A Comparative Study of Statistical and Machine Learning Methods for Credit Card Default Prediction

Zhiwen Wang

University of Minnesota

STAT 4052: Statistical Machine Learning II

Instructor: Sara Algeri

Introduction

Credit card default prediction can help financial institutions identify high-risk clients and avoid potential losses. The primary goal of this project is to determine which variables contribute the most to predicting credit card default, and construct a classification model which can help predicting possible defaults.

Initial modeling results show that variables related to past repayment status (e.g., PAY_0, PAY_5) and the amounts of bill statement (BILL_AMT3, BILL_AMT5) and previous payment (e.g., PAY_AMT1, PAY_AMT2) are among the most important predictors of default. We will use several statistical methods, including logistic regression, KNN and ensemble method, to predict the probability of default and explore the influence of missing data imputation on model performance.

Through this analysis, we expect to explore the most influential factors of default and provide insights that are beneficial to financial decision making and management strategies.

Methods

There are two different datasets that we plan to use: one in which the missing values are handled by simple imputation, the other in which missing values are imputed by iterative regression. Through multiple modeling approaches, we can evaluate differences in predictive performance and model flexibility.

Data Preprocessing, Missing Value Imputation and Predicting Default

The dataset includes categorical variables such as SEX, EDUCATION, MARRIAGE and repayment status (PAY_0 to PAY_6), as well as quantitative variables such as LIMIT_BAL and the amount of bill statement and previous payment. First, we preprocess data through simple imputation. Then we construct a full logistic regression model including all predictors, and select variables through three different methods: Backward Stepwise, Forward Stepwise and LASSO. Then we compare the predictive performances of these three methods, which helps us get a more concise logistic model. To follow the procedures discussed in Handout 8, we plan to use KNN classifier. We will use standardized quantitative predictors and dummy categorical predictors, and tune k to optimize performance. Additionally, to further explore the effect of model flexibility, we use random forest, bagging and boosting to predict Default and compare their performances.

Iterative Regression

We will also apply iterative regression imputation to estimate the missing values. Then we analyze the dataset obtained through this procedure using the same methods described above. Finally, we compare the resulting error rates with those from the first round of analysis to assess if iterative regression improves predictive performance.

Notes: The complete R code, corresponding outputs, and brief notes about the use of AI in coding are included in the Appendix.

Results

Logistic Regression Models on the Mean-Imputed Dataset

After replacing missing values in quantitative variables with their sample means and removing the observations with missingness, we first fitted a full logistic regression model including all predictors. The summary is as follows.

```
> summary(logit_full)

Call:
glm(formula = Default ~ ., family = "binomial", data = train)

Coefficients: (7 not defined because of singularities)
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.662e+01  9.502e+02  -0.017  0.98604
LIMIT_BAL   -4.509e-07  6.785e-07  -0.665  0.50636
SEX2         2.289e-02  1.285e-01   0.178  0.85859
EDUCATION2   5.775e-02  1.469e-01   0.393  0.69412
EDUCATION3   1.989e-01  1.949e-01   1.020  0.30763
EDUCATION4  -1.461e+01  1.038e+03  -0.014  0.98877
EDUCATION5  -1.129e-01  8.689e-01  -0.130  0.89662
EDUCATION6  -1.429e+01  1.313e+03  -0.011  0.99132
MARRIAGE1    1.510e+01  9.502e+02   0.016  0.98732
MARRIAGE2    1.494e+01  9.502e+02   0.016  0.98745
MARRIAGE3    1.505e+01  9.502e+02   0.016  0.98736
AGE         -4.095e-03  8.496e-03  -0.482  0.62983
PAY_0-1      1.137e+00  4.293e-01   2.648  0.00809 **
PAY_00       3.600e-01  4.718e-01   0.763  0.44540
PAY_01       1.492e+00  3.410e-01   4.376 1.21e-05 ***
PAY_02       2.653e+00  4.306e-01   6.161 7.25e-10 ***

......

PAY_AMT1    -4.058e-05  1.494e-05  -2.717  0.00659 **
PAY_AMT2    -7.353e-06  7.436e-06  -0.989  0.32274
PAY_AMT3     5.729e-06  3.330e-06   1.721  0.08532 .
PAY_AMT4     9.646e-07  7.051e-06   0.137  0.89119
PAY_AMT5    -1.383e-05  9.847e-06  -1.404  0.16021
PAY_AMT6     1.291e-06  3.991e-06   0.324  0.74631
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1998.3  on 1886  degrees of freedom
Residual deviance: 1669.8  on 1822  degrees of freedom
AIC: 1799.8

Number of Fisher Scoring iterations: 15
```

