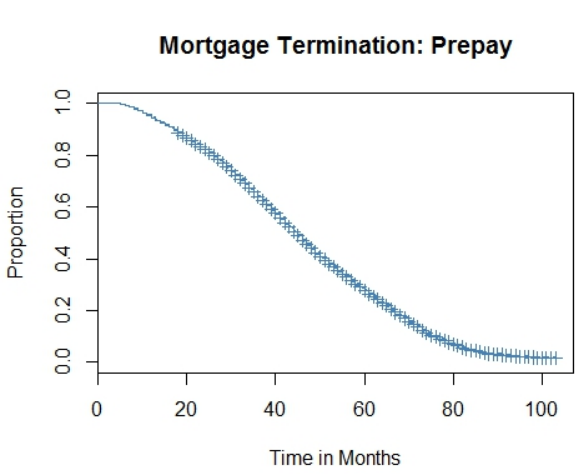


(a) Survival Curve

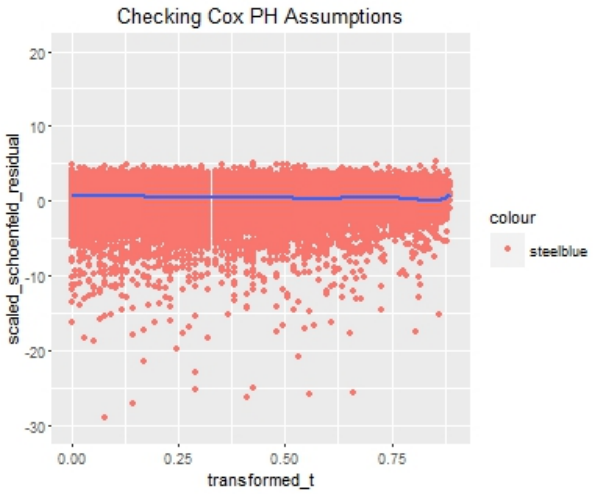


(b) Checking Assumptions

Figure 5: Cox PH Model for Owner-Occupied Homes - Default Only

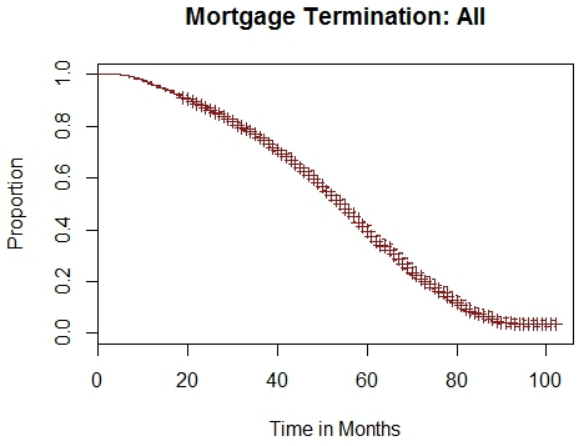


(a) Survival Curve

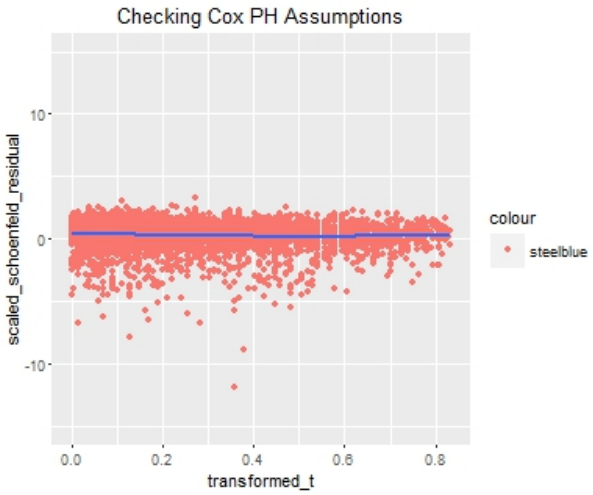


(b) Checking Assumptions

Figure 6: Cox PH Model for Owner-Occupied Homes - Prepay Only

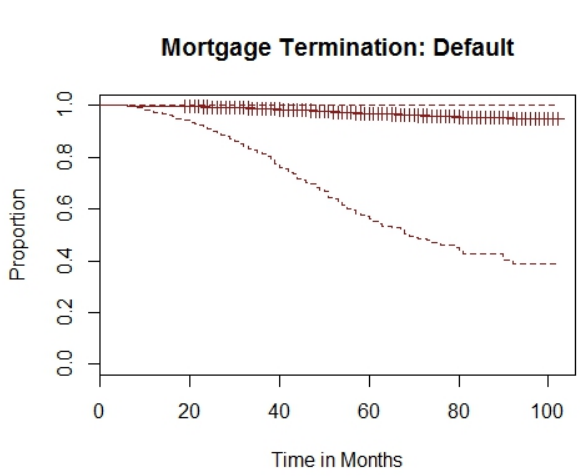


(a) Survival Curve

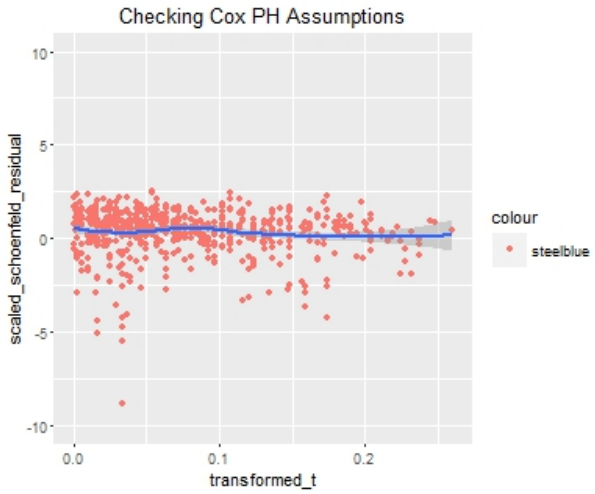


(b) Checking Assumptions

Figure 7: Cox PH Model for Investment Homes - Default and Prepay

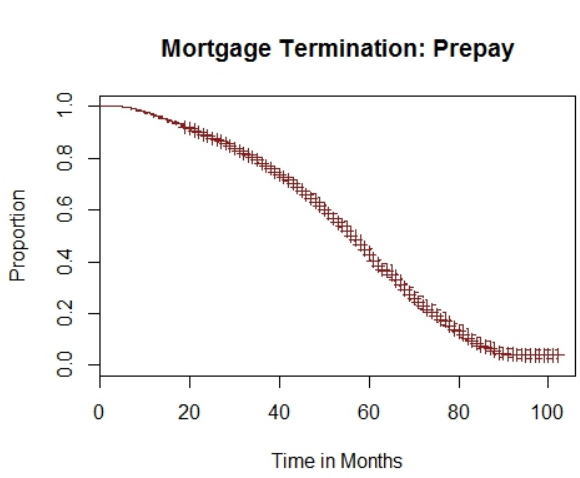


(a) Survival Curve

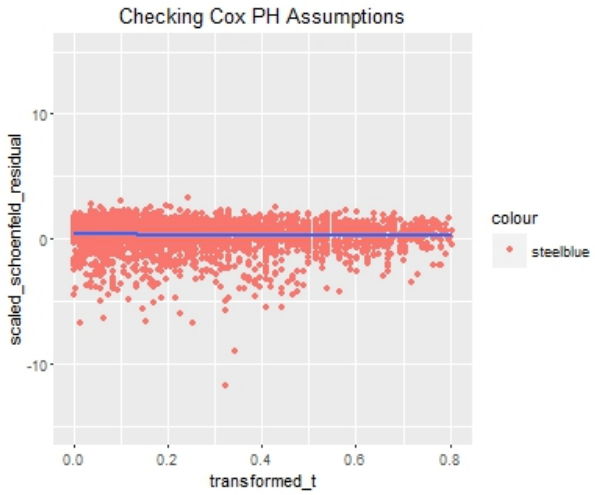


(b) Checking Assumptions

Figure 8: Cox PH Model for Owner-Occupied Homes - Default Only



(a) Survival Curve



(b) Checking Assumptions

Figure 9: Cox PH Model for Owner-Occupied Homes - Prepay Only

9 Discussion

This paper applies numerous machine learning and survival analysis techniques on mortgage data in order to examine homeowner choices to continue payment, default, or prepay. Machine learning methods used include Binary Logit, Multinomial Logit (MNL), K-Nearest Neighbors (KNN), K-fold Cross Validation (KCV), and Random Forest. The simple logit and MNL are run on the full 5.5 million dataset, while KNN, KCV, and Random Forest are run on a randomized susbset for computational purposes. Fifty KNN models and fifty

KCV models are performed to identify the optimal K values; similarly, one hundred and fifty Random Forest models, each using a unique combination of number of trees and number of entries, are performed to find the model with highest overall accuracy (93%).

