CS 4701: Implementing a credit risk prediction and assessment model

Ayian Shariff (as2964)
Jaehyung Joo (jj534)
Nguyen Vo (ntv4)

December 14, 2023

1. Introduction

Accurate credit risk assessments allow financial institutions to lend to individuals who are likely to repay on time, which is critical for ensuring the institutions' profits and maintaining the wider stability of financial systems. Furthermore, for loan applicants who lack previous credit history and would normally be excluded from loan consideration, the ability to identify low-risk borrowers facilitates broader access to credit among traditionally underserved groups at risk levels that financial institutions could tolerate. The resulting expansion of traditional credit would also disincentivize lower income households from taking extortionate loans, which are not regulated uniformly in the US and can reach annual percentage rates up to 32% for $2000 loans without being considered as predatory by the median state.[1] These reasons motivate the development of models that can assess credit risk based on applicants' comprehensive demographic and financial data, which this paper investigates.

2. Literature Review

2.1 Handling Credit and Financial Data

There are several aspects of credit and financial data that merit special attention, namely its tendency to be unbalanced and its inclusion of dynamic variables. Furthermore, the potential role of credit risk models' interpretability in maintaining financial institutions' transparency should also be considered.

Unbalanced data:

Credit risk data are characterized by the fact that a large majority of customers of financial institutions repay their debts on time, skewing the data towards good behavior and thus potentially causing models to not learn enough from the minority of "bad" samples.

[1] Carter, "Predatory Installment Lending in the States: How Well Do the States Protect Consumers Against High-Cost Installment Loans? (2022)."

Marceau et al. (2020) compared the ability of multiple statistical learning models, which were Logistic Regression (LR), Support Vector Machines (SVM), Random Forests (RF), and Extreme Gradient Boosting (XGBoost), as well as deep learning (DL) models such as a hand-tuned Neural Network (NN) and the Auto-Keras Neural Network. to predict credit scores given an unbalanced dataset where most people are good payers.[2] They found that XGBoost outperformed all other classifiers and both deep learning models in accuracy, recall, and area under the curve (AUC), and also had lower runtimes than the deep learning models.[3]

Temporal effects:
Credit risk is dynamic and can change over time due to sudden shocks or structural shifts in the economy. As historical data may not adequately represent current or future risk landscapes, it is important to consider the dynamics of temporal aspects and potential shifts in economic conditions.

Chatzis et al. (2018)'s application of models to forecast stock market crisis events found that the deep learning feedforward network MXNET had the highest predictive ability among an assortment of statistical and DL methods that also included LR, Classification and Regression Tree (CART), RF, SVM, XGBoost, and NN.[4]

In addition, results from Teng, Lin, and Lu (2023) indicate that deep learning models such as Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) can potentially be used to improve the predictive ability of existing machine learning models. By extracting new features from temporally dynamic variables and combining them with static variables, CNN and RNN were able to improve the mutual information score of the data, thus lowering uncertainty.[5] The authors claimed that this induced slight improvement in the AUC of all examined ML models, which spanned LR, Decision Tree (DT), Light Gradient Boosting Machine (LightGBM), and NN.[6] Among these models, they also found that LightGBM performed the best in the training set while LR performed the best in the test set.[7]

---

[2] Louis Marceau et al., "A Comparison of Deep Learning Performances with Other Machine Learning Algorithms on Credit Scoring Unbalanced Data," 2020, 4-5.
[3] Ibid.
[4] Sotirios P. Chatzis et al., "Forecasting Stock Market Crisis Events Using Deep and Statistical Machine Learning Techniques," *Expert Systems with Applications* 112 (December 1, 2018): 353–71, https://doi.org/10.1016/j.eswa.2018.06.032, 367-70.
[5] Huei-Wen Teng, Lin, Lu, "Enhancing Credit Score Predictions with Dynamic Feature Engineering using Deep Learning," March 2, 2023, 16-8.
[6] Ibid, 19.
[7] Ibid, 21.

Interpretability:

In the financial industry, interpretability may contribute to the appearance of transparency to stakeholders and the public. In this sense, black-box models may struggle to provide transparent explanations for credit risk predictions, costing creditors credibility.

