SCSD 4009 Project Report

Credit Card Approval Prediction

Chan Yuk Yee 56230549

Leung Tsan Lim 56211178

Table of Context

Abstract

Credit card is a widely accepted method of payment. There is a lot of risk management system for credit cards in commercial banks and financial institutions. However, manually processing applications for credit cards is error-prone and a waste of time. In this paper, we use individual information and data submitted by credit card applicants to predict the possibility of future default. Before that, we conduct exploratory data analysis to explore the discovery of data. Then, we analyzed and compared the data with several supervised machine-learning algorithms. In addition, the most important features that determine whether a customer will become a 'bad' client are extracted from all features. Furthermore, the best-performing algorithm was chosen for predicting credit card approval. The algorithm gave around 90% accuracy in prediction.

1. Introduction

Credit card is a widely accepted method of payment. There is a lot of risk management system for credit cards in commercial banks and financial institutions. However, manually processing applications for credit cards is error-prone and a waste of time. The staff will spend too much time and work inefficiently. In order to increase productivity and efficiency, credit card approval predictions in machine learning are very valuable.

With the advancement of technology and digitalization coming in the 20s, the ability of machines to perform the predictive analysis is very significant in this era of big data. Machine learning is widely used in a lot of applications. For instance, financial institutions and commercial banks sometimes face the challenge of which risk factors to consider when providing credit to customers. To lower the risk of advancing credit to customers, machine learning tools should be used to identify the low-risk customer rather than analysis the person with human feelings. In this paper, we use individual information and data submitted by credit card applicants to predict the possibility of future default. Before that, we conduct exploratory data analysis to explore the discovery of data. Then, we analyzed and compared the data with several supervised machine-learning algorithms. In addition, the most important features that determine whether a customer will become a 'bad' client are extracted from all features. Furthermore, the best-performing algorithm was chosen for predicting credit card approval. The algorithm gave around 90% accuracy in prediction.

After the completion of the project, a finding of the detection of the credit risk of the customer by machine learning was reported in the following.

1.1 Introduction to dataset

Figure 1. Application record dataset

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	ID	CODE_GEI	FLAG_OW	FLAG_OW	CNT_CHIL	AMT_INC(	NAME_IN(	NAME_ED	NAME_FA	NAME_HC	DAYS_BIR	DAYS_EMI	FLAG_MO	FLAG_WO	FLAG_PHC	FLAG_EM/	OCCUPATI	CNT_FAM_	MEMBERS
2	5008804	M	Y	Y	0	427500	Working	Higher ed	Civil marr	Rented ap	-12005	-4542	1	1	0	0		2	
3	5008805	M	Y	Y	0	427500	Working	Higher ed	Civil marr	Rented ap	-12005	-4542	1	1	0	0		2	
4	5008806	M	Y	Y	0	112500	Working	Secondar	Married	House / a	-21474	-1134	1	0	0	0	Security s	2	
5	5008808	F	N	Y	0	270000	Commerci	Secondar	Single / n	House / a	-19110	-3051	1	0	1	1	Sales staf	1	
6	5008809	F	N	Y	0	270000	Commerci	Secondar	Single / n	House / a	-19110	-3051	1	0	1	1	Sales staf	1	
7	5008810	F	N	Y	0	270000	Commerci	Secondar	Single / n	House / a	-19110	-3051	1	0	1	1	Sales staf	1	
8	5008811	F	N	Y	0	270000	Commerci	Secondar	Single / n	House / a	-19110	-3051	1	0	1	1	Sales staf	1	
9	5008812	F	N	Y	0	283500	Pensioner	Higher ed	Separatec	House / a	-22464	365243	1	0	0	0		1	
10	5008813	F	N	Y	0	283500	Pensioner	Higher ed	Separatec	House / a	-22464	365243	1	0	0	0		1	

Figure 2. Monthly credit card account status dataset

	A	B	C
1	ID	MONTHS_	STATUS
2	5001711	0	X
3	5001711	-1	0
4	5001711	-2	0
5	5001711	-3	0
6	5001712	0	C
7	5001712	-1	C
8	5001712	-2	C
9	5001712	-3	C
10	5001712	-4	C

The data was collected from the Kaggle Platform[1]. Figure 1 showed the credit card application records and it contains around 430,000 rows & 18 features. Figure 2 show monthly credit card account status respectively and it contains around 1,000,000 rows & 3 features.

