IDENTIFYING RISK THROUGH CUSTOMER TRANSACTION XGBOOST TECHNIQUE IN SUPPLY CHAIN FINANCE

Lakshmi Sri Lasya Tatiraju, Krishna Venkat Chowdary Dashti, Tejaswi K, Sumathi M
SASTRA Deemed to be University, Tamil Nadu, India
Assistant Professor, SASTRA Deemed to be University, Tamil Nadu, India
Email: 125156063@sastra.ac.in, 125156161@sastra.ac.in, 125015118 sumathi@it.sastra.edu

Abstract

SCF is a type of supplier finance that enables the supplier to serve their receivables earlier than the actual payment date, thereby freeing up its working capital and also benefits the buyer as the buyer can obtain short-term credit at a lesser cost. Delayed payments by buyers pose a significant threat to supply chain stability. Thus, identifying potential supplier liquidity is crucial. This work addresses this issue by developing a finance risk prediction model using XG Boost and examining customer transaction behavior. The single and hybrid models are constructed for a comparative analysis of their performance using ROC, area under the ROC curve (AUC), and F1-Score. Feature importance and partial dependence plots (PDPs) are applied to interpret the model's predictions. The model's effectiveness and accessibility are further explored by a web-based tool, enabling users to directly interact with the model and obtain personalized risk predictions. The single models to be implemented are XG Boost, Random Forest (RF), Gradient Boosting Decision Tree (GBDT), and Light GBM. Hybrid models are made by combining these single models with Linear Regression (LR). Among all the above-mentioned models, the XG Boost model demonstrates superior performance, effectively predicting potential risks and uplifting managerial payment practices. This clear understanding is revolutionizing risk management strategies and fostering more robust supply chains. This study opens new paths for future exploration of practical models for financial risk assessment within SCF.

Keywords: Supply chain stability, Delayed Payment, Buyer transaction behavior, financial risk prediction, XG Boost model

1. Introduction

SCF is a type of supplier finance that enables the supplier to serve their receivables earlier than the actual payment date. Thereby freeing up its working capital and also benefits the buyer as the buyer can obtain short-term credit at a lesser cost. The credit payment is adopted by 80% of UK companies. This credit payment is grabbing increasing attention, especially after COVID-19.

When the delayed payment passes from customer to customer, there will be a liquidity crisis that arises which in turn will invoke SCF Financial risk. When large enterprises fail to pay their suppliers, there will be a parallel increment in loss to suppliers indicating large enterprises as a key attribute in delayed payment.

Delayed payments by buyers pose a significant threat to supply chain stability. Hence there has to be a clear study of several trends of delayed payment by buyers. Thus, identifying potential supplier liquidity is crucial.

2. Related Work

In this section, the authors have displayed a summary of the literature review. [1] Addresses the concern regarding risk prediction using Low-Certainty-Need (LCN) supply chains. LCN SCs prioritize flexibility, structural variety, and parametrical redundancy, enhancing disruption resistance and resource allocation during recovery. [2] Is based on a time-based payment contract. It compares delayed payment and on-time payment contracts in an assembly system with stochastic production times. Most of the papers taken as references for the base paper include exploring various time series data, particularly over DPS.

Hybrid mechanisms are also explored, they proposed integrated machine learning models exhibiting higher accuracy than individual models like decision trees. In contrast, the authors' study focuses on credit risk predictive models for SMEs by combining both customer transaction behavior data and ML models to study the degree of DPS for each customer. They emphasize model interpretability by constructing both a single tree-based model and a hybrid model. The aim is to select the optimal model to enhance financial risk prediction capabilities.

2.1 Problem Statement

Supply Chain Finance is prone to risk with delayed payments and hence analysis, identification and assessment of imperative reasons are crucial for smooth transactions. Several methods of delayed payments are key factors for liquidity crises. Hence there has to be a collaborative approach of combining delayed payment analysis with machine learning and deep learning approaches.

2.2 Objective

- This work addresses this issue by developing a finance risk prediction model using XGBoost and examining customer transaction behavior.
- In this work, there will be a comparative analysis for risk assessment using various machine learning algorithms like Random Forest, Gradient Boosting Decision Tree (GBDT), and LightGBM.
- Hybrid models are made by combining single models with linear regression as the target variable here is binary
- CNN model implementation.

