CSC 540 Final Project Report

Megh Thakkar (2111824)

Credit Card Approval Prediction

Abstract:

This project uses advanced machine learning algorithms to improve the credit risk assessment process in the financial industry. The goal is to create a more accurate model that can identify "good" and "bad" credit risks. This model uses a wider range of data than traditional financial indicators, giving a more comprehensive picture of borrowers. The project involves data cleaning, new feature creation, and model optimization. The result is a model that can predict credit risk more accurately, reducing the chance of borrowers defaulting on loans.

Dataset link:

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

Google colab link:

https://colab.research.google.com/drive/1qB78DD3H2zSchQCnn3lGmQxVE3DIhfKj?usp=sharing

1) Introduction:

In a world where smart money matters, predicting whether someone will repay a loan is essential. This project combines two linked datasets (application_record.csv and credit_record.csv) to build a credit risk model. The unique "ID" field connects these datasets, showing us each person's credit history and details from their loan application. The first dataset (application_record.csv) has a lot of information about people's demographics and finances. It includes things like their gender, whether they own a car or house, how many people depend on them for money, and how much money they make each year. This gives us a good picture of their financial situation. The application process involves gathering detailed information about the applicant, including their income sources, education level, relationship status, living situation, age, work experience, and ability to communicate. This comprehensive data provides a nuanced understanding of the applicant's way of life and financial circumstances. The credit_record.csv dataset tracks clients' credit transactions over time, detailing their monthly financial activities. This dataset focuses specifically on payment behaviors, such as when payments were made, how often, and how consistently. Combined with other datasets, this information helps lenders identify customers with strong and weak credit histories. It combines personal, financial, and behavioral data, allowing algorithms to make highly accurate predictions about creditworthiness. The analysis conducted provides valuable insights that will enhance lending practices. Instead of relying solely on past information, lenders will now consider a broader range of factors to assess risk. This approach creates a more comprehensive foundation for informed and fair financial decisions.

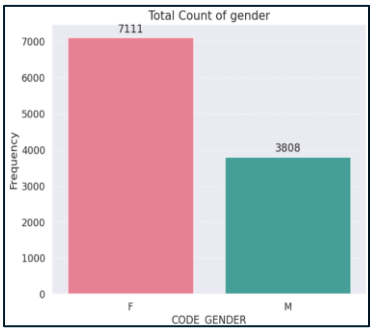
2) Methodology:

2.1) Data Cleaning and Feature Engineering

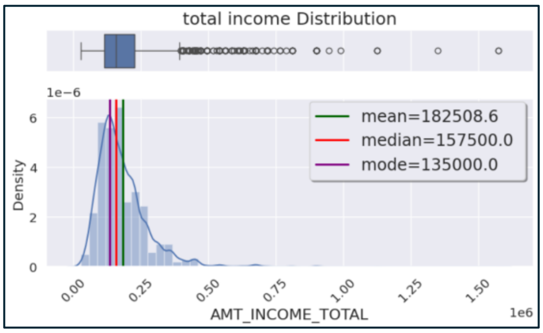
The preliminary stage of the analysis was dedicated to enhancing the quality and reliability of the dataset through meticulous data cleaning procedures. This phase commenced with the identification and removal of 94 duplicate records, each distinguished by a unique identifier ('ID'). This initial purging of duplicates was crucial for eliminating redundancy, thereby ensuring the integrity and accuracy of the dataset—a foundational step for any robust predictive modelling effort. The subsequent task involved the integration of two pivotal datasets, namely application_record.csv and credit_record.csv, through a unifying 'ID' linkage. This integration was imperative for consolidating comprehensive financial data, essential for creating an elaborate analytical framework conducive to a thorough evaluation of credit risk. Upon merging, the dataset underwent an intensive preprocessing phase to address anomalies and inconsistencies. For example, an encountered issue was the presence of 1,482 null entries within the 'OCCUPATION_TYPE' field, which were strategically filled with a 'Others' placeholder to maintain data completeness. Additionally, a composite 'INFO' identifier was devised to unveil and subsequently eliminate hidden duplicates not discernible through 'ID' analysis alone. This step further refined the dataset's quality. Anomalies in the 'DAYS_EMPLOYED' field, specifically those reflecting positive values indicative of unemployment, were normalized to zero, enhancing the dataset's logical consistency. The 'OCCUPATION_TYPE' for these entries was adjusted to 'not_working', providing a clearer representation of employment status. The feature engineering phase also involved transforming 'DAYS_BIRTH' data into 'AGE', converting it into a more intuitive and analytically useful format. Corrections were made to illogical entries in 'CNT_FAM_MEMBERS', ensuring data consistency. Moreover, a significant transformation entailed reclassifying the 'STATUS' feature into a binary 'Status' category, effectively distinguishing clients into 'good' (0) and 'bad' (1) creditworthiness based on their payment histories. These extensive preprocessing and feature engineering efforts were instrumental in preparing the dataset for the application of sophisticated machine learning algorithms aimed at predicting creditworthiness with enhanced accuracy and reliability.

