Explainable Federated Learning for U.S. State-Level Financial Distress Modeling

Lorenzo Carta
cartal@rpi.edu
Rensselaer Polytechnic Institute
Troy, New York, USA

Fernando Spadea
spadef@rpi.edu
Rensselaer Polytechnic Institute
Troy, New York, USA

Oshani Seneviratne
senevo@rpi.edu
Rensselaer Polytechnic Institute
Troy, New York, USA

Abstract

We present the first application of federated learning (FL) to the U.S. National Financial Capability Study, introducing an interpretable framework for predicting consumer financial distress across all 50 states and the District of Columbia without centralizing sensitive data. Our cross-silo FL setup treats each state as a distinct data silo, simulating real-world governance in nationwide financial systems. Unlike prior work, our approach integrates two complementary explainable AI techniques to identify both global (nationwide) and local (state-specific) predictors of financial hardship, such as contact from debt collection agencies. We develop a machine learning model specifically suited for highly categorical, imbalanced survey data. This work delivers a scalable, regulation-compliant blueprint for early warning systems in finance, demonstrating how FL can power socially responsible AI applications in consumer credit risk and financial inclusion.

Our open source code repository is available at: https://github.com/brains-group/xfl-for-loan-eligibility.

CCS Concepts

• **Information systems** → *Collaborative and social computing systems and tools*; • **Computing methodologies** → **Neural networks**; **Distributed artificial intelligence**; • **Applied computing** → **Economics**.

Keywords

Federated Learning, Financial Distress Prediction, Explainable AI, Consumer Credit Risk, Financial Inclusion, Survey Data Analysis, Fairness in Finance, Responsible AI

1 Introduction

Financial distress, often culminating in contact from a debt collection agency (DCA), is a leading indicator of economic instability and a precursor to long-term credit damage, bankruptcy, and social hardship. In the U.S. alone, tens of millions of consumers experience this annually [4, 5]. For financial institutions, regulators, and social welfare organizations, early identification of individuals at risk is critical to enabling targeted, proactive interventions that mitigate harm and promote financial inclusion. While predictive modeling has shown great promise for anticipating such outcomes, real-world deployment in financial services remains constrained by two key factors: (1) the sensitive nature of personal financial data makes centralizing consumer-level information a regulatory and ethical challenge, and (2) black-box machine learning models undermine trust, particularly in high-stakes financial decision-making. Bridging these gaps requires techniques that safeguard data sovereignty while supporting interpretability.

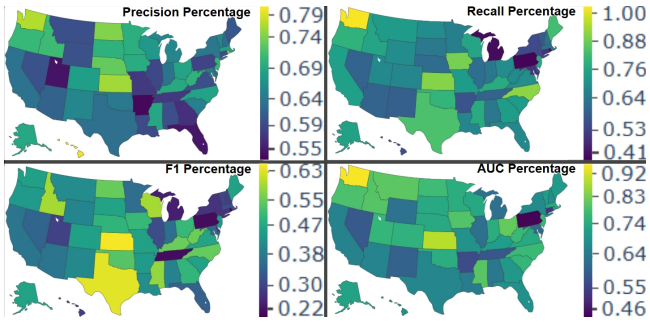


Figure 1: State-Level Evaluation of Federated Model Performance. Choropleth maps showing the performance of the global federated model across U.S. states using four evaluation metrics: precision, recall, F1-score, and AUC. Each state acts as an independent client in the cross-silo FL setup, revealing heterogeneity in model effectiveness and highlighting regional variation in financial distress prediction.

In this work, we present a novel application of federated learning (FL) to predict a key signal of consumer financial distress, whether a person has been contacted by a DCA, based on demographic and behavioral data from the U.S. National Financial Capability Study (NFCS) [18], a large-scale, representative survey managed by the FINRA Foundation. Rather than pooling data centrally, we simulate cross-silo FL across U.S. states, enabling localized state-level model development while capturing regional variation in financial hardship dynamics. To support actionable insights, we augment our approach with explainable AI (XAI) techniques, integrating SHAP [17] and Owen [11] values for richer explanations, identifying both nationwide and state-specific predictors of distress. Our findings show that our combined technique using FL and XAI surfaces interpretable patterns that can inform risk assessments, financial counseling strategies, and policy design. To our knowledge, this is the first such federated approach applied to financial distress prediction on survey data across U.S. states, combining highly interpretable models with cross-state granularity.

Although our study focuses on U.S. survey data, the framework is broadly applicable to global financial ecosystems, particularly in countries with strong privacy regulations (e.g., GDPR in the EU) or fragmented data institutional infrastructures. In regions where credit bureaus are nascent or incomplete, government surveys or mobile financial data could serve as analogs to power similar decentralized risk prediction tools. FL enables such models to be trained across institutions (e.g., microfinance lenders, public sector agencies) or geographies without violating data sovereignty or exposing individuals to surveillance.

1.1 Contributions

The primary contributions of this work are:

(1) **Design and evaluation of a novel FL framework for financial distress prediction.** We designed and evaluated a cross-silo FL pipeline to predict DCA risk using data from the NFCS [18]. The model was trained on data partitioned across the 50 U.S. states and the District of Columbia.
(2) **Development of a tailored architecture for handling imbalanced survey data in FL settings:** We developed a solution for the challenges of real-world survey data by combining an 8-layer Highway Network [6] with a targeted class weighting strategy.
(3) **Application of multi-faceted XAI for comprehensive model explanation:** We applied and compared SHAP [17] and Owen value [11] techniques to interpret model behavior at both global (nationwide) and local (state-specific) levels. This analysis identifies critical risk factors like age and income and demonstrates how distributing the total model prediction value among features in a way that respects their cooperative contributions offers a more intuitive explanation for interdependent, categorically binned features.

