MGMT 590: Using R for Analytics
Shiny Final Project

Risk Control in Consumer Lending

Team Bubble Tea Rocks:
Jou Tzu (Rose) Kao | kao21@purdue.edu
Rong (Alice) Liao | rliao@purdue.edu
Hongxia Shi | shi395@purdue.edu
Shenyang Yang | yang1469@purdue.edu

Purdue University,
Department of Management,
403 W. State Street, West Lafayette, IN 47907

Abstract

As the only business that sometime rejects customers, consumer lending is known for taking a lot of risk. When a transaction is generated, the company will invest money at first, then the return is at the mercy of the customers. Therefore, risk control is considered the core competence of companies in consumer lending business. With the help of machine learning technology, companies could use mathematical approach to evaluate the risk of a new customer and make better decisions.

In this project, we design an R-Shiny app to help consumer lending companies on risk control improvement. The app will allow companies to evaluate the probability of whether a customer will payback or not. After inputting customer's profile, the app will give a result as if the customer is a "good' or "bad customer" based on prediction of the customer's probability of paying back, i.e. whether the company should lend money or not. A logistic regression model is used behind the app. The model was first trained (75%) and tested (25%) with 1000 records with payment status, and then used for prediction. In the following report, all the design procedures will be discussed, as well as the corresponding considerations and evaluations.

Keywords: Credit Risk, cost matrix, logistic regression, Random Forest, Cluster modeling

Business Problem

The business objective of this project is to help consumer lending companies to decide whether to give a loan to a customer or not.

Because of the nature of consumer lending industry, the profitability of a consumer lending company largely depends on the probability of the customers repaying the money. When creating a transaction, the company will first pay money to the customer and hoping the customer will repay the debt. If the customer fails to pay the debt, the company will suffer a lost when the transaction happens. In order to mitigate the risk of loss, before the transaction, intervention is needed if the customer is expected to default. How to predict the probability of repay becomes the key problem for the company to control the risk.

Fortunately, full credit record and demographic information of an individual can be purchased from the Credit Bureau. But because of the volume and dimension of the records, how to process the data effectively and efficiently separates the successful companies from the rest. Moreover, how to use the data to predict the profitability of a new customer is even more important. By using all the information beyond the FICO score, the companies will have a more accurate prediction of an individual's probability to repay the debt. And by using machine learning technology instead of manually evaluating the risk, the app will help consumer lending companies to improve efficiency and profitability to a great extent.

This is especially true for the consumer lending companies for two reasons: (1)since most of their loans are unsecured, default frisk is a bigger concern for the consumer lending companies; (2)as the amount of every transaction is relatively small compared to the commercial lending companies or other financial services, their need to use machine learning to help increase the processing efficiency is much urgent.

Being said, using analytics approach is not magic. The app cannot guarantee 100% accuracy in prediction. Company may mistakenly accept a bad customer or deny a good customer. The app will output the probability of a customer repaying the debt to the company. How to decide whether to lend money to the particular customer all depends on how the company conceive the potential risk and benefits. With a cost matrix provided by the company, i.e. the measurement used to associate the risk with predict accuracy, this app could help the company to find the customer classification criteria with the lowest risk.

Analytics Problem

The analytical objective of this project is to find the best model to fit the customer profile with highest accuracy and to use the cost matrix to optimize the prediction with the lowest cost.

To achieve the analytical objective, a few technical problems are needed to be solved. (1)13 of the dependent variables are categorical, and the independent variable is also a categorical variable. How to select a machine learning model that could fit the variables is very important. The second part of this problem is how to choose the model with the best fit and highest accuracy; (2) the output of the model should be probabilities and the decision-making process should be based on selected criteria by the user that could fit into different strategies. To help the users to make the decision, the cost matrix should be integrated with the output accuracy to generate an optimization in the prediction model; (3) transform the algorithm into an interactive interface that could let the user to generate the result in a straightforward way is needed. To solve this problem an R-Shiny app will be developed.

In this process, we assume that we have all the data we need for a new customer, and we only classify the customers in two groups: good or bad, i.e. accept the loan or not. Every companies should have their own risk preference in terms of mistakenly accepting a bad customer or denying a good customer. In response, the app provides a risk preference matrix which enable users to customize the combination that best fit their risk preference.

To help the user predict the classification of a new observation, the app that should have input area for keying in observation attributes and an output that tells the classification of the observation. Also, the app should also tell the user what is the best decision-making criteria and allow the user to change it when necessary. The KPIs of the mechanism will be the accuracy of the prediction on good/bad classification and the specificity of the prediction of detecting the negative class.

