Lost interest? Using artificial intelligence to predict if low-income households are at risk of meeting loan repayments

Ed Bayes, Cole Bateman

December 10, 2020

1. Introduction

Traditional approaches to managing credit risk are outdated and have been shown to perpetuate bias. Many lenders still assess applications using a small number of static features like address and occupation or credit scores, which are opaque and difficult to understand. [1] Studies show these are not only ineffective in predicting defaults [2] but have been used as proxies to discriminate based on protected characteristics like race and gender [3] despite laws against such behavior. [4]

Artificial Intelligence (AI) based credit risk models provide an opportunity to make lending more equitable [5] by better predicting future repayment issues, allowing for tailored interventions to support borrowers in financial distress. Additionally, fintech lenders using AI-based models have been found to discriminate less in application decisions while charging minority borrowers 40% less on average compared against face-to-face lenders. [6] Yet, research shows only 5.5 percent of banks use AI in risk modelling. [7]

For this paper, we built a bespoke deep learning credit risk model that predicts whether low income borrowers will default on current loans based on previous payments. The results of our experiments show that it significantly outperforms baseline models based on traditional credit scores, producing an ROC AUC of up to 0.882, compared to a baseline of 0.678. In a real-time risk prediction setting, this could be used to proactively flag nearly 50% more borrowers at risk of non-payment.

We modelled our work on LEAP (Lstm rEal-time Adherence Predictor), a long short-term memory (LSTM) recurrent neural network (RNN) model developed as part of a study by Tambe, Killian et al. to prescribe interventions for TB patients using real-time digital medication adherence data. [8] Our model - named BUST (Bank Underwriting lStm predicTor) - is trained on data from Home Credit, a global consumer lending company [9], covering over 800,000 loans and totaling over 7 million payments.

We argue that such a model can not only address biases in credit risk—by basing risk on the ability to pay rather than proxies for protected characteristics—but can also benefit both borrowers and lenders by flagging those at risk of default so lenders can more efficiently allocate resources to at-risk borrowers, consequently boosting overall repayments from their client base. [34]

2. Problem Specification

In the US, there is a long history of credit being used as a tool of discrimination, [22] with the term "redlining" originating in the lending industry to segregate neighborhoods based on race. [12] Despite laws passed in recent years to prevent such behavioral, evidence shows minority borrowers and those with low incomes still face discrimination in lending. For example:

- UCLA researchers analysed nearly 7 million 30-year mortgages and found that Black and Latino applicants were charged higher interest and refinance fees compared with white borrowers. [13]
- Loan comparison company LendingTree found through the analysis of Home Mortgage Disclosure Act (HMDA) data that Black borrowers are denied refinance loans 30% of the time compared to an average of 17%. [33]
- The Federal Reserve Bank of Chicago found in 2017 that redlining had a persistent adverse impact on neighborhoods, impacting homeownership rates, home values and credit scores. [14]

One reason for this is that current models rely on static features like occupation or address that are used as proxies for characteristics like gender and race. AI's ability to avoid the traditional credit reporting and scoring system provides a rare, if not unique, opportunity to alter the status quo and create fairer, more inclusive economic systems.

There are potential macroeconomic consequences to these problems. Unsustainable high interest loans create vulnerability to financial shocks and undermines economic growth and raises inequality. [15] Increasing rates of unemployment due to the COVID-19 crisis, given this problem another layer of urgency. [16] Behind these macroeconomic figures lie human stories, with debt leading to mental health issues and the breakdown of relationships.

