Bachelor Thesis in Computer Science
June 2022

Comparison of Machine learning algorithms on Predicting Churn within Music streaming service

Lahari Gaddam
Sree Lakshmi Hiranmayee Kadali

Faculty of Computing, Blekinge Institute of Technology, 371 79 Karlskrona, Sweden

This thesis is submitted to the Faculty of Computing at Blekinge Institute of Technology in partial fulfilment of the requirements for the degree of Bachelor of Computer science. The thesis is equivalent to 10 weeks of full time studies.

The authors declare that they are the sole authors of this thesis and that they have not used any sources other than those listed in the bibliography and identified as references. They further declare that they have not submitted this thesis at any other institution to obtain a degree.

Contact Information:
Author(s):
Lahari Gaddam
E-mail: laga21@student.bth.se

Sree Lakshmi Hiranmayee Kadali
E-mail: srkd21@student.bth.se

University advisor:
Dr. Shahrooz Abghari
Department of Computer Science

Faculty of Computing
Blekinge Institute of Technology
SE–371 79 Karlskrona, Sweden

Internet : www.bth.se
Phone : +46 455 38 50 00
Fax : +46 455 38 50 57

Abstract

Background: Customer churn prediction is one of the most popular part of big businesses and often help the companies in customer retention and revenue generation.Customer churn may lead to huge loss of revenue and is important to analyze and determine the cause for churn. Moreover, it is easier to retain an existing customer rather than acquiring new clients.Therefore, to get a better understanding on churn prediction, this research work focuses on finding the best performing machine learning model after effective comparision among four machine learning models. The research also gives a brief report of latest literature work done in churn analysis of music streaming services.

Objectives: In this thesis work, we aim to research about churn prediction done in music streaming services. We focus on two main objectives, first one includes literature review on the latest research work done in churn prediction of music streaming services. Secondly, we aim in comparing the performance of four supervised machine learning algorithms, to find out the best performing algorithm for churn prediction.

Methods: This thesis involves two methods literature review and experimentation to answer our research questions. We chose to use literature review for RQ1 so it can give a better understanding on our selected problem and works as base work for our research and helps in clear and better comprehension. Experimentation is chosen for RQ2 to to build and train the selected machine learning model to validate the performance of algorithms. Experimentation is chosen because it gives better results and prediction compared to surveys and reviews.

Results: We have selected four classification supervised machine learning algorithms namely, Logistic regression, Naive Bayes, K-nearest neighbors, and Random forest in this research. Upon experimentation and training the models using the algorithms with a preprocessing the KKBox's dataset, Random forest achieved highest accuracy of 97% compared to other models.

Conclusions: We have trained four models using the four machine learning algorithms for the prediction of churn in music streaming service domain. Upon training the models with the KKBox's dataset and upon experimentation, we came to a conclusion that random forest has the best performance with better accuracy and AUC score.

Keywords: Churn Prediction, Classification Algorithms, Machine Learning, Music Stream, Supervised Learning.

Acknowledgments

We are extremely grateful for our supervisor Dr. Shahrooz Abghari for helping and meeting us throughout our thesis work. We sincerely thank our supervisor, friends and family who have been supportive throughout our research work.

Authors
Lahari Gaddam
Sree Lakshmi Hiranmayee Kadali

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In the generation of digital services, most of the growing businesses are using customer churn metrics as their key metrics to predict whether the customer is going to continue with the company or not. Customer churn plays a significant role in the world of digital services. There are a large number of companies that offers services to the customers. Customers can choose any company or the service that they are interested in. If the customer is not satisfied with the services provided by the company, the customer may cancel his/her subscription. Churn is nothing but the percentage of the customers who discontinued the services from a company or a stream at a particular period.

Churning is a process where customers stop using the services or cancel subscriptions provided by a company. Customer Relationship Management (CRM), is the strategy that is used for establishing, managing and enhancing long-term customer connections. The main goal is to identify signals of customer churn earlier [33]. Any minor change in the level of customer retention can lead to a major change in the shares or the profits of the companies. The company should try to maintain customers by satisfying their expectations through their services. This makes customers happy, and they will continue using the services. Companies need to accurately identify the percent of churn customers and non-churn customers as, wrong prediction might lead to a major loss in the company's profits. It's a waste of resources for the organization, if the company, identifies a non-churn customer as a churned customer. Here the customer may not leave the company but because of the wrong identification, the churn rate will raise. To overcome such problems, companies need to use highly accurate methods to identify customer churn so that they can avoid the risk of future churning [17].

Customer churn can be managed by two approaches. These two approaches are reactive and proactive. When the company receives a request from the customer to cancel the service then this approach is said to be a reactive approach. By this, the company offers an incentive to the customer to continue the services provided by them. The other approach that companies adapt to support the customers is the proactive approach. Using the proactive approach, most of the companies predict the customer churn by applying various accurate machine learning algorithms. Here, the company strives to identify the customers who are thinking of churning before they stop the services from the company. Through this, companies come up with special incentives to keep the customers from churning away from the services of the

company [30].

Customer churn can be predicted by using various machine learning algorithms. Machine learning (ML) is a segment of Artificial intelligence (AI). It helps an application or a software to predict accurate results without any explicit programming. Here various models are built and trained with data for an accurate prediction as output [10]. The machine learning approaches are of three types. The approaches are supervised, semi-supervised and unsupervised [1]. In Supervised learning approach, the training data is labelled that is every input has a tagged output. There are two categories of supervised learning approaches. They are Regression and Classification. In this project we decided to use classification supervised learning approaches like Naive Bayes, Logistic Regression, K-Nearest Neighbors, and Random Forest. These algorithms are trained with a music data set and then we will find the accuracy of each of the algorithm for comparison. We chose a public music data set "KKBOX" from Kaggle. The most accurate algorithm is used to predict the churn percentage of a company [25].

1.1 Aim and Objectives

1.1.1 Aim

The main aim of this project involves building churn predictive model for a music streaming data through machine learning using Random forest (RF), K-Nearest Neighbors (KNN), Logistic regression and Naive Bayes classifier. Later the performance of algorithms is formulated based on performance metrics like F1,precision, accuracy and recall and propose the best algorithm churn prediction based on accuracy. The models stability is observed by ROC/AUC curves respectively. The data is analysed by exploratory data analysis (EDA).

1.1.2 Objectives

1. To perform literature review on churn prediction algorithms done on music streaming service domain and giving a brief report.
2. To study the given data set through EDA to discover patterns and visualize features.
3. To apply and build selected machine learning algorithms to train, test and evaluate the best performing among the selected algorithms. We have considered Random forest, K-Nearest Neighbors, Logistic regression, and naive Bayes classifier for our thesis work.
4. To select best performing algorithm among the selected algorithms by comparing performance metrics specifically accuracy for churn prediction in music streaming service.

