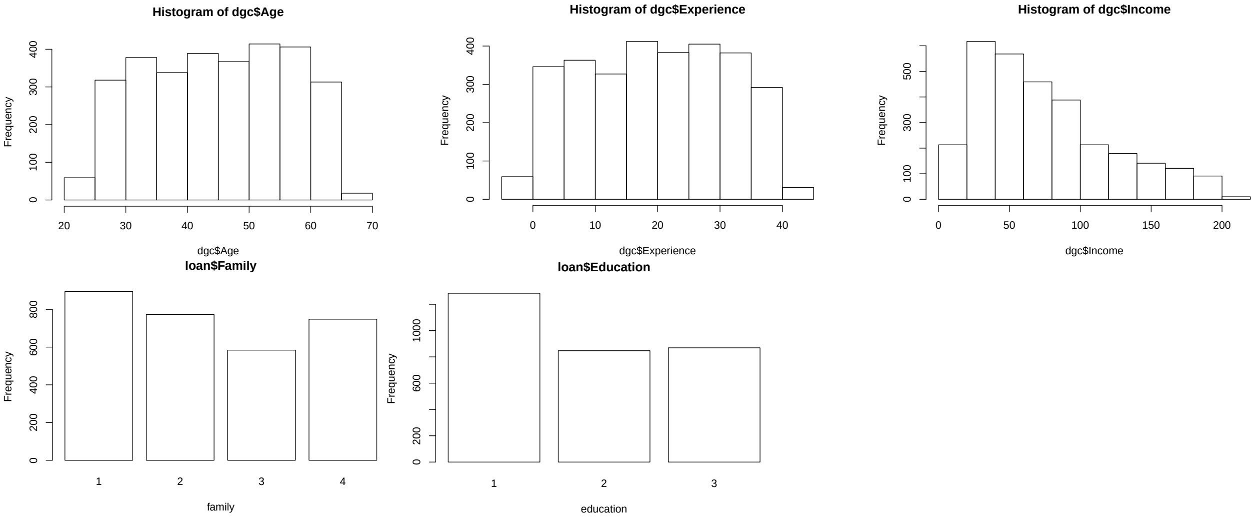
Problem2

(a)Explore the demographic information of the customers.

After analysis, we regard **Age, Experience, Income, Family** and **Education** as demographic variables. Below are our results.



As we can see from the histograms and bar plots, the distributions of the age and experience of the bank's customers are smooth, varying mainly from 25-65 years and 0-40 years respectively. And the proportion of each part is similar. It shows that the bank's customers from all backgrounds of age and working experience and there is no specific feature.

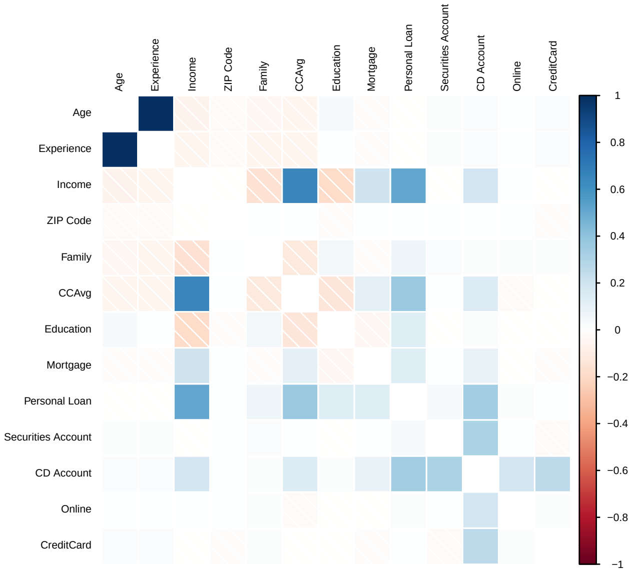


Figure 1: Correalation Among Variables

As for the income, it is obvious that the main customer group is people with annual income level from 20-100 thousand. And there is also a decreasing trend, which means with income increasing, people prefer not to open accounts in this bank. Maybe it has something to do with the bank's positioning.
In term of family size, the largest percentage of customers is people with family size of 1, followed by family size of 4. There is no special situation observed.
About education, most of its customers are with bachelor's degree, about 1/3 higher than customers with master's degree and advanced/professional degree. The last two kinds have nearly the same amount.
From the correlation plot of the five variables, it is obvious that age and experience have a strong positive correlation, which is the same with our expectation.
However, it seems that family and education both have a slight negative correlation with income. Maybe people with higher income level have more advanced life insights, so some of them prefer to have fewer children. Another possible reason is that they

are busy doing their career, thus have less time to take good care of families.
As for education and income, it is possible that many people may think earning money is more important than obtaining a higher degree. So they prefer to earn money first and then they don't have enough time to get a higher degree. In addition, some people may choose to learn by themselves in order to rich themselves. As a consequence, even though they don't have a high degree, their income level is satisfied.