2 Related Work

Financial Distress Prediction and Credit Risk Modeling. Predicting financial distress, whether personal default or corporate bankruptcy, has been a long-standing focus in both finance and machine learning [2, 8]. Classical models such as logistic regression have dominated this space for decades due to their interpretability and alignment with financial theory, consistently identifying key risk factors like liquidity, solvency, and firm size [20]. In the consumer domain, datasets like the FINRA Foundation's NFCS have been used to study individual-level financial distress at a macro scale [18]. However, the the sensitivity and governance restrictions of the underlying financial data often limit its use to centralized or pre-aggregated analyses, preventing more granular, jurisdiction-aware modeling. This is especially limiting given that access to even more sensitive or fine-grained data, such as transaction histories, credit scores, or behavioral signals, at the local level could yield deeper insights into region-specific patterns of financial distress. These signals are often obscured or lost in centralized datasets due to anonymization, aggregation, and privacy-preserving constraints. Our work addresses this gap by applying FL to support fine-grained, cross-state modeling, while aligning with the data sovereignty and access control policies typical of large-scale aggregated datasets. Importantly, we demonstrate that even when using publicly available survey data like FINRA's, federated training yields good predictive performance, highlighting the potential of this approach for applications involving richer and more sensitive data sources.

Decentralized and Federated Learning in Finance. First introduced by McMahan et al. [15], FL is a decentralized machine learning paradigm that enables multiple parties to collaboratively train a model without sharing raw data. In the financial services domain, FL is seen as a promising solution to the "data silo" problem where sensitive customer data cannot be centrally pooled, for applications such as financial crime detection [1, 9] and credit scoring [7, 19]. Lee et al. (2023) developed an FL prototype for credit risk modeling across smaller financial institutions, enabling them to "compete by training in a cooperative fashion" on a shared mortgage dataset [13]. Similarly, Wang et al. (2024) proposed an FL method for credit scoring, addressing data imbalance concerns [19]. Our work builds on this foundation by applying cross-silo FL to a nation-wide social finance survey, treating each U.S. state as a silo, a scenario that, to our knowledge, has not been previously explored for financial distress prediction.

Explainable AI for Financial Distress Models. Regulatory and practical constraints in finance require that predictive models offer clear explanations, especially in high-stakes decisions like credit approvals, where lenders must justify adverse actions. As a result, there is growing interest in integrating XAI into financial risk modeling. One strategy involves using inherently interpretable models, such as linear classifiers or decision rules, that provide transparent reasoning at the cost of reduced complexity [20]. Alternatively, post-hoc explanation techniques like Shapley Additive Explanations (SHAP) [17] have been applied to complex models to retain both performance and interpretability. For example, Nguyen et al. (2024) used SHAP to uncover the key financial drivers (e.g., leverage, cash flow) behind corporate distress predictions made by neural networks and gradient-boosted trees [16]. Similarly, Aljunaid et al. [1] propose an explainable FL model for bank fraud detection. Additionally, Jovanovic et al. [7] integrate blockchain with explainable FL for credit risk modeling. We extend this line of work by applying explainable FL to the domain of personal financial distress, a setting where interpretability is essential for both regulators and consumers. Our contribution is distinct in combining SHAP with Owen values, enabling interpretation of both global (national) and local (state-level) model behavior. This dual-layer explainability also addresses the challenge of making complex FL models interpretable for each data-holding client, particularly when categorical features exhibit strong interdependencies.

3 Methodology

3.1 Dataset and Preprocessing

We used the FINRA Investor Education Foundation's 2021 NFCS [18] for our analysis. This comprehensive dataset is derived from a 126-question multiple choice survey that captures various statistics such as income, education, and financial education from more than 25,000 U.S. respondents, with approximately 500 respondents per state, plus the District of Columbia. The raw data, available in comma-separated-value (CSV) format categorically numbers each respondent's multiple-choice answer (i.e., labeled as 0 for the first choice, 1 for the second, etc.). We extract 11 of the 126 features, along with 1 designated learning objective. To facilitate binary classification, respondents who gave a non-binary response to the chosen learning objective were filtered out.

Prediction Goal. The specific learning objective extracted came from question G38 of the NFCS survey: "Have you been contacted by a debt collection agency in the past 12 months?" Respondents answered with a 'Yes', 'No', 'Don't know', and 'Prefer not to say'; the latter 2 of which were filtered out for binary classification. A

notable challenge with this prediction goal was the significant class imbalance: having over 78.5% of respondents answering 'No' and only 18% answering 'Yes'. This imbalance initially led to the model collapsing into always predicting 'No', resulting in high accuracy at the cost of trivial predictions; thus, the implementation of class weighting was crucial to address this issue and enable effective learning of the minority class.

Feature Selection. Of the 126 available features in the NFCS dataset, only 12 were selected for this model. This decision was driven by concerns regarding the large number of subjective questions and the potential for biased data within the full survey. Out of all the features, the 12 selected ones were shown to be the most objective and relevant features, focusing on less sensitive financial and demographic information such as income range and living arrangements. Additionally, they represent information that could be realistically and ethically queried from willing participants in alternative data collection scenarios.

Data Cleaning and Encoding. All features used in the model were categorical and derived from a questionnaire. During data loading, certain responses were flagged for exclusion (e.g., non-binary answers to a binary question) and removed from the model's learning structure.

3.2 FL Setup

Our model employs a cross-silo, Flower-FL setup [3]. The dataset is initially partitioned into a training and testing set, each holding 80% and 20% of the raw data respectively. The training data is further divided among 51 separate clients, each representing one of the 50 U.S. states and the District of Columbia. These clients independently train their local models and send their updates to a global server, which then utilizes Federated Averaging (FedAvg) [15] to aggregate the local updates and compute a single, server-wide model update. After the training is complete, the final global model's performance is evaluated against the partitioned testing data, and overall metrics are extracted. Furthermore, the extracted metrics are compared to a centralized model with identical attributes to demonstrate that FL achieves similar performance. Additionally, to show that the state-specific results with FL are better than without, we train local models as a baseline to compare against. The FL state-wide data partitioning is made possible due to the inclusion of the 'STATEQ' variable within the NFCS dataset, which numerically represents each respondent's state of origin alphabetically (e.g., 1 = Alabama, 2 = Alaska, etc).

