



UNIVERSIDAD DE INVESTIGACIÓN DE TECNOLOGÍA EXPERIMENTAL YACHAY

Escuela de Ciencias Matemáticas y Computacionales

TÍTULO: Customer churn prediction by using support vector machine

Trabajo de integración curricular presentado como requisito para la obtención del título de Ingeniero en Tecnologías de la Información

Autor:

Santacruz Nagua Alfredo Fabian

Tutor:

Ph.D Chang Tortolero Oscar Guillermo

Urcuquí, marzo 2021

SECRETARÍA GENERAL
(Vicerrectorado Académico/Cancillería)
ESCUELA DE CIENCIAS MATEMÁTICAS Y COMPUTACIONALES
CARRERA DE TECNOLOGÍAS DE LA INFORMACIÓN
ACTA DE DEFENSA No. UITEY-ITE-2021-00008-AD

A los 26 días del mes de mayo de 2021, a las 14:30 horas, de manera virtual mediante videoconferencia, y ante el Tribunal Calificador, integrado por los docentes:

Presidente Tribunal de Defensa	Dr. CUENCA LUCERO, FREDY ENRIQUE , Ph.D.
Miembro No Tutor	Dr. MAYORGA ZAMBRANO, JUAN RICARDO , Ph.D.
Tutor	Dr. CHANG TORTOLERO, OSCAR GUILLERMO , Ph.D.

El(la) señor(ita) estudiante **SANTACRUZ NAGUA, ALFREDO FABIAN**, con cédula de identidad No. **0705020873**, de la **ESCUELA DE CIENCIAS MATEMÁTICAS Y COMPUTACIONALES**, de la Carrera de **TECNOLOGÍAS DE LA INFORMACIÓN**, aprobada por el Consejo de Educación Superior (CES), mediante Resolución **RPC-SO-43-No.496-2014**, realiza a través de videoconferencia, la sustentación de su trabajo de titulación denominado: **CUSTOMER CHURN PREDICTION BY USING SUPPORT VECTOR MACHINEE**, previa a la obtención del título de **INGENIERO/A EN TECNOLOGÍAS DE LA INFORMACIÓN**.

El citado trabajo de titulación, fue debidamente aprobado por el(los) docente(s):

Tutor	Dr. CHANG TORTOLERO, OSCAR GUILLERMO , Ph.D.
--------------	--

Y recibió las observaciones de los otros miembros del Tribunal Calificador, las mismas que han sido incorporadas por el(la) estudiante.

Previamente cumplidos los requisitos legales y reglamentarios, el trabajo de titulación fue sustentado por el(la) estudiante y examinado por los miembros del Tribunal Calificador. Escuchada la sustentación del trabajo de titulación a través de videoconferencia, que integró la exposición de el(la) estudiante sobre el contenido de la misma y las preguntas formuladas por los miembros del Tribunal, se califica la sustentación del trabajo de titulación con las siguientes calificaciones:

Tipo	Docente	Calificación
Tutor	Dr. CHANG TORTOLERO, OSCAR GUILLERMO , Ph.D.	9,0
Miembro Tribunal De Defensa	Dr. MAYORGA ZAMBRANO, JUAN RICARDO , Ph.D.	5,0
Presidente Tribunal De Defensa	Dr. CUENCA LUCERO, FREDY ENRIQUE , Ph.D.	8,5

Lo que da un promedio de: **7.5 (Siete punto Cinco)**, sobre 10 (diez), equivalente a: **APROBADO**

Para constancia de lo actuado, firman los miembros del Tribunal Calificador, el/la estudiante y el/la secretario ad-hoc.

SANTACRUZ NAGUA, ALFREDO FABIAN
Estudiante



Dr. CUENCA LUCERO, FREDY ENRIQUE , Ph.D.
Presidente Tribunal de Defensa



Firmado electrónicamente por:
FREDY ENRIQUE
CUENCA LUCERO

Firmado electrónicamente por:
OSCAR GUILLERMO
CHANG TORTOLERO

Dr. CHANG TORTOLERO, OSCAR GUILLERMO , Ph.D.
Tutor



Firmado electrónicamente por:
**JUAN RICARDO
MAYORGA
ZAMBRANO**

Dr. MAYORGA ZAMBRANO, JUAN RICARDO , Ph.D.
Miembro No Tutor

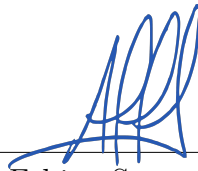
DAYSY MARGARITA MEDINA BRITO  Firmado digitalmente por DAYSY
MARGARITA MEDINA BRITO
Fecha: 2021.06.14 15:08:02 -05'00'

MEDINA BRITO, DAYSY MARGARITA
Secretario Ad-hoc

Autoría

Yo, **Alfredo Fabian Santacruz Nagua**, con cédula de identidad **0705020873**, declaro que las ideas, juicios, valoraciones, interpretaciones, consultas bibliográficas, definiciones y conceptualizaciones expuestas en el presente trabajo; así cómo, los procedimientos y herramientas utilizadas en la investigación, son de absoluta responsabilidad de el autor del trabajo de integración curricular. Así mismo, me acojo a los reglamentos internos de la Universidad de Investigación de Tecnología Experimental Yachay.

Urcuquí, Marzo del 2021.



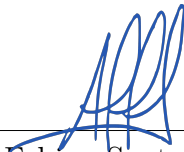
Alfredo Fabian Santacruz Nagua
CI: 0705020873

Autorización de publicación

Yo, **Alfredo Fabian Santacruz Nagua**, con cédula de identidad **0705020873**, cedo a la Universidad de Tecnología Experimental Yachay, los derechos de publicación de la presente obra, sin que deba haber un reconocimiento económico por este concepto. Declaro además que el texto del presente trabajo de titulación no podrá ser cedido a ninguna empresa editorial para su publicación u otros fines, sin contar previamente con la autorización escrita de la Universidad.

Asimismo, autorizo a la Universidad que realice la digitalización y publicación de este trabajo de integración curricular en el repositorio virtual, de conformidad a lo dispuesto en el Art. 144 de la Ley Orgánica de Educación Superior.

Urcuquí, Marzo del 2021.



Alfredo Fabian Santacruz Nagua
CI: 0705020873

Dedication

“A mi dos madres, mi abuela que ha cuidado de mí y me ha permitido seguir estudiando gracias a sus labores dentro del hogar y a mi madre materna, persona que a la distancia me ha ayudado moralmente como económicamente. También a mis compañeros de estudio, a mis maestros y a todos los que me apoyaron en la revisión de esta tesis.”

Acknowledgments

I would like to express my special thanks of gratitude to my supervisor Oscar Chang, Lecturer and Coordinator for guiding me throughout the research. Research itself helped me quite a lot in building up the fundamental knowledge of the field. I would like to thank the Departments of Bienestar Estudiantil & Administrativo for helping me with different academic problems.

Resumen

La predicción de la fuga de clientes es importante debido a la intensa competencia de marketing. Con el propósito de retener clientes, las empresas aplican modelos de predicción de abandono para determinar el abandono de clientes analizando su comportamiento y tratando de poner esfuerzo y dinero en retenerlos. En esta tesis, desarrollamos y probamos un modelo para estimar la propensión de un cliente a abandonar la empresa en un futuro próximo. Este estudio aplica máquinas de vectores de soporte (SVM), una técnica de aprendizaje automático utilizada en la clasificación binaria. Se comparó SVM con diferentes kernel: lineal, función de base radial (RBF) y polinomial. El experimento se llevó a cabo en Python con la herramienta de aprendizaje automático, junto con una base de datos real de Kaggle. Posteriormente, el rendimiento predictivo de los tres núcleos muestra que SVM con kernel polinomial y RBF tiene la mejor tasa de precisión y proporciona una medida eficaz para la predicción de fuga de clientes del banco. Los resultados se mostraron en diferentes medidas de evaluación.

Palabras Clave: Predicción de fuga de clientes, Maquinas de vectores de soporte, Kernel, Análisis de datos.

Abstract

Prediction of customer churn is important due to intensive marketing competition. With the purpose of retaining customers, companies apply churn prediction models to determine the customers churn by analyzing their behavior and trying to put effort and money into retaining them. In this thesis, we develop and test a model to estimate the propensity of a customer to abandon the company in a near future. This study applies support vector machines (SVM), a machine learning technique used in binary classification. SVM was compared with different kernels: linear, radial basis function (RBF) and polynomial. The experiment was carried out in Python with machine-learning tools, along with a real database from Kaggle. Afterward, the predictive performance of three kernel show that SVM with polynomial and RBF have the best accuracy rate and provide an effective measurement for the bank's customer churn prediction (CCP). The results were shown in different evaluation measures.

Keywords: Prediction Churn, SVM, Kernel, Data Analysis.

Contents

Dedication	v
Acknowledgments	vii
Resumen	ix
Abstract	xi
Contents	xiii
List of Tables	xv
List of Figures	xvii
1 Introduction	1
1.1 Background	1
1.2 Problem statement	2
1.3 Justification	2
1.4 Objectives	3
1.4.1 General Objective	3
1.4.2 Specific Objectives	3
2 Theoretical Framework	5
2.1 Customer churn	5
2.1.1 Churn matter	5
2.1.2 Churn and retention rate	6
2.2 Predictive Machine Learning models	7
2.2.1 Artificial Neural Network (ANN)	7
2.2.2 Support vector machine	7
2.2.3 Decision tree	7
2.3 Proposed CCP model	8
2.3.1 Support vector machine	8
2.3.2 Optimal separation hyperplane (OSH)	9
2.3.3 Construction of the optimization problem	9

2.3.4	Analytical resolution of the optimization problem	11
2.3.5	Nonlinear SVM	13
3	Methodology	15
3.1	Dataset	15
3.2	Data preprocessing	15
3.3	Feature selection	17
3.4	Creation and transformation of variables	18
3.5	Classification	19
3.6	Evaluation measures	19
3.7	Experimental Setup	21
4	Results and Discussion	25
5	Conclusions	33
	Bibliography	35

List of Tables

3.1	Dataset Type	16
3.2	Description Dataset	16
3.3	Confusion matrix for classifier evaluation	20
4.1	Acc: Accuracy, Prs: Precision, Rec: Recall, F-m: F-measure	31

List of Figures

2.1	Most Significant Retail Revenue [1]	6
2.2	SVM-Optimal Separating Hyperplane	11
3.1	Knowledge discovery in database process [2]	17
3.2	CV basic approach [3]	19
4.1	Correlation Matrix	25
4.2	Multicollinearity using VIF	26
4.3	Churn Distribution	27
4.4	Gender Distribution	28
4.5	Age Distribution	28
4.6	Geography Distribution	29
4.7	Churn Rate by Tenure	30
4.8	Credit Card and Active Member Distribution	30
4.9	Roc Auc	32

Chapter 1

Introduction

1.1 Background

Normally, each year some clients decide to stop consuming products or services and go over to the competition [4]. And since it is closely related to the economic health of companies, the churn rate should be one of the first metrics to analyze. Each customer will have their reasons for leaving a company and companies will need to determine those reasons to avoid losing customers. In this case, banking companies have a database of their clients, where information can be extracted and the reasons for abandonment can be interpreted. Therefore, analyzing the data would reduce customer churn.

