FINAL REPORT FOR A HYBRID ATTENTION-BASED DEEP LEARNING MODEL TO IMPROVE CREDIT RISK PREDICTION

Arnav Talwani
Student# 1008233662
arnav.talwani@mail.utoronto.ca

Dhanishika Antony Jotheeswaran
Student# 1010131073
d.antony@mail.utoronto.ca

Dharuniga Antony Jotheeswaran
Student# 1010102112
dharuniga.antony@mail.utoronto.ca

Vishwajith Subhashraj
Student# 1010009688
vishwa.subhashraj@mail.utoronto.ca

ABSTRACT

This final report outlines the development of a Hybrid Attention and MLP Ensemble for credit risk prediction on structured financial data. The Lending Club dataset (855,969 records, 81 features) was processed to address missing values, class imbalance, and feature leakage, and then split into training, validation, and test sets (70, 15, 15) using stratified sampling. A logistic regression model was implemented as the baseline, achieving ROC-AUC scores of 0.813 (train), 0.817 (validation), and 0.812 (test), with F1 scores around 0.740 (test). Performance dropped in a reduced feature subset (ROC-AUC: 0.773, F1: 0.719) and on a new dataset (ROC-AUC: 0.722, F1: 0.663), revealing sensitivity to distribution shifts and missing key variables. The Hybrid Attention and MLP Ensemble outperformed logistic regression across all splits, achieving ROC-AUC scores of 0.829 (train), 0.830 (validation), 0.822 (test), 0.821 (subset), and 0.748 (new dataset), with F1 scores ranging from 0.689 to 0.778. Gains were largest in recall for the minority class, confirming stronger generalization under data imbalance. Attention heatmap analysis showed that purpose_debt_consolidation consistently interacted with other loan purposes and credit indicators, aligning with domain intuition and supporting interpretability. These findings highlight the ensemble's improved accuracy and transparency while underscoring the need for continuous monitoring and retraining to maintain robustness in deployment.

—-Total Pages: 10

1 INTRODUCTION

Credit risk prediction is critical for financial institutions to assess the likelihood of borrower default. Traditional models like logistic regression, decision trees, and ensemble methods use financial and demographic inputs (e.g., income, credit history, loan amount, debt-to-income ratio). However, these models often fail to capture non-linear relationships and complex feature interactions, especially with growing data complexity or limited information in underrepresented populations.

Accurate predictions can reduce default rates, improve loan allocation, and enhance financial stability. A reliable model can also reduce the time, effort, and subjectivity required by analysts, enabling them to focus on edge cases and improving operational efficiency.

To address these challenges, this project introduces a Hybrid Attention-Based Deep Learning Model that processes tabular loan applicant data to predict default risk. The model architecture (Figure 1) integrates feature embeddings, multi-layer perceptrons (MLPs), and attention mechanisms to learn both low and high-level patterns. The input is vectorized applicant data, which includes numerical data (income, loan amount) and categorical features (loan grade), encoded via ordinal one-hot encoding (see Section 3). Then the best models from training are ensembled with weighting based on their individual validation AUC score. The ensemble model outputs a probability between 0 and 1 indicating the likelihood of default.

This hybrid deep learning approach offers several advantages over traditional methods:

- Non-linear modeling: Deep networks can capture complex interactions (e.g., how income modifies the impact of loan grade).
- Attention mechanisms: Help prioritize relevant features or feature groups, boosting performance and interpretability.
- Automatic feature learning: Embedding layers and MLPs reduce reliance on manual feature engineering.
- Group-wise attention: Offers transparency by highlighting how feature categories (e.g., loan vs. borrower info) influence predictions.

This report summarizes our research, data preparation, the results of the primary model against the baseline, and an analysis and discussion of our findings.

2 Background & Related Work

Addo et al. (2018) applied logistic regression and decision trees to loan default prediction, selecting ten top features for simplicity. This concept creates abstraction, making it easier to train, but also risks losing nuanced patterns which we want to keep for model.

