ISSN: 2473-4901
v10 n6186

A Comparative Analysis of Oversampling Methods for Predicting Credit Card Default with Logistic Regression

Dara Tourt
Dara.tourt@my.metrostate.edu
Department of Management Information Systems
Metropolitan State University
Minneapolis, MN 55403 USA

Queen E. Booker
Queen.booker@metrostate.edu
Department of Management Information Systems
Metropolitan State University
Minneapolis, MN 55403 USA

Carl Rebman
carlr@sandiego.edu
Department of Management Information Systems
University of San Diego
San Diego, CA 92110 USA

Simon Jin
Simon.jin@metrostate.edu
Department of Management Information Systems
Metropolitan State University
Minneapolis, MN 55403 USA

Abstract

In the era of big data, the prevalence of imbalanced datasets has emerged as a significant challenge in machine learning and data analytics. Analysts often employ two primary techniques - undersampling and oversampling - to overcome the imbalance problem. This study explores the multiple oversampling techniques in addressing these imbalances, focusing on how appropriate sampling methods can enhance model performance, improve predictive accuracy, and facilitate better decision-making. The results affirm that oversampling does improve the predictive power for the minority class when compared to building a model with unbalanced data. However, the additional contribution is that the type of balancing technique matters to the overall performance and accuracy of the predictive model.

Keywords: Data Balancing, Predictive Modeling, Logistic Regression, Credit Card Fraud.

1. INTRODUCTION

In the era of big data, the prevalence of imbalanced datasets has emerged as a significant challenge in machine learning and data analytics. Imbalanced datasets occur when one class significantly outnumbers another which is common in financial modeling to address issues such as credit decisions, fraud detection, and default predictions (He & Garcia, 2009). For example, in a dataset used to build models to detect credit card fraud, fraudulent transactions may represent less than 1% of the total data. Traditional classification algorithms, such as logistic regression and support vector machines, often perform inadequately on such datasets because they tend to favor the majority class, leading to high overall accuracy but poor sensitivity for the minority class (Saito & Rehmsmeier, 2015).

Analysts often employ two primary techniques - undersampling and oversampling - to overcome the imbalance problem. Undersampling involves reducing the number of instances in the majority class to create a more balanced dataset. This technique can lead to simpler models that generalize better, as it prevents the model from becoming overwhelmed by the sheer volume of majority class instances (Kotsiantis, 2006; Dube & Verster, 2023). However, undersampling carries the risk of losing potentially valuable information, which can negatively impact model performance (Batista et al., 2004).

Conversely, oversampling increases the number of instances in the minority class. Techniques such as SMOTE (Synthetic Minority Over-sampling Technique) generate synthetic instances based on the existing minority class data, helping to mitigate the risk of overfitting associated with simple duplication of minority instances (Chawla et al., 2002). By enhancing the representation of the minority class, oversampling can significantly improve the model's ability to learn relevant patterns.

Given the emphasis prior research has made regarding the significance of balancing imbalanced datasets, many balancing methods have been introduced to accomplish the goal. However, few studies have compared and contrasted the various methods of balancing datasets and measured the difference in the results of the different approaches. This research study aims to address this gap, applying different oversampling methods to the credit card default problem using a logistic regression model as the comparative tool. We report the difference in Type I and Type II errors, and significant difference between each model built using the different balancing methods. The paper continues with the literature review of oversampling methods, followed by the research methodology, results, conclusions, limitations, and next steps.

2. LITERATURE REVIEW

In the field of machine learning, addressing class imbalance remains a critical challenge that can significantly impact the performance of predictive models. This literature review explores various oversampling techniques reported in the literature and used in data analysis, their theoretical foundations, and their benefits and challenges.

Random Oversampling

Random oversampling serves as a fundamental technique for addressing class imbalance in machine learning. By increasing the representation of minority class instances, random oversampling helps mitigate bias and improve the performance of predictive models on imbalanced datasets. Random oversampling involves randomly duplicating instances from the minority class until a balanced distribution is achieved (Chawla et al., 2002). Random oversampling is easy to implement and does not require complex algorithms or parameter tuning compared to other oversampling techniques. Random oversampling also retains all instances from both classes, thereby preserving the overall information content of the dataset. While simple, it may lead to overfitting and increased computational costs. Random oversampling is based on the premise of increasing the minority class instances randomly until the class distribution is balanced with the majority class. Random oversampling preserves all instances from both classes but duplicates minority class instances. By doing so, it aims to provide the model with more examples of the minority class, thereby reducing bias and improving the model's ability to generalize to minority class instances. (Yang et al, 2024)

Synthetic Minority Over-sampling Technique (SMOTE)

SMOTE generates synthetic instances for the minority class by interpolating between existing instances (Chawla et al., 2002). This technique preserves the underlying data structure better than random oversampling and reduces the risk of overfitting. SMOTE, proposed by Chawla et al. (2002), tackles class imbalance by generating synthetic instances for the minority class. It works by interpolating between existing minority class instances to create new synthetic samples in the feature space, thereby balancing the dataset without blindly duplicating existing data points.

