'InvisiFraud': Fraud Intelligence System

Analytics Capstone Project

Sai Sandip Bhere[1], Krystyn Gutu[2]

Received: December 10, 2024

Modern financial systems face a significant problem in detecting fraud, as the use of credit cards and internet payments increases the likelihood of fraudulent behavior. Machine learning (ML) and deep learning (DL) are crucial techniques for more dynamic detection because traditional rule-based systems find it difficult to keep up with changing fraud strategies.

The InvisiFraud system uses CNN-based deep learning models and supervised machine learning to accurately detect fraudulent transactions. The system investigates undersampling (lowering the majority class) and oversampling (raising the minority class) in order to address the problem of class imbalance, where fraudulent transactions are substantially fewer than valid ones. In terms of recall, precision, and F1-score, oversampling regularly performs better than undersampling.

Accuracy, precision, recall, and F1-score are used to assess machine learning models, such as Random Forest, Logistic Regression, K-Nearest Neighbors (KNN), and Decision Trees. The CNN improves the ability to identify intricate fraud patterns. In order to optimize fraud detection, balance precision and recall, and suggest future enhancements, this study examines the methodology, results, and limitations of InvisiFraud. The system shows how resampling, strong machine learning models, and suitable metrics may greatly improve the precision and effectiveness of fraud detection in practical financial applications.

1. Introduction

The necessity for efficient fraud detection systems has increased due to the rise in online financial transactions. The complexity and number of transactions increase with the growth of digital payments, e-commerce, and cashless banking, giving cybercriminals more possibilities to take advantage of weaknesses. For financial institutions, this presents a double challenge: providing seamless user experiences while guaranteeing fraud prevention.

Fraudulent transactions harm customer trust and brand reputation in addition to causing financial losses. Conventional rule-based systems, which depend on predetermined "if-else" conditions, find it difficult to adjust to complex, constantly changing fraud strategies. These static models lose their effectiveness as fraudsters come up with new techniques, underscoring

[1]Student at Seidenberg School of Computer Science & Information Systems, Pace University, NY,
[2]Professor at Student at Seidenberg School of Computer Science & Information Systems, Pace University, NY

the necessity for adaptive, machine-learning-driven solutions that can identify new fraud trends.

These issues are addressed by the InvisiFraud system using a deep learning (DL) and machine learning (ML) approach. Models are able to identify non-linear trends that rule-based systems overlook by examining big transaction datasets. To capture intricate transaction patterns, important models include Random Forest, Logistic Regression, K-Nearest Neighbors (KNN), Decision Trees, and Convolutional Neural Networks (CNNs).

Class imbalance, in which fraudulent transactions are substantially less frequent than legitimate ones, is a major problem in fraud detection. Recall may be decreased by ML models favoring predictions that are not fraudulent as a result of this imbalance. Oversampling (raising the fraud class) and undersampling (lowering the non-fraud class) are used to counteract this. The model's recall, accuracy, and F1-score all improve with oversampling.

The design, evaluation, and benchmarking of the InvisiFraud system are described in this study. It contrasts CNN and ML models, highlighting the effect of data resampling, and stresses striking a balance between precision and recall. In order to improve financial systems' openness and credibility, it concludes by suggesting future enhancements in real-time detection and explainable AI (XAI).

2. Literature Review

Financial system fraud detection has changed dramatically, moving from conventional rule-based techniques to more sophisticated machine learning (ML) techniques. Earlier systems used pre-established rules created by subject-matter specialists to identify transactions that were deemed suspicious, like those that exceeded a specific threshold or came from odd places. Although rule-based systems were straightforward, they were rigid, unadaptable, and needed regular modifications to stay up with changing fraud strategies. Contrarily, machine learning makes it possible to automatically recognize intricate patterns in data and adjust to novel fraud behaviors, giving it a more accurate and dynamic solution.

The following machine learning models are utilized in fraud detection: Neural Networks, Random Forest, and Logistic Regression. Despite being straightforward and easy to understand, logistic regression has trouble identifying intricate patterns in fraud detection. On the other hand, the ensemble approach Random Forest provides great accuracy and robustness and is excellent at handling high-dimensional data and capturing both linear and non-linear correlations. Fraud detection has seen a growing use of Neural Networks, such as Convolutional Neural Networks (CNNs), especially for huge, unstructured datasets. Although CNNs have potential, particularly in identifying temporal or sequential fraud trends, their high computational costs make them more appropriate for batch processing.

