XAI Applications in Automobile Financing Default Modeling

Kevin Sweet†
Data Science Graduate
University of Washington
Seattle WA USA
ksweet1@uw.edu

ABSTRACT

Consumer loan underwriting requires a high level of transparency in the output of predictive models. This constraint has limited the techniques possible for predicting the creditworthiness of an applicant. This work seeks to apply an explainable AI framework to a black box model that predicts occurrence of a debtor defaulting on their auto loan. This framework and model will be compared with the explainability and performance of a baseline interpretable model.

KEYWORDS

Consumer Finance, XAI, Credit Risk Modeling

1. INTRODUCTION

Financial institutions have a strong incentive to accurately predict whether a borrower will pay back their loan or not. This process is known as credit risk modeling. Traditionally, the process of assessing risk and reaching a credit decision has been largely manual, typically done through trained professionals in loan underwriting. More recently, this process has become increasingly automated through the use of artificial intelligence and increased access to data. There are, however, regulatory requirements in place that limit the capabilities of artificial intelligence. In the U.S. in particular, there is the Fair Credit Reporting Act (FCRA) of 1970 [1] that requires that banks provide adverse action notices to consumers that are denied credit. These action notices, among other requirements, must include a specific reason as to why their credit application was denied. This comes into conflict with the black-box nature of many powerful machine learning methods such as artificial neural networks which provide no explanation for the reason why they might predict a consumer will default.

Explainable artificial intelligence (XAI) frameworks address this issue by providing explanations for *why* the model makes a specific prediction and/or how it makes predictions across the entire problem domain. This work explores two different models for predicting default in consumer automobile lending and applies an XAI framework to assess the viability of using a black-box model in compliance with FCRA. In particular, this study hopes to address two key questions:

Q1. Does the selected black-box model outperform logistic regression in predicting default in auto loans? If so, is that performance increase worth the loss in interpretability?

Q2. Does the selected XAI framework provide suitable explanations to allow black-box models to comply with regulatory requirements?

2. BACKGROUND & RELATED WORK

Predictive models, in particular logistic regression, have been used to model credit risk in the U.S. since the 1980s. [2] The models used in practice for underwriting have been limited to only models inherently interpretable such as logistic regression or decision trees. Since the advent of explainable AI methods, there have been attempts to apply black-box methods with multiple explainable AI frameworks to consumer loans in Malta [3] as well as on the Lending Club dataset [4], but no known efforts have been made in the U.S. in practice. Researchers at the University of Malta found that all three of the XAI frameworks tested produced valid explanations when assessed by experts in finance and other workers in the finance industry.

3. METHOD & APPROACH

Loan-level data is not normally publicly available. The exception to this is when loans are bundled into securities and sold. This financial product is known as an asset-backed security. Asset-backed securities have public filing requirements from the U.S. Securities and Exchange Commission (SEC). Data was collected from these SEC filings and preprocessed for analysis. This process involved scripting a program that downloaded all of the relevant files which were then filtered to only automobiles and merged into a single dataset for analysis. This singular dataset was further cleaned and transform until it consisted of only numeric and indicator variables that are suitable for predictive models. The dataset was heavily imbalanced, so the majority class (loans that did not default) was down sampled to achieve equal representation of the classes.

Two models were selected for evaluation: logistic regression and an artificial neural network (ANN). These two models were chosen to more easily compare the change in performance from using a black box model (ANN) against a more interpretable model (logistic regression). For the neural network, a hyperparameter

search was performed using grid search to optimize the model architecture. This was compared against a L2 regularized logistic regression model.

Of the state-of-the-art XAI frameworks, the most applicable to this use case was Anchors. [4] Anchors provides feature-based local explanations that are model-agnostic. This means either method of modeling can be explained at the per-loan level, which allows for a detailed adverse action notice in the event of a predicted credit default. These explanations take the form of a set of one or more conditional statements that result in a probability of default that exceeds a pre-defined threshold. In this case, the threshold is set to 95%. Of the possible conditions that meet this requirement, anchors will select the condition that has the fewest number of conditions as well as the largest *coverage.* Coverage is defined as the probability that the same condition applies to other samples from the data, which allows a human to better understand how well the sufficient conditions can be generalized to other data points.

In the case of logistic regression, the anchors explanations are compared with the log-odds of the coefficients to determine if the two explanations are in alignment. If there is a difference between these two explanations, it is indicative that there may be an issue with one of the interpretations. This comparison was made through visual inspection of explanations.

