Econ 265

Professor Anderton

Different Factors Affecting Loan Default Rate and Income

Minh Q. Vu

Clark University

1. **Introduction:**

As a student who is double majoring in Economics and Computer Science and minoring in Data Science, I am fascinated with data analysis and data science. I want to know more about how we can harness the power of data in the economics fields. One of the most popular applications of data in the banking industry in the last couple of decades has been credit risk models and early warning indicators. In the banking industry, all the decisions of approving loans, opening new credit cards, or extending deadlines can be risky without information of the customers. As a result, data from customers will be collected and exchanged by banks to know whether a customer is reliable. Credit risk models were among the first examples of using data in risk management, having been developed and applied widely in the past 70 years. Since the 1960s, when the number of customers started growing rapidly, banks around the world desired a system that helps them make quick or even automate decisions and credit risk models provided exactly that.

1. **Project and Data:**

My project is inspired by a competition held by Home Credit two summers ago on Kaggle. More details of the project/competition can be found here:

<https://www.kaggle.com/c/home-credit-default-risk>

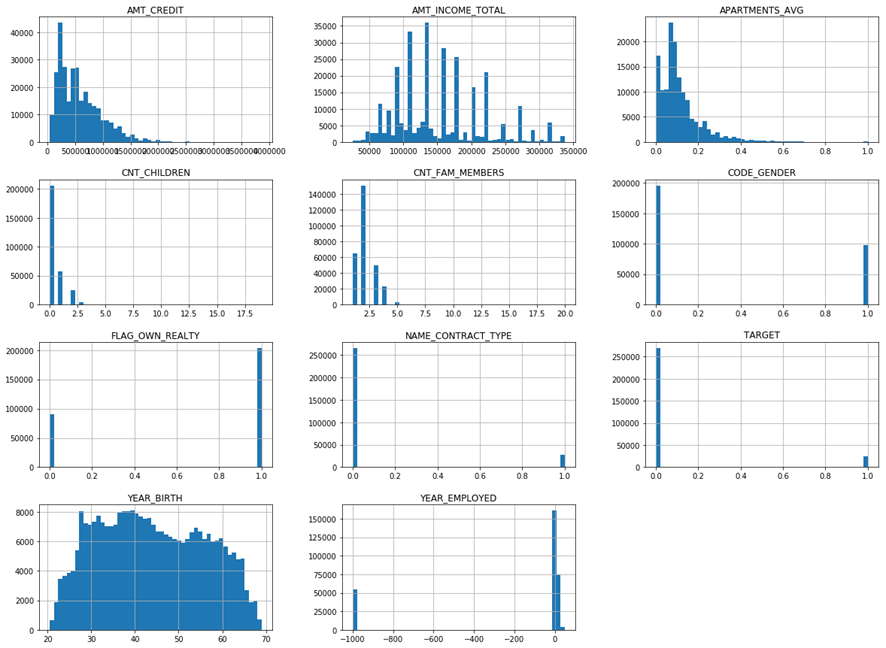
A couple of questions I want to answer:

1. Should we approve this loan to this particular applicant?
2. Which factors affect the repayment ability of a client?
3. Can we predict the income of an applicant with the information he or she provides?

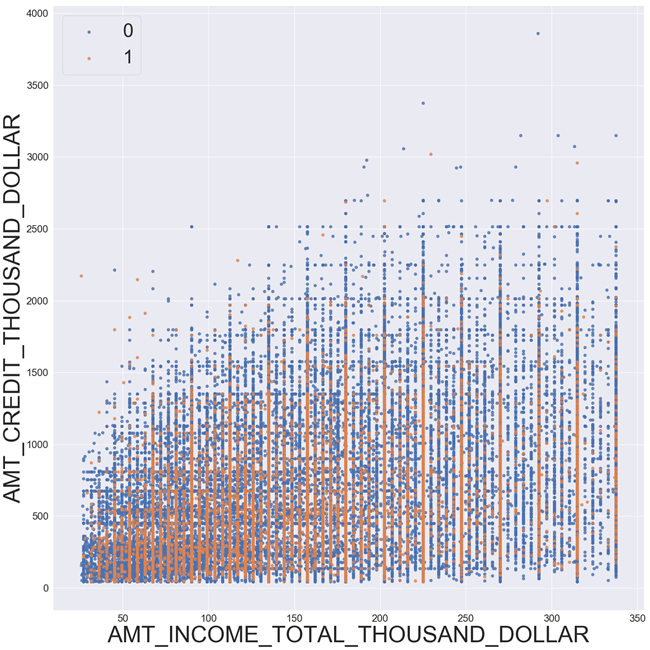
This project will revolve around a dataset that has more than 300,000 observations (each represents one loan application) and 127 features. Our job is to predict a customer’s default likelihood using this useful information, ranging from specific information of the application such as income, gender, number of house/children, etc. to more general information such as details of the building one lives, the forms they submit, etc. However, I will only use 10 to 15 most useful (in my opinion) features. A list of variables and their descriptions can be found on the next page.

