Sirui Feng

Machine Learning for Public Policy

April 13, 2016

Homework 2

The purpose of this analysis is to identify potential serious delinquency in two years. A summary of statistics and histograms show that our sample is very skewed—97 percent of the sample have negative labels for serious delinquency in two years. Furthermore, the correlation table shows that delinquency has the highest correlation with number of time30-59 days past due not worse.

I used the logistic regression model to predict delinquency score. I have conducted cross validation with a test size of 0.2 and used all available features as explanatory variables. According to the evaluation on the logit model’s performance, it has an accuracy of 93 percent, precision of 61 percent, but only 0.09 for f1 score (1 is the perfect score). A revisit of the sample distribution reminds us that 93 percent of the sample has a negative label, which implies that without any models, if we label all observations as negative, we will obtain a 93 percent of accuracy score. This explains why the f1 score is low—the logit model does not provide much information to our knowledge.

The prediction results of the test file have been saved to output/logit\_result.csv.

**Field Name: Serious Dlqin2yrs**

Mean: 0.07

Standard Deviation: 0.25

Median: 0.0

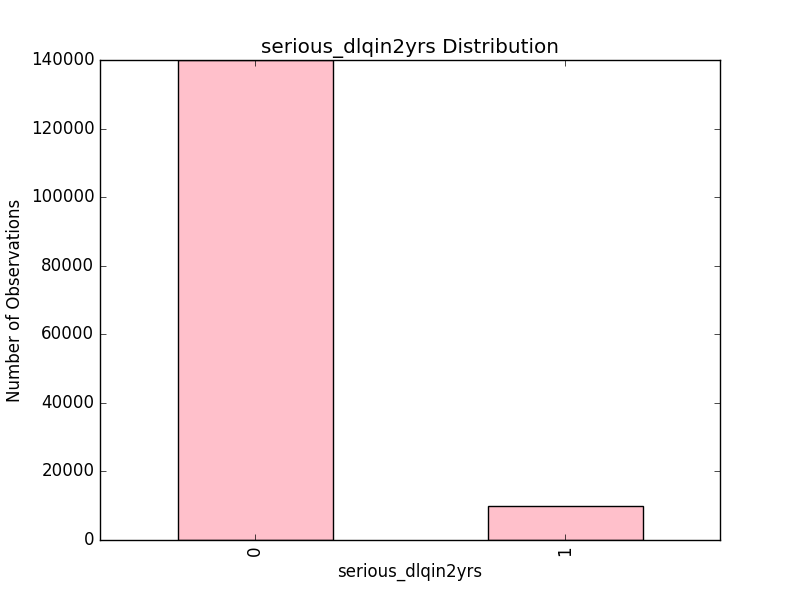
Min: 0

Max: 1

Mode: 0 0

Non-missing Value Count: 150000

Missing Value Count: 0



**Field Name: Revolving Utilization of Unsecured Lines**

Mean: 6.05

Standard Deviation: 249.76

Median: 0.154180737

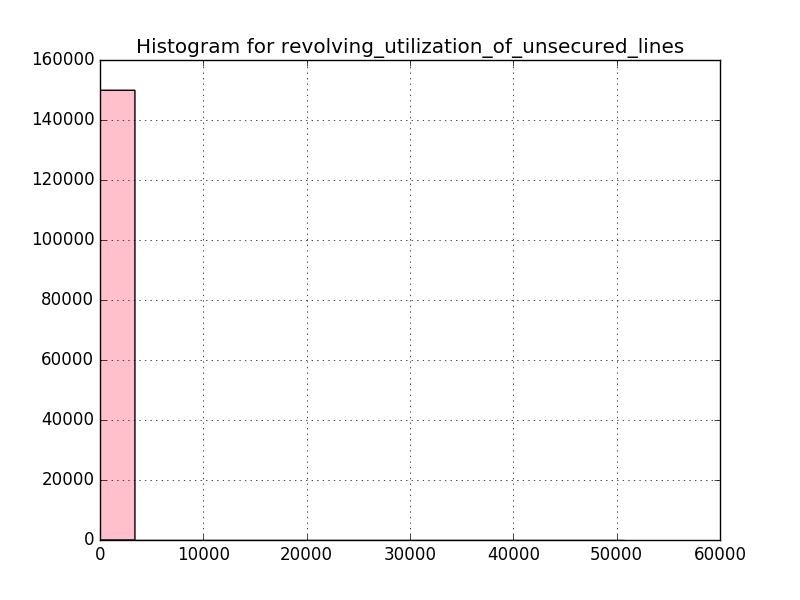
Min: 0.0

Max: 50708.0

Mode: 0 0.0

Non-missing Value Count: 150000

Missing Value Count: 0