Class imbalance, where fraudulent transactions are significantly less common than non-fraudulent ones, is a major problem in fraud detection. Models that prioritize the majority class may result from this imbalance, frequently overlooking fraudulent transactions. Techniques like undersampling and oversampling, like SMOTE, are employed to address this. Oversampling might result in overfitting even while it enhances memory. The goal of hybrid approaches is to minimize the

disadvantages of both strategies while maximizing their advantages. Furthermore, as accuracy by itself is insufficient to evaluate model performance, evaluation metrics like as precision, recall, and F1 score are essential.

In order to improve fraud detection accuracy and explainability while tackling the issues of class imbalance, this work integrates resampling techniques and evaluates several machine learning models.

3. Methodology

To ensure the best possible performance in fraud detection, the InvisiFraud system is developed using a methodical approach. The crucial processes of feature engineering, model selection, and data preparation are described in this section. Each of these processes is essential to creating a reliable and efficient fraud detection system.

3.1. Preparing Data

Transaction amount, duration, and anonymised behavioral variables are among the features that the InvisiFraud system cleans, encodes, and scales to prepare data. Categorical variables are transformed using label or one-hot encoding, and missing data is imputed. While oversampling uses Random Oversampling to increase fraud cases, undersampling decreases non-fraud samples to correct class imbalance. By ensuring a balanced dataset, these procedures help machine learning models—like KNN and Logistic Regression—identify fraud more successfully by learning from samples of both the majority and minority classes.

3.2 Exploratory Data Analysis

EDA is carried out to comprehend the structure of the dataset and find patterns or anomalies prior to using machine learning models. Univariate analysis is a crucial phase in EDA, in which boxplots and histograms are used to analyze the distribution of each feature (such as transaction amount and time). This aids in identifying any outliers or skewness in the data.

To investigate the connections between features and the goal variable (fraud or non-fraud), bivariate analysis is then performed. This is accomplished by using pair plots, correlation matrices, and scatter plots, which aid in locating pertinent characteristics and possible correlations that might point to fraudulent conduct.

3.3 Feature Engineering

Principal Component Analysis (PCA) and Feature Importance Analysis are two methods that InvisiFraud uses to improve feature quality. By breaking down features into uncorrelated components, PCA lowers dimensionality, prevents overfitting, and lowers complexity. By keeping only the most pertinent features, Random Forest's feature importance analysis finds the most useful ones for fraud detection. These techniques ensure that the system trains efficiently and accurately by concentrating on the key features and removing redundancy, which enhances model efficiency, interpretability, and predictive performance.

3.4 Model Selection

InvisiFraud integrates deep learning methods with conventional machine learning models to provide an efficient fraud detection system. Every model has special advantages that are appropriate for certain data kinds and levels of complexity. Random Forest, Logistic Regression,

Decision Tree Classifier, K-Nearest Neighbors (KNN), and Convolutional Neural Networks (CNNs) are among the models that were chosen. Metrics including accuracy, precision, recall, and F1 score are used to compare these models.

3.4.1 Random Forest Classifier
An ensemble model called Random Forest reduces overfitting and increases robustness by combining predictions from several decision trees. Because of its high accuracy and capacity to spot significant data features, it is frequently employed in fraud detection.

3.4.2. Logistic regression
A basic linear model called logistic regression calculates the likelihood that a transaction is fraudulent. It is a suitable starting point for fraud detection since it is interpretable, computationally economical, and performs well in high-dimensional areas.

3.4.3. Decision Tree Classifier
To generate predictions, decision trees divide the data according to features. Although they are simple to understand, if not adjusted, they are prone to overfitting. In spite of this, they provide decision-making openness, which is useful for detecting fraud.

3.4.4. KNN (K-Nearest Neighbors)
The proximity to nearby data points determines the class that a KNN, a distance-based classifier, assigns. KNN is more appropriate for smaller, evenly distributed data sets since, despite its intuitiveness, it is computationally costly and may not work well with huge datasets.