1.2 Research Questions

RQ1: What is the latest research work, and churn prediction techniques used in music streaming services?
Method opted: Literature review.
Justification: There are different domains in which churn prediction is surveyed on and several combinations of machine learning algorithms used. As our research paper is focused on music streaming service domain, literature review works as base work for our research and helps in clear and better comprehension. We would like to perform literature review on existing machine learning algorithms used for churn prediction in our selected domain i.e music streaming service since this domain is very less researched on compared to telecommunication domain. Our main aim from this research question is to give a brief report based on performance evaluation of churn prediction algorithms used in our selected domain.

RQ2:In terms of performance, which machine learning algorithm is the most accurate for the churn prediction in music streaming domain?
Method opted: Experimentation.
Justification: We have chosen experimentation as our approach to build and train the selected machine learning model to validate the performance of algorithms. Experimentation is chosen because it gives better results and prediction compared to surveys and reviews. It helps in developing an approach and predicting best results for our prediction. Here, our main goal is to select an algorithm among the selected, to find out the most accurate and robust algorithm for churn prediction.

1.3 Outline of Thesis

The structure of our thesis work is given in this section. The first chapter includes introduction of our thesis along with Aim and Objectives. We conclude the first chapter with two research questions, we would like to research on. The next chapter consists of background work including several topics we used in this thesis, giving a brief discussion on selected machine learning algorithms. Third chapter involves all the related work done in churn analysis citing all the papers which helped us in selecting our research work. Chapter four includes methodology of the study from literature review, data analysis, to coming up with model building using selected algorithms. In chapter five we cover the results of our experimentation. Chapter six and seven followed by summary, conclusion and future work respectively.

Chapter 2

Background work

2.1 Churn Prediction

The term churn is often used to describe the tendency of customers who want to stop doing business with the service provider or the company. For the companies, that earn by providing services to customers, it is very crucial to prevent customer from churning the service. Churning traditionally originated from Customer Relation Management (CRM). Churn prediction is used in the fields of games, management, and internet services. Since,it is used intensively in various domains,there is a big difference between definition and utilization. In this thesis, we focus on one single domain i.e music streaming to explain the churn study and predictive models used in previous research work.

2.2 Motivation for Churn Prediction

As Customer retention costs less when compared to gaining new customers to a company. By analysing various studies we can say that gaining a new customer costs nearly four to twenty five times more of retaining a customer who is already a previous user of that company [4]. Just by minimizing the percentage of churn by 5%, the profits of a company could get increased up to 75% [11]. We can see that retention of customer has more impact on companies profit than gaining new customers. The customers who churns the company are divided into two categories. They are

- Voluntary Churners
- Non-voluntary Churners

Identifying non-voluntary churners are easy when compared to voluntary churners. Because non-voluntary churners are those churned by the company itself due to random reasons like misuse of the service or not paying to the service in the right time. Voluntary churners are difficult to identify as the customers consciously terminate the service from the company. Voluntary churning can be categorised into two, Incidental churning and Deliberate churning. Incidental churning is a churning where the services is terminates due to changes in situations where the customer could not continue with the service. For example, the financial situation of the customers might not allow them to continue with the service and results in churning. The geographical location of the customer can also become a reason for incidental churning

if the company might not be able to provide the service in that particular location. Deliberate churning is the most important and difficult as the customer terminates the service due to various reasons. It can be finding a company that provides better services with better coverage and price, technological issues where a competitor company provides latest technology whereas the current provider does not or, when the customers probably experience bad feedback from the call centers. [13].

2.3 Machine Learning Models

Few concepts we work in this thesis are discussed below:

Machine learning: Machine learning is one the fastest growing sector of computer science engineering. It is a sub field of computer science and a branch of artificial intelligence which has wide range of applications. Machine learning is an automated way of detecting patterns in huge data. Since, analysis of large data by humans is more prone to errors, data using machine learning can help in finding new patterns and better insights in real life. The curiosity of finding hidden patterns in data lead for the development of a taxonomy of machine learning algorithms which retrieve data according to the respective algorithm. Types of machine learning include:

Supervised learning: Supervised machine learning includes supervision. The main goal of supervised learning is to map input variable to the output variable. The technique involves training the machine using labelled data-set. A supervised learning algorithm then infers a function based on training data set and then it can be used to map new examples, which predicts the "output". It includes two types, classification and regression.

Unsupervised learning: As the name suggests unsupervised machine learning predicts the output without any supervision. The machine is given an unlabelled data-set and is focused to search for hidden patterns, similarities and differences. It consists of clustering and association.

Semi Supervised learning: To deal with the drawbacks of supervised and unsupervised machine learning, the concept of semi supervised learning is introduced. The main aim is to effectively use available data which is a combination of both labelled and unlabelled data and predict the output.

Reinforcement learning: Reinforcement learning has a similar human being like approach which has an AI agent which explore its surrounding by hitting/trail, taking action, learning experiences and finally improving its performance by experience. It works on a feedback-based technique. It is employed in few fields like information technology, game theory, operation research etc.

2.3.1 Supervised machine learning models:

Supervised learning is part of machine learning taxonomy which is most popular among classification problems. Supervised learning approaches the data by mapping the input data into desired outputs. The learning approach is to learn from the given set of data i.e initially the model is trained with training set. Later, the model is tested with test data in the testing phase. Supervised learning primary goal is to get the computer learn our classification system and find out better insights from the test data given. Few approaches in supervised learning is as follows, Linear classifiers, Artificial neural networks, support vector classifier, K- means, decision trees etc.

Our thesis deals with the following supervised machine learning algorithms:

2.3.1.1 Random Forest

Random forest is a decision tree method which helps in both regression and classification.

Approach: Random Forest decision is made by a method named random subspace method. This method finalises the decision by a final decision tree which is based on the net average weights. The method has the ability to handle missing values inside the data set and helps in preventing the deletion when working with thousand of input values. The name random forest comes from the behaviour of selecting random attributes for decision.

Advantages: It can be an excellent selective algorithm while working with categorical data. Random forest works good in the case of classification. It performs well with large data set handles missing variable without variable deletion which can be helpful in case of churn prediction. It works well with noise and outliers and helps in better prediction. This algorithm is stable compared to all the other algorithms.

Disadvantages: The algorithm is little complex to work with because it requires more computational time and more resources.

2.3.1.2 K-Nearest Neighbor:

K-nearest neighbor is an non parametric technique which approaches a lazy learning method. K-nearest neighbor is an attractive approach while working problems of classification and regression. Since churn prediction is a classification problem, K-nearest neighbor can be one of the best suitable model for this thesis.

