Credit Card Approval Prediction Intermediate Report

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

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

I. Introduction

Our project, "Credit Card Approval Prediction," aims to improve the accuracy and justice of the credit card approval process in the continually evolving banking industry. This project is important because it could make it easier for customers to obtain credit while also protecting financial institutions' interests in the rapidly changing digital landscape. Our team member Hasan Taha Bağcı, who manages project management, data organization, model construction, and fairness assurance, and Yusuf Faruk Güldemir, who handles data cleaning, preparation, feature engineering, and model evaluation, is dedicated to using data mining techniques, especially Support Vector Machines (SVM) and other models, to accomplish these goals.

The project is based on a dataset that contains vital client data such as gender, marital status, income details, car and property ownership, education level, and more. Taking advantage of the knowledge contained in the utility functions in our `utils` folder—such as `isCategorical`, `labelEncoding`, `gridsearch`, and others—we hope to implement strong preprocessing, encoding, and model assessment techniques. These features make it easier to apply machine learning models, find categorical columns, and explore data distributions.

We provide an overview of the project's goals, the dataset that was used, and the fundamental procedures that were followed to clean up the data in preparation for later model training in this report. We also explore the specifics of the utility functions created to facilitate the modeling and data exploration procedures. The following sections will go into detail on relevant literature, our suggested methodology, experimental outcomes, and the implications of our discoveries. We will wrap up with some thoughts on the project's problems and potential directions for further investigation.

Our project aims to make a significant contribution to the ongoing discussion about credit risk assessment, which is necessary due to the dynamic nature of the financial domain.

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

II. Related Work

A. Application of Deep Learning for Credit Card Approval: A Comparison with Two Machine Learning Techniques

In 2021, Kibria and Mehmet [7] undertook a comparative analysis, pitting deep learning against logistic regression and support vector machines in predicting credit card approval. The evaluation metrics encompassed accuracy, precision, recall, and F1 score. Deep learning achieved the highest accuracy at 87.10%, while the other two classifiers exhibited identical accuracy levels of 86.23%. The results imply that the deep learning model marginally outperformed the other two machine learning techniques.

B. Credit Card Approval Predictions Using Logistic Regression, Linear SVM and Naïve Bayes Classifier

In 2022, Yiran Zhao [8] explored the effectiveness of various machine learning classifiers for predicting credit card approval. The classifiers utilized included LRC (Logistic Regression Classifier), a linear SVM (Support Vector Machine), and a naive Bayes classifier. Performance evaluation for each algorithm was based on balanced accuracy. The balanced accuracy values were approximately 88.4% for LRC, around 89.0% for the linear SVM, and about 83.4% for the naive Bayes classifier. The results suggested that, among the models, the linear SVM exhibited the most efficient prediction performance.

C. A Classification Approach in the Probability of Credit Card Approval using Relief-Based Feature Selection

In a study conducted in 2022, Rowell, Lysa, Celinne, and Maricel [9] undertook a comparative analysis of various machine learning classifiers. Their investigation involved utilizing an open credit card approval dataset comprising 19 variables and 304,356 instances. Following data cleaning and dimensional reduction through relief-based feature selection, it was observed that the dataset still retained 12 attributes and 132,492 instances. The researchers employed 10-fold cross-validation for both testing and training data, employing the RF, KNN, and NN algorithms. The evaluation of results utilized a confusion matrix and ROC curve. The RF algorithm exhibited the highest accuracy at 95.76%, followed by KNN with 94.37%, and NN with 71.56%. The findings indicated

that among the three classifiers, RF demonstrated superior accuracy, precision, recall, specificity, and AUC.

III. Proposed Work

Our proposed work revolves around a systematic and comprehensive approach to credit card approval prediction, encompassing data preprocessing, feature engineering, and model training. In this section, we provide a detailed insight into the methodology, tools, and techniques employed to achieve the project's objectives.

A. Data Preprocessing and Exploration

A crucial part of our project is data preprocessing, which involves carefully organizing and cleaning the dataset in order to get it ready for further analysis and model training. We start by carefully analyzing the structure of the dataset, looking for any anomalies, and fixing any missing or inconsistent values. We employ functions from the `utils` folder, like `isCategorical`, to methodically evaluate every column in order to ascertain its nature; we concentrate on categorical attributes in particular. We differentiate between categorical and non-categorical features by setting a threshold for uniqueness within a column, which guides subsequent encoding choices.

