Credit Card Defaulting in Taiwan & Related Observations and Analyses

By: Nachiket Kulkarni

TABLE OF CONTENTS

I. ABSTRACT AND PURPOSE

The following paper studies the dataset containing numerical and categorical information pertaining to the credit card crisis of Taiwan. Every feature column outlines specific population-based data and financial data, and these factors collectively determine the outcome of whether a credit card holder will default on a payment for the coming month. The primary purpose of this paper was to determine which variables of the given dataset were the strongest predictors of default payment. The tests and models that were used to determine these variables were Random Forest Classification, Naive Bayes and Gradient Boosting. It is suggested that gradient boosting model has an advantage in determining the variables that were used and whether the discarding of other specific variables was justified.

II. INTRODUCTION

In the final decade of the twentieth century, the government of Taiwan allowed the formation of new banks.[1] With excessive lending to real estate businesses, the market became saturated and profits did not increase.[1] Banks then focused on credit cards as another source of revenue, easing restrictions and allowing low or non earners, such as the youth, to receive these cards.[1] By the year 2006, credit card debt was $268 billion in U.S. dollars.[1] Many people were only able to pay the monthly minimum on their card balances, thus making them “credit card slaves”[1], with 700,000 citizens earning the unfortunate term.[2] Many credit card holders eventually defaulted, in some cases finding themselves bankrupt and homeless. The result was that families were unable to pay off debts accrued through extensive credit card usage, leading to major increases in crime, anxiety and even depression.[1]

In this paper, the dataset, the factors and variables that influence whether defaulting occur are studied, with an overview and analysis of data and conclusions being drawn regarding the most crucial variables indicating defaulting.

III. DATA OVERVIEW

In the given dataset, there are 30000 rows of data, with each row representing a credit card holder. Twenty-three feature columns are present, labeled as such:

'X1','X2','X3','X4','X5','X6','X7','X8','X9','X10','X11','X12','X13','X14','X15','X16','X17','X18', 'X19','X20','X21','X22','X23'.

The target column, labeled 'Y'. The descriptions of the variables are outlined in the following table.

LIMIT_BAL = X1	SEX = X2	EDUCATION = X3	MARRIAGE = X4	AGE = X5
PAY_0 = X6	PAY_2 = X7	PAY_3 = X8	PAY_4 = X9	PAY_5 = X10
PAY_6 = X11	BILL_AMT1 = X12	BILL_AMT2 = X13	BILL_AMT3= X14	BILL_AMT4=X15
BILL_AMT5 = X16	BILL_AMT6 = X17	PAY_AMT1 = X18	PAY_AMT2 = X19	PAY_AMT3 = X20
PAY_AMT4 = X21	PAY_AMT5 = X22	PAY_AMT6 = X23	default payment next month = Y	

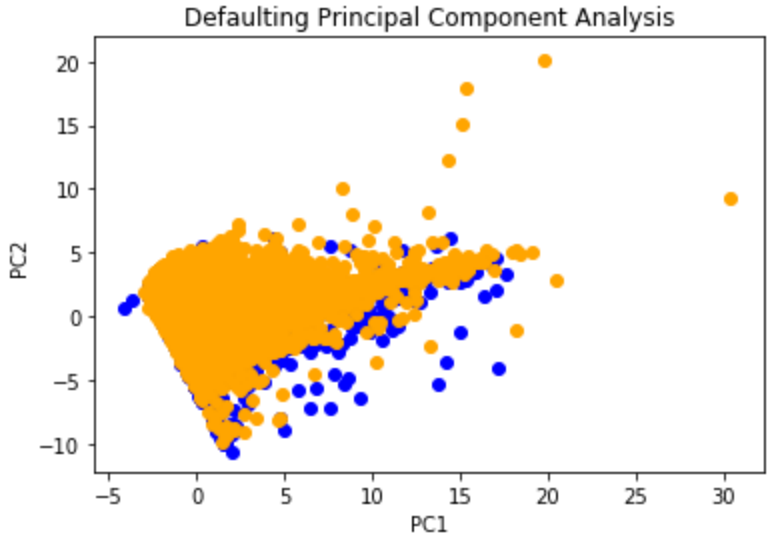
Furthermore, the description of the vital statistics of the data columns is displayed in the APPENDIX section of this document.

IV. METHODOLOGY

The data was cleaned and prepared through comparing the vital statistics of each predictor variable in the dataset using exploratory data analysis. Correlation charts were made for the data before and after the removal of data determined to be irrelevant. For each testing algorithm employed in this study, accuracy, precision, and recall (the ability to find relevant data to work with)[4], were observed in scenarios where each predictor variable was withheld from being tested against the dependent variable "Y". The AUC values of the ROC Curves for the models were also taken into consideration. For each variable withheld, the three aforementioned metrics were calculated for the tested data. Through comparing these metrics with each other, it was concluded that the variables of "PAY_2", "PAY_4", "PAY_AMT3", "PAY_AMT4", and "PAY_AMT6" decreased the accuracy of the machine learning models. All other variables, at the time of testing, appeared to be relevant to the outcome of possible credit card defaulting. For the specific data regarding the accuracy, precision and recall metrics for each test that withheld a variable, please refer to the data of "Relevant Metrics Data of the Data Withholding Specific Predictor Variables" in the APPENDIX section.

