Statistical Analysis of Credit Score based on Bayesian Methods

Seturaj Matroja
University of Waterloo, 21064444

I. INTRODUCTION

Assessing individuals' creditworthiness is a critical function in the financial sector. Traditional credit scoring systems typically rely on naive approaches, providing single-point estimates that overlook uncertainty and important factors. This report introduces innovative Bayesian Statistical approach to credit score analysis. By implementing a Bayesian framework, this methodology enables the integration of prior knowledge and the generation of probability distributions, thereby enhancing the accuracy and reliability of categorizing credit scores as 'Good' or 'Bad'. The analysis is conducted on a dataset obtained from OpenML website, an open source repository containing a variety of demographic and financial variables associated with creditworthiness. My primary goal is on analyzing the credit score using , "Credit Score Class" as the outcome variable and how multiple factors like 'credit amount', 'age', 'savings account', checking accounts' influence the credit score.

II. METHODS

A. Data Acquisition and Feature Selection

The dataset, source from OpenML website, comprises of features such as age, employment, duration, debts, credit history, and others pertinent to determining credit scores. Each record corresponds to an individuals' credit profile and is labeled with a credit score, which we aim to predict as 'good' or 'bad'.

The feature selection is crucial when building a strong statistical model, especially with datasets that have many features. The dataset used for the anlaysis has 20 features, which could make the model too complicated and overfit. The feature selection was used to pick out the most important features by applying generalized linear models (GLMs). GLMs help to see which features are most likely to predict creditworthiness. The p-values of each feature were looked at and the p-values < 0.05 were chosen and kept in the model.

The generalized linear model with a logit link function for the logistic regression is given by:

$$\begin{aligned}
z_i = {} & \beta_0 + \beta_1 \cdot \text{checking\_status}_i + \beta_2 \cdot \text{duration}_i + \beta_3 \cdot \text{credit\_history}_i \\
& + \beta_4 \cdot \text{purpose}_i + \beta_5 \cdot \text{credit\_amount}_i + \beta_6 \cdot \text{savings\_status}_i \\
& + \beta_7 \cdot \text{employment}_i + \beta_8 \cdot \text{installment\_commitment}_i + \beta_9 \cdot \text{age}_i \\
& + \beta_{10} \cdot \text{other\_parties}_i + \beta_{11} \cdot \text{residence\_since}_i + \beta_{12} \cdot \text{job}_i \\
& + \beta_{13} \cdot \text{personal\_status}_i + \beta_{14} \cdot \text{other\_payment\_plans}_i + \\
& + \beta_{15} \cdot \text{housing}_i + \beta_{16} \cdot \text{existing\_credits}_i + \\
& + \beta_{17} \cdot \text{property\_magnitude}_i + \beta_{18} \cdot \text{num\_dependents}_i \\
& + \beta_{19} \cdot \text{own\_telephone}_i + \beta_{20} \cdot \text{foreign\_worker}_i \\
q_i = {} & \frac{1}{1 + \exp(-z_i)} \\
y_i \sim {} & \text{Binomial}(n, q_i)
\end{aligned}$$

where β_0 is the intercept, and y_i is the class.

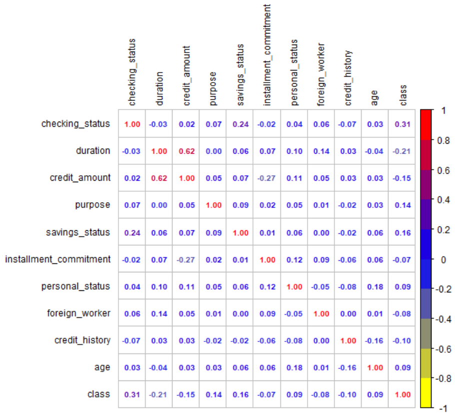
After training the glm model the features having p-value (<0.05) were chosen as the potential features including age and job. However, through stepwise regression analysis revealed that the inclusion of job did not enhance the model. This was reflected in a superior Akaike Information Criterion (AIC) value when the job variable was omitted. Consequently, the optimized feature set for predicting the target variable, class, was finalized to include checking_status, duration, credit_amount, purpose, savings_status, installment_commitment, personal_status, foreign_worker, credit_history, and age.