Data

Source:

To realize the objectives mentioned previously, we need to find a data that has enough observations for model training, and enough attributes for prediction accuracy. The data used in this project is found on UCI Machine Learning Repository. It was originally created by Professor Dr. Hans Hofmann from Institut f"ur Statistik und "Okonometrie Universit" at Hamburg.

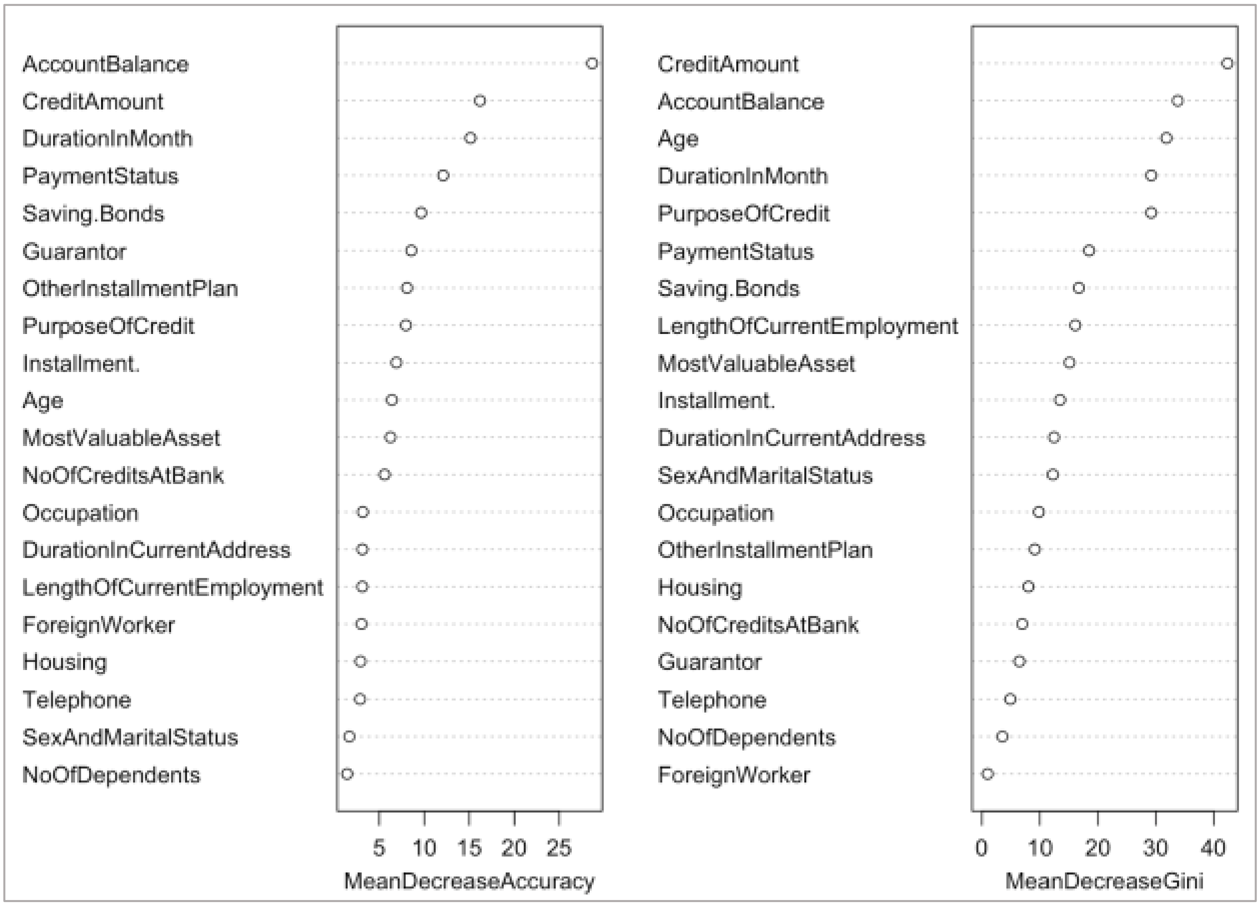
The dataset contains 1000 observations and 20 attributes, which include demographics and financial status of loan applications. Within the attributes, 13 of them are categorical, and 7 of them are numerical. The data classifies every observation as a default customer or not, and come with a cost matrix that defines the decision risk level.

Data cleaning:

First, missing and abnormal values were checked before exploring the dataset. Fortunately, the dataset used in the project is well cleaned.

Second, all the data types were standardized before model building. This dataset has two data types: number (i.e. integer) and category data. For all the categorical features, they were changed into factor data type. For example, the variable Account Balance has four statuses. Factors 1,2,3 and 4 are used to identify each status.