To verify the relevance of the remaining variables, K-Means clustering was employed. With K-Means clustering, the existence of clusters suggests that the data utilized for the clustering mechanism share similarities with each other, thus deeming the data necessary to use for future analysis. A K-Means Cluster plot comparing the "LIMIT_BAL" and the "AGE" columns and an accompanying elbow plot verifying the cluster selection are available in the APPENDIX section. The cluster value is equivalent to '3'.

Furthermore, Principal Component Analysis was also performed to ensure that the removal of the aforementioned columns was necessary. This dimension reduction facilitates the streamlining of the data, enabling concentration on essential information provided by the dataset. The resulting Principal Component Analysis chart was rendered by comparing the status of defaulting payment to the next month with the remaining data columns. The strong clustering demonstrated in the chart indicates that the feature columns do indeed support the trend of the target variable, and the chart does not suffer from the unnecessary columns. The PCA chart is shown below.



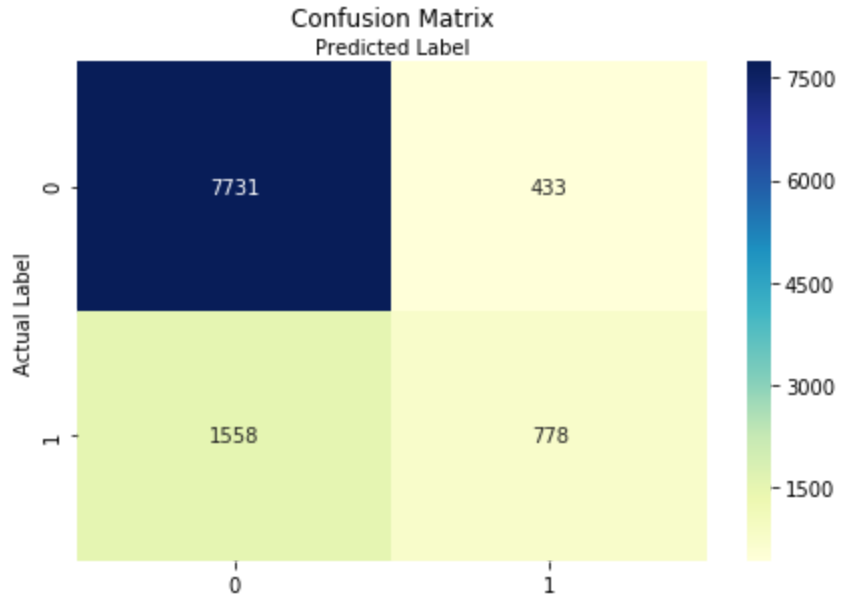
V. MODELING PERFORMANCE MEASURES

Random Forest

Two modeling algorithms were used in this study; random forest classification and gradient boosting. Random Forest Classification utilizes multiple decision trees based on a set amount of estimators, and calculates the average decision tree based on all collected tree results. For this dataset, the number of estimators was twenty, meaning that the random forest classification model created twenty different decision trees and created an average decision tree using their results. Using the selected predictor variables and the target variable, the accuracy, precision and recall were measured. The confusion matrix for the model was also rendered, along with the ROC Curve of the model.

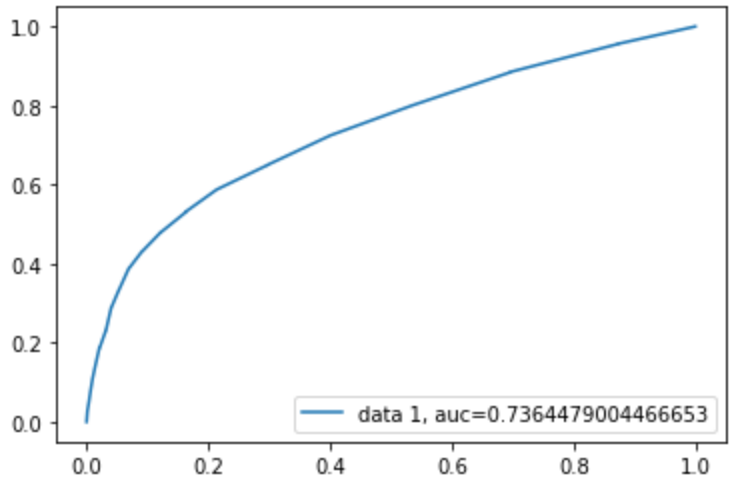
Accuracy	Precision	Recall
0.8103809523809524	0.6424442609413707	0.3330479452054795

Random Forest Confusion Matrix



True Positive: 778; False Positive: 433; True Negative: 7731; False Negative: 1558

