*University of Essex*
**Department of Mathematical Sciences**

MA981: DISSERTATION

# A Comparative Analysis of Machine Learning Models for the Credit Card Default Assessment: Unveiling Performance Variations and Factors Impacting Predictive Accuracy

**Nichamon Chanphithak**
**2207472**

Supervisor: Dr Georgios Papamikos

August 24, 2023
Colchester

# Contents

# List of Figures

# List of Tables

# Abstract

Assessing credit card default is crucial for financial institutions seeking effective risk management. Accurate prediction of credit defaults can offer valuable insights for informed lending decisions, optimizing portfolio performance, and maintaining a healthy financial ecosystem. This paper aims to evaluate machine-learning models - traditional, deep-learning, and the Tab Transformer (a recent state-of-the-art model) - on different scenarios using three distinct datasets with varying sizes and features. To ensure result reliability and model robustness, evaluations are performed using various metrics. Additionally, experiments were conducted to understand the effects of pre-processing steps on each model's performance by applying them individually.

Our study indicates that the Tab transformer shows significant potential in outperforming other models when applied to two datasets: the American Express Credit Card and Default Credit Card dataset. The ROC scores for these datasets are 0.867 and 0.706, while the F1 scores are 0.749 and 0.532, respectively. However, for smaller datasets like the German Credit Card dataset, it appears that Random Forest or LSTM models may be better choices based on their higher ROC and F1 scores.

These findings underscore the significance of choosing a suitable model. When deciding on a model, considerations such as interpretability, complexity, dataset characteristics, and desired performance trade-offs should inform the decision-making process. It is essential to have an in-depth understanding of the strengths and limitations associated with each available model option. For instance, LSTMs are proficient at handling sequential data with temporal dependencies but may encounter challenges when dealing with the complex American Express Credit Card dataset due to issues like vanishing gradient problems that hinder their learning capability. In addition to selecting the appropriate model, it is important to consider preprocessing steps such as standardization and addressing the class imbalance. Our research indicates that these steps can greatly improve the accuracy of credit risk evaluation

results.

CHAPTER 1

# Introduction

Credit card default prediction is a vital area of study in finance and risk management. It involves the use of advanced statistical and machine learning techniques to determine the likelihood that a credit card holder will fail to make their required payments within specified timeframes. Default, in this context, refers to the inability to meet payment obligations on schedule. This understanding holds great significance across various aspects of the financial domain, playing a role in both individual financial security and overall economic sustainability. Accurate credit card default prediction plays a pivotal role in effective risk management for financial institutions such as banks and credit card companies. When customers fail to make timely payments, these entities face significant risks. By accurately predicting defaults, these institutions can adopt proactive strategies to manage their risk exposure, allocate resources efficiently, and formulate contingency plans aimed at minimizing potential losses. This proactive approach enhances stability within these institutions and contributes to the overall resilience of the financial ecosystem.

The integration of a reliable and precise machine learning model in credit card default prediction has significant advantages for the financial industry. This advanced model improves risk assessment, profitability, regulatory compliance, and customer satisfaction. One primary benefit is the transformation it brings to risk management strategies. Traditional methods often struggle to capture the complexity of customer behavior and economic trends accurately. In contrast, a robust and accurate machine learning model analyzes diverse data points, revealing hidden patterns and relationships. This thorough analysis empowers

financial institutions to detect early indicators of potential defaults promptly. As a result, resources can be allocated wisely, targeted interventions can be implemented effectively, and overall financial losses are minimized.

In the field of credit card default prediction, both traditional machine learning and deep learning models have played a significant role in improving predictive accuracy and assessing risk. These models provide different strategies that contribute to the challenging task of identifying possible default situations. Credit risk assessment has traditionally relied on conventional machine learning models such as logistic regression, decision trees, and random forests [1, 2, 3]. These models utilize established statistical techniques to analyze historical credit data to identify patterns and relationships. In recent works, deep learning models such as neural networks, convolutional neural networks and recurrent neural networks have gained popularity for predicting credit card defaults [4, 5]. These models are capable of processing large amounts of complex data and extracting intricate patterns. CNNs are well-suited for analyzing image and sequence data, making them useful in detecting anomalies or identifying patterns that may indicate potential defaults within transaction sequences. On the other hand, RNNs excel at capturing temporal dependencies by utilizing their memory of past inputs, allowing them to effectively model sequences of credit transactions over time.

A notable example of a successful application of deep learning in credit card default prediction is the work by [6], which demonstrated the effectiveness of artificial neural networks in capturing patterns and features within transaction data to predict defaults. While traditional machine-learning models are often preferred for their interpretability and ease of implementation, the deep-learning model offers automated feature extraction and representation learning from raw data. However, it should be noted that the deep-learning models can pose challenges due to their increased complexity and resource requirements, as well as potential overfitting concerns. As shown in some studies [7, 8], there may not always be a direct correlation between the performance of deep learning models and their outcomes in this particular context.

In addition to selecting the model architecture, pre-processing plays a vital role in credit card default prediction. Pre-processing involves transforming raw data into a structured format suitable for analysis by machine learning algorithms. This step is crucial as it improves the quality and effectiveness of predictive models. By cleaning and preparing the data, removing noise and addressing imbalances, we lay the foundation for training robust and

accurate predictive models. Models built on pre-processed data are more likely to generalize well to unseen data, enabling lenders to identify potential credit card defaults with greater accuracy and make informed decisions.

One widely used dataset in credit card default prediction research is the UCI Credit Card Default Dataset [9]. With its large number of instances, this dataset proves to be significant for analyzing patterns of default behavior and evaluating the performance of predictive models across different algorithms and techniques. Another frequently employed dataset is the UCI German Credit Dataset [10]. This dataset categorizes individuals based on a set of attributes as either good or bad credit risks. However, due to its limited size, knowledge derived from this dataset may also be restricted. In recent years, American Express has made available datasets for researchers to analyze credit risk scenarios. These datasets [11] contain a wide range of attributes and payment histories, providing a realistic representation of how credit card users behave. The use of diverse datasets in previous studies highlights the multi-faceted nature of predicting credit card default. This variety allows researchers to test their models on various scenarios and demographics, leading to a more comprehensive understanding of credit risk.

In this study, a comparative analysis will be conducted using eight machine learning models. The models selected for the analysis include traditional, deep-learning, and transformer models. The performance of these models will be evaluated using various metrics - accuracy, precision, recall, F1 score, and ROC - to ensure their reliability and robustness. Furthermore, an examination of the pre-processing steps will be carried out to understand their impact on model performance.

The paper is structured as follows: Section 2 presents a literature review on the topic. relevant literature. Section 3 describes the methodology and model architecture used in the study. In Section 4, we discuss the data employed, our experimental design, and present the results of our model predictions. Finally, in Section 5, we analyze and discuss our empirical findings.

CHAPTER 2

# Literature Review

This literature review delves into the wide array of classification models utilized in classification analysis. By examining different methodologies, this review seeks to offer a comprehensive comprehension of how predictive modeling has progressed in evaluating credit risk. The reviewed models are grouped into three primary classes: traditional models, deep learning models, and transformer models. Each class presents a distinct approach to predicting the classification datasets, influenced by advancements in machine learning techniques and data accessibility. In this review, we assess the strengths and limitations of various model categories for the classification tasks. We analyze their performances, interpretability, and computational complexities with academic rigor. In the upcoming sections, we aim to provide a comprehensive analysis of the existing literature on credit risk analysis. We will discuss studies that have explored traditional models as well as deep learning architectures and transformer models. By synthesizing knowledge from these studies, our goal is to offer an overview of advancements in classification datasets, specifically focusing on credit card default predictions. This analysis aims to facilitate informed decision-making and inspire future research in this important field.

## 2.1 Introduction to Classification Models

Classification is a fundamental concept in machine learning and data mining. It has numerous applications across various domains such as image recognition, natural language

processing, the financial industry, e-commerce product prediction, and medical diagnosis. The primary purpose of classification is to categorize input data into different classes based on their inherent patterns and features. This process plays a crucial role in gaining insights, making informed decisions, and automating complex tasks by enabling computers to accurately recognize and label data similar to human capabilities. Accurate classification and categorization are crucial in the current era where there is a vast amount of available data. These models play a significant role in differentiating between legitimate emails and spam, detecting potential fraud within financial datasets, utilizing images for medical diagnostics, or predicting consumer preferences in e-commerce scenarios. For instance, tasks like object detection and facial recognition heavily rely on classification models to assign images into various categories. In the field of natural language processing, these models find applications in sentiment analysis, spam identification, and text classification processes. These well-developed models serve as fundamental pillars for intelligent systems that drive our modern technological landscape.

## 2.2 Traditional Classification Models

Classification is a fundamental task in the field of machine learning, involving the assignation of data points to pre-defined categories or classes. Traditional classification models are foundational algorithms used widely across various domains for many years [12]. These models are highly regarded for their established methodologies and often serve as the foundation upon which more advanced techniques are built.

Traditional classification models encompass different types, each possessing unique characteristics and applications. In the subsequent paragraphs, we will provide descriptions of each classification model alongside potential areas where they can be applied.

Logistic regression [13] is a widely utilized statistical method that allows for the modeling of associations between binary dependent variables and independent variables. Its applicability spans various fields, including medical diagnosis and credit risk assessment. Though logistic regression encompasses multiple techniques, one means of assessing predictor effects involves the calculation of standardized regression coefficients. These coefficients offer insight into how each predictor influences the outcome while keeping all other predictors constant [14]. Nonetheless, it is important to acknowledge that these coefficients may not necessarily

reflect conventional measures of predictor importance. To evaluate its strengths compared to alternative modeling methods, researchers have undertaken comparisons between logistic regression and other approaches. In a comparative analysis conducted [15], the accuracy of the random forest model was compared to that of the logistic regression approach. Results revealed that it is possible to directly assess whether or not the random forest model outperforms logistic regression in terms of accuracy by utilizing both techniques simultaneously. Consequently, this indicates that there are scenarios where employing logistic regression for modeling binary outcomes may not yield optimal results and alternative approaches such as random forest should be contemplated instead.

Decision trees [16] and random forests [17] are widely used classification models due to several advantages they offer. One key advantage is their ability to handle both categorical and numerical features, making them suitable for a wide range of datasets. Moreover, these models exhibit interpretability, allowing decision-making processes to be represented clearly and intuitively. Decision trees provide an easily understandable representation of the underlying decision rules, enhancing model transparency. On the other hand, random forests build upon this concept by addressing overfitting issues commonly associated with individual decision trees. By combining multiple decision trees through ensemble learning techniques, random forests improve generalization performance while maintaining interpretability. To further optimize the performance and generalization capabilities of these models, pruning techniques can be applied effectively. It is crucial to keep in mind that computational requirements should be considered when applying tree-based models practically. Tree-based models have proven to be highly useful in a wide range of fields. One such example is the application of a random forest algorithm in leveraging molecular structure data [18] for accurate predictions. Another study [19] has shown that Random Forest performs exceptionally well in compound classification and quantitative structure-activity relationship modeling, demonstrating its effectiveness across different domains within cheminformatics. The inherent strengths of Random Forest lie in its ability to handle both numerical and categorical variables, contributing to its high accuracy and making it an invaluable tool for predictive analysis.

Support Vector Machines (SVMs) are a widely utilized machine learning algorithm employed for classification tasks. They construct a linear decision boundary in feature space with high dimensions, enabling the separation of different classes [20]. SVMs possess notable

capabilities in handling both nonlinear relationships between features and high-dimensional data. This is achieved by utilizing diverse kernel functions, such as polynomial, radial basis function, or sigmoid kernels. The successful application of SVMs encompasses various domains including medicine and finance. In the medical field, SVMs have been effectively utilized for disease diagnosis, prediction of treatment outcomes, as well as analysis of medical images. Furthermore, within the realm of finance, they have exhibited their value in credit scoring and fraud detection tasks [2]. By employing SVM methodologies to evaluate bank customers' creditworthiness accurately and assessing risks associated with loan approval. There have been efforts made to compare Support Vector Machines with other modeling techniques. For instance, a study by [21] investigated various machine learning algorithms, including SVM, for the detection of five typical stenches using an electronic nose. Additionally, in another research paper [22], different machine learning techniques were compared, and SVM was one of them used for predicting disruptions in tokamak reactors. Overall, SVMs are known as powerful machine learning algorithms that can be applied across a wide range of domains. They excel at handling high-dimensional data and capturing nonlinear relationships. The performance of SVMs is greatly influenced by the choice of kernel function and tuning parameters; therefore it is crucial to carefully select and optimize these aspects to achieve accurate and reliable results.