This method is effective in improving the generalization ability of machine learning models by providing more balanced training data. Despite its advantages, SMOTE may struggle with datasets where the minority class is not uniformly distributed or when instances of the minority class overlap with those of the majority class. This can lead to synthetic samples that do not adequately represent the true characteristics of the minority class, potentially affecting model performance. (Kimbrell, 2014)

Synthetic Minority Over-sampling Technique for Nominal and Continuous (SMOTE-NC)

SMOTE-NC is used for datasets that contain both numerical and categorical features. SMOTE-NC first identifies instances belonging to the minority class. For each minority instance, the algorithm calculates the k-nearest neighbors (typically using Euclidean distance). Synthetic samples are generated by taking a minority instance and one of its nearest neighbors. A random point is then created along the line segment joining the two instances. This process is repeated until the desired level of balance is achieved in the dataset. SMOTE-NC uses the most common value among the k-nearest neighbors for categorical attributes when generating synthetic samples. Instead of calculating a weighted average (as with continuous data), it identifies the mode (most frequent category) for categorical features. The synthetic instance is formed by combining the continuous attributes generated as described earlier and the most frequent values for nominal attributes. (Gök et al, 2021)

Safe-level-SMOTE (SL-SMOTE)

SL-SMOTE builds upon the original SMOTE framework by incorporating a safety level mechanism to control the generation of synthetic samples. SL-SMOTE evaluates the density of instances surrounding minority class samples. It utilizes the concept of safe regions, where synthetic samples can be generated without risking overfitting or misclassifying noisy data points. Instead of randomly selecting neighbors as in traditional SMOTE, SL-SMOTE identifies "safe neighbors" that lie within a certain threshold of distance from the minority instance.

This selective approach minimizes the chances of generating synthetic samples that may lead to decision boundary distortions (He & Ma, 2009). Once safe neighbors are identified, synthetic samples are generated similarly to traditional SMOTE, using linear interpolation. By focusing on safe regions for sample generation, SL-SMOTE significantly reduces the risk of overfitting to noisy or outlier data points that may exist within the minority class. Several studies have demonstrated that SL-SMOTE can lead to improved classification performance compared to traditional SMOTE, particularly in scenarios with extreme class imbalance (Shing et al, 2023).

Borderline-SMOTE (BSMOTE)

This variant of SMOTE focuses on generating synthetic instances near the decision boundary between classes (Han et al., 2005). It addresses classification errors that occur near the class boundaries. Borderline SMOTE focuses on generating synthetic samples specifically in the "borderline" areas where the minority class instances are most vulnerable to misclassification (Han et al., 2005). The algorithm first identifies minority class instances that lie close to the decision boundary between classes. These borderline instances are critical because they are often misclassified or underrepresented, making them essential for model training. For each borderline minority instance, the algorithm identifies its k-nearest neighbors within the minority class. The choice of k can be adjusted based on the dataset's characteristics. Synthetic samples are generated by interpolating between a borderline instance and its nearest minority neighbors. The interpolation is performed similarly to traditional SMOTE, using a weighted combination of the instances to create new synthetic samples in the feature space. (Han et al, 2004; Chen et al, 2023).

K-Means SMOTE

K-Means SMOTE is an approach that integrates K-Means clustering with SMOTE to enhance the representation of the minority class. By leveraging K-Means clustering, K-Means SMOTE generates synthetic instances that better capture the distribution and structure of the minority class, leading to improved model performance (Batista et al., 2004). The localized generation of synthetic instances helps in reducing overfitting by preserving the diversity within the minority class and avoiding excessive duplication of instances (Sun et al., 2007). Models trained on datasets augmented with K-Means SMOTE synthetic samples are better able to generalize to unseen data, as they have learned from a more balanced and representative dataset (He & Ma, 2013).

Support Vector Machine SMOTE (SVM SMOTE)

To enhance the performance of predictive models on imbalanced datasets, SVM SMOTE has emerged as an advanced approach that integrates Support Vector Machines (SVM) with SMOTE to improve the representation of minority class instances. SVMs are powerful supervised learning models used for classification tasks. SVM SMOTE integrates the strengths of SVM and SMOTE by selectively applying the oversampling technique to minority class instances that are support vectors or are close to the SVM decision boundary. By focusing on instances near the SVM decision boundary, SVM SMOTE generates synthetic samples that are more relevant to the SVM classifier's learning process, thereby enhancing its ability to generalize (He & Ma, 2013). SVM SMOTE aims to mitigate the impact of class imbalance on SVM classifiers, resulting in improved accuracy and robustness in predicting minority class instances (Sun et al., 2007). Empirical studies and applications across various domains, such as healthcare diagnostics, fraud detection in finance, and image classification, have demonstrated the efficacy of SVM SMOTE in addressing class imbalance and improving predictive model performance (Batista et al., 2004; Zhang & Mani, 2003).