3. Proposed System

This work addresses this issue by developing a finance risk prediction model using XG Boost and examining customer transaction behavior. In this work, there will be a comparative analysis for risk assessment using various machine learning algorithms like Random Forest, Gradient Boosting Decision Tree (GBDT), LightGBM and hybrid models mixed up with Linear Regression. Deep neural network like Convolution neural networks is implemented with necessary regularization techniques. Data pre-processing is made the same for every algorithm implementation to have a justified comparative analysis. For hybrid models, each model output is transferred to subsequent linear regression. The authors identified the architecture as shown in Figure 1.1.

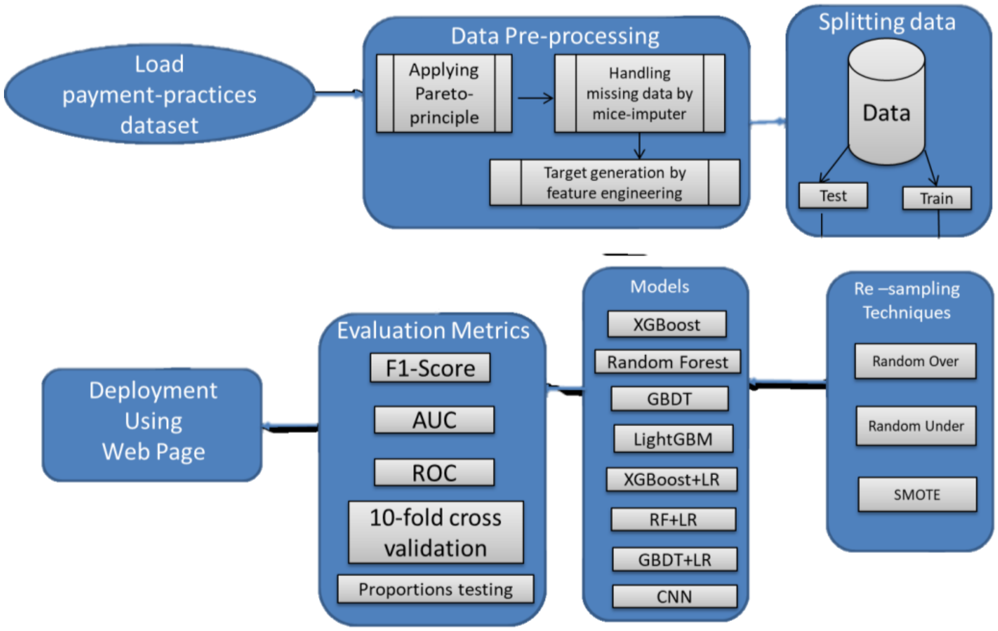


Fig 1.1 System Architecture

The payment practices dataset is imported and preprocessed to remove noise. Exploratory data analysis is applied to understand the features dependencies and relations. The payment behavior of customers is used as a variable and the dependent variable target is identified by the pareto principle. Identify imbalances in the data and apply random over- sampling. The payment practices dataset is split for training and testing. XG Boost, Random Forest (RF), Gradient Boost Decision Tree (GBDT), Light GBM, XG Boost + LR, RF + LR, and GBDT + LR are implemented. The performance of applied techniques is evaluated. The outperformed method is used for financial risk assessment. Web-based tool is developed to explore effectiveness and accessibility. In addition to this deep learning models are also explored.

XGBOOST Algorithm:

XGBoost is a machine learning algorithm that belongs to the ensemble learning category, specifically the gradient boosting framework. It utilizes decision trees as base learners and employs regularization techniques to enhance model generalization.

1.Initialization: Initialize the model with a constant value. For binary classification, it can be 0.5 for all instances

2.Compute the Similarity Scores: For each node in the tree, calculate the similarity score. The similarity score for a node is given by the formula:

$$similarity\ =\ Gradient \wedge 2\ /\ (Hessian\ +\ \lambda)$$

where λ is the regularization term.

3.Split Finding: For each node, find the best split that maximizes the reduction in the loss function. The reduction is given by the formula:

$$Gain\ =\ Left\ Similarity\ +\ Right\ Similarity\ -\ Similarity\ of\ parent$$

If the Gain is less than γ, stop growing the tree further. Here, γ is a regularization parameter.

4.Pruning: Prune the tree, remove the splits for which Gain is less than γ.