In summary, a wide range of traditional machine-learning models are available for classification tasks. These include Logistic Regression, Support Vector Machines, Decision Trees, and Random Forests. Each model has its unique characteristics and can adapt to different challenges. Many studies [23, 24] have been conducted to evaluate the performance of these traditional models to determine their suitability. These evaluations have utilized various datasets that represent the complexity and diversity found in real-world applications. By employing a comprehensive set of metrics including accuracy, precision, recall, and F1-score, a detailed analysis was carried out regarding the capabilities of each model. The findings from these extensive assessments provide valuable insights into how well each model performs under different circumstances. This combination of research efforts enhances our understanding of the strengths and weaknesses of traditional models, enabling us to choose the most suitable approach for classification tasks. Consequently, we can develop more precise, efficient, and contextually appropriate classification solutions.

## 2.3 Deep Learning Models

The emergence of deep learning has resulted in notable advancements in the domains of machine learning and artificial intelligence. These sophisticated models are capable of autonomously extracting complex features from data, resulting in exceptional performance when it comes to classifying various types of information. The objective of this paper is to amalgamate critical findings derived from numerous studies that examine and evaluate deep learning models for classification objectives. Specific attention will be given to their methodologies, applications, and comparative analyses.

Deep learning models, specifically Convolutional Neural Networks, have shown impressive capabilities in image classification. Previous work [25] paved the way for utilizing deep CNNs and achieving remarkable results on the ImageNet dataset. This breakthrough sparked a significant amount of research focused on improving CNN architectures, resulting in advancements like VGGNet, GoogLeNet, and ResNet. These models, with their complex hierarchical structures, consistently outperform traditional methods by accurately identifying objects, scenes, and patterns in images. Unlike traditional methods relying on manual feature engineering, CNNs can automatically learn discriminative features through multiple layers of convolution and pooling. In recent research, the authors of [26] investigate the potential use of convolutional neural networks in classifying non-image data. They propose a unique method that adapts CNN architectures originally designed for image analysis to handle sequential, time-series, and tabular datasets. By transforming non-image data into two-dimensional formats, they exploit the feature extraction capabilities of CNNs. This study contributes to broadening the applications of CNNs beyond images and highlights how deep learning techniques can be leveraged for diverse data modalities in classification tasks.

Recurrent Neural Networks and their specialized variant, Long Short-Term Memory Networks [27], are highly effective for classifying sequential data. They excel at capturing temporal dependencies and intricate patterns within sequences. This literature review examines the utilization and performance of RNNs and LSTMs in classification tasks, synthesizing key findings from various studies. It provides insights into the methodologies used, applications explored, as well as comparative analyses conducted. Recurrent neural networks have proven to be effective in various classification tasks that involve sequences. They are equipped with memory cells that store information about previous inputs, enabling them

to retain long-range dependencies. The potential of RNNs was originally highlighted by research, which introduced a novel architecture for recognizing handwritten text. This breakthrough spurred the development of different variants of RNNs, including long short-term memory networks, which address the vanishing gradient problem and improve learning capabilities for sequence data. LSTMs, proposed by [27], introduce memory cells and gating mechanisms that enable selective retention and updating of information, effectively mitigating the vanishing gradient issue. RNNs and LSTMs have wide-ranging applications in various domains, including natural language processing tasks like sentiment analysis, time-series forecasting, and speech recognition. A thorough review conducted by experts [28] provides insights into the performance of different RNN and LSTM models in sequence prediction tasks, highlighting their effectiveness in capturing complex relationships within sequences. Moreover, comparative studies such as the one carried out by researchers [29] have compared different variants of LSTM models in terms of their ability to handle long-term dependencies. In the field of medical diagnostics, RNNs and LSTMs show promise for disease prediction and patient monitoring based on electronic health records. These models excel at analyzing sequential data to extract subtle temporal patterns that contribute to accurate predictions. In conclusion, recurrent neural networks and long short-term memory models have become essential tools for sequence-based classification tasks in various domains. Their capacity to capture temporal dependencies and nuances within sequential data has led to their widespread application in fields such as natural language processing and medical diagnostics. Comparative analyses have identified differences in performance and architecture, offering valuable insights for practitioners aiming to utilize these models effectively. Ongoing research efforts continue to enhance the capabilities of RNNs and LSTMs, making them indispensable contributors to advancing classification tasks involving sequential data.

Deep learning models have significantly transformed classification efforts in various domains such as images, text, and medical data. Through detailed comparative studies, researchers have illuminated the strengths and limitations of different architectural paradigms, enabling professionals to choose suitable models for specific classification needs. Ongoing exploration continues to propel deep learning models forward in enhancing classification capabilities by autonomously identifying complex patterns and representations from diverse datasets.

## 2.4 Transformer Models

Transformer models have gained significant attention in recent years due to their groundbreaking impact on various natural language processing tasks, particularly classification. This literature review offers an overview of studies that explore the use and effectiveness of transformer models for classification purposes, providing insights into their methodologies, applications, and comparative analyses. The introduction of the Transformer architecture by [30] revolutionized NLP through its attention mechanism, enabling the model to capture contextual relationships in sequences. This innovation quickly expanded to include classification tasks as researchers capitalized on transformer models' ability to efficiently process sequential data and capture long-range dependencies. BERT and its subsequent variants emerged as prominent models introducing masked language modeling techniques while enhancing contextual embeddings for classification tasks. GPT-3 further showcased the power of transformers, achieving state-of-the-art performance on various language tasks, including text classification.

The use of transformer models expands beyond text, as they have proven to be useful in image classification tasks. One notable example is the Vision Transformer, which was introduced by [31]. ViT partitions images into patches and treats them as tokens, allowing transformers to process image-based data. This application has opened up possibilities for using transformer-based models in multimodal classification scenarios that involve diverse data types.

Notably, the impact of transformer models has expanded beyond text and images to include tabular datasets. Researchers have demonstrated the effectiveness of transformer architectures in capturing relationships within tabular data, which is typically addressed through gradient boosting and linear models. The TabTransformer [32] serves as a prominent example, as it applies transformer principles to improve tabular classification tasks. This extension highlights the versatility of transformers, enabling them to uncover complex patterns within structured datasets.

In summary, the rise of transformer models has revolutionized classification tasks in both natural language processing and image processing. These models excel at capturing context and dependencies, leading to superior performance in classification compared to traditional approaches. Notable examples include BERT, GPT-3, and Vision Transformer. Comparative

studies have provided valuable insights into the differences between transformer variants and their traditional counterparts, assisting practitioners in choosing suitable models for their needs. As research progresses, transformer models are expected to continue shaping the classification field by offering improved accuracy and generalization across various data domains.

## 2.5 Challenges in Classification and Strategies for Handling Normalization and Imbalance

In classification tasks, there are various challenges that arise in data preprocessing and model performance. This literature review specifically examines the challenges encountered in classification tasks, with a focus on two important aspects: data normalization and imbalance handling. It discusses the methodologies proposed to address these challenges and highlights their significance and effectiveness.

Data normalization is a challenge due to the varying scales and distributions of input features. Normalization techniques aim to standardize features by ensuring they have similar ranges, which helps improve model convergence and stability. Commonly used methods such as z-score normalization and min-max scaling show promise in enhancing classification performance. Studies like [33] emphasize the impact of normalization on classification models, highlighting its role in promoting uniform convergence during optimization.

Imbalance, characterized by significantly unequal class distributions, poses a challenge in classification tasks. This issue can result in biased training and predictions. Various specialized techniques have been proposed to address this problem. Resampling strategies such as oversampling and undersampling aim to adjust the class proportions. Hybrid approaches that combine normalization techniques with imbalance handling methods have shown comprehensive effectiveness. For example, a recent study [34] proposed combining synthetic oversampling with Z-score normalization to simultaneously address both imbalance and variance issues. Additionally, deep learning models like Generative Adversarial Networks offer promise in generating synthetic samples for addressing data imbalances [35].

In conclusion, addressing class imbalance and data normalization is crucial in classification tasks. Extensive literature offers various techniques to tackle these challenges, highlighting their importance in improving classification outcomes. The combination of normalization and

imbalance handling methods has the potential to enhance classification performance across different applications, shaping the future direction of classification research and practice.

## 2.6 Credit Card Default Prediction

Credit card default prediction is a critical task in the financial industry, intending to forecast the probability of borrowers failing to make their credit card payments. This literature review explores the body of research dedicated to using machine learning models for credit card default prediction. It investigates various methodologies employed in early studies, including decision trees, support vector machines, and artificial neural networks [36], as well as ensemble methods that combine multiple models to improve accuracy. These techniques have become essential tools for proactive risk management within financial institutions. Boosting techniques like AdaBoost [37] have been utilized to enhance both predictive accuracy and generalization performance. The emergence of deep learning has had a profound impact on credit card default prediction. Deep neural networks, specifically Long Short-Term Memory networks [38], have shown promising results in capturing the temporal patterns and complex relationships present in credit data. Effective feature engineering continues to play a crucial role in improving predictive models. Recent research studies [39, 40] highlight the significance of selecting and constructing relevant features to enhance the accuracy of default predictions. It is important to consider the impact of imbalanced data, where non-default instances outnumber defaults. Various techniques such as oversampling and hybrid models [41] have been used by researchers to address this class imbalance issue and ensure balanced model training. Comparative studies in the literature have provided valuable insights into the effectiveness of different machine learning models for credit card default prediction. For example, [3, 42] compared SVMs, decision trees, and ensemble methods on datasets related to default credit cards, highlighting their respective strengths and weaknesses. In conclusion, extensive research has been conducted on various machine-learning models for credit card default prediction. From traditional approaches such as decision trees and support vector machines to more advanced techniques like ensemble methods and deep learning architectures, a wide range of strategies have been utilized to improve the accuracy of default predictions. Comparative analyses and feature engineering efforts also demonstrate the practical importance of these models in assisting financial institutions with credit risk

management and informed lending choices.

With the emergence of the Transformer and the Tab Transformer models in classification tasks, specifically on credit card default datasets. These models have emerged as state-of-the-art for various classification tasks, but a comparative analysis is lacking. By conducting experiments using multiple credit card default datasets and comparing traditional and deep-learning models with different evaluation metrics, we can provide a comprehensive comparison. The analysis will shed light on the performance of these models and help us understand the underlying factors contributing to their effectiveness.

CHAPTER 3

# Methodology

In this section, we present the research methodology used in our study and provide an overview of the machine learning models selected for comparison along with their corresponding preprocessing steps. The objective of this section is to offer a comprehensive understanding of the experimental setup and technical specifics associated with each model.

## 3.1 Preprocessing Steps

The preprocessing phase plays a critical role in preparing the dataset for effective model training and classification. In our study, we employed two key preprocessing steps: standardization and imbalance handling through class-weight training.

### 3.1.1 Standardization

Standardization [43, 44], also known as z-score normalization, is a commonly used preprocessing technique in machine learning. The goal of standardization is to transform the features of a dataset onto a common scale by subtracting their mean and dividing by their standard deviation. This process helps ensure that all features have comparable magnitudes and exhibit a mean of zero and a standard deviation of one.

$$z = \frac{x - \mu}{\sigma} \tag{3.1.1}$$

where: $z$ is the z-score, $x$ is the raw score, $\mu$ is the mean of the distribution, and $\sigma$ is the standard deviation of the distribution.

The benefits of standardization are particularly evident when using gradient-based optimization algorithms for machine learning tasks. These algorithms tend to be sensitive to the scale of input features, so using standardized data prevents features with larger scales from dominating the optimization process. By doing so, it enables the model to converge more quickly and efficiently during training.

Another advantage offered by standardization is improved stability and generalizability across different features in a dataset. When applying this preprocessing step consistently throughout all variables, it ensures that models perform consistently regardless of individual feature characteristics.

### 3.1.2 Imbalance Handling

Class imbalance [45, 46, 47] is a common problem in classification tasks where one class greatly outweighs the others. This can cause biased model performance as the algorithm tends to prioritize the majority class due to its higher frequency. To handle this issue, we utilized class-weight training, a technique that assigns different weights to classes during training.

