Credit Card Approval Prediction

1st Yusuf Faruk Güldemir
Artificial Intelligence and Data Engineering
Istanbul, Turkey
guldemir21@itu.edu.tr

2nd Hasan Taha Bağcı
Artificial Intelligence and Data Engineering
Istanbul, Turkey
bagcih21@itu.edu.tr

***Abstract*—Our project, "Credit Card Approval Prediction," aims to revolutionize the credit approval process by developing a predictive model that enhances accuracy, fairness, and efficiency. We will leverage a diverse dataset to train and test our models while implementing fairness-aware techniques. This project, outlined in this proposal, provides a comprehensive methodology, evaluation methods, and a GitHub repository, fostering collaboration and progress tracking. With a team of dedicated members, we are committed to making regular commits and ensuring transparency throughout the project. The successful execution of this project has the potential to significantly impact individuals seeking credit and the financial institutions making lending decisions.**

***Index Terms*—Predictive Modeling, Data Preprocessing, Fairness, Machine Learning, Project Collaboration, Financial Inclusion.**

I. Introduction

We're starting the "Credit Card Approval Prediction" project in the rapidly changing world of banking. Our team's main objective is to improve the process of credit card approval, making it more precise and equitable. Yusuf Faruk Güldemir will lead the effort by cleaning and preparing the data, engineering features, and assessing the models. Meanwhile, Hasan Taha Bağcı is in charge of setting a project timetable, gathering and organizing data, constructing models, and guaranteeing system fairness. We'll employ data mining techniques like Support Vector Machines (SVM) to improve the precision of credit approval. This idea goes beyond banking, perhaps making it simpler for consumers to get credit while protecting financial institutions' interests in our becoming more digital surroundings.

II. Dataset

We received our dataset [1] for our "Credit Card Approval Prediction" research from Kaggle, a well-known platform for data science and machine learning tools. The data is sent in CSV (Comma Separated Values) format, which makes it easy to maintain and compatible with data analysis tools.

The dataset is comprised of 438,557 rows and 18 columns, with each row representing a client and the columns containing various features and information. These features include essential information such as client number (ID), gender, car and realty ownership, number of children, annual income, income category, education level, marital status, housing type, age, employment history, and more. Some columns provide binary information, such as the presence of a mobile phone, work phone, phone, email, and others. Additionally, the dataset features an "Occupation Type" column and a count of family members.

The column names and explanation are shown below

- **ID** : Client number
- **CODE_GENDER** : Gender
- **FLAG_OWN_CAR** : Is there a car
- **FLAG_OWN_REALTY** : Is there a property
- **CNT_CHILDREN** : Number of Children
- **AMT_INCOME_TOTAL** : Annual income
- **NAME_INCOME_TYPE** : Income category
- **NAME_EDUCATION_TYPE** : Education level
- **NAME_FAMILY_STATUS** : Marital status
- **NAME_HOUSING_TYPE** : Way of living
- **DAYS_BIRTH** : Birthday Count backwards from current day (0), -1 means yesterday
- **DAYS_EMPLOYED** : Start date of employment. Count backwards from current day(0). If positive, it means the person currently unemployed.
- **FLAG_MOBIL** : Is there a mobile phone
- **FLAG_WORK_PHONE** : Is there a work phone
- **FLAG_PHONE** : Is there a phone
- **FLAG_EMAIL** : Is there an email
- **OCCUPATION_TYPE** : Occupation
- **CNT_FAM_MEMBERS** : Number of members in family

To guarantee the quality and consistency of the dataset, preliminary cleaning and preparation will be performed as part of the data preprocessing process. This could include dealing with outliers, handling missing values, and, if required, encoding categorical variables. Keeping the enormous dataset organized and making sure it continues to be appropriate for the machine learning models we want to use is one possible problem. Like any real-world dataset, there can also be inconsistent or missing data that needs to be handled carefully. These problems will be settled by our data preprocessing procedures, which will give our credit card approval prediction model a good foundation.

III. Methodology

Our methodical approach to resolving the "Credit Card Approval Prediction" problem includes feature engineering, data processing, and the application of several machine

Identify applicable funding agency here. If none, delete this.

learning models. [2] Effective data preprocessing is essential when dealing with a large dataset. To guarantee data quality, we will start by managing missing values, dealing with outliers, and encoding categorical variables as necessary. For our models to be accurate and reliable, this step is essential.

We will then do extensive feature engineering with the goal of obtaining relevant information from the dataset. For example, we can add additional attributes to the "DAYS_BIRTH" column to display the client's age, or the "DAYS_EMPLOYED" column to show the client's length of employment. Our models will get richer input from these engineered features, which will improve their predictive power.

Moreover, we'll use a variety of algorithms in our modeling to predict credit card approval. Strong classification powers are provided by classification modelse like Support Vector Machines (SVM), and interpretability [5] is provided by decision trees. In addition, we may look into deep learning methods such as neural networks to identify complex patterns in the data. We can evaluate the models' performance and choose the most accurate one thanks to their variety.

To make sure that our model makes credit decisions fairly, we will use fairness-aware modeling techniques [3] in addition to traditional modeling. Fairness metrics, replication methods, or fairness-aware algorithms will all be used in this, with the goal of reducing bias and prejudice in the hiring process. We hope to develop a highly accurate, equitable, and effective credit card approval system with this complex approach.