Data relationships



The first graph MeanDecreaseAccuracy shows how much would accuracy decrease when. The second graph MeanDecreaseGini represents how much would the purity decrease when each of the predictors is removed. Reading from the plot, there are many variables play an important role in the regression. The variables are AccountBalance, CreditAmount, DurationInMonth, PaymentStatus, Saving.Bonds, PurposeofCredit, Installment and age. These variables have a significant direction for the bank to decide whether a customer who wishes to loan is a good customer or a bad customer.

Methodology Selection & Model Building

Model building, testing & comparison:

After cleaning the data, four regression model are built and tested: logistic regression, KNN, GBM and Random Forest regression. All the 20 variables are used in the the model building. Within the 1000 observations in the dataset, 750 observations are randomly chosen to train the model and the remaining 250 are saved for model testing.

Model evaluation:

The four models are then compared between accuracy, sensitivity, specificity. The following table summarizes the comparison of the results:

	Logistic regression	**KNN**	**Gradient Boosting**	**Random Forest**
Accuracy	76.4%	66.4%	76%	78.4%
Sensitivity	86%	83.1%	87.2%	91.3%
Specificity	55.1%	20.9%	51.3%	43.3%

Cost matrix	Customer Will Repay	Customer Will Default
Accept Loan	0	Monetary Cost (High)
Reject Loan	Opportunity Cost (Low)	0

Based on the nature of the industry, the cost of the company lending money to a bad customer is much higher than losing a good customer. In response, we are selecting a model which has a well balance between specificity and accuracy. Therefore, the Logistic model is selected for developing the app.

Using R:

After comparing all possible environments: R, Python, and SAS, R is selected for realization of the objectives. This is mainly because R has the Shiny app function that could enable users without programming knowledge to easily generate the result they want. Also, R modeling functions return an object which can be modified and manipulated by the programmer to adapt to new modeling situations and generate predictions, summaries, and etc. [1]

Limitations:

Because there are only 1000 observations indata. The fitted model may not have very high accuracy. Another limitation is that all of the input of the observations must be completed (i.e. 20 attributes) in order to generate the output.

R packages:

The following R packages are used.

Package	Description
randomForest	Classification and regression based on a forest of trees using random inputs
shiny	To build interactive web applications with R
rgl	Provides medium to high level functions for 3D interactive graphics, including functions modelled on base graphics as well as functions for constructing representations of geometric objects
useful	A set of little functions that have been found useful to do little odds and ends such as plotting the results of K-means clustering, substituting special text characters, viewing parts of a data.frame, constructing formulas from text and building design and response matrices.

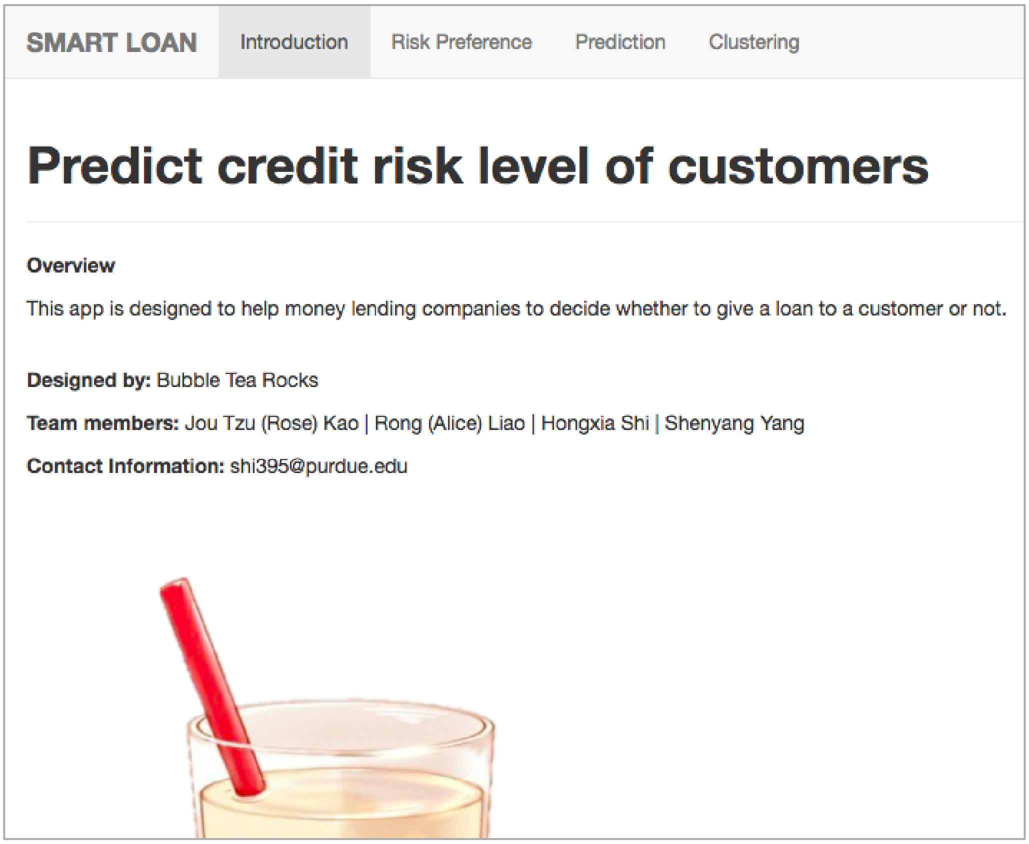
ggplot2	A system for 'declaratively' creating graphics, based on ''The Grammar of Graphics''. You provide the data, tell 'ggplot2' how to map variables to aesthetics, what graphical primitives to use, and it takes care of the details.

