MA334-Final Project

The study of the relationship between Mortgage application approval and Home Mortgage Disclosure Act (HMDA) data

Created by: Khwanchanok Chumkhun
Registration Number: 2110866
Date: April 21, 2022.

Table of Contents

Introduction

This report investigates the predictors that influence mortgage approval and design the model to predict whether the mortgage application will be denied or approved based on the given Home Mortgage Disclosure Act (HMDA) data set. The study is divided into 3 main parts:

I.) Data set introduction.

The chosen data set and each feature (variable) in this data set will be introduced in this part.

II.) Data analysis methodologies.

The analysis performed in this report will be explained in this section. Data cleaning, exploratory data analysis, then feature selection and modelling and model performance evaluation will be implemented and compared. Moreover, the result given by each analysis will be visualized and discussed.

III.) Conclusion.

Finally, the overall result of the study of mortgage approval prediction by using HMDA data will be briefly discussed in the conclusion section.

All analysis and statistical studies in this report will be conducted using the R language. And the code used in this report is included in the appendices.

Data set introduction

The data set used in this report is a minimized version of Cross-section data on the Home Mortgage Disclosure Act (HMDA) used by Stock and Watson (2007). To reduce the complexity of the study only 2,380 observations on 8 variables are selected to be the sample data of this report. (1) 'deny' variable, which will be used as a response variable, is a nominal binary categorical variable (yes and no) that shows the mortgage approval result whether the mortgage application was denied or approved. The other 7 variables are the potential predictors that may influence mortgage approval. There are 3 continuous variables; (2) 'pirat' represents a ratio of consumer payment to their income, while (3) 'lvrat' explains the ratio between loan to value of the house, and (4) 'unemp' which shows 1989 Massachusetts unemployment rate in applicant's industry. 4 remaining variables are in binary categorical form, (5) 'phist' shows the public poor credit record, it records as 'yes' or 'no', while 'yes' means that consumer has the bad public credit record and 'no' means contradiction. (6) 'insurance' is used to explain if the individual has denied the mortgage insurance or not, (7) 'condomin' describes the property type whether it is a condominium or not, and the last variable is 'single', which defines the marriage status of the individual whether they are single or not.

By using these 8 variables in the chosen data set, the statistical analysis will be proceeded to study which predictor variables are the most influential on the response, which model is the best fit for this type of data and how much accuracy the prediction model can provide.

Data analysis methodologies

1. Data cleaning

The first step is to confirm whether the data set contains any missing values or not. The record with any missing value should be removed or replaced by the mean or median of that column. Then check if the type of each variable is matched with what it should be, if it is not, the data type conversion method will be applied. All categorical variables were assigned as a character type in this selected data set, so we need to convert them into factors. And since the response variable is a binary categorical variable, before starting the analysis, it is required to balance the proportion of each factor level. All procedures are explained in the coding area of the appendices.

2. Exploratory Data Analysis.

2.1 Univariate analysis

For the continuous variables (‘pirat’, ‘lvrat’ and ‘unemp’), first we will analyse the data using a boxplot to make understand the Central Tendency and the distribution of data.

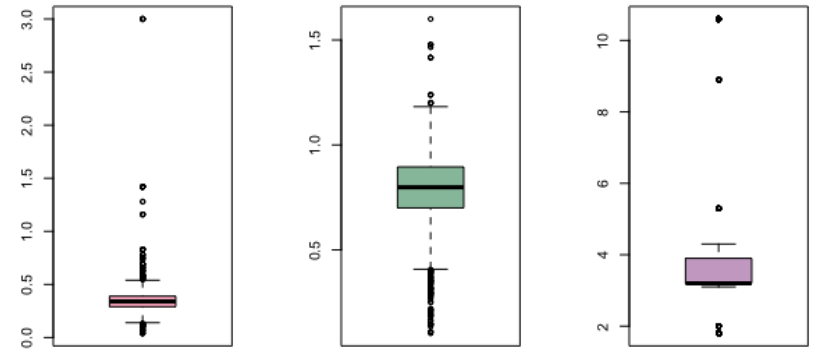

Figure1: Boxplots of 3 continuous variables

As shown in the above figure, we can see that all 3 variable sample data contain several outliers. For ‘pirat’ and ‘lvrat’ which represent a ratio of consumer payment to their income and the ratio between loan to value of the house respectively, the distribution (excluding the outliers) looks normal, while the distribution of ‘unemp’ which shows the unemployment rate in applicant's industry is significantly non-normal. We will use the histogram to study more.