B. Correlation between the variables

The dataset featured a substantial number of categorical variables, requiring their conversion to a numerical format through encoding techniques. Classes were labeled; with 'Good' assigned a value of 1 and 'Bad' designated as 0. This encoding scheme was consistently applied across all categorical data within the dataset to facilitate quantitative analysis. To understand the dependencies among variables, a correlation matrix was constructed, revealing the degree to which one variable may influence another. This step was crucial for identifying potential confounders for the subsequent modeling process.

C. Mediators, Moderators and Confounders

The mediator is a variable that forms part of the casual pathway between a predictor and an outcome. The mediators are essential to explain how or why a



	checking_status	duration	credit_amount	purpose	savings_status	installment_commitment	personal_status	foreign_worker	credit_history	age	class
checking_status	1.00	-0.03	0.02	0.07	0.24	-0.02	0.04	0.06	-0.07	0.03	0.31
duration	-0.03	1.00	0.62	0.00	0.06	0.07	0.10	0.14	0.03	-0.04	-0.21
credit_amount	0.02	0.62	1.00	0.05	0.07	-0.27	0.11	0.05	0.03	0.03	-0.15
purpose	0.07	0.00	0.05	1.00	0.09	0.02	0.05	0.01	-0.02	0.03	0.14
savings_status	0.24	0.06	0.07	0.09	1.00	0.01	0.06	0.00	-0.02	0.06	0.16
installment_commitment	-0.02	0.07	-0.27	0.02	0.01	1.00	0.12	0.09	-0.06	0.06	-0.07
personal_status	0.04	0.10	0.11	0.05	0.06	0.12	1.00	-0.05	-0.08	0.18	0.09
foreign_worker	0.06	0.14	0.05	0.01	0.00	0.09	-0.05	1.00	0.00	0.01	-0.08
credit_history	-0.07	0.03	0.03	-0.02	-0.02	-0.06	-0.08	0.00	1.00	-0.16	-0.10
age	0.03	-0.04	0.03	0.03	0.06	0.06	0.18	0.01	-0.16	1.00	0.09
class	0.31	-0.21	-0.15	0.14	0.16	-0.07	0.09	-0.08	-0.10	0.09	1.00

Fig. 1. Correlation Matrix between the features.

relationship exists between two variables. The 'Mediation' library in R was used to find out potential mediators in the model. The Average Causal Mediation Effect (ACME) p-value served as a criterion to determine potential mediators among the features, with a threshold set at less than 0.005. Further examination of a correlation matrix indicated that 'checking_status' acts as a mediator between 'savings_status' and the outcome variable.

The Moderators are variables that alter the strength or direction of the relationship between a predictor and an outcome, without being affected by the predictor. Moderators are important in the analysis to ensure a comprehensive understanding of the relationships within the data. The statistical approach was used to detect the interaction effects between pairs of features in a regression model. In this method, generalized linear models (GLMs) are fitted for each possible combination of two features from the dataset, with the outcome variable 'class'. The model include an interaction term between the paired features. After fitting the models, the p-values for these interaction terms are check to determine the statistically significant moderation effect. When the p-value is below the cutoff point of 0.05, we look closer at the pair of features to see if they might be moderators. Additionally, we check how strongly these possible moderators are related to each other. If their relationship is very weak (with a correlation of less than 0.1), then it's likely that they really are moderators.

