DANA LITOVSKY ALCALA Machine Learning, report 2

In order to understand what factors make individuals more likely to not be able to pay debt, I will will analyze and evaluate data with the following features.

"SeriousDlqin2yrs', 'RevolvingUtilizationOfUnsecuredLines', 'age', 'NumberOfTime30-59DaysPastDueNotWorse', 'DebtRatio', 'MonthlyIncome', 'NumberOfOpenCreditLinesAndLoans', 'NumberOfTimes90DaysLate', 'NumberRealEstateLoansOrLines', 'NumberOfTime60-89DaysPastDueNotWorse', 'NumberOfDependents"

Where `SeriousDlqin2yrs` is 1 if an individual experienced 90 days past due delinquency, or worse, and 0 otherwise..

To better understand the data, I first look at some summary statistics.
For example, the **mean** of all columns:

```
SeriousDlqin2yrs                          0.066840
RevolvingUtilizationOfUnsecuredLines      6.048438
age                                      52.295207
NumberOfTime30-59DaysPastDueNotWorse      0.421033
DebtRatio                               353.005076
MonthlyIncome                          6670.221237
NumberOfOpenCreditLinesAndLoans           8.452760
NumberOfTimes90DaysLate                   0.265973
NumberRealEstateLoansOrLines              1.018240
NumberOfTime60-89DaysPastDueNotWorse      0.240387
NumberOfDependents                        0.757222
```

And the **mode**:

```
SeriousDlqin2yrs                      0       0.0
RevolvingUtilizationOfUnsecuredLines  0       0.0
age                                   0      49.0
NumberOfTime30-59DaysPastDueNotWorse  0       0.0
DebtRatio                             0       0.0
MonthlyIncome                         0    5000.0
NumberOfOpenCreditLinesAndLoans       0       6.0
NumberOfTimes90DaysLate               0       0.0
NumberRealEstateLoansOrLines          0       0.0
NumberOfTime60-89DaysPastDueNotWorse  0       0.0
NumberOfDependents                    0       0.0
```

The **median** for most other features is similar, but here is the median for age and income, which are different than the mean because the mean is more sensitive to outliers.

```
age                                          52.000000
MonthlyIncome                              5400.000000
```

The number of dependents could also be an interesting feature. Here is the count for number of dependents. This shows that most people have no dependents.

0	86902
1	26316
2	19522
3	9483
4	2862
5	746
6	158
7	51
8	24
10	5
9	5
20	1
13	1

Missing values pose a challenge to work with data. Monthly Income and Number of dependents have many missing values (other features have no missing values):

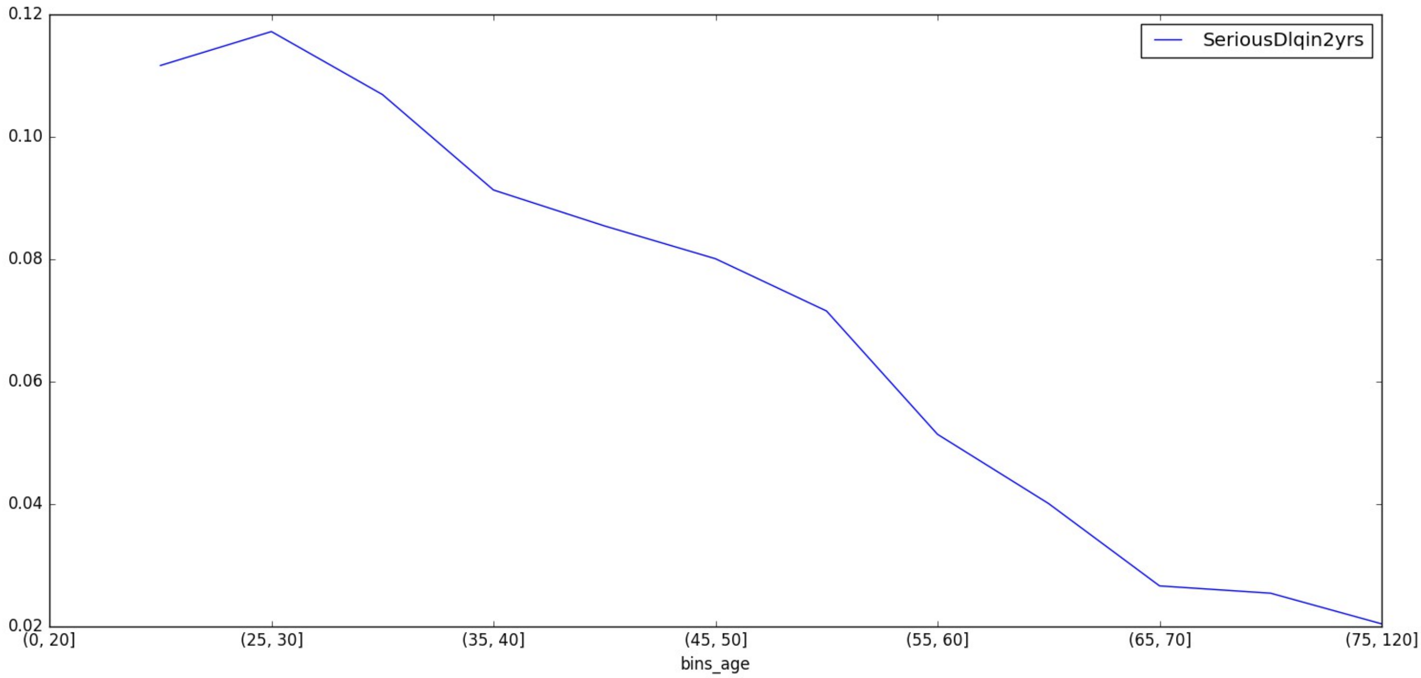
MonthlyIncome	29731
NumberOfDependents	3924