Manzo & Qiao (2021) combined deep learning with the unscented Kalman filter, achieving R^2 values above 95%. However, treating all features equally reduced interpretability, making real-world adoption difficult under regulatory constraints.

S&P Global (Vidovic & Yue, 2020) showed ML models like SVM outperform traditional methods by capturing nonlinear patterns. However, nonlinear kernels make them black boxes, offering little insight into how predictions are made, which is necessary for human interpretation when evaluating credit risk.

Shen et al. (2021) used LSTM with AdaBoost and an improved SMOTE method to address imbalanced datasets, improving minority class performance. However, temporal modeling is more complex and we seek to use undersampling instead to address class imbalance.

Emmanuel et al. (2024) stacked Random Forest, Gradient Boosting, and XGBoost with feature selection to achieve high AUCs. This sequential structure, however, reduced interpretability and risked discarding subtle feature interactions.

3 Data Processing

3.1 Primary Dataset

Our main primary dataset for our model is the Lending Club dataset (Mehta, 2020) which contains 855,969 records and 73 features, with severe class imbalance and multiple missing values.

Our first step was checking features with missing values. Those with a large amount of missing/null values (over 98%) were removed entirely from the dataset. Features with moderate missingness were either imputed or a missing indicator feature was created when missingness was informative. This is a new binary feature which represents if the value for the original feature was null. Other steps for feature cleaning and transformation include:

- Removing 27 features, including those with excessive missing data, and redundant categorical fields with high cardinality. Also removed an additional 22 due to a data leakage issue.
- Converting date columns into days since a chosen reference date (earliest date in the dataset).
- Ordinal encoding features with ordered values like `grade` and one-hot encoding categorical features with low cardinality like `purose`.
- Normalizing numerical features with PowerTransformer (Yeo-Johnson) for improved model convergence and using StandardScaler for highly varied data.

The final dataset now has 47 features including the target `default_ind`.

The original dataset was highly imbalanced so we applied random undersampling on the majority class to balance the dataset for training, which mitigates bias towards non-default predictions. After removing the target variable, the balanced dataset was split into training (70%), validation (15%), and test (15%) sets with stratified sampling to maintain class proportions.

Table 1 shows a sample record from the cleaned, processed training set after encoding and normalization. All features are numerical (scaled floats) or binary indicators, with no missing values.

Table 1: Cleaned data sample (first 3 rows with selected features)

Index	**term_60 months**	**emp_length**	**pymnt_plan_y**	**loan_amnt**	**dti**	**default_ind**
0	0	1.03	0	-1.157	0.547	0
1	1	-1.69	0	-1.453	-0.983	1
2	0	1.03	0	-1.465	-0.540	0

3.2 Subset Dataset with Removed Features

To assess model robustness when key predictive features are unavailable, we created a subset dataset derived from the original test set by removing loan grade-related features (`grade` and `sub_grade`). These features are known to be highly predictive but may not always be available in practice, especially for new loan applicants.

The subset dataset was constructed after splitting the data into training, validation, and testing. Then we selected all samples from the original test set while dropping loan grade features entirely from the feature matrix. Then aligned columns to match the original training feature space by reindexing and filling missing columns with zeros to maintain input dimensionality. Finally we used the same scaling as the test set for the main model.

3.3 New Dataset

An entirely new dataset of loan applications was obtained from Kaggle (Tse, 2020) with 32 581 records and 12 features before processing. This dataset also has class imbalance and some missing values. The processing for the dataset was the same as the original dataset except for some additional key preprocessing for alignment:

- The dataset contained some the same features as the original dataset but with different names which were changed to align.
- Columns were reindexed to match the original feature set; any missing columns were added with zeros.
- Categorical features were encoded using the same mappings.
- Normalization and undersampling was similar to the original dataset.

The final dataset now has 2 features including the target `default_ind`. No retraining or fine-tuning was done and the entire dataset was used as a test set. This approach ensures no data leakage and provides a rigorous test of model generalization to truly unseen real-world data.