4. EXPERIMENT DETAILS & RESULTS

4.1 LOGISTIC REGRESSION

The L2 regularized logistic regression model predicted default accurately in 83.9% of cases. It maintained relatively equal false positive and false negative rates with a precision of 87% and a recall of 84%. Figure 1 details the confusion matrix for the predictions made on the test set and Figure 2 shows the receiver operator characteristic (ROC) curve. The total area under the ROC curve was 0.84.

Figure 1: Logistic Regression Confusion Matrix

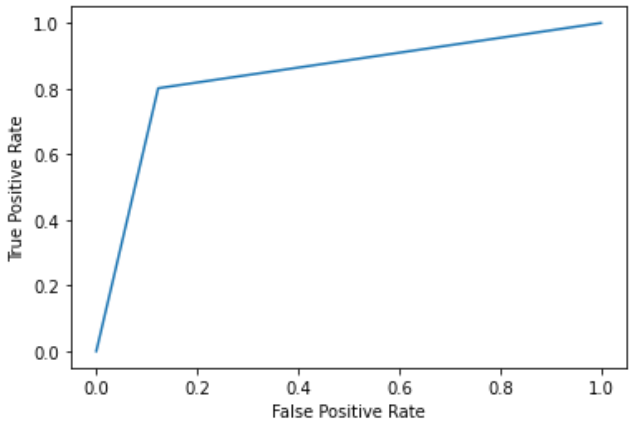

Figure 2: Logistic Regression ROC Curve

A vast majority of the explanations given by anchors for the logistic regression predictions include interest rate as a primary factor for predicting probability of default. Figure 3 highlights two explanations, it shows that high interest rates, low credit scores, and a lack of income verification explains the prediction of defaulted with 99% precision. For the second explanation, the very low interest rate alone is enough to provide a sufficient condition.

Logistic Regression Prediction: Defaulted

Anchor: interestRate > 0.51 AND creditScore <= 0.07 AND incomeVerifiedIndicator = -1

Precision: 0.99

Coverage: 0.35

Logistic Regression Prediction: Not Defaulted

Anchor: interestRate <= 0.18

Precision: 1.00

Coverage: 0.25

Figure 3: Two Anchor Explanations Logistic Regression

It should be noted that all of these values are normalized using min-max normalization, so “interestRate > 0.51” should be interpreted as the min max normalized interest rate is greater than 51% between the minimum and maximum observed values.

Comparing the majority of interest rate explanations to the logistic regression coefficient for interest rate found two opposite explanations. The coefficient for interest rate found by logistic regression was -5.83, which implies an inverse relationship between interest rate and default likelihood. This is counter-

intuitive to a conceptual understanding of high interest rates reflecting high risk, and thus high likelihood of default.

4.2 ARTIFICIAL NEURAL NETWORK

The selected architecture for the neural network was a multi-layer perceptron with 2 hidden layers containing 75 and 50 nodes respectively. The first hidden layer's activation function was sigmoid, and the second was the rectified linear unit (ReLU). The model predicted default accurately in 80.6% of cases while preferring to minimize false positives over false negative rates with a precision of 94% and a recall of 66%. Figure 4 details the confusion matrix for the predictions made on the test set and Figure 5 shows the receiver operator characteristic (ROC) curve. The total area under the ROC curve was 0.80.