In class-weight training [48, 49], instances from the minority class are given higher weights while instances from the majority class are given lower weights. This encourages the algorithm to correctly classify instances from the minority class because misclassifying them would result in larger penalties. By adjusting these weight assignments, we effectively balance each class's influence during training which leads to a more sensitive and accurate model specifically designed for capturing patterns of underrepresented classes.

This preprocessing step is particularly crucial when dealing with imbalanced datasets, as it helps the model achieve better precision and recall for the minority class. It enhances the model's ability to correctly identify instances from the underrepresented class, thereby improving overall classification performance and reducing the bias towards the majority class.

## 3.2 Classification Models

In our study, we compared a range of classification models to predict and classify instances within the given datasets. Each model brings distinct strengths and characteristics to the table, allowing us to comprehensively assess their performance across various classification tasks.

### 3.2.1 Logistic Regression

Logistic Regression [13], a straightforward and easily interpretable algorithm commonly employed for binary classification tasks, estimates the probability of an instance belonging to a particular class by applying the logistic function to a linear combination of input features. This results in a score that falls between 0 and 1. To make the final decision on a class assignment, Logistic Regression applies a threshold. The simplicity and interpretability of this algorithm have contributed greatly to its popularity as it serves as an ideal baseline model. Logistic Regression exhibits excellent performance when there exists an approximately linear relationship between features and target classes, making it particularly suitable for problems with relatively simple decision boundaries.

### 3.2.2 Decision Tree and Random Forest

Decision Trees [16] are powerful models used for decision-making based on a set of attribute tests. They divide the feature space into subsets, capturing complex interactions among features and effectively handling both numerical and categorical data. Each branch in the tree represents a decision path that leads to a leaf node with an assigned class label prediction. However, Decision Trees can be prone to overfitting when trained on small datasets due to their ability to create complex decision boundaries.

Random Forests [17], on the other hand, is an ensemble method that improves predictive performance and mitigates overfitting by combining multiple Decision Trees. Each tree is trained on a different subset of the data, and their predictions are averaged or voted upon for final output. Random Forests excel at handling high-dimensional data and non-linear relationships more effectively than individual Decision Trees while providing better generalization capabilities.

### 3.2.3 Support Vector Machine (SVM)

Support Vector Machines [20] are robust algorithms used for classifying data into binary or multi-class categories. SVM aims to identify the hyperplane that maximizes the separation between distinct classes while minimizing classification errors. This algorithm is particularly effective in high-dimensional spaces and can accurately capture intricate, non-linear decision boundaries by utilizing kernel functions. It should be noted that SVMs might be sensitive to hyperparameters and may necessitate meticulous tuning to achieve optimal performance.

### 3.2.4 Neural Networks

Neural Networks [50, 51], inspired by the human brain, are composed of interconnected layers of nodes called neurons that process and transform input data. They excel at capturing intricate relationships and patterns within data, which makes them suitable for complex tasks. Deep Neural Networks take this capability further by utilizing multiple hidden layers to learn hierarchical representations. However, their effectiveness comes at the cost of significant computational resources and a higher risk of overfitting when dealing with small datasets.

### 3.2.5 Convolutional Neural Networks (CNNs)

Convolutional Neural Networks [52] are a specific type of neural network that is specifically designed to analyze and process grid-like data, such as images. Their architecture includes convolutional layers which are capable of automatically extracting hierarchical features from the input data. Originally developed for image processing tasks, CNNs can also be adapted to handle tabular data with patterns resembling spatial structures. They excel in capturing and understanding spatial relationships between different elements within the dataset, as well as identifying interactions between various features present in the data.

### 3.2.6 Long Short-Term Memory (LSTM)

Long Short-Term Memory models [27] belong to the family of Recurrent Neural Networks, which are designed to handle sequential data and temporal dependencies. LSTMs can capture long-range relationships and memory patterns, making them particularly well-suited for tasks involving time series or sequences of data. However, in scenarios where high-

dimensional tabular data is involved, LSTMs may face certain challenges like the vanishing gradient problem.

### 3.2.7 TabTransformer

TabTransformer [32] is a recent development in machine learning that offer a novel methodology for dealing with tabular data. They draw inspiration from the effectiveness of attention mechanisms used in natural language processing and utilize self-attention techniques to capture detailed connections and dependencies within tabular datasets. By implementing this innovative architecture, TabTransformer has proven to be particularly successful in handling intricate tasks related to complex tabular data, such as credit risk assessment, customer churn prediction, and fraud detection.

The key idea behind TabTransformer is the use of attention. Unlike conventional machine learning models that rely on predefined feature engineering, TabTransformer can automatically learn the importance of each feature and how it interacts with others. This is made possible through self-attention mechanisms, which enable the model to evaluate and assign weights to different features based on their contextual relevance within the data. This flexibility allows TabTransformer to excel in handling high-dimensional tabular datasets with a large number of features, where capturing intricate feature interactions using traditional methods can be challenging.

The TabTransformer architecture, as shown in Figure 3.1, is composed of several layers, each containing attention and feedforward neural networks. Through the training process, the model acquires the ability to assign higher attention weights to relevant features while diminishing the influence of less informative ones. This dynamic feature weighting capability empowers TabTransformer to effectively capture intricate patterns and relationships that might not be easily identified through conventional feature engineering techniques.

One important advantage of TabTransformer is their ability to effortlessly handle both categorical and numerical features. In traditional models, extensive preprocessing is often required to convert categorical variables into numeric representations. However, with TabTransformer, categorical features can be processed directly using learned embeddings. This eliminates the need for manual feature engineering and simplifies the modeling process.

Additionally, TabTransformer demonstrate remarkable resilience in dealing with noisy data and missing values. The attention mechanisms employed by these models enable them

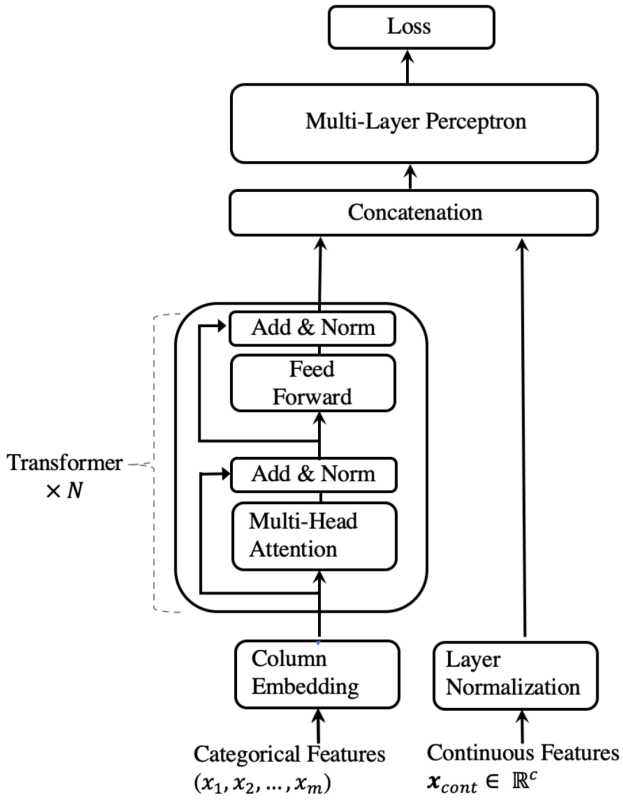


Figure 3.1: The TabTransformer model architecture.

to focus on relevant features even when confronted with noise or incomplete data points. As a result of this robustness, these models exhibit improved generalization and performance when applied to real-world datasets that may contain imperfections.

Recent benchmark studies have demonstrated the impressive performance of TabTransformer in various classification tasks. These transformers outperform traditional models and achieve state-of-the-art results. The effectiveness of TabTransformer lies in their ability to capture complex relationships, handle high-dimensional data, and adapt to diverse tabular datasets. As a result, they offer great potential for practitioners dealing with intricate tabular classification problems.

CHAPTER 4

# Result

In this section, we provide a thorough overview of our experiments and analysis. We examine the relationship between various classification models, preprocessing methods, and evaluation metrics using three different credit card default datasets. Our study aims to determine the most effective machine learning model for predicting credit card defaults while also investigating the influence of preprocessing techniques and handling class imbalance issues. The following subsections outline our experimental setup, discuss the performance of each model, and highlight key findings from each dataset.

## 4.1 Experimental Setting

This subsection presents the key components of our experimental framework that facilitated a systematic and rigorous evaluation of machine learning models in predicting credit card defaults. Our methodology entailed carefully selecting datasets, refining hyperparameters, and employing a comprehensive set of evaluation metrics. By establishing an extensive experimental setup, we ascertain the robustness and reliability of our analysis of model performance.

### 4.1.1 Datasets

Our study leverages three distinct credit card default datasets, each sourced from real-world credit card transactions, providing a diverse range of credit-related attributes and default

occurrences. The datasets utilized are as follows.

**Default of Credit Card Clients Dataset**

The Default of Credit Card Clients Dataset [9] is a comprehensive collection that encompasses diverse attributes. It includes demographics, credit information, repayment history, and bill statements of credit card clients from Taiwan between April 2005 to September 2005. This dataset has been widely used in binary classification tasks to predict instances of defaulting on credit card payments. With a total of 30,000 instances and spanning 23 features, this dataset provides a detailed representation of individual profiles and transaction histories. By considering these features collectively, it becomes possible to gain insights into the financial behavior patterns exhibited by credit card holders. The target variable of interest in this study is a binary outcome that indicates whether or not a credit card holder defaulted on their payments the following month. When examining the dataset, it becomes apparent that there exists an imbalance between default instances and non-default instances, with more instances falling into the non-default category. This class imbalance poses challenges to model training and evaluation as it highlights the importance of employing appropriate techniques for handling such imbalances.

**American Express Credit Card Dataset**

The dataset of American Express Credit Cards [11] comprises a total of 5,531,451 distinct instances that correspond to unique credit profiles and transaction histories. This dataset includes various characteristics related to Delinquency, spending patterns, repayment behavior, balance information as well as risk indicators in a total count of 191 features. The purpose behind this competition is to build predictive models aimed at determining the probability that a customer will not fulfill their payment obligation for the outstanding credit card balance in future months based on their monthly customer profile data. To create the target variable for this task, a performance window lasting 18 months after the most recent credit card statement was examined. If a customer fails to pay off the due amount within 120 days following their latest statement date, it is considered an instance of default.

To streamline the model training and evaluation process given limited time and resources, a downsampling technique was employed in this study. By using this technique, an expanded dataset with 300,000 instances was obtained which is ten times larger than the original Default

| Dataset | Instances | Features | Label distribution | |
|---|---|---|---|---|
| | | | Positive | Negative |
| Default of Credit Card Clients | 30000 | 23 | 6636 | 23364 |
| American Express Credit Card | 300000 | 191 | 74007 | 225993 |
| South German Credit | 1000 | 20 | 700 | 300 |

Table 4.1: Summary of Dataset Characteristics and Label Distribution.

Credit Card Clients Dataset. The aim of creating such an enlarged dataset was to enhance its size to achieve robust performance of the model. To ensure that the reduced dataset remains representative despite being downsized, stratified sampling techniques were applied during data reduction procedures. This approach guarantees that the proportionality between default cases is preserved.

#### South German Credit Dataset

The South German Credit dataset [10] consists of 1,000 instances, representing 700 good credits and 300 bad credits. It includes 20 predictor variables and encompasses data from the years 1973 to 1975. This dataset offers valuable insights into predicting credit card default by including two crucial attributes - personal demographics like age and gender, along with various financial indicators. The dataset was created using a stratified sampling technique that heavily oversampled bad credit instances from actual credits. Furthermore, it is important to acknowledge that despite its relatively smaller size compared to other available datasets, this dataset holds significance in understanding model performance as small variations within these cases can potentially influence predictions due to factors such as initializations and random seeds.

### 4.1.2 Dataset Splitting: Training, Validation, and Test Sets

This section presents the process of dividing the dataset into three subsets: a training set, a validation set, and a test set [53, 54]. Each subset has its specific role in our research methodology. The purpose of this allocation is to ensure the development of robust models, effective refinement processes, and unbiased evaluations. This approach greatly improves the reliability and generalizability of our findings by carefully reasoning and determining

appropriate ratios for data allocation within each subset.

**Training Set (70%)**

The training set is a crucial component of model training and acquisition as it constitutes the majority of the dataset. By having a large number of data instances, models can capture essential patterns, associations, and attributes that characterize credit card default tendencies. The sizeable nature of the training set facilitates precise parameter estimation and allows for effective adjustments in modeling.