I chose to impute MonthlyIncome with the median because I believe it is a statistic that is less susceptible to large earning outliers. For example, the largest value for MonthlyIncome is `3008750.0`, much larger than both mean and median.

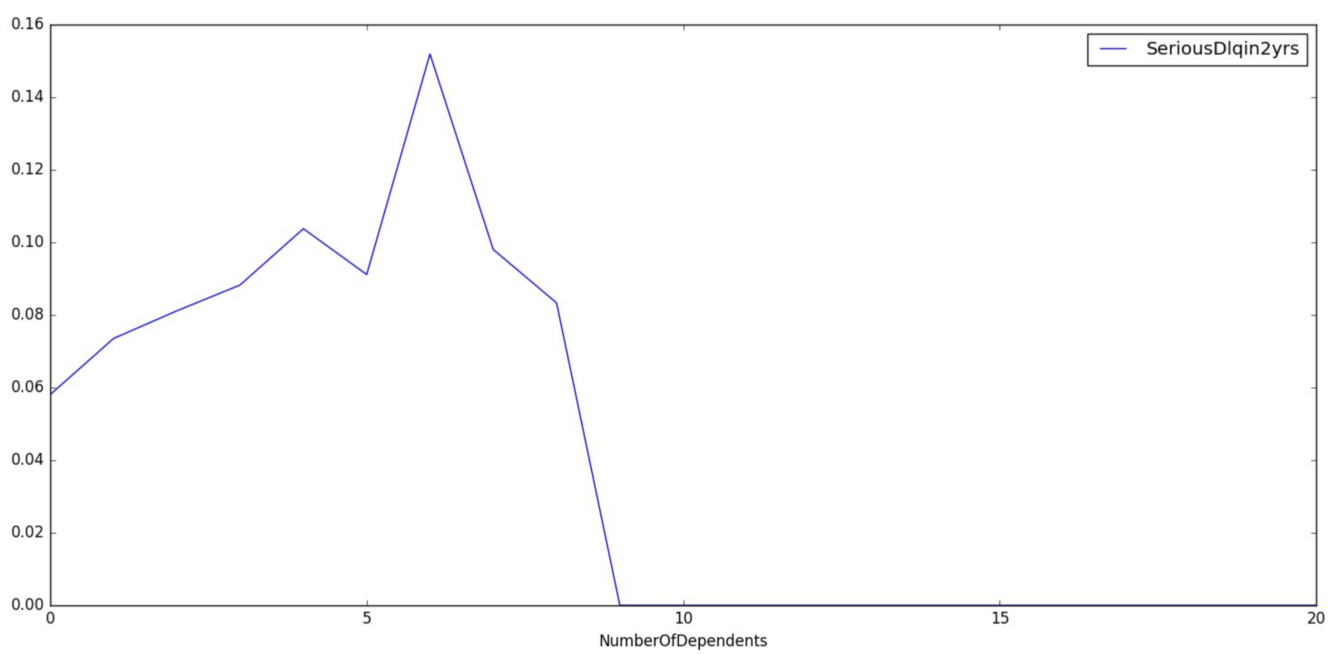
For Number of Dependents, I imputed with the mode (also the median) of 0 dependents, because more than 60% of people indicated they had no dependents.

To better understand the data, here are some graphs.