Using Log Rank Test, the survival time (age in months) is compared between Primary and Investment loans. The difference is statistically significant, meaning that investment homeowners tend to pay mortgages longer than primary homeowners. In addition, owner-occupied homes have a greater probability of default and lower probability of prepay relative to investment homes. Finally, the proportionality assumption is checked before fitting six Cox Proportional Hazard models onto subsets of Primary and Investment loans, and the corresponding survival curves are examined.

Future research could be conducted with models that address the problem of loan observations greatly exceeding the number of predictors (N >> p). Monte Carlo simulations would also be helpful in predicting how changes in the economy (i.e. the Fed raising interest rates) could influence future mortgage performance.

References

[1] Ali, Jehad, et al. "Random Forests and Decision Trees." *International Journal of Computer Science* Issues 3rd ser. 9.5 (2012).

[2] James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: With Applications in R.* Print.

[3] Kahneman, D., Knetsch, J.L., and Thaler, R.H. (1991). "Anomalies: The Endowment Effect, Loss Aversion, and Status Quo Bias." *The Journal of Economic Perspectives.* 5:1, 193-206.

[4] Hosmer, David W., Stanley Lemeshow, and Rodney X. Sturdivant. *Applied logistic regression.* Hoboken, New Jersey: Wiley, 2013. Print.

[5] Quercia, R. G., and Stegman, M. A. (1992). "Residential Mortgage Default: A Review of the Literature." *Journal of Housing Research.* 2:3, 341-379.

