Federated Learning for Credit Risk Assessment

Chul Min Lee
charles.lee@uni.lu

Joaquín Delgado Fernández
joaquin.delgadofernandez@uni.lu

Sergio Potenciano Menci
sergio.potenciano-menci@uni.lu

Alexander Rieger
alexander.rieger@uni.lu

Gilbert Fridgen
gilbert.fridgen@uni.lu

SnT - Interdisciplinary Center for
Security, Reliability and Trust
University of Luxembourg

Abstract

Credit risk assessment is a standard procedure for financial institutions (FIs) when estimating their credit risk exposure. It involves the gathering and processing quantitative and qualitative datasets to estimate whether an individual or entity will be able to make future required payments. To ensure effective processing of this data, FIs increasingly use machine learning methods. Large FIs often have more powerful models as they can access larger datasets. In this paper, we present a Federated Learning prototype that allows smaller FIs to compete by training in a cooperative fashion a machine learning model which combines key data derived from several smaller datasets. We test our prototype on an historical mortgage dataset and empirically demonstrate the benefits of Federated Learning for smaller FIs. We conclude that smaller FIs can expect a significant performance increase in their credit risk assessment models by using collaborative machine learning.

Keywords: federated learning, artificial intelligence, credit risk assessment, financial collaboration

1. Introduction

Most financial institutions (FIs) employ comprehensive credit risk models to estimate their exposure to credit risk. These models typically employ either traditional or advanced methods. Traditional methods rely on induction principles to make mathematical and statistical inferences from curated data. They facilitate the creation of static models that build on a range of assumptions, such as linearity, independence, and normality. Advanced methods, in turn, are more data-driven and less reliant on these assumptions (Chen et al., 2016; Galindo & Tamayo, 2000). Like traditional methods, they infer information from curated data but they enable the creation of flexible models that adapt to the curated data. As a result, credit risk models that employ advanced methods typically perform better at to extracting patterns from complex real-world datasets that are replete with noise, nonlinearity, and idiosyncrasies.

Both methods depend strongly on data inputs (Altman, 2002; Heitfield, 2009). Models trained with more and better data can estimate real word situations more accurately. In effect, data availability is crucial for FIs and can translate into a competitive advantage (Bansal et al., 1993; Walczak, 2001). Limited data, in turn, can lead to less reliable predictions. For smaller FIs with limited data access, this effectively means that 'data sharing' with other FIs could have a material impact on the performance of their credit risk models (Bansal et al., 1993; Walczak, 2001). However, data sharing is often challenging due to concerns about privacy, control and legal recourse (Borgman, 2012; Ekbia et al., 2015).

A more feasible alternative could be the use of Federated Learning (FL) to create joint credit risk models. FL is an ML technique that allows models to train on a distributed basis without the need to move raw data (McMahan et al., 2016). In other words, financial institutions would not need to reveal their data as they gain insights from its processing, allowing every participating FI to benefit from use of each other's information.

In this paper, we thus ask the following two research questions:

RQ1 How does FL based credit risk assessments perform?

RQ2 Will FL help to reduce the disparities in risk calculations between financial institutions?

URI: https://hdl.handle.net/10125/102676
978-0-9981331-6-4

To answer these research questions, we developed an FL-based credit risk assessment prototype. We tested our prototype on Freddie Mac's Single Family Loan-Level Dataset (Freddie Mac, 2021b) to simulate collaboration between FIs when assessing the credit risk of mortgage portfolios. Mortgages are an important financial instrument, but their typically long time horizons complicate the task of making accurate forecasts. Specifically, we compared the performance of credit risk models under different scenarios to evaluate and quantify the impact of information sharing. These comparisons indicate that FL can offer significant performance gains for smaller FIs with limited in-house datasets. To the best of our knowledge, this paper is the first to examine FL in assessing the credit risk of mortgages with real-world FI divisions.

The research paper is structured as follows. Section 2 provides an overview of relevant literature on credit risk assessment and federated learning. Moreover, it presents previous research that studies the application of FL in financial services. Section 3 describes the implementation of our FL prototype. Section 4 details the hypotheses, scenarios, and evaluation metrics we used to examine the performance of our FL prototype. Section 5 presents the results of our evaluation. Section 6 discusses the limitations of our study as well as future research directions. Section 7 offers concluding remarks.

2. Related Work

2.1. Credit Risk Assessment

Credit risk assessment methods have evolved over time from traditional to advanced methods, but essentially start and end in the same fashion. At the start, data or information about the prospective mortgage is gathered systematically. Subsequently, the newly collected information is used to measure the likelihood of the mortgage to experience credit risk events. The likelihood of these events results in a score representing the credit risk of the mortgage.

Performance of credit risk models is not only dependent on the method used but also on data inputs (Altman, 2002; Heitfield, 2009). Models trained on larger datasets, for example, when sharing data, allow for more real world data to be represented in the trained model. As a result, changes to the quantity and quality of data inputs have material impacts on the performance of credit risk models.

