Default of Credit Card Clients – Predictive Models

Jennifer Zhang

***Abstract*—Financial institutions mitigate risks by modeling the chances of defaults and credit risk. This paper presents three different machine learning classification models using the Scikit-learn python packages – K-Nearest Neighbors, Logistic Regression, and Random Forest - based on past credit card default data. The data includes various features which lead to the target value of whether there would default payments in the next month. The Random Forest method seems to be most appropriate for this data set, yielding an accuracy of about 82 percent.**

These classification models may serve as guidelines for future decisions of credit card issuers. The data also shows that the top features that determine default payments in the next month include: repayment status in the first month, age, and the bill amount in the first month.

***Index Terms*—Credit card default refers to delinquent credit card payments. Consumers may pay duly or delay payments; The amount of payment may also differ from the bill amount.**

I. INTRODUCTION

This paper examines different classification methods – K-Nearest Neighbors, Logistic Regression, and Random Forest to predict default payment in the next month. There are various features, or independent variables included in the dataset.

II. DATA

This dataset contains information on default payments, demographic factors, credit data, payment history, and bill history of credit card clients in Taiwan from April 2005 to September 2005. The dataset includes 30,000 entries and does not have any missing values.

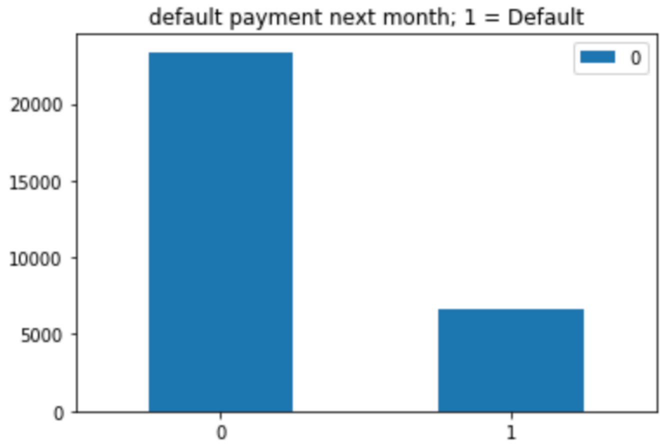
A. Variables

ID	ID of each client
LIMIT_BAL	Amount of given credit in NT dollars (includes individual and family/supplementary credit
SEX	Gender (1=male, 2=female)
EDUCATION	(1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)
MARRIAGE	Marital status (1=married, 2=single, 3=others)
AGE	Age in years
PAY_0	Repayment status in September, 2005 (-1=pay duly, 1=payment delay for one month, 2=payment delay for two months, ... 8=payment delay for eight months, 9=payment delay for nine months and above)
PAY_2	Repayment status in August, 2005 (scale same as above)
PAY_3	Repayment status in July, 2005 (scale same as above)
PAY_4	Repayment status in June, 2005 (scale same as above)
PAY_5	Repayment status in May, 2005 (scale same as above)
PAY_6	Repayment status in April, 2005 (scale same as above)
BILL_AMT1	Amount of bill statement in September, 2005 (NT dollar)
BILL_AMT2	Amount of bill statement in August, 2005 (NT dollar)
BILL_AMT3	Amount of bill statement in July, 2005 (NT dollar)
BILL_AMT4	Amount of bill statement in June, 2005 (NT dollar)
BILL_AMT5	Amount of bill statement in May, 2005 (NT dollar)
BILL_AMT6	Amount of bill statement in April, 2005 (NT dollar)
PAY_AMT1	Amount of previous payment in September, 2005 (NT dollar)
PAY_AMT2	Amount of previous payment in August, 2005 (NT dollar)
PAY_AMT3	Amount of previous payment in July, 2005 (NT dollar)
PAY_AMT4	Amount of previous payment in June, 2005 (NT dollar)
PAY_AMT5	Amount of previous payment in May, 2005 (NT dollar)
PAY_AMT6	Amount of previous payment in April, 2005 (NT dollar)
default payment next month	Default payment (1=yes, 0=no)

Based on the dataset, most values take on discrete or categorical values.

B. Default Payment Next Month

Default payment next month takes on a value of 1 if the consumer defaults on his/her payment in the next month.



