Group 11 report

Stock default probability prediction and machine learning application

Group member	Student ID
Xuanying Yu	**1155208081**
Ruoyan Tang	**1155205755**
Jiahui Ding	**1155201690**
YiYun Wu	**1155208167**
Yujie Li	**1155203136**

1.Introduction

The partial valuation method estimates risk by valuing the asset portfolio once in its initial state and inferring potential asset changes using local derivatives. The delta-normal distribution method is an example of this approach. The full valuation method measures risk by re-pricing the portfolio under various scenarios. Historical simulation and Monte Carlo simulation are examples of the full valuation method. Historical simulation requires a large amount of data, typically no less than 1500 samples. VaR (Value at Risk) represents the lower percentile of the return distribution within a certain period and measures the maximum potential loss with a confidence level. There are two broad categories of VaR calculation methods: partial valuation and full valuation. This paper uses the sorting method to calculate VaR for stocks and trains models like logistic regression, SVM, and random forest to predict default events based on the results.

1.1 Data Processing

1.1.1 Stock Data Retrieval

In this article, the "pedquant" package is used to select the stock with the code "000008.sz" for analysis. Firstly, the stock data from its listing date until 2023-01-01, is selected and stored in a dataframe. The data is then cleaned by removing irrelevant and missing information.According to appendix 1.1.1.The following is the result of processing the first ten rows of stock data.

1.1.2 VaR Data Retrieval

To predict the default probability between 2020 and 2021, the VaR for the period of 2019-2020 is processed in descending order using the sort function. The return rate at the 25th percentile is selected as the stock return rate for the first trading day of 2020. Then, through a loop operation, it is determined whether a default occurs. If the daily return rate is higher than the VaR at a 75% confidence level for the next one-year period, as calculated based on sorting, it is considered as non-default and assigned a value of 0 in the "if default" column. Otherwise, it is assigned a value of 1.

Among them, the number of defaults was 237, and the number of non-defaults was 735.According to appendix 1.1.2.

1.1.3 Indicator Calculation

After obtaining the default data, the stock data with default records will be stored in a new data frame. The following data indicators will be calculated in the original data frame, and the data from 2019 to 2021 will be stored in the new data frame with default records.According to appendix 1.1.3

(1) Trading volume, returns, volatility

Trading volume refers to the total number of shares or contracts traded during a specified time period. It indicates market activity and liquidity. Returns represent the profit or loss generated on an investment over a specific period and are often expressed as a percentage of the initial investment. Volatility measures the degree of variation in a trading price series over time and is important for assessing potential risk and return, with higher volatility indicating higher potential risk.

(2) MACD

The Moving Average Convergence Divergence (MACD) is calculated using the TTR package and is derived from the double exponential moving average. It indicates the current bullish or bearish state and potential price trend development based on the convergence or divergence of fast and slow averages. Changes in MACD represent changes in market trends, and MACD at different candlestick levels reflects buying and selling trends within the current timeframe.

(3) KDJ

KDJ is calculated using the xts package to convert the required data into time series data. It assesses the statistical system and calculates the Unsmoothed Stochastic Value (RSV) and values for K, D, and J using smoothed moving averages to determine stock trends.

(4) OBV

On-Balance Volume (OBV) predicts price trends by analyzing volume changes. It quantifies volume as a trend line and combines it with a price trend line to infer market sentiment. OBV emphasizes that market price changes must be accompanied by corresponding volume changes.

(5) CCI

The Commodity Channel Index (CCI) measures whether stock prices have deviated from their normal distribution range and identifies overbought and oversold conditions. CCI fluctuates between positive and negative infinity and is calculated based on the statistical concept of the deviation between prices and the average price range over a fixed period.

2.Logit Model

2.1 Introduction about logit

The logit model is a logistic regression-based model that transforms the results of a linear regression to a model with the output between 0 and 1 through the logistic function, indicating the probability of default.