Regulators and policymakers have called for increased disclosure of credit risk related data (on Banking Supervision, 2018) and developed infrastructure to encourage voluntary sharing (Bank, 2010; Israël et al., 2017). However, concerns about sharing data remain and are two-fold. Firstly, data privacy laws, such as the EU's General Data Protection, prohibit data sharing without an appropriate legal basis. Secondly, data typically offers a competitive advantage to its holder (Kearns & Lederer, 2004; Redman, 1995; Zuiderwijk et al., 2015). Therefore, companies are often reluctant to share their data to avoid risking the disclosure of valuable information.

2.2. Federated Learning

FL was introduced as a collaborative ML technique in (McMahan et al., 2017; McMahan et al., 2016) and might help to mitigate privacy and competitiveness concerns. In FL, data remains decentralized across collaborating clients. These clients collaborate through share information (or inferences) about the data rather than the data itself. FL typically builds on one of two algorithms: Federated Averaging (Fed-Avg) and Federated Stochastic Gradient Descent (Fed-SGD). The first algorithm shares model while the second model gradients.

In FL (McMahan et al., 2017; McMahan et al., 2016), there are typically two roles: clients and the central server. The roles are the same for both Fed-Avg and Fed-SGD. Clients host and locally compute ML models locally using their own data. The central server coordinates the sharing of the locally computed information from clients by aggregating, averaging and then distributing the averaged information back to the collaborating clients.

There are four FL variants in standard practice, with these based on the data structures and features used to train the FL models: Horizontal, Vertical, Transfer Learning, and Assisted Learning. Horizontal FL requires that the data used to train each client have the same data structure and features. Vertical FL requires that each client has the same structure but different data features. Transfer learning allows each client to have different structures and features in their data. Assisted Learning allows each client to train using other clients' errors.

The training of an ML model with FL follows an iterative process as depicted in Figure 1. The training steps are as follows: initially, in (1), the central server selects a list of collaborating clients and an ML model to be run by each of the selected clients. Subsequently, in (2), the central server communicates the selected ML model to each randomly selected client. After receiving the selected ML model, in (3), each client simultaneously trains the selected ML model on their

data and produces a newly trained model. In (4), each client communicates to the central server the computed results of their new ML models. Once the central server has aggregated the information from all clients, in (5), it will average the aggregated information. Lastly, in (6) the aggregated information is relayed back to each newly randomly selected subset of clients. This collaborative training process continues repetitively between steps (2) and (6) until a prescribed number of rounds are complete, or a target goal is reached.

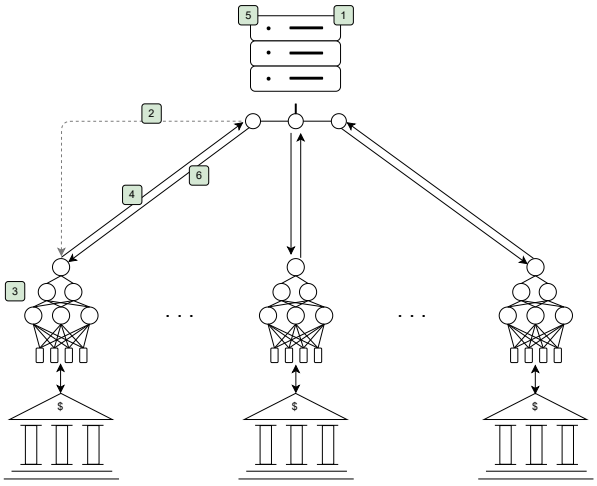

Figure 1. Diagram of Federated Learning process.

2.3. Federated Learning for Financial Services

FL is a relatively novel ML method that allows disconnected entities to train ML models without sharing their raw data. At the same time, there is a burgeoning quantity of literature in various fields which study the potential impact of FL on analytical capabilities, such as medical imaging (Kaissis et al., 2021), the Internet of Things (Aïvodji et al., 2019), and energy demand optimization (Saputra et al., 2019), which study the potential impact of FL on analytical capabilities.

FL is also gaining traction amongst financial services businesses. For instance, Yang et al. (2019) proposes a FL framework to train fraud detection models. They use an anonymized real-world dataset of credit card transactions from European cardholders provided by the Université Libre de Bruxelles (ULB) ML Group. Their framework demonstrates an increase in performance of approximately 10% when implementing FL versus conventional ML approaches. Zheng et al. (2020) propose an FL framework to train meta-learning based models. They test their proposed framework on four publicly available credit card transaction datasets. These tests demonstrate an increase in performance compared to conventional meta-learning approaches. Shingi (2020) applies an FL model to predict loan defaults using a modified learning algorithm for FL and a Feed-Forward Network exhibiting a 3.88% increase in performance. However, the current literature still lacks real portfolio divisions of small, medium, and large FI to create credit risk models that can consume large datasets. Thus, we contribute to reducing this gap in our paper.

3. Prototype

To evaluate empirically the effectiveness of FL in assessing the credit risk of mortgages, we developed an FL prototype. The developed FL prototype estimates the probability of default of the underlying mortgages. We use historical data of mortgage transactions and real division of entities for our federated clients.