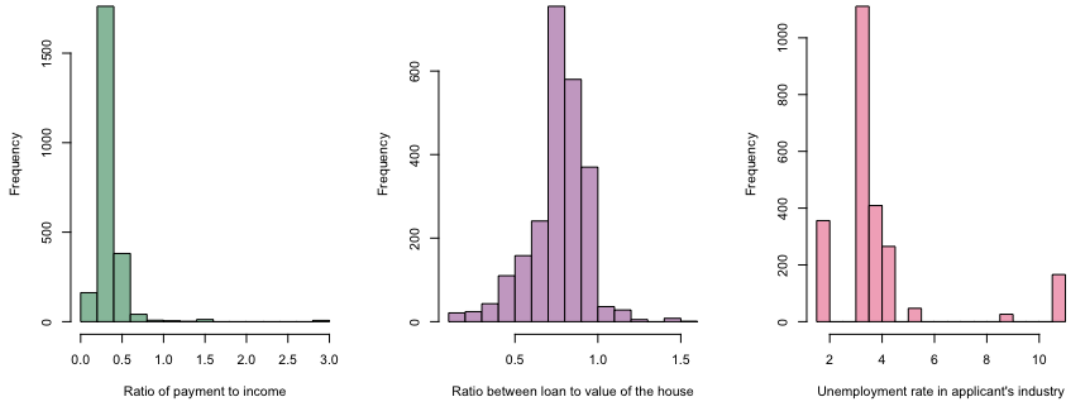

Figure2: Histograms of 3 continuous variables

The first and second histograms show the skewed normal distribution which corresponds to the displayed boxplots above, while the histogram of the unemployment rate can explain that the selected sample data does not cover the entire distribution. Some ranges of data are missing from the sample data set, we might need a bigger size of the sample to cover all ranges. We will investigate more in a further step.

Next for binary categorical variables, we will use ‘table’ to show how often each category of variable occurs and confirm if the data is balanced between categories. And as shown in the table below, apart from ‘deny’ which is intentionally balanced, there is only a ‘single’ variable that the number of samples in each category is slightly balanced. We will use bivariate analysis to explore more whether each variable is associated with the response variable ’deny’ or not.

Table 1: Frequency table of 5 categorical variables

	deny	**phist**	**insurance**	**condomin**	**single**
no	1239	2022	2214	1654	1300
yes	1141	358	166	726	1080

2.2 Bivariate analysis

Again, the boxplot is used, this time, to check the relation between each continuous variable and each category of the response. The mean of ’pirat’ and ‘lvrat’ are slightly different for each group of the ‘deny’ (approval result). These 2 variables might influence the approval. While the means of ‘unemp’ for both categories of ‘deny’ are almost the same.

