# Predicting Credit Card Default using Binary Classifiers and Imbalanced Learning

Priscila Moreira
pmoreira@nd.edu
University of Notre Dame

Katherine Dearstyne
kdearsty@nd.edu
University of Notre Dame

Tony Alarcon
palarcon@nd.edu
University of Notre Dame

## 1 INTRODUCTION

According to the Federal Reserve Economic Data (FRED), delinquency rates on credit card loans have been increasing across all commercial banks since 2016 [3]. The climbing default rates not only result in significant loss for lending institutions but additionally jeopardizes the credit history and future buying power of the borrower. As such, a predictive data analytical model that classifies the common characteristics and patterns indicative of an individual with the highest probability of credit card default can be a crucial tool to mitigate the current phenomenon.

This paper aims to analyze an individual's personal demographics and payment history to identify the most common factors impacting credit card default. We will attempt to solve the imbalance issue of the data set by employing a k-means SMOTE algorithm to change the training data distribution. Additionally, since features are measured at different scales, we will begin by standardizing the data and encoding categorical values as required. We will then construct various machine learning algorithms and analyze their accuracy to ascertain the most optimal model for predicting credit card default.

## 2 RELATED WORK

Ying Cheng and Ruirui Zhang propose a prediction model based on k-means SMOTE and BP neural network that achieves an accuracy of 92% [4]. In this model, the importance of data features is calculated by using random forest, and then it is substituted into the initial weights of BP neural network for prediction. The model effectively solves the problem of sample data imbalance. Therefore, this paper will explore this technique.

## 3 EXPERIMENTAL DATA AND PRELIMINARY ANALYSIS

### 3.1 Source

Our data was compiled by Yei and Lien [5] for analyzing costumer defaults on credit card payments in Taiwan. This data was obtained from UCI ML respiratory [2].

### 3.2 Preliminary Analysis of Data

Our data includes 25 attributes and uses default payment (**Yes** = 1, **No** = 0) as the target variable. Of the 30,000 customers, only 6,636 defaulted, which constitutes only 22.12% of the data. The first feature column contains unique identifying (**ID**) values for each sample. It can be safely assumed that this feature has no impact on our target variable. Thus we remove and exclude it from further analysis. The remaining 24 features, their types, and possible values are summarized in Table 1.

| Description | Possible Values |
|---|---|
| Amount of Given Credit Line (X1) | [10000, 1000000] |
| Gender (X2) | 1=male; 2=female |
| Education (X3) | 1=graduate school;<br>2=university;<br>3=high school;<br>4=other |
| Marital Status (X4) | 1=married;<br>2=single;<br>3=others |
| Age in Years (X5) | [21, 79] |
| History of Payment (X6-11) | -1=pay duly;<br>1=delayed 1 month;<br>2=delayed 2 months<br>...<br>9=delayed 9+ months |
| Amount of Bill Statement (X12-17) | (-∞, ∞) |
| Amount of Previous Payment (X18-23) | (0, ∞) |

**Table 1: Data Attributes**

While our team verified no missing values in the data, the **Education** feature contained seven unique categorical values, from 0 to 6, of which meanings for values 0, 5, and 6 were unknown. These unknown feature values are binned together into a single category called **unknown** whose categorical value is 5, as displayed in Figure 1.

Two multi-density charts are shown in Figure 2 and 3, which displays the normalized distribution of the **Limit Balance** and **Age** features, respectively, according to the **default** type. Figure 2 demonstrates that the probability of **default** is greater for the customers whose **credit limit** is below 150,000. This observation is consistent with the notion that customers who have shown good debt management (via good credit score or similar metrics) are given higher credit limits by the lending institution. Figure 3 indicates that customers between the age group of 25 to 40 are more capable of repaying their debt.