|  |  |
| --- | --- |
| **Variable name** | **Description** |
| TARGET | Target variable (1 - client with payment difficulties: 0 - all other cases) |
| AMT\_INCOME\_TOTAL | Income of the client |
| NAME\_CONTRACT\_TYPE | Identification if loan is cash or revolving (0 - cash loan: 1 - revolving loan) |
| CODE\_GENDER | Gender of the client (1 - male: 0 - female) |
| CNT\_CHILDREN | Number of children the client has |
| FLAG\_OWN\_REALTY | Flag if client owns a house or flat (1 - Y: 0 - N) |
| AMT\_CREDIT | Credit amount of the loan |
| APARTMENT\_AVG | Apartment size |
| OCCUPATION\_TYPE | What kind of occupation does the client have |
| CNT\_FAM\_MEMBER | How many family members does client have |
| NAME\_FAMILY\_STATUS | Family status of the client |
| DAYS\_EMPLOYED | How many days before the application the person started current employment |
| DAYS\_BIRTH | Client's age in days at the time of application |

I have modified some features to run a better regression. For DAYS\_EMPLOYED and DAYS\_BIRTH, I divide both by 365.25 to get the age and employment in years. With regard to AMT\_INCOME\_TOTAL, after visualizing it using Python, I observe some outliers. I decided to get rid of all all observations that are above Q3\*1.5. I also generate two new variables logincome and logcredit, which are the natural logarithmic of amt\_income\_total\_in\_thousand\_dollar and amt\_credit\_thousand\_dollar respectively. Here is the histogram of all features after I deal with outliers from income.



The two most important features I will care about are Target and Income. We will run multiple regressions with these two features as dependent variables. Before using Stata, I plot a graph so we can have a sense of how the Target, Income, and Credit relate to each other.



Surprisingly, most applicants with difficulties in repaying their loan apply for a small loan.

I could not find the perfect paper discussing my research questions. Most of the papers concerning this topic focus more on either how to build a credit risk model using different data sources or how loan default rate impacts profitability in financial institutions. In addition, we can educationally guess the impact of aforementioned variables on the default rate and income (for example, it is safe to say that the higher one’s income is, the less likely that person will default his/her loan, or the more experience one has, the higher his/her income will be). As a result, the literature review will focus more on how banks can harness different sources of data to improve risk management. Additionally, we will pay more attention to the magnitude of different coefficients and different hypotheses in this paper.

1. **Literature Review:**

An increase in the number of customers and information has made risk management a more challenging field for financial institutions. In the last decade, banks have identified 13 different types of risks including the main three types (credit risk, operational risk, market risks), of which liquidity and credit risk have arisen to be among the two most complicated and unpredicted risks. Consequently, the traditional data set of columns and rows is proved to be no longer useful enough. Instead, big data is now desired more than ever by banks to minimize losses by mitigating risks and increase revenue by tracking emerging opportunities.

In the research paper of big data, the Economist journal discusses how banks are using, or planning to use, big data in their risk management and how big data can be used to pursue new sources of revenue. So what is Big Data, how is it different from traditional data and why it is so desired by banks or financial institutions in general? Data used by retail banks before was just numbers of pass transactions, which will be reviewed by credit and financial officers to track down patterns or any outliner. Instead, big data is a massive amount of information from a wide range of forms, such as website visits, social media, or mobile devices (the Economist, 2014). Big data promises to provide valuable information about customers at an individual level, which will be far more efficient than traditional data. In the reviewed journey “Applying Big Data To Risk Management” in 2014, the two authors Arnold Veldhoen and Stéphan De Prins agree with the idea of big data being the future of risk management. With the increase in velocity, variety and volume of data, the traditional data warehouse is being replaced by data lakes, which can contain both structured data, such as relational database and spreadsheet, and unstructured data, such as emails and social media (Veldhoen and Prins, 2014). Therefore, big data is invested more by businesses, considering the advantage edge it brings forth.