Reducing customer churn is one of the main problems for companies because it causes direct loss of revenue and marketing costs. Since investing in marketing is necessary to replace customers churn with new customers, it is more difficult and expensive to acquire new customers than retain them.

There are two types of churn: voluntary and involuntary. A voluntary churn is when customers choose to cancel their service with the company, where the company does not interfere in this decision. This type of abandonment could be preventable. On the other hand, involuntary churn occurs when the company is directly responsible. This thesis studies the prediction of voluntary churn by using a machine learning model fed with customer information.

Customer churn's information is useful, thanks to the fact that customers who leave the company often leave a trace of their passage through the company in the databases. Therefore, it is possible for companies' decision-makers to exploit such a data to prevent or minimize future churns, analyzing customer's behavior by using predictive models [4]. Concretely, Customer Churn Prediction (CCP) allows the estimation of the probability for future customer churn based on the historical behavior of the customer [5].

CCP is one of the best tools to identify customers who will abandon the company. This will allow companies to make decisions to proactively reduce churn. Various Machine Learning (ML) techniques, such as logistic regression [6] [7] [8], non-parametric statistical models such as K-nearest neighbors (KNN) [9], decision trees [10], and neural networks [11], have been proposed which can efficiently classify the customers as churn or non-churn. In this thesis SVM be used SVM for predicting customer churn.

In Chapter II, we will introduce a definition about customer churn or retention churn, as well as a description about the proposed CCP model. Chapter 3 describes the methodology for the model proposed. The code implementation and tools needed for this project have been discussed in Chapter 4. Chapter 5 will talk about the outcomes. Finally, Chapter 6 shows the conclusions and future work.

1.2 Problem statement

Customer churn analysis and prediction play an important role in the benefit of an enterprise. It is easier for a company to implement strategies to retain an existing customer as opposed to attracting a new customer which costs 6 times more [12]. Taking into account that the defection of clients that companies face has an economic impact, it is important to have a predictive model that reduces this impact.

The creation of a predictive model allows companies to monitor target customers, who stop using their services. Once the target customers have been determined, the company can implement strategies to retain them. The best customer is the one who always comes back. The returning customer is cheaper for the company because it spends less on advertising or incentives such as price cuts or gifts. A happy customer also tells other potential customers about the product or service. So determining a method considered optimal to retain customers is difficult since it involves many aspects according to the variables managed by the company, which are not the same in other companies. Therefore, the big problem to take into account is the fact that analyzing the different factors involved in the the loss of customers make its classification a difficult task.

1.3 Justification

The service sector in a country is varied; it includes companies dedicated to telephony, transport, hotels, insurance, credit, and educational services among many others. These types of companies are generally judged by the poor quality of the service they provide, which influences customers to look for other companies that offer better service. Therefore, the quality of service has an important role in any company, which will allow them to retain their existing and attract new customers.

Financial institutions are not the exception, in recent years with the advancement of digital transactions and the growing competition they face. Financial institutions are committed to improving the quality of their services in order to retain customers who are more selective and informed. Taking into account that customers are important to companies, this study attempts on developing a CCP by using SVM.

Recently, SVM have become increasingly popular and have been applied to a variety of tasks. However, it is difficult to select the adequate model for a specific application because of the tuning parameters. One of the most important tuning parameters is the kernel. Therefore, this work presents an evaluation of different kernels to provide a viewpoint to select the most appropriate model for churn prediction, regarding accuracy and precision in terms of churn or non-churn.

Due to the importance of customers, companies are forced to do an analysis of their data in order to avoid the loss of their customers. Because of this, this thesis will focus on developing a model based on SVM to predict customer churn in a banking institution from a database obtained on the Kaggle platform. Thus, this thesis will show companies about the use of machine learning, in addition to offering a look at this new trend in data analysis.

1.4 Objectives

1.4.1 General Objective

The general objective of this thesis is to build a churn prediction model which can identify churn and non-churn customers from banking companies using the support vector machine based on customer behavior.

1.4.2 Specific Objectives

- To analyze and pre-process customer data.
- To select the most relevant predictor variables for the training and testing of the model.
- To design a model to classify clients as churn and non-churn.
- To compare the results and performance of different kernels

Chapter 2

Theoretical Framework

2.1 Customer churn

Customer churn, characterized by the inclination of customers to stop using goods or services with a company, has become a major problem and is one of the main challenges faced by many companies around the world [13].

The churn behavior can be classified into the following sub-categories [14]:

- Voluntary customer churn, in which a customer decides to stop using the service from an company on their own responsibility, and
- Involuntary customer churn, in which the company or service provider decides to terminate a contract with their customers.

2.1.1 Churn matter

The explanation below, where the two reasons for the impact of customer churn, is based on Zarema Plaksij (Content Editor & Senior Copywriter at Self-employed. Self-EmployedUniversity of Nottingham) retrieved from [1].

- First, customer churn causes financial problems.
- The second reason lies in the fact that the more customers a business retains, the more revenue it makes!

According to Gartner, a staggering 80% of a companies future revenue will come from just 20% of its existing customers. Losing customers not only leads to opportunity costs because of reduced sales but also leads to an increased need for attracting new customers [15].

Besides the direct loss of revenue that results from a customer abandoning the business, it costs roughly 5-6 times as much to sign on a new customer as to retain an existing one [16](Lovelock and Wright, 1999, cited in McIlroy, A., & Barnett, S., 2000). Furthermore, it is always more difficult and costly to acquire a new customer than preserving an existing

one [17]. Marketing Metrics claims that the probability of selling to an existing customer is 60-70%, and only 5-20% to sell to a new prospect.

The same truths are revealed by KPMG, an organization dedicated to offer audit services, who found that customer retention is the main driver of a company's revenue Fig. 2.1.

MOST SIGNIFICANT RETAIL REVENUE DRIVERS

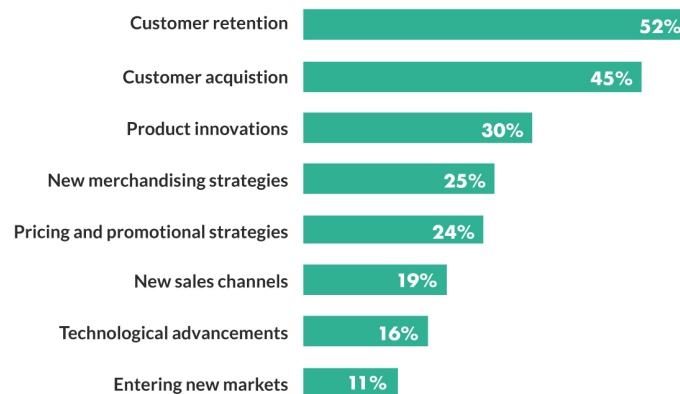


Figure 2.1: Most Significant Retail Revenue [1]

2.1.2 Churn and retention rate

Hwang et al. [18] define churn like the number or percentage of regular customers who abandon relationship with service provider and that is closely related to the customer retention rate and loyalty.

Churn rate and retention rate terms that can be confusing to understand. Molly Galetto [19] explain how the calculate churn rates vary (CCR) from industry to industry. First by economic, as churn rate may represent the total number of customers lost over the company's total customer count or the percent of recurring value lost. The second way to calculate churn is based on time, organizations calculate churn rate for a certain period of time, such as quarterly periods or fiscal years. Third, one of the most commonly used methods for calculating customer churn is to divide the total number of clients a company has at the beginning of a specified time period by the number of customers lost during the same period. CCR is calculated as follows:

- Customer Churn Rate (CCR):
Definition: The percentage of customers that are lost over a given period of time.
Calculation: $(\text{Customers churned in period} / \text{Customers at the start of the period})$
- Customer Retention Rate (CRR):
Definition: The percentage of customers retained over a given period of time.
Calculation: $(1 - (\text{customers churned in period} / \text{customers at the start of the period}))$

ReliaBills mentions in [20] that the goals of a company should be to have a low churn rate rather than a high churn rate. If the churn rate of its customers is high, this indicates that the company has problems with its product or service. Also, StartUpDonut [21] also mentions the use of new customers in the metrics for calculating customer retention, as opposed to metrics that only deal with existing customers.

2.2 Predictive Machine Learning models

According to the aforementioned, the churn rate is one of the markers that determines the proper functioning of a company. The company must be aware of the value of its customers in its economic activity. Companies must analyze the causes of abandonment of their clients, understand the behavior that their clients are carrying out, and the reason for the transfer to the competition. With this in mind, developing a predictive model is one of the most important steps [22]. In this section, three important and practical techniques for predicting the churning of customers are briefly introduced, whose reliability, performance and functionality have been proved in a great number of studies [22], [23], [24], [25].

2.2.1 Artificial Neural Network (ANN)

ANN is one of the most successful machine-learning methods for solving complex problems including the prediction of customer churning [25]. Neural networks such as multi-layer perceptron (MLP) and radial basis function (RBF) are among the most popular and the most commonly used feed forward neural networks with supervised learning [25]. ANN consists of a number of layers which are completely connected and the information moves from a layer to another layer (from left to right). For the problem of predicting customer churning, Keramati et al [26] show that ANN has a significantly better performance compared to DT and SVM methods.

2.2.2 Support vector machine

The support vector machine was proposed by Cortes and Vapnik [27]. SVM is a supervised learning method which is capable of solving linear and non-linear classification problems [28], [29]. In SVM, the new unlabeled samples are classified into two classes and according to the side of the hyperplane they lies. SVM is applicable for purposes especially in cases where classification needs to be done in exactly two classes. SVM is considered to be a very promising method for predicting customer churning [11], [30], and it has gained very successful experimental experiences in this field in a way that many studies show that this method provides a better performance compared to DT [30], [25], [31].

2.2.3 Decision tree

The Decision tree is one of the most popular and common classification techniques [32], [24], and one of the most important reasons behind its popularity is the fact that it is

flexible and easy to understand, as stated in [33]. And considering the type of the data, it can provide a good performance as well as accurate models for churning prediction [25].

2.3 Proposed CCP model

The problem posed was raised with a binary classification approach [34]. This type of procedure is based on the determination of a classifying function that allows assigning each object to one of the two classes defined a priori. In this case, each client will be assigned to one of the “Churn” or “Non Churn” classes. The construction of the model is carried out in two stages: training and testing. For each of the stages, a subset of the total objects (customers) to be classified is considered. These subsets of objects form a partition of the total set of objects and are called the training set and the test set, respectively. In the training stage, the best classifying function is estimated considering some criterion (for example, the classification error) in the training set. In the test stage, the effectiveness of the model is validated with respect to objects not used in training. For this, the model obtained is used to classify the elements of the test set. The model assigns each object to one of the defined classes, which are called the “predicted class” as opposed to the “real class”. Considering the “misclassified” objects (those whose predicted class is different from its real class) a classification error is estimated. Depending on this error, the proposed model is revised. There are several techniques that deal with the binary classification problem. Among these we can mention: artificial neural networks, decision trees and SVM [34]. For this work, motivated by the effectiveness and robustness in classification problems reported in [35], [36], [34], SVMs method was selected. This technique is briefly described below and the way in which the separation function is obtained.