Functions like `labelEncoding` and `oneHotEncoding` in the `utils` folder make encoding categorical variables easier. These functions make sure that numerical data is properly converted from categorical format, which is necessary for machine learning models to be trained.

Furthermore, we utilize the `plotdist` and `plotbox` functions to obtain a thorough comprehension of the attributes of the data. The distributions of important features, like `AMT_INCOME_TOTAL`, are shown in these visualizations, enabling us to evaluate skewness, spot anomalies, and decide on possible changes. In order to reduce the impact of outliers on model training and increase the predictive model's robustness, this step is essential.

As we go forward, we will continue to be dedicated to data preprocessing by creating new features and transforming old ones into new ones through a process called feature engineering. We plan to investigate intricate relationships within the dataset by utilizing the `utils` functions, producing features that capture significant patterns associated with creditworthiness.

In conclusion, the goal of our data preprocessing work is to provide a strong framework for the project's later phases. Through the implementation of appropriate categorical variable encoding, thorough exploratory data analysis, and dataset integrity assurance, we create a strong foundation for the creation of a credit card approval prediction model that is both fair and accurate.

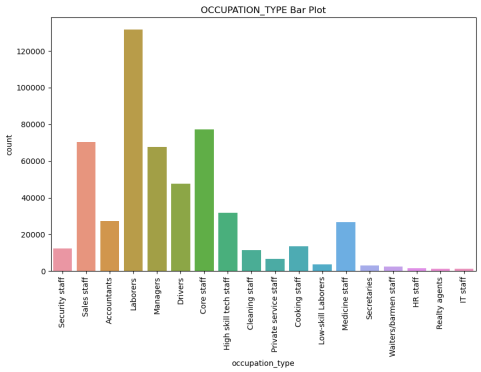

Fig. 1. Confusion Matrix

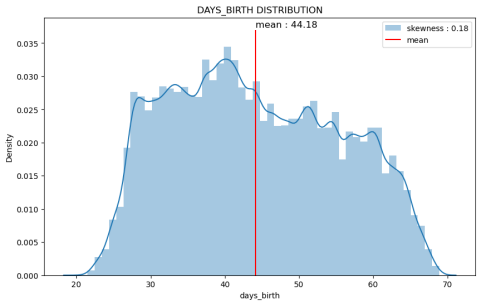

Fig. 2. Age Distribution

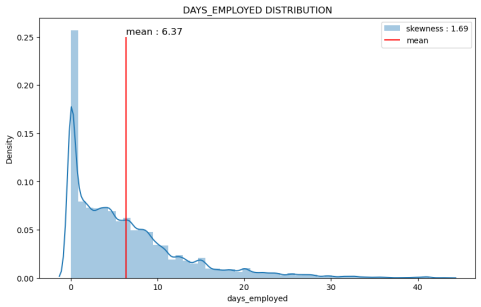

Fig. 3. Confusion Matrix

The figures that were shown above played a crucial role in our data exploration efforts in the sections that came before them. They provided insightful information that made it easier to fully comprehend the dataset.

B. Feature Engineering

A crucial stage of our project is feature engineering, which aims to turn raw data into a feature set that is more representative and informative. Through the introduction of new dimensions and the capture of complex relationships within the dataset, this process is essential to improving the predictive performance of the model.

1) Combining Temporal Features: Combining temporal features like `DAYS_EMPLOYED` and `DAYS_BIRTH` is one way to achieve feature engineering. We search to gain new

TABLE I
RELATED WORKS SUMMARY

Authors	Algorithms	Efficient Algorithm	Accuracy
Kibria Md, Şevkli Mehmet	Logistic regression, deep learning, SVM	Deep learning	87.10%
Yiran Zhao	Linear SVM, LRC, naive Bayes	Linear SVM	89.0%
Leonard, Rowell, Lysa, Celinne, Maricel	RF, KNN, NN	RF	95.76%

insights into the stability and dependability of an individual's source of income in relation to their life expectancy by developing new features that express the ratio of employment duration to age. The following could be used to indicate the derived features:

$$\text{EMPLOYMENT\_AGE\_RATIO} = \frac{\text{DAYS\_EMPLOYED}}{\text{DAYS\_BIRTH}}$$

This ratio offers a nuanced viewpoint on the connection between age and work history and may contain important information for estimating creditworthiness.