Approach: K-nearest neighbor initially stores all the available data and doesn't make any assumption out of the underlying data. And then when new data is given it is classified based on the similarity present in available cases. New cases are easily classifies into available category.

Advantages: 1.It is very easier to work with and doesn't need any prior knowledge about the structure of the data.

2.The performance is asymptotically better compared to naive Bayes.

Disadvantages: 1.The computation cost is high while working with large data for calculating the K point.
2.When the data is not preprocessed and has huge amounts of noise and missing values, the algorithm can turn imprecise and inefficient.

2.3.1.3 Naive Bayes

Based on the background research, we have found out that naive Bayes is very simple and easy to implement while classification.

Approach : Naive Bayes classifier is a probabilistic model in machine learning which assumes the correlation of a particular feature in a class with a presence of class variable. Naive Bayes is based on Bayes theorem which is "naive" to assumed the conditional dependency of the features.

Advantages : While considering the data which is continuous, Naive Bayes work well in finding the likelihood among the data present. Since, churn prediction is all about finding the pattern of how customers can churn, we found Naive Bayes suitable for it. It takes very less storage while working with this algorithm. Naive Bayes is one the best machine learning algorithms to work with. It is the fastest compared to all other sophisticated models.

Disadvantages : 1)Naive Bayes approaches every variable, thinking it is independent in nature. This approach is very rare in real life application.
2)The estimations from naive Bayes is not reliable in terms of estimation.

2.3.1.4 Logistic Regression

Logistic regression which is quite similar to linear regression helps in finding discrete outcomes. It is tend to be less likely in over fitting which is important while working with huge data sets in churn prediction. It helps in relating coefficients and predict outcomes of dependable variables.

Approach: Logistic regression is a class estimation model which uses a single estimator to build a logistic regression model. It usually defines a boundary between classes so as to state the probabilities of the classes depending on the distance of the boundary. There exists two extremes (0 & 1) which help in defining the probabilities and grew larger when the data set is big. It works well in predicting categorical value and continuous values.

Advantages: 1.Logistic regression is one of the popular technique in predicting churn because they provide great prediction and good comprehensibility.
2.Logistic regression works well in making strong and detailed predictions.

Disadvantages : The estimations and predictions are not very reliable. Logistic regression is not used for linear classification and works well for non linear data. Data needs to be transformed properly before prediction because logistic regression requires good data transformation for good prediction.

While every algorithm has their own pros and cons, we want to study on how accurately this four algorithms can predict churn in music service. The algorithm with the most accuracy in prediction of churn is chosen.

2.4 Performance metrics

After a model is trained and is made ready for prediction, the model should be analysed for performance. There are various performance metrics that used in machine learning to analyse performance. Classification metrics like Confusion matrix, Classification Report, Classification accuracy, ROC Curve, AUC Curve are some of the performance metrics used for classification problems. Also metrics like recall and precision can be used in performance analysis of sorting algorithms [31].

2.4.1 Confusion matrix

In Machine learning, models behavior in supervised classification scenarios can be analysed or illustrated by using confusion matrix. Generally, confusion matrix is a two by two matrix that consists of rows and columns. Where rows represents the instances in the actual class and the predicted class represented by the column. The matrix results in 4 possible outcomes namely TP, FP, TN and FN.

- TP - True Positives
- FP - False Positives
- TN - True Negatives
- FN - False Negatives

We can check if the predictions are correct or not by using the above outcomes. [6].

2.4.2 Classification Report

In classification algorithms the predictions quality can be measured by using classification report. Classification report consists of four main metrics. All the four metrics are calculated based on true positive, false positive, true negative and false negatives. The four metrics are [27]:

- Precision
- Recall
- F1-score
- Support

2.4.3 Classification accuracy

For classification algorithms, classification accuracy is the most frequently used performance metrics. It can be defined that the ratio of correctly predicted results obtained out of total predicted results.

2.4.4 ROC Curve (Receiver Operating Characteristic)

ROC Curve is a graphical representation of the classification models performance at all the thresholds of classifications. The ROC curve plots TP rate and FP rate at each threshold of the classification.

2.4.5 AUC Curve (Area Under ROC curve)

AUC refers to "Area Under the ROC Curve". AUC curve is a measure which has the ability to act as classifier to distinguish between classes. The higher the AUC value the better it's ability to work as classifier to distinguish between positive and negative classes.

Chapter 3

Related Work

The following literature review is done before we have decided on our selected domain i.e music streaming industry. The literature review consists of churn prediction done in three major and important industries, telecommunications, gaming industry and music streaming service. Churn prediction is done in almost 12 industries to predict the factors leading to churn. Previously, churn prediction used to be done from management perspective, which used to focus on customer attraction, customer retention,customer identification and customer development. Later, Churn prediction is mostly predicted by CRM (Customer Relation Management).

Most of the previous studies accompany churn prediction in telecommunications industry. The following papers consists of work done in telecommunications industry.

Brânduşoiu et al. [5], researched on churn prediction using three machine learning approaches neural networks, Bayesian classifier and K-Nearest Neighbors etc. They worked on preprocessing of the data using PCA analysis i.e Principle Component Analysis. This method help in finding continuous independent variables using linear combination. Overall, SVM has given best performance with an precision of 95%.Optimization algorithm is a feature selection process which is used for increasing classification accuracy. ROC curve and gain measure is used as performance metrics.

Dahiya and Bhatia [9], researches applied two approaches namely decision trees and logistic regression the data set to find out best performing algorithm between both performance. The work used KDD(Knowledge discovery data) process for preprocessing the data. Finally, the paper estimated that the performance of decision tree is far more accurate compared to logistic regression respectively. Decision tree was more accurate with a precision of 95% compared to regression. The research paper used Weka tool for data mining.

Rodan et al. [28], worked on Support vector approach in churn prediction. The researchers compared the working of Support Vector Machine with traditional machine learning approaches and declared how SVM can be powerful with an accuracy of 98.7%. The proposed model SVM has higher prediction power after optimizing the validation parameters through Grid search.

Ying and Xie [34], researched on churn prediction with random forest and proposed a novel method called IBRF(Improved balanced random forest) to predict churn. The researchers worked with banking industry data set to produce results showing how their method(IBRF) is greater in prediction compared to weighted random forest, balanced random forest. Finally, the thesis aimed to show that limitations in future experimentation with this method might lead to develop a cost effective and enhanced method.

Nie and Zhang [22], presented a paper on showing application of two machine learning algorithms namely, logistic regression and decision trees on a real life data set taken from a Chinese bank to predict credit card churn rate. The research focused on developing a new criterion called mis-classification, which estimates the cost of the model for economic evaluation. The final evaluation showed that logistic regression worked better than decision tree.

