Analysis Report of Customer Loan Application

Cindy

2022-04-29

1. Introduction

1.1 Background

The loan is a fast-growing and intense competition industry. Decreasing the default rate to improve loan quality and control risk is very important for a loan company's business. The company needs to learn from the past, analyzing existing loan data to direct future business. The report is based on the company loan data set, including 1000 clients and 21 features, such as checking balance, age, job, and others. Refer to the appendix Code Book for detail.

1.2 Analysis Purpose

According to the provided data, the default percentage of the company loan business is 30%, which is high. The report aims to mine data information, assist the company in deeply understanding its customers and their loan behaviors, and figure out the particular clients and behavior that should be noticed. Meanwhile, based on the data set, build a model to estimate the default probability, classify clients' risk level, make a reference for lower the default loss, and improve the loan quality.

1.3 Report Brief

The report has three parts: introduction, analysis report, and conclusion. The analysis includes client and loan feature analysis, default loan analysis, default probability estimate, client risk level classification and requested amount, and loan issuance amount analysis.

By client and loan feature analysis, we find that most customers are not married foreign workers whose bank account balance is not much but have their own house or another property. And most of the loans for daily consumption, the average request amount is 3271 dollars, and the average loan term is about 21 months. The default loan analysis explores and presents the association between loan default and other features. For instance, people with fewer checking balances have more defaults; people with better history credit levels have more defaults, which is not normal.

After exploring the data, we build a logistical regression model on existing data and try to predict the default probability of a new loan application. The prediction accuracy is about 75%, which could employ as an assistant method for making the decision. At the same time, classifying the client's risk level by default probability is also a helpful way to help the company take some risk control methods, such as adjusting the loan origination rate, adding guarantees, etc.

2.Analysis Report

2.1 Client and Loan Feature Analysis

We compiled some summary statistics of borrowers' traits and their application features. Table 1 summarized the categorical features, and table 2 summarized the numerical features. From the two tables, we can see that 96.3% of borrowers are foreign workers, 85.8% of them are not married, and they are from age 19 to 75, 88.5% are from age 20 to 50; all have no more than two dependents,84.5% have one dependent. Most customers have work experience, account for 93.8%, and only 2.2% are unemployed non-residents.

Many customers do not have much-floating capital in the bank account, only 6.3% of customers have more than 1000 dollars in a checking account, and 4.8% have more than 2000 in their savings account. While 82.1% of them have their own house or fully paid for the housing,84.6% have another property. Some people have more than one existing loan, accounting for 36.4%, and 8.8% of customers have a delayed credit history. Almost 80% of clients borrow loans for consumption. Meanwhile, 81.4% of clients did not have an installment plan,47.4% paid quarterly, and 15.7% paid monthly.

default			foreign worker			landline		
no	700	70%	no	37	3.7%	yes	404	40.4%
yes	300	30%	yes	963	96.3%	none	596	59.6%
housing			installment plan			other_debtors		
fully paid	108	10.8%	bank	139	13.9%	co-applicant	41	4.1%
own	713	71.3%	stores	47	4.7%	guarantor	52	5.2%
rent	179	17.9%	none	814	81.4%	none	907	90.7%
checking balance			personal_status			property		
< $0	274	27.4%	single	548	54.8%	society savings	232	23.2%
$1 - $1000	269	26.9%	commonlaw	310	31.0%	real estate	282	28.2%
> $1000	63	6.3%	married	92	9.2%	other	332	33.2%
unknown	394	39.4%	single	548	54.8%	unknown/none	154	15.4%
installment rate			residence history			existing loans		
1	136	13.6%	1	130	13.0%	1	633	63.3%
2	231	23.1%	2	308	30.8%	2	333	33.3%
3	157	15.7%	3	149	14.9%	3	28	2.8%
4	476	47.6%	4	413	41.3%	4	3	0.3%
saving_balance			employment_length			credit_history		
< $500	603	60.3%	unemployed	62	6.2%	critical	293	29.3%
$501 - $1000	63	6.3%	0 - 1 yrs	172	17.2%	delayed	88	8.8%
$1001 - $2000	103	10.3%	1 - 4 yrs	339	33.9%	repaid	530	53.0%
> $2000	48	4.8%	4 - 7 yrs	174	17.4%	fully repaid this bank	49	4.9%
unknown	183	18.3%	> 7 yrs	253	25.3%	fully repaid	40	4.0%
job			purpose			*note: installment rate*		
non-resident	22	2.2%	electronics	280	28.0%	*1=weekly,2=bi_weekly,3=monthly,*		
unskilled resident	200	20.0%	new vehicle	234	23.4%	*4=quarterly*		
skilled employee	630	63.0%	used vehicle	103	10.3%			
self-employed	148	14.8%	furniture	181	18.1%			
dependents			business	97	9.7%			
1	845	84.5%	education	50	5.0%			
2	155	15.5%	(Other)	55	5.5%			

The shortest loan term is four months, and the most extended loan is six years, about 80% of loan terms from 6 months to 36 months. The minimum request amount is 250 dollars, the maximum is 18424, and about 91% of loans amount are fewer than 7500 dollars.

	loan month	request amount	age
Min	4	250	19

	loan month	request amount	age
Median	18	2320	33
Mean	20.9	3271	35.55
Max	72	18424	75

The following distribution plots of the loan months, the requested amount, and borrower age (Figure 1) show that these features' distribution skew to the right, presenting the former numerical statistics about these features.

