International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

FRAUD DETECTION IN BANKING DATA BYMACHINE LEARNING

V.Srinivas, Professor, Department Of Data Science, SICET, Hyderabad
R.Sandeep, K.Sai Neeraj, M.Abhinav, E.Srujana
UG Student, Department Of Data Science, SICET, Hyderabad

ABSTRACT

With the spread of technology and ecommerce services, credit cards have become one of the most popul ar payment methods, causing an increase in bank accounts. Additionally, a significant amount of fraud e ntails high bank transaction fees. Therefore, fraud detection has become an interesting topic. In this wor k, we consider using class weight tuning hyperparameters to control the weight of fraudulent and legitim ate transactions. We specifically use Bayesian optimization to optimize hyperparameters while preservin g welldefined problems such as nonlinear data. We recommend using CatBoost and XGBoost as well as prioritizing weights for unbalanced data to improve the performance of the LightGBM method consideri ng the selection process. Finally, to further improve performance, we use deep learning to finetune the h yperparameters, specifically our parameter weights. We conduct some experiments on realworld data to test the proposed method. We use the regression model to add accuracy to the ROCAUC model to better cover data inconsistencies. CatBoost, LightGBM, and XGBoost were valuated separately using the 5fol d crossvalidation method. Additionally, most of the learning set selection is used to evaluate the perform ance of the clustering algorithm. The results show that LightGBM and XGBoost achieve bestinmodel wi th ROCAUC = 0.95, precision 0.79, recall 0.80, F1 score 0.79, and MCC 0.79. Using deep learning and Bayesian optimization methods to tune the hyperparameters, we also achieved ROCAUC = 0.94, precisi on = 0.80, recall = 0.82, F1 score = 0.81, MCC = 0.81. This is a significant improvement over the stateof theart method we compared.Index terms Bayesian optimization, data mining, deep learning, integrative l earning, hyperparameters, imbalanced data, machine learning

I. Introduction

In recent years, the volume of financial transactions has increased significantly due to the expansion of f inancial institutionsand the popularity of ecommerce websites. Business fraud has become a growing pro blem in online commerce, and fraud is always difficult to detect [1], [2].With the evolution of credit card s, credit card fraud patterns are also constantly evolving. Scammers go to great lengths to make themselv es known, and credit card scams are always new. Therefore, researchers are constantly trying to find ne w ways or improve the performance of existing methods [3].Criminals often exploit security, control and monitoring weaknesses in businesses to achieve their goals. However, technology can be a tool in the fi ght against fraud [4]. To prevent further fraud, it is important to detect fraud as it occurs [5].Fraud can b e defined as an illegal act of deception or fraud committed for the purpose of financial or personal gain. Credit card fraud refers to the illegal use of credit card information to make physical or digital purchases

International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

. In the digital economy, fraud can occur offline or online, as cardholders often provide their cards, expiration dates, and card verification codes over the phone or online [6].It has two strategies: antifraud and antifraud, which can be used to prevent fraud.Fraud prevention is one way to prevent fraud from happening in the first place. On the other hand, fraud detection is needed in cases where fraudsters try to commit commercial crimes [7].Detection of fraud in the banking industry is considered a dual classification problem where information is classified as legitimate or fraudulent [8]. Due to the vast amount of transaction data contained in numerous bank records and databases, it is impossible or takes a long time to review the books and find fraud patterns. Therefore, machine learningbased algorithms play an important role in fraud detection and prediction [9]. Machine learning algorithms and advanced processing power increase the ability to process big data and detect fraud more effectively. Machine learning algorithms and deep learning also provide fast and effective solutions to pressing problems [10].In this paper, we propose an effective method for credit card fraud detection, evaluated on publicly available data with the most common selection methods as well as LightGBM, XGBoost, CatBoost and logistic regression optimization. such as deep learning and hyperparameter tuning. The best fraud detection system should be able to detect many frauds, and the detection accuracy of fraud data should be very high, that is, all results must be checked accurately, which, among other things, will increase the customer's trust in the bank. It can also improve the bank's creditworthiness. Banks do not suffer losses due to false discoveries.