3.3 Model Architecture

We employ an 8-layer highway network because its gated architecture enables training of deep neural networks by mitigating vanishing-gradient problems [6]. Our model design, inspired by highway networks' success in other domains (e.g. sequence labeling [14] and language modeling [12]), is well-suited to our tabular survey data with many categorical features. The gating mechanism allows the model to effectively learn complex feature interactions while skipping unnecessary transformations, which we found beneficial and yielded stronger results than standard architectures (demonstrating an approximate 9% increase in accuracy to a standard LeakyReLu model), given the high dimensionality and sparsity of the NFCS features.

Furthermore, to address the substantial class imbalance in the learning objective, where positive cases (i.e., individuals contacted by a DCA) are significantly underrepresented, class weighting was applied to prevent the model from biasing toward the majority class. The weighting factor was determined empirically after training the model multiple times using various different factors. After evaluating several configurations, we found that increasing the standard minority class weight by a factor of 1.1835 yielded the best balance between precision and recall, enabling the model to learn meaningful patterns from the minority class while maintaining stable overall performance. Our model implementation also incorporates various training tools, including the Adam optimizer [10], and early stopping.

3.4 Evaluation Metrics

To assess the overall training results, this model utilizes four distinct evaluation metrics:

- Precision, which denotes the accuracy of the model's positive predictions across the entire testing dataset.
- Recall, which represents the model's accuracy in identifying all relevant instances of the positive minority class (i.e., individuals who were contacted by a DCA).
- F1-Score, which provides the harmonic mean of precision and recall and offers a balanced measure of the model's accuracy.
- Area Under the Receiver Operating Characteristic Curve (AUC), which quantifies the model's ability to distinguish between the two classes.

In addition to these metrics, we also analyze the individual feature summaries used to discern feature importance, using SHapley Additive exPlanations (SHAP) [17] and Owen values [11]. These two methods, while related, employ distinct attribution mechanisms: SHAP based on marginal contribution across permutations, and Owen values based on structured coalition games, which can lead to both overlapping and divergent insights about feature importance.

4 Experimental Results

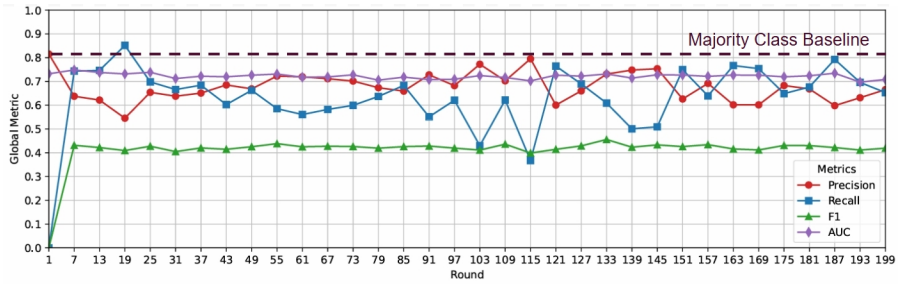
4.1 Global Model Performance

As can be seen in Figure 2a, the global model converges rapidly in the early rounds, with relatively stable performance thereafter, with only minor trade-offs between precision and recall, However, throughout the training, a seemingly negative correlation can be observed between precision and recall, which have an inversely proportional relationship. This trade-off arises because similar feature profiles may map to both positive and negative outcomes, so optimizing for one metric often comes at the cost of the other.

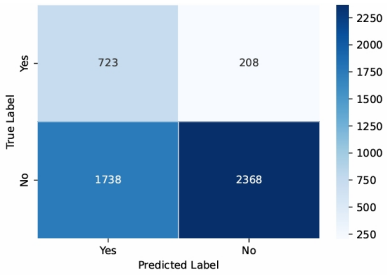
Figure 2b further reflects the challenge of class imbalance in the global model. Despite this, the model achieved 61.4% precision while maintaining 77.7% recall, indicating that it successfully identified most true positive cases while avoiding excessive false positives.

As mentioned in Section 3.1, the dataset was heavily imbalanced, with over 78.5% of samples belonging to the negative class.

As a consequence of the empirically-tuned class weighting, the model achieved a more balanced performance between overall



(a) Evolution of key evaluation metrics over 200 training rounds.



(b) Confusion matrix

Figure 2: Global performance metrics of the federated model over training rounds and final confusion matrix on the test set.

precision and recall, with some training rounds demonstrating over 65% accuracy in both metrics.

4.2 Local (State-Level) Model Performance

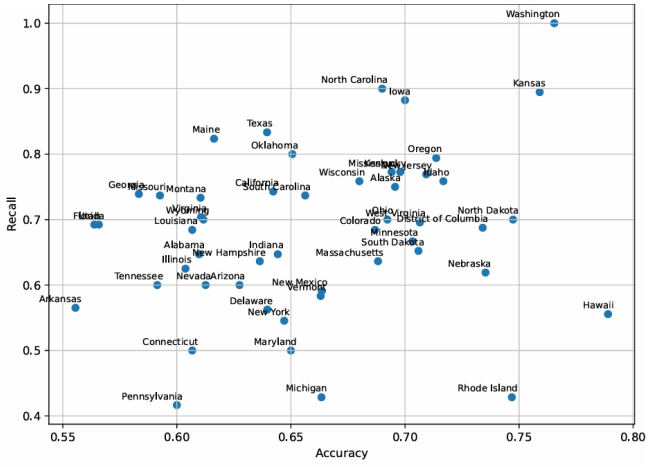
While the global federated model is trained collaboratively across all clients, its performance can vary significantly when evaluated independently on each state's local test data. Below, we analyze how well the shared global model generalizes to individual states.

Each point in Figure 3a represents a state's performance, highlighting variation in the model's ability to identify true positives (recall) versus its selectivity (precision). Notable outliers include Washington (high recall and precision) and Hawaii (high precision, low recall), reflecting diverse local patterns.