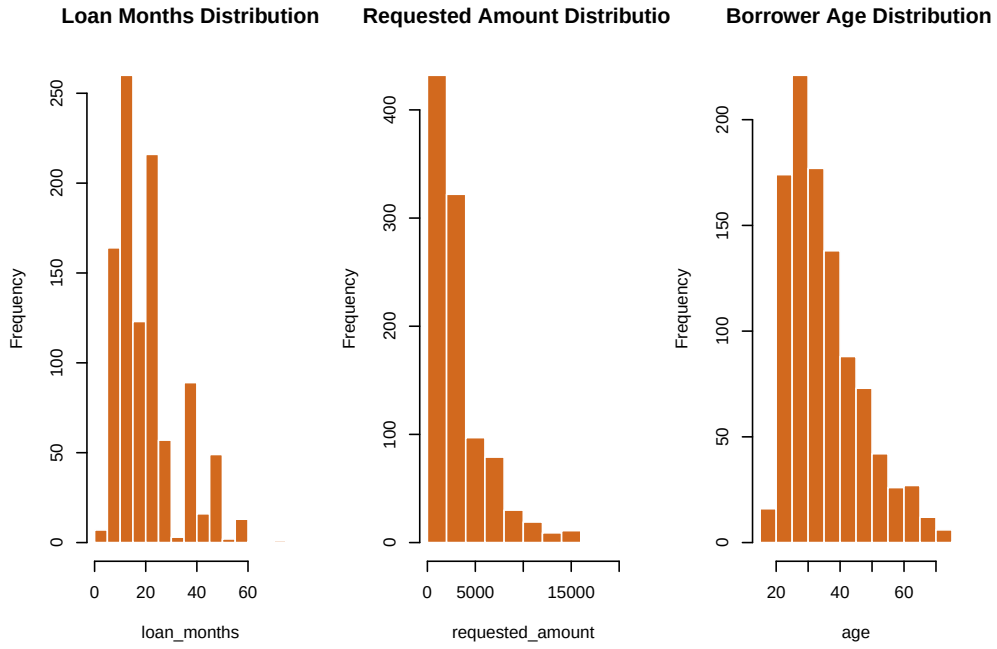

Figure 1: The Distribution Plots

2.2 Default Loan Analysis

There are 30% of loan applications defaulted, and 70% did not default in the analysis sample data set . The total default amount is about 1.18 million, accounting for 36.12% of the total loan amount. The default rate is high. If the profit can not cover the default loss, the business will lose money. This part analyzes the influencing factors of problem loans and intends to infer the characteristics of people prone to problem loans. At the same time, the default amount is discussed.

2.2.1 Correlation of Features

The figure 2 variables correlation matrix plot shows the correlation of all analysis features, including the correlation of default and other features. The number in the square is the correlation coefficient. A positive value means two features related positively, and a negative value means negatively related, the absolute value closer to 1 means more related, and 0 represents not related. The correlation coefficients of default and dependent, default, and residence history are 0, which means there is no relationship between them. The correlation coefficients of checking balance, credit history, and loan month are the greatest, so the report will analyze them first.

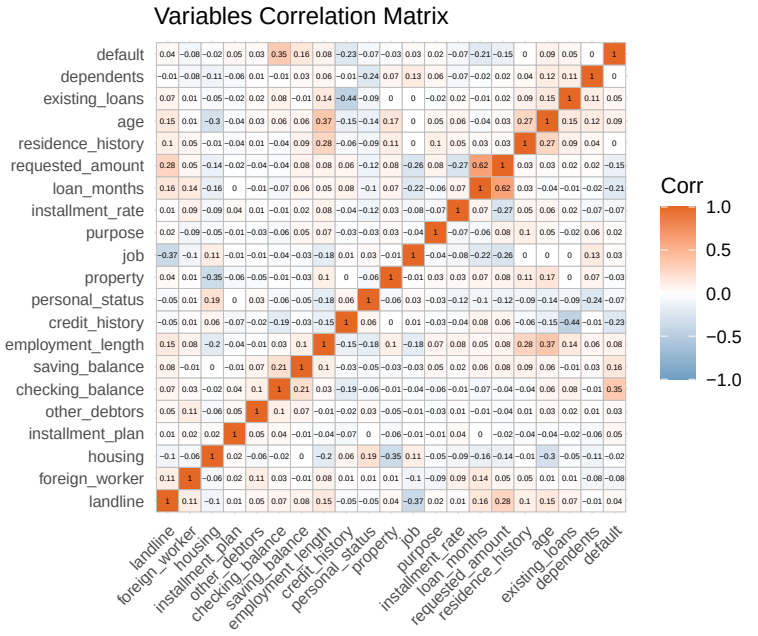

Figure 2: Variables Correlation Matrix

2.2.2 Default Loan and Checking Balance

The following table and figure 3 Plots of Default Loan and Checking Balance indicate that people with less checking balance have more defaults and account for a higher default proportion. Although negative balance customers have more defaults than 1-1000 balance people, the default amount is less. The default amount of people with more than 1000 dollars is the fewest.