2) Family Size Normalization: A person's financial obligations may be impacted by the size of their family. A relative measure of financial cost can be obtained by normalizing the number of children in the family by the total number of family members: In this case, N denotes the total number of income records for a given person.

$$\text{FAMILY\_SIZE\_NORMALIZED} = \frac{\text{CNT\_CHILDREN}}{\text{CNT\_FAM\_MEMBERS}}$$

This normalized feature helps determine whether a person is capable of taking on more financial responsibilities by providing information about the financial responsibilities of each family member.

3) Income Status:

$$x \in \text{AMT\_INCOME\_TOTAL}$$

$$\text{INCOME STATUS}(x) = \begin{cases} 1 & \text{if } 0 \leq x < 100000 \\ 2 & \text{if } 100000 \leq x < 200000 \\ 3 & \text{if } 200000 \leq x < 300000 \\ 4 & \text{if } 300000 \leq x < 400000 \\ 5 & \text{if } 400000 \leq x < 500000 \\ 6 & \text{if } 500000 \leq x < 600000 \\ 7 & \text{if } 600000 \leq x < 700000 \\ 8 & \text{if } 700000 \leq x < 800000 \\ 9 & \text{if } 800000 \leq x < 900000 \\ 10 & \text{if } 900000 \leq x \leq 1000000 \end{cases}$$

C. Model Training and Evaluation

Our work's main contribution is the use of machine learning models to forecast credit card approval. Training a Random Forest model, an ensemble learning method recognized for its flexibility and adaptability, has been the first step in this process. However since model optimization is necessary, we will be focusing on thorough grid search methods in the future to optimize hyperparameters and improve model performance.

This optimization process is supported by the `utils` function `gridsearch`, which explores the hyperparameter space methodically in order to find the configuration that maximizes predictive accuracy. Furthermore, we employ numerous metrics for model evaluation, as contained in functions such as `plotroc`, `plotpr`, and `confusionmatrix`, to evaluate model's performance concerning ROC AUC, accuracy, precision, and recall.

D. Ensuring Fairness and Interpretability

We understand the significance of fairness and interpretability in credit-related decision-making as responsible machine learning practitioners. In order to guarantee the fairness and transparency of the credit card approval model, our next research will examine various approaches and strategies. This calls for careful investigation.

IV. EXPERIMENTAL RESULTS

Our project's experimental phase consists of using the preprocessed dataset to train and assess a Random Forest model at first step. It will be supported more models and processing. We provide the initial results in this section, along with model performance metrics and learnings from the assessment procedure.

A. Model Training

Using the preprocessed dataset, we trained a Random Forest classifier to start the modeling process. Because of its flexibility in credit risk prediction, the Random Forest algorithm—which is renowned for its capacity to manage intricate relationships and preserve robustness—was chosen as the foundation.

To adjust the Random Forest model's hyperparameters, we will use a grid search using the `GridSearchCV` function from the `utils` folder in next weeks. The chosen hyperparameters minimize overfitting and maximize the accuracy of the model. For the next assessment, the configuration that performed the best overall, as determined by cross-validation, was selected.

