BUILDING A SEMI SUPERVISED FRAMEWORK TO IMPROVE CREDIT RISK PREDICTION

A PROPOSAL SUBMITTED TO FACULTY OF SCIENCE, TECHNOLOGY AND INNOVATION DEPARTMENT OF INFORMATION AND COMMUNICATION TECHNOLOGY IN PARTIAL FULFILMENT OF THE AWARD OF BACHELOR OF DEGREE IN DATA SCIENCE

SUBMITTED BY

SAMUEL CHIZUMA(BSDS0821)

ELIZABETH MFUNE (BSDS1922)

QUEEN SOSOLA (BSDS2822)

JEREMIA NKOSI (BSDS2422)

MAXWELL MWALA(BSDS2322)

SUPERVISOR

MR. REUBEN MOYO

Contents

Abstract: 3

1. Introduction 3

2. Problem Statement 4

3. Research objectives and questions 4

How do generative models improve predictions compared to traditional models 5

How do generative models handle class imbalance in credit risk prediction? 5

3. Literature review 7

5. Methodology 8

Data Acquisition and Preparation: 8

Preprocessing: 9

Exploratory Data Analysis (EDA): 9

Model Development: 9

Model Evaluation: 10

Performance metrics: 10

Accuracy and Deployment: 10

6. **Results and findings** 10

Chi-square test 11

T-test 11

7. Timeline 12

8. Resources Required 12

9. Conclusion 13

REFERENCES 14

ENHANCING CREDIT RISK PREDICTION USING SEMI SUPERVISED LEARNING

Abstract:

The project investigates the application of semi-supervised learning techniques to improve credit risk prediction. Traditional credit risk models heavily rely on supervised learning where labeled data is used in the modelling process. Supervised learning requires large volumes of labeled data. Acquiring and labelling such financial data is both time consuming and expensive. Semisupervised learning addresses the challenge by utilizing both labeled and unlabeled data. The proposed project aims to utilize self-training and co-training semi-supervised learning techniques to overcome the data scarcity challenge by generating pseudo labels. By utilizing these techniques, models will be trained and compared with a target accuracy metric of at least 80%. The methodology involves data acquisition, preprocessing and exploratory data analysis, model development using algorithms, not limited to: self-training and graph based models, and model evaluation as well as model deployment. The context of this project mainly explores the application of self-training and co-training semi-supervised learning techniques in credit risk prediction in African financial environments with limited labeled data.

1. Introduction

The demand to acquire loans in African financial markets is on a rapid increase due to adverse financial conditions. It would be a very good case if lending institutions would provide loan opportunities to people provided they have the standard credit worthiness based on carefully picked criteria. It is therefore necessary to have good technological tools for assessing credit risk and chances of loan default. K.Dennis et al (2024). To date, supervised machine learning models and statistical models have been used to assess credit risk. Xu Zhu et al (2023). These models mainly utilize labeled data to map relationships between predictor features and the target variable which is, in most cases loan default. Shi et al (2022). In practice however, labeled data for training supervised machine learning models is scarce and expensive. This has led to reliance on models which fail to capture nonlinear relationships and less powerful prediction models. Semi-supervised learning techniques make good use of unlabeled data provided there is a small amount

of labeled data. With this in mind, a question “can we utilize semi-supervised learning to improve credit risk prediction?” comes into the picture. We look to answer this question and discuss the inclusion of semi-supervised learning techniques in credit risk prediction and whether we can utilize it to maximize good results and if so, how we are going to implement at least one semi-supervised learning technique to improve credit risk prediction and quantifying a borrower’s chances of default.

2. Problem Statement

The limited availability of labelled data in credit risk assessment presents a significant challenge, particularly in African financial markets, where access to comprehensive borrower information is often restricted. Traditional credit risk models, which rely heavily on labelled data, struggle to accurately assess loan delinquency risk, leading to higher default rates, and limited financial inclusion and reduced lending opportunities for businesses and individuals. These challenges hinder economic growth as financial institutions are forced to adopt conservative lending practices to mitigate risks associated with data scarcity. To address this issue, this research explores the use of semi-supervised learning techniques to enhance credit risk prediction by leveraging both labeled and unlabeled data. By incorporating advanced machine learning approaches such as self-training, graph-based label propagation, and generative models, the study aims to develop a more robust and efficient credit scoring system. This approach has the potential to significantly improve **loan** approval accuracy, expand credit access to underserved populations, and optimize risk management strategies for financial institutions.

