IBN HALDUN UNIVERSITY
SCHOOL OF GRADUATE STUDIES
DEPARTMENT OF MANAGEMENT

MASTER THESIS

A MACHINE LEARNING APPROACH TO PREDICT CUSTOMER CHURN IN THE BANKING SECTOR USING CRISP-DM AND KNIME ANALYTICS

MORO HUSSEINI

THESIS SUPERVISOR
PROF. MUSTAFA KEMAL YILMAZ

ISTANBUL, 2024

IBN HALDUN UNIVERSITY
SCHOOL OF GRADUATE STUDIES
DEPARTMENT OF MANAGEMENT

MASTER THESIS

A MACHINE LEARNING APPROACH TO PREDICT CUSTOMER CHURN IN THE BANKING SECTOR USING CRISP-DM AND KNIME ANALYTICS

by
MORO HUSSEINI

A thesis submitted to the School of Graduate Studies in partial fulfilment of the requirements for the degree of Master of Arts in Management

THESIS SUPERVISOR
PROF. MUSTAFA KEMAL YILMAZ

ISTANBUL, 2024

APPROVAL PAGE

This is to certify that we have read this thesis and that, in our opinion, it is fully adequate, in scope and quality, as a thesis for the degree of Master of Arts in Management.

Thesis Jury Members

Title - Name Surname	Opinion	Signature
_____	_____	_____
_____	_____	_____
_____	_____	_____

This is to confirm that this thesis complies with all the standards set by the School of Graduate Studies of Ibn Haldun University:

Date of Submission Seal/Signature

ACADEMIC HONESTY ATTESTATION

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name Surname:

Signature:

ÖZ

CRISP-DM VE KNIME ANALİTİĞİ KULLANARAK BANKACILIK SEKTÖRÜNDE MÜŞTERİ KAYBINI TAHMİN ETMEYE YÖNELİK BİR MAKİNE ÖĞRENMESİ YAKLAŞIMI

Husseini, Moro
İşletme Tezli Yüksek Lisans Programı
Öğrenci Numarası: 214040010
Open Researcher and Contributor ID (ORC-ID): 0000-0030-0296-1625
Ulusal Tez Merkezi Referans Numarası: 10595577

Tez Danışmanı: Prof. Dr. Mustafa Kemal Yılmaz
Ocak 2024, 70 Sayfa

Bankacılık sektöründe yeni müşteri edinmenin maliyeti mevcut müşteriyi elde tutmanın maliyetinde daha fazla olup, bankalar mevcut müşterilerin ayrılmalarını önlemek için ayrılma olasılıklarını önceden tahmin etmeye çalışmakta ve bu potansiyeli taşıyan müşterilerini elde tutmaya yönelik stratejiler geliştirmektedirler. Bu çalışmanın amacı, CRISP-DM ve KNIME Analytics yöntemini birlikte kullanarak, beş farklı makine öğrenimi modeli (karar ağacı, rassal orman, lojistik regresyon, destek vektör makinası, yapay sinr ağları) ile çok uluslu ABC Bankası'nın Ağustos 2022 dönemi verilerini kullanarak bir tahminleme model oluşturmaktır. Elde edilen sonuçlar, rassal orman modelinin bankanın müşteri kaybını tahmin etmekte %78,91 genel doğruluk puanı ile en iyi sonucu verdiğini, karar ağacı ile lojistik regresyon modellerinin ise sırasıyla %71.55 ve %71.3 genel doğruluk puanı ile en düşük performansı gösteren modeller olduğunu göstermiştir. Ayrıca çalışma sonuçları, müşterinin banka ile olan geçmiş ilişkisinin, kredi notunun ve yaşının ayrılma potansiyeli taşıyan müşterileri tespit etmekte en etkin faktörler olduğunu ortaya koymuştur. Bu açılardan çalışma, finansal kuruluşlar için değerli bilgiler sunmaktadır.

Anahtar Kelimeler: Bankacılık, CRISP-DM, KNIME Analitiği, Müşteri Kaybı.

ABSTRACT

A MACHINE LEARNING APPROACH TO PREDICT CUSTOMER CHURN IN THE BANKING SECTOR USING CRISP-DM AND KNIME ANALYTICS

Husseini, Moro

MA in Management

Student ID: 214040010

Open Researcher and Contributor ID (ORC-ID): 0000-0030-0296-1625

National Thesis Centre Reference Number: 10595577

Thesis Supervisor: Prof. Mustafa Kemal Yılmaz

January 2024, 70 Pages

Attracting new customers is more expensive than maintaining the existing ones. One way of preventing customers from churning is to develop techniques for predicting their likelihood to churn. This study aims to forecast the customer churn of a multinational bank by using one-month period data, from July 31, 2022, to August 29, 2022. We employed CRISPM-DM, in conjunction with KNIME Analytics, to build several predictive models, i.e., decision tree, random forest, logistic regression, artificial neural networks, support vector machine, and ensemble models to predict customer churn. The results show that the random forest model has the highest performance in accurately predicting the churn of bank clients by its high overall accuracy of 78.91% and AUC score of 85.3%. The decision tree model, with an overall accuracy of 71.55%, and the logistic regression model, with an overall accuracy of 71.3%, are the least-performing predictive models. The findings also show that the customer's historical record with the bank (product_number), credit score, and age have the highest predictive power for customer churn. This study offers valuable insights for financial institutions. Using reliable predictive models, banks may identify potential clients likely to switch to other financial institutions. This identification would allow banks to design innovative marketing strategies and powerful customer relationship management to prevent consumers from churning.

Keywords: Banking Sector, CRISP-DM, Customer Churn, KNIME Analytics.

DEDICATION

I dedicate this thesis to Mr. Abdul Karim Salifu, who sponsored my education from high school until I got my bachelor's degree in 2019.

ACKNOWLEDGEMENT

In the name of Allah, the beneficent, the merciful. All praise is due to Allah, the Lord of the world. May the blessings and peace of Allah be on the final messenger, Muhammed (SAW), his companions and all the rightly guided Prophets.

To begin with, I would like to acknowledge the support of Professor Mustafa Kemal Yilmaz and Dr. Melike Zehir as my thesis supervisor and co-supervisor, respectively. I completed this enormous work on time with their constant support and encouragement. Their supervision was unique. They gave me guidelines on how to improve my work from the beginning till the end. I also acknowledge the enlightenment I received through studying under Professor Umit Hacioglu, Assistant Professor Omar Kachkar and Professor Dursun Delen, who influenced my interest in analytics research.

Moreover, I also acknowledge the support of all my family members, especially my beloved mother and friends, who kept encouraging me to work hard to finish on time. Mr. Abdul Aziz Husseini, Suleiman Husseini and Dr. Abdullah Husseini deserve special mention for their advice and moral support. Special acknowledgement also goes to Mr. Musah Abdulai, my friend who never ceases to advise me when needed.

Finally, I acknowledge my coursemate at Ibn Haldun University, Ibrahim Bague, for the time he spent helping me with the final editing and organisation of the thesis.

Moro Husseini
ISTANBUL, 2024

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

AI	Artificial Intelligence
AUC	Area Under the Curve
CCP	Customer Churn Prediction
CM	Confusion Matrix
CNN	Convolution Neural Networks
CRISP-DM	Cross-Industry Standard Process for Data Mining
CRM	Customer Relations Management
DM	Data Mining
DNN	Deep Neural Network
DT	Decision Tree
ETL	Extract, Transform, and Load
FN	False Negative
FP	False Positive
GUI	Graphical User Interface
IS	Information System
JDBC	Java Database Connectivity
KDD	Knowledge Discovery in Databases
KNIME	Konstanz Information Miner
k-NN	k-Nearest Neighbours
LR	Logistic Regression
MART	Multiple Additive Regression Trees
ML	Machine Learning
MS	Management Science
NB	Naïve Bayes
NN/ANN	Neural Network/Artificial Neural Network
RBM	Restricted Boltzmann Machine
RF	Random Forest
ROC	Receiver Operating Characteristic Curve
SEMMA	Sample, Explore, Modify, Model and Assess
SGB	Stochastic Gradient Boosting
SHAP	SHapely Additive Explanations
SVM	Support Vector Machines
TN	True Negative
TP	True Positive

CHAPTER I

INTRODUCTION

1.1. Background of the Study

The banking sector in the world has been facing a significant problem of customer churn over the last decade due to remarkable changes in financial technologies and diverse financial services offered by financial companies (Jahan & Farah Sanam, 2023; Kumar & Dhandapani, 2016; Seid & Woldeyohannis, 2022). Accenture (2021) reports that the rate at which bank customers churn is around 11%, while the churn rate of new clients is 20-25% throughout the first year, with half of the customers leaving in the first 90 days. The CallMiner Churn Index 2020 shows that US firms suffer an annual loss of USD 136.8 billion due to preventable customer attrition (CallMiner, 2020).

Many studies show that getting new customers is costly and requires a substantial financial commitment, whereas retaining existing ones is relatively less expensive (Hegde & Mundada, 2019; Iranmanesh et al., 2019; Jahan & Farah Sanam, 2023; Vafeiadis et al., 2015; Verma, 2020; Zoric, 2016). Li and Wang (2018) claim that it costs six times as much to acquire a new customer as it does to keep an existing one, while Vafeiadis et al. (2015) argue that the cost of acquiring a new customer could be 20 times higher than the cost of retaining an existing one. Verma (2020) and Tran et al. (2023) reveal that a 25% to 95% boost in profit could be possible with just a 5% improvement in client retention across different sectors. Dalbah et al. (2022) found that a 5% decrease in customer attrition may result in a significant profit increase for commercial banks, ranging from 25% to 85. Kumar & Ravi (2008) argue that boosting customer retention by 5% may lead to a decrease of 18% in operating expenses. Thus, customer churn significantly affects the financial performance of banks.

Poor customer experiences (Benoit & Van Den Poel, 2012), dissatisfaction with the standardisation of services (Charandabi, 2023), and unsatisfied expectations (Guliyev

& Tatoğlu, 2021) are among the major causes of customer attrition in the banking sector. Customers are prone to switch to a competitor when they face challenges such as long waiting times, intricate procedures, and insufficient assistance. They are also inclined to change their banks when they do not receive enough attention, encounter frequent problems and have inadequate personalised services (Charandabi, 2023).

Customer churn may have negative consequences for companies, including significant premium losses, reduced profit margins, and potential loss of referral business from loyal customers (Tran et al., 2023). Several studies have shown that using customer churn prediction algorithms can enhance companies' revenue and market position (Hou & Tang, 2010; Jahan & Farah Sanam, 2023; Tran et al., 2023). Sophisticated analytical tools present novel opportunities for understanding and forecasting customer behaviour (Karvana et al., 2019). Using customer data, financial institutions can identify preliminary indications of customer attrition and take pre-emptive actions to execute customer retention tactics (Benoit & Van Den Poel, 2012; Karvana et al., 2019; Zoric, 2016). According to Huang et al. (2012), machine learning algorithms possess the capability to scrutinise customer data, encompassing transaction history, demographics, and interaction patterns, with the objective of unearthing latent patterns and forecasting the probability of customer churn.

Banks employ analytical tools and machine learning algorithms to forecast customer attrition rates to enhance customer retention (Dalbah et al., 2022; Kaur et al., 2013; Zoric, 2016). Business Wire (2019) declared that banks that employ predictive analytics for churn prediction were able to curtail the churn rate from 10% to 3%. They can also enhance customer retention rate by 85% and increase the return on investment by 70%. Consequently, these financial institutions experience noteworthy reductions in expenses and increases in income. This evidence shows a strong relationship between customer retention and financial performance (Tran et al., 2023).

Understanding customer churn within the banking industry is also vital for optimising customer relationship management (Dalbah et al., 2022) and formulating efficient customer retention tactics (Charandabi, 2023). Financial institutions can optimise resource allocation, tailor product offerings, and deliver proactive customer services to mitigate the likelihood of customer attrition (Guliyev & Tatoğlu, 2021). Effective

customer churn prediction models allow banks to create customised customer retention strategies, including individualised communication, loyalty schemes, and focused marketing efforts.

This research aims to develop a model that can foretell bank client attrition using machine learning algorithms. Its primary goal is to help the bank track down clients who are considering leaving so that it can devise targeted methods to retain them and make them happier to remain competitive in the market.

1.2. Scope of the Study and Research Objectives

The rate of customer attrition has shown an upward trend in recent years, as reported by Alizadeh et al. (2023). Organisations now rank reducing client churn as one of their top priorities, given the substantial rise in competition across industries (Troncoso, 2018). In this research, we hope to construct a resilient machine-learning algorithm that can anticipate client attrition in the banking sector by employing CRISP-DM and the KNIME Analytics platform. The research is conducted by using data[1] from ABC Multinational Bank. We obtained the data from the Kaggle data platform, and we covered a one-month period from July 31, 2022, to August 29, 2022.

The objectives of the study can be summarised as follows:

I. To use the CRISP-DM methodology and KNIME Analytics to develop a machine learning model for predicting customer churn in the banking sector.

II. To identify the main signs that a customer will churn in the banking sector and to enable banks to target these customers with the right customer retention strategies.

III. To conduct a comparative analysis of the performance of the developed models against current approaches in the domain of customer churn prediction.

[1] https://www.kaggle.com/datasets/gauravtopre/bank-customer-churn-dataset

1.3. Significance of the Study

The use of machine learning in banking for the purpose of churn prediction is of great importance to multiple stakeholders (Belém, 2018; Jahan & Farah Sanam, 2023; Vafeiadis et al., 2015). This study provides a data-driven approach for banks to predict customer churn, allowing them to proactively identify and retain at-risk customers. By identifying potential churners, banks can execute customer retention tactics, including customised communication, individualised incentives, and enhanced customer support (Vafeiadis et al., 2015). By investigating the factors that lead to churn, financial institutions may get valuable knowledge regarding customer inclinations and areas of satisfaction, formulate customer-centric strategies, enhance service quality, and build sustainable customer relationships. The precise anticipation of the customer churn also empowers banks to optimise resource allocation and minimise costs.

The research employs several predictive models, such as DTs, RF, LR, artificial neural networks, SVM, and an ensemble model, to forecast bank clients who are prone to churn. With a minor alteration, the constructed models have the potential to forecast customer churn in other sectors, including telecommunications, e-commerce, and insurance.

While a lot of research has focused on forecasting banking customers' likelihood of leaving, to the best of our knowledge, none of them used the CRISPM DM in conjunction with KNIME Analytics except Kumar and Ravi (2008). By using machine learning theory, customer lifetime value theory, social network theory, and behavioural economics theory, this research endeavours to highlight the interaction between customer attrition and bank performance.

1.4. Structure of the Study

This thesis comprises five chapters that are summarised below:

Chapter 1 provides an overview of the purpose and inspiration of the study, as well as the specific goals of the research.

Chapter 2 provides the theoretical background for customer retention in the banking sector and reviews the literature. It also discusses the factors that influence customer attrition.

Chapter 3 presents the data, methodology, and research tools.

Chapter 4 provides the results and identifies potential areas for improvement in customer retention management in the banking sector.

Chapter 5 concludes and discusses the study's implications and makes recommendations for banks. The study's weaknesses are also discussed in Chapter 5, and suggestions for further research are made therein.

CHAPTER II

LITERATURE REVIEW

The emergence of technology and the availability of diverse products and services have led to intense competition in several sectors (Shirazi & Mohammadi, 2019). Lemos et al. (2022) stated that as customers embrace new technologies, their expectations for banking services rise. Over the years, the banking sector has experienced significant intensification in competition due to the challenges arising from rivals and inventive participants like Apple and Google (An et al., 2022). The primary goal of subscription-based entities like banks is to secure new subscribers and to maintain the loyalty of existing ones (Bahnsen et al., 2015). The reason behind this is that subscribers directly impact a company's profitability. The phenomenon of customer churn has exhibited an upward trend in recent years (Alizadeh et al., 2023). To optimise profitability, companies should expand their customer base by reducing customer churn.

Customer attrition is a major issue in the competitive markets (Guliyev & Tatoğlu, 2021). Clients are the lifeblood of the banking sector. Therefore, it stands to reason that if banks can predict which clients are likely to churn, they might devise stronger retention measures (Seid & Woldeyohannis, 2022). Iranmanesh et al. (2019) claim that customer churn has significant implications for firms across sectors, as the business landscape witness increases in how much it costs to bring in new clients compared to how much it costs to retain existing ones (Dalbah et al., 2022; Hegde & Mundada, 2019; Liu, Li & Huang, 2022; Liu et al., 2022; Seid & Woldeyohannis, 2022; Vafeiadis et al., 2015). How much it costs to bring in a new client can be 5 to 10 times higher than how much it costs to retain existing ones (Bahnsen et al., 2015; Liu et al., 2022).

How much it costs to bring in a new client can be attributed to several factors, including the entry of new rivals, the emergence of innovative business models, and the enhancement of services (Kumar and Dhandapani, 2016). This issue has motivated

many researchers to identify the category of clients that are more likely to leave a firm and to formulate the right retention strategies for keeping them from churning. This chapter examines the primary elements that contribute to customer churn in the banking industry and investigates the use of machine learning methods in forecasting customer churn and enhancing the accuracy of churn prediction models.

2.1. Conceptual Framework

2.1.1. Definition of Customer Churn and Its Business Impact

Customer churn, also called customer attrition (Agarwal et al., 2023; Alizadeh et al., 2023; Belém, 2018; Bharathi et al., 2022; Dalbah et al., 2022; Kaur & Kaur, 2020; Khine & Myo, 2019; Tran et al., 2023), turnover (Belém, 2018; Tran et al., 2023), or defection (Belém, 2018; Bharathi et al., 2022), is characterised by a gradual yet persistent decline in the number of customers over a period (Agarwal et al., 2023). It refers to the occurrence wherein a client expresses his or her desire to discontinue his/or her engagement with the services of a company (Kumar & Dhandapani, 2016).

In the banking sector, banks often experience the loss of clients (Dalbah et al., 2022). In such cases, customers close their accounts and discontinue engaging in any further transactions with the bank (Karvana et al., 2019). Poor customer experience (Benoit & Van Den Poel, 2012), dissatisfaction with the standardisation of services (Charandabi, 2023), and unsatisfied expectations (Guliyev & Tatoğlu, 2021) are among the major causes of customer attrition in the banking sector. Customers are prone to switch to a competitor when they face challenges such as long waiting times, intricate procedures, and insufficient assistance. They are also inclined to change their banks when they do not receive enough attention, encounter frequent problems, and receive inadequate services (Charandabi, 2023). Customer churn may have negative consequences for companies, including reduced profit margins and loss of business from loyal customers (Tran et al., 2023).

The analysis of customer attrition can be extended to encompass several sectors including telecommunications (Amin et al., 2017, 2019), marketing (De Caigny et al., 2018; Sabbeh, 2018), banking (Coser et al., 2020; Khine & Myo, 2019, 2023), and

insurance (Nagaraju & Vijaya, 2022), where the success of operations relies heavily on customer engagement and collaboration (Hegde & Mundada, 2019). The accessibility that information technology provides has rendered the act of transitioning between banking services an exceedingly effortless endeavour due to radical changes in customer preferences (Dalbah et al., 2022; Shirazi & Mohammadi, 2019). A report issued by Forbes suggests that in North America, bank customer churn is 11% each year on average, and banks are forced to use substantial marketing resources to acquire new clients to maintain the level of clients (Dube, 2020). Many studies reported an inverse relationship between customer churn rate and profitability (Karvana et al., 2019a; Kaur & Kaur, 2020; Kaur et al., 2013; Sabbeh, 2018). Charandabi (2023) claims that a mere 5% reduction in churn rate has the potential to lead to an 85% increase in profitability, while Verma (2020) indicates that the same 5% reduction in churn rates can lead to a massive 95% increase in firm profitability.

2.1.2. Significance of Customer Churn Prediction and Customer Retention

Customer churn analysis is important as it allows companies to identify current customers who display a propensity to discontinue their relationship with firms (Guliyev & Tatoğlu, 2021). This, in turn, enables entities to create focused retention campaigns to retain these customers. According to Iranmanesh et al. (2019), the detrimental impact of customer churn on a bank's revenue streams, particularly earnings and fee incomes, makes predicting it more important for banks. Moreover, the financial resources invested in acquiring a new client are five times higher than the resources invested in retaining existing ones (Bahnsen et al., 2015; Guliyev & Tatoğlu, 2021; Kumar & Dhandapani, 2016; Verma, 2020). This could be up to 20% (Vafeiadis et al., 2015). Dalbah et al. (2022) and Verma (2020) identified that a 25% to 85% boost in profits could be achieved with just a 5% decrease in client churn rates. Considering these statistics and the belief that more than 1.5 million instances of client attrition are recorded across several sectors annually (Karvana et al., 2019), customer churn prediction is quite important, given the fact that investors increasingly use churn rates to explore the viability of companies.

