We need to come up with error rates by using *Naïve Bayes* Algorithm. When the data is examined, there are numerical and categorical variables. "*Zip Code*" and "*ID Number"* features are eliminated.

- Even if numerical variables will turn into categorical variables, I scaled the numerical data into [0,1] for convenience. "Age", "Experience", "Income", "CCAvg" features are scaled.
- Mortgage feature is turned into categorical variables as follows:
  If there is no payment for mortgage, i.e. if feature is "0", it is treated as there is no mortgage, i.e. it is 0. If there is an amount of payment for mortgage, the amount can be ignored, it fits into category 1. Results and reasons are explained later.
- Other numerical features are turn into categorical variables by using *discretize* function. Equal length method is used to categorize numerical variables. Different number of category is used. Different combinations are constructed and corresponding error rate is held in error matrix in the code section.
- To use Naïve Bayes to predict validation set classifiers, I coded *naivebayes* function which can be found in the folder.
- Rng function is used to get the same partitioning.
- Different number of categories can be used, however I tried 2 to 4 categories. It can be increased. After trying different combinations, the best combination and corresponding error rate as follows:

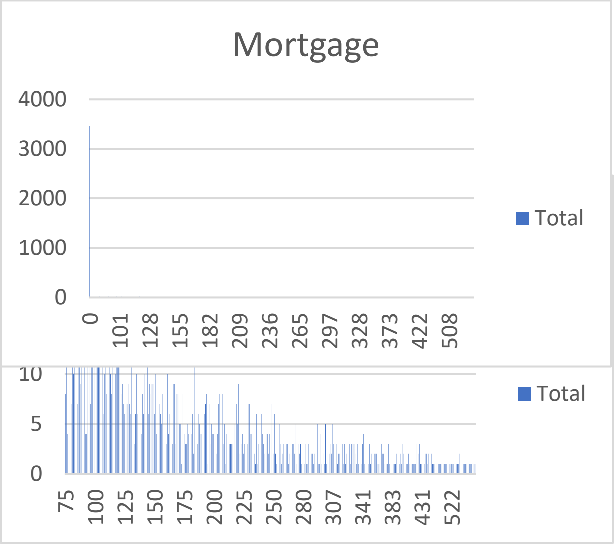
Age	Experience	Income	CCAvg	Mortgage	Error Rate
2	2	2	4	2	0.5

The minimum error rate can be found as 0.5, and related combination is on the table.

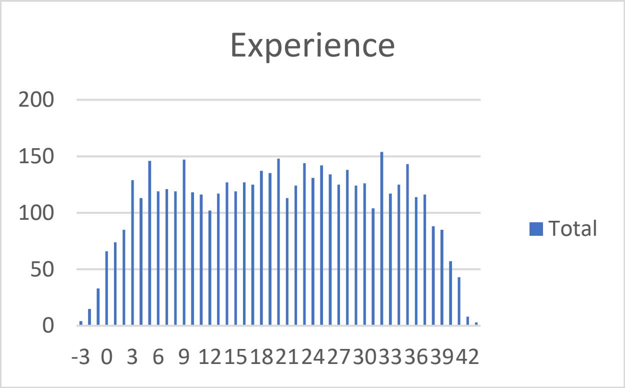
I tried another option for mortgage, and that is to discretize the mortgage as other numerical variables. I calculated error rate and corresponding categorized combination. From different combinations, I found minimum and average error rates for two options. As it can be seen in the following table, no need to discretize the mortgage variables.

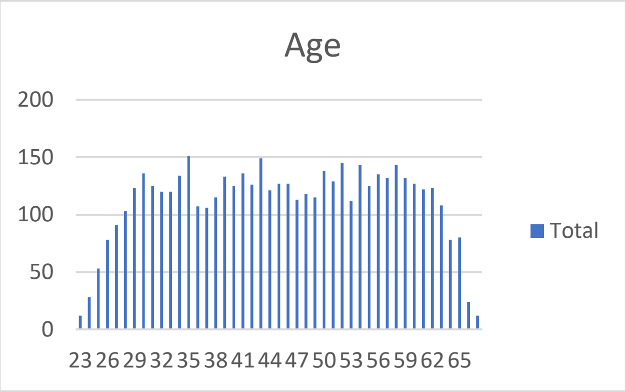
	Minimum	Average
Mortgage-not discretized	0.05	0.0664
Mortgage discretized	0.057	0.0705

Values of mortgage for frequency can be seen as above. Left-upper chart is together with the zero value, and the left-bottom chart is without zero values. I come up with that 0 values are dominating other values. Hence, I do not need to discretize this feature, instead, I separated directly into two categories as 0 and 1.



Age and Experience features seems as follows:





Even if I did not calculate probabilities of each values of experience and age, method of equal length can be used. Also, other numerical features are left-skewed, using less number of categories is appropriate for categorized.

Lastly, when we use to calculate error rate of Naïve Rule, it is 0.096. Then, we can say that my algorithm is better than that.