4 Architecture

The proposed Hybrid Attention Network predicts credit risk from tabular financial data by combining feature embeddings, multi-layer perceptrons (MLPs), and attention mechanisms to weight features like loan amount, term, and interest rate. Raw inputs pass through two parallel paths: an attention branch for capturing critical relationships and an MLP branch for numerical features (ex. annual income) with ReLU and dropout. They then merge into a sigmoid output for default probability.

Planned experiments include varying MLP depth (3 vs. 5 layers) and an ablation study to measure attention's impact, targeting a +5% AUC gain over simpler models. A logistic regression baseline

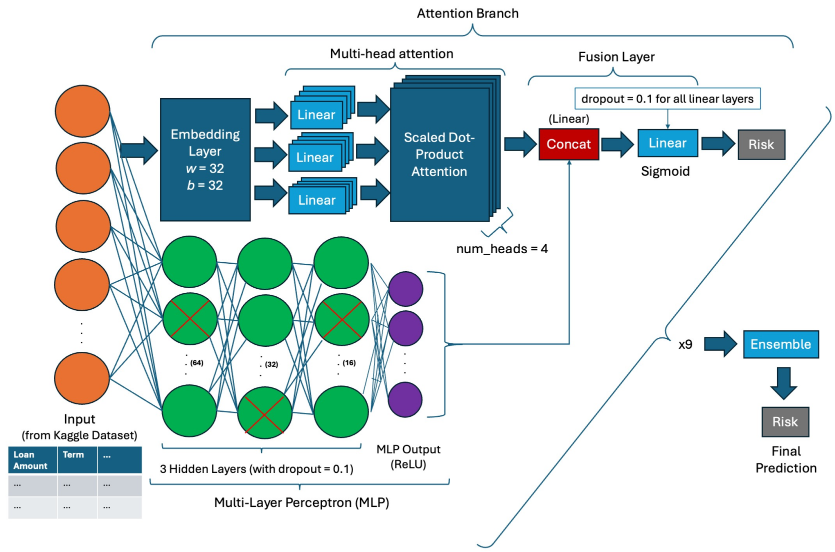

Figure 1: Detailed architecture of the model showing dataset input, the Hybrid Attention Network with parallel Multi-Head Attention and Multi-Layer Perceptron streams, and the output credit risk probability with ensembling. "Linear" layers in the attention stream represent multiple fully-connected layers, abstracted for simplicity.

is used for comparison, the industry standard for credit scoring. To boost accuracy and stability, multiple models with different hyperparameters are trained via random search, with the best models ensembled with weighting proportional to its validation performance. This approach helps to smooth out individual model biases and variance, resulting in more stable and accurate credit risk predictions.

Performance is evaluated using AUC-ROC (targeting > 0.85) and precision-recall balance to minimize false approvals while capturing true defaults. Attention weights also enhance transparency, enabling lenders to see which features drive predictions.

4.1 Primary Model

As mentioned above, the primary model design is that of a Hybrid Attention Network consisting of both a multi-head attention stream reminiscent of that employed in transformer models as well as a standard multi-layer perceptron (MLP) stream that form the basis of all deep learning models. Figure 1 shows the detailed architecture of the model, including an example of the number of layers in each branch, the number of weights, and other hyperparameters which will be tuned in training. In summary:

- The attention branch includes an embedding layer (size treated as a hyperparameter) and a multi-head attention layer, which consists of three linear projection streams (V, K, and Q) followed by a scaled dot-product attention mechanism. The number of attention heads is also a hyperparameter.
- The MLP stream has a configurable number of hidden layers, each with corresponding bias terms; the number of layers and their sizes are treated as hyperparameters.
- The fusion layer, which concatenates the outputs from both branches and passes them through a fully connected layer followed by a sigmoid output layer, contains $2048 \times \texttt{input\_dim} + 1153$ parameters in the default configuration.
- The model outputs a single probability representing the predicted risk of a candidate defaulting on their credit.