Adaptive Synthetic Sampling (ADASYN)

ADASYN adjusts the density distribution of the minority class by focusing synthetic instance generation on instances that are harder to classify (He & Ma, 2013). It emphasizes regions of the feature space where the classifier performs poorly. ADASYN is another extension of Synthetic Minority Over-sampling Technique (SMOTE), which addresses class imbalance by oversampling the minority class. SMOTE generates synthetic samples along line segments joining minority class instances. However, SMOTE does not consider the distribution of minority class instances, potentially leading to overfitting in dense minority regions and underfitting in sparse regions. ADASYN improves upon SMOTE by adaptively generating synthetic samples based on the density distribution of minority class instances. Specifically, it focuses more on generating samples in regions where the class distribution is sparser, thereby making the classifier more robust and reducing the risk of overfitting. (Mitre et al, 2023)

Self-adaptive Oversampling (SAOM)

The Self-Adaptive Oversampling Method (SAOM) introduces a dynamic approach to the oversampling process, allowing it to adjust based on the characteristics of the data at hand. Unlike static oversampling techniques that apply a uniform strategy across the dataset, SAOM adapts its sampling strategy according to the local distribution of minority and majority classes, thereby enhancing the quality of the synthetic samples generated. SAOM continuously evaluates the data distribution and adjusts the oversampling strategy based on local density estimates.

By incorporating an adaptive mechanism, SAOM strikes a balance between exploring underrepresented regions of the feature space and exploiting areas where the minority class is already well-represented. This dual strategy improves the diversity of synthetic samples while ensuring they remain relevant to the underlying data distribution. Studies have shown that models trained using SAOM exhibit superior performance compared to those utilizing traditional oversampling methods. The self-adaptive mechanism allows SAOM to be tailored to a wide range of applications and datasets, making it a versatile tool in the machine learning toolbox. Its scalability ensures that it can be applied effectively in both small and large datasets. (Tao et al, 2023)

3. RESEARCH METHODOLOGY

The purpose of this study was to evaluate the performance of a predictive modeling technique using different oversampling techniques. The application for the study is logistic regression to predict customer default on credit card payments. There are many machine learning methods used to predict default behavior. Logistic regression was used for this study for exploration of the oversampling method. According to Yeh and Lien (2009) and Sperandei (2014), logistic regression is specifically tailored for scenarios with a binary response variable and is typically the first or baseline technique to compare subsequent models for performance. Logistic Regression’s strength lies in its ability to offer a straightforward probabilistic framework for classification.

The study compares eight different oversampling methods to the unbalanced dataset. The oversampling techniques studied were Random Over-Sampling, SMOTE, SMOTENC, ADASYN, BSMOTE, SVM SMOTE, K-Means SMOTE, and SL-SMOTE.

The research questions for the study were:

1. *Does oversampling improve the performance of the logistic regression predictive model for*

identifying potential credit card accounts that default?

2. *Is there an oversampling method that improves the performance of the logistic regression predictive model for identifying potential credit card accounts that default?*

Based on the literature, oversampling methods improve the performance of data mining algorithms. However, there was no indication through the literature review process that any method significantly outperformed another for the credit card default application. The hypotheses for the study are as follows:

H0: A logistic regression model for predicting credit card payment default built using an unbalanced dataset will not perform significantly better than a model built using a balanced dataset.

H1: When compared to other oversampling methods to balance the dataset, no logistic regression model performs significantly better than another.

All models were built and evaluated using Python. To test the significant difference between each model, a t-test was performed comparing error rates. Each model was evaluated using standard suitability measures. According to the literature, there is general agreement how they are defined and are listed as follows (Chen et al., 2021; Demraoui et al., 2022; Karthiban et al., 2019; Lusinga et al., 2021; Li et al., 2017; Ndayisenga, 2021; Orji et al.; Peiris, 2022; Pimcharee & Surinta, 2022, Booker & Rebman, 2024):

- Accuracy score over 90%
- Specificity score over 85%
- Type I Error score under 10%
- Type II Error score under 10%
- Recall score over 85%
- Precision score over 85%
- F measure score over 85
- AUC near to 1

Dataset

The dataset used in the study contains information on customer default payments in Taiwan. Figure 1 illustrates that the number of accounts not expected to default the following month vastly outnumbers those that are at risk of default and shows the class imbalance between defaults and non-defaults, with 6,636 accounts classified as defaults and 23,364 as non-defaults. It is a multivariate dataset with 30,000 instances and 23 features, including both categorical and nominal data types. The dataset is hosted by the UCI Machine Learning Repository and can be accessed directly at https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients.