Santharam and Krishnan [29], surveyed on different techniques used in churn prediction. Main aim and purpose is to project may attributes and techniques in churn prediction. Almost 18 modelling techniques have been present and predicted that churn prediction can help in customer retention.

Jain et al. [14], worked on churn prediction and retention in Telecom, banking and IT sectors using machine learning techniques. The group worked with data related to three different domains namely Telecom, banking, IT and compared four different algorithms' performance for best performance. The paper works on four algorithms, SVM, XGBoost, Random forest and logistic regression. Exploratory data analysis is used for data preprocessing. The paper main aim is to declare that different domains requires specific algorithm to work with to give best performance respectively. So, it can improve company's profit.

Ahn and Hwang [3], worked on the trends in churn prediction in previous works and at present. The paper compared different churn prediction techniques using log data. It gives an overview of all the domains which works with churn prediction like games, insurance, and management and internet services.The paper outlined comparison of performance analysis of all the algorithms used in churn prediction from all the previous research works. Main aim of the paper is to provide an overview of research done in churn prediction so that future researches can understand what algorithms, data set and features can be selected for their work.

Osisanwo et al. [24] researched on classification and comparison of supervised machine learning algorithms. This paper helped us in selecting our respective algorithms for our research. The paper shows the strengths and weaknesses of supervised algorithms and predicted that the data variables and features selected for one algorithm cannot work for other algorithms for giving best performance.

Chapter 4

Methodology

In this thesis, we mainly focus on concepts of supervised machine learning, churn prediction and explanatory data analysis. The two main methods we are working on are, literature review and experimentation. Literature review in this thesis, we mainly focused on selected domain music streaming service. The research papers are mainly retrieved from Google Scholar and IEEE explore. The relevant publications are extracted and compared. An overview on research work done in churn prediction in music streaming device is given below.

Research Question 1	**Research Question 2**
Inclusion Criteria: 1. Selecting all the latest research papers done in churn prediction. 2. Filtering and finalising research papers related to music streaming service.	**Inclusion Criteria:** 1. Selecting research papers experimenting with churn prediction using machine learning experimentation. 2. Studying all the research papers which related to churn prediction.
Exclusion Criteria: 1. Literature work done in a different domains is not considered.	**Exclusion Criteria:** 1. Excluding literature work that is not available in full text.

Table 4.1: Inclusion & Exclusion Criteria

4.1 Literature Review

Primary goal of literature review is to understand the work done in existing literature work done in music streaming device.

To select research work done in churn prediction in music streaming service, we perform the following steps to work with:

1) We identify paper by searching keywords like machine learning, churn prediction and music streaming service.

2) After the first step, the results are filtered according to our selection i.e churn prediction in music streaming service.

3) We finally attain all the related research works from Google scholar, IEEE Explore, Research Gate etc.

4) We reviewed each and every paper for our literature review, and selected few research papers which can be helpful in our work.

4.1.1 Related work of Churn studies in music streaming service:

Chen worked on map reduced based artificial neural network for churn prediction in music streaming service [7]. KKBOX data-set is considered in the research paper and the paper deals with a paralleled Artificial Neural Network (ANN) which is created to deal with churn prediction when considering a huge data-set. The scalability and effectiveness of the algorithm is also studied.

Lee and Lee worked on a hybrid model for churn prediction [18]. The paper dealt with a hybrid model which the researchers called SEPI. The paper shows the outperformance of SEPI when compared to Artificial Neural Network (ANN),logistic regression, random forest. A real life customer data-set from Korean music streaming service is taken.

Stojanovski [32], researched that Random Forest is the best model among Random Forest, Recurrent Neural Network (RNN) and Long Short-term Memory (LSTM) models, for churn prediction on the eight hours data provided. It consists of best accuracy and F1-Score, highest PR AUC and lowest run time .

Nimmagadda et al. [23] concluded that they tested Logistic Regression, Neural Networks and XGBoost for the binary classification on customer churning problem. The quality of the data plays very crucial part to predict the output accurately. The resultant model for the customer churn problem is XGBoost as it performed better than Neural Networks and Logistic Regression.

Junior and Dinis [16], worked on RF, Artifical Neural Networks (ANN) and Neural Augmented Trees (NAT) models to compare the best performing model among them. ANN is concluded as the best performing algorithm based on all the metrics like precision, recall, accuracy and etc.

Martins [20], researched on predicting customer churn on streaming services. Comparison among the three models namely Logistic Regression, Random Forest and Long short-term memory (LSTM) performed. After performing the experiment the output was LSTM and Random Forest as they both performs best. It is clear that the Logistic Regression is not suitable for churn prediction.

Prey [26], wrote an article on how one should look at music streaming services and compared two major music streaming services Pandora Internet Radio and Spotify. The article gives an overview on how personalisation works and how algorithmic individualization can help in offering right services to the customer.

McCarthy and Oblander, worked on scalable data fusion with selection correlation for better churn prediction in customer based analysis [21]. The paper deals with showing the importance of how data fusion from various sources can lead to better insights. The paper worked with two data-sets credit card bank data-set and spotify real life data-set. The final prediction shows that the customers stayed longer with spotify who used data acquisition from multiple sources.

Maasø and Spilker [19], researched on finding the hidden logic behind music recommendation system. The paper deals with gate keeping mechanisms which helps in customer retention using music service.

Zhang [35], worked on understanding user behaviour in spotify. The dataset is taken from the year 2010-2011 form spotify. An overview of user behaviour while using a music streaming service is given in the research paper.

4.1.2 Churn Prediction Models

Authors	**Selected Algorithms**	**Best Algorithm**	**Year of Publication**
Stojanovski and Filip	Logistic Regression, Random Forest	Random Forest	2017 [32]
Nimmagadda et.al	Logistic Regression, Neural Network, XG-Boost	XGBoost	2017 [23]
Junior and Dinis	Logistic Regression, Decision Trees, Artificial Neural Networks, Neural Augmented Trees, Meta-heuristic Methods	Artificial Neural Networks	2017 [16]
Martins and Helder	Logistic Regression, Decision Trees, Random Forests, Recurrent Neural Networks, Long Short-Term Memory	Random Forests, Recurrent Neural Networks	2017 [20]
Min Chen	Artificial Neural Networks	Artificial Neural Networks	2019 [8].
Jae Sik Lee and Jin Chun Lee	Bayesian-inference-based decision tree, Decision Tree	Bayesian-inference-based decision tree	2006 [18].
Bryan Gregory	Decision Trees, XG-Boost	XGBoost	2018 [12]
Ahmed and Kachkach	Decision Tree, Gaussian Naive Bayes, Random Forest, Logistic Regression, Gradient Boosting Trees	Random Forest	2016 [2]

Table 4.2: Table of Previous Research Papers

4.1.3 Customer Data set:

The data sets used in the research papers that are analysed for churn prediction models are KKBox's challenge data and spotify data. KKBox is a most popular music streaming service that contains nearly 400 million user's data. On average of 14TB per day, user log data is stored by spotify and which can be expanded up-to 140TB. The data collected to work for research, is mainly sample data that is in a form of CSV file that consists 0.74 million playbacks from 1,644 unique users.