**Field Name: Age**

Mean: 52.3

Standard Deviation: 14.77

Median: 52.0

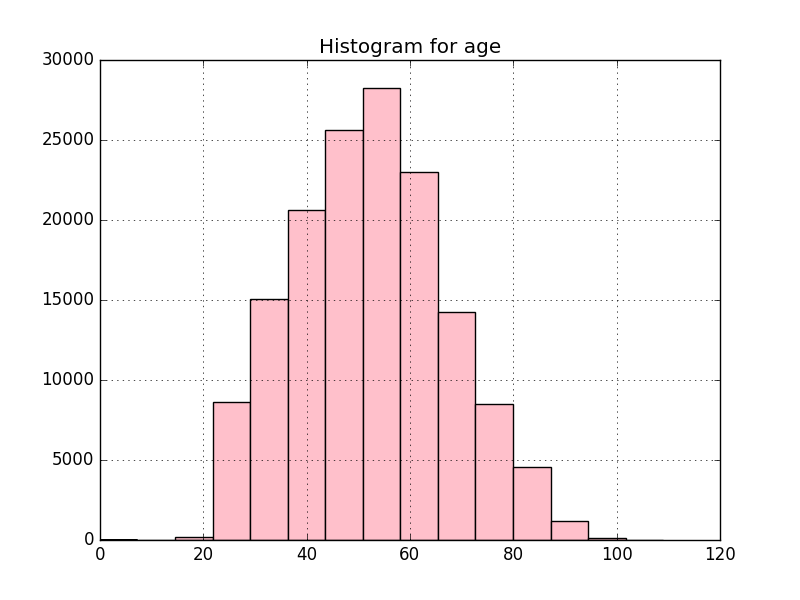
Min: 0

Max: 109

Mode: 0 49

Non-missing Value Count: 150000

Missing Value Count: 0



**Field Name: Number of Time 30-59 Days Past Due Not Worse**

Mean: 0.42

Standard Deviation: 4.19

Median: 0.0

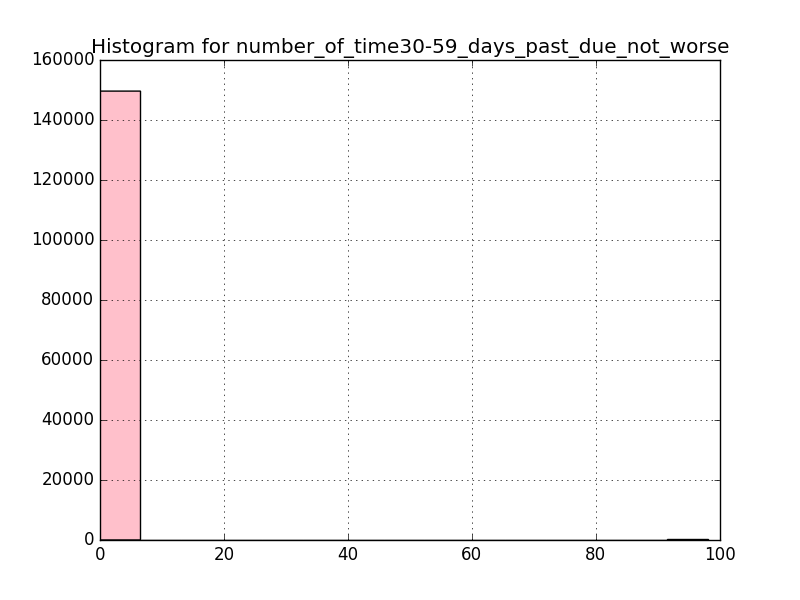
Min: 0

Max: 98

Mode: 0 0

Non-missing Value Count: 150000

Missing Value Count: 0



**Field Name: Debt Ratio**

Mean: 353.01

Standard Deviation: 2037.82

Median: 0.366507841

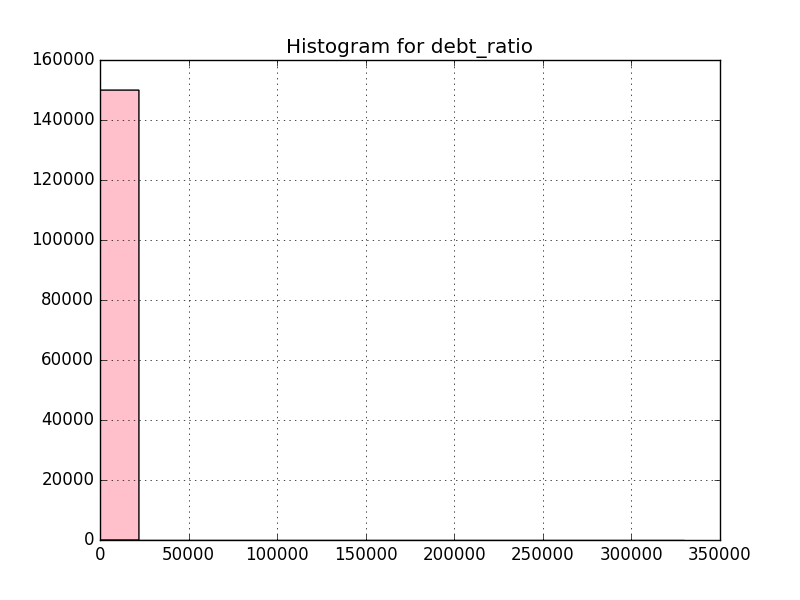
Min: 0.0

Max: 329664.0

Mode: 0 0.0

Non-missing Value Count: 150000

Missing Value Count: 0



**Field Name: Monthly Income**

Mean: 6670.22

Standard Deviation: 14384.67

Median: 5400.0

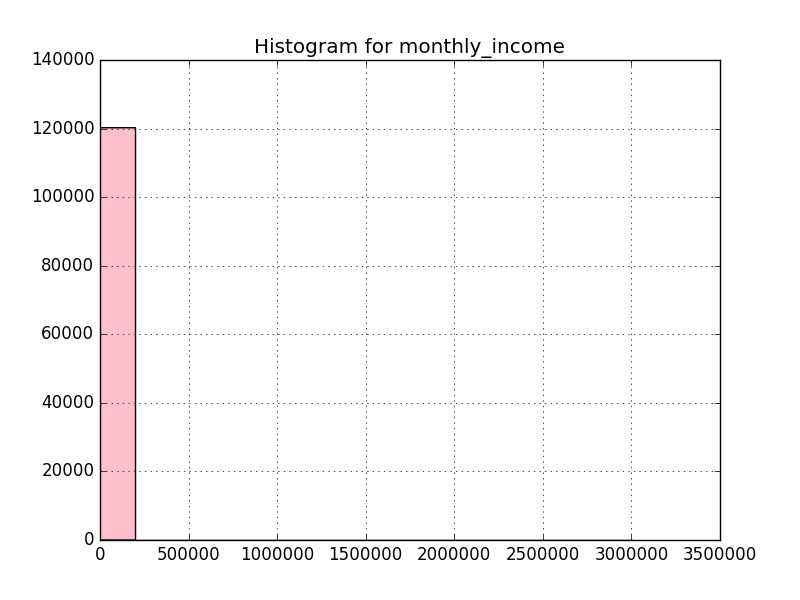
Min: 0.0

Max: 3008750.0

Mode: 0 5000.0