The logistic function has shown below:

$$logit(p) = \ln\left(\frac{p}{1-p}\right)$$

Where p is standard for the probability of default

There are several reasons for choosing logit model:

1. Linear Relationship: The Logit model assumes a linear relationship between the independent variables and the probability of default, which enhances the interpretability of the model. We can understand the impact of each independent variable on the probability of default through the regression coefficients.
2. Probability Output: The output of the Logit model is a probability value ranging from 0 to 1, representing the likelihood of default. This allows us to directly classify and make decisions based on the probability values, such as setting a threshold to distinguish between default and non-default.
3. Handling Outliers: The Logit model is not sensitive to outliers, which means it can maintain robustness even in the presence of exceptional values in the data, without having a significant impact on the overall performance of the model.
4. Interpretability: The regression coefficients of the Logit model provide information about the independent variables. We can explain the contribution of each independent variable to the probability of default, thereby gaining a better understanding of the factors influencing stock default.

2.2 Logit regression result

In the logit model, the indicate J in KDJ is calculated based on K and D, so it has high linear relationship, it may cause overfitting. So we delete this independent variable.

The logit regression model has shown below. According to Appendix 2.2

The pseudo R2 is a measure of the goodness-of-fit of the logit regression model. In this case,

the pseudo R2 is 0.9744. This indicates that the independent variables included in the model can explain 97.44% of the variation in the dependent variable. It suggests that the model is a good fit for the data and has a strong explanatory power.According to appendix 2.2

When the p-value < 0.05, we can conclude that there is significant relationship between independent variable and dependent variable. Only yield rate satisfy this condition.

Based on the p-value above, we focus on the coefficient. The coefficient of yield rate is less than 0, so it has negative feedback of the result. The absolute value of coefficient is 88. We can think that this independent variable contributes a lot to the probability of default

3.SVM model

SVM (support vector machine) is generally used for binary classification tasks. Here are some important terminologies. One is hyperplane and another is support-vector. Usually, we have more than one independent variable, so here comes to hyperplane for classifying the data set into two groups. Support vectors are the points which are nearest to the hyperplane. Although there can be a lot of hyperplanes, our goal is to find a hyperplane which makes the distance between these support vectors and hyperplane largest. According to appendix 3.1

Hinge Loss function is needed to express the degree that the sample does not meet the constraint.

$$L(y) = \max(0, 1 - \hat{y}\, y)(0 < \hat{y} < 1)$$

SVM also has kernel function which makes itself become a nonlinear classifier. Mapping the data points to a higher dimension, in this case, transferring a curve into a plane, which is easier to solve.

In this project, we have seven independent variables, and one binary dependent variable. We use svm_classification = SVC() to initialize the SVM model, and use svm_classification.fit(train_x, np.ravel(train_y)) to fit the model. Also get the accuracy score using accuracy_score(test_y, y_pred). The following is the result table.According to Appendix 3.2

Because SMV is a black-box model, so we can only know some parameters. In the above table, the kernel function is Gaussian distribution, gamma is a kernel function parameter that controls the range of influence the data point has on the decision boundary, C is the penalty parameter, and score is the accuracy.

4.Random forest

The basic idea of random forest is to use the same data to grow many decorrelated trees, each of which is very imperfect, and then combine those many imperfect trees into one prediction. The method that implements that basic idea is called Bootstrap aggregation.It helps reduce the risk of overfitting by smoothing out individual tree biases and errors and mitigate the impact of outliers, thereby enhancing predictive performance.

As the target variable is binary, Random forest classifier model with parameter n_estimators=100, random_state=0,is used. The classification report suggests that the model is performing exceptionally well on the given data set. Precision in this case, reveals that when the model predicts a certain class (either 0 or 1), it is correct 100% of the time. Recall figure illustrates its ability to capture 99% of the actual instances of class 1. The F1-score, representing the harmonic mean of precision and recall, providing a balance between the two metrics, further

underscores the robustness of the model. Accuracy represents the overall correctness of the model across all classes. Notably, the model is achieving perfect accuracy. According to appendix 4.1

However, that better performance comes at a cost. Because of the random forest's black box nature, it's hard to tell what patterns of association between y and x are important for the prediction, we need to carry out extensive diagnostic evaluation. To address this, I conducted an analysis using importance scores and partial dependence plots. The importance scores plot highlights the features with stronger predictive contributions, aiding in the identification of influential variables. Meanwhile, the partial dependence plot for the selected feature 'RRE' reveals a distinct relationship, indicated by a sharp decreasing line, signifying that higher 'RRE' values are associated with elevated predicted probabilities of the target outcome 0. On the contrary, the MACD plot exhibits a complex curve rather than a simple linear trend, it suggests that the relationship is not straightforward. There may be more complex interactions between the selected features and others.According to appendix 4.2

5.Model evaluation

The effectiveness of the classification model is evaluated as the accuracy of the model, and the model that can accurately predict default and non-default is better.It involves using various techniques to measure how well the model is performing in making predictions .Three classification model NP, ROC and CAP .

