Application of Supplemental Sampling and Interpretable AI in Credit Scoring for Canadian Fintechs: Methods and Case Studies

Yi Shen(✉)

Data and Analytics, Decision Insights, Equifax, Toronto, Canada
shawnie.shen@equifax.com

Abstract. Over the last decade, the fintech industry has witnessed fast growth, where online or digital lending has appeared as an alternative and flexible form of financing in addition to popular forms of financing, such as term loans, and expedited the entire lending process. High-tech online platforms have provided better user experiences and attracted more consumers and SMBs (small and medium-sized businesses). Partnerships and collaborations have also been forged between financial startups and incumbents or traditional lending institutions to further differentiate products and services to niche markets and accelerate the way of innovations. With the evolving landscape of the fintech sector, risk management tools such as credit scoring have become essential to assess credit risks such as default or delinquency based on debtor credit history or status. Many fintech companies are relatively young and sometimes serve only a small portfolio with a relatively scarce delinquency history. How can they predict default risk when making financing decisions on new applications? In this paper, we document a framework of leveraging supplemental samples of consumer or business credit information from the credit bureau that can be augmented with fintech applications for credit scoring based on theoretical and empirical studies of credit application data from a Canadian online auto leasing corporation. We also provide and compare credit scoring modeling solutions utilizing interpretable AI and machine learning methods such as logistic regression, decision tree, neural network, and XBGoost.

Keywords: Credit scoring · Sample selection bias · Machine learning · Interpretable AI · Validation

1 Introduction

Between 2010 and 2020 about 20 fintech hubs have been formulated around the globe with strong growth trends and activities or investments [1]. Canada has become home to several of the leading hubs benefiting from good quality of national regulation and mature business environment, attracting a strong talent pool and embracing technology innovation [2]. Consumers and SMBs are

W. Chen et al. (Eds.): ADMA 2022, LNAI 13725, pp. 3–14, 2022.
https://doi.org/10.1007/978-3-031-22064-7_1

now taking advantage of innovation capabilities of fintechs with more flexible financing terms and programs, accelerated customer experiences from digital banking and fast and transparent funding opportunities with less fees and costs. Traditional credit risk assessment tools such as credit scoring conditional on applicants' previous credit history and current credit status are still commonly used by lending fintechs to assess borrowers' creditworthiness and delinquency risk or to make prediction of their probability of default after application. Various generic credit scores such as credit bureau scores or FICO score [3] have existed for many years to evaluate consumer or commercial credit risk and almost become universal rules of standard for underwriting. They can be easily acquired by fintech financers from credit bureaus at the point of applications and usually have good and robust performance in helping risk judgment. Sometimes customized credit scores have also been developed by lenders based on their own existing credit portfolios that make better prediction or risk classification.

One observation with this approach is that for fintechs with relatively short history of establishment, small portfolios of customers from niche markets and prudent credit decision process or market strategies, the scarcity of delinquency history or low portfolio default rate has sometimes caused generic credit scores difficult to build and unreliable overtime even they have been built. This could also be contributed by the lack of funded applications in the case of commercial credit seekers such as SMBs. Section 2 proposes a framework of supplementary sampling of consumers or businesses with similar credit quality and their proxy credit trades from credit bureau reported by various partner lenders, which is based on application credit bureau scores and default rate distributions of original fintech applications. It can be used to build credit scoring models tailored to specific fintech companies to overcome aforementioned model data issues.

In the era of big data, interpretable AI and machine learning have gained more popularity over the years and become almost ubiquitous in credit scoring literature. While traditional approach logistic regression has still been widely applied in industry practices due to its simplicity, robustness and parsimonious form and easy interpretability for regulatory purposes, many sophisticated generic classification algorithms have also been implemented in credit risk modeling and proven to provide better risk prediction and validation results in cases such as e-commerce or retail banking [4,5]. Section 3 reviews several credit risk modeling techniques including logistic regression, decision tree, constrained neural network and XGBoost for consumer or commercial credit scoring. Section 4 compares their performances based on real data analysis of application and proxy samples from a Canadian fintech company and credit bureau.