Conditional logic:

In the app, the users have the choice to set the cutoff point, which will directly affect the classification result (i.e. good/bad) of the customer. In addition, by providing the risk preference control bars, users can get a recommended cutoff point from us based on their risk preference and the affected level of accuracy of the model.

Functionality, GUI Design & Quality

The name of the app is 'Smart Loan'. The User Interface looks like the following:



Predict credit risk level of customers

Overview

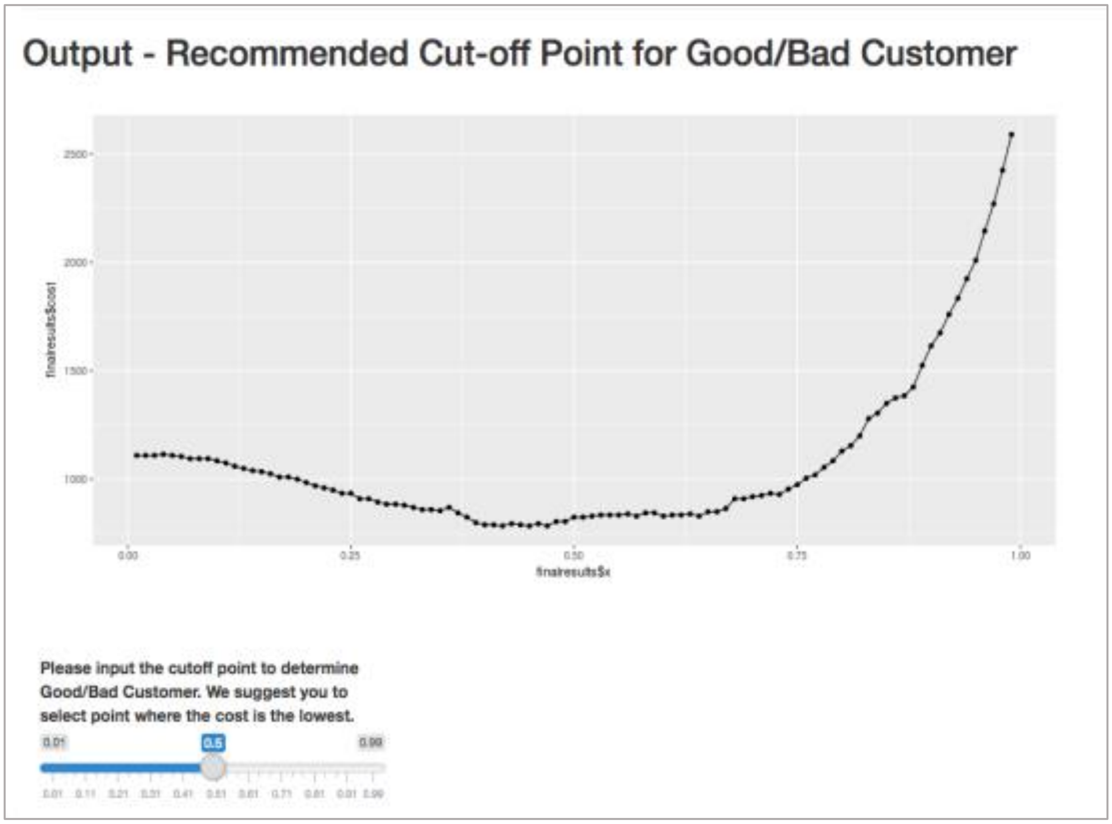
This app is designed to help money lending companies to decide whether to give a loan to a customer or not.