5.1 NP principle

The NP principle is an evaluation principle where r1 represents an assessment of accuracy, which involves discarding true errors and is the cost of misclassifying positive cases as negative cases. R2 represents an assessment of sensitivity, which involves preserving false errors and is the cost of misclassifying negative cases as positive cases.what's more,A larger NP value in np principle indicates that the model has a better effect.

$$\frac{1-r1}{r2} > \theta_0$$

Based on the results provided, we can compare and evaluate these models.Logit:Based on the results provided, we can compare and evaluate these model.Logit:Accuracy (r1) = 0.1726384 Sensitivity (r2) = 0.6240602 Normalized score (NP) = 1.325772 SVM.NP: Accuracy (r1) = 0.4104235 Sensitivity (r2) = 0.6842105 Normalized score (NP) = 0.8616888 Random forest. NP: Accuracy (r1) = 0.1791531 Sensitivity (r2) = 0.7894737,Normalized score (NP) = 1.039739.According to appendix 5.1

In this part,In terms of accuracy, svm.NP has the highest value, followed by lg.NP and rp.NP.In terms of sensitivity, rp.NP has the highest value, followed by svm.NP and lg.NP ranks last. In terms of normalized scores, rp.NP has the highest value, followed by lg.NP, and svm.NP ranks last.

From the comprehensive evaluation, rp.NP performs well in accuracy and sensitivity, and has a high normalization score, so it can be considered as a relatively good model among the three models. Next up is svm.NP, which also has a good performance in accuracy and sensitivity. lg.NP is weak in these indicators and needs to be further optimized or adjusted.

5.2 ROC curve

The ROC curve is a graphical representation of the sensitivity and specificity trade-off in a binary classification model. It plots the true positive rate (sensitivity) against the false positive rate, illustrating the model's performance at different decision thresholds.

Each point on the curve represents a unique decision threshold, showing how the model responds to the same stimulus under different criteria. The horizontal axis represents the false alarm rate (FAR), indicating the ratio of non-targets falsely identified as targets. The vertical axis represents the hit rate (HR), representing the ratio of correctly identified targets. The ROC curve is formed by connecting these points, and the area under the curve (AUROC) measures the overall model performance.

The higher the AUROC,the better the model.In this case,the ROROC of the logit model,SVM model and Random forest model are 0.47704946285976907,:0.5097661860171195 and 0.5070086746711093 respectively.In conclusion，based on AUROC,SVM model performs best,then random forest model , logit model is the last.According to appendix 5.2

5.3 CAP curve

The Cumulative Accuracy Profile (CAP) curve assesses a classification model's risk detection ability. Customers are sorted by default probabilities, with the x-axis showing the top x% of customers with high default rates and the y-axis representing the cumulative defaulting customers. Three curves are plotted: perfect model (a straight line with a slope of 1/PD), random model (a diagonal line with a 45-degree slope), and the actual model.

The closer the actual model curve is to the perfect model curve, the stronger its predictive ability.The Accuracy Ratio (AR) is the proportion of the shaded area under the curve.According to appendix 5.3

The CAP curve measures the ability of a classification model to detect risk,AR is accuracy, and the higher the accuracy rate, the more accurate the prediction.In conclusion， based on AR,Random forest model performs best,then logit model , SVM model is the last.In this case,the AR ofthe logit model,SVM model and Random forest model are 0.940586124401914,0.7941746411483 and 0.9681459330143541 respectively.In conclusion,Based on AR,Random forest model performs best,then logit model , SVM model is the last.According to appendix 5.3.

In conclusion,the logit model is low accuracy, average sensitivity, low AUROC and high accuracy. From the comprehensive evaluation, its overall performance is slightly below the average level.

The SVM model showed high accuracy, average sensitivity, medium AUROC and average accuracy. From the comprehensive evaluation, the comprehensive performance under a number of indicators is medium, slightly better than logistic regression model.

