Credit Default Factor Analysis

Jiaye Li, Jingxi Zhao, Maohe Wang, Zimeng Wang

Section 1. Business Problem

Credit default risk analysis is a process of assessing the likelihood that a borrower will default on their debt obligations. It is a critical problem for banks and other financial institutions that lend money, as a high default rate can lead to significant financial losses. The present study aims to leverage supervised machine-learning algorithms to identify the key factors that contribute to credit card defaults. Specifically, this research seeks to assist commercial banks in predicting credit defaults through the analysis of customer demographics and payment history. By utilizing this information, banks can create effective credit risk mitigation strategies that will help maximize profits.

To achieve this objective, we have developed multiple automated models using advanced techniques such as Logistic Regression, Random Forest, and Support Vector Machines. These models have demonstrated high accuracy and provide valuable insights into the factors that can be used in risk assessment.

Section 2. Exploratory Data Analysis (EDA)

2.1 Data Description

The dataset used in this study was obtained from the UCI Machine Learning Repository and is known as the "*Default of Credit Card Clients*" dataset. It consists of 30,000 observations and 24 variables, with each instance representing a client and containing information on their demographic characteristics and credit transaction history. The 24 variables in the dataset can be categorized into four groups, including the client's demographic information, repayment status, credit billing amount, and prior payment records. The response variable is binary, indicating whether or not the client will default on their credit in the upcoming month.

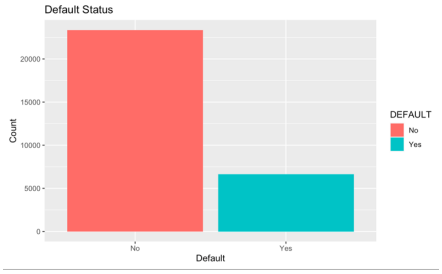

Chart 2.1 (Response Variable Distribution)

As illustrated in the graph above, the binary decision variable "default" is relatively imbalanced, with a greater proportion of instances falling into the non-default category.

2.2 Feature Analysis

From the relationships between default and given credit limit, we could see that non-defaulters have a relatively higher given credit limit.

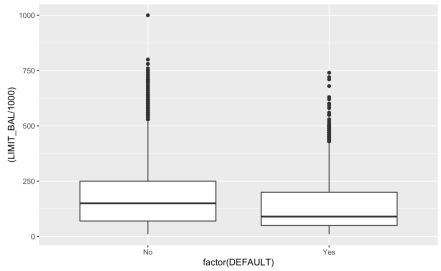

Chart 2.2 (Given Credit Limit and Default Status)

From the relationships among sex, education, and given credit, we could find that the higher level of education is, the higher the given credit would be. Also, higher education people are less likely to default.

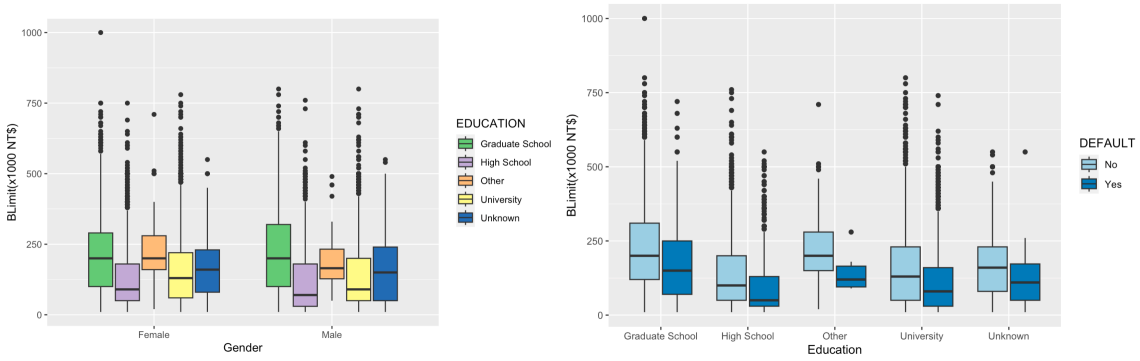

Chart 2.3 (Education Level, Gender, and Default Status)

By adding the feature of marital status, we could find that married people are more likely to be non-defaulters. Married people with a higher education level have a higher given credit limit.