(b)Analyze what variables possibly make a customer more likely to accept a personal loan?

In this question, we decide to analyze it in both qualitative and quantitative way.

(i)Qualitative Analysis:
Regarding the possibility of customers to accept personal loan, the graph clearly demonstrates the correlation between personal loan and other variables. It is apparent that personal loan is positively correlated to 7 variables, which are:

- Income
- Family
- CCAvg
- Education
- Mortgage
- Securities Account
- CD account

For each variable correlated to "Personal Loan", there could be some possible explanation for it. To demonstrate it more clearly, possible explanation for every variable is concluded in the table below.

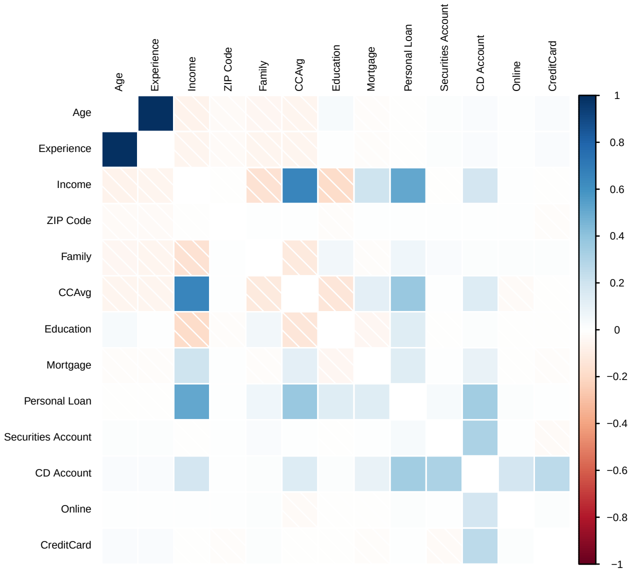


Figure 2: Correalation Among Variables

Table 1: Summary of Personal Loan's Correlated Variables

Variable	Coefficience	Possible Explanation
Income	0.517	(1)People who earn higher salaries will have more funding needs for investment. (2)People use personal loan for investment and consequently gain a higher income
CCAvg	0.376	People with a higher average spending in credit card probably has a stronger sense in ”pre-expenditure”. This might make them: (1)run out of money so they have to use personal loan for payback. (2)get loan for future.
CD Account	0.347	People with enough deposit are willing to accept personal loan because they at least can guarantee they have money to pay the loan partially when they CD account is due.
Education	0.137	Education: (1) People with higher degree may have more advanced consumption concept, thus to accept personal loan and spend in advance.(2) These people may have the demand for some research equipment which is very costly.
Mortgage	0.136	The reason seems obvious. People who need to pay for the mortgage are more likely to accept loan so that they can have enough money to maintain their daily life.
Family	0.063	It is quite possible that people with a larger family size have stronger motivation to use personal loan. Generally, they have heavier life burden. They need to borrow money from the bank to support their family, and then work hard to pay off it.
Securities Account	0.033	As we know, buying securities is a kind of investment. So, these people need to have enough current fund to deal with some big variance of their investments. And when they believe the market is good they want to borrow money to increasing their investment amount.