1.2 Challenge

In this project, there are the following challenges:

- The dataset is highly imbalanced.
- The definition of 'good' or 'bad' is not given. The target must be defined.

1.3 Objective

Our achievement in this project is to:

- Determine the importance features that affect the credit card approval result.
- Use predictors to predict the credit card approval result with high accuracy

2. Methods

2.1 Data Processing

2.1.1 Data Integration – defining target and merging

In the dataset, due payment days can be observed. Assume that a person is considered a 'Bad' client if he has unpaid payments for more than 60 days. Status is defined by the table 1.

Table 1. The definition of "Status" variable

Status	Meaning
0	1-29 days past due
1	30-59 days past due
2	60-89 days overdue
3	90-119 days overdue
4	120-149 days overdue
5	Overdue or bad debts, write-offs for more than 150 days
C	Paid off that month
X	No loan for the month

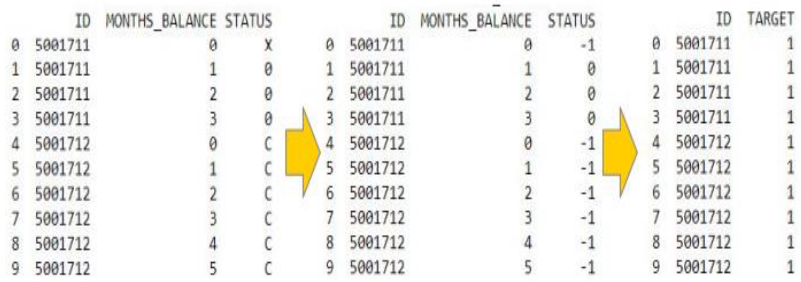
[1] Credit card approval prediction. Kaggle. (n.d.). Retrieved November 22, 2021, from https://www.kaggle.com/rikdifos/credit-card-approval-prediction/metadata.

The target is determined by below logic:

Target definition

procedure TargetDefinition
    for i in id:
        if status(i) >= 2:
            target(i) ← 0 (It means a person is considered to be a bad client)
        else:
            target(i) ← 1 (It means a person is considered to be a good client)

Figure 3. The flow of target definition



	ID	MONTHS_BALANCE	STATUS
0	5001711	0	X
1	5001711	1	0
2	5001711	2	0
3	5001711	3	0
4	5001712	0	C
5	5001712	1	C
6	5001712	2	C
7	5001712	3	C
8	5001712	4	C
9	5001712	5	C

	ID	MONTHS_BALANCE	STATUS
0	5001711	0	-1
1	5001711	1	0
2	5001711	2	0
3	5001711	3	0
4	5001712	0	-1
5	5001712	1	-1
6	5001712	2	-1
7	5001712	3	-1
8	5001712	4	-1
9	5001712	5	-1

	ID	TARGET
0	5001711	1
1	5001711	1
2	5001711	1
3	5001711	1
4	5001712	1
5	5001712	1
6	5001712	1
7	5001712	1
8	5001712	1
9	5001712	1

2.1.2 Data Cleaning

a. Remove duplicate

The duplicate data will be deleted. After duplicate removal by using pandas (pd.drop_duplicates(...)), the data left 32713 rows.

b. Filter unwanted outliers

99.7 rules applied in CNT_CHILDREN, AMT_INCOME_TOTAL, CNT_FAM_MEMBERS since they are continuous features.

c. Handle missing value

One variable has 14% missing value, and that is categorical feature. The solution is to ignore the missing value.

2.1.3 Data Wrangling

- One hot encoder is used to divide the Categorical data into different column datasets.
- Train-test Split are applied in 7:3 rate to avoid overfitting.
- SMOTE is used to handle imbalanced data. Note that Applying oversampling after train-test split because the observations from the minority class in the training dataset might end up in the testing dataset. This is in a way allows the algorithm to cheat since it learned from something similar.