The main contributions of this paper are summarized as follows:III We adopt Bayesian optimization for fraud detection and propose the use of weighted modified hyperparameters to solve the problem of inconsistent data according to the previous step. We also recommend using CatBoost and XGBoost with LightGBM to improve performance. We use the XGBoost algorithm because of the high speed and constant processing time of large datasets, it overcomes overfitting by measuring the complexity of the tree and does not need to spend a lot of time tuning hyperparameters. We also used the Catboost algorithm since there is no need to tune hyperparameters for control and it achieves good results without changing hyperparameters compared to other machine learning algorithms.â We propose a majority vote on the working implementation combining CatBoost, We also recommend using deep learning to unpack and fine-tune hyperparameters.We conduct extensive testing on real-world data to evaluate the effectiveness of the plan. We use inverse precision data in addition to the commonly used ROCAUC to better cover inconsistent data. We also use F1_score and MCC metrics to measure performance. As a result, thecurrent way the request is issued and the process accordingly. We use publicly available data for our analysis and publish publicly available data for use by other researchers.The report of this article is organized as follows: In Part II, we examine the current situation. Part III introduces the credit card fraud detection process, including data, prioritization, extraction and specific selection, algorithms, framework, and evaluation. Section 4 discusses the evaluation of experiments and finally Section 5 concludes the paper.

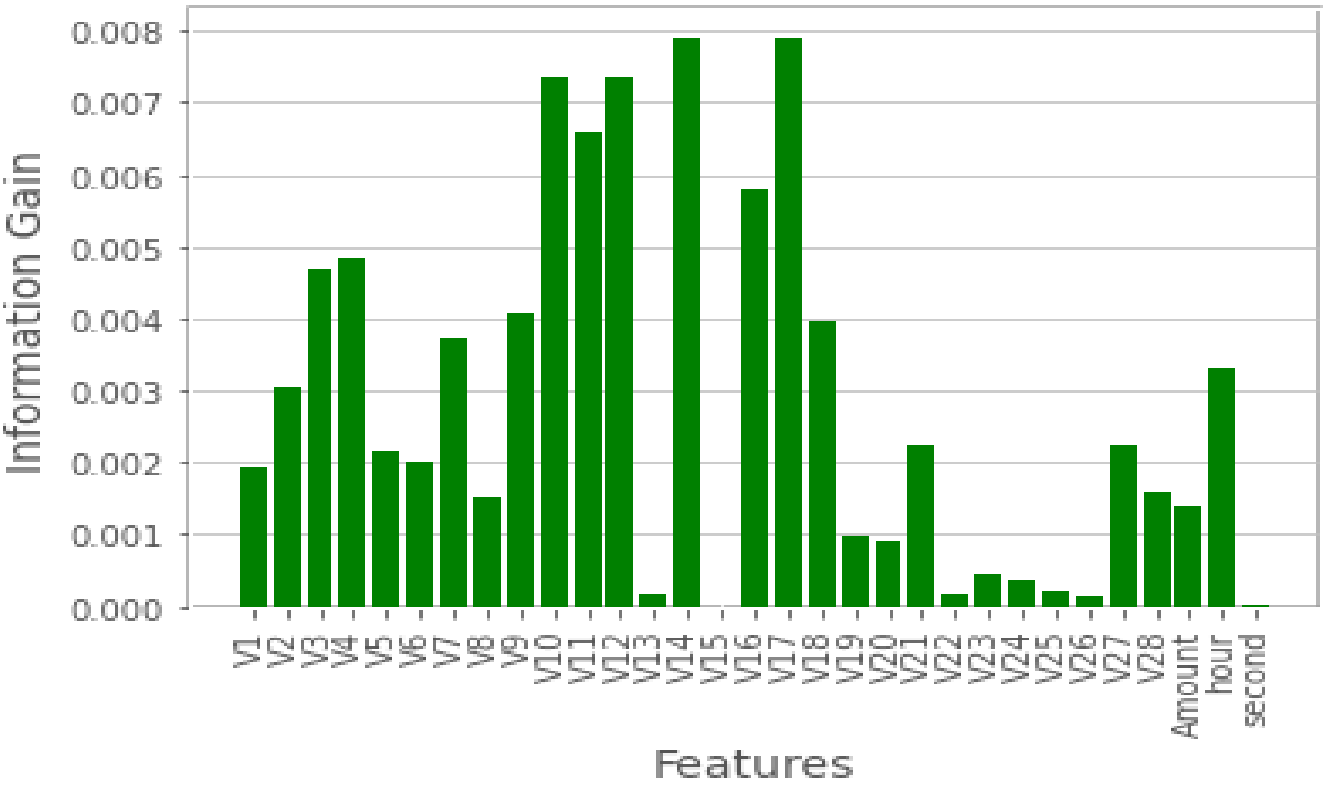
They evaluated five classification algorithms and found that supervised vector classifiers and logistic regression classifiers outperformed other algorithms on unbalanced data [20]. A summary of the literature review is shown in Figure 1.