5.Tree Weight Calculation: Calculate the output value for each leaf node which minimizes the loss function. The output value is given by the formula:

$$-\ Gradient\ /\ (Hessian\ +\ \lambda)$$

6.Update the Tree: Add this tree to the ensemble of trees. The contribution of this tree to the final prediction is controlled by a learning rate η.

7.Repeat Steps 2-6: Repeat Steps 2 to 6 for a fixed number of iterations (n_estimators) or until the loss function does not improve on an external validation dataset.

8.Prediction: Make predictions by summing the predictions from all the trees. The final prediction is the class with the highest sum

4. Experimental Results

4.1 Dataset Details

The payment practices dataset from the website:
https://check-payment-practices.service.gov.uk/export

The dataset consists of 77681 entries and 23 attributes.
Data is from the UK government website

1. **Data Pre-processing:**
   a. Bar plot for '% Invoices not paid within agreed terms' and 80th percentile in figure 4.1

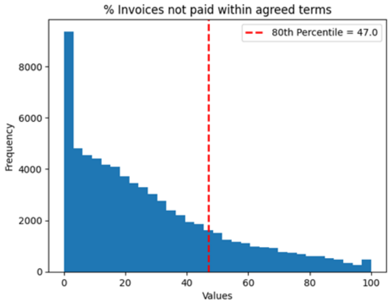


Fig.4.1 80th percentile in '% Invoices not paid within agreed terms'

The figure explains about the 80th percentile in '% Invoices not paid within agreed terms' to identify the proportion of high risk and low risk samples in target variable. Initial consideration was 20 % high risk and 80% low risk, later there is an analysis made for different proportions.

b. Correlation Analysis in figure 4.2

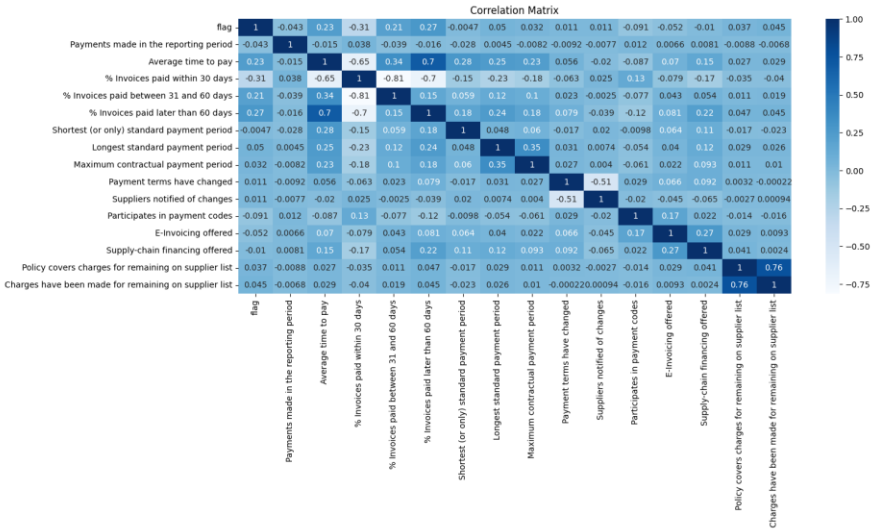


Fig.4.2 Correlation Analysis

The correlation matrix explains about correlation between features. Highly correlated features are analyzed and are removed to avoid unnecessary computations. We observed that

1.% Invoices paid within 30 days is highly correlated with % Invoices paid between 31 and 60 days (correlation: -0.81)

2.% Invoices paid between 31 and 60 days is highly correlated with % Invoices paid within 30 days (correlation: -0.81)

3.Policy covers charges for remaining on supplier list is highly correlated with Charges have been made for remaining on supplier list (correlation: 0.76)

4.Charges have been made for remaining on supplier list is highly correlated with Policy covers charges for remaining on supplier list (correlation: 0.76)

c. Pie chart of high and low risk after sampling in figure 4.4

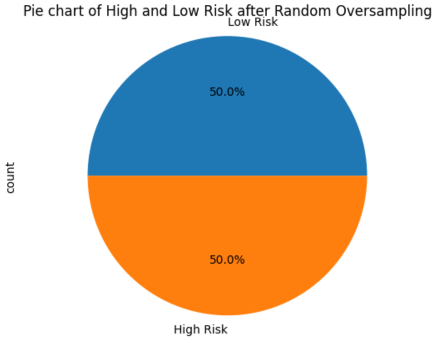


Fig.4.4 Pie chart of high and low risk after sampling

After observing the results of different re-sampling techniques, random over sampling out performs with F1 -score of 54.9 which is highest among random under sampling and smote.

