UMBC

AN HONORS UNIVERSITY IN MARYLAND

Department of Computer Science and Electrical Engineering

CMSC 491 – Introduction to Data Science
Assignment 3
Student: Sanaa Mironov

Background: The Lending Club dataset predict whether an individual will default on their current loan. The dataset features are from credit bureaus that evaluate an individual's worthiness to receive credit. Usually, this kind of data is collected to analyze the person and develop a credit score; this score helps different agencies that give out credit to individuals make a more informed decision.
The bases for scores are from different criteria such as payment records, frequency of payments, amount of debts, income to debt ratio, income before taxes, credit charge-offs, number of credit cards held, and number of inquiries an individual has requested the past 6,12,24,36 months. Different agencies use different models to come up with a score.

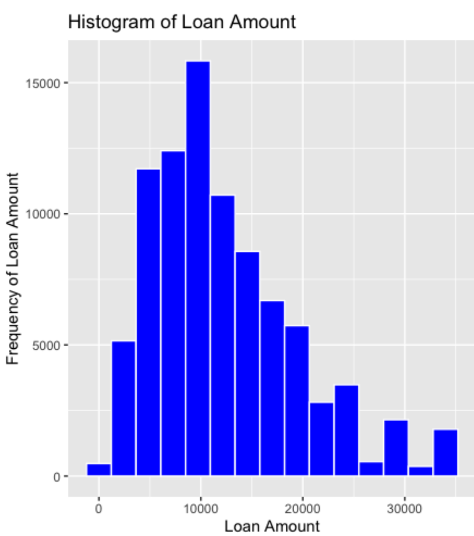
Interpretation: There is no discernable trend that would allow us to connect any independent variables to the know loan_default(dependent variable).

```
Observations: 1,000
Variables: 12
$ loan_default              <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
$ loan_amnt                 <int> 7200, 10000, 6000, 15000, 10000, 12000, 12000, 17800, 5600, 10650, 10000, 9950, 14800, 10575, …
$ adjusted_annual_inc       <dbl> 24272, 34812, 123928, 48340, 24560, 39100, 55668, 76784, 31456, 54560, 35472, 18588, 31344, 29…
$ pct_loan_income           <dbl> 0.17560976, 0.23809524, 0.04225352, 0.27272727, 0.28571429, 0.23076923, 0.19047619, 0.20342857…
$ dti                       <dbl> 20.79, 19.63, 32.23, 25.70, 20.81, 33.23, 30.21, 17.55, 7.77, 11.08, 10.82, 22.28, 16.10, 17.2…
$ residence_property        <chr> "Own", "Own", "Rent", "Own", "Own", "Rent", "Own", "Own", "Own", "Own", "Rent", "Own", "Rent",…
$ months_since_first_credit <dbl> 268, 155, 211, 100, 178, 207, 231, 80, 160, 430, 163, 124, 81, 74, 108, 144, 205, 118, 266, 87…
$ inq_last_6mths            <int> 1, 2, 1, 0, 0, 1, 0, 1, 0, 2, 1, 1, 2, 0, 0, 0, 0, 2, 0, 0, 0, 1, 1, 1, 1, 0, 0, 1, 0, 0, 1, 0…
$ open_acc                  <int> 6, 15, 11, 12, 6, 15, 11, 10, 6, 9, 15, 9, 9, 12, 6, 10, 16, 9, 10, 9, 22, 10, 16, 11, 8, 13, …
$ bc_util                   <dbl> 102.5, 61.1, 93.5, 57.1, 82.0, 89.6, 89.6, 75.5, 84.7, 52.8, 27.5, 62.5, 100.3, 96.3, 99.5, 97…
$ num_accts_ever_120_pd     <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 2, 0, 0, 1, 0, 0, 0, 0…
$ pub_rec_bankruptcies      <int> 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1…
```