International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

TABLE 1. The features of the credit-card fraud dataset that is used in this paper.

Variable Name	Description	Type
$V_1, V_2, \ldots, V_{28}$	Transaction feature after PCA transformation	Integer
Time	Seconds elapsed between each transaction with the first transaction	Integer
Amount	Transaction Value	Integer
Class	Legitimate or Fraudlent	0 or 1

Performance issues in machine learning algorithms and ranking with most models will affect evaluation results [6]. Therefore, many studies have adopted undersampling and oversampling methods to solve the problem of incons istent data [15]. Using a low sample size may result in data loss [21]. Additionally, using oversampling techniques can result in overlapping of nonshared data (data and data are different, this topic is discussed in “Entropy”). Som e researchers use lowenergy object (SMOTE) as a solution, which overcomes the disadvantages of undersampling and oversampling [5], [17], [22]. However, SMOTE's approach leads to an increase in false positives, which is un acceptable in the consumer market. To solve this problem, in this work, we use class weight tuning hyperparamet ers to solve the above problems [5], [17], [22]. However, SMOTE's approach leads to an increase in false positive s, which is unacceptable in the consumer market. To solve this problem, in this work, we use class parameters to t une hyperparameters to solve the above problems.

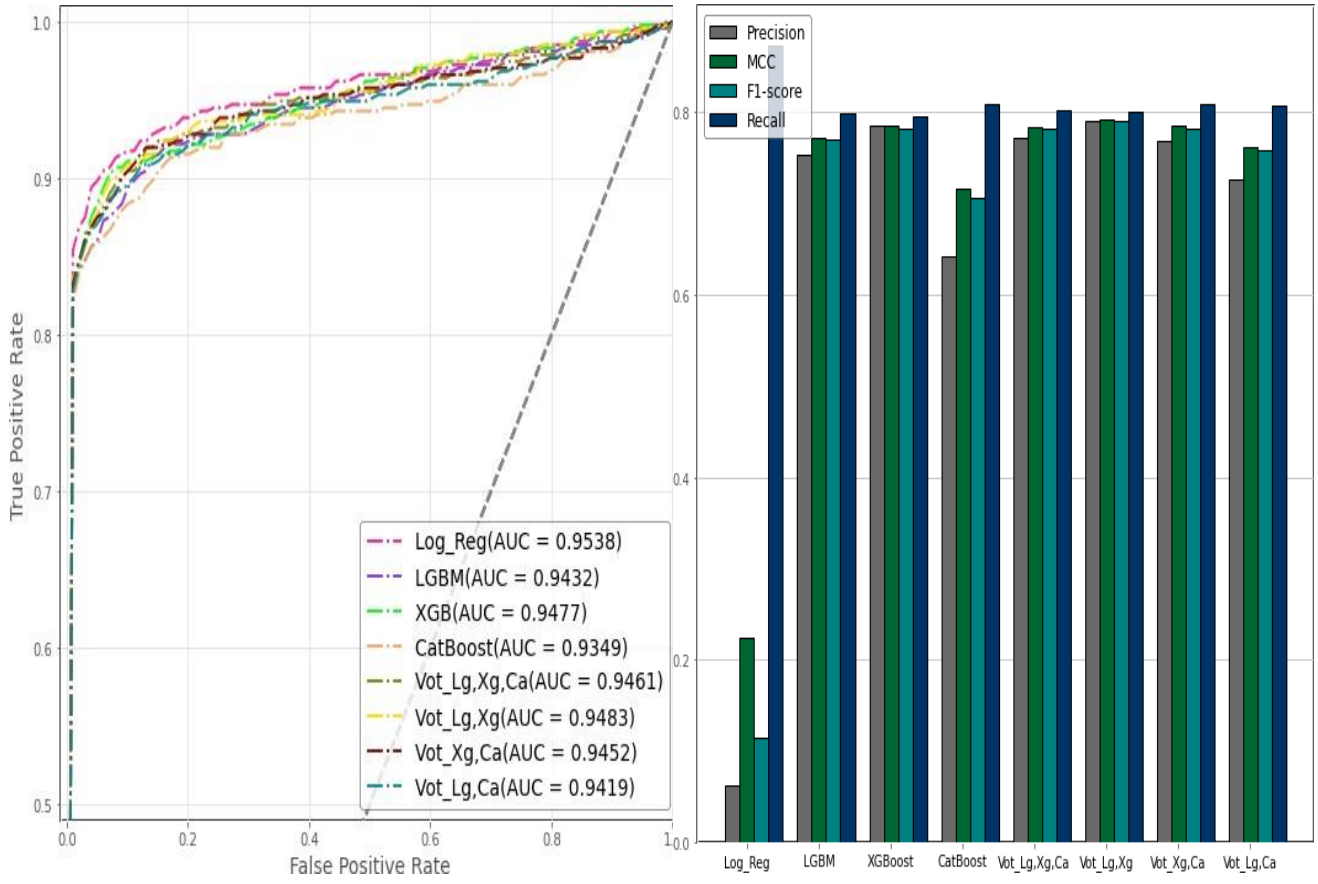


International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

FIGURE 3. Feature importance diagram that shows the IG for the unknown features of the "creditcard" dataset. The top six features are used in evaluations.

IV Experimental results and discussion

We evaluate the performance of the proposed framework using a boosting algorithm with a 5-fold layered cross-validation method and a Bayesian optimization method. Before using the majority voting method, we extract meta-parameters and evaluate each algorithm individually. Check the algorithm with triple and double precision. The comparison results are shown in Table 5.Most studies in the literature rely on AUC plots to evaluate performance. However, as can be seen from the ROC-AUC curve in Figure 4, the AUC value is highly unbalanced for the data.F1-Score