In order to have more insights into the dataset, we applied the feature importance technique to assign a score to input features based on how useful they are at predicting a target variable. There are different ways to obtain the scores, in that paper we have used scores based on an extra-trees classifier. The Figure 4 shows a bar chart created for the feature importance scores. The top four important features are **Pay_0**, **Age**, **LIM_BAL**, and **BILL_AMT1**; on the other hand, the less essential feature is **MARRIAGE**.

*Data Science Project, Notre Dame, US*
2021. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

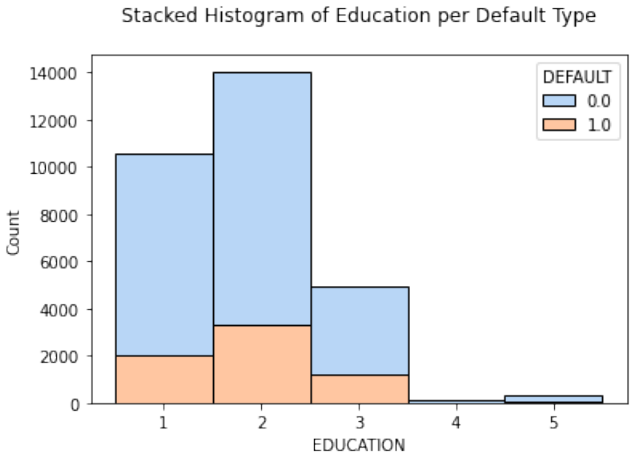



**Figure 1: Stacked Histogram of Education per Default Type**

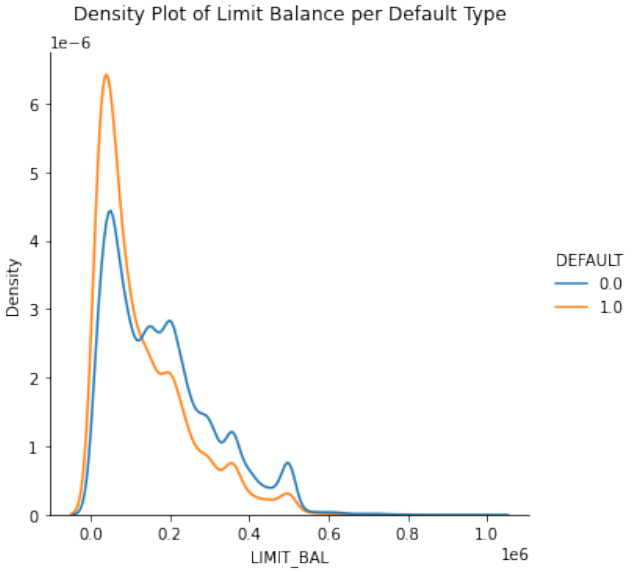



**Figure 2: Density Plot of Limit Balance per Default Type**

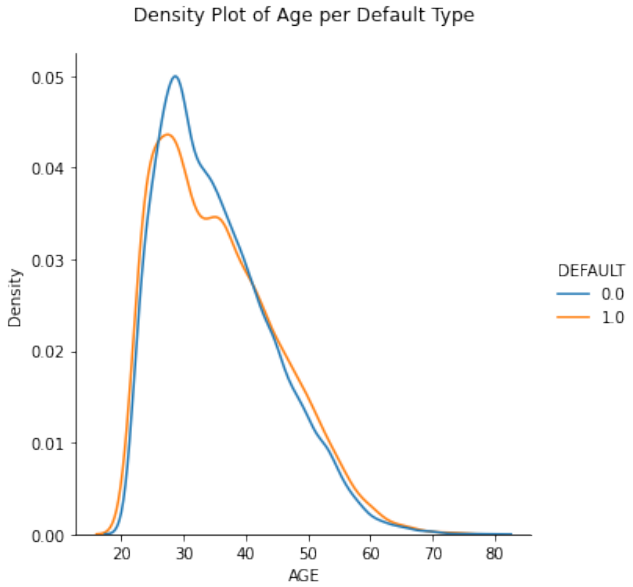



**Figure 3: Density Plot of Age per Default Type**

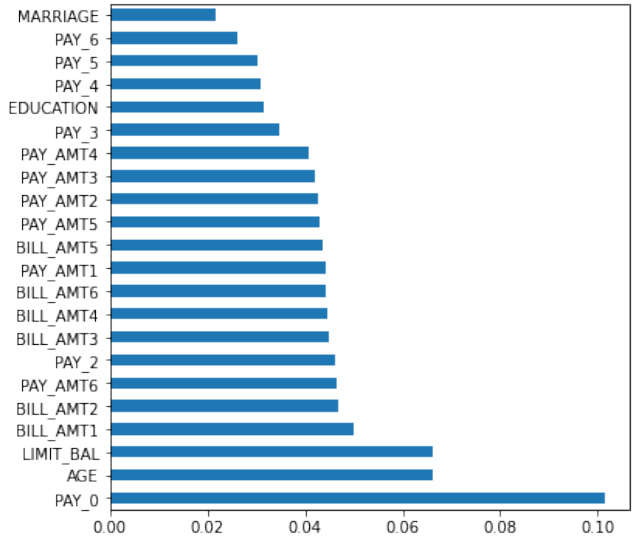



**Figure 4: Feature importance for each feature in the dataset.**

# 4 SOLUTION & METHODS

## 4.1 Model Generation

As previously noted, our data set is extremely imbalanced, with the target of interest (customers who default) consisting of only 22.12% of the observations. This skewed class distribution results in unequal miss-classification costs during the training process, as such the machine learning models are ill-equipped to learn meaningful characteristic that are indicative of customer defaults. Therefore this issue must be addressed in order to avoid poor predictability performance on the minority class. This paper attempts to address this problem by employing the SMOTE oversampling technique on the training data. This sampling technique synthesizes new instances that are close in feature space as those found in the minority class. The class distribution after employing this technique is shown in Figure 5.

To select hyperparameters for each model, we utilized a randomized search cross validation, with k=5. We opted to use the randomized search because it tends to produce comparable results to complete grid search with a substantial increase in speed [1]. For scoring, we used the $F_1$ measure due to our imbalanced dataset. The hyperparameters that were selected can be found in Table 2.

Finally, we constructed 6 different machine learning algorithms, namely, Decision Tree, Random Forest, Adaboost, Neural Network, and SVM with RBF Kernel, in order to analyze their predictability performance and ascertain the most optimal model.