Random Forest ROC Curve

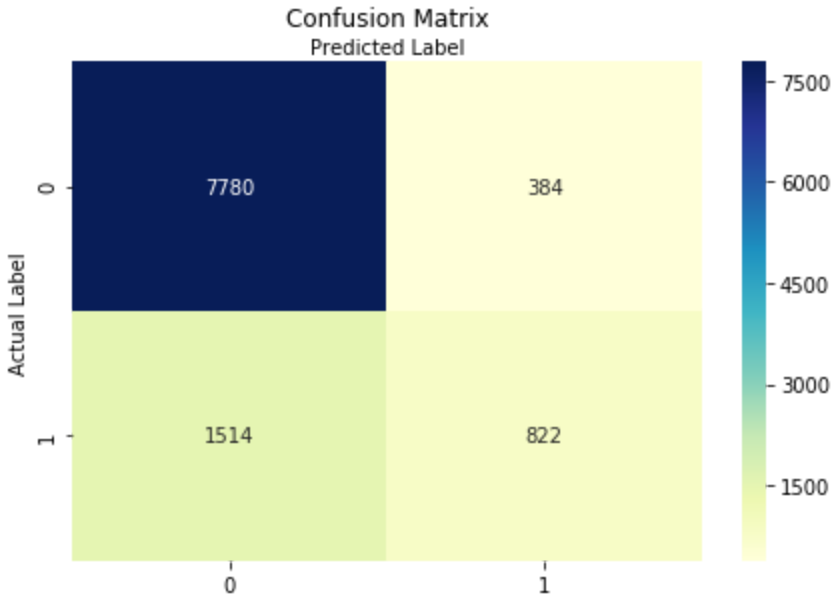


Gradient Boosting

Gradient Boosting is a classification model that also uses decision trees, but produces a tree for every version of a dataset that is being modified by the model.[3] It then adds together and creates a composite result for predictions of categorical labels. Using the selected predictor variables and the target variable, the accuracy, precision and recall were measured. The confusion matrix for the model was also rendered, along with the ROC Curve of the model.

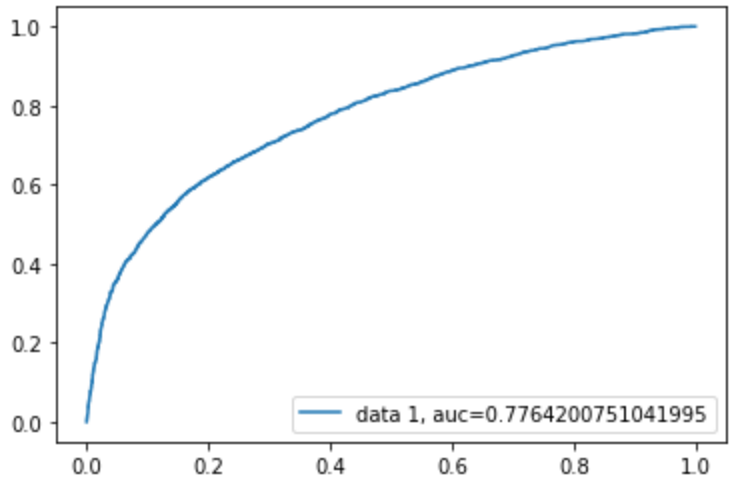
Accuracy	Precision	Recall
0.8192380952380952	0.681592039800995	0.3518835616438356

Gradient Boosting Confusion Matrix



True Positive: 822; False Positive: 384; True Negative: 7780; False Negative: 1514

Gradient Boosting ROC Curve

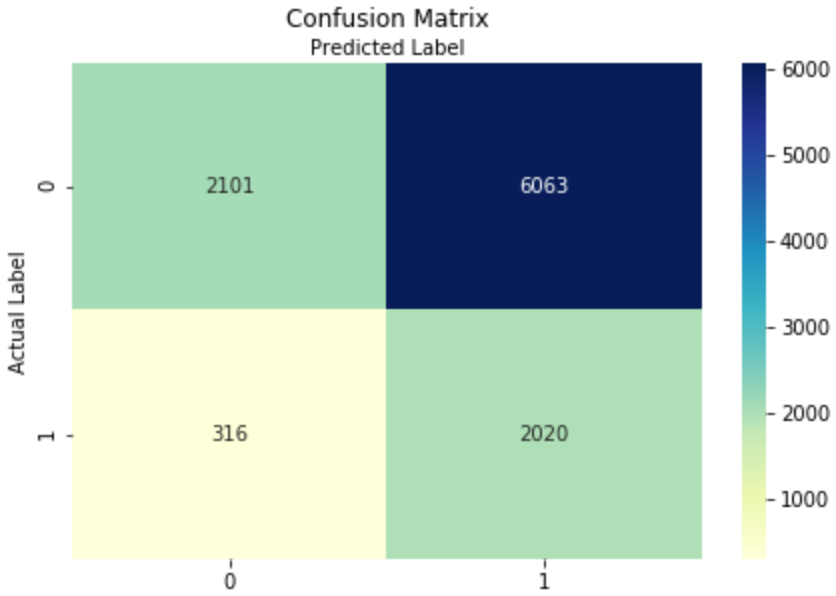


Naive Bayes Classification

Naive Bayes is a classification model which utilizes the Bayes Theorem.[5] This theorem functions on the conjecture that the attributes used in the analysis are not dependent on each other; subsequently, the theorem presents a probability of a certain event occurring, as long as another event was already executed.[5] Using the selected predictor variables and the target variable, the accuracy, precision and recall were measured. The confusion matrix for the model was also rendered, along with the ROC Curve of the model.

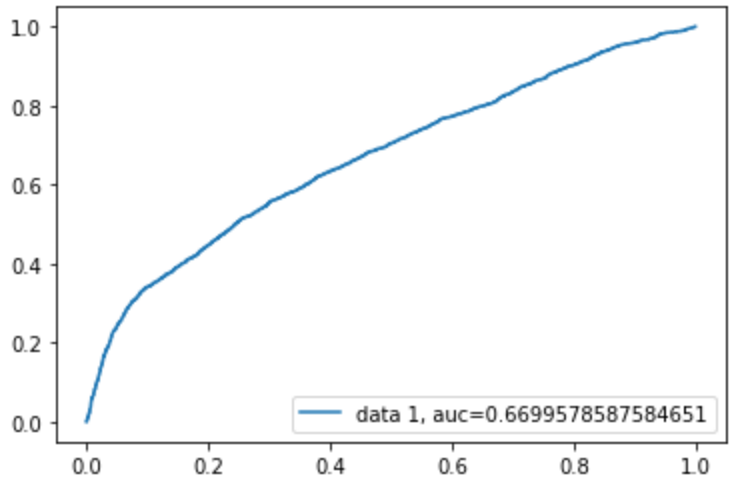
Accuracy	Precision	Recall
0.3924761904761905	0.24990721266856364	0.8647260273972602