So how the emergence of big data has provided opportunities for credit application? One problem with the credit industry is that the number of clients and the number of data sources of each of them has increased dramatically in the last decades. Consequently, it is not as accurate as it used to be to predict someone’s default rate using the old credit risk scoring models. In this new era, big data will provide new sources of data, such as social media and marketing databases, which can give decision-makers a fresh insight into their clients at an individual level. For instance, big data contains monthly expensive purchases or gambling problems via credit card history, which can be extra useful to decide if a bank should open for him or her a new bank account (Veldhoen and Prins, 2014). Another problem with the traditional credit scoring model is that data inevitably goes stale. When you apply for a mortgage, you provide banks with the current information of your employment and assets. However, if any change happens and that information is not updated, those data will no longer be accurate. (The Economist, 2014).Bart Baesens also agrees with this point in his 2016 journal “Big Data for Credit Scoring: Opportunities and Challenges”. The traditional credit scoring model uses historical information to score applicants, which can be troublesome when it comes to a customer who lacks borrowing experience. For example, if someone applies to a mortgage for the first time or someone just moved to a new country with no credit history, he or she will automatically be scored as a risky loan with a high default rate. Alternatively, if banks use various data sources, the credit risk assessment can be more accurate. Having agreed with Veldhoen and Prins, Baesens claims that having access to payment history via mobile payment can be beneficial in quantifying his or her creditworthiness (Baesens, 2016).

Big data has shown to also be effective in fraud detection. According to a 2010 study by the Association of Fraud Examiners, more than 16% of the fraud cases happen to the banking and services industry (Tandulwadikar, 2011). The first advantage of big data over traditional data in the fraud detection field is its size. It can be challenging to predict something far from the center of the distribution unless one has a data set of all the outliers. Since fraud happens as infrequently as approximately five out of every 1000 banking transactions, banks need a huge sample of data of transactions to have a moderate sample of fraud cases. Big data, consequently, fits perfectly into this description (The Economist, 2014). Secondly, the traditional anti-fraud approach is still heavily based on manual work of structured data. An example of money laundering was mentioned in Veldhoen and Prins’ article to highlight the advantage of using big data. Without big data analytics, traditional approaches for anti-money laundering involves manual pattern tracking from a vast lake of structured data, which is resource-consuming and potentially high risk. Big data analytics can improve this approach by using algorithms to analyze both structured and unstructured data and track down any suspicious transaction activity (Veldhoen and Prins, 2014).

1. **Regression for Target:**
2. *1st and 2nd regression:*

For the first regression, I will include all the variables in the dataset so we can have a better look at the big picture. With regard to name\_family\_status and occupation\_type, I leave out name\_family\_status\_married and occupation\_type\_accountant as the controlling group. Looking at the report, name\_family\_status\_unknown and name\_family\_status\_widow are omitted due to collinearity. For the sake of simplicity, I will only include name\_family\_status\_civilmarriage in the second regression. The controlling group for name\_family\_status are all the statuses but civil marriage.

Looking at the t-statistic of all the coefficients, we can see that the coefficients of name\_family\_status\_separated, name\_family\_status\_singlenotmarr and a lot of occupation types are not significant. We will do an F-test to see if they are jointly insignificant. If they do, we can leave them out of my final regression. The null hypothesis states that the coefficients of all variables in our test equal to zero. The alternative states that at least one of them is not zero (at least one of them is significant). The F-test gives me an F score of 1.77, which is lower than the critical F value:

F-score = 1.77 < F-critical = 1.96

This means that we cannot reject the null hypothesis. As a result, we can safely remove all of these variables from our equation because they insignificantly affect the Target value.