2.2 Exploratory data analysis

2.2.1 Findings of data visualization

Figure 4. The distribution of different variables

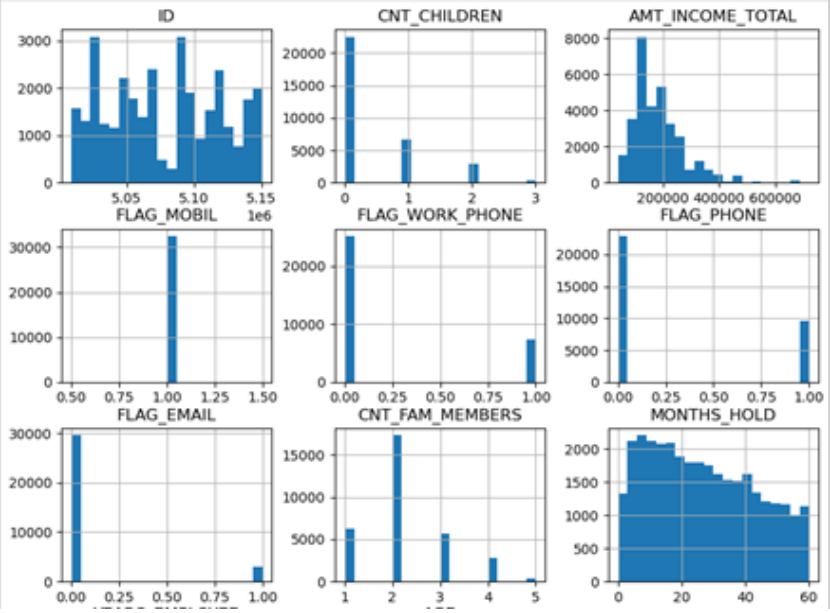
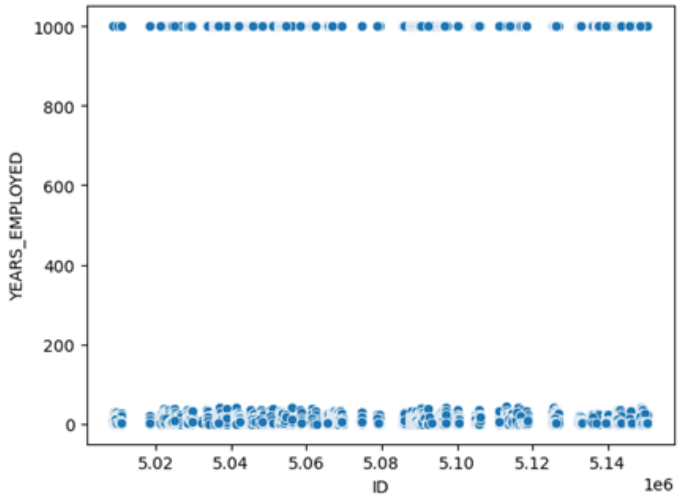


Figure 5. The distribution of YEARS_EMPLOYED



1. Figure 4 showed that FLAG_MOBIL is useless data since the data record that all people have mobile phone. This column needs to be drop.

2. Figure 5 showed that some people have around 1000 YEARS_EMPLOYED. After checking, we found that they are both pensioners. Assume that most people around 40-45 YEARS_EMPLOYED are going to become pensioners. ➔ Solution: transform 1000 to [40,45] uniform randomly.

3. MONTHS_HOLD should not be a predictor since it refers to the month the user holds the credit card. It may lead to overfitting. This column also needs to be dropped.

