Predicting Credit Card Approvals with Supervised Learning

Aika Aldayarova
Cornell Tech MSIS

Joo Kim
Cornell Tech MBA

Maya Powell
Cornell Tech MEng in ORIE

Abstract

This project focuses on predicting credit card approval based on socio-demographic applicant features, such as gender, age, income, marital status, credit score, and employment history. We aim to explore which machine learning algorithms and applicant attributes offer the best predictive power for determining approval outcomes. Our analysis compares several supervised classification models, including Logistic Regression and K-Nearest Neighbors, using a pre-processed dataset of credit card applicants. The results highlight the importance of feature engineering, such as adding quadratic features, and the role of regularization in managing model complexity, especially when working with small datasets. Our findings also emphasize the need for fairness and transparency in machine learning-based credit scoring, as certain features, such as ethnicity, can introduce biases. Ultimately, we present a model that balances bias and variance, providing insights into improving credit card approval processes while considering the ethical implications of automated decision-making.

1 Introduction

The problem we are addressing in this project is the prediction of credit card approval based on socio-demographic applicant features (gender, age, outstanding debt, marital status, ethnicity, years employed, prior default, current employment status, credit score, driver's license ownership, and income). Our goal is to explore which machine learning algorithm and applicant attributes have the highest predictive power in terms of determining credit card approval. The topic of credit card approval prediction is critically important in today's financial landscape, as it addresses key challenges faced by both financial institutions and potential credit applicants. By developing a transparent machine learning model that can accurately predict credit approvals, we can help mitigate financial risk for banks while simultaneously promoting more equitable and data-driven lending practices. Ultimately, this project is an application of machine learning to a real-world problem, where we aim to develop a practical solution that improves credit card approval processes. In the sections that follow, we will be detailing our analysis on an existing credit card approval dataset in hopes of better understanding how such predictive models are built and which features are most influential.

2 Background

Recent machine learning research in credit approval has seen a focus on bias mitigation and explainable model development. Researchers have been compiling more representative datasets and developing sophisticated models to address inherent historical bias in data. With regulatory pressures to make commercial models more transparent to users, researchers have also focused on building model-agnostic explanation techniques like Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) to interpret model results. As an explorative project, our

work expands on both areas by experimenting with different supervised classification algorithms on a toy dataset and making the best model's results interpretable to the general public.

2.1 How Credit Card Approvals Work

The credit card approval process is a rigorous evaluation of an applicant's credit management capabilities. Applicants submit personal information, and lenders thoroughly examine their credit report, focusing on critical financial indicators. The credit score, derived from models like FICO or VantageScore, serves as a primary metric for assessing debt repayment reliability. Lenders analyze multiple financial indicators beyond the credit score, including debt-to-income ratio and credit utilization. The credit utilization ratio is particularly significant, as high utilization may signal financial strain. Additional factors such as credit history length, recent inquiries, and past delinquencies also inform the decision-making process. Based on this comprehensive assessment, lenders determine approval, potentially offering modified terms for applicants with limited or challenging credit histories. Success in obtaining favorable credit card terms hinges on maintaining a strong financial profile, characterized by a high credit score, low credit utilization, and consistent payment history.

2.2 Supervised Learning Explained

Supervised classification is a type of machine learning task where the goal is to map input features to specific predefined output labels based on training data. It relies on labeled datasets, where each data point consists of input variables (features) and a corresponding output label (target). The model learns the relationship between features and labels during training and uses this understanding to predict labels for unseen data. In the context of our project, supervised classification involves predicting whether a credit card application will be approved (label 1) or denied (label 0) based on socio-demographic and financial features of applicants. The process of supervised classification typically involves three stages: training, validation, and testing. During training, the model learns patterns from labeled data by minimizing a loss function that quantifies prediction errors. Validation data is used to tune hyperparameters and monitor performance to avoid overfitting or underfitting, while testing assesses the final performance of the model on unseen data.

2.3 Underfitting and Overfitting

Underfitting and overfitting are the two problems machine learning models face. A model underfits when it cannot capture patterns in data well enough to make predictions (high bias). Some common ways to curb underfitting include adding/changing type of features, reducing regularization, and adjusting model architecture. A model overfits when it is too expressive for the amount of data (high variance). In addition to adding more data, some common ways to reduce overfitting include adding regularization, dropping/changing type of features, and reducing model size.