2. Models implementation and comparison:

a. ROC (Receiver Operating Characteristic Curves) in figure 4.5

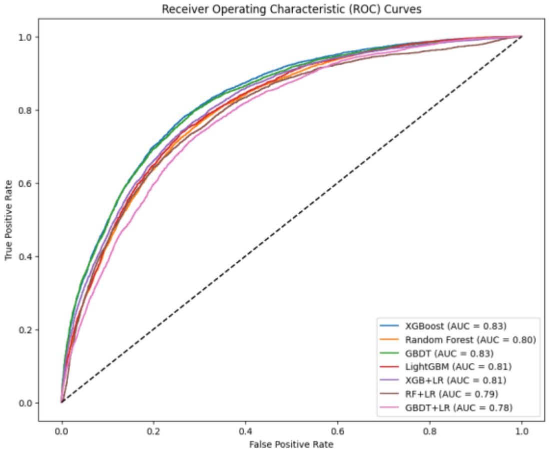


Fig.4.5 ROC curves for all the models

By plotting ROC curve, we can identify that XGBOOST has highest AUC of 0.83, followed by GBDT with AUC of 0.83 and Random Forest with AUC of 0.80. We can observe that single models perform well than hybrid models.

b. Model comparisons is figure 4.6

	Model	**Accuracy**	**F1 Score**	**AUC**
1	**Random Forest**	**0.700144**	**0.509270**	**0.802509**
2	**GBDT**	**0.774579**	**0.552037**	**0.826370**
3	**LightGBM**	**0.717474**	**0.517679**	**0.806924**
4	**XGB+LR**	**0.746640**	**0.532741**	**0.812604**
5	**RF+LR**	**0.802477**	**0.493552**	**0.792003**
6	**GBDT+LR**	**0.671528**	**0.486486**	**0.782524**
7	**XGBOOST**	**0.762955**	**0.551432**	**0.830329**

Fig.4.6 Model comparisons

c. Feature Importance using XGBOOST in-built function in figure 4.7

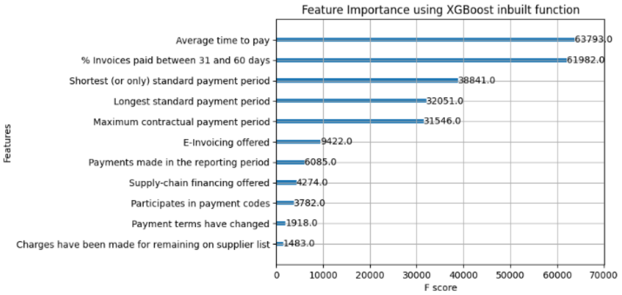


Fig.4.7 Feature Importance by XGBOOST in-built function

This is feature importance plot made using inbuilt XGBOOST function to make understand what features are essentially prevailing risk in supply chain. This is an XAI method brought here to make model interpretable and explainable. Here it tells about target variable dependence on each feature.

d. Feature importance by IV in figure 4.8

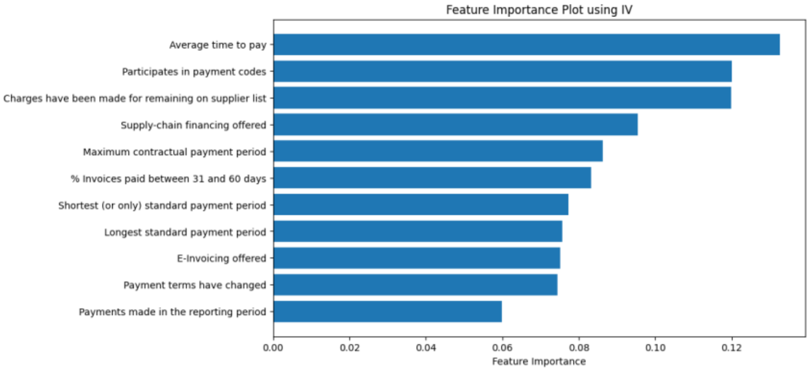


Fig.4.8 Feature importance by information value method

This is also a partial dependence plot which interprets the dependence of target on each feature but here it is made using information value method in which features importance are obtained by XGBOOST in-built function and the values are sorted for knowing the highly important feature.