Designed by: Bubble Tea Rocks

Team members: Jou Tzu (Rose) Kao | Rong (Alice) Liao | Hongxia Shi | Shenyang Yang

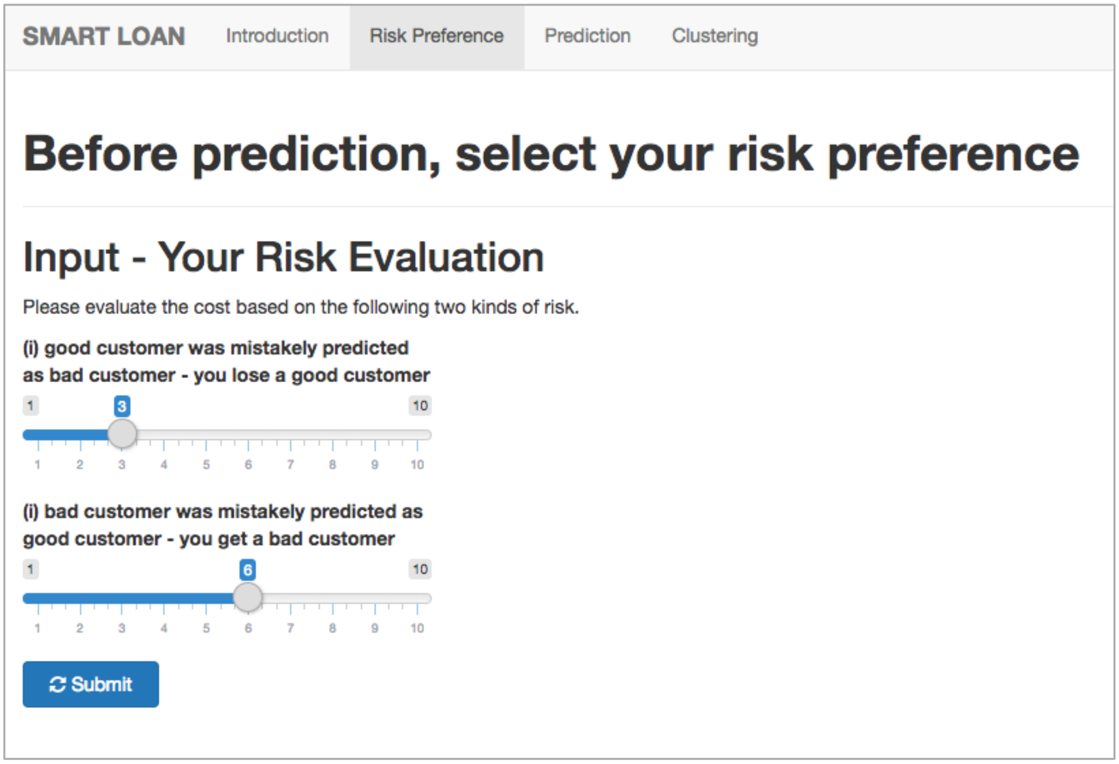
Contact Information: shi395@purdue.edu

Main page: it contains the name and a brief introduction of the app with the name of all the developers. On the top of the main page, there are 4 tabs, within which there are different functions.



Output - Recommended Cut-off Point for Good/Bad Customer

Please input the cutoff point to determine Good/Bad Customer. We suggest you to select point where the cost is the lowest.