	< $0	$1 - $1000	> $1000	unknown	sum
Default Times	135	105	14	46	300
Default Amount	460,837	499,249	24,160	197,192	1,181,438

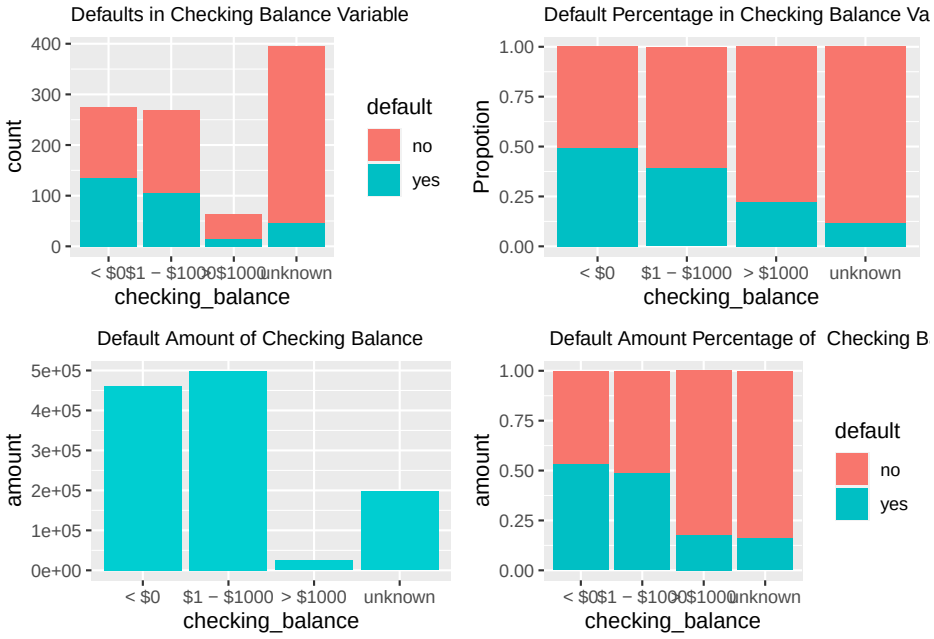

Figure 3: Plots of Default Loan and Checking Balance

2.2.3 Default Loan and History Credit

Figure 4 Plots of Default loans and Credit History show that the default probability and the default amount increased as the clients' history credit level improved, which is not normal. The proportion of critical credit's default amount is 22.6%, and the ratio of fully repaid credit's default amount is 65.1%. It is better for the company to re-check the history credit and related loan issuance criterion. The left side plots show that the company offered the most loans to clients who have repaid history credit, whose default times and amount are the most. The proportion of default amount of repaid clients is about 39.5%.

Credit	fully repaid	fully repaid this bank	repaid	delayed	critical
Default Times	25	28	169	28	50
Default Amount	138156	94405	636380	107750	204747

Figure 4: Plots of Default Loan and Credit Histiory

2.2.4 Default Loan and Loan Month

Figure 5 Plots of Default Loan and Loan Term shows that the default percentage of times and the amount increased with the loan term extended, which means the possibility of default increases with the longer loan term. The shorter the loan term, the less default probability. Most clients borrow money for 13 to 24 months (2 years), in which the default times and default amount are the most, which the following table value verified. The highest default probability is located in year 4, from 37 to 48 months. More than four years' loan also had a high default probability, about 50%.

loan Month	1-12 months	13-24 month	24-36 months	37-48 month	more than 48 months
Default Times	76	122	57	37	8
Default Amount	153,583	373,436	316,309	259,932	78,178

Figure 5: Plots of Default Loan and Loan Term

2.2.5 Default Loan and Other Features

Figure 6 Default Percentage Plot of Client Features describes the default times percentage of age, personal status, job, and work length. From the plots, we could find figure out people aged 31 to 50, single or married, skilled, and worked more than four years had relatively lower default probability. Clients aged less than 30 or from 50 to 60, in a common-law relationship or divorced, self-employed or unemployed, with work experience less than one year, had higher default probabilities. The company needed to pay more attention.

Figure 7 Default Percentage Plot of Client Property and Debit indicates that clients have a lower default probability if they have houses and real estate. Clients with three existing loans have the lowest percentage of defaults, and guarantor of other debtors' default percentage is the lowest.

Figure 8 Default Percentage Plot of Other Features states that the clients whose saving balance is less than 500 dollars saving balance, or the loan for education purposes, default times percentage are higher. The probability of default increases when the installment rate increase. There are more application amounts, and the likelihood of default is higher.