3.1. Data Source

The datasets used to train and test our FL prototype are Freddie Mac's Single-Family Loan-Level (FMSFLL) (Freddie Mac, 2021b), Freddie Mac's House Price Index (FMHPI) (Freddie Mac, 2021a), United States Bureau of Labor Statistics' Local Area Unemployment Statistics (LAUS) (United States Bureau of Labor Statistics, 2021), and Federal Reserve Economic Data (FRED) (Federal Reserve, 2021). While the FMSFLL dataset provides data directly related directly to mortgage transactions, the FMHPI, LAUS, and FRED datasets provide complementary data related to economic and environmental factors.

The FMSFLL dataset holds historical records of credit performance data on all mortgages that Freddie Mac has purchased or guaranteed since 1999 and covers approximately 45.5 million mortgages. The dataset has two tables: origination and monthly. The origination table has 35 variables and includes data relevant to when the FI granted the mortgage to the applicant. The monthly table concerns data relevant to the status of the mortgage granted at monthly intervals. The "Seller Name" describes the originating financial institution that initially funded the mortgage transaction at its inception. It allowed us to isolate the mortgages that originated from each FI as true portfolio holdings before Freddie Mac acquired them.

We complemented the FMSFLL dataset with the FMHPI, LAUS, and FRED datasets to include relevant 'environmental' factors for mortgage defaults. Intuitively, the factors considered are general levels of housing price, unemployment, delinquency, charge-off, and interest rates. The FMHPI dataset holds historical

housing price levels in the US by state. The LAUS dataset holds historical unemployment levels in the US by state. Lastly, the FRED dataset holds delinquency, charge-off, and interest rate levels in the US at the national level. As a result, the FL prototype considers descriptive information about mortgages and relevant 'environmental' factors for mortgage defaults.

3.2. Data Pre-processing

We pre-processed our combined dataset to make it more accessible for our prototype. Firstly, due to the large amount of variables, we reduced the number of those we used to 31 (Table 1). Secondly, to reduce the scope of our evaluation, we worked only with data points from 2006 until 2009. We chose this time frame because the US mortgage markets had a high rate of defaults in those years, which provided an ideal period to test our prototype. Thirdly, to have a consistent terminating state, we only considered mortgage records in the FMSFLL dataset which had default and non-default "termination events" in the Zero Balance Code variable. This only includes mortgage records that experienced credit events such as "Third Party Sale", "Short Sale or Charge Off", "Repurchase prior to Property Disposition", and "REO disposition".

After pre-processing, the combined dataset resulted in 9.6M observations from 250k unique mortgages previously held by 14 FIs.

3.3. Prototype Implementation

The combined dataset after pre-processing is time-stamped at monthly intervals. As time-stamped datasets allow for the consideration of temporal patterns, we chose a Neural Network (NN) with four layers of Long Short-Term Memory (Hochreiter & Schmidhuber, 1997) interspersed with dropout layers and two fully connected layers. Unlike fully connected layers, LSTM layers have feedback connections with previous neurons. These connections allow neurons to access information about their former states so they can make inferences about the future based on previous data.

In NNs, the frequent use of dropout layers mitigates overfitting (Srivastava et al., 2014). During the training phase, dropout will deactivate some neurons at random, encouraging the network to find ways around previously established patterns and preventing some neurons from becoming a bottleneck in the architecture. As a result, the selected architecture can find temporal patterns in the development of the mortgage market (Sezer et al., 2020).

We implemented the FL prototype using the algorithm presented in (McMahan et al., 2017). In a first step, the central server initializes the baseline model and distributes it to the FIs. In a second step, the FIs start training the model sent by the central server on their local data. In the training they conduct, they use Stochastic Gradient Descend (SGD) (Robbins & Monro, 1951) as the model optimizer with the parameters defined in Table 2. As we are implementing Fed-Avg, the communication rounds happen after one or more complete pass through the dataset, We found 10 internal rounds before averaging the optimal number

Table 1. List of variables.

Variable	Dataset	Data-type
Channel	FMSLL	Discrete
Charge-Off Rate	FRED	Continuous
Combined Unemployment Rate	LAUS	Continuous
Credit Score	FMSLL	Continuous
Current Actual Unpaid principal balance	FMSLL	Continuous
Current Loan Delinquency Status	FMSLL	Discrete
Delinquency Due to Disaster	FMSLL	Discrete
Delinquency Rate	FRED	Continuous
Estimated Loan-to-Value (ELTV)	FMSLL	Continuous
First Time Homebuyer Flag	FMSLL	Discrete
Fixed Rate Mortgage Average	FRED	Continuous
House Price Index	FMHPI	Continuous
Loan Age	FMSLL	Continuous
Loan Purpose	FMSLL	Discrete
Loan Sequence Number	FMSLL	Discrete
Mortgage Insurance Percentage (MI %)	FMSLL	Continuous
Number of Borrowers	FMSLL	Continuous
Number of Units	FMSLL	Continuous
Occupancy Status	FMSLL	Discrete
Original Debt-to-Income (DTI) Ratio	FMSLL	Continuous
Original Interest Rate	FMSLL	Continuous
Original Loan Term	FMSLL	Continuous
Original Loan-to-Value (LTV)	FMSLL	Continuous
Property State	FMSLL	Discrete
Property Type	FMSLL	Discrete
Property Valuation Method	FMSLL	Discrete
Remaining Months to Legal Maturity	FMSLL	Continuous
Seller Name	FMSLL	Discrete
Super Conforming Flag	FMSLL	Discrete
Unemployment at origination	FRED	Continuous
Zero Balance Code	FMSLL	Discrete

of rounds. Once all the FIs have finished their training, they share their models weights' with the central server. The central server averages the weights, creating a new model. This model is then shared again with the FIs until a pre-determined number of rounds is reached. In our case, after 100 rounds, there was no improvement in the performance metrics.