e. Partial dependence plot for ‘average time to pay’ in figure 4.9

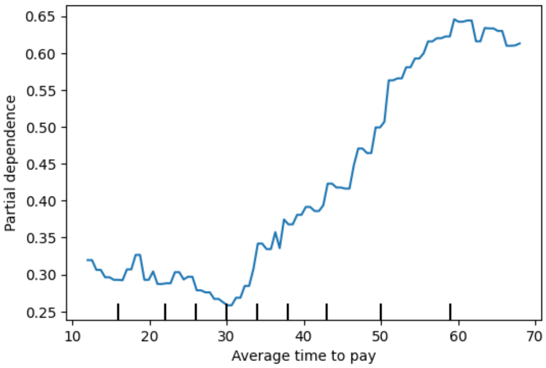


Fig.4.9 Partial dependence plot for ‘average time to pay’

This is the individual trend of dependence of output on the feature ‘average time of time’ which is identified as most important by partial dependence plot.

f. Partial dependence plot for ‘supply-chain financing offered’ in figure 4.10

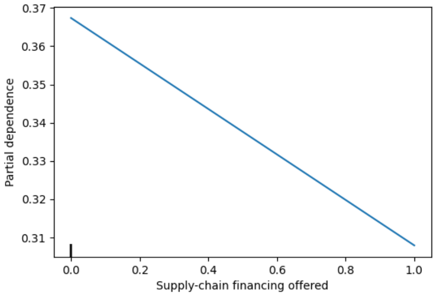


Fig.4.10 Partial dependence plot for ‘supply-chain financing offered’

There is a negative dependence of target on the feature ‘supply-chain financing offered’ yet this feature is important in risk assessment.

g. Partial dependence plot for ‘payment terms have changed’ in figure 4.11

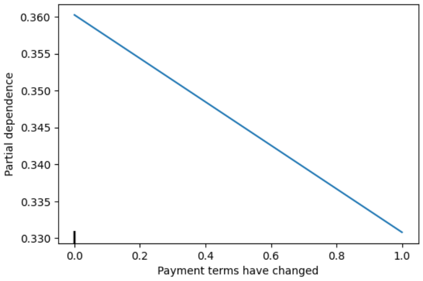


Fig.4.11 Partial dependence plot for ‘Payment terms have changed’

There is a negative dependence of target on the feature ‘payment terms have changed’.

h. Partial dependence plot for ‘charges have been made for remaining on supplier list’ in figure 4.12

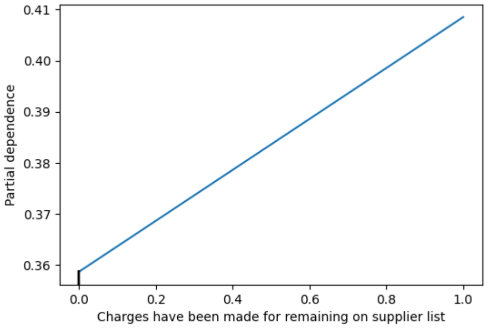


Fig.4.12 Partial dependence plot for ‘Charges have been made for remaining on supplier list’

There is a positive dependence of target on the feature ‘Charges have been made for remaining on supplier list’.

i. Partial dependence plot for ‘% Invoices paid between 31 and 60 days’ in figure 4.13

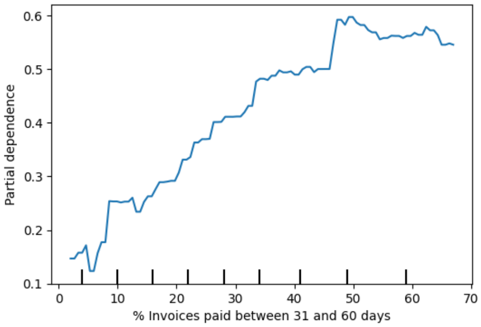


Fig.4.13 Partial dependence plot for ‘% Invoices paid between 31 and 60 days’