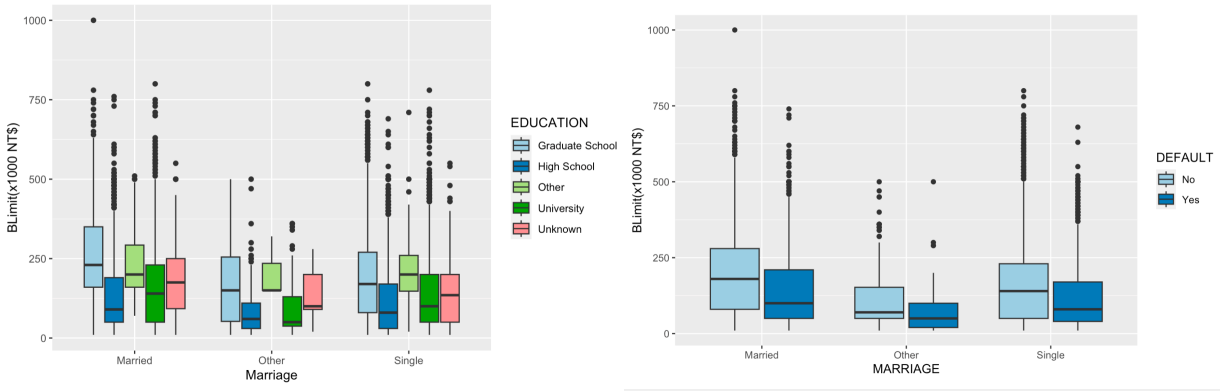

Chart 2.4 (Marriage, Education, and Default Status)

By adding the feature of age, we could find that older people are less likely to default.

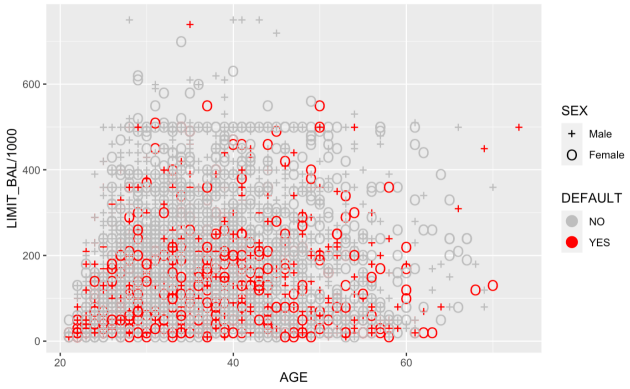

Chart 2.5 (Age, Sex, Given Credit Limit, and Default Status)

From the relationships between default and amount of bill statements in September, August, July, and June, we could find that non-defaulters usually have a lower amount of bill payment. From the relationships between default and the number of previous payments in September, August, July, and June, we could find that non-defaulters usually have a relatively low amount of previous payments each month.

Chart 2.6 (Billing Statement, Payment Amount, and Default Status)

Then, we further explore the amount of bill payment and the amount of previous payment. In a fixed previous month's payment, people with default will have a higher bill payment. Also, the amount of bill payment has a positive relation with the amount of the previous payment.

Chart 2.7 (Linear Relationship in Default Status)

From the relationship among repayment status in September (-2=pay duly, 0=payment delay for one month, 2=payment delay for two months, …8=payment delay for nine months and above), default and the amount of bill statement in September, we could find that if the repayment status is worse and late, the default probability and bill statement is higher.

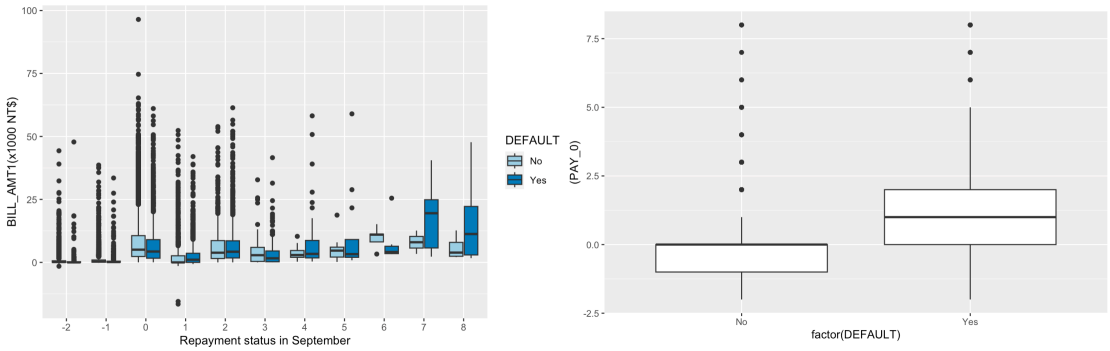

Chart 2.8 (Repayment in September and Default Status)

From the correlation heatmap, we could see that the payment status in each month and the amount of bill payment in each month have a very high correlation within themselves.