2 Supplementary Sampling

It has been well known that credit scoring model exploiting only booked applications might introduce selection bias due to reasons such as cherry picking in credit decisioning process [6]. Many reject inference methods and strategies have been evaluated under various assumptions of missing mechanism, which include

statistical inference based on regression models or reclassification, variations of Heckman's 2-step sample bias correction procedure and nonparametric methods, and recently proposed machine learning methods such as SMOTE and graph-based semi-supervised learning algorithm [7]. This paper does not focus on these reject inference methods, instead we take an approach of utilizing supplementary samples and using proxy trade performances for rejects or non approvals as documented by Barakova et al. (2013) [8]. A proxy trade is a trade from similar credit product but funded by other lenders. Using massive credit bureau database, rejected or non-funded applications from a specific lender can be identified if they were able to get funds from other competitors and their performances on similar trades can also be extracted for risk modeling.

Building a credit scoring model on small application data can be unstable or unrealistic even when there could exist abundant credit attributes at the point of application due to the problem of overfitting or curse of dimensionality [9]. Independent consumers or SMBs with similar credit quality measured by credit bureau scores and their proxy consumer or commercial credit trades from the similar credit products funded by other lenders can be extracted and augmented to the original fintech credit applications so that enough bad observations and sufficient model sample size can be achieved. We prove that under certain assumptions of distributions of application credit scores, credit attributes and default rate, the conditional probability estimation of default can be unbiased using the augmented samples from both fintech applications and proxy population.

2.1 Notations

We introduce mathematical notations for credit scoring modeling as following: Let $X = \{X_1, ..., X_p\}$ be the available credit attributes from credit bureau at point of application, these are usually composed of hundreds or even thousands of predictive attributes related to applicant's credit characteristics such as payment or delinquency history in various credit products or trades, e.g., credit card, line of credit, installment loans, mortgages or telco trades; derogatory public records such as charge-off, collection or bankruptcy. They can also be related to applicant's credit appetite such as trade balance, credit limit or utilization and credit history such as number of trades opened and their ages or new credit.

Let $S = \{S_1, ..., S_s\}$ be credit scores at point of application, they could be multiple bureau risk scores, or bankruptcy scores and FICO scores.

Let $Y = \{0, 1\}$ be the indicator of default or non-default of a credit trade or loan in a performance window, e.g. 12 or 24 months after application.

Let $Z = \{F, P\}$ be the indicator of whether an application is from a specific fintech and funded by the fintech itself or by a competitor after reject inference, or an independent proxy trade from similar credit product or portfolio booked by other lenders reported to credit bureau, P represents that the trade is a proxy, F represents a trade related to a specific fintech.

Let $P(Y = 1|X, S, Z = F)$ or $P(Y = 1|X, S, F)$ be the conditional probability of default of fintech applications given application credit bureau scores and credit attributes that we want to estimate using credit scoring models.

2.2 Theories

Theorem 1. *Assuming that the distribution of applicants credit attributes related to a credit trade given default status and the same application credit bureau scores is independent of whether the trade is from a specific fintech or an independent proxy trade from credit bureau report pool, i.e.,* $P(X|Y,S,F) = P(X|Y,S,P) = P(X|Y,S)$*, if a random sample from the proxy population has the same joint distribution of default and application credit scores as that of the applications from a specific fintech, i.e.,* $P(Y,S|P) = P(Y,S|F), Y = 1$ *or* 0*, then we have the conditional probability of default as following:*

$$P(Y|X,S,F) = P(Y|X,S,P) = P(Y|X,S) \tag{1}$$

Corollary 1. *Under the same assumption as in Theorem 1, if a random sample from the proxy population has the same conditional credit score distribution* $P(S|Y,P)$ *and marginal default distribution* $P(Y|P)$ *as the fintech applications, i.e.,* $P(S|Y,P) = P(S|Y,F)$ *and* $P(Y|P) = P(Y|F)$*, then we have the conditional probability of default as following:*

$$P(Y|X,S,F) = P(Y|X,S,P) = P(Y|X,S) \tag{2}$$

Corollary 2. *Under the same assumption as in Theorem 1, if a random sample from the proxy population has the same conditional default probability distribution* $P(Y|S,P)$ *and the marginal application credit scores distribution* $P(S|P)$ *as the fintech applications, i.e.,*$P(Y|S,P) = P(Y|S,F)$ *and* $P(S|P) = P(S|F)$*, then we have the conditional probability of default as following:*

$$P(Y|X,S,F) = P(Y|X,S,P) = P(Y|X,S) \tag{3}$$

Proof. See proof in Appendix.