4.1.4 Limitations and Issues:

1. Churn studies in Music streaming service domain is very less researched on. Finding previous research works related to machine learning experimentation in music streaming domain was limited.

2. Not every platform provides and reports the data in same manner. Understanding and analysing the data is time consuming.

4.1.5 Conclusion of Literature review:

A comprehensive view of a research problem can be analysed by studying all the work done in the respective area. Analysing churn studies which involves feature selection, predictive models and data sets selected, can help future researches in better model selection. A researcher can work better when they can understand all the limitations and studies done in a particular model. Therefore literature work, can help as a base work for future research.
After our thesis literature review is done, we have understood, churn prediction in music streaming is less researched and studied. Throughout the study and review, experimentation with machine learning algorithms for churn prediction in this domain is very limited.

4.2 Experimentation

Our Research Question2, is explored in this section. Throughout the section we will go through our research work experimentation environment, Exploratory Data Analysis (EDA), Framework, and finally the performance analysis of our model. The predicted results is given in the results section.

4.2.1 Experimental Environment

The environment we used for the research is as follows.

4.2.1.1 Google Colab:

Google colab is a product of Google research. It is a jupyter notebook environment that totally runs on cloud. It allows you to write and execute any python code through the browser and is highly suitable for machine learning and data science.

4.2.2 Python:

Python is one of the most popular programming language at present. It supports multiple paradigms like structure oriented, object oriented, and function oriented programming. It is a general-purpose high-level interpreted and programming language.

4.2.2.1 Numpy:

Numpy is an open source python library that is used to work with arrays. It stands for Numerical Python. It has all the mathematical functions for working with domains like linear algebra and matrices. Numpy library can be imported using the statement **"import numpy"**.

4.2.2.2 Pandas:

Pandas is an open source python library that is used to work with data sets. It has number of functions for analysing, exploring, cleaning and manipulating data. The pandas library can be imported using the statement **"import pandas"**.

4.2.2.3 Sklearn:

Sklearn (Scikit-learn) is one of the most useful python library that can be used for machine learning. Sklearn includes many functions for machine learning and modelling such as Classification, Clustering, Dimensionality Reduction, and Regression. Sklearn can be installed using the pip command **"pip install -U scikit-learn"**.

4.2.2.4 Matplotlib:

Matplot is an open source graph plotting library used for visualization. It is generally written in python, some of the segments can be compatible with C, Java script and Objective-C. We can import the library with the statement **"import matplotlib"**.

4.2.3 Data Set

The data that we used for thesis work is provided for competition purpose by KKBox. KKBox is a music streaming service from Taiwan. Music service is available and widely used in selected countries only. They are Japan and Southeast Asian countries. The challenge or competition consists of tasks to identify whether the customer continues his subscription or terminates it based on the listening habits of the user. Based on the data provided by KKBox we have to predict customer churn. The data that is provided by KKBox's in Kaggle was of 8.33GB which is available in the form

of compressed zip folder that consists of subscribers data from three various sources. The three sources are member data, transactions, and user activity logs. [15]

4.2.4 Exploratory Data Analysis (EDA)

Exploratory Data analysis is an approach used by data scientists to investigate the given data set to analyse the raw characteristics of the data which cannot be perceived by a machine. Its help preprocessing the data-set from a human perspective to see what the data can tell us beyond formal modelling. It can manipulate the data for better results. Data in raw form is full of noise, out-liners and missing values, therefore it is important for the data to be preprocessed before fitting it to model. Data preprocessing of raw data involves the following steps:

1. Data cleaning: Raw data naturally contains noisy, missing values. Data cleaning involves handling this inconsistent data and noise. Data cleaning can be handled through filling the missing values,ignoring the tuples,clustering and regression.

2. Data Transformation: Data is transformed into appropriate forms suitable for mining process so as to visualize and use the data effectively. Data transformation is done by normalization, discretization and feature selection etc.

3. Data Reduction: Data reduction is a process applied with the help of methods like dimensionality reduction, feature subset selection etc to handle huge data. Since working with huge data can become difficult while modelling, data reduction is opted.We have chosen Principle component analysis(PCA) for dimensionality reduction.

```
is_churn                 int64
city                   float64
registered_via         float64
payment_method_id      float64
payment_plan_days      float64
plan_list_price        float64
actual_amount_paid     float64
is_auto_renew          float64
is_cancel              float64
num_unq                float64
total_secs             float64
percent_25             float64
percent_50             float64
percent_100            float64
registration_day       float64
transaction_day        float64
membership_expire_day  float64
last_play_day          float64
dtype: object
```

Figure 4.1: Data Type

4.2.5 Framework

A pictorial representation of the system framework is shown in the given figure below:

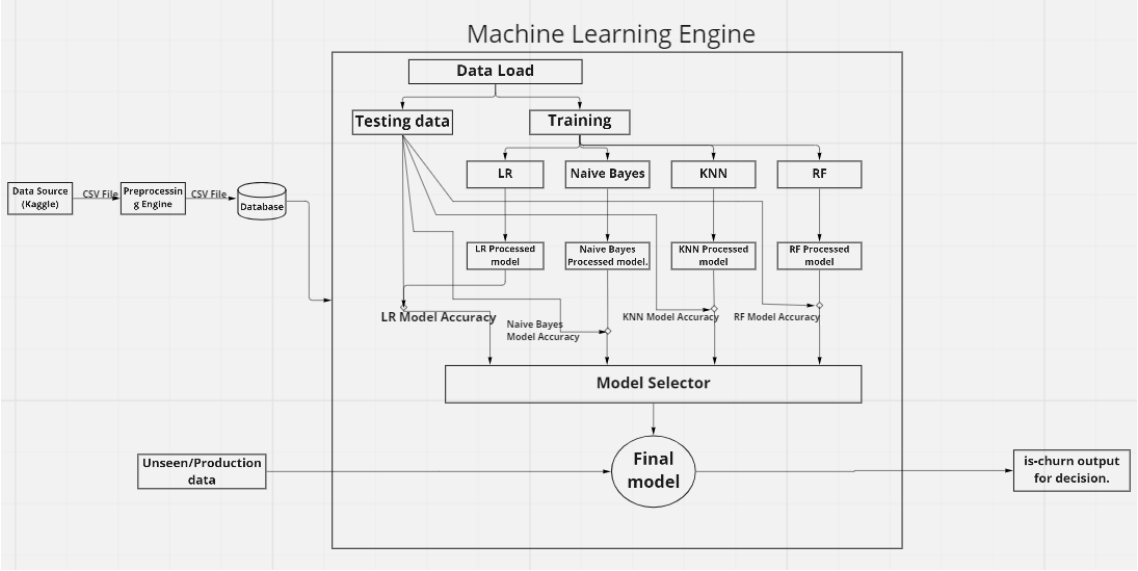

Figure 4.2: Framework

Data Acquisition: Initially, all the data is received through primary and secondary sources. Our research work uses a secondary data source named "Kaggle". The data set named "KKBox" is taken from a churn prediction challenge in music streaming service.