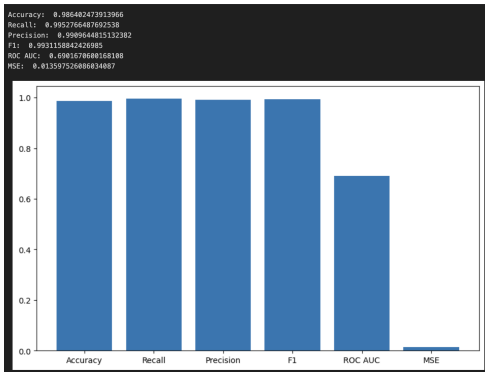

Fig. 4. Random Forest Model Metrics

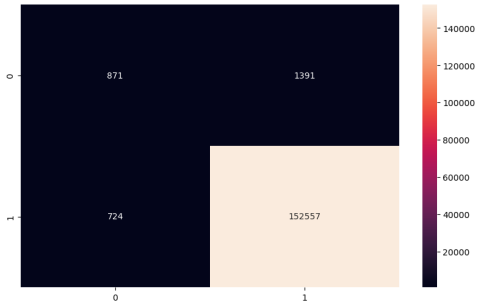

Fig. 5. Confusion Matrix

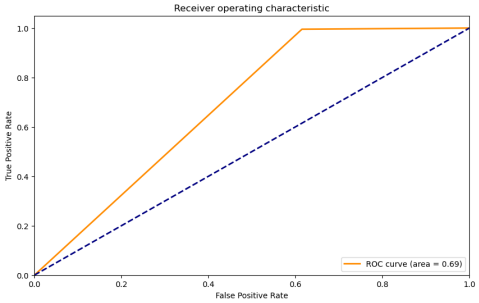

Fig. 6. ROC-Curve

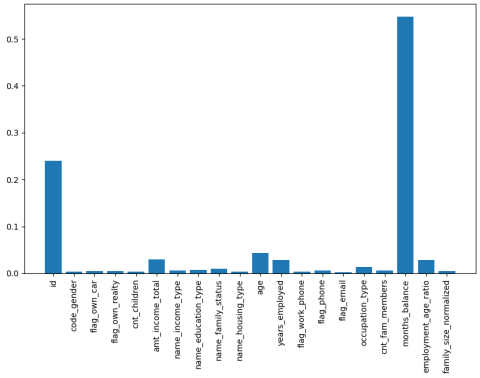

Fig. 7. Feature Importance of Random Forest Model

B. Model Evaluation

A variety of evaluation metrics were employed to thoroughly evaluate the trained model's performance. We extracted important metrics, such as accuracy, recall, precision, F1 score, ROC AUC, and mean squared error, using the `printmetrics` function from the `utils` folder.

A high degree of overall prediction accuracy is indicated by the Random Forest model's impressive performance metrics, which include an accuracy of 98.6%. With a recall of 99.5%, the model captures a good percentage of relevant instances. Its precision of 99.1% indicates that it can correctly identify positive cases. The model's efficacy is further confirmed by the robust 99.3% F1 score, which strikes a balance between precision and recall.

The ROC AUC score of 69.0%, which indicates that there is still opportunity for improvement in the model's capacity to distinguish between positive and negative instances, is of note. In light of this, a proactive plan to improve the model's performance is laid out for the next few weeks. With an emphasis on iterative development and a dedication to continuous improvement, the Random Forest model is guaranteed to adapt to changing needs in the near future, resulting in predictions that are even more durable and trustworthy.

1) Confusion Matrix: The `confusionmatrix` function is used to visualize the confusion matrix, which offers a comprehensive analysis of the predictive accuracy of the model. The matrix provides insights into the model's capacity to distinguish between credit applications that are approved and denied by showing the number of true positive, true negative, false positive, and false negative predictions.

2) Receiver Operating Characteristic (ROC) Curve: At different threshold settings, the trade-off between true positive rate and false positive rate is represented graphically by the ROC curve, which is produced by the `plotroc` function. The ROC AUC, or area under the curve, is a quantitative indicator of how well the model can differentiate between positive and negative examples.

C. Feature Importance

Interpretability depends on each feature's contribution to the model's predictions being understood. We can visualize the feature importance and determine the main factors influencing credit card approval predictions with the aid of the `plotfeatureimportance` function.

V. Conclusion

In conclusion, our credit card approval prediction project has produced encouraging results, with the Random Forest model achieving a high level of accuracy (98.6%). Notably, the model demonstrates impressive recall (99.5%), precision (99.1%), and F1 score (99.3%), indicating its effectiveness in correctly identifying creditworthy individuals.

The ROC AUC of 69.0% suggests reasonable discriminative power, but we acknowledge the need for further model improvements. Future iterations will involve rigorous optimization, exploration of alternative algorithms, and the inclusion of additional features to enhance predictive accuracy and address any inherent biases

These preliminary findings underscore the potential of data mining techniques in reshaping credit risk assessment processes. As we move forward, our commitment remains steadfast to refining and expanding our models, ensuring they align with evolving industry needs and ethical considerations.

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.

[7] M. Kibria and M. Şevkli, "Application of Deep Learning for Credit Card Ap- proval: A Comparison with Two Machine Learning Techniques," vol. 11, pp. 286–290, Jun. 2021.

[8] Zhao, Y. (2022, February). Credit card approval predictions using logistic regression, linear SVM and Naïve Bayes classifier. In 2022 International Conference on Machine Learning and Knowledge Engineering (MLKE) (pp. 207-211). IEEE.

[9] L. Flores, R. M. Hernandez, L. C. Tolentino, C. A. Mendez, and M. G. Z. Fer- nando, "A Classification Approach in the Probability of Credit Card Approval using Relief-Based Feature Selection," in 2022 2nd Asian Conference on Inno- vation in Technology (ASIANCON), Aug. 2022, pp. 1–7.