5 Baseline Model

We selected Logistic Regression (LR) as our baseline model due to its interpretability, simplicity, and common use in credit risk assessment (Stojiljković, 2024). As a binary classifier, LR produces probabilistic outputs indicating the model's confidence in each prediction. These probabilities support evaluation using the Receiver Operating Characteristic – Area Under the Curve (ROC-AUC), which measures the model's ability to distinguish between classes across all thresholds (scikit-learn developers, 2024). A higher ROC-AUC indicates better separability. LR also provides transparency through its learned coefficients, revealing which features most influence loan default predictions.

The baseline model was configured as follows:

- `penalty='l2'`: L2 regularization- penalizes large coefficients to prevent overfitting, default
- `C=1.0`: inverse strength of regularization- balances between under- and overfitting, default
- `class_weight='balanced'`: mitigates class imbalance by automatically adjusting weights inversely proportional to class frequencies to ensure minority classes aren't ignored
- `solver='liblinear'`: algorithm used to fit model, for small/medium datasets, default
- `max_iter=1000`: ensures convergence during training

`default_ind` is the binary target variable where 1 indicates a defaulted loan and 0 otherwise.

```
from sklearn.linear_model import LogisticRegression

logreg_model = LogisticRegression(class_weight='balanced', max_iter=1000)
logreg_model.fit(X_train, y_train)
```

Listing 1: LR model configuration and training (Stojiljković, 2024), full code in GitHub repository

6 Quantitative Results

We evaluated the final ensemble model against the logistic regression baseline model using three key metrics, ROC-AUC, accuracy, and F1 score (macro), to provide a comprehensive assessment of performance from different perspectives.

- **ROC-AUC** measures the model's ability to distinguish defaulters from non-defaulters across all classification thresholds. As a threshold-independent metric, it is especially useful for imbalanced datasets by reflecting overall ranking quality.
- **Accuracy** represents the overall proportion of correct predictions. While intuitive, accuracy can be misleading when classes are imbalanced, as it may overstate performance on the majority class.
- **F1 Score (Macro)** is the harmonic mean of precision and recall averaged equally across classes. It balances the trade-off between false positives and false negatives, providing a better measure of performance on both classes than accuracy alone.

Overall, the final ensemble model consistently outperforms the baseline logistic regression model with a $> 1\%$ gain across all datasets and metrics, except for the accuracy of the new test dataset. There was also a $\sim 5\%$ increase in the ensemble model for all metrics of the test subset, indicating improved generalization and performance with limited information (Table 2). The improvements in ROC-AUC indicate better class discrimination, while gains in F1 score, which generally saw the greatest increase, reflect a more balanced handling of false positives and negatives. Accuracy improvements confirm more correct predictions overall.

The lower performance on the new test dataset reflects distributional differences compared to training data, such as changes in borrower profiles or economic conditions, which can challenge model generalization. This highlights the importance of continuous monitoring and potential retraining to maintain robustness in real-world deployment.

Table 2: Performance of Baseline (Logistic Regression) vs. Ensemble Model

Dataset	**ROC-AUC**		**Accuracy**		**F1 (Macro)**	
	Baseline	Ensemble	Baseline	Ensemble	Baseline	Ensemble
Train	0.8128	0.8288 **(+0.016)**	0.7384	0.7587 **(+0.020)**	0.7391	0.7780 **(+0.019)**
Validation	0.8167	0.8296 **(+0.013)**	0.7406	0.7593 **(+0.019)**	0.7406	0.7779 **(+0.037)**
Test	0.8120	0.8223 **(+0.010)**	0.7383	0.7540 **(+0.016)**	0.7394	0.7737 **(+0.034)**
Test Subset	0.7732	0.8213 **(+0.048)**	0.7030	0.7553 **(+0.052)**	0.7192	0.7738 **(+0.055)**
New Data Test	0.7216	0.7477 **(+0.026)**	0.6391	0.6469 **(+0.008)**	0.6632	0.6894 **(+0.026)**

7 Qualitative Results

7.1 Learning Curves (ROC/AUC)

Similar to the previous trials in the progress report, the updated model was again evaluated on both full training, test, and validation sets as well as on a small class of data known as the subset test set (which is the same subset as defined in the progress report). This time, as mentioned earlier in the report, the model was tested on a new set as well, and Figure 2 depicts the learning curves of the model on all of these sets. As was for the iteration of the model seen in the progress report, the training and validation curves rise sharply toward the top-left, indicating strong separation and good generalization without overfitting, and the full test ROC curve closely matches this, confirming model robustness on unseen data. The model maintains the same sharper edge on the subset test; however, this time it does falter a bit when exposed to an entirely new set, with the trend not being as sharp as with the other sets. Potential reasons for this are explored further below.

