Assessing Default Risk in Credit Card Payments: A Statistical Approach

Binitha Chandrasena

04 December, 2024

I. Introduction

In a well-developed financial system, crisis management is on the downstream and risk prediction is on the upstream. The major purpose of risk prediction is to use financial information, such as business financial statement, customer transaction and repayment records, etc., to predict business performance or individual customers' credit risk and to reduce the damage and uncertainty. In this project I will utilize the Taiwan Credit Card Default dataset from the UC Irvine Machine Learning Repository, which has information regarding client demographics, payments, bank balances during different periods etc. upto 25 predictors to address this pressing issue. To predict default risk, this analysis employs two advanced econometric techniques: Lasso regression with cross-validation and k-nearest neighbors (KNN) classification with cross-validation. Lasso regression is particularly effective for variable selection and addressing multicollinearity, while KNN offers a flexible, non-parametric approach to classification. A key focus of this study is to evaluate the role of potentially discriminating variables, such as gender and age, in predicting default risk. By comparing model performance both excluding and including these variables this analysis seeks to understand their impact on classification error and accuracy, while also addressing broader implications for fairness and ethical decision-making in credit scoring.This research aims to provide actionable insights into the design of predictive models that balance accuracy with fairness. The findings will contribute to the development of more equitable credit risk assessment frameworks, offering guidance for financial institutions and policymakers alike.

II. Research Question

How do Lasso regression and k-nearest neighbors (KNN) classification compare in terms of predictive accuracy for credit card default risk prediction in the Taiwan Credit Card Default dataset?

The primary aim of this study is to compare the predictive accuracy of Lasso regression and k-nearest neighbors (KNN) classification in forecasting credit card default risk using the Taiwan Credit Card Default dataset. To address this, I will first apply both models with the inclusion of demographic variables—specifically age and gender—and evaluate their performance. Following this, I will re-assess the models excluding these variables and compare the differences in predictive accuracy, classification errors, and overall performance. By analyzing the impact of including or excluding these sensitive demographic features, this study aims to provide insights into how such variables influence model outcomes and to explore the trade-off between accuracy and fairness in predictive modeling.

III. Methods

Method 1: K-nearest neighbor classifiers (KNN)

K-nearest neighbor (KNN) classifiers are based on learning by analogy. When given an unknown sample, a KNN classifier searches the pattern space for the KNN that are closest to the unknown sample. Closeness is defined in terms of distance. The unknown sample is assigned the most common class among its KNN. The major advantage of this approach is that it is not required to establish predictive model before classification. The disadvantages are that KNN does not produce a simple classification probability formula and its predictive accuracy is highly affected by the measure of distance and the cardinality k of the neighborhood.

Method 2: Lasso Regression with Cross-Validation

Lasso regression, or Least Absolute Shrinkage and Selection Operator, is a linear modeling technique that performs both variable selection and regularization to enhance predictive accuracy and interpretability. By penalizing the absolute size of the regression coefficients, Lasso forces some coefficients to become exactly zero, effectively removing less important predictors from the model. This makes it particularly useful for high-dimensional datasets with many predictors. Cross-validation is commonly employed with Lasso to determine the optimal penalty parameter (lambda) that minimizes prediction error. A key advantage of this approach is its ability to handle multicollinearity and reduce model complexity. However, Lasso regression may struggle when dealing with datasets where relevant predictors are highly correlated, as it tends to select one and ignore the others.

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|$$

IV. Data

The Taiwan credit card default data set takes payment data in October 2005 from a bank which has 25,000 observations, amongst the observations 22.2% are the cardholders with default payments (1=Yes, 0=No) as the response.

- X1: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.
- X2: Gender (1 = male; 2 = female).
- X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).
- X4: Marital status (1 = married; 2 = single; 3 = others).
- X5: Age (year).
- X6-X11: History of past payment. We tracked the past monthly payment records (from April to September, 2005) as follows: X6 = the repayment status in September, 2005; X7 = the repayment status in August, 2005; . . . ; X11 = the repayment status in April, 2005. The measurement scale for the repayment status is: -1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; . . . ; 8 = payment delay for eight months; 9 = payment delay for nine months and above.
- X12-X17: Amount of bill statement (NT dollar). X12 = amount of bill statement in September, 2005; X13 = amount of bill statement in August, 2005; . . . ; X17 = amount of bill statement in April, 2005.
- X18-X23: Amount of previous payment (NT dollar). X18 = amount paid in September, 2005; X19 = amount paid in August, 2005; . . . ; X23 = amount paid in April, 2005.