Naive Bayes Confusion Matrix



True Positive: 2020; False Positive: 6063; True Negative: 2101; False Negative: 316

Naive Bayes ROC Curve



VI. CONCLUSIONS

The final observations and outcomes of this analysis strongly suggest that the following variables are essential towards determining credit card default in Taiwan:

LIMIT_BAL, SEX, EDUCATION, MARRIAGE, PAY_0, PAY_3, PAY_4, PAY_5, PAY_6, BILL_AMT1, BILL_AMT2, BILL_AMT3, BILL_AMT4, BILL_AMT5, BILL_AMT6, PAY_AMT1, PAY_AMT2, and PAY_AMT5.

The modeling mechanisms yielded varied results, with one performing better than the others. Observing the results of the metrics and the AUC values of the models, it is suggested that the gradient boosting model outperforms the random forest model and the Naive Bayes model in determining which variables of the dataset are important in determining credit card default. Due to its larger values for accuracy, precision and the AUC value of the corresponding ROC curve, the gradient boosting model appears to be the superior model. Not only does the model predict defaulting more accurately and reliably, it also justifies the discarding of the variables that were discarded more strongly than both the random forest classification model and the Naive Bayes model.

VII. APPENDIX

Vital Statistics

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
count	30000.000000	30000.000000	30000.000000	30000.000000	30000.000000	30000.000000	30000.000000	30000.000000	30000.000000	30000.000000
mean	167484.322667	1.603733	1.853133	1.551867	35.485500	-0.016700	-0.133767	-0.166200	-0.220667	-0.266200
std	129747.661567	0.489129	0.790349	0.521970	9.217904	1.123802	1.197186	1.196868	1.169139	1.133187
min	10000.000000	1.000000	0.000000	0.000000	21.000000	-2.000000	-2.000000	-2.000000	-2.000000	-2.000000
25%	50000.000000	1.000000	1.000000	1.000000	28.000000	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000
50%	140000.000000	2.000000	2.000000	2.000000	34.000000	0.000000	0.000000	0.000000	0.000000	0.000000
75%	240000.000000	2.000000	2.000000	2.000000	41.000000	0.000000	0.000000	0.000000	0.000000	0.000000
max	1000000.000000	2.000000	6.000000	3.000000	79.000000	8.000000	8.000000	8.000000	8.000000	8.000000

	X11	X12	X13	X14
count	30000.000000	30000.000000	30000.000000	3.000000e+04
mean	-0.291100	51223.330900	49179.075167	4.701315e+04
std	1.149988	73635.860576	71173.768783	6.934939e+04
min	-2.000000	-165580.000000	-69777.000000	-1.572640e+05
25%	-1.000000	3558.750000	2984.750000	2.666250e+03
50%	0.000000	22381.500000	21200.000000	2.008850e+04
75%	0.000000	67091.000000	64006.250000	6.016475e+04
max	8.000000	964511.000000	983931.000000	1.664089e+06

X15	X16	X17	X18	X19	X20	X21	X22	X23	Y
30000.000000	30000.000000	30000.000000	30000.000000	3.000000e+04	30000.00000	30000.000000	30000.000000	30000.000000	30000.000000
43262.948967	40311.400967	38871.760400	5663.580500	5.921163e+03	5225.68150	4826.076867	4799.387633	5215.502567	0.221200
64332.856134	60797.155770	59554.107537	16563.280354	2.304087e+04	17606.96147	15666.159744	15278.305679	17777.465775	0.415062
-170000.000000	-81334.000000	-339603.000000	0.000000	0.000000e+00	0.00000	0.000000	0.000000	0.000000	0.000000
2326.750000	1763.000000	1256.000000	1000.000000	8.330000e+02	390.00000	296.000000	252.500000	117.750000	0.000000
19052.000000	18104.500000	17071.000000	2100.000000	2.009000e+03	1800.00000	1500.000000	1500.000000	1500.000000	0.000000
54506.000000	50190.500000	49198.250000	5006.000000	5.000000e+03	4505.00000	4013.250000	4031.500000	4000.000000	0.000000
891586.000000	927171.000000	961664.000000	873552.000000	1.684259e+06	896040.00000	621000.000000	426529.000000	528666.000000	1.000000

Relevant Metrics Data of the Data Withholding Specific Predictor Variables (Random Forest)

Relevant Data Withholding Variable: X1
Accuracy: 0.8074285714285714
Precision: 0.6218944099378882
Recall: 0.3428938356164384

Relevant Data Withholding Variable: X2
Accuracy: 0.8072380952380952
Precision: 0.6232227488151659
Recall: 0.3377568493150685

Relevant Data Withholding Variable: X3
Accuracy: 0.8100952380952381
Precision: 0.6354992076069731
Recall: 0.3433219178082192

Relevant Data Withholding Variable: X4
Accuracy: 0.8123809523809524
Precision: 0.6473429951690821
Recall: 0.3441780821917808

Relevant Data Withholding Variable: X5
Accuracy: 0.8084761904761905
Precision: 0.6335250616269515
Recall: 0.3300513698630137