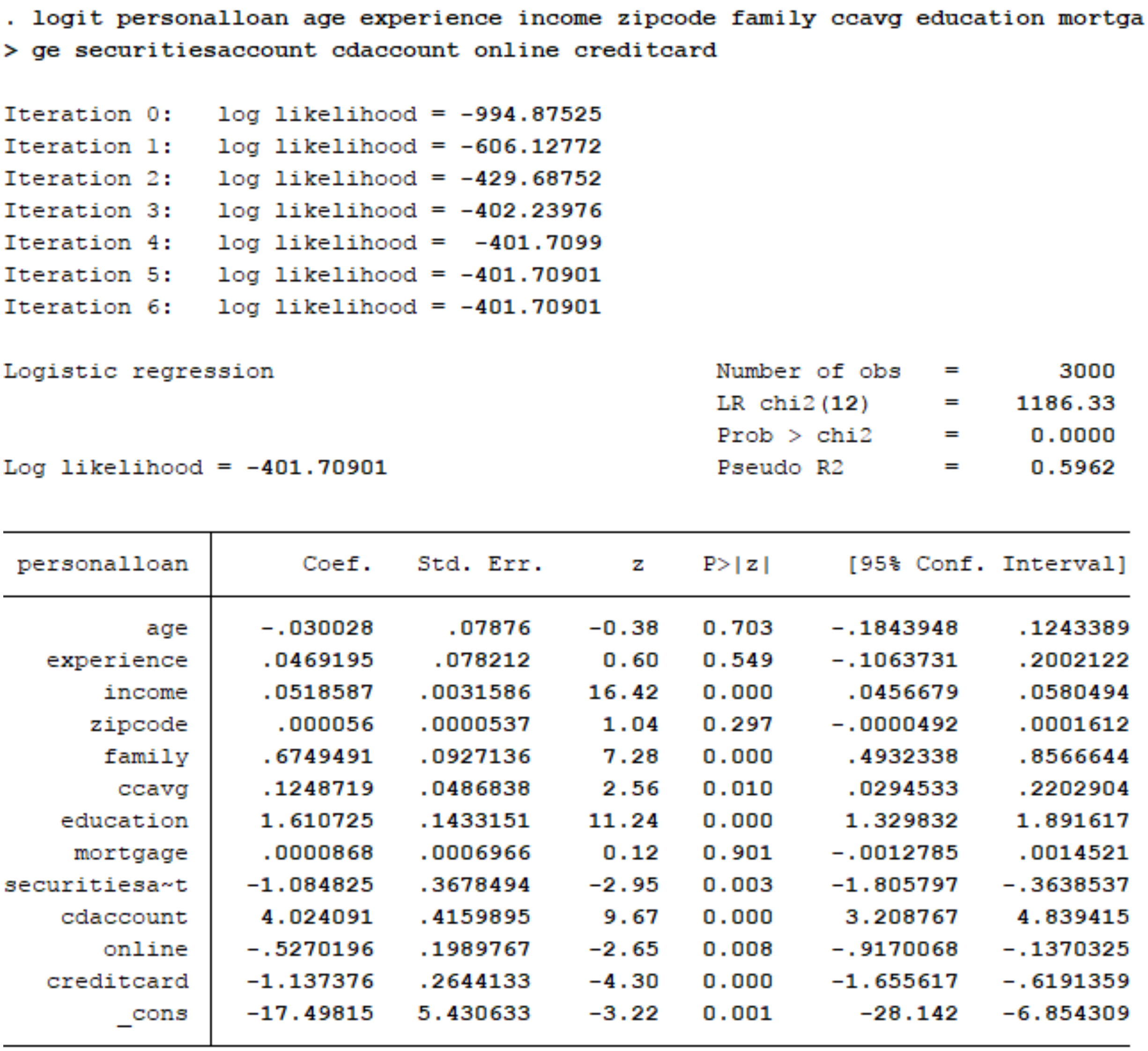
But we need to pay attention to the fact that the correlation can only show the linear relation between the variables. Our logistic model can reveal a more reliable relation between personal loan and other variable, not just focus on linear relation. Therefore we used **quantitative model** for further discovery.

(ii)Quantitative Analysis

For quantitative analysis, Logistic regression model is applied. The reason why we choose the Logistic model is that:

- The dependent variable is binary variable (Y=1 or Y=0)
- The linear probability cannot better measure since the probability should be between 0 and 1. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

Regression1:
All variables in the model, model is shown in the screen shot from STAT.