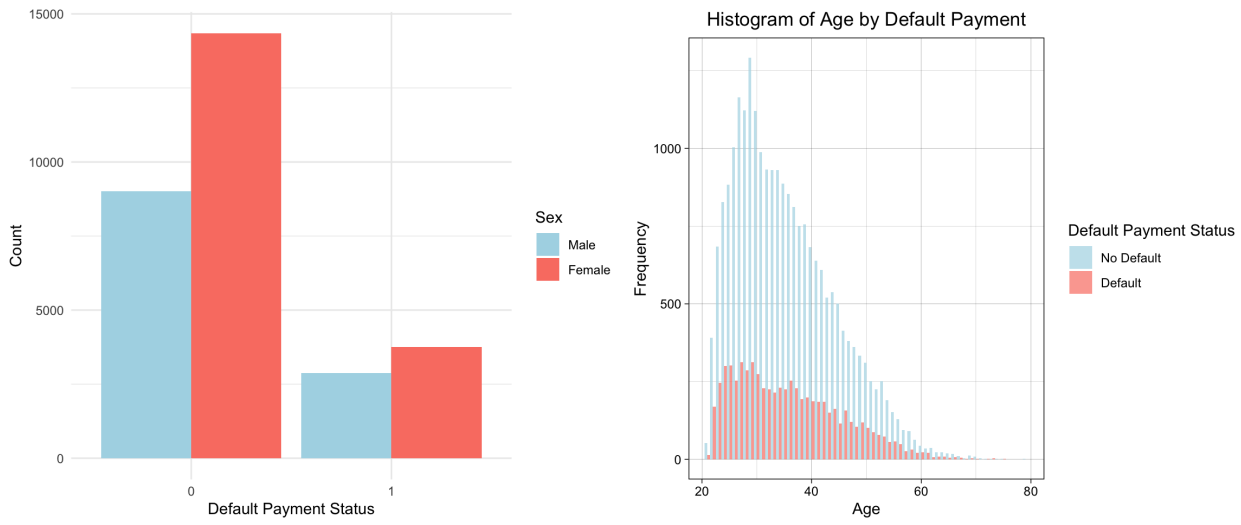

Figure 1: Sex and Age Plot

- Graph on the left: Female payments significantly outnumber male payments, and non-defaults dominate in both sexes. However, males seem to have a slightly higher proportion of defaults.
- Graph on the right: Younger individuals (around 20-30 years) are more likely to have non-default payments, but defaults tend to increase around the 30-40 age range.

From the graphs, there is no clear, substantial evidence that age or sex uniquely influence the likelihood of the payment defaulting in a way that would justify their exclusion from predictive models. The distribution of both variables shows no strong or obvious trends that decisively separate defaulted individuals from non-defaulted ones. This suggests that while these variables may have some association with the payment defaulting, their effect is not enough to remove them from the next step which is building the model. Therefore, in the context of building a predictive model for credit payment default risk, it is important to retain these variables, especially when considering fairness and ethical decision-making. It is also important to note that there seems to be a inbalance in the representation of the males to females ratio in this datset. Another interesting feature of the graph to the right is that people in their early and mid 20's default greatly in proportion to the the non defaulters and the default rates reduce as clients become older.

V. Results

KNN with Cross Validation

The two KNN models, one including SEX and AGE features and the other excluding them, show similar performance characteristics, with only a marginal 0.17 percentage point difference in overall accuracy. The model with demographic features (SEX and AGE) achieved 77.03% accuracy, compared to 76.86% for the model without these features. This suggests that demographic variables have minimal impact on the model's predictive power in this specific credit default scenario.