3.5 Convolutional Neural networks

A popular deep learning model for image identification, the Convolutional Neural Network (CNN) has been modified for use with tabular transaction data. CNNs are able to identify temporal and spatial patterns in the transaction data by converting the tabular data into an image-like organized representation. CNNs are very helpful for huge, high-dimensional datasets and are able to simulate more complex relationships than typical machine learning models. CNNs, in contrast to simpler models, are able to recognize non-linear dependencies that linear models find challenging. CNNs, however, need longer training time, are computationally demanding, and might not be explainable. However, they have a great deal of promise for identifying extremely intricate fraud patterns that are beyond the scope of more straightforward models.

A strong technique is used by the InvisiFraud system to identify fraudulent transactions. Using undersampling and oversampling techniques, transaction data is preprocessed, scaled, and balanced at the start of the process. PCA is used to reduce feature dimensions during the feature engineering phase, and feature importance analysis is carried out to determine which characteristics are most pertinent. Lastly, a range of machine learning models are trained and assessed, such as CNNs, Random Forest, KNN, Decision Trees, and Logistic Regression.

The approach guarantees that the InvisiFraud system is ready to tackle the difficulties presented by unbalanced datasets and intricate fraud trends. The system is capable of achieving high precision, recall, and overall performance in fraud detection by utilizing dimensionality reduction, data balancing approaches, and a combination of

classical and deep learning models. The experimental results, including a comparison of the models' performances and the efficacy of data resampling methods, will be presented in the next section.

4. Results

Addressing the imbalance between fraudulent and non-fraudulent transactions in the dataset is a crucial component of fraud detection. The dataset for the InvisiFraud system shows a notable imbalance, with a substantially greater number of non-fraudulent transactions than fraudulent ones. This imbalance is a frequent problem in fraud detection since the model may start to anticipate transactions that aren't fraudulent, making it harder to spot fraudulent situations.

Techniques like undersampling and oversampling were used to balance the dataset to solve this problem, enhancing the model's capacity to detect fraud effectively without being overpowered by the majority class. A better-balanced distribution for model training is ensured by oversampling, which raises the number of fraudulent samples while undersampling lowers the number of non-fraudulent transactions.

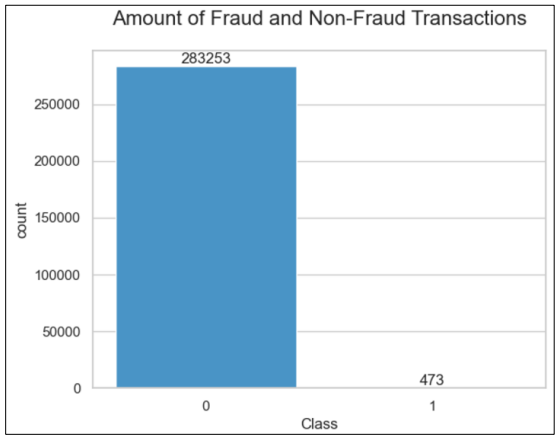

Figure 1: Fraudulent and Non-Fraudulent Transactions

To assess the relationship between each characteristic in the dataset and the "fraud" class label, a feature correlation plot was created. Which characteristics are more suggestive of fraud are shown by the correlation analysis. The feature selection method was influenced by some features that showed substantial positive or negative relationships with fraudulent transactions. To avoid noise in the model, features having a high association to fraud were kept, while those with little to no correlation were eliminated.
By utilizing the correlation data, we made sure that the models were trained with the most important attributes, maximizing the system's predictive ability and enhancing fraud detection precision and recall.

Comparing undersampling and oversampling methods reveals significant variations in the InvisiFraud system's performance. Better overall performance was shown by the increased accuracy of 0.99 obtained with oversampling as opposed to 0.94 with undersampling. However, the model's ability to detect fraud is not entirely captured by accuracy alone, particularly in datasets that are unbalanced.

With a score of 0.99 versus 0.96 in precision, oversampling once more performed better than undersampling, lowering false positives and enhancing fraud detection. Better identification of fraudulent transactions was made possible by oversampling, which demonstrated a small improvement (1.0) over undersampling (0.93) for recall.