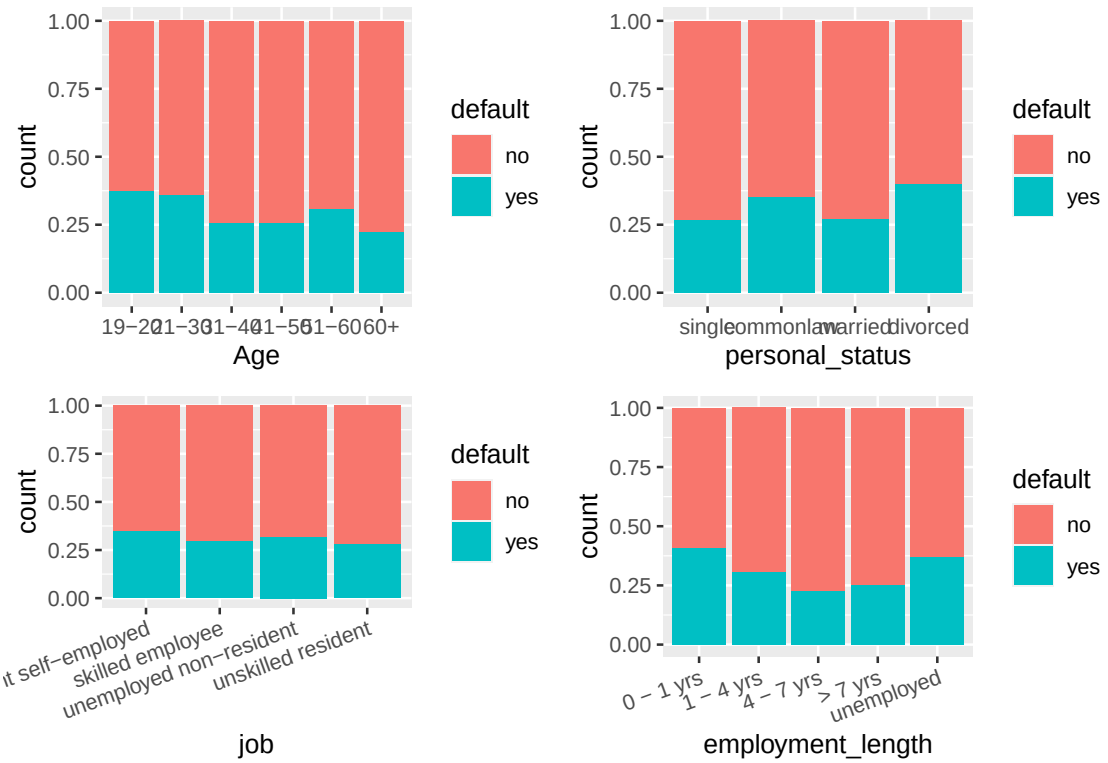

Figure 6: Default Percentage Plot of Client Features

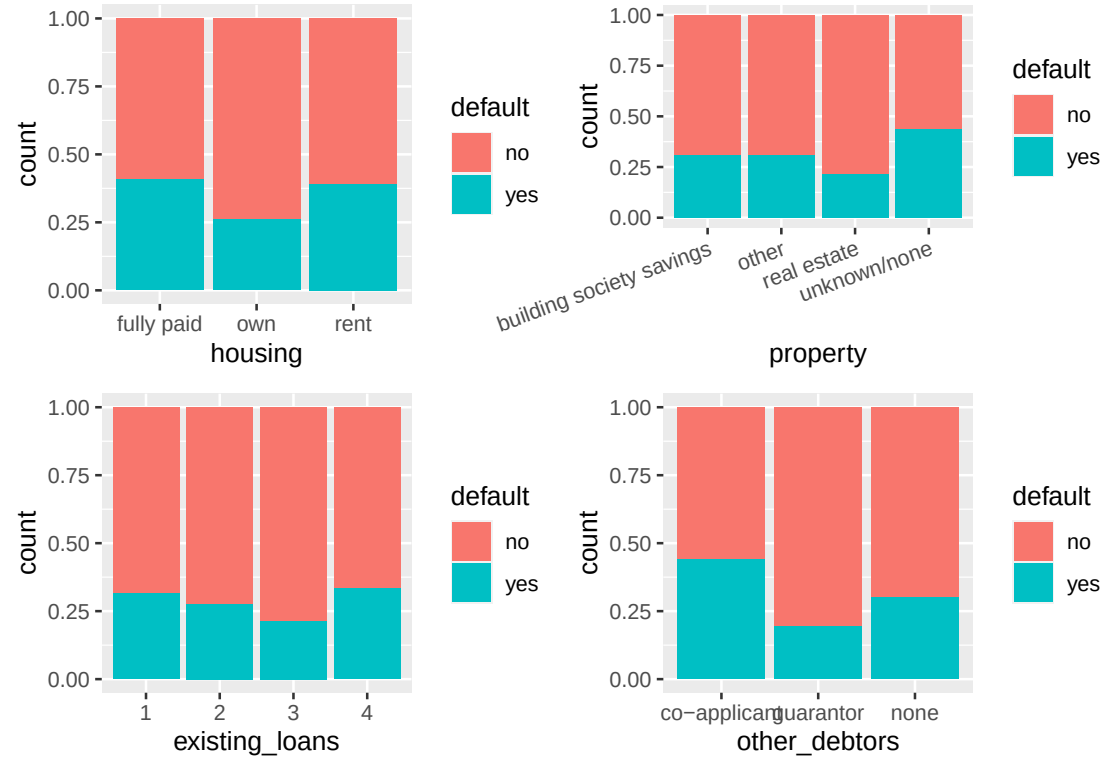

Figure 7: Default Percentage Plot of Client Property and Debet

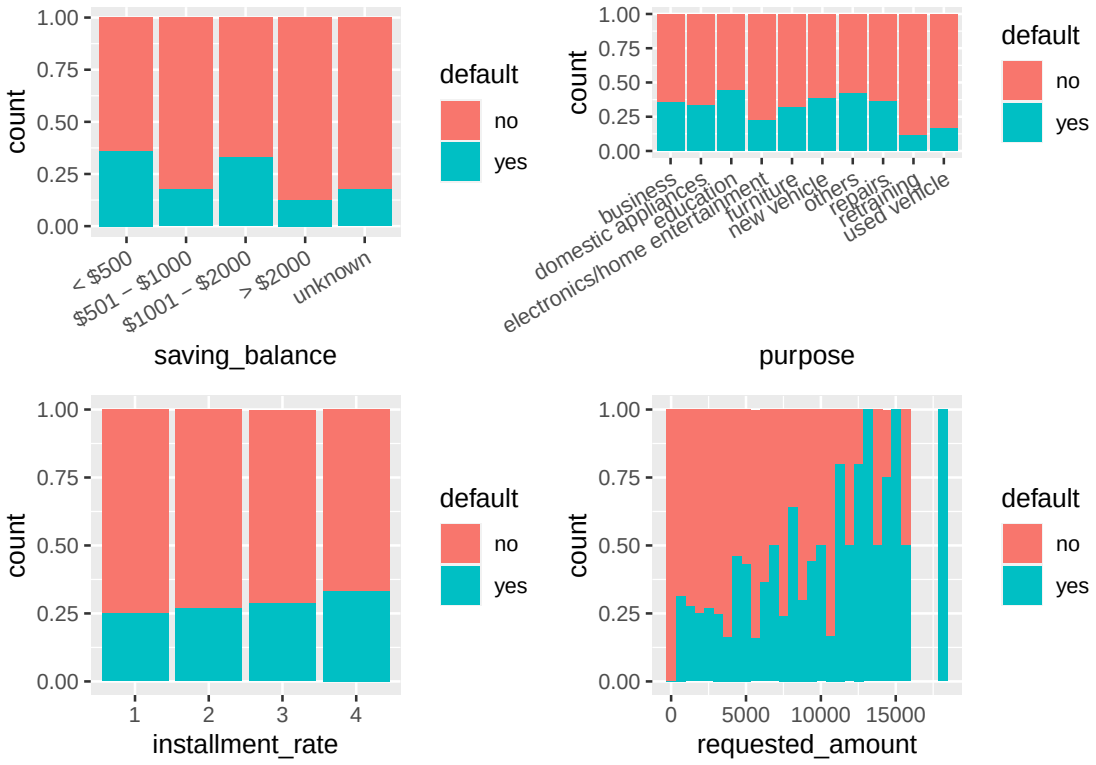

Figure 8: Default Percentage Plot of Other Features

2.3 Special Clients Analysis

As we are in the loan business, focusing on problem loans and particular clients is more significant than showing regular loans. We classify customers into different groups to figure out whether specific clients exist. We utilize the Manhattan method to measure the similarity of the 1000 customers by some of their characteristics: checking balance, saving balance, loan months, requested amount, installment rate, purpose, job, housing,credit_history, default, and age.And, using the average linkage method to compute the distance between clusters, employ the PAM algorithm and Silhouette method to calculate the optimal number of groups, and then cluster all observations into groups.

Figure 9 *Distribution Plot of Clients Similarity* plot presents the similarity of all clients, and each point represents a client. The closer distance means they are more similar, and further means more various. As the plot showed, most of the points cluster together, and a few points are far from the cluster center, which means they are dissimilar from the clustered points.

Theoretically, the observations can be divided into any number of groups. According to the PAM algorithm, as the top right-hand plot indicates, dividing the observation into two groups is the optimal choice. So, we cluster two groups, as the *Cluster Plot of Observations k=2* plot showed. There are 983 observations divided into group1 representing typical clients. The 17 observations in or around the blue circle are particular and more dissimilar from the red group. Compared to group1 and group 2, We could find that the group 2 clients requested higher loan amounts. The average value of the requested amount of group 1 is about 14718, but the average requested amount of group 1 is about 3109. More loan money means more risk, and the company needs to pay more attention to group 2 clients.