Examining the confusion matrices reveals critical nuances in model performance. Both models exhibit significant class imbalance and poor discriminative ability for default cases. The sensitivity (true positive rate) is high at around 95% for the non-default class (0), indicating the models are excellent at identifying non-default instances. However, the specificity is extremely low (12-13%), meaning the models struggle dramatically to correctly identify actual default cases. The positive predictive value (around 79%) and negative predictive value (around 42-43%) further underscore this challenge. The No Information Rate of 0.7788 reveals a heavily skewed dataset where simply predicting all cases as non-default would yield similar performance.

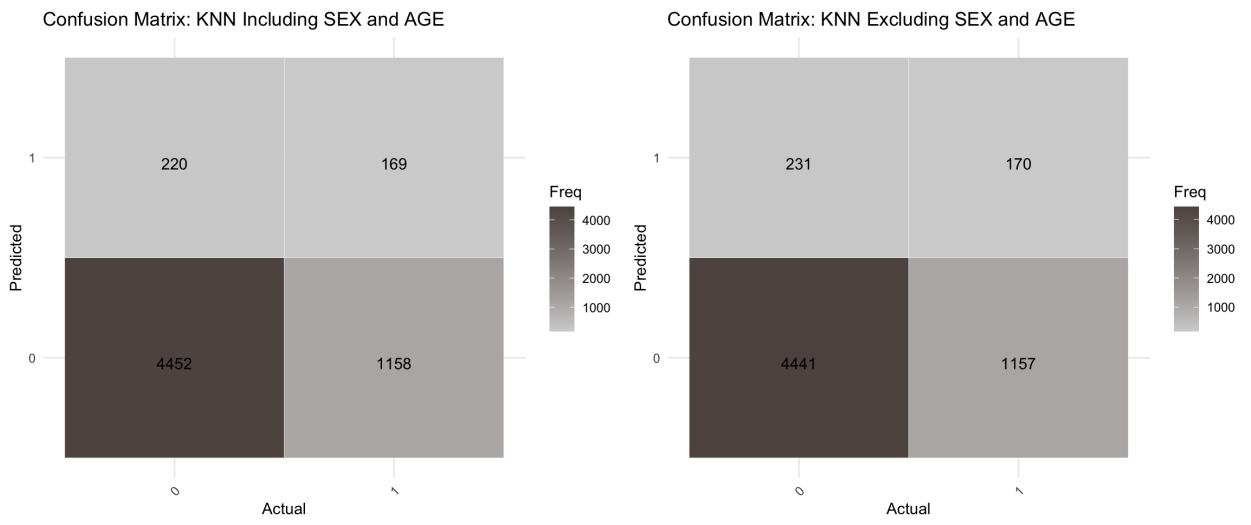

Figure 2: Sex and Age Plot

In this context specificity is more important than sensitivity. In credit default prediction, specificity measures the model's ability to correctly identify actual default cases (true positives) among all the predicted defaults. A high specificity is crucial because:

1. False Negatives are Costly: Missing a potential default (false negative) can result in significant financial losses for a lending institution. Each undetected high-risk borrower represents a potential unrecovered loan.
2. Risk Mitigation: The primary goal of default prediction is to accurately identify high-risk borrowers who are likely to default. Specificity directly measures this ability to pinpoint potential defaults.

In these models, the extremely low specificity (around 12-13%) is a major concern. This means the models are very poor at correctly identifying actual default cases, which is precisely the most critical aspect of credit risk assessment. The high sensitivity (95%) indicates the models are good at identifying non-default cases, but this is less valuable when the primary objective is to detect potential defaults. The low specificity suggests these KNN models would be unreliable for real-world credit risk decision-making, as they would fail to effectively flag high-risk borrowers. A more effective model would prioritize improving specificity, even if it means sacrificing some overall accuracy or sensitivity.

Lasso with Cross Validation

Minimum MSE: The minimum MSE values observed in the two graphs are also quite close, suggesting that the models achieve comparable predictive performance regardless of whether the AGE and SEX features are included or excluded.

Optimal λ: The optimal value of λ, determined by the lowest MSE, is around -4 for both models. This indicates that the optimal level of regularization is similar when using the full set of variables versus the reduced set excluding AGE and SEX.

Curve Positioning: While the overall shapes are alike, the curve in the graph excluding AGE and SEX appears to be slightly shifted upwards compared to the one including these features. This slight difference in MSE values could suggest that the demographic variables provide a marginal improvement in the model's predictive ability, though the impact seems minimal based on the closeness of the curves.

