Literature Review

1 Part I

Credit risk is also referred to as default risk. In financial area, debt, loan and some financial derivatives all come with credit risk. The review mainly focus on the credit risk comes with loan. Failure of loan repayment will hurt the borrowers (mostly referred to financial institutes). Being aware of this, most financial institutes perform strict background investigation before loaning money to market participators. However, fairness of predicting risk is an unavoidable issue. Interpretability and trade-offs between accuracy and fairness are two main challenges.

Some traditional machine learning techniques have shown good performance in predicting credit risk like SVM, Random Forest, KNN. These are also popular methods applied in industry [1]. However, these techniques are often seen as black box models. To achieve high accuracy, institutes may increase the number of layers in neural network and they may not care much about how these models interpret their reasons for loan denial. A black box model only wants a True or False result. Without fully understanding of why a loan request is denied will lead to unconscious discrimination. For example, the SVM model applied in predicting the credit risk is just a process of minimizing the loss function, during which a borrower might have been denied simply due to some "mysterious" information comes with his or her family name or age. The challenge is you will never know the denial reasons. Financial institutes may not make credit scoring unfair consciously, but unfairness just happens. Thinking of this, an ideal machine learning method tells you why the borrower is denied and you, as a financial institute officer, will decide whether the reason is fair.

The other challenge in predicting credit risk using machine learning is that bad feature selection will lead to unfairness even if the machine learning technique is interpretable. For example, gender is a factor that is considered as one of the discriminatory features in predicting credit risk. However, women are largely discriminated as borrower [2] .

The Equal Credit Opportunity Act of 1974 (ECOA) [3] lists some factors that cannot be used to make decision in terms of whether a borrower is eligible for the loan. These factors include race, sex, national origin, age—as well as less common factors, like if the individual receives public assistance [4].

Aside from that, Congress has passed The Fair Credit Reporting Act (FCRA) [5], which aims to maintain fairness in credit reporting and ensure consumer's privacy. However, it is difficult for machine learning techniques to strictly obey these laws and regulations. One of the biggest challenges here is that even if the AI engineer who trains the model cautiously chooses features and performs good Exploratory Data Analysis (EDA), it is unavoidable that the machine techniques will learn to know about these discriminative features.

So, even if conventional machine learning techniques have already shown great performance in improving the accuracy in predicting credit risk, they should be carefully

used.

2 Part II

With machine learning applied more in financial area, some conventional machine learning approaches, like (Support Vector Machine) SVM, Random Forest or (K-Nearest Neighbour) KNN, have already been found shown good performance in measuring credit risk. However, these traditional techniques may sacrifice fairness. The features considered and the interpretability are two main challenges now. As for interpretability, there is a trade-off between accuracy and interpretability. To achieve high accuracy, interpretability is often sacrificed. Higher accuracy also requires some discriminative features to be included. The novel techniques discussed in this review mainly target on these two points.

2.1 Target on Interpretability

The improvement of interpretability with the application of GANs and two-layer additive risk model with the application of neural network are two typical novel techniques two solve the problem of unfairness.

Some scholars promoted a User-friendly technique using Generative Adversarial Network (GANs) [6] to improve the interpretability. The big picture of GANs' application in predicting credit risk is to introduce humans' words to generate interpretations. Th authors designed GAN model to generate explanations for loan denial using limited training data (approximately thousands of sentences).

The paper firstly incorporate broad and specific reasons for loan denials with consideration of two-level reasoning for each sentence. This hierarchical manner considers two-level reasons for each sentence and then combine them into a single vector using neural network. The hierarchy way of reasoning takes humans' tendency to make over-hypothesis into consideration. Human tend to make further hypothesis with incomplete information [7]. After the first step, the author applied loss function proposed by Kocaoglu et al. [8] to estimate possible reasons for loan denial. The authors consider two classifiers and they learn simultaneously. One is for estimation of real sentences (labeler), while the other is for estimation of generated sentences (anti-labeler). What should be noticed is that during the process, the authors also apply Gaussian mixture model (GMM) to generate noise input. The workflow may refer to the Figure 1 in Appendix.

The GAN model aiming at design a user-friendly way to represent the explanations for why the loan is denied. As mentioned before, one of the most important factors that leads to unfairness in measuring credit risk is the lack of interpretability. In this paper, interpretability has been improved accordingly, thus the fairness has been improved.