3 Method

The main machine learning method that we applied in this project is supervised classification. We applied several algorithms to the dataset so that we can compare their results. We selected Ordinary Least Squares (OLS), Logistic Regression, Gaussian Discriminant Analysis (GDA), Naive Bayes, and K-Nearest Neighbors (KNN).

3.1 Dataset

The dataset has been preprocessed to ensure it is suitable for the models, and multiple models are evaluated using balanced accuracy as the primary metric. We dropped irrelevant features like BankCustomer, Industry, Citizen, and ZipCode. The target variable is Approved, which indicates whether a customer was approved or not.

3.2 Data Preprocessing and Splitting

We applied z-score standardization to the numerical columns to ensure that all features are on the same scale. This helps prevent features with larger ranges from dominating the model performance.

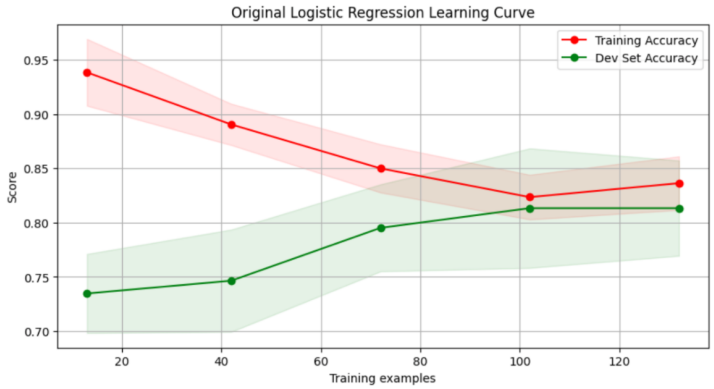

Figure 1: Original accuracy curve.

Additionally, the categorical feature Ethnicity was encoded using one-hot encoding to represent each category as a binary vector, enabling the model to handle categorical data. After encoding, all the columns were converted to integer types. The dataset was split into three subsets: training, testing, and development sets. The test set accounts for 30% of the data and is used for final evaluation. The training set comprises 70% of the data and is used for model training. 20% of the training data is set aside as the dev set, which is used to adjust the model and check its performance during training. This split helps ensure the model is evaluated correctly and reduces the risk of overfitting.

3.3 Metrics

We selected balanced accuracy to evaluate model performance for our classification task. Unlike standard accuracy, balanced accuracy provides a nuanced assessment by averaging sensitivity (correctly identifying approved applicants) and specificity (correctly rejecting unqualified applicants). This metric is crucial in credit risk assessment, where both false positives and false negatives can have significant financial and reputational implications, ensuring a fair and comprehensive evaluation of the model's predictive power.

4 Experimental Analysis

Five different models were built using default parameters and their accuracies were observed using balanced accuracy. The model that performed the best and made most sense within the context of the dataset – small dataset with two classes and 16 features – was chosen, logistic regression. Error analysis was conducted by reviewing data points that the model misclassified. However, pinpointing the source of the errors visually was difficult due to the complexity introduced by the 16 features, so alternative methods were explored, including a bias/variance analysis. In the first iteration as seen in Figure 1, the model yielded a train error of 16.8% and a dev error of 15.4%, with an optimal error of around 10%. The errors were similar, indicating that the model was generalizing reasonably well to the dev set and not overly sensitive to the training data. However, the absolute error levels were still relatively high, suggesting potential underfitting (high bias). This indicated that the model might be too simple to capture the underlying patterns in the data. Since the dataset likely contains non-linear relationships, adding polynomial features, specifically quadratic ones, was considered to reduce bias.

After adding quadratic features in the second iteration, the train error improved to 12.5%, but the dev error remained at 15.4%. This indicated a potential overfitting problem, likely due to the small size of the dataset, which led the model to learn features too quickly. This overfitting could also be due to an overly expressive model relative to the sample size, or insufficient regularization. Given that adding more data was not an option, and that quadratic features were the least expressive form of polynomial features, regularization was identified as the next potential solution. The default logistic regression in scikit-learn uses L2 regularization, which explained the results seen thus far. Thus, L1 regularization was tested in the third iteration.