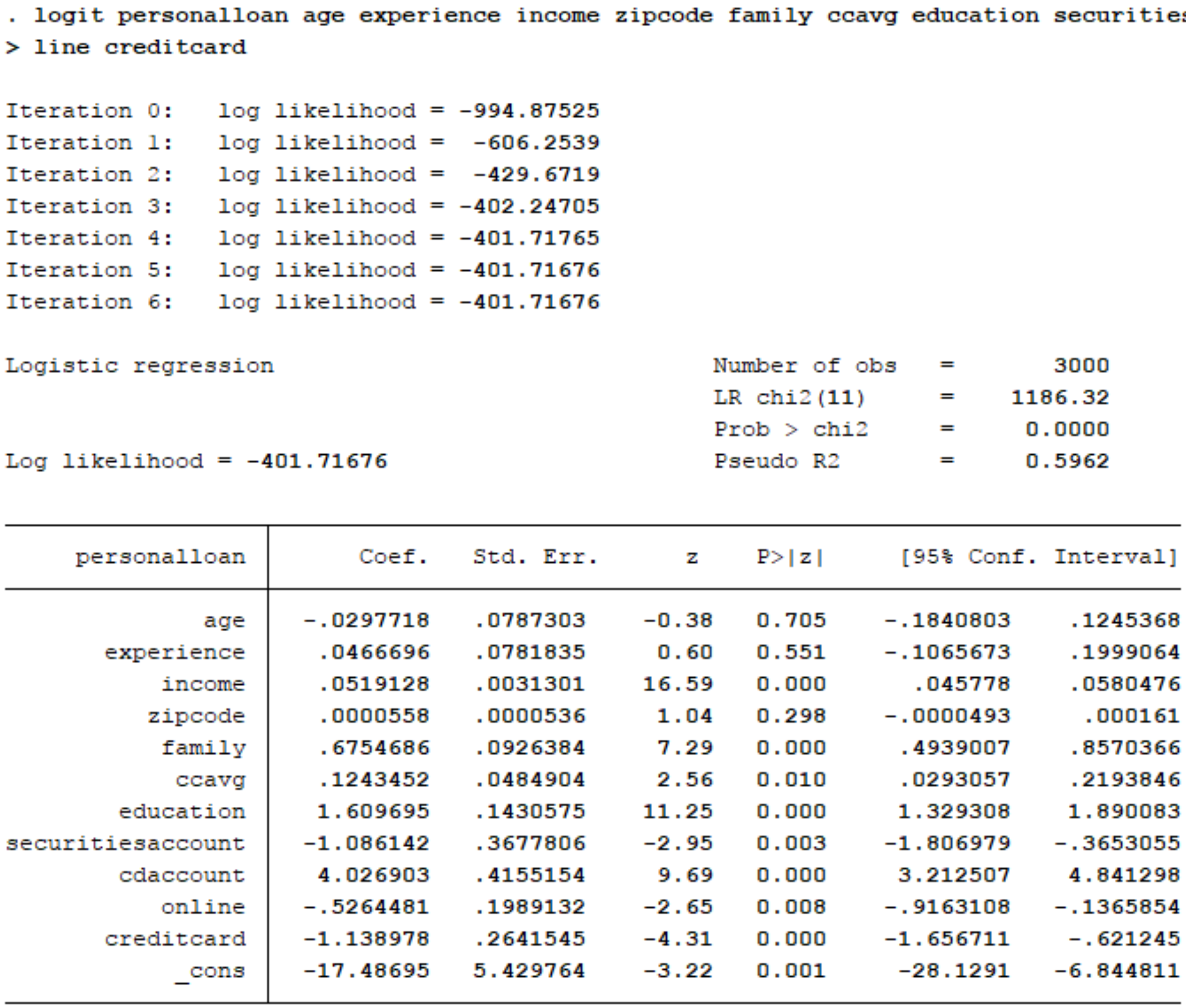


```
. logit personalloan age experience income zipcode family ccavg education mortga
> ge securitiesaccount cdaccount online creditcard

Iteration 0:   log likelihood = -994.87525
Iteration 1:   log likelihood = -606.12772
Iteration 2:   log likelihood = -429.68752
Iteration 3:   log likelihood = -402.23976
Iteration 4:   log likelihood =  -401.7099
Iteration 5:   log likelihood = -401.70901
Iteration 6:   log likelihood = -401.70901

Logistic regression                             Number of obs   =       3000
                                                LR chi2(12)     =    1186.33
                                                Prob > chi2     =     0.0000
Log likelihood = -401.70901                     Pseudo R2       =     0.5962

------------------------------------------------------------------------------
personalloan |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   -.030028     .07876    -0.38   0.703    -.1843948    .1243389
  experience |   .0469195    .078212     0.60   0.549    -.1063731    .2002122
      income |   .0518587   .0031586    16.42   0.000     .0456679    .0580494
     zipcode |    .000056   .0000537     1.04   0.297    -.0000492    .0001612
      family |   .6749491   .0927136     7.28   0.000     .4932338    .8566644
       ccavg |   .1248719   .0486838     2.56   0.010     .0294533    .2202904
   education |   1.610725   .1433151    11.24   0.000     1.329832    1.891617
    mortgage |   .0000868   .0006966     0.12   0.901    -.0012785    .0014521
securitiesa~t |  -1.084825   .3678494    -2.95   0.003    -1.805797   -.3638537
   cdaccount |   4.024091   .4159895     9.67   0.000     3.208767    4.839415
      online |  -.5270196   .1989767    -2.65   0.008    -.9170068   -.1370325
  creditcard |  -1.137376   .2644133    -4.30   0.000    -1.655617   -.6191359
       _cons |  -17.49815   5.430633    -3.22   0.001      -28.142   -6.854309
------------------------------------------------------------------------------
```

Figure 3: STAT Screen Shot for REG1

We figure out some of the variables are not significant because their p-value is more than 0.05. It may have imperfect multicollinearity, irrelevant variable or omitted variable.

