Summary Report: Serious Delinquency
Classification Evaluation using Credit Data

CAPP 30254

May 7, 2015

THE UNIVERSITY OF CHICAGO

PRIMARY INVESTIGATOR:

JOSHUA MAUSOLF

Executive Summary.

In this report, I compare the performance of various machine-learning algorithms for the provided credit data, particular to the problem of predicting serious delinquency in the past two years, N = 150,000 observations for N = 11 variables, including the desired classification variable, serious delinquency.

After reviewing the models, evaluation methods, and evaluation, I conclude that for the given dataset and desired classification, the top overall recommended model is Random Forest with Boosting, depth = 5.

Models.

For this comparison, I tested the given data using a variety of machine learning algorithms, namely: logistic regression, decision trees, random forest, random forest with bagging, random forest with boosting, gradient boosting, KNN, and ANOVA linear SVM.

While perhaps mysterious in name, the bottom line of testing these numerous models is to determine which ones perform best for the given question and data.

Evaluation Method and Metrics.

To address this question, I evaluate the models on the given dataset, using five-fold cross-validation for random train-test splits. In essence, this method replicates the performance of the data using random subsets, and I average the results for each metric. In this way, I can confidently make recommendations on the average performance of each model.

Regarding the evaluation metrics, I consider the following: Accuracy, Average precision, Precision, Recall, Area under precision recall curve (AUC – PR), ROC-AUC, F1, Log-Loss, Mean Squared Error (MSE), and R2. In addition, I also consider both the initial model formulation time and the total time to run the five-fold cross validation.

Arguably, the key metrics to examine are overall accuracy relative to the baseline (no model), precision, recall, AUC-PR, and total evaluation time. I consider how the models compare in the next section.

Evaluating the Models.

With respect to the given models, the general classifier category that performed best were tree models (both decision trees and random forests), particularly when bagging or boosting was applied. In Table 1, I display the evaluation metric comparisons, highlighting in yellow the best performing classifier for various metrics. In red, I show models whose accuracy fell below the baseline.

TABLE 1 **Model Evaluation Table – Five-Fold Cross Validation**
- Serious Delinquencies Past Two Years -

	Logistic Regression	Decision Trees, Depth 4	Random Forest, Depth= 5	Random Forest with Bagging, Depth = 5	Random Forest with Boosting, Depth = 5	Gradient Boosting, Depth = 3	KNN, K=3	ANOVA Linear SVM
Baseline	0.933	0.933	0.933	**0.933**	**0.933**	**0.933**	0.933	0.933
Accuracy	0.934	0.936	0.936	**0.935**	**0.937**	**0.931**	0.927	0.933
Avg. Precision	0.218	0.370	0.385	**0.386**	**0.395**	**0.323**	0.128	0.314
Precision	0.559	0.573	0.624	**0.759**	**0.593**	**0.485**	0.258	0.551
Recall	0.029	0.153	0.119	**0.041**	**0.191**	**0.298**	0.045	0.015
AUC - PR	0.180	0.370	0.380	**0.380**	**0.390**	**0.420**	0.130	0.310
ROC-AUC	0.696	0.829	0.856	**0.853**	**0.861**	**0.822**	0.559	0.742
F1	0.065	0.245	0.185	**0.035**	**0.290**	**0.358**	0.077	0.029
Log-Loss	-0.228	-0.186	-0.183	**-0.187**	**-0.606**	**-0.880**	-1.833	-0.243
MSE	-0.066	-0.064	-0.064	**-0.066**	**-0.063**	**-0.069**	-0.073	-0.067
R2	-0.068	-0.290	-0.023	**-0.052**	**-0.009**	**-0.114**	-0.163	-0.069
Runtime Base Model	1.239	0.437	0.730	**3.236**	**16.044**	**10.270**	0.658	1479.967
Runtime 5-Fold Cross Validation	**39.843**	**11.404**	**29.531**	**170.861**	**713.542**	**491.975**	**114.866**	**38788.619**

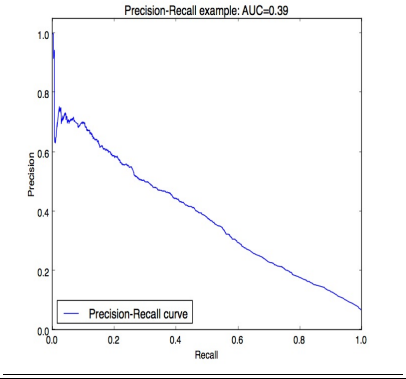
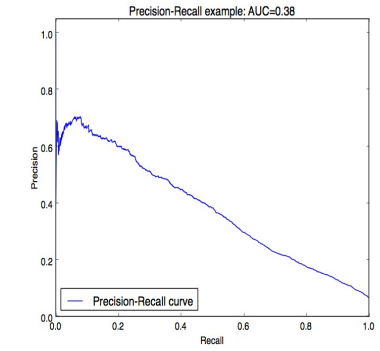
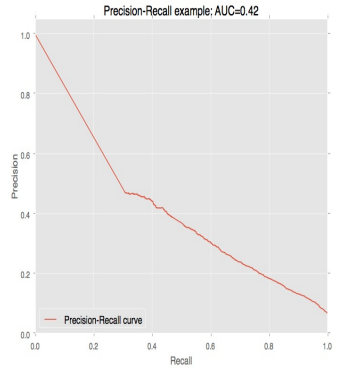
Note: Runtime is displayed in seconds. Baseline proportion of no serious delinquencies to serious delinquencies is listed. Accuracy shows model improvement over baseline.
Original Data Source: cs-training.csv.
Data Used: Preprocessed Data using cs-training.csv, imputing missing values for monthly income and number of dependents.

From this table, we can see that the model with most top-performing metrics was the random-forest with boosting model. This model had the highest overall accuracy 0.937 versus the baseline 0.933, the best average precision (0.395), and the best ROC-AUC (0.861). Moreover, this model had the most second-best evaluation metrics, for recall (0.191), AUC-PR (0.390), and F1 (0.290). While slower than most of the models (713.542 seconds), this model's high performance across accuracy, precision, and recall justifies the time cost in most situations. Thus, in comparison to other models, random forest with boosting, depth = 5 was the top overall-performing algorithm.

With respect to gradient boosting, depth = 3, although this model was a top performer in as many categories as random forest with boosting, it underperformed the mean accuracy compared to the baseline, in essence yielding no improvement compared to not running a model at all. Consider using gradient boosting only if accuracy is a low priority and you wish to maximize recall.

As a function of time, decision trees were by far the fastest (11.404 seconds), followed by simple random forests, depth = 5 (29.531 seconds). The most cost-prohibitive model was the ANOVA linear-SVM, which took a staggering 38789 seconds or approximately 10.77 hours. To balance time cost with performance, select a simple random-forest, which performs well but much more quickly than random forest with boosting.

For each of these respective models, I display their area under the precision recall curve and their respective time to run.

FIGURE 1	**AUC – Precision and Recall**	
Random Forest with Boosting Depth = 5	**Random Forest Depth= 5**	**Gradient Boosting Depth = 3**
Best for Top Overall Performance ***First Choice, Slower***	**Best for Mid-Grade Performance** ***Fast Alternative***	**Best for Maximizing Recall** ***Slower, Accuracy Suffers***
		

To see these figures in greater detail, please see appendix A0. For further information on the descriptive statistics of each dataset, please see appendix A1.