2.2. Data Mining

Data mining, which has its roots in the latter part of the 1970s and earlier part of the 1980s, is the process of discovering useful information from big datasets. From a technical standpoint, it is a procedure that employs AI, statistics, and mathematics to sift through enormous datasets in searching for relevant information. Data mining is conveniently located at the crossroads of numerous fields, such as databases, statistics, AI, ML, MS, and IS (Delen et al., 2018; Sharda et al., 2020). As shown in Figure 2.1, data mining covers several fields.

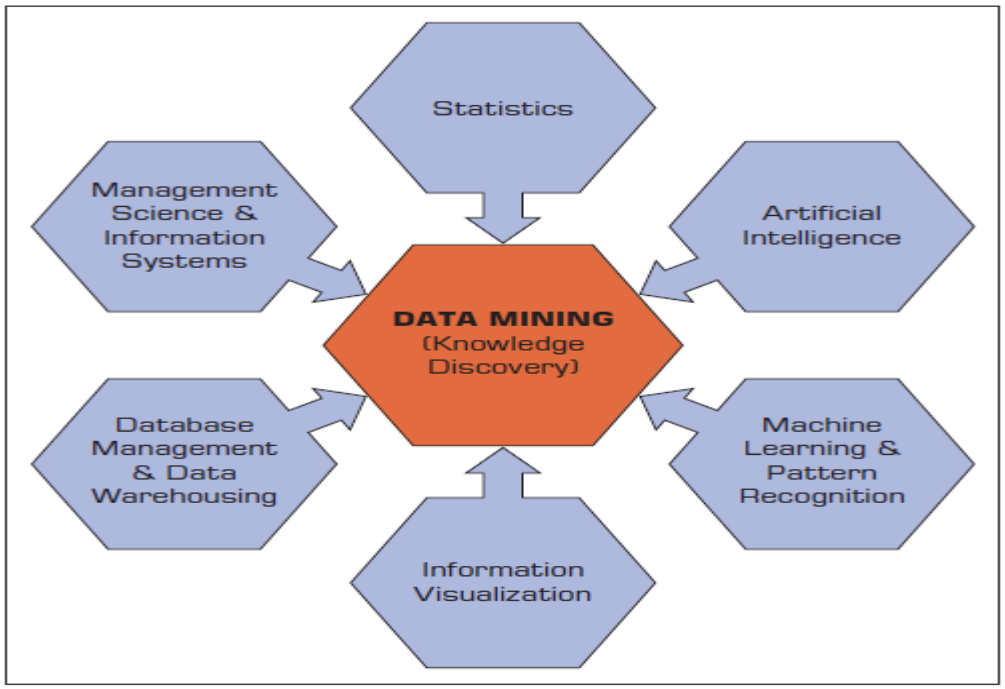


Figure 2.1. The Disciplines of Data Mining

Source: Sharda et al. (2020)

2.3. Predictive Modelling and Machine Learning Techniques

Predictive modelling enables decision-makers to anticipate the future by drawing insights from existing data. According to Bahnsen et al. (2015), customer churn predictive modelling involves the use of historical, behavioural, and socio-economic data to forecast the potential of customer attrition. It is a proactive endeavour to forecast the likelihood of a client terminating the relationship with a firm (Karvana et

al., 2019). Predictive models are used to predict churn (Kaur & Kaur, 2020). These models are designed to detect early indicators of customer churn and identify individuals who are likely to terminate their relationship with a firm.

According to Kumar and Dhandapani (2016), in machine learning, it is pertinent to acknowledge the existence of three approaches: unsupervised learning, semi-supervised learning, and supervised learning. Supervised learning involves the diligent exploration and identification of concealed patterns in the datasets that are labelled. Unsupervised learning involves the exploration and identification of latent patterns in data that lack explicit labelling. Semi-supervised learning is a category of learning tasks and methodologies in supervised learning that involves the utilisation of unlabelled data in addition to labelled data during the training process. This approach involves a relatively limited collection of labelled data accompanied by a larger collection of unlabelled data. Semi-supervised learning occupies an intermediary position in the broader spectrum of the machine learning paradigm, positioning between unsupervised learning and supervised learning. The subsequent section provides an overview of the five widely employed techniques in churn prediction, as captured by Sharda et al. (2020). These techniques are evaluated based on their credibility, effectiveness, and prevalence.

2.3.1. Artificial Neural Network (ANN)

Neural networks (NNs) serve as a metaphor for information processing that draws upon the structure and functioning of the human brain. NNs are derived from biological inspiration rather than being a precise replication of the brain's actual functioning. Their generalizability, capacity to glean knowledge from data and lack of dependence on strict assumptions make these models ideal for use in prediction and business classification applications. In the context of machine learning, "neural computing" refers to a pattern recognition paradigm. The model generated using neural computing is referred to as an artificial neural network (ANN). Pattern recognition, forecasting, prediction, and classification are some of the many business domain applications of NNs. NN computing is an essential element inside the toolkit of data science and business analytics and has extensive use in finance, marketing, manufacturing, information systems, and other fields.

ANNs are comprised of interconnected artificial neurons that bear a resemblance to the structure of their biological counterparts. During the information processing phase, the processing units in an ANN function in a collective manner, resembling the behaviour of biological neurons. These include the capacity to acquire knowledge, self-organize, and maintain resilience in the faults. Over the last decades, many studies have extensively examined ANNs. The formal investigation of ANN started with the ground-breaking research conducted by McCulloch and Pitts (1943). They proposed a rudimentary model of an artificial neuron with binary functionality, drawing inspiration from biological experiments. They constructed a neural network model by employing interconnected artificial binary neurons that were designed to simulate the brain's functioning by using information-processing devices.

In the last two decades, there has been a notable development in the field of ANN studies. The development of new learning algorithms, activation functions, and network topologies, as well as progress in cognitive science and neurology, are responsible for this. Moreover, significant progress in theoretical frameworks and research methodologies has successfully addressed numerous challenges that previously impeded the advancement of ANN studies. The increasing acceptability of ANNs is supported by the compelling findings of multiple studies.

The efficacy of ANN applications has sparked interest in the business environment. The rise of deep NNs, a key component of the recent deep learning trend, has led to a surge in NNs with more complex architectures and improved analytical capabilities, generating significant anticipation for the potential of this new generation of NNs. An extremely limited set of ideas borrowed from biological NNs constitute the basis of neural computing. The model’s naming is more of a metaphor than a precise depiction of the human brain. Figure 2.2 shows how ANNs function in processing information.

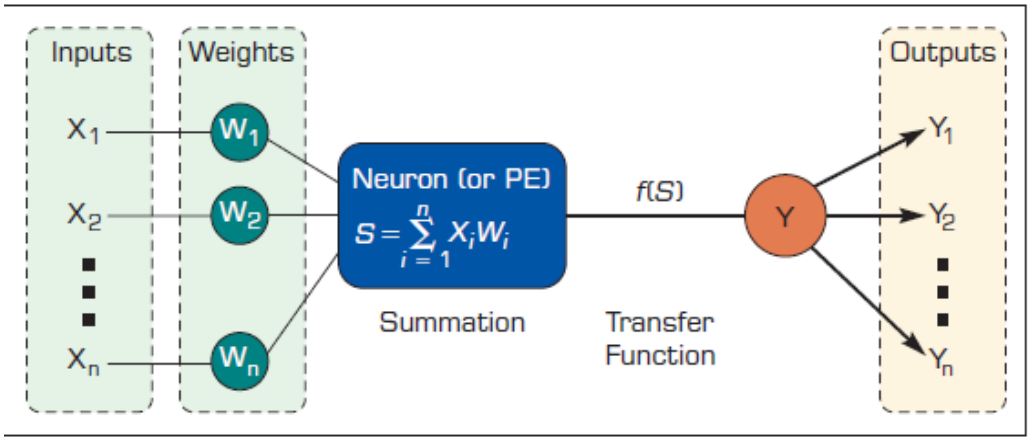


Figure 2.2. The Process of Information Processing in an Artificial Neuron

Source: Sharda et al. (2020)

2.3.2. Support Vector Machines

SVMs have become popular as machine learning techniques due to their exceptional prediction capabilities and strong theoretical underpinnings. SVMs are a type of supervised learning algorithm that generates input-output functions based on a given collection of labelled data for training. One way to look at the input-output relationship is as a classification, wherein cases are assigned to certain classes, or as a regression, wherein the intended output's continuous numerical value is estimated. In classification, it is common to employ non-linear kernel functions to convert input data that inherently captures intricate nonlinear relationships into a feature space with a higher dimensionality. This transformation facilitates the linear separability of the input data. Subsequently, the hyperplanes with maximal margins are formed to isolate each output class effectively within the training set.

In the context of a classification-type prediction issue, it is commonly observed that multiple linear classifiers, i.e., hyperplanes, can effectively partition the data into distinct subsets, with each subset corresponding to a specific class. This can be visualised in Figure 2.3a, in which the two groups are represented by the shapes of circles and squares. It is important to note that there is only one hyperplane that can accomplish the maximum separation between the classes. This is shown in Figure 2.3b,

where the two groups are divided by that hyperplane and two other hyperplanes with maximum margins.

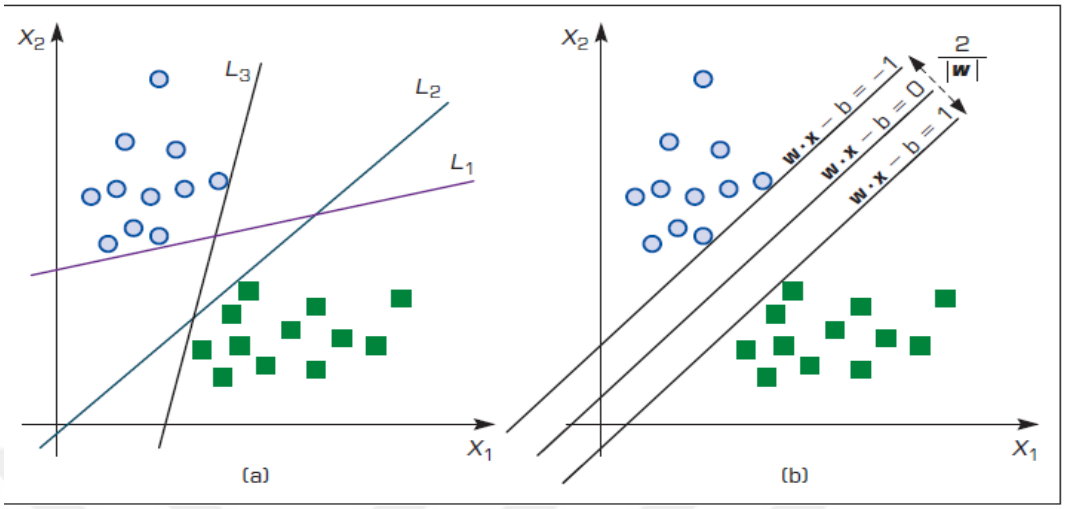


Figure 2.3. Separating Classes with Hyperplanes

Source: Sharda et al. (2020)

SVM, like ANN, can approximate any multivariate function with desired accuracy. Therefore, SVM is ideal for modelling complex, nonlinear systems and processes. Figure 2.4 shows a simple SVM model development process.

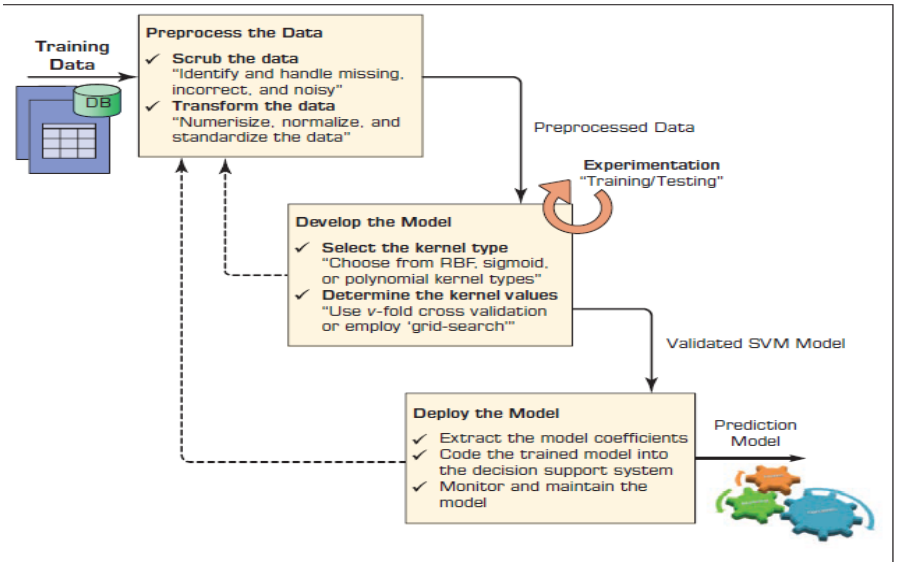


Figure 2.4. Simple SVM Model Development Process

Source: Sharda et al. (2020)

2.3.3. Nearest Neighbour Prediction Method

Data mining algorithms are computationally intensive. ANNs and SVMs need time-consuming and computationally difficult mathematical derivations, while the k-nearest neighbour algorithm (k-NN) appears simple for competitive prediction. K-NN is a technique for making predictions in tasks that are classification and regression tasks in their nature. As a learning method by instances, k-NN is a sort of lazy learning which approximates functions locally and delays computations until prediction. A majority vote from nearby cases determines a case's classification in predictions, which are classification in nature. The idea is illustrated in Figure 2.5 using a space with two dimensions that represent the values of the variables (x, y).

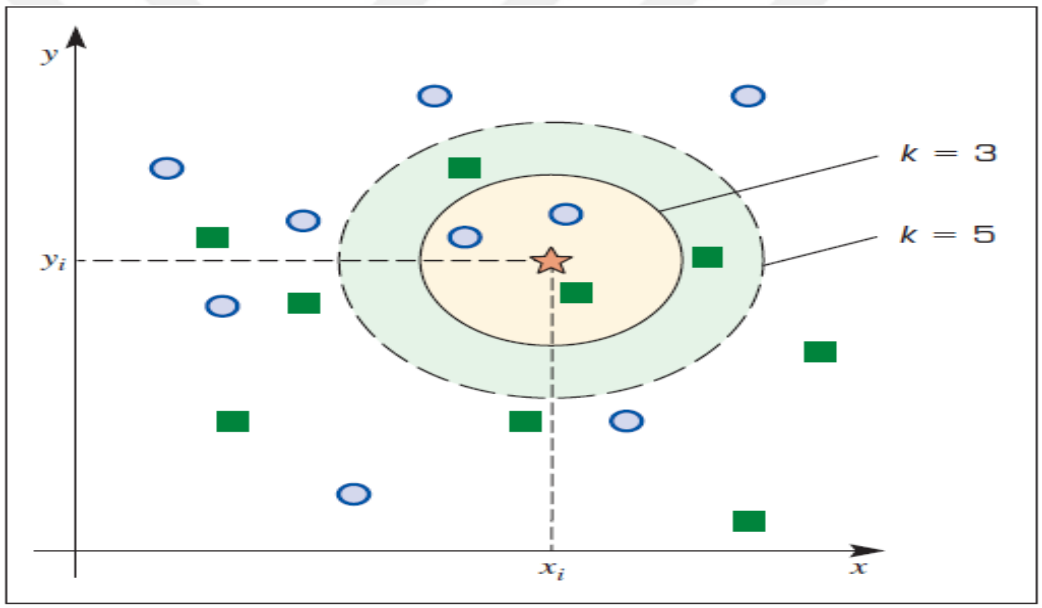


Figure 2.5. The Significance of k in the k-NN Algorithm

Source: Sharda et al. (2020)

The star in Figure 2.5 represents a new case, whereas the circles and the squares show known occurrences. Here, when there is a new item, it will be assigned to either the circles or the squares based on which one the new item is similar to. For instance, if we set k to 1, i.e. (k = 1), the new item will be assigned to the squares, which is the closest example to the star. Regression-type prediction tasks can employ a similar strategy as well.

2.3.4. Naïve Bayes Method

NB is a technique for classifying data that relies on probability. It is derived from Bayes' theorem and it is used for prediction problems that are classification by nature. The name "Naïve" comes from the idea that it is based on strong and unreasonable assumptions about inputs' independence. It is the underlying assumption of NB classifiers that all inputs are completely independent of one another and that the existence or omission of any one variable has no bearing on the existence or the absence of any of the others. Supervised machine learning is the ideal setting for creating NB classification models. Due to the independence assumption, NB models can be developed without fully adhering to all the requirements of the Bayes theorem.

2.3.5. Ensemble Modelling

Ensembles combine the outputs of multiple analytics models to create a single output. They are commonly employed in prediction modelling to improve predictions by combining the scores of multiple models. The kind of prediction might be a classification or a regression, as stated by Sharda et al. (2020). Regression estimates a numerical output, whereas classification predicts a class label. Not only can ensembles be utilised for prediction modelling, but they are also applicable to clustering and association rule mining, two more analytics tasks. Model ensembles are versatile enough to be employed for ML jobs requiring either supervised or unsupervised learning.

Researchers and practitioners construct ensembles to improve accuracy and enhance the stability, robustness, consistency, and reliability of outcomes. Ensembles enhance predictive accuracy for a given problem (Sharda et al., 2020). Ensembles became popular for winning data mining and predictive modelling competitions in the early to mid-2000s. Researchers and practitioners were invited to participate in the prestigious Netflix prize, an open competition, to forecast user ratings of films based on ratings from the past. The ultimate winning team of the USD 1 million prize, as well as all the other teams who made it to the top in the rankings, used model ensembles in their prediction. According to Sharda et al. (2020), the winners used hundreds of models to build their ensemble models.

Model ensembles have consistently shown the ability to enhance model accuracy as well as improve model robustness, stability, and reliability. Ensemble models combine multiple models into a single prediction outcome by using some form of averaging. This approach prevents any single model from dominating the final prediction, thereby reducing the likelihood of making inaccurate or extreme predictions. Figure 2.6 displays the graphical presentation of model ensembles for classification-based prediction problems.

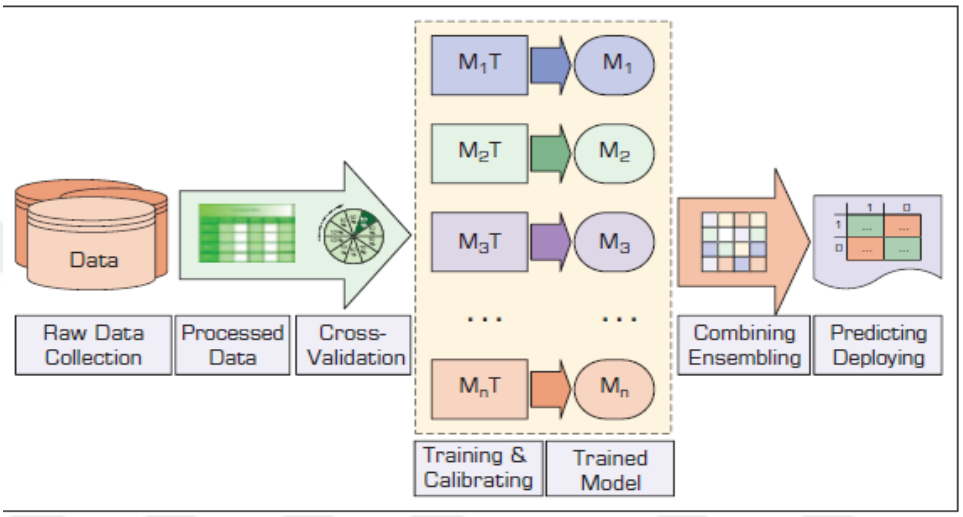


Figure 2.6. Ensemble Modelling for Prediction: A Visual Presentation

Source: Sharda et al. (2020)

The bias-variance trade-off is a prominent topic in predictive modelling that holds significant relevance to the use of model ensembles. Hence, prior to exploring many categories of ensembles, it is imperative to examine the notion of bias-variance trade-off in the ML. In predictive data analytics, when comparing prediction accuracy across various data sets, "variance" refers to the degree of consistency (or lack thereof) and "bias" refers to the existence of error. It is expected that the best models will have low variance, showing consistency in accuracy across different datasets, and little bias, showing high accuracy. When building predictive models, there is an unfortunate trade-off between these two metrics since improving one statistic degrades the other.