The second regression is now much easier to interpret. While the controlling group for family status is all statuses except for civil marriage, that for occupation types are accountant, human resource staff, IT staff, manager, medicines staff, private services staff, realty agent, and secretary, which are mostly highly skilled occupations. This regression does not have a high R-squared value because the dependent variable is binary. However, we can still get a lot of insight from this regression. Some variables with a noticeably significant coefficient are income, contract type, gender, apartment average size, age, family status and many more occupation types. Surprisingly, variable logcredit, with a t-statistic of 1.95, is not significant in this regression. With the value of one indicating the applicant has some difficulties in repaying his or her loan and the value of zero indicating all other cases, we can interpret some coefficients as below:

* HOVF, if the income increases by 1%, the Target value is predicted to decrease by 0.000138. This makes economic sense as the higher your income is, the less difficulties you will face in repaying your loan.
* HOVF, an applicant applying for a revolving loan is predicted to have 0.03 less in the Target value than an applicant applying for a cash loan.
* HOVF, a male applicant is predicted to have 0.02 more in the Target value than a female applicant. This means on average, male have more difficulties in repaying their loan, which comes off as quite surprising.
* HOVF, if the age increases by 1 year, the Target value is predicted to decrease by 0.0016. This means the older an applicant gets, the more likely that applicant will repay his or her loan.
* HOVFA, if an applicant is in a civil marriage, he or she is predicted to have 0.02 more in the Target value than an applicant who has a different family status.

The coefficients of different occupation types make a lot of economic sense.

* HOVF, an applicant who works as a cleaning staff is predicted to have 0.033 more in the Target value than an applicant who has one of the jobs in the controlling group.
* HOVF, an applicant who is a low skill laborer is predicted to have 0.081 more in the Target value than an applicant who has one of the jobs in the controlling group.
* HOVF, an applicant who works as a waiter/barmen staff is predicted to have 0.033 more in the Target value than an applicant who has one of the jobs in the controlling group.

\* HOVF: Holding other variables fixed

The p-value of all the coefficients also contain information. All of our coefficients are significant at 90 percent at confidence level. At a 95 percent confidence level, most of the coefficients are still significant. Some exceptions are logcredit, occupation\_type\_corestaff, and occupation\_type\_highskilltechstaff. This means that at 95 percent confidence level, we fail to reject the null hypothesis of these three coefficients.

1. *3rd regression:*

In the third regression, I want to look closer into the return to income and return to the amount of credit loan between male and female using interaction terms. We will generate two new variables:

1. maleincome = code\_gender \* logincome
2. malecredit = code\_gender \* logcredit

The coefficients of all variables are significant except for logcredit, malelogcredit, and year\_employed. This implies that the return to income is not the same for male and female.

* *Target* = 0.191 + 0.15\**code\_gender* - 0.0108\**logincome* - 0.0241\**maleincome* + … + u

If the applicant is male, the regression for Target value will be:

* *Target* = (0.191 + 0.15) - (0.0108 + 0.0241)*logincome* + … + u

If the applicant is female, the regression for Target value will be:

* *Target* = 0.191 - 0.0108\**logincome* + … + u

We will interpret some differences between these two regressions:

* If an applicant is male and all of the dependent variables are zero, the expected mean value of Target will be 0.341.
* If an applicant is female and all of the dependent variables are zero, the expected mean value of Target will be 0.191. .
* If an applicant is male, for every 1% increase in income, the Target value is predicted to decrease by 0.000349.
* If an applicant is female, for every 1% increase in income, the Target value is predicted to decrease by 0.000108.

Since the coefficient of malelogcredit is not significant, the slope of logcredit is the same for both male and female. We can also interpret some coefficients as below:

* HOVF, an applicant who owns a realty is predicted to have 0.004 more in the Target value than an applicant who does not own a realty. This is saying that people with realty have more trouble repaying their loan, which does not make much economic sense. However, since this coefficient only has a t-statistic of 3.23, the relationship is not strongly indicated.
* HOVF, if the apartment size increases by 1 unit, the Target value is predicted to decrease by 0.064. This makes economic sense as the larger your apartment is, the less difficulties you will face in repaying your loan.
* HOVF, if the age of an applicant increases by 1 year, the Target value is predicted to decrease by 0.00158. This is indicating the same relationship as the second regression: the older an applicant is, the less likely he or she is to default his or her loan. The absolute value of the t-statistic of this variable is also high (21.24) just like in the second regression.