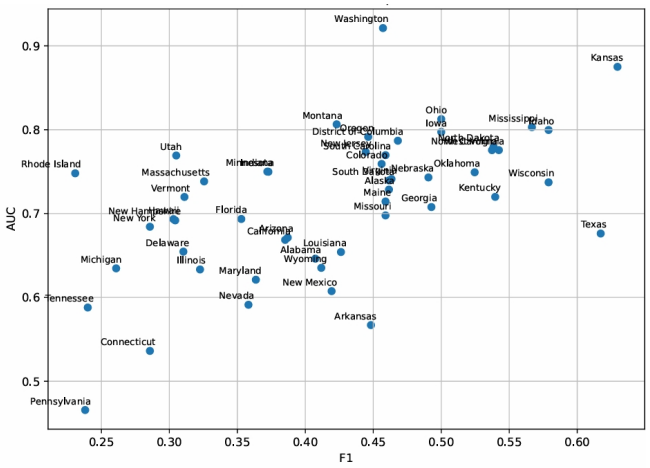
Contrary to initial expectations, the data in Figure 3a did not appear inversely proportional, but instead appears to be fairly clustered with a slight positive correlation in its distribution. Most states appear to reside within this cluster, excluding some outliers such as Washington (which consistently outperformed all other clients' recall), Hawaii, Rhode Island, Michigan, and Pennsylvania. In general, every client produced a favorable precision > 55%, and a decent recall accuracy: with only 3 states demonstrating suboptimal results. Overall, it is very interesting to observe such variation within the scatterplot, where some states can have very similar precision but vastly different recalls (e.g., Washington compared to Hawaii), illustrating that the two metrics do not necessarily correlate to each other.

Each point in Figure 3b indicates how well the model balances precision and recall (F1) and its overall discriminative ability (AUC). While most states cluster in a moderate performance range, states like Washington and Kansas show strong performance on both metrics. Figure 3b followed our initial expectations and demonstrates a strong proportional relationship between the two metrics. Overall, all clients except for Pennsylvania demonstrate an AUC above 50%, but an F1 score below 65%. The low F1 score highlights the inherent difficulty in precisely identifying all instances of DCA contact within an imbalanced real-world dataset, even with applied class weighting.

Furthermore, per-client metric results for the final round can also be observed in Figure 1, which shows the same results as in Figure 3a and Figure 3b, but in a more geographically evident manner, with states like Washington and New Jersey colored to represent a high precision and states like Texas and Mississippi showing a higher F1 score.



(a) State-Level Precision vs. Recall



(b) State-Level F1 Score vs. AUC

Figure 3: State-level scatterplots of model performance metrics across all 51 clients (50 states + District of Columbia).

4.3 Centralized and Local Model Comparison

To establish a benchmark for FL performance, we additionally trained a centralized version of the DCA prediction model. We incorporated the exact same parameters as those that were used in the FL setting, with the only change being the empirical adjustment of the minority class weight factor from 1.1835 to 1.182 in order to account for the un-intuitive differences in data modeling. As can be seen in Figure 4a, over the course of 200 epochs, the centralized model exhibited convergence behavior very closely aligned with that of the federated model, with its performance trends nearly identical across all 4 metrics. After computing the confusion matrix after the final epoch (Figure 4b), we found the centralized model achieved an F1 score of 42.4% and an AUC score of 74.1%. Compared to the federated model's scores of 42.2% and 71.4% in F1 and AUC respectively.

Furthermore, we trained a localized model across each individual state and took the average metric values for comparison. An important note to address is the fact that some states had very biased datasets, with a couple of states (NJ, ND) not even having any positive data points, which majorly impacted the average metric results. Overall, over 20 epochs of training, the average per-state centralized training model reached an average F1 score of 33.31%, and an AUC score of 69.96%, highlighting a much less favorable overall performance compared to FL, which was trained to be generalized to every state. These findings highlight that FL not only matches the accuracy benchmarks of centralized learning in single-model cases, but also surpasses it in aggregate performance while preserving data privacy and security.

4.4 Communication Cost Analysis

While our cross-silo FL framework preserves data locality, it necessarily incurs communication overhead due to repeated synchronization between the central server and participating clients. To manage this cost while ensuring representative learning, we adopt a partial participation strategy: in each communication round, 12 out of the 51 clients (approximately 25%) are randomly selected for training. This participation rate follows common practice in the FL literature [15], balancing the diversity of updates with reduced per-round bandwidth requirements. Since clients do not use the model for local tasks when they are not actively training, the server only needs to send the global model to the 12 selected clients and receive their locally updated models. Therefore, in each round, the server sends the global model to 12 clients and receives their locally updated models. Given that our 8-layer highway network occupies approximately 1,078 KB, each round requires 12,936 KB in upload and an equal amount in download, yielding a total of 25,873 KB per round. Over 200 training rounds, the total communication volume amounts to roughly 4.94 GB.

In comparison, a naïve approach in which all 51 clients receive and return models in every round would result in a communication cost of approximately 12.95 GB. Thus, our strategy achieves more than a 60% reduction in bandwidth consumption without compromising model performance or fairness across clients.

4.5 Feature Importance

To understand the contribution of each input feature to the model's predictions, we analyzed feature importance using both SHapley Additive exPlanations (SHAP) [17] and Owen values [11], as introduced in Section 3.4. Our analysis revealed varying degrees of influence among the 11 selected features. The `Gender/Age` bin feature was included twice, once grouped by gender and once by age, to separately assess the contribution of each component within the binned categorical structure and highlight how feature interactions influence model attribution across explanation methods in classifying whether a respondent was contacted by a DCA.