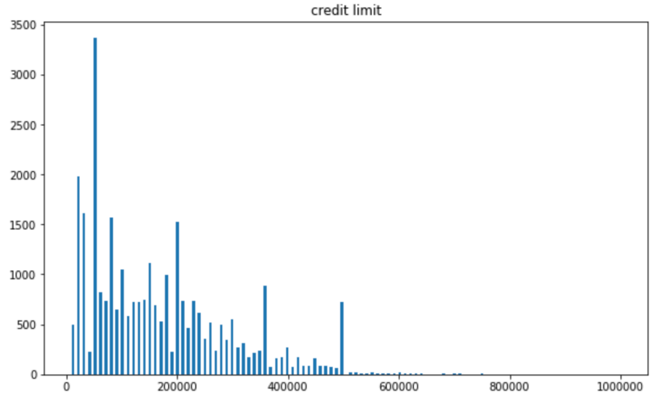
The chance of default next month for the entire dataset is about 0.2212.

This paper is submitted for review on March 19, 2020 to Professor Xiaobai Liu, PhD for Economics 425, Machine Learning, under the Master of Applied Economics Program at University of California, Los Angeles (UCLA).

Jennifer Zhang is a student of the Master of Applied Economics Program at UCLA (email: jennifer.zhang@berkeley.edu).

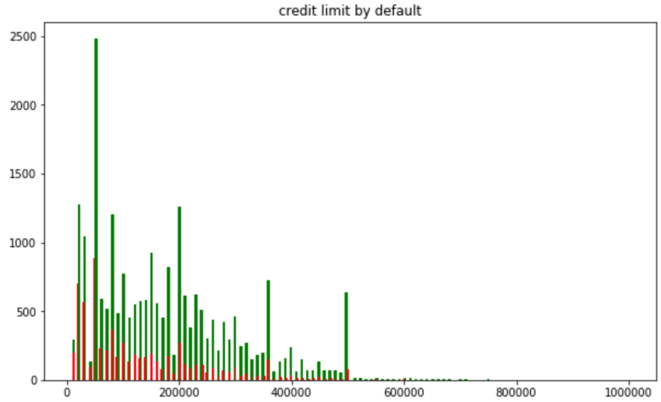
C. *Credit Limit*

A credit card credit limit is the maximum outstanding balance allowed on a credit card in a given amount of time. For this dataset, the credit limit should span month. Credit limit is a crucial factor that could be associated with default risk, alongside demographic factors.



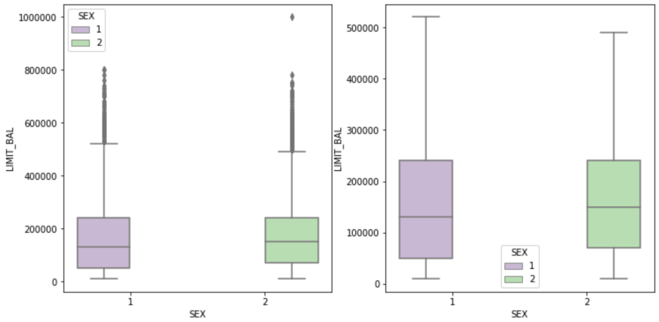
Credit limits for this dataset are skewed to the right, meaning that most consumers have relatively lower credit limits.

Credit limit by default:



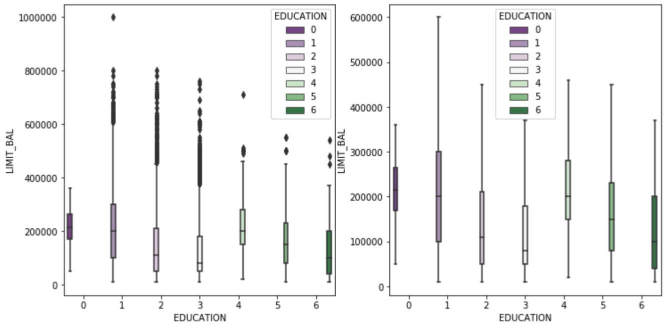
After separating by default, it does not seem that there is any significant difference in credit limits for those who do or do not default payment in the next month.