\* HOVF: Holding other variables fixed

1. **Regression for Income:**
2. *1st regression:*

For the first regression, I will do the same thing as when I run regression for Target and include all the variables (with an exception of Target variable) in the dataset. Just like above, I will leave out occupation\_type\_accountant as the controlling group for occupation\_type and I will only include name\_family\_status\_civil\_marriage in the regression. Based on the t-statistic of all the coefficients, we can see that the coefficients of year\_birth and occupation\_type\_itstaff are not significant. That is only two variables out of 27 variables included in the regression, so for the sake of simplicity, I will not do an F-test and keep them in the regression.

Looking at the scatter plot of the relationship between income and credit produced by Python above, we can see that there is a high chance there exists heteroskedasticity in the regression of logincome. I will test for heteroskedasticity using the Breusch-Pagan test in Stata. The test gives the probability of 0.00. This indicates that on 100 percent confidence level, the regression has heteroskedasticity. We can double check this using the alternative White test. After running a regression with uhatsq as the dependent variable and yhat and yhatsq as independent variables, we do an F-test for the two independent variables. The test generates an F-score of 1233.48, which is much higher than the critical F-value. As a result, we can safely reject the null hypothesis (there is no heteroskedasticity) and confirm the existence of heteroskedasticity. In order to deal with heteroskedasticity, I run the robust regression for the first regression.

After including all variables as dependent variables (with an exception of the Target variable), deciding not to do an F-test with the two insignificant coefficients, checking for heteroskedasticity and deal with it using robust regression, we can now take a closer look at the robust regression. Among all variables, only year\_birth and occupation\_type\_itstaff are not significant based on their t-statistics. All of the variables are also different from 0 on the 95 percent level, with the exception of the two variables mentioned above. We can interpret some of the coefficients as below:

* HOVF, if the credit increases by 1%, the income is predicted to increase by 0.224%. This makes economic sense as the bigger loan you are applying for, the higher your income should be. This interpretation agrees with the relationship plotted using Python above.
* HOVF, an applicant applying for a revolving loan is predicted to have a 3.71% higher income than an applicant applying for a cash loan.
* HOVF, a male applicant is predicted to have 16.9% higher income than a female applicant. This also makes economic sense because men tend to earn more than women, holding all factors fixed.
* An interesting discovery is cnt\_fam\_member (number of family members). HOVF, if the number of family members increases by 1, the income is predicted to decrease by 5%.
* HOVF, if the number of years employed increases by 1, the income is predicted to increase by 0.02032%.

Many occupation\_type variables also make a lot of economic sense. The omitted occupation is account.

* HOVF, a cleaning staff is predicted to have a 19.8% lower income than an accountant.
* HOVF, a manager is predicted to have a 14.6% higher income than an accountant.
* HOVF, a realty agent is predicted to have a 13.08% higher income than an accountant.
* HOVF, a security staff is predicted to have a 17.03% lower income than an accountant.

\* HOVF: Holding other variables fixed

Overall, although the R squared of our regression is only around 0.25, the sign and magnitude of most of the coefficients make a lot of economic sense.

1. *2nd regression:*

In the second regression, we gets the following coefficients:

amt\_income\_total\_in\_thousand\_dollars = 157.50 + 0.33\*year\_employed + 0.0003\*year\_employed\_sq

This indicates that there exists no point where an additional year of employment lowers your income. Thus, the effect of experience will always be positive.

I will examine more closely the effect of experience (years employed) on income in the second regression. The hypothesis is that the relationship between income and experience is polynomial, meaning that there exists a turning point where an additional year of experience actually lowers the income. We can test this hypothesis by first generating a new variable called years\_employed\_sq, which the the square of variable years\_employed, and then running the regression below:

amt\_income\_total\_in\_thousand\_dollars = *beta0* + *beta1*\*years\_employed + *beta2*\*year\_employed\_sq

1. *3rd regression:*

In the third regression, I want to take a closer look into the return to the size of the loan between an applicant who does and an applicant who does not own a realty, using interaction terms. We will generate a variable:

realtycredit = flag\_own\_realty \* logcredit

After getting a new equation:

logincome = beta0 + alpha0\*flag\_own\_realty + beta1\*logcredit + alpha1\*realtycredit + … + u

, I will test for the null hypothesis:

H0 : alpha0 = 0, alpha1 = 0

H1: at least one of them is not 0

If I cannot reject the null hypothesis, it means that the whole income equation is the same for applicants who do and applicants who do not own realty.