2.3.1 Support vector machine

SVM is a nonlinear programming model that searches for the hyperplane that best separates the data set into two regions or given classes, for this study such as “Churn” or “Non-churn”.

On the one hand, hyperplanes that are farther from the boundaries of the object classes correspond to larger margins of separation. And, hyperplanes that are more accurate in assigning objects to the classes to which they belong have a smaller classification error. Therefore, an ideal separation hyperplane should maximize the separation margin and minimize the classification error [27]. However, it is not always possible to meet the two objectives simultaneously. To overcome this difficulty, an optimization problem is posed whose objective function combines both objectives. This optimization problem turns out to be a convex quadratic minimization problem [36]. In the case that the number of objects to be classified is greater than the number of attributes of each object, which usually happens, this problem has a single optimal solution. We will call the separation hyperplane associated with this solution the optimal separation hyperplane (HOS). What is described above corresponds to the case that there is a hyperplane of separation of the classes. In that case the classes are said to be linearly separable. For problems that are not linearly separable, the kernel functions are used.

Next, the optimization problem used to determine the separation hyperplane is presented for classifying the behavioral patterns studied (“Churn” - “Non Churn”).

2.3.2 Optimal separation hyperplane (OSH)

Let us consider a binary classification problem for which the training set has already been defined. In our case, the objects to be classified are the clients. Suppose that for each client n variables have been defined to study and that there are m clients in the training set. Then, each client is represented by a characteristic vector of dimension $n + 1$ whose first n coordinates correspond to the variables of the study and the last corresponds to the class to which the client belongs. We will denote these vectors as a pair (\vec{x}, y) where $\vec{x} = (x^1, \dots, x^n)$ are the variables of the study, and the last coordinate $y \in \{-1, +1\}$ indicates what class belongs to the client. In particular, we will denote by (\vec{x}_i, y_i) the characteristic vector corresponding to customer $i (i = 1, \dots, M)$. In this sense, when it is clear from the context, if the client i is represented by a vector of the form $(\vec{x}_i, +1)$, we will say that this client is in the positive class or that the vector \vec{x}_i is in the positive class. Analogously, for a client i represented by a vector of the form $(\vec{x}_i, -1)$ we will say that the client’s vector \vec{x}_i is in the negative class.

An initial assumption for this work is that the training set is linearly separable. That is, there is a hyperplane at \mathbb{R}^n that leaves all the vectors \vec{x} associated with the clients of a class on one side of the hyperplane and to those of the other on the other side. Formally, there is a pair $(\vec{\alpha}, b) \in \mathbb{R}^{n+1}$ as $\alpha \cdot \vec{x}_i + b > 0$ if client i is in the positive class and $\alpha \cdot \vec{x}_i + b < 0$, if it is in the negative class.

In this way, given a nonzero vector normal to the separation hyperplane $\alpha \in \mathbb{R}^n$, then if $f(\vec{x}_i) > 0$ client i is in the positive class and if $f(\vec{x}_i) < 0$, i is in the negative class.

To define the classification margin, we consider the distances d^+ and d^- . The distance d^+ is the Euclidean distance between the hyperplane and the positive class. That is, the distance between the hyperplane and the point in the positive class closest to it. In an analogous way, but with respect to the negative class d^- , it is the distance between the hyperplane and the negative class. Once these distances are determined, we define the margin of separation (of the hyperplane) as the sum $d^+ + d^-$.

2.3.3 Construction of the optimization problem

From the previous discussion we know that to correctly classify all clients in the training set we must impose the restrictions:

$$\begin{aligned} \vec{\alpha} \cdot \vec{x}_i &> 0 && \text{for a customer in positive class} \\ \vec{\alpha} \cdot \vec{x}_i &< 0 && \text{for a customer in negative class} \end{aligned} \tag{2.1}$$

It is not a good practice to include strict inequalities as constraints in optimization problems. Therefore, it would be important to be able to find conditions equivalent to the previous ones but that are of the type “less than or equal” or “greater or equal”. Remembering that if a hyperplane is defined by the pair $(\vec{\alpha}, b)$, then any pair of the form $(\lambda\vec{\alpha}, \lambda b)$ with $\lambda > 0$ also defines the same hyperplane, so we can say that if $\vec{\alpha} \cdot \vec{x} + b > 0$ for a particular pair $(\vec{\alpha}, b)$, then there exists a pair $(\vec{\alpha}_i, b_i)$ that defines the same hyperplane

and such that $\vec{\alpha}_i \cdot \vec{x} + b_i \geq 1$. For the negative class, one can reason in an analogous way. Since the training set is finite, if there exists a pair $(\vec{\alpha}, b)$ that satisfies conditions (2.1) for all clients in the training set, then there exists a pair $(\vec{\alpha}', b')$ that defines the same hyperplane such that the left side of the expressions in (2.1) has an absolute value greater than or equal to 1. In this way, the conditions of equation (2.1) can be included in the optimization problem in the form:

$$\begin{aligned} \vec{\alpha} \cdot \vec{x}_i + b &\geq 1 && \text{for customer in positive class} \\ \vec{\alpha} \cdot \vec{x}_i + b &\leq -1 && \text{for customer in negative class} \end{aligned} \quad (2.2)$$

The points (customers) for which one of these constraints is active, that is $|\vec{\alpha} \cdot \vec{x}_i + b| = 1$, are especially important. In particular, these points lie exactly on the so-called canonical hyperplanes that are defined by equations [27]:

$$\begin{aligned} \vec{\alpha} \cdot \vec{x} + b &= +1 \\ \vec{\alpha} \cdot \vec{x} + b &= -1 \end{aligned} \quad (2.3)$$

We will denote by x^+ any client that is in the canonical hyperplane of the positive class (right side equal to +1) and by x^- any client that is in the canonical hyperplane of the negative class. The hyperplane defined by the pair $(\vec{\alpha}, b)$ is parallel to the canonical hyperplanes. With this notation, the margin of separation of the hyperplane defined by the pair $(\vec{\alpha}, b)$, and denoted by γ , can be expressed as [27]:

$$\gamma = \frac{1}{2} \left[\left(\frac{\vec{\alpha} \cdot x^+}{\|\vec{\alpha}\|} \right) - \left(\frac{\vec{\alpha} \cdot x^-}{\|\vec{\alpha}\|} \right) \right] \quad (2.4)$$

From the previous expression we get:

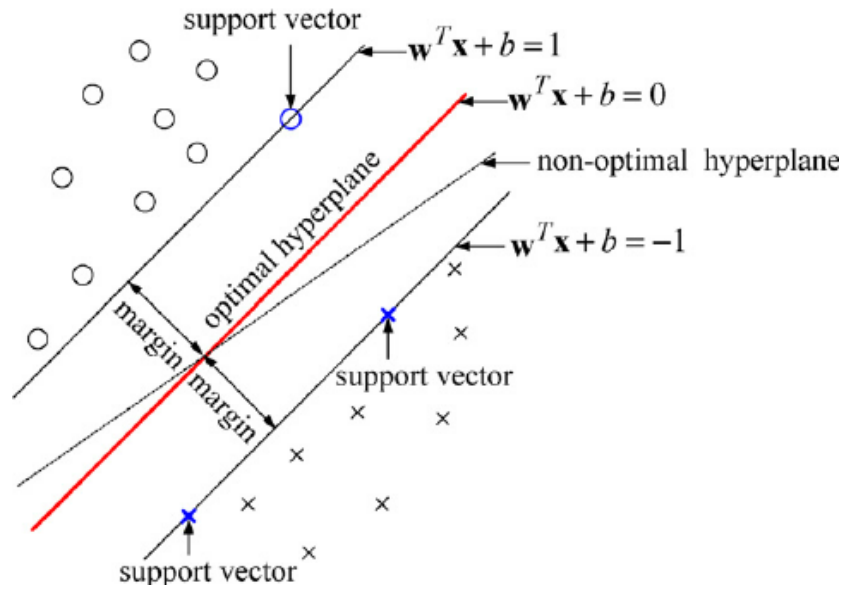
$$\gamma = \frac{1}{2\|\vec{\alpha}\|} \left[(\vec{\alpha} \cdot x^+) - (\vec{\alpha} \cdot x^-) \right] = \frac{1}{\|\vec{\alpha}\|} \quad (2.5)$$

Note that the distance between the hyperplane defined by the pair $(\vec{\alpha}, b)$ to each of the canonical hyperplanes is equal to $1/\|\vec{\alpha}\|$. Also, the closest points to this hyperplane in the training set are on the canonical hyperplanes. Therefore, the separation margin of this hyperplane is equal to $2/\|\vec{\alpha}\|$. In figure 9 this situation is illustrated: H_1 and H_2 are the canonical hyperplanes, Fig. 2.2.

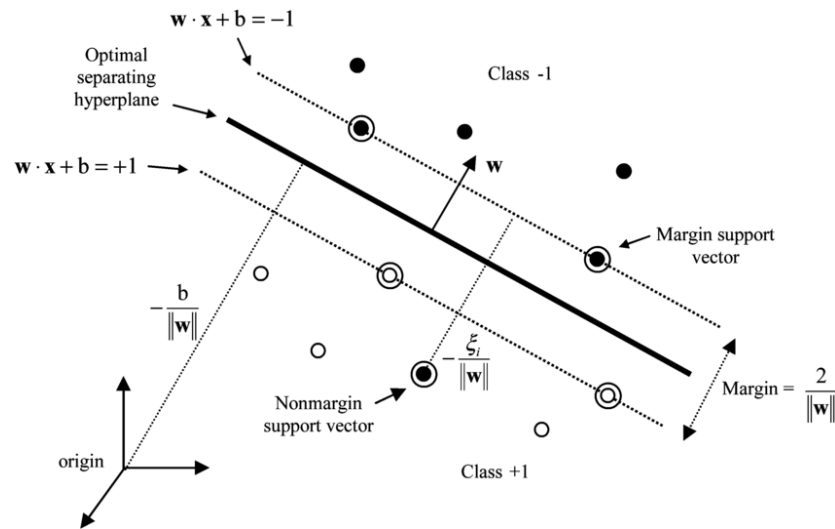
To maximize this margin is equivalent to minimizing the Euclidean norm of $\vec{\alpha}$. In this way, the optimization problem that we are going to solve is the following [27]:

$$\begin{aligned} \max_{\vec{\alpha}, b} \quad & \frac{1}{2} \|\vec{\alpha}\|^2 \\ y_i \cdot (\vec{x}_i \cdot \vec{\alpha} + b) - 1 & \geq 0 \quad \text{para } i = 1, \dots, m \end{aligned} \quad (P)$$

Solving this problem, an $\vec{\alpha}^*$ optimal vector is obtained that allows to calculate the maximum geometric margin of a separation hyperplane, $\gamma^* = 2/\|\vec{\alpha}^*\|$. In the next section, this optimization problem is analyzed analytically and a system of equations is described that allows us to calculate the optimal vector $\vec{\alpha}^*$ and the right side b .