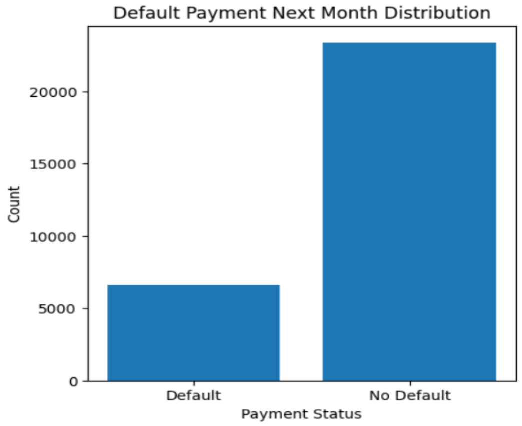

Figure 1: Default Instances in the Dataset

The creators of this dataset, led by I-Cheng Yeh compiled it for business applications, specifically within the subject area of risk management associated with credit card default payments. The dataset does not contain any missing values. Variables were recoded as necessary to ensure categorical data was represented as binary variables. The decision variable was whether a customer defaulted with 1 for defaulted and 0 for non-default.

The variables in the dataset were:

- X1: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.
- X2: Gender (1 = male; 2 = female).
- X3: Education (1 = graduate school; 2 = university; 3 =high school; 4 = others).
- X4: Marital status (1 = married; 2 = single; 3 = others).
- X5: Age (year).
- X6–X11: History of past payment. We tracked the past monthly payment records (from April to September 2005) as follows:
    - X6 = the repayment status in September, 2005; X7=the repayment status in August, 2005;...;X11 = the repayment status in April, 2005.
        - The measurement scale for the repayment status is:

  - -1 =no pay delay
  - 1=payment delay for one month
  - 2 =payment delay for two months...;
  - 8 = payment delay for eight months;
  - 9 = payment delay for nine months and above.

- X12–X17: Amount of bill statement (NT dollar).
  - X12 =amount of bill statement in September, 2005; X13 =amount of bill statement in August, 2005;...;X17 = amount of bill statement in April, 2005.
- X18–X23: Amount of previous payment (NT dollar).
  - X18 =amount paid in September, 2005; X19 = amount paid in August, 2005;...;X23 = amount paid in April, 2005.

Model Development

Each model was built using 10,000 instances. For the unbalanced dataset, all the observations were drawn from the dataset. Following the recommendation of Gholamy et al. (2018), the training used 80% of the dataset. The remaining 20% was used for testing. All models were built using logistic regression which is a widely used statistical method for predictive modeling, particularly suited for binary classification tasks. It models the relationship between one or more independent variables (features) and a binary dependent variable (outcome) using a logistic function. Logistic regression is designed to predict the probability of a binary outcome, typically coded as 0 and 1. Each model, including the application of sampling techniques were built using the Python software application.

4. RESULTS

This section presents the results of the validation stage of the analysis. Each model was applied to the full dataset. Table 1 summarizes the performance metrics of accuracy, precision, recall, and F-measure across the different oversampling methods using the results from the validation of the models.

Based on the results, the unbalanced model appears to perform better than most of the models using oversampling methods. Each of the oversampling methods had at least two measures that met the suitability standards. Random oversampling met all four standards. In comparing the suitability measures, it would seem that the unbalanced trained model and the random oversampling model would provide the best predictive power.

Model	Precision	Recall	Accuracy	F1 Score
Unbalanced	0.8631	0.8917	0.8055	0.8771
Random Over-Sampling	0.9729	0.9649	0.9517	0.9689
SMOTE	0.9658	0.7771	0.8050	0.8612
SMOTENC	0.9677	0.7806	0.8089	0.8642
ADASYN	0.9662	0.7803	0.8076	0.8634
Borderline SMOTE	0.9665	0.7793	0.8070	0.8628
SVM SMOTE	0.9676	0.7774	0.8064	0.8621
KMeans SMOTE	0.9667	0.7757	0.8045	0.8607
SL-SMOTE	0.9066	0.7803	0.7662	0.8387

Table 1: Validation Data Results Logistic Regression on Various Over-sampling Methods to Deal with Imbalance Class (Default, Non-Default)

However, a review of the confusion matrices in Figures 2 through 10 show that the unbalanced model predicts the majority instances well but falters when predicting the defaults, providing a 50/50 predictive power. For credit card default, a client is likely interested in having more potential default cases predicted than fewer.

In examining the matrices, the unbalanced model has the worst performance with regards to correctly identifying default instances, predicting approximately 50% of the instances correctly. The best method of those tested was SVM-SMOTE, correctly identifying more than 90% of the default instances. However, the model with the best predictive power for the majority class was random oversampling.