3. Related Work

Underwriters are waking up to the fact that AI can help in credit risk modelling [20] with researchers and companies incorporating larger types of data into credit calculations in recent years. These include using digital footprint variables like social media profiles, email addresses, what type of computer you are using, where you buy your clothes. [21] and even suspected infidelity. [22] This has the potential to tackle financial exclusion by allowing borrowers to access credit even if they don't have previous credit histories. [23] For example, Omdena has developed a credit scoring AI system for individuals without a previous bank account. [24] However, these methods are still not commonly adopted, as research shows only 5.5 percent of banks use AI in risk modelling. [25]

An indicator of an increasing focus on big data and alternative lending models in the financial services sector is a proliferation of competitions and challenge prizes at the intersection of AI and credit risk. One such competition was hosted in 2018 by Home Credit, a consumer loan company that focuses on lending primarily to people with little or no credit history, [9] on Kaggle, a data science subsidiary of Google. [26] Over 7,000 teams analyzed a variety of alternative data, including telco and transactional information, to better predict their clients' repayment abilities.

Methodologically, our work is related to a large body of research that deals with artificial intelligence and machine learning in credit risk. [27] Neural networks are one of the most used ML methods due to the size of the data available in financial services. We leverage this fact in our own model.

4. Data

We have utilized open-source anonymized datasets from the Kaggle competition to build our model. Because labels are included in the training data and our goal is to train a model to learn to predict labels from features, this is a supervised classification task. Figure 1 summarizes the data.

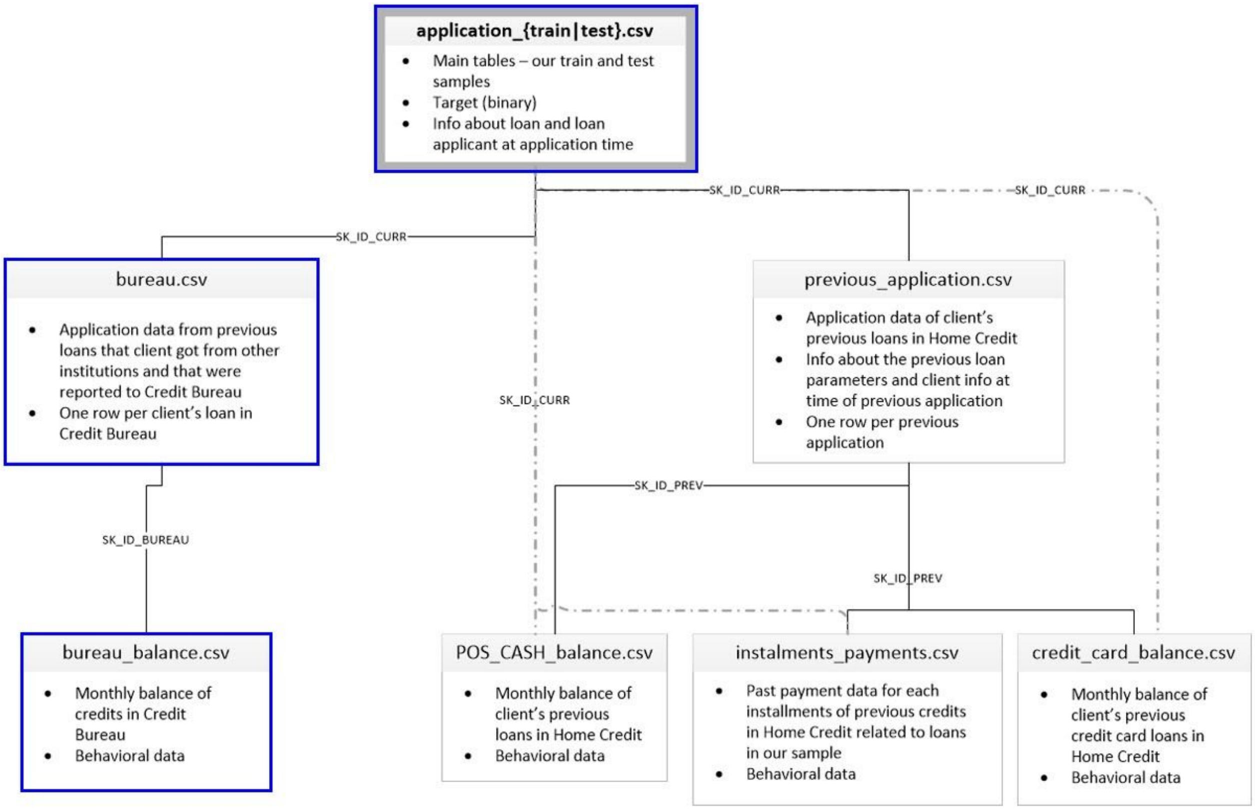
application_{train|test}.csv
Main tables – our train and test samples
Target (binary)
Info about loan and loan applicant at application time
SK_ID_CURR
SK_ID_CURR
SK_ID_CURR
SK_ID_CURR
bureau.csv
Application data from previous loans that client got from other institutions and that were reported to Credit Bureau
One row per client's loan in Credit Bureau
previous_application.csv
Application data of client's previous loans in Home Credit
Info about the previous loan parameters and client info at time of previous application
One row per previous application
SK_ID_PREV
SK_ID_BUREAU
SK_ID_PREV
bureau_balance.csv
Monthly balance of credits in Credit Bureau
Behavioral data
POS_CASH_balance.csv
Monthly balance of client's previous loans in Home Credit
Behavioral data
instalments_payments.csv
Past payment data for each installments of previous credits in Home Credit related to loans in our sample
Behavioral data
credit_card_balance.csv
Monthly balance of client's previous credit card loans in Home Credit
Behavioral data