The minimal bias on the data used for training a model has the potential to cause the model to exhibit large variance on the hold-out or validation data due to potential

overtraining of the models. For example, the k-NN algorithm with a value of k equal to 1 can be seen as a model with low bias, meaning it performs exceptionally well on the training data. However, it is prone to significant variation when applied to a separate test or validation data set. The prevailing approach in addressing the trade-off between bias and variance in predictive modelling is the use of cross-validation in line with the appropriate model ensembles.

2.3.5.1. Different Ensemble Model Types

Figure 2.7 shows that there are typically four types of model ensembles that are defined by two distinct dimensions. The x-axis in Figure 2.7 represents the method used to categorise the ensembles into bagging and boosting types, respectively. According to Abbott (2014), ensembles are classified as either homogeneous or heterogeneous based on the second dimension, which represents the type of model.

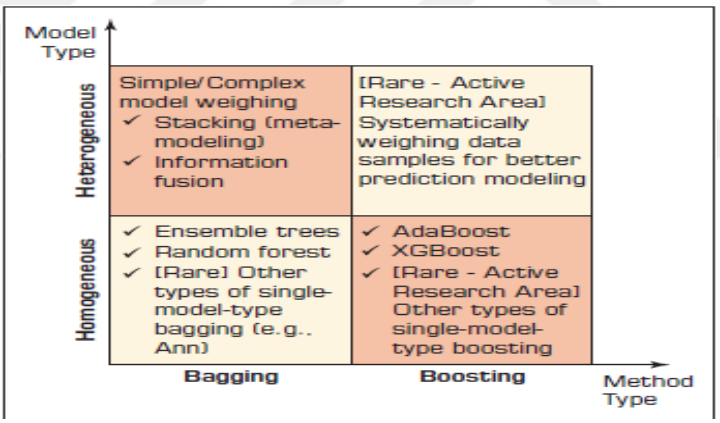


Figure 2.7. Simple Model Ensemble Taxonomy

Source: Sharda et al. (2020)

2.3.5.1.1. Bagging

The most popular and straightforward ensemble method is bagging (Abbot, 2014). Bagging, also known as bootstrap aggregating, was first introduced at UC Berkeley by Breiman (1996). The process is straightforward but efficient: it uses resampled data to generate multiple decision trees, which then average or vote to combine predicted values. Bootstrap resampling, which involves duplicating records in the training data,

was utilised by him. According to Abbott (2014), this method of selection typically leaves out 37% of the data from the training set. Although it was originally designed for decision trees, bagging can be applied to any other predictive modelling system with significant diversity in anticipated values. While uncommon, additional predictive modelling methods such as NNs, NB, k-NN, and logistic regression may be suitable for bagging-type model ensembles. Bagging with k-NN is not recommended for big k values as the algorithm already averages or votes on predictions, resulting in consistent and low variance predictions.

2.3.5.1.2. Boosting

The next most common ensemble technique after bagging is boosting. Freund and Schapire (1996) developed the AdaBoost boosting algorithm in the beginning of the 1990s. Boosting is as simple as bagging. Start by creating a simple classification model that is a little bit better than an arbitrary chance for a 50% correct classification for binary classification. In the first step, similar to a standard prediction model, the algorithm uses each record with an equal weight for each case. The errors in the projected values are noted for each scenario. Case weights for accurately categorised records will either stay the same or go down, while those for incorrectly classified records will increase. With the transformed/weighted-training dataset set, a second, simpler model is subsequently constructed. The second model uses case weights to give improperly classified records more weight in the prediction model. In each iteration, hard-to-classify records receive higher case weights, signalling the algorithm to focus on them until they are correctly classified.

2.3.5.2. Different Variations of the Bagging and Boosting Algorithms

2.3.5.2.1. Random Forest

The RF model was introduced by Breiman (2001) as a variation to the conventional bagging procedure. The algorithm starts with a dataset that was sampled using bootstrapping and creates a DT for each of the samples. For each tree split, beginning with the first split, random forest evaluates a different subset of input variables, setting it apart from bagging. In the RF, bootstrap sampling is used for random picking of

cases and features. RF model construction involves determining the cases, variables, and trees. Typically, the default number of variables at each point of splitting is the square root of the entire number of candidate inputs. In a model with 100 candidate inputs, a random of 10 inputs is chosen for each split. Prediction outputs using the RF model are, in most cases, more accurate when compared to prediction outputs using basic bagging and boosting methods like the AdaBoost.

2.3.5.2.2. Stochastic Gradient Boosting

AdaBoost remains the most popular boosting approach in commercial software, while other boosting variations are available in open-source software packages. The SGB, which was developed at Stanford University by Friedman (1999), has gained prominence due to its improved performance. In 2001, Friedman created a more advanced technique called MART that Salford Systems eventually marketed as TreeNet' in their software package. MART, like other boosting algorithms, creates basic trees repeatedly and adds them together. After creating the initial tree, residuals, which are also referred to as errors, are calculated. The second, as well as all the subsequent trees, use the residuals as their target variable. Poor prediction of initial errors leads to huge errors in the following tree, while good prediction leads to minor errors. After building hundreds of trees, the final forecasts are obtained by combining the hundreds of trees, which are all piecewise constant models, into an additive form. Individual tree details are often overlooked due to the huge number of trees involved in the model ensemble (Abbott, 2014). TreeNet technique won numerous DM competitions due to its excellent prediction and minimal data cleaning requirement.

2.4. Customer Relationship Management (CRM)

CRM refers to the systematic approach of obtaining, retaining, and expanding the client base in a profitable manner (Kaur et al., 2013). It is an extensive approach to developing and sustaining long-term customer relationships (Vafeiadis et al., 2015). During the last two decades, companies have transitioned their marketing approach from product-centred strategies to customer-centred strategies (Shirazi & Mohammadi, 2019). Consequently, the dynamics of customer-firm connection have undergone significant transformation, leading to the emergence of numerous novel

marketing prospects and making the retention of customers a top priority in the CRM (Iranmanesh et al., 2019; Kaur et al., 2013).

2.5. Cross-Industry Standard Process for Data Mining (CRISP-DM)

To systematically execute data mining tasks, it is important to adhere to a standardised procedure (Azevedo & Santos, 2008; Sharda et al., 2020; Wirth & Hipp, 2000). Various processes have been developed by data mining researchers and practitioners to enhance the likelihood of success. These processes can be in the form of workflows or in the form of systematic approaches. The most prominent among them are the CRISP-DM, SEMMA, and KDD (Delen et al., 2018; Sharda et al., 2020). Figure 2.8 shows the CRISP-DM technique that we use in this study due to its widespread acceptance in analytics by many studies (Belém, 2018; Hegde and Mundada, 2019; Karvana et al., 2019; Kumar and Ravi, 2008; Troncoso, 2018). The following section describes the processes of this framework.

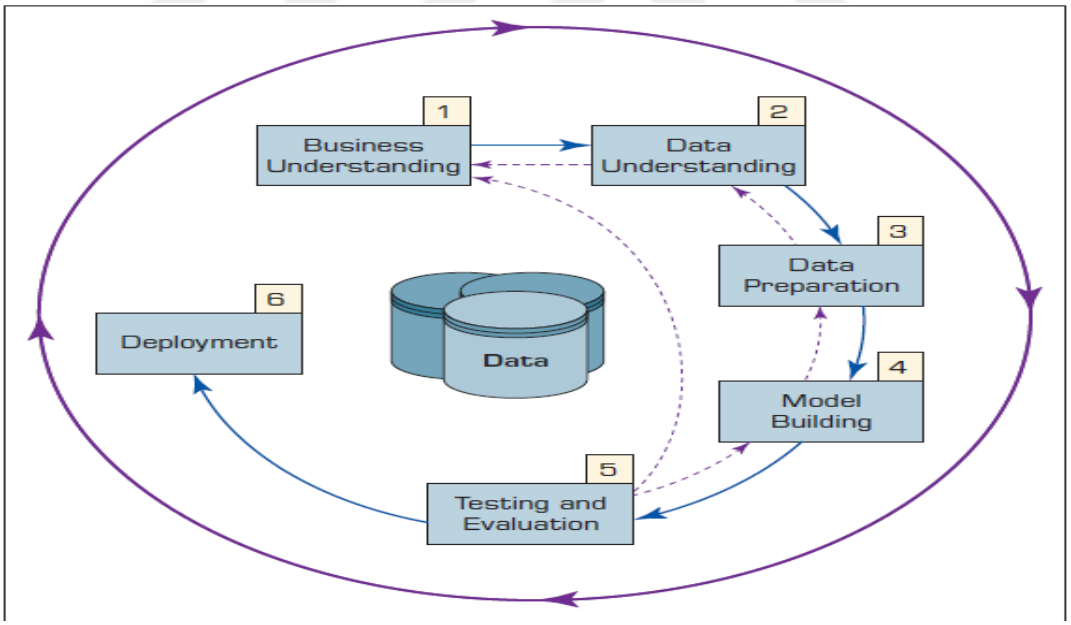


Figure 2.8. CRISP-DM Data Mining Process

Source: Delen et al. (2018) and Sharda et al. (2020)

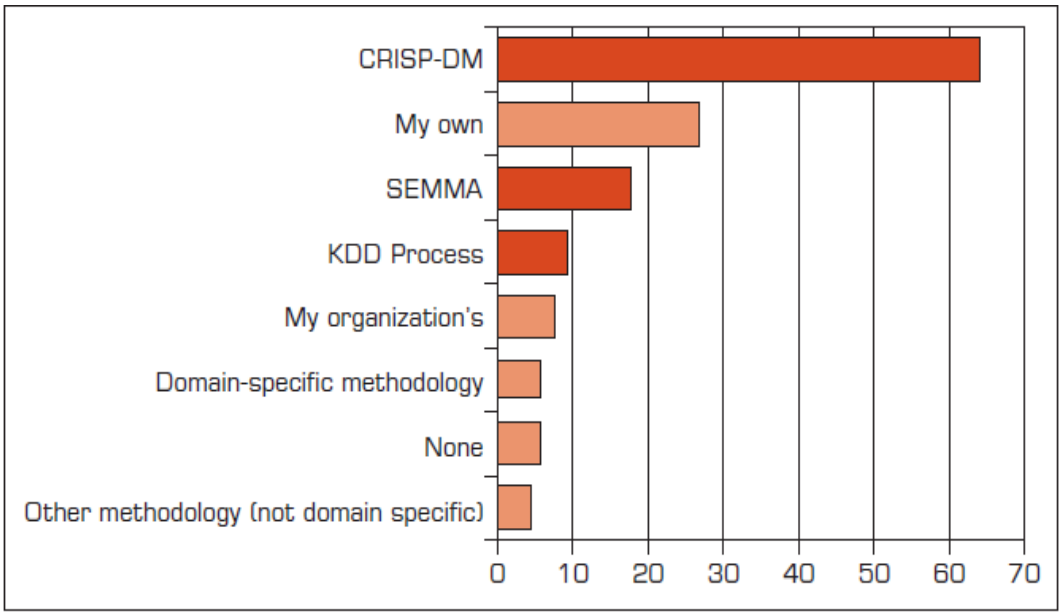


Figure 2.9. The Ranking of the Most Widely Used DM Methodologies

Source: Sharda et al. (2020)

2.5.1. Business Understanding

The fundamental aspect of every data mining investigation entails understanding the purpose of the study (Delen et al., 2018; Sharda et al., 2020). This identification assists in formulating a precise definition of the business objective. At this stage, the study addresses certain inquiries, such as the identification of shared attributes among customers, i.e., the typical customer profiles and their respective value contributions, that have recently switched to the competitors.

2.5.2. Data Understanding

In line with the initial stage, the main objective of a data mining process entails the identification of pertinent data from a multitude of accessible databases (Sharda et al., 2020). To gain a deeper understanding, analysts frequently utilise various graphical and statistical methods. These techniques include the summary statistics of the various variables in the data, such as the mean, median, standard deviation, and maximum and minimum values, and generating mode and frequency tables for categorical variables. Moreover, to understand the data, we can also make use of scatterplots, correlation

analysis, box plots, and histograms. By identifying data sources and pertinent variables, DM algorithms can more efficiently cover valuable patterns of knowledge.

2.5.3. Data Preparation

Another name for data preparation is data pre-processing. This step involves taking previously identified data and making it ready for analysis using DM techniques. It is the most time-consuming and effort-intensive step in the entire CRISP-DM process. It constitutes approximately 80% of the time dedicated to a DM project due to the incomplete, noisy, and inconsistent nature of the real-world data (Delen et al., 2018; Sharda et al., 2020).

2.5.4. Model Building

In model building, a number of modelling techniques are built and applied to a dataset to meet a specific business requirement. As part of this process, we also compare and contrast the different models. Since there isn't a single, best way to do data mining, it's necessary to employ a number of different models in accordance with a clearly defined strategy for testing and evaluation in order to find the optimal approach. Verma (2020), Kaur & Kaur (2020), and Lemos et al. (2022) identified the RF to be the best-performing model, while other studies identified different models as the best model: SVM (Karvana et al., 2019), ANN (Charandabi, 2023), XgBoost model (Guliyev & Tatoğlu, 2021) and NB (Agarwal et al., 2023).

In this study, we specifically used the DT, SVM, ANN, LR, and RF models to facilitate the comparison of their performance with the results reported in the prior studies, as these models are the most extensively employed models in the literature.

2.5.5. Testing and Evaluation

Here, we evaluate to which extent the chosen model or models align with the aims of the company and determine whether further models need to be produced (Delen et al., 2018). An alternative action is to do a practical evaluation of the produced model(s) in a real-world context, provided that there are no limitations imposed by time and budget

constraints. While it is anticipated that the results of the models will be relevant to the initial business objectives, it is not uncommon to uncover additional findings that may not directly align with these goals but could provide valuable information or suggestions for future endeavours.

2.5.6. Deployment

The complexity of the deployment phase varies depending on the requirements (Sharda et al., 2020). It can range from a straightforward task of generating a report to a more intricate process of implementing a DM process that can be replicated throughout the company. In numerous instances, the responsibility of executing deployment processes lies with the customer rather than the data analyst. Nevertheless, when the analyst does not undertake the deployment endeavour, it is imperative for the client to have a clear understanding of the necessary steps that must be executed to use the models generated. The deployment phase could encompass maintenance tasks for the models deployed, particularly as the maintenance and monitoring of these models are crucial for integrating DM outcomes into the daily operations and overall business.

2.6. Literature Review

The churn predictive modelling has received extensive attention from the data mining and machine learning communities (Bahnsen et al., 2015). Researchers have focused on employing classification algorithms to identify the patterns exhibited by churners. In recent years, scholars have used different data mining methodologies to anticipate customer attrition in the banking industry (Dalbah et al., 2022). Shirazi and Mohammadi (2019) developed a predictive churn model in the financial sector by leveraging big data. For this, they had to combine organised archival data with unstructured data from a variety of sources, such as internet web pages, records of website visits, and conversation logs from phone calls. Notably, this approach represents the first instance of using such a dataset in the financial sector. The authors also investigated the impact of several dimensions of consumers' behaviour on their decisions to churn. In order to analyse the retirement trajectory of clients and create a model for churn prediction, they utilised the big data analytics tool called Datameer in

conjunction with the Hadoop platform and applied predictive approaches through the business intelligence system.

Verma (2020) examined the prediction of customer churn for savings accounts by using statistical and machine-learning models. He incorporated under-sampling techniques to enhance the predictive accuracy of the models by considering the imbalanced nature of the customer churn rate in the dataset. Model accuracy, receiver operating characteristic (ROC) curve, AUC, and Gini coefficient were employed to compare the models. The results show that it is evident that RF, among other machine learning models, exhibits the highest predictive accuracy of 78% for churn. Bahnsen et al. (2015) examined the prediction of customer churn in commercial banks by using the SVM model. To enhance the performance of the model, the authors employed a random sampling technique. The results showed that the technique significantly improved the predictive accuracy of the model. Kaur and Kaur (2020) forecasted customer churn in a bank by using a variety of machine learning models, including logistic regression (LR), decision tree (DT), k-NN, and RF. The work was carried out by using Python programming on the Google Colab platform. Based on criteria of sensitivity, specificity and accuracy, the authors assessed the performance of the models. They also looked at ensembling strategies to improve the performance of the models with lesser accuracy, such as averaging and max voting. The results show that RF performed better than other models.

Karvana et al. (2019) repeatedly evaluated five classification models by employing cross-class comparisons and using a dataset with 57 variables. The results indicate that customers at a private bank in Indonesia may best be predicted to leave by using data collected via class sampling. Guliyev and Tatoğlu (2021) examined the utilisation of explainable machine learning models, particularly focusing on the application of SHapely Additive explanations (SHAP) values. These values are employed to enhance the evaluation of machine learning models in the customer churn analysis. The results show that the XgBoost model has superior performance in comparison to the other ML techniques. Iranmanesh et al. (2019) presented a customer attrition prediction model designed for the retail customers segment of a commercial bank in Iran. Through the application of advanced data analysis techniques to customer-specific transaction and operational data, they present a suitable categorisation of customers based on the churn

rate. The authors employed the ANN algorithm in Python for their prediction. The findings indicate that occupations in the food services sector, along with technical services, exhibit the highest rate of customer turnover in the banking sector. The sports centres and households are subsequent categories experiencing high turnover rates in the banking services. The bank's lowest risky customers were kindergartens, governmental organisations, and counselling centres. The largest churn was observed among retail customers aged 30-40 years. Agarwal et al. (2023) conducted a study on the utilisation of ML algorithms for identifying banking customers who exhibit a propensity to switch to other financial institutions. The authors showed the efficacy of employing machine learning models, particularly LR and NB, in accurately predicting customer churn in the bank. This prediction is based on various variables, including age, location, gender, credit card information, balance, and other pertinent factors.

Seid and Woldeyohannis (2022) implemented a machine learning algorithm to forecast customer churn in the Commercial Bank of Ethiopia by using a sample size of 204,161 datasets, each including eleven features. In this study, the assessment metric employed to ascertain the optimal classifier was the accuracy of the model. In the Commercial Bank of Ethiopia, several supervised machine learning techniques were employed to forecast client churn, including LR, RF, SVM, k-NN, and DNN. The selection of features was conducted by using feature importance and a correlation matrix. The SMOTE technique was also employed to achieve data balance, and the outcomes for the selected algorithm were assessed. In the series of trials, it was observed that a DNN showed superior performance, achieving an accuracy of 79.32%, precision of 85.08%, and recall of 78.19%.

Alizadeh et al. (2023) utilised a supervised ML technique, specifically a DT, along with the change mining approach, to construct a model based on hard data. The utilisation of K-means clustering, an unsupervised machine learning approach, is commonly observed in conjunction with data pretreatment techniques. This work also examines the Dempster-Shafer theory and various methodologies for modelling soft data. The findings indicate that this model results in a more dynamic CRM in the banking sector. Hegde and Mundada (2019) developed an accurate predictive model in the banking domain by using the Enhanced Deep Feed Forward Neural Network Model to anticipate customer churn. The outcome is contrasted with other categories

of machine learning techniques, including LR, DT, Gaussian naïve Bayes algorithm, and ANN. The results indicate that the Enhanced Deep Feed Forward Neural Network Model outperforms the existing machine learning model in accurately predicting client churn rates in the banking industry.

Mahajan and Gangwar (2017) employed a hybrid combination of SVM and RF to predict customer attrition. The objective is to maximise the margin by finding the hyperplane that effectively separates the classes in a high-dimensional space. The larger the margin, the more significant the impact on the accuracy of the predicted outcome. In the RF algorithm, the pursuit of ensembling systems involves the assembly of multiple models instead of constructing a single model. By combining the accuracies of these models, a more reliable forecast can be obtained. The work is conducted by using the MATLAB tool, wherein the dataset comprises 3,333 rows and encompasses a total of 21 properties. The model ensemble achieved a better outcome in comparison to the individual model. Kaya et al. (2018) developed a dynamic behavioural model to predict the churn rate in the financial sector. The model's execution hinged on the inclusion of behavioural characteristics and spatiotemporal patterns. The experiment used credit card transactional data obtained from a prominent financial institution. The authors implemented a novel method for feature selection by using the concept of entropy of choice. This method aims to identify the most relevant features from a given dataset. The findings indicate that the dynamic behavioural model exhibited notably superior performance in comparison to the conventional approach for forecasting churn rates in the financial sector.

