**Team 004 Progress Report**

GitHub Link: [MGT-6203-Spring-2024-Canvas/Team-4: Team 4's group project GitHub repository for MGT 6203 (Canvas) Spring of 2024 semester. (gatech.edu)](https://github.gatech.edu/MGT-6203-Spring-2024-Canvas/Team-4)

**Enhancing the Dataset:**

In response to the valuable feedback emphasizing the need for a more complex dataset, our team has strategically augmented our data assets to deepen the analytical depth and broaden the scope of our predictive models. Recognizing the multifaceted nature of loan default risk, we have undertaken the following enhancements to ensure our dataset meets the project's rigor:

1. **Integration of Comprehensive Economic Indicators and Borrower Behavior Metrics:**
   * We have also incorporated detailed borrower behavior metrics from credit bureaus, encompassing transaction histories, payment patterns, and overall financial behavior. These metrics provide a granular view of individual financial habits, which is crucial for predicting default risks.
2. **Advanced Data Manipulation and Variable Creation:**
   * Our data manipulation efforts have been extensive, involving merging core datasets such as **application\_train**, **bureau**, and **bureau\_balance**, with a deliberate focus on maintaining data integrity and relevance. Adding the **previous\_application** dataset has been instrumental in capturing applicants' historical loan behaviors, offering insights into their financial reliability.
   * Recognizing the importance of nuanced financial indicators, we created an interaction term for income categories derived from the U.S. Census Bureau's guidelines. This categorization allows us to explore the relationship between income levels and default risks in a structured manner.
   * We introduced a novel variable, the debt-to-income ratio, standardized across different income levels to assess its impact on default risk. This variable provides a relative measure of financial burden, enhancing our model's ability to identify risk patterns across diverse income brackets.
   * We identified a critical issue where individuals lacking credit reports from any of the three major credit agencies were excluded from our dataset. To address this, we introduced a binary column named 'no\_credit'. This column assigns a value of '1' to individuals without credit information, and '0' to those whose credit data is available through at least one agency. This adjustment ensures the inclusion of all individuals in our analysis, enhancing the comprehensiveness and diversity of our dataset.
3. **Rigorous Data Quality Assurance:**
   * Our data enhancement process has been accompanied by stringent data cleaning and preprocessing measures to ensure the highest quality and completeness of the augmented dataset. This includes addressing missing values, outliers, and inconsistencies and ensuring our models are built on reliable data.

These comprehensive data enhancement strategies have significantly increased our dataset's complexity, ensuring it meets the project's analytical rigor. These enhancements address the feedback received and fortify our project's foundation, enabling more sophisticated analyses and robust predictive modeling.

**Refining Business Justification:**

In refining our business justification, we emphasize the multifaceted benefits of accurately predicting loan defaults, which extend beyond a mere reduction in credit losses to encompass strategic business advantages and societal impacts.

1. **Economic Ramifications:**
   * Beyond the direct financial benefit of minimizing credit losses, accurately predicting loan defaults supports more efficient capital allocation. Financial institutions can optimize their reserve holdings, adhering to regulatory requirements while freeing up capital for additional lending opportunities and enhancing profitability.
   * Moreover, the predictive insights gained can inform pricing strategies, allowing institutions to adjust interest rates based on risk profiles, balancing competitiveness with risk management.
2. **Strategic Business Advantages:**
   * Enhanced risk prediction enables the development of more nuanced borrower risk profiles, facilitating the creation of personalized loan products. This customization can cater to a broader range of customer needs, potentially unlocking new customer segments and market opportunities.
   * The ability to accurately assess risk also enhances an institution's reputation for responsible lending, fostering trust and loyalty among customers and stakeholders, which is invaluable in today's competitive financial landscape.
3. **Societal Impacts:**
   * By ensuring that loan default predictions are fair and unbiased, our project aligns with broader societal goals of financial inclusion and equality. This approach mitigates the risk of discriminatory lending practices, contributing to a more equitable financial ecosystem.
   * Furthermore, by reducing the incidence of loan defaults, the project can contribute to financial stability for individuals and communities, reducing the social and economic stresses associated with debt and default.
4. **Innovative Data Utilization and Ethical Considerations:**
   * Our project's commitment to incorporating diverse and complex datasets while addressing the challenges of imbalanced data showcases our dedication to innovation and ethical responsibility in data analysis. By leveraging advanced analytics and machine learning techniques, we aim to set new standards in predictive modeling within the financial sector.