(a) Optimal Hyperplane [3]



(b) OSH in SVMs for a linearly nonseparable classification [37]

Figure 2.2: SVM-Optimal Separating Hyperplane

2.3.4 Analytical resolution of the optimization problem

The optimization problem posed in the previous section is a convex quadratic problem. In the case that interests us, this problem has a single global optimal solution. This solution can be obtained by finding a solution for the first order optimality conditions, known as the Karush, Kuhn and Tucker conditions [27]. For, this we define the Lagrangian of (P).

$$L_P = \frac{1}{2} \|\vec{\alpha}\|^2 - \sum_{i=1}^m \beta_i y_i (\vec{x}_i \cdot \vec{\alpha} + b) + \sum_{i=1}^m \beta_i \quad (2.6)$$

Where the β_i are the Lagrange multipliers associated with the constraints of (P). From this Lagrangian we can propose the optimal conditions [38].

$$\frac{\partial L_P}{\partial \alpha_j} = \alpha_j - \sum_{i=1}^m \beta_i y_i x_i^j = 0 \quad j = 1, \dots, n \quad (2.7)$$

$$\frac{\partial L_P}{\partial b} = - \sum_{i=1}^m \beta_i y_i = 0 \quad (2.8)$$

$$y_i(\vec{x}_i \cdot \vec{\alpha} + b) - 1 \geq 0 \quad i = 1, \dots, m \quad (2.9)$$

$$\beta_i \geq 0 \quad i = 1, \dots, m \quad (2.10)$$

$$\beta_i(y_i(\vec{\alpha} \cdot \vec{x}_i + b) - 1) = 0 \quad i = 1, \dots, m \quad (2.11)$$

From the conditions (2.7) and (2.8), the optimum are obtained in the following relations:

$$\vec{\alpha} = \sum_{i=1}^m \beta_i y_i \vec{x}_i \quad (2.12)$$

$$\sum_{i=1}^m \beta_i y_i = 0 \quad (2.13)$$

Replacing these expressions in the equation (2.6) we can take the Lagrangian to the form:

$$L_D = \sum_{i=1}^m \beta_i - \frac{1}{2} \sum_{i=1}^m \beta_i \beta_s y_i y_s \vec{x}_i \cdot \vec{x}_s \quad (2.14)$$

From this expression, applying Lagrangian duality, a dual problem known as the Wolfe Dual [39, 38] of (P) can be obtained:

$$\begin{aligned} \max_{\beta} L_D &= \sum_{i=1}^m \beta_i - \frac{1}{2} \sum_{i=1}^m \beta_i \beta_s y_i y_s \vec{x}_i \cdot \vec{x}_s \\ &\sum_{i=1}^m \beta_i y_i = 0 \\ &\beta_i \geq 0 \quad i = 1, \dots, m \end{aligned} \quad (D)$$

In the case of regular convex optimization problems, such as problem (P), the Wolfe Dual (D) is a strong dual. The resolution is reduced, then, to obtain the optimal values for the multipliers β_i . Once these are known, the optimal values of the primary variables $\vec{\alpha}$ are obtained from the relation X.

Although the term b or *bias* does not appear explicitly in the formulation, it can be obtained from the optimally conditions [27]:

$$b^* = -\frac{1}{2} \left(\max_{y_i=-1} (\vec{\alpha}^* \cdot \vec{x}_i) + \max_{y_i=+1} (\vec{\alpha}^* \cdot \vec{x}_i) \right) \quad (2.15)$$

Summarizing, to find the pair $(\vec{\alpha}^*, b^*)$ that defines the optimal separation hyperplane, the dual problem (D) is solved, obtaining the optimal multipliers $\vec{\beta}^*$. Using these values in the expressions (2.12) and (2.15) we obtain $\vec{\alpha}^*$ and b^* , respectively.

2.3.5 Nonlinear SVM

Since real life is not perfectly linearly separable, then the kernel trick is used for this, which mapped the data in a higher alternative dimension in order to arrive at a linearly separable solution [40]:

$$x = (x_1, \dots, x_n) \longrightarrow \phi(x) = (\phi_1(x), \dots, \phi_n(x)) \quad (2.16)$$

In new data, a new representation that does not affect computation [40]:

$$F = \{\phi(x) : x \in X\} \quad (2.17)$$

Where ϕ is the embedding map and x the vector containing the feature values. A kernel is a function k for all $x_i, x_j \in X$ satisfies [40]:

$$K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle \quad (2.18)$$

Then the optimal classification is [40]:

$$f(x) = \text{sgn}\left(\sum_{i=1}^m \alpha_i y_i K(x_i, x) + b^*\right) \quad (2.19)$$

Common kernel functions are as follows [31]:

- Linear: $K(x, x_i) = (x \cdot x_i)$
- Polynomial: $K(x, x_i) = (x \cdot x_i)^d, K(x, x_i) = ((x \cdot x_i)^d + 1)$
- RBF: $K(x, x_i) = \exp(-\|x - x_i\|^2 / \sigma^2)$

Chapter 3

Methodology

The main goal of this thesis is to evaluate the prediction capability by using SVM with different kernel functions for customer churn prediction. The Kaggle dataset "Bank's customer churn prediction" was selected as the study area. The dataset was partitioned into training and testing datasets through random selection. Three SVM kernel types such as linear (LN), polynomial (PL), and radial basis function (RBF), were utilized to check the robustness of the SVM model.

3.1 Dataset

The data set was obtained from an online platform called Kaggle created on March 19, 2018. This data set presents a class imbalance problem [6] that makes it difficult to recognize minority classes "Churn" [33]. To handle this minority class, K-Fold Cross Validation was applied to the data set. The data provided by the database includes the client's personal information: age, salary, geographic location, and data associated with the banking service for a total of 10,000 customers. The information associated with each customer is shown in Table 3.1.

Int the table 3.1, the first 13 columns are the independent variable, and the last column is the dependent variable that contains a binary value of 1 or 0, where 1 refers to the customer who left the company, and 0 refers to customers who still remain with the company. This implies that the customer are divided into two classes, Churn and Non-Churn. Therefore, we can model the problem as a binary classification problem.

The subject database shown in Table 3.2, where the database containing customers are Churn and Non Churn.

3.2 Data preprocessing

Data is always incomplete, noisy and inconsistent due to human or computer error at data entry, shortcomings in data transmission etc. Therefore, data cleaning, data integration, data transformation, data reduction and data scaling are major tasks in data preprocessing

Column	Type
RowNumber	int64
CustomerId	int64
Surname	object
CreditScore	int64
Geography	object
Gender	object
Age	int64
Tenure	int64
Balance	float64
NumOfProducts	int64
HasCrCard	int64
IsActiveMember	int64
Estimated Salary	float64
Exited	int64

Table 3.1: Dataset Type

Description	Dataset
Independent variables	12
Dependent variables	1
Class Label	2
Total number of samples	10.000
No. Churn Samples	7.960
No. Non Churn Samples	2.040

Table 3.2: Description Dataset

[41]. The steps taken in this study are in accordance with the KDD process described by Fayad [2], Fig 3.1 .

Data preprocessing is the one of the basic and important processes in machine learning. In this study, three tasks carried out for preprocessing the data. Firstly, we have performed the missing value analysis and it was found that, fortunately, there was no missing value present in the dataset. We have also checked for duplication of customer ID and it shows that all the rows are unique. Second, the conversion of data types i.e., categorical data type to numerical data type was implemented. Finally, feature selection was needed in order to eliminate variables that do not have a significant impact on our model.

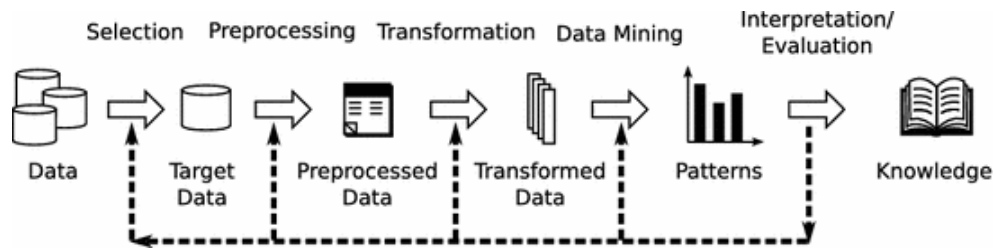


Figure 3.1: Knowledge discovery in database process [2]

3.3 Feature selection

The features in this dataset include the following:

- Demographic data: Age, Gender, Geography
- Customer account information: CustomerId, CreditScore, Tenure, Balance, NumOfProducts, HasCrCard, IsActiveMember, Estimated Salary
- Exited Target is Churn, which has binary classes 1 and 0.

Making a feature selection is important for the predictive model, since it allows us to take important features being relevant for the model, for this, irrelevant attributes are eliminated. Then, The relevant feature are:

- RowNumber, CustomerId, Surname: These are identifiers that every company uses to identify its customers. These records should not influence the bank's customer churn prediction model. These columns will be removed.
- CreditScore: A credit score plays a key role in financial institutions, which use these metrics to calculate the customer's credit score. The higher the score, the better the borrower will look. Customers who have high credit scores will get loans and thus remain customers because they're repaying a loan therefore it's a relevant variable to the study of customer churn.
- Geography: The location of a customers can differentiate the bank's treatment of them. This is because certain banks carry out different strategies depending on the region and this affects customers in one way or another. So, it can be a variable that influences the predictive model.
- Gender: It would be interesting to see the influence of gender in financial institutions and how women are perceived in these institutions, since in many Latin American countries most people perceive men as those who earn a higher salary and are responsible for the household finances. Therefore, it will be interesting to see how gender influences the predictive model.
- Age: Age can certainly be a relevant variable because people who are old enough to be in a stable position to work are less likely to leave the company, unlike young people or elderly people who do not have a stable financial situation.

- **Tenure:** Determines the time that the client is a partner of the financial company, in this case it is determined in years. It is a variable that should show that clients with more years in the company will continue with the company since this generates a level of trust that has been given for years between the company and the client, meaning that the higher the tenure the less likely to leave the company.
- **Balance:** This is a report of the current situation of the client in the financial institution. This variable may reflect the abandonment of customers, since people with a low balance are more likely to leave the company.
- **NumOfProducts:** This variable deals with the number of services that the customer has purchased. Therefore, the higher this variable, the less likely its abandonment could be.
- **HasCrCard:** This is a variable that only indicates if a customer uses the credit card service. Being a binary indicator, it is much easier to handle and determine if it is relevant that the client has a credit card or not when leaving the company.
- **IsActiveMember:** Variable that establishes if the clients have an active behavior with the company. Therefore it would imply that active clients are less likely to leave the bank.
- **Estimated Salary:** Variable closely related to Balance, since people with lower salaries have lower balance. And as explained previously, it could influence the predictive model.
- **Exited:** This is a variable with binary values, and simply here it is simply determined whether or not the client remains in the company. This is what the model has to predict.