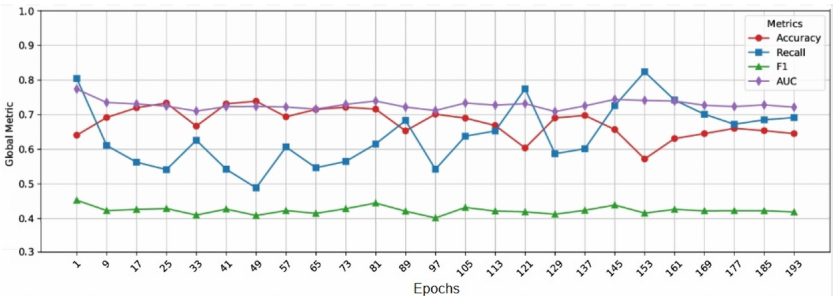
4.5.1 SHAP Analysis. As can be seen in Figure 5a, demographic features such as `Gender/Age` bins and income exhibit the highest impact on the model's output, while features like `Financial Education` and `Stock Investments` contribute less. Following the `Gender/Age` bins, `Annual Income` was found to be a highly influential feature, with a mean SHAP value of approximately 0.16. This finding intuitively aligns with expectations, as income is directly related to financial capacity and the ability to manage debt. `Employment Status` (around 0.105) and `Education` (around 0.11) also demonstrated considerable importance, reflecting their well-established correlations with financial stability and access to resources. Conversely, `Stock Investments` was identified as the feature with the lowest mean absolute SHAP value (approximately 0.02) among all analyzed features, indicating its minimal average contribution to the prediction.

However, as demonstrated by the large standard deviation values and surprising difference between `Gender/Age Bin (Age)` and `Gender/Age Bin (Gender)`, despite only differing categorically, we conclude that SHAP values alone are not a very good metric to use when evaluating the largely categorical features in the NFCS dataset.

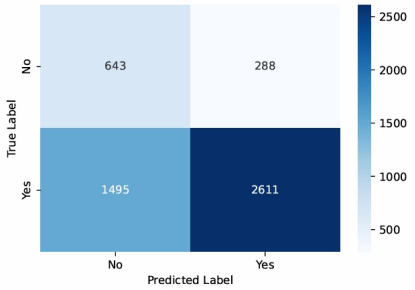
4.5.2 Owen Value Analysis. In contrast to SHAP values, Owen values offer a more consistent representation for categorical variables, showing stronger separation between key features (e.g., `Gender/Age`, `Employment`) and highlighting their relevance to model output, as can be seen in Figure 5b. This is because Owen values, unlike SHAP, are designed to distribute the total model prediction value among features in a way that respects their cooperative contributions, making them robust when dealing with interdependencies often found in categorical data.

A crucial observation from the Owen value analysis (Figure 5b) is the equal contribution of `Gender/Age Bin (Gender)` and `Gender/Age Bin (Age)`, both showing Owen values concentrated around 0.5. This equality in Owen values provides a more intuitive and consistent representation of their combined importance, especially considering that they are categorical permutations derived from the same underlying demographic information. The discrepancy highlights how Owen values can be a better indicator for this type of binned categorical data, providing a more balanced view of interdependent feature groups.

Beyond the `Gender/Age` bins, the Owen values further clarify the relative importance of other features. Features such as `Employment Status`, `Marital Status`, and `Financial Children` demonstrated significant impact on the model, showing denser distributions of



(a) Centralized model: evaluation metrics over 200 training epochs.



(b) Centralized model: Confusion Matrix.

Figure 4: Centralized Model Performance

Owen values further from zero. These features intuitively carry substantial weight, as an individual's employment stability, marital status, and financial dependents can be thought to directly influence their household's financial obligations and capacity to manage debt.

In contrast, features like `Stock Investments`, `Financial Education`, and `Health Insurance` consistently exhibited Owen values clustered closer to zero, indicating a considerably lower impact on the model's predictions. This is also intuitively explainable, because although these aspects relate to an individual's broader financial landscape, it is not directly visible or immediately actionable information for a DCA.

4.5.3 Bin-Level Owen Distributions. To gain a granular understanding of how each feature influences the model's predictions, we analyzed the Owen values for specific hand-picked individual feature bins, as presented in Figure 5c. These plots illustrate the distribution of Owen values across different categories within each feature.

Figure 5c uses bee-swarm feature summary to plot the Owen summary values for each specific category or bin within a feature (e.g., "Male 25-34" from the `Gender/Age` feature, "Retired" from `Employment`, etc.), with the most interesting categories being selected for visualization. The spread and concentration of points along each line, colored by feature value (with red indicating positive respondents within the category and blue representing negative ones), demonstrate the degree to which each category influences the model's output. An important aspect to note is that Owen values only compare the data with itself, and thus does not produce a value when there is nothing to compare to. For example, as observed in the 'Prefer not to say (Living Arrangements)' row in Figure 5c, categories that did not receive any positive (red) datapoints resulted in a zeroing of negative points, which could not be compared.

Consistent with the aggregate Owen value analysis in Figure 5b, the `Gender/Age` bins had the most pronounced influence on model predictions. Both male and female age categories displayed highly spread-out values across nearly all points in the bee-swarm plots in Figure 5c. This suggests that these specific demographic profiles are strong indicators of whether an individual is likely to be contacted by a DCA. In addition to the `Gender/Age` bins, employment status also emerged as a key predictor: retired individuals exhibited significantly lower Owen values compared to non-retired individuals, while full-timers showed a less distinct but visually separable trend.

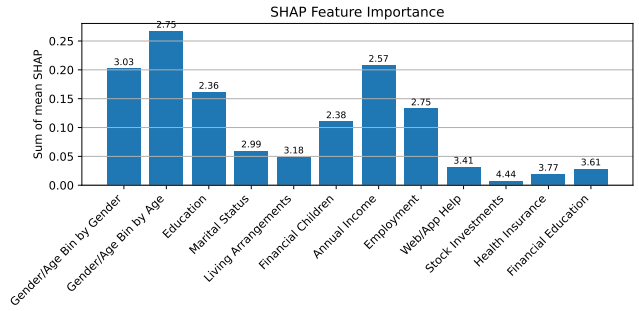
Marital status and health insurance status also had similar patterns of influence: certain marital status categories, such as divorced or widowed individuals, displayed similar bee-swarm distributions, with red points tending to show negative Owen values. Likewise, the presence of health insurance generally resulted, albeit with less effect, in a lower Owen score and vice versa.