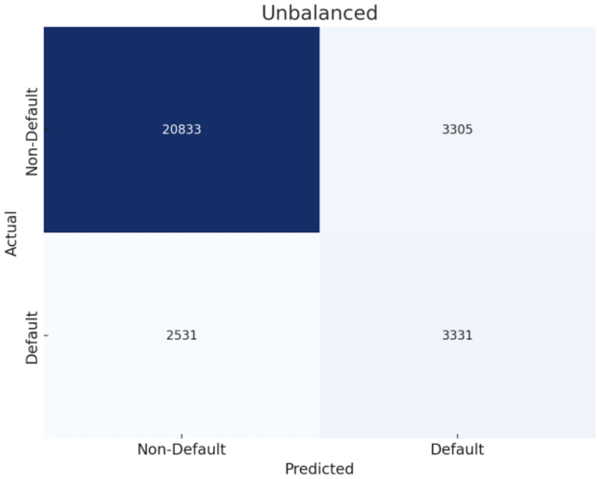

Figure 2: Unbalanced Confusion Matrix

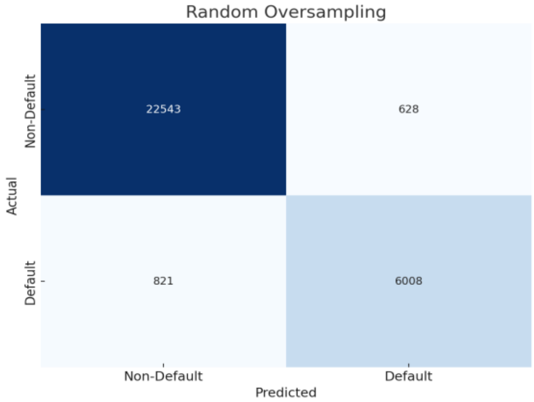

Figure 3: Random Oversampling Confusion Matrix

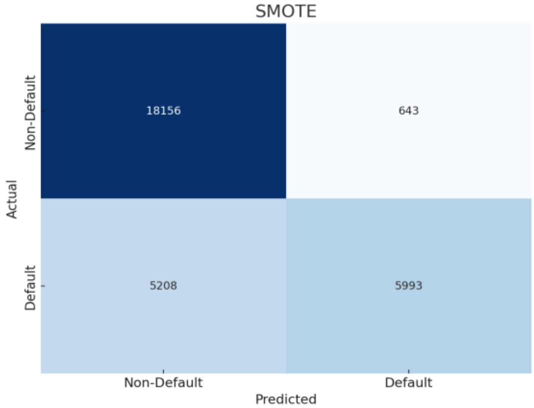

Figure 4: SMOTE Confusion Matrix

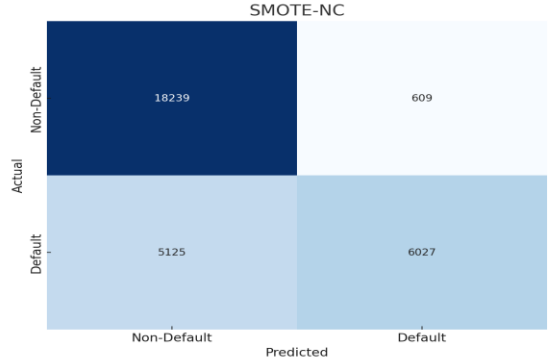

Figure 5: SMOTE-NC Confusion Matrix

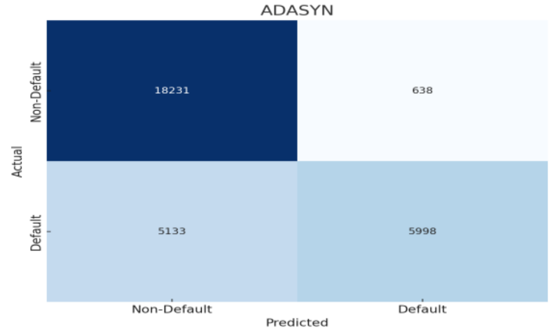

Figure 6: ADASYN Confusion Matrix

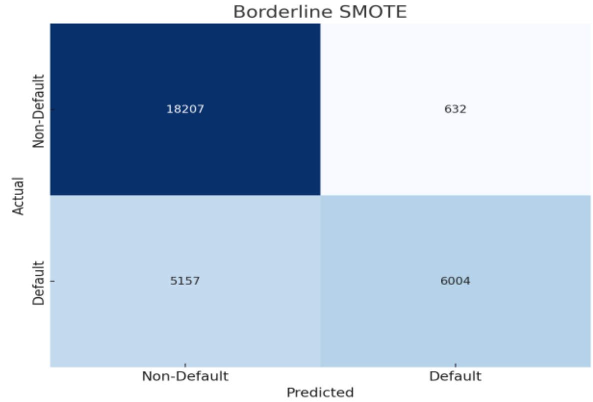

Figure 7: Borderline SMOTE Confusion Matrix

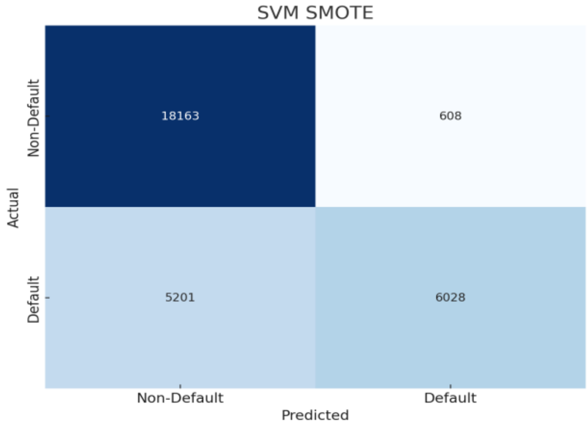

Figure 8: SVM SMOTE Confusion Matrix

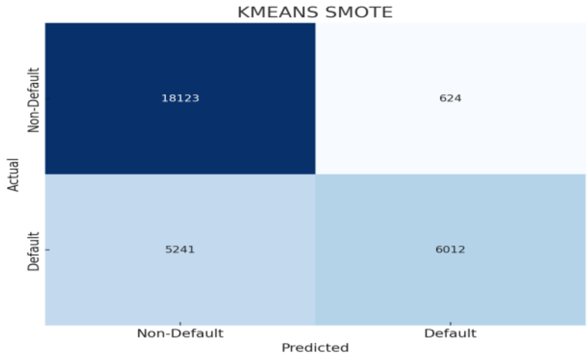

Figure 9: KMEANS SMOTE Confusion Matrix

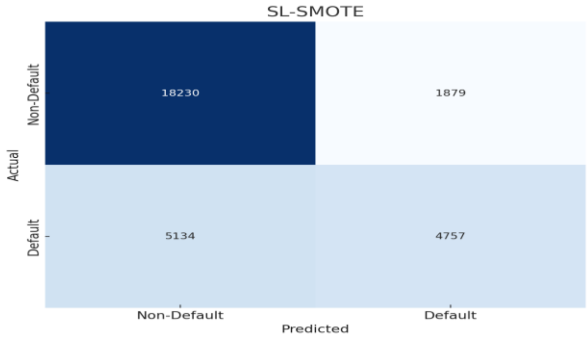

Figure 10: SL-SMOTE Confusion Matrix

The suitability measures and the confusion matrices indicate that the oversampling models perform better than the unbalanced model when predicting default instances. The next step was to determine if the differences were significant. Paired t-tests were performed for the unbalanced model and each of the oversampling models, and between each of the oversampling models. The results are shown in Table 2 in Appendix A.

Based on the t-tests, all the models that used oversampling methods performed with significant difference from the model using an unbalanced dataset when evaluating the prediction for default. Within the oversampling methods, SL-SMOTE was significantly different from the other methods, with SL-SMOTE performing worse rather better.

The final step in the analysis was to evaluate the hypotheses and research questions. Recall the primary research hypotheses:

H0: *A logistic regression model for predicting credit card payment default built using an unbalanced dataset will not perform significantly better than a model built using a balanced dataset.*

H1: *When compared to other oversampling methods to balance the dataset, no logistic regression model performs significantly better than another.*

H0 is accepted because all the oversampling models performed better, based on the t-test results. H1 is partially accepted as the SL-SMOTE performed better than the other models.