We simulated all the FIs and the communications on the High-Performance Computing facilities of the University of Luxembourg's (Varrette et al., 2014). The hardware used was 256Gb of RAM and one 16Gb/32Gb NVIDIA Tesla V100 depending on the allocation. We implemented the FL architecture in TensorFlow Federated "TensorFlow Federated" (2018) while for the Deep Learning modules we used Keras (Chollet et al., 2015)

Table 2. Hyperparameters for FL models.

Parameter	Value
Rounds before averaging	10
Baseline architecture	4x (LSTM + Dropout) + 2 Dense
Total number of FIs	14
Optimizer	SGD
Optimizer Learning rate (L_r)	0.01
Optimizer Momentum (v_{t+1})	0.9
Optimizer Decay (λ)	1e-2/100
Batch size	128
Number of communication rounds	100

4. Evaluation

To analyze the performance effects of the FL prototype, we formulated a null and an alternative hypothesis, defined metrics to measure performance, and designed a series of scenarios to test the hypotheses.

4.1. Hypotheses

We formulated our null and alternative hypotheses as follows: Given an FI's dataset F_i: $\{F_1, F_2, ..., F_n\}$ and a global dataset $D = \{F_1 \cup F_2 \cup ... \cup F_n\}$ with n being each individual FI, the null hypothesis (H_0) is that the performance of models trained collaboratively through FL on F_i is better than the performance of the same model trained on F_i. The alternative hypothesis (H_1), in turn, is that the performance of models trained collaboratively through FL is worse than the same model trained on F_i.

4.2. Evaluation Metrics

We used a range of standard metrics to measure the performance levels of the models: Accuracy, Recall, Precision, and F1. The performance of any classification task could be summarized using four main indicators: True positive (TP) representing the instances correctly classified, similarly true Negatives (TP) where the model correctly predicts the negative class. On the other hand false positive (FP) represents the instances incorrectly predicted as positive class. False negative (FN) represents the instances incorrectly predicted as negative class. Eq. 1 Accuracy describes the proportion of correct predictions as opposed to the total number of predictions. Eq. 2 Recall describes the proportion of positive classifications that were correctly classified over the all the positive instances Eq. 3 Precision describes the proportion of positive classifications that were correctly classified among all instances. Eq. 4 F1 is the equal weighted harmonic average of Eq. 2 and Eq. 3.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$F1 = \frac{2 * TP}{2 * TP + FP + FN} \tag{4}$$

4.3. Evaluation Scenarios

We designed a series of five scenarios to test the null hypothesis: 1) Local Model, 2) Central Model , 3a) FL, 3.b) FL without the biggest FI (n-1), and 3.c) FL without the two biggest FIs (n-2). Each scenario represents a hypothetical instance of credit risk assessment. Each scenario uses data from the years 2006 to 2008 as training data, and data from year 2009 as testing data. We calculated the performance metrics for the five scenarios in relation to the observed loan termination status.

1. **Local Model** This scenario explores independent FIs that only use the data at their immediate disposal and without collaboration for credit risk assessment. It denotes an 'imperfect information' scenario in which FIs do not have access to each other's data. In this scenario, we trained one model for each FI.

Table 3. Scenario and the data they ingest.

Scenario Name	Train data	Tested data
1. Local Model	Own	Own
2. Central	All	All
3a. FL	FL	Local
3b. FL n-1	FL	Local
3c. FL n-2	FL	Local

2. **Central** This scenario represents a hypothetical data lake in which all data is pooled together in a singular or centralized data silo for credit risk assessment. It represents a 'perfect information' scenario in which all the data of every FI is available. In this scenario, we trained one model in a 'centralized' manner.

3. (a) **Federated Learning** This scenario represents collaboration using Horizontal FL. Each FI stores their own data while their data structure remains identical among clients.
   (b) **Federated Learning without the biggest bank** This scenario explores collaboration without Wells Fargo Bank, N.A. Wells Fargo is the FI with the largest number of unique mortgages in our dataset. It holds 40.21% of the total number of unique mortgages.
   (c) **Federated Learning without the two biggest FIs** This scenario is similar to the previous one and explores the impact of removing the two biggest FIs. Wells Fargo and Chase Home Finance are the two FIs with the largest number of unique mortgages. The two account for 52.59% of the total number of unique mortgages.

To ensure that the results are robust and to normalize the effects of NN's random nature, we utilized a Monte Carlo simulation (Kroese et al., 2014). The n in Table 4 presents the number of simulations for every particular scenario. Additionally, we summarized the metrics by the mean μ and their standard deviation σ. In the **Local Model** scenario, we compute μ and σ across the different FIs for each Monte Carlo simulation.