The coefficients of all variables are significant except for flag\_own\_realty and realtycredit. This implies that the whole income equation is the same for applicants who do and applicants who do not own realty.

1. **Regression results:**
2. Target regression:

|  |  |  |  |
| --- | --- | --- | --- |
|  | (1) | (2) | (3) |
|  | target | target | target |
| logincome | -0.0135\*\*\* | -0.0138\*\*\* | -0.0108\*\*\* |
|  | (-7.21) | (-7.43) | (-5.04) |
|  |  |  |  |
| logcredit | 0.00212 | 0.00214 | 0.000148 |
|  | (1.93) | (1.95) | (0.11) |
|  |  |  |  |
| name\_contract\_type | -0.0317\*\*\* | -0.0318\*\*\* | -0.0338\*\*\* |
|  | (-13.58) | (-13.63) | (-14.49) |
|  |  |  |  |
| code\_gender | 0.0230\*\*\* | 0.0224\*\*\* | 0.150\*\*\* |
|  | (13.35) | (13.17) | (7.55) |
|  |  |  |  |
| cnt\_children | 0.00661\* | 0.0108\*\*\* | -0.00290\*\* |
|  | (1.98) | (5.26) | (-2.72) |
|  |  |  |  |
| flag\_own\_realty | 0.00480\*\* | 0.00478\*\* | 0.00484\*\* |
|  | (3.21) | (3.20) | (3.23) |
|  |  |  |  |
| apartments\_avg | -0.0607\*\*\* | -0.0606\*\*\* | -0.0645\*\*\* |
|  | (-9.51) | (-9.49) | (-10.09) |
|  |  |  |  |
| cnt\_fam\_members | -0.00803\*\* | -0.0123\*\*\* |  |
|  | (-2.58) | (-7.84) |  |
|  |  |  |  |
| year\_employed | -0.0000135\*\*\* | -0.0000143\*\*\* | -0.00000220 |
|  | (-4.99) | (-5.73) | (-0.97) |
|  |  |  |  |
| year\_birth | -0.00161\*\*\* | -0.00164\*\*\* | -0.00158\*\*\* |
|  | (-20.72) | (-21.65) | (-21.24) |
|  |  |  |  |
| name\_family\_status\_civilmarriage | 0.0192\*\*\* | 0.0192\*\*\* | 0.0164\*\*\* |
|  | (8.20) | (8.22) | (7.22) |
|  |  |  |  |
| name\_family\_status\_separated | 0.00653 |  |  |
|  | (1.71) |  |  |
|  |  |  |  |
| name\_family\_status\_singlenotmarr | 0.00453 |  |  |
|  | (1.31) |  |  |
|  |  |  |  |
| name\_family\_status\_unknown | 0 |  |  |
|  | (.) |  |  |
|  |  |  |  |
| name\_family\_status\_widow | 0 |  |  |
|  | (.) |  |  |
|  |  |  |  |
| occupation\_type\_cleaningstaff | 0.0325\*\*\* | 0.0333\*\*\* |  |
|  | (5.88) | (6.13) |  |
|  |  |  |  |
| occupation\_type\_cookingstaff | 0.0271\*\*\* | 0.0280\*\*\* |  |
|  | (5.06) | (5.32) |  |
|  |  |  |  |
| occupation\_type\_corestaff | -0.00595\* | -0.00491 |  |
|  | (-2.14) | (-1.90) |  |
|  |  |  |  |
| occupation\_type\_drivers | 0.0279\*\*\* | 0.0295\*\*\* |  |
|  | (7.87) | (8.74) |  |
|  |  |  |  |
| occupation\_type\_hrstaff | 0.0242 |  |  |
|  | (1.68) |  |  |
|  |  |  |  |
| occupation\_type\_highskilltechsta | -0.00733\* | -0.00615 |  |
|  | (-2.02) | (-1.76) |  |
|  |  |  |  |
| occupation\_type\_itstaff | -0.0226 |  |  |
|  | (-1.48) |  |  |
|  |  |  |  |
| occupation\_type\_laborers | 0.0230\*\*\* | 0.0243\*\*\* |  |
|  | (9.77) | (11.50) |  |
|  |  |  |  |
| occupation\_type\_lowskilllaborers | 0.0795\*\*\* | 0.0810\*\*\* |  |
|  | (8.07) | (8.26) |  |
|  |  |  |  |
| occupation\_type\_managers | -0.00695\* |  |  |
|  | (-2.20) |  |  |
|  |  |  |  |
| occupation\_type\_medicinestaff | -0.000276 |  |  |
|  | (-0.06) |  |  |
|  |  |  |  |
| occupation\_type\_privateservicest | 0.00476 |  |  |
|  | (0.67) |  |  |
|  |  |  |  |
| occupation\_type\_realtyagents | 0.00995 |  |  |
|  | (0.77) |  |  |
|  |  |  |  |
| occupation\_type\_salesstaff | 0.0209\*\*\* | 0.0219\*\*\* |  |
|  | (7.82) | (8.83) |  |
|  |  |  |  |
| occupation\_type\_secretaries | 0.00942 |  |  |
|  | (0.98) |  |  |
|  |  |  |  |
| occupation\_type\_securitystaff | 0.0346\*\*\* | 0.0361\*\*\* |  |
|  | (6.68) | (7.11) |  |
|  |  |  |  |
| occupation\_type\_waitersbarmensta | 0.0413\*\*\* | 0.0422\*\*\* |  |
|  | (3.89) | (3.99) |  |
|  |  |  |  |
| maleincome |  |  | -0.0241\*\*\* |
|  |  |  | (-5.98) |
|  |  |  |  |
| malecredit |  |  | 0.000297 |
|  |  |  | (0.13) |
|  |  |  |  |
| \_cons | 0.196\*\*\* | 0.206\*\*\* | 0.192\*\*\* |
|  | (16.93) | (20.80) | (17.59) |
| *N* | 142497 | 142497 | 142498 |
| *R*2 | 0.014 | 0.014 | 0.011 |