From the analysis checking_status shows moderating effect on installment_commitment and duration shows moderating effect on savings_status The generalized linear model for finding moderators is given by:

$$\begin{aligned} z_1 &= \beta_0 + \beta_1 \cdot \text{checking\_status} + \beta_2 \cdot \text{installment\_commitment} \\ &\quad + \beta_{12} \cdot \text{checking\_status*installment\_commitment} \\ z_2 &= \beta_0 + \beta_1 \cdot \text{duration} + \beta_2 \cdot \text{savings\_status} \\ &\quad + \beta_{12} \cdot \text{duration*savings\_status} \end{aligned}$$

The confounder is an unseen factor that influences both the outcome variable and the predictor leading to a spurious association which is not casual. To identify potential confounders, a statistical method known as the 'change-in-estimate' procedure combined with a 'change in AIC value' approach was employed. This technique assesses the impact of introducing a second feature, 'feature2', on the coefficient of an initial feature, 'feature1'. Should the coefficient's change exceed 10% and the AIC value decrease, suggesting an improved model fit, 'feature2' is flagged as a potential confounder. Pairs of features with strong correlations were then considered confounders. Through this analytical process, 'duration' emerged as a confounding factor for 'credit_amount'.

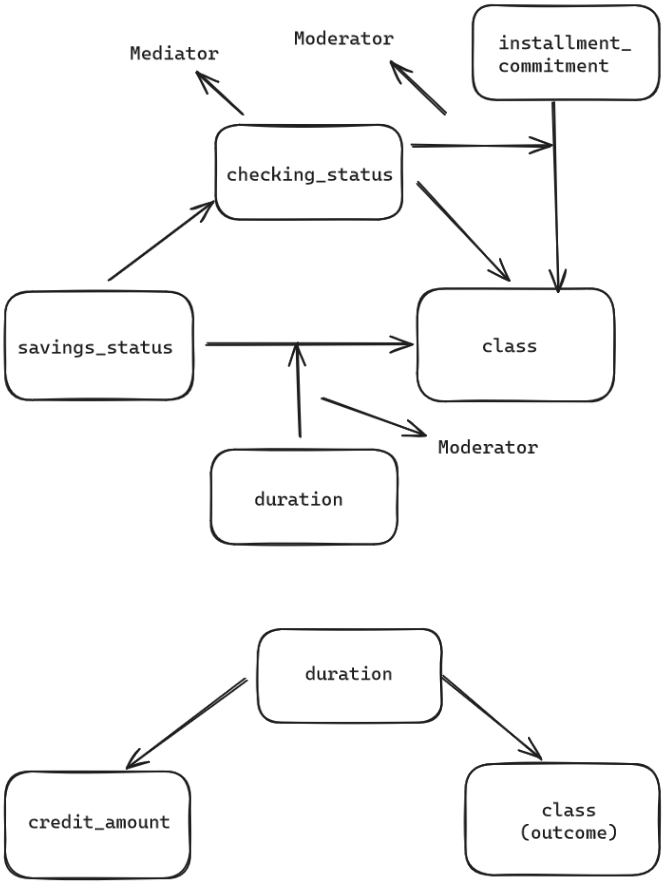


Fig. 2. Mediators , Moderators and Confounding behaviour

D. Logistic Regression and Multilevel Logistic Regression

The 'brms' package was used to perform Bayesian modeling using Markov Chain Monte Carlo (MCMC) simulation. The two models were fitted using brms and then they were compared. The number of Markov chains chosen to run was four. The formula of brms Logistic Regression is given by:

$$\begin{aligned} \text{class} \sim\ & \text{checking\_status} + \text{duration} + \text{credit\_amount} + \text{purpose} \\ & + \text{savings\_status} + \text{installment\_commitment} \\ & + \text{personal\_status} + \text{foreign\_worker} + \text{age} \end{aligned}$$

The formula of brms Multilevel Logistic Regression (Random Intercept Model) is given by:

$$\begin{aligned}\text{class} \sim\ & (1|\text{checking\_status}) + \text{duration} + \text{credit\_amount} \\ & + (1|\text{purpose}) + (1|\text{savings\_status}) \\ & + \text{installment\_commitment} + (1|\text{personal\_status}) \\ & + (1|\text{foreign\_worker}) + \text{age}\end{aligned}$$

The ’(1|variable)’ denotes a random intercept for the group defined by ’variable’ , where variable is a categorical feature.