**Validation Set (10%)**

In the model selection and hyperparameter tuning process, the validation set holds significant importance. This particular subset is kept separate from the training phase to establish an unbiased evaluation framework for assessing various model configurations. The performance of models on the validation set helps determine the best parameters that result in improved predictive accuracy while preventing overfitting issues.

**Test Set (20%)**

During the training and validation phases, the test set remains unchanged to serve as an unbiased measure for evaluating model performance in real-life situations. This final evaluation provides an impartial estimation of the model's predictive capacity for credit card defaults. The larger allocation given to the test set ensures a more accurate representation of the model's ability to generalize.

### 4.1.3 Hyperparameter Tuning

Hyperparameter optimization is crucial in maximizing the effectiveness of machine learning models. Each chosen model's hyperparameters were carefully adjusted to strike a balanced trade-off between complexity and generalization ability. Through an iterative approach, various combinations of hyperparameters were systematically explored and evaluated on an independent validation set. This process ensured that the selected machine learning model achieved optimal performance in credit card default prediction. To determine the best combination of hyperparameters, we trained each model using a designated training subset

and assessed its performance on another subset called a validation set. The configuration that yielded the highest performance metrics on the validation set was considered optimal for that specific model. In this study, a process called hyperparameter tuning was conducted for each of the models. This process involves adjusting various parameters that control how the model learns and makes predictions to find their optimal values. Below are the best-tuned hyperparameters obtained for each model:

- Logistic Regression: regularization strength (C) = 1.0 with an L2 penalty.
- Decision Tree: minimum sample split = 2, minimum samples per leaf = 1.
- Random Forest: number of estimators = 100, minimum sample split = 2, minimum samples per leaf = 1.
- SVM: regularization parameter (C) = 1.0, kernel type = rbf.
- Neural Networks: hidden layers = [32, 8], activation function = relu.
- CNN: filter sizes = 32, kernel sizes = 3.
- LSTM: number of LSTM units = 32, dropout rate = [0.1].
- TabTransformer: number of transformer layers = 4, attention heads = 8, embedding size = 16, dropout rate = [0.2].

### 4.1.4 Evaluation Metrics

To comprehensively assess the predictive abilities of our models, we utilized a varied range of evaluation metrics. Each metric offers distinct perspectives on different aspects of model performance, allowing for a comprehensive understanding. These metrics provide valuable insights into how well our models can generalize to real-world credit card default scenarios.

#### Accuracy

Accuracy is a measure that indicates the percentage of correct predictions made by a model in comparison to the total number of instances in a dataset. It is calculated by dividing the number of accurate predictions by the overall number of instances. A higher accuracy value suggests that the model has better predictive capabilities. Nonetheless, when dealing with

imbalanced datasets where one class predominates over others, relying solely on accuracy as an evaluation metric can result in misinterpretations because it tends to favor majority classes. Therefore, other metrics such as precision, recall and F1 score should also be taken into account for evaluating models' performance under such circumstances.

$$\text{Accuracy (\%)} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \times 100 \tag{4.1.1}$$

**Precision**

Precision is a metric that measures the ratio of correctly predicted positive instances to all positive predictions made by the model. Precision becomes especially important in situations where false positives incur significant costs. A higher precision signifies a lower occurrence of false positives, indicating the model's competence in accurately identifying positive cases. It can be calculated using the formula:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{4.1.2}$$

**Recall (Sensitivity or True Positive Rate)**

The recall is a performance metric that measures the ability of a model to correctly identify positive instances. It is defined as the ratio between true positives (the number of correct positive predictions) and all actual positive cases (true positives plus false negatives). This metric becomes especially important when it is crucial to capture as many positive instances as possible due to potential consequences associated with missing them. A higher value for recall indicates that a model has a greater capability to accurately detect genuine positive instances. In other words, a higher recall score suggests that the model effectively minimizes false negatives by identifying more true positives.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{4.1.3}$$

**F1 score**

The F1 score is a metric that provides a balanced measure of the performance of a model. It represents the harmonic mean between precision and recall, which are two important

evaluation measures in classification tasks. The formula to calculate the F1 score is as follows:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.1.4}$$

This metric considers both false positives and false negatives, making it particularly useful when maintaining a balance between precision and recall is crucial. By incorporating information from true positives, true negatives, false positives, and false negatives into its calculation, the F1 score provides an overall assessment of how well a model performs across all categories or classes in imbalanced datasets.

**ROC (Receiver Operating Characteristic) Curve and AUC (Area Under the Curve)**

The ROC curve graphically depicts the trade-off between the true positive rate (recall) and the false positive rate. By adjusting the classification threshold values, this curve is constructed, capturing these rates at various thresholds. The area under this curve serves as a summary measure of discrimination capabilities: higher AUC values indicate better proficiency in distinguishing positive from negative instances. An AUC value of 0.5 signifies random guessing ability, while an AUC of 1.0 denotes perfect discriminatory performance.

$$\text{AUC-ROC} = \int_0^1 \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \, d\left(\frac{\text{False Positives}}{\text{False Positives} + \text{True Negatives}}\right) \tag{4.1.5}$$

To comprehensively assess the predictive capabilities of each model in credit card default prediction across multiple datasets, we can employ a diverse set of evaluation metrics. These metrics will allow us to gain valuable insights into the strengths and constraints exhibited by each model. In subsequent sections, we will present an analysis of the performance outcomes attained by each model, thereby facilitating the identification of the most suitable approach for credit card default prediction.

## 4.2 Model Performance

In this section, we thoroughly evaluate the performance of different machine learning models on diverse datasets with unique credit risk scenarios. Through careful evaluation using important metrics, we gain valuable insights into their predictive abilities and limitations.

By examining various evaluation metrics, we determine the best-performing model for predicting credit card defaults in each subsection of the dataset.

### 4.2.1 Default of Credit Card Clients Dataset, as shown in Table [4.2]

#### Original Data

In the initial evaluation using the original dataset, the performance of different models paints a diverse landscape. The Random Forest model emerges as a strong contender, attaining an accuracy of 81.4%, precision of 64.0%, recall of 36.1%, F1-score of 46.1%, and a ROC value of 65.2%. The Decision Tree model exhibits a respectable accuracy of 72.8%, precision of 39.1%, recall of 41.1%, F1-score of 40.1%, and a ROC value of 61.5%. While these models showcase varying degrees of effectiveness in capturing credit card defaults, the Neural Network faces challenges in this endeavor. With an accuracy of 77.8%, precision of 11.1%, recall of 0.1%, F1-score of 0.1%, and a ROC value of 50.0%, the Neural Network's struggle to capture positive instances is evident, reflecting the complexities of the credit card default prediction task. TabTransformer steps into the spotlight as the top performer, boasting an accuracy of 81.7%, precision of 63.4%, recall of 41.1%, F1-score of 49.9%, and a ROC value of 67.2%.

#### Standardization

Upon standardizing the features, the Tab Transformer becomes the top performer with an accuracy of 81.8%, precision of 64.8%, recall of 38.7%, F1-score of 48.4%, and a ROC value of 66.3%. The Decision Tree model showcases marginal improvements, achieving an accuracy of 72.7%, precision of 38.7%, recall of 40.1%, F1-score of 39.4%, and a ROC value of 61.0%. It's noteworthy that the LSTM model experiences gains in the precision, recall, and F1-score after standardization, reaching values of 66.7%, 35.6%, and 46.5% respectively.

#### Standardization

In the context of addressing the class imbalance, we observe noteworthy shifts in model performances, indicating the positive influence of specialized techniques. The Random Forest model maintains consistent performance after imbalance handling, attaining an accuracy of 81.3%, precision of 64.3%, recall of 34.6%, F1-score of 45.0%, and a ROC value of 64.6%. Interestingly, the Neural Network responds positively to imbalance handling, yielding a

| Method | Model | Testing Perf. | | | | |
|---|---|---|---|---|---|---|
| | | Acc | Precision | Recall | F1 score | ROC |
| Original | Logistic regression | 0.779 | 0.000 | 0.000 | 0.000 | 0.500 |
| | Decision tree | 0.728 | 0.391 | 0.411 | 0.401 | 0.615 |
| | Random forest | 0.814 | 0.640 | 0.361 | 0.461 | 0.652 |
| | SVM | 0.779 | 0.000 | 0.000 | 0.000 | 0.500 |
| | Neural network | 0.778 | 0.111 | 0.001 | 0.001 | 0.500 |
| | CNN | 0.802 | 0.620 | 0.276 | 0.382 | 0.614 |
| | LSTM | **0.818** | **0.646** | 0.391 | 0.487 | 0.665 |
| | Tab transformer | 0.817 | 0.634 | **0.411** | **0.499** | **0.672** |
| Standardization | Logistic regression | 0.809 | **0.691** | 0.248 | 0.365 | 0.608 |
| | Decision tree | 0.727 | 0.387 | 0.401 | 0.394 | 0.610 |
| | Random forest | 0.815 | 0.646 | 0.359 | 0.462 | 0.652 |
| | SVM | 0.816 | 0.662 | 0.338 | 0.448 | 0.645 |
| | Neural network | 0.816 | 0.641 | 0.384 | 0.480 | 0.662 |
| | CNN | 0.816 | 0.655 | 0.356 | 0.462 | 0.652 |
| | LSTM | 0.818 | 0.667 | 0.356 | 0.465 | 0.653 |
| | Tab transformer | **0.818** | 0.648 | **0.387** | **0.484** | **0.663** |
| Imbalance Handling | Logistic regression | 0.681 | 0.369 | **0.621** | 0.463 | 0.660 |
| | Decision tree | 0.740 | 0.410 | 0.395 | 0.402 | 0.617 |
| | Random forest | **0.813** | **0.643** | 0.346 | 0.450 | 0.646 |
| | SVM | 0.777 | 0.497 | 0.560 | 0.526 | 0.699 |
| | Neural network | 0.776 | 0.495 | 0.561 | 0.526 | 0.699 |
| | CNN | 0.790 | 0.527 | 0.508 | 0.517 | 0.689 |
| | LSTM | 0.797 | 0.555 | 0.427 | 0.482 | 0.665 |
| | Tab transformer | 0.769 | 0.483 | 0.592 | **0.532** | **0.706** |

Table 4.2: Model Performance on Default of Credit Card Clients Dataset.

precision of 49.5%, recall of 56.1%, F1-score of 52.6%, and a ROC value of 69.9%. The SVM model, benefiting from imbalance handling, achieves a precision of 49.7% and a recall of 56.0%, contributing to an enhanced F1-score of 52.6% and a ROC value of 69.9%.

### 4.2.2 American Express Credit Card Dataset, as shown in Table [4.3]

#### Original Data

The initial assessment using the original dataset showcases a diverse array of performance outcomes. The Random Forest model asserts its dominance with an accuracy of 87.3%, precision of 74.7%, recall of 73.0%, F1-score of 73.9%, and a ROC value of 82.5%. Amidst the complexity of the dataset, the LSTM model grapples with a recall and F1-score of 0.1%, underlining the intricate challenge of capturing relevant information from this expansive data landscape. An expansive volume of data and a high number of columns, characterized by a multitude of columns, introduce challenges that can profoundly impact certain models, such as the Long Short-Term Memory (LSTM) model, as they strive to distill meaningful patterns from intricate credit risk scenarios.

#### Standardization

The transformative impact of standardization becomes evident as the Tab Transformer emerges as the top performer, achieving an accuracy of 87.5%, precision of 74.1%, recall of 75.7%, F1-score of 74.9%, and a ROC value of 83.5%. Standardization also yields subtle yet impactful improvements for the Decision Tree model, which demonstrates accuracy of 81.5%, precision of 62.4%, recall of 63.2%, F1-score of 62.8%, and a ROC value of 75.3%. Additionally, the SVM model consistently maintains its performance with precision of 75.9%, recall of 67.3%, F1-score of 71.4%, and ROC of 80.2%.