10 Appendix

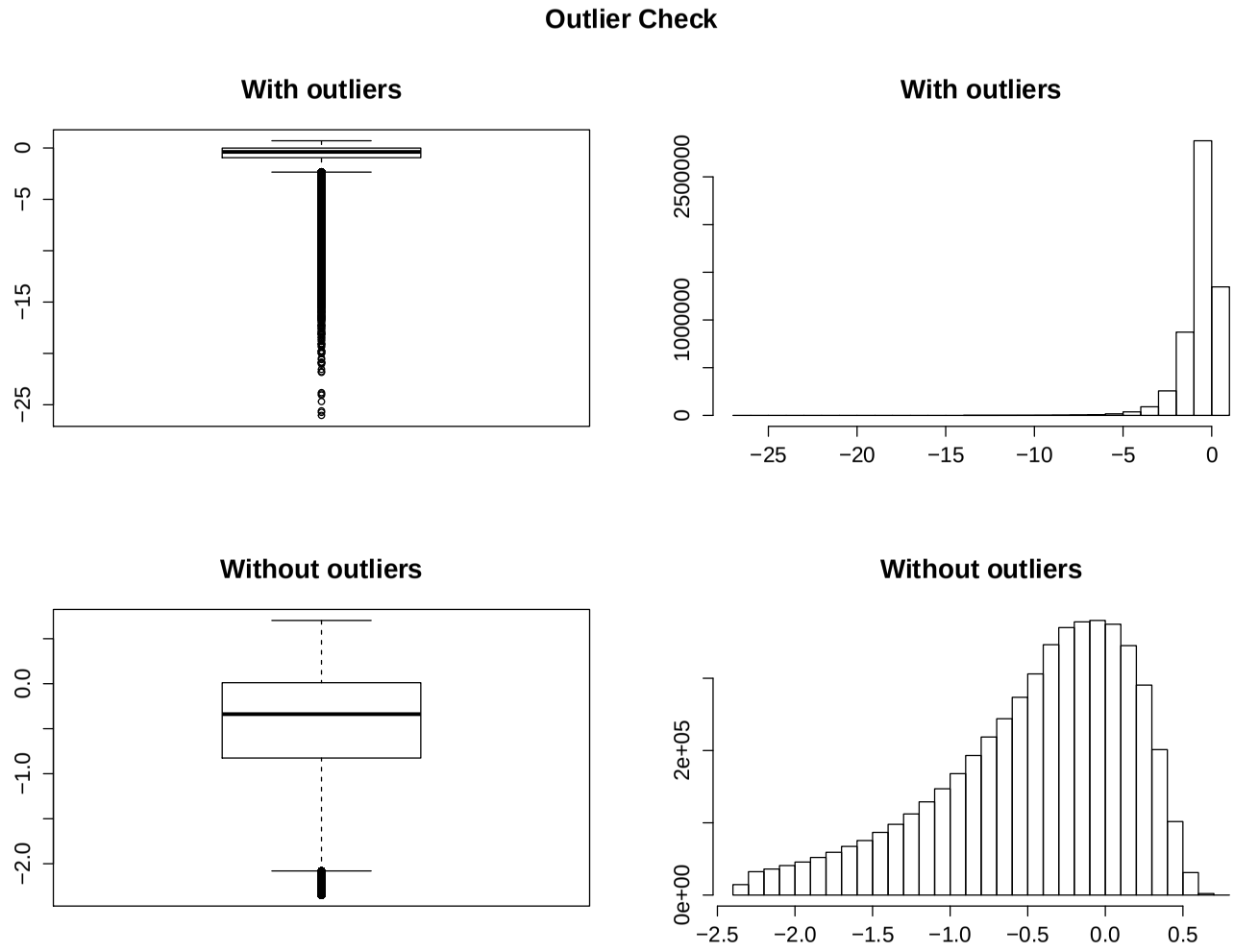