5. Results

Based on our simulations, we fail to reject the H_0 (the FL model is better than the local model). The hypothesis holds even for scenarios where the largest and two largest FIs do not collaborate in training a forecasting model to predict credit risk. To support our rejection, we provide our simulation results in Table 4.

Table 4. Performance comparison between scenarios.

Scenario	Accuracy	Recall	Precision	F1 score
1. Local Model	μ = 95.04%	96.97%	89.76%	92.65%
$n = 10$	σ = 0.0667	0.0249	0.1359	0.0879
2. Central	μ = 98.59%	99.8%	95.56%	97.49%
$n = 10$	σ = 0.0263	0.0041	0.0845	0.0468
3a. FL	μ = 99.06%	98.81%	97.69%	98.25%
$n = 10$	σ = 0.0002	0.0005	0.0008	0.0004
3b. FL n-1	μ = 99.04%	98.74%	97.69%	98.21%
$n = 10$	σ = 0.0002	0.0006	0.0011	0.0005
3c. FL n-2	μ = 99.04%	98.72%	96.82%	98.22%
$n = 10$	σ = 0.0004	0.0007	0.0015	0.0007

We found that the local model results offer an average performance worse than the other models. Moreover, we found that the performance of the models was proportional to the number of mortgages on which they were trained. To quantify this effect, we fitted a linear regression between the number of mortgage records and the performance of the models. We found that a 1% increase in the number of loans increased the performance by an average of 0.06% with $p = 0.02 \leq 0.05$.

This relationship between data quantity and model performance explains the variability in the evaluation metrics. For example, the model for Metlife Home Loans, a division of Metlife Bank, N.A., that holds 45340 mortgage observations (0.33% of the total number of observations in our dataset), had 91.12% recall, 75.97% accuracy, a F1 score of 69.44%, and 56.56% precision. Meanwhile, the Wells Fargo Bank NA model with almost 4m mortgage observations (40.20% of the total observations) had 98.97%, 97.95%, 97.35%, and 97.65% accuracy, precision, recall, and F1 score, respectively. In effect, we can conclude that the higher the number of records per financial institution, the higher their performance levels.

On the contrary, the central model created by all FIs sharing their data in a silo outperforms the **Local Model** scenario by a relatively average performance increase of 4.27 percentage points (pp). However, the difference is not significant between the central and FL models. The difference is 0.57 pp in favor of the FL model; this difference being negligible mainly due to the stochasticity of the models.

Surprisingly, even when we excluded the FIs with the most mortgage records (**FL n-1** and **FL n-2**), the performance levels still matched those of the **Central** and **FL** scenario. Even without the 40.21% and 52.59% respectively of the total mortgage observations,

performance did not drop significantly.

The standard deviation over our Monte Carlo simulations (see Table 4 under σ) indicates that there are only minor variations across simulations, exemplifying the robustness of the results. These findings demonstrate that small-to-medium FIs could significantly improve their credit risk assessments by joining forces with others to create a collaborative FL model.

Overall, each FI holds different data. This variation in data induces each FI to estimate credit risk differently, sometimes creating overexposure and other times underexposure. First, we explored how these differences between models develop over time. Thus, as an example, we considered two different loans, F09Q10036282 and F09Q10037931, and explore how different models estimate their risk throughout the life cycle of the mortgage. We illustrate the results of these two loans in Figure 2, where the former defaulted (upper), whereas the latter did not (lower). For instance, we can observe in the predictions for F09Q10036282 that even though all models correctly estimate its default at the end, during the middle years, the default estimations vary across FIs. Even in these two instances, in Figure 2, the difference between the risk estimation between **Central** and **FL** remains negligible.

Secondly, we measured these differences in risk estimation over time. To do so, we calculated the kernel density approximations (Rosenblatt, 1956) of the differences in risk estimation. Individual FIs do not have a complete view of the market and tend to perform poorly in estimating risk distributions. As a complementary step, we add Figure 3, which visualizes how the **Local Model** estimates risk compared to both **Central** and **FL** models. FIs deviate from both **Central** and **FL** when calculating their risk. For example, Metlife Home Loans, a Division of Metlife Bank, N.A., tends to underestimate risk (-25%). Another example is Chase Home Finance LLC which overestimates by around 7%.

Furthermore, Table 5 collects the simple average of the default probability deviations and complements Figure 3. In simple average terms, the **Local Model** scenario underestimates the probability of default compared to the **Central Model** and **FL** scenario at both the initial (2.7% and 5.6%) and final month of mortgages (6% and 4.3%). These results can be explained since a single institution's data has less variability than the rest of the datasets combined. For instance, individual FIs seem to better estimate risk at the beginning of the mortgage but underestimate the default probability at the end. These results are reasonable since an accurate initial default estimation may be simpler to make than analyzing the changing environmental and