2.3 Sampling Strategies

Based on Theorem 1 and corollaries, we propose three strategies of proxy sampling for credit scoring modeling of Fintechs' applications:

Let N be the total number of funded fintech applications and reject inferences.

Let $I(Y_i)$ be the indicator of applications being default or not default.

Let $I(S_{ik})$ be the indicator of the ith application falling into a score band k. If there exists multiple application scores from the credit bureau, the score band could be extended to a grid.

Let $I(Y_i, S_{ik})$ be the indicator of the ith application falling into an application credit score band k and being default or not default.

Strategy 1. Stratified sampling based on joint distribution of default and application credit scores $P(Y,S|F)$:

1. Find the empirical joint distribution of default and application credit scores $\hat{P}(Y,S_k|F) = \frac{\sum_{i=1}^{N} I(Y_i,S_{ik})}{N}$, where $Y_i = 1$ or 0.

2. Break the proxy population into strata based on combination of application scores bands and default status, draw a stratified sample from proxy population using proportional rate at each stratum according to $\hat{P}(Y, S_k|F)$.

Strategy 2. Stagewise sampling based on marginal default distribution $P(Y|F)$ and posterior conditional distribution of application credit scores given default status $P(S|Y, F)$:

1. Find the empirical marginal default rate $\hat{P}(Y|F) = \frac{\sum_{i=1}^{N} I(Y_i)}{N}$, $Y_i = 1$ or 0 and the empirical posterior conditional distribution of application credit scores given default status $\hat{P}(S_k|Y, F) = \frac{\sum_{i=1}^{N} I(Y_i, S_{ik})}{\sum_{i=1}^{N} I(Y_i)}$, where $Y_i = 1$ or 0.
2. Do a two stage sampling from proxy population by breaking the sample into default or not default according to empirical marginal default distribution $\hat{P}(Y|F)$ from step 1 at the 1st stage, then draw a stratified sample using proportional rate at each stratum according to $\hat{P}(S_k|Y, F)$.

Strategy 3. Stagewise sampling based on marginal application credit score distribution $P(S|F)$ and posterior conditional distribution of default given application credit scores $P(Y = 1|S, F)$:

1. Find the empirical marginal distribution of application credit score $\hat{P}(S_k|F) = \frac{\sum_{i=1}^{N} I(S_{ik})}{N}$, where S_{ik} falls in score band k and the empirical posterior conditional distribution of default given application credit scores $\hat{P}(Y|S_{ik}, F) = \frac{\sum_{i=1}^{N} I(Y_i, S_{ik})}{\sum_{i=1}^{N} I(S_{ik})}$, where $Y_i = 1$ or 0.
2. Do a two stage sampling from proxy population by breaking the sample into score bands or grids according to empirical marginal score distribution $\hat{P}(S_k|F)$ from step 1 at the 1st stage, then draw a stratified sample using proportional rate at each stratum according to $\hat{P}(Y|S_k, F)$.

It is worth to notice that using the same estimation methods of empirical distributions of default and application credit scores, the three sampling strategies are essentially equivalent.

3 Techniques of Credit Scoring

A wide variety of credit scoring techniques have been used to build credit scoring models. Hand and Henley (1997) [10] offer an excellent review of the statistical techniques used in building credit scoring models. Abdou and Pointon (2011) [11] have extensively reviewed both of the traditional and advanced techniques of credit scoring. In this paper we mainly investigate and compare four of the credit scoring techniques:

Logistic regression is a type of statistical regression analysis often used to predict the outcome of a binary target variable, e.g., default or not, conditional on a set of independent predictive variables such as the credit attributes and

credit scores, it is assumed the probability of default is linked to the predictive credit attributes through a logit function [12].

Decision tree is a type of classification methods based on recursive partitioning of predictive attributes or features depending on the information gain or purity measures such as entropy or GINI after splits [13]. There have been variations of decision trees, such as CART or CHAID based on different types of target variables or splitting criteria. They are a convenient tool that can automatically handle large amounts of predictive attributes and feature selection.