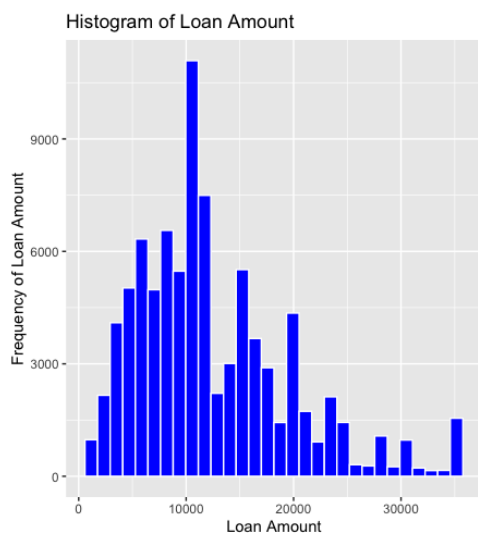
The chart below produce summary statistics of the dataset:

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew
loan_default	1	88451	0.13	0.33	0.00	0.03	0.00	0	1.00	1.00	2.26
loan_amnt	2	88451	12435.17	7186.94	10500.00	11655.23	6671.70	1000	35000.00	34000.00	1.01
adjusted_annual_inc	3	88451	57013.98	55124.23	47028.00	50666.37	28934.42	-14540	7135346.00	7149886.00	41.71
pct_loan_income	4	88451	0.20	0.10	0.19	0.20	0.11	0	0.45	0.45	0.43
dti	5	88451	16.90	7.71	16.49	16.74	8.47	0	34.99	34.99	0.16
residence_property*	6	88451	NaN	NA	NA	NaN	NA	Inf	-Inf	-Inf	NA
months_since_first_credit	7	88451	183.32	85.19	166.00	174.73	71.16	36	750.00	714.00	1.12
inq_last_6mths	8	88451	0.78	1.02	0.00	0.60	0.00	0	7.00	7.00	1.46
open_acc	9	88451	10.87	4.55	10.00	10.48	4.45	1	62.00	61.00	0.98
bc_util	10	88451	66.71	26.22	72.10	69.29	28.02	0	173.20	173.20	-0.70
num_accts_ever_120_pd	11	88451	0.32	0.94	0.00	0.10	0.00	0	29.00	29.00	5.38
pub_rec_bankruptcies	12	88451	0.09	0.30	0.00	0.00	0.00	0	7.00	7.00	3.55

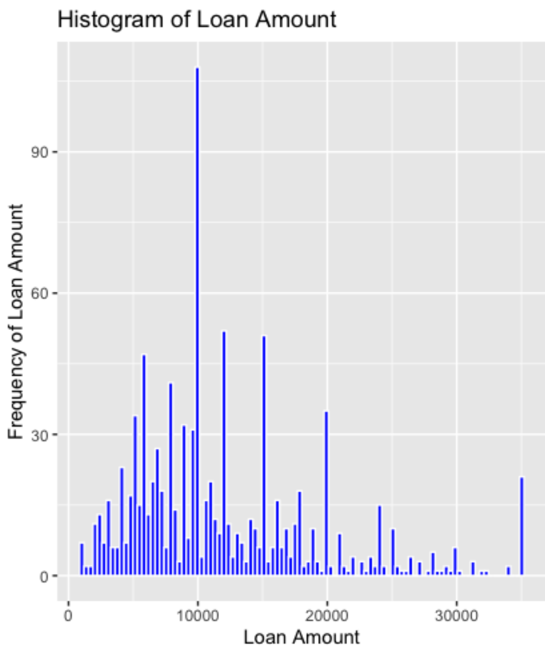
	kurtosis	se
loan_default	3.12	0.00
loan_amnt	0.81	24.17
adjusted_annual_inc	4720.98	185.35
pct_loan_income	-0.48	0.00
dti	-0.70	0.03
residence_property*	NA	NA
months_since_first_credit	1.77	0.29
inq_last_6mths	2.26	0.00
open_acc	1.89	0.02
bc_util	-0.39	0.09
num_accts_ever_120_pd	53.29	0.00
pub_rec_bankruptcies	19.38	0.00



Bin = 15



Bin = 30



This is with Bin 100 because our observation is over 1000
mean(loan): 12435.17
median(loan):10500

2. Please view my A3.R file for all the different models built in R
3. Multiple regression and Logistic regression and Naïve Bayes model compared below

##
Multiple regression: Multiple regression models predict a variable's value based on the value of two or more other variables. In our case, we have one dependent variable, loan default, and all other independent variables in the data set. With the multiple regression, there are many assumptions we would have about our dataset.