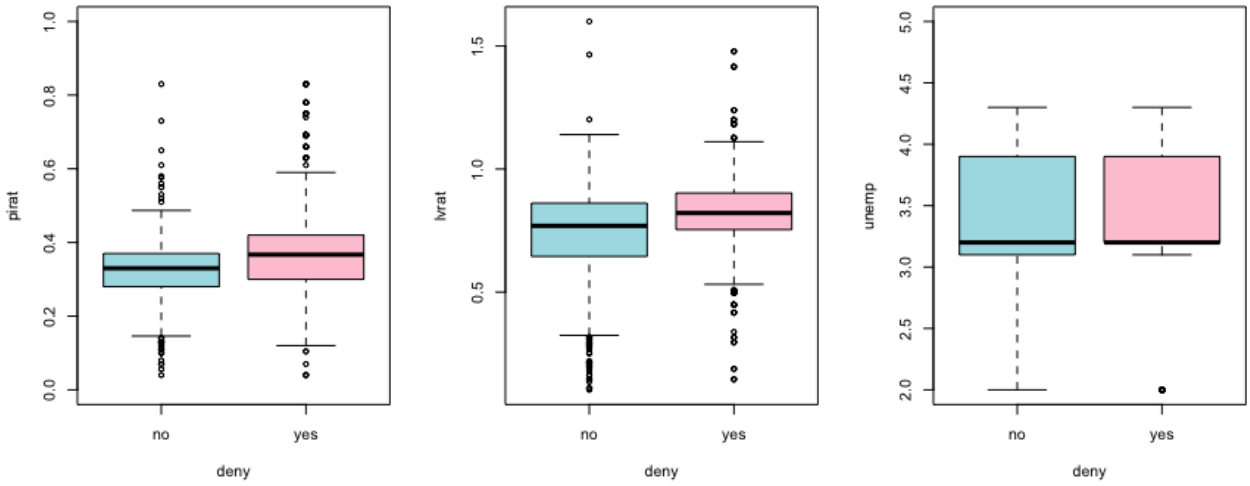

Figure3: Boxplots of 3 continuous predictors for each category of the response

To be more precise, the t-test is implemented to confirm the hypothesis for equality of means between each continuous variable and each group of the mortgage approval result. The result given in the below table shows that, with the p-value of approximate 0 of ‘pirat’ and ‘lvrat’, we have important evidence to reject the hypothesis for equality of means. Whereas the p-value for ‘unemp’ is higher than 0.05 which means that we have no evidence to reject the equality of means hypothesis. Then in the next procedure of the study, we will create a model by leaving this variable out.

Table 2: Welch Two Sample t-test table

data	t	df	p-value
pirat[deny == "yes"] and pirat[deny == "no"]	9.4298	1363	< 2.2e-16
lvrat[deny == "yes"] and lvrat[deny == "no"]	13.8	2375.7	< 2.2e-16
unemp[deny == "yes"] and unemp[deny == "no"]	2.3096	2318.2	0.021

As for categorical variables, we will use a 2 by 2 contingency table with, a log-likelihood ratio to test the hypothesis of the independence of 2 variables (response variable and each factor variable). in addition, the chi-square will be used to confirm the result.

Table 3: 2 by 2 contingency table

		phist		**insurance**		**condomin**		**single**	
		no	yes	no	yes	no	yes	no	yes
deny	no	1183	56	1238	1	877	362	748	491
	yes	839	302	976	165	777	364	552	589

Table 4: log-likelihood ratio table

	G	**X-squared df**	**p-value**
deny~phist	240.54	1	< 2.2e-16
deny~insurance	244.95	1	< 2.2e-16
deny~condomin	2.0187	1	0.1554
deny~single	34.532	1	4.193e-09

The contingency table can explain the probability of the approved mortgage for the individual without a bad credit record ('phist' = no) is 1183/(1183+56) = 0.95, while the probability of one with a bad credit record is 56/(1183+56) = 0.05. And these can show that having a bad credit record or not is dependent on the approval of the mortgage. And from the p-value of the log-likelihood result also demonstrates that we did not get this by chance. So in this case, we have significant evidence to reject the hypothesis of the independence of these 2 features. The 'insurance' and 'condomin' give the same result as 'phist', which means we have strong evidence to confirm that 'phist', 'insurance' and 'single' are associated to 'deny'. Whereas the p-value of the log-likelihood ratio between 'deny' and 'condomin' is more than 0.05, then we can conclude that we do have not enough evidence to reject the null hypothesis.

3. Feature selection and modelling

First, we will use the analysis results from the previous step to select the features (variables) to include in the model and 5 predictors variables are selected which are 'pirat', 'lvrat', 'phist', 'insurance' and 'single'. Then after splitting sample data into 2 sets, Train data and Validation data, we will use the Train data to fit the model. Since the response variable in this study is a nominal binary categorical variable (yes or no question), we will use logistic regression to create the model.

Model_1: glm(formula = deny ~ pirat + lvrat + phist + insurance + single, family = binomial(link = "logit"), data = xTrain) **AIC**: 2091.5

Coefficients:	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-4.2870	0.3333	-12.860	< 2e-16 ***
pirat	4.0968	0.5688	7.203	5.89e-13 ***
lvrat	2.7720	0.3518	7.880	3.28e-15 ***
phistyes	1.8179	0.1781	10.208	< 2e-16 ***
insuranceyes	4.8108	1.0085	4.770	1.84e-06 ***
singleyes	0.5765	0.1065	5.415	6.13e-08 ***

The result above shows that we have robust evidence to reject the null hypothesis of the independence between the response and each predictor. It means that all selected features strongly influence the response variable.