3. Research objectives and questions

1.To predict loan delinquency with at least 80% accuracy by utilizing semi supervised learning techniques.

Questions:

How do generative models improve predictions compared to traditional models*?*

Generative models refer to the subset of state-of-the-art artificial intelligence techniques and models that are designed to generate new data samples or outputs that are similar to those seen in the training data. They create new content such as images, text, audio or other types of data. They are able to learn and capture the underlying patterns and structures in the training data and use this knowledge to produce realistic outputs. Traditional models focus on tasks like classification or prediction. Traditional models are prone to having limited historical data from existing clients which brings difficulties when predicting the credit worthiness of new market segments while generative models offers a future forward perspective by generating synthetic data for untapped customer profiles (Auro Boquin jan 1 2024). They can also handle missing values in datasets reducing theneed for manual data preprocessing and improving overall prediction accuracy.

How do generative models handle class imbalance in credit risk prediction? A problem where the number of instances in one class significantly outnumber the instances in another class in the case of credit risk is where the number of instances of good credit far exceeds the number of instances of bad borrowers. The use of cost sensitive learning is a method that assigns different weights or cost to different classes or instances based on their importance or impact for example, the cost of misclassifying a default as a non default. Cost sensitive learning can penalize the models more for making false negatives then false positives and adjust the decision threshold. (sam navin,Sanjay kumar).Use sampling techniques. They involve modifying the original dataset by either adding, removing or changing instances of the classes to create more balanced distribution.

2.Evaluate the effectiveness of semi supervised learning compared to traditional learning models Questions.

a. What is the computational efficiency of Semi-supervised learning models compared to traditional learning models?

- Semi supervised learning take full advantage of available information in the data and obtain the most accurate prediction. Semi supervised can give high accuracy ranging from 90-98 with just half of the training data. (research target 2024).
- Traditional models might be efficient but do not perform well with complex datasets.
- Semi supervised models have the ability to handle complex datasets (Nicolas suhadolnik ,2023)

b. Can semi-supervised learning models out perform traditional learning models in credit risk prediction when labelled data is scarce?

- Semi supervised learning models can indeed outperform traditional learning models even when labelled data is scarce. They are able to handle limited labelled data which is often the case in credit risk prediction.
- They have the ability to leverage both labelled and unlabeled data. In Traditional models they can be effective when labelled data is scarce but may not perform well.

3. Provide insights into the applicability of semi supervised learning techniques for financial modelling in southern Africa and broader African markets

Questions

a. What are the unique challenges and opportunities of applying semi supervised learning techniques to financial modelling in southern Africa?

- Limited labelled data and labelling is time consuming
- High dimensional heterogeneous data. The data may include demographics, financial and transactional data which may be difficult to handle
- Limited expertise and resources. Access to expertise and resources and funding might be limited

b. How do semi supervised learning techniques perform in predicting credit risk for small to medium sized enterprises in southern Africa?

- Semi supervised learning techniques have good results in small to medium sized enterprises.
- The techniques are useful when dealing with labelled data that is limited, and there are challenges like high uncertainty, difficult to reason, imbalanced data.

3. Literature review

Traditional methods are inadequate when faced with large scale, high dimensional and complex data. This is also the case when labelled data is scarce hence their performance is limited. Semi-supervised learning techniques aim to improve algorithms' ability to utilize unlabeled data conditions. This approach enhances model robustness by streamlining kernel selection and parameter tuning, addressing dimensionality challenges.

M. Li and Y. Fu (2022). In further research Shi et al (2022), introduced a Support Vector Regression (SVR) model with kernel function, demonstrating superior accuracy and efficiency compared to logistic regression (LR) and NN counterparts. And also iu (2018) conducted research on credit card clients and compared Support Vector Machine, kNearest Neighbors, Decision Tree, and Random Forest with Feedforward

Neural Network and Long Short-Term Memory

In other literature, a self-training method is adopted and combined with a convolutional neural network for classification and continuously improve the model prediction performance through an iterative process, B. Yan et al (2022). And also Mahajan et al (2022). Implemented the Light-GBM algorithm in credit risk assessment, showcasing enhanced accuracy compared to single decision tree models, along with efficient training and minimal memory consumption.