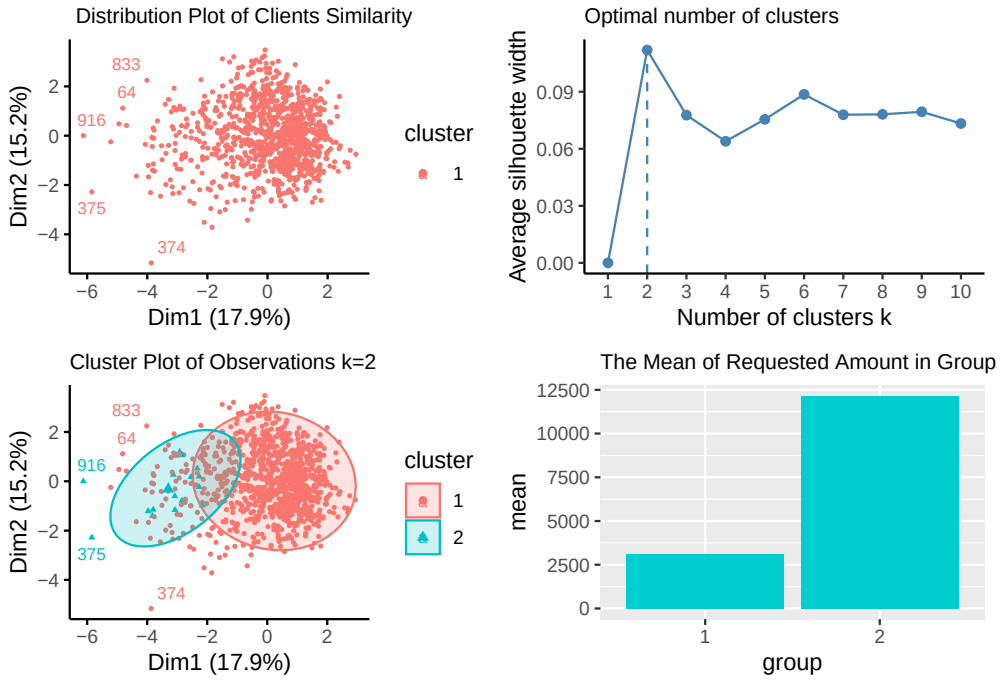

Figure 9: Analysis Plots of Clients' Cluster

2.4 Estimating Default Probability

We are interested in whether the application will default or not for a loan application and the default probability for a new application or a new client. We already had some history loan data, so it is possible to model based on the existing data set to predict a new application's default probability.

2.4.1 Model Selection

There are many model methods available to estimate whether a loan will default or not, such as KNN(k-nearest neighbors), Regularized Regression, and Logistical Regression.

We experimented with three modeling methods in the report, and their experimental accuracy is very similar.As the Knn model is challenging to show the correlation between default and other features, We choose the Logistical Regression model to do deep analysis. For the process of modeling, see the appendix for detail.

	k-nearest Neighbors	Regularized Regression	Logistical Regression
Training Accuracy	0.78	0.7714286	0.7757143

The logistic regression model algorithm as the following formula shows:

$$p(X) = \frac{1}{1 + e^{-(B_0 + B_1 x_1 + B_2 x_2 + B + B_q x_q)}}$$

To prevent the model from overfitting, we split 1000 observations into training data sets with 700 records and testing data sets with 300 observations. The model builds on the training data set. The logistical regression model accuracy is about 0.78, which is not very high. Still, we could also reference its outcome to help us decide, which is much better than randomly guessing probability, 0.5.

2.4.2 Experiment Case

To compare the predicted outcome and the actual result, we chose three loans from the test data set as new applications, which were not used to train the model. The estimated value and actual value as the following table shows:

	Observation 1	Observation 2	Observation 3
Real Default	no (1)	no(1)	yes(0)
Estimate Default	no (1)	yes(0)	yes(0)
Estimate Default Probability (0)	0.039	0.605	0.587

Generally speaking, if the probability of default is bigger than 0.5, we will consider it default. The default probability of Observation 1 is less than 0.5, which means have a higher likelihood of not default, considered as not default. The actuary of the predicted outcome is about 0.67, two records estimated correctly, and one misestimated.We can experiment with more applications and employ 300 observations not used in the training model as new application data to predict and test. There are 222 records correctly estimated. The accuracy is about 0.74.In these two case experiments, we used all features as predictors. Theoretically, we could also use partial features to estimate the default probability.

2.4.3 Feature Importance Analysis

Meanwhile, by analyzing the features' regression coefficient in the models, we could find that features' importance is different. The following plots show the most critical ten features relevant to the logistic regression and regularized regression models' default. The first ten important feature ranks of the two models are the same. The clients' checking balance, credit history, saving balance, loan month, and installment rate affect the default relatively more than others, which we need to pay more attention to.

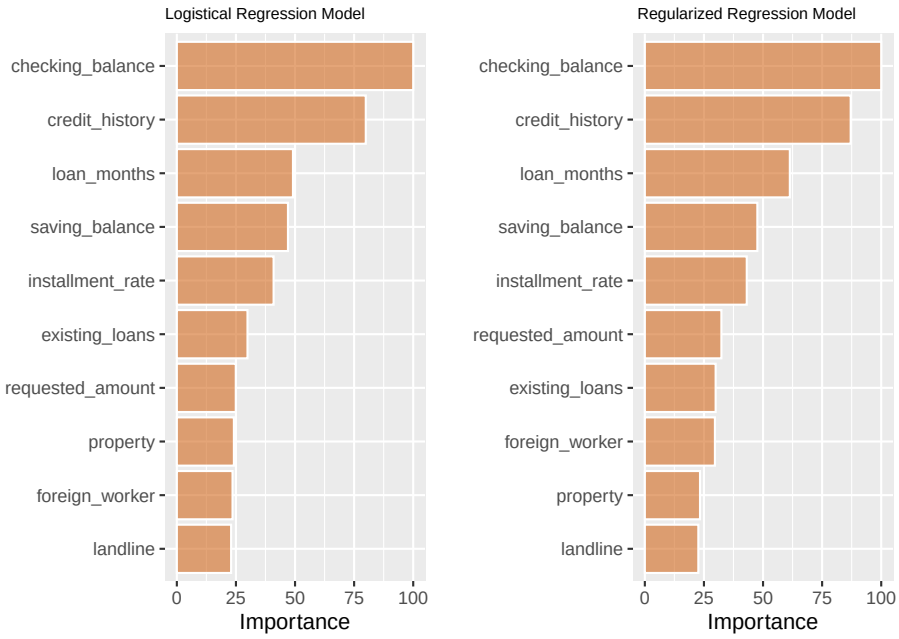

Figure 10: The Feature Importance Rank

2.5 Client Risk Level

According to the historical loan data and model, we could classify the client's default risk level to help us make better loan granting decisions. The classified rule is as follows:

Risk Level	High	Medium	Low
Default Probability	>= 0.8	<0.8 and >=0.5	<0.5

The report experiments with the rule to the 700 observations and gets the following result.