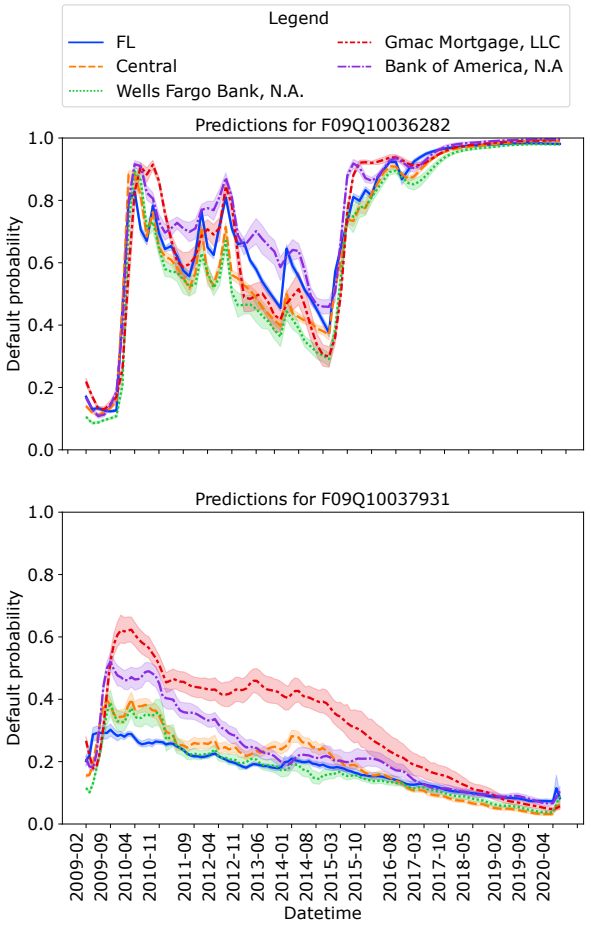

Figure 2. Different models used to estimate default probability.

macroeconomic conditions mortgages are subject to. Hence, FIs could be failing to estimate the fat tails of the market's risk distribution (Taleb, 2020). In other words, smaller FIs may not have access to a *big-picture* view of the risk distribution.

Table 5. Simple average, standard deviation, min and max values for Figure 3. For the local models, the average is across mortgages and then across FIs.

Deviations	μ_s	σ	min	max
Central at initial month	-0.027	0.119	-0.579	0.639
Central at final Month	-0.060	0.153	-0.980	0.929
FL at initial month	-0.056	0.104	-0.710	0.465
FL at final Month	-0.043	0.156	-0.968	0.971
FL to Central at initial month	0.028	0.061	-0.117	0.283
FL to Central at final month	-0.017	0.048	-0.474	0.765

In essence, there are significant benefits to collaboration through FL for smaller FIs. FIs whose

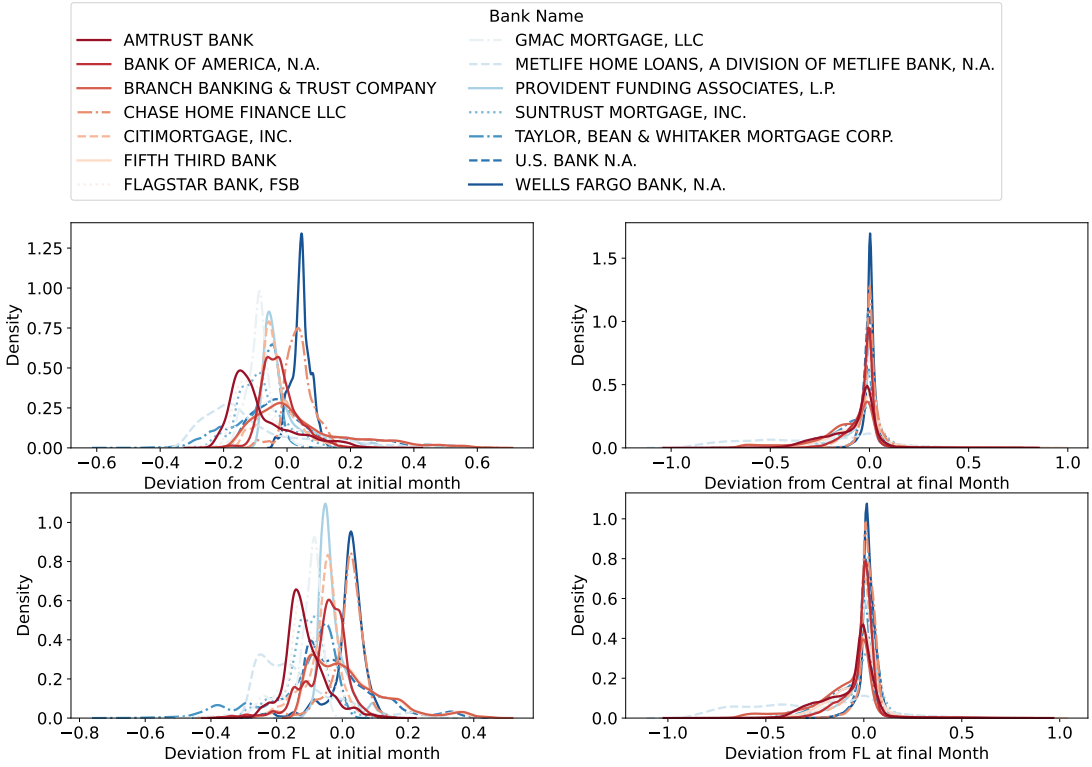

Figure 3. Relative deviation by the local models compared with central and FL scenarios. On the right, the comparison is between FL and Central.

datasets are large enough to approximate the overall variability in the market, in turn, do not significantly benefit from collaboration. Finally, we observe different estimations of default probability by each FIs. These differences vary during a mortgage. Therefore, each FI should individually assess the benefits of applying FL to the credit risk assessment of mortgages.