Non-missing Value Count: 120269

Missing Value Count: 29731



**Field Name: Number of open credit lines and loans**

Mean: 8.45

Standard Deviation: 5.15

Median: 8.0

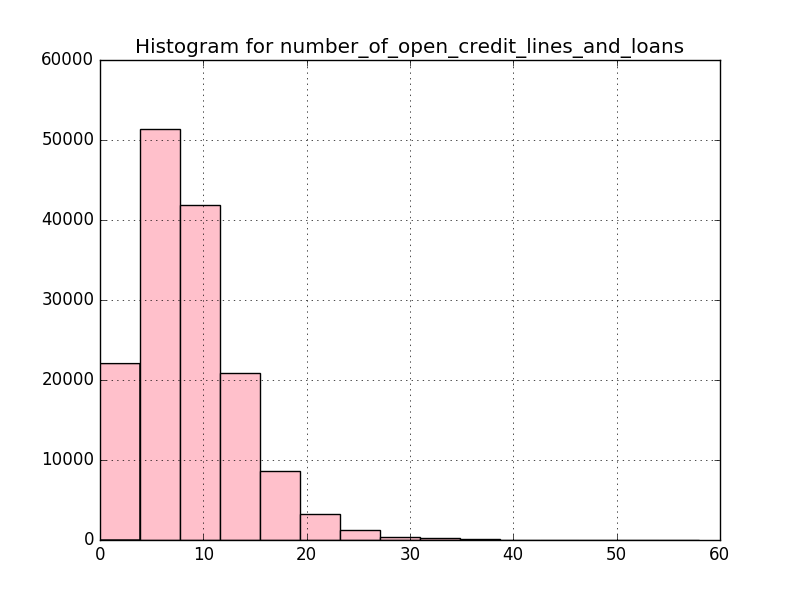
Min: 0

Max: 58

Mode: 0 6

Non-missing Value Count: 150000

Missing Value Count: 0



**Field Name: Number of times 90 days late**

Mean: 0.27

Standard Deviation: 4.17

Median: 0.0

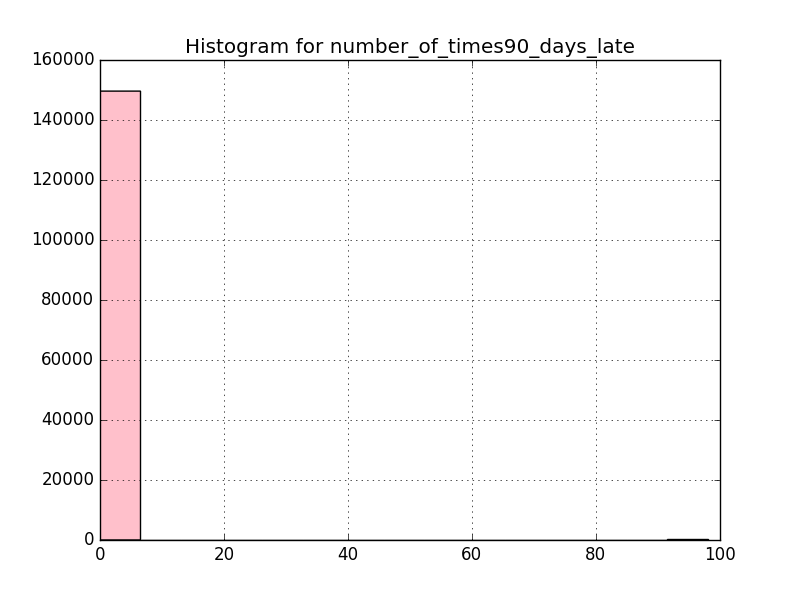
Min: 0

Max: 98

Mode: 0 0

Non-missing Value Count: 150000

Missing Value Count: 0



**Field Name: Number real estate loans or lines**

Mean: 1.02

Standard Deviation: 1.13

Median: 1.0

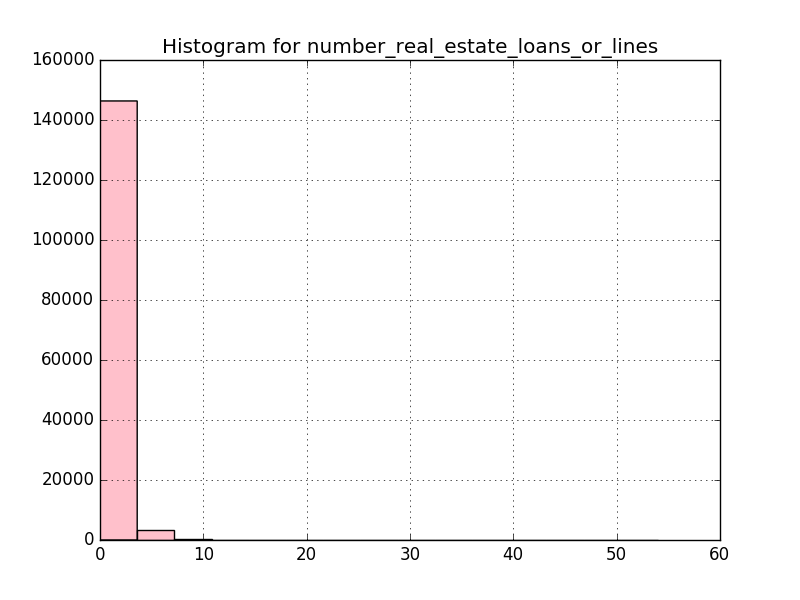
Min: 0

Max: 54

Mode: 0 0

Non-missing Value Count: 150000

Missing Value Count: 0



**Field Name: Number of time60-89 days past due not worse**

Mean: 0.24

Standard Deviation: 4.16