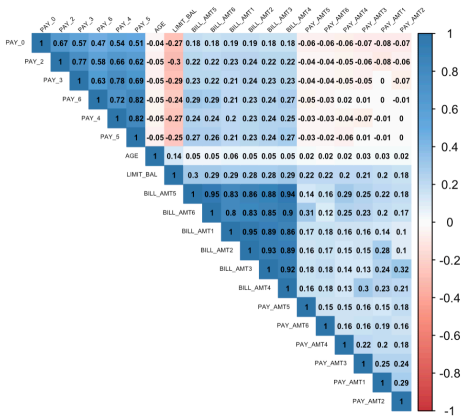

Chart 2.9 (Correlation Heatmap)

Section 3. Data Preprocessing

3.1 Data Cleaning

During the data cleaning stage, we first approached to identify missing values in the dataset, as missing values may impact the accuracy and validity of our analysis. We used descriptive statistics to identify the number and percentage of missing values for each variable in the dataset, and luckily, there aren't any *as the graph shown in Exhibit 2*.

3.2 Data Transformation

The data transformation stage is an important part of data preprocessing, as it helps to improve the quality of the data and ensure that the data is suitable for use in machine learning algorithms. In this study, we performed several transformations on the dataset to address issues such as encoding, skewness, and outliers.

First, we performed one-hot encoding on categorical label variables to convert them into a format that can be used in machine learning algorithms. We then dropped the last encoded column to prevent multicollinearity between variables, which can lead to instability in the models.

3.2.1 Skewness Correction & Outliers Removal

Next, we segregated the dataset into categorical and quantitative features and performed skewness correction on the quantitative features. For variables with positive skewness of larger than one, we deployed log transformation, and exponential transformation for negative skewness with less than -1. Since log transformation is impossible on a negative number, we offset all values in that column by a constant to move all values to positive territory. We observed a significant improvement in the distributions of some variables, such as `PAY_AMT2`, *as table shown in Exhibit 3 and 4*.

However, we identified the `BILL_AMT6` variable as performing poorly after skewness correction, and thus, excluded it from the dataset.

We then addressed outliers in the quantitative features by drawing boxplots, which displayed the distribution of data based on number summaries and showed the existence of many outliers. We identified outliers using Tukey's Rule, which represents values below Q1-1.5 IQR or above Q3+1.5 IQR, where Q1 and Q3 are the first and third quartiles, respectively, and IQR is the interquartile range (Q3-Q1). We then winsorized outliers, replacing the smallest and largest number with less extreme values, as the sample result *shown in Exhibit 5*. In winsorizing, we used 90% as the threshold, meaning all data below the 5th percentile and above the 95th percentile were set to the 5th and 95th percentile, respectively.

3.2.2 Undersampling

To address the class imbalance issue and improve the performance of our classification model, we performed undersampling on the training dataset. The purpose of undersampling is to prevent the model from being biased towards the majority class, which in our case corresponds to non-default cases, and instead focus on accurately classifying the minority class, representing default cases.

Undersampling involves randomly selecting a subset of observations from the majority class to balance the class distribution in the dataset. By doing so, we minimized false negatives (actual defaulters misclassified as non-defaulters), which carry higher risk and cost to the bank in terms of the credit assessment.

Through our back-testing with logistic regression, we observed a significant improvement in the true positive rate (TPR) when using the balanced dataset obtained through undersampling. The TPR increased from 36.2% in the unbalanced dataset to 62.3% in the balanced dataset. This indicates that the undersampling technique successfully enhances our model's ability to identify default cases, reducing the risk of misclassifying them as non-defaulters.

Section 4. Model Building and Evaluation

4.1 Model Building

4.1.1 Decision Tree

The decision tree analysis was performed on a dataset of 10,432 instances to identify factors influencing credit defaults. The tree consists of multiple nodes representing splits based on different parameters. The root node splits the data based on the payment status in the previous month (PAY_0). The analysis indicates that payment status variables (PAY_0, PAY_3, PAY_4)

and payment amount variables (PAY_AMT2, PAY_AMT4) play significant roles in predicting credit defaults. Instances with specific combinations of these variables lead to terminal nodes indicating either non-defaulting or defaulting outcomes.