Credit limit versus gender:

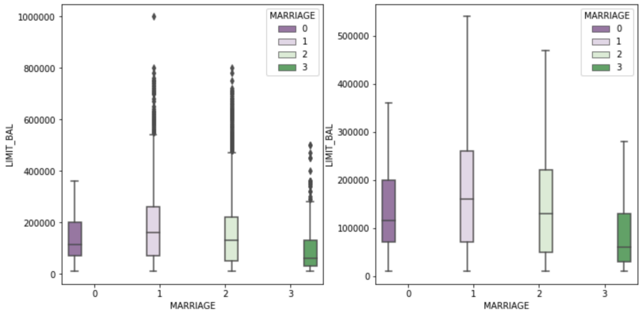


It seems that females have a slightly higher credit limits.

Credit limit versus education:



There is no clear relationship between levels of education and credit limit.



It seems that married individuals have higher credit limits.

III. Data Splitting

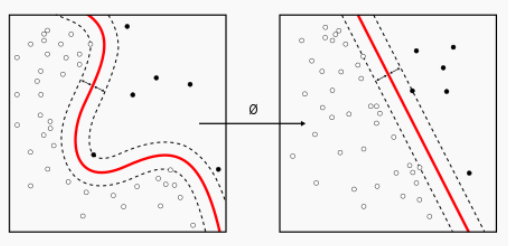
The data is split into training, validation, and testing sets. The training set is used to fit the model; The validation set is used to evaluate model fit; The testing set is used to provide an unbiased evaluation of the model fit. There are two methods of splitting:

1. The first method of splitting is first split into training and testing sets, with the testing set containing 30 percent of data. The training set is then further split into training and validation sets, with the validation set containing 30 percent of the original training set. The training, validation, and testing sets contain 14700, 6300, and 9000 rows, respectively.
2. The second method of splitting is first split into training and testing sets, with the testing set containing 15 percent of data. The training set is then further split into training and validation sets, with the validation set containing 15 percent of the original training set. The training, validation, and testing sets contain 21675, 3825, and 4500 rows, respectively.

The training and validation sets are compared to reduce over-fitting, if it is existent.

IV. Method 1: K-Nearest Neighbors

The K-nearest neighbors (KNN) method is a simple supervised method which classifies samples based on a plurality vote of n-number of neighbors.



(The figure is not representative of the dataset.)

We begin with n=5, for 5 neighbors. We then increase and decrease the number of neighbors to find the most suitable number of neighbors. It seems that for the most constant and optimal training, validation, and testing scores, our model should implement 9 neighbors. The result of the test is as follows:

for n = 9:
Training set score: 0.7979591836734694
Validation set score: 0.7614285714285715
Testing set score: 0.7708888888888888

The confusion matrix for the KNN method is represented by:

	$\hat{y} = 0$	$\hat{y} = 1$
$y = 0$	6636	424
$y = 1$	1638	302

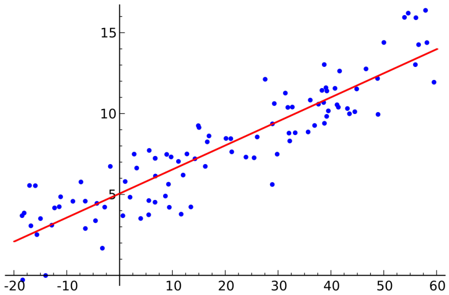
accuracy: 0.7708888888888888
precision class 0: 0.8020304568527918
precision class 1: 0.41597796143250687
recall class 0: 0.939943342776204
recall class 1: 0.1556701030927835

V. Method 2: Linear models

We attempt linear models to classify default payment in the next month.

A. Linear Regression

The Linear Regression model assumes a linear relationship between the features and the target value.



(The figure is not representative of the dataset.)

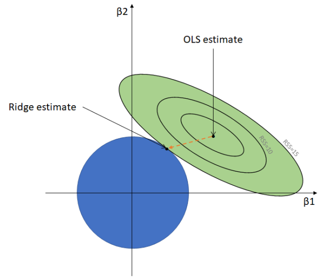
The result of the linear regression model gives the following results:

Training set score: 0.12092337391942576
Validation set score: 0.126314507368683
Testing set score: 0.12129366731987268

The target value is binary (0 or 1), so the Linear Regression model is not quite efficient.

B. Ridge Regression

The Ridge Regression model assumes a linear relationship between the features and the target value. There is possibly multicollinearity between the feature variables.



(The figure is not representative of the dataset.)