XGBoost provides a regularizing gradient boosting framework for various computing languages through an open-source library [14]. The method contains rounds of iteration that create a weighted summation of learners or ensemble through gradient descent search that minimizes the empirical loss function [15]. It's a process of merging all weak classifiers together to have a model with better performance [4]. Sometimes it can also suffer from imbalanced data [16,17]. XGBoost is a slightly different version of boosting in that the optimization is not directly based on gradient descent but based on approximation. Like decision trees, XGBoost provides a convenient feature selection process through iteration and often achieves higher accuracy than a single decision tree. However, it also sacrifices the intrinsic interpretability of decision tree diagrams through aggregation of multiple tree learners.

Constraint Neural Network or NDT is a refined version of neural networks with additional constraints such as monotonic relationship between predictive attributes and target variable. A set of credit attributes are fed into multiple layers of neuron nodes through activation functions and output the final prediction through the last hidden layer. By adding monotonic constraints, it offers more interpretability for regulatory purposes and creates logical reason codes for credit decisioning while still maintaining the machine learning structure and the accuracy from artificial intelligence [18]. The invention is credited to Turner M., Jordan, L. and Joshua, A. (2021) [19], it requires stringent feature selection and normalization of data before training.

4 Empirical Studies

4.1 Data Source and Sample Facts

Auto lease application data submitted between Nov 2015 and Dec 2019 from a recently established Canadian online auto leasing corp has been analyzed. Bad or default of an auto lease is defined as 90dpd+ or worse in 24 months post application. There are 47,818 consumer applications, out of which 31,754 consumers have been funded by the fintech lender, 152 of them are defaults (0.48% bad rate); 9,093 consumer applications can also be qualified as commercial SMBs, out of which 6,035 commercial applications have been funded by the fintech lender, 33 of them are defaults (0.55% bad rate).

For the rest of non-funded applications, reject inferences are performed based on auto lease trades from credit bureau consumer or commercial trade pools opened within 2 months after the original submission of fintech application, only 559 of non-funded consumer applications can be found opened a consumer auto

lease with 9 defaults (1.6% bad rate); 156 of non-funded commercial applications can be found opened a commercial auto lease with 3 defaults (1.9% bad rate).

Three benchmark consumer credit scores and three commercial credit scores have been selected from bureau: ERS2 or Equifax Risk Score predicts consumer tendency of delinquency; BNI3 or Bankruptcy Navigation Indicator predicts consumer tendency of bankruptcy; BCN9 or FICO8 score is the Fair Issac consumer credit score. BFRS2 or bankruptcy financial risk score predicts tendency of businesses to go bankruptcy in commercial trades; FTDS2 or financial trade delinquency score predicts tendency of businesses being delinquent in financial commercial trades; CDS2 or commercial delinquency score predicts tendency of businesses being delinquent in industry commercial trades.

For the consumer applications, stratified sampling strategy 2 from only ERS2 score band strata has been performed due to the lack of bad observations in some of the combined score grids, which resulted in an independant supplementary proxy sample of 104,922 consumer auto leases with 3,888 defaults (non-defaults were down-sampled 1 out of 10, 0.38% weighted bad rate). For the commercial applications, the entire commercial proxy population of 150K commercial auto leases (1.10% bad rate) has been used without further stratified sampling because of very limited fintech observations and the fintech's business strategy and risk tolerance or appetite.

4.2 Model Development and Comparisons

The fintech applications and proxy samples are matched at bureau and appended consumer and commercial credit scores and attributes at the point of application separately based on their individual categories. There are around 2K trended or static consumer credit attributes and 800 commercial credit attributes available for modeling. Before passing them to modeling, several procedures have been completed to ensure the modeling quality including data integration and cleansing, segmentation analysis, data filtering and transformation. The consumer applications have been splitted based on a credit attribute related to the worst ever rating of all credit trades from a CART tree, which results in two segments of ever 30dpd or worse, or intuitively clean or dirty.

In the model development stage, the full samples are divided into 70% training and 30% validation set for consumer and commercial applications individually. Four credit risk modeling techniques have been utilized including: logistic regression, decision tree, XGBoost and constraint neural network (NDT). The related feature selection methods for the four procedures are as following:

Logistic Regression utilizes stepwise selection according to attribute statistical significance from maximum likelihood estimation after prescreening of credit attributes. To cope with collinearity, sometime VIF (variance inflation factor) filtering has also been applied after model fit to reduce any model attributes that are highly correlated with other predictors.