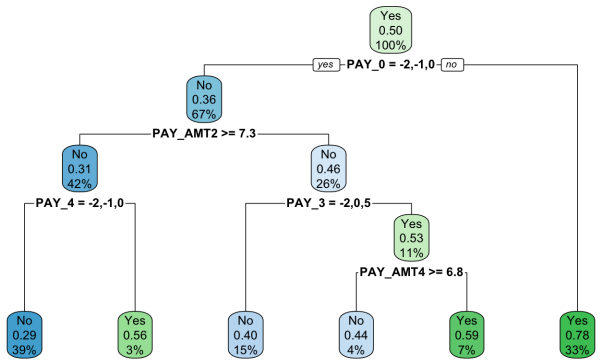

Chart 4.1 (Decision Tree Model)

4.1.2 Random Forest

The random forest model was applied to analyze a dataset consisting of 10,432 instances, aiming to identify key factors influencing credit defaults. The model, composed of 500 classification trees, investigated various variables such as payment amounts, payment statuses, age, education, marital status, and bill amounts.

The analysis revealed that the payment status in the previous month (PAY_0) played a significant role in determining default probabilities. Individuals with payment statuses of -2, -1, or 0 had a higher likelihood of non-defaulting. For those with payment statuses of -1, 2, 3, or 4, the payment amount in the fourth month (PAY_AMT4) emerged as a crucial factor.

4.1.3 Logistic Regression

A logistic regression model was applied to the dataset, incorporating various predictor variables to predict credit defaults. The coefficients obtained from the model indicate the strength and direction of the relationships between each predictor and the likelihood of default. Significant predictors include variables such as SEX.Male, EDUCATION.High.School, EDUCATION.Unknown, MARRIAGE.Single, and several PAY_AMT variables. Additionally, payment status variables (e.g., PAY_02, PAY_03, PAY_04) displayed significant coefficients. The model's goodness of fit was assessed using deviance measures, with the residual deviance showing a lower value, indicating a better fit.

4.1.4 Support Vector Machines (SVM)

The SVM model was trained using a linear kernel and a cost parameter of 0.1. It was applied to the "DEFAULT" variable in the dataset, using features such as "BILL_AMT1," "SEX," "AGE," "EDUCATION," and "MARRIAGE." The model aimed to classify instances into two categories: "Yes" and "No" for the "DEFAULT" variable.

4.2 Model Evaluation Metrics

	Accuracy	Precision	Recall	F1
Decision Tree	0.739	0.454	0.634	0.529
Random Forest	0.752	0.474	0.645	0.546
Logistic Regression	0.776	0.628	0.513	0.565
Support Vector Machines (SVM)	0.781	0.529	0.528	0.528

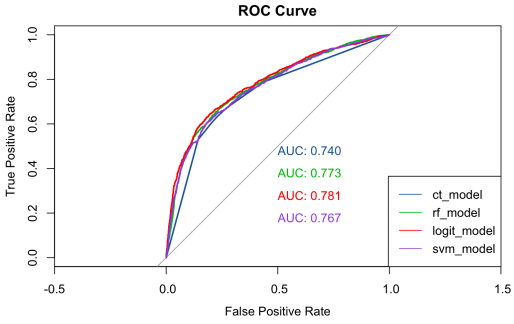

Chart 4.2 (Performance Matrix and ROC Curve)

We used four different machine learning models, including Decision Tree, Random Forest, Logistic Regression, and Support Vector Machines (SVM), to predict credit card defaults. The performance of these models was evaluated using various metrics, including accuracy, precision, recall, F1 score, and the area under the Receiver Operating Characteristic (ROC) curve.

Among the models, Logistic Regression exhibited the highest area under the ROC curve, indicating its superior ability to distinguish between positive and negative instances. Additionally, Logistic Regression achieved competitive results in terms of F1 score, which considers both precision and recall. It demonstrated a higher precision value of 0.628, suggesting a better ability to correctly identify positive instances. The recall value of 0.513 suggests that there is potential for improvement in identifying all true positive instances.

Overall, Logistic Regression outperformed other models in terms of ROC curve area and achieved a balanced F1 score. This model is recommended for credit default prediction due to its strong discriminative power and reasonably balanced precision and recall. Further, we would conduct model-tuning methods on logistic regression.

4.3 Feature Engineering & Model Tuning

To improve the logistic regression model's performance, we performed feature engineering by choosing the significant variables, including 'SEX.Male', 'EDUCATION.Other', 'PAY_0', 'PAY_AMT1', 'PAY_AMT2', 'PAY_AMT3', and 'PAY_AMT4'. We also optimized the model by varying the threshold and selecting the threshold that maximized the F1 score.