Risk Level	High	Medium	Low
Count	27	118	555
Percentage	0.04	0.17	0.79

The following table shows the ten observation's default probability and risk level and related information.

default	level	default_P	checking	credit_history	amount	saving	loan_months
1	low	0.4384723	1	3	806	1	15
1	medium	0.5007349	4	2	15653	1	60
0	medium	0.5126982	2	3	2671	3	36
1	low	0.1110625	1	3	3763	5	21
1	low	0.1755617	4	3	1768	1	12
1	low	0.1723677	3	3	3049	1	18
0	low	0.4741889	2	2	8086	3	36
0	low	0.1208459	1	1	2625	1	16
1	low	0.2619958	1	1	1382	3	24
0	low	0.3042938	2	3	1922	1	12

Note

default: 0=default, 1= no default

checking balance: 1= < $0,2=$1 - $1000,3= >$1000,4=unknown

credit history: 1=critical,2=delayed,3=repaid,4=fully repaid this bank, 5=fully repaid

saving balance: 1=< $500,2=$501 - $1000,3=$1001 - $2000,4=>$2000,5=unknown

The company needs to pay more attention to high-risk clients. For the 27 high-risk level clients in 700 observations, there are 23 defaulted and four not defaulted, and the average requested amount is 7932, higher than the total average amount. The loan month of high-risk clients range from 12 months to 60 months, and the median month is 42 months, much longer than the total average level. Meanwhile, 26 clients' checking balance is less than 1000 dollars, 23 clients' saving balance is less than 500 dollars, but most of them are fully repaid past loans before.

2.6 Request Amount and Loan origination Amount Analysis

Except for loan default probability, we are also interested in the loan amount of the loan application, which could help the company decide how much money needs to be prepared and how much money will be lost as default. By building the regularized regression model, we could estimate the possible requested amount of the clients and find the most related important features: loan month and installment rate, as the following figure 10 shows.

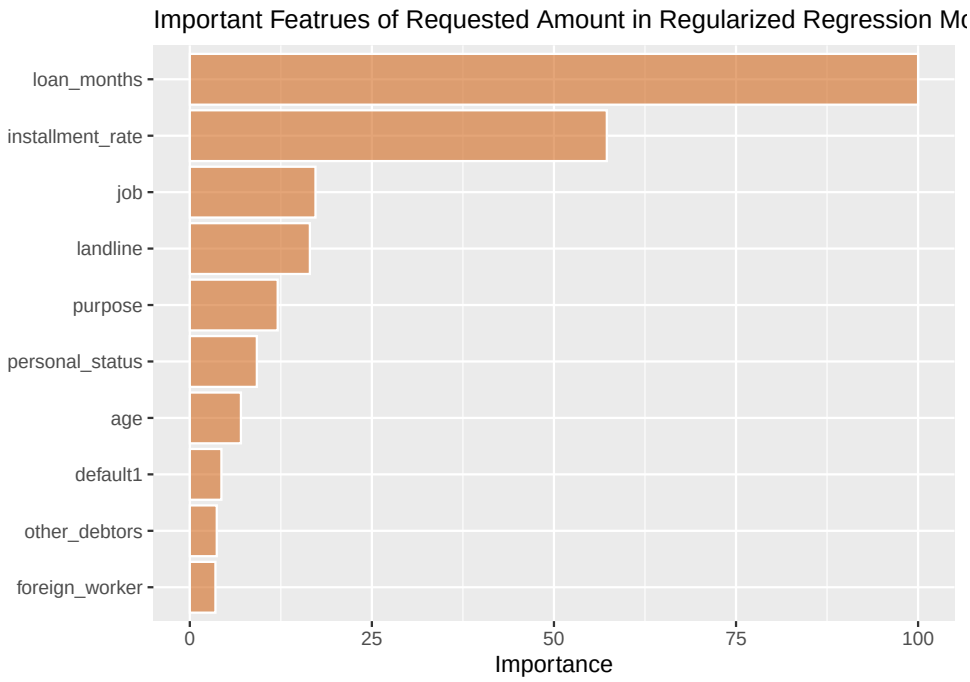

Figure 11: Important Featrues of Requested Amount

As the previous analysis showed, some customers are high-risk clients and always request more money. Suppose the company gives them all the requested money, which means more money would be a loss. For example, taking the previous 27 high-risk clients, if the company only issuance 60% requested money, it will decrease the 70632 dollars loss. It is a helpful way to take some guarantee or lower the issuance percentage of requested money by clients risk level.Correctly estimating clients' risk levels is essential in the default risk analysis and control method.

Risk Level	Requested Amount	Assurance Amount (60%)	Difference
no default	37573	22543.8	15029.2
default	176582	105949.2	70632.8

3. Conclusion

The company default rate is 30%. The most current clients have low deposits, are not married foreign workers, and have their own house or other property. The average request amount is 3271 dollars, and most of the loan is for consumption. The most loan term is not longer than 30 month. The clients' checking balance, credit history, saving balance, loan month, and installment rate affects the default relatively more than others, which we need to pay more attention to. The loan's loan term, installment rate, land-line or not, job situation, and loan purpose impact the amount people want to request, which also means the amount the company needs to prepare.