Decision tree and XGBoost automatically select attributes at each partition or iteration based on information gain or variable importance when optimizing the loss functions such as impurity or negative likelihood etc. It is similar to

stepwise selection in some sense but it would not update the previous selection and estimation when new attribute entered through iteration.

Constrained neural network (NDT) does not provide automatic feature selection, instead it requires preselection of attributes before model fitting. Top 20 to 50 attributes selected from XGBoost have been used based on attribute importance. Further reduction or addition of attributes seem not to improve the model performances.

The generic algorithms from machine learning and artificial intelligence usually requires tuning of hyperparameters that regularize the model fit, such as the number of features, tree depth, leaf size, learning rate, number of iterations, number of nodes, number of hidden layers, L1 and L2 regularization etc. Combination of these hyperparameters and grid search have been tested to find the best fits through cross validation.

For model interpretability, the credit attributes relative importance from the algorithms can be easily extracted.

4.3 Model Evaluation

We have used the following common performance measures for credit scoring model including:

KS (Kolmogorov-Smirnov) test statistic: It measures the separation ability of classification, which captures the maximum difference in the cumulative distributions of the good/bad samples. Larger KS represents better separation.

GINI or AUROC: The two metrics measure the discriminatory power of the classification models. The GINI Coefficient is the summary statistic of the Cumulative Accuracy Profile (CAP) chart. A ROC curve shows the trade-off between true positive rate (TPR) and false positive rate (FPR) across different decision thresholds. The AUROC is calculated as the area under the ROC curve. GINI is equal to 2AUROC-1. Larger GINI represents better discrimination.

Gains or Lift: A gain or lift chart graphically represents the improvement that a model provides when compared against a random guess. Gain is the ratio between the cumulative number of bad observations up to a decile to the total number of bad observations in the data. Lift is the ratio of the number of bad observations up to kth decile using the model to the expected number of bads up to kth decile based on a random or benchmark model.

Risk ordering and accuracy: As model prediction improves, the bad rate should improve in an orderly and predictable fashion. The estimated probability of bad decreases when the default risk decreases, so the observed bad rate should decrease as well. Z-statistics can compare the actual and predicted bad rate for each of the decile ranks so the direction of estimation discrepancies can be checked.

The model performance has been assessed on both training set and validation set by sample categories of the full (Proxy+Fintech+RI), fintech applications (Fintech+RI) and fintech funded, by applicant type of consumer or commercial and by combined or segments. The detailed comparison can be found in Table 1, Fig. 1 and 2. Key findings of validation include:

Table 1. Separation and discriminatory power

Consumer

Proxy+Fintech+RI	KS						GINI					
Segment	Combined		Clean		Dirty		Combined		Clean		Dirty	
	Train	Validation	Train	Validation	Train	Validation	Train	Validation	Train	Validation	Train	Validation
BCN9	38.33	37.18	35.45	35.53	37.63	36.3	47.76	46.22	42.24	42.01	49.62	44.82
ERS2	38.55	37.51	36.28	35.44	36.66	35.9	47.12	45.33	41.87	40.65	47.46	44.28
BNI3	35.06	27.53	22.44	21.51	26.94	26.76	35.73	35.21	27.99	27.54	36.48	35.31
Logistic Regression	44.2	41.52	41.78	40.92	43.77	40.04	58.59	55.03	55.97	53.53	57.7	51.45
Decision Tree	54.41	36.19	55.77	35.47	49.47	31.44	69.06	43.64	70.58	41.66	63.66	39.48
XGBoost	51.45	42.02	49.42	41.68	49.65	40.69	60.7	55.86	56.44	49.46	66.1	51.96
NDT	44.17	40.87	42.53	40.4	44.34	38.55	58.83	53.29	55.66	51.49	58.75	50.24
Fintech+RI	**KS**						**GINI**					
Segment	**Combined**		**Clean**		**Dirty**		**Combined**		**Clean**		**Dirty**	
	Train	**Validation**	**Train**	**Validation**	**Train**	**Validation**	**Train**	**Validation**	**Train**	**Validation**	**Train**	**Validation**
BCN9	30.99	34.33	32.79	29.75	27.38	38.92	28.45	42.34	38.14	34.64	31.32	44.47
ERS2	35.06	31.32	38.2	26.13	31.99	32.3	38.95	37.97	36.62	32.53	33.84	40.2
BNI3	31.63	34.56	38.19	31.65	25.46	31.81	39.96	38.78	40.01	39.49	34.34	35.77
Logistic Regression	38.27	43.59	45.19	42.86	32.86	52.47	48.52	53.53	48.84	43.61	42.04	57.11
Decision Tree	53.82	47.79	59.75	43.37	46.41	45.58	65.14	53	68.62	44.36	55.98	50.58
XGBoost	48.8	45.38	47.45	44.76	51.23	48.8	60.7	55.3	63.2	54.72	62.8	58.14
NDT	36.91	43.24	36.23	44.81	36.27	45.03	45.73	48.87	42.48	35.23	45.61	53.66
Fintech Funded	**KS**						**GINI**					
Segment	**Combined**		**Clean**		**Dirty**		**Combined**		**Clean**		**Dirty**	
	Train	**Validation**	**Train**	**Validation**	**Train**	**Validation**	**Train**	**Validation**	**Train**	**Validation**	**Train**	**Validation**
BCN9	30.27	33.11	32.62	28.21	27.82	37.15	36.75	40.75	37.37	32.97	30.27	33.11
ERS2	35.17	29.28	37.79	25.41	32.55	32.71	32.55	36.59	35.76	30.88	35.17	29.28
BNI3	32.34	31.79	39.45	30.56	25.76	28.83	40.18	34.67	40.82	37.3	32.34	31.79
Logistic Regression	37.67	44.43	44.77	41.66	34.27	52.15	48.52	53.31	48.31	43.44	37.67	44.43
Decision Tree	53.73	46.38	59.7	41.58	46.18	45.79	65.38	51	69	41.74	53.73	46.38
XGBoost	48.26	43.67	47.33	44.27	51.23	49.56	60.62	54.08	56.34	47.28	48.26	43.67
NDT	36.26	42.71	36.02	43.04	36.46	44.79	45.6	46.08	42.1	31.83	36.26	42.71

Commercial

Proxy+Fintech+RI	KS		GINI	
Score	Train	Validation	Train	Validation
BFRS2	18.08	16.78	19.49	16.16
FTDS2	17.34	16.09	19.1	17.33
CDS2	11.44	10.39	12	10.03
Logistic Regression	31.13	28.56	43.79	40.38
Decision Tree	31.36	28.25	44.26	40
XGBoost	36.5	30.76	52.36	44.12
NDT	36.41	28.52	51.16	39.93
Fintech+RI	**KS**		**GINI**	
Score	**Train**	**Validation**	**Train**	**Validation**
BFRS2	24.53	17.28	22.9	4.09
FTDS2	19.55	17.33	18.9	3.8
CDS2	14.61	24.4	9.89	12.56
Logistic Regression	43.86	16.69	44.24	4.18
Decision Tree	36.88	36.3	48.83	30.71
XGBoost	40.06	24.63	46.98	16.7
NDT	40.07	22.61	53.95	14.06
Fintech Funded	**KS**		**GINI**	
Score	**Train**	**Validation**	**Train**	**Validation**
BFRS2	25	16.38	24.18	6.67
FTDS2	19.9	15.74	19.19	4.6
CDS2	14.75	21.56	8.48	13.45
Logistic Regression	43.57	14.43	41.59	2.81
Decision Tree	36.52	35.61	46.56	30.71
XGBoost	40.36	19.27	44.64	9.32
NDT	38.71	18.63	51.27	13.38