Regression2:
From the correlation plot, the correlation between mortgage and income is positive relationship, and the p-value of mortgage is greater than 0.05. Thus, we decide to remove the mortgage.



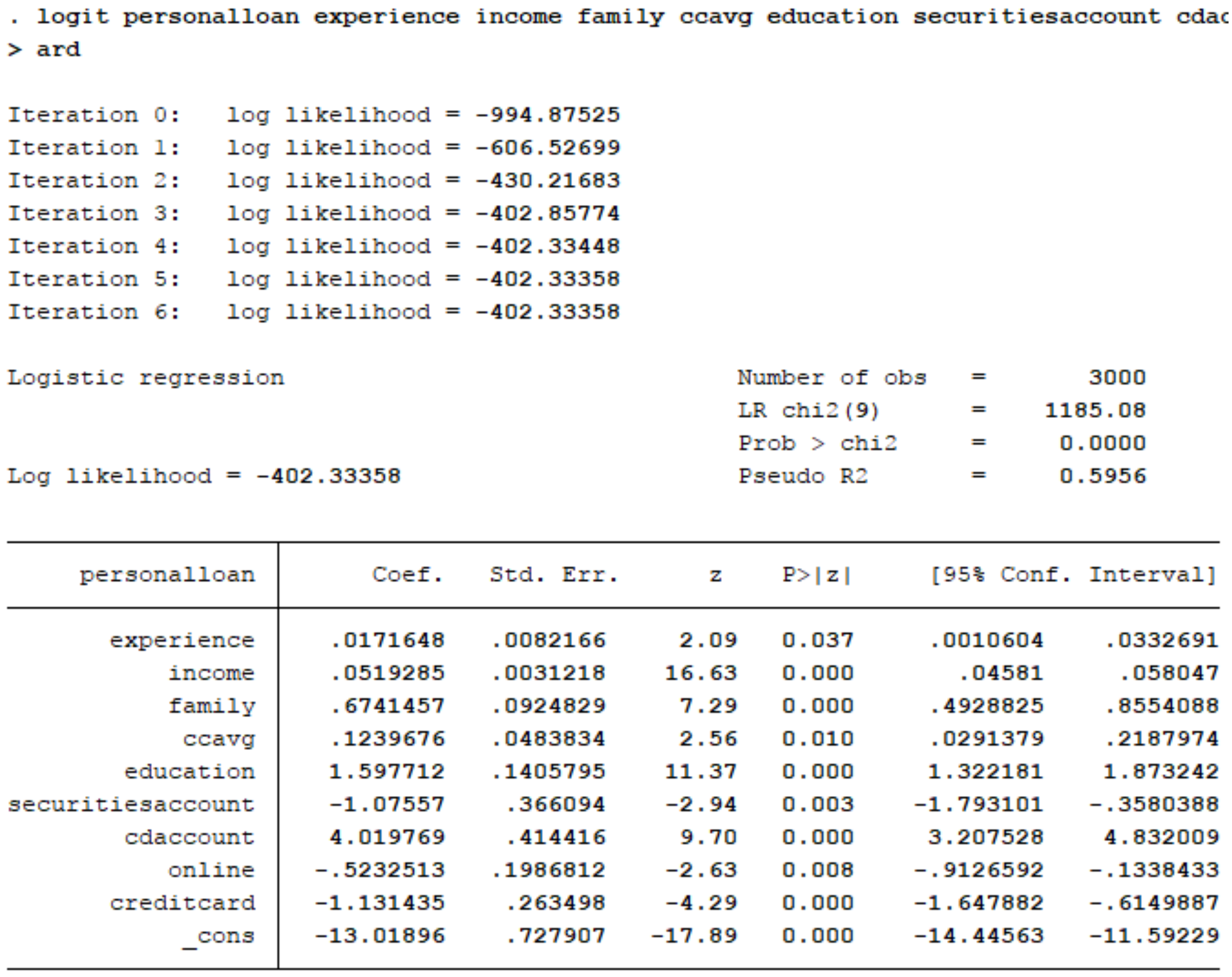
```
. logit personalloan age experience income zipcode family ccavg education securitie
> line creditcard

Iteration 0:   log likelihood = -994.87525
Iteration 1:   log likelihood =  -606.2539
Iteration 2:   log likelihood =  -429.6719
Iteration 3:   log likelihood = -402.24705
Iteration 4:   log likelihood = -401.71765
Iteration 5:   log likelihood = -401.71676
Iteration 6:   log likelihood = -401.71676

Logistic regression                               Number of obs   =       3000
                                                  LR chi2(11)     =    1186.32
                                                  Prob > chi2     =     0.0000
Log likelihood = -401.71676                       Pseudo R2       =     0.5962

------------------------------------------------------------------------------
    personalloan |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----------------+------------------------------------------------------------
             age |  -.0297718   .0787303    -0.38   0.705    -.1840803    .1245368
      experience |   .0466696   .0781835     0.60   0.551    -.1065673    .1999064
          income |   .0519128   .0031301    16.59   0.000      .045778    .0580476
         zipcode |   .0000558   .0000536     1.04   0.298    -.0000493     .000161
          family |   .6754686   .0926384     7.29   0.000     .4939007    .8570366
           ccavg |   .1243452   .0484904     2.56   0.010     .0293057    .2193846
       education |   1.609695   .1430575    11.25   0.000     1.329308    1.890083
securitiesaccount |  -1.086142   .3677806    -2.95   0.003    -1.806979   -.3653055
       cdaccount |   4.026903   .4155154     9.69   0.000     3.212507    4.841298
          online |  -.5264481   .1989132    -2.65   0.008    -.9163108   -.1365854
      creditcard |  -1.138978   .2641545    -4.31   0.000    -1.656711    -.621245
           _cons |  -17.48695   5.429764    -3.22   0.001     -28.1291   -6.844811
------------------------------------------------------------------------------
```