Gregory (2018) conducted a study to predict customer churn by using the extreme gradient boosting algorithm, commonly referred to as XGBoost. The input to the model consists of transactional and subscription data. The collection of information was partitioned into two distinct sets: the training set and the test set. The model is checked for accuracy by using a cross-validation method. The dataset is also checked for accuracy with the utilisation of a log loss model. The dataset comprised 208 distinct features. The characteristics that enhance the accuracy of the model were retained, while those that were undesirable were discarded. The model was implemented by using the XGBoost library in conjunction with Python programming language, yielding an accuracy of 79.7% with the provided information.

Dingli et al. (2017) examined customer attrition in the retail sector by employing a model based on deep learning techniques. The learning models are crafted by using a restricted Boltzmann machine (RBM) and a convolution neural network (CNN). The experiment is conducted by using the POS value-based data sets. The dataset has undergone the process of Extract, Transform, and Load (ETL). Anomalies were removed from the dataset prior to partitioning it into training and testing subsets. Following the successful evacuation of exceptions, the informational collection underwent a partitioning process that was carried out in an arbitrary manner, resulting in a distribution ratio of 75:25. 75% of the data is allocated to the training set, while the remaining 25% is designated for the test set. The input to the RBM and CNN consists of the informational dataset. The aim is to ascertain whether the level of accuracy can be attributed to the presence of historical data. A total of 30 iterations were conducted on the training set by using the sigmoid activation function, and an accuracy rate of 74% was obtained. Using the RBM algorithm, a level of accuracy of 83% is achieved.

Zoric (2016) employed NNs in the Alyuda NeuroIntelligence software package for forecasting client churn in the banking industry. The findings indicate a positive correlation between client loyalty and a greater number of bank services, thereby suggesting that the bank ought to prioritise its attention towards those clients that currently engage with fewer than three products. By tailoring product offerings to cater to the needs of these clients, the bank can enhance its loyalty and foster a more robust CRM. Charandabi (2023) evaluated the effectiveness of six supervised classification methods to identify a suitable model for predicting customer turnover in the banking sector by using a dataset of 10 demographic and personal traits collected from a sample of 10,000 customers from several European banks. The results show that the ANN configuration consisting of a single hidden layer with five nodes was the most effective classifier since it did not exhibit any significant concerns related to overfitting.

Finally, Lemos et al. (2022) investigated customer churn prediction in the banking sector by using a customer-level dataset obtained from a prominent Brazilian bank. A horserace was conducted to compare multiple supervised machine learning algorithms using the same cross-validation and evaluation setup, ensuring a fair comparison. The RF technique showed superior performance compared to DT, k-NN, elastic net, LR,

and SVM models across various metrics. The findings indicate that customers who have a stronger relationship with the bank use more services, borrow more from the bank, and have a lower likelihood of closing their accounts. The model had the capacity to predict potential losses amounting to approximately 10% of the operating result reported by the largest Brazilian banks in 2019. This suggests that the model holds considerable economic significance. The results support the allocation of resources towards cross-selling and upselling strategies for the existing customers.

2.7. Gap in the Literature

Numerous studies have employed the CRISP-DM methodology (Belém, 2018; Hegde and Mundada, 2019; Karvana et al., 2019; Kumar and Ravi, 2008; Troncoso, 2018). Moreover, Kumar and Ravi (2008) have used the KNIME Analytics tool. The present study observed a dearth of citations pertaining to the prediction of customer churn in the banking sector using the CRISP-DM and KNIME Analytics together, except for the work of Kumar and Ravi (2008). The current study contributes to the literature by applying the most widely accepted data mining framework, i.e., CRISP-DM and the latest version of KNIME Analytics (KNIME 5.2.1), to predict customer churn for ABC Multinational Bank. We use open-access data that we retrieve from Kaggle to build, train, and test our predictive models.

CHAPTER III

DATA AND METHODOLOGY

This chapter describes the data and provides the methodology used in this study to predict customer churn in the banking sector. It also offers in-depth information about the tool for building, training, and testing the predictive models. In this context, we give detailed descriptions of the CRISP-DM methodology and KNIME Analytics tool. We also discuss the accuracy measurement metrics.

3.1. Data

3.1.1. Data Sample

In this study, we use the data of ABC Multinational Bank[2] (Bank ABC), which has been in business for over 40 years in the MENA region. It has a long history of cultivating long-term partnerships with its clients. The bank employs around 4,000 people. As a global bank, it serves 25 markets through its presence in 15 nations and financial centres, including New York, London, Singapore, Sao Paulo, Dubai International Financial Centre, Cairo, and Amman. Figure 3.1 and Figure 3.2 depict the global outlook and the locations of the Bank ABC branches.

We extracted the data from Kaggle. It is one of the world's most prominent online platforms that fosters collaboration among data scientists and machine learning practitioners. It operates under the Google LLC. It provides users with the ability to discover and share datasets, engage in model development within a web-based data platform, and collaborate with fellow machine learning experts. The platform has over 261,000 publicly available datasets, 887,000 notebooks and 1,900 models[3], making it one of the biggest data science platforms in the world. The data contains 120,000 data

[2] https://rpubs.com/Rvge_mvsrter/939193
[3]https://www.kaggle.com/

points (12 columns and 10,000 rows) and covers 31st July 2022 to 29th August 2022. The data was obtained from Spain, Germany, and France.

Figure 3.1. An Overview of the Bank ABC

Source: Bank ABC

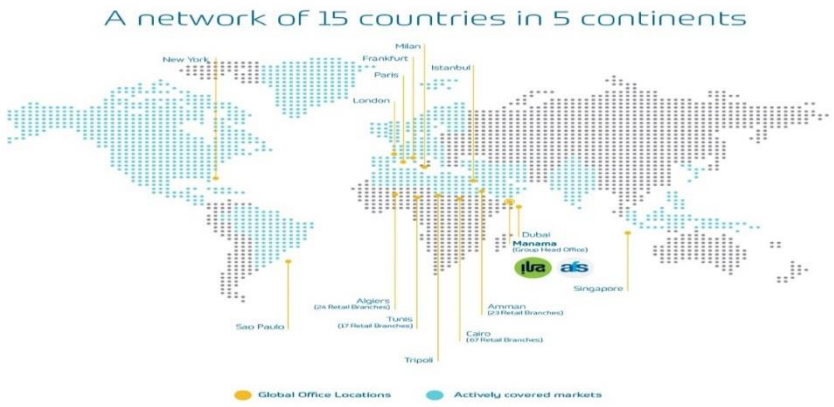


Figure 3.2. The Locations of the Branches of Bank ABC Across the World

Source: Bank ABC

3.1.2. Definition of the Variables

To measure the customer churn, we used the following variables in this study:

(i) *Customer ID:* The special Identity number assigned to each customer.

(ii) *Credit_score:* Numerical representation of a customer's creditworthiness.

(iii) *Country:* The nation from which the customer banks

(iv) *Gender:* The customer's gender

(v) Age: The customer's age

(vi) *Tenure:* How long has the customer been with the bank?

(vii) *Balance:* The amount of money the customer has in his or her account at the time of data collection.

(viii) *Products_number:* The number of products purchased or subscribed to by the customer.

(ix) *Credit_card:* It indicates whether the customer has a credit card. YES indicates that he or she has a credit card, while NO indicates that he or she does not have one.

(x) *Active_member:* It indicates whether or not the client is an active member of the bank.

(xi) *Estimated_salary:* It shows the estimation of the bank for the income of the customer.

(xii) *Churn:* It denotes whether the consumer terminated its relationship with the bank.

Of the 12 variables, only customer ID will be eliminated during the data preparation stage since it is merely a statistical property and has no effect on the churn prediction. The final variable (Churn) is our target variable. YES indicates that the customer churned at some point, while NO indicates that the customer never churned.

3.2. Methodology

3.2.1. CRISP-DM

The process of DM requires many skills. According to Wirth and Hipp (2000), the implementation of DM requires the utilisation of a standardised strategy that serves

the purpose of converting business challenges into DM tasks, recommending suitable DM techniques, and facilitating the assessment of the findings.

The CRISP-DM was introduced during the mid-1990s to establish a universally applicable and non-proprietary standard methodology for DM (Azevedo & Santos, 2008; Delen et al., 2018). As presented in Figure 3.3., the CRISP-DM methodology consists of six sequential steps. It starts with an understanding of the business and necessity of the DM project, also known as the application domain. The process concludes with the implementation of a solution that fulfils business requirements. Despite the sequential pattern of the processes, it is common for a significant amount of backtracking to occur. Due to its reliance on experience and experimentation, DM is characterised by an iterative and time-consuming process. The extent of iteration is contingent upon the problem, necessitating repeated navigation through the various phases. Given the interdependence of subsequent stages on preceding ones, it is imperative to exercise heightened vigilance towards the initial steps to prevent the entire study from being led astray at the outset.

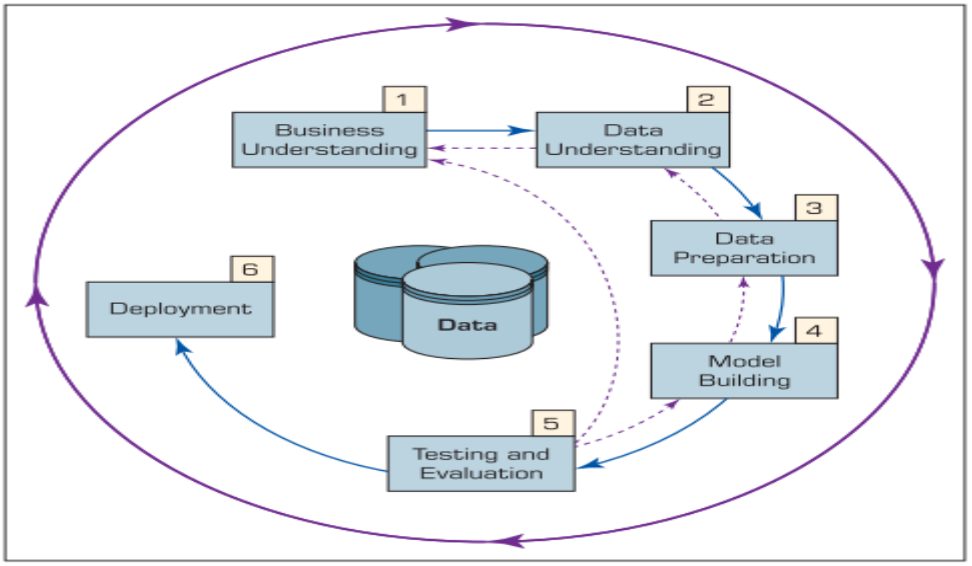


Figure 3.3. CRISP-DM

Source: Delen et al. (2018) and Sharda et al. (2020)

3.2.2. The Six Stages of the CRISP-DM Methodology

3.2.2.1. Business Understanding

The first step is to understand the business goals and use this information to formulate a strategy for achieving these goals (Azevedo & Santos, 2008). At this stage, there is a need for clearly defined objectives that address inquiries such as "What are the prevailing attributes exhibited by customers that switched to our competitors?" or "What are the profiles of the customers, and what is the magnitude of value contributed by each of them?" Subsequently, a project plan is set up to facilitate the acquisition of knowledge, delineating the individuals accountable for data collection, data analysis, and dissemination of results. During this phase, it is imperative to develop a budget that provides financial support for the study.

3.2.2.2. Data Understanding

The data understanding phase starts by collecting the initial data and subsequently engaging in activities to become acquainted with the data, identifying data quality, gaining preliminary insights, and detecting noteworthy subsets that may lead to hypotheses regarding concealed information (Azevedo & Santos, 2008; Wirth & Hipp, 2000). The relationship between business understanding and data understanding is closely intertwined. To enhance understanding of the data, analysts frequently employ a range of statistical methodologies such as calculating average, median, and standard deviation for numeric variables and generating mode and frequency tables for categorical variables. Moreover, correlation analysis, scatterplots, histograms, and box plots are used. The meticulous identification and selection of data sources, along with the most pertinent factors, can facilitate the expeditious discovery of valuable knowledge patterns by data mining algorithms (Delen et al., 2018).

3.2.2.3. Data Preparation

The data preparation step encompasses all tasks involved in constructing the ultimate dataset from the original data. In comparison to the other stages, data pre-processing is the most time-consuming and resource-intensive step. It encompasses 80% of the

time invested in DM (Delen et al., 2018; Sharda et al., 2020). The rationale behind dedicating a significant amount of effort to this stage is due to the inherent nature of the real-world data. This data typically suffers from incompleteness, where attribute values are missing or certain attributes of interest are absent. The real-world data is also noisy, containing errors or outliers that can distort the accuracy of the information. Furthermore, inconsistencies in codes or names can be found in the real-world data, further complicating the analysis process. It is likely that the data preparation task will be repeated. Selecting tables and records, cleaning data, building new attributes, and transforming data for modelling tools are all part of the activities at this stage.

3.2.2.4. Model Building

During this phase, a range of modelling methodologies are implemented. The process of model-building includes evaluating and comparing several models. Due to the absence of a universally acknowledged optimal algorithm for DM, it is advisable to employ a diverse range of viable models, accompanied by a well-defined testing and assessment plan, to choose the most suitable approach for a certain objective. The proposed models in this study include DT, RF, naïve Bayes, and ANN.

3.2.2.5. Testing and Evaluation

Before deciding to move forward with the deployment of the model, it is imperative to conduct an evaluation of the model and examine the actions undertaken in its construction to ensure that it effectively fulfils business goals. The primary aim is to ascertain whether a significant business concern exists that is not adequately addressed. Upon the conclusion of this phase, it is imperative to arrive at a definitive determination regarding the utilisation of the outcomes obtained from the DM process.

3.2.2.6. Deployment

The completion of the model creation does not signify the conclusion of the project. The acquired knowledge must be structured in a manner that is applicable to the consumer. The complexity of the deployment phase might vary depending on the specific requirements, ranging from a straightforward generation of a report to the

implementation of a sophisticated and replicable DM process. Maintenance actions for the deployed models may also be included in the deployment process. Due to the dynamic nature of business operations, the data pertaining to business activities undergo continuous fluctuations. Over time, the models may eventually lose their relevance, become obsolete, or lead to false outcomes. Hence, the upkeep of the models assumes significance for the outcomes of DM to be integrated into the company's regular operations. Thoroughly planning a maintenance strategy is essential to prevent prolonged periods of incorrect utilisation of DM outcomes.

An important point to note is that, in many instances, the responsibility of executing the deployment procedures lies with the customer rather than the data analyst. Thus, it is crucial for the customer to possess a clear understanding of the necessary actions required to effectively use the generated models.

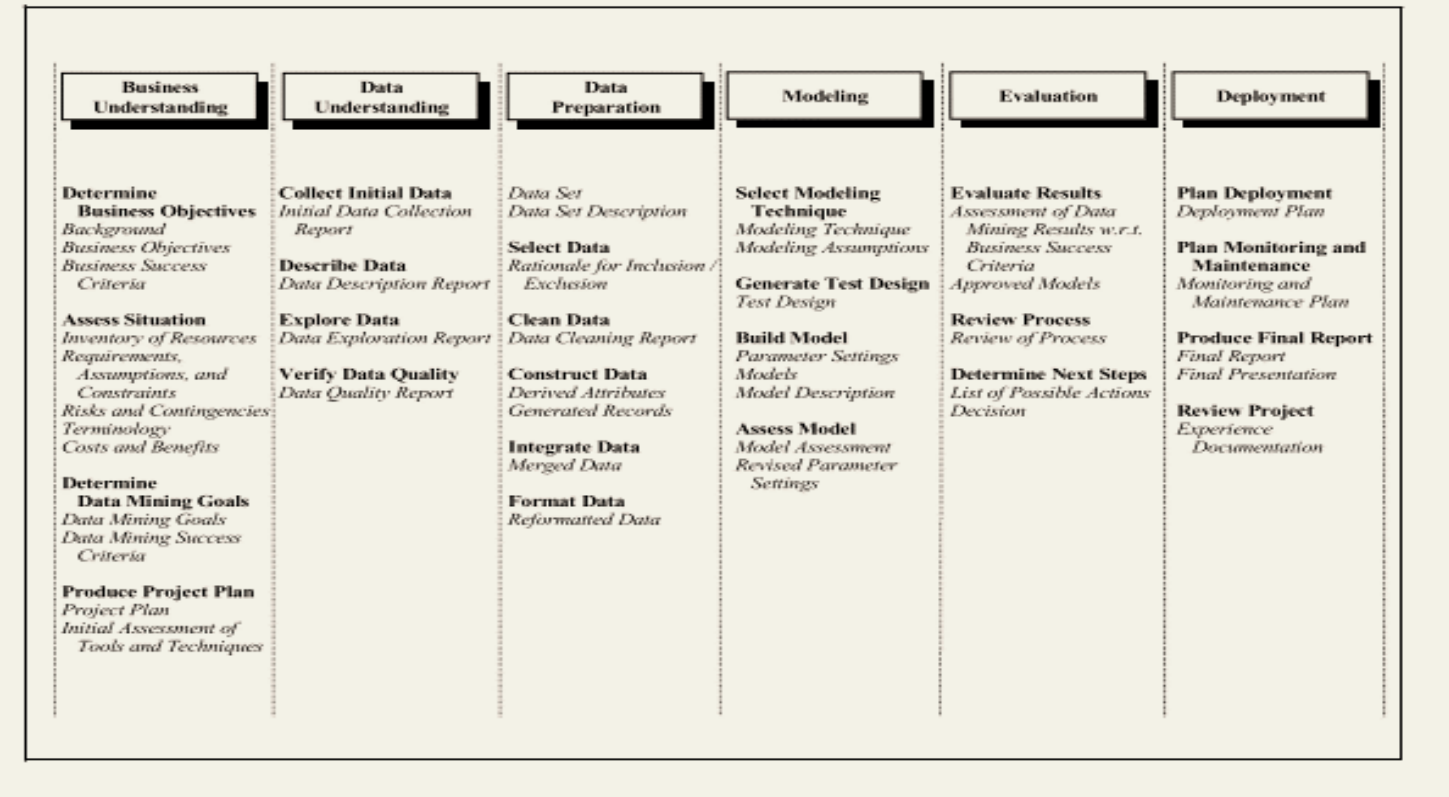


Figure 3.4. CRISP-DM Task Summary and Results

Source: Wirth & Hipp (2000)

3.2.3. KNIME Analytics

In recent years, there has been a proliferation of proficient DM tools that are both open-source and freely available. Notably, Konstanz Information Miner (KNIME) and RapidMiner have emerged as frontrunners in this domain (Dwivedi et al., 2016). KNIME is a data analytics, reporting, and integration tool that incorporates a range of components designed for machine learning and data mining, using a modular data pipelining approach known as the "Building Blocks of Analytics". The integration of a graphical user interface (GUI) and the utilisation of Java database connectivity (JDBC) facilitate the amalgamation of nodes that combine many data sources. This includes the preparation stages of extraction, transformation, and loading (ETL), enabling the creation of models, data analysis, and visualisation tasks.

3.2.3.1. Internal Capabilities of KNIME

The KNIME platform enables users to construct data flows (pipelines) through a visual interface. Users have the ability to selectively execute individual or multiple analytic processes in these pipelines. Furthermore, users may examine the outcomes and models through interactive widgets and views. KNIME software is implemented using the Java programming language, and it is built upon the Eclipse framework. The system uses an extension mechanism to incorporate plugins that offer supplementary functionality. The core version of the software encompasses a wide range of modules that facilitate data integration. These modules include functionalities such as file input/output and database nodes that support various database management systems (e.g., SQLite, MS-Access, SQL Server, MySQL, Oracle, PostgreSQL, Vertica, and H2) through JDBC or native connectors. Moreover, the core version offers modules for data transformation, including filtering, converting, splitting, combining, and joining. It provides commonly utilised statistical methods, data mining techniques, analytical tools, and text analytics capabilities. The utilisation of the free report designer plugin facilitates the incorporation of visualisation techniques. The utilisation of KNIME workflows as data sets enables the creation of report templates that can be exported to various document formats, including doc, ppt, xls, pdf, and other compatible formats.