#### Imbalance Handling

Imbalance handling techniques unveil models' adaptability to the dataset's complexities. The Neural Network capitalizes on the class-weight training, achieving a precision of 67.0%, recall of 85.9%, F1-score of 75.3%, and a ROC value of 86.0%. The Random Forest model sustains its dominance with an accuracy of 86.6%, precision of 75.7%, recall of 67.4%, F1-score of 71.3%,

| Method | Model | Testing Perf. | | | | |
|---|---|---|---|---|---|---|
| | | Acc | Precision | Recall | F1 score | ROC |
| Original | Logistic regression | 0.862 | 0.742 | 0.674 | 0.706 | 0.798 |
| | Decision tree | 0.814 | 0.622 | 0.632 | 0.627 | 0.753 |
| | Random forest | 0.873 | 0.747 | 0.730 | 0.739 | 0.825 |
| | SVM | 0.827 | 0.765 | 0.433 | 0.553 | 0.695 |
| | Neural network | 0.874 | 0.741 | 0.751 | 0.746 | 0.832 |
| | CNN | **0.876** | **0.768** | 0.713 | 0.739 | 0.821 |
| | LSTM | 0.753 | 0.400 | 0.001 | 0.001 | 0.500 |
| | Tab transformer | 0.873 | 0.731 | **0.764** | **0.747** | **0.836** |
| Standardization | Logistic regression | 0.870 | 0.750 | 0.711 | 0.730 | 0.817 |
| | Decision tree | 0.815 | 0.624 | 0.632 | 0.628 | 0.753 |
| | Random forest | 0.873 | 0.749 | 0.728 | 0.739 | 0.824 |
| | SVM | 0.867 | **0.759** | 0.673 | 0.714 | 0.802 |
| | Neural network | 0.875 | 0.743 | 0.754 | 0.748 | 0.834 |
| | CNN | **0.876** | 0.752 | 0.745 | 0.748 | 0.832 |
| | LSTM | 0.755 | 0.648 | 0.017 | 0.033 | 0.507 |
| | Tab transformer | 0.875 | 0.741 | **0.757** | **0.749** | **0.835** |
| Imbalance handling | Logistic regression | 0.853 | 0.647 | 0.887 | **0.748** | 0.864 |
| | Decision tree | 0.812 | 0.623 | 0.605 | 0.614 | 0.742 |
| | Random forest | **0.866** | **0.757** | 0.674 | 0.713 | 0.802 |
| | SVM | 0.841 | 0.624 | 0.896 | 0.736 | 0.860 |
| | Neural network | 0.861 | 0.670 | 0.859 | 0.753 | 0.860 |
| | CNN | 0.873 | 0.746 | 0.737 | 0.742 | 0.827 |
| | LSTM | 0.711 | 0.415 | 0.418 | 0.416 | 0.612 |
| | Tab transformer | 0.845 | 0.628 | **0.910** | 0.743 | **0.867** |

Table 4.3: Model Performance on American Express Credit Card Dataset.

and a ROC value of 80.2%. Notably, the SVM model harnesses imbalance handling, achieving an impressive recall of 89.6%, contributing to an F1-score of 73.6% and a ROC value of 86.0%.

### 4.2.3 South German Credit Dataset, as shown in Table [4.4]

#### Original Data

Our preliminary evaluation of the original dataset unveils diverse performance outcomes. The Decision Tree model takes the lead with an accuracy of 72.5%, precision of 80.6%, recall of 80.0%, F1-score of 80.3%, and a ROC value of 67.5%. The Neural Network and SVM models exhibit competitive performance, leveraging the small dataset to achieve high precision, recall, and F1-score values. This limited data environment highlights the potential for models to capitalize on subtle patterns, albeit with a degree of variability stemming from the data's size.

#### Standardization

After the standardization, the Tab Transformer emerges as a robust performer with an accuracy of 77.0%, precision of 79.0%, recall of 91.4%, F1-score of 84.8%, and a ROC value of 67.4%. The Decision Tree model maintains its consistency, showcasing an accuracy of 71.5%, precision of 79.4%, recall of 80.0%, F1-score of 79.7%, and a ROC value of 65.8%. Notably, the LSTM model, despite challenges posed by limited data, exhibits potential, reaching a precision of 74.7% and recall of 95.0%.

#### Imbalance Handling

In the context of handling class imbalance within the South German dataset, several classification models exhibit significant improvements when subjected to imbalance handling techniques. Notably, the Logistic Regression model showcases enhanced precision of 81.4% and recall of 68.6%, achieving a balanced trade-off between minimizing false positives and maximizing true positives. The Decision Tree model demonstrates notable gains in both precision of 79.6% and recall of 77.9%, reflecting its capacity to accurately classify positive instances and capture actual positives. The Random Forest model excels in the recall of 94.3%, effectively identifying the minority class, with a balanced precision of 76.7% and an F1 score of 84.6%. The SVM model exhibits strong performance in the precision of 81.2% and the

| Method | Model | Testing Perf. | | | | |
|---|---|---|---|---|---|---|
| | | Acc | Precision | Recall | F1 score | ROC |
| Original | Logistic regression | 0.720 | 0.756 | 0.886 | 0.816 | 0.610 |
| | Decision tree | 0.725 | 0.806 | 0.800 | 0.803 | 0.675 |
| | Random forest | 0.740 | 0.756 | 0.929 | 0.833 | 0.614 |
| | SVM | 0.710 | 0.709 | 0.993 | 0.827 | 0.521 |
| | Neural network | 0.710 | 0.707 | **1.000** | 0.828 | 0.517 |
| | CNN | **0.740** | 0.747 | 0.950 | **0.836** | 0.600 |
| | LSTM | 0.715 | 0.717 | 0.979 | 0.828 | 0.539 |
| | Tab transformer | 0.710 | **0.836** | 0.729 | 0.779 | **0.698** |
| Standardization | Logistic regression | 0.710 | 0.747 | 0.886 | 0.810 | 0.593 |
| | Decision tree | 0.715 | **0.794** | 0.800 | 0.797 | 0.658 |
| | Random forest | 0.740 | 0.759 | 0.921 | 0.832 | 0.619 |
| | SVM | 0.740 | 0.734 | **0.986** | 0.841 | 0.576 |
| | Neural network | 0.700 | 0.741 | 0.879 | 0.804 | 0.581 |
| | CNN | 0.745 | 0.780 | 0.886 | 0.829 | 0.651 |
| | LSTM | 0.740 | 0.747 | 0.950 | 0.836 | 0.600 |
| | Tab transformer | **0.770** | 0.790 | 0.914 | **0.848** | **0.674** |
| Imbalance handling | Logistic regression | 0.670 | 0.814 | 0.686 | 0.744 | 0.660 |
| | Decision tree | 0.705 | 0.796 | 0.779 | 0.787 | 0.656 |
| | Random forest | **0.760** | 0.767 | **0.943** | **0.846** | 0.638 |
| | SVM | 0.715 | 0.812 | 0.771 | 0.791 | 0.677 |
| | Neural network | 0.675 | 0.810 | 0.700 | 0.751 | 0.658 |
| | CNN | 0.740 | 0.778 | 0.879 | 0.826 | 0.648 |
| | LSTM | 0.720 | **0.862** | 0.714 | 0.781 | **0.724** |
| | Tab transformer | 0.730 | 0.795 | 0.829 | 0.811 | 0.664 |

Table 4.4: Model Performance on South German Credit Dataset.

recall of 77.1%, maintaining equilibrium between precision and recall with the F1 score of 79.1%. Finally, the Tab Transformer stands out with competitive precision of 79.5% and a recall of 82.9%, emphasizing its resilience in effectively addressing class imbalance while maintaining a balanced performance profile. These models showcase varying strengths in handling imbalanced data, offering insights into their suitability for scenarios where class distribution is a challenge.

Figure 4.1 illustrates the visualization of the top-performing models across various datasets. The evaluation metric employed in this analysis is the ROC value, which surpasses the F1 score due to its ability to encompass both true positive and false positive rates within a singular measurement. This figure compellingly demonstrates how tab transformer showcases strong performance on diverse datasets.

In conclusion, the evaluation of various machine learning models on multiple datasets consistently demonstrates the TabTransformer as a top-performing model for different scenarios. This can be attributed to its adaptability and robustness in capturing complex patterns within tabular data. With its unique architecture that includes adaptive attention mechanisms, the TabTransformer effectively weighs feature interactions and relationships, making it well-suited for diverse datasets with intricate patterns. Its adaptability enables it to capture both linear and non-linear relationships, resulting in high accuracy, precision, recall, F1-score, and ROC values across different datasets and preprocessing techniques. The TabTransformer's impressive performance underscores its potential as a reliable choice for a wide range of classification tasks, making it an excellent option for practitioners seeking a model that can consistently deliver strong results across different datasets and preprocessing scenarios. As the field of machine learning continues to advance, the TabTransformer's capabilities and adaptability position it as a promising candidate for achieving state-of-the-art performance in various domains.

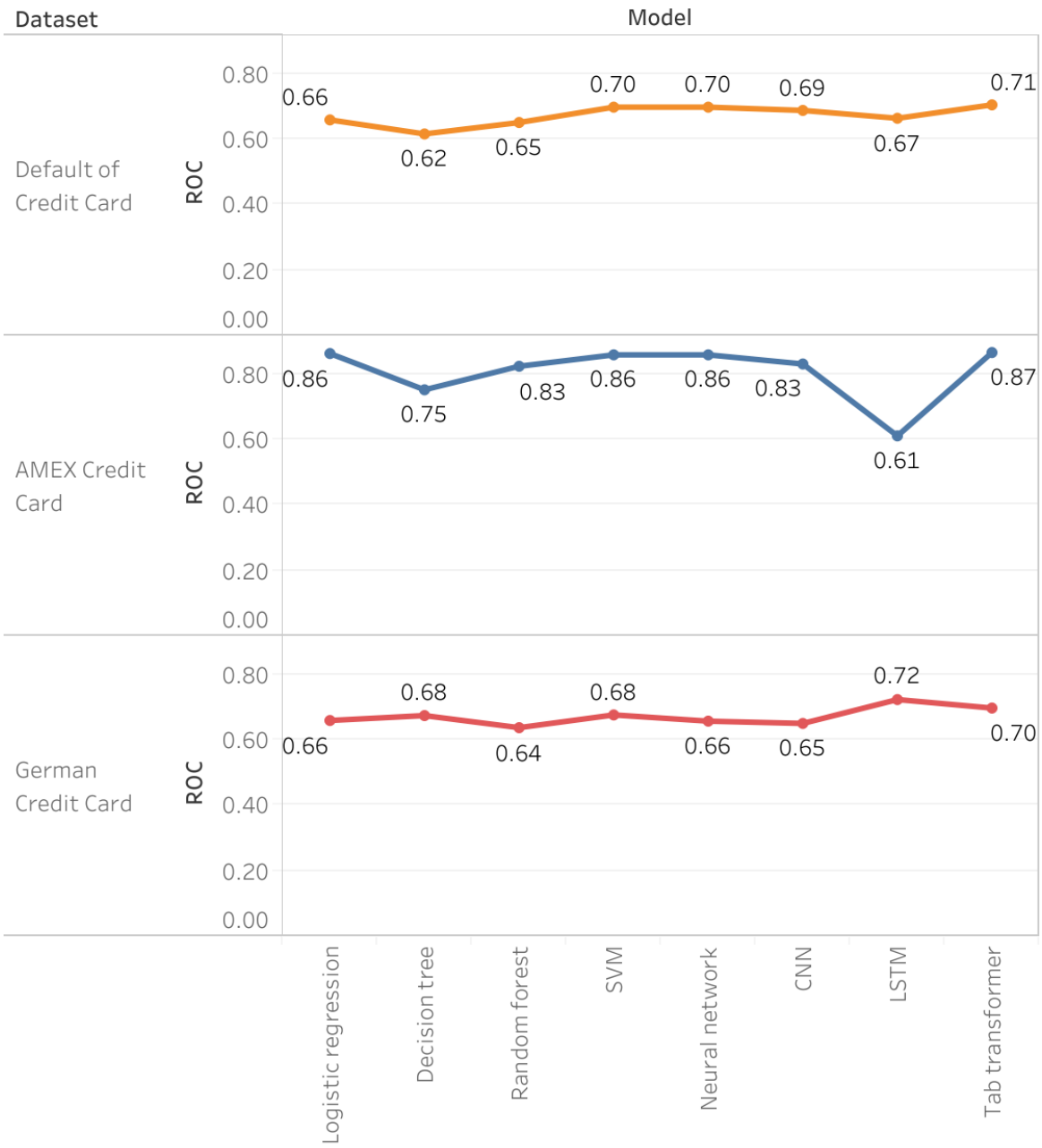


Figure 4.1: A comparison of various machine learning models on three different datasets: Default of Credit Card Clients, American Express Credit Card, and South German Credit. The ROC value is utilized as the primary evaluation metric to assess the performance of these models.