Some assumptions about the independent and dependent variables include whether the problem's dataset is the right choice. Such as the need to be a linear relationship between the dependent variable(loan default) and each of your independent variables is a linear relationship between the dependent variable (loan default and each of our independent variables.

F-statistic: with a value of 189 on 11 and 88439 DF, p-value: < 2.2e-16, indicating that there is strong evidence that at least one of our predictor variables is related to the response

Once we find that at least one of our predictors is related to our response variable, we can look at our R^2 value, 0.02296 for this model, and RSE, 0.3273 on 88439 degrees of freedom in this case.

```
Call:
lm(formula = loan_default ~ ., data = LendingClub)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.3497 -0.1502 -0.1128 -0.0695  1.0561 

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)    
(Intercept)                1.816e-03  5.737e-03   0.317  0.75157    
loan_amnt                 -3.516e-06  2.459e-07 -14.301  < 2e-16 ***
adjusted_annual_inc        2.031e-08  2.683e-08   0.757  0.44894    
pct_loan_income            3.481e-01  1.713e-02  20.318  < 2e-16 ***
dti                        1.520e-03  1.655e-04   9.188  < 2e-16 ***
residence_propertyRent     3.222e-02  2.333e-03  13.812  < 2e-16 ***
months_since_first_credit -7.912e-05  1.370e-05  -5.775 7.71e-09 ***
inq_last_6mths             2.500e-02  1.104e-03  22.646  < 2e-16 ***
open_acc                   8.301e-04  2.731e-04   3.039  0.00237 ** 
bc_util                    6.203e-04  4.382e-05  14.157  < 2e-16 ***
num_accts_ever_120_pd      3.224e-03  1.191e-03   2.707  0.00679 ** 
pub_rec_bankruptcies       8.681e-04  3.702e-03   0.235  0.81458    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3273 on 88439 degrees of freedom
Multiple R-squared:  0.02296,	Adjusted R-squared:  0.02284 
F-statistic:   189 on 11 and 88439 DF,  p-value: < 2.2e-16
```

Based on P value test these X variable have something in common: loan_amnt, pct_loan_income, dti, residence_propertyRent, inq_last_6mths, bc_util.

Now lets look at our reduced Model with just those datasets as our X for our Y.

```
Analysis of Variance Table

Model 1: loan_default ~ loan_amnt + pct_loan_income + dti + residence_property + 
    inq_last_6mths + bc_util
Model 2: loan_default ~ loan_amnt + adjusted_annual_inc + pct_loan_income + 
    dti + residence_property + months_since_first_credit + inq_last_6mths + 
    open_acc + bc_util + num_accts_ever_120_pd + pub_rec_bankruptcies
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)    
1  88444 9478.8                                  
2  88439 9473.9  5    4.8897 9.1291 1.078e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
```

We obtained an F statistic value of 9.1291 and a very small *p*-value. With these results, we conclude that there is strong evidence that at least one of β_2 or β_3 is related to the response, *loan_default*.

##
Logistic Regression

Model 1:

LendingClub = rbind(sample_n(filter(LendingClub, loan_default==1), 1000), sample_n(filter(LendingClub, loan_default==0), 1000))

Each of the coefficient estimates below can be interpreted in the following way: a negative value represents a decrease in the odds of the event Loan_default = 1 occurring. The larger the negative value, the larger the decrease. A positive value represents an increase in the odds of the event Loan_default = 1 occurring. The larger the value, the larger the increase in odds.