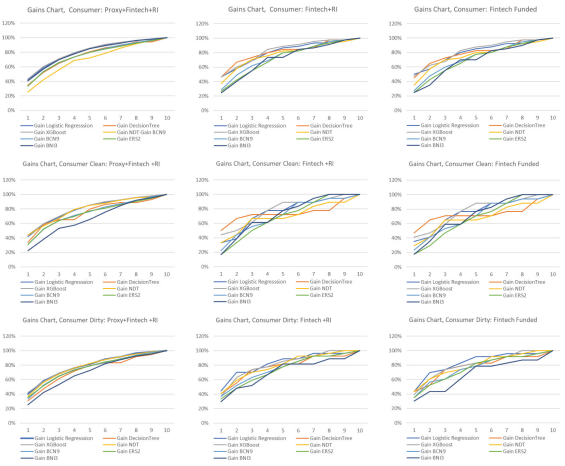

Fig. 1. Gains and lift chart:consumer scores

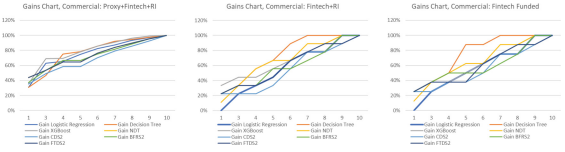

Fig. 2. Gains and lift chart:commercial scores

– Customized new scoring models almost always outperform the three benchmark bureau scores regardless of the sample categories in the validation.

- For consumer applications, XGBoost significantly outperforms the other three methods in the clean segment of Fintech applications, also has best GINI in combined segments validation. Decision tree has the best separation KS or gains at the 1st decile in the full validation data of Fintech applications. Logistic regression has the best separation KS or gains in the dirty segment of Fintech applications. NDT can have better seperation in clean segment.
- For commercial applications, there is more volatility in their performances in the fintech applications: generic algorithms always outperform logistic regression and the benchmark scores in the sample of Fintech+RI, however logistic regression and they sometimes underperform some of the benchmarks in the fintech booked applications. Decision tree seems to provide more stable separation and discriminatory power and XGBoost has the highest gains in the 3rd decile.
- For the consumer applications, the new scores seem to slightly underestimate the bad rate with most of the Z-statistics being positive but not severe (<1.96) and the risk ordering is in the right direction. For commercial applications, the new scores seem to slightly overestimate the bad rate with most of the Z-statistics being negative for fintech applications due to the higher bad rate in the proxy sample. This problem does not appear to be severe for generic algorithms, but more severe for logistic regression so further calibration may be needed even though the risk ordering is in the right direction.

5 Conclusion

Supplemental samples are effective and useful for both reject inference and predictive modeling in credit scoring. By leveraging the tailored supplemental samples, we have demonstrated that traditional credit scoring tools and innovative generic algorithms, which are from machine learning and interpretable AI, can both be utilized to assess default risk for fintech loan applications when the original volume and number of bad observations are small and not sufficient for modeling. They can provide significantly better performances than the existing benchmark scores. Generic algorithms can sometimes overfit so additional out-of-time validations need to be required for further investigation in practice.

Acknowledgements. The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

Appendix. Proof of Theroem and Corollaries

Proof of Theorem 1: According to Bayes' rules, we have the conditional probability of default given applications from a specific fintech:

$$P(Y = 1|X, S, F) = P(Y = 1, S|F)P(X|Y = 1, S, F)/P(X, S|F) \qquad (4)$$

In the above equation, the joint distribution of credit attributes and credit scores given applications from a specific fintech can be expanded as:

$$P(X, S|F) = P(X|Y = 1, S, F)P(Y = 1, S|F) + P(X|Y = 0, S, F)P(Y = 0, S|F) \tag{5}$$

we also have the conditional probability of default given an independent proxy trade in a proxy sample from credit bureau:

$$P(Y = 1|X, S, P) = P(Y = 1, S|P)P(X|Y = 1, S, P)/P(X, S|P) \tag{6}$$

And the joint distribution of application credit attributes and credit scores given an independent proxy trade in a proxy sample from credit bureau can be expanded as:

$$P(X, S|P) = P(X|Y = 1, S, P)P(Y = 1, S|P) + P(X|Y = 0, S, P)P(Y = 0, S|P) \tag{7}$$

Hence under the assumption $P(X|Y, S, F) = P(X|Y, S, P) = P(X|Y, S)$, if there exists a proxy sample that has the same joint distribution of probability of default and credit scores at the point of application as that of a trade related to a specific fintech, then the conditional probabilities of default from proxy or fintech are equivalent.

Proof of Corollary 1: Notice that according to Bayes' rules and Theorem 1, we have the conditional joint distributions in Eq. (4) and (5):

$$P(Y = i, S|Z) = P(S|Y = i, Z)P(Y = i|Z), i = 0 \text{ or } 1, Z = F \text{ or } P \tag{8}$$

By plugging (8) back into Eqs. (4)–(7), we have the desired result.