Data Points: The red data points along the curves are also very similar between the two graphs, further reinforcing the conclusion that the inclusion or exclusion of AGE and SEX features does not significantly alter the model's fit to the observed data.

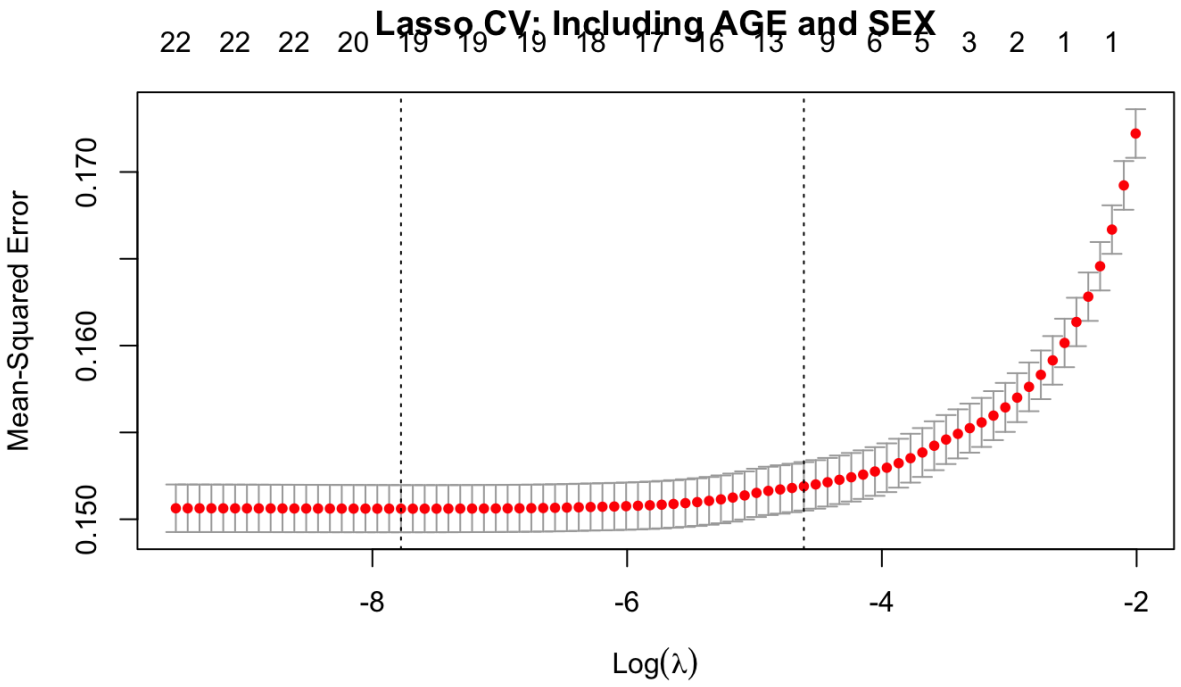

Figure 3: Sex and Age Plot

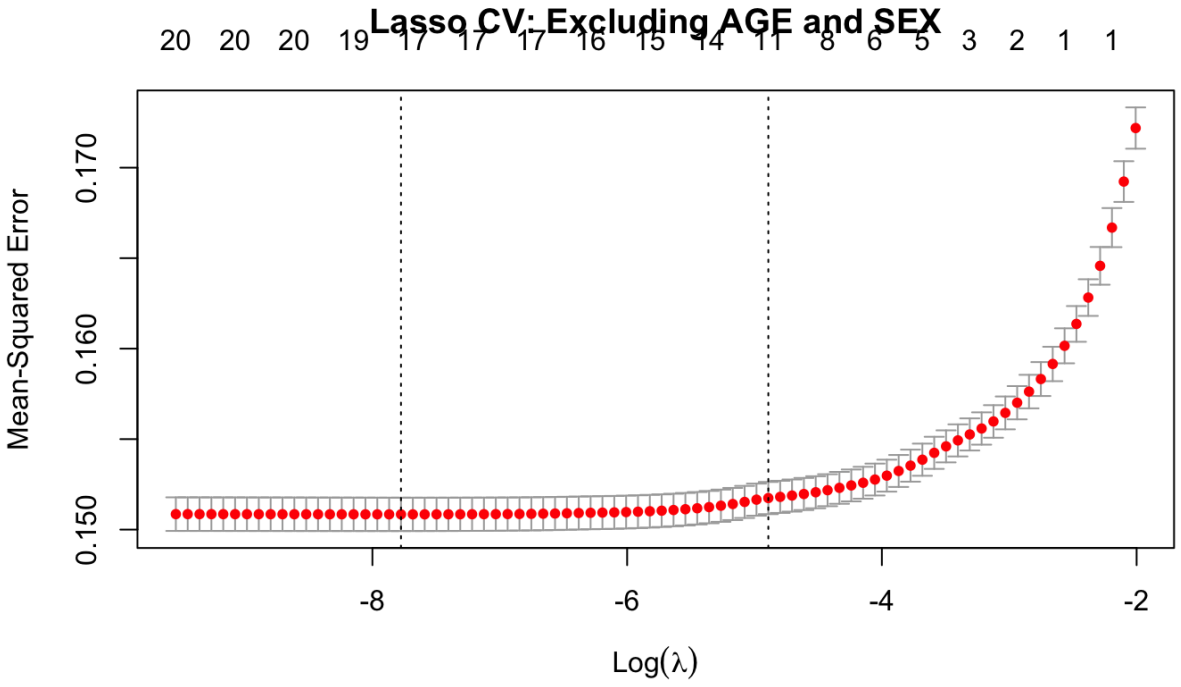

Figure 4: Sex and Age Plot

In summary, the two Lasso CV graphs demonstrate that the predictive performance of the credit default model is not heavily influenced by the presence or absence of the AGE and SEX features. The model appears to be able to achieve similar levels of accuracy regardless of whether these demographic variables are used as inputs.

comparison_table <chr>	cv_summary_stats_min_incl <chr>	cv_summary_stats_min_excl <chr>
model type	Model (Including SEX and AGE)	Model (Excluding SEX and AGE)
MSE	0.151872951180385	0.152040633373749
Test MSPE	0.151872951180385	0.152040633373749

Comparing the two sets of metrics, we can see that the inclusion of SEX and AGE features results in a slightly lower MSE and Test MSPE compared to the model that excludes these demographic variables. The differences, however, are quite small, suggesting that the SEX and AGE features do not have a significant impact on the overall predictive performance of the Lasso CV model for this dataset.

This observation aligns with the insights gained from the previous Lasso CV graphs, which showed very similar curves and optimal regularization parameters regardless of whether SEX and AGE were included or excluded. The marginal improvements in performance when using the demographic features indicate that the other variables in the dataset may be more crucial drivers of the credit default prediction task.