Angelov et al. (2021) asserted that among typical ML algorithms, DL possessed the lowest interpretability, followed by SVM, KNN, RF, Naive Bayes (NB), and DT in improving order.[8] They also mentioned several risk assessment situations as critical applications that have driven a recent rise in interest in the development of explainable AI (XAI).[9]

2.2 Existing Survey Papers

In addition to individual manuscripts, we also examined two recent large-scale literature review papers.

Shi et al. (2022) analyzed 76 studies relating to the application of ML models in credit risk assessment. They concluded that DL methods outperform statistical learning methods in credit risk estimation, and that ensembles of methods outperform single methods.[10] These findings broadly concur with most of the papers examined in our own readings of recent works.

Noriega, Rivera, Herrara (2023) reviewed 52 ML studies relating to the microfinance credit industry to investigate patterns of model, dataset, evaluation criteria usage and practices used to improve predictive ability. They found that models that feature boosting are used most frequently, with the XGBoost model receiving the most attention.[11] Additionally, they observed that AUC is the most frequently used evaluation metric, followed by accuracy, F1, precision, recall, and others.[12] K-Fold Cross Validation was the most commonly used hyperparameter tuning method, while SMOTE was the most frequently used dataset balancing technique.[13]

---

[8] Plamen P. Angelov et al., "Explainable Artificial Intelligence: An Analytical Review," *WIREs Data Mining and Knowledge Discovery* 11, no. 5 (September 1, 2021): e1424, https://doi.org/10.1002/widm.1424, 3.
[9] Ibid, 8.
[10] Si Shi, Rita Tse, Wuman Luo, Stefano D'Addona, and Giovanni Pau, "Machine Learning-Driven Credit Risk: A Systemic Review," *Neural Computing and Applications* 34, no. 17 (September 1, 2022): 14327–39, https://doi.org/10.1007/s00521-022-07472-2, 14333.
[11] Jomark P. Noriega, Luis A. Rivera, and José A. Herrera, "Machine Learning for Credit Risk Prediction: A Systematic Literature Review," *Data* 8, no. 11 (2023), https://doi.org/10.3390/data8110169, 8.
[12] Ibid.
[13] Ibid, 12.

3. Problem Definition

Data Provenance:

Data used in this project are taken from the Home Credit Group's 2018 "Home Credit Default Risk" Kaggle competition.[14] As using the data in an educational setting does not comply with Home Credit's specific instructions on data usage, which limits usage to the competition only,[15] we recognize that Home Credit reserves the right to disqualify our team from the contest[16] as well as the $70000 prize that was given in 2018. Of the data files provided in the Data section of the competition portal, only *application_train.csv* is used. *application_train.csv* features 163704 rows and 122 columns, with one column being the TARGET column. Provided below is a list of input and output variables, where the inputs are the 12 variables chosen after data processing and the output is the TARGET variable.

Inputs:

NAME_CONTRACT_TYPE , denoting whether the loan is a cash loan or a revolving loan

CODE_GENDER , denoting the gender of the client

FLAG_OWN_CAR , denoting whether the client owns a car

CNT_CHILDREN , denoting the number of children a client has

AMT_INCOME_TOTAL , denoting the income of the client

AMT_CREDIT , denoting the credit amount of the loan

NAME_EDUCATION_TYPE , denoting the highest level of education that the client achieved

NAME_FAMILY_STATUS , denoting the family status of the client (married or not)

NAME_HOUSING_TYPE , denoting the housing situation of the client (whether they're renting, living with parents, etc)

DAYS_EMPLOYED , denoting how many days before the loan application the client started employment

DAYS_BIRTH , denoting the client's age in days at the time of application

CNT_FAM_MEMBERS , denoting how many family members a client has

Outputs:

TARGET , denoting the target variable. 1 denotes that the client had difficulty paying back the loan on time, 0 denotes that the loan was repaid.

---

[14] Anna Montoya, inversion, KirillOdintsov, Martin Kotek, "Home Credit Default Risk," accessed December 14, 2023, https://kaggle.com/competitions/home-credit-default-risk.

[15] Anna Montoya, inversion, KirillOdintsov, Martin Kotek, "Home Credit Default Risk," accessed December 14, 2023, https://kaggle.com/competitions/home-credit-default-risk/rules, section A.

[16] Anna Montoya, inversion, KirillOdintsov, Martin Kotek, "Home Credit Default Risk," accessed December 14, 2023, https://kaggle.com/competitions/home-credit-default-risk/rules, section B, subsection 11.