*t* statistics in parentheses

\* *p* < 0.05, \*\* *p* < 0.01, \*\*\* *p* < 0.001

1. Income regression:

|  |  |  |  |
| --- | --- | --- | --- |
|  | (1) | (2) | (3) |
|  | logincome | amt\_income\_total\_thousand\_dollar | logincome |
| logcredit | 0.224\*\*\* |  | 0.224\*\*\* |
|  | (154.17) |  | (86.51) |
|  |  |  |  |
| name\_contract\_type | 0.0371\*\*\* |  |  |
|  | (11.27) |  |  |
|  |  |  |  |
| code\_gender | 0.169\*\*\* |  | 0.162\*\*\* |
|  | (71.56) |  | (76.19) |
|  |  |  |  |
| cnt\_children | 0.0496\*\*\* |  |  |
|  | (17.04) |  |  |
|  |  |  |  |
| flag\_own\_realty | 0.0410\*\*\* |  | 0.0260 |
|  | (19.18) |  | (1.37) |
|  |  |  |  |
| apartments\_avg | 0.250\*\*\* |  | 0.277\*\*\* |
|  | (27.86) |  | (30.11) |
|  |  |  |  |
| cnt\_fam\_members | -0.0500\*\*\* |  | -0.0156\*\*\* |
|  | (-22.51) |  | (-13.40) |
|  |  |  |  |
| year\_employed | 0.000203\*\*\* | 0.336\*\*\* | 0.000165\*\*\* |
|  | (50.87) | (17.15) | (51.24) |
|  |  |  |  |
| year\_birth | 0.00000263 |  | -0.000653\*\*\* |
|  | (0.03) |  | (-6.18) |
|  |  |  |  |
| name\_family\_status\_civilmarriage | 0.0539\*\*\* |  | 0.0333\*\*\* |
|  | (16.67) |  | (10.11) |
|  |  |  |  |
| occupation\_type\_cleaningstaff | -0.198\*\*\* |  |  |
|  | (-24.00) |  |  |
|  |  |  |  |
| occupation\_type\_cookingstaff | -0.132\*\*\* |  |  |
|  | (-17.42) |  |  |
|  |  |  |  |
| occupation\_type\_corestaff | -0.00775\* |  |  |
|  | (-1.99) |  |  |
|  |  |  |  |
| occupation\_type\_drivers | -0.0240\*\*\* |  |  |
|  | (-5.20) |  |  |
|  |  |  |  |
| occupation\_type\_hrstaff | 0.0562\*\* |  |  |
|  | (3.12) |  |  |
|  |  |  |  |
| occupation\_type\_highskilltechsta | -0.0146\*\* |  |  |
|  | (-2.91) |  |  |
|  |  |  |  |
| occupation\_type\_itstaff | -0.0132 |  |  |
|  | (-0.68) |  |  |
|  |  |  |  |
| occupation\_type\_laborers | -0.0709\*\*\* |  |  |
|  | (-22.02) |  |  |
|  |  |  |  |
| occupation\_type\_lowskilllaborers | -0.236\*\*\* |  |  |
|  | (-17.19) |  |  |
|  |  |  |  |
| occupation\_type\_managers | 0.147\*\*\* |  |  |
|  | (35.84) |  |  |
|  |  |  |  |
| occupation\_type\_medicinestaff | -0.0713\*\*\* |  |  |
|  | (-11.07) |  |  |
|  |  |  |  |
| occupation\_type\_privateservicest | 0.0513\*\*\* |  |  |
|  | (5.40) |  |  |
|  |  |  |  |
| occupation\_type\_realtyagents | 0.131\*\*\* |  |  |
|  | (7.80) |  |  |
|  |  |  |  |
| occupation\_type\_salesstaff | -0.0782\*\*\* |  |  |
|  | (-21.03) |  |  |
|  |  |  |  |
| occupation\_type\_secretaries | -0.0698\*\*\* |  |  |
|  | (-5.05) |  |  |
|  |  |  |  |
| occupation\_type\_securitystaff | -0.170\*\*\* |  |  |
|  | (-22.94) |  |  |
|  |  |  |  |
| occupation\_type\_waitersbarmensta | -0.0761\*\*\* |  |  |
|  | (-5.11) |  |  |
|  |  |  |  |
| year\_employed\_sq |  | 0.000308\*\*\* |  |
|  |  | (15.65) |  |
|  |  |  |  |
| realtycredit |  |  | 0.00267 |
|  |  |  | (0.87) |
|  |  |  |  |
| \_cons | 3.648\*\*\* | 157.5\*\*\* | 3.597\*\*\* |
|  | (355.08) | (865.09) | (216.09) |
| *N* | 142497 | 293476 | 142497 |
| *R*2 | 0.254 | 0.033 | 0.227 |