Figure 4: STAT Screen Shot for REG2

Regression3

Age and experience have highly positive relationship. If we put both age and experience in the model, it will affect the regression. Compared these two variables, the work experience is more meaningful to the personal loan. The age doesn't have the clear stable relationship with personal loan since it much depends on the need of different age period. For example, when you are an adult, you may more likely to have the personal loan as the age increases. However, when you get old, you don't want the loan in order to avoid the risk. Thus, we decide remove the age.

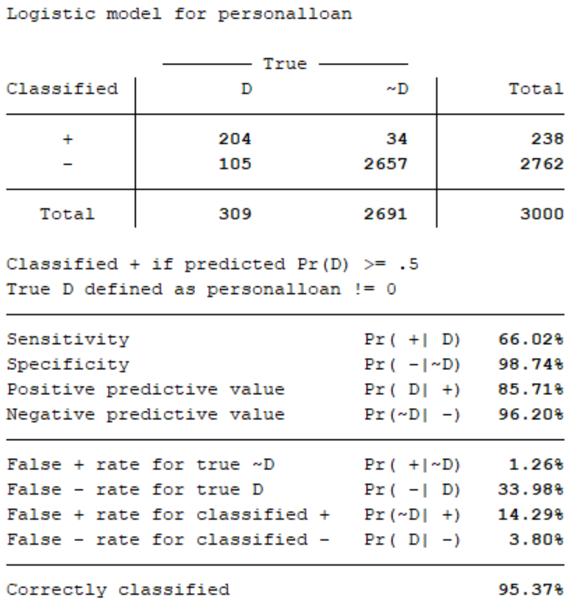
Zip code is not necessary because the coefficient is small, p-value is great than 0.05. Zip code doesn't have the economics meaning to the personal loan.



```
. logit personalloan experience income family ccavg education securitiesaccount cdac
> ard

Iteration 0:   log likelihood = -994.87525
Iteration 1:   log likelihood = -606.52699
Iteration 2:   log likelihood = -430.21683
Iteration 3:   log likelihood = -402.85774
Iteration 4:   log likelihood = -402.33448
Iteration 5:   log likelihood = -402.33358
Iteration 6:   log likelihood = -402.33358

Logistic regression                             Number of obs     =       3000
                                                LR chi2(9)        =    1185.08
                                                Prob > chi2       =     0.0000
Log likelihood = -402.33358                     Pseudo R2         =     0.5956

------------------------------------------------------------------------------------
     personalloan |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------------+-----------------------------------------------------------------
       experience |   .0171648   .0082166     2.09   0.037     .0010604    .0332691
           income |   .0519285   .0031218    16.63   0.000       .04581     .058047
           family |   .6741457   .0924829     7.29   0.000     .4928825    .8554088
            ccavg |   .1239676   .0483834     2.56   0.010     .0291379    .2187974
        education |   1.597712   .1405795    11.37   0.000     1.322181    1.873242
securitiesaccount |   -1.07557    .366094    -2.94   0.003    -1.793101   -.3580388
        cdaccount |   4.019769    .414416     9.70   0.000     3.207528    4.832009
           online |  -.5232513   .1986812    -2.63   0.008    -.9126592   -.1338433
       creditcard |  -1.131435    .263498    -4.29   0.000    -1.647882   -.6149887
            _cons |  -13.01896    .727907   -17.89   0.000    -14.44563   -11.59229
------------------------------------------------------------------------------------
```