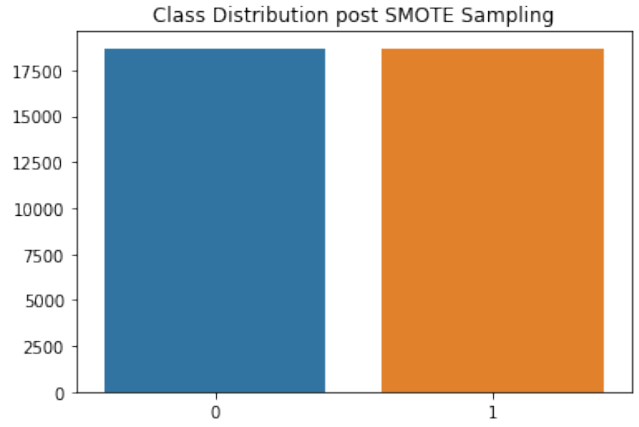



**Figure 5: Histogram of Class Distribution After Employing the SMOTE Algorithm in Training Data**

## 4.2 Model Evaluation

To evaluate our machine learning models, we perform stratified k-fold cross validation on the entire dataset, where k=5. This procedure will allow us to quantify the reliability of our measurements by reporting the mean metric scores along with their corresponding standard deviations. Setting our cross validation hyper-parameter to 5 entails that at each fold, the data will be split in a 80-20 standard partitions, such that they are randomly stratified in order to conserve the same class distributions in our testing data. Due to the severity of our imbalanced dataset, the standard accuracy metric may produce misleading evaluation results as the algorithms can simply achieve a high accuracy by simply classifying all samples as the majority class. Therefore, we will evaluate and compare the models utilizing the *F1 Score*, *Recall*, and *Precision* Macro values. In particular, this paper prioritizes the optimization of the recall score value in order to minimize the number of False Negatives in each model. False Negatives within the context of our study indicates the model predicts a customer will not default, when in fact they actually do. This is harmful for all parties involved as the lending institutions is unable to provide default mitigation services to reduce the overall impact. Furthermore, our results will be illustrated with a *confusion matrix* to visualize the number of true positives/negatives as compared to the number of false positives/negatives. Finally, we create a *AUC-ROC curve* to demonstrate the diagnostic ability of our model.

Lastly, in order to ascertain the efficacy of our sampling technique, this paper will develop several binary classifiers with identical hyper-parameters trained on two independent datasets: (1) The original dataset and (2) the over sampled SMOTE dataset. Analysis will then be performed on both dataset using the evaluation techniques described above.

# 5 RESULTS & ANALYSIS

As described in the proceeding section, five machine learning models were trained on both the original and over-sampled dataset, whose results on the test data are summarized in Table 3. This table indicates that the resulting F1 Macro score on both datasets do not differ by a significant margin. However, the recall macro score increased for all machine learning algorithms trained on the SMOTE dataset. For example, in random forest, the recall value jumped from .65 to .70 when utilizing the SMOTE dataset. This suggests that employing oversampling techniques results in better predicting performance of the minority class.

The top three models were identified as Random Forest (trained on SMOTE), Random Forest (trained on Original) .67, and Neural Networks (trained on Original), achieving a Macro F1 Score of .69, .67, and .67 respectively. Random Forest trained on the SMOTE dataset deserves an extra layer of attention as it also achieved the highest recall value of .70. It is clear that all machine learning algorithms employed in this paper far outperforms a naive solution of random guessing which would achieve an accuracy of 50%.

To further evaluate our models, we plot the ROC curves for each model(See Figure 6). The AUC is $\geq 75$ for all models except for the Decision Tree.

To better understand the performance of Random Forest, which had the greatest F1 macro, we conducted a more detailed analysis using the two most important features, AGE and PAY 0.

Figure 7a shows the recall by age groups of customers for the Random Forest trained on the original dataset. We can see that this model has a good recall when predicting the class NO DEFAULT. On the other hand, it has lower recall predicting the class DEFAULT, mainly for groups 1, 2, and 3. Therefore, that model is mispredicting some customers between 25 and 45 years old that will DEFAULT.

Figure 7b shows the recall by age groups of customers for the Random Forest model trained on SMOTE data. That model shows

higher recall for all the groups and it can correctly predict some DEFAULT cases that the previous SMOTE-less model has mispredicted as NO DEFAULT.

We also made analyses by group of samples for the feature PAY_0. For that, we analyzed the recall by PAY_0 groups of customers for the Random Forest model trained on SMOTE data. That model is performing very well predicting if a person that has a history of delaying the first payment for more than 2 months will DEFAULT, although it mispredicted all the NON DEFAULT cases when that type of delay happened (hence the NON DEFAULT zero recall). This is reflected in Figure 8, where there are no blue bars for the delays beyond 2 months. Another effect of that feature is that the model has a low DEFAULT recall for cases where the customers have paid either on time or in advance (groups 0, -1 and -2).