After implementing these changes, the logistic regression model achieved an accuracy of 0.816, precision of 0.724, recall of 0.330, and an F1 score of 0.453 on the test set. Compared to the original model, the accuracy and precision increased, while the recall and F1 score decreased. This indicates that the optimized model is better at correctly classifying non-defaults, but at the expense of correctly identifying defaults. We also reduced the number of variables from 24 to 7, cutting down the model complexity substantially. Overall, the feature engineering and threshold optimization improved the model's performance, resulting in a more accurate and precise model.

Section 5. Conclusions and Limitations

5.1 Conclusions

From our EDA analysis, individuals with higher education and those who are married may have a lower probability of defaulting on their loans, as they are more likely to have stable employment. Elderly individuals also tend to have a lower default rate, while younger individuals may be more likely to default due to unstable job situations. The payment status is a direct reflection of an individual's likelihood of defaulting.

From our model analysis, the sex, and education level factors are significant variables used when predicting the defaulting, and our final logistic model also includes the payment status and payment amount in a certain time period to make predictions. Our research findings suggest that commercial banks should take these factors into consideration when assessing a client's default risk. By analyzing an individual's education level, marital status, sex, and age, banks can more accurately evaluate their potential for default and make informed decisions about approving loan applications.

In conclusion, the optimized models, namely Logistic Regression, Random Forest, and Support Vector Machines, effectively address the business problem of identifying key factors contributing to credit card defaults. By leveraging these models, commercial banks can predict credit defaults and implement effective credit risk mitigation strategies. This study highlights the potential of supervised machine learning algorithms in improving risk assessment and decision-making in the

financial industry, ultimately maximizing profitability and minimizing the impact of credit defaults.

5.2 Limitations and Suggestions

Classifying clients into only two categories is too simplistic. To achieve greater accuracy, we could categorize clients into several groups based on their default risk levels, and determine the appropriate loan amounts and credit approvals accordingly. Furthermore, we could implement stricter payment schedules or other policy mechanisms to minimize credit risk. This approach would enable us to engage with more clients and benefit both parties involved.

Finally, it is important to continually monitor and update credit risk models to ensure that they remain accurate and effective. This requires regularly collecting and analyzing new data, and using feedback from real-world outcomes to refine the models and improve their performance over time.

Appendix to Project Write-up

```
  LIMIT_BAL          SEX            EDUCATION         MARRIAGE          AGE            PAY_0             PAY_2
Min.   :  10000   Length:30000       Length:30000       Length:30000       Min.   :21.00   Min.   :-2.0000   Min.   :-2.0000
1st Qu.:  50000   Class :character   Class :character   Class :character   1st Qu.:28.00   1st Qu.:-1.0000   1st Qu.:-1.0000
Median : 140000   Mode  :character   Mode  :character   Mode  :character   Median :34.00   Median : 0.0000   Median : 0.0000
Mean   : 167484                                                           Mean   :35.49   Mean   :-0.0167   Mean   :-0.1338
3rd Qu.: 240000                                                           3rd Qu.:41.00   3rd Qu.: 0.0000   3rd Qu.: 0.0000
Max.   :1000000                                                           Max.   :79.00   Max.   : 8.0000   Max.   : 8.0000
     PAY_3             PAY_4             PAY_5             PAY_6           BILL_AMT1         BILL_AMT2        BILL_AMT3
Min.   :-2.0000   Min.   :-2.0000   Min.   :-2.0000   Min.   :-2.0000   Min.   :-165580   Min.   :-69777   Min.   :-157264
1st Qu.:-1.0000   1st Qu.:-1.0000   1st Qu.:-1.0000   1st Qu.:-1.0000   1st Qu.:   3559   1st Qu.:  2985   1st Qu.:   2666
Median : 0.0000   Median : 0.0000   Median : 0.0000   Median : 0.0000   Median :  22382   Median : 21200   Median :  20088
Mean   :-0.1662   Mean   :-0.2207   Mean   :-0.2662   Mean   :-0.2911   Mean   :  51223   Mean   : 49179   Mean   :  47013
3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.:  67091   3rd Qu.: 64006   3rd Qu.:  60165
Max.   : 8.0000   Max.   : 8.0000   Max.   : 8.0000   Max.   : 8.0000   Max.   : 964511   Max.   :983931   Max.   :1664089
  BILL_AMT4         BILL_AMT5         BILL_AMT6          PAY_AMT1         PAY_AMT2          PAY_AMT3         PAY_AMT4
Min.   :-170000   Min.   :-81334   Min.   :-339603   Min.   :     0   Min.   :      0   Min.   :     0   Min.   :     0
1st Qu.:   2327   1st Qu.:  1763   1st Qu.:   1256   1st Qu.:  1000   1st Qu.:    833   1st Qu.:   390   1st Qu.:   296
Median :  19052   Median : 18104   Median :  17071   Median :  2100   Median :   2009   Median :  1800   Median :  1500
Mean   :  43263   Mean   : 40311   Mean   :  38872   Mean   :  5664   Mean   :   5921   Mean   :  5226   Mean   :  4826
3rd Qu.:  54506   3rd Qu.: 50190   3rd Qu.:  49198   3rd Qu.:  5006   3rd Qu.:   5000   3rd Qu.:  4505   3rd Qu.:  4013
Max.   : 891586   Max.   :927171   Max.   : 961664   Max.   :873552   Max.   :1684259   Max.   :896040   Max.   :621000
   PAY_AMT5          PAY_AMT6          DEFAULT
Min.   :     0.0   Min.   :     0.0   Length:30000
1st Qu.:   252.5   1st Qu.:   117.8   Class :character
Median :  1500.0   Median :  1500.0   Mode  :character
Mean   :  4799.4   Mean   :  5215.5
3rd Qu.:  4031.5   3rd Qu.:  4000.0
Max.   :426529.0   Max.   :528666.0
```