Figure 1: a diagram [28] detailing the connection between the datasets (with used datasets in blue)

We used four of the available datasets in our project:

- **Application train** and **Application test** provides data on 356,255 separate loan applications at Home Credit. Each loan is identified with an ID number and 122 features including variables such as age, occupation, educational level and a final outcome training label ('TARGET'), which indicates 0 if the loan was repaid or 1 if it was not. This dataset is split into a training set of 307,511 loans with a training label and a test set of 48,744 loans without.

- **Bureau** and **bureau balance** provides data on Home Credit loan applicant’s previous loans from other financial institutions. This totals 799,459 loans, where some applicants can have multiple previous loans. It includes detailed time series data regarding monthly payments on these other loans, totalling 7,299,925 payments, with each payment labelled according to the table in Figure 2.

4.1 Sample Generation

We generated samples from these datasets by engineering time series data into two binary sequences - an attention label sequence (‘Attention_Sequence’) and repayment sequence (‘Adherence_Sequence’) - and undertook data analysis to select static features.

4.2 Time Series Data - repayment sequence

To create our sequences, we used payment profile data from the bureau balance dataset, which provides monthly data about payments on previous loans with other lenders. The attention label sequence consists of integers representing whether payment was made or not, where ‘0’ represents successful payment that month while ‘1’, ‘2’ or ‘5’ represent a missed payment of varying severity (see Figure 2). In places where the payment status is unknown (‘X’), we substitute ‘0’. This decision introduces the least amount of influential data into our sequences. We then filtered out months where the account was closed (‘C’) and sequences where the borrower did not miss any repayments, as these aren’t useful in training the model. This generated 103,264 samples.

Label	Description
C	Account closed
X	Payment status unknown
0	Didn’t miss payment
1	Missed payment in last 1 and 30 days
2	Missed payment for 31 and 60 days
5	Missed payment for over 60 days or loan written off (i.e., proxy for default)

Figure 2 – status labels for payments

We then split the repayment sequence into fiscal quarters to discretize a continuous variable, so it could be processed by our model. We chose this segmentation because six months of missed payments is the industry standard for writing off the loan or defaulting according to the competition description.

4.3 Time Series Data - attention label sequence

We created an attention label sequence by approximating a credit risk value in {MEDIUM, HIGH} representing a borrower’s risk of non-payment for that quarter. If they pay or miss under 6 payments in the last 6 months, they are changed to ‘MEDIUM’ attention. If they reach 6 missed payments within the last 6 months, they are changed to ‘HIGH’ attention. This screening process generated 5,204 samples out of 103,264 with a ‘HIGH’ risk label.

Using this as a screening system, we can identify sequences of months during which a borrower is at risk of default and therefore in need of intervention. Interventions could include lenders restructuring loans if the borrower is unable to make payments and/or providing information on support services. We accomplish this with our formulation of the real-time risk prediction model in Section 5.