From the output we can find that most demographic variables such as SEX, EDUCATION, MARRIAGE and AGE are not statistically significant in predicting Default since they have large p-values, while some repayment status variables in recent months such as PAY_01, PAY_2 and PAY_AMT1 show strong significance, indicating that the default is closely related to the payment status in recent months.

Variable Selection and Logistic Model Comparison

We apply three different variable selection methods – Backward Stepwise, Forward Stepwise and LASSO.

From the output of Backward Stepwise, we can find that the procedure sequentially removed predictors that don't contribute to explaining the response. The AIC of the final model is 1750.

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1998.3  on 1886  degrees of freedom
Residual deviance: 1706.0  on 1865  degrees of freedom
AIC: 1750

Number of Fisher Scoring iterations: 14
```

According to the output of Forward Stepwise, we can find that the procedure fails to produce a useful model, without introducing any predictors, indicating that none of the predictors provide improvement. The AIC of the model remains at 2000.

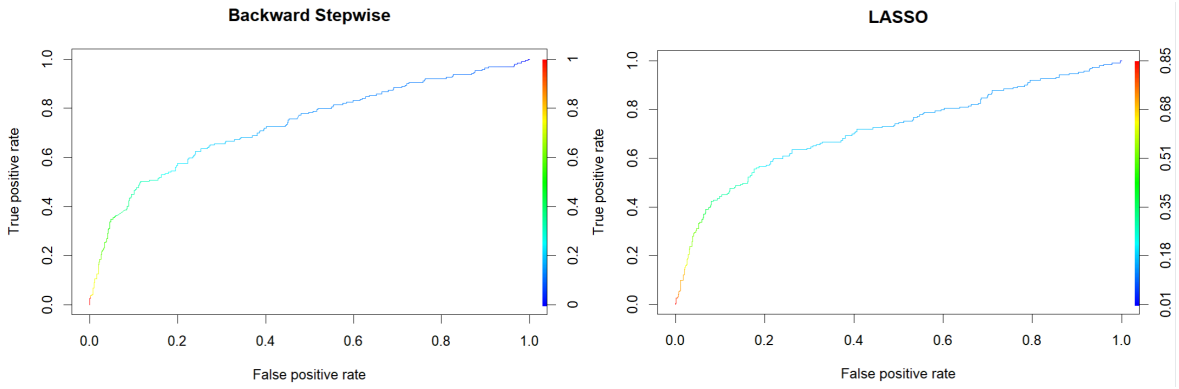
```
Start:  AIC=2000.28
Default ~ 1


Call:  glm(formula = Default ~ 1, family = "binomial", data = train)