While these models are not perfect classifiers, depending on you main priorities and needs, these are the three recommended models for your classification problem of *serious delinquency* given the current data.

Appendix A0: AUC-Precision and Recall Curves

FIGURE A0-1 **Random Forrest with Boosting, Depth = 5**

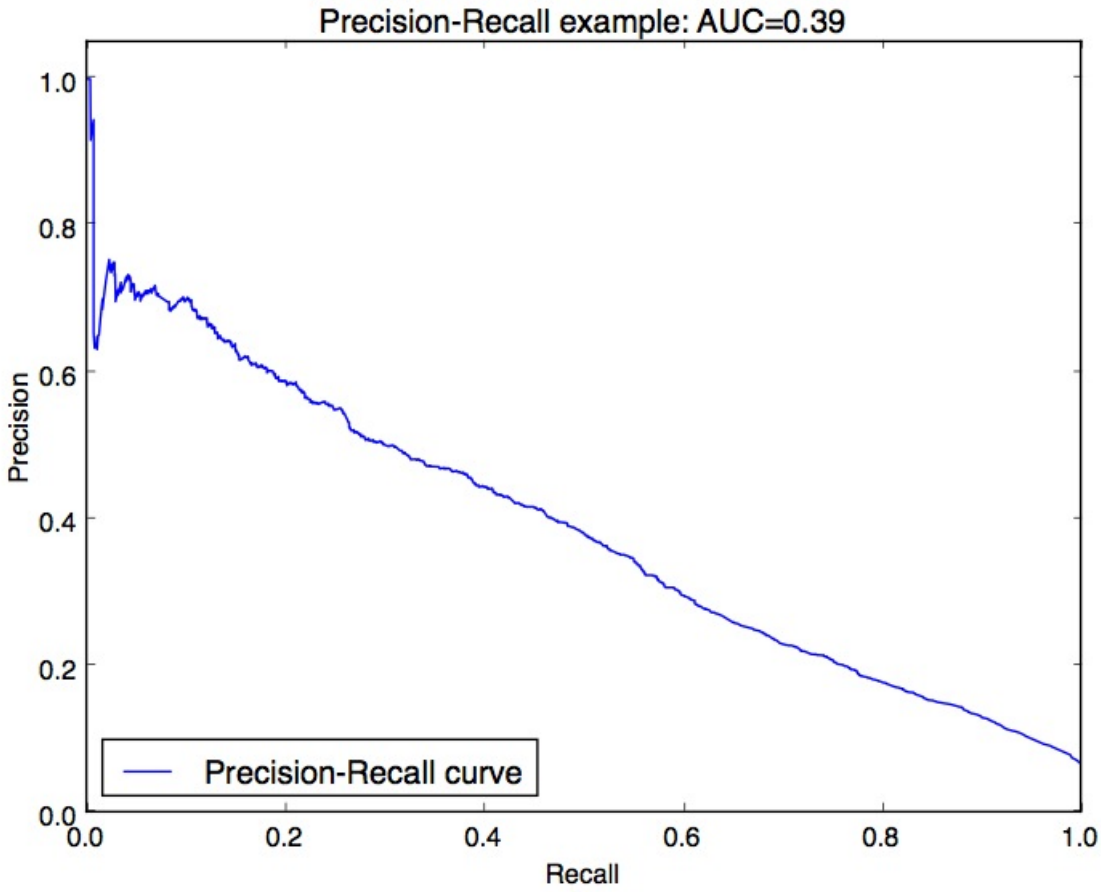


FIGURE A0-2 **Random Forrest, Depth = 5**

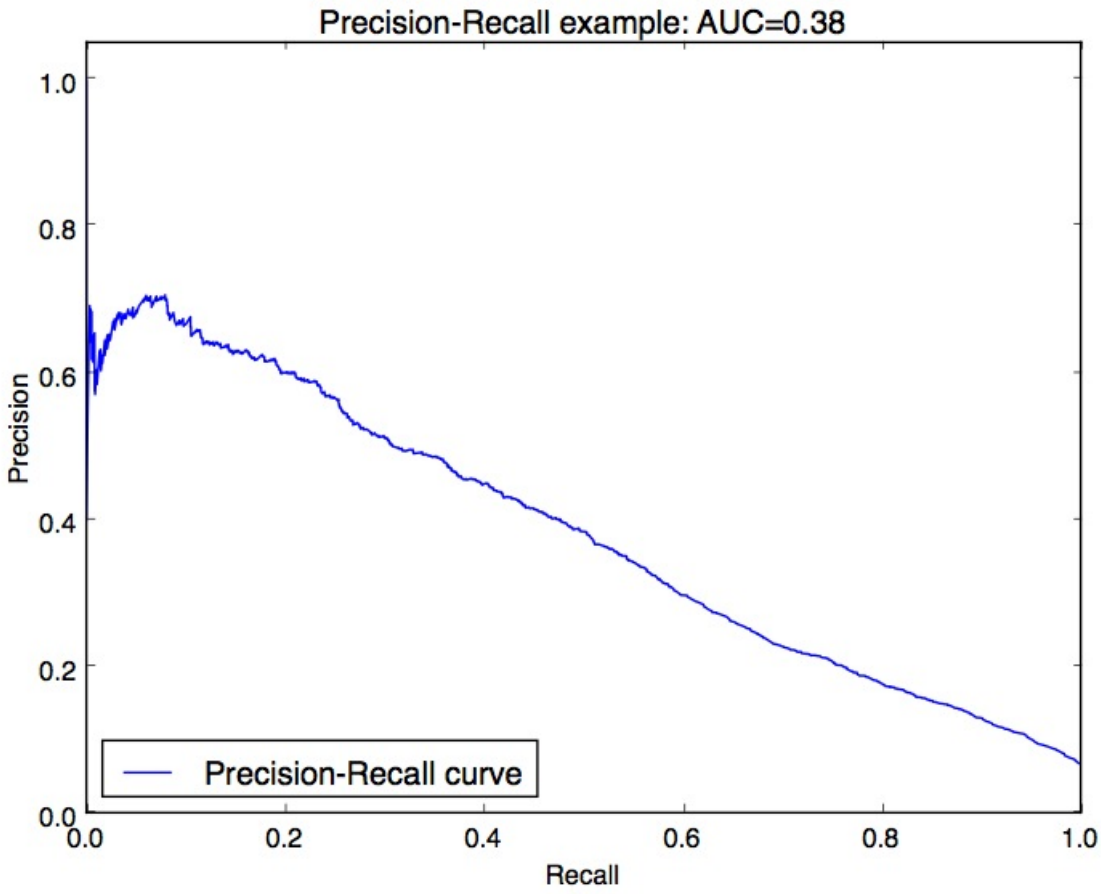
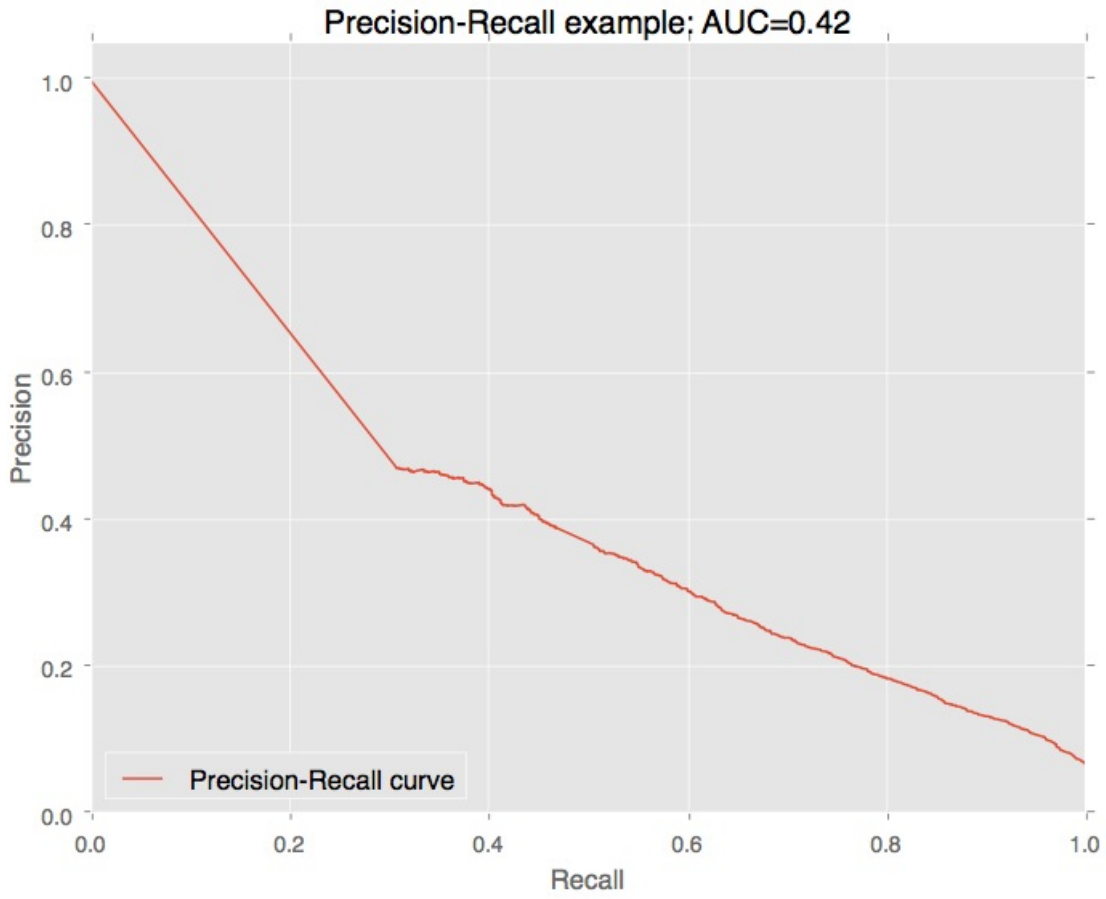