Exhibit 1 (Summary of features)

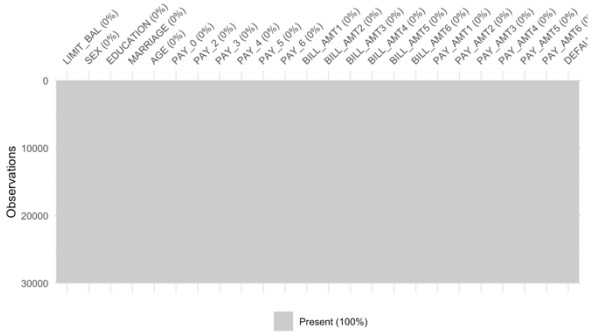

Exhibit 2 (Missing Values)

Feature Name	Skewness	Skewness after Correction
LIMIT_BAL	0.9928173	No Transformation
AGE	0.7322093	No Transformation
BILL_AMT1	2.663728	-0.5006624
BILL_AMT2	2.705086	0.7948691
BILL_AMT3	3.087676	-0.4483056
BILL_AMT4	2.821824	-1.087994
BILL_AMT5	2.876236	0.7495502
PAY_AMT1	14.66763	-1.119563
PAY_AMT2	30.45229	-1.063536

PAY_AMT3	17.21577	-0.899868
PAY_AMT4	12.90434	-0.7839622
PAY_AMT5	11.12686	-0.7681229
PAY_AMT6	10.6402	-0.6896096

Exhibit 3 (Table of Skewness Correction)

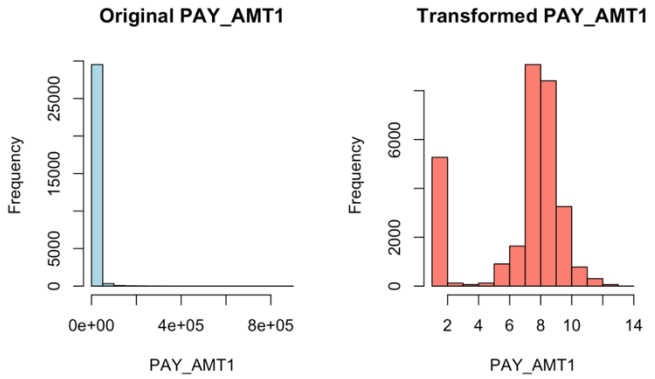

Exhibit 4 (Sample Distribution After Skewness Correction)

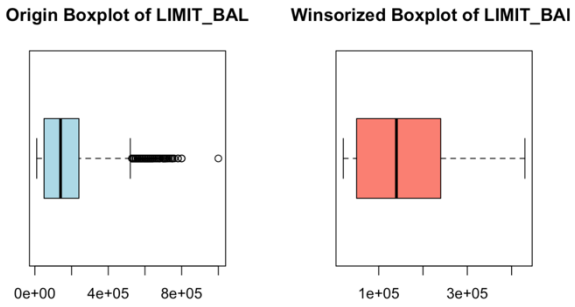

Exhibit 5 (Sample Distribution After Outlier Correction)