Coefficients:
(Intercept)
     -1.254

Degrees of Freedom: 1886 Total (i.e. Null);  1886 Residual
Null Deviance:      1998
Residual Deviance: 1998         AIC: 2000
```

As for the LASSO model, it selected several repayment status predictors, including PAY_00, PAY_01, PAY_AMT1 etc. Most demographic variables are shrunk to zero, indicating their limited predictive power.

To evaluate the predictive performances of the two logistic regression models – Backward Stepwise and LASSO (Forward Stepwise has been excluded according to previous procedure), we use their accuracy, ROC curves and AUC values.

The accuracy of the Backward Stepwise model is 0.8096 and its AUC value is 0.7329.

The accuracy of the LASSO model is 0.8010 and its AUC value is 0.7172.

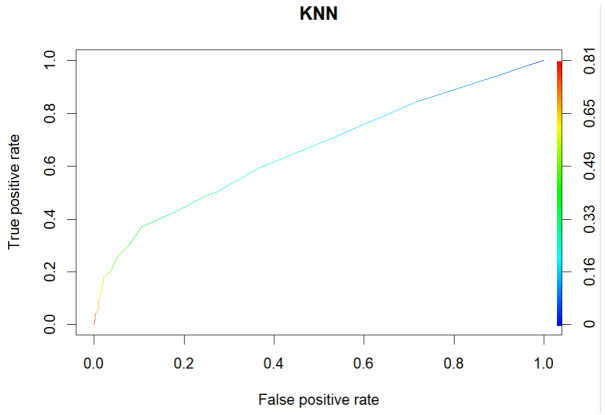


Through comparing, we can find that both models have similar predictive ability, but the accuracy and AUC value of Backward Stepwise are both higher than those of LASSO, indicating that the Backward Stepwise provides the best predictive performance in Default.

KNN (procedure discussed in Handout 8)

As required by Handout 8, we also use KNN classifier to help us explore prediction. Before fitting the model, we standardize all quantitative predictors. Since applying Educlidean distance to categorical predictors is not appropriate, we convert the categorical variables into dummy variables.

Through evaluating k from 1 to 50, we find that when k = 23, the model has the highest accuracy. Using k = 23, we can get the accuracy (0.7923) and AUC value (0.6619).



Although the accuracy is competitive, the AUC value is relatively low, indicating that KNN is limited in classifying defaulters. Since the dataset contains many dummy categorical variables and has higher dimensions, the limitation of KNN for this dataset is reasonable. These results also indicate that KNN is less suitable for this credit default prediction compared to logistic regression.

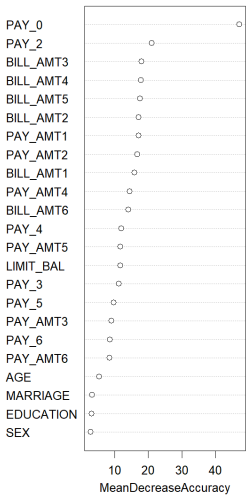
Ensemble Models

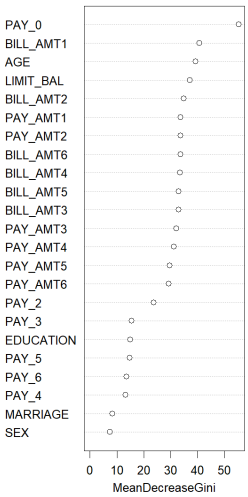
To further improve predictive performance beyond logistic regression, we plan to use three ensemble methods: Random Forest, Bagging and Boosting. These methods can help us find the relationships and complex interactions between predictors.

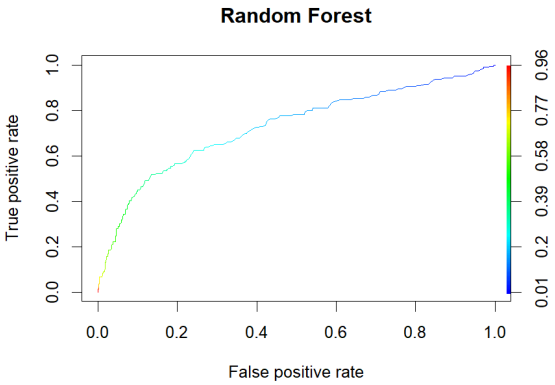
Random Forest:

```
Call:
 randomForest(formula = Default ~ ., data = train, mtry = floor(sqrt(ncol(train) -      1)), import
ance = T)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 4