Relevant Data Withholding Variable: X6
Accuracy: 0.7927619047619048
Precision: 0.575187969924812
Recall: 0.261986301369863

Relevant Data Withholding Variable: X7
Accuracy: 0.8135238095238095
Precision: 0.6551724137931034
Recall: 0.3416095890410959

Relevant Data Withholding Variable: X8
Accuracy: 0.8108571428571428
Precision: 0.6448675496688742
Recall: 0.3334760273972603
Relevant Data Withholding Variable: X9
Accuracy: 0.8095238095238095
Precision: 0.6414141414141414
Recall: 0.3261986301369863

Relevant Data Withholding Variable: X10
Accuracy: 0.8093333333333333
Precision: 0.6384742951907131
Recall: 0.3296232876712329

Relevant Data Withholding Variable: X11
Accuracy: 0.810952380952381
Precision: 0.6434995911692559
Recall: 0.3369006849315068

Relevant Data Withholding Variable: X12
Accuracy: 0.8098095238095238
Precision: 0.6363636363636364
Recall: 0.3386130136986301

Relevant Data Withholding Variable: X13
Accuracy: 0.8094285714285714
Precision: 0.6362896663954435
Recall: 0.3347602739726027

Relevant Data Withholding Variable: X14
Accuracy: 0.8072380952380952
Precision: 0.6270358306188925
Recall: 0.3296232876712329

Relevant Data Withholding Variable: X15
Accuracy: 0.8080952380952381
Precision: 0.6334164588528678
Recall: 0.3261986301369863

Relevant Data Withholding Variable: X16
Accuracy: 0.8094285714285714
Precision: 0.6374077112387203
Recall: 0.3326198630136986

Relevant Data Withholding Variable: X17
Accuracy: 0.8085714285714286
Precision: 0.632952691680261
Recall: 0.3321917808219178
Relevant Data Withholding Variable: X18
Accuracy: 0.8095238095238095
Precision: 0.6337579617834395
Recall: 0.3407534246575342

Relevant Data Withholding Variable: X19
Accuracy: 0.8095238095238095
Precision: 0.6374795417348609
Recall: 0.3334760273972603

Relevant Data Withholding Variable: X20
Accuracy: 0.8102857142857143
Precision: 0.6402936378466558
Recall: 0.3360445205479452

Relevant Data Withholding Variable: X21
Accuracy: 0.8081904761904762
Precision: 0.6306818181818182
Recall: 0.3326198630136986

Relevant Data Withholding Variable: X22
Accuracy: 0.807047619047619
Precision: 0.6236044657097288
Recall: 0.3347602739726027

Relevant Data Withholding Variable: X23
Accuracy: 0.8079047619047619
Precision: 0.629780309194467
Recall: 0.3313356164383562

Relevant Metrics Data of the Data Withholding Specific Predictor Variables (Gradient Boosting)

Relevant Data Withholding Variable: X1
Accuracy: 0.8191428571428572
Precision: 0.678951678951679
Recall: 0.3548801369863014

Relevant Data Withholding Variable: X2
Accuracy: 0.8178095238095238
Precision: 0.6732186732186732
Recall: 0.3518835616438356

Relevant Data Withholding Variable: X3
Accuracy: 0.8187619047619048
Precision: 0.6773136773136773
Recall: 0.3540239726027397

Relevant Data Withholding Variable: X4
Accuracy: 0.818
Precision: 0.6763485477178424
Recall: 0.3488869863013699

Relevant Data Withholding Variable: X5
Accuracy: 0.817047619047619
Precision: 0.6702214930270713
Recall: 0.3497431506849315

Relevant Data Withholding Variable: X6
Accuracy: 0.8014285714285714
Precision: 0.6217264791464597
Recall: 0.2744006849315068

Relevant Data Withholding Variable: X7
Accuracy: 0.8185714285714286
Precision: 0.6767842493847416
Recall: 0.3531678082191781

Relevant Data Withholding Variable: X8
Accuracy: 0.8180952380952381
Precision: 0.6740196078431373
Recall: 0.3531678082191781

Relevant Data Withholding Variable: X9
Accuracy: 0.8194285714285714
Precision: 0.6833333333333333
Recall: 0.351027397260274

Relevant Data Withholding Variable: X10
Accuracy: 0.8185714285714286
Precision: 0.6773662551440329
Recall: 0.3523116438356164

Relevant Data Withholding Variable: X11
Accuracy: 0.818
Precision: 0.6769358867610324
Recall: 0.3480308219178082

Relevant Data Withholding Variable: X12
Accuracy: 0.8193333333333334
Precision: 0.6794766966475879
Recall: 0.355736301369863

Relevant Data Withholding Variable: X13
Accuracy: 0.8182857142857143
Precision: 0.6745513866231647
Recall: 0.3540239726027397

Relevant Data Withholding Variable: X14
Accuracy: 0.8187619047619048
Precision: 0.6776045939294504
Recall: 0.3535958904109589

Relevant Data Withholding Variable: X15
Accuracy: 0.8181904761904762
Precision: 0.6748566748566749
Recall: 0.3527397260273973

Relevant Data Withholding Variable: X16
Accuracy: 0.8177142857142857
Precision: 0.6723856209150327
Recall: 0.3523116438356164

Relevant Data Withholding Variable: X17
Accuracy: 0.8181904761904762
Precision: 0.6740016299918501
Recall: 0.3540239726027397

Relevant Data Withholding Variable: X18
Accuracy: 0.818952380952381
Precision: 0.6784249384741592
Recall: 0.3540239726027397

