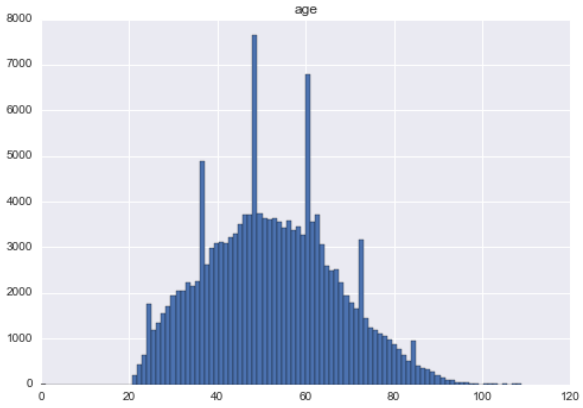
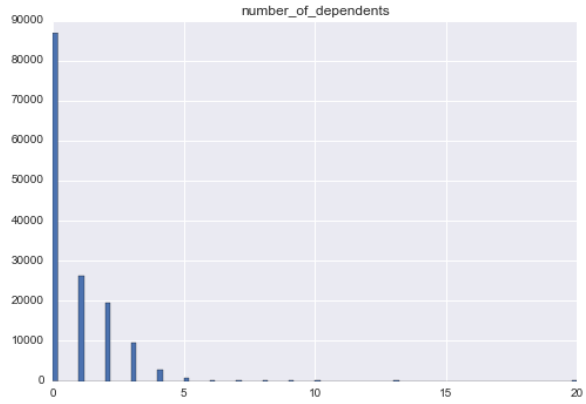
Abhi Gupta
CAPP 30254
HW 2

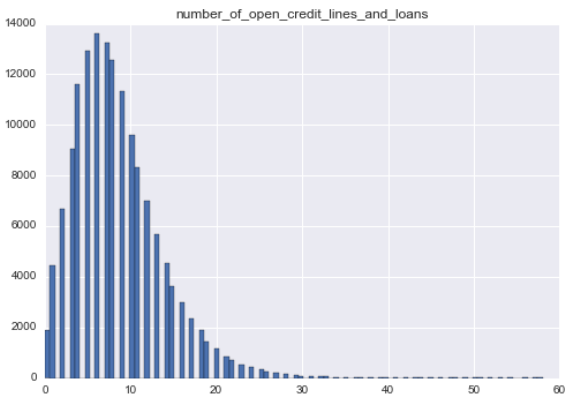
The dataset presented here is used to model the risk of a serious delinquency as a function of various borrow characteristics- including age, income, number of dependents, debt ratio, and other relevant measures. The training dataset is summarized in the table below:

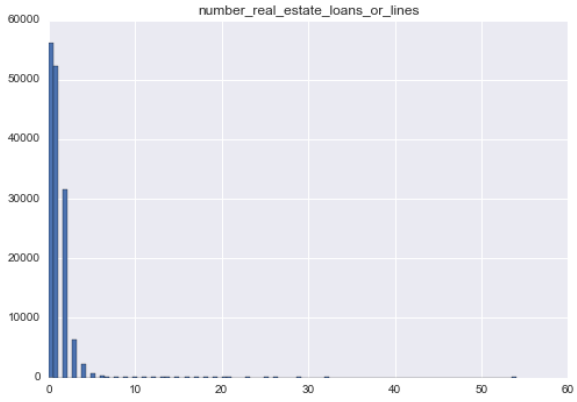
	count	mean	std	min	25%	50%	75%	max
serious_dlqin2yrs	150000	0.06684	0.249746	0	0	0	0	1
revolving_utilization_of_unsecured_lines	150000	6.048438	249.7554	0	0.029867	0.154181	0.559046	50708
age	150000	52.29521	14.77187	0	41	52	63	109
number_of_time30-59_days_past_due_not_worse	150000	0.421033	4.192781	0	0	0	0	98
debt_ratio	150000	353.0051	2037.819	0	0.175074	0.366508	0.868254	329664
monthly_income	120269	6670.221	14384.67	0	3400	5400	8249	3008750
number_of_open_credit_lines_and_loans	150000	8.45276	5.145951	0	5	8	11	58
number_of_times90_days_late	150000	0.265973	4.169304	0	0	0	0	98
number_real_estate_loans_or_lines	150000	1.01824	1.129771	0	0	1	2	54
number_of_time60-89_days_past_due_not_worse	150000	0.240387	4.155179	0	0	0	0	98
number_of_dependents	146076	0.757222	1.115086	0	0	0	1	20