        OOB estimate of  error rate: 20.56%
Confusion matrix:
     0   1 class.error
0 1388  80  0.05449591
1  308 111  0.73508353
```

The OOB error rate is 20.56%, indicating a misclassification rate of about 21% during training. On the test set, the misclassification rate is 0.2077. From the variable importance plot, we can find that PAY_0 is the most influential predictors. The AUC value is 0.7286.



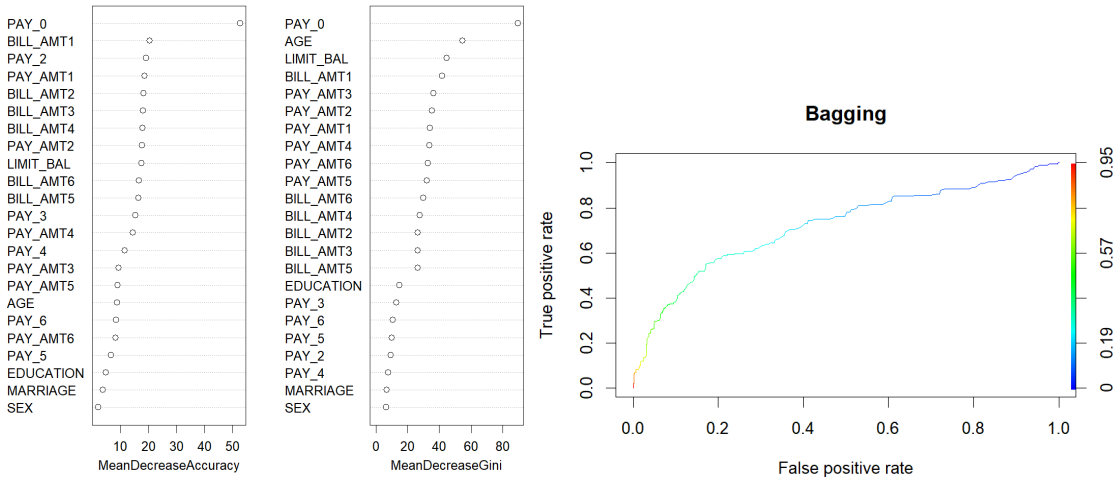




Bagging:

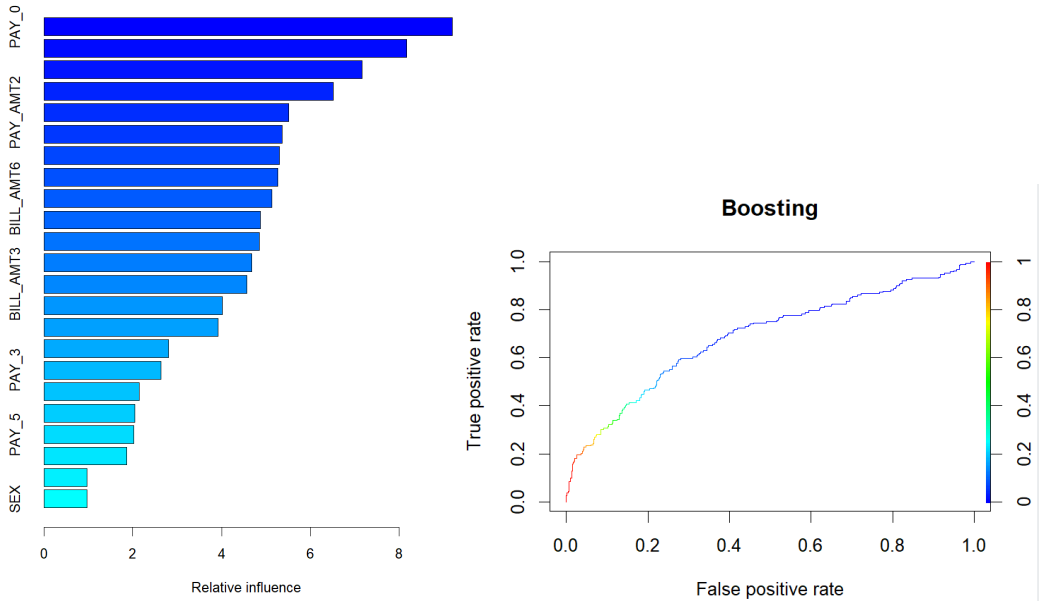
```
Call:
 randomForest(formula = Default ~ ., data = train, mtry = ncol(train) -      1, importance = T)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 23