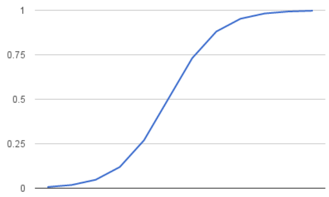
The result of the ridge regression model gives the following results:

Training set score: 0.12092337349629922
Validation set score: 0.1263141550243242
Testing set score: 0.12129283924333345

The target value is binary (0 or 1), so the Ridge Regression model is not quite efficient as well. Moreover, the dataset is mostly normalized due to being categorical values.

C. Logistic Regression

The Logistic Regression model is used for classification problems based on probability concepts. It limits the cost function between 0 and 1.



(The figure is not representative of the dataset.)

The result of the logistic regression model gives the following results:

Training set score: 0.7756462585034014
Validation set score: 0.7777777777777778
Testing set score: 0.7845555555555556

The confusion matrix for the Logistic Regression method is represented by:

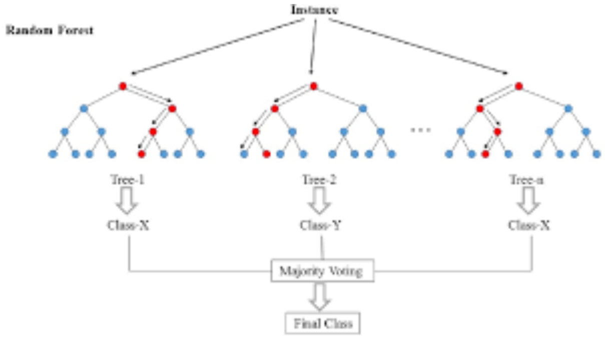
	$\hat{y} = 0$	$\hat{y} = 1$
$y = 0$	7060	0
$y = 1$	1939	1

accuracy: 0.7845555555555556

precision class 0: 0.7845316146238471
precision class 1: 1.0
recall class 0: 1.0
recall class 1: 0.0005154639175257732

VI. Method 3: Random Forest

The Random Forest Classifier is a set of decision trees from randomly selected subsets of training sets. For our model, we set the number of trees in the forest to 100.



(The figure is not representative of the dataset.)

The result of the random forest model gives the following results:

Training set score: 0.9998639455782313
Validation set score: 0.8177777777777778
Testing set score: 0.8187777777777778

To prevent overfitting issues, we use the second method of splitting the training, validation, and testing sets, which is mentioned earlier the paper. The result of the random forest model with adjusted training, validation, and testing sets, gives the following results:

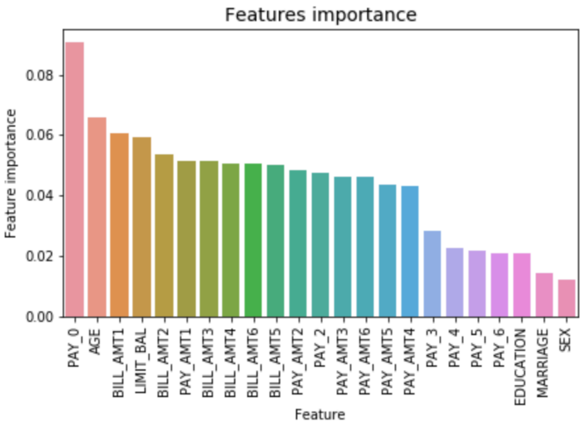
Training set score: 0.9994925028835063
Validation set score: 0.8177777777777778
Testing set score: 0.8215555555555556

The confusion matrix for the Random Forest method is represented by:

	$\hat{y} = 0$	$\hat{y} = 1$
$y = 0$	3343	186
$y = 1$	617	354

accuracy: 0.8215555555555556
precision class 0: 0.8441919191919192
precision class 1: 0.6555555555555556
recall class 0: 0.9472938509492774
recall class 1: 0.364572605561277

Lastly, we sort the features by their relevance. In order of importance, the most important features for predicting default next month are: repayment status in the first month, age, and the bill amount in the first month.



VII. Conclusion

Of the three methods, the Random Forest Classifier is the most accurate method. Since we determined the most important factors, future research could reduce the number of features implemented in the model. Moreover, this dataset may not be applicable to current cases since it is drawn from financial data from Taiwan in 2005.

References

[1] https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html

[2] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

[3] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[4] https://www.kaggle.com/gpreda/default-of-credit-card-clients-predictive-models