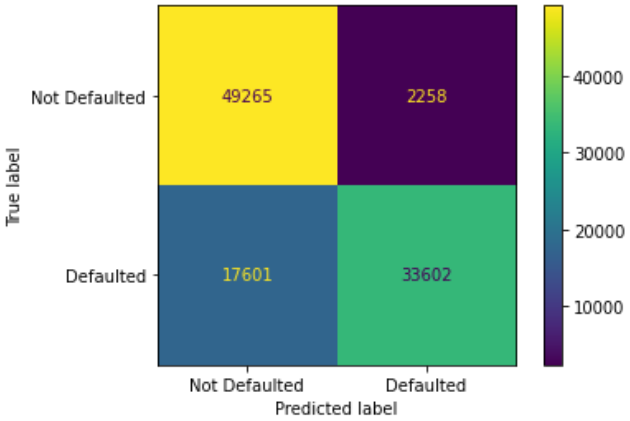

Figure 4: Multi-layer Perceptron Confusion Matrix

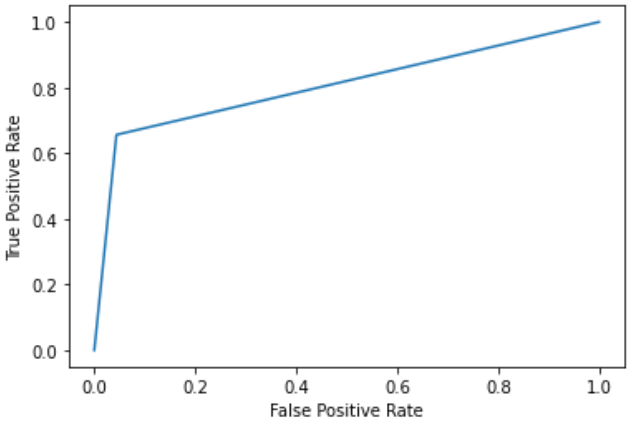

Figure 5: Multi-layer Perceptron ROC Curve

Like the logistic regression model, the vast majority of anchors explanations for the neural network were explained by one variable: interest rate. Figure 6 you can see the explanation for the prediction of the same two data points as the logistic regression. These two data points use very similar anchors to explain the reason for its prediction, with the neural network using rateSubvention in favor of credit score and income verification.

Neural Network Prediction: Defaulted

Anchor: interestRate > 0.51 AND rateSubvention = No

Precision: 0.97

Coverage: 0.49

Neural Network Prediction: Not Defaulted

Anchor: interestRate <= 0.18

Precision: 0.95

Coverage: 0.24

Figure 6: Two Anchor Explanations ANN

5. DISCUSSION

Comparing logistic regression to our chosen neural network architecture shows that logistic regression outperformed the neural network. The overall accuracy was higher, and the model was more balanced in its false positive and false negative rates. Depending upon how this model is applied, the higher precision shown by the neural network could make it a more suitable candidate, but overall, it does not outperform to the logistic regression model.

In terms of the explanations provided by anchors, both models were closely aligned. Both heavily relied on interest rate for explaining the prediction outcome. However, when validating the explanations against the log-odds of the logistic regression coefficient for interest rate, we see the inverse relationship. This is interesting because traditionally higher interest rates are offered to candidates with a perceived higher risk, but the model has found that candidates with higher interest rates have a lower risk of default. While not in the scope of this work, the next step would be to determine whether this effect is statistically significant. This can be done through hypothesis testing such as a one-sample Z test.

Assuming interest rate is a valid differentiator between the classes, these explanations could be sufficient for regulators. In the event that an individual is denied a credit application, an adverse action notice can be sent informing the consumer that due to their credit profile, the applicant is unable to qualify for an interest rate that places the risk of default in an acceptable threshold. Future work

can be done to integrate a model like this into a full credit risk model that incorporates the expected loss given default of the loan.

6. LIMITATIONS & CONCLUSION

There are many areas of this problem that can be further explored that are outside of the scope of this study. In particular, the models can be improved by exploring different model types, performing a feature selection process to remove highly collinear features, search for more complex neural network architectures, and many more. To improve the XAI framework's explainability, future works can create a process for inversing the normalization of the input features to return explanations to their true values.

This work set out to address two key questions. For Q1, it was discovered that for the architectures searched, the logistic regression model performed better on both model accuracy and explainability. Q2 found that there is potential for the Anchors framework to be used to provide adverse action notices to consumers from decisions reached using black-box models. To assess the viability, a continuation of this study in partnership with legal counsel would be required.

REFERENCES

[1] Fair Credit Reporting Act, 15 U.S.C § 1681 (1970). https://www.consumer.ftc.gov/articles/pdf-0111-fair-credit-reporting-act.pdf

[2] Wiginton, J. C. (1980). A Note on the Comparison of Logit and Discriminant Models of Consumer Credit Behavior. *The Journal of Financial and Quantitative Analysis*, *15*(3), 757–770. https://doi.org/10.2307/2330408

[3] Demajo, L.M., Vella, V., & Dingli, A. (2020). Explainable AI for Interpretable Credit Scoring. ArXiv, abs/2012.03749.Why is this important?

[4] Ribeiro. (2018). Model-Agnostic Explanations and Evaluation of Machine Learning. ProQuest Dissertations Publishing.