        OOB estimate of  error rate: 20.4%
Confusion matrix:
     0   1 class.error
0 1380  88   0.0599455
1  297 122   0.7088305
```

Similarly to the Random Forest, the OOB error rate of the Bagging model is 20.4%. The misclassification rate on the test set is 0.2089. From the variable importance plot, we can also find that PAY_0 is the most influential predictors. The AUC value is 0.7207.



Boosting:

From the summary of the Boosting model, we can find that PAY_0 contributes most to the prediction. The misclassification rate on the test set is 0.2435 and the AUC value is 0.6880.



Iterative Regression

We apply iterative regression imputation to address missing values in quantitative and categorical variables. First, we impute quantitative variables by mean substitution, while categorical variables are imputed by mode substitution. Then we iteratively imputed values through fitting predictive models based on the data. For example, linear regression is used

for quantitative variables; binary logistic regression is used for two-level categorical variables; multinomial logistic regression is used for categorical variables with more than two levels. We repeated the iterative procedure ten times in order to stabilize the imputations.

Part of the summary for the dataset after iteration is as follows.

```
> summary(data_iter)
   LIMIT_BAL       SEX      EDUCATION MARRIAGE      AGE            PAY_0          PAY_2
 Min.   :  10000   1:1245   1:1122    0:   7   Min.   :21.00   0      :1416   0      :1533
 1st Qu.:  50000   2:1720   2:1332    1:1189   1st Qu.:28.00   -1     : 623   -1     : 635
 Median : 150000            3: 489    2:1732   Median :34.00   1      : 403   2      : 391
 Mean   : 163030            4:   7    3:  37   Mean   :35.34   2      : 266   -2     : 359
 3rd Qu.: 230000            5:  10             3rd Qu.:41.00   -2     : 221   3      :  32
 Max.   :1000000            6:   5             Max.   :75.00   3      :  20   7      :   6
                                                               (Other):  16   (Other):   9
     PAY_3          PAY_4          PAY_5          PAY_6        BILL_AMT1        BILL_AMT2
 0      :1514   0      :1632   0      :1615   0      :1525   Min.   :-14386   Min.   :-24704
 -1     : 648   -1     : 603   -1     : 605   -1     : 655   1st Qu.:  3097   1st Qu.:  2894
 2      : 387   -2     : 423   -2     : 436   -2     : 465   Median : 21148   Median : 20313
 -2     : 376   2      : 263   2      : 275   2      : 284   Mean   : 49850   Mean   : 47836
 3      :  10   3      :  22   3      :  14   3      :  23   3rd Qu.: 61189   3rd Qu.: 59011
 4      :   9   5      :   8   4      :  12   6      :   5   Max.   :964511   Max.   :983931
 (Other):  21   (Other):  14   (Other):   8   (Other):   8
```

From the output we can find that the imputation process doesn't impact the data patterns. Categorical predictors maintain valid levels, indicating that the imputations are applied properly.

Overall, the iterative regression imputation can incorporate information from all available predictors and estimate missing values through the relationship between variables. This imputed dataset will be used in the following modeling procedures, which can help us explore the predictive performance of different methods under a complete and consistent dataset.

Analysis on the Iteratively Imputed Dataset

Similarly to the previous procedure, we fit a full logistic regression model based on the iteratively imputed dataset. Part of the summary is as follows.