We separate each quarter according to the risk level assigned to it while also keeping track of the cumulative missed payments. Finally, we merged each set of quarters (‘HIGH’ and ‘LOW’) with a set of static features to be fed into our LSTM model (see 4.4).

4.4 Static features

In addition to time series data to train our LSTM model, we also selected the 29 static features in order to align with the methodology in the LEAP model. We did this by calculating the Pearson correlation coefficient [29] between variables in the application data and the final outcome training label (‘TARGET’).

The highest correlations included age, education levels, gender and three features which represent a "*normalized score from external data source*", according to the documentation. We have taken the last three to be credit ratings, based on discussions on Kaggle. We isolated these as part of our baseline model in order to compare how a real-time risk prediction model compares to current industry benchmarks.

5. Models

5.1 Real-time risk prediction

5.1.1 Motivation

First, we built a model for real-time risk prediction which leverages our training labels set out in 4.3 (‘MEDIUM’ or ‘HIGH’ risk labels). Our goal was to develop a model that predicts the likelihood of someone defaulting on the a loan based on their repayment history of that same loan in order to flag high risk borrowers for intervention before they miss critical repayments.

5.1.2 Baseline

We used a random forest model [30] (with 100 trees and a max depth of 5) inputting only static features, because it performed better than Linear Regression. We show the results of this model in comparison to BUST in Section 7.

5.1.3 Implementation

To leverage the time series in the bureau balance dataset we built a deep network called BUST (Bank Underwriting lStm predicTor), based on LEAP (Lstm rEal-time Adherence Predictor) from Tambe, Killian et al’s paper. [8]

BUST was implemented with Keras and takes both the time series and static features as input. It used the same hyperparameters as LEAP, building two input layers: 1) a LSTM with 64 hidden units for the time series input and 2) a dense layer with 100 units for the static feature input. We concatenated the outputs of these two layers to feed forward into another dense layer with 16 units, followed by a single sigmoid activation unit. We used a batch size of 128 and trained for 20 epochs.

We then compared the implementation of BUST on two separate datasets, the first containing only ‘MEDIUM’ risk level borrowers and second only ‘HIGH’ according to our methodology in 4.3 to see how it performed depending on the risk level of borrowers.

5.2 Outcome prediction

5.2.1 Motivation

We also built a model with a training label set to application train dataset to investigate how payment data on previous loans can be used to predict defaults on current ones. our goal was to understand how the first 6 months of a borrower's payment profile on a previous loan can enable more accurate, personalized outcome predictions on current ones.

5.2.1 Baseline and Implementation

We used the same baseline and BUST model as in 5.1, except we changed the training label from the 'MEDIUM' or 'HIGH' risk labels created in 4.3 to the final outcome ('TARGET') labels in the application train and test data, which indicates 0 if the loan was repaid or 1 if it was not.

We formalized the outcome prediction task as follows: given the last 3 months of repayments, predict the final binary default outcome. The static features and sequence inputs were the same as for the real-time risk prediction model and it was trained on the same set of 103,264 samples (5,204 positive).

Like for the real risk prediction model set out in 4.1, we also compared the implementation of BUST on two separate datasets, the first containing borrowers flagged as 'MEDIUM' risk and the second 'HIGH' level risk to understand how previous risk impacts current risk.

6. Evaluation

We used the Receiver Operating Characteristic Area Under the Curve (ROC AUC, also sometimes called AUROC) as the metric to measure our success, which graphs the true positive rate versus the false positive rate as well as the same evaluation criteria as was used in Tambe, Killian et al's paper [8], which we repeat below.

6.1 Real-time risk prediction

To evaluate models we randomized all data and used a 4-fold grid search to determine the best model parameters. To deal with class imbalance, we used SMOTE to over-sample the training set [31] implemented with the Python library imblearn. [32]

6.2 Outcome prediction

We used the same models, grid search design, training process, and evaluation procedure as for real-time risk prediction. For the Random Forest we used 150 trees and no max depth. For BUST, we used 64 hidden units for the LSTM input layer, 48 units for the dense layer input, and 4 units in the penultimate dense layer.