Data Preprocessing: Since the data cannot be directly applied for modelling and experimentation. The data is preprocessed through different phases like data cleaning, data transformation, data reduction. All the inconsistencies, missing values and noise of raw data is effectively transformed into data which is consistent and normalized. The data collected from various sources goes through different steps like data variance analysis, data cleaning, filtering and feature selection. Finally, Principle Component Analysis (PCA) is done for analysing the principle components and saves the data in terms of preprocessed data.

Developing predictive models: This phase consists of applying the preprocessed data to fit the models to makes predictions. We have applied four selected supervised machine learning algorithms, Logistic regression, Naive Bayes, Random forest, K-Nearest Neighbors to predict our output "is-churn" variable. The results of each model is analysed through performance metrics and later compared to find out the best performing algorithm.

Evaluation of results: Model evaluation is the key task to analyse the performance of the respective models. We used various performance metrics like ROC curve, Classification report, Accuracy of each model to evaluate the results. Best performing algorithm after comparison is taken.

4.2.6 Performance Analysis of the Models

After splitting the data into training data and testing data we obtained experiment results. Performance metrics which are commonly used are accuracy, precision, recall, support and f1 score. Along with these we consider confusion matrix, correlation heat map, ROC curve and AUC score. After performing data preprocessing a balanced data set contains 970,960 rows and 11 columns is given to train and test the algorithms to predict the variable output "is-churn". The data set is fit into the model perfectly with the given transform variables. The results of each and every algorithm, namely Logistic regression, Random Forest, Naive Bayes, K-nearest neighbor are noted and compared to get the best performing algorithm. All the results of each and every model is presented in terms of Confusion Matrix, Classification report, Heat-map and ROC curve.

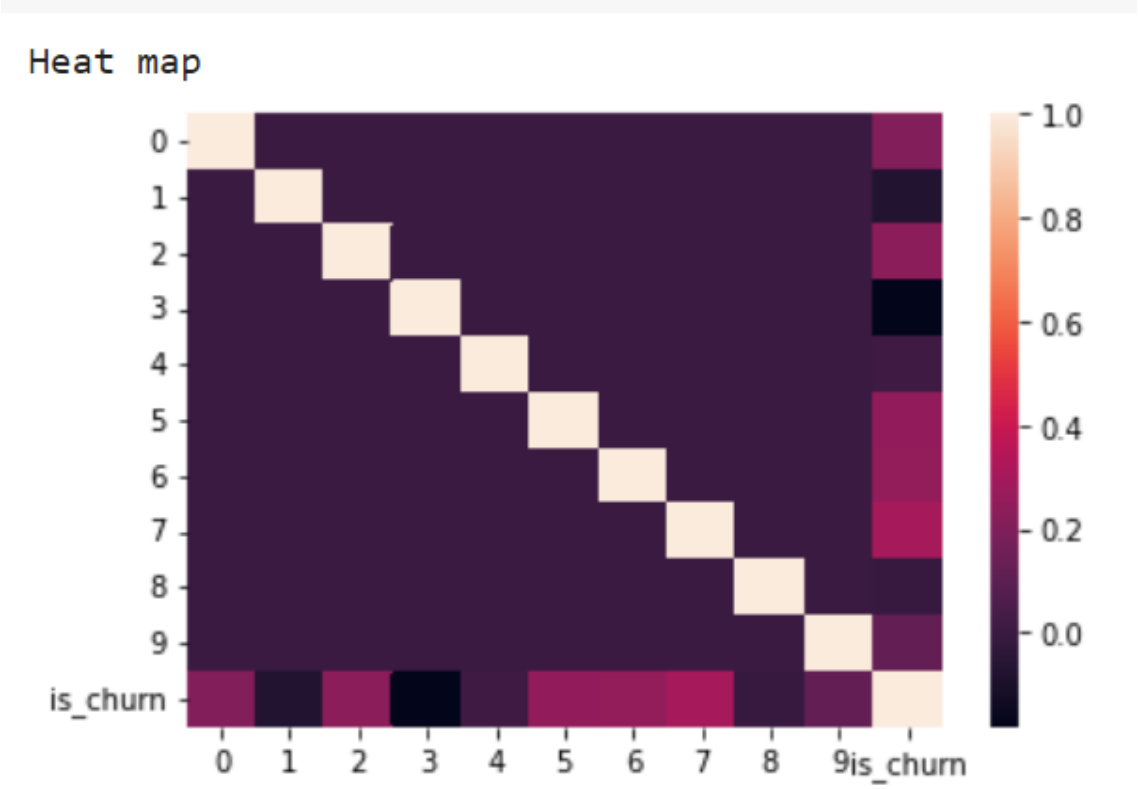

Figure 4.3: Heat-map

The accuracy of each model is noted and compared to find out the best predicted model.

Accuracy: Accuracy is the best measurement in determining the best machine learning algorithms. It can be defined as the total number of correct predictions of data out of total data in the test data set. Usually accuracy is the best performing metrics while comparing classification models.

$$Accuracy = \frac{Total number of correct predictions}{Total number of predictions} \tag{4.1}$$

Classification Report: Precision essentially answers to *What proportion of the variables indicate to the members of reference class actually belong to the graph?* The value is a fraction of true positives overall all the positively classified variables i.e., sum of true positives and false positives.

$$Accuracy = \frac{TP}{TP + FP} \tag{4.2}$$

Recall: Recall answers the question *How many variables belonging to the reference class are correctly identified?*

$$Accuracy = \frac{TP}{TP + FN} \tag{4.3}$$

Support: Support is the actual number of occurrences of the reference class of the specified data set.

F1-Score: F1-Score is a metric which is a harmonic mean of precision and recall. The metric is used to measure an overall model performance.

$$Accuracy = \frac{2TP}{2TP + FP + FN} \tag{4.4}$$

Receiver Operating Characteristic: The performance of binary classifiers is described by ROC. ROC is a curve which plots false positives on X-axis and true positives on y-axis. We generally interested in AUC measure as a metric. The Auc of ROC curve gives us the probability of randomly drawn pair of sample variables from positive and negative classes. An AUC value of 0.5 indicates the model does not perform better than random chance. Higher value of 1.0 indicates the model is performed.