Before prediction, select your risk preference

Input - Your Risk Evaluation

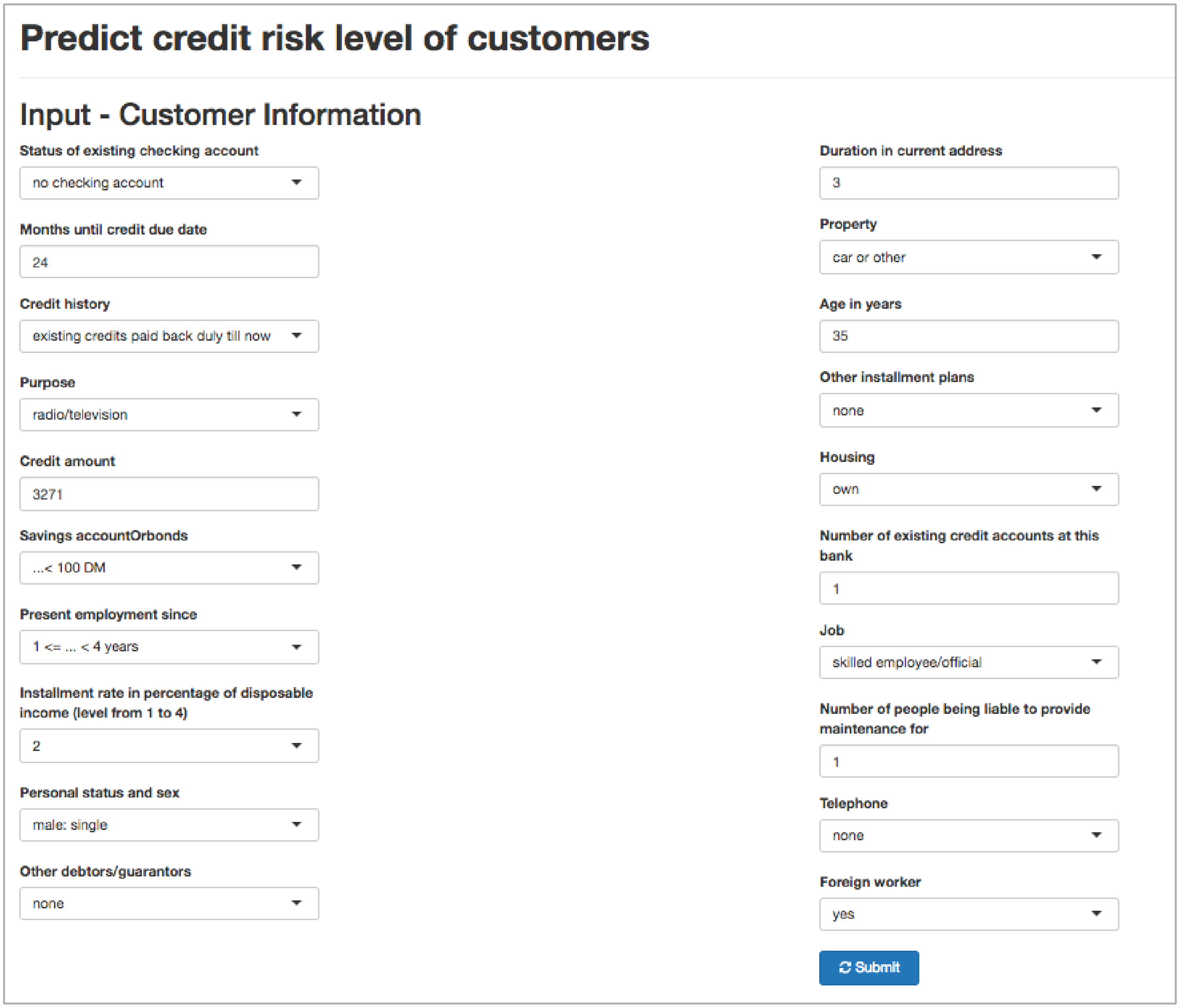
Please evaluate the cost based on the following two kinds of risk.

(i) good customer was mistakely predicted as bad customer - you lose a good customer

(i) bad customer was mistakely predicted as good customer - you get a bad customer

Risk preference:

On the second tab the user is able to choose their risk preferences. Refer to the cost matrix mentioned previously, the app could save the input and generate the cost curve with an optimized cut-off value. The user is then asked to choose the cut-off value for their decision of whether to approve the loan of the customer or not. Choosing the cut-off value at the lowest point of the cost graph is recommended, but the user could always go back and change the cut-off value based on their business strategy.



Predict credit risk level of customers

Input - Customer Information

Status of existing checking account
no checking account

Months until credit due date
24

Credit history
existing credits paid back duly till now

Purpose
radio/television

Credit amount
3271

Savings accountOrbonds
...< 100 DM

Present employment since
1 <= ... < 4 years

Installment rate in percentage of disposable income (level from 1 to 4)
2

Personal status and sex
male: single

Other debtors/guarantors
none

Duration in current address
3

Property
car or other

Age in years
35

Other installment plans
none

Housing
own

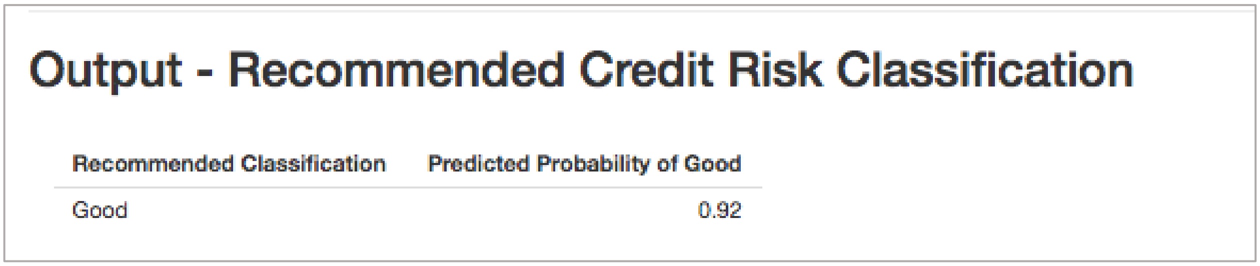
Number of existing credit accounts at this bank
1