International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

FIGURE 4. ROC_AUC Curve

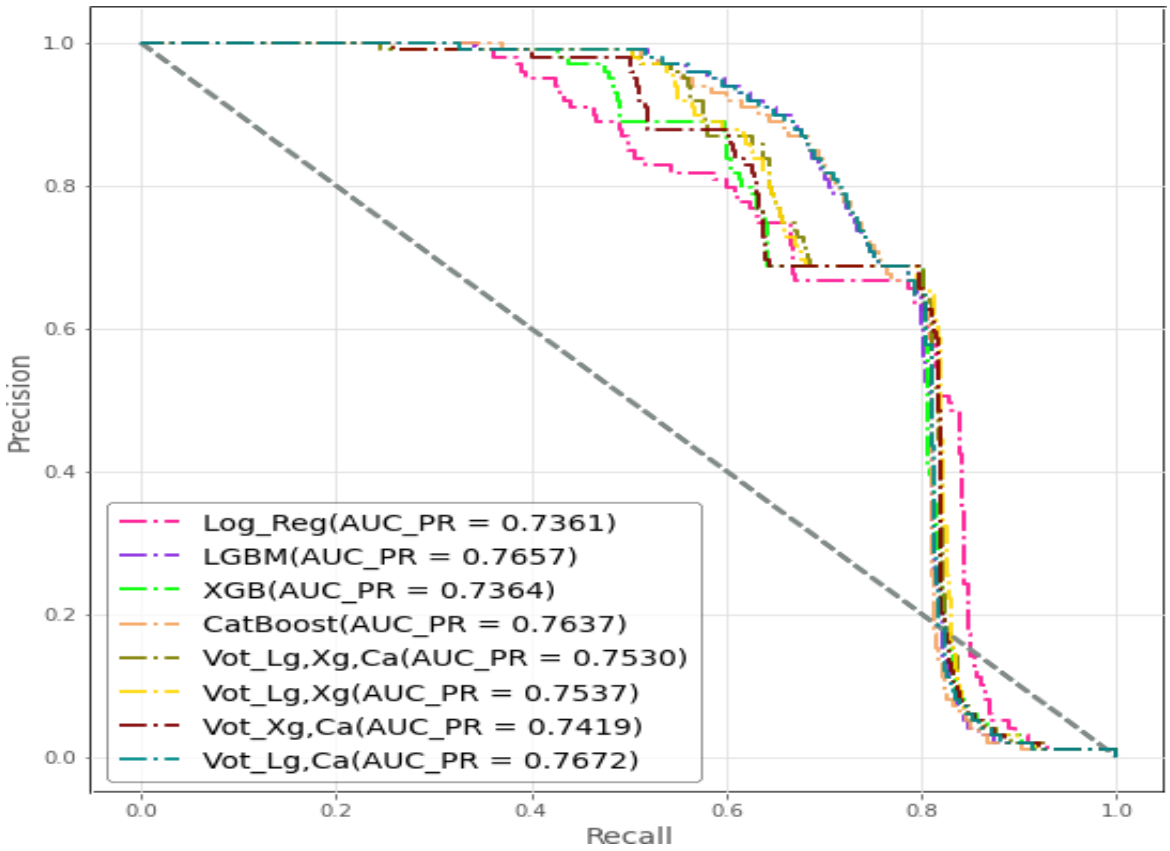


FIGURE 5. Precision_Recall Curve .

The method using signals and publicly available information outperforms the intelligent meth od proposed in [17].

V. Conclusion and Future Work

In this paper, we studied the problem of credit card fraud detection in real unbalanced data. We recommend machine learning to improve fraud performance. We used the publicly availa ble “Credit Card” dataset, which contains 28 features and 0.17% fake data. We offer two opti ons. In the LightGBM implementation, we use the weight class modifier to select appropriate hyperparameters.We use statistical metrics including accuracy, precision, recall, F1 score and AUC. Our experimental results show that the proposed LightGBM method improves fraud d etection by 50% and F1 score by 20% compared to the last model in [17]. We use most algori thms to improve the performance of the algorithm. We also improved the model using deep le arning. MCC results are validated for unbalanced data, proving to be more reliable than other

measurements. In this article, by combining the LightGBM and XGBoost method, we obtain t he learning depth0.79 and 0.81. In addition to reducing the memory and time required to eval uate the algorithm, using hyperparameters to solve for nonlinear data can lead to better results compared to standard models. For future research and studies, we propose the use of other hy brid models with specific studies