7.2 Individual Sample Cases

Some sample cases have also been provided in Tables 3 and 4 below. While it is true that the model does generalize well and does not overfit, it does seem to have certain unstable areas based on these sample cases that generate false positives and negatives.

Specifically, the model seems to have set for itself a cutoff probability percentage between what is and isn't considered a default range, with that cutoff being at or around 50%. This is evident in cases such as Loans #2, #8, and #10, for which the model calculates a default probability of just over 50% and generates false positive results, as well as for Loans #7113, #7129, and #7133, for which the model calculates a default probability of just below 50% and generates false negative results.

Additionally, these probability calculations themselves seem to be heavily influenced by a few key categories, which, while in isolation remain good generalizations by both human knowledge and over the entire dataset, lead to false predictions in these complex cases. For example, the model seems to often overweigh high interest rates and loan amounts in its calculations, such that in situations where the additional factors (for example, annual income supporting a high loan amount, or a low loan amount with high interest rate) in sum may balance out these individual traits, the model will still tend towards a prediction characteristic of only that individual feature (e.g., for the previously described examples in this sentence, the model would likely predict a default when there was none). This is especially true for cases such as Loans #5 and #12, where the interest rate and loan amount were already quite high, and the borrower's income wasn't that low (in the case of Loan #5), or the loan amount and interest rate were a bit low but their annual income was especially low; however, they owned their home and thus had a significant detractor from default that the model ignored (in the case of Loan #12).

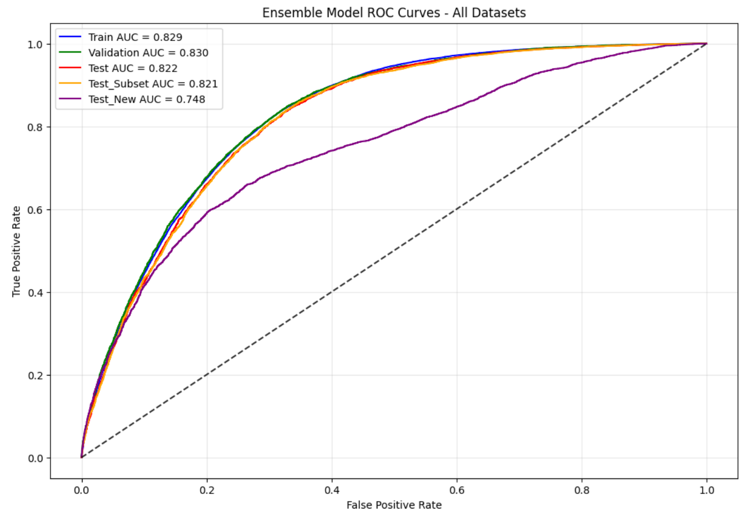

Figure 2: ROC Curves for train, validation, test, test subset, and test new dataset plotted on the same axes.

Table 3: Loan Data Sample — Features

Loan #	**Annual Income**	**Employ. Length**	**Grade**	**Loan Amount**	**Interest Rate**	**Home Ownership**	**Purpose**
1	0.1240	1.5380	1.3633	2.5373	1.4964	RENT	other
2	-3.2664	0.3180	0.0753	-2.3953	0.2649	OWN	educational
3	-3.2664	-1.1143	0.8333	-0.5371	0.7456	MORTGAGE	medical
4	0.3075	0.0175	0.8333	2.5373	1.3207	RENT	medical
5	-0.0194	1.0936	0.8333	2.5373	1.0965	RENT	medical
8	0.6326	0.3179	0.0753	2.5373	0.2560	RENT	medical
12	-3.1859	-0.6847	-1.2505	-0.7967	-0.5563	OWN	other
7113	-0.1132	0.5947	0.0754	-0.8831	0.0990	MORTGAGE	educational
7129	-0.1063	0.0175	0.0753	-0.1688	0.4298	MORTGAGE	educational
7133	0.2941	-1.6370	0.0753	-0.2084	0.1358	RENT	small business