2.2) Data Analysis

The analytical exploration of the dataset yielded insightful revelations about the credit profiles of applicants, encompassing both demographic and financial aspects. Notably, the analysis highlighted that a substantial majority of the applicants were female (approximately 65%) which can be seen in Figure 1, and a significant proportion did not own a car (67%), suggesting a preference for urban living or reliance on alternative transportation methods. The average age of applicants was determined to be 43 years, indicating a demographic potentially at a stage of financial stability—crucial for accurate credit risk assessment. The financial analysis uncovered disparities among applicants, with an average annual income of approximately \$182,508.6 and a median income of \$157,500, signaling varied financial health and stability across the applicant pool which can see in Figure 2. The majority of applicants resided in either houses or apartments, with about two-thirds possessing real estate, a factor positively reflecting on their credit assessments. Furthermore, the analysis of family compositions revealed that most applicants did not have children, indicating lower financial dependencies and potentially higher disposable incomes. These findings are vital for comprehensively understanding an applicant's creditworthiness, taking into consideration factors such as age, gender, family status, housing history, financial stability, and obligations.



(Figure 1)



(Figure 2)

2.3) Pre-processing (Creating Dummies)

In preparation for model training, the raw dataset underwent several transformative processes. Initially, features deemed irrelevant to the predictive modeling task, such as 'ID', 'CODE_GENDER', and various flags, were excluded from further analysis. The refined dataset was then subjected to a thorough examination for missing values, followed by the identification of categorical variables requiring conversion for analytical suitability. Notable among these were 'FLAG_OWN_CAR', 'FLAG_OWN_REALTY', 'NAME_INCOME_TYPE', 'NAME_EDUCATION_TYPE', 'NAME_FAMILY_STATUS', 'NAME_HOUSING_TYPE', and 'OCCUPATION_TYPE'. These categorical variables were one-hot encoded, a process converting them into a numerical format amenable to machine learning algorithms. This encoding was pivotal in ensuring the dataset comprised solely numerical features, a prerequisite for the effective application of many machine learning techniques.

2.4) Train-Test Split

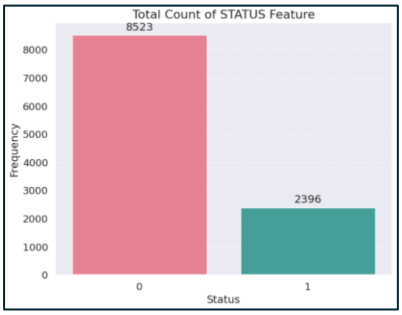
The dataset was subsequently divided into training and testing sets, employing a stratified approach via the train_test_split() function from scikit-learn to maintain a consistent class distribution across both sets. Specifically, 70% of the data was allocated to the training set, with the remaining 30% designated for testing. This strategic division is imperative for evaluating the predictive model's performance on unseen data, a critical measure of its generalization capability.