There is an increasing trend observed in identifying dependence of target on ‘% Invoices paid between 31 and 60 days’

j. **10-fold cross-validation results on XGBOOST in figure 4.14**

Average Accuracy	**0.8129338704884527**
Average F1 Score	**0.8217140268262384**
Average ROC AUC Score	**0.8129096994521376**

Fig.4.14 10-fold cross validation results on XGBOOST

10-fold cross-validation gives a good F1-score of 82.17 which is 38.9% higher than train-test split of data for XGBOOST.

k. **Results of sampling techniques in figure 4.15**

	Model	**Accuracy**	**F1 Score**	**AUC**
0	**Random Oversampling**	**0.774966**	**0.549779**	**0.814912**
1	**Random Undersampling**	**0.706352**	**0.524085**	**0.802064**
2	**SMOTE**	**0.797724**	**0.516050**	**0.805171**

Fig.4.15 Results of sampling techniques

Random over sampling shows a good increment in F1-score with 4.9% over random under sampling and an increment of 6.5% over SMOTE procedure.

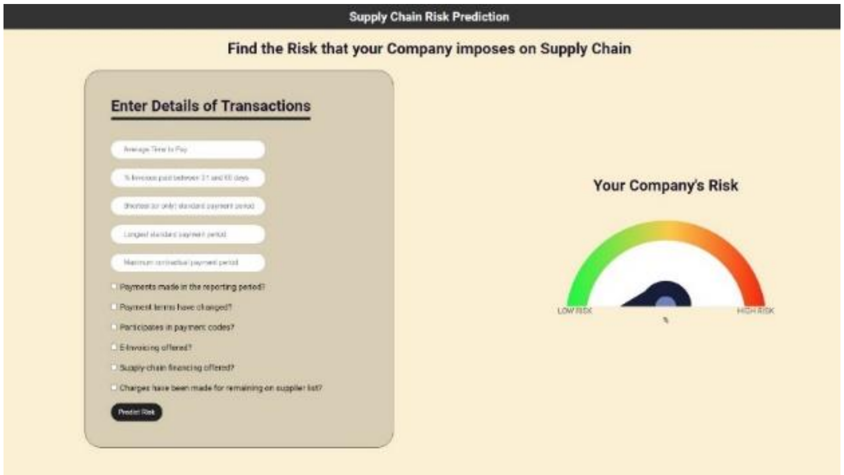
l. **Results of various risk proportion in figure 4.16**

Risk Proportion	**Accuracy**	**F1 Score**	**AUC**
10%	**0.860599**	**0.430976**	**0.831349**
20%	**0.801072**	**0.553533**	**0.825286**
30%	**0.759830**	**0.620794**	**0.814654**
40%	**0.740995**	**0.687769**	**0.812988**
50%	**0.750069**	**0.753123**	**0.824772**

Fig.4.16 Results of various risk proportions

By observing several proportions of high risk and low risk samples, 20% gives stable results and is also relevant with data.

m. **GUI interface in figure 4.17**



Supply Chain Risk Prediction

Find the Risk that your Company imposes on Supply Chain

Enter Details of Transactions

Average Time to Pay

% Invoices paid between 31 and 60 days

Shortest (or only) standard payment period

Longest standard payment period

Maximum contractual payment period

□ Payments made in the reporting period?

□ Payment terms have changed?

□ Participates in payment codes?

□ E-Invoicing offered?

□ Supply-chain financing offered?

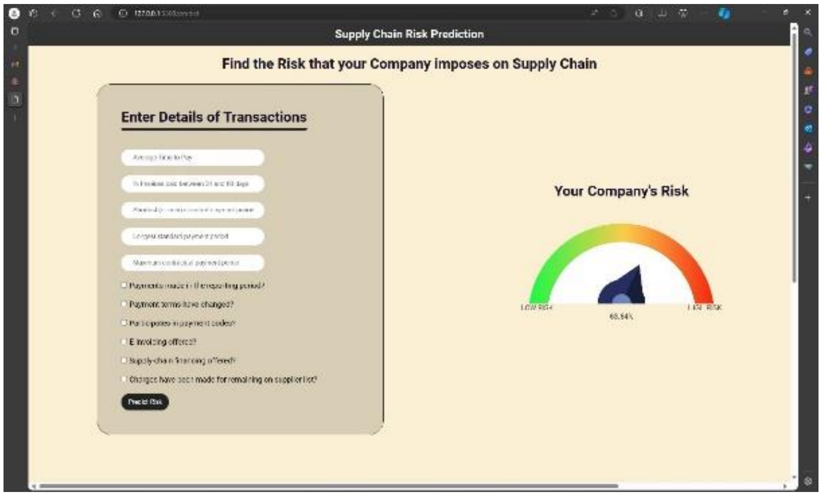
□ Charges have been made for remaining on supplier list?