3.4 Creation and transformation of variables

In this stage, the data were transformed into another form which is more appropriate for the predictive model, switching them to numerical variables.

On the other hand, we should also explore the type of variables present in the dataset. Categorical variables cannot be handled directly. For this database, we have transformed the gender variable. The variable Gender is categorical. It turns into two binary variables: Male taking the value 1 and Female taking the value 0.

And finally, the data are transformed through feature scaling, which is essential for machine learning algorithms. Feature scaling through standardization is a common requirement for many machine learning algorithms. Standardization involves re-scaling the features such that they have the properties of a standard normal distribution with a mean of zero and a standard deviation of one.

3.5 Classification

The splitting of the dataset into training and test sets is a common process of ML for building a predictive model [42]. The dataset is split into two groups: the training group and the testing group. The training group consists of 70% of the dataset and aims to train the algorithm. The test group contains 30% of the dataset and is used in order to estimate how well the proposed model has been trained.

Training with the same data over and over again creates a problem called overfitting [43]. Scikit-learn [3] explained it as a problem in which a model, which only shows the same input data, is unable to recognize new input data, leading the model to only identify the identical data from a training set as reliable. Another problem we have in our database is bias, which occurs when we have unbalanced data, leading our model to classify minority classes incorrectly. To reduce these two problems were implemented a tool called Cross-Validation (CV).

As explained on the same page in the Scikit-learn library [3], CV prevents the results from depending on a random choice only of a pair of sets (training and test). CV also called K-Fold CV, where the training set is split into k smaller sets, Fig. 3.2b. That is, the same database is used k times as if they were different databases resulting in a less biased estimate as opposed to a simple training and test split. For a better understanding, a flow char and a diagram of the CV-operations in the database are presented Fig 3.2.

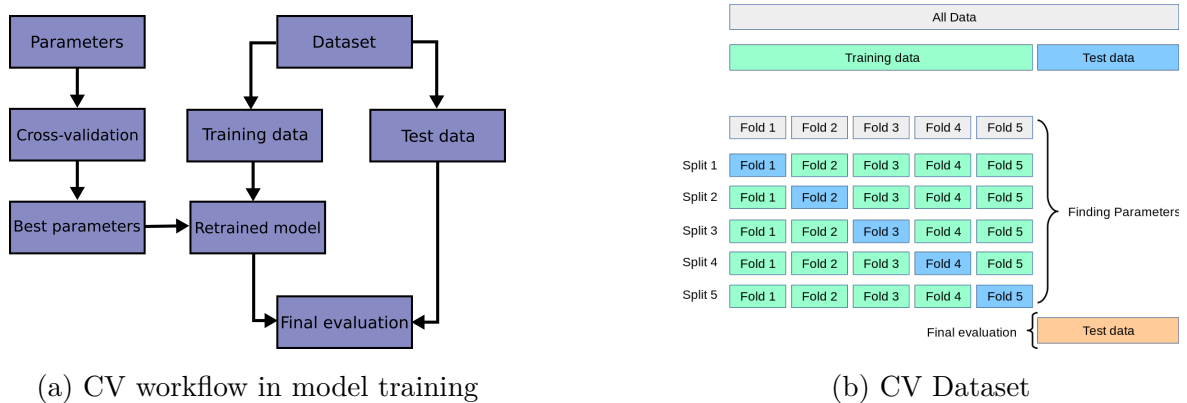


Figure 3.2: CV basic approach [3]

3.6 Evaluation measures

The confusion matrix is usually applied to evaluate the performance of the binary classification or predictive classifiers. It contains the following corresponding measures seen in Table 3.3:

- True Positive (TP): customer churn that has been correctly predicted. The number of customers that should be in the churning category and the prediction algorithm has determined their category correctly as churning

		Actual Class	
		Churn	Non Churn
Predicted Class	Churn	TP	FP
	Non Churn	FN	TN

Table 3.3: Confusion matrix for classifier evaluation

- True Negative (TN): customer non-churn has been correctly predicted. The number of customers that should be in the non-churner category and the prediction algorithm has determined their category correctly as non-churner.
- False Positive (FP): customer non-churn that has been incorrectly predicted to be customer churn. The number of customers who are non-churners but the algorithm incorrectly categorized them as churners, and
- False Negative (FN): customer churn that has been incorrectly predicted to be customer non-churn. The number of customers who are churners but the algorithm incorrectly categorized them as non-churner

Evaluation criteria are important to measure the accuracy of a predictive model by using the confusion matrix [44]. There are well-known criteria for evaluating the accuracy or the performance of a predictor model. These criteria include accuracy, precision, recall, and the F-measure, which can successfully show the performance of the predictor [44] [26]. To evaluate the performance of machine-learning methods in the customer churn prediction, we have used these criteria, which are calculated based on the confusion matrix shown in Table 3.3. These criteria are defined as [45]:

$$Accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)} \quad (3.1)$$

Recall (true positive rate or sensitivity) is the ratio of real churners which are correctly identified, recognizing a positive class instance, and it is calculated:

$$TPR = Recall = \frac{TP}{(TP + FN)} \quad (3.2)$$

Precision is the ratio of predicted churners which are correct, a predicted positive class instance is a true positive, and it is calculated:

$$Precision = \frac{TP}{(TP + FP)} \quad (3.3)$$

False Positive Rate indicates the probability of failure in recognizing a negative class instance, and it is calculated:

$$FPR = \frac{FP}{(TN + FP)} \quad (3.4)$$

F-measure is the harmonic mean of precision and recall, and it is calculated:

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (3.5)$$

Area under the curve (AUC)

AUC is a metric that is used to evaluate the performance of a model in machine learning. The AUC measures the area under the receiver operating characteristics (ROC) curve. This ROC curve allows a comparison of TPR versus FPR. The area of that curve shows the probability that a randomly drawn positive example will have a higher decision function value than a random negative example. The higher the AUC, the more accurate the test will be [46].

3.7 Experimental Setup

To empirically validate the proposed approach, the experiment was conducted by comparing the credibility of the proposed CCP model. The implemented model used the Python and analysis libraries. Its performance is evaluated using the accuracy score, F-measure and ROC_AUC criteria.

Below is a small outline of the steps applied in the model:

- Importing the libraries
- Loading the dataset
- Selecting relevant features
- Converting categorical type to numeric type
- Preprocessing the data
- Training and tuning a machine learning algorithm
- Testing and evaluate the proposed model.

Importing the libraries

To carry out the data analysis, The follows functions was imported:

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns

from sklearn import svm
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split, cross_val_score,
    GridSearchCV, validation_curve, StratifiedKFold
from sklearn.metrics import confusion_matrix, roc_curve, precision_score,
```

```
precision_recall_curve, auc, recall_score, plot_precision_recall_curve
```

Loading the dataset

After the libraries are imported, Panda is used to manage the database. This library loads the .csv file into a set called dataframe:

```
data = pd.read_csv('../Churn%20Modeling.csv')
```

Selecting relevant features

In the previous chapter it was explained which variables influence in the predictive model. So, the irrelevant characteristics of the database were eliminated:

```
data = data.drop(['RowNumber', 'CustomerId', 'Surname'], axis=1)
```

Converting categorical to numeric

Table 3.1 shows two categorical variables: Geography and Gender. They are variables that contain qualitative data but that can be converted to quantitative variables.

The gender variable has 2 categories: male or female. Therefore, it will simply be changed to 0 for female and 1 for male.

The Geography variable has three categories (France, Spain, Germany). We could have done the same as it was done with the gender by placing numerical values such as 0 to France, 1 to Spain and 3 to Germany. But doing it this way, we would have given a certain weight to the categories. So, to avoid this problem, we decided a better alternative would be to delete a category. For example, we removed Spain, so to identify France, we assign a 1 to France and a 0 to Germany, to identify Germany, we assign a 0 to France and a 1 to Germany, and to identify Spain, we assign a 0 to France and a 0 to Germany.

The functions for transforming categorical variables to numerical ones are as follows:

```
data=data.replace({'Male':1, 'Female':0})
Geography = pd.get_dummies(data.Geography, drop_first=True)
data = data.drop(['Geography'], axis=1)
data = data.concat([data, Geography], axis=1)
```

Preprocessing the data

The dependent variables are separate from the independent variable, in order to have the predictor variable isolated from the dependent variables.

```
X = dataset.drop(['Exited'], axis=1)
y = dataset['Exited']
```

The variable X contains all the dependent variables, while Y contains the predictor variable. X and y will be divided into a training and test set. For this example X_{test} and y_{test} will have 30% of the total samples for testing and X_{train} and y_{train} 70% for training. The test samples will be used to evaluate how good the model is.

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y,
                                                    test_size= 0.3, random_state= 0)
```

Feature scaling

To normalize the range of independent variables used the function `StandardScaler`.

```
from sklearn.preprocessing import StandardScaler
scale = StandardScaler()
X_train = scale.fit_transform(X_train)
X_test = scale.transform(X_test)
```

Machine learning algorithm training and evaluation

The model is fed on the previously processed training data. After, the model is ready to predict on the test data:

```
from sklearn import svm
svmR = svm.SVC(kernel='rbf', C=1, gamma=0.1)
svmR.fit(X_train, y_train)
y_pred = svmR.predict(X_test)
```

Evaluate Model Results on Test Set:

```
acc = accuracy_score(y_test, y_pred )
prec = precision_score(y_test, y_pred )
rec = recall_score(y_test, y_pred )
f1 = f1_score(y_test, y_pred )
cm = confusion_matrix(y_test, y_pred)
```


Chapter 4

Results and Discussion

Data correlation

Before building a predictive model, the correlation among the features is identified and sorted in descending order. Fig 4.1 shows the correlation values of the dependent variables with the independent variable. Based on this assumption, it shows clearly that there is a moderate correlation between the attributes Exited and Age.