In some literature most writers heavily rely on linear regression models for example Shi et al (2022). Introduced a Support Vector Regression (SVR) model with kernel function, demonstrating superior accuracy and efficiency compared to logistic regression (LR) and NN counterparts which may not fully capture the complex and potentially non-linear relationships between risk preferences. We aim to address this by heavily relying on non-linear models such as

classifiers, ensemble methods and neural networks together with semi-supervised learning techniques such as co-training and graph labelling techniques the results showed that these semi-supervised models outperform these traditional methods. Some research conducted to analyze biasness in credit scoring for instance, Lyn et al. (2002) used logistic regression, a statistical technique for credit scoring that has proven successful and has replaced linear discriminant analysis.

Liu (2018) conducted research on credit card clients and compared Support Vector Machine, k-Nearest Neighbors, Decision Tree, and Random Forest with Feedforward Neural Network and Long Short-Term Memory. Their research was conducted in order to improve on earlier risk predictions. findings by Lien and Yeh (2009). Liu (2018) proposed to add two important factors, drop-out and long short-term memory, to neural networks in order to find their effect on improving accuracy and also solving the problem of overfitting.

Overall main gap as observed from existing research is that most researchers relied on linear models which may not fully capture the complex and potentially non-linear relationships between risk preferences. We aim to address the gap by using semi-supervised learning techniques including self-training and co-training to label the unlabeled data by using the concept of pseudo labels achieved by applying the techniques on nonlinear baseline models including, but not limited to, classifiers and ensemble methods.

5. Methodology

The following is an outline of the project methodology and methods as well as techniques we are going to use:

Data Acquisition and Preparation:

Data source*:* The dataset is an augmentation of collected financial datasets containing both labelled and unlabeled data with borrower attributes such as age, marital status, education level, household size.

Data size and Quality: Synthetic data will be added to the original dataset which has 356 rows and 30 columns. This combination aims to improve model generalization ability since it is improved as the data size increases. The generated data will be validated by comparing distributions of the synthetic data and original dataset obtained through statistical methods not limited to, chi-square test.

Preprocessing:

- Handling missing values and outliers
- Ensuring consistency.
- Perform feature encoding for example, one hot encoding to convert categorical features into numerical features.
- Normalize numerical features to improve model performance
- Split the data into labelled and unlabeled subsets while maintaining class distribution balance.

Exploratory Data Analysis (EDA):

Conduct analysis to identify trends and relationships between features and the target variable. This includes visualizations such as learning curves to assess the effect of additional labelled data, model comparison, pseudo label quality evaluation to verify the reliability of predicted labels, decision boundary visualization to assess learning capability.

Model Development:

2. Implement semi-supervised learning algorithms including self-training and cotraining to generate pseudo labels.

   i. We will use self-training because it is simple to use and will require less computational resources and co-training because it is more suited to our dataset which is an augmentation

of data from multiple sources and it will improve model robustness by combining different predictions from multiple classifier models trained to a different view of data.

ii. We will use both techniques for validation purposes and a range of choice in order to choose the method which produces better results.

iii. Finally, we will Train the model by combining labelled and unlabeled data to optimize learning.

Model Evaluation:

Performance metrics:

- Assess the model's performance using metrics like precision, recall, F1 score, and AUC-ROC.
- Benchmark results against supervised learning models to evaluate improvements.
- Ablation studies to evaluate the contribution of different components (e.g., selftraining vs. co-training).

Accuracy and Deployment:

- Fine-tune the model for optimal performance.
- Create a prototype system to demonstrate practical applications of the model.
- Implement interpretability tools to ensure model decisions are explainable including a visualization dashboard.

6. **Results and findings**

Statistical significance testing such as chi square tests and t-tests, can be used to determine if observed improvements in accuracy when enhancing credit risk prediction using semi supervised learning to determine if the observed improvements in accuracy are statistically significant

Chi-square test

A chi-square test is useful for categorical outcomes for a functional semi supervised model capable of accurately predicting credit risk. It helps assess whether a model's predictions are significantly associated with actual credit outcomes of the Statistical significance tests like the Chi-square to observe improvements in credit risk accuracy which are statistically significant by comparing the results of a new credit risk model to a baseline model, allowing you to assess whether the observed differences are likely due to chance or represent a true improvement in predictive power, based on a chosen significance (university, 2024). When analyzing categorical variables like loan approval status across different credit risk categories, a Chi-square test can be used to determine if the observed differences in classification accuracy between the new and old model are statistically significant.