Job
skilled employee/official

Number of people being liable to provide maintenance for
1

Telephone
none

Foreign worker
yes



Output - Recommended Credit Risk Classification

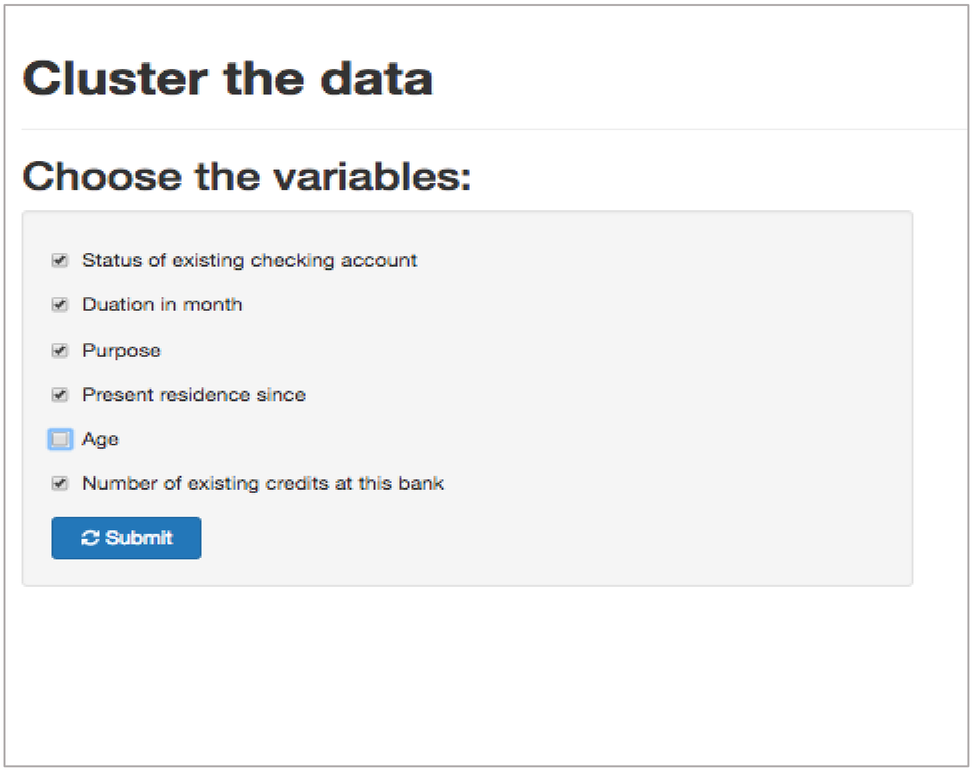
Recommended Classification	Predicted Probability of Good
Good	0.92

Prediction:

After setting the cut-off criteria, the user is then asked to input the loan applicant's demographic and financial information. After submitting the information, the app will give out the probability of payback of this applicant and a recommendation of whether the user should accept the loan or not.

Clustering:

The app also contains a clustering feature that could allow it's users to classify the loan applicants based on different attributes. On the left side, the user can select some attributes that are at their interest, and the right side is the classification plot, in which the dots in the same color represents the applicants with the similar attributes. It is aimed to help the user to draw the customer profiles and identify the potential customer segmentations. The users could then use these information for marketing purpose.



Cluster the data

Choose the variables:

☑ Status of existing checking account

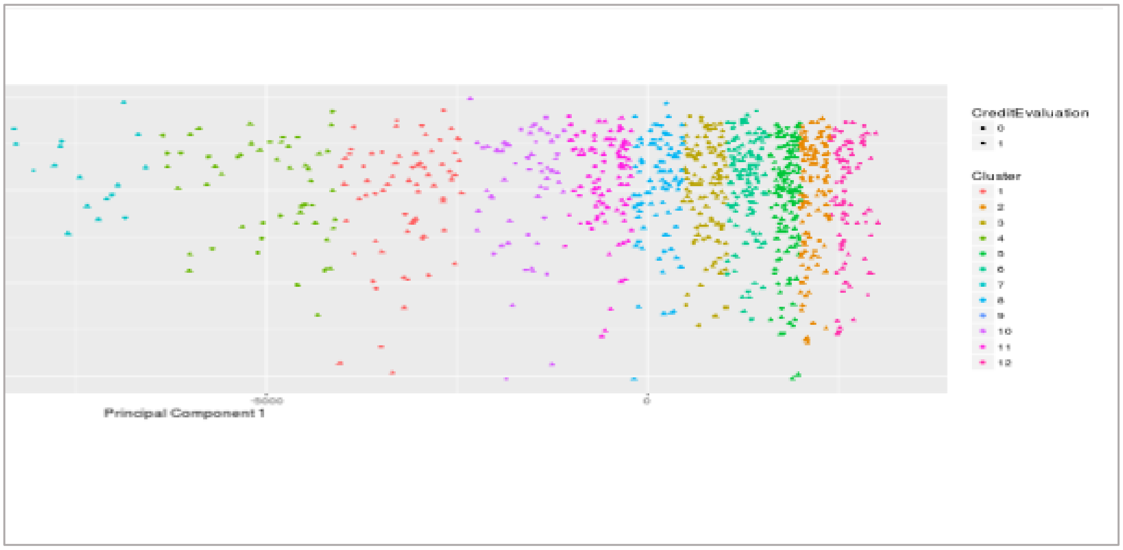
☑ Duation in month

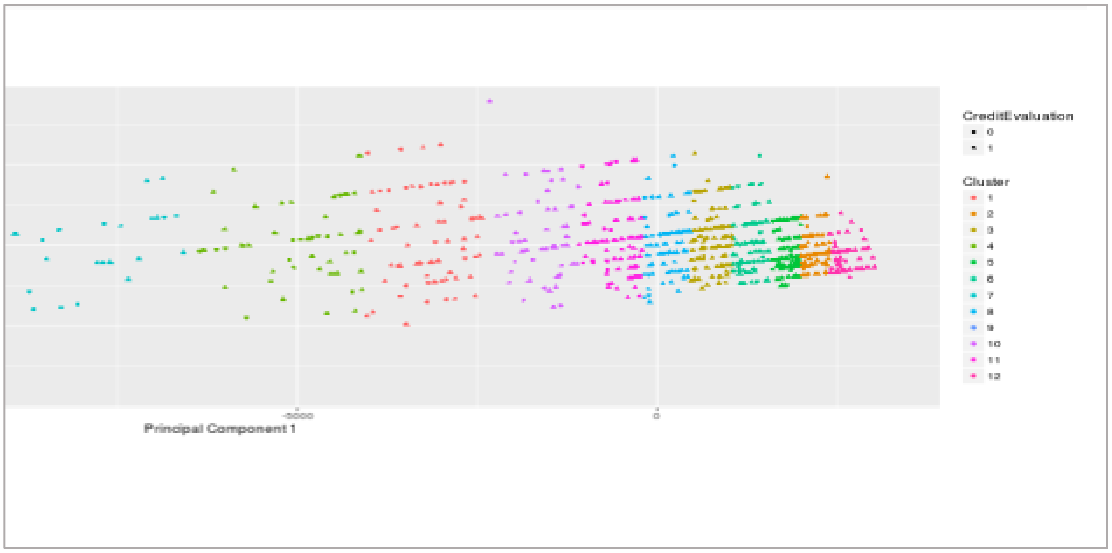
☑ Purpose

☑ Present residence since

□ Age

☑ Number of existing credits at this bank





Improvement

If more time is allowed for this project, the variable used for modeling could be further selected and refined. Especially when the dataset is big, this would help improve process efficiency. Also, nonlinear-logistic model can be tested to increase the model fitness in the future.

Moreover, the model used in this project is the logistic regression. Even though the specificity at the default cut-off value is the highest, when the cut-off value is adjusted based on the user's requirement, the specificity is also changed, as well as the model accuracy. Therefore, future action is needed to taking different models into consideration every time the cut-off value is changed. One proposal here is to calculate the specificity of all the model with every cut-off value.

A minimum cost will be then selected as output for constructing the potential cost v.s. cut-off value.

Also, in the clustering tab, the app cannot output specific features of data of good or bad customer to app users. Additional algorithms are needed here to provide more meaningful insights to the users.

Lastly, the user interface could be made more colorful and interactive, some instructions about how to fill in the survey can be added.

Conclusions

In this project, a R shiny app is successfully designed and developed. It works the function as assisting consumer lending companies with identifying the default risk of the loan applicants and help them decide whether to accept the loan. The accuracy of the method behind the app is good enough for the real application. The app has the ability to integrate a large amount of applicant data as well as the ability to tailor to the risk preference of every user.

The app is designed only as a reference, the users should have the full responsibility of the decisions they made. Team Bubble Tea Rocks reserves the right of final explanation and revision for the terms. The app will not be responsible for any loses in revenue or business opportunities caused by using the app.

References

1.How ANZ uses R for credit risk analysis

http://blog.revolutionanalytics.com/2011/08/how-anz-uses-r-for-credit-risk-analysis.html

2. Decoding the probabilities in Random Forest
https://www.linkedin.com/pulse/decoding-probabilities-random-forest-sanchit-tiwari/

3. Confusion Matrix and Cost Matrix
http://dni-institute.in/blogs/confusion-matrix-and-cost-matrix/

4.PREDICTIVE ANALYTICS FOR CONSUMER LENDING: INTRODUCTION
http://www.pi-cube.com/2015/03/07/predictive-analytics-for-consumer-lending-intro/

5. Data source
https://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data)