## 6 CONCLUSION

In this paper, we examined an individual's personal demographics and payment history in order to find the most prevalent factors influencing credit card default. We trained five machine learning models on both the original and over-sampled dataset using SMOTE. The top three models were identified as Random Forest (trained on SMOTE), Random Forest (trained on Original) .67, and Neural Networks (trained on Original), achieving a Macro F1 Score of .69, .67, and .67 respectively

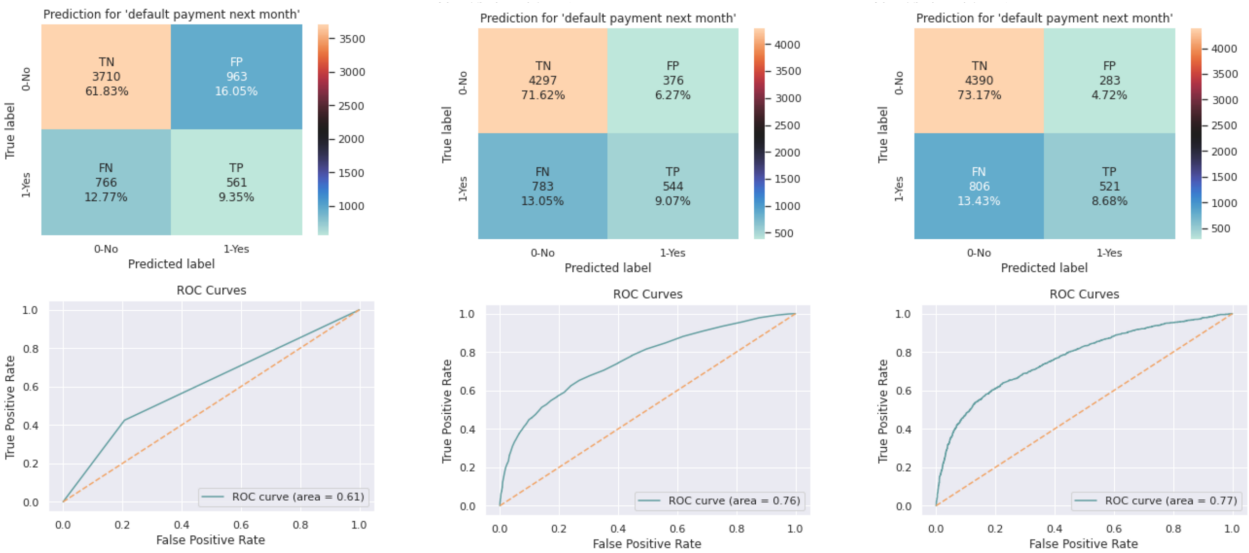
We made more detailed analyses for two features of high importance, PAY_0, and AGE, on the prediction of our best model, the Random Forest model trained on SMOTE data. As a result of that analysis, we realize that for future work, we should focus on improving the F1 macro score for groups of customers that will DEFAULT even if they have duly paid their first credit card bill. We could see by the analyses of the feature PAY_0 that the model needs to improve the performance since it has a low recall for NON DEFAULT cases. Also, we could ensure that the recall by age groups of customers had really improved the F1 macro score with SMOTE when compared to the recall of the model trained on the original data.

| Model | Parameter | Value |
|---|---|---|
| Decision Tree | max_depth | 7 |
| | max_samples_split | 8 |
| | max_samples_leaf | 4 |
| | criterion | entropy |
| | max_features | sqrt |
| Random Forest | n_estimators | 500 |
| | max_depth | 7 |
| | max_samples_split | 8 |
| | max_samples_leaf | 2 |
| | criterion | gini |
| | max_features | sqrt |
| Adaboost | n_estimators | 100 |
| | learning_rate | 1.5 |
| | algorithm | SAMME |
| SVM with RBF Kernel | C | 1 |
| | gamma | 0.1 |
| | probability | 1 |
| Neural Network | hidden_layer_sizes | (20, 10) |
| | max_iter | 700 |
| | learning_rate | invalscaling |
| | alpha | 0.0001 |
| | activation | logistic |