in the following areas: CatBoost by changing more hyperpa rameters, especially the number of hyperparameters of the tree. Additionally, due to the hard ware limitations in this study, the use of more powerful and better equipment may lead to bett er results comparable to the results of study #1.

References

Allen, I. E., & Seaman, J. (2008). ***Staying the course: Online education in the united states, 2008.*** ERIC.

Bambrick-Santoyo, P. (2010). ***Driven by data: A practical guide to improve instruction***. John Wiley & Sons.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2023). ***lme4: Linear mixed-effects models using eigen and S4***. https://github.com/lme4/lme4/

Betebenner, D. W. (2021). ***randomNames: Generate random given and surnames***. https://CenterForAssessment.github.io/randomNames

Bransford, J. D., Brown, A. L., Cocking, R. R., et al. (2000). ***How people learn*** (Vol. 11). Washington, DC: National academy press.

Breiman, L. (2001). Random forests. ***Machine Learning***, ***45***(1), 5–32.

Bryan, J. (2017). ***Project-oriented workflow***. https://www.tidyverse.org/blog/2017/12/workflow-vs-script/

Bryan, J. (2019). ***Reproducible examples and the 'reprex' package***. https://community.rstudio.com/t/video-reproducible-examples-and-the-reprex-package/14732

Bryan, J. (2020). ***Happy git with r***. https://happygitwithr.com/

Bryk, A. S., Gomez, L. M., Grunow, A., & LeMahieu, P. G. (2015). ***Learning to improve: How america's schools can get better at getting better***. Harvard Education Press.