4. Data Exploration

We compared the correlation of the predictor columns to the target column, whether the client had difficulty paying the loan back on time. The columns that were most correlated to the target were name_contract_type (denoting whether the loan is a cash loan or a revolving loan) and cnt_children (the number of children a client has). We then also plotted a heatmap of the correlations between every column to get a sense of how the variables in the dataset are correlated.

We also graphed the distribution of how many years a client was employed. The distribution was right skewed, and most clients who borrowed the loan were employed for 1-2 years.

We also graphed the education level of the clients in the dataset. Most people in the dataset had attended up to secondary school or finished their higher education. Furthermore, we graphed the age distribution of the clients in the dataset. A significant majority of the clients were between the ages of 20 to 50.

We then plotted a KDE plot showing how the number of years of employment affects loan repayment. We found that, counter-intuitively, the less time a person was employed for, the more likely they were to repay their loans on time. We also plotted a KDE plot for the education level of the clients in the dataset, getting similar results to the previous education level graph. Finally, we graphed a KDE plot showing how the number of years of employment affects loan repayment. The loan repayment graph is skewed toward younger ages, which means that younger people are more likely to repay their loans on time.

5. Training, Testing, and Model Evaluation

We split our dataset into 80% training data and 20% testing data, as well as the target (predicted) and the predictor variables for credit risk. So in total we had 4 data sets, X_train (training dataset for the non predictor variables), X_test (the testing dataset for the non predictor variables), Y_train (training dataset for the predicted variable), and Y_test (testing dataset for the predicted variable). We then made sure that the training data and testing data is as expected by checking the sizes of each of X_train, Y_train, X_test, and Y_test and making sure of the following:

The number of rows in X_train and Y_train are matching
The number of rows in X_test and Y_test are matching

The number of columns in X_test and X_train are matching
The number of columns in Y_train and Y_test are matching

We then started to fit models to determine which is the best at predicting credit risk. After reading about which models would be good on our high dimensional data, we narrowed our model choices down to the following 5 models:

Logistic Regression
XG Boost
Random Forest
Gaussian Naive Bayes
K-Nearest Neighbors

For each of the models we fit, we evaluated how good the model was by calculating the raw accuracy, F1 score, recall, and confusion matrix. We thought that calculating the F1 score would be useful because in our dataset, we have uneven class distribution. Many more target results are '0' than '1' - we have significantly less targets where someone had trouble paying back their loan compared to not. We thought that calculating the recall score would be helpful because we can see how sensitive the model is. Also, since our dataset is imbalanced, we wanted to make sure that the accuracy metric was not misleading. The recall metric is sensitive to the minority class and provides us a better understanding of how the model performed for that class. Finally, we decided to calculate the confusion matrix to see the overall distribution of the model. The confusion matrix provided insights into how many true positives, true negatives, false positives, and false positives were predicted.

We first decided to fit a logistic regression model as our baseline model, making sure that the class weight parameter is set to "balanced" because we wanted the model to not be biased toward a particular class. Logistic regression is good for binary classification problems, and in this case, the outcome we are predicting is binary - whether the individual defaulted or did not. After we fit the logistic regression model and calculated various statistics, we obtained the following:

Accuracy : 0.5453738910012674
F1 Score : 0.18372145945535917
Recall : 0.5557595227168426
Confusion Matrix : [[11698 9793]

[968 1211]]

This confusion matrix can be interpreted as follows:
11698 true negatives
9793 false positives
968 false negatives
1211 true positives

The logistic regression model was not very accurate in predicting whether one had trouble paying back their loan, so we decided to fit other models on our dataset to see whether we could achieve higher accuracy.

We then fit an XGBoost model for the reasons described in the next few sentences. The XGBoost model is an ensemble method, and according to Shi et al. (2022), ensemble methods tend to outperform other statistical models when used to predict credit risk data.[17] Hence, we believed that it would be useful in helping predict whether one had trouble paying back their credit risk loan or not. We tuned the hyperparameters n_estimators, max_depth, learning_rate, scale_pos_weight. For the best model, n_estimator is 2, max_depth is 10, learning_rate is 0.5, scale_pos_weight 3. We used a small number of weak learners. A max_depth of 10 means that more complicated relationships in the data are captured, which is especially important because this dataset has tons of rows and multiple columns. We gave a little bit more weight to correctly predicting the positive class because it is more important to predict the positive cases rather than the negative. This is because it is more important to capture cases that accurately classify someone having trouble paying back their loans. We set the learning rate to moderate because we wanted a trade off between speed and accuracy. After we fit the XGBoost model and calculated various statistics, we obtained the following:

Accuracy : 0.9034220532319391
F1 Score : 0.014655172413793103
Recall : 0.007801743919229004
Confusion Matrix : [[21367 124]
[2162 17]]

[17] Si Shi, Rita Tse, Wuman Luo, Stefano D'Addona, and Giovanni Pau, "Machine Learning-Driven Credit Risk: A Systemic Review," *Neural Computing and Applications* 34, no. 17 (September 1, 2022): 14327–39, https://doi.org/10.1007/s00521-022-07472-2, 14333.

This confusion matrix can be interpreted as follows:
21367 true negatives
124 false positives
2162 false negatives
17 true positives

We then fit a Random Forest model for the reasons described in the next few sentences. The random forest model is an ensemble method, and ensemble methods are good at predicting credit risk data, according to Shi et al. 2022.[18] Additionally, random forest takes into account feature importances, indicating which features are the most influential in identifying the key factors that contribute to an individual's creditworthiness. Also, the random forest method can handle high dimensional data really well and random forest has mechanisms, such as adjusting class weights, to counteract the fact that in our dataset, the number of loans in which the client had difficulty paying it back is significantly less than the number of loans in which the client did not have difficulty paying it back. After we fit the Random Forest model and calculated various statistics, we obtained the following:

Accuracy: 0.9072665821715251
F1 Score : 0.0009103322712790169
Recall : 0.0004589261128958238
Confusion Matrix : [[21474 17]
[2178 1]]

This confusion matrix can be interpreted as follows:
21474 true negatives
17 false positives
2178 false negatives
1 true positives

We then fit a Gaussian Naive Bayes model for the reasons described in the next few sentences. Gaussian Naive Bayes is not as computationally expensive as the other methods and is less prone to overfitting. Additionally, the Gaussian Naive Bayes model performs well on imbalanced datasets like ours. After we fit the Gaussian Naive Bayes model and calculated various statistics, we obtained the following:

[18] Ibid.

Accuracy: 0.90794254330376
F1 Score: 0.0
Recall: 0.0
Confusion Matrix: [[21491 0]
[2179 0]]

This confusion matrix can be interpreted as follows:
21491 true negatives
0 false positives
2178 false negatives
0 true positives

It is worth noting that Gaussian Naive Bayes method did not yield any positive predictions, assigning '0' as the predicted class for every row, hence the F1 score and recall came out to be 0. This method is misleading and not a good fit for our data, despite what we thought originally.

We then fit a KNN model, setting the n_neighbors parameter equal to 4. Our model worked best when the 4 nearest neighbors were considered when making predictions for a new data point, as it balanced both model complexity and robustness to noise in the data. We believed that KNN would be a good fit for our data for the reasons described in the next few sentences. KNN is a non-parametric algorithm and is hence suitable for capturing non-linear relationships in our dataset. Additionally, KNN performs well on imbalance datasets like ours. Furthermore, KNN is a flexible model as the number of neighbors considered can be easily tuned. After we fit the KNN model and calculated various statistics, we obtained the following:

Accuracy: 0.9040135192226447
F1 Score : 0.007860262008733625
Recall : 0.004130335016062414
Confusion Matrix: [[21389 102]
[2170 9]]

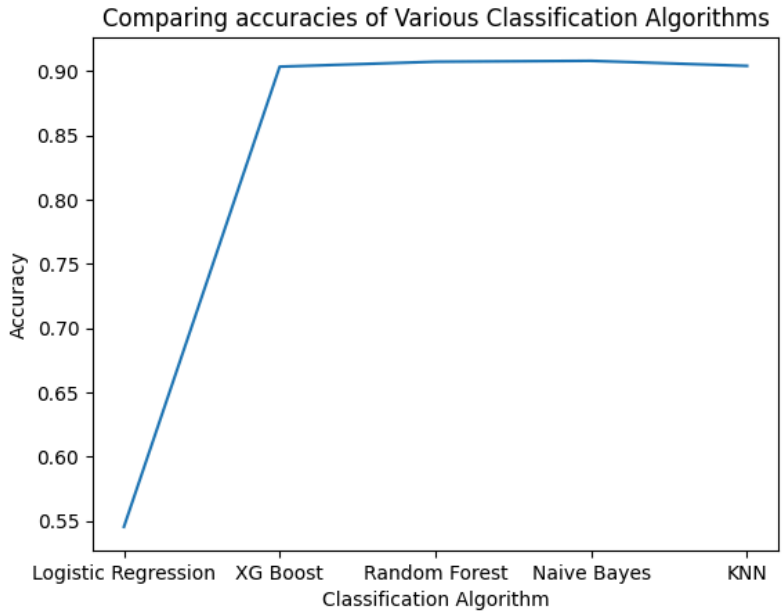
This confusion matrix can be interpreted as follows:
21389 true negatives
102 false positives