Relevant Data Withholding Variable: X19
Accuracy: 0.8191428571428572
Precision: 0.678951678951679
Recall: 0.3548801369863014

Relevant Data Withholding Variable: X20
Accuracy: 0.8183809523809524
Precision: 0.6759639048400328
Recall: 0.3527397260273973

Relevant Data Withholding Variable: X21
Accuracy: 0.8192380952380952
Precision: 0.6798029556650246
Recall: 0.3544520547945205

Relevant Data Withholding Variable: X22
Accuracy: 0.8186666666666667
Precision: 0.6773399014778325
Recall: 0.3531678082191781

Relevant Data Withholding Variable: X23
Accuracy: 0.8185714285714286
Precision: 0.6773662551440329
Recall: 0.3523116438356164

Relevant Metrics Data of the Data Withholding Specific Predictor Variables (Naive Bayes)

Relevant Data Withholding Variable: X1
Accuracy: 0.38876190476190475
Precision: 0.25278585271317827
Recall: 0.8934075342465754

Relevant Data Withholding Variable: X2
Accuracy: 0.3924761904761905
Precision: 0.24990721266856364
Recall: 0.8647260273972602

Relevant Data Withholding Variable: X3
Accuracy: 0.39266666666666666
Precision: 0.24996906323474818
Recall: 0.8647260273972602

Relevant Data Withholding Variable: X4
Accuracy: 0.3924761904761905
Precision: 0.24990721266856364
Recall: 0.8647260273972602

Relevant Data Withholding Variable: X5
Accuracy: 0.39
Precision: 0.24929143561306222
Recall: 0.8660102739726028

Relevant Data Withholding Variable: X6
Accuracy: 0.3862857142857143
Precision: 0.24908380161250918
Recall: 0.8728595890410958

Relevant Data Withholding Variable: X7
Accuracy: 0.3871428571428571
Precision: 0.2492966360856269
Recall: 0.872431506849315

Relevant Data Withholding Variable: X8
Accuracy: 0.3871428571428571
Precision: 0.2488050006128202
Recall: 0.8690068493150684

Relevant Data Withholding Variable: X9
Accuracy: 0.38685714285714284
Precision: 0.24859034076979653
Recall: 0.8681506849315068

Relevant Data Withholding Variable: X10
Accuracy: 0.3878095238095238
Precision: 0.24895705521472392
Recall: 0.8685787671232876

Relevant Data Withholding Variable: X11
Accuracy: 0.38771428571428573
Precision: 0.2488648913977175
Recall: 0.8681506849315068

Relevant Data Withholding Variable: X12
Accuracy: 0.3921904761904762
Precision: 0.24962871287128713
Recall: 0.8634417808219178

Relevant Data Withholding Variable: X13
Accuracy: 0.39266666666666666
Precision: 0.24996906323474818
Recall: 0.8647260273972602

Relevant Data Withholding Variable: X14
Accuracy: 0.3923809523809524
Precision: 0.24981440237564959
Recall: 0.8642979452054794

Relevant Data Withholding Variable: X15
Accuracy: 0.3922857142857143
Precision: 0.24984539270253556
Recall: 0.8647260273972602

Relevant Data Withholding Variable: X16
Accuracy: 0.3924761904761905
Precision: 0.2497833890332962
Recall: 0.8638698630136986

Relevant Data Withholding Variable: X17
Accuracy: 0.3922857142857143
Precision: 0.24984539270253556
Recall: 0.8647260273972602

Relevant Data Withholding Variable: X18
Accuracy: 0.43066666666666664
Precision: 0.25602893890675243
Recall: 0.8180650684931506

Relevant Data Withholding Variable: X19
Accuracy: 0.4336190476190476
Precision: 0.2554517133956386
Recall: 0.8073630136986302