7. Results

Our results in Table 1 show a significant improvement of nearly 50% over the baseline model when using the real-time risk prediction BUST model. Using the outcome prediction model we got an extremely high number of true positives (4451), which we believe may be an anomaly.

Method	True Positive	# of False Negatives	# of True Negatives
baseline (Real-time, M)	214	85261	22683
BUST (Real-time, M)	447	2542	105288
Improvement	47.9%	33.5%	21.5%

Table 1: Numeric results of our testing

While further analysis is required to understand these results better, we were still able to obtain reliable ROC AUC for each of the models.

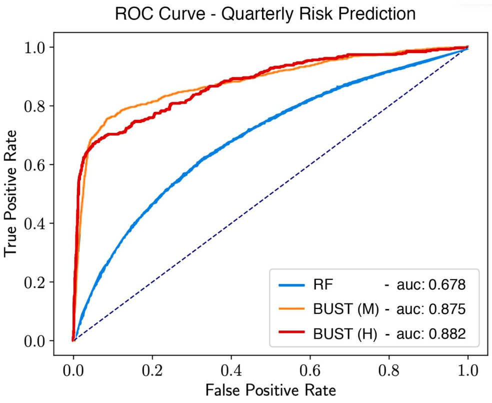

Figure 3: ROC Curve for the quarterly risk prediction task comparing the random baseline (blue), BUST MEDIUM (M) (orange) and BUST HIGH (H) (red). Numbers under the blue curve give thresholds used to calculate the baseline's ROC curve.

7.1 Real-time risk prediction

Figure 3 shows the ROC curve of our real-time risk prediction model compared to the baseline. It shows that our BUST model, based on previous payments, significantly outperforms the baseline, which uses traditional credit scores.

7.2 Outcome prediction

Figure 4 shows the ROC curve of our outcome prediction model compared to the baseline. Like the previous model, it also significantly outperforms the baseline, which uses traditional credit scores.

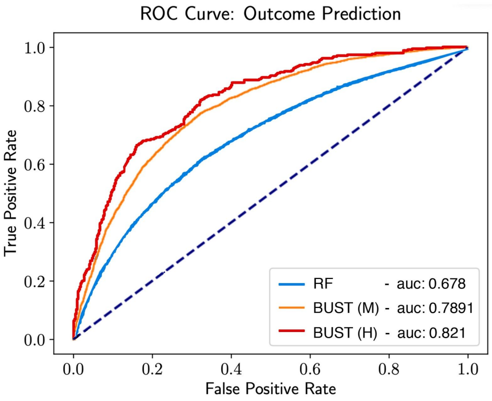

Figure 4: ROC curves for outcome prediction models

7.3 Comparison of models

By comparing the results, we can see that the real-time risk prediction model performed better on both the 'MEDIUM' and 'HIGH' risk datasets, suggesting our model LSTM model may be better at real-time predictions over longer term outcome predictions. We discuss the implications of these findings in Section 8.

8. Discussion

The results of our experiments show that an AI-based system trained on historical data can significantly outperform traditional means of prediction, such as credit scores. Our models produce ROC AUCs of up to 0.882, compared to a baseline of 0.678. In a real-time risk prediction setting, this could be used to proactively flag nearly 50% more borrowers at risk of non-payment. This is promising as it validates the thesis that more intelligence credit risk modelling can reduce discrimination in lending.

Comparing our results to the 7,000+ team entrants to the Kaggle competition we obtained the Home Credit dataset from, our results are even more promising. The winner of the competition obtained an ROC AUC of 0.806, training the application train and test datasets on the final outcome label ('TARGET'). [26] Our outcome prediction model obtained an ROC AUC of 0.821 training the bureau and bureau balance datasets on the same labels. Given the slightly different datasets, it is too early to draw conclusions. However, because we only used standard features and competition winners used engineered features, there is reason to believe that further research could demonstrate the efficacy of BUST as compared to alternative AI solutions in the market.