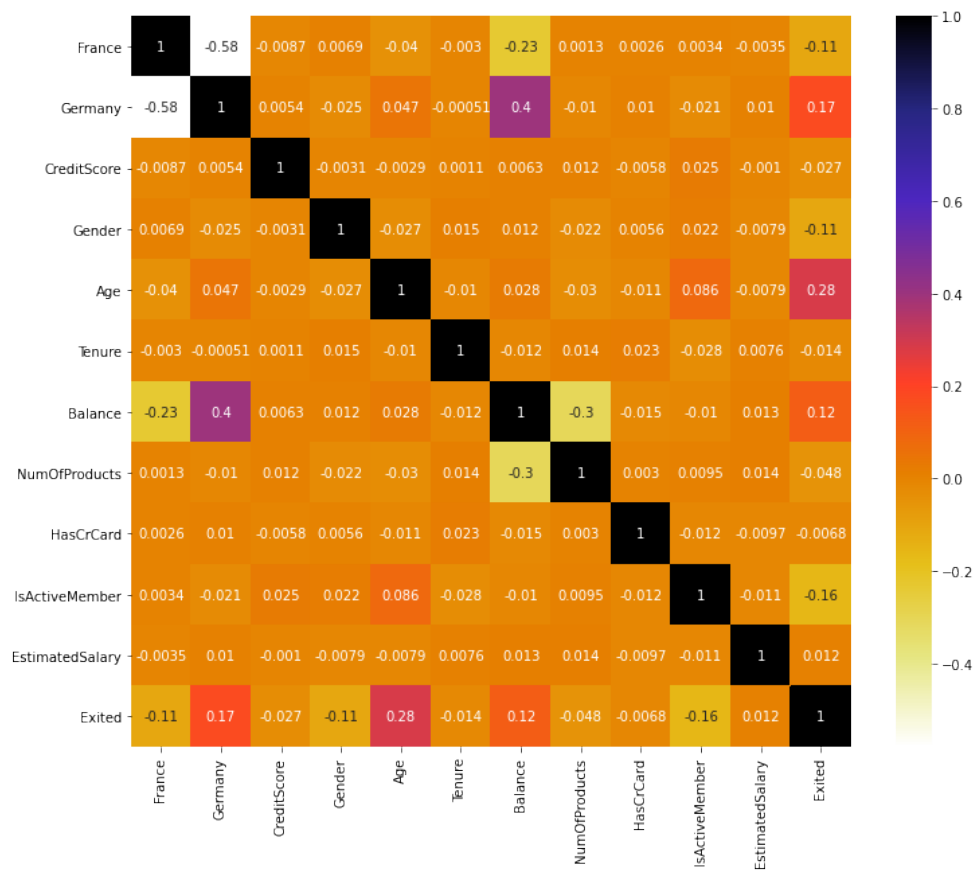


Figure 4.1: Correlation Matrix

Multicollinearity using Variance Inflation Factors (VIF)

Multicollinearity is one of serious problems that should be resolved before starting the process of modeling the data [47]. The dependence between features leads to instability of the model and redundancy of the feature set. The VIF is a tool to measure and quantify how much the variance is inflated.

Kumari in “Multicollinearity : Estimation and Elimination” shows a model to predict a dependent (Y) variable from two or more independent (X) variables using the formula: $VIF = \frac{1}{1-R^2}$. Where R^2 is the correlation coefficient between two variables, and its value falls between -1 and $+1$. When $r = 0$, there is no multicollinearity, and the $VIF = 1$. As R increases in absolute terms, the variances for the estimated coefficients increase too. As R approaches $+1$, the inflation factor approaches infinity [48]. Kleinbaum considers that there are collinearity problems if some IVF are greater than 10 [49]. The results of these analyses are presented in Table 4.2.

	variables	VIF		variables	VIF
0	CreditScore	20.990014	0	Gender	2.119203
1	Gender	2.165341	1	Age	9.247795
2	Age	12.295280	2	Tenure	3.697013
3	Tenure	3.871195	3	Balance	2.445583
4	Balance	2.628334	4	NumOfProducts	6.311506
5	NumOfProducts	7.708681	5	HasCrCard	3.166297
6	HasCrCard	3.289272	6	IsActiveMember	2.049454
7	IsActiveMember	2.074279	7	EstimatedSalary	3.713499
8	EstimatedSalary	3.886482			

(a) Before VIF

(b) After VIF

Figure 4.2: Multicollinearity using VIF

Fig. 4.2a shows that the “CreditScore” and “Age” have $VIF > 10$. By dropping the variable “CreditScore” the multicollinearity decreases in all other variables, Fig 4.2b. New VIF values are below 10. Thus, the measures selected for assessing independent variables in this study do not reach levels that indicate multicollinearity.

Churn Rate Data Analysis

A preliminary look at the overall churn rate shows that around 79.6% of the customers are “Non-Churn” and 20.4% “Churn”. As shown in Fig. 4.3, this is an imbalanced classification problem. In extreme cases, model may classify all instances like majority class, resulting in overall high accuracy but unacceptably low precision with respect to the minority class [33]. This could lead to a lot of false negatives.

Then, to better understand the patterns in the data, the predictor variable (Churn) will be compared with all the other variables.

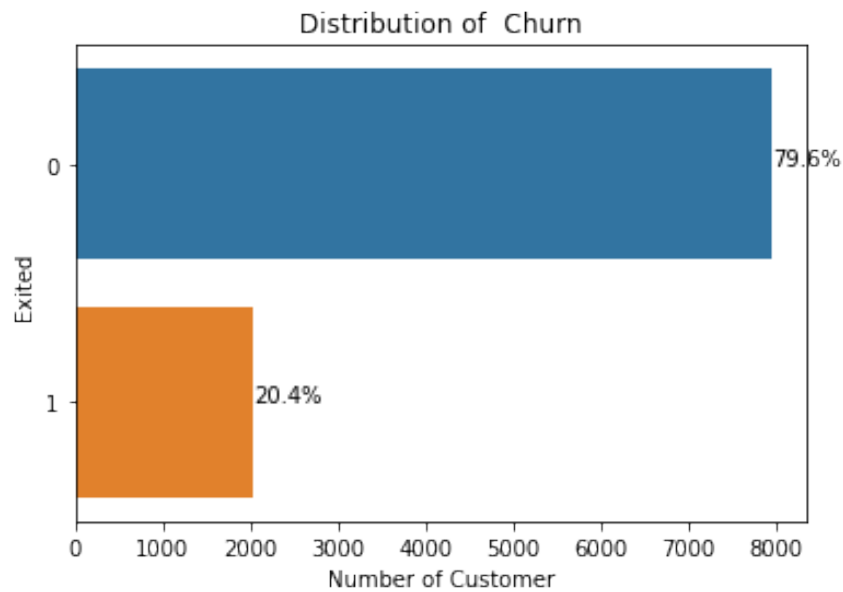
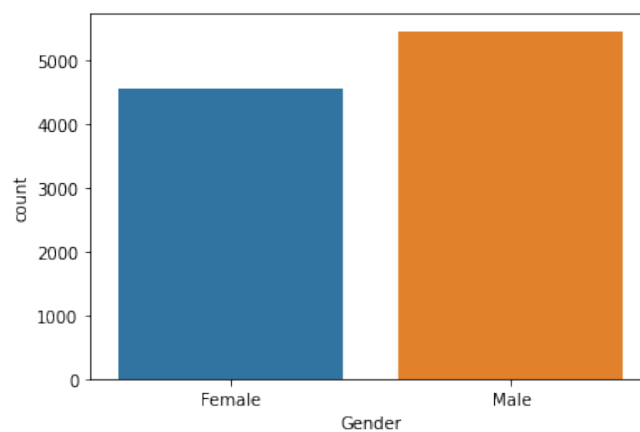


Figure 4.3: Churn Distribution

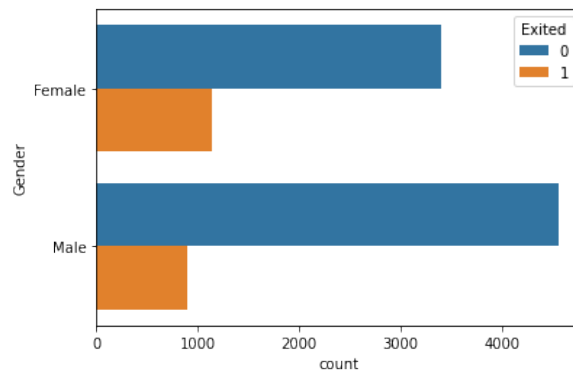
A.) Demographics: Gender, Age and Geography of the customers will be compared with regard to the predictor variable “Churn”.

Gender Distribution: Customers’ gender is a balanced variable, Fig. 4.4a. Female customers tend to leave the bank more than male, Fig.4.4b. So, female customers are more likely to churn than male customers, with the difference ($\approx 10\%$). A quarter of all of the churned customers in dataset are female while the other 15% are male, Fig.4.4c.

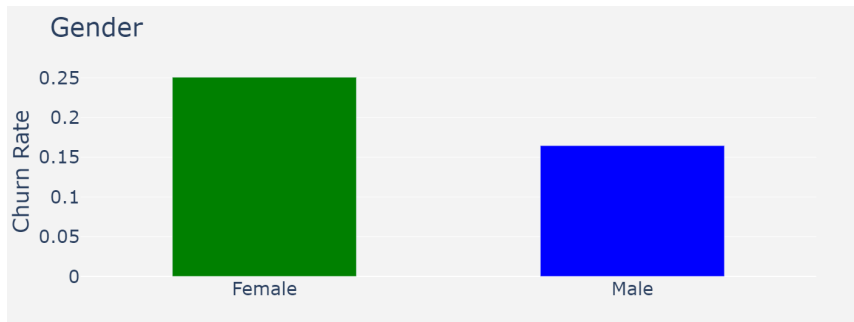


(a) Number of Male and Female

Age Distribution: Most of the clients are between 30 and 40 year old, Fig. 4.5a. And the customers most likely to churn are around 50 years old, Fig. 4.5b.