Your Company's Risk

Fig 4.17 GUI Interface

This is the GUI interface made for the user convenience. The web page consists of features essential for evaluating. Few features like ‘average time to pay’, ‘% invoices between 31 and 60 days’, ‘shortest or only standard payment period’, ‘longest standard payment period’, ‘maximum contractual payment period’ are taken numerical input which other binary features are given as options to select.

n. GUI interface with input and prediction in figure 4.18



Supply Chain Risk Prediction

Find the Risk that your Company imposes on Supply Chain

Enter Details of Transactions

Average Time to Pay

% Invoices paid between 31 and 60 days

Shortest (or only) standard payment period

Longest standard payment period

Maximum contractual payment period

□ Payments made in the reporting period?

□ Payment terms have changed?

□ Participates in payment codes?

□ E invoicing offered?

□ Supply-chain financing offered?

□ Changes have been made for remaining on supplier list?

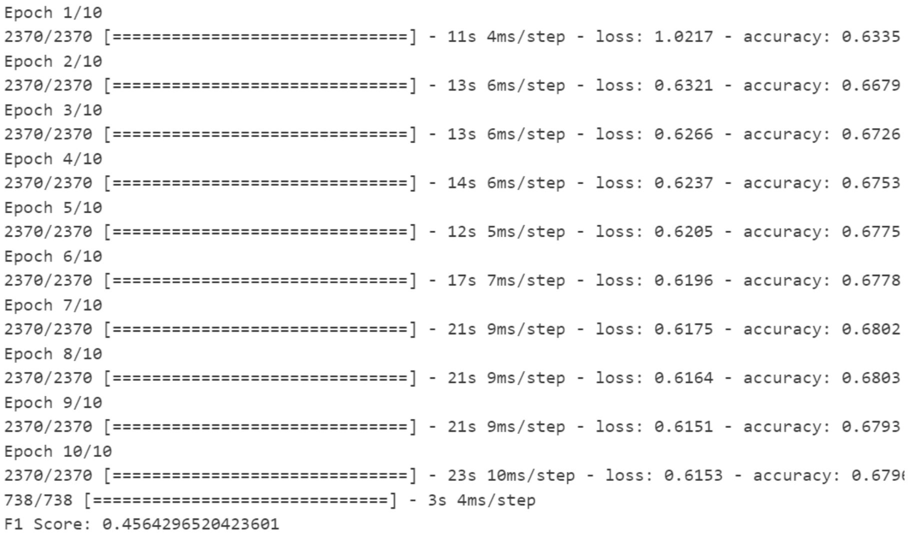
Predict Risk

Your Company's Risk

Fig 4.18 GUI Interface with input and prediction

This is the final web page with the input and prediction. The risk prediction obtained here is 68.54% which is also indicated with mild red which resembles the impact of risk.

o. CNN output with epochs count is in figure 4.19



```
Epoch 1/10
2370/2370 [==============================] - 11s 4ms/step - loss: 1.0217 - accuracy: 0.6335
Epoch 2/10
2370/2370 [==============================] - 13s 6ms/step - loss: 0.6321 - accuracy: 0.6679
Epoch 3/10
2370/2370 [==============================] - 13s 6ms/step - loss: 0.6266 - accuracy: 0.6726
Epoch 4/10
2370/2370 [==============================] - 14s 6ms/step - loss: 0.6237 - accuracy: 0.6753
Epoch 5/10
2370/2370 [==============================] - 12s 5ms/step - loss: 0.6205 - accuracy: 0.6775
Epoch 6/10
2370/2370 [==============================] - 17s 7ms/step - loss: 0.6196 - accuracy: 0.6778
Epoch 7/10
2370/2370 [==============================] - 21s 9ms/step - loss: 0.6175 - accuracy: 0.6802
Epoch 8/10
2370/2370 [==============================] - 21s 9ms/step - loss: 0.6164 - accuracy: 0.6803
Epoch 9/10
2370/2370 [==============================] - 21s 9ms/step - loss: 0.6151 - accuracy: 0.6793
Epoch 10/10
2370/2370 [==============================] - 23s 10ms/step - loss: 0.6153 - accuracy: 0.679
738/738 [==============================] - 3s 4ms/step
F1 Score: 0.4564296520423601
```

Fig 4.9 CNN output

The CNN model implemented with regularization is obtained with the F1 score of 45.6 and there were 10 epochs.