CHAPTER 5

# Discussion

In this section, we provide a thorough examination of our extensive inquiry into machine learning models used for predicting credit card defaults. We conduct an evaluation of the performance of each model on various datasets and suggest potential avenues for improvement. Our analysis seeks to deepen understanding regarding the factors influencing the performance of these models by exploring their advantages, limitations, and possible future advancements.

## 5.1 Model performance analysis

### 5.1.1 Logistic regression

Logistic Regression demonstrates both strengths and weaknesses in predicting credit card defaults using the Default of Credit Card Clients dataset. Although achieving a relatively high accuracy rate of 0.779, the model's precision, recall, and F1 score exhibit noticeable deficiencies. This indicates that while the model accurately classifies a significant proportion of instances overall, it faces challenges in identifying positive class instances specifically due to potential data imbalance issues within the dataset. The decreased precision and recall values further suggest that many actual occurrences of credit card defaults are incorrectly classified by the model. This misclassification is concerning as accurate identification of high-risk cases is vital for effective prediction models targeted at detecting credit card defaults. The lower precision and recall values can primarily be attributed to an underrepresentation

of minority class instances (i.e., default cases) within the dataset.

On the American Express credit card dataset, Logistic Regression demonstrates significantly better performance compared to the previous dataset. The model achieves a commendable accuracy of 0.862, signifying an overall accurate classification. Moreover, it exhibits well-balanced precision, recall, and F1 score values which indicate fair treatment of positive and negative class instances. This implies that the linear decision boundaries established by the model align effectively with both feature distribution and class representation within this particular dataset. The larger size and higher dimensionality of this specific dataset may enhance Logistic Regression's ability to discern patterns accurately. These factors potentially offer increased discriminatory power for capturing underlying data patterns more effectively. It is worth noting that some level of class imbalance might still be present in the dataset; however, LogisticRegression manages to harness its inherent characteristics advantageously resulting in a comprehensive performance across multiple evaluation metrics.

Logistic regression performs relatively well on the South German Credit dataset, achieving an overall classification accuracy of 0.720. However, there is a trade-off between precision and recall measures observed in this model. While the recall value is high, indicating that the model successfully identifies many instances of default occurrences, its precision value is low. This means that a significant portion of the predicted labels by this logistic regression model are incorrect when compared to the actual training labels.

The limitations of Logistic Regression become particularly significant in scenarios such as credit risk assessment where correctly identifying instances from both classes is crucial. Various factors can influence its performance, including feature distributions and separability across different classes within the dataset. Nevertheless, Logistic Regression may not be as effective at capturing underlying patterns related to positive class instances which results in lower recall scores.

### 5.1.2 Decision tree

The model performance on the Default of Credit Card Clients dataset can be characterized as a balance between its strengths and limitations. Initially, the model achieves a reasonable accuracy level of 0.728, indicating satisfactory classification performance. However, when evaluating other metrics such as precision, recall, and F1 score, it becomes evident that there is a trade-off in achieving well-balanced performance. In particular, the relatively low

precision value highlights an issue where the model tends to misclassify a significant number of negative instances as positive ones. This phenomenon stems from decision trees' inherent nature to divide feature space based on predetermined thresholds. While this partitioning approach allows decision trees to effectively handle imbalanced datasets by accommodating minority class instances within separate branches or leaves; capturing intricate relationships present in complex datasets may pose challenges for these models.

The performance of the Decision Tree classifier on the American Express credit card dataset demonstrates significant improvement compared to the previous dataset. The model achieves an accuracy score of 0.814, suggesting that it effectively partitions its feature space based on class labels within the data's distribution and characteristics. This observation showcases one advantage of decision trees: their ability to identify distinct regions corresponding to different classes, a structure evident in this particular dataset. Furthermore, balanced precision and recall values indicate that this model is adaptable to varying class distributions and can handle imbalanced datasets up to a certain extent. It is worth noting that achieving a balance between precision (predictive accuracy) and recall (sensitivity) is crucial for effective classification models when dealing with imbalanced data.

The Decision Tree model's performance on the South German Credit dataset provides a different perspective. Initially, it achieves a high accuracy of 0.725, indicating good classification capability. The balanced precision and recall demonstrate the model's effectiveness in capturing positive and negative class instances equally well. This is because decision trees are suitable for datasets with distinct feature partitioning, which is true in this case as the model adapts to these characteristics while maintaining a balance between precision and recall. However, it is important to acknowledge that decision trees may struggle in accurately capturing complex patterns within the data set. While they excel when there is a clear separation between classes, their ability to comprehend intricate relationships can be limited. Hence, when dealing with datasets containing intricate patterns, decision trees may exhibit suboptimal performance.

### 5.1.3 Random Forest

Overall, the performance of the Random Forest model on the Default of Credit Card Clients dataset demonstrates a trade-off between accuracy and capturing class imbalance. The initial form of the model achieves an accuracy score of 0.814, indicating reasonably accurate

classification results. However, there is a compromise between correctly identifying positive class instances and overall classification performance as evidenced by precision, recall, and F1 score metrics. Precision and recall exhibit a relatively balanced distribution which suggests that this ensemble approach effectively addresses class imbalance by combining multiple decision trees' outputs. Nevertheless, it should be noted that due to this inherent trade-off nature in these metrics, achieving higher values for the F1 score might not be possible.

Random Forest exhibits strong and balanced performance on the American Express credit card dataset, achieving an accuracy of 0.873. This indicates its proficiency in effectively classifying instances from different classes while maintaining a balance between precision, recall, and F1 score. The ensemble approach employed by Random Forest allows it to comprehend complex patterns inherent in the data and adapt accordingly to varying class distributions. By leveraging its ensemble of decision trees, this algorithm captures a wider range of relationships and interactions within the dataset, resulting in well-rounded performance across multiple evaluation metrics. Furthermore, Random Forest's robustness towards class imbalance enhances its effectiveness on this specific dataset. In conclusion, Random Forest excels at handling imbalanced datasets due to its exceptional ability to capture intricate data patterns.

The performance of the Random Forest model on the South German Credit dataset demonstrates its versatility in handling different data characteristics. When used in its original form, the model achieves an accuracy of 0.740 and maintains balanced precision and recall measures. This indicates that the ensemble of decision trees effectively captures instances from both positive and negative classes. By employing an ensemble approach, Random Forest can generate predictions that generalize well across various relationships present in the data.

This adaptability is particularly valuable for credit risk assessment tasks where accurately capturing cases from both positive (low-risk) and negative (high-risk) classes holds significant importance. The balanced precision and recall further highlight how Random Forest addresses class imbalance by making informed predictions on all classes, thereby mitigating any biases introduced due to imbalance.

### 5.1.4 SVM

The performance of the Support Vector Machine model on the Default of Credit Card Clients dataset is varied. While it achieves an accuracy rate of 0.779, its precision, recall, and F1 score

are relatively low. This suggests that although the SVM model correctly identifies a substantial number of instances, it struggles when identifying positive class instances specifically. The presence of class imbalance significantly affects SVM's performance in this case since its linear decision boundaries may not effectively separate both classes. Additionally, SVM's inherent sensitivity to class imbalance and emphasis on maximizing the margin between classes contribute to the difficulties in capturing positive class instances accurately.

In the American Express credit card dataset, support vector machines are found to have a balanced performance compared to the previous dataset. The original form of SVM achieves an accuracy of 0.827 with reasonable precision, recall, and F1 score. This suggests that SVM's optimization process effectively distinguishes between classes in this particular dataset by aligning decision boundaries with feature and class distributions. When data patterns exhibit linear separability as observed here, SVM excels at identifying hyperplanes that maximize class separation. Moreover, due to the larger size and higher dimensionality of this dataset in comparison to others previously evaluated, it presents ample opportunities for creating well-separated decision boundaries which significantly contribute towards its comprehensive performance.

In evaluating the performance of the SVM model on the South German Credit dataset, it becomes evident that the model maintains a balanced approach in capturing instances from both positive and negative classes. With an accuracy score of 0.710, this indicates that the model is effective in handling class imbalances and can accurately identify linear decision boundaries within this particular dataset. However, there is a slight difference between precision and recall metrics. While precision remains reasonably balanced, indicating accurate identification of positive class instances among all detected cases, recall shows very high values. This suggests that the model can capture most of the positive class instances while potentially sacrificing some level of precision.

### 5.1.5 Neural network

The performance of the Neural Network model on the Default of Credit Card Clients dataset demonstrates both its capabilities and limitations. Despite initially achieving an accuracy rate of 0.778, it shows low precision, recall, and F1 score values. This suggests that the model struggles to effectively identify positive class instances and achieve a well-balanced overall performance. Although neural networks have an inherent ability to capture complex

relationships in data, in this specific case study, either the architecture or training process employed by the model appears inadequate for learning discriminative patterns between the two classes. Moreover, these issues can be partially attributed to the limited sample size and intricate interactions present within the dataset itself.

The Neural Network model demonstrates superior performance on the American Express credit card dataset. With an accuracy of 0.874, along with well-balanced precision, recall, and F1 score values, the original form of this model proves effective in capturing complex patterns within the data while accurately classifying both positive and negative instances. This success can be attributed to neural networks' suitability for high-dimensional datasets such as this one since they are capable of learning intricate relationships between features. Moreover, owing to the larger size and higher dimensionality of this dataset compared to others, there exists a greater potential for the neural network model to acquire representative features enabling proficient classification.

The performance of the Neural Network model on the South German Credit dataset is worth mentioning. The accuracy rate achieved by the model is 0.710, indicating its ability to classify instances correctly. Moreover, it demonstrates a recall score of 1.000, signifying that it effectively identifies positive cases from actual positive ones. Neural networks have shown their effectiveness in capturing complex patterns and handling high-dimensional data present in this particular dataset. Their capability to handle such characteristics makes them suitable for this task. Overall, these findings shed light on how neural networks can be useful in accurately classifying instances from both positive and negative classes within imbalanced datasets like the South German Credit Dataset.

### 5.1.6 CNN

The performance of the Convolutional Neural Network model on the Default of Credit Card Clients dataset provides valuable insights into its adaptability to non-image data. Although in its current form, the model achieves an accuracy score of 0.802, it is evident that there is a noticeable imbalance between precision, recall, and F1-score values. This discrepancy suggests that while CNNs can capture certain patterns within tabular data, they may struggle with accurately identifying positive class instances. It should be noted that CNN architectures are primarily designed for image data analysis and applying them to tabular datasets like this might not fully harness their capabilities.

The CNN model demonstrates promising performance on the American Express credit card dataset, achieving an impressive accuracy of 0.876 and demonstrating reasonably balanced precision, recall, and F1 score. This suggests that the architecture of the CNN effectively captures pertinent features and patterns from the high-dimensional data in its original form. Although typically used for image data, CNNs can also be advantageous in situations where intricate interactions among features are present, as illustrated by this dataset. The larger size and higher dimensionality of the dataset provide additional opportunities to identify meaningful patterns through modeling, ultimately contributing to its comprehensive overall performance. The feature extraction capabilities exhibited by CNNs align well with the inherent characteristics encompassed within this particular dataset.

The performance of the CNN model on the South German Credit dataset provides valuable insights. The original form of the model achieves an overall accuracy of 0.740, exhibiting balanced precision and recall values. This indicates that CNN is capable of effectively capturing instances from both positive and negative classes. Despite being unconventional for tabular data, CNNs exhibit great potential due to their ability to extract relevant features and detect nonlinear patterns in this particular dataset. However, it is worth noting that there seems to be a trade-off between precision and recall, suggesting that not all positive class instances are successfully captured by the model given the intricacy inherent in its relationships.

### 5.1.7 LSTM

The performance of the LSTM model on the Default of Credit Card Clients dataset is a combination of positive and negative outcomes. The original form shows an accuracy rate of 0.818, indicating a moderate ability to accurately classify both positive and negative instances. However, the F1 score indicates limitations in achieving balanced precision and recall values. LSTMs are designed for sequential data analysis such as natural language processing or time-series prediction tasks, making their application to tabular datasets like this less effective due to the absence of inherent sequential structure.