Further demonstrating oversampling's improved performance, the F1 score, which balances precision and recall, was greater for oversampling (0.99) than undersampling (0.94). Overall, oversampling was the

preferable method for the InvisiFraud system because it was more successful in addressing class imbalance, providing superior fraud detection, and lowering mistakes.

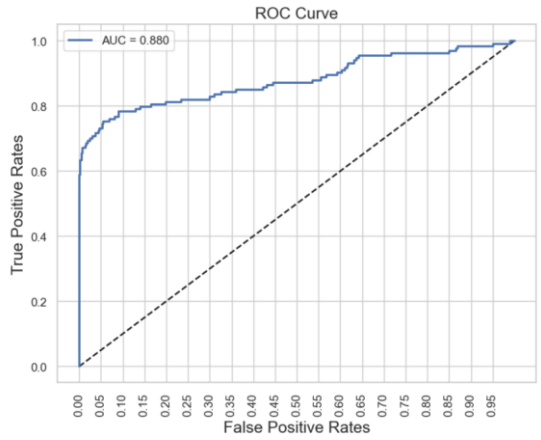

Figure 2: ROC-AUC Curve for Logistic Regression

With an accuracy of 99.91%, logistic regression was able to classify the majority of transactions properly. While its recall of 58.21% shows its limitations in recognizing all fraudulent cases, its precision of 78.79% shows a reasonable capacity to correctly categorize fraudulent transactions. An F1-score of 66.95% reflects a balance between precision and recall, capturing this trade-off. The model's ability to differentiate between fraud and non-fraud situations is demonstrated by its AUC score of 0.880. The comparatively low recall, however, raises the possibility that some fraudulent transactions may go unnoticed using logistic regression, which could be problematic for fraud detection.

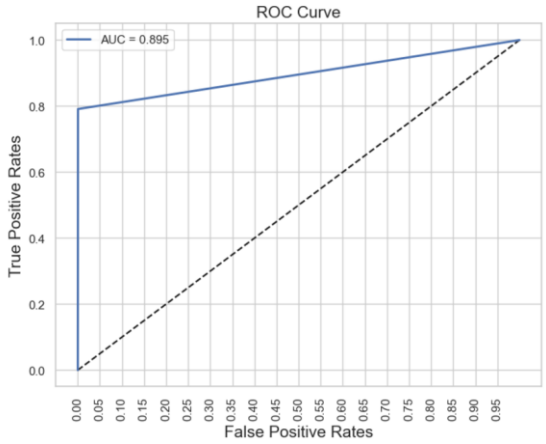

Figure 3: ROC-AUC Curve for Decision Tree

With an accuracy of 99.92%, the Decision Tree Classifier marginally outperformed Logistic Regression. With a recall of 79.10%, it detects more instances of fraud. With a precision of 71.62%, it has a little higher false-positive rate than Random Forest. When compared to Logistic Regression, the F1-score of 75.18% indicates a superior balance between precision and recall. Improved discrimination of the model is confirmed by an AUC of 0.895. Although decision trees are more interpretable, overfitting is a risk. Given the Decision Tree's strong performance, overfitting appears to be well-managed in this instance.

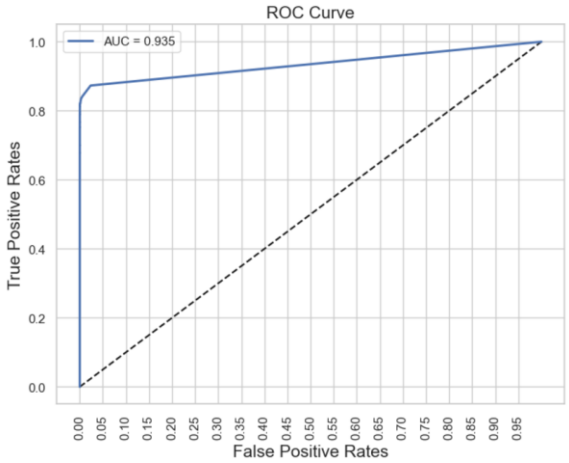

Figure 4: ROC-AUC Curve for Random Forest

With an accuracy of 99.95% and an AUC of 0.935, Random Forest is the best-performing model, demonstrating great discriminative power. It greatly lowers false positives and has the highest precision (95.05%). Its 71.64% recall rate demonstrates how well it can detect a significant percentage of fraudulent transactions. With an F1-score of 81.70%, precision and recall are well-balanced. Random Forest is the best choice for fraud detection since it can reduce false positives while keeping a high recall.