FIGURE A0-3 **Gradient Boosting, Depth = 3**



Appendix A1: Summary of Dataset Variables

1-i. Summary Data for Serious Delinquencies in the Last Two Years:

FIGURE A1 **Serious Delinquencies in the Last Two Years**

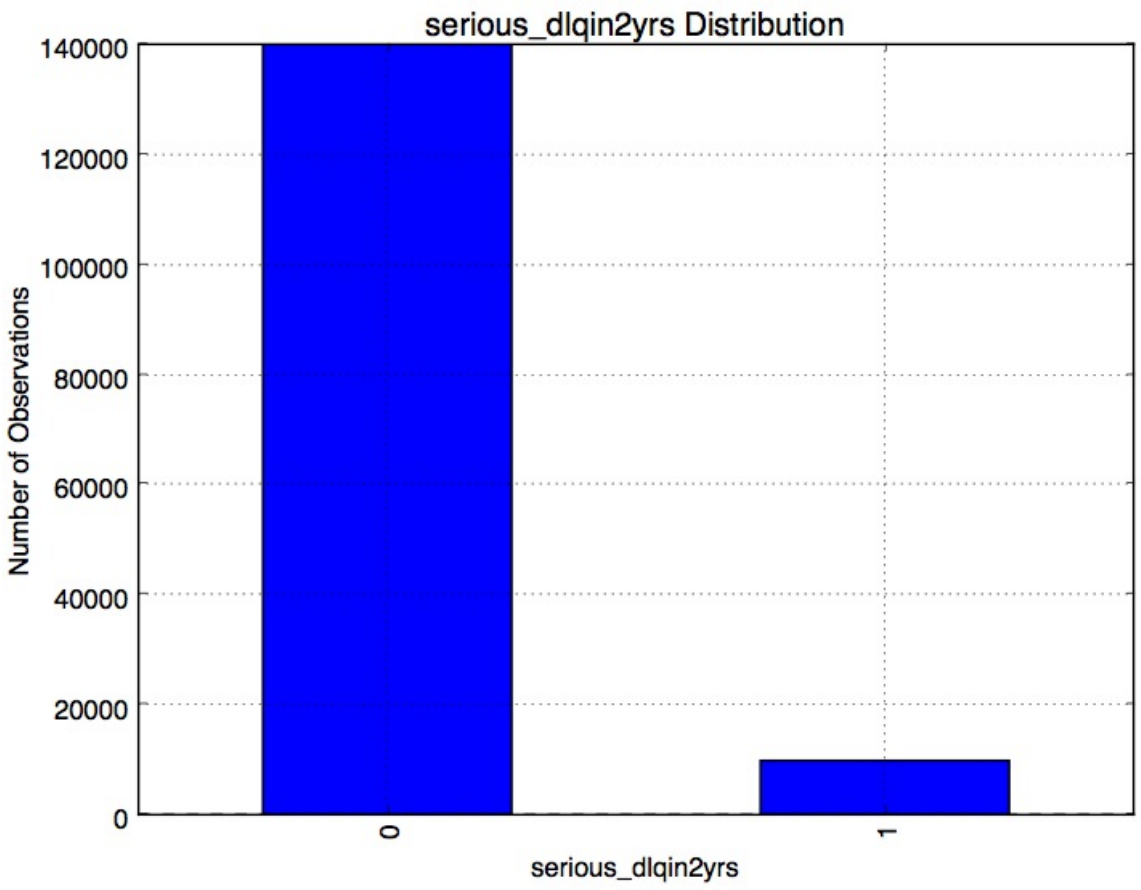


TABLE A1 **Serious Delinquencies in the Last Two Years**

N =	150,000
Missing =	0
Percent Delinquent =	6.68%
Mean =	0.067
Median =	0.00
Min, Max =	(0, 1)

Source: cs-training.csv

1-ii. Summary Data for Revolving Utilization of Unsecured Lines:

FIGURE A2 **Revolving Utilization of Unsecured Lines**

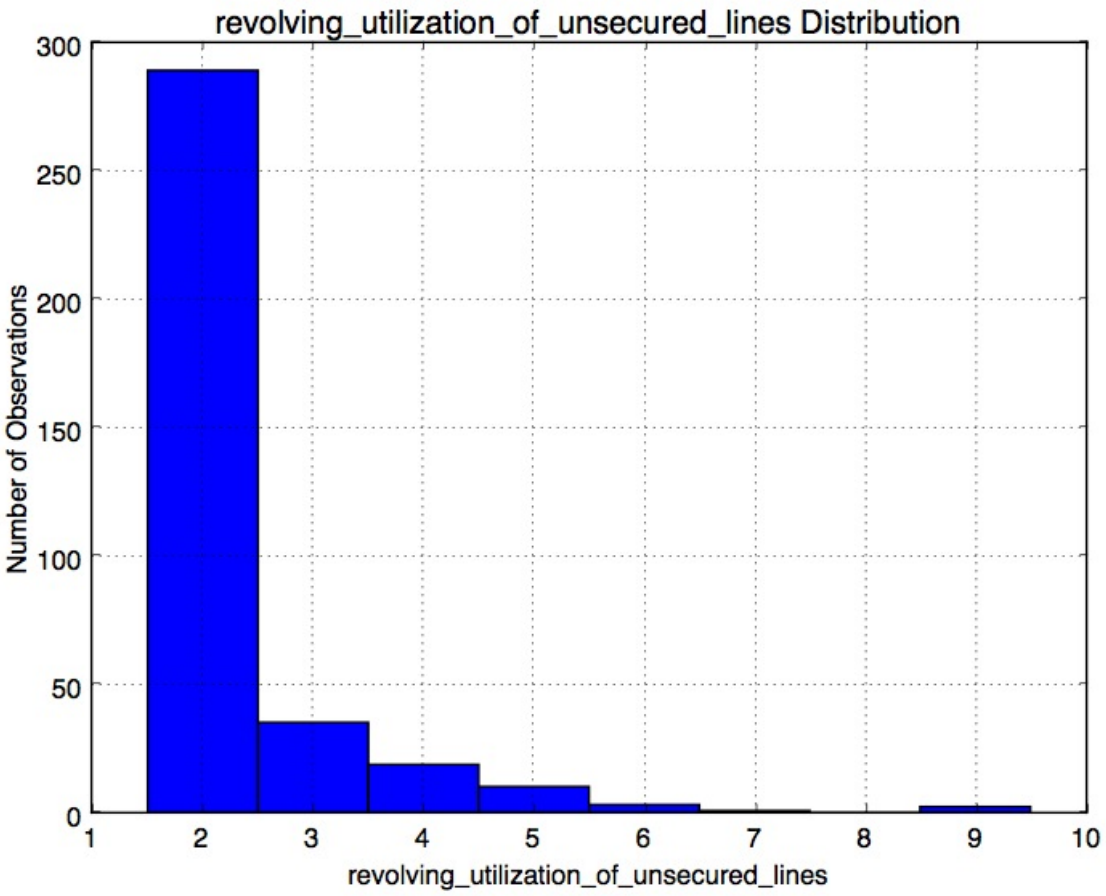


TABLE A2 **Revolving Utilization of Unsecured Lines**

N =	150,000
Missing =	0
Mean =	6.05
Median =	0.15
Standard Deviation =	249.76
Min, Max =	(0, 50,708)

Source: cs-training.csv

1-iii. Summary Data for Age:

FIGURE A3 **Age**

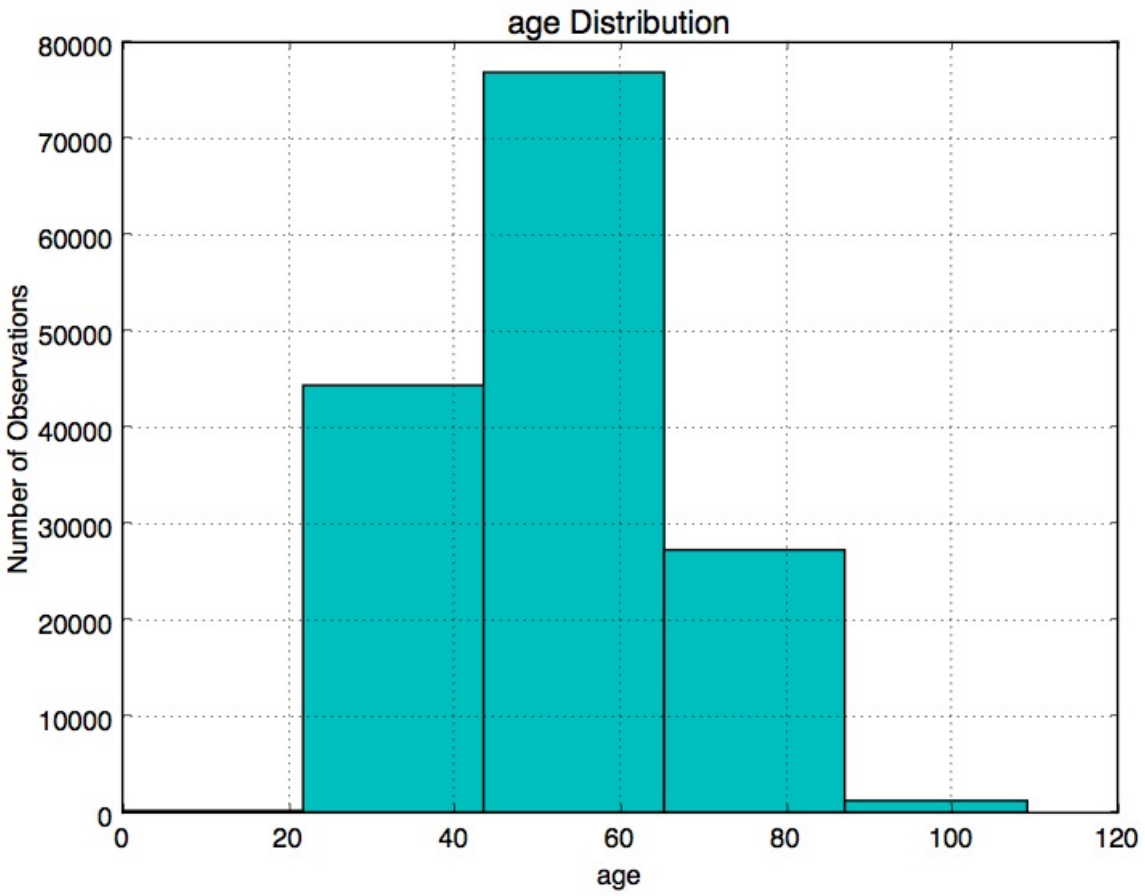


TABLE A3 **Age**

N =	150,000
Missing =	0
Mean =	52
Median =	52
Mode =	49
Standard Deviation =	15
Min, Max =	(0, 109)

Source: cs-training.csv

1-iv. Summary Data for Number of Times 30-59 Days Past Due (Not Worse):