7.3 RESULTS

Attention heatmap analysis shows that `purpose_debt_consolidation` consistently interacts with other loan purposes and credit indicators, including `purpose_educational`, `purpose_small_business`, `revol_bal`, `open_acc`, and `pub_rec`. This indicates the model is prioritizing relationships between a borrower's debt obligations, credit activity, and loan purpose. These interactions align with traditional domain knowledge, thereby supporting the

Table 4: Loan Data Sample — Prediction Outputs

Loan #	True Label	Ensemble Probability	Predicted Label	Confidence Gap	Classification
1	0	0.6947	1	0.1947	FP
2	0	0.5663	1	0.0663	FP
3	0	0.1845	0	0.3155	TN
4	0	0.5989	1	0.0989	FP
5	0	0.7922	1	0.2922	FP
8	0	0.5511	1	0.0511	FP
12	0	0.8782	1	0.3782	FP
7113	1	0.4757	0	0.0243	FN
7129	1	0.417	0	0.083	FN
7133	1	0.4563	0	0.0437	FN

8 Evaluation of Model on New Data

To assess our models' predictive ability and generalizability, we evaluated both our baseline model and our main model on multiple test datasets:

1. A holdout test set from our original dataset (15%) — represents data from the same distribution as training and validation.
2. A test subset using the test set above with key features (loan grades) removed — designed to evaluate model predictability when certain important information is missing.
3. An entirely new dataset — comprised of an entire dataset that was never used during model development with only a few feature (7 vs. 24 before encoding) simulating real-world unseen data with limited information.

To ensure valid comparison and avoid data leakage we processed our new dataset to align with our trained model as described in Data Processing (3). We also used consistent metrics (AUC-ROC, accuracy, precision, recall, and F1-score) for fair comparison.

8.1 Results

The logistic regression baseline performs well on the original data but drops on the test subset with missing loan grade features and on the new dataset, likely due to distribution shifts and missing key variables (Table 2). The ensemble consistently outperforms logistic regression across all datasets, especially in positive-class recall, and shows greater robustness to missing features and shifting distributions, with smaller performance losses on both the subset and new dataset. Performance drops on the new dataset for both models underscore real-world challenges with data distribution shifts and limited applicant information. Testing across varying missingness levels and data shifts confirms the ensemble model's improved generalizability and highlights common deployment challenges. interpretability of the ensemble model.

9 Discussion

The hybrid attention ensemble shows clear performance gains over logistic regression, particularly when predictive features are missing, which supports its ability to capture deeper interactions. Attention heatmaps confirm that the model emphasizes meaningful relationships between debt consolidation and credit indicators, reinforcing interpretability. While results on the new dataset were weaker, this reflects distributional shifts in borrower behavior and highlights the limits of static models. These findings underline the need for continuous monitoring, threshold calibration, and retraining to maintain robustness in deployment. Overall, the model balances predictive strength with transparency, but long-term reliability will depend on adapting to changing data.

In cases where the model assigned an arbitrary probability cutoff, the confidence gap on these predictions is low, which means that the model was equally drawn to both predictions and was only slightly more confident in the final (incorrect) one, and thus it may have been nudged in that direction by this arbitrary cutoff that it may have learned itself through the data.

The model's feature biases may have even played a role in nudging the probability calculations for the uncertain cases either just over or just under 50%, as seen in Loan #2, where while the loan amount is low, the annual income is even lower, and thus the model may be employing that to push the probability of default just over 50%, even though other factors, such as the fact that the borrower owns their home or has a higher than average credit score are clearly larger factors in the real-life result of them not defaulting. This then, is also indicative of a potential systemic bias in the model towards these features, and future iterations of the model may require corrections to avoid these biases.