2.5) Feature Selection

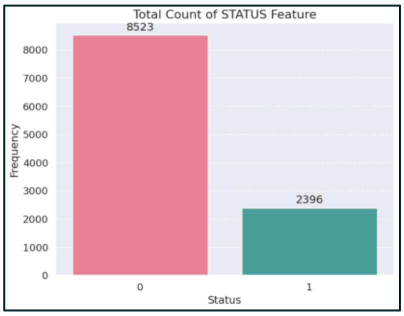
Feature selection was undertaken to identify and retain the most impactful features for predicting the target variable. This process was facilitated by employing a Random Forest algorithm, renowned for its efficacy in determining feature importance. The algorithm's feature_importances_ attribute was utilized to rank features based on their significance, with the SelectFromModel technique applied to filter out features with importance scores above the mean threshold. This methodological approach effectively reduced the dataset's dimensionality, potentially enhancing the model's performance by concentrating on the most informative features. Such a streamlined feature set not only improves computational efficiency but also mitigates the risk of overfitting, thereby increasing the predictive model's accuracy and generalizability.

2.6) Class Imbalance

In the investigation of creditworthiness, an initial challenge presented itself in the form of class imbalance, as evidenced by the predominant frequency of Class 0 (8,523 instances) over Class 1 (2,396 instances) within the 'STATUS' feature of the dataset which can be seen Figure 3. Such disproportion often leads to biased model performance favoring the majority class. To rectify this issue, Synthetic Minority Over-sampling Technique (SMOTE) was employed, as indicated by the implementation SMOTE(random_state=42). SMOTE algorithmically generates synthetic samples from the minority class, thus artificially augmenting the dataset to achieve parity between classes. The result of this intervention is depicted in the second bar graph, where post-SMOTE application, an equitable distribution is observed with both classes reflecting an equal count of 5,994 instances which can be seen Figure 4. This balanced dataset is instrumental for training unbiased models, enabling them to learn from a dataset that accurately represents the diverse outcomes they are intended to predict, thus enhancing the models' predictive performance and validity in practical applications.



(Figure 3)



(Figure 4)

2.7) Model Training and Evaluation

In this study, various machine learning algorithms were explored to find the best approach for predicting creditworthiness. A range of classifiers—Logistic Regression, Decision Tree, Support Vector Machine (SVM), and Random Forest—were trained and evaluated on a carefully prepared dataset. Each model underwent thorough fine-tuning of its hyperparameters using GridSearchCV, which systematically explores parameter combinations to optimize predictive performance. A variety of parameters, from regularization strength to tree-specific settings like maximum depth and minimum samples per leaf, were tinkered with in this exhaustive process to extract the best performance from each algorithm while guarding against overfitting.

To ensure the models were robust and reliable, a comprehensive cross-validation strategy was employed. By splitting the dataset into five equally sized subsets and training the models on different combinations of these subsets, a thorough assessment of how well they would perform on unseen data was conducted. This helped in obtaining more accurate estimates of performance metrics like accuracy, precision, recall, and F1-score. By averaging these metrics across multiple rounds of validation, the risk of bias or instability from any single data split was minimized, making the findings more trustworthy.

Additionally, model development and deployment were streamlined using a structured pipeline approach. This allowed for the integration of everything from data preprocessing to hyperparameter tuning in a seamless workflow. For instance, techniques like Synthetic Minority Over-sampling Technique (SMOTE) for handling class imbalances were smoothly incorporated into the models' training process. This not only made model building more efficient but also facilitated easier comparison of different algorithms. By combining hyperparameter optimization, cross-validation, and pipeline structuring, the study aimed to provide actionable insights for predicting creditworthiness, thereby aiding decision-making in financial contexts.

2.8) Model Comparison Metrics

A comparative analysis of the performance metrics was conducted across the various models to determine the most suitable algorithm for the task at hand. Key metrics including accuracy, precision, recall, and F1-score, in addition to ROC-AUC scores, served as the basis for this comparison. These metrics provided a multifaceted view of each model's performance, with a particular focus on their ability to accurately classify and discriminate between different levels of creditworthiness. The outcomes of this analysis were instrumental in unveiling the relative advantages and limitations of each model, guiding the selection process towards the most appropriate model for deployment in real-world credit risk assessment applications. This methodical comparison not only underscored the nuanced capabilities of each algorithm but also illuminated the path towards optimizing predictive accuracy in assessing applicants' creditworthiness.