Proof of Corollary 2: Notice that according to Bayes' rules and Theorem 1, we have the conditional joint distributions in Eq. (4) and (5):

$$P(Y = i, S|Z) = P(Y = i|S, Z)P(S|Z), i = 0 \text{ or } 1, Z = F \text{ or } P \tag{9}$$

By plugging (9) back into Eqs. (4)–(7), we have the desired result.

References

1. Accenture: Collaborating to win in Canada's Fintech ecosystem, Accenture 2021 Canadian Fintech Report (2021). https://www.accenture.com/_acnmedia/PDF-149/Accenture-Fintech-report-2020.pdf
2. Goulard, B., Lake, K.T., Reynolds M.: Canadian Fintech Review, Torys LLP, November 2021. https://www.torys.com/our-latest-thinking/publications/2021/11/canadian-fintech-review
3. Fair-Isaac: FAQs-About-FICO-Scores-Canada-2019.pdf (2019). https://www.ficoscore.com/ficoscore/pdf/FAQs-About-FICO-Scores-Canada-2019.pdf

4. Tian, Z.Y., Xiao, J.L., Feng H.N., Wei, Y.T.: Credit risk assessment based on gradient boosting decision tree. In: 2019 International Conference on Identification, Information and Knowledge in the Internet of Things (IIKI2019)
5. Ma, Z., Hou, W., Zhang, D.: A credit risk assessment model of borrowers in P2P lending based on BP neural network. PLoS ONE **16**(8), e0255216 (2021). https://doi.org/10.1371/journal.pone.0255216
6. Hand, D.J.: Reject inference in credit operations: theory and methods. In: The Handbook of Credit Scoring, pp. 225–240. Glenlake Publishing Company (2001). /art00177
7. Kang, Y., Cui, R., Deng, J., Jia, N.: A novel credit scoring framework for auto loan using an imbalanced-learning-based reject inference. In: 2019 IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr), pp. 1–8. IEEE (2019)
8. Barakova, I., Glennon, D., Palvia, A.: Sample selection bias in acquisition credit scoring models: an evaluation of the supplemental-data approach. J. Credit Risk **9**, 77–117 (2013)
9. Surrya, P.D., Radcliffea, N.J.: Why size does matter in credit scoring. In: Proceedings of Credit Scoring and Credit Control V, Edinburgh (1997) (1997)
10. Hand, D.J., Henley, W.E.: Statistical classification methods in consumer credit scoring: a review. J. R. Stat. Soc. A **160**, 523–541 (1997)
11. Abdou, H., Pointon, J.: Credit scoring, statistical techniques and evaluation criteria: a review of the literature. Intell. Syst. Account. Finance Manag. **18**(2–3), 59–88 (2011)
12. Hosmer, D.W., Lemeshow, S.: Applied Logistic Regression, 2nd edn. Wiley, New York (2000)
13. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees. The Wadsworth, Belmont (1984)
14. Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. In: Krishnapuram, B., Shah, M., Smola, A.J., Aggarwal, C.C., Shen, D., Rastogi, R. (eds.) Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016, pp. 785–794. ACM (2016). arXiv:1603.02754. https://doi.org/10.1145/2939672.2939785
15. Hastie, T., Tibshirani, R., Friedman, J.H.: Boosting and additive trees. In: Hastie, T., Tibshirani, R., Friedman, J.H (eds.) The Elements of Statistical Learning, 2nd edn., pp. 337–384. Springer, New York (2009). https://doi.org/10.1007/978-0-387-84858-7_10. ISBN 978-0-387-84857-0
16. Buja, A., Stuetzle, W., Shen, Y.: Loss functions for binary class probability estimation: structure and applications, Technical report, The Wharton School, University of Pennsylvania, January 2005
17. Shen, Y.: Loss functions for binary classification and class probability estimation, Ph.D. dissertation, The Wharton School, University of Pennsylvania (2005)
18. McBurnett, M., Sembolini, F., Turner, M., Jordan, L., Hamilton, H., Torres, S.R.: Comparative Analysis of Machine Learning Credit Risk Model Interpretability: Model Explanations, Reasons for Denial and Routes for Score Improvements, Credit Scoring and Credit Control XVII, University of Edinburgh, UK, August 26 2021 (2021)
19. Turner, M., Jordan, L., Joshua, A.: Machine-learning techniques for monotonic neural networks, Equifax, US patent 11010669 (2021)