These biases also seem to be in line with the feature-attention analysis that was performed on the model during the progress report, validating those results and further emphasizing the importance of architectural corrections necessary to mitigate these biases, alongside the very real fact that false negatives, such as those predicted by the model, prove costly to businesses who require default predictions for their loan risk analyses.

10 Ethical Considerations

The deployment of our credit risk prediction model raises several ethical concerns:

- **Deployment Bias:** Using the model outside its intended scope — for example, in employment screening, insurance underwriting, or marketing — may result in unfair or harmful decisions. The model is specifically designed for credit risk assessment and should not be repurposed without rigorous validation.
- **Training Data Limitations:** The Lending Club dataset contains historical biases related to socioeconomic factors, race, and geography, such as higher interest rates for low-income ZIP codes or ethnic enclaves victims of redlining to propagate. These biases can be inadvertently learned by the model, perpetuating unfair treatment of certain groups.
- **Fairness:** The model may exhibit disparate impacts on different demographic groups due to biased training data or feature correlations, raising concerns about equity and discrimination. Fairness-aware evaluation and mitigation strategies are necessary to minimize such risks.
- **Class Imbalance and Missing Data:** Although techniques like undersampling and imputation were used to address imbalanced classes and missing values, these approaches may limit the model's generalizability, especially for underrepresented populations or in cases with incomplete applicant information.
- **Robustness to Incomplete Data:** The model's performance decreases when key features are missing, highlighting vulnerability in real-world scenarios where data may be incomplete or unreliable.
- **Impact:** False negatives (missed defaults) result in financial losses for lenders, while false positives (unfair denials) may unjustly exclude qualified applicants. This can be addressed by using SHAP values to provide clear explanations for denials (e.g. high debt-to-income ratio).

To address these concerns, deployment must be accompanied by transparency about the model's limitations, ongoing fairness monitoring, and human oversight to prevent misuse and unintended harm.

References

Peter Martey Addo, Dominique Guegan, and Bertrand Hassani. Credit risk analysis using machine and deep learning models. *Risks*, 6(2), 2018. ISSN 2227-9091. doi: 10.3390/risks6020038. URL https://www.mdpi.com/2227-9091/6/2/38.

Ileberi Emmanuel, Yanxia Sun, and Zenghui Wang. A machine learning-based credit risk prediction engine system using a stacked classifier and a filter-based feature selection method. *Journal of Big Data*, 11(1):23, 2024. ISSN 2196-1115. doi: 10.1186/s40537-024-00882-0. URL https://doi.org/10.1186/s40537-024-00882-0.

Gerardo Manzo and Xiao Qiao. Deep learning credit risk modeling. *The Journal of Fixed Income*, 31:jfi.2021.1.121, 08 2021. doi: 10.3905/jfi.2021.1.121. URL https://openreview.net/pdf?id=F88KOHKRY3.

Ramesh Mehta. Credit risk analysis dataset. https://www.kaggle.com/datasets/rameshmehta/credit-risk-analysis, 2020. Accessed: 2025-06-12.

scikit-learn developers. roc_auc_score. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_auc_score.html, 2024. Accessed: 2025-07-10.

Feng Shen, Xingchao Zhao, Gang Kou, and Fawaz E. Alsaadi. A new deep learning ensemble credit risk evaluation model with an improved synthetic minority oversampling technique. *Applied Soft Computing*, 98:106852, 2021. ISSN 1568-4946. doi: https://doi.org/10.1016/j.asoc.2020.106852. URL https://www.sciencedirect.com/science/article/pii/S1568494620307900.

Mirko Stojiljković. Logistic regression in python. https://realpython.com/logistic-regression-python/, 2024. Accessed: 2025-07-10.

Lao Tse. Credit risk dataset. https://www.kaggle.com/datasets/laotse/credit-risk-dataset, 2020. Accessed: 2025-07-09.

Luka Vidovic and Lei Yue. Machine learning and credit risk modelling. Technical report, S&P Global, 2020. URL https://www.spglobal.com/marketintelligence/en/documents/machine_learning_and_credit_risk_modelling_november_2020.pdf.