The model describes that considering all 5 predictors are equal to 0, the individual has a probability of mortgage rejection equal to exp(-4.287)/(1+exp(-4.287)) = 0.014 or they have a probability of getting approval of 1-0.014 = 0.986. The estimate coefficient of 'pirat' explains that a 1 unit increase of the ratio of payment to income ('pirat') is associated with an augment of the odds of mortgage approval (deny versus approve) by exp(4.0968) = 60.15. As for 'lvrat', it shows that 1unit increase of loan to value ratio augments the odds of mortgage rejection by exp(2.772) = 16.

Holding other variables at a fixed value, the odds of getting a rejection on a mortgage application for the individual with a bad credit record ('phistyes') over the odds of getting rejection on a mortgage application for the individual without a bad credit record ('phistno') is exp(1.8179) = 6.16. Insurance is the most influential predictor, the probability of mortgage rejection for the individual who rejects the insurance will be exp(-4.287+4.8108)/1+exp(-4.287+4.8108) = 0.628. While the probability of mortgage rejection of the single individual ('singleyes') will be exp(-4.287+0.5765)/1+exp(-4.287+0.5765) = 0.024, the probability increases by (0.024-0.014) = 0.01 from the one who is not single (in control of all other variables). *See more about probability calculation in the appendices.*

Next, we try to use the AIC feature selection tool in r to select the variables for the model and again use the Train data set to create the logistic regression model, then predict the results and verify the accuracy of the model.

Model_2: glm(formula = yTrain ~ pirat + lvrat + phist + ***unemp*** + insurance + single, family = binomial(link = "logit"), data = xTrain) **AIC**: 2085.6 (*see more detail in appendices*).

Model_2 includes 1 more variable 'unemp' which makes the AIC score becomes a bit better but in exchange, it makes the model more complex, and it may impact the study cost. So, the first model has been selected for further investigation.

4. Evaluating model performance

After getting the model, we use the Validation data to confirm and compare the accuracy of the models. The result is shown in terms of the confusion table and the average prediction accuracy.

Table 4: confusion table

Model_1		Actual ('deny' =)		
		no	yes	Total
Predict ('deny' =)	no	201	81	282
	yes	46	147	193
	Total	247	228	**475**

From 475 sample cases, there are actual 247 approved cases, the model correctly predicted 201 cases and incorrectly predicted (as rejected) 46 cases. While the actual rejected cases are 228, our model accurately forecasted 147 cases. **The accuracy average of this model is 73%**

Then, by using the whole data set, we use, in this study, the non-exhaustive cross-validation method called k-fold cross-validation and the leave-one-out exhaustive cross-validation method to assess how the prediction results of the models will generalize to this data set to prevent the overfitting and underfitting. And it shows that the model accuracy has not changed from the previous prediction by the Validation data set at about 73%.

10-fold: 0.728	Leave 1 out: 0.727

Conclusion

This study concludes that we can use a logistic regression model to predict whether the mortgage application will be approved or rejected and the payment to income ratio, loan to value rate, public credit record, insurance status and marital status should be provided. The insurance is the most influential variable, if the individual rejects the insurance they have a high possibility to get a mortgage denied, while the marital status (single or not) impacts the model the less. In case their payment to income ratio and loan to value ratio are 0 and they have no bad credit record (phistyes=0), they accept the insurance (insuranceyes = 0), and they are not single (singleyes =0), the probability of getting approval of the mortgage will be 0.986.

The average accuracy of using this model to predict is about 73%. We can improve the accuracy by adjusting the predictors, but we need to compare it with other factors (model complexity, cost, ...).