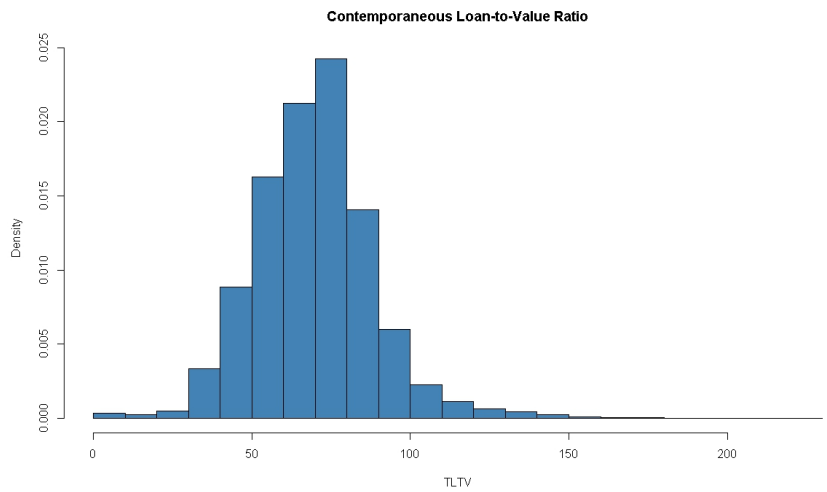
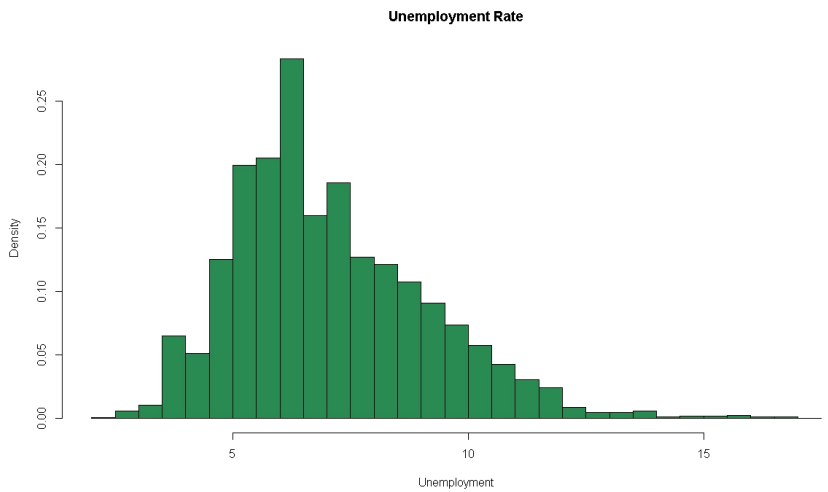