When returning to the research questions RQ1 "*does oversampling improve the performance of the logistic regression predictive model for identifying potential credit card accounts that default?" and RQ2 "is there an oversampling method that improves the performance of the logistic regression predictive model for identifying potential credit card accounts that default?"*, the results indicate that oversampling does improve the performance of the logistic regression predictive model for identifying potential credit card accounts that default and of the oversampling methods tested, all the other models performed better than the SL-SMOTE method.

In summary, the results indicate that there is value in comparing data balancing methods when developing predictive modeling as such a comparison can improve the performance of the predictive model.

5. LIMITATIONS AND CONCLUSIONS

This study examined only oversampling methods in the context of predicting credit card default for a specific dataset using a specific modeling method-logistic regression. The results of the study cannot be generalized as there are many factors to consider when building predictive models including but not limited to the variables, data balancing methods, and predictive modeling techniques. Therefore, additional analysis is needed to determine the conditions best suited for each sampling method, dataset configuration, and predictive modeling tool.

However, oversampling techniques represent a critical approach to addressing class imbalance in data analysis. While each technique has its strengths and weaknesses, their application depends heavily on the specific characteristics of the dataset and the objectives of the analysis. Continued research and development in this area aim to improve the robustness, scalability, and applicability of oversampling methods across diverse domains and applications in machine learning and statistical modeling. Oversampling serves as a viable strategy to address the challenges posed by imbalanced datasets. The selection of the appropriate method hinges on the specific requirements of the task, the nature of the dataset, and the criticality of predictive accuracy in the minority class. As machine learning continues to evolve, ongoing research into sampling approaches that combine the strengths of multiple methods may provide further avenues for improvement in managing imbalanced datasets.