3) Results

When comparing the performances of four different machine learning models—Logistic Regression, Decision Tree, Support Vector Machine (SVM), and Random Forest—a variety of strengths and weaknesses came to light. The Decision Tree model stood out for its high accuracy rate of 74%, showcasing its ability to classify instances effectively. However, it fell short in identifying positive instances, as evidenced by its low recall rate for Class 1 (0.03). On the other hand, Logistic Regression, while less accurate overall at 51%, demonstrated a noteworthy capability to classify negative instances accurately (precision for Class 0 at 0.78) and had a significantly better recall for Class 1 (0.53). The SVM and Random Forest models offered a middle ground in terms of performance, with accuracy rates of 62% and 72% respectively, and displayed a mix of precision, recall, and F1-scores for Class 1. This detailed comparison sheds light on each model's specific advantages and challenges, providing valuable guidance for selecting the right model for practical applications.

4) Discussion

The study also highlighted two critical issues: the lack of normalization techniques like MinMax and the computational challenges faced with the SVM model. The absence of such normalization techniques might have contributed to the underwhelming performance of some models, especially SVM, which relies on feature scaling for optimal operation. Additionally, the SVM model's training time, which stretched to about two hours on a Google Colab TPU, poses a considerable challenge for its practical application, especially when swift decision-making or real-time processing is necessary. Despite this lengthy training period, SVM's precision was not as high as hoped, indicating a tendency to misclassify positive instances. These observations underscore the importance of both computational efficiency and the ability to perform well across different metrics, suggesting a need to explore more scalable and effective alternatives.

5) Conclusion

Considering the overall findings, Logistic Regression emerges as the standout model for predicting creditworthiness in this context. It might not have the highest accuracy compared to the Decision Tree and Random Forest models, but its superior precision for negative instances and recall for positive instances make it particularly valuable for credit risk assessment. Its strengths in balancing precision and recall—key for identifying credible applicants while minimizing false positives—alongside its interpretability and computational efficiency, mark it as an ideal choice for real-world applications. While the Decision Tree showed high accuracy, its poor performance in correctly identifying positive instances limits its usefulness. SVM, despite its computational demand and suboptimal precision, remains a potential option if its issues can be addressed. Ultimately, Logistic Regression offers a well-rounded solution that adeptly balances performance, understandability, and efficiency, making it the preferred option for assessing credit risk.

6) Future Work

Looking ahead, there's much room for improvement and exploration in the predictive modeling of creditworthiness. Advancing feature engineering could unlock deeper insights into applicant behaviors and enhance model accuracy. Delving into additional socio-economic factors, optimizing normalization techniques, and experimenting with ensemble learning are just a few pathways to potentially boost model effectiveness. Moreover, exploring scalable and efficient machine learning algorithms, such as gradient boosting machines or deep learning, could offer significant advancements without sacrificing accuracy. Integrating alternative data sources, like social media activities or transaction histories, could also enrich the assessment framework, offering a more nuanced view of an applicant's financial stability. Pursuing these directions could significantly refine and advance the methodologies for credit risk assessment, benefiting both financial institutions and borrowers.

7) Literature Review

1) Enhanced system for revealing fraudulence in credit card approval.

The study by B. Subashini and Dr. K. Chitra on credit card approval and fraud detection in the International Journal of Engineering Research & Technology can be considered a landmark paper in this realm. The research, 'Enhanced System for Revealing Fraudulence in Credit Card Approval,' orbits around the fact that a secure system to prevent fraudulent activities is instrumental in the financial sector, especially in jurisdictions such as India plagued by banking fraud. To do this, the authors use more advanced data mining techniques to investigate the potential model. They compare different classification models, including but not limited to C4.5, C5.0 & CART, Support Vector Machine using Sequential Minimal Optimization, BayesNet, and Logistic Regression, exploring the possibility of integrating them for improved security protocols in automatic credit card approval to achieve their goal of fraud prevention. The paper also describes the significance of performance metrics such as accuracy and precision for predictive modeling in financial institutions. They demonstrate how these data mining devices prevent fraudulent approval, arguing that the appropriate model must be chosen after an extensive analysis of all these performance metrics.