Overall, the comparison of these two models suggests that the exclusion of SEX and AGE features does not severely compromise the model's predictive capabilities, making it a viable option if the goal is to simplify the model or avoid potential bias or fairness issues related to the use of demographic variables.

VI. Conclusion

Both the Lasso regression and K-Nearest Neighbors (KNN) models performed similarly in predicting credit card default risk, regardless of whether the demographic variables of age and sex were included or excluded. For the KNN models, the inclusion of these variables only marginally improved the overall accuracy by 0.17 percentage points. However, the models struggled with class imbalance, exhibiting very low specificity (around 12-13%) in correctly identifying actual default cases, a critical metric for credit risk assessment. Similarly, Lasso regression models showed minimal differences in performance when age and sex were included versus excluded. The optimal regularization parameter (λ) remained the same for both cases, and the minimum Mean Squared Error (MSE) values were close, suggesting that demographic variables had little impact on predictive capabilities.

The findings indicate that for this specific credit card default dataset, other variables, such as payment history and billing amounts, are more critical drivers of the predictive models' performance than age and sex. While the slight upward shift in the Lasso regression curve when excluding these variables suggests they may provide a marginal improvement, the difference is negligible compared to the overall model performance. Thus, the exclusion of age and sex variables does not significantly compromise the predictive accuracy of either model. This suggests that financial institutions can develop more equitable credit risk assessment frameworks without heavily relying on potentially sensitive demographic characteristics, provided other relevant predictors are included.