4. If a person is at a work, the probability of being a good client will be higher than the person who is unemployed.

Figure 6: The distribution of MONTHS_HOLD

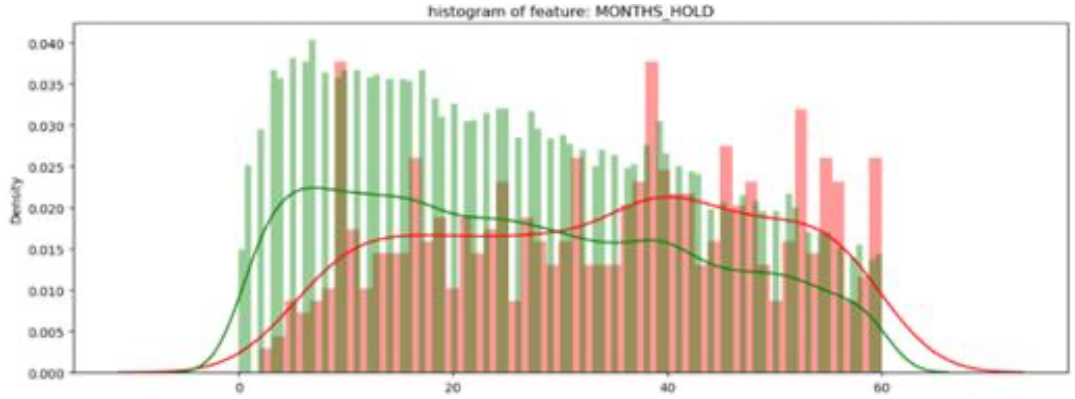


Figure 7: The distribution of AGE

2.2.2 Analysis importance features

Information value (IV) is measured in the powers among the predictors, which determine the important variables in predictive model. It ranks variables on the basis of the importance[2]. Figure 8 shows the result.

Figure 8: Iv value of features

	VAR_NAME	IV
0	AGE	2.085125
15	YEARS_EMPLOYED	1.060689
14	OCCUPATION_TYPE	0.052802
11	NAME_FAMILY_STATUS	0.030403
7	FLAG_OWN_REALTY	0.029616
12	NAME_HOUSING_TYPE	0.017450
13	NAME_INCOME_TYPE	0.016912
4	CODE_GENDER	0.011656
10	NAME_EDUCATION_TYPE	0.008853
6	FLAG_OWN_CAR	0.001867
9	FLAG_WORK_PHONE	0.001840
3	CNT_FAM_MEMBERS	0.000803
5	FLAG_EMAIL	0.000459
1	AMT_INCOME_TOTAL	0.000417
8	FLAG_PHONE	0.000227
2	CNT_CHILDREN	0.000067

Table 2: Rules related to Information Value

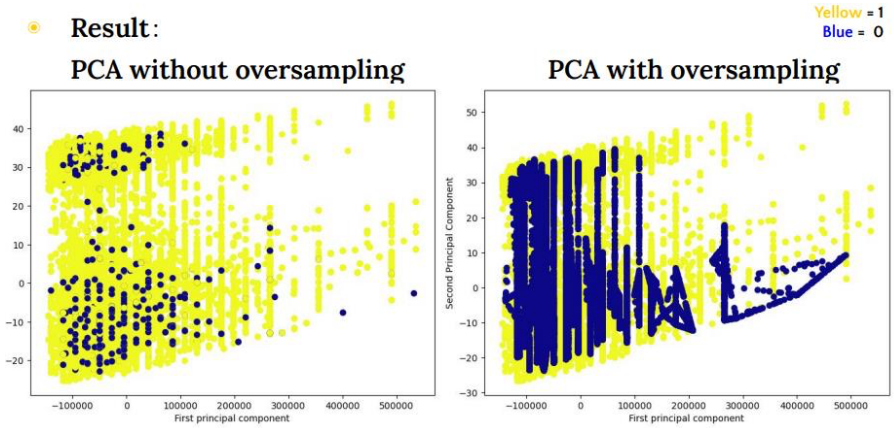
Information Value	Predictive Power
< 0.02	Useless
0.02 - 0.1	Weak
0.1 - 0.3	Medium
0.3 - 0.5	Strong
> 0.5	Suspiciously good; too good to be true

Table 2 showed that AGE and YEARS_EMPLOYED are top 2 importance features.

