**Predictive Analytics for Credit Default Risk Management**

BY

**Venkata siva ram pande Kalivela**,

MSc in economics, University of North Texas.

**Sri sai Durga Katreddi**,  
MSc in Data Science, University of North Texas

30 JUNE 2024.

**Introduction**

This project is fundamentally aimed at predicting whether customers will default on their upcoming payments, an endeavor that is vitally important for financial stability and economic health. By leveraging a variety of demographic and financial behavior factors, from age and education to credit card payment histories, we provide financial institutions with the tools to proactively manage risk and reduce financial losses. Our interdisciplinary team, consisting of risk analysts, economists, and data scientists, has applied advanced machine learning algorithms—Random Forest, Gradient Boosting, and Support Vector Machines (SVM)—to create a robust predictive model. Following extensive data collection and preparation, our team refined a dataset that effectively supports our analytical needs. This process involved rigorous data cleaning and the conversion of categorical variables into dummy variables suitable for machine learning applications. Our best-performing model, the Random Forest, achieved an impressive accuracy of 81%, with an Area Under the Curve (AUC) of 0.78 and an F-measure of 0.79, which are strong indicators of the model’s predictive performance.

During the project, we explored using the SMOTE method to tackle the challenge of class imbalance, a common issue that can skew predictive accuracy in such applications. However, concerns about potential overfitting and misleadingly high accuracy figures led us to refine our methodology, ensuring the integrity and reliability of our predictions.

The project culminated with the successful deployment of our model into a production environment, enabling real-time analysis and decision-making that help financial institutions anticipate and mitigate potential defaults. This proactive approach not only aids individual institutions but also contributes significantly to broader economic stability, showcasing the transformative potential of applying data science to financial risk management.

**Data Description**

The dataset used in this project, titled "Default of Credit Card Clients," was sourced from the reputable UCI Machine Learning Repository, ensuring a reliable foundation for our analyses. Originally containing 30,000 samples, this dataset includes 23 independent variables that provide a comprehensive view of individual credit card clients. These variables encompass a range of demographic and financial information, such as age, education, payment history, and monthly bill amounts, which are crucial for predicting credit card payment defaults.

During the data preparation phase, our team undertook meticulous cleaning efforts to ensure data accuracy and reliability. This process involved the removal of records with missing values and the identification and exclusion of outliers. Outliers, which can result from data entry errors or unusual but genuine variations in credit behavior, were removed to prevent potential distortions in our predictive models. After these adjustments, the dataset was refined to 22,455 samples, maintaining the original 23 variables. This cleaned and processed dataset not only aids in training robust machine learning models but also enhances the reliability of our predictions, providing valuable support for financial institutions in their risk management strategies.  
  
Dataset link:   
<https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients>.

**Data Processing**

After the initial cleaning phase, we conducted thorough data visualizations for each variable to identify any visible abnormalities. During this examination, we noticed irregularities in the "EDUCATION" variable, which contained unnecessary values such as 0, 5, and 6, despite the expected categories being 1 (graduate school), 2 (university), 3 (high school), and 4 (others). We removed rows containing these erroneous entries. Similarly, the "MARRIAGE" variable included some instances of value 0, which did not comply with the defined categories (1 = married, 2 = single, 3 = others); these rows were also excluded.

For the variables ranging from "PAY\_0" to "PAY\_6," we identified values of -2, which did not align with the expected range where -1 indicates pay duly; values from 1 to 9 indicate varying degrees of payment delay. In response, we standardized these entries by converting all instances of -1 (pay duly) to 0, thus creating a consistent, ordered sequence indicating no delay in repayment status. Additionally, we removed negative values from the "BILL\_AMT" and "PAY\_AMT" variables, as negative amounts in these contexts were illogical.

Transformations were applied to certain variables to better suit machine learning models. The "SEX" variable was transformed into a dummy variable named "Sex\_dummy." The "MARRIAGE" variable was divided into three dummy variables: "MARRIAGE\_1," "MARRIAGE\_2," and "MARRIAGE\_3," to effectively handle its categorical nature. The continuous variable "LIMIT\_BAL" was log-transformed into "LIMIT\_BAL\_log" to normalize its distribution.

Upon reassessment, we performed a correlation matrix analysis, particularly noting that the "PAY" variables showed strong correlations due to their reliance on historical payment behavior. With the data now fully prepared and cleansed of inaccuracies and inconsistencies, we will proceed to the model building and validation phases. This next stage involves developing and tuning various predictive models to determine which best captures the complexities of our dataset while maintaining robustness and accuracy in predictions.

**Model Development and Validation.**

During the model development phase, we initially applied the Random Forest algorithm and evaluated its performance through metrics such as ROC-AUC. The importance of each variable was assessed, revealing that features like "MARRIAGE" and "Sex\_dummy" had low feature importance, contributing less than 1% each. Based on this insight, these variables were removed to streamline the model. Satisfied with the initial results yet aiming for further refinement, we employed techniques such as Grid Search to optimize the model's parameters and Cross-Validation to ensure the model's performance was consistent across different data subsets. This approach led to an improved accuracy of 0.81 and an AUC of 0.80, indicating slight benefits from the model refinement process.