4.5.4 Features with Minimal Predictive Value. In contrast, several features consistently had little to no impact on the model's predictions, with their Owen values clustered around zero. These include:

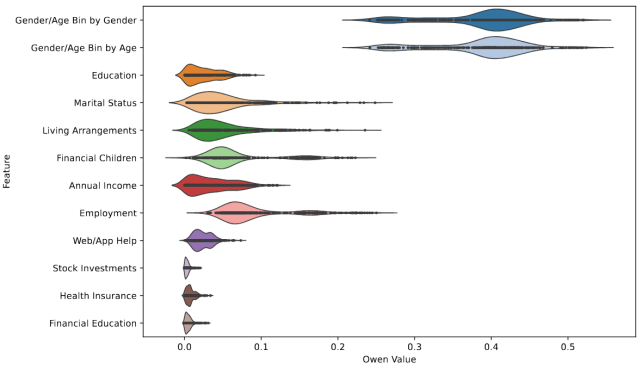
- Stock/Bond Investments: The wide spread of identically-colored points across both ends of the value spectrum suggests that this feature has minimal impact, which intuitively shows that DCAs generally lack access to individuals' investment data.
- Annual Income: Surprisingly, this feature did not strongly correlate with a person's likelihood of being contacted by a DCA. This indicates that a person's payment history and existing debt obligations are not closely tied to their earning potential.
- Financial Education: Despite some outliers, the general distribution of data points indicates that having formal financial education alone is not strongly associated with a reduced likelihood of being contacted by a DCA. In contrast, individuals who reported receiving financial help from dynamic sources such as websites or mobile apps exhibited greater separability in the model's predictions, suggesting that fresh, accessible financial guidance at the point of need tends to be more impactful than static or outdated knowledge acquired through past educational experiences.

4.6 Comparative Model Behavior Identifying Outlier and Representative States

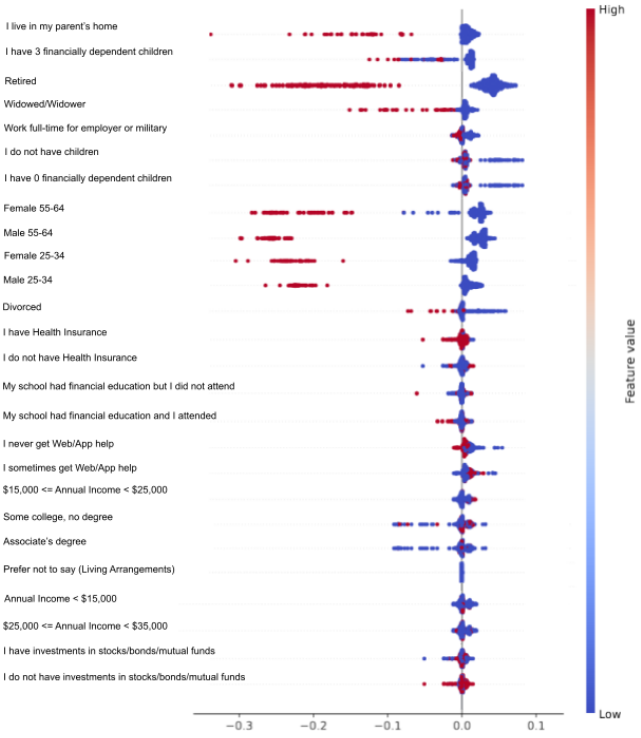
Throughout the training rounds the FL model was subjected to, a small subset of states consistently emerged as outliers in terms of their individual performance metrics. While the specific set of outlier states varied in every session (due to data heterogeneity and stochasticity of the training process), we chose to highlight two common outlier states, Washington and Hawaii, for which we calculated and plotted the individual state's performance metrics, as well as their SHAP feature summary. We also generated the individual metrics and feature summary for New York, a populated state whose final metrics closely mirrored those of the global model.



(a) SHAP value summary showing the average contribution of each input feature to the model's predictions.



(b) Owen value distribution plot, capturing the cooperative influence of each feature under interdependence.



(c) Grouped Owen value summary plot.

Figure 5: Global feature importance analysis using SHAP and Owen values.

Finally, we compare each plot to provide a comprehensive view of the state-level model behavior as can be seen in Figure 6.

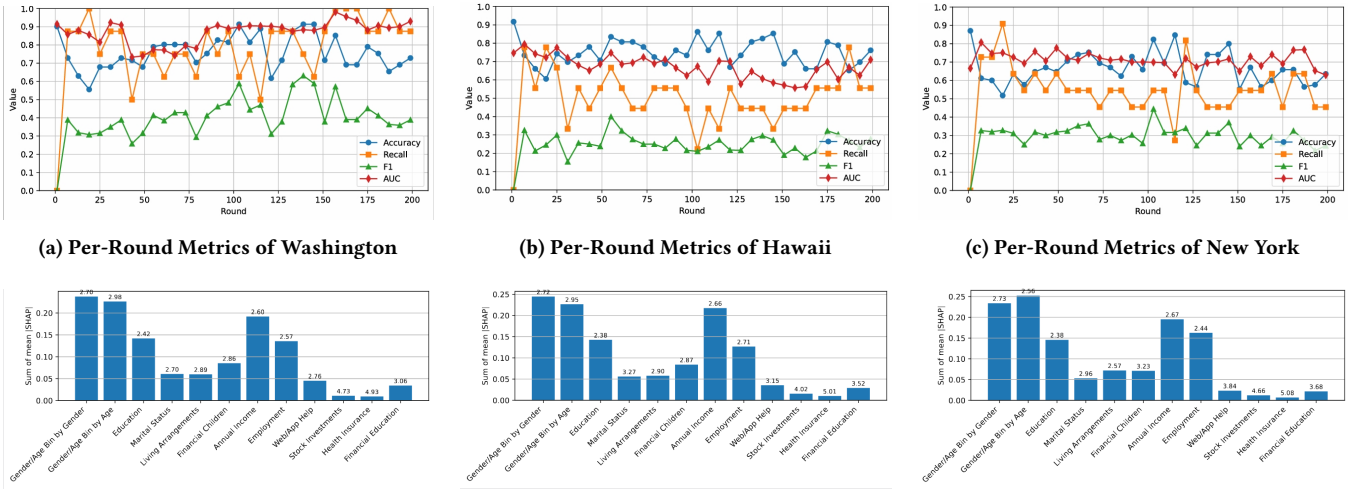
Looking at our first individual metrics in Figure 6a, Washington consistently demonstrated exceptional performance, outperforming the global model in nearly every training round. Washington's local model consistently achieved high scores across all evaluation metrics, often reaching values close to 0.9 or higher during its training. This sustained high performance suggests that the data distribution within Washington were particularly well aligned with the learning objective.