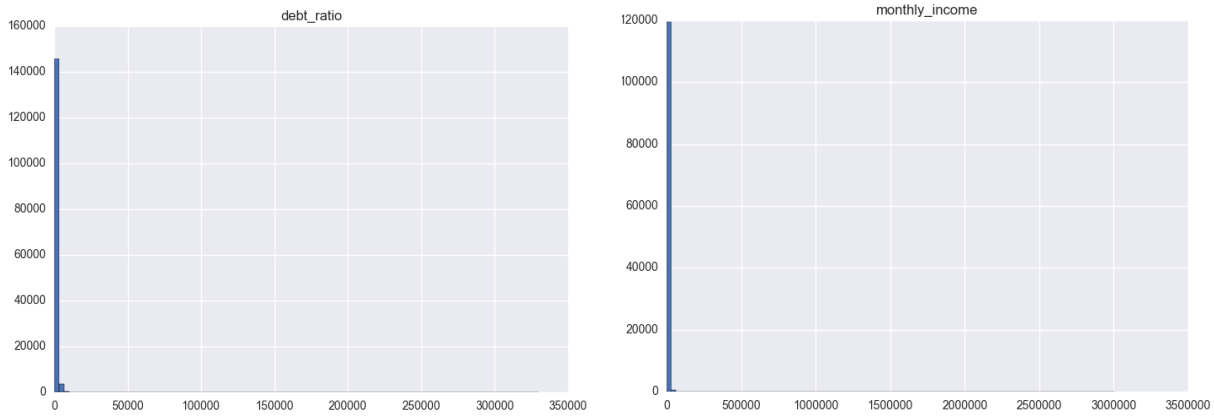
Additionally, many of these variables can be visualized meaningfully (the variables concerning numbers of days past due are not plotted, as they are not particularly illuminating):



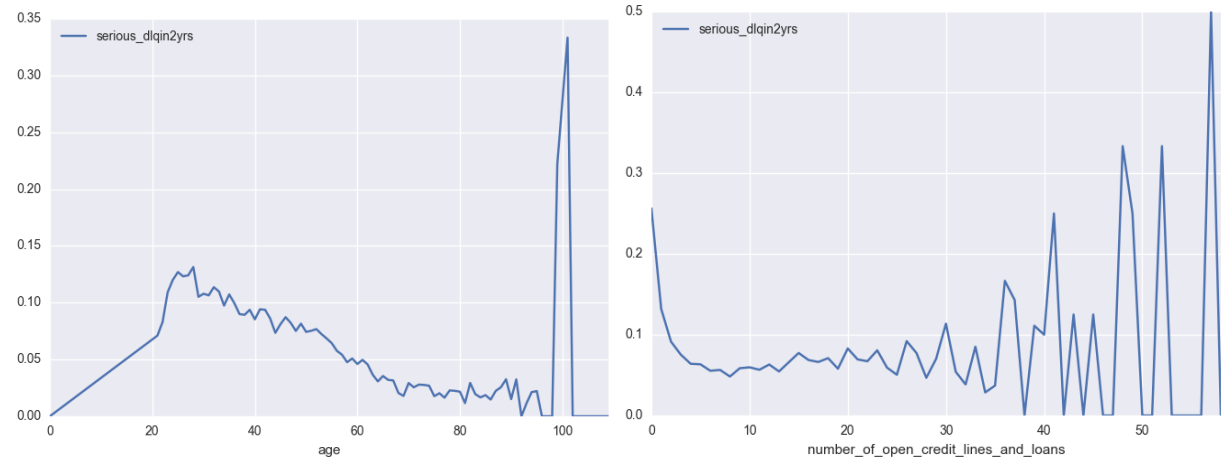








Note that for many of them, there are a few outlying observations that are meaningfully separated from the bulk of the observations. To clean the data, we fill in missing values with the median observation (this avoids assigning fractions of a dependent to those with missing data, for example) and trim the outliers to the value of the 90th percentile for revolving_utilization_of_unsecured_lines, debt_ratio, and monthly_income, as it is unclear if these outlying values provide any insight into the behavior of the majority of the sample.
To gain insight into how various variables interact with serious_dlqin2yrs, it is helpful to plot various features against the conditional mean delinquency:



The non-smoothness results from the lack of observations at certain values of these features. In any event, the non-linear relationship seen here suggests it may be worthwhile to bin these variables or to fit some sort of non-linear model.

As a baseline, I ran a simple logistic regression using all of the original features (post trimming). Evaluated on the training set, the model had an accuracy of 93.43%. To get a sense of how the model may or may not be succeeding, consider the following:

Actual\Predicted	0	1
0	139649	372
1	9593	433

Thankfully, this model is not just assigning 0 to all data points. A linear SVM was fit as well, which also gave an accuracy of 93.39%. Its results were essentially identical:

Actual\Predicted	0	1
0	139746	228
1	9686	340

The next model considered substitutes binned versions of age and number_of_open_credit_lines_and_loans for their original continuous versions, while keeping the logistic regression approach. This approach is somewhat better, with an accuracy of 93.87%.

Actual\Predicted	0	1
0	139859	115
1	9077	949

Since this model is significantly better at identifying delinquencies, it is the best of the models considered here.