Because our real-time risk prediction model performed better than our outcome prediction model, this suggests an LSTM credit risk model may be better at shorter, over longer, term outcome prediction based on previous repayments. In the context of lending, this could be because default often arises due to a change in circumstances, such as job loss or relationship breakdown, [35] which are hard to predict on previous repayments alone. To this end, using additional, complementary datasets could complement this approach.

Two challenges we envisage in the implementation of such a model at scale is explainability and interpretability, which are well-documented challenges within AI, [11] and unintended consequences of models that flag high risk borrowers that could reinforce bias and discrimination further. [22]

Regarding the first, when a borrower is denied credit lenders are required by federal law to provide a reasonable explanation. [21] If we are seeking to create a more equitable model, interpretability is thus a key factor in our model's usefulness. Because BUST is a black-box network, rather than a natively interpretable model like linear regression, a visualization tool such as the SHapley Additive exPlanations (SHAP) python library could help users draw insights about our model's reasoning. [8]

Regarding the second, there are concerns that AI can actually exacerbate biases in the financial services industry if not properly accounted for in models. [17][18][19] For example, there is a risk that our models could negatively impact the creditworthiness of borrowers on low incomes and/or with protected characteristics. In order to avoid these problems, further analysis needs to be undertaken on the types of borrowers that the model flags as risky. Careful selection of unbiased training data is also key as we hope that our system, were it ever to be used in practice, could help to more efficiently allocate resources to those in need, rather than be used as a tool for further discrimination.

End Notes

1. Wasik, John (2011) ‘Why are credit scores such a mystery?’, Reuters.

2. Bhardwaj, Geetesh and Sengupta, Rajdeep (2015) ‘Credit Scoring and Loan Default’, Federal Reserve Bank of Kansas City Research Working Papers.

3. Diana Olick (Aug 19, 2020), ‘A troubling tale of a Black man trying to refinance his mortgage’, CNBC.

4. https://www.justice.gov/crt/equal-credit-opportunity-act-3

5. Bureau of Consumer Financial Protection (July 8, 2019) ‘Fair Lending Report of the Bureau of Consumer Financial Protection’, June 2019, 84 C.F.R. 32420.

6. Bartlett, R.; Morse, A.; Stanton, R.; Wallace, N. (Nov 2019). ‘Consumer-Lending Discrimination in the FinTech Era’, Berkeley.

7. Manyika, James; Silberg, Jake. (June 2019). ‘Tackling bias in artificial intelligence (and in humans)’, McKinsey Global Institute.

8. Jackson A. Killian, Bryan Wilder, Amit Sharma, Vinod Choudhary, Bistra Dilkina, and Milind Tambe (2019) ‘Learning to prescribe interventions for tuberculosis patients using digital adherence data’, in Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD ’19, pages 2430–2438, New York, NY, USA.

9. Home Credit has operations in nine countries: Czech Republic, Slovakia, Kazakhstan, Russian Federation, China, India, Indonesia, Philippines and Vietnam. Available at: https://www.homecredit.net/.

10. Brafman, O. & Brafman, R. (2008) ‘Sway: The Irresistible Pull of Irrational Behavior’, Broadway Books: New York, NY. Chapter 6: In France, the Sun Revolves around the Earth.

11. Johnson, Jonathan (July 2020) ‘Interpretability vs Explainability: The Black Box of Machine Learning’, BMC Blogs.

12. Gross, Terry (March 2017) ‘A 'Forgotten History' Of How The U.S. Government Segregated America’, NPR.

13. Badger, Emily (2017-08-24) ‘How Redlining's Racist Effects Lasted for Decades’, The New York Times.

14. Mazumder, Bhashkar; Hartley, Daniel A.; Aaronson, Daniel (2017). ‘The Effects of the 1930s HOLC "Redlining" Maps"’, FRB of Chicago Working Paper No. WP–2017–12.Sahay, R., M. Čihák, P. N’Diaye, A. Barajas, R. Bi, D. Ayala, Y. Gao, A. Kyobe, L. Nguyen, C. Saborowski, K. Svirydzenka, and S.R. Yousefi (2015). ‘Rethinking Financial Deepening: Stability and Growth in Emerging Markets. International Monetary Fund Staff Discussion Note’, SDN/15/08.

