Final Project Proposal

Group 1*: Antonio Recalde, Hassan Ali, Omar Sagoo, Alejandro Marchini*

About the dataset:

URL: https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients

Instances: 30000

Features: 23

a. Problem Discussion, Algorithms, and System Design

The problem we're addressing is predicting credit card defaults for a financial institution, aiming to minimize the risk of approving loans or credit lines to high-risk customers. This requires accurately predicting the probability of default while providing interpretable insights into what drives these predictions, which is crucial for regulatory and operational transparency.

We plan to investigate several machine learning and AI algorithms:

1. **Artificial Neural Network (ANN)** integrated with the Sorting Smoothing Method (SSM) can enhance accuracy in predicting credit card defaults and estimating default probabilities by effectively capturing complex nonlinear relationships in the data while reducing noise and improving model stability
2. **XGBoost** provides robust and efficient classification for credit card default prediction, while pairing it with SHAP (SHapley Additive ExPlanations) ensures interpretability by highlighting the importance and contribution of each feature to the model's predictions.
3. **Naive Bayes** can be used to predict credit card defaults by calculating the probability of default based on features like payment history, billing amounts, and customer demographics, under the assumption of conditional independence between these features

The system will handle data preprocessing, model training, evaluation, and interpretability outputs. This includes balancing the dataset using SMOTE techniques, tuning parameters, and comparing models on metrics like accuracy, recall, and interpretability.

b. Related Course Topics

1. **Classification**: For credit scoring and default risk assessment.
2. **Deep Learning**: ANN for high-dimensional and non-linear relationships.
3. **Naive Bayes**: A probabilistic classifier based on Bayes' theorem that makes a strong conditional independence assumption between features given the class label, which allows it to multiply individual feature probabilities to make predictions, making it computationally efficient and surprisingly effective for classification tasks despite its simplifying assumptions.
4. **Clustering**: Involved in SMOTE balancing, particularly KMeansSMOTE for synthetic sample generation.
5. **Experimental Comparison and Parameter Tuning**: To analyze and optimize algorithmic performance.

---

c. Expected System Behaviors and Problem Types

- The system should classify customers into high-risk and low-risk groups based on their likelihood of default, providing probabilities rather than binary classifications.
- **ANN and XGBoost**: Produce accurate classifications with SHAP explaining feature impact in XGBoost. The ANN should use SSM for smoother probability estimates.
- **Naive Bayes** predicts credit card defaults by modeling probabilistic relationships between customer features and default outcomes, assuming conditional independence, and providing a fast, interpretable solution for binary classification problems with well-structured data.
- **Imbalance Handling**: Models should accurately capture patterns in an imbalanced dataset, maintaining high recall for the minority class (defaults).

These behaviors address challenges in finance, where accurate default prediction is critical, and interpretability allows for regulatory compliance and improved trust in AI systems.

---

d. Expected Focus Areas

1. **Data Imbalance**: Effective balancing techniques like SMOTE and KMeansSMOTE to improve model sensitivity for default cases.
2. **Interpretability**: Using SHAP layers to make model predictions transparent and actionable for real-world applications.
3. **Model Comparison and Tuning**: Systematically comparing ANN, XGBoost, and Naive Bayes on accuracy, recall, interpretability, and robustness.
4. **Probability Estimation**: Implementing SSM to refine the probability predictions of default beyond binary classification.

---

e. Core Resources and Papers

Running list (will be adding more as we continue research)

1. **Credit Card Score Prediction Using Machine Learning Models: A New Dataset:** This paper introduces a novel dataset for credit card score prediction and evaluates the effectiveness of different machine learning models in predicting credit card defaults (https://arxiv.org/abs/2310.02956)
2. **Research on Credit Card Default Prediction Based on k-Means SMOTE and BP Neural Network**: This paper proposes a prediction model based on k-means SMOTE and BP neural network. In this model, k-means SMOTE algorithm is used to change the data distribution, and then 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. (https://onlinelibrary.wiley.com/doi/10.1155/2021/6618841)
3. **Toward interpretable credit scoring: integrating explainable artificial intelligence with deep learning for credit card default prediction:** This study presents a new method for predicting credit card default using a combination of deep learning and explainable artificial intelligence (XAI) techniques. Integrating these methods aims to improve the interpretability of the decision-making process involved in credit card default prediction (https://link-springer-com.sandiego.idm.oclc.org/article/10.1007/s00521-023-09232-2)
4. **Credit Card default prediction using ML and DL techniques:** This research paper compares the performance of Deep Learning (specifically ANN) against traditional Machine Learning models (Decision Tree and AdaBoost) (https://www.sciencedirect.com/science/article/pii/S2667345224000087?via%3Dihub)
5. **Credit Card Default Prediction Based on XGBoost:** The study develops a credit card default prediction model using XGBoost combined with RandomUnderSampler to address data imbalance. (https://www.sciencedirect.com/science/article/pii/S2667345224000087?via%3Dihub)
6. **Explainable AI for Credit Assessment in Banks:** This paper proposes an explainable artificial intelligence (XAI) model for predicting credit default on a unique dataset of unsecured consumer loans provided by a Norwegian bank. (https://www.proquest.com/docview/2756732835?accountid=14742&parentSessionId=8vGinUvmTy0MKYlLwVgF1wQ3LMp167gdbOVI38VEy70%3D&pq-origsite=primo&sourcetype=Scholarly%20Journals)

Team contributions on next page…

f. Equal Team Contributions

1. **Data Preparation and Preprocessing**:
    - ○ Team Member 1 will handle data cleaning and initial exploration.
    - ○ Team Member 2 will manage class balancing using SMOTE and KMeansSMOTE.
2. **Model Development and Testing**:
    - ○ Team Member 3 will develop the ANN model with SSM, focusing on tuning and testing.
    - ○ Team Member 4 will implement XGBoost and integrate SHAP for interpretability.
    - ○ Team Members 1 and 2 will collaborate on building and testing Bayes Nets
3. **Comparative Analysis and Reporting**:
    - ○ Team Member 3 will lead model comparison, evaluating accuracy, recall, AUC, and interpretability.
    - ○ Team Member 4 will compile results, handle report writing, and ensure the integration of visualizations and summary statistics.