6. Limitations and Future Research

Our study is subject to two limitations: 1) in that we used only a subset of our dataset; 2) in that we only used the FMSFLL holding division; and 3) in that we worked only with mortgages with final status.

While the full FMSFLL dataset contains approximately 45 million unique mortgages spanning from 1999 to 2022, we used only those 250k that where active from 2006 to 2009. We focused on these years as they had a high number of defaults (Murphy, 2008). Time frames with less defaults, in turn, might lead to different results for the five scenarios. Further research should thus extend our study also to such other time frames.

Moreover, the FMSFLL dataset only includes mortgages that Freddie Mac has bought. In reality, the US FIs' mortage portfolio holdings contains more mortgages than just the ones Freddie Mac bought. We used the FMSFLL dataset because Freddie Mac is one of the major players in the US residential mortgage market. Furthermore, the portfolio holding divisions by each US FI in the sample subset are not arbitrary or random but based on true values and reflect mortgages that each US FI originated or had once held. However, the study could improve by including a hypothetical yet more realistic representation of each FI's mortgage portfolio holdings.

In addition, due to the high volume of loans originated over these years, we limited the number of mortgages by filtering out those with a final status so that the ML models could accurately predict them. The study could improve by modifying the models to handle a higher number of loans.

7. Conclusions

In this research paper, we present an FL prototype for the credit risk assessment of mortgages. We evaluate this prototype with an empirical dataset and a scenario analysis consisting of five scenarios. We find that smaller financial institutions could benefit significantly from collaboration with others through FL. On average,

our FL prototype improved accuracy, recall, precision, and F1 scores by 4.02, 1.84, 7.93, and 5.59 percentage points respectively.

The work presented in this paper contributes to the existing literature on the use of FL in financial services. In particular, our study contributes the following:

1. We present a prototype that uses FL for credit risk assessment of mortgages;

2. We demonstrate empirically the potential benefit of using FL for the credit risk assessment of mortgages.

Acknowledgements

We would like to thank to Daniel Howard Maclennan and Mohammad Ansarin for their valuable feedback.

This work has been supported by the European Union (EU) within its Horizon 2020 programme, project MDOT (Medical Device Obligations Taskforce), Grant agreement 814654, from the Kopernikus-project "SynErgie" by the German Federal Ministry of Education and Research (BMBF), from PayPal and the Luxembourg National Research Fund FNR (P17/IS/13342933/PayPal-FNR/Chair in DFS/Gilbert Fridgen), as well as from the Luxembourg National Research Fund (FNR) – FiReSpARX Project, ref. 14783405. The experiments presented in this paper were carried out using the HPC facilities of the University of Luxembourg (Varrette et al., 2014)– see hpc.uni.lu

References

Aïvodji, U. M., Gambs, S., & Martin, A. (2019). Iotfla : A secured and privacy-preserving smart home architecture implementing federated learning. *2019 IEEE Security and Privacy Workshops (SPW)*, 175–180. https://doi.org/10.1109/SPW.2019.00041

Altman, E. I. (2002). Managing credit risk: A challenge for the new millennium. *Economic Notes*, *31*(2), 201–214. https://doi.org/10.1111/1468-0300.00084

Bank, E. C. (2010). *Memorandum of understanding on the exchange of information among national central credit registers for the purpose of passing it on to reporting institutions* (tech. rep.).

Bansal, A., Kauffman, R. J., & Weitz, R. R. (1993). Comparing the modeling performance of regression and neural networks as data quality varies: A business value approach. *Journal of Management Information Systems*, *10*(1), 11–32. http://www.jstor.org/stable/40398029

Borgman, C. L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, *63*(6), 1059–1078. https://doi.org/10.1002/asi.22634

Chen, N., Ribeiro, B., & Chen, A. (2016). Financial credit risk assessment: A recent review. *Artificial Intelligence Review*, *45*. https://doi.org/10.1007/s10462-015-9434-x

Chollet, F. et al. (2015). Keras.

Ekbia, H., Mattioli, M., Kouper, I., Arave, G., Ghazi, A., Bowman, T., Suri, V., Tsou, A., Weingart, S., & Sugimoto, C. (2015). Big data, bigger dilemmas: A critical review. *Journal of the Association for Information Science and Technology*, *66*. https://doi.org/10.1002/asi.23294

Federal Reserve. (2021). Federal Reserve Economic Data (FRED) [Accessed: 2021-08-11]. https://fred.stlouisfed.org/

Freddie Mac. (2021a). Freddie Mac's House Price Index (FMHPI) [Accessed: 2021-08-11]. https://www.freddiemac.com/research/indices/house-price-index

Freddie Mac. (2021b). Single Family Loan-Level Dataset [Accessed: 2021-08-11]. http://www.freddiemac.com/research/datasets/sf_loanlevel_dataset.page

Galindo, J., & Tamayo, P. (2000). Credit risk assessment using statistical and machine learning: Basic methodology and risk modeling applications. *Computational Economics*, *15*, 107–43. https://doi.org/10.1023/A:1008699112516