The performance of the LSTM model on the American Express credit card dataset is limited. In its original form, the model achieves an accuracy of 0.753 but demonstrates low precision, recall, and F1 score. These results suggest that the LSTM struggles to effectively capture both positive and negative class instances in this dataset which lacks a clear sequential

structure. One reason for this suboptimal performance could be attributed to the architecture of the LSTM model itself. Designed to capture sequential dependencies, it might not align well with datasets like this one that do not exhibit strong temporal patterns or sequencing characteristics. Another factor that could contribute to these limitations is the vanishing gradient problem, particularly when dealing with high-dimensional datasets such as this one. The presence of a large number of features can exacerbate this issue and hinder learning long-range dependencies accurately.

The LSTM model's performance on the South German Credit dataset provides valuable insights. Initially, the model achieved an accuracy of 0.715 with a balanced precision and recall, indicating its ability to effectively capture instances from both positive and negative classes to some extent. However, there is room for improvement in achieving a better trade-off between precision and recall as indicated by the F1 score. It should be noted that like previous datasets, the architecture of LSTM might not be optimally suited for tabular data analysis.

In conclusion, due to misalignment with non-sequential data structures and challenges posed by high dimensionality and vanishing gradients, optimizing or exploring alternative models might be necessary for better performance on such datasets.

### 5.1.8 Tab Transformer

The original version of the Tab Transformer achieved an accuracy score of 0.817 when evaluated on the Default of Credit Card Clients dataset. This suggests that the model demonstrates proficiency in capturing instances belonging to both positive and negative classes, thereby showcasing its competence in understanding intricate relationships within tabular data. The presence of attention mechanisms within the architecture enables the Tab Transformer to effectively assign weights to various features when making predictions about classifications.

The Tab Transformer model stands out in its performance on the American Express credit card dataset. With an accuracy of 0.873, it demonstrates robustness in handling diverse and high-dimensional data while maintaining balanced precision and recall. The use of attention mechanisms within the Tab Transformer plays a crucial role in achieving this level of performance by effectively capturing important feature interactions. It is noteworthy that the larger size and increased complexity of the dataset do not negatively impact the model's

ability to make accurate predictions. This underscores its capacity to learn intricate patterns within tabular data, thus making it a powerful choice for scenarios involving such datasets.

The Tab Transformer demonstrates impressive performance on the South German Credit dataset, achieving an accuracy of 0.770 with balanced precision and recall for the standardization scenario. This indicates its effectiveness in capturing class instances accurately. The attention mechanisms implemented in the model play a vital role in comprehending patterns and relationships present within the dataset, enabling informed predictions to be made. Notably, even though the South German Credit dataset is relatively small in size, the Tab Transformer efficiently leverages feature interactions to adapt to different data sizes and complexities.

## 5.2 Strengths and limitations

Machine learning models offer diverse capabilities for handling classification tasks, each with its own strengths and limitations.

Logistic regression, known for its simplicity and interpretability, suits binary tasks with linear decision boundaries and clear feature-target relationships. It serves as an excellent baseline model for initial exploration and feature importance interpretation. However, it may struggle with complex, non-linear relationships and high-dimensional data. The logistic regression can be effective in scenarios where the relationship between features and the target follows a linear pattern, such as in simple binary classification problems. The decision trees can capture non-linear interactions, making them well-suited for tasks demanding feature interactions and mixed data types. They provide easy interpretability through visualizations and are robust in handling both numerical and categorical features. However, they are prone to overfitting, particularly on small datasets with intricate decision boundaries. Decision Trees are useful when we need to understand feature interactions and relationships in a classification problem, especially when there is a mix of different types of features.

Random Forests extend decision trees by ensemble averaging, effectively addressing overfitting and complexity while maintaining interpretability. They excel in handling high-dimensional data and can capture complex relationships. However, their computational intensity and reduced interpretability compared to individual decision trees might be limitations. Random Forests are valuable when dealing with large datasets that have many

| Method | Model | Confusion Matrix | | | |
|---|---|---|---|---|---|
| | | TN | FP | FN | TP |
| Original | Logistic regression | 4673 | 0 | 1327 | 0 |
| | Decision tree | 3824 | 849 | 781 | 546 |
| | Random forest | 4403 | 270 | 848 | 479 |
| | SVM | 4673 | 0 | 1327 | 0 |
| | Neural network | 4665 | 8 | 1326 | 1 |
| | CNN | 4449 | 224 | 961 | 366 |
| | LSTM | 4389 | 284 | 808 | 519 |
| | Tab transformer | 4358 | 315 | 781 | 546 |
| Standardization | Logistic regression | 4526 | 147 | 998 | 329 |
| | Decision tree | 3831 | 842 | 795 | 532 |
| | Random forest | 4412 | 261 | 850 | 477 |
| | SVM | 4444 | 229 | 878 | 449 |
| | Neural network | 4387 | 286 | 817 | 510 |
| | CNN | 4424 | 249 | 854 | 473 |
| | LSTM | 4437 | 236 | 854 | 473 |
| | Tab transformer | 4394 | 279 | 814 | 513 |
| Imbalance handling | Logistic regression | 3262 | 1411 | 503 | 824 |
| | Decision tree | 3918 | 755 | 803 | 524 |
| | Random forest | 4418 | 255 | 868 | 459 |
| | SVM | 3920 | 753 | 584 | 743 |
| | Neural network | 3915 | 758 | 583 | 744 |
| | CNN | 4069 | 604 | 653 | 674 |
| | LSTM | 4219 | 454 | 761 | 566 |
| | Tab transformer | 3831 | 842 | 541 | 786 |

Table 5.1: Confusion Matrix for the Default of Credit Card Clients Dataset.

| Method | Model | Confusion Matrix | | | |
|---|---|---|---|---|---|
| | | TN | FP | FN | TP |
| Original | Logistic regression | 41725 | 3474 | 4828 | 9973 |
| | Decision tree | 39504 | 5695 | 5448 | 9353 |
| | Random forest | 41546 | 3653 | 3997 | 10804 |
| | SVM | 43229 | 1970 | 8398 | 6403 |
| | Neural network | 41308 | 3891 | 3692 | 11109 |
| | CNN | 42011 | 3188 | 4254 | 10547 |
| | LSTM | 45187 | 12 | 14793 | 8 |
| | Tab transformer | 41041 | 4158 | 3492 | 11309 |
| Standardization | Logistic regression | 41691 | 3508 | 4280 | 10521 |
| | Decision tree | 39555 | 5644 | 5450 | 9351 |
| | Random forest | 41595 | 3604 | 4026 | 10775 |
| | SVM | 42042 | 3157 | 4835 | 9966 |
| | Neural network | 41332 | 3867 | 3636 | 11165 |
| | CNN | 41552 | 3647 | 3769 | 11032 |
| | LSTM | 45064 | 135 | 14552 | 249 |
| | Tab transformer | 41289 | 3910 | 3592 | 11209 |
| Imbalance handling | Logistic regression | 38026 | 7173 | 1676 | 13125 |
| | Decision tree | 39773 | 5426 | 5849 | 8952 |
| | Random forest | 42003 | 3196 | 4828 | 9973 |
| | SVM | 37210 | 7989 | 1542 | 13259 |
| | Neural network | 38927 | 6272 | 2084 | 12717 |
| | CNN | 41492 | 3707 | 3895 | 10906 |
| | LSTM | 36485 | 8714 | 8620 | 6181 |
| | Tab transformer | 37218 | 7981 | 1337 | 13464 |

Table 5.2: Confusion Matrix for the American Express Credit Card Dataset.

features, as they can effectively capture complex interactions and patterns.

Support Vector Machines (SVMs) are powerful in high-dimensional spaces, effectively capturing non-linear relationships. With appropriate kernel functions and class imbalance handling, SVMs are adept at scenarios requiring distinct class separation. However, they can be computationally demanding and sensitive to hyperparameter choices. SVMs are suitable for tasks where class separation is crucial and feature relationships are complex, such as in text classification or image recognition problems.

Neural Networks offer unparalleled capacity for capturing complex patterns across large datasets. They excel in tasks with non-linear relationships and intricate feature interactions. However, they demand substantial computational resources, careful hyperparameter tuning, and are prone to overfitting on small datasets. Neural Networks are well-suited for tasks involving large amounts of data and complex relationships, such as natural language processing and image classification.

Convolutional Neural Networks (CNNs), originally designed for image data, can adapt their feature extraction capabilities to tabular data with spatial-like patterns. They are effective in capturing spatial relationships and feature interactions. However, their optimal performance might be hindered without the inherent spatial structure found in image data. CNNs are particularly effective when dealing with data that exhibits spatial patterns, such as images or structured data with inherent spatial relationships.

Long Short-Term Memory (LSTM) Models shine in sequential data with temporal dependencies, capturing long-range relationships and memory patterns. However, challenges arise when dealing with high-dimensional tabular data, as vanishing gradient problems may hinder the model's ability to learn effectively. LSTMs are well-suited for tasks involving sequential data, such as time series forecasting or natural language processing, where capturing temporal patterns is essential.

Tab Transformers, equipped with adaptive attention mechanisms, excel in capturing intricate patterns within tabular data. Their ability to weigh feature interactions effectively makes them a reliable choice across diverse datasets. However, complex architectures and careful hyperparameter tuning might be necessary for optimal performance. Tab Transformers are a strong choice when dealing with tabular data that has complex relationships and interactions between features, making them suitable for tasks like credit risk assessment or customer churn prediction.

In conclusion, selecting an appropriate model depends on aligning its strengths with specific task requirements. Interpretability, complexity, dataset characteristics, and desired performance trade-offs guide the choice of model, highlighting the importance of understanding each model's capabilities and limitations.

## 5.3 Further improvements

This study offers valuable insights into the performance of machine learning models on diverse datasets. However, there is potential for further improvement to achieve more comprehensive and robust results. An effective approach to enhance model performance is through the utilization of ensemble methods such as Gradient Boosting or Stacking. These techniques harness the diversity of model outputs and address individual models' weaknesses, resulting in significant improvements in classification outcomes across various datasets. Exploring a range of ensemble strategies, experimenting with different ensemble sizes, and investigating various combinations of base models would be worthwhile endeavors that can contribute to a deeper understanding of how these techniques contribute to enhanced classification outcomes. Another area of improvement is to ensure the reliability and robustness of research findings by conducting cross-validation and performing sensitivity analyses. This enables us to gain insights into how well the models perform under different conditions such as variations in data distribution, noise levels, and feature perturbations. By incorporating comprehensive robustness tests along with proper cross-validation protocols, we can significantly enhance the reliability of these models' results and improve our understanding of their real-world performance.

CHAPTER 6

# Conclusions

Credit card default prediction has become an important issue globally, particularly for financial institutions and risk management. Various machine-learning methods have been employed to predict the probability of credit card holders failing to meet their payment obligations within specific timeframes. Default, which refers to the failure to make timely payments, holds great importance for individual financial stability and economic sustainability as a whole. Predicting defaults allows these entities to manage risk exposure proactively, allocate resources efficiently, and create backup plans in order to minimize potential losses. This proactive approach strengthens institutional stability and enhances the overall resilience of the financial ecosystem.

In this study, we conducted a comprehensive analysis of different machine learning models for predicting credit card default. We utilized three datasets: Default of Credit Card Clients, American Express Credit Card, and South German Credit. These datasets were chosen to provide diverse instances, attributes, and statistical information. Our evaluation focused on assessing the performance of each model using metrics such as accuracy, precision, recall, F1 score, and AOC. These metrics are vital in classification tasks. The models considered included Logistic Regression, Decision Trees, Random Forests, SVMs, Neural Networks, CNNs, LSTMs, and the Tab Transformer which represents state-of-the-art methods used in prior research studies. To ensure the reliability of our analyses, we performed necessary pre-processing steps before evaluating the models. We utilized standardization to scale feature values and make them consistent, enabling effective capture of data patterns. Moreover,

addressing class imbalance was critical due to one class being more prevalent than the other. Techniques like oversampling and undersampling were employed to balance the distribution of classes, enhancing accuracy in accurately classifying instances from minority classes. The performance of the models varied across the datasets. Logistic Regression showed high accuracy rates on certain datasets but faced challenges in identifying positive class instances due to data imbalance. Decision Trees achieved a balance between strengths and limitations, performing well with imbalanced data while struggling with complex pattern recognition. Random Forests demonstrated adaptability to different data characteristics by balancing accuracy and handling class imbalance. SVM had mixed performance, excelling in some cases but encountering difficulties with precision and recall for positive class instances. Based on deep-learning approaches, Neural Networks, CNNs, and LSTMs showed different performance patterns with strengths and limitations across various datasets. These models demonstrated their ability to capture complex patterns effectively when the dataset characteristics matched their strengths. The Tab Transformer consistently excelled in capturing intricate relationships within tabular data, making it a promising model for credit card default prediction and establishing itself as a state-of-the-art solution across all analyzed datasets. This study emphasizes the significance of accurate credit card default prediction in financial risk management. By analyzing various machine learning models, valuable insights are provided to aid in selecting and implementing appropriate models for credit card default prediction tasks. Given the dynamic nature of the financial landscape, informed model selection and comprehensive performance evaluation will continue to play a crucial role in improving risk management practices and maintaining stability within financial institutions and the economy as a whole.