*t* statistics in parentheses

\* *p* < 0.05, \*\* *p* < 0.01, \*\*\* *p* < 0.001

1. **Conclusion:**

The regressions of Target value give us some valuable information to answer the first two questions:

1. Should we approve this loan to this particular applicant?
2. Which factors affect the repayment ability of a client?

Most of the variables included in this project are significant, with a few exceptions of dummy family status and dummy occupation types variables, in determining whether or not an applicant will default his or her loan. Some crucial factors are gender, age, income and the type of loan. However, these are only basic information of any applicant so banks cannot make an automatic decision based only on these factors. Rather, this information only gives banks a broad view on which group of customers is more likely to default their loan, leading to a better strategy in distributing loans and mortgages.

We can now also estimate the range of one’s income based on the information provided by applicants. All of the variables included are significant in estimating one’s income, with only a surprising exception of age. The size of the loan a customer is applying for is strongly correlated with his or her income, which can help banks more easily group their customers. The coefficients of different occupation type variables also make a lot of economic sense. Although most of the signs of those coefficients are straightforward and can be educationally guessed, the regression shows us the magnitudes of those coefficients, indicating how much more or less a particular job makes.

To conclude, these are all valuable information in building a good credit risk model. I am sorry for not being able to find any paper that is close enough to the topic. Although the literature review was not close to the research questions, I thought I could actually learn more about the industry by reviewing broader papers. I will keep working on this topic as my honor thesis, and hopefully with the equipped knowledge from this final project, I can find some more closely related papers as I have more time.

1. **References:**

Hormozi, Amir, and Giles, Stacy. Data Mining: A Competitive Weapon for Banking and Retail Industries. 2004.

Veldhoen, Arnold, and Stéphan De Prins. *APPLYING BIG DATA TO RISK MANAGEMENT: Transforming Risk Management Practices within the Financial Services Industry*. Dec. 2014.

Baesens, Bart.”Big Data for Credit Scoring: Opportunities and Challenges” *DataMiningApps*. Dec. 2014.

The Economist. “*Retail Banks and Big Data: Big Data as the Key to Better Risk Management*”, 2014.

1. **Appendix:**

\*\* see uploaded STATA output