2.2.3 PCA

Unsupervised learning is applied to test does PCA fit the data well. Figure 9 shows the result.

Figure 9: PCA result



[2] Bhalla, D. (n.d.). Weight of evidence (WOE) and information value (iv) explained. ListenData. Retrieved November 22, 2021, from https://www.listendata.com/2015/03/weight-of-evidence-woe-and-information.html.

Although unsupervised learning could be applied to the dataset, a clear classification point cannot be found in the PCA graph, it does not result in a clear classification area.

Therefore, as the datasets consist of a certain number of predictors, supervised machine learning should be used.

3. Results

3.1 Logistic Regression

Information Value is designed mainly for the binary logistic regression models[3]. Features' IV value < 0.01 is decided to remove. Therefore, the remaining features will be used in the Logistic Regression model. AGE, YEARS_EMPLOYED, OCCUPATION_TYPE, NAME_FAMILY_STATUS, FLAG_OWN_REAITY, NAME_HOUSING_TYPE, NAME_INCOME_TYPE, CODE_GENDER are included. The result shows Logistic Regression has an Accuracy of 60%.

	precision	recall	f1-score	support
0	0.60	0.63	0.61	9557
1	0.61	0.58	0.59	9557
accuracy			0.60	19114
macro avg	0.60	0.60	0.60	19114
weighted avg	0.60	0.60	0.60	19114

3.1 KNN

Feature selection via IV value and using them in KNN and XGBoost model might not produce the most accuracy. IV value should not be used in KNN and XGBoost models. The result shows KNN has an Accuracy of 72%.

	precision	recall	f1-score	support
0	0.87	0.52	0.65	9557
1	0.66	0.93	0.77	9557
accuracy			0.72	19114
macro avg	0.77	0.72	0.71	19114
weighted avg	0.77	0.72	0.71	19114

*** the correlation graph between parameter (n) and the error rate will be tested in validation part.*

---

[3] Bhalla, D. (n.d.). Weight of evidence (WOE) and information value (iv) explained. ListenData. Retrieved November 22, 2021, from https://www.listendata.com/2015/03/weight-of-evidence-woe-and-information.html.

3.1 XGBoost

The full name of XGboost is eXtreme Gradient Boosting. It is currently the most common algorithm in Kaggle competitions, and it is also the model used by most winners. Result shows XGBoost has Accuracy with 92%.

	precision	recall	f1-score	support
0	0.96	0.87	0.91	9557
1	0.88	0.97	0.92	9557
accuracy			0.92	19114
macro avg	0.92	0.92	0.92	19114
weighted avg	0.92	0.92	0.92	19114

4. Results

4.1 5-fold cross validation

After performing 5-fold cross-validation, accuracy rate has changed.

Logistic Regression: 60%→58% (slightly decrease)

KNN: 70%→94% (considerably increase)

XGBoost: 90%→94% (slightly increase)

4.2 KNN parameters error rate

Figure 10: The graph of KNN error rate

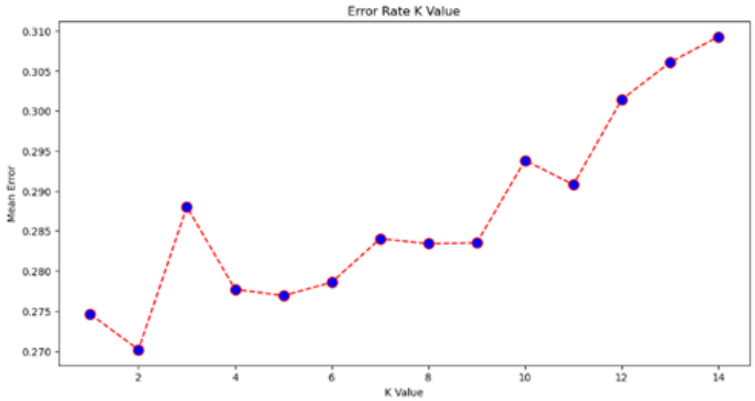


Figure 10 showed that k values bigger, Error Rate bigger.