3.2.3.2. The KNIME Workbench

Figure 3.5 presents the KNIME graphical user interface. In the following part, we provide the primary features of the interface and how they work.

(i) KNIME explorer

It provides an overview of the workflows that are available in the open KNIME workspaces, including local workspace, KNIME servers, and your own personal KNIME hub space.

(ii) Workflow coach

It displays suggested nodes from the KNIME community's shared workflows. If you disable KNIME's ability to track your usage patterns, the feature will be deactivated.

(iii) Node repository

In this feature, you can see a complete catalogue of all nodes that are both part of the base KNIME Analytics Platform and any add-ons you may have installed. A search bar sits atop the node repository, but browsing the categories is also an option.

(iv) Workflow editor

This is a canvas for editing the workflow that is presently active.

(v) Description

This displays detailed information about a chosen node (in the workflow editor or node repository) or the currently ongoing workflow.

(vi) Outline

This displays an overview of the workflow that is currently being used.

(vii) Console

It displays execution messages that show what is happening behind the scenes.

(viii) KNIME HUB[4]

Users can cooperate and commercialise analytical solutions with the KNIME Analytics Platform through KNIME Hub. KNIME Hub comes in two forms: KNIME Business Hub, which is put into your private infrastructure, and KNIME Community Hub, which is accessible to the entire world.

[4] https://www.knime.com/knime-hub

3.2.4. Accuracy Measurement Metrics

The confusion matrix is widely regarded as the fundamental method for estimating accuracy in classification problems (Delen et al., 2018). The confusion matrix in Figure 3.6 represents the outcomes of a two-class classification task. The values on the diagonal running from the top left to the bottom right of the matrix indicate accurate decisions, whereas the values outside this diagonal show instances of error.

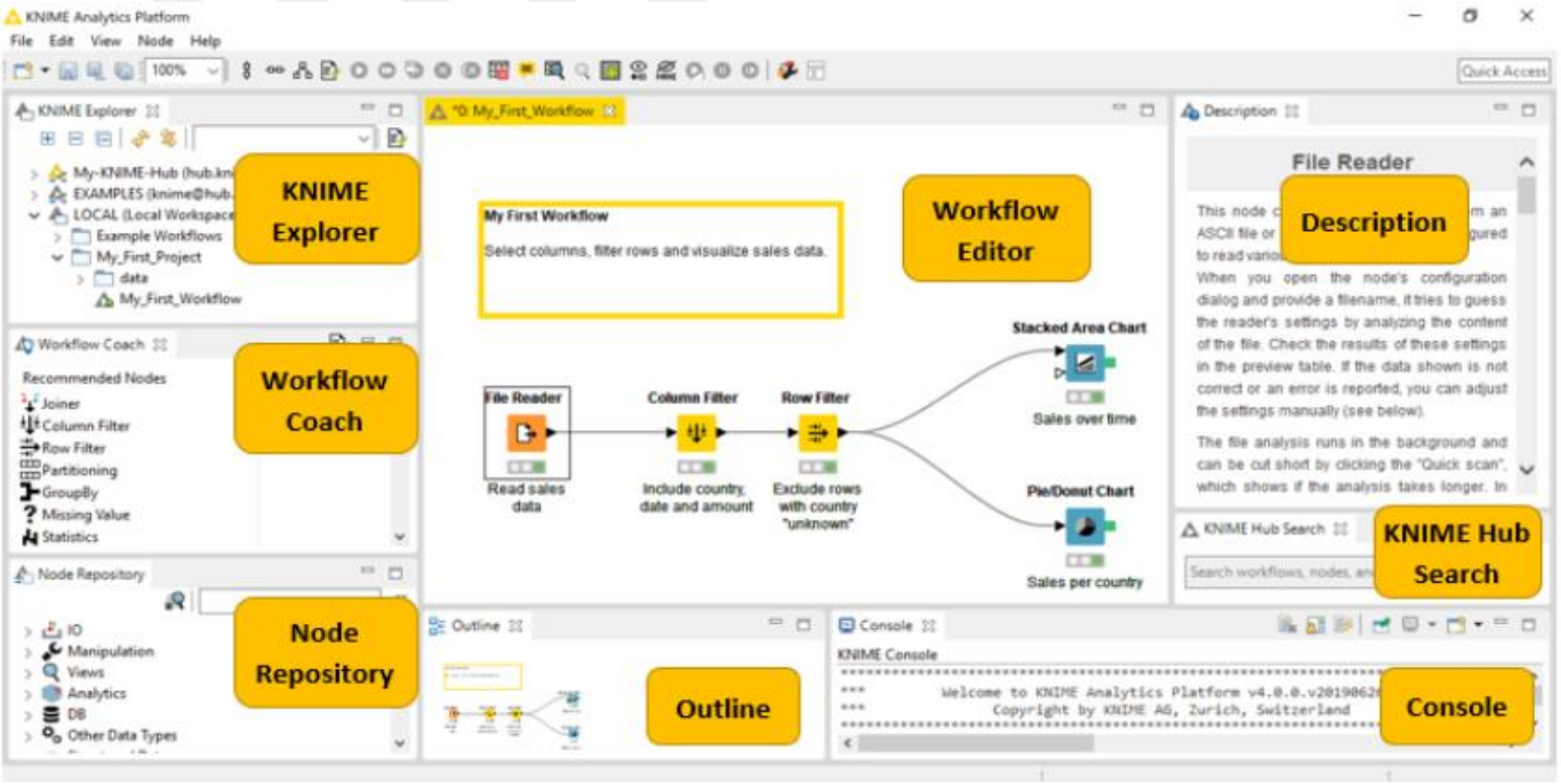


Figure 3.5. KNIME User Interface[5]

Source: Delen et al. (2018)

[5] https://docs.knime.com/2020-07/analytics_platform_quickstart_guide/index.html#start-knime-analytics-platform

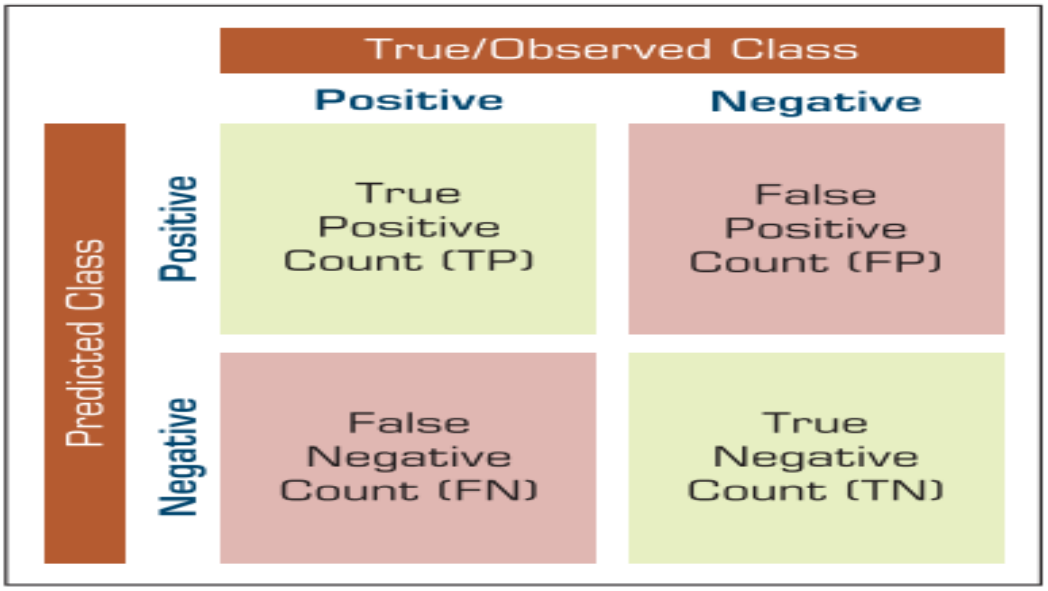


Figure 3.6. A Basic Confusion Matrix for Two-Class Classification Results Tabulation

Source: Delen et al. (2018)

3.2.4.1. Predictive Accuracy

This measures how well the model can assign a label to unknown data. For classification models, prediction accuracy is employed as a metric of quality. This metric is calculated by comparing the model's predicted class labels to the actual class labels from a test set. Accuracy can be measured in terms of the rate at which a model properly classifies examples from a test data set. It can be measured by calculating the ratio of correctly categorised cases, both positive and negative, divided by the total number of cases as shown in the following formula:

$$\textbf{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

where TP = True positive; TN= True negative; FP=False positive and FN=False negative

3.2.4.2. True Positive and Negative Rate (a.k.a Sensitivity)

The true positive rate, also known as the hit rate or recall, is defined as the ratio of all the positives which were rightly identified as such to the total number of positives. It measures the likelihood that an actual positive will test positive.

True Positive = $\frac{TP}{TP+FN}$

The ratio of all the negatives which were rightly identified as such to the total number of negatives. This is also known as the false alarm rate. It is calculated as follows:

True Negative = $\frac{TN}{TN+FP}$

This is the likelihood that a true negative will test negative.

3.2.4.3. Precision and Recall

In mathematical terms, precision is usually defined as the ratio of true positives to the sum of true positives and false positives.

Precision = $\frac{TP}{TP+FP}$

Recall is defined mathematically as the ratio of true positives to the sum of true positives and false negatives.

Recall = $\frac{TP}{TP+FN}$

While all the accuracy measures are essential, this study will focus on the overall accuracy of the models as well as their sensitivity and specificity as a measure of the performance of the models. However, the others will be displayed in the outputs of the KNIME Analysis.

3.2.4.4. 10-Fold Cross-Validation

This study will use a technique called 10-fold cross-validation to compare the predicted performance of our models while minimising the bias introduced by the random sampling of the training and holdout data samples. With this technique, the data set is randomly split into 10 independent subsets of similar size using 10-fold cross-validation. There are ten iterations of training and testing the categorisation model. Every time, we train on all but one fold and then test on that last fold. Figure 3.7. is a visual representation of a 10-fold cross-validation.

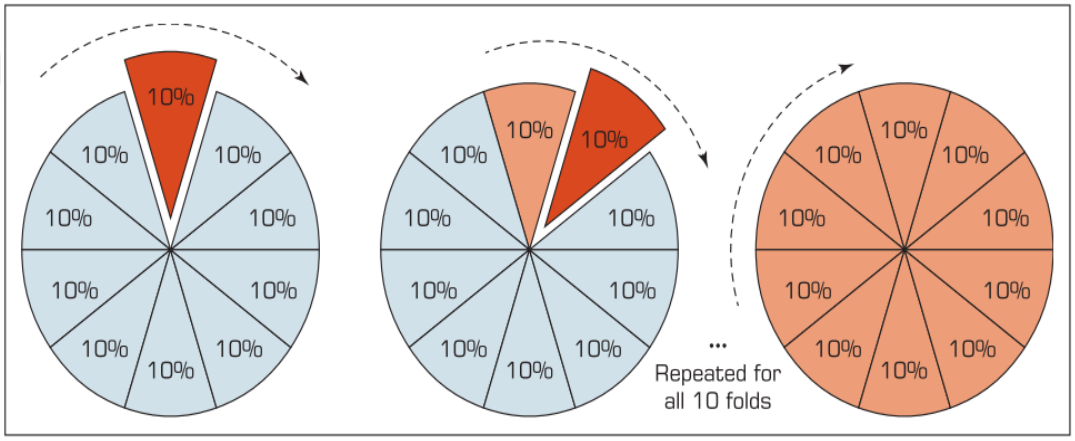


Figure 3.7. A Visual Representation of a 10-Fold Cross-Validation

3.2.5. The Receiver Operating Characteristic (ROC) Curve

With the ROC curve, the true positive rate is plotted on the y-axis and the false positive rate on the x-axis in a graphical evaluation method known as the Area Under the ROC Curve (AUC). A classifier's efficacy is quantified by its area under the receiver operating characteristic (ROC) curve. A score of 1 would represent a flawless classifier, while a score of 0.5 would imply a performance no better than chance. In practice, the scores would fall somewhere in between these two extremes.

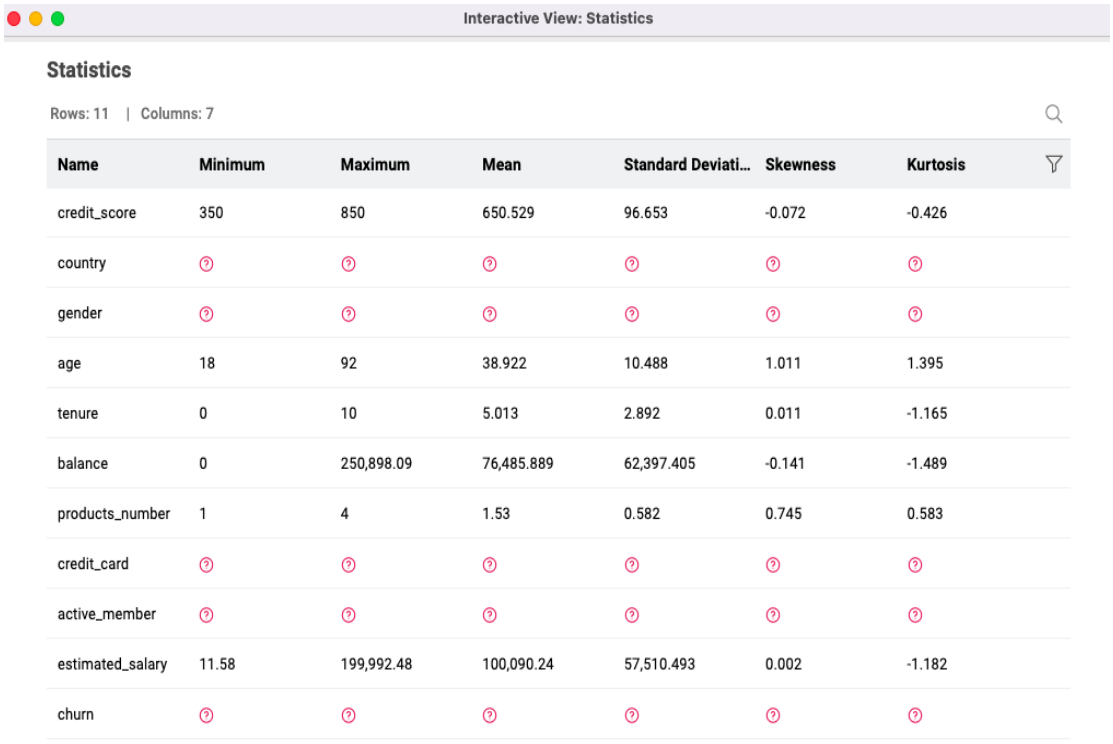
CHAPTER IV

EMPIRICAL FINDINGS

4.1. Business and Data Understanding

In the business understanding stage, we used the data of Bank ABC to build, train, and test models to predict customers who are likely to churn from the bank. Then, in the understanding stage, we employed KNIME statistics to view the important properties of data. Table 4.1. shows the summary of the descriptive statistics. The red question marks indicate missing values. These values are non-numerical, i.e., they are categorical variables. For instance, the churn column has two categories, YES and NO. The former represents those that churned, while the latter represents those that did not.

Table 4.1. Descriptive Statistics


Interactive View: Statistics

Statistics

Rows: 11 | Columns: 7

Name	Minimum	Maximum	Mean	Standard Deviati...	Skewness	Kurtosis
credit_score	350	850	650.529	96.653	-0.072	-0.426
country	?	?	?	?	?	?
gender	?	?	?	?	?	?
age	18	92	38.922	10.488	1.011	1.395
tenure	0	10	5.013	2.892	0.011	-1.165
balance	0	250,898.09	76,485.889	62,397.405	-0.141	-1.489
products_number	1	4	1.53	0.582	0.745	0.583
credit_card	?	?	?	?	?	?
active_member	?	?	?	?	?	?
estimated_salary	11.58	199,992.48	100,090.24	57,510.493	0.002	-1.182
churn	?	?	?	?	?	?

4.2. Data Preparation

In the data preparation stage, we pre-process the data by cleaning and transforming it to make it ready for building the predictive models. This stage is the most time-consuming one. We used the Column filter to remove the customer ID since each ID is unique and IDs do not have any generalisable information. The removal of the Customer ID reduced the number of columns to 11, while the number of rows remained 10,000. The data that we used had no missing values; thus, after using the column filter, the next node is the colour manager. We used the colour manager to label all those that churned as *red* and all those that did not churn as *green*. Table 4.2 shows the results of this step. Matching the first column with the last column shows that all the YES in the churn column correspond to Red in the Row ID column, while all the NO in the churn column correspond to Green in the first column.

Table 4.2. Churned Customers and Non-Churned Customers

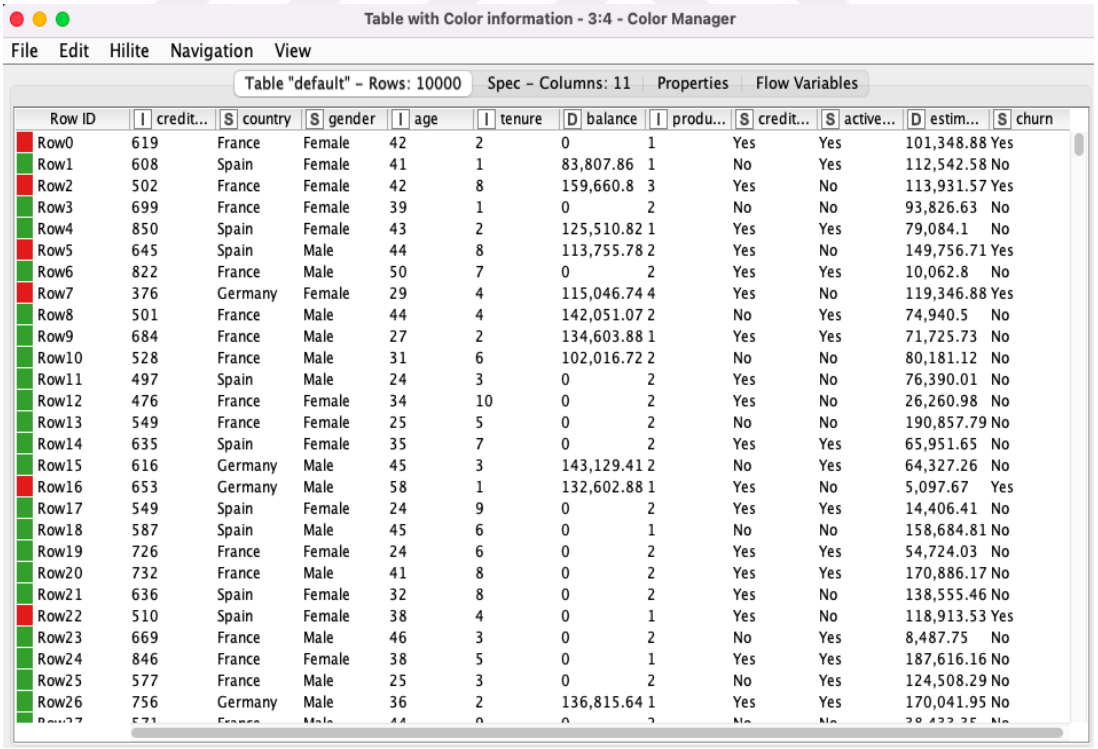


Table with Color information - 3:4 - Color Manager

File Edit Hilite Navigation View

Table "default" – Rows: 10000 | Spec – Columns: 11 | Properties | Flow Variables

Row ID	credit...	country	gender	age	tenure	balance	produ...	credit...	active...	estim...	churn
Row0	619	France	Female	42	2	0	1	Yes	Yes	101,348.88	Yes
Row1	608	Spain	Female	41	1	83,807.86	1	No	Yes	112,542.58	No
Row2	502	France	Female	42	8	159,660.8	3	Yes	No	113,931.57	Yes
Row3	699	France	Female	39	1	0	2	No	No	93,826.63	No
Row4	850	Spain	Female	43	2	125,510.82	1	Yes	Yes	79,084.1	No
Row5	645	Spain	Male	44	8	113,755.78	2	Yes	No	149,756.71	Yes
Row6	822	France	Male	50	7	0	2	Yes	Yes	10,062.8	No
Row7	376	Germany	Female	29	4	115,046.74	4	Yes	No	119,346.88	Yes
Row8	501	France	Male	44	4	142,051.07	2	No	Yes	74,940.5	No
Row9	684	France	Male	27	2	134,603.88	1	Yes	Yes	71,725.73	No
Row10	528	France	Male	31	6	102,016.72	2	No	No	80,181.12	No
Row11	497	Spain	Male	24	3	0	2	Yes	No	76,390.01	No
Row12	476	France	Female	34	10	0	2	Yes	No	26,260.98	No
Row13	549	France	Female	25	5	0	2	No	No	190,857.79	No
Row14	635	Spain	Female	35	7	0	2	Yes	Yes	65,951.65	No
Row15	616	Germany	Male	45	3	143,129.41	2	No	Yes	64,327.26	No
Row16	653	Germany	Male	58	1	132,602.88	1	Yes	No	5,097.67	Yes
Row17	549	Spain	Female	24	9	0	2	Yes	Yes	14,406.41	No
Row18	587	Spain	Male	45	6	0	1	No	No	158,684.81	No
Row19	726	France	Female	24	6	0	2	Yes	Yes	54,724.03	No
Row20	732	France	Male	41	8	0	2	Yes	Yes	170,886.17	No
Row21	636	Spain	Female	32	8	0	2	Yes	No	138,555.46	No
Row22	510	Spain	Female	38	4	0	1	Yes	No	118,913.53	Yes
Row23	669	France	Male	46	3	0	2	No	Yes	8,487.75	No
Row24	846	France	Female	38	5	0	1	Yes	Yes	187,616.16	No
Row25	577	France	Male	25	3	0	2	No	Yes	124,508.29	No
Row26	756	Germany	Male	36	2	136,815.64	1	Yes	Yes	170,041.95	No

After using the colour manager, we used the x-partitioner to split the data into 80% for training and 20% for testing using the stratified sampling method on churn. We chose the x-partitioner for this task because it produces better results than the single-split

partitioner. The x-partitioner node helps divide the dataset into ten equal-sized samples, and during each iteration of the 10-fold cross-validation process, we used nine partitions to train the model while the remaining one partition was used to test it. This process takes place ten times until each partition of the data serves as a test set at least once. This process is advantageous as it facilitates the evaluation of a model's ability to generalise new data and prevents complications such as overfitting during the model deployment. Finally, we applied the equal-sized sampling node to the training dataset to avoid the dominance of the majority class affecting the predictive accuracy of models.