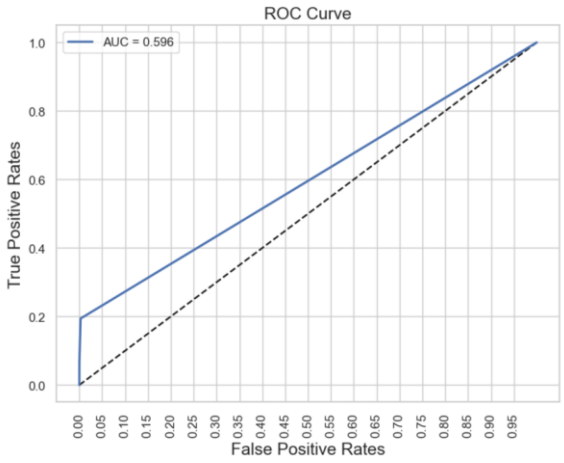

Figure 5: ROC-AUC Curve For KNN

KNN has the lowest performance. Because of the unbalanced dataset, its accuracy of 99.85% is deceptive. It misses the majority of fraudulent cases, as evidenced by its incredibly low recall of 2.98% and precision of 80%. Its F1-score of 5.75% likewise reflects this. Poor discrimination between fraud and non-fraud is indicated by the AUC of 0.595. KNN is not appropriate for this purpose because of its dependence on distance-based categorization, which renders it unsuccessful in huge, unbalanced datasets.

Figure 6: ROC-AUC Curve for CNN

With an accuracy of 99.94% and the greatest AUC of 0.967 of any model, CNN performs admirably. Its accuracy in identifying fraudulent instances is demonstrated by its 85.60% precision, and its 75.37% recall shows that it can identify a sizable percentage of real fraud cases. With an F1-score of 80.86%, the performance is well-balanced. CNN is a potent tool for fraud detection because of its ability to recognize intricate, non-linear patterns, but it uses more processing power than more straightforward models.

Model Name	Accuracy
Logistic Regression	0.99909
Decision Tree	0.99917
Random Forest	0.99949
KNN	0.99846
CNN	0.99941

Table 1: Comparison of Models

Despite its impressive performance, the InvisiFraud system has drawbacks that impact its effectiveness, scalability, and practicality. Managing data imbalance, where fraudulent transactions are far less often than genuine ones, is one of the main challenges. In order to improve recall and F1-score, synthetic fraud samples are

created via oversampling. But doing so raises the possibility of overfitting, which would make the model learn artificial patterns rather than extrapolating to new data. This problem can be lessened by using sophisticated methods like SMOTE and ensemble models, which combine predictions from several models and provide more varied synthetic samples.

The computational complexity of the system is another drawback. To train on huge datasets, CNNs need a lot of memory and processing power, especially if they don't have GPUs or TPUs. CPU training lengthens processing times, which delays model updates. Using pre-trained models to lower computational costs, solutions include dimensionality reduction (PCA), cloud-based GPUs, and transfer learning.

Because the system presently employs batch processing, which causes delays in detecting fraudulent transactions, real-time fraud detection is still difficult. By moving to event-driven architectures or stream processing, incoming transactions may be analyzed instantly. Furthermore, detection delay can be decreased by employing lightweight models for real-time inference, such as Logistic Regression. By resolving these problems, InvisiFraud will become more responsive, scalable, and efficient.

5. Future Work

Although the InvisiFraud system exhibits potential in detecting fraud, it still needs to be improved in order to increase transparency, scalability, and robustness. Managing the class imbalance present in fraud datasets, when fraudulent transactions are substantially less frequent than valid ones, is one crucial topic. Advanced techniques like SMOTE (Synthetic Minority Oversampling Technique), which interpolates between existing minority class data points to create more varied synthetic samples, could be used to improve current methods like oversampling and undersampling. This lessens overfitting and improves the model's ability to generalize to novel fraud patterns. Improved precision, recall, and F1 scores can result from Hybrid sampling methods that combine SMOTE with undersampling.