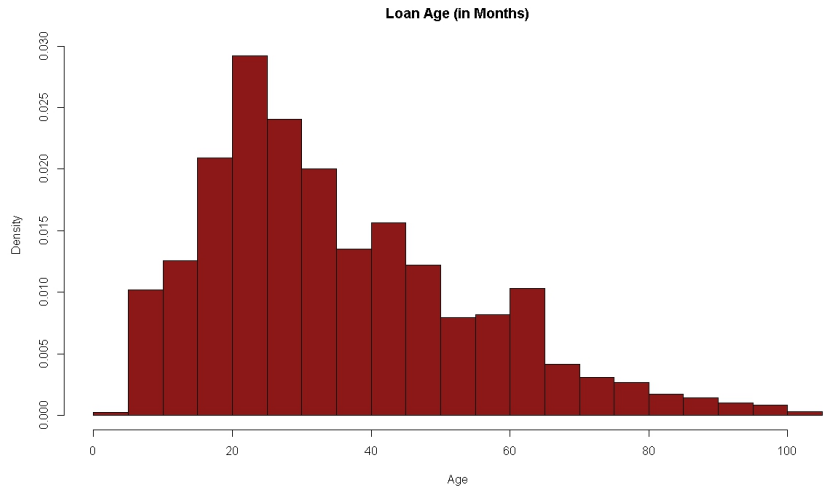
Figure 10: Outlier Check - Rent Ratio



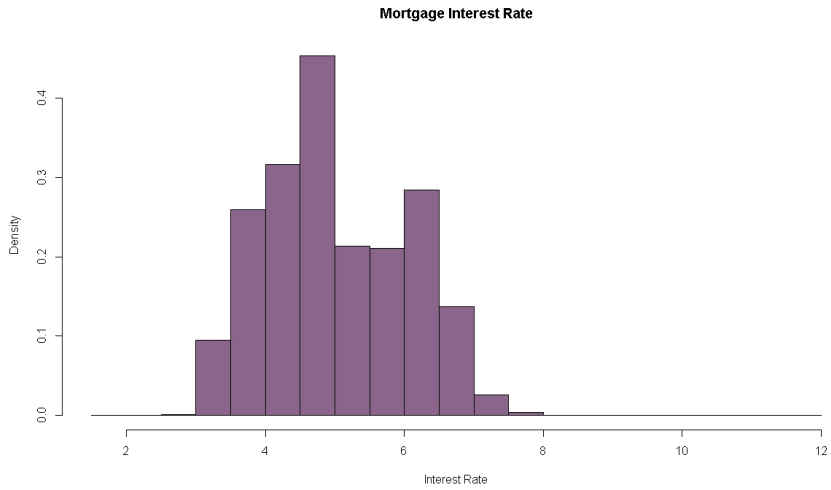
(a)



(b)



(c)



(d)

Figure 11: Histograms - TLTV, Unemployment Rate, Loan Age, and Interest Rate