Figure 5: STAT Screen Shot for REG3

Finally, the pseudo R square is 0.5956, which means all variables explain almost 60% of the variance in personal loan. All the p-value are less than 0.05. They are significant. Experience, income, family, credit card average spending, education and CD account have positive relationship with the personal loan. Securities account, online banking and credit card have negative relationship with the personal loan. CD account is much more significant than the other variables because if you have CD account, you have more ability to pay back the personal loan in the future. What's more, the bank may promote the advertisement of personal loan to the customers who have the CD account.



```
Logistic model for personalloan

              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |       204            34  |        238
     -     |       105          2657  |       2762
-----------+--------------------------+-----------
   Total   |       309          2691  |       3000

Classified + if predicted Pr(D) >= .5
True D defined as personalloan != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   66.02%
Specificity                     Pr( -|~D)   98.74%
Positive predictive value       Pr( D| +)   85.71%
Negative predictive value       Pr(~D| -)   96.20%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)    1.26%
False - rate for true D         Pr( -| D)   33.98%
False + rate for classified +   Pr(~D| +)   14.29%
False - rate for classified -   Pr( D| -)    3.80%
--------------------------------------------------
Correctly classified                        95.37%
```

Figure 6: Model

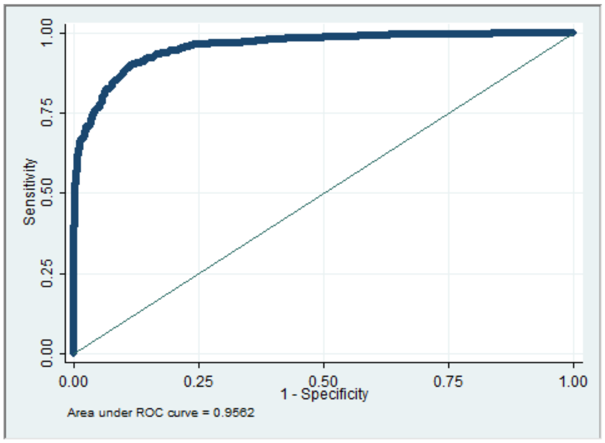


Figure 7: ROC Curve

A receiver operating characteristic curve, (ROC curve), is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied.

The best possible prediction method would yield a point in the upper left corner or coordinate (0,1) of the ROC space, representing 100% sensitivity and 100% specificity. The (0,1) point is also called a perfect classification. A random guess would give a point along a diagonal line from the left bottom to the top right corners.

In this case, our correctly classified is 95.37%. our ROC curve closes to the (0,1) point, and the area under ROC curve is 0.9552.

In conclusion, from our model, **experience, income, family, credit card average spending, education, securities account, cd account, online banking, credit card** possibly makes a customer more likely to accept a personal loan.

(c)There are several products the bank offers like CD and security accounts, online services, credit cards, etc. Can we identify any association among these for finding cross-selling opportunities?

To identify cross-selling opportunities, here we try to use **shopping basket analysis** to discover cross-selling rules for different products. During the analysis, we regard there are 6 products :**Mortgage, Personal Loan, Securities Account, CD Account, Online** and **Credit Card**. Among the variables, mortgage is the only numeric variable. Therefore, we change it as a nominal variable, which represents whether have mortgage.

Before running advanced model, we could use bar plot to check the most frequent bought item

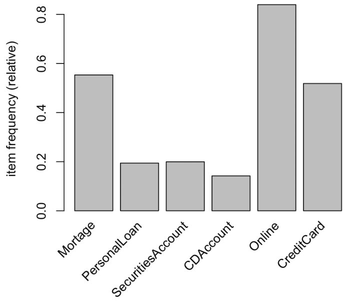


Figure 8: Item Frequency Plot

As shown in the graph, online services has the highest relative frequency, indicating that it is the best selling product among other products. **This means that online services could probably be sold together with other products.**
On the other hand, CDAccount has the lowest relative Frequency, indicating that might be the least popular product compared to other products offered by bank.