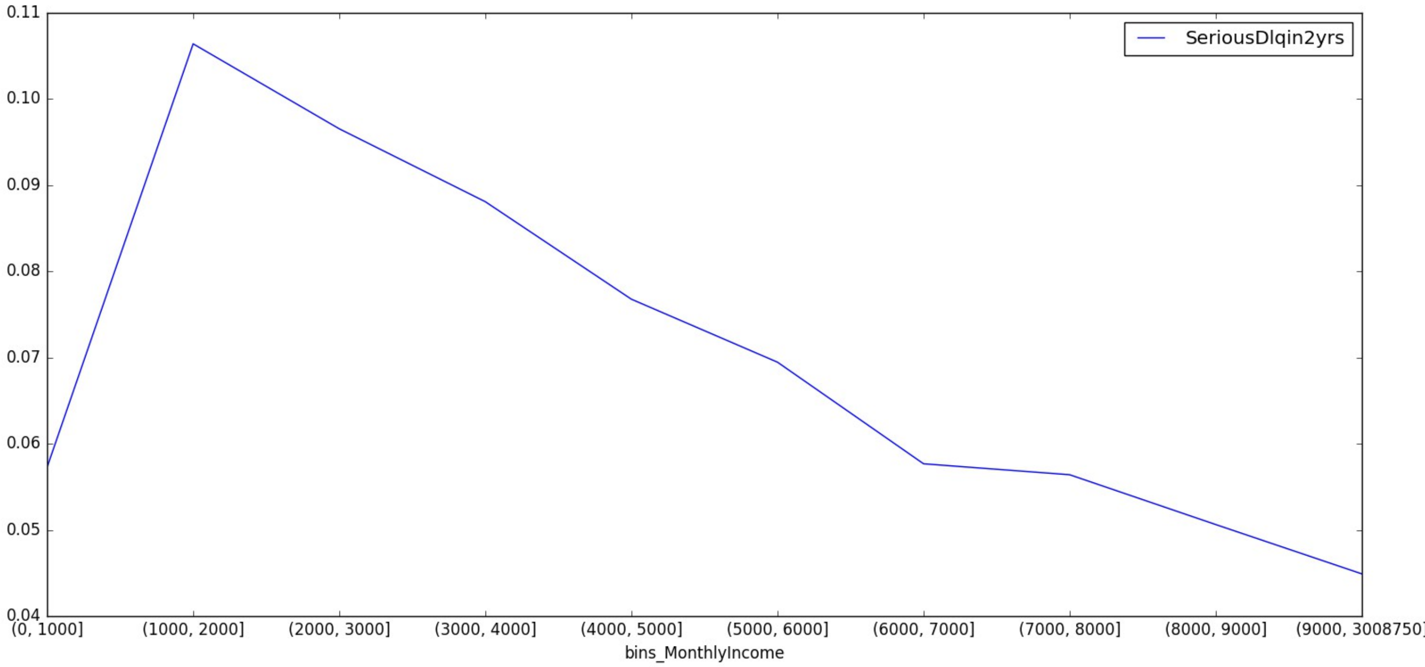
I created bin categories for age. This image graphs the mean `SeriousDlqin2yrs` per age category. This shows that most people who had serious issues paying debt are in their 20s.



The following graph shows mean `SeriousDlqin2yrs` for number of dependents. I see that the more dependents an individual has, the more trouble she has repaying debt, as we would suspect. The large drop in the data after 9 dependents could be a data anomaly, but it most likely reflects the fact that most people in our dataset did not have many, if any, dependents.



I also created category bins for income, such as income [0, 1000], [1000, 2000] … with the last bin containing most of the outliers. This graph of mean `SeriousDlqin2yrs` for income shows a very clear relationship. It appears that people with income around $2000 a month have a tough time paying back loan, and more monthly income steadily declines in rates of `SeriousDlqin2yrs`.



Creating and Evaluating a Model to predict loan default

To predict which individuals would be more likely to default on debt, I created a logistic regression model which will give category values of 0 or 1 for `SeriousDlqin2yrs` for new data. In general, the logistic regression model estimates the probability of an outcome (in our case `SeriousDlqin2yrs`) being 1.

I built the model using 80% of the data, and left 20% to test the model's predictions. 80-20 is a very common split of data, but a more robust approach could include creating multiple “folds”, or splits, of the data and repeatedly creating and testing a model.

To give the model the best chance of predicting accurately, I created additional features. These include the age and income bins, the log of monthly income, and a scaled monthly income.

I evaluated my model's predictions with accuracy score, which compares the true value of `SeriousDlqin2yrs` to my model's prediction. This is certainly not the best evaluation model because most of the data has `SeriousDlqin2yrs` = 0, so even if I randomly assigned

0 to every prediction, I would get a very high accuracy score.
It is very common to test

While we have many features to work with, there is often a best group of features that does a better job of predicting. It is very common to manually test what group of features predicts best. I decided to use an approach called recursive feature elimination, which recursively considers smaller sets of features. There are many other approaches to do this kind of task, but I found this approach to be effective in increasing the accuracy score, albeit by 0.001 points.

If I used all features, including the ones I created, my accuracy score is:

`0.931566666667`

While using the recursive feature elimination method, which created the best model with the following features:

```
'NumberOfTime30-59DaysPastDueNotWorse',
'NumberOfTimes90DaysLate',
'NumberOfTime60-89DaysPastDueNotWorse',
'NumberOfDependents',
'log_MonthlyIncome',
'bins_age',
'bins_MonthlyIncome'
```

Returned an accuracy score of: **`0.931866666667`**

Overall, the logistic regression model is a simple way to predict loan default, but there can be more appropriate models and evaluation metrics for this type of data and task.