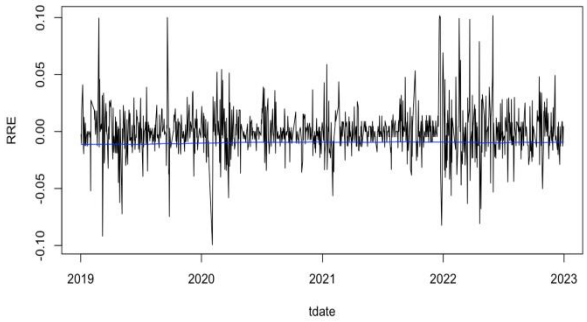
The accuracy of the random forest model is low, but the sensitivity is high, the AUROC is at the medium level, and the accuracy is the highest. The comprehensive evaluation shows that the performance is good in terms of accuracy and sensitivity, but the AUROC is slightly lower and has the highest accuracy.

In conclusion， the random forest model performs well in several indexes, especially in the correct rate.

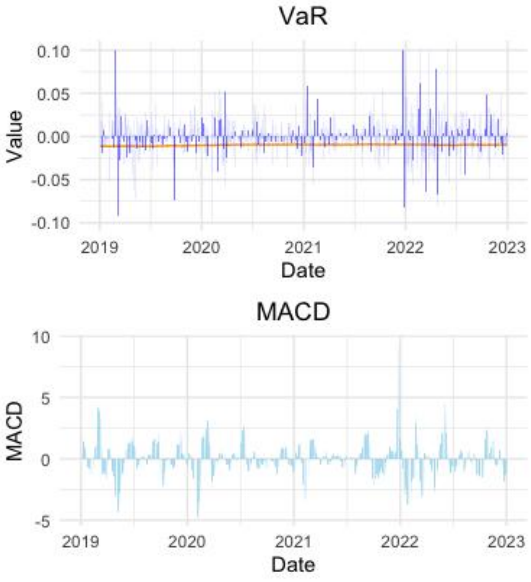
Appendix:1.1.1

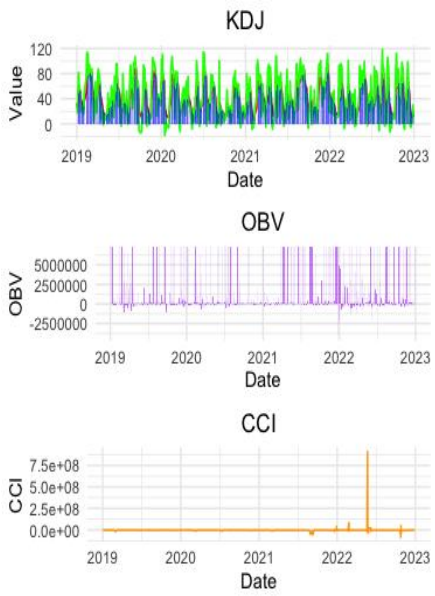
	Date	Highest.Price	Lowest.Price	Close_yesterday	Swing	Volumn	Amount
1	2010-03-19	12.35	11.98	NA	NA	9501	11649720
2	2010-03-22	12.60	12.25	12.35	-0.003238866	11630	14385451
3	2010-03-23	12.40	12.01	12.31	-0.002437043	8310	10174749
4	2010-03-24	12.55	12.18	12.28	0.013843648	8337	10379812
5	2010-03-25	12.60	12.10	12.45	-0.015261044	13460	16639106
6	2010-03-26	12.46	12.13	12.26	0.002446982	8238	10141977

Appendix:1.1.2



Appendix 1.1.3:



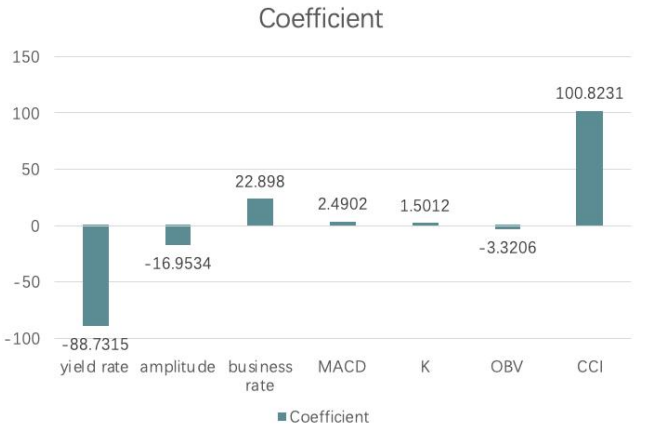


Appendix 2.2:

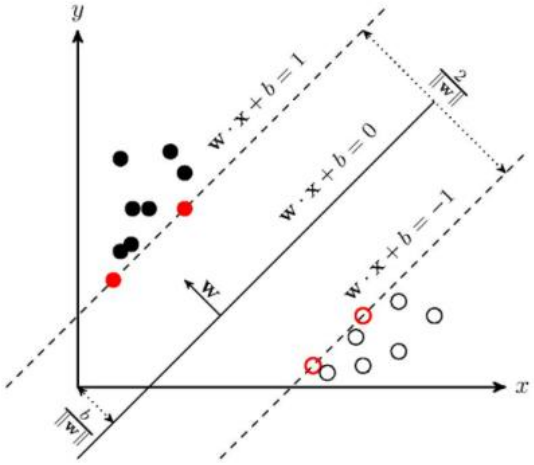
Logit Regression Results

Dep. Variable:	default or not	No. Observations:	441
Model:	Logit	Df Residuals:	433
Method:	MLE	Df Model:	7
Date:	Mon, 13 Nov 2023	Pseudo R-squ.:	0.9744
Time:	11:27:22	Log-Likelihood:	-6.1427
converged:	True	LL-Null:	-239.70
Covariance Type:	nonrobust	LLR p-value:	9.342e-97

	coef	std err	z	P>|z|	[0.025	0.975]
const	-34.2113	16.593	-2.824	0.030	-66.813	-12.610
yield rate	-88.7315	39.222	-2.847	0.026	-167.886	-28.577
amplitude	-16.9534	10.325	-1.869	0.109	-37.389	0.482
business rate	22.8980	12.439	2.188	0.077	-1.238	22.558
MACD	2.4902	3.015	-0.243	0.464	-4.439	3.458
K	1.5012	2.187	0.422	0.604	-1.825	2.828
OBV	-3.3206	8.187	1.716	0.668	-0.045	0.687
CCI	100.8231	78.010	0.549	0.197	-38.116	67.762



Appendix 3.1:



Appendix 3.2:

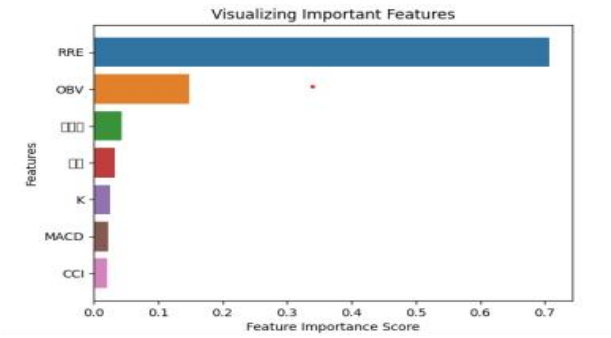
Parameters:

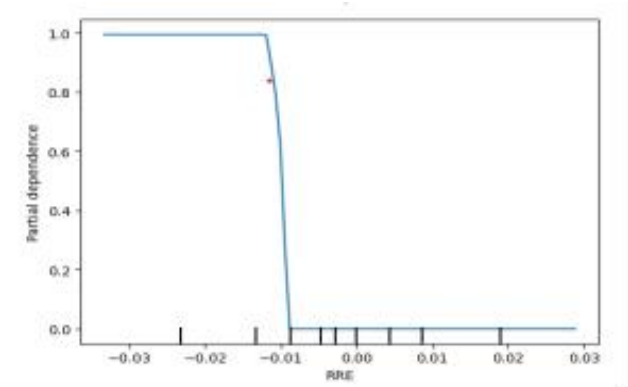
SVM-Type:	one-classification
SVM-Kernel:	RBF
gamma:	1/7
C	1
score	0.9546

Appendix 4.1:

	precision	recall	F1-score	support
0	1.00	1.00	1.00	308
1	1.00	0.99	1.00	133
accuracy			1.00	441
macro avg	1.00	1.00	1.00	441
weighted avg	1.00	1.00	1.00	44

Appendix 4.2:



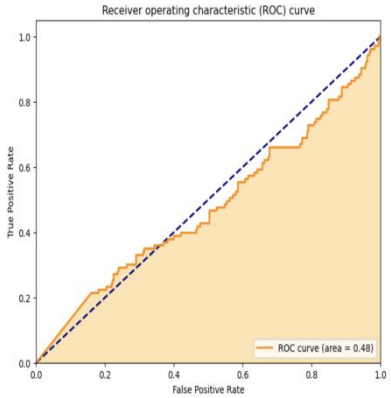


Appendix 5.1

	logit	SVM	Random forest
r1(accuracy)	0.1726384	0.4104235	0.1791531
r2(sensitivity)	0.6240602	0.6842105	0.7894737
NP	1.325772	0.8616888	1.039739

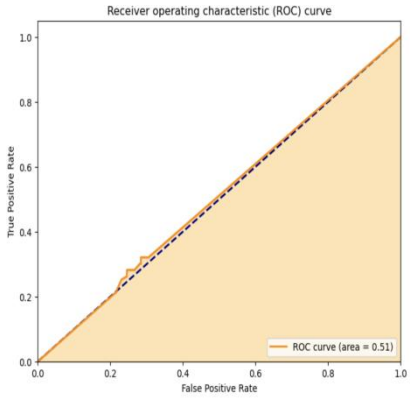
Appendix 5.2:

logit



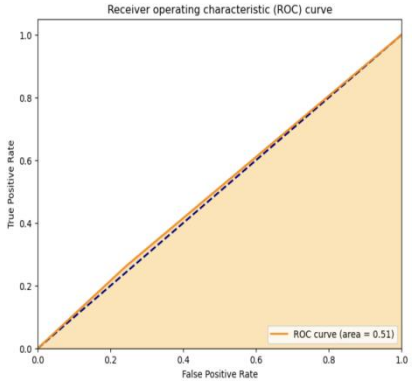
AUROC:0.477

SVM



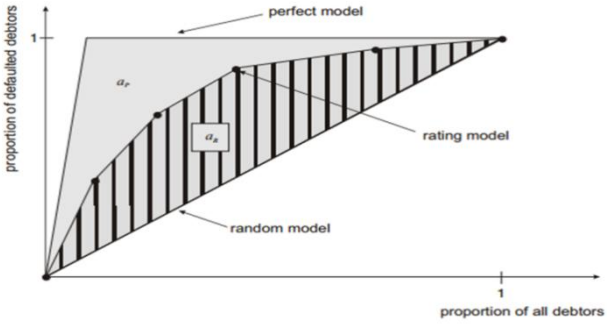
AUROC:0.510

Random

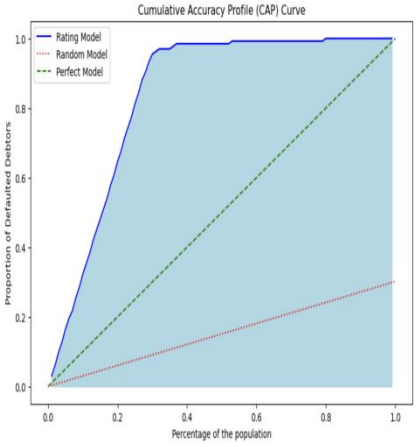


AUROC: 0.507

Appendix 5.3:

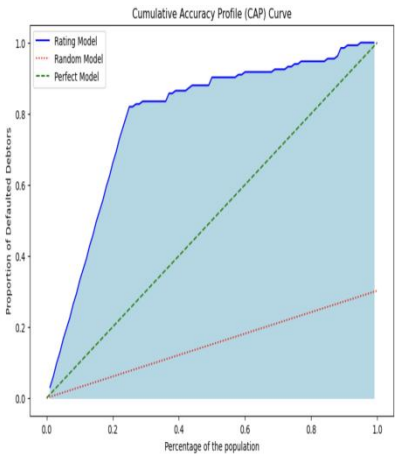


logit



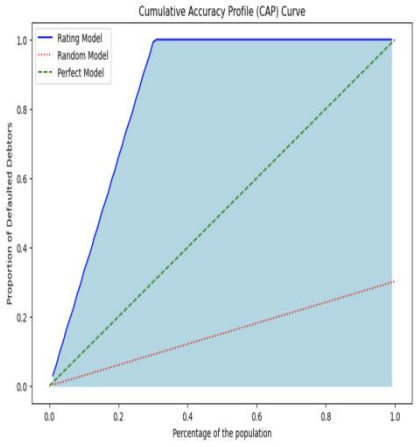
AR:0.94

SVM



AR:0.79

Rando



AR:0.97