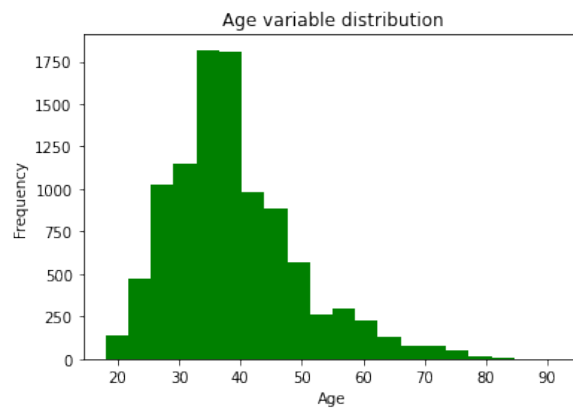


(b) Churn and Non-Churn by Gender

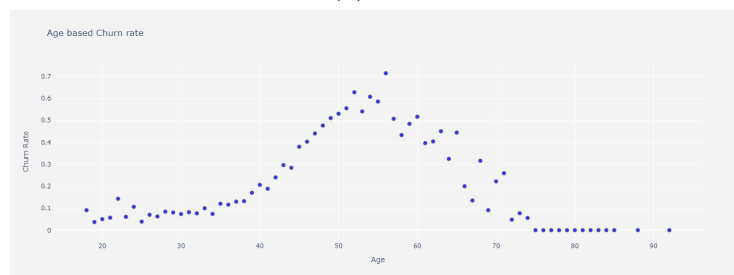


(c) Churn Rate by Gender

Figure 4.4: Gender Distribution



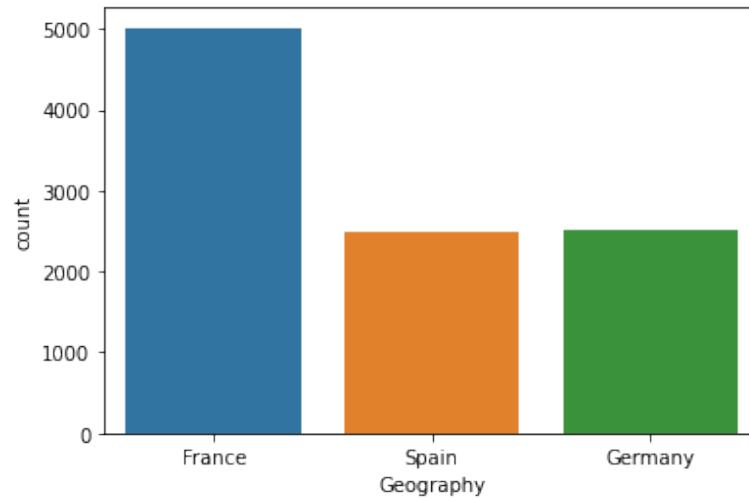
(a) Age



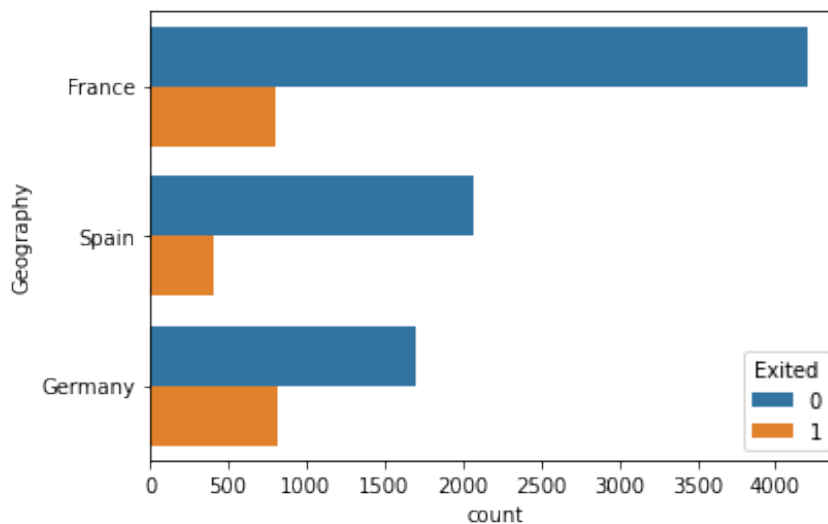
(b) Rate Churn By Age

Figure 4.5: Age Distribution

Geography Distribution: Customers mainly live in France, while the number of customers living in Germany and Spain is very close, Fig. 4.6a. Average loss of customers is highest in Germany, Fig. 4.6b.



(a) No. Customer by Geography



(b) Churn and Non-Churn by Geography

Figure 4.6: Geography Distribution

B.) Customer Account Information: To see the relation between Tenure and average Churn Rate a scatter plot was built, Fig. 4.7. A lot of customers have been with the company for just a year, while quite a many have been there for about 2-8 years.

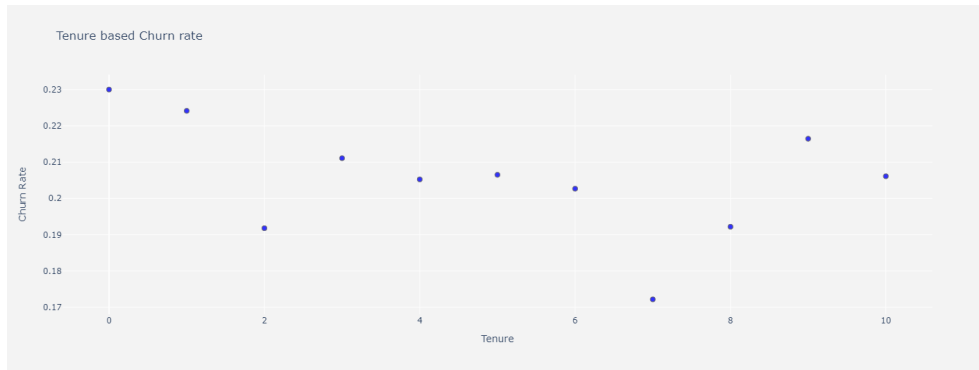
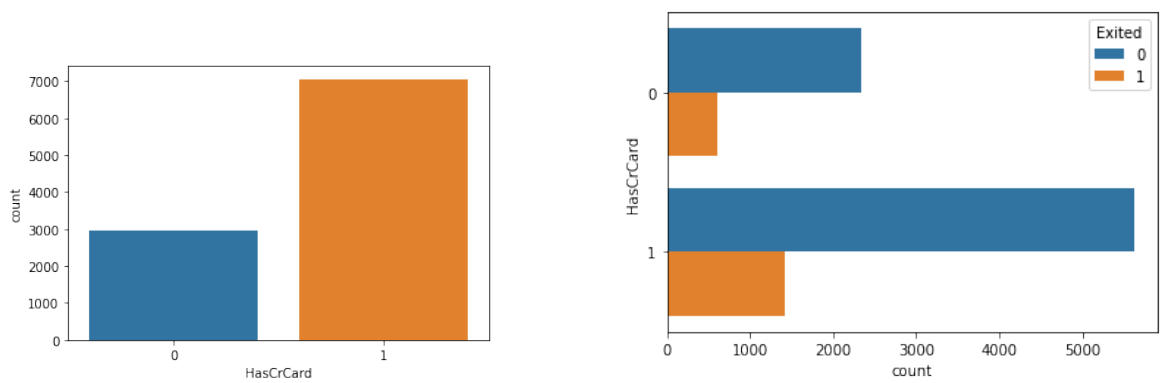


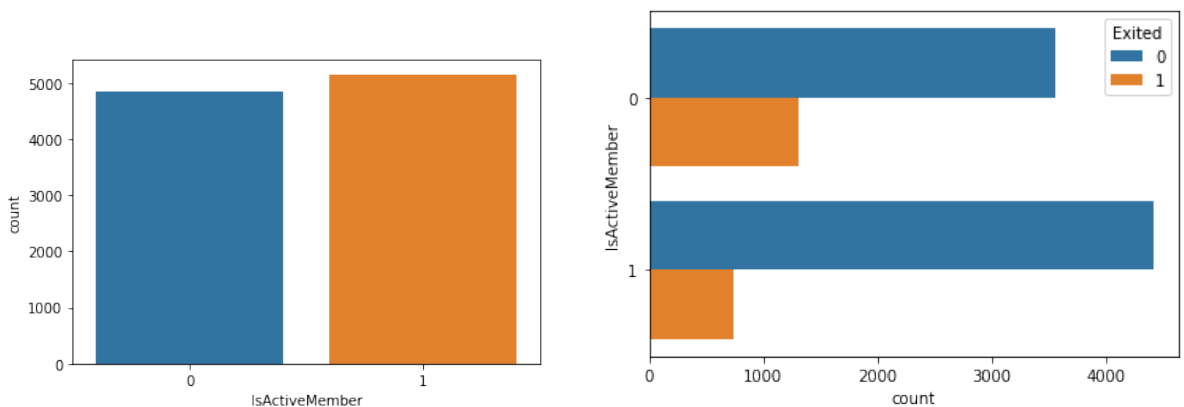
Figure 4.7: Churn Rate by Tenure

Another aspect of staying in the bank could be for being an active member and having a credit card, Fig. 4.8. Customers who do not actively use the bank tends to leave the bank, Fig. 4.8d.



(a) No. Customer has Credit Card

(b) Churn and Non-Churn by Credit Card



(c) No. Customer are Active Member

(d) Churn and Non-Churn by Active Member

Figure 4.8: Credit Card and Active Member Distribution

Machine Learning Algorithm Evaluation

As previously established, the metrics to be used to evaluate the performance of the model are the F1 MEASURE, PRECISION, RECALL, AND ACCURACY.

Table 4.1 shows the values of the confusion matrix and the performance measures for the best performances of the algorithms after adjusting their respective parameters. The best performance is shown by the Support Vector Machine adjusted with Polynomial kernel and parameters $\gamma = 0.02$ and $c = 62$. The second best performance is shown by the Support Vector Machine adjusted with RBF kernel and parameters $\gamma = 0.02$ and $c = 36$.

MODELS	TP	FP	FN	TN	Acc	Prs	Rec	F-m
SVM Linear	0	0	509	1991	0.80	0.80	1.00	0.89
						0	0	0
SVM Rbf	197	37	312	1954	0.86	0.86	0.98	0.92
						0.84	0.39	0.53
SVM Poly	209	39	300	1952	0.86	0.87	0.98	0.92
						0.84	0.41	0.55

Table 4.1: Acc: Accuracy, Prs: Precision, Rec: Recall, F-m: F-measure

Accuracy is one of the most common classification evaluation metrics to compare algorithms, since it is the number of correct predictions made over the total predictions. However, it is not the ideal metric when the database is imbalance. Therefore, To evaluate the model, we will use another metric called ROC_AUC curve. This metric discriminates between positive and negative classes.

From confusion matrix, the false positive rate (FPR) was obtained, and the true positive rate (TPR) for the three models. Then, The ROC curve were plotted as shown in Fig 4.9. This plot shows that both SVM-Poly and SVM-Rbf better performace.

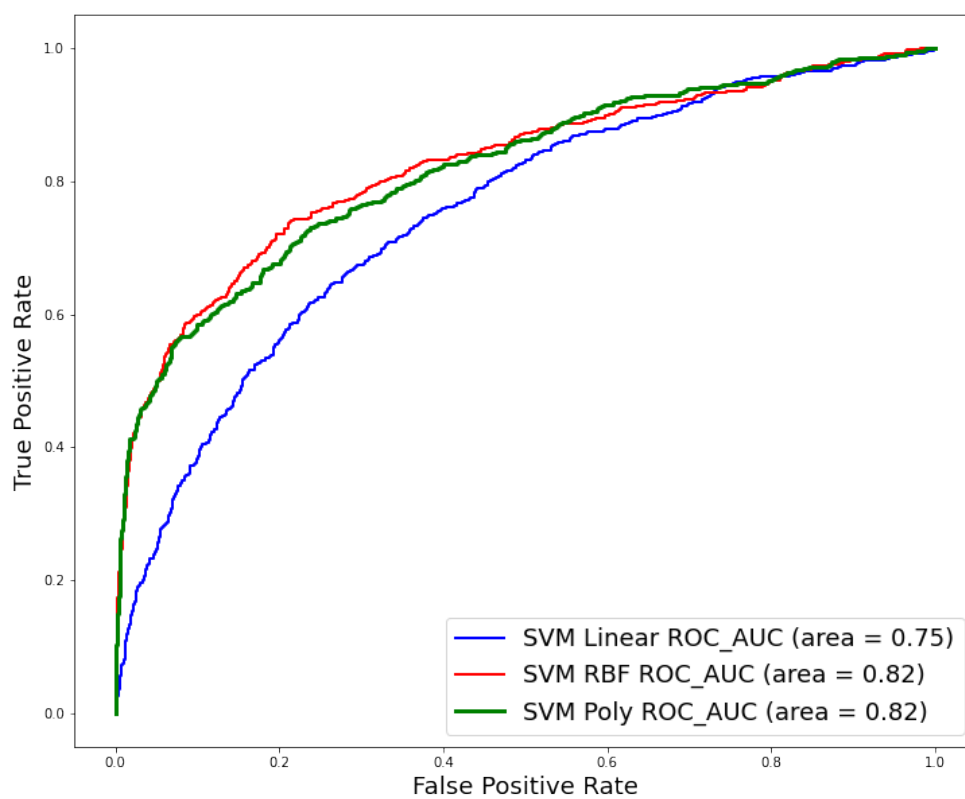


Figure 4.9: Roc Auc

Chapter 5

Conclusions

A comparison study has been made by using three different kernels: linear, radial basis function and polynomial. The three models are used for predicting customer churn in a financial institution. They show how to handle higher dimensional and linearly inseparable data. From the three models, polynomial and RBF show better performance according to the ROC_AUC metric. It also is supported by confusion matrix having relatively the highest combination of precision, recall and F1 scores in the two models giving the highest number of correct positive predictions while minimizing the false negatives.