4.3 5-fold cross validation with SMOTE

Using SMOTE to avoid inaccurate evaluation metrics when using 5-fold cross-validation. After performing 5-fold cross-validation with SMOTE, accuracy rate has changed.

Performance (simple train-test-split v.s. 5-fold with SMOTE):

Logistic Regression: 60%→58% (slightly decrease)
KNN: 70%→89% (considerably increase)
XGBoost: 90%→93% (slightly increase)

4.4 Comparing accuracy

Table 3. Accuracy comparison

Accuracy	Simple train-test	5-fold	5-fold with SMOTE
Logistics Regression	60%	58%	58%
KNN	70%	94%	89%
XGBoost	90%	94%	93%

The performance of the simple train-test spilt is the poorest. The performance of the KNN and XGBoost models via 5-fold cross-validation with SMOTE is lightly decreased compared to the models via 5-fold cross-validation.

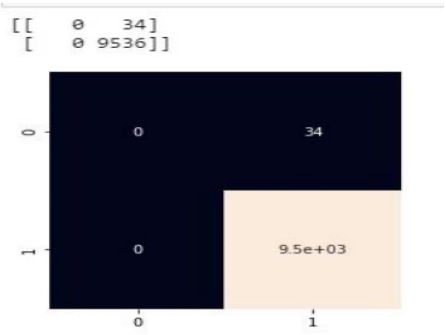
5. Discussion

5.1 Further discussion on balanced data

Unbalanced data is mentioned at the report's introduction, and this report used SMOTE to simulate target "0" data. The effect of imbalanced data on model training and testing is also examined in this research.

The accuracy of logistic regression is 99% when imbalanced data is fitted into the model after the train-test split, seems that it is a perfect predicting model in this situation. However, Figure 11 is shown, 9536 of the predicted data are accurate and 34 are inaccurate. There are 34 data with "0" labeled in the testing data since the matrix predicts all 1 in the predicted result.

Figure 11. Confusion Matrix for Imbalanced Data Classification



This showed that if extremely unbalanced data is fitted to the data set, the training data set and the test data set continue to be unbalanced, causing it to be meaningless modeling with high accuracy. Therefore, data balance should be ensured before the modeling process.

5.2 Further discussion on cross-validation for imbalanced datasets

5-fold cross-validation (performing oversampling before executing cross-validation) and 5- fold cross-validation with SMOTE (performing oversampling during cross-validation) are both adopted for model validation in this paper. We found that performing oversampling before executing cross-validation seems "efficient". However, the fact is that we are overestimating the classification performance while adopting this method[4]. Therefore, performing oversampling during cross-validation is the correct way of handling imbalanced data.

6. Conclusion

In conclusion, AGE and YEAR_EMPLOYED are the important features in the predictors, which means they are easily interpreted the credit card approval result. On the other hand, XGBoost has the highest performance with 90-95% accuracy. Therefore, XGBoost should be applied to the business situation in order to optimize the prediction result.

---

[4] Santos, M. S., Soares, J. P., Abreu, P. H., Araujo, H., & Santos, J. (2018). Cross-validation for imbalanced datasets: Avoiding overoptimistic and overfitting approaches [research frontier]. IEEE Computational Intelligence Magazine, 13(4), 59–76. https://doi.org/10.1109/mci.2018.2866730

References

1. Credit card approval prediction. Kaggle. (n.d.). Retrieved November 22, 2021, from https://www.kaggle.com/rikdifos/credit-card-approval-prediction/metadata
2. Bhalla, D. (n.d.). Weight of evidence (WOE) and information value (iv) explained. ListenData. Retrieved November 22, 2021, from https://www.listendata.com/2015/03/weight- of-evidence-woe-and-information.html
3. Santos, M. S., Soares, J. P., Abreu, P. H., Araujo, H., & Santos, J. (2018). Cross- validation for imbalanced datasets: Avoiding overoptimistic and overfitting approaches. IEEE Computational Intelligence Magazine, 13(4), 59–76. https://doi.org/10.1109/mci.2018.2866730