**Introduction**

**Research Problem & Significance**

Credit card approval is an important decision process for financial institutions that directly impacts profitability, risk management, and customer satisfaction. With the increasing volume of applications, automating the approval process using data-driven models can enhance efficiency, reduce human bias, and minimize default risk. However, inaccurate predictions can also lead to financial losses or customer dissatisfaction.

**Main Research Questions**

* Which applicant features strongly influence credit card approval decisions?
* How accurately can machine learning models predict approval outcomes?
* What challenges arise when modeling approval decisions, and how can they be addressed?

**Challenges**

* **Data Imbalance:** The number of approved applications may exceed rejections, making it difficult to properly predict the rejections.
* **Feature Diversity:** The dataset includes both numerical, income & age and categorical such as education & marital status variables which will require careful preprocessing.
* **Interpretability:** Financial institutions will need to understand and justify their approval decisions, so model interpretability is important.
* **Privacy and Bias:** Ensuring that models do not reinforce social biases or violate privacy is a growing concern.

**Brief Literature Review**

Research has shown that machine learning models, including logistic regression, decision trees, and ensemble methods, can significantly improve credit approval prediction accuracy over traditional rule-based systems. Feature selection and handling class imbalance are key to robust performance ([Lessmann et al., 2015](https://www.sciencedirect.com/science/article/abs/pii/S095741741400799X)).

**Initial Hypotheses**

1. **Income and Employment Status:** Applicants with higher income and stable employment are more likely to be approved, as these factors may indicate lower default risk ([Abdou & Pointon, 2011]).
2. **Credit History:** A positive credit history with no delinquencies and longer credit history strongly correlates with approval ([Lessmann et al., 2015]).
3. **Age and Marital Status:** Middle-aged and married applicants may have higher approval rates due to perceived stability.
4. **Education Level:** Higher education may be associated with higher approval rates, reflecting better earning potential.
5. **Gender:** Gender may also potentially affect approval rates.

**Data-Driven Hypotheses**

As we explore the data, we plan to develop new hypotheses such as:

* **Interaction Effects:** The impact of income on approval may differ by education or employment status.
* **Nonlinear Relationships:** There may be thresholds such as minimum income above which approval probability increases sharply.
* **Feature Importance:** Some features such as number of dependents, housing status or marital status may be more predictive than initially assumed.
* **Imbalance Impact:** If approvals vastly outnumber rejections, the models may require resampling or class weighting to avoid bias toward the majority class.

**Proposed Work**

* **Exploratory Data Analysis (EDA):** Visualize distributions, correlations, and class balance. Identify missing values and outliers.
* **Preprocessing:** Encode categorical variables, scale numerical features, and address missing data.
* **Modeling:** Implement three regression models (e.g., linear regression, ridge regression, decision tree regression) and three classification models (e.g., logistic regression, random forest, gradient boosting).
* **Evaluation:** Compare models using accuracy, precision, recall, F1-score, and ROC-AUC. Analyze feature importance and model interpretability.
* **Bias and Fairness Checks:** Assess whether sensitive attributes (e.g., gender) influence predictions unfairly.
* **Contextualization:** Compare findings with existing literature and discuss implications for real-world credit approval systems.

**Discussion**

The results will be contextualized within the broader literature on credit scoring and approval prediction. Emphasis will be placed on model performance, interpretability, and fairness. If the findings align with previous research, this will reinforce the validity of current best practices; if not, it may suggest new avenues for improving credit risk assessment.

**References**

* Hussein A. Abdou & John Pointon, 2011. "Credit Scoring, Statistical Techniques And Evaluation Criteria: A Review Of The Literature," Intelligent Systems in Accounting, Finance and Management, John Wiley & Sons, Ltd., vol. 18(2-3), pages 59-88, April.
* Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research, 247*(1), 124–136. <https://doi.org/10.1016/j.ejor.2015.05.030>
* Chang, V., Sivakulasingam, S., Wang, H., Wong, S. T., Ganatra, M. A., & Luo, J. (2024). Credit Risk Prediction Using Machine Learning and Deep Learning: A Study on Credit Card Customers. *Risks*, *12*(11), 174. <https://doi.org/10.3390/risks12110174>
* Wahab, F., Khan, I., & Sabada, S. (2024). Credit card default prediction using ML and DL techniques. *Internet of Things and Cyber-Physical Systems, 4*, 293–306. <https://doi.org/10.1016/j.iotcps.2024.09.001>

<https://www.kaggle.com/datasets/rikdifos/credit-card-approval-prediction>