15. Koulouridi, Efstathia; Kumar, Sameer; Nario, Luis; Pepanides, Theo; and Vettori, Marco (July 2020) ‘Managing and monitoring credit risk after the COVID-19 pandemic’, McKinsey.

16. Solon Barocas & Andrew D. Selbst, (2016) ‘Big Data’s Disparate Impact’, 104 CALIF. L. REV. 67.

17. Matthew Adam Bruckner (2018) ‘The Promise and Perils of Algorithmic Lenders’ Use of Big Data, 93 CHI.-KENT L. REV. 3, 25–29.

18. Mikella Hurley & Julius Adebayo (2016) ‘Credit Scoring in the Era of Big Data’, 18 YALE J.L. & TECH. 148, 168.

19. Townson, Sian (November 2020) ‘AI Can Make Bank Loans More Fair’, HBR.

20. Puri, M.; Gombović, A.; Burg, V.; Berg, T. (2020). On the Rise of FinTechs: Credit Scoring Using Digital Footprints, The Review of Financial Studies, vol 33(7), pages 2845-2897.

21. Klein, A. (April 22 2019). ‘Credit denial in the age of AI’, Brookings Institute.

22. Escobar de Nogales, Ximena (April 2018) ‘Fintech for the Financially Excluded?’, SSIR.

23. Omdena (July 26 2020). ‘AI For Financial Inclusion: Credit Scoring for Banking the Unbankable’, Omdena. Accessed at: https://omdena.com/blog/credit-scoring-ai/.

24. Brighterion (November 2 2020). ‘Assess today’s credit risk and prevent tomorrow’s delinquency: a concise guide’, Brighterion, Inc. Accessible at: https://brighterion.com/wp-content/uploads/2020/10/ebook-ai-to-predict-credit-risk-and-prevent-credit-delinquency-a-concise-guide.pdf

25. Home Credit Group (2018) ‘Home Credit Default Risk, Kaggle’. Available at: https://www.kaggle.com/c/home-credit-default-risk/overview.

26. Breeden Joseph (June 2020) ‘Survey of Machine Learning in Credit Risk’, SSRN.

27. Koehrsen, Will (2018) ‘Start Here: A Gentle Introduction’, Kaggle. Available at: https://www.kaggle.com/willkoehrsen/start-here-a-gentle-introduction.

28. Pearson correlation coefficient. (n.d.). In Wikipedia from https://en.wikipedia.org/wiki/Pearson_correlation_coefficient.

29. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011) ‘Scikit-learn: Machine Learning in Python’, Journal of Machine Learning Research 12 (2011), 2825–2830.

30. Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer (2002) ‘SMOTE: synthetic minority over-sampling technique’, Journal of artificial intelligence research 16 (2002), 321–357.

31. Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas (2017) ‘Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning’m Journal of Machine.

32. Ceizyk, Denny (Sept 2020) ‘HMDA: What Is It and Why Is It Important?’, LendingTree. Available at: https://www.lendingtree.com/home/mortgage/hmda-what-is-it-and-why-is-it-important/.

33. White, Martha (July 2020) ‘Banks set aside billions of dollars in expectation of massive losses as consumers default on loans’, NBC. Accessed at: https://www.nbcnews.com/business/economy/banks-earnings-reflect-massive-hit-coronavirus-show-importance-regulations-n1233748.

34. Financial Conduct Authority, UK (2014) ‘Consumer credit and consumers in vulnerable circumstances’, FCA. Accessed at: https://www.fca.org.uk/publications/research/consumer-credit-and-consumers-vulnerable-circumstances.