Median: 0.0

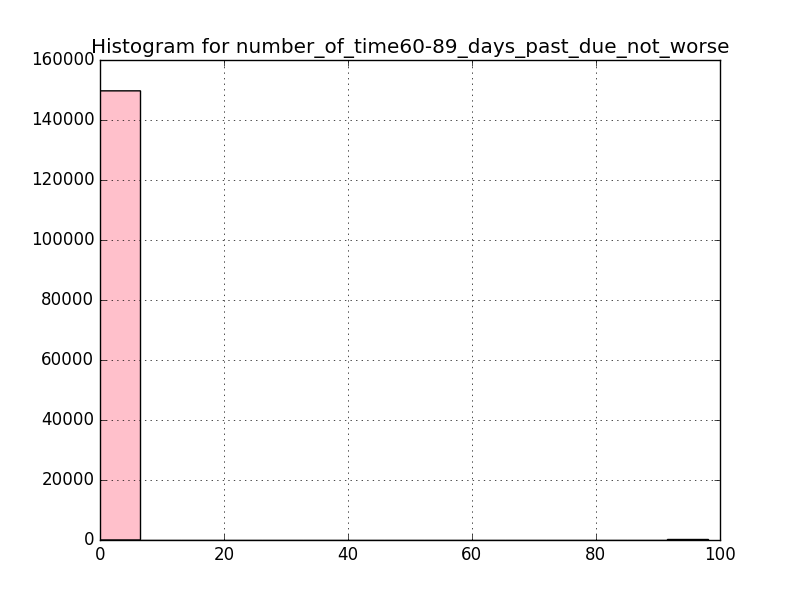
Min: 0

Max: 98

Mode: 0 0

Non-missing Value Count: 150000

Missing Value Count: 0



**Field Name: Number of Dependents**

Mean: 0.76

Standard Deviation: 1.12

Median: 0.0

Min: 0.0

Max: 20.0

Mode: 0 0.0

Non-missing Value Count: 146076

Missing Value Count: 3924

**Correlation Table**

|  |  |
| --- | --- |
|  | Serious dlqin2yrs |
| Revolving utilization of unsecured lines | -0.001801503 |
| Age | -0.115385518 |
| Number of time30-59 days past due not worse | 0.125586965 |
| Debt ratio | -0.00760212 |
| Monthly income | -0.019745547 |
| Number of open credit lines and loans | -0.029668568 |
| Number of times90 days late | 0.117174613 |
| Number real estate loans or lines | -0.007038116 |
| Number of time60-89 days past due not worse | 0.102260861 |
| Number of dependents | 0.046047944 |