Campaign, D. Q. (2018). ***Teachers see the power of data - but don't have the time to use it***. https://dataqualitycampaign.org/wp-content/uploads/2018/09/DQC_DataEmpowers-Infographic.pdf

Conway, D. (2010). The data science venn diagram. ***Drew Conway***, ***10***. http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Datnow, A., & Hubbard, L. (2015). Teachers' use of assessment data to inform instruction: Lessons from the past and prospects for the future. ***Teachers College Record***, ***117***(4), n4.

Dirksen, J. (2015). ***Design for how people learn***. New Riders.

Dweck, C. (2015). Carol dweck revisits the growth mindset. ***Education Week***, ***35***(5), 20–24.

Education Statistics U.S. Department of Education, N. C. for. (2019). Concentration of public school students eligible for free or reduced-price lunch. ***The Condition of Education 2019***. https://nces.ed.gov/fastfacts/display.asp?id=898

Elbers, B. (2020). ***Tidylog: Logging for dplyr and tidyr functions***. https://github.com/elbersb/tidylog/

Emdin, C. (2016). ***For white folks who teach in the hood... And the rest of y'all too: Reality pedagogy and urban education***. Beacon Press.

Estrellado, R. A., Bovee, E. A., Motsipak, J., Rosenberg, J. M., & Vel'asquez, I. C. (2019). ***Taylor and francis book proposal for data science in education***. https://github.com/data-edu/DSIEUR_support_files/blob/master/planning/T%26F%20Book%20Proposal%20for%20Data%20Science%20in%20Education.docx

Estrellado, R., Bovee, E., Mostipak, J., Rosenberg, J., & Vel'asquez, I. (2024). ***Dataedu: Package for data science in education using r***. https://github.com/data-edu/dataedu

Firke, S. (2023). ***Janitor: Simple tools for examining and cleaning dirty data***. https://github.com/sfirke/janitor

for Education Statistics, N. C. (2018). ***Public elementary/secondary school universe survey***. https://nces.ed.gov/programs/digest/d17/tables/dt17_204.10.asp?current=yes

Gelman, A., & Hill, J. (2006). ***Data analysis using regression and multilevel/hierarchical models***. Cambridge university press.

Grimm, K. J., Ram, N., & Estabrook, R. (2016). ***Growth modeling: Structural equation and multilevel modeling approaches***. Guilford Publications.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). ***The elements of statistical learning: Data mining, inference, and prediction***. Springer Science & Business Media.

Hattie, J. (2012). ***Visible learning for teachers: Maximizing impact on learning***. Routledge.

Healy, K. (2019). ***Data visualization: A practical introduction***. Princeton University Press.

Hill, A. (2017). ***Up and running with blogdown***. https://alison.rbind.io/post/2017-06-12-up-and-running-with-blogdown/

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. ***Science***, ***349***(6245), 261–266.

Ismay, C., & Kim, A. Y. (2019). ***Statistical inference via data science***. CRC Press.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). ***An introduction to statistical learning*** (Vol. 112). Springer.

Jarvis, C. (2019). ***Creating calling***. HarperCollins.

Jordan, R. (2015). ***High-poverty schools undermine education for children of color***. https://www.urban.org/urban-wire/high-poverty-schools-undermine-education-children-color

Kahneman, D. (2011). ***Thinking fast and slow***.

Kearney, Michael W. (2016). Rtweet: Collecting twitter data. ***Comprehensive R Archive Network. Available at: Https://Cran. R-Project. Org/Package= Rtweet***.

Kearney, Michael W., Revilla Sancho, L., & Wickham, H. (2023). ***Rtweet: Collecting twitter***

data. https://docs.ropensci.org/rtweet/

Kleon, A. (2012). ***Steal like an artist: 10 things nobody told you about being creative***. Workman Publishing.

Kozol, J. (2012). ***Savage inequalities: Children in america's schools***. Broadway Books.

Krist, C., Schwarz, C. V., & Reiser, B. J. (2019). Identifying essential epistemic heuristics for guiding mechanistic reasoning in science learning. ***Journal of the Learning Sciences***, ***28***(2), 160–205.

Kuhn, M. et al. (2008). Building predictive models in r using the caret package. ***Journal of Statistical Software***, ***28***(5), 1–26.

Kuhn, M. (2023). ***Caret: Classification and regression training***. https://github.com/topepo/caret/

Kurz, S. (2019). ***Statistical rethinking with brms, ggplot2, and the tidyverse***.