The experimental findings presented in this study have significant implications for credit risk assessment. The insights gained from evaluating various machine learning models provide valuable guidance on improving the accuracy and effectiveness of credit card default prediction processes. These findings impact credit risk assessment in multiple ways. The varied performance among different machine learning models emphasizes the importance of careful model selection and customization based on specific dataset characteristics and objectives of credit risk assessment. Institutions can utilize this knowledge to tailor their model choices to the unique aspects of their credit portfolios, thereby increasing the precision of default predictions. Additionally, it is demonstrated how a state-of-the-art classification model can be applied to credit card default datasets. Moreover, the examination of pre-processing

techniques, including standardization and imbalance handling, highlights their crucial role in enhancing credit risk evaluation results. Standardization ensures consistent scaling of features, enabling models to effectively detect patterns within the data and make more precise predictions. Additionally, our study on methods for addressing class imbalance reveals that these techniques can rectify biases present in imbalanced datasets. This allows models to accurately identify instances belonging to minority classes and minimize the chances of misclassification. The variability in model performance across different datasets underscores the importance of a nuanced approach to credit risk assessment. Financial institutions can utilize this understanding to select models that are specifically tailored to individual data distributions and complexities, thus optimizing their predictive capabilities. For instance, the Tab Transformer's ability to capture intricate relationships within tabular data provides a promising avenue for institutions when dealing with complex credit portfolios. By harnessing the strengths of each individual model, financial institutions can create ensemble methods that combine diverse outputs. This enhances prediction accuracy and overall reliability in assessing risks. These approaches empower institutions to proactively identify customers with high-risk profiles, allocate resources more efficiently, and develop contingency plans as safeguards against potential losses. Ultimately, these measures strengthen financial stability and resilience.

This study provides valuable insights into the performance of machine learning models on diverse datasets. However, there is a need for further advancement and refinement to achieve a comprehensive understanding. One potential approach to enhance model performance is by using ensemble methods like Gradient Boosting or Stacking. These methods utilize the diversity of individual model outputs to address weaknesses and improve classification outcomes on different datasets. To better understand the impact of ensemble methods, it is recommended to thoroughly investigate various strategies for creating ensembles, including different ensemble sizes and combinations of base models. This analysis will provide insights into how ensemble techniques influence model performance and contribute to an improved understanding of their ability to enhance classification outcomes. An additional area for future improvement is enhancing the reliability and robustness of our research findings. This can be achieved by implementing cross-validation methods and conducting sensitivity analyses to better understand how the models perform in different scenarios. These scenarios may involve variations in data distribution, noise levels, and changes in feature attributes.

By incorporating comprehensive tests of robustness along with clear protocols for cross-validation, we can greatly increase the trustworthiness of our model's results and gain a deeper understanding of their practicality in real-world situations.

In conclusion, our study explains the potential of using machine learning in credit risk assessment and its implications for risk management. By leveraging advanced models, financial institutions can enhance stability while safeguarding individuals' financial well-being. Our research emphasizes the importance of considering model performance and preprocessing techniques to gain a comprehensive understanding of how machine learning can be applied effectively in credit risk assessment. In this era of technological advancement, our findings offer valuable insights for future research endeavors and encourage the utilization of ensemble methods and rigorous validation approaches to navigate complex credit risk landscapes efficiently with machine-learning capabilities. As the financial sector undergoes continuous transformation, insights from our study will serve as a guiding light for stakeholders in developing accurate and comprehensive credit risk assessment strategies that are adaptable to the ever-evolving challenges of today's financial ecosystem.

# Bibliography


[1] Talha Mahboob Alam, Kamran Shaukat, Kamran Shaukat, Ibrahim A. Hameed, Suhuai Luo, Muhammad Umer Sarwar, Muhammad Umer Sarwar, Muhammad Sarwar, Shakir Shabbir, Jiaming Li, and Matloob Khushi. An investigation of credit card default prediction in the imbalanced datasets. *IEEE Access*, 2020.

[2] Classification and estimation of high-risk factors to low-risk factors in approving loan through creditworthiness of bank customers using svm algorithm and analyze its performance over logistic regression in terms of accuracy. *pnr*, 13, 2022.

[3] Alzbeta Bacova, FrantiÅ¡ek BabiÄ, and Frantisek Babic. Predictive analytics for default of credit card clients. *International Symposium on Applied Machine Intelligence and Informatics*, 2021.

[4] Shigeyuki Hamori, Minami Kawai, Takahiro Kume, Yuji Murakami, and Chikara Watanabe. Ensemble learning or deep learning? application to default risk analysis. *null*, 2018.

[5] Jui-Yu Wu and Pei-Ci Liu. Identifying a default of credit card clients by using a lstm method: A case study. *Artificial Intelligence Trends*, 2022.

[6] I-Cheng Yeh and Che-Hui Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems With Applications*, 2009.

[7] Peter Martey Addo, Dominique Guegan, Bertrand Hassani, and Bertrand K. Hassani. Credit risk analysis using machine and deep learning models. *null*, 2018.

[8] Vijay S. Desai, Vijay Desai, Vijay S. Desai, Jonathan Crook, and George A. Overstreet. A comparison of neural networks and linear scoring models in the credit union environment. *European Journal of Operational Research*, 1996.

[9] I-Cheng Yeh. default of credit card clients. UCI Machine Learning Repository, 2016. DOI: https://doi.org/10.24432/C55S3H.

[10] Ulrike GrÃ¶mping. South german credit data: Correcting a widely used data set. 11 2019.

[11] William Greene. Sample selection in credit-scoring models. *Japan The World Economy*, 10(3):299–316, 1998.

[12] S. B. Kotsiantis. Supervised machine learning: A review of classification techniques. In *Proceedings of the 2007 Conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in EHealth, HCI, Information Retrieval and Pervasive Technologies*, page 3â24, NLD, 2007. IOS Press.

[13] John Doe. Logistic regression: Description, examples, and comparisons. *Journal of Marriage and Family*, Jan 1990.

[14] R. Azen and N. Traxel. Using dominance analysis to determine predictor importance in logistic regression. *Journal of Educational and Behavioral Statistics*, 34:319–347, 2009.

[15] J. Levy and A. O'Malley. Donât dismiss logistic regression: the case for sensible extraction of interactions in the era of machine learning. *BMC Medical Research Methodology*, 20, 2020.

[16] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1:81–106, 1986.

[17] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[18] Yanjun Qi. Random forest for bioinformatics. *null*, 2012.

[19] V. Svetnik, A. Liaw, C. Tong, J. Culberson, R. Sheridan, and B. Feuston. Random forest:â a classification and regression tool for compound classification and qsar modeling. *Journal of Chemical Information and Computer Sciences*, 43:1947–1958, 2003.

[20] Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer Publishing Company, Incorporated, 1st edition, 2008.

[21] W. Jiang and D. Gao. Five typical stenches detection using an electronic nose. *Sensors*, 20:2514, 2020.

[22] J. Croonen, J. Amaya, and G. Lapenta. Tokamak disruption prediction using different machine learning techniques. 2020.

[23] Ahmed Shihab Ahmed and H. Salah. A comparative study of classification techniques in data mining algorithms used for medical diagnosis based on dss. *Bulletin of Electrical Engineering and Informatics*, 2023.

[24] Kishansingh Rajput and Bhavesh A. Oza. A comparative study of classification techniques in data mining. *null*, 2017.

[25] KrizhevskyAlex, SutskeverIlya, and E HintonGeoffrey. Imagenet classification with deep convolutional neural networks. *Communications of The ACM*, 2017.

[26] Anuraganand Sharma, Dinesh Kumar, Dinesh Kumar, and Dinesh Kumar. Non-image data classification with convolutional neural networks. *null*, 2020.

[27] Ralf C. Staudemeyer and Eric Rothstein Morris. Understanding LSTM - a tutorial into long short-term memory recurrent neural networks. *CoRR*, abs/1909.09586, 2019.

[28] Zachary Chase Lipton. A critical review of recurrent neural networks for sequence learning. *CoRR*, abs/1506.00019, 2015.

[29] Klaus Greff, Rupesh Kumar Srivastava, Jan Koutník, Bas R. Steunebrink, and Jürgen Schmidhuber. LSTM: A search space odyssey. *CoRR*, abs/1503.04069, 2015.

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[31] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020.

[32] Xin Huang, Ashish Khetan, Milan Cvitkovic, and Zohar S. Karnin. Tabtransformer: Tabular data modeling using contextual embeddings. *CoRR*, abs/2012.06678, 2020.

[33] Henderi Henderi, Henderi Henderi, Tri Wahyuningsih, and Efana Rahwanto. Comparison of min-max normalization and z-score normalization in the k-nearest neighbor (knn) algorithm to test the accuracy of types of breast cancer. *null*, 2021.

[34] Mikel Galar, Alberto FernÃ¡ndez, Edurne Barrenechea, Humberto Bustince, Humberto Bustince, and Francisco Herrera. A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems Man and Cybernetics Part C (Applications and Reviews)*, 2012.

[35] Vignesh Sampath, Inaki Maurtua, J. Aguilar, and Aitor Gutierrez. A survey on generative adversarial networks for imbalance problems in computer vision tasks. 07 2020.

[36] I-Cheng Yeh and Che-Hui Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems With Applications*, 2009.

[37] Esteban Alfaro, Noelia GarcÃa, MatÃas GÃ¡mez, and David Elizondo. Bankruptcy forecasting: An empirical comparison of adaboost and neural networks. *Decision Support Systems*, 2008.

[38] Esteban Alfaro, Noelia GarcÃa, MatÃas GÃ¡mez, and David Elizondo. Bankruptcy forecasting: An empirical comparison of adaboost and neural networks. *Decision Support Systems*, 2008.

[39] Lucia F. Dunn, TaeHyung Kim, and Taehyung Kim. An empiricial investigation of credit card default. *null*, 1999.

[40] Joanna Stavins. Credit card borrowing, delinquency, and personal bankruptcy. *New England Economic Review*, 2000.

[41] 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*, 2020.

[42] Begum Cigsar and Deniz Ãnal. Comparison of data mining classification algorithms determining the default risk. *Scientific Programming*, 2019.

[43] Peshawa Muhammad Ali and Rezhna Faraj. Data normalization and standardization: A technical report. 01 2014.

[44] S. Gopal Krishna Patro and Kishore Kumar Sahu. Normalization: A preprocessing stage. *CoRR*, abs/1503.06462, 2015.

[45] Aida Ali, Siti Mariyam Shamsuddin, and Anca Ralescu. Classification with class imbalance problem: A review. 7:176–204, 01 2015.

[46] Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *CoRR*, abs/1710.05381, 2017.

[47] Justin Johnson and Taghi Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6:27, 03 2019.

[48] Ravi Prakash and Karmanya Kumar. Class weight technique for handling class imbalance. 07 2022.

[49] Ziyu Xu, Chen Dan, Justin Khim, and Pradeep Ravikumar. Class-weighted classification: Trade-offs and robust approaches. 2020.

[50] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *CoRR*, abs/1404.7828, 2014.

[51] Bernhard Mehlig. Artificial neural networks. *CoRR*, abs/1901.05639, 2019.

[52] Keiron O'Shea and Ryan Nash. An introduction to convolutional neural networks. *CoRR*, abs/1511.08458, 2015.

[53] Yun Xu and Royston Goodacre. On splitting training and validation set: A comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. *Journal of Analysis and Testing*, 2, 10 2018.

[54] Kevin Dobbin and Richard Simon. Optimally splitting cases for training and testing high dimensional classifiers. *BMC medical genomics*, 4:31, 04 2011.