As shown in Figure 6d, we observe a local feature importance profile that broadly mirrors the global model's trends (Figure 5a). Excluding some differences like the overperformance of `Gender/Age Bin (Gender)` and the underperformance of the `Financial Children` feature. This consistency suggests that the underlying predictive relationships in Washington's data are similar to the national average, but perhaps with clearer signals or less noise, allowing the model to learn more effectively. The robust performance could also be attributed to a more distinct separation of classes within Washington's dataset, demographics, and financial characteristics that make the predictive task inherently simpler for the model.

In contrast to Washington, Hawaii exhibited a distinct performance signature characterized by a generally low recall value but high precision. Figure 6b illustrates this pattern, where precision often remains high (above 0.6), while recall frequently dips below 0.5, sometimes even falls to 0.2. This indicates that the general model, when applied to the individual state, tends to make many negative predictions that catch the majority of the respondents correctly but in turn misses out on a large amount of positive cases.

Hawaii's SHAP (Figure 6e), on the other hand, demonstrates a more skewed impact, with `Gender/Age Bin (Gender)` and `Living Arrangements` having a much higher impact compared to the global model, while other features like `Financial Children`, `Education`, and `Health Insurance` have a below-average impact. This is likely due to the client's lower recall score, which causes the model to exhibit less confidence in the features with a more nuanced separation; When a model struggles to identify all positive instances (low recall), it often becomes more reliant on the stronger, most unambiguous signals, and less on subtle contributions from features with smaller or more complex relationships, leading to a more polarized importance distribution.

As seen in Figure 6c and Figure 6f, The trends of a typical near-average state like New York generally follow the global model's behavior more closely than the outliers above, with only features like `Gender/Age Bin (Gender)`, `Living Arrangements`, and `Employment` having a noticeable difference. Looking back to the state's metrics, while showing surprising variation, they tend to oscillate around the global model's metrics (Figure 2a). The consistent dominance of `Gender/Age` features and `Annual Income` in populated states like New York reinforces the uniformity of these critical demographic and financial indicators in predicting DCA contact across significant populations, highlighting their broad relevance in the predictive task and thus allow the FL model to generalize effectively.



(a) Per-Round Metrics of Washington

(b) Per-Round Metrics of Hawaii

(c) Per-Round Metrics of New York

(d) Washington SHAP feature importance

(e) Hawaii SHAP feature importance

(f) New York SHAP feature importance

Figure 6: A comparison of per-round training metrics and SHAP feature importance across Washington, Hawaii, and New York.

4.7 Global vs. Local Insights Summary

A key strength of our FL framework lies in its ability to simultaneously reveal both global (nationwide) patterns and local (state-specific) nuances in financial distress predictors.

At the global level, the most influential predictors identified by both SHAP and Owen value analysis include demographic features such as gender-age groupings, annual income, and employment status. These variables consistently ranked among the top contributors to model output across most clients, indicating their generalizability as risk signals for consumer financial distress. For example, lower income brackets and unemployment status were associated with higher model-predicted likelihood of DCA contact, aligning with long-standing findings in credit risk literature [2, 8, 20].

In contrast, state-specific patterns exhibited meaningful deviations from these global trends. For instance, the model performed particularly well in Washington state, where both recall and precision were high, suggesting clearer separability of at-risk respondents. SHAP analysis for Washington mirrored global feature rankings but showed stronger emphasis on employment status and income, indicating that the local data distribution may accentuate these effects. Conversely, Hawaii demonstrated high precision but low recall, and its SHAP profile prioritized different features, such as living arrangements and gender, while underweighting financial dependents and education. These differences suggest that localized economic factors, cultural norms, or even variations in state DCA practices may influence model performance and feature relevance.

Furthermore, Owen values revealed more stable and interpretable patterns for binned categorical features at the state level. For example, grouped features like `Gender/Age Bin` showed balanced contributions across states with sufficient class diversity, whereas states with low within-group variance (e.g., rural states with fewer younger respondents) showed skewed distributions.

These observed differences open up several plausible policy- and behavior-driven interpretations. For example, Hawaii's lower recall may stem from its higher cost of living and relatively strong social safety nets, which could result in moderate financial strain without triggering formal DCA, making distress harder to detect through surface-level indicators like employment status or income. In contrast, Washington's strong performance may reflect clearer reporting patterns or more direct links between income levels and DCA contact. Meanwhile, in rural or lower-density states, features like financial dependents or employment status were stronger indicators, likely reflecting household-level vulnerability rather than cost-of-living constraints. These distinctions suggest that interventions, such as financial counseling or hardship assistance, are more effective when tailored to the region's dominant risk factors.

5 Conclusion

Early identification of at-risk individuals is crucial for enabling proactive interventions by financial institutions and social welfare organizations. However, assembling large, centralized datasets for this purpose is often impractical due to privacy concerns, regulatory barriers, and incomplete coverage, making FL a more viable and collaborative alternative for building robust predictive models. Furthermore, FL can be applied on a much larger scale than traditional centralized learning. In addition to addressing problems typically handled by centralized methods, FL can also be implemented across a variety of real-world decentralized systems — for instance, collaborative disease prediction among hospitals or genomic research across laboratories that are unwilling to share raw data directly. Given the high stakes of financial hardship, explainability is crucial for transparency and trust, especially in policy and aid decisions.