By creating a contingency table showing the counts of correctly and incorrectly classified loans for both models and then performing a chi square test of independence with the null hypothesis and p value you can improve the sensitivity analysis to assess the model validation in data. (colade, 2017)

T-test

When comparing continuous variables like credit score distributions between the new and old model's classifications, a T-test can be used to assess if the mean difference in credit scores is statistically significant using

1. Null Hypothesis to typically state that there is no significant difference between the new and old models, which are aimed to reject if the test result is significant.
2. P-value which represents the probability of observing the data if the null hypothesis is true, where a low p-value will indicate that the observed difference is unlikely to be due to chance.

3. Data Splitting will ensure reliable results, by splitting your data into a training set to develop the models and a testing set to evaluate performance and perform the statistical tests on the testing data. (schechter, 2014)

T-test compares the average credit scores of the loans classified by the new model versus the old model using to see if there is a statistically significant difference, T-tests will help determine whether the difference in accuracy or other performance metrics between two credit risk models are statistically significant (nigam, 2024)

After model, training results will show only target variable that will bring overall conclusion on is loan delinquency. The results will be yes or no based on input parameter of the borrower that is model will be assessing different factors and come up with final decision

7. Timeline

week	**task**
Week 2 - 4	Literature review
Week 8 - 10	Data preprocessing
Week 10 - 13	Model training and evaluation

Table 1: project timelime.

8. Resources Required

1. Access to datasets containing labeled and unlabeled financial data.
2. Computing infrastructure for model training and testing.

3. Programming tools, including Python with libraries like Scikit-learn, PyTorch. And tensorflow

4. Reference materials, such as research papers, articles and books on semi-supervised learning techniques.

9. Conclusion

This project aims to address the challenges of credit risk prediction by leveraging semi supervised learning. The combination of labeled and unlabeled data promises to improve predictive performance and reduce reliance on extensive labeled datasets. The outcomes are expected to provide financial institutions with an innovative and efficient solution for credit risk assessment, with potential applications in other domains requiring similar data strategies. The project also highlights the importance of utilizing semi supervised learning techniques in the financial sector not only for credit risk prediction but other economic problems and solutions.

REFERENCES

M. Sun Jiang and Z. Xu. (2024). Applying Hybrid Graph Neural Networks to Strengthen Credit Risk Analysis.

M. Zhang, H. W. (2024). Leveraging Semi-Supervised Learning and Convolutional Neural Networks for Enhanced Image Analysis in Healthcare Applications. *Journal of Advanced Computational*, 45-62.

S. Lu, Z. L. (2023). Engineering Applications of Artificial Intelligence. *Scaling-up Medical Vision-and Language Representation Learning with Federated Learning*.

Xiao, Y. (2024). Self-Supervised Learning in Deep Networks. *: A Pathway to Robust Few-Shot Classification*.

Y. Wei, K. X. (2024). Financial Risk Analysis Using Integrated Data and Transformer-Based Deep Learning. *Journal of computer science and software application*, 1-8.

Karasan, A. (2021). Machine Learning for Financial Risk Management with Python. *United Kingdom: O'Reilly Media.*

Hájek, P. (2010). Credit Rating Modelling by Neural Networks. *United States: Nova Science Publishers.*

Chen, C.; Lin, K.; Rudin, C.; Shaposhnik, Y.; Wang, S.; Wang, T. A holistic approach to interpretability in financial lending:

Shih, D.H.; Wu, T.W.; Shih, P.Y.; Lu, N.A.; Shih, M.H.(2023) A Framework of Global Credit Scoring Modeling Using Outlier Detection

Zhang, Z.; Jia, X.; Chen, S.; Li, M.; Wang, F.(2022) Dynamic Prediction of Internet Financial Market Based on Deep Learning. Comput.

https://www.altexsoft.com/blog/semisupervised-learning/