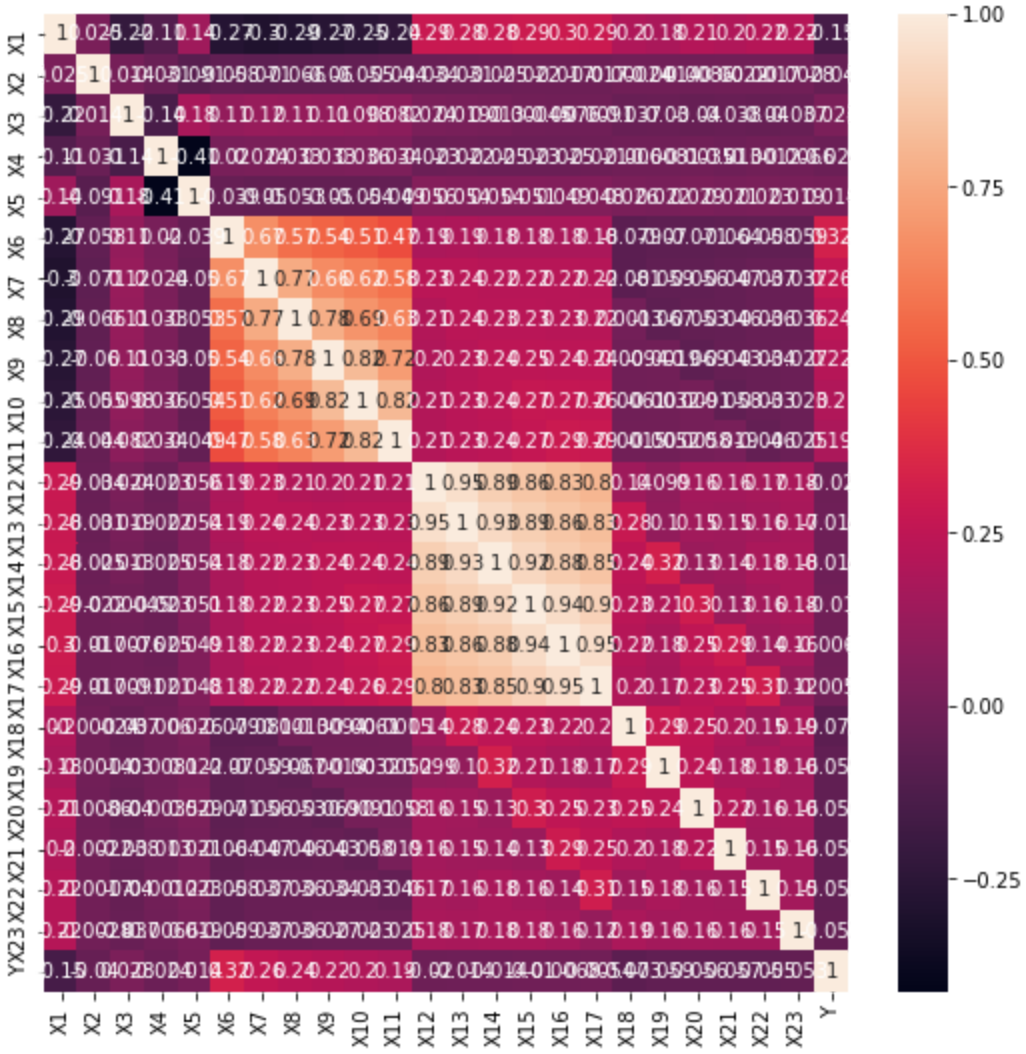
Relevant Data Withholding Variable: X20
Accuracy: 0.42333333333333334
Precision: 0.25490971398444706
Recall: 0.8279109589041096

Relevant Data Withholding Variable: X21
Accuracy: 0.4086666666666667
Precision: 0.2531548757170172
Recall: 0.8501712328767124

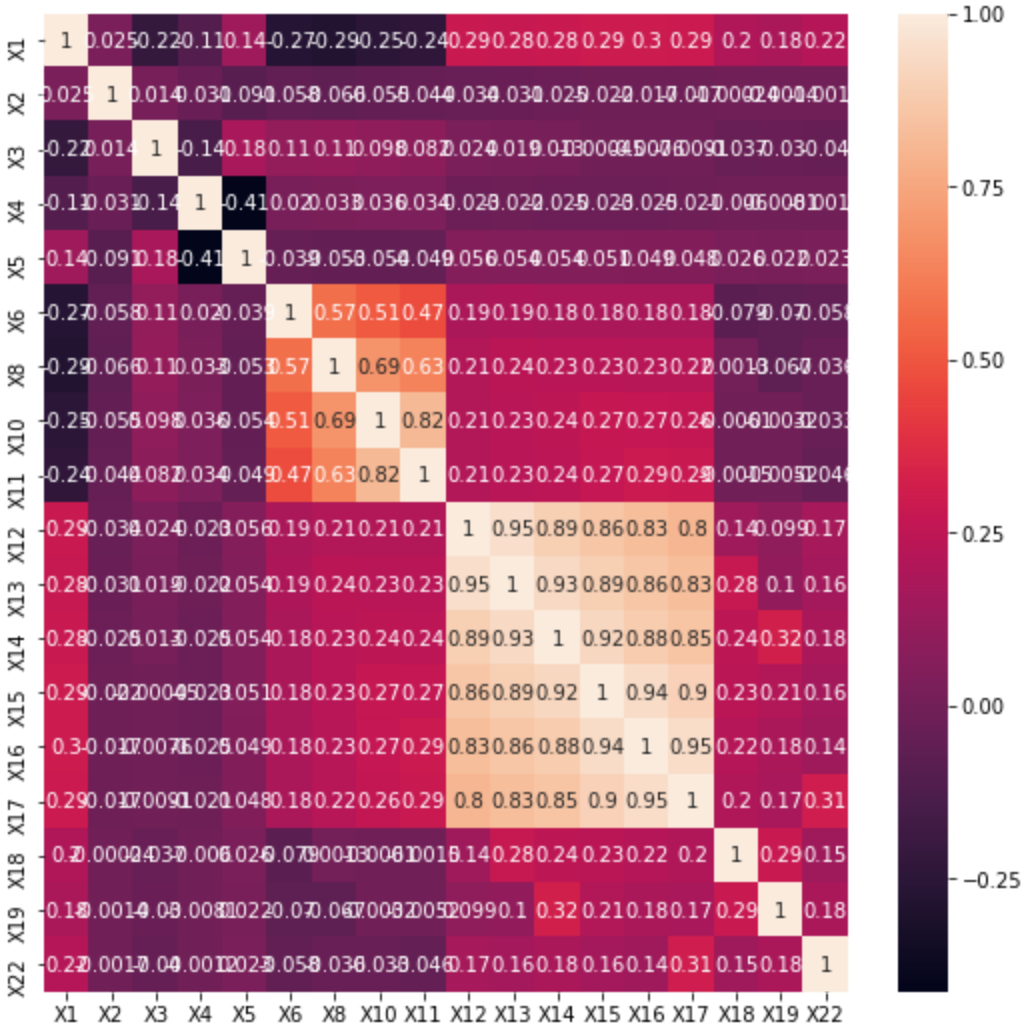
Relevant Data Withholding Variable: X22
Accuracy: 0.4147619047619048
Precision: 0.25346278317152104
Recall: 0.8381849315068494