We set we set the **Confidence level at 0.1, Support at 0.1**. After running Shopping Basket Analysis Model, we found 20 rules in total. Sorted by **lift**, the top 6 is demonstrated as below:

Table 2: Summary of Association Rules Among Different Products

No.	LHS	RHS	Support	Confidence	Lift
1	Online, CreditCard	CDAccount	0.1005291	0.2504708	1.762622
2	CDAccount, Online	CreditCard	0.1005291	0.7471910	1.441011
3	CDAccount	CreditCard	0.1058201	0.7446809	1.436170
4	CreditCard	CDAccount	0.1058201	0.2040816	1.436170
5	CDAccount, CreditCard	Online	0.1005291	0.9500000	1.131278
6	Online	CDAccount	0.1345427	0.1602160	1.127478

In shopping basket model, lift indicates the possibly that the presence of items in LHS will trigger the occurance of items in RHS. For lift greater than 1, it suggests that the precense of the items on the LHS increase the probability that the items on the right hand side will occur on this transaction.
For instance, the first rule {Online,CreditCard}=>{CDAccount} has a lift near 1.76, meaning that client who use Online Services and Credit Card will tend to apply for a Cash Deposit Account. To demonstrate the association rules more clearly, we can use network graph and parallel coordinates plot.

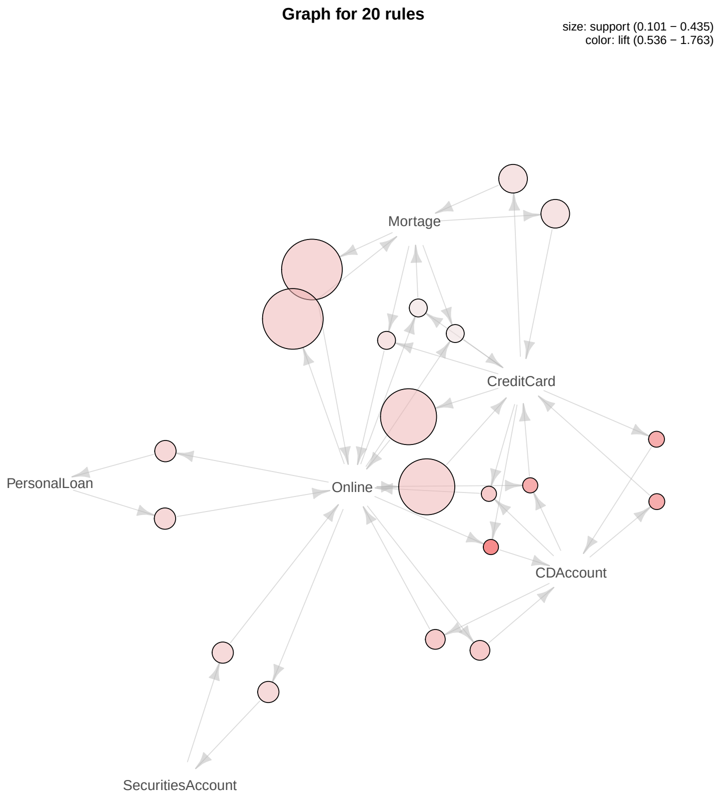


Figure 9: Network Graph

As shown in the graph, the size of the circle represents represents the rules' support(the larger, the greater), while the color of the circle represents rules' lift (the deeper, the greater)
From the network graph, we can see that **Online Services, Mortgage, CreditCard** and **CDAccount** are most frequently related to other products. **This might indicate that these products could be bonded to a wide range of products for cross-selling.**

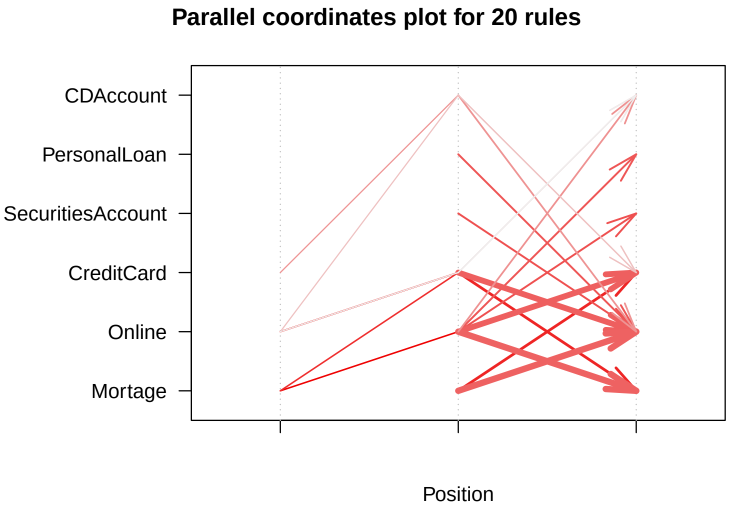


Figure 10: Parallel Coordinate Plot

As shown in the graph, different directed lines represent different rules(from LHS to RHS). Thicker line represents a higher support, while a deeper color represents a higher lift.
From the network graph, we can see that **Online Services, CreditCard** and **Mortgage** are most frequently directed to by other products. And some of the rules lines are and in deep color. **This might indicate that these products could be sold as add-in or promotion when selling other products to customers.**