In summary, the detailed business justification for our project encompasses not only the direct financial benefits of reducing loan defaults but also the strategic and societal advantages of more accurate and ethical risk prediction. These benefits collectively underscore the value and necessity of our project in today's financial landscape.

**Addressing Imbalanced Data:**

In addressing the critical challenge of imbalanced data inherent in loan default predictions, we are implementing a comprehensive suite of techniques to ensure our models are robust, unbiased, and capable of accurately predicting rare default events. Our multifaceted approach includes the following strategies:

1. **PCA (Principal Component Analysis) for Feature Relevance:**
   * We are employing PCA to distill our dataset to its most informative features. This reduction not only mitigates the risk of overfitting but also helps reveal the underlying structure of the data, allowing our models to focus on the most impactful predictors.
2. **VIF (Variance Inflation Factor) Test to Reduce Collinearity:**
   * Recognizing that high collinearity can distort the predictive power of our models, we are using the VIF test to identify and mitigate multicollinearity among predictors. This ensures that our models are built on independent variables that provide unique and valuable information.
3. **Incorporation of Interaction Terms and Polynomial Features:**
   * Understanding that relationships in financial data can be complex and non-linear, we include interaction terms, particularly with income levels, to capture the nuanced effects of various predictors on loan default risk. Moreover, polynomial terms for variables like 'age of car' are introduced to model quadratic relationships, enhancing our models' ability to capture real-world complexities.
4. **Outlier Management:**
   * To ensure outliers do not skew our models, we have instituted a 'child cap' at seven children for the variable indicating the number of children. This approach moderates the influence of extreme values while retaining the integrity of our data.
5. **Advanced Modeling with XGBoost:**
   * XGBoost, renowned for its performance and flexibility, is utilized for its inherent capability to incorporate regularization techniques like Lasso and Ridge regression. This not only helps in feature selection but also in preventing overfitting, making our models more generalizable and robust.

Through these strategic interventions, we aim to comprehensively address the imbalanced data challenge, ensuring our predictive models are equitable, accurate, and reflective of the true dynamics of loan default risks. Each technique has been carefully chosen for its proven effectiveness in similar analytical contexts, and together, they form a robust methodology poised to deliver insightful and reliable predictions.

**Progress on Methodology:**

Our team has successfully completed a thorough exploratory data analysis in the methodology domain, ensuring our dataset's integrity and suitability for advanced modeling. We have pinpointed critical predictors such as income, credit amounts, and employment tenure, which empirical evidence suggests play pivotal roles in loan default likelihood. Our explorative journey into various classification models has led us to focus on Logistic Regression and Random Forests for their balance between interpretability and predictive power. A notable advancement in our methodology is the integration of interaction terms, particularly within income brackets, to capture the nuanced effects of financial standing on default risks. This refinement is in direct response to our analysis of median household income data, which informed our segmentation strategy.

**Timeline Adjustments:** Our project timeline remains on track, with key milestones achieved in dataset preparation and model training. We anticipate finalizing our predictive models by the previously set deadline and are preparing for the subsequent analysis phase, where we will draw insights and implications from our findings.

**Conclusion:** Our team is committed to addressing the feedback received and enhancing our project's scope and depth. By enriching our dataset, refining our business justification, addressing imbalanced data, and progressing in our methodological approach, we are confident in our project's potential to predict loan defaults while upholding ethical standards accurately.

**Literature References/Citations**:

1. Xu Zhu, Qingyong Chu, Xinchang Song, Ping Hu, Lu Peng, Explainable prediction of loan default based on machine learning models, Data Science and Management, Volume 6, Issue 3, 2023, Pages 123-133, ISSN 2666-7649, doi:10.1016/j.dsm.2023.04.003
2. S. K. Shaheen and E. ElFakharany, "Predictive analytics for loan default in banking sector using machine learning techniques," 2018 28th International Conference on Computer Theory and Applications (ICCTA), Alexandria, Egypt, 2018, pp. 66-71, doi: 10.1109/ICCTA45985.2018.9499147
3. Chen H., “Prediction and Analysis of Financial Default Loan Behavior Based on Machine Learning Model”. Computational Intelligence and Neuroscience. Volume 2022, Article ID: 7907210. doi: 10.1155/2022/7907210