Approximately outline of methods is shown as below:

- **Data Preprocessing:**
  - Handle missing values.
  - Address outliers.
  - Encode categorical variables.
- **Feature Engineering:**
  - Create new features, e.g., client's age and employment length.
- **Model Selection:**
  - Utilize Support Vector Machines (SVM) for robust classification.
  - Employ decision trees for interpretability.
  - Explore deep learning techniques, like neural networks, for capturing complex patterns.
- **Fairness Integration:**
  - Implement fairness metrics to assess bias.
  - Employ re-sampling techniques to mitigate bias.
  - Utilize fairness-aware algorithms to ensure equitable credit decisions.

IV. Evaluation Methods

We plan to use a set of key performance metrics [4] and requirements that are in line with the project's specific objectives in order to evaluate the success of our "Credit Card Approval Prediction" project. We will be able evaluate the precision, equity, and effectiveness of our solution thanks to these metrics.
[6]

Our project's success will be evaluated based on obtaining a high accuracy rating, proving that credit decisions are fair, and making sure that the decision-making process is effective. Using these metrics, we will compare our models to the baseline and perform a thorough assessment of how well our solution improves credit card approval predictions.

Evaluation methods are shown as below:

A. Accuracy Metrics

We will employ standard accuracy metrics to assess the performance of our credit card approval prediction model:

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}} \tag{1}$$

B. Fairness Metrics

To ensure fairness in our predictions, we will use the following fairness metrics:

$$EO\ Disparity = \frac{\text{True Positive Rate (Group A)}}{\text{True Positive Rate (Group B)}} \tag{2}$$

$$SPD = \text{P(Group A)} - \text{P(Group B)} \tag{3}$$

C. Efficiency Metrics

Efficiency will be evaluated based on the time taken for the model to process a credit card approval request.

D. Baseline Model Comparison

To gauge the effectiveness of our models, we will establish a baseline model, such as logistic regression, and compare the performance of our selected machine learning models against this baseline.

V. Github Repository

A GitHub repository for our project has been established and can be accessed at the following link:

YZV311E Data Mining Group 17 Github Repository

The project code, documentation, and related materials will be regularly committed and pushed to this repository. Additionally, a "Progress.txt" file within the repository will be consistently updated on a weekly basis to detail the tasks completed, promoting clear communication and progress tracking.

VI. Time Plan & Distribution of Work

Project Duration: 2.5 months

Project Time Plan:

- Week 1 : Define project objectives and goals
- Week 2 : Data collection and initial data preprocessing
- Week 3 : Further data preprocessing, handle missing values, and outliers
- Week 4 : Feature engineering and data enrichment
- Week 5 : Model selection and initial model development
- Week 6 : Hyperparameter tuning and cross-validation
- Week 7 : Implement fairness-aware techniques
- Week 8 : Fine-tune and optimize models
- Week 9 : Evaluate models using metrics (accuracy, precision, recall, F1, ROC AUC)
- Week 10 : Compile project results, insights, and recommendations
- Week 11 : Prepare project documentation and final report

Working Distribution:

Yusuf Faruk Güldemir:

- Data preprocessing and cleaning
- Feature engineering
- Model evaluation and metric analysis

Hasan Taha Bağcı:

- Create project timeline and task allocation
- Data collection and integration
- Model development and optimization
- Implement fairness-aware techniques

VII. Conclusion

Our project, driven by data mining and predictive analytics, stands at the forefront of transforming credit card approval processes. It addresses the industry's pressing need for accuracy, efficiency, and risk mitigation.

The project's essence lies in the dataset, continually updated and preprocessed to reflect the ever-evolving financial landscape. It champions fairness, transparency, and compliance with evolving regulations.

In a nutshell, our project is more than an initiative; it's a commitment to reshaping credit card approvals for the better. It's an invitation to embrace this forward-looking approach and usher in an era of secure, efficient, and customer-centric credit card approvals. The time to act is now, and we are poised to lead the way.

References

[1] Kaggle Dataset for Credit Card Approval https://www.kaggle.com/datasets/rikdifos/credit-card-approval-prediction

[2] S. Lahmiri, S. Bekiros, A. Giakoumelou, and F. Bezzina, "Performance assessment of ensemble learning systems in financial data classification," *Intelligent Systems in Accounting, Finance and Management*, vol. 27, no. 1, pp. 3-9, 2020.

[3] T. P. Pagano, R. B. Loureiro, F. V. Lisboa, R. M. Peixoto, G. A. Guimarães, G. O. Cruz, et al., "Bias and unfairness in machine learning models: a systematic review on datasets, tools, fairness metrics, and identification and mitigation methods," *Big Data and Cognitive Computing*, vol. 7, no. 1, p. 15, 2023.

[4] D. L. Olson and D. Delen, "Performance evaluation for predictive modeling," in *Advanced Data Mining Techniques*, pp. 137-147, 2008.

[5] F. Poursabzi-Sangdeh, D. G. Goldstein, J. M. Hofman, J. W. Wortman Vaughan, and H. Wallach, "Manipulating and measuring model interpretability," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1-52, May 2021.

[6] Vujović, Ž. (2021). Classification model evaluation metrics. International Journal of Advanced Computer Science and Applications, 12(6), 599-606.