6. REFERENCES

Batista, G. E. A. P. A., Monard, M. C., & Silva, J. C. P. (2004). A study of data preprocessing and classifiers for imbalanced datasets. *Proceedings of the Brazilian Conference on Neural Networks*. https://doi.org/10.5220/0005201103820389

Batista, G.E., Prati, R. C. & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data," ACM SIGKDD Explorations Newsletter, vol. 6, no. 1, pp. 20–29. https://doi.org/10.1145/1007730.1007735

Booker, Q. & Rebman, C. (2024). Applying Heterogeneous Ensemble Models to Detect Credit Card Fraudulent Transactions. *Proceedings of the 2024 Southwest Decision Sciences Conference (SWDSI),* Galveston, TX.

Bunkhumpornpat, C., Sinapiromsaran, K., & Lursinsap, C. (2009). Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In *Advances in knowledge discovery and data mining: 13th Pacific-Asia conference, PAKDD 2009 Bangkok, Thailand, April 27-30, 2009 proceedings 13* (pp. 475-482). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-01307-2_43

Bunkhumpornpat, C., Sinapiromsaran, K., & Lursinsap, C. (2011) DBSMOTE: Density-based synthetic minority over-sampling technique. Applied Intelligence, vol. 36, pp. 1–21. https://doi.org/10.1007/s10489-011-0287-y

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321-357. https://doi.org/10.1613/jair.953

Chen, C., Shen, W., Yang, C., Fan, W., Liu, X., & Li, Y. (2023). A New Safe-Level Enabled Borderline-SMOTE for Condition Recognition of Imbalanced Dataset. *IEEE Transactions on Instrumentation and Measurement*, *72*, 1–10. https://doi.org/10.1109/TIM.2023.3289545

Chen, R.C., Luo, S.T., Liang, X., and Lee, V. (2005). Personalized Approach Based on SVM and ANN for Detecting Credit Card Fraud. *International Conference on Neural Networks and Brain, IEEE,* 810-815. https://doi.org/10.1109/icnnb.2005.1614747

Demraoui, L., Eddamiri, S., & Hachad, L. (2022). Digital Transformation and Costumers Services in Emerging Countries: Loan Prediction Modeling in Modern Banking Transactions. *Lecture Notes on Data Engineering and Communications Technologies*, 627–642. https://doi.org/10.1007/978-3-030-90618-4_32

Dube, L. & Verster, T. (2023). Enhancing classification performance in imbalanced datasets: A comparative analysis of machine learning models. Data Science in Finance and Economics. 3. 354-379. https://doi.org/10.3934/DSFE.2023021.

Fernández, A., García, S., Galar, M., Prati, R.C., Krawczyk, B., Herrera, F. (2018). Foundations on Imbalanced Classification. In: Learning from Imbalanced Data Sets. Springer, Cham.

https://doi.org/10.1007/978-3-319-98074-4_2

Gholamy, A., Kreinovich, V., and Kosheleva, O. (2018) "Why 70/30 or 80/20 Relation Between Training and Testing Sets: A Pedagogical Explanation" (2018). Departmental Technical Reports (CS). 1209. https://scholarworks.utep.edu/cs_techrep/1209

Gök, E. C., & Olgun, M. O. (2021). SMOTE-NC and gradient boosting imputation based random forest classifier for predicting severity level of covid-19 patients with blood samples. *Neural Computing & Applications*, *33*(22), 15693–15707. https://doi.org/10.1007/s00521-021-06189-y

Han, H., Wang, W. Y., & Mao, B. H. (2005) Borderline-smote: A new over-sampling method in imbalanced data sets learning,” in Advances in Intelligent Computing, (Hefei, China), vol. 3644, pp. 878–887, Springer-Verlag. https://doi.org/10.1007/11538059_91

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263-1284. https://doi.org/10.1109/tkde.2008.239

He, H., & Ma, Y. (Eds.). (2013). Imbalanced learning: foundations, algorithms, and applications. https://doi.org/10.1002/9781118646106

Karthiban, R., Ambika, M., & Kannammal, K. E. (2019). A Review on Machine Learning Classification Technique for Bank Loan Approval. *International Conference on Computer Communication and Informatics*. https://doi.org/10.1109/iccci.2019.8822014

Kimbrell, J. (2014). Smote. *Ploughshares*, *40*(1), 137–138. https://doi.org/10.1353/plo.2014.0004

Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. *Informatica*, 30(3), 301-320. https://doi.org/10.1007/s10462-007-9052-3