Subsequently, we explored another algorithm, Gradient Boosting, which resulted in a slightly lower performance with an accuracy of 0.80 and an AUC of 0.79 compared to the Random Forest. Considering the potential to enhance performance due to class imbalance, we experimented with the SMOTE technique, which adjusted the class distribution in our training dataset from a heavily imbalanced state (Class 0: 639, Class 1: 61) to an equal distribution (Class 0: 639, Class 1: 639). This experimentation with SMOTE led to exceptionally high results in subsequent testing: the Random Forest model reached an accuracy of 95.33%, an AUC of 1.00, and similar high performance was observed with the Support Vector Machine (SVM), which achieved an accuracy of 95.67%. However, a cross-validation with 100 folds revealed a surprisingly low standard deviation of 0.0268, prompting concerns about potential data leakage or overfitting. Further checks, such as analyzing training accuracy (1.00) against testing accuracy (0.9922), confirmed these suspicions, indicating an unusually high prediction accuracy that could be suspect.

Upon reviewing the correlation of variables with the dependent variable, results appeared normal, yet the exceptional performance metrics led us to suspect overfitting. Given these considerations, we decided to deploy the Random Forest model without the use of SMOTE, prioritizing reliability and generalizability over the maximization of accuracy metrics that could potentially be misleading due to overfitting. This decision underscores our commitment to deploying a robust and trustworthy model that genuinely reflects the underlying data patterns without being overly tailored to the training set.  
  
**Model Deployment:**   
The deployment section of our project involves setting up a Flask web application that allows users to interact with a pre-trained Random Forest model to make predictions. The application includes an HTML form where users can input various features required by the model, such as the amount of credit given, education level, age, repayment status for several months, bill amounts, and payment amounts.

Upon submission, the form data is validated and processed to ensure that all inputs are in the correct format and within the expected range. For instance, the education input must be one of the predefined categories, and repayment statuses must fall within a specific range. The validated inputs are then transformed as necessary (e.g., logarithm of the credit amount) and scaled using a pre-fitted scaler before being fed into the Random Forest model for prediction.

The prediction result is displayed back to the user on the same page. The entire Flask application runs in a separate thread to ensure smooth and continuous operation. This deployment method enables a user-friendly and interactive way to leverage our machine learning model for real-time predictions, providing a practical interface for end-users to engage with our solution effectively.

**Conclusion:**

In this project, we successfully developed and deployed a predictive model to determine the likelihood of credit card default among customers, addressing a critical need for enhanced risk management in the financial sector. By integrating a variety of demographic and financial data points into our analyses, we were able to provide financial institutions with robust tools for proactive risk assessment and decision-making.

Throughout the process, we engaged in extensive data cleaning and preparation to ensure the accuracy and relevance of our models. This meticulous approach enabled us to refine our dataset effectively, resulting in a solid foundation for the application of advanced machine learning algorithms. The Random Forest algorithm, in particular, demonstrated strong performance with an accuracy of 81% and an AUC of 0.80, bolstered by further refinement through techniques like Grid Search and Cross-Validation. Although alternative algorithms like Gradient Boosting and the application of SMOTE were explored, the latter raised concerns about potential overfitting, leading us to opt for the more reliable and generalizable model setup without SMOTE for final deployment. The deployment of our model through a Flask-based web application marks a significant achievement in making sophisticated risk assessment tools accessible to end-users. This platform not only facilitates real-time predictions but also underscores the practical applicability of our research and development efforts.

As we look to the future, the scope for further enhancements remains vast. We are committed to continuous improvement of our models and the exploration of new methodologies that could provide even more precise predictions and richer insights into customer behavior. The ongoing evolution of machine learning and its applications in finance promises great potential for transforming how institutions manage risk and enhance their operational efficiency.

This project not only supports the financial sector but also contributes to broader economic stability by improving the predictability and management of financial risks. We are proud to have delivered a solution that not only meets the technical demands of risk analysis but also aligns with the strategic needs of economists and financial institutions globally.  
  
  
**References:**

* Yeh,I-Cheng. (2016). Default of Credit Card Clients. UCI Machine Learning Repository. <https://doi.org/10.24432/C55S3H>.
* . A. Subasi and S. Cankurt, "Prediction of default payment of credit card clients using Data Mining Techniques," 2019 International Engineering Conference (IEC), Erbil, Iraq, 2019, pp. 115-120, doi: 10.1109/IEC47844.2019.8950597. keywords: {Default of credit card payment;Artificial Neural Networks (ANN);Decision Trees;Random Forest}.
* Rigatti, Steven J. "Random forest." Journal of Insurance Medicine 47.1 (2017): 31-39.
* Speiser, Jaime Lynn, et al. "A comparison of random forest variable selection methods for classification prediction modeling." Expert systems with applications 134 (2019): 93-101.
* Natekin, Alexey, and Alois Knoll. "Gradient boosting machines, a tutorial." Frontiers in neurorobotics 7 (2013): 21.
* Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., Aljaaf, A.J. (2020). A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science. In: Berry, M., Mohamed, A., Yap, B. (eds) Supervised and Unsupervised Learning for Data Science . Unsupervised and Semi-Supervised Learning. Springer, Cham. <https://doi.org/10.1007/978-3-030-22475-2_1>
* McFadden, Daniel L. "Econometric analysis of qualitative response models." *Handbook of econometrics* 2 (1984): 1395-1457.
* Myers, Robert J. "Time series econometrics and commodity price analysis: a review." Review of Marketing and Agricultural Economics 62.2 (1994): 167-181.
* Hsu, Chih-Wei, Chih-Chung Chang, and Chih-Jen Lin. "A Practical Guide to Support Vector Classification." BJU International, vol. 101, no. 1, 2008, pp. 1396-1400.
* Chang, Jun-Jie, Zhifu Mi, and Yi-Ming Wei. "Temperature and GDP: A review of climate econometrics analysis." *Structural Change and Economic Dynamics* 66 (2023): 383-392.