4.3. Model Building

In this stage, we built the models that we use for the analytics. In this study, we used DT, RF, LR, SVM and ANN in the customer churn prediction model development. Then, we combined these models into an ensemble model that is expected to be a higher-performance predictive model (Azevedo & Santos, 2008; Delen et al., 2018; Sharda et al., 2020; Vafeiadis et al., 2015). We carried out this activity with the help of the Column Appender node.

4.3.1. The Performance Measurement of the Predictive Models

In our target column (Churn column), YES means the customer churned, and NO means the customer did not churn. Thus, YES represents the positive category as far as the churn is concerned, while NO represents the negative category. This distinction is useful in interpreting the results. It helps understand the confusion matrix. Deciding on which class to call positive and which class to call negative can be done arbitrarily (Carter et al., 2016). That means it is left to the researcher's discretion[6]. Hence, we decided to refer to those who churned as the positive class and vice versa, and we have been consistent throughout the work.

[6] https://www.knime.com/blog/from-modeling-to-scoring-confusion-matrix-and-class-statistics

4.3.2. The Confusion Matrix and Accuracy Statistics

The scoring process of a predictive model consists of a match count. That is, how many data rows have been correctly classified and how many data rows have been incorrectly classified by the model? These counts are summarised in the confusion matrix. This matrix shows the numbers of correct and incorrect predictions. In the matrix, the rows are the Actual values, and the Columns are the predictions of the actual values.

Initially, we implemented the confusion matrix to assess the performance of binomial classification (Delen et al., 2018). The aim of this step is to designate one of the two classes, namely the positive class (those who churned in our case), as the class of interest. It is necessary to arbitrarily[7] select one value as the positive class in the target column (the churn column in our case). The alternative value (those who did not churn in our case) is subsequently classified as the negative class by default. It should be noted that while this assignment is arbitrary, certain class statistics may exhibit varying values contingent upon the positive class chosen[8]. In this study, we considered those who churned as Positive and those who did not as Negative.

In the customer churn prediction, we need to answer several different questions:

I. How many of the actual churned customers were predicted as churned? These are referred to as *True Positives (TP)***.**
II. How many of them were predicted as not churned even though they churned? They are known as *False Negatives (FN).*
III. Were some customers who did not churn predicted by the models as churned? These are referred to as *False Positives (FP).*
IV. How many of those who did not churn were predicted correctly as not churned? They are known as *True Negatives (TN).*

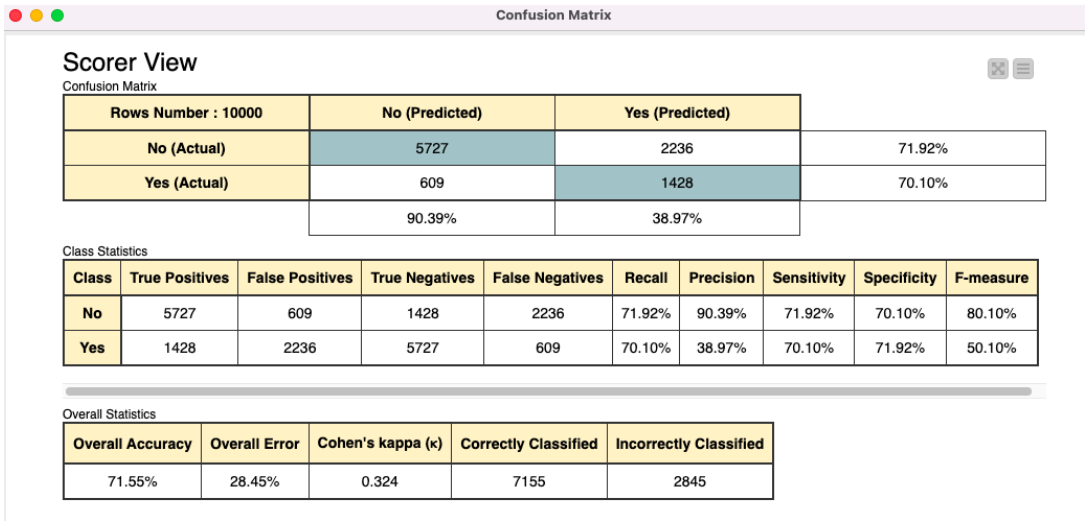
The answer to all these questions can be found in the confusion matrix.

[7] https://www.knime.com/blog/from-modeling-to-scoring-confusion-matrix-and-class-statistics
[8] https://www.knime.com/blog/from-modeling-to-scoring-confusion-matrix-and-class-statistics

4.3.2.1. Sensitivity, Specificity, and Overall Accuracy

To explain sensitivity, specificity, and overall accuracy, we need to display the output of one of the models that we will refer to. Table 4.3 shows the scorer (JavaScript) interactive view for the DT model. In Table 4.3, the numbers with the deep-coloured background were the correctly predicted outcomes, i.e., 5727 and 1428 were the correctly predicted negatives and positives, respectively. Likewise, the number 609 is the false negative, while the number 2236 is the false positive. Moreover, in Table 4.3., the second table represents the class prediction statistics, while the third table represents the overall prediction statistics. In the class prediction statistics, the first four columns are just a repetition of what is in the confusion matrix, while the second four columns show the accuracy measures, including sensitivity and specificity for both the positive class (YES) and the negative class (NO). We extracted the confusion matrix from Table 4.3 and showed how sensitivity, specificity and overall accuracy are computed in Table 4.4.

Table 4.3. The Scorer (JavaScript) Interactive View for the Decision Tree Model


Confusion Matrix

Scorer View

Confusion Matrix

Rows Number : 10000	No (Predicted)	Yes (Predicted)	
No (Actual)	5727	2236	71.92%
Yes (Actual)	609	1428	70.10%
	90.39%	38.97%	

Class Statistics

Class	True Positives	False Positives	True Negatives	False Negatives	Recall	Precision	Sensitivity	Specificity	F-measure
No	5727	609	1428	2236	71.92%	90.39%	71.92%	70.10%	80.10%
Yes	1428	2236	5727	609	70.10%	38.97%	70.10%	71.92%	50.10%

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
71.55%	28.45%	0.324	7155	2845

The sensitivity of a model indicates how well it identifies occurrences belonging to the positive category (YES category in the churn column). Therefore, sensitivity in customer churn prediction is the extent to which the predictive model can predict those who churn. A model that is 100% sensitive will detect all churned clients. Any model

is unlikely to be completely sensitive[9]. A model with 85% sensitivity will identify 85% of churned consumers but will miss 15%. A highly sensitive model can be useful for ruling out a customer who is projected not to churn. The reasoning is that it is highly accurate in forecasting who will churn. This means that if it claims this customer would not churn, we can potentially rule them out. Our DT model achieved the sensitivity value of 0.701, meaning that 70.1% of the customers in the dataset who churned were correctly predicted by the model as churned.

Table 4.4. Sensitivity, Specificity, and Prediction Accuracy for the Decision Tree Model

	Predicted Class Negative (NO)	Predicted Class Positive (YES)	
Actual Class Negative (NO)	**TN = 5727**	FP = 2236	**Specificity** = $\frac{TN}{FP+TN}$ $=\frac{5727}{2236+5727}$ = 71.92%
Actual Class positive (YES)	FN = 609	**TP = 1428**	**Sensitivity** = $\frac{TP}{FN+TP}$ $=\frac{1428}{609+1428}$ = 70.10%
			Overall Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$ $=\frac{1428+5727}{1428+5727+2236+609}$ = 71.55%

Specificity, on the other hand, is the extent to which a model can predict occurrences belonging to the negative category (the NO category in our case). The DT has a specificity of 0.7192, meaning that 71.92% of those that did not churn were correctly identified as not churned.

[9] https://towardsdatascience.com/how-do-you-measure-if-your-customer-churn-predictive-model-is-good-187a49a9eee3

Finally, the DT model had an overall prediction accuracy of 71.55%, meaning that 71.55% of the 10,000 cases were correctly predicted.

4.3.2.2. The AUC of the ROC Curve

The total capacity of a test to accurately differentiate positive from negative or normal from abnormal can be quantified using the AUC of an ROC curve. According to Carter et al. (2016), a flawless test will yield an AUC score of 1.0, indicating the absence of both false positives and false negatives. Conversely, a value of 0.5 shows that the test result is no more accurate than if it were determined randomly. The AUC cannot be less than 0.5. The interpretation of the AUC values is as follows: a value of 1.0 indicates a perfect test, a range of 0.9-0.99 shows an excellent test, a range of 0.8-0.89 indicates a good test, a range of 0.7-0.79 shows a fair test, a range of 0.51-0.69 indicates a poor test and a value of 0.5 indicates no value.

4.4. The Evaluation of the Performance of the Models

In this section, we analysed the confusion matrix and the ROC curves of each model, including the ensemble model.

4.4.1. Decision Tree Model

Table 4.3 shows that the DT model had a sensitivity of 70.1%, specificity of 71.9%, and overall accuracy of 71.6%, meaning that the DT's accuracy at predicting the churners (sensitivity) is 70.1% and non-churners (specificity) is 71.9%, and it has 71.6% accuracy at predicting. Figure 4.1 displays the ROC curve of the DT model presenting the AUC score, showing that the DT has an AUC score of 72.6%. This score falls within the range of 0.7 - 0.79 and, thus, indicates a fair test. Since the closer the AUC is to 1, the better the model is at distinguishing between classes, a 72.6% means the DT model is good at separating churners and non-churners. The DT has an accuracy of 72.6% according to the AUC of the ROC curve.

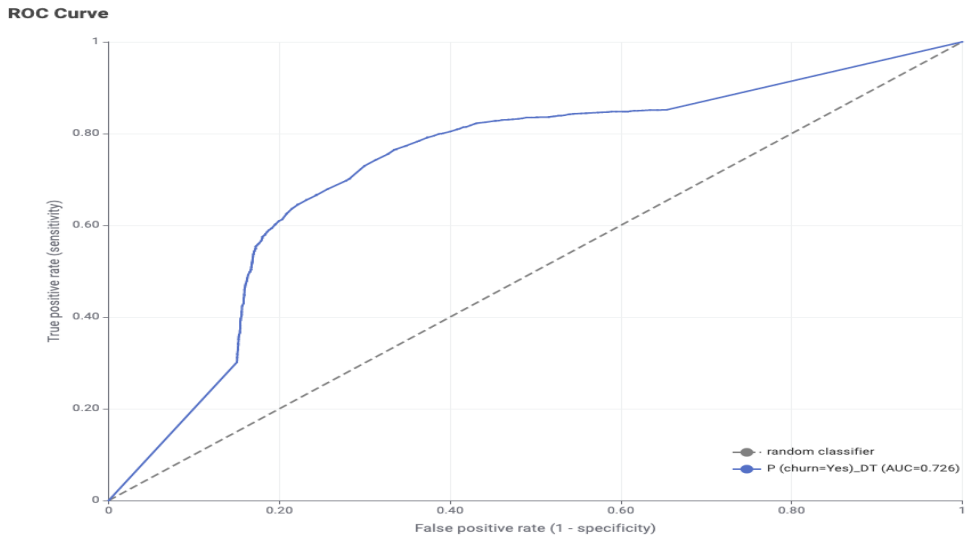


Figure 4.1. The ROC Curve of the Decision Tree Model

4.4.2. Artificial Neural Network (ANN) Model

Before building the NN, we used the one-to-many node to convert all non-numeric variables to numeric ones except the target variable, i.e., churn, since NNs, LR, SVM, and nearest neighbour algorithms use numeric derivations. We also normalised the data using the normaliser node in KNIME so that the value of every variable is between 0 and 1, and we replicated it for all subsequent models. Table 4.5 shows the scorer (JavaScript) interactive view for the ANN. Table 4.5 shows that the ANN had a sensitivity of 72.7%, specificity of 77.71%, and overall accuracy of 76.69%, meaning that the ANN's accuracy at predicting the churners (Sensitivity) is 72.7%, non-churners (specificity) is 77.71%, and it has a 76.69% overall accuracy.

Figure 4.2 presents the ROC curve of the ANN model, showing the AUC score of the model. Figure 4.2 shows that the ANN has an AUC score of 83.4%. This score falls within the range of 0.8-0.89 and, thus, indicates a good test. Since the closer the AUC is to 1, the better the model is at distinguishing between classes, an 83.4% AUC score means that the ANN is good at separating churners and non-churners. The ANN has an accuracy of 83.4% according to the AUC of the ROC curve.

Table 4.5. The Scorer (JavaScript) Interactive View for the ANN Model

Confusion Matrix

Scorer View

Confusion Matrix

Rows Number : 10000	No (Predicted)	Yes (Predicted)	
No (Actual)	6188	1775	77.71%
Yes (Actual)	556	1481	72.70%
	91.76%	45.49%	

Class Statistics

Class	True Positives	False Positives	True Negatives	False Negatives	Recall	Precision	Sensitivity	Specificity	F-measu
No	6188	556	1481	1775	77.71%	91.76%	77.71%	72.70%	84.15%
Yes	1481	1775	6188	556	72.70%	45.49%	72.70%	77.71%	55.96%

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
76.69%	23.31%	0.412	7669	2331

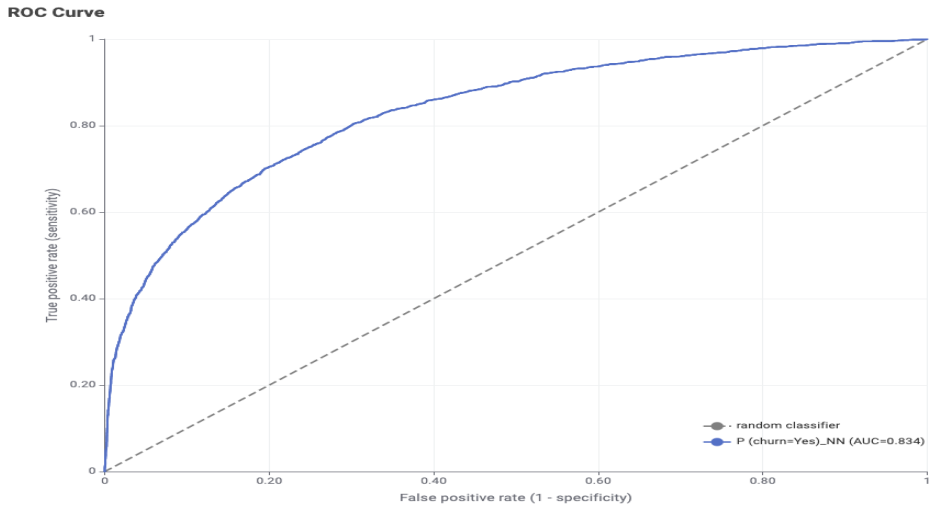


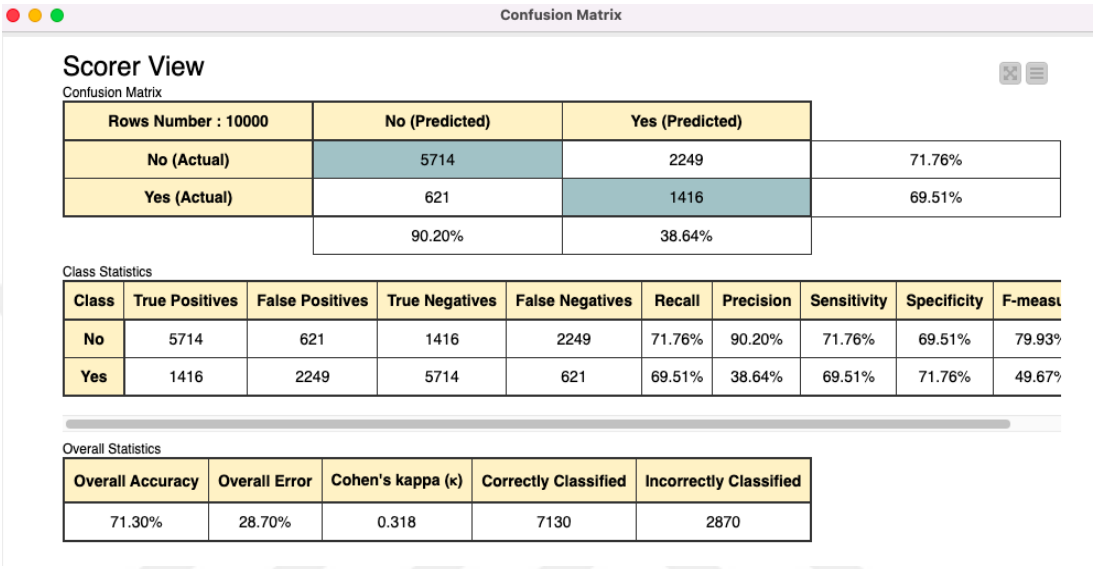
Figure 4.2. The ROC Curve of the ANN Model

4.4.3. Logistic Regression (LR) Model

Table 4.6. shows the scorer (JavaScript) interactive view for the LR model. Table 4.6 shows that the LR model had a sensitivity of 69.51% for the positive class, a specificity of 71.76%, and an overall accuracy of 71.3%, meaning that the LR's accuracy at

predicting the churners (sensitivity) is 69.51%, non-churners (specificity) is 71.76%, and it has a 71.3% overall accuracy in making the right predictions.

Table 4.6. The Scorer (JavaScript) Interactive View for the Logistic Regression Model


Confusion Matrix

Scorer View

Confusion Matrix

Rows Number : 10000	No (Predicted)	Yes (Predicted)	
No (Actual)	5714	2249	71.76%
Yes (Actual)	621	1416	69.51%
	90.20%	38.64%	

Class Statistics

Class	True Positives	False Positives	True Negatives	False Negatives	Recall	Precision	Sensitivity	Specificity	F-measu
No	5714	621	1416	2249	71.76%	90.20%	71.76%	69.51%	79.93%
Yes	1416	2249	5714	621	69.51%	38.64%	69.51%	71.76%	49.67%

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
71.30%	28.70%	0.318	7130	2870

Figure 4.3 is the ROC curve of the LR model, showing the AUC score of the model. In Figure 4.3, it is obvious that the LR has an AUC score of 76.9%. This score falls within the range of 0.7-0.79 and, thus, indicates a fair test. Since the closer the AUC is to 1, the better the model is at distinguishing between classes; a 76.9% AUC score means the LR is good at separating churners and non-churners. The LR has an accuracy of 76.9% according to the AUC of the ROC curve.