Li, Z., Tian, Y., Li, K., Zhou, F., & Yang, W. (2017). Reject Inference in Credit Scoring Using Semi-Supervised Support Vector Machines. *Expert Systems with Applications*, *74*, 105–114. https://doi.org/10.1016/j.eswa.2017.01.011

Lusinga, M., Mokoena, T., Modupe, A., & Mariate, V. (2021). Investigating Statistical and Machine Learning Techniques to Improve the Credit Approval Process in Developing Countries. *AFRICON*. https://doi.org/10.1109/africon51333.2021.9570906

Mitra, R., Bajpai, A., & Biswas, K. (2023). ADASYN-assisted machine learning for phase prediction of high entropy carbides. *Computational Materials Science*, *223*, 112142-. https://doi.org/10.1016/j.commatsci.2023.112142

Naboureh, A.; Li, A.; Bian, J.; Lei, G.; Amani, M. A Hybrid Data Balancing Method for Classification of Imbalanced Training Data within Google Earth Engine: Case Studies from Mountainous Regions. *Remote Sens.* 2020, *12*, 3301. https://doi.org/10.3390/rs12203301

Ndayisenga, T. (2021). Bank Loan Approval Prediction Using Machine Learning Techniques. *[Doctoral dissertation, University of Rwanda]*. http://www.dr.ur.ac.rw/handle/123456789/1437

Orji, U. E., Ugwuishiwu, C. H., Nguemaleu, J. C. N., & Ugwuanyi, P. O. (2022). Machine Learning Models for Predicting Bank Loan Eligibility. *2022 IEEE Nigeria 4th International Conference on Disruptive Technologies for Sustainable Development (NIGERCON)*. https://doi.org/10.1109/nigercon54645.2022.9803172

Peiris, M. P. C. (2022). *Credit Card Approval Prediction by Using Machine Learning Techniques* [Doctoral dissertation, University of Colombo School of Computing]. https://dl.ucsc.cmb.ac.lk/jspui/handle/123456789/4593

Pimcharee, K., & Surinta, O. (2022). Data Mining Approaches in Personal Loan Approval. *Engineering Access*, *8*(1), pp. 15-21. doi: 10.14456/mijet.2022.2. https://ph02.tcithaijo.org/index.php/mijet/article/view/244392

Saito, K., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Shing, T. W., Sudirman, R., Daud, S. N. S. S., Razak, M. A. A., Zakaria, N. A., & Mahmood, N. H. (2023). Multistage Anxiety State Recognition based on EEG Signal using Safe-Level SMOTE. *Journal of Physics. Conference Series*, *2622*(1), 12010-. https://doi.org/10.1088/1742-6596/2622/1/012010

Sperandei S. (2014). Understanding Logistic Regression analysis. *Biochemia medica*, *24*(1), 12–18. https://doi.org/10.11613/BM.2014.003

Sun, L., Hu, N., Ye, Y., Tan, W., Wu, M., Wang, X., & Huang, Z. (2022). Ensemble stacking rockburst prediction model based on Yeo–Johnson, K-means SMOTE, and optimal rockburst feature dimension determination. *Scientific Reports*, *12*(1), 15352–15352. https://doi.org/10.1038/s41598-022-19669-5

Tao,X., Guo,X., Zheng, Y., Zhang,X, & Chen, Z. (2023) Self-adaptive oversampling method based on the complexity of minority data in imbalanced datasets classification. Know.-Based Syst. 277, C. https://doi.org/10.1016/j.knosys.2023.110795

Yeh, I., & Lien, C. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Syst. Appl*., 36, 2473-2480. https://doi.org/10.1016/j.eswa.2007.12.020

Yang, C., Fridgeirsson, E. A., Kors, J. A., Reps, J. M., & Rijnbeek, P. R. (2024). Impact of random oversampling and random undersampling on the performance of prediction models developed using observational health data. *Journal of Big Data*, *11*(1), 7–17. https://doi.org/10.1186/s40537-023-00857-7

ISSN: 2473-4901
v10 n6186

Appendix A

Model	Unbalanced	Random Over-Sampling	SMOTE	SMOTENC	ADASYN	Borderline SMOTE	SVM SMOTE
Random Over-Sampling	56.544 *						
SMOTE	56.239 *	0.444					
SMOTENC	57.518 *	0.572	1.006				
ADASYN	56.357 *	0.295	0.146	0.864			
Borderline SMOTE	57.346 *	0.118	0.323	0.689	0.177		
SVM SMOTE	57.173 *	0.598	1.032	0.030	0.905	0.720	
KMeans SMOTE	56.745 *	0.118	0.569	0.453	0.418	0.238	0.578
SL-SMOTE	25.838 *	28.810 *	28.134 *	29.390 *	28.612 *	28.737 *	39.647 *

Table 2: T-test Results Comparing Accuracy of the Models