Delays in detecting fraudulent transactions can lead to substantial financial loss, making real-time fraud detection another crucial topic. The system would be able to evaluate transactions as they happen if it switched from batch processing, which now increases latency, to stream processing. Event-driven architectures are made possible by technologies like Apache Kafka and Apache Flink, which enable the real-time detection and blockage of fraudulent transactions. Furthermore, decision-making can be accelerated without compromising accuracy by combining quantization and pruning of deep learning models with lightweight models for real-time inference.

Lastly, user trust and compliance depend on explainability and interpretability. Since many AI models are "black boxes," it can be challenging to explain why a transaction was reported as fraudulent. Through the use of Explainable AI (XAI) techniques such as SHAP and LIME, stakeholders can determine which characteristics had an impact on a choice. Clear, understandable choice paths are also offered by transparent models, including choice Trees and Logistic Regression. Through these initiatives, InvisiFraud will become more responsive, transparent, and effective, guaranteeing a strong fraud detection system that satisfies legal requirements and inspires user confidence.

6. Conclusions

An important development in fraud detection is the InvisiFraud system, which increases accuracy by using machine learning models and data resampling techniques. By addressing issues such as class imbalance and model overfitting, it has significant promise for preventing financial fraud in the real world. InvisiFraud's usage of machine learning and data resampling, which improve fraud detection performance, is one of its major accomplishments. When undersampling and oversampling are compared, the results show that oversampling produces greater accuracy, precision, recall, and F1 scores.

After evaluating the performance of several categorization models, the best-performing models were Random Forest and Convolutional Neural Networks (CNNs). By integrating several decision trees, balancing bias and variance, and providing information on feature relevance, the Random Forest classifier demonstrated efficacy. Originally created for image recognition, CNNs have shown potential in identifying intricate fraud patterns in larger datasets. They are more appropriate for batch processing, nevertheless, because they demand a lot of computing power.

Class imbalance, overfitting, and generalization to larger datasets are some of the issues that InvisiFraud encounters despite its excellent performance. Overfitting may result from the current methodology, which uses simple oversampling and undersampling. To produce synthetic samples that are more realistic, sophisticated methods like SMOTE and hybrid resampling are advised. While training on bigger datasets and using domain adaptation approaches can improve model generalization and scalability, regularization techniques such as L2 regularization and dropout can assist prevent overfitting.

InvisiFraud can be incorporated into mobile wallets, e-commerce platforms, and fraud detection systems used by financial institutions. The system may offer transparent fraud analysis by utilizing explainability techniques like SHAP and LIME and real-time detection frameworks, which are essential for both customer trust and regulatory compliance. InvisiFraud is in a strong position to lower fraud in a variety of industries thanks to continuous advancements in resampling, real-time streaming, and explainable AI.

7. References

- Ahmed, M., Mahmood, A. N., & Hu, J. (2016). A survey of network anomaly detection techniques. Journal of Network and Computer Applications, 60, 19-31.
- Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. Expert Systems with Applications, 36(3), 4626-4636.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321-357.
- Gupta, R., & Ziegler, C. (2019). Fraud detection in financial systems using machine learning techniques: A review. International Journal of Advanced Computer Science and Applications, 10(6), 337-344.
- He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263-1284.
- Iglewicz, B., & Hoaglin, D. C. (1993). How to detect and handle outliers. Sage Publications.

- Kim, J., & Park, H. (2021). A comparison of machine learning algorithms for credit card fraud detection. Journal of Financial Engineering, 8(2), 124-135.
- Krawczyk, B. (2016). Learning from imbalanced data: Open challenges and future directions. Progress in Artificial Intelligence, 5(4), 221-232.
- McCluskey, L. (2016). Fraud detection using ensemble learning algorithms. International Journal of Computer Applications, 139(4), 13-17.
- Saberi, M., & Varma, S. (2019). Fraud detection in financial systems using machine learning models. International Journal of Machine Learning and Computing, 9(3), 407-413.
- Yao, Y., & Zhang, Y. (2020). A deep learning-based approach to fraud detection in e-commerce. Proceedings of the IEEE International Conference on Machine Learning and Data Mining, 2020, 58-67.
- Zong, Y., & Lu, X. (2021). Detecting fraudulent transactions using machine learning models: A review. Journal of Financial Technology, 1(2), 43-56.