III. RESULTS

	Logistic Regression	Multilevel Regression
Accuracy	74%	75%
R^2	0.2019	0.2324
loo diff	-14.3	0
WAIC	1039.0	1010.3
Precision	0.74	0.76
Recall	0.90	0.91
F1-score	0.81	0.83

TABLE I
COMPARISON BETWEEN MODELS

Terminology:

- **Accuracy** : Proportion of correct predictions out of all predictions made.
- **R²** : Proportion of outcome variability explained by the model (used cautiously in logistic regression).
- **loo diff** : Indicates increased error when each data point is sequentially left out, suggesting potential overfitting.
- **WAIC** : A lower score generally indicates better model prediction; higher scores may reflect greater uncertainty or poorer model fit.

The results compare a Logistic Regression model and a Multilevel Regression model:

- **Logistic Regression**: It shows 74% accuracy and explains about 20% of the variance in the dependent variable. The leave-one-out cross-validation suggests potential overfitting with a higher predictive uncertainty as indicated by the WAIC of 1039.0.
- **Multilevel Regression**: This model performs slightly better with 75% accuracy and explains approximately 23% of the variance. It appears more stable in cross-validation (loo diff of 0) and shows slightly better predictive accuracy with a WAIC of 1010.3.

Analysis: While both models show comparable performance, the Multilevel Regression model edges out with higher accuracy, a better fit, and more stability in cross-validation. The choice of model might be influenced by the data structure, with Multilevel Regression being preferable for hierarchical data.

III. DISCUSSION

A comparative analysis of Logistic Regression and Multilevel Regression models in the context of credit score evaluation is presented. The utilization of a Bayesian Statistical approach marks a significant improvement over traditional single-point estimate methods by incorporating prior knowledge and generating probability distributions, which results in more accurate and reliable classification of credit scores.

Both models demonstrated robust performance on the dataset sourced from the OpenML website. However, the Multilevel Regression model showed a marginally higher accuracy (75%) compared to the Logistic Regression model (74%). This increase in accuracy can be attributed to the model’s ability to account for hierarchical data structures and its flexibility in handling the complexities of the data.

In addition to the slight increase in accuracy, the Multilevel Regression model also showcased greater stability in cross-validation (loo diff) and an improved WAIC score (1010.3), indicating a better predictive accuracy and lower uncertainty in its estimations.Precision and recall metrics for the Multilevel Regression model were better compared to Logistic Regression, leading to a better F1-score. The improved F1-score for the Multilevel Regression model, which considers both precision and recall, confirms its strong performance in making accurate predictions. This means the model is good at identifying relevant outcomes and minimizing incorrect predictions.

In the analysis, understanding how different factors interact with each other was key. Mediators were used to explain the processes connecting the predictors with the credit score outcome. Moderators were identified to understand which variables might change the impact of the predictors on the outcome. Lastly, the role of confounders was considered to make sure that the relationships being studied were true and not due to some other hidden factor. This careful consideration of mediators, moderators, and confounders helped to ensure a more accurate analysis.

A decision tree analysis on the same dataset yielded a lower accuracy of 72.5%, reinforcing the Multilevel Regression model’s superiority. Decision trees may not capture the same level of detail in data relationships as multilevel models.

Looking ahead, with the ongoing advancements in Artificial Intelligence, neural networks are poised to become a game-changer in predictive analytics. Their ability to learn complex patterns through deep learning offers the potential to significantly enhance the accuracy of credit scoring models. As computational power and algorithmic innovations continue to grow, neural networks could provide even more accurate insights, far surpassing the capabilities of traditional statistical models.

REFERENCES

[1] Tingley, D., Yamamoto, T., Hirose, K., Keele, L., Imai, K. (2014). Mediation: R package for causal mediation analysis.

[2] Friendly M. Visualizing GLMs for binary outcomes.

[3] https://sciwiki.fredhutch.org/gendemos/Confounding/.

[4] https://advstats.psychstat.org/book/moderation/index.php.