FIGURE A4 **Number of Times 30-59 Days Past Due (Not Worse)**

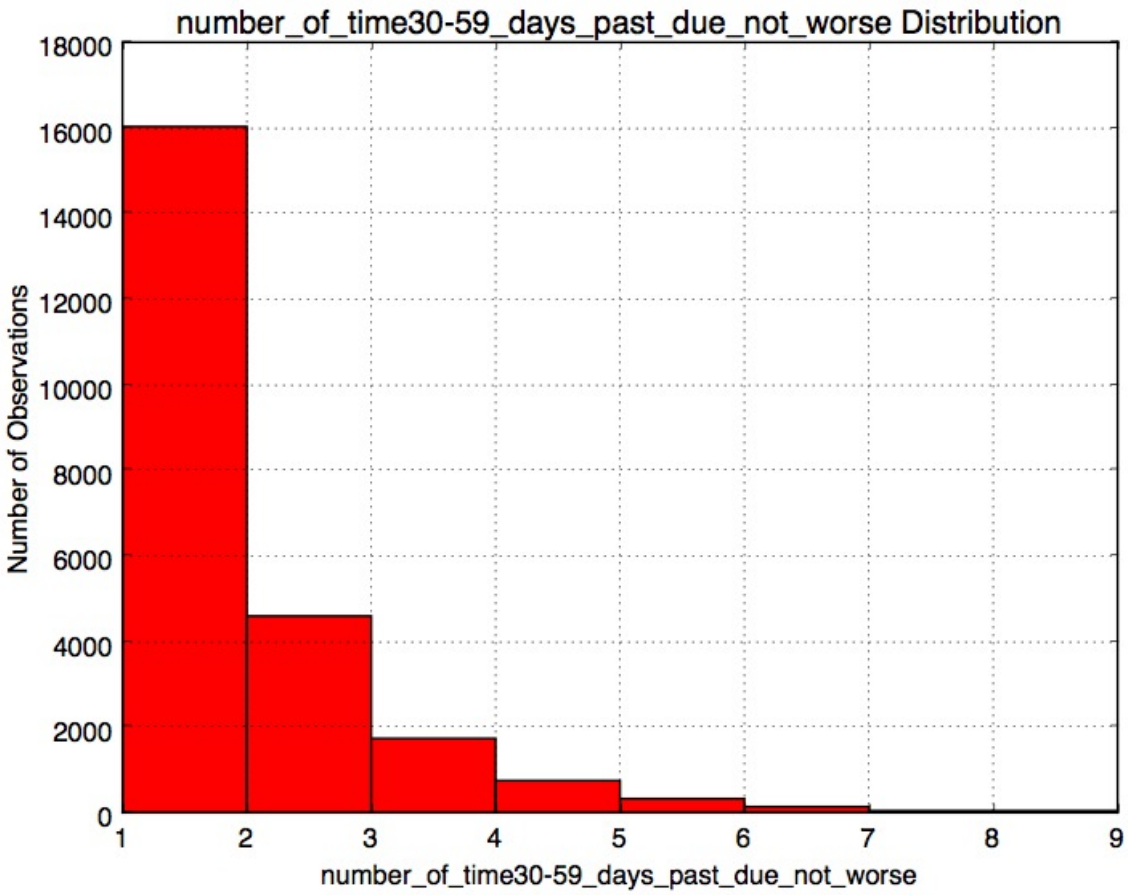


TABLE A4 **Number of Times 30-59 Days Past Due (Not Worse)**

N =	150,000
Missing =	0
Mean =	0.42
Median =	0.00
Standard Deviation =	4.19
Min, Max =	(0, 98)

Source: cs-training.csv

1-v. Summary Data for Number of Times 60-89 Days Past Due (Not Worse):

FIGURE A5 **Number of Times 60-89 Days Past Due (Not Worse)**

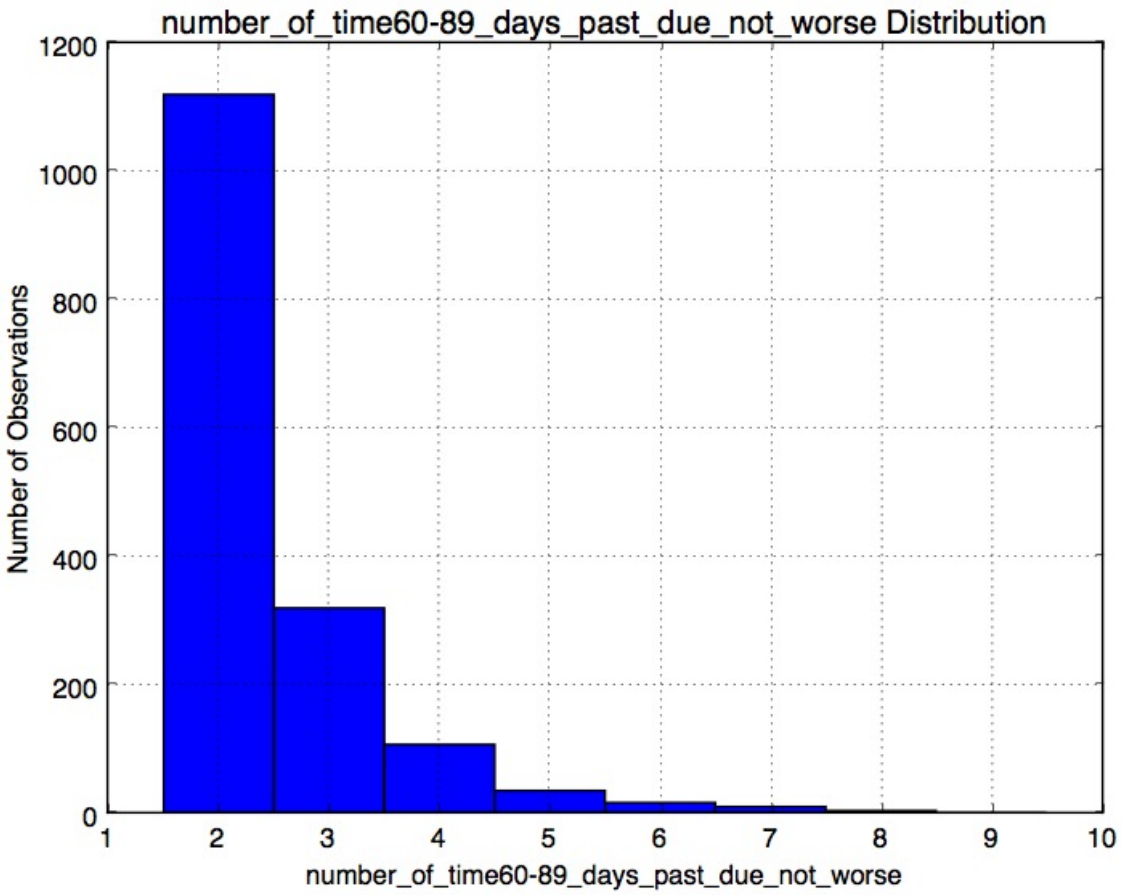


TABLE A5 **Number of Times 60-89 Days Past Due (Not Worse)**

N =	150,000
Missing =	0
Mean =	0.24
Median =	0.00
Standard Deviation =	4.16
Min, Max =	(0, 98)

Source: cs-training.csv

1-vi. Summary Data for Debt Ratio:

FIGURE A6 **Debt Ratio**

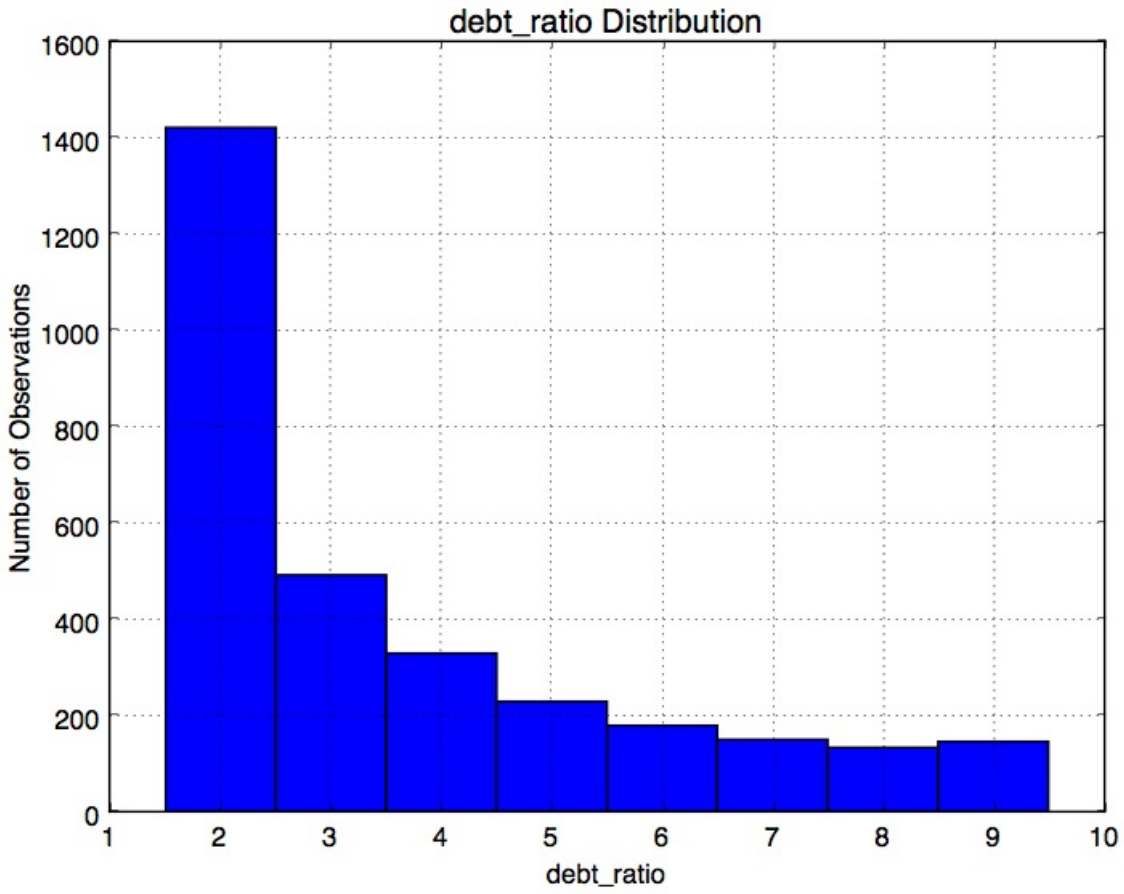


TABLE A6 **Debt Ratio**

N =	150,000
Missing =	0
Mean =	353.01
Median =	0.37
Standard Deviation =	2037.82
Min, Max =	(0, 329,664)

Source: cs-training.csv

1-vii. Summary Data for Monthly Income:

FIGURE A7 **Monthly Income**

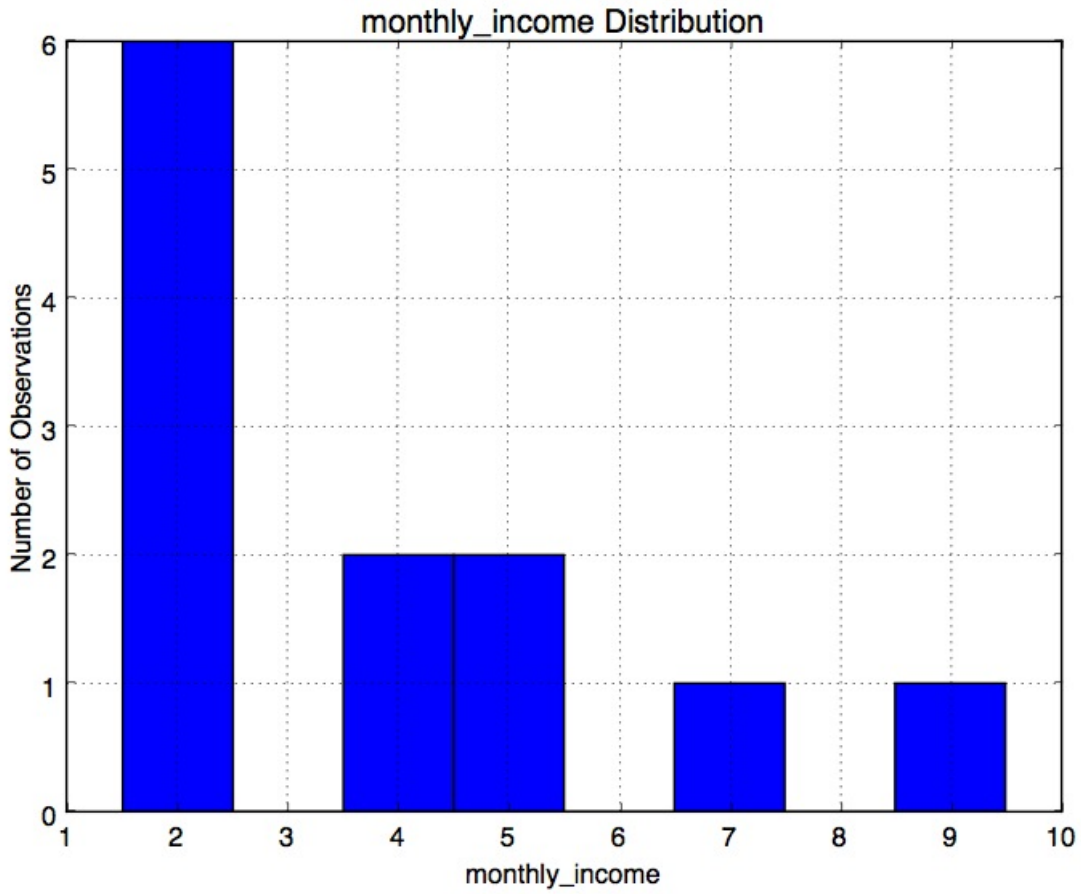


TABLE A7 **Monthly Income**

N =	120,269
Missing =	29,271
Mean =	6,670.22
Median =	5,400.00
Standard Deviation =	14,384.67
Min, Max =	(0, 3,008,750)

Source: cs-training.csv

1-viii. Summary Data for Number of Open Credit Lines and Loans:

FIGURE A8 **Number of Open Credit Lines and Loans**

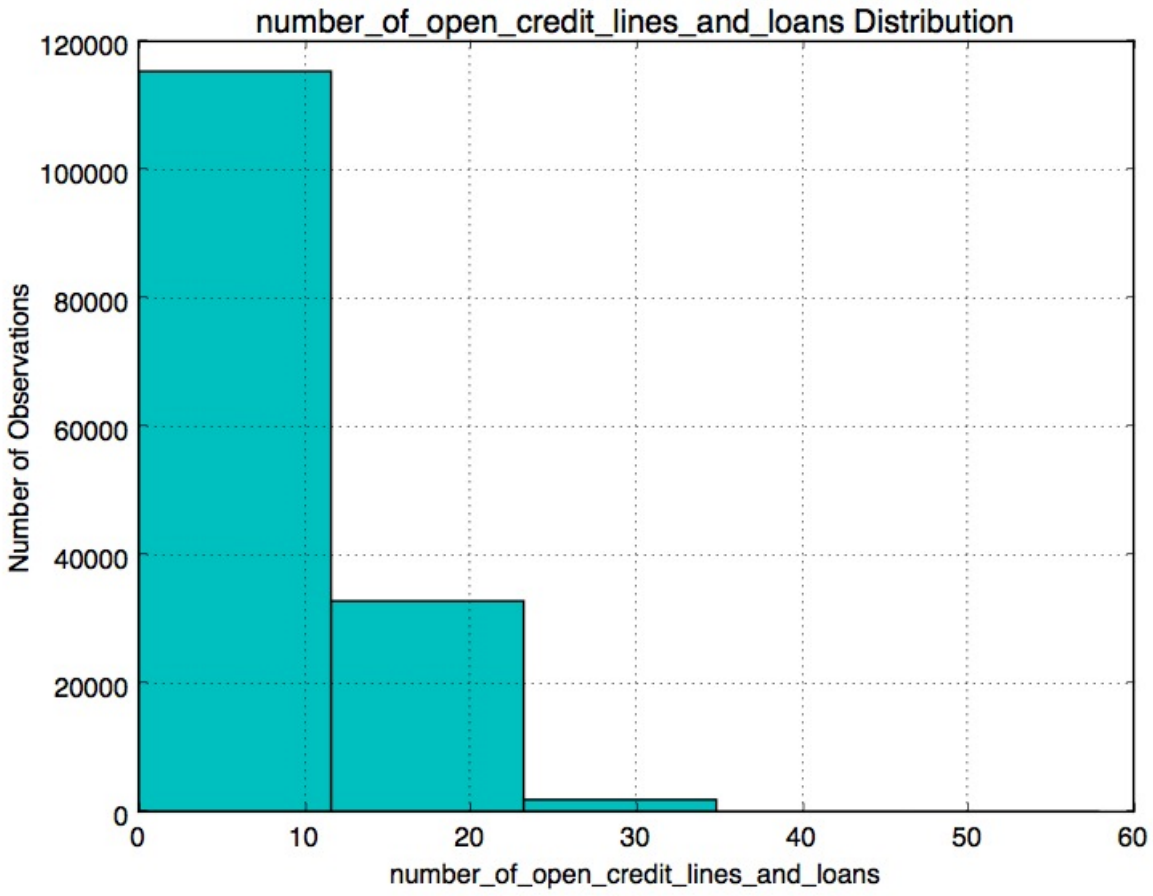


TABLE A8 **Number of Open Credit Lines and Loans**

N =	150,000
Missing =	0
Mean =	8.45
Median =	8.00
Mode =	6.00
Standard Deviation =	5.15
Min, Max =	(0, 58)

Source: cs-training.csv

1-ix. Summary Data for Number of Times 90 Days Late:

FIGURE A9 **Number of Times 90 Days Late**

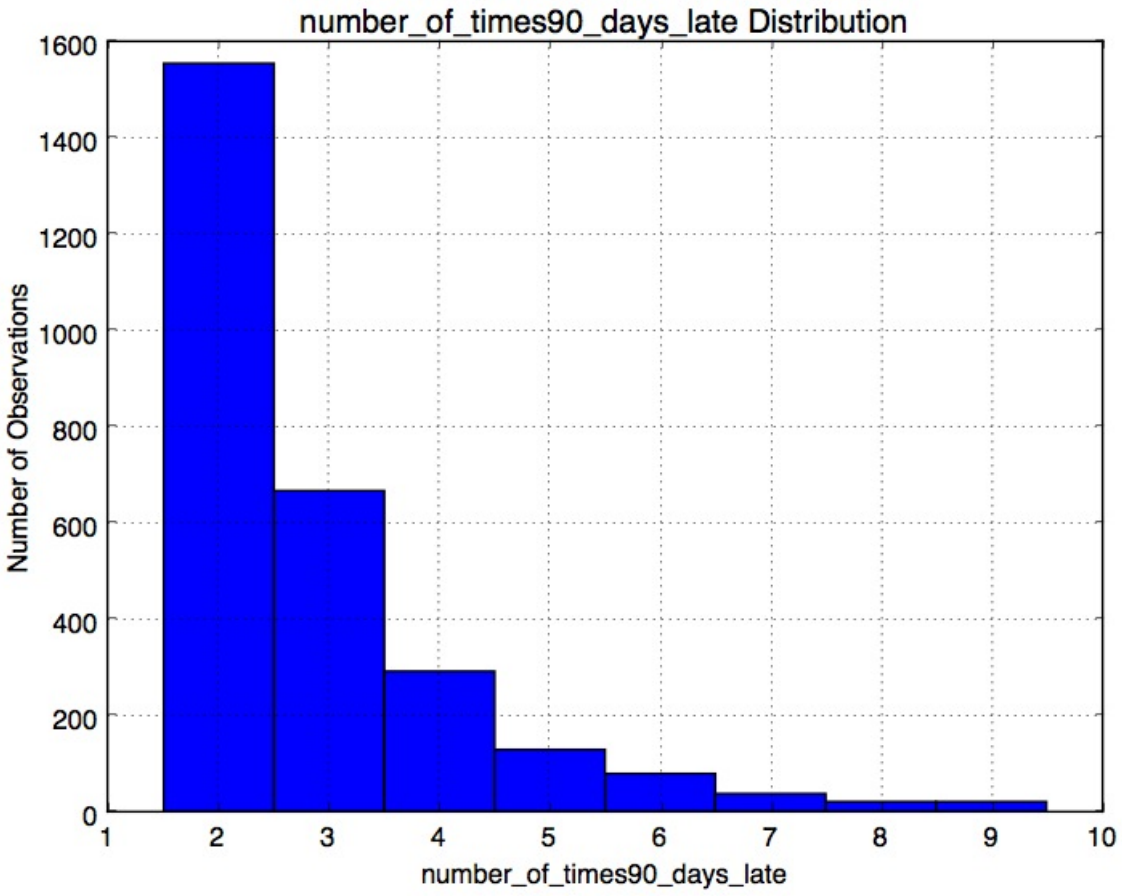


TABLE A9 **Number of Times 90 Days Late**

N =	150,000
Missing =	0
Mean =	0.27
Median =	0.00
Standard Deviation =	4.17
Min, Max =	(0, 98)

Source: cs-training.csv

1-x. Summary Data for Number of Real Estate Loans or Titles:

FIGURE A10 **Number of Real Estate Loans or Titles**

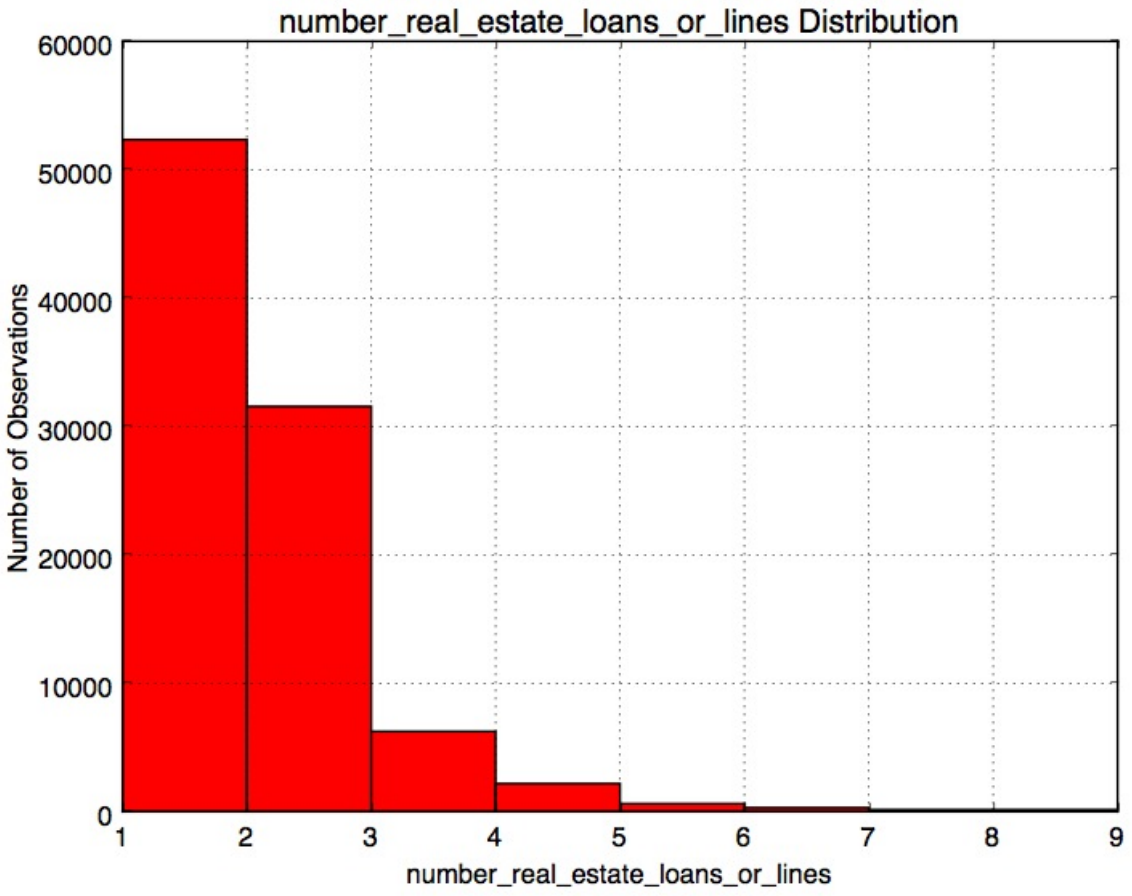


TABLE A10 **Number of Real Estate Loans or Titles**

N =	150,000
Missing =	0
Mean =	1.02
Median =	1.00
Standard Deviation =	1.13
Min, Max =	(0, 54)

Source: cs-training.csv

1-xi. Summary Data for Number of Dependents:

FIGURE A11 **Number of Dependents**

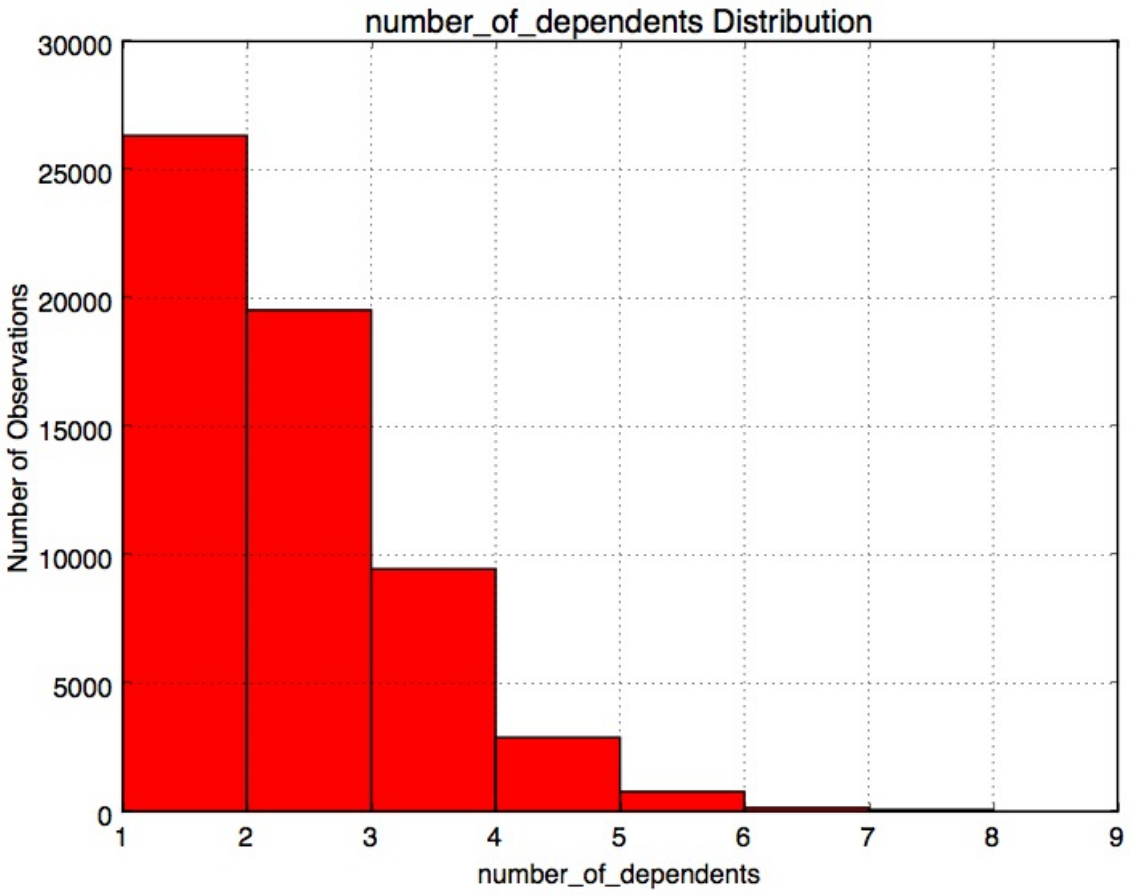


TABLE A11	Number of Dependents
N =	146,076
Missing =	3,924
Mean =	0.76
Median =	0.00
Standard Deviation =	1.12
Min, Max =	(0, 20)

Source: cs-training.csv