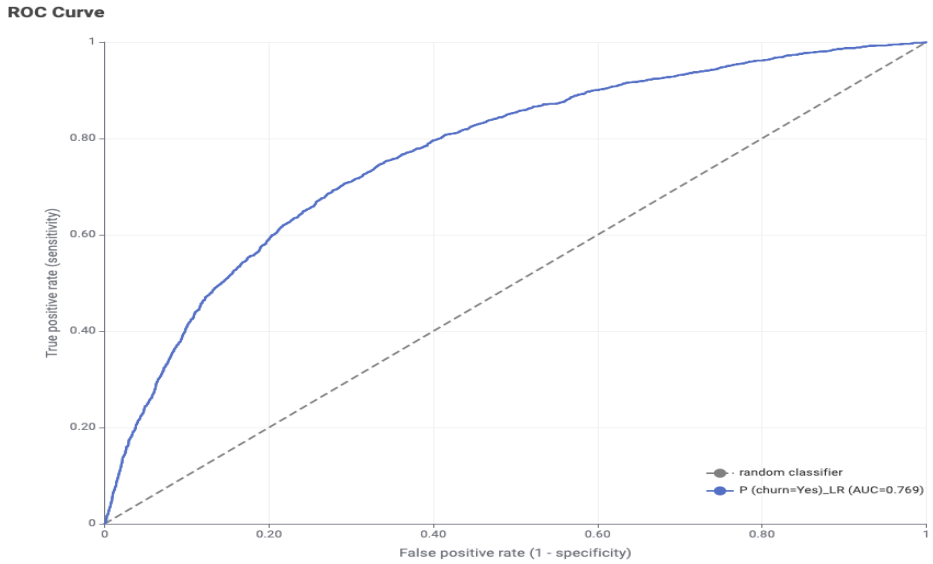


Figure 4.3. The ROC Curve of the Logistic Regression Model

4.4.4. Support Vector Machine (SVM) Model

Table 4.7 shows the scorer (JavaScript) interactive view for the SVM model. Table 4.7 shows that the SVM model had a sensitivity of 70.1% for the positive class, a specificity of 72.4%, and an overall accuracy of 71.93%, meaning that the accuracy of SVM at predicting the churners (sensitivity) is 70.1%, non-churners (specificity) is 72.4%, and it has a 71.93% overall accuracy in making the right predictions.

Table 4.7. The JavaScript Scorer Interactive View for the SVM Model

Confusion Matrix

Scorer View

Support Vector Machine (SVM)

Confusion Matrix

Rows Number : 10000	No (Predicted)	Yes (Predicted)	
No (Actual)	5765	2198	72.40%
Yes (Actual)	609	1428	70.10%
	90.45%	39.38%	

Class Statistics

Class	True Positives	False Positives	True Negatives	False Negatives	Recall	Precision	Sensitivity	Specificity	F-measure
No	5765	609	1428	2198	72.40%	90.45%	72.40%	70.10%	80.42%
Yes	1428	2198	5765	609	70.10%	39.38%	70.10%	72.40%	50.43%

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
71.93%	28.07%	0.329	7193	2807

Figure 4.4 is the ROC curve of the SVM model, showing the AUC score of the model. In Figure 4.4, it is obvious that the SVM has an AUC score of 77%. This score falls within the range of 0.7-0.79 and, thus, indicates a fair test. Since the closer the AUC is to 1, the better the model is at distinguishing between positive and negative classes; a 76.9% AUC score means that the SVM is good at separating churners and non-churners, and the SVM has an accuracy of 76.9% according to the AUC of the ROC curve.

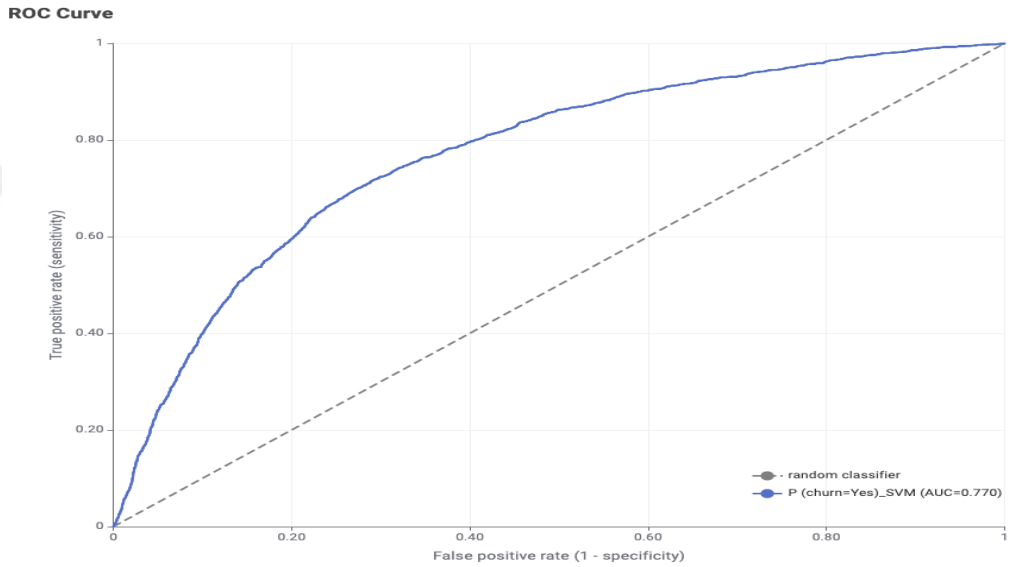


Figure 4.4. The ROC Curve of the SVM Model

4.4.5. Random Forest (RF) Model

Table 4.8 shows the scorer (JavaScript) interactive view for the RF model. Table 4.8 shows that the RF model had a sensitivity of 73.83% for the positive class, a specificity of 80.21%, and an overall accuracy of 78.91%, meaning that the RF's accuracy at predicting the churners (sensitivity) is 73.83%, non-churners (specificity) is 80.21%, and it has a 78.91% overall accuracy in making the right predictions. These scores make the RF model better than all the other models.

Table 4.8. The Scorer (JavaScript) Interactive View for the RF Model

Confusion Matrix

Scorer View

Random Forest (RF)

Confusion Matrix

Rows Number : 10000	No (Predicted)	Yes (Predicted)	
No (Actual)	6387	1576	80.21%
Yes (Actual)	533	1504	73.83%
	92.30%	48.83%	

Class Statistics

Class	True Positives	False Positives	True Negatives	False Negatives	Recall	Precision	Sensitivity	Specificity	F-measure
No	6387	533	1504	1576	80.21%	92.30%	80.21%	73.83%	85.83%
Yes	1504	1576	6387	533	73.83%	48.83%	73.83%	80.21%	58.78%

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
78.91%	21.09%	0.454	7891	2109

Figure 4.5 shows that the RF has an AUC score of 85.3%. This score falls within the range of 0.8-0.89 and, thus, indicates a good test. Since the closer the AUC is to 1, the better the model is at distinguishing between positive and negative classes; an 85.3% AUC score means the RF is good at separating churners and non-churners. The RF has an accuracy of 85.3% according to the AUC of the ROC curve. By the AUC metrics, the RF is the best-performing model according to the overall accuracy measure from the confusion matrix.

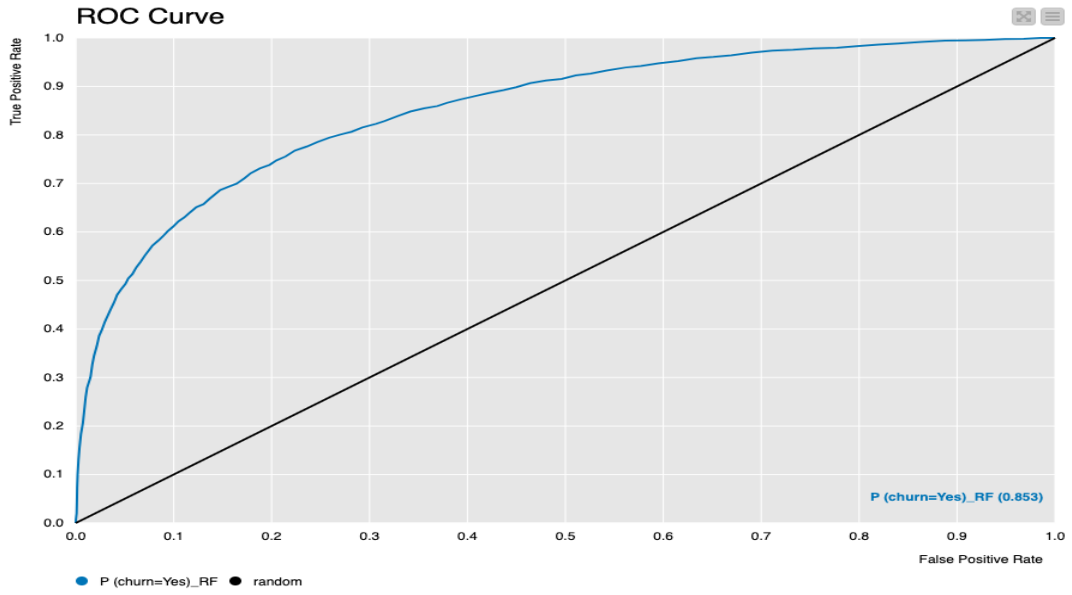
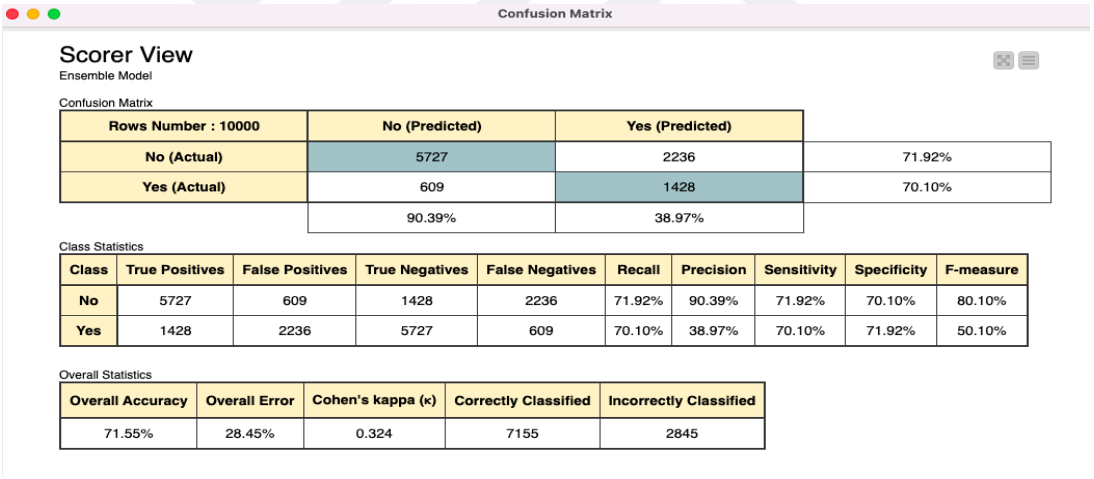


Figure 4.5. The ROC Curve of the RF Model

4.4.6. Ensemble Model

Table 4.9 shows the scorer (JavaScript) interactive view for the Ensemble model. Table 4.9 shows that the Ensemble model had a sensitivity of 70.1%, specificity of 71.9%, and overall accuracy of 71.6%, meaning that the accuracy of the model at predicting the churners (sensitivity) is 70.1%, non-churners (specificity) is 71.9%, and it has 71.6% accuracy at predicting. Figure 4.6 is the ROC curve of the Ensemble model showing the AUC scores of all the models in a single visual representation. Figure 4.6 indicates that the RF model with an AUC score of 85.3% is the best-performing model. This is followed by the ANN model, which has an AUC score of 83.4%. Finally, with a score of 72.6%, the least-performing model, according to the AUC scores, is the DT.

Table 4.9. The Scorer (JavaScript) Interactive View for the Ensemble Model



Rows Number : 10000	No (Predicted)	Yes (Predicted)	
No (Actual)	5727	2236	71.92%
Yes (Actual)	609	1428	70.10%
	90.39%	38.97%	

Class	True Positives	False Positives	True Negatives	False Negatives	Recall	Precision	Sensitivity	Specificity	F-measure
No	5727	609	1428	2236	71.92%	90.39%	71.92%	70.10%	80.10%
Yes	1428	2236	5727	609	70.10%	38.97%	70.10%	71.92%	50.10%

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
71.55%	28.45%	0.324	7155	2845

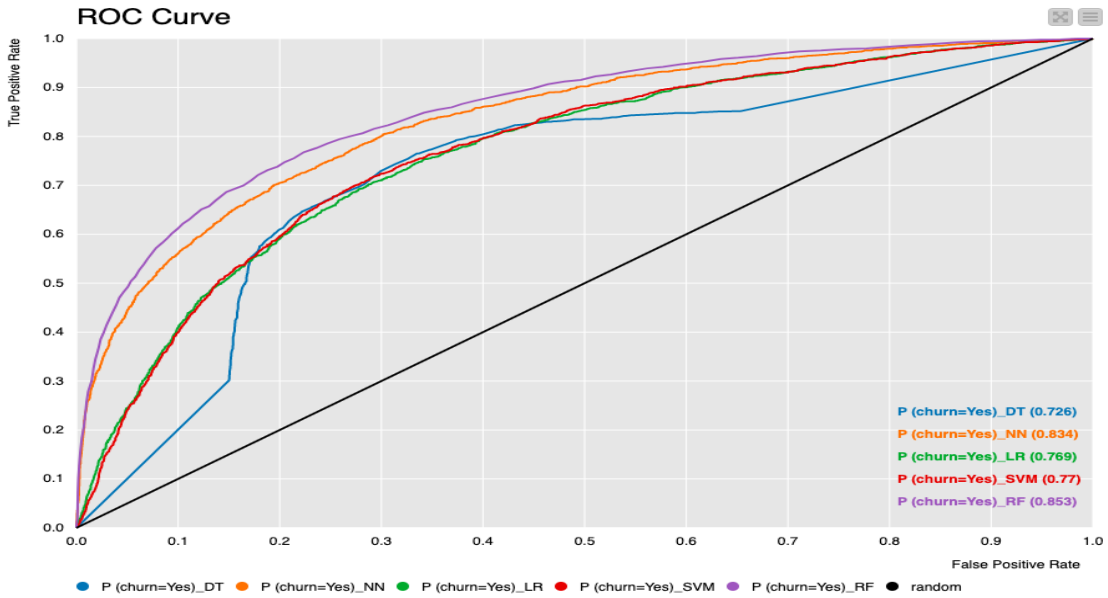


Figure 4.6. The ROC Curve of the Ensemble Model

4.5. The Selection of the Model for Deployment

The deployment stage is the final stage of the CRISP-DM methodology. In this stage, we select the best-performing one among the models for deployment. Table 4.10 shows the summary of the performance of the predictive models. In Table 4.10, it is evident that the RF is the best predictive model by all standards, followed by the ANN model. Although the scores of the DT, LR and SVM models are not so low, they are not good enough and need improvement.

Table 4.10. The Summary of the Performance of the Predictive Models

Model	Sensitivity (True Positive Rate)	Specificity (True Negative Rate)	Overall Accuracy	AUC
Decision Tree	70.10%	71.92%	71.55%	72.60%
ANN	72.70%	77.71%	76.69%	83.40%
LR	69.51%	71.76%	71.30%	76.90%
SVM	70.10%	72.4%	71.93%	77.00%
RF	73.83%	80.21%	78.91%	85.30%
Ensemble Model	70.10%	71.92%	71.55%	72.60%

Finally, although the RF is the best model by all standards and is recommended for deployment, its performance could be improved. To improve the performance of the models in general and the RF in particular, one should focus on how to improve the sensitivity and the overall accuracy of the models. The sensitivity determines the ability of a model to predict those that will churn from the list of customers in the dataset, i.e., the objective of the churn prediction. The sensitivity should be as close to 100% as possible. If a model has a 100 % sensitivity, it correctly predicts every person who will churn from the services of the bank. Although specificity is also important and should be as close to 100% as possible, it is not as important as sensitivity in customer churn prediction. The reason is that specificity only tells us those who will not churn so that the bank ignores them in designing marketing campaigns for customers at the risk of churning.

4.6. Variable Importance Graphic

After building and evaluating the predictive models, we used the Excel writer node in KNIME to create a variable importance graphic by using the output data of the RF predictor. We do this to ascertain the ranking order of how important each of the variables in the data is in building the RF predictor that happened to be the best-performing model in this study. Figure 4.7 shows the graphical representation of the variable importance graphic that we generated in Excel after exporting the output data of the RF predictor and by using the Excel Reader node in KNIME.

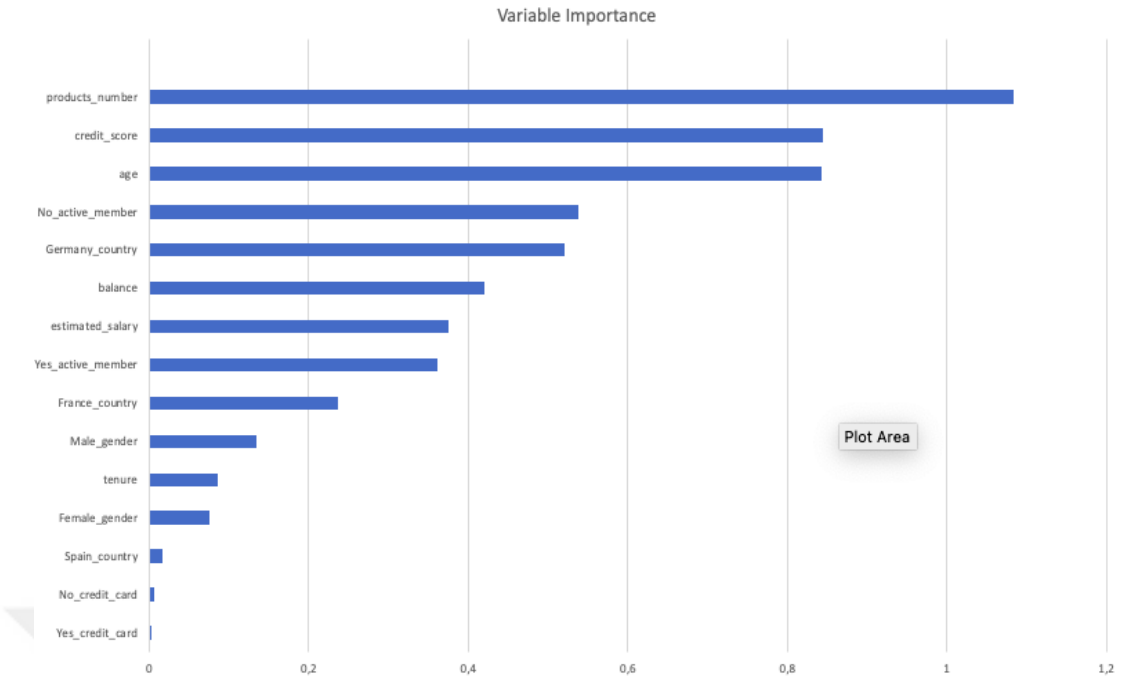


Figure 4.7. The Variable Importance Graphic

Figure 4.7 shows that the most important variables in building the best predictive model were the product_number, followed by the credit score and age. Similarly, the least important variables were Yes_credit_card and No_credit_card. This result indicates that having a credit card or not does not matter much in using the RF model to predict the customers that are likely to leave the bank's services.