In this work, we demonstrate the viability of FL for predicting consumer financial distress using state-siloed data from the NFCS. We developed an 8-layer Highway Network architecture tailored to handle highly categorical, imbalanced real-world data, using class weighting to address the significant imbalance (only 18% of respondents had been contacted by a DCA), and showed through comparison that our results were on-par with the centralized model, while also outperforming localized ones. We also optimized model update communication, reducing costs from 12.954 GB to 4.935 GB over 200 training rounds. A key contribution of this work is the integration of XAI techniques, such as SHAP and Owen values, to

provide actionable insights. Our analysis identified critical predictors of financial distress, including Gender/Age and Annual Income. This interpretability fosters trust in financial decision-making and supports targeted interventions.

Our results highlight the value of FL in uncovering regional disparities in financial risk signals without requiring centralized data pooling. This enables policymakers, regulators, and financial institutions to tailor interventions, such as financial education programs, hardship assistance, or regulatory enforcement, based on state-level drivers while still benefiting from the statistical power of a national model. Our approach thus bridges the gap between macro-level analysis and micro-level action, offering a robust framework for equitable, data-driven financial inclusion strategies.

References

[1] Saif Khalifa Aljunaid, Saif Jasim Almheiri, Hussain Dawood, and Muhammad Adnan Khan. 2025. Secure and transparent banking: explainable AI-driven federated learning model for financial fraud detection. *Journal of Risk and Financial Management* 18, 4 (2025), 179.

[2] Edward I Altman, Małgorzata Iwanicz-Drozdowska, Erkki K Laitinen, and Arto Suvas. 2017. Financial distress prediction in an international context: A review and empirical analysis of Altman's Z-score model. *Journal of international financial management & accounting* 28, 2 (2017), 131–171.

[3] Daniel J Beutel, Taner Topal, Akhil Mathur, Xinchi Qiu, Javier Fernandez-Marques, Yan Gao, Lorenzo Sani, Kwing Hei Li, Titouan Parcollet, Pedro Porto Buarque De Gusmão, et al. 2020. Flower: A friendly federated learning research framework. *arXiv preprint arXiv:2007.14390* (2020).

[4] Alexander Carther, Caleb Quakenbush, and Signe-Mary McKernan. 2022. *The Number of Americans with Debt in Collections Fell during the Pandemic to 64 Million.* Urban Institute. https://www.urban.org/urban-wire/number-americans-debt-collections-fell-during-pandemic-64-million

[5] Julia Fonseca, Katherine Strair, and Basit Zafar. 2017. *Access to Credit and Financial Health: Evaluating the Impact of Debt Collection.* Technical Report 814. Federal Reserve Bank of New York. https://www.newyorkfed.org/research/staff_reports/sr814.html

[6] Klaus Greff, Rupesh K Srivastava, and Jürgen Schmidhuber. 2016. Highway and residual networks learn unrolled iterative estimation. *arXiv preprint arXiv:1612.07771* (2016).

[7] Zorka Jovanovic, Zhe Hou, Kamanashis Biswas, and Vallipuram Muthukkumarasamy. 2024. Robust integration of blockchain and explainable federated learning for automated credit scoring. *Computer Networks* 243 (2024), 110303.

[8] Kevin Keasey and Robert Watson. 2019. Financial distress prediction models: a review of their usefulness 1. *Risk Management* (2019), 35–48.

[9] Md Saikat Islam Khan, Aparna Gupta, Oshani Seneviratne, and Stacy Patterson. 2024. Fed-RD: Privacy-preserving federated learning for financial crime detection. In *2024 IEEE Symposium on Computational Intelligence for Financial Engineering and Economics (CIFEr)*. IEEE, 1–9.

[10] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[11] Konstandinos Kotsiopoulos, Alexey Miroshnikov, Khashayar Filom, and Arjun Ravi Kannan. 2023. Approximation of group explainers with coalition structure using Monte Carlo sampling on the product space of coalitions and features. *arXiv preprint arXiv:2303.10216* (2023).

[12] Gakuto Kurata, Bhuvana Ramabhadran, George Saon, and Abhinav Sethy. 2017. Language modeling with highway LSTM. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 244–251.

[13] Chul Min Lee, Joaquín Delgado Fernández, Sergio Potenciano Menci, Alexander Rieger, and Gilbert Fridgen. 2023. Federated Learning for Credit Risk Assessment.. In *HICSS*. 386–395.

[14] Liyuan Liu, Jingbo Shang, Xiang Ren, Frank Xu, Huan Gui, Jian Peng, and Jiawei Han. 2018. Empower sequence labeling with task-aware neural language model. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[15] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*. PMLR, 1273–1282.

[16] Minh Nguyen, Thanh Ngo, Bang Nguyen, and Sukhwa Hong. 2024. Using Machine Learning and Counterfactual Explanations for Financial Distress Prediction. *Available at SSRN 5032226* (2024).

[17] Mukund Sundararajan and Amir Najmi. 2020. The many Shapley values for model explanation. In *International conference on machine learning*. PMLR, 9269–9278.

[18] Olivia M Valdes, Gary R Mottola, Judy T Lin, and Christopher Bumcrot. 2024. The FINRA Foundation's National Financial Capability Study: Unpacking 12 years of data on US financial capability. *Journal of Financial Literacy and Wellbeing* 2, 1 (2024), 79–90.

[19] Zhongyi Wang, Jin Xiao, Lu Wang, and Jianrong Yao. 2024. A novel federated learning approach with knowledge transfer for credit scoring. *Decision Support Systems* 177 (2024), 114084.

[20] Kun Yang, Nikhil Krishnan, and Sanjeev R Kulkarni. 2025. *Financial Data Analysis with Robust Federated Logistic Regression.* Technical Report. arXiv. org.