Relevant Data Withholding Variable: X23
Accuracy: 0.41695238095238096
Precision: 0.25421968319916904
Recall: 0.8381849315068494

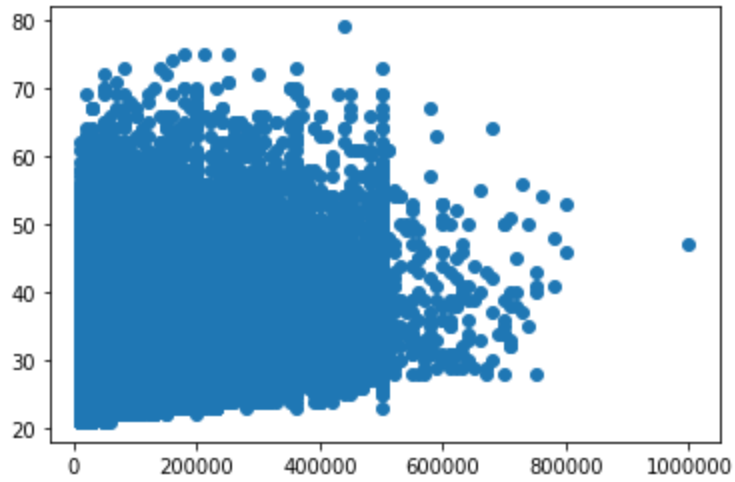
Correlation Before Data Preparation



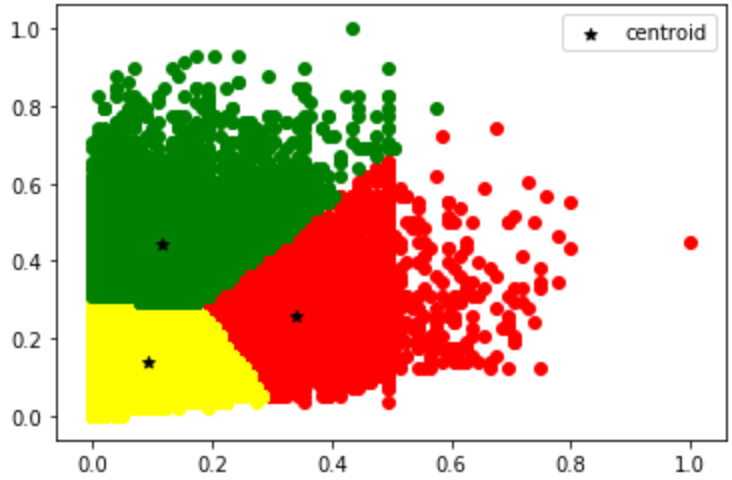
Correlation After Data Preparation

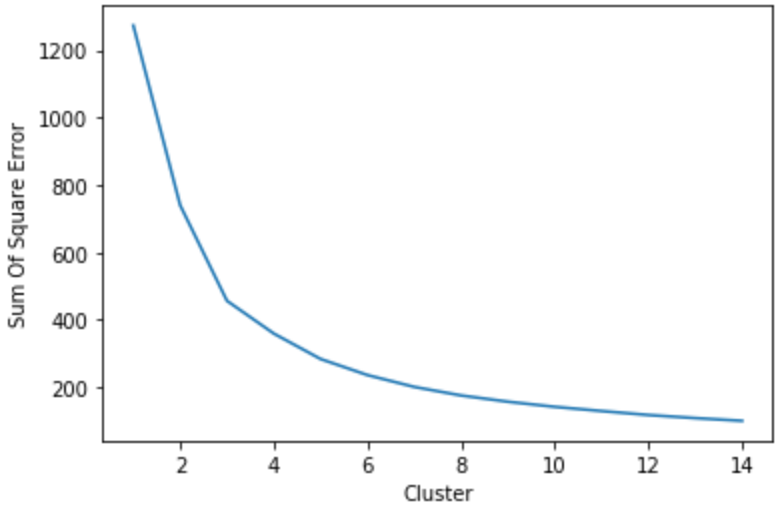


K-Means CLustering Plots and Elbow Plot



(the ability to find relevant data to work with)[4]





VII. REFERENCES

[1] *Sevenpillarsinstitute.org*, sevenpillarsinstitute.org/case-studies/taiwans-credit-card-crisis/.

[2] Chang, Chih Hsiung, et al. “A Study on the Coping Strategy of Financial Supervisory Organization under Information Asymmetry: Case Study of Taiwan's Credit Card Market.” *Horizon Research Publishing Corporation*, www.hrpub.org/download/20171030/UJM3-12110203.pdf.

[3] Singh, Harshdeep. “Understanding Gradient Boosting Machines.” *Medium*, Towards Data Science, 4 Nov. 2018, towardsdatascience.com/understanding-gradient-boosting-machines-9be756fe76ab.

[4] Koehrsen, Will. “Beyond Accuracy: Precision and Recall.” *Medium*, Towards Data Science, 10 Mar. 2018, towardsdatascience.com/beyond-accuracy-precision-and-recall-3da06bea9f6c.

[5] Gandhi, Rohith. “Naive Bayes Classifier.” *Medium*, Towards Data Science, 17 May 2018, towardsdatascience.com/naive-bayes-classifier-81d512f50a7c.