CHAPTER V

CONCLUSIONS AND DISCUSSIONS

The rate of customer attrition has shown an upward trend in recent years, given the increasing competition across sectors. To deal with customer churn, companies formulate customer-centric strategies to enhance service quality and ensure sustainable customer relationships. The anticipation of customer churn also empowers companies to optimise resource allocation and minimise costs. Although there are many studies held on the prediction of customer attrition in the banking sector (Alizadeh et al., 2023; An et al., 2022; Belém, 2018; Bharathi et al., 2022; Charandabi, 2023; Guliyev & Tatoğlu, 2021; Karvana et al., 2019; Kaur & Kaur, 2020; Kaur et al., 2013; Khine & Myo, 2019, 2023; Li & Wang, 2018, 2018; Muneer et al., 2022; Seid & Woldeyohannis, 2022; Tran et al., 2023; Verma, 2020; Zoric, 2016), there are relatively few studies that employ CRISPM DM in conjunction with KNIME Analytics (Kumar & Ravi, 2008). In this study, we employed a data-driven approach and several predictive models, i.e., decision tree, random forest, logistic regression, artificial neural networks, support vector machine, and ensemble models, to forecast the customer churn of a multinational bank, i.e., ABC Multinational Bank, by using one month period data, from July 31, 2022, to August 29, 2022.

The results show that the random forest model has the highest performance in accurately predicting the churn of bank clients by its high overall accuracy of 78.91% and AUC score of 85.3%. This result is in line with the findings of the prior studies (Kaur & Kaur, 2020; Lemos et al., 2022; Rahman & Kumar, 2020; Verma, 2020). We also find that the decision tree model, which has an overall accuracy of 71.55% and an AUC score of 72.6, and the logistic regression model, which has an overall accuracy of 71.3% and an AUC score of 76.9%, are the least performing predictive models. Using the AUC score, we find that our model accurately identifies all customers that are likely to churn with an accuracy of 85.3%. Similarly, based on the overall accuracy metric, our model has the capability to identify customers who are likely to churn with an accuracy rate of 78.91%. The findings also identify the factors that have the highest

predictive power for potential customer churn. Historical records of the customer with the bank, its credit score, and age are more effective predictors than other factors related to transaction volume. Hence, enhancing the ties with customers may serve to mitigate customer churn in the banks.

5.1. Implications of the Study

This study offers valuable insights for financial institutions. By using reliable predictive models, banks may identify potential clients that are likely to switch to other financial institutions. This identification would allow banks to design innovative marketing strategies, powerful customer relationship management, and rewarding loyalty programs to prevent consumers from churning. In this frame, banks may act proactively in managing the preferences of customers to foster a strong relationship with them. Banks can also use their resources more efficiently and decrease their costs related to promotional activities, as they would commit fewer resources to retain clients.

5.2. Limitations of the Study and Future Research

We acknowledge that this study has some limitations. First, the validity and reliability of the findings are contingent upon the representativeness of the data obtained from Kaggle. Future studies may collect the data directly from the banks to produce more refined results. Second, the results rely on the selected model implemented in the KNIME platform. More sophisticated deep learning models like Convolutional Neural Networks and Gradient Boosting Machines like XGBoost, LightGBM, and CatBoost can produce better results. Third, we do not consider external factors, such as economic conditions or industry-specific variables, that may affect customer churn prediction. Future research may include these variables to measure their effects on the findings of the predictive models. Moreover, the data covers only a short period, i.e., one month, disregarding the changing behaviour of customers in the long run. Future studies may use long-term data to build predictive models for producing more accurate results. Finally, the findings of this study show the reflection of customers in a single international bank and cannot be generalised to encompass all multinational banks. Future research may assess the applicability of the results to other multinational banks operating under the same circumstances.

REFERENCES

Abbott (2014). *Applied predictive analytics: Principles and techniques for the professional data analyst.* Wiley.

Accenture. (2021). *Purpose-driven banking: The path to powerful digital transformation.* https://www.accenture.com/content/dam/accenture/final/industry/banking/document/Accenture-Purpose-Driven-Banking-2021.pdf

Agarwal, V., Taware, S., Yadav, S. A., Gangodkar, D., Rao, A. L. N., & Srivastav, V. K. (2023). Customer–churn prediction using machine learning. *2nd International Conference on Technological Advancements in Computational Sciences (ICTACS),* October 2022, 893-899. https://doi.org/10.1109/ICTACS56270.2022.9988187

Alizadeh, M., Zadeh, D. S., Moshiri, B., & Montazeri, A. (2023). Development of a customer churn model for banking industry based on hard and soft data fusion. *IEEE Access*, *11*(January), 1–1. https://doi.org/10.1109/access.2023.3257352

Amin, A., Anwar, S., Adnan, A., Nawaz, M., Alawfi, K., Hussain, A., & Huang, K. (2017). Customer churn prediction in the telecommunication sector using a rough set approach. *Neurocomputing*, *237*, 242–254. https://doi.org/10.1016/j.neucom.2016.12.009

Amin, A., Shah, B., Masood, A., Joaquim, F., & Moreira, L. (2019). Cross-company customer churn prediction in telecommunication: A comparison of data transformation methods. *International Journal of Information Management*, *46*, 304–319. https://doi.org/10.1016/j.ijinfomgt.2018.08.015

An, Z., Song, Z., & Wang, X. (2022). Bank customer churn based on different models, oversampling, and encoding methods. *BCP Business &Management, 26,* 703-713. https://api.semanticscholar.org/CorpusID:252934467

Azevedo, A., & Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: A Parallel overview. *IADIS European Conference on Data Mining 2008, Amsterdam, The Netherlands, July 24-26.*

Bahnsen, A. C., Aouada, D., & Ottersten, B. (2015). A novel cost-sensitive framework for customer churn predictive modelling. *Decision Analytics, 2*(5), 1-15. https://doi.org/10.1186/s40165-015-0014-6

Bank ABC. (2023). https://www.bank-abc.com/en

Belém, N. (2018). *Gauging and foreseeing customer churn in the banking industry: A neural network approach*. Master Thesis. Universidade Nova de Lisboa.

Benoit, D. F., & Van Den Poel, D. (2012). Improving customer retention in financial services using kinship network information. *Expert Systems with Applications, 39*(13), 11435–11442. https://doi.org/10.1016/j.eswa.2012.04.016

Bharathi, S. V., Pramod, D., & Raman, R. (2022). An ensemble model for predicting retail banking churn in the youth segment of customers, *Data, 7*(5), 61, 1-15.

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*, 123-140.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. doi: 10.1023/A:1010933404324

Business Wire. (2019). Churn analysis engagement helped a German banking firm to improve customer Retention rate by 85%. https://www.businesswire.com/news/home/20191001005871/en/Churn-Analysis-Engagement-Helped-a-German-Banking-Firm-to-Improve-Customer-Retention

CallMiner. (2020). *The CallMiner Churn Index 2020.* us-callminer-churn-index-2020-report.pdf.

Carter, J., Pan, J., Rai, S., & Galandiuk, S. (2016). ROC-ing along: Evaluation and interpretation of receiver operating characteristic curves. *Surgery, 159*(6). DOI: 10.1016/j.surg.2015.12.029

Charandabi, S. E. (2023). Prediction of customer churn in banking industry. *Researchgate.Net*.https://www.researchgate.net/profile/Sina-Esmaeilpour-Charandabi2/publication/342424673_Prediction_of_Customer_Churn_in_Banking_Industry/links/5ef39d534585153fb1b3852b/Prediction-of-Customer-Churn-in-Banking-Industry.pdf

Coser, A., Aldea, A., Maer-Matel, M. M., & Besir, L. (2020). Propensity to churn in banking: What makes customers close the relationship with a bank? *Economic Computation and Economic Cybernetics Studies and Research, 54*(2), 77–94. https://doi.org/10.24818/18423264/54.2.20.05

Dalbah, L. M., Ali, S., & Al-Naymat, G. (2022). An interactive dashboard for predicting bank customer attrition. *International Conference on Emerging Trends in Computing and Engineering Applications, ETCEA 2022-Proceedings*, 1–5. https://doi.org/10.1109/ETCEA57049.2022.10009818

De Caigny, A., Coussement, K., & De Bock, K. W. (2018). A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. *European Journal of Operational Research, 269*(2), 760–772. https://doi.org/10.1016/j.ejor.2018.02.009

Delen, D., Sharda, R., & Turban, E. (2018). *Business intelligence, analytics, and data science: A managerial perspective* (Fourth Edition). Pearson.

Dingli, A., Marmara, V., & Fournier, N. S. (2017). Comparison of deep learning algorithms to predict customer churn within a local retail industry. *International Journal of*

Machine Learning and Computing 7(5), 128-132. https://doi.org/10.18178/ijmlc.2017.7.5.634

Dube, D. (2020). Why retaining customers for banks is as important as winning new ones. Forbes. https://www.forbes.com/sites/forbestechcouncil/2020/05/27/why-retaining-customers-for-banks-is-as-important-as-winning-new-ones/?sh=2a2400683f98

Dwivedi, S., Kasliwal, P., & Soni, S. (2016). Comprehensive study of data analytics tools (RapidMiner, Weka, R tool, Knime). *2016 Symposium on Colossal Data Analysis and Networking, CDAN 2016*, 1–8. https://doi.org/10.1109/CDAN.2016.7570894

Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm. *In Machine Learning: Proceedings of the Thirteenth International Conference*, 148–156.

Friedman, J. H. (1999). *Stochastic gradient boosting.* Technical Report, Stanford University.

Gregory, B. (2018). Predicting customer churn: Extreme gradient boosting with temporal data. *First-place Entry for Customer Churn Challenge in WSDM Cup 2018.* https://arxiv.org/ftp/arxiv/papers/1802/1802.03396.pdf

Guliyev, H., & Tatoğlu, F. Y. (2021). Customer churn analysis in banking sector: Evidence from explainable machine learning models. *Journal of Applied Microeconometrics, 1*(2), 85–99. https://doi.org/10.53753/jame.1.2.03

Hegde, S., & Mundada, M. R. (2019). Enhanced deep feed forward neural network model for the customer attrition analysis in banking sector. *International Journal of Intelligent Systems and Applications*, *11*(7), 10–19. https://doi.org/10.5815/ijisa.2019.07.02

Hou, L., & Tang, X. (2010). Customer churn identifying model based on dual customer value gap. *Management Science and Financial Engineering*, *16*(2), 17–27.

Huang, B., Kechadi, M. T., & Buckley, B. (2012). Customer churn prediction in telecommunications. *Expert Systems with Applications*, *39*(1), 1414–1425. https://doi.org/10.1016/j.eswa.2011.08.024

Iranmanesh, S. H., Hamid, M., Bastan, M., & Shakouri, H., & Nasiri, M. M. (2019). Customer churn prediction using artificial neural network: An analytical CRM application. *3rd European International Conference on Industrial Engineering and Operations Management,* 2214–2226.

Jahan, I., & Farah Sanam, T. (2023). An improved machine learning based customer churn prediction for insight and recommendation in e-commerce. *2022 25th International Conference on Computer and Information Technology (ICCIT), December 2022*, 1–6. https://doi.org/10.1109/iccit57492.2022.10054771

Karvana, K. G. M., Yazid, S., Syalim, A., & Mursanto, P. (2019). Customer churn analysis and prediction using data mining models in banking industry. *2019 International*

Workshop on Big Data and Information Security, IWBIS 2019, 33–38. https://doi.org/10.1109/IWBIS.2019.8935884

Kaur, I., & Kaur, J. (2020). Customer churn analysis and prediction in banking industry using machine learning. *PDGC 2020 - 2020 6th International Conference on Parallel, Distributed and Grid Computing*, 434–437. https://doi.org/10.1109/PDGC50313.2020.9315761.

Kaur, M., Singh, K., & Sharma, N. (2013). Data mining as a tool to predict the churn behaviour among Indian bank customers. *International Journal on Recent and Innovation Trends in Computing and Communication*, *1*(9), 720-725 http://www.ijritcc.org

Kaya, E., Dong, X., Suhara, Y., Balcisoy, S., & Bozkaya, B. (2018). Behavioral attributes and financial churn prediction. *EPJ Data Science, 7*(41), 1-18. https://doi.org/10.1140/epjds/s13688-018-0165-5

Khine, S. T., & Myo, W. W. (2019). Customer churn analysis in banking sector. *University Journal of Research and Innovation 1*(1), 191–195.

Khine, S. T., & Myo, W. W. (2023). Mining customer churns for banking industry using K-means and multi-layer perceptron. *2023 IEEE Conference on Computer Applications (ICCA)*, 220–225. https://doi.org/10.1109/ICCA51723.2023.10182152

Kumar, A., & Ravi, V. (2008). Predicting credit card customer churn in banks using data mining. *International Journal of Information Systems in Service Sector (IJISS) and the International Journal of Information Technology Project Management, 1*(1), 4–28.

Kumar, S., & Dhandapani, C. (2016). A survey on customer churn prediction using machine learning techniques. *International Journal of Computer Applications*, *154*(10), 13–16. https://doi.org/10.5120/ijca2016912237

Lemos, R. A. de L., Silva, T. C., & Tabak, B. M. (2022). Propension to customer churn in a financial institution: a machine learning approach. *Neural Computing and Applications, 34*, 11751-11768.

Li, Y., & Wang, B. (2018). A study on customer churn of commercial banks based on learning from label proportions. *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*, 1241–1247. https://doi.org/10.1109/ICDMW.2018.00177

Liu, Y., Li, Z., & Huang, L. (2022). The application of blockchain technology in smart sustainable energy business model. *Energy Reports*, *8*, 7063–7070. https://doi.org/10.1016/j.egyr.2022.05.002

Liu, Y., Shengdong, M., Jijian, G., & Nedjah, N. (2022). Intelligent prediction of customer churn with a fused attentional deep learning model. *Mathematics*, *10*(24), 4733, 1–16.

Mahajan, D., & Gangwar, R. (2017). Improved customer churn behaviour by using SVM. *International Research Journal of Engineering and Technology (IRJET)*, *4*(8), 2364–2368.

McCulloch, W. & Pitts, W. (1943). A logical calculus of ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, *5*(4), 115–133.

Muneer, A., Ali, R. F., Alghamdi, A., Taib, S. M., Almaghthawi, A., & Ghaleb, E. A. A. (2022). Predicting customers churning in banking industry: A machine learning approach. *Indonesian Journal of Electrical Engineering and Computer Science*, *26*(1), 539–549. https://doi.org/10.11591/ijeecs.v26.i1.pp539-549

Nagaraju, J., & Vijaya, J. (2022). Boost customer churn prediction in the insurance industry using meta - heuristic models. *International Journal of Information Technology*, *14*(5), 2619–2631. https://doi.org/10.1007/s41870-022-01017-5

Sabbeh, S. F. (2018). Machine-learning techniques for customer retention: A comparative study. *International Journal of Advanced Computer Science and Applications*, *9*(2), 273–281.

Seid, M. H., & Woldeyohannis, M. M. (2022). Customer churn prediction using machine learning: Commercial Bank of Ethiopia. *2022 International Conference on Information and Communication Technology for Development for Africa, ICT4DA 2022*, 7–12. https://doi.org/10.1109/ICT4DA56482.2022.9971224.

Sharda, R., Delen, D., & Turban, E. (2020). *Analytics, Data Science, & Artificial Intelligence systems for Decision Support*. Pearson.

Shirazi, F., & Mohammadi, M. (2019). A big data analytics model for customer churn prediction in the retiree segment. *International Journal of Information Management, 48*, 238–253. https://doi.org/10.1016/j.ijinfomgt.2018.10.005

Tran, H., Le, N., & Nguyen, V. (2023). Customer churn prediction in the banking sector Using machine learning-based classification models. *Interdisciplinary Journal of Information, Knowledge, and Management 18*, 87–105.

Troncoso, C. A. M. (2018). *Predicting customer churn using voice of the customer. A text mining approach*. PhD Thesis, University of Manchester, Faculty of Humanities.

Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., & Chatzisavvas, K. C. (2015). A comparison of machine learning techniques for customer churn prediction. *Simulation Modelling Practice and Theory, 55*, 1–9. https://doi.org/10.1016/j.simpat.2015.03.003

Verma, P. (2020). Churn prediction for savings bank customers: A machine learning approach. *Journal of Statistics Applications and Probability*, *9*(3), 535–547. https://doi.org/10.18576/JSAP/090310

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, Manchester, 11-13 April 2000,* 29-40.

Zoric, A. B. (2016b). Predicting customer churn in banking industry using neural networks. *Interdisciplinary Description of Complex Systems*, *14*(2), 116–124. https://doi.org/10.7906/indecs.14.2.1

https://www.kaggle.com/datasets/gauravtopre/bank-customer-churn-dataset

https://www.knime.com/knime-hub

APPENDIX

APPENDIX A

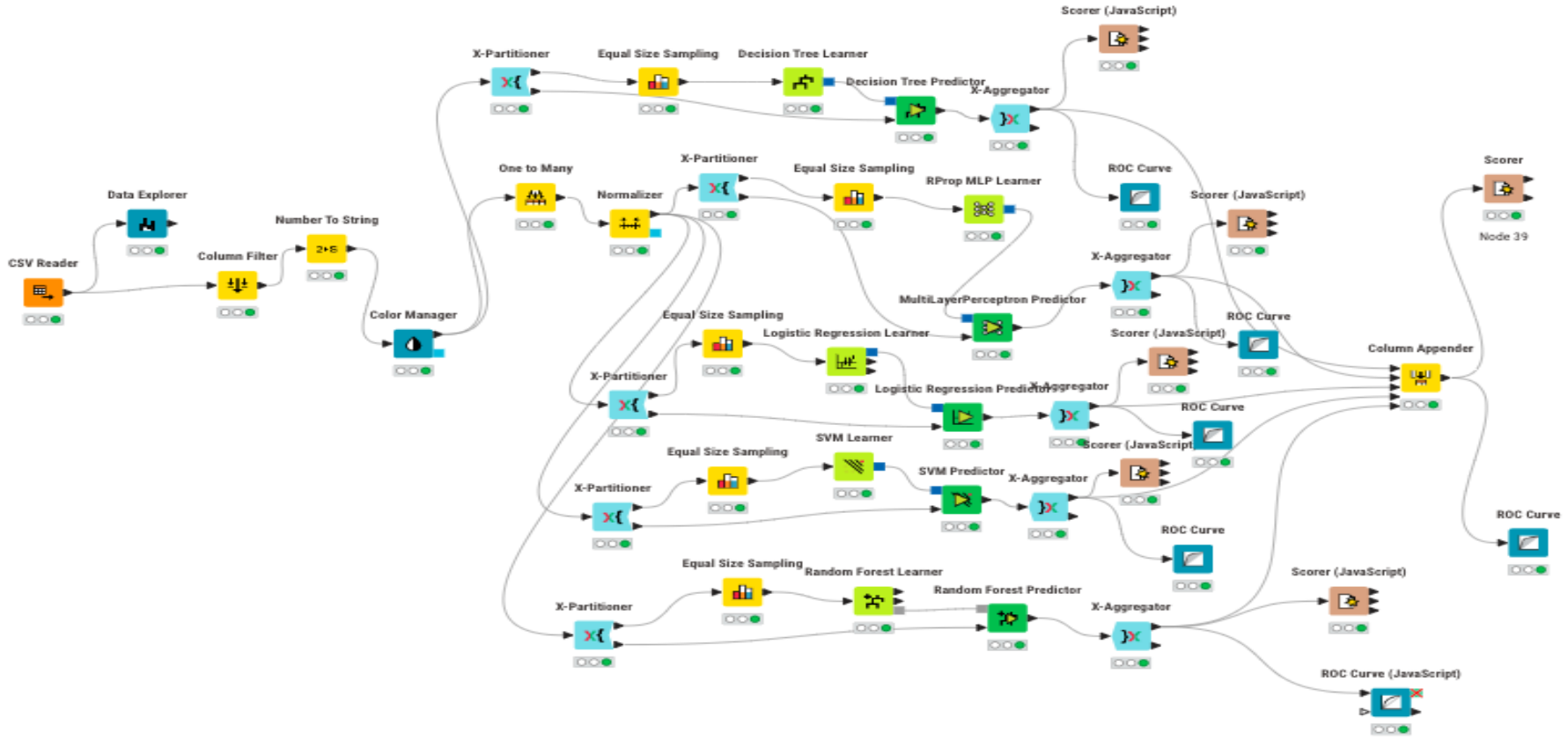


Figure A.1. Complete KNIME Workflow

CURRICULUM VITAE

Personal Information

Moro Husseini

E-mail (1):
E-mail (2):

Education:

February 2022 – February 2024	MA in Management, Ibn Haldun University
August 2015 – May 2019	BCOM Procurement and Supply Chain Management

Experience:

February 2022 – February 2024	Research Assistant, Ibn Haldun University
September 2019 – August 2020	Asst. Procurement Officer, KEEA Municipal Assembly