Chapter 5

Results

5.1 Prediction

The results of model prediction from the preprocessed data after fitting the data set into each selected algorithms is given in this section. Four algorithms namely Logistic Regression, Naive Bayes, K-Nearest Neighbors and Random Forest is considered for comparison. The performance metrics of each model is noted and studied to compare the models to find out the best performing algorithm.

5.1.1 Logistic Regression Results:

The model is fit by the training data set using a logistic regression algorithm to aim to predict the churn rate.
The configuration of Logistic Regression is:

```
model = lir.LogisticRegression()
model.fit(X_train,y_train)
y_pred  = model.predict(X_test)

results = metrics.confusion_matrix(y_test, y_pred)
print("\nconfusion_matrix : \n\n",results)
print("\nAccuracy is : ",metrics.accuracy_score(y_test, y_pred))
print("\nClassification Report\n",metrics.classification_report(y_test, y_pred))


list_of_modals_and_accuracy["Logistic Regression"] = metrics.accuracy_score(y_test, y_pred)
y_pred_proba = model.predict_proba(X_test)[::,1]
fpr1, tpr1, _ = metrics.roc_curve(y_test,  y_pred_proba)
auc = metrics.roc_auc_score(y_test, y_pred_proba)

plt.plot(fpr1,tpr1,label="AUC="+str(auc))
plt.ylabel('True Positive Rate')
plt.xlabel('False Positive Rate')
plt.legend(loc=4)
plt.show()
```

Figure 5.1: Code of Logistic Regression

The predictions of logistic regression models in terms of our performance metrics is given below:

```
confusion_matrix :

 [[261805   3321]
 [ 14988  11174]]

Accuracy is :  0.9371446815522781

Classification Report
               precision    recall  f1-score   support

           0       0.95      0.99      0.97    265126
           1       0.77      0.43      0.55     26162

    accuracy                           0.94    291288
   macro avg       0.86      0.71      0.76    291288
weighted avg       0.93      0.94      0.93    291288
```

Figure 5.2: Performance metrics of Logistic Regression

5.1.2 Naive Bayes Results:

The configuration of Naive Bayes and the performance metrics of noted is given below:

```
nv = GaussianNB()
nv.fit(X_train,y_train)
y_pred = nv.predict(X_test)

results = metrics.confusion_matrix(y_test, y_pred)
print("\nconfusion_matrix : \n\n",results)
print("\nAccuracy is : ",metrics.accuracy_score(y_test, y_pred))
print("\nClassification Report\n",metrics.classification_report(y_test, y_pred))


list_of_modals_and_accuracy["Naïve bayes"] = metrics.accuracy_score(y_test, y_pred)

y_pred_proba = nv.predict_proba(X_test)[::,1]
fpr2, tpr2, _ = metrics.roc_curve(y_test,  y_pred_proba)
auc = metrics.roc_auc_score(y_test, y_pred_proba)

plt.plot(fpr2,tpr2,label="AUC="+str(auc))
plt.ylabel('True Positive Rate')
plt.xlabel('False Positive Rate')
plt.legend(loc=4)
plt.show()
```

Figure 5.3: Code of Naive Bayes

```
confusion_matrix :

 [[246393  18733]
 [ 11120  15042]]

Accuracy is :  0.8975138007744913

Classification Report
              precision    recall  f1-score   support

           0       0.96      0.93      0.94    265126
           1       0.45      0.57      0.50     26162

    accuracy                           0.90    291288
   macro avg       0.70      0.75      0.72    291288
weighted avg       0.91      0.90      0.90    291288
```

Figure 5.4: Performance metrics of Naive Bayes

5.1.3 K-Nearest Neighbors:

The model is fit by the training data set using a K-Nearest Neighbors algorithm to aim to predict the churn rate.

The performance metrics and configuration of K-Nearest Neighbors is given below:

```
knn = KNeighborsClassifier(n_neighbors=3)
knn.fit(X_train, y_train)
y_pred = knn.predict(X_test)

results = metrics.confusion_matrix(y_test, y_pred)
print("\nconfusion_matrix : \n\n",results)
print("\nAccuracy is : ",metrics.accuracy_score(y_test, y_pred))
print("\nClassification Report\n",metrics.classification_report(y_test, y_pred))

list_of_modals_and_accuracy["KNN"] = metrics.accuracy_score(y_test, y_pred)


y_pred_proba = knn.predict_proba(X_test)[:,1]
fpr3, tpr3, _ = metrics.roc_curve(y_test,  y_pred_proba)
auc = metrics.roc_auc_score(y_test, y_pred_proba)



plt.plot(fpr3,tpr3,label="AUC="+str(auc))
plt.ylabel('True Positive Rate')
plt.xlabel('False Positive Rate')
plt.legend(loc=4)
plt.show()
```

Figure 5.5: Code of K-Nearest Neighbors

```
confusion_matrix :

 [[261496   3630]
 [  7616  18546]]

Accuracy is :  0.9613921617093736

Classification Report
              precision    recall  f1-score   support

           0       0.97      0.99      0.98    265126
           1       0.84      0.71      0.77     26162

    accuracy                           0.96    291288
   macro avg       0.90      0.85      0.87    291288
weighted avg       0.96      0.96      0.96    291288
```

Figure 5.6: Performance metrics of K-Nearest Neighbors

5.1.4 Random Forest:

The model is fit by the training data set using a Random Forest algorithm to aim to predict the churn rate.

The configuration of Random Forest and performance metrics is given below:

```
model = RandomForestClassifier(n_estimators = 1000, random_state = 42)
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
print(y_pred)
y_pred = np.where(y_pred > 0.5, 1, 0)

results = metrics.confusion_matrix(y_test, y_pred)
print("\nconfusion_matrix : \n\n",results)
print("\nAccuracy is : ",metrics.accuracy_score(y_test, y_pred))
print("\nClassification Report\n",metrics.classification_report(y_test, y_pred))

list_of_modals_and_accuracy["Random forest"] = metrics.accuracy_score(y_test, y_pred)


y_pred_proba = model.predict_proba(X_test)[:,1]
fpr4, tpr4, _ = metrics.roc_curve(y_test,  y_pred_proba)
auc = metrics.roc_auc_score(y_test, y_pred_proba)

plt.plot(fpr4,tpr4,label="AUC="+str(auc))
plt.ylabel('True Positive Rate')
plt.xlabel('False Positive Rate')
plt.legend(loc=4)
plt.show()
```

Figure 5.7: Code of Random Forest

```
confusion_matrix :

 [[262605   2521]
 [  5697  20465]]

Accuracy is :  0.9717873719480377

Classification Report
              precision    recall  f1-score   support

           0       0.98      0.99      0.98    265126
           1       0.89      0.78      0.83     26162

    accuracy                           0.97    291288
   macro avg       0.93      0.89      0.91    291288
weighted avg       0.97      0.97      0.97    291288
```

Figure 5.8: Performance metrics of Random Forest

5.1.5 Comparison Results:

After we apply our desired training data set into four selected machine learning algorithms naming, logistic regression, Naive Bayes, K-nearest Neighbor, Random Forest, we compare four algorithms to obtain results of accuracy prediction in terms of graph representation.