5.1 CONCLUSIONS

Today supply chain finances face several challenges, especially concerning delayed payments by customers. Understanding customer transaction behavior and collaborating that knowledge with machine learning and deep learning models can enhance the chances of predicting risk in the supply chain. Displaying the results with proper interpretations and explanations can secure the company from risk and also develop trust in the reason for the risk that was predicted. When tree-based machine learning models are used, the results are quite explanatory as the decision-making is transparent to understand. This gives companies a good precaution for the risk that is going to prevail and gives a hint for the method to avoid risks by making an idea over features that are contributing to the risk factor.

The proposed supply chain risk prediction system puts forward risk prediction through a website that takes relevant features to evaluate the risk and display the amount of risk using a risk meter for user convenience and feasibility. Features like 'average time to pay','% Invoices paid between 31 to 60 days', 'shortest payment method' and 'longest payment method' are asked to enter numerical values and there is a check box given for user convenience to select features like 'Payments made in the reporting period', 'Payment terms have changed',' Participates in payment codes', 'E-Invoicing offered', 'Supply-chain financing offered', 'Charges have been made for remaining on supplier list' is of binary (yes or no). All these features are identified by the feature importance model and XGBOOST outperforms all the machine learning algorithms with an F1-score of 0.55 which is 8.2% higher than Random Forest, and 3.5% higher than XGBOOST + LR. As XGBOOST outperforms, it is improvised with a different pre-processing technique of imputing data with 'iterative imputer' which showed 0.59 as the F1-score which is 7.9% higher than the previous XGBOOST model. The approach also includes a CNN model implementation with the same data pre-processing to identify the F1-score. XGBOOST outperforms CNN by an increment of 12% in the F1-score (CNN F1-score: 0.49). Hence the inputs from the website are passed to the XGBOOST algorithm as a test case and the prediction percentage evaluated is displayed numerically and also in the risk meter.

The data considered for analysis is from the UK government and it is dynamic data. Hence an analysis is made and model implementation is done over both the base paper dataset (till 2018) and the updated data set (till 2023). XGBOOST acquired an F1-score of 0.547 and AUC of 0.812 on old data while on the new data set F1-score is about 0.551 and AUC is about 0.830.

In conclusion, the proposed application overcomes the challenges of unclear results and complex methods of implementation for predicting risk in the supply chain. This helps for the smooth balance of the financial aspects of the company. This is a user-interactive system that helps non-technical users to easily evaluate their risk rate.

5.2 FUTURE PLAN

The following methods can be employed in future:

1. Direct the methods towards data independent approach which gains good attention when there is lack of data.
2. Model independent results

REFERENCES

[1] Ivanov, D., & Dolgui, A. (2019). Low-Certainty-Need (LCN) supply chains: a new perspective in managing disruption risks and resilience. International Journal of Production Research, 57(15–16), 5119–5136. https://doi.org/10.1080/00207543.2018.1521025

[2] Guan, X., Li, G. & Yin, Z. The implication of time-based payment contract in the decentralized assembly system. Ann Oper Res 240, 641–659 (2016). https://doi.org/10.1007/s10479-014-1579-5

[3] Zelong Yi, Zhuomin Liang, Tongtong Xie, Fan Li, Financial risk prediction in supply chain finance based on buyer transaction behavior, Decision Support Systems, Volume 170, 2023, 113964, ISSN 0167-9236, https://doi.org/10.1016/j.dss.2023.113964.

[4] R. Qin, The construction of corporate financial management risk model based on XGBoost algorithm, J. Math. 2022 (2022) 1–8, https://doi.org/10.1155/2022/2043369.

[5] M. Bahrami, B. Bozkaya, S. Balcisoy, using behavioral analytics to predict customer invoice payment, Big Data 8 (1) (2020) 25–37, https://doi.org/10.1089/ big.2018.0116.