Appendices

<Probability calculation for Model_1>

Variable	**Coefficient**						
(Intercept)	-4.287	1	1	1	1	1	1
pirat	4.0968	0	1	0	0	0	0
lvrat	2.772	0	0	1	0	0	0
phistyes	1.8179	0	0	0	1	0	0
insuranceyes	4.8108	0	0	0	0	1	0
singleyes	0.5765	0	0	0	0	0	1
	sumproduct	-4.287	-0.1902	-1.515	-2.4691	0.5238	-3.7105
Probability	exp()/exp()+1	0.014	0.453	0.180	0.078	0.628	0.024

<Model_2 detail>

Model_2: glm(formula = deny ~ pirat + lvrat + phist + unemp + insurance + single, family = binomial(link = "logit"), data = xTrain)

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-4.57094	0.35099	-13.023	< 2e-16 ***
pirat	4.07071	0.57311	7.103	1.22e-12 ***
lvrat	2.79696	0.35439	7.892	2.97e-15 ***
phistyes	1.81811	0.17785	10.223	< 2e-16 ***
unemp	0.07078	0.02517	2.812	0.00492 **
insuranceyes	4.79870	1.00866	4.757	1.96e-06 ***
singleyes	0.57773	0.10672	5.414	6.18e-08 ***

AIC: 2085.6

<Code>

```
library(ggplot2)
library(ggrepel)
library(dplyr)
library(GGally)
library(dsEssex)
library(tidytext)
library(tidyverse)
library(AICcmodavg)
library(psych)
library(MASS)
library(caret)
library(tidyr)
library(nnet)
library(ROSE)
```

```
library(DescTools)

#Load data and remove all rows that contain NA
full.data <- read.csv("/Volumes/Macintosh HD - Data/00_Study
folder/MA334/Final/Mortgage/HMDA2.csv")
#check data structure
str(full.data)
#confirm if there is some missing value
sum(is.na(full.data))
#convert all variables to be factor
p.data <- as.data.frame(lapply(full.data, as.factor))
#check data structure
str(p.data)
#convert back the variables which are not factor to their type
p.data$pirat <-full.data$pirat
p.data$lvrat <- full.data$lvrat
p.data$unemp <- full.data$unemp
#check data structure
str(p.data)
#check the balance of response variable (binary categorical variable)
table(p.data$deny)
# Over or under sampling to unbalance data
p.data <- ovun.sample(deny~.,data=p.data, p=0.5,seed=1, method="both")$data
#check the balance of response variable again
table(p.data$deny)
#univariate analysis
# summary statistics for continuous variable
summary(p.data[,c(2:3,5)],digits = 2)
#boxplot
par(mfrow = c(1,3))
boxplot(p.data$pirat,xlab = "Ratio of payment to income",ylim = c(0,1.5),col = "#97C1A9")
boxplot(p.data$lvrat,xlab = "Ratio between loan to value of the house",col = "#CBAACB")
boxplot(p.data$unemp,xlab = "Unemployment rate in applicant's industry",col = "#F3B0C3", ylim
= c(2,6))
#histogram for continuous variables
par(mfrow=c(1,3))
hist(p.data$pirat,xlab = "Ratio of payment to income",col = "#97C1A9" , main = "Histogram of
Payment ot Income Ratio")
hist(p.data$lvrat,xlab = "Ratio between loan to value of the house",col = "#CBAACB", main =
"Histogram of Loan to Value Ratio")
hist(p.data$unemp,xlab = "Unemployment rate in applicant's industry",col = "#F3B0C3", main =
"Histogram of Unemployment rate in applicant's industry")
# summary statistics for categorical variables
summary(p.data[,-c(2:3,5)],digits = 2)
```

```
# checks levels of a response variable (binary categorical variable )
levels(p.data$deny)
#Bivariate analysis of categorical variables (table, g-test,fisher and chi-square)
table(p.data$deny,p.data$phist)
GTest(table(p.data$deny,p.data$phist))
fisher.test(table(p.data$deny,p.data$phist),simulate.p.value = T)
chisq.test(table(p.data$deny,p.data$phist),simulate.p.value = T)
table(p.data$deny,p.data$insurance)
GTest(table(p.data$deny,p.data$insurance))
fisher.test(table(p.data$deny,p.data$insurance),simulate.p.value = T)
chisq.test(table(p.data$deny,p.data$insurance),simulate.p.value = T)
table(p.data$deny,p.data$condomin)
GTest(table(p.data$deny,p.data$condomin))
fisher.test(table(p.data$deny,p.data$condomin),simulate.p.value = T)
chisq.test(table(p.data$deny,p.data$condomin),simulate.p.value = T)
table(p.data$deny,p.data$single)