The results after comparison of four algorithms is shown below:

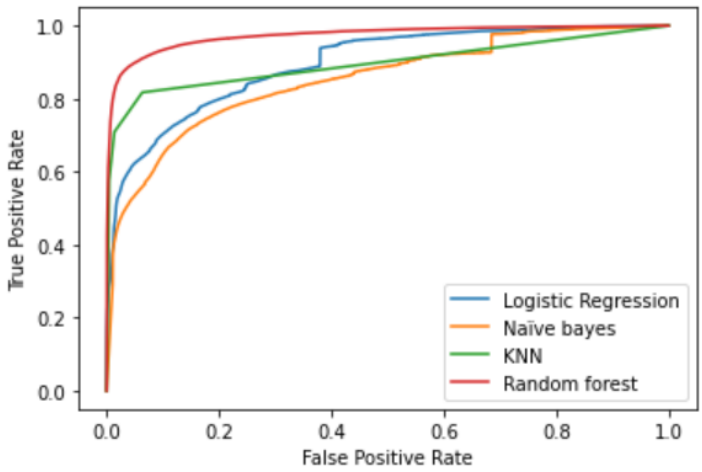

Figure 5.9: Comparison graph

5.2 Observation on Models:

Data preprocessing and feature selection are the two most important steps while experimentation in-order to get better results. We used correlation heat-map to visualize the features after data preprocessing. Accuracy and AUC score is given in the first table. After comparing the accuracy predicted of every algorithms, we conclude that Random forest is the best performing algorithm with an accuracy of 97%. K-Nearest neighbors is followed next with an accuracy of 96%. Logistic regression works with an accuracy of 93%, and finally Naive Bayes with an accuracy of 89%. AUC score range from 0-1.Looking at the AUC score, it is depicted, the higher the AUC, the better the performance of the model. We have Random forest with the best AUC score of 0.97%, depicting that our algorithm results are valid.

The classification report of algorithms is given in a tabular form below. The

Models	**Accuracy**	**AUC score**
Logistic Regression	0.93	0.89
Naive Bayes	0.89	0.85
K-Nearest Neighbors	0.96	0.89
Random Forest	0.97	0.97

Table 5.1: Table of Accuracy and AUC.

table depicts the precision, recall, and F1-score of all the algorithms.

Models	**Precision**	**Recall**	**F1-Score**
Logistic Regression	0.77	0.43	0.55
Naive Bayes	0.45	0.57	0.50
K-Nearest Neighbors	0.85	0.71	0.77
Random Forest	0.89	0.78	0.83

Table 5.2: Table of Precision, Recall and F1-Score.

Chapter 6

Summary

RQ1:
What is the latest research work, and churn prediction techniques used in music streaming services?
Answer:
To answer the first question, a literature review is conducted in Google scholar, IEEE Explore etc using precised keywords "churn prediction" "machine learning", "music streaming service". The results are then filtered according to our selected category "churn prediction in music streaming service" and then studied for better analysis on the domain work. A report on existing literature review done in music streaming service is given in the Method section.

RQ2:
In terms of performance, which machine learning algorithm is the most accurate for the churn prediction in music streaming domain?
Answer:
Research question 2 is answered by the Experimentation process. Thesis main aim is to build a churn predictive model for a music streaming data using various classification algorithms. We have built the models for Logistic regression, Random forest, Naive Bayes and K- Nearest neighbor algorithms. The algorithms are trained by preprocessed data which is obtained by preprocessing the KKBox's dataset. The algorithms are trained with 70% of the data and tested with 30% of the data. After training and testing the algorithms we obtained the results for each model namely classification accuracy, classification report, confusion matrix, ROC and AUC. We considered the accuracy and AUC of the models as a main metrics to evaluate the best performing model. The accuracy of Logistic Regression is 93% and AUC score is of 89%, Naive Bayes accuracy is 89% and AUC is 85% K-Nearest Neighbors accuracy is 96% and AUC is 89%, finally Random Forest accuracy is 97% and AUC is 97%. The final result we obtained is Random forest with the highest accuracy of 97% compared to all other algorithms.

Chapter 7

Conclusions and Future Work

7.1 Conclusions:

With a drastic increase in service providers, in 21st century, it is very important for every company to understand why customers turn down their services to keep up with the growth of their company in high competition. Business loss is unavoidable, but it can be reduced by customer attraction and customer retention. Good methods and improvising present methods can help any area in better performance in terms of managing a company. Churn prediction is one such method which helps in reducing business loss and customer churn rate. So, in this thesis we have decided to work on churn prediction in music streaming service domain. We have collected our data from an open source data resource called Kaggle. The data is analysed by EDA(Exploratory data analysis). Almost 70% of data is trained and 30% of data is used for testing after preprocessing. Initially, literature review is done in our selected domain and an overview is given thoroughly. Later the thesis includes comparative study of supervised machine learning algorithms namely, Logistic regression, Random forest, Naive Bayes, K- Nearest neighbor to predict the best accurate performing algorithm. The experimental results shows that the random forest is the best performing algorithm with an accuracy of 97%.

7.2 Future Work:

Music streaming domain is a very less research domain in compared to all the domains which does churn prediction. In future, we intend to do better research by constructing various churn prediction models in deep learning like neural networks. Conducting experiments and evaluating each model can help in giving better insights for future researchers.

7.3 Limitations:

Churn prediction is a vastly researched topic and can be applied in different fields. Churn studies in these domains can help us in giving a comprehensive view on the subject but limits researchers to give common performance. Researchers can be in anticipation if their model has been widely researched on and can be suitable for best performance in their study.

References

[1] A. K. Ahmad, A. Jafar, and K. Aljoumaa, "Customer churn prediction in telecom using machine learning in big data platform," *Journal of Big Data*, vol. 6, no. 1, pp. 1–24, 2019.

[2] K. Ahmed, "Analyzing user behavior and sentiment in music streaming services," 2016.

[3] J. Ahn, J. Hwang, D. Kim, H. Choi, and S. Kang, "A survey on churn analysis in various business domains," *IEEE