```
    Null deviance: 2188.4  on 2075  degrees of freedom
Residual deviance: 1787.4  on 2010  degrees of freedom
AIC: 1919.4

Number of Fisher Scoring iterations: 15
```

The model has a residual deviance of 1787.4 on 2010 df and AIC of 1919.4. Compared to the full model in the previous analysis (AIC = 1799.8), the AIC on the iteratively imputed dataset is higher, indicating that the iterative regression may introduce uncertainty and doesn't necessarily improve the fit of the logistic regression model. However, we can still

find that the most influential predictors are the recent payment status variables (e.g., PAY_0, PAY_AMT1), which aligns with the results obtained from the earlier analyses.
For model comparison, we repeated the same procedures as before – Backward, Forward, LASSO, Random Forest, Bagging and Boosting. The results are as follows.

	Backward	Forward	LASSO	KNN	Random Forest	Bagging	Boosting
AIC	1876.4	2190	NA	NA	NA	NA	NA
Accuracy	0.7975	NA	0.7975	0.7863	1-0.2103	1-0.2081	1-0.2565
AUC	0.7142	NA	0.7005	0.6645	0.6950	0.6912	0.6396

From the output we can find that the backward stepwise model continues to perform well in logistic regression, though the accuracy and AUC value have slightly declined. And LASSO has similar accuracy and AUC value. Compared with other methods, KNN has lower accuracy and AUC value, indicating its limitation in classifying. As for ensemble methods, the Random Forest and Bagging can provide competitive predictive performance with higher accuracy, indicating that both methods equip consistent performance across imputations. However, Boosting produces lower accuracy and AUC value compared to Random Forest and Bagging. This reduction may reflect the model's sensitivity to noise introduced during the iterative imputation process.
Additionally, almost all the accuracy and AUC values slightly declined compared to those from the simple-imputation dataset. This is likely because iterative regression reduces natural variability. The imputed values may introduce additional noise, leading to slightly lower predictive performance across models.
To sum up, though the iterative regression imputation can produce a complete dataset, it doesn't change the comparative performance of the models. The predictive patterns remain stable across different imputation strategies, indicating that the model's conclusions are stable to the choice of missing-data handling.

Discussion

This analysis aims to identify the key predictors of credit card default and compare the performances of different models under different imputation methods. Across all models, recent payment status variables and the amount variables of bill statement and previous payment such as PAY_0, PAY_AMT1 are the most influential predictors, while demographic variables contribute little to prediction.

In comparing model performance, Backward Stepwise and LASSO show similar accuracy and AUC values, indicating that both methods capture main predictive relationships in the data. Though KNN achieves moderate accuracy, it has the lower AUC value, indicating its limitation in classifying due to categorical variables and high dimensions. Among ensemble methods, Random Forest and Bagging have higher accuracy, while Boosting performs slightly worse, likely because of its higher sensitivity to noise and the uncertainty introduced in iterative imputation.

When we repeat all analyses on the iteratively imputed dataset, we can find that most models have slight reduction in the accuracy and AUC values, since the process might reduce natural variability. The overall performances of models show that the main conclusions are stable in different imputation methods.

For future analyses, we can improve model performance through more advanced imputation methods such as multiple imputation, and more flexible classification algorithms such as support vector machines, which may help identify nonlinear patterns that are hard to capture.

To sum up, predictors related to payment behavior are strongest in predicting default risk, and this finding is consistent across models. Through exploring more sophisticated approaches and machine-learning algorithms, we can provide deeper insights into client risk and support more accurate decision making in credit assessment.

References

Ghimire, B., Rogan, J., Galiano, V. R., Panday, P., & Neeti, N. (2012). An evaluation of bagging, boosting, and random forests for land-cover classification in Cape Cod, Massachusetts, USA. *GIScience & Remote Sensing, 49*(5), 623-643.

Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data*. John Wiley & Sons.

Yeh, I. C., & Lien, C. H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert systems with applications, 36*(2), 2473-2480.

Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003, November). KNN model-based approach in classification. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"* (pp. 986-996). Berlin, Heidelberg: Springer Berlin Heidelberg.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning: with applications in R*. Vol. 103. New York: springer, 2013.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models.* Cambridge university press.

Algeri, S. (2025). Course materials: Labs 2, 3, 4, 5, 8, 9, 11, and 12. STAT 4052: Statistical Machine Learning II, University of Minnesota.

Algeri, S. (2025). Course materials: Handouts 3, 8, 9, 11. STAT 4052: Statistical Machine Learning II, University of Minnesota.