Based on the information below

```
Call:
glm(formula = loan_default ~ ., family = "binomial", data = LendTrain)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7133  -1.1417   0.7035   1.1306   1.8563

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)               -1.124e+00  3.130e-01  -3.592 0.000328 ***
loan_amnt                 -9.430e-06  1.663e-05  -0.567 0.570667
adjusted_annual_inc       -2.044e-06  2.666e-06  -0.767 0.443263
pct_loan_income            1.576e+00  1.103e+00   1.428 0.153272
dti                        8.955e-03  8.451e-03   1.060 0.289339
residence_propertyRent     3.344e-01  1.161e-01   2.881 0.003963 **
months_since_first_credit -1.930e-03  6.820e-04  -2.829 0.004665 **
inq_last_6mths             2.583e-01  5.395e-02   4.788 1.69e-06 ***
open_acc                   3.194e-02  1.344e-02   2.377 0.017477 *
bc_util                    6.870e-03  2.154e-03   3.190 0.001421 **
num_accts_ever_120_pd      1.610e-02  6.277e-02   0.256 0.797630
pub_rec_bankruptcies       2.437e-01  1.967e-01   1.239 0.215231
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1940.8  on 1399  degrees of freedom
Residual deviance: 1863.4  on 1388  degrees of freedom
AIC: 1887.4

Number of Fisher Scoring iterations: 4
```

Confusion Matric on the logistic regression

Model 1:

LendingClub = rbind(sample_n(filter(LendingClub, loan_default==1), 1000), sample_n(filter(LendingClub, loan_default==0), 1000))

```
Confusion Matrix and Statistics


     0   1
0   42  22
1  100 136

               Accuracy : 0.5933
                 95% CI : (0.5354, 0.6494)
    No Information Rate : 0.5267
    P-Value [Acc > NIR] : 0.01186
```

Sample Size = 1000 , Probability .4
Summary: Higher sample size resulted in more accuracy of my model

Model 2:

LendingClub = rbind(sample_n(filter(LendingClub, loan_default==1), 500), sample_n(filter(LendingClub, loan_default==0), 500))

```
     0   1
0   35  27
1  112 126

               Accuracy : 0.5367
                 95% CI : (0.4784, 0.5942)
    No Information Rate : 0.51
    P-Value [Acc > NIR] : 0.1932
```

Sample Size = 500 , Probability .4
Summary: As my sample size decreased so did my accuracy.

##
 # Naïve Bayes:
Naïve Bayes assumes that features of measurement are independent of each other. By merely taking each feature separately and determine the proportion of previous measurements that belong to class A or class B that have the same value for this feature only. It is good at handling missing values by ignoring the instance during probability estimate calculations. Naive Bayes algorithm calculates the probability of a variable is a

value given some other variables. The Naive Bays model is used to calculate the probability of loan_default given all other variables in the dataset.

```
> NVmodel$tables
$loan_amnt
   loan_amnt
Y       [,1]     [,2]
  0 11995.75 7069.925
  1 12273.64 7206.003

$adjusted_annual_inc
   adjusted_annual_inc
Y       [,1]     [,2]
  0 55107.62 46231.93
  1 46064.65 30363.96

$pct_loan_income
   pct_loan_income
Y        [,1]      [,2]
  0 0.2025006 0.1050182
  1 0.2228172 0.1004296

$dti
   dti
Y       [,1]     [,2]
  0 17.37455 7.695384
  1 18.32337 7.449519

$residence_property
   residence_property
Y         Own      Rent
  0 0.5209424 0.4790576
  1 0.5054348 0.4945652

$months_since_first_credit
   months_since_first_credit
Y       [,1]     [,2]
  0 182.3377 83.17246
  1 170.1467 80.23963

$inq_last_6mths
   inq_last_6mths
Y        [,1]      [,2]
  0 0.7225131 0.9286719
  1 0.9592391 1.1077982
```

```
$open_acc
   open_acc
Y       [,1]     [,2]
  0 10.64921 4.607706
  1 11.37228 4.792705

$bc_util
   bc_util
Y       [,1]     [,2]
  0 63.63822 27.89734
  1 69.61957 25.46661

$num_accts_ever_120_pd
   num_accts_ever_120_pd
Y        [,1]      [,2]
  0 0.3298429 0.9887023
  1 0.2092391 0.5595069

$pub_rec_bankruptcies
   pub_rec_bankruptcies
Y        [,1]      [,2]
  0 0.1099476 0.3132347
  1 0.1141304 0.3184025
```

###

Comparing all 3 models: All three models depict different things about the dataset. Based on our desire to have Loan_default as the dependent variable (Y), we were trying to predict that Y is based on our independent variable X.

The Naive Bayes algorithm is the most accurate of the three models because it is the most forgiving algorithm than the other two. It was successful in predicting the correct loan_default value majority of the time based on the confusion matrix.