Some other scholars form a two-layer additive risk model [9] which is globally interpretable rather than a black box. The model in the paper applies the skeleton of neural

network. Neural network is often considered as a black box with low interpretability, though. To improve the fairness to measuring credit risk while still maintaining certain level accuracy, the authors cautiously choose features which are all interpretable. The authors do not focus much on the part of feature selection (which will be discussed later). They simply use the data promoted by Fair Isaac Corporation (FICO). The features chosen are interpretable and explainable. Then, they decide their monotone step functions as initial transformation of the features:

$$f_p(x_{\cdot,p}) = \beta_{p,1} b_{p,1}(x_{\cdot,p}) + \beta_{p,2} b_{p,2}(x_{\cdot,p}) + \beta_{p,3} b_{p,3}(x_{\cdot,p}) + \beta_{p,0} b_{p,0}(x_{\cdot,p}) \tag{1}$$

where $b_{p,1} = 1[x_{\cdot,p} < 10]$, $b_{p,2} = 1[x_{\cdot,p} < 50]$, $b_{p,3} + 1[x_{\cdot,p} < 75]$, $b_{p,1} + 1[else]$.

This step functions is linear and monotonic piecewise. The model does not include quadratic terms as Lou and Caruana et al. [10] did. Quadratic neurons are difficult to interpret, especially in multilayer neural network [11]. The authors then choose sigmoid as the activation function. Sigmoid function inserts some nonlinearity to make the model more accurate and flexible. The paper then denote the feature subsets and send to subscales. The 23 features yield 10 subscales where each subscale can be interpreted as a miniature model to predict the credit score. The subscale scores are then linearly combined and transformed to final probability of bad loan using nonlinearly. The neural network is shown in Figure 2 in Appendix. The colors in the results indicate the contribution of each feature, each subscale to the combined score. Red indicates highest contribution while green indicates the lowest contribution.

Instead of training the neural network with a large number of layers and nodes, the paper cautiously chooses input features and the number of layers. It also keeps the initial transformation function and the combination function linear. This enables the model to be interpretable rather than a black box. Machine learning can predict the quality of loan faster than human. But they often learning deeper. It indeed sacrifices some accuracy and flexibility, but it avoids overfitting and largely guarantees fairness.

In addition to the two novel techniques mentioned above, some other cutting-edge methods are also focusing more on fairness of credit risk assessment. To ensure the fairness for the prediction of credit risk, scholars have introduced mathematical fairness formulation like equal odds, positive predictive parity, and counterfactual fairness [12]. Counterfactual explanations also provide us with a useful way to interpret models. The difference is that counterfactual credit risk assessment provides a way to interpret black box models. It does not care about what the model is applied, SVM, Random Forest or other. Counterfactual reasoning will also describe the importance of features, which will help people make decision how to refine the assets [13]. Data scientists at IBM also promoted interpretable models (Boolean Rules Column Generation) BRCG and GLRM. Some other techniques like making a copy for training data are also helpful.

2.2 Target on Feature Selectionn

Feature selection is also another important point that is emphasized to improve the fairness of credit risk. Based some the laws and regulations, “family name”, “gender”, “date of birth” are identified as potentially discriminatory features. However, they were

traditionally regarded as related features to measure the credit risk. To avoid unfairness, these features should be deleted in the feature selection part. As mentioned in Part I, some other features are strongly related to these sensitive features.

For example, social media are sometimes applied by scholars to predict credit risk [14], which should be used carefully. Social media may easily reveal the gender, age and other sensitive information of the users sine machine may learn more than we believe. This may also lead to unfairness and discrimination.

2.2.1 Conclusion

In summary, the novel techniques, especially those appears after the promotion of FCRA, are mainly aiming at solving the problem on interpretability and feature selection. With the two sub-problems solved, the credit risk assessment will probably be much fairer. This is not just good news for those micro businesses, but also for people need money for emergencies.