2170 false negatives

9 true positives

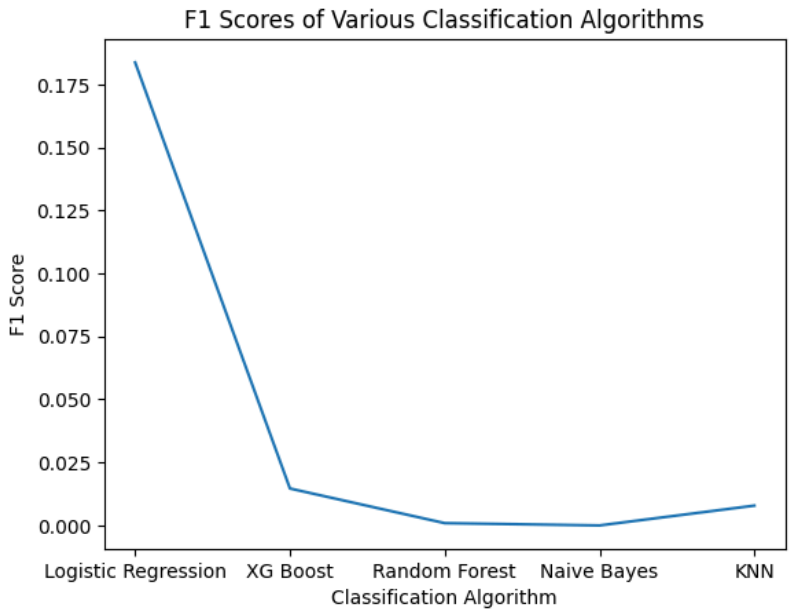
We then graphed the models' accuracy scores, F1 scores, and recall scores, comparing them against one another to determine which model performs the best.

Below is a graph of the various accuracy scores of the 5 classification algorithms



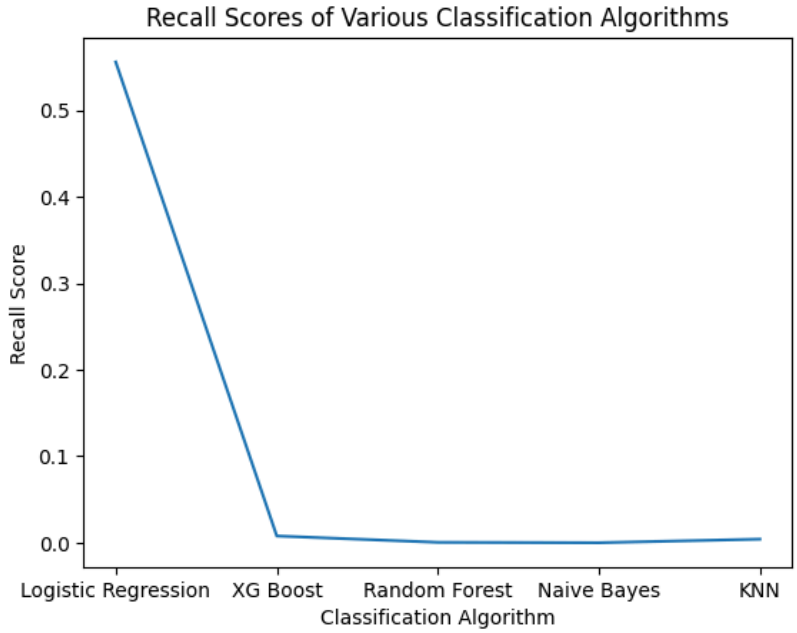
As can be observed, logistic regression is the least accurate classification algorithm out of the 5. XGBoost, Random Forest, Naive Bayes, and KNN are about equally accurate in predicting whether the client had difficulty paying back the loan on time.