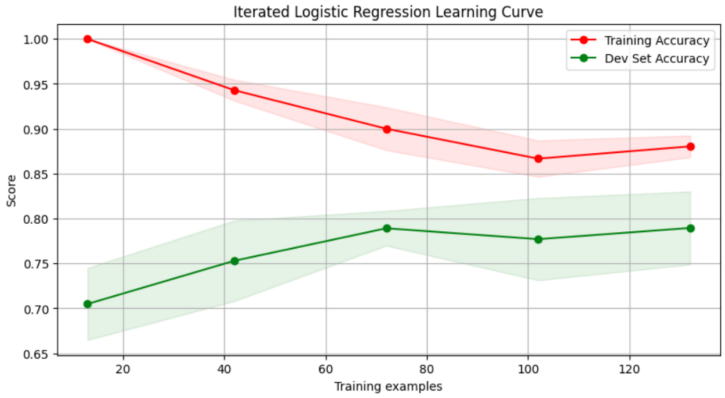

Figure 2: Final accuracy curve.

In the third iteration, using L1 regularization, the train error increased to 15%, and the dev error decreased to 10%. The model was now underfitting significantly, likely due to L1 regularization's sparsity effect, which eliminated less important features. Some important features might have been removed as a result, so a combination of L1 and L2 regularization, with a stronger emphasis on L2, was tried in subsequent iterations.

Iterations four through seven involved adjusting the L1 ratio: the train error ranged from 12.4% to 13.6%, and the dev error remained at 15.4%. Despite the changes, the model did not significantly improve beyond the 12.4% train error rate achieved in iteration six, where a L1 ratio of 0.3 was used. After exploring various regularization techniques and adding polynomial features, it was concluded that the best performance, achieving the lowest possible error rate, was attained using logistic regression with quadratic feature addition and Elastic Net regularization with an L1 ratio of 0.3, as seen in Figure 2.

5 Discussion and Prior Work

Our experimental analysis revealed the complexities of building credit approval prediction models. One of the key takeaways from our experiments is that even with a relatively small and straightforward dataset, achieving an optimal balance between bias and variance can be difficult. Logistic regression, a simple yet widely used algorithm, proved to be a suitable baseline model. However, its default implementation struggled to capture the underlying complexity of the data, as evidenced by the initial underfitting observed in Iteration 1. By incorporating quadratic features, we were able to account for non-linear relationships between the socio-demographic and financial attributes, however, this approach led to overfitting. We addressed this challenge through regularization techniques, ultimately finding an optimal balance with Elastic Net regularization and an L1 ratio of 0.3. Our best-performing model achieved a train error of 12.4% and dev error of 15.4%, demonstrating significant improvement from the initial linear model while effectively managing the bias-variance trade-off.

Another interesting result from our model is the importance our model assigned to various features in determining approval, as shown in Figure 3. The model demonstrates disparate predictive probabilities based on employment duration and ethnicity, with white applicants with stable employment receiving higher approval predictions, while applicants identifying as "Other" or carrying debt face lower approval probabilities.

These findings underscore critical ethical considerations in algorithmic credit scoring, particularly regarding potential demographic discrimination. While machine learning models offer data processing advantages, the presence of potentially discriminatory features necessitates careful scrutiny to ensure fair and equitable decision-making processes.

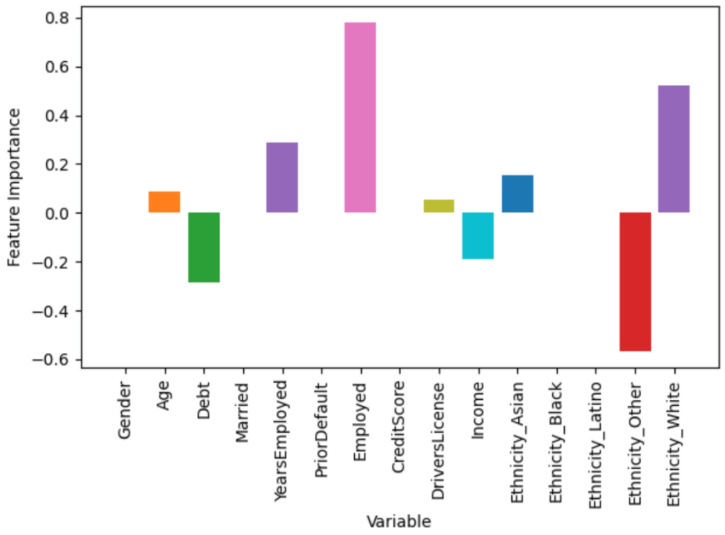

Figure 3: Feature importance graph.

6 Conclusion

Our work contributes to the growing research on transparent and fair machine learning models for credit scoring by demonstrating how feature engineering and regularization techniques can improve predictive performance while revealing potential biases. Future research should focus on using this research to develop bias mitigation and data collection strategies to enhance the interpretability and equity of algorithmic credit decision-making.