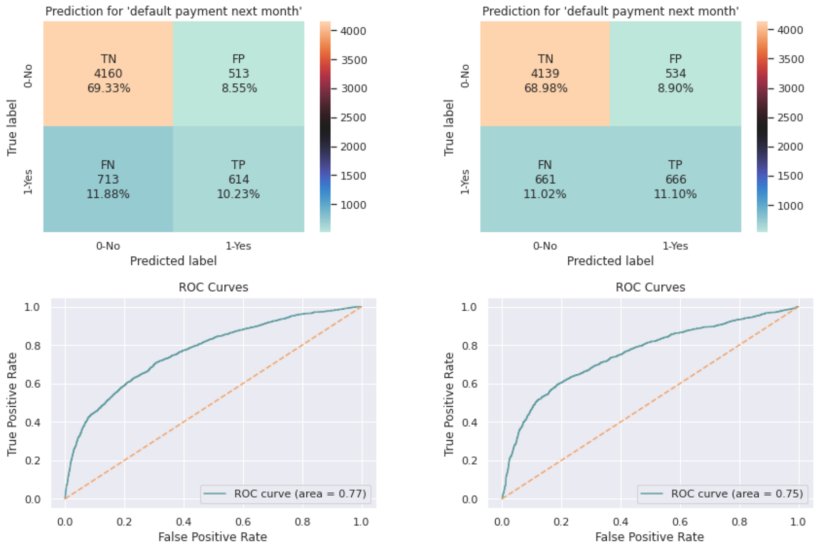
**Table 2: Selected Hyperparameters**

| Classification Model | Sampling Method | F1 (Macro) | Precision (Macro) | Recall (Macro) | Accuracy (Macro) |
|---|---|---|---|---|---|
| Decision Tree | Original | 0.664 +/- 0.019 | 0.73 | 0.63 | 0.805 +/- 0.007 |
| | SMOTE | 0.657 +/- 0.022 | 0.63 | 0.67 | 0.693 +/- 0.027 |
| Random Forest | Original | 0.67 +/- 0.015 | 0.76 | 0.65 | 0.820 +/- 0.009 |
| | SMOTE | 0.69 +/- 0.020 | 0.68 | 0.70 | 0.772 +/- 0.22 |
| Adaboost | Original | 0.668 +/- 0.014 | 0.76 | 0.64 | 0.818 +/- 0.008 |
| | SMOTE | 0.660 +/- 0.017 | 0.65 | 0.69 | 0.735 +/- 0.020 |
| Neural Network | Original | 0.683 +/- 0.015 | 0.76 | 0.66 | 0.820 +/- 0.009 |
| | SMOTE | 0.658 +/- 0.12 | 0.64 | 0.69 | 0.716 +/- 0.024 |
| SVM with RBF Kernel | Original | 0.670 +/- 0.016 | 0.76 | 0.65 | 0.818 +/- 0.009 |
| | SMOTE | 0.65 +/- 0.019 | 0.66 | 0.69 | 0.744 +/- 0.023 |