Below is a graph of the various F1 scores of the 5 classification algorithms



As can be observed, logistic regression has the highest F1 score out of the 5 classification algorithms, and XGBoost has the second highest F1 score out of the 5 classification algorithms. Random Forest, Naive Bayes, and KNN have roughly the same F1 scores.

Below is a graph of the various recall scores of the 5 classification algorithms



Logistic regression has the highest recall score out of the 5 classification algorithms. XGBoost, Random Forest, Naive Bayes, and KNN have roughly the same recall scores.

We chose the XGBoost model as the best one for our dataset because it has a very high accuracy and the second highest F1 score out of the classification algorithms. While logistic regression has a high recall and F1 score, its accuracy is significantly lower than that of any of the other classification algorithms, hence we did not choose it as the best model. We did not choose Random Forest, Naive Bayes, or KNN as the best models because, while their accuracy and recall scores are similar to that of XGBoost, their F1 scores are visibly lower than that of XGBoost.

We then extracted features for the XGBoost model, and found out that the 4 most important predictors of credit risk (in order) are the NAME_EDUCATION_TYPE, CODE_GENDER, FLAG_OWN_CAR, and DAYS_BIRTH variables. The higher the level of education that one achieved, the more likely they are to repay their loan on time, and that is the most important predictor. The gender of the client was the second most important predictor of their ability to repay their loan on time. Determining whether the client owns a car was the third most important predictor of their ability to repay their loan on time. Finally, the client's age at the time of application was the fourth most important predictor in determining their ability to repay their loan on time.

6. Conclusion

Corroborating existing literature, we found that XGBoost performs the best on credit risk data. The accuracy of the XGBoost model stands out, particularly in its ability to effectively identify positive samples. The primary indicators, in descending order, include the applicant's attained education level, gender, car ownership status, and age.

References

Angelov, Plamen P., Eduardo A. Soares, Richard Jiang, Nicholas I. Arnold, and Peter M. Atkinson. "Explainable Artificial Intelligence: An Analytical Review." *WIREs Data Mining and Knowledge Discovery* 11, no. 5 (September 1, 2021): e1424. https://doi.org/10.1002/widm.1424.

Anna Montoya, inversion, KirillOdintsov, Martin Kotek. (2018). Home Credit Default Risk. Kaggle. https://kaggle.com/competitions/home-credit-default-risk

Carter, Carolyn. "Predatory Installment Lending in the States: How Well Do the States Protect Consumers Against High-Cost Installment Loans? (2022)." NCLC. Accessed December 13, 2023. https://www.nclc.org/resources/predatory-installment-lending-in-the-states-2022/.

Chatzis, Sotirios P., Vassilis Siakoulis, Anastasios Petropoulos, Evangelos Stavroulakis, and Nikos Vlachogiannakis. "Forecasting Stock Market Crisis Events Using Deep and Statistical Machine Learning Techniques." *Expert Systems with Applications* 112 (December 1, 2018): 353–71. https://doi.org/10.1016/j.eswa.2018.06.032.

Marceau, Louis, Lingling Qiu, Nick Vandewiele, and Eric Charton. "A Comparison of Deep Learning Performances with Other Machine Learning Algorithms on Credit Scoring Unbalanced Data," 2020.

Noriega, Jomark P., Luis A. Rivera, and José A. Herrera. "Machine Learning for Credit Risk Prediction: A Systematic Literature Review." *Data* 8, no. 11 (2023). https://doi.org/10.3390/data8110169.

Shi, Si, Rita Tse, Wuman Luo, Stefano D'Addona, and Giovanni Pau. "Machine Learning-Driven Credit Risk: A Systemic Review." *Neural Computing and Applications* 34, no. 17 (September 1, 2022): 14327–39. https://doi.org/10.1007/s00521-022-07472-2.

Teng, Huei-Wen and Lin, Jui-Yu and Lu, Kung-Wei, Enhancing Credit Score Predictions with Dynamic Feature Engineering using Deep Learning (March 2, 2023). Available at SSRN: https://ssrn.com/abstract=4375313 or http://dx.doi.org/10.2139/ssrn.4375313