The loan industry is a kind of risk management business. So the company needs to notice a larger amount of loans, problem loans, and high-risk clients and adjust feature weight, not rely on one or two features too much. According to the previous analysis about the default loan and credit history, people with better history credit levels have more defaults, which is not normal. History credit situation is an important index to judge people's credit status, but it does not mean people with a good history will not default.

Meanwhile, we can use historical data to build a model to predict the interesting variable to help us decide. But how to improve the actuary of the model is the key. One way to improve the model's performance is to improve the quality of the fitting value and predictor features. We could collect and treat feature data from three aspects: customer characteristics, repayment ability, and repayment willingness. To improve the model and its performance, we can consider adding features such as deposit loan ratio and collect the original numerical data, such as checking balance and saving balance; it is better to use numerical data to analyze. Also, the client risk level is a critical analysis index.

APPENDIX

Part A Code Book

Description

The data of *Loan Application Data* with 1000 observations and 21 variables, provided by the company, including the clients and loan application information.

Variables information

Variable Name	Variable Label	Missing Data	Range	Data Type	Value
default	default			character	no/yes
landline	landline			character	none/yes
foreign worker	foreign_worker			character	no/yes
housing	housing			character	fully paid/own/rent
installment plan	installment_plan			character	bank/none/stores
other debtors	other_debtors			character	co-applicant/guarantor/none
checking balance	checking_balance			character	< $0 / $1 - $1000/> $1000/unknown
saving balance	saving_balance			character	< $500 / $501 - $1000/$1001 - $2000/ > $2000/unknown
employment length	employment_length			character	0 - 1 yrs/ 1 - 4 yrs/ 4 - 7 yrs/ > 7 yrs/unemployed
credit history	credit_history			character	critical/delayed/repaid/fully repaid this bank/fully repaid
personal status	personal_status			character	single/ common-law/ married /divorced
property	property			character	building society savings/other/real estate/unknown/none

Variable Name	Variable Label	Missing Data	Range	Data Type	Value
job	job			character	management self-employed/skilled employee/unemployed non-resident/ unskilled resident
purpose	purpose			character	electronics/home entertainment/ new vehicle/ furniture/ used vehicle /business/education/Other
installment rate	installment_rate			character	1/2/3/4
loan months	loan_months		4-72	numerical	months
requested amount	requested_amount		250-18424	numerical	dollars
residence history	residence_history		1-4	integer	
age	age		19 - 75	integer	
existing loans	existing_loans		1-4	integer	
dependents	dependents		1-2	integer	

Part B Modeling Part

Build Model to Estimate Default Probabilty

```
# treat data in different way

set.seed(666)



library(rsample )
library(caret)
# split data set

loannew$default<-as.factor(loannew$default)

set.seed(666)
loansplit<-initial_split(loannew,prop=0.7,stata="default")
loan_train<- training(loansplit)
loan_test<- testing(loansplit)

# knn

set.seed(666)
```

```
knn_model<- train(default~.,data = loan_train,
method = "knn",
preProc = c("zv", "center", "scale"),
trControl = trainControl(method = "cv", number = 10),
tuneLength = 10
)



# predict in train data and get the accuracy
pre_knntrain<- predict(knn_model,loan_train)

cm<-caret::confusionMatrix(pre_knntrain,loan_train$default)
cm$overall['Accuracy']
```

```
## Accuracy
##     0.78
```

```
#regularized regression
set.seed(666)

glmnet_modle<- train(default~.,data = loan_train,
                     method="glmnet",
                     preProc = c("zv", "center", "scale"),
                     trControl = trainControl(method = "cv", number = 10),
                     tuneLength = 10)

# predict in train data and get the accuracy
pre_glmnettrain<- predict(glmnet_modle,loan_train)

cm<-caret::confusionMatrix(pre_glmnettrain,loan_train$default)
cm$overall['Accuracy']
```

```
##  Accuracy
## 0.7714286
```

```
# logistic regression

set.seed(666)
glm_model<-train(default~.,data = loan_train,
                 method="glm",family="binomial",
                 preProc = c("zv", "center", "scale"),
                 trControl=trainControl(method="cv",number=10))

# predict in train data and get the accuracy
pre_gtrain<- predict(glm_model,loan_train)

cm<-caret::confusionMatrix(pre_gtrain,loan_train$default)
cm$overall['Accuracy']
```

```
##  Accuracy
## 0.7757143
```

```r
# predict a new application

new<- as.data.frame(loan_test[1:3,])

pre_new<-predict(glm_model,new);pre_new
```

```
## [1] 1 0 0
## Levels: 0 1
```

```r
loan_test$default[1:3]
```

```
## [1] 1 1 0
## Levels: 0 1
```

```r
pre_new_p<-predict(glm_model,new,type="prob");pre_new_p
```

```
##            0         1
## 1 0.03943415 0.9605658
## 2 0.60519712 0.3948029
## 3 0.58741832 0.4125817
```

```r
# predict in test data set

pre_gtest<- predict(glm_model,loan_test)

cm<-caret::confusionMatrix(pre_gtest,loan_test$default)
cm$overall['Accuracy']
```

```
##  Accuracy
## 0.7433333
```