Heitfield, E. (2009). Parameter uncertainty and the credit risk of collateralized debt obligations. *Risk Management*.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Israël, J.-M., Damia, V., Bonci, R., & Watfe, G. (2017). *The Analytical Credit Dataset - A magnifying glass for analysing credit in the euro area* (Occasional Paper Series No. 187). European Central Bank. https://ideas.repec.org/p/ecb/ecbops/2017187.html

Kaissis, G., Ziller, A., Passerat-Palmbach, J., Ryffel, T., Usynin, D., Trask, A., Lima, I., Mancuso, J., Jungmann, F., Steinborn, M.-M., Saleh, A., Makowski, M., Rueckert, D., & Braren, R. (2021). End-to-end privacy preserving deep learning on multi-institutional medical

imaging. *Nature Machine Intelligence*, 1–12. https://doi.org/10.1038/s42256-021-00337-8

Kearns, G. S., & Lederer, A. L. (2004). The impact of industry contextual factors on it focus and the use of it for competitive advantage. *Information & Management*, *41*(7), 899–919. https://doi.org/https://doi.org/10.1016/j.im.2003.08.018

Kroese, D. P., Brereton, T. J., Taimre, T., & Botev, Z. I. (2014). Why the monte carlo method is so important today. *Wiley Interdisciplinary Reviews: Computational Statistics*, *6*.

McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. *Artificial Intelligence and Statistics*, 1273–1282.

McMahan, H. B., Moore, E., Ramage, D., & y Arcas, B. A. (2016). Federated learning of deep networks using model averaging. *CoRR*, *abs/1602.05629*. http://arxiv.org/abs/1602.05629

Murphy, A. (2008). An analysis of the financial crisis of 2008: Causes and solutions. *An Analysis of the Financial Crisis of*.

on Banking Supervision, B. C. (2018). *Pillar 3 disclosure requirement - updated framework* (tech. rep.).

Redman, T. C. (1995). Improve data quality for competitive advantage [Copyright - Copyright Sloan Management Review Association, Alfred P. Sloan School of Management Winter 1995; Last updated - 2021-09-09; SubjectsTermNotLitGenreText - United States–US]. *Sloan management review*, *36*(2), 99. https://www.proquest.com/scholarly-journals/improve-data-quality-competitive-advantage/docview/224971040/se-2?accountid=41819

Robbins, H., & Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics*, 400–407.

Rosenblatt, M. (1956). Remarks on Some Nonparametric Estimates of a Density Function. *The Annals of Mathematical Statistics*, *27*(3), 832–837. https://doi.org/10.1214/aoms/1177728190

Saputra, Y. M., Hoang, D. T., Nguyen, D. N., Dutkiewicz, E., Mueck, M. D., & Srikanteswara, S. (2019). Energy demand prediction with federated learning for electric vehicle networks. *2019 IEEE Global Communications Conference (GLOBECOM)*, 1–6. https://doi.org/10.1109/GLOBECOM38437.2019.9013587

Sezer, O. B., Gudelek, M. U., & Ozbayoglu, A. M. (2020). Financial time series forecasting with deep learning : A systematic literature review: 2005–2019. *Applied Soft Computing*, *90*, 106181. https://doi.org/https://doi.org/10.1016/j.asoc.2020.106181

Shingi, G. (2020). A federated learning based approach for loan defaults prediction. *2020 International Conference on Data Mining Workshops (ICDMW)*, 362–368. https://doi.org/10.1109/ICDMW51313.2020.00057

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*(56), 1929–1958. http://jmlr.org/papers/v15/srivastava14a.html

Taleb, N. N. (2020). Statistical consequences of fat tails: Real world preasymptotics, epistemology, and applications. *arXiv preprint arXiv:2001.10488*.

Tensorflow federated. (2018). https://github.com/tensorflow/federated

United States Bureau of Labor Statistics. (2021). United States Bureau of Labor Statistics' Local Area Unemployment Statistics (LAUS) [Accessed: 2021-08-11]. https://www.bls.gov/lau/

Varrette, S., Bouvry, P., Cartiaux, H., & Georgatos, F. (2014). Management of an academic hpc cluster: The ul experience. *Proc. of the 2014 Intl. Conf. on High Performance Computing & Simulation (HPCS 2014)*, 959–967.

Walczak, S. (2001). An empirical analysis of data requirements for financial forecasting with neural networks. *Journal of Management Information Systems*, *17*(4), 203–222. https://doi.org/10.1080/07421222.2001.11045659

Yang, W., Zhang, Y., Ye, K., Li, L., & Xu, C. (2019). Ffd: A federated learning based method for credit card fraud detection. *BigData Congress*.

Zheng, W., Yan, L., Gou, C., & Wang, F.-Y. (2020). Federated meta-learning for fraudulent credit card detection. *IJCAI*.

Zuiderwijk, A., Janssen, M., Poulis, K., & van de Kaa, G. (2015). Open data for competitive advantage: Insights from open data use by companies. *Proceedings of the 16th Annual International Conference on Digital Government Research*, 79–88. https://doi.org/10.1145/2757401.2757411