Credit Scoring For Everyday Borrowers

Avneet Kaur & Karsten Assoua

https://github.com/avneetk194/Credit-Scoring-for-Everyday-Borrowers
https://corporatefinanceinstitute.com/resources/commercial-lending/default-rate/
https://www.kansascityfed.org/documents/7727/rwp15-02.pdf
https://www.stlouisfed.org/on-the-economy/2025/may/broad-continuing-rise-delinquent-us-credit-card-debt-revisited
Fairness and Machine Learning: Limitations and Opportunities. 2019, https://fairmlbook.org.
"Credit Scoring in the Era of Big Data." *Yale Journal of Law and Technology*, vol. 18, no. 1, 2017, pp. 148–216. https://digitalcommons.law.yale.edu/yjolt/vol18/iss1/5.
"Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads." *Management Science*, vol. 65, no. 7, 2019, pp. 2966–2981. INFORMS, https://doi.org/10.1287/mnsc.2018.3093.

Highlights

-Identified **top 10 features** influencing default risk (call duration ranked #1).
-Gradient Boosting achieved the strongest performance across accuracy and feature interpretability.
-Risk distribution chart shows most borrowers fall into low-risk category, with small percentages in medium and high risk.
-Confusion matrix analysis reveals higher accuracy for predicting "No Default" than "Default" cases, suggesting room for recall improvement.

Background

Default Rates

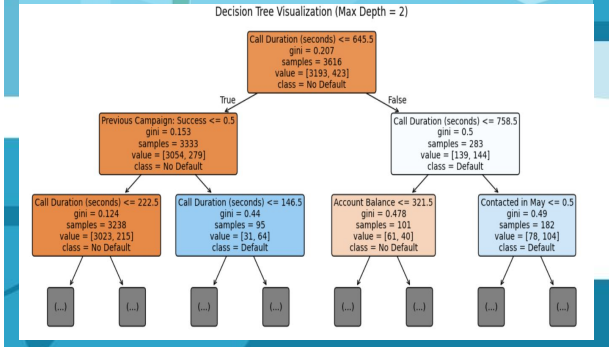
- The rate of all loans issued by a lender or financial institution that is left unpaid by the borrower and declared to be in default

Credit Scoring

- Credit score is a significant factor in determining default rates. Individuals with low credit scores are more likely to default on their loans compared to those with high credit scores

Data

- Dataset: Bank marketing dataset with borrower information such as call duration, account balance, age, and past campaign outcomes.
- Target variable: Loan default (Yes/No)
- Features: Demographics, financial details, and interaction history with the bank.



Model

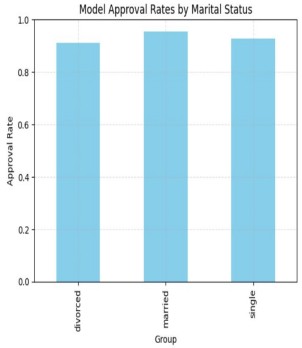
Algorithms Applied: Logistic Regression, Decision Tree, and Gradient Boosting which are all supervised learning classification models.

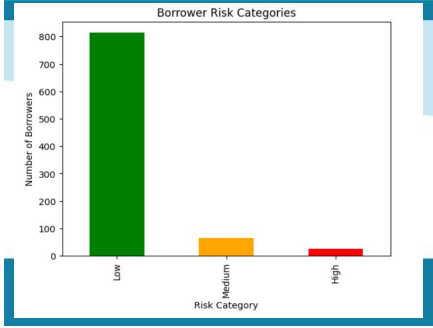
Inputs: Borrower demographics, financial account data, and marketing interaction details.

Outputs: Predicted probability and classification of borrower default risk.

What Does This Have to Do With Bias?

- Call duration is analyzed for possible bias because longer calls might be associated with particular borrower profiles.
- Indirect Bias Risk: Age and balance are two characteristics that could subtly represent demographic or socioeconomic factors.
- Mitigation: Understanding potential proxy variables that might inadvertently impact equity.





Next Steps

- Perform additional bias testing on important attributes such as age, account balance, and call duration.
- To better identify borrowers who are at danger, raise the recall score for "Default" cases.
- To make sure models are robust, test them on fresh, varied borrower datasets.
- Install a predictive model with recurring retraining and integrated fairness monitoring.

Analysis

The baseline model of logistic regression, which has an accuracy of about 80%, is interpretable but has simpler feature interactions.

Decision Tree: Clearly defined "if-then" rules for forecasting; primary splits highlight call duration and prior campaign performance as the most important indicators.

Among models, gradient boosting has the highest accuracy and is best at capturing complicated feature combinations and non-linear correlations.