GTest(table(p.data$deny,p.data$single))
fisher.test(table(p.data$deny,p.data$single),simulate.p.value = T)
chisq.test(table(p.data$deny,p.data$single),simulate.p.value = T)
#Bivariate analysis of continuous variables (boxplot)
par(mfrow = c(1,3))
boxplot(pirat~deny,ylim=c(0,1),col = c("#ABDEE6","#FEC8D8"))
boxplot(lvrat~deny,col = c("#ABDEE6","#FEC8D8"))
boxplot(unemp~deny,ylim=c(2,5),col = c("#ABDEE6","#FEC8D8"))
# t test for difference of mean see p value
t.test(pirat[deny=="yes"],pirat[deny=="no"])
t.test(lvrat[deny=="yes"],lvrat[deny=="no"])
t.test(unemp[deny=="yes"],unemp[deny=="no"])
#Split data to train and test
set.seed(4321)
trainIndex <- createDataPartition(deny, p = .8, list = FALSE,times = 1)
Train <- p.data[trainIndex,]
str(Train)
table(Train$deny)
xTrain <- Train[,-1]
yTrain <- Train[,1]
Valid <- p.data[-trainIndex,]
xValid <- Valid[,-1]
yValid <- Valid[,1]
table(Valid$deny)
str(Valid)
#logistic model fit using train data
manual.model <-
glm(yTrain~pirat+lvrat+phist+insurance+single,family=binomial(link='logit'),data=xTrain)
```

```
summary(manual.model)
summary(manual.model)$coefficients[,c(1,4)]
manual.model$coefficients
#predict using validation data
contrasts(yValid)
f.glm.probs2 <- predict(manual.model ,xValid,type = "response")
f.glm.predicted2 <- rep("no",475)
f.glm.predicted2[f.glm.probs2>0.5]="yes"
table(f.glm.predicted2,yValid)
round(mean(f.glm.predicted2==yValid),2)
# fit model by using step (AIC) to select the features
full.model <- glm(yTrain ~.,family=binomial(link='logit'),data=xTrain) # logistic model fit
min.model <- step(full.model) # this finds a best fit model by selecting and rejecting variables
summary(min.model) # min_model is the best model found check p values for each variable
#CV MODEL_1
set.seed(333)
# defining training control
# as cross-validation and
# value of K equal to 10
train_control1 <- trainControl(method = "cv",number = 10)
# k-fold CV
model.111 <- train(deny~ pirat + lvrat + phist + insurance +
                     single, data = p.data,
                   method = "glm",
                   family=binomial(),
                   trControl = train_control1)
print(model.111)
res1 <-model.111$results[,2:3]
round(res1,3)
#Leave One Out CV
set.seed(888)
ctrl1 <- trainControl(method = "LOOCV")
model111 <- train(deny~ pirat + lvrat + phist + insurance +
                    single, data = p.data,
                  method = "glm",
                  family=binomial(),
                  trControl = ctrl1)
print(model111)
#Reference
#predict using validation data
contrasts(yValid)
f.glm.probs <- predict(min.model ,xValid,type = "response")
f.glm.predicted <- rep("no",475)
f.glm.predicted[f.glm.probs>0.5]="yes"
```

```
table(f.glm.predicted,yValid)
round(mean(f.glm.predicted==yValid),2)
#CV MODEL_2
set.seed(333)
# defining training control
# as cross-validation and
# value of K equal to 10
train_control <- trainControl(method = "cv", number = 10)
# k-fold
model <- train(deny~ pirat + lvrat + phist + unemp + insurance +
                 single, data = p.data,
               method = "glm",
               family=binomial(),
               trControl = train_control)
print(model)
model$results
res2 <-model$results[,2:3]
round(res2,3)
#Leave One Out
set.seed(888)
ctrl <- trainControl(method = "LOOCV")
modelll <- train(deny~ pirat + lvrat + phist + unemp + insurance +
                 single, data = p.data,
               method = "glm",
               family=binomial(),
               trControl = ctrl)
print(modelll)
```

References

Munnell, A. H., Tootell, G. M. B., Browne, L. E. and McEneaney, J. (1996). Mortgage Lending in Boston: Interpreting HMDA Data. American Economic Review, 86, 25–53.

Stock, J. H. and Watson, M. W. (2007). Introduction to Econometrics, 2nd ed. Boston: Addison Wesley.