3 Appendix

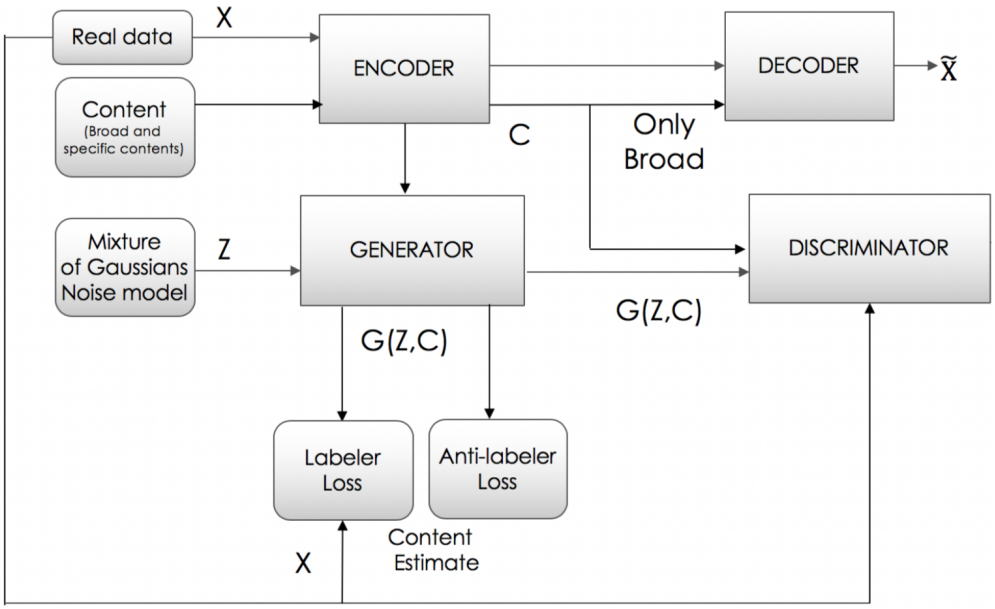

Figure 1: Block diagram illustrating the system architecture.

Figure 2: Left: Neural Network visualization, with colors indicating contribution to the final score. The 23 feature values are entered on the left. Right: Final combined score pop-up.

References

[1] D. J. Z. By Dinesh Bacham, "Machine learning: Challenges, lessons, and opportunities in credit risk modeling." https://www.moodysanalytics.com/risk-perspectives-magazine/managing-disruption/spotlight/machine-learning-challenges-lessons-and-opportunities-in-credit-risk-modeling.

[2] NAWRB, "The gender gap: Women as mortgage consumers." https://www.nawrb.com/the-gender-gap-women-as-mortgage-consumers/.

[3] Department of Justics, the United States, "The equal credit opportunity act [ECOA]." https://www.justice.gov/crt/equal-credit-opportunity-act-3.

[4] A. Klein, "Credit denial in the age of ai." https://www.brookings.edu/research/credit-denial-in-the-age-of-ai/.

[5] Federal Trade Commission, "Fair credit reporting act." https://www.ftc.gov/enforcement/statutes/fair-credit-reporting-act.

[6] R. Srinivasan, A. Chander, and P. Pezeshkpour, "Generating user-friendly explanations for loan denials using gans," *arXiv preprint arXiv:1906.10244*, 2019.

[7] C. Kemp, A. Perfors, and J. B. Tenenbaum, "Learning overhypotheses with hierarchical bayesian models," *Developmental science*, vol. 10, no. 3, pp. 307–321, 2007.

[8] M. Kocaoglu, C. Snyder, A. G. Dimakis, and S. Vishwanath, "Causalgan: Learning causal implicit generative models with adversarial training," *arXiv preprint arXiv:1709.02023*, 2017.

[9] C. Chen, K. Lin, C. Rudin, Y. Shaposhnik, S. Wang, and T. Wang, "An interpretable model with globally consistent explanations for credit risk," *arXiv preprint arXiv:1811.12615*, 2018.

[10] Y. Lou, R. Caruana, J. Gehrke, and G. Hooker, "Accurate intelligible models with pairwise interactions," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 623–631, 2013.

[11] F. Fan and G. Wang, "Fuzzy logic interpretation of quadratic networks," *Neurocomputing*, vol. 374, pp. 10–21, 2020.

[12] M. S. A. Lee and L. Floridi, "Algorithmic fairness in mortgage lending: from absolute conditions to relational trade-offs," *Available at SSRN*, 2020.

[13] R. M. Grath, L. Costabello, C. L. Van, P. Sweeney, F. Kamiab, Z. Shen, and F. Lecue, "Interpretable credit application predictions with counterfactual explanations," *arXiv preprint arXiv:1811.05245*, 2018.

[14] J. R. KOREN, "What does that web search say about your credit?." https://www.latimes.com/business/la-fi-zestfinance-baidu-20160715-snap-story.html.