Choosing the tuning parameters is not an easy task because their effect on the model is hard to see. Many experiments with different parameters need to be done to determine the best parameters for each model. For example, SVM-linear takes a lot of time for training with larger c parameters and the same happens with SVM-Poly when the degree of the polynomial is increased. Thus, it is hard to chose the final model since the changes are minimal from one to another.

A recommendation to solve the previous problems would be to have a graphics card. The computing power was a limitation, for which it would be good to have a graphics card that allows more experiments in the tuning parameters and thus obtain better results, as well as being able to accelerate data processing using parallel programmin

For future studies, real data from a financial institution could be used, since the database that was used for the development of this thesis was obtained from a platform. The platform may have modified the data, in addition to having limited variables.

The metrics selected to evaluate the best model are suggested in several investigations for the prediction of customer churn. It would be good to investigate alternative metrics, which could yield a more efficient model in the prediction, as well as to investigate how to deal with an imbalanced database.

Finally the implementation of machine learning algorithm by using SVM will allow companies to identify those customers with the highest probability of churn, and thus establish actions for their retention.

Bibliography

- [1] Z. Plaksij, “Customer churn: 12 ways to stop churn immediately,” April 2020. [Online]. Available: <https://www.superoffice.com/blog/reduce-customer-churn/>
- [2] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “The kdd process for extracting useful knowledge from volumes of data,” *Communications of the ACM*, vol. 39, no. 11, pp. 27–34, 1996.
- [3] “3.1. cross-validation: evaluating estimator performance.” [Online]. Available: https://scikit-learn.org/stable/modules/cross_validation.html
- [4] Y.-H. Tao and C.-C. R. Yeh, “Simple database marketing tools in customer analysis and retention,” *International Journal of Information Management*, vol. 23, no. 4, pp. 291–301, 2003.
- [5] A. De Caigny, K. Coussement, and K. W. De Bock, “A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees,” *European Journal of Operational Research*, vol. 269, no. 2, pp. 760–772, 2018.
- [6] J. Burez and D. Van den Poel, “Handling class imbalance in customer churn prediction,” *Expert Systems with Applications*, vol. 36, no. 3, pp. 4626–4636, 2009.
- [7] A. Lemmens and C. Croux, “Bagging and boosting classification trees to predict churn,” *Journal of Marketing Research*, vol. 43, no. 2, pp. 276–286, 2006.
- [8] S. A. Neslin, S. Gupta, W. Kamakura, J. Lu, and C. H. Mason, “Defection detection: Measuring and understanding the predictive accuracy of customer churn models,” *Journal of marketing research*, vol. 43, no. 2, pp. 204–211, 2006.
- [9] P. Datta, B. Masand, D. R. Mani, and B. Li, “Automated cellular modeling and prediction on a large scale,” *Artificial Intelligence Review*, vol. 14, no. 6, pp. 485–502, 2000.
- [10] C.-P. Wei and I.-T. Chiu, “Turning telecommunications call details to churn prediction: a data mining approach,” *Expert systems with applications*, vol. 23, no. 2, pp. 103–112, 2002.
- [11] S.-Y. Hung, D. C. Yen, and H.-Y. Wang, “Applying data mining to telecom churn management,” *Expert Systems with Applications*, vol. 31, no. 3, pp. 515–524, 2006.

- [12] M. Hassouna, A. Tarhini, T. Elyas, and M. S. AbouTrab, "Customer churn in mobile markets a comparison of techniques," *arXiv preprint arXiv:1607.07792*, 2016.
- [13] M. Chandar, A. Laha, and P. Krishna, "Modeling churn behavior of bank customers using predictive data mining techniques," in *National conference on soft computing techniques for engineering applications (SCT-2006)*, 2006, pp. 24–26.
- [14] N. Lu, H. Lin, J. Lu, and G. Zhang, "A customer churn prediction model in telecom industry using boosting," *IEEE Transactions on Industrial Informatics*, vol. 10, no. 2, pp. 1659–1665, 2012.
- [15] M. C. Mozer, R. Wolniewicz, D. B. Grimes, E. Johnson, and H. Kaushansky, "Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry," *IEEE Transactions on neural networks*, vol. 11, no. 3, pp. 690–696, 2000.
- [16] C. Bhattacharya, "When customers are members: Customer retention in paid membership contexts," *Journal of the academy of marketing science*, vol. 26, no. 1, pp. 31–44, 1998.
- [17] J. Hadden, A. Tiwari, R. Roy, and D. Ruta, "Computer assisted customer churn management: State-of-the-art and future trends," *Computers & Operations Research*, vol. 34, no. 10, pp. 2902–2917, 2007.
- [18] H. Hwang, T. Jung, and E. Suh, "An ltv model and customer segmentation based on customer value: a case study on the wireless telecommunication industry," *Expert systems with applications*, vol. 26, no. 2, pp. 181–188, 2004.
- [19] M. Galetto, "What is customer churn?" May 2016. [Online]. Available: <https://www.ngdata.com/what-is-customer-churn/>
- [20] Reliabilis, "Churn rate vs retention rate – comparison guide," Jul 2020. [Online]. Available: <https://www.reliabilis.com/blog/churn-rate-vs-retention-rate/>
- [21] Startupdonut, "Understanding customer retention and churn," Jun 2018. [Online]. Available: <https://www.startupdonut.co.uk/sales-and-marketing/looking-after-your-customers/understanding-customer-retention-and-churn>
- [22] W. Buckinx and D. Van den Poel, "Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual fmcg retail setting," *European journal of operational research*, vol. 164, no. 1, pp. 252–268, 2005.
- [23] S. Khodabandehlou and M. Z. Rahman, "Comparison of supervised machine learning techniques for customer churn prediction based on analysis of customer behavior," *Journal of Systems and Information Technology*, 2017.
- [24] A. Keramati, H. Ghaneei, and S. M. Mirmohammadi, "Developing a prediction model for customer churn from electronic banking services using data mining," *Financial Innovation*, vol. 2, no. 1, pp. 1–13, 2016.

- [25] T. Vafeiadis, K. I. Diamantaras, G. Sarigiannidis, and K. C. Chatzisavvas, “A comparison of machine learning techniques for customer churn prediction,” *Simulation Modelling Practice and Theory*, vol. 55, pp. 1–9, 2015.
- [26] A. Keramati, R. Jafari-Marandi, M. Aliannejadi, I. Ahmadian, M. Mozaffari, and U. Abbasi, “Improved churn prediction in telecommunication industry using data mining techniques,” *Applied Soft Computing*, vol. 24, pp. 994–1012, 2014.
- [27] V. Vapnik, I. Guyon, and T. Hastie, “Support vector machines,” *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.
- [28] W. S. Noble, “What is a support vector machine?” *Nature biotechnology*, vol. 24, no. 12, pp. 1565–1567, 2006.
- [29] J. A. Suykens and J. Vandewalle, “Least squares support vector machine classifiers,” *Neural processing letters*, vol. 9, no. 3, pp. 293–300, 1999.
- [30] B. Huang, M. T. Kechadi, and B. Buckley, “Customer churn prediction in telecommunications,” *Expert Systems with Applications*, vol. 39, no. 1, pp. 1414–1425, 2012.
- [31] B. He, Y. Shi, Q. Wan, and X. Zhao, “Prediction of customer attrition of commercial banks based on svm model,” *Procedia Computer Science*, vol. 31, pp. 423–430, 2014.
- [32] L. Bin, S. Peiji, and L. Juan, “Customer churn prediction based on the decision tree in personal handyphone system service,” in *2007 International Conference on Service Systems and Service Management*. IEEE, 2007, pp. 1–5.
- [33] B. Zhu, G. Xie, Y. Yuan, and Y. Duan, “Investigating decision tree in churn prediction with class imbalance,” in *Proceedings of the International Conference on Data Processing and Applications*, 2018, pp. 11–15.
- [34] N. Cristianini, J. Shawe-Taylor *et al.*, *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.
- [35] C. J. Burges, “A tutorial on support vector machines for pattern recognition,” *Data mining and knowledge discovery*, vol. 2, no. 2, pp. 121–167, 1998.
- [36] N. Cristianini, “Support vector and kernel methods for pattern recognition,” 2003.
- [37] F. Melgani and L. Bruzzone, “Classification of hyperspectral remote sensing images with support vector machines,” *IEEE Transactions on geoscience and remote sensing*, vol. 42, no. 8, pp. 1778–1790, 2004.
- [38] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty, *Nonlinear programming: theory and algorithms*. John Wiley & Sons, 2013.
- [39] R. Fletcher, *Practical methods of optimization*. John Wiley & Sons, 2013.
- [40] A. Mammone, M. Turchi, and N. Cristianini, “Support vector machines,” *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 1, no. 3, pp. 283–289, 2009.

- [41] J. Han, M. Kamber, and J. Pei, "Data mining concepts and techniques third edition," *The Morgan Kaufmann Series in Data Management Systems*, vol. 5, no. 4, pp. 83–124, 2011.
- [42] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, "Practical machine learning tools and techniques," *Morgan Kaufmann*, p. 578, 2005.
- [43] A. W. Moore, "Cross-validation for detecting and preventing overfitting," *School of Computer Science Carnegie Mellon University*, 2001.
- [44] K. Coussement and K. W. De Bock, "Customer churn prediction in the online gambling industry: The beneficial effect of ensemble learning," *Journal of Business Research*, vol. 66, no. 9, pp. 1629–1636, 2013.
- [45] A. Maratea, A. Petrosino, and M. Manzo, "Adjusted f-measure and kernel scaling for imbalanced data learning," *Information Sciences*, vol. 257, pp. 331–341, 2014.
- [46] A. Mishra, "Metrics to evaluate your machine learning algorithm," *Towards Data Science*, 2018.
- [47] J. I. Daoud, "Multicollinearity and regression analysis," in *Journal of Physics: Conference Series*, vol. 949, no. 1. IOP Publishing, 2017, p. 012009.
- [48] S. Kumari, "Multicollinearity: Estimation and elimination," *Journal of Contemporary research in Management*, vol. 3, no. 1, pp. 87–95, 2008.
- [49] W. Yoo, R. Mayberry, S. Bae, K. Singh, Q. P. He, and J. W. Lillard Jr, "A study of effects of multicollinearity in the multivariable analysis," *International journal of applied science and technology*, vol. 4, no. 5, p. 9, 2014.