References

Subashini, B., & Chitra, K. (2013). Enhanced system for revealing fraudulence in credit card approval. Int. J. Eng. Res. Technol, 2(8), 936-949.

2) Application of deep learning for credit card approval: A comparison with two machine learning techniques

The issue of credit card defaulters is discussed in the article published in the International Journal of Machine Learning and Computing, titled “Application of Deep Learning for Credit Card Approval: A Comparison with Two Machine Learning Techniques” by Md. Golam Kibria and Mehmet Sevkli. To address this problem, the authors have proposed a novel deep learning model to assist firms in approving credit cards and validated their proposal using the UCI dataset. Further, the study compared the efficiency of the deep learning model with two machine learning models, namely Logistic Regression and Support Vector Machine. As noted in the study, the deep learning model yields a slightly better performance than the other two and permits a little improvement in identifying the creditworthiness of an individual. The current research highlights the edge of exploiting a deep learning model over traditional methods to ameliorate the accuracy of credit card approval systems and alleviate the risk of default.

References

Kibria, M. G., & Sevkli, M. (2021). Application of deep learning for credit card approval: A comparison with two machine learning techniques. International Journal of Machine Learning and Computing, 11(4), 286-290.

3) An empirical study for credit card approvals in the Greek banking sector

This paper describes factors that make a bank decide on whether to approve or reject a credit card application and helps the financial institution improve their credit card applicants' retention strategy. According to this source, the paper uses a logistic regression model to analyze a Greek bank’s application data. The document, therefore, reflects on the variables which are relevant for the bank’s credit card decision-making process, contributing to approving or rejecting applications. The source aided this current study to identify scenarios like property ownership and financial credibility as some of the substantial factors the bank considers in their decision-making processes, while on the other hand, gender and family status are not very significant. Finally, the document checks the actual results against the logistic regression model’s predictions.

References

Mavri, M., & Ioannou, G. (2004). An empirical study for credit card approvals in the Greek banking sector. Operational Research, 4, 29-44.

4) Credit Card Approval Verification Model

This Master’s thesis by Umabhanu Tanikella, titled "The Credit Card Approval Verifier Model," reports on the study of the automation of the credit card approval process. The primary objective of the model is to forecast the decision of financial institutions regarding the approval of credit cards by analyzing customer details, including credit-worthiness, repayment history, and income. The project is designed to provide a tool to financial institutions for making precise, fraud-free decisions in the approval process, enhancing the experience for both the institutions and their clients. This project has focused on data-related challenges, such as cleaning, manipulation, visualization, and the application of various machine learning classifiers. It concludes with the significant factors affecting the issuance of credit cards as observed from the study. The use of logistic regression and random forest classifiers, along with the assessment of their predictive accuracy using a confusion matrix and feature importance rankings, contributes to the body of knowledge in predictive analytics and decision-making automation in the finance sector.

References

Tanikella, U. (2020). Credit Card Approval Verification Model (Doctoral dissertation, California State University San Marcos).

5) Prediction of Credit Card Approval

In the work titled “Prediction of Credit Card Approval,” published in the International Journal of Soft Computing and Engineering, Peela Harsha Vardhan et al. provide a machine learning algorithm to predict the approval of credit cards by financial institutions. The topic is paramount in the field of credit risk management. The model developed helps distinguish the likelihood of a customer's inclination towards a credit card by learning specific variables. In particular, the researchers prepared the AI model well, alongside exploratory data analysis to materialize their theorem before executing a predictive model. They use a variety of learning tools whereby they prove their 86% accuracy in discerning credit card approval. Although the accuracy figure is high, the study conducted a grid search to implement optimization as a further step. To perceive further details, a grid search method was used to evaluate the algorithm, aiming to evaluate the first model on balanced accuracy, and lastly, the grid search sought the most stable parameters in logistic regression, highlighting the socio-economic significance of predictive modeling in financial decision-making.

References

Peela, H. V., Gupta, T., Rathod, N., Bose, T., & Sharma, N. Prediction of Credit Card Approval. International Journal of Soft Computing and Engineering, 11(2), 1-6.