**Table 3: Comparison Results of Machine Learning Models**

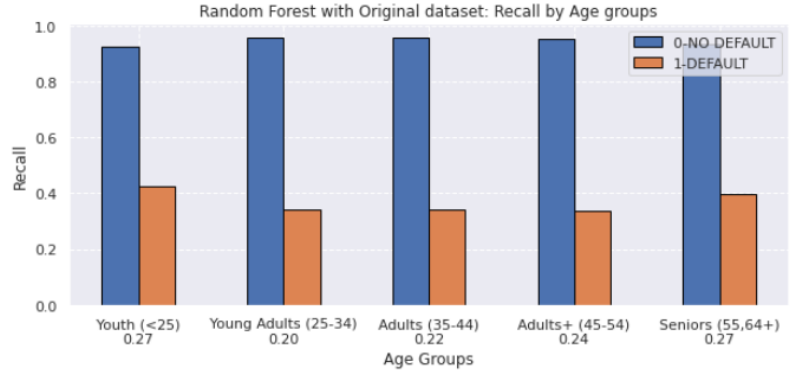



(a) Decision Tree (b) Random Forest (c) Adaboost

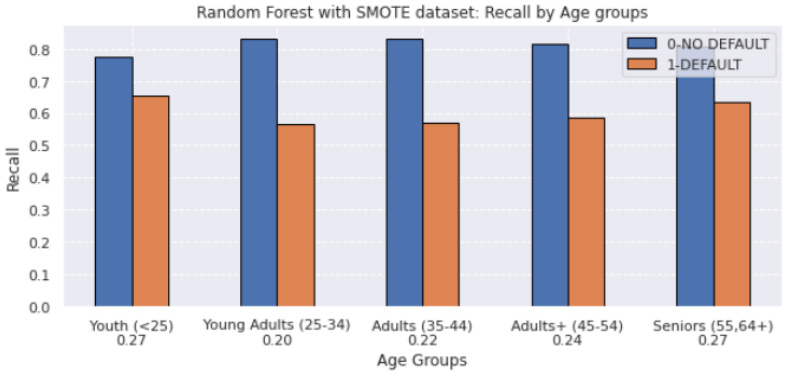



(d) Neural Network (e) SVM with RBF Kernel

**Figure 6: Comparison Results of Machine Learning Models.**




**(a) Random Forest with Original dataset.**




**(b) Random Forest with SMOTE dataset.**

**Figure 7: Recall by Age groups using Random Forest model. (a) Training with Original data and (b) with SMOTE data. At the bottom of each bar graph, a real number $in[0, 1]$ expresses the percentage of group-wise samples belonging to class DEFAULT. Note that all groups are unbalanced, with class DEFAULT being the least represented class.**

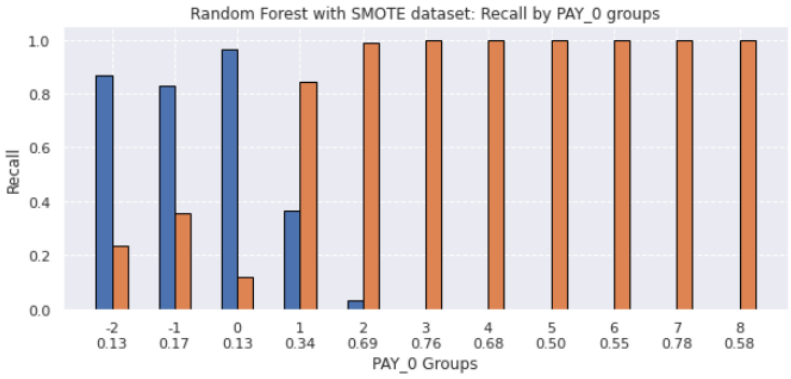



**Figure 8: Random Forest with SMOTE dataset. Recall by PAY_0 groups. At the bottom of each bar graph, a real number $in[0, 1]$ expresses the percentage of group-wise samples belonging to class DEFAULT. Note that all groups are unbalanced, with class DEFAULT being the least represented class.**

# REFERENCES


[1] [n.d.]. sklearn.model_selection.RandomizedSearchCV. https://scikit-learn/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html

[2] 2016. Default of credit card clients. UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients.

[3] 2021. Board of Governors of the Federal Reserve System (US), Delinquency Rate on Credit Card Loans, All Commercial Banks, retrieved from Federal Reserve Bank of St. Louis. https://fred.stlouisfed.org/series/DRCCLACBS.

[4] Ying Chen and Ruirui Zhang. 2021. Research on Credit Card Default Prediction Based on k-Means SMOTE and BP Neural Network. *Complexity* 2021 (2021), 1–13. https://doi.org/10.1155/2021/6618841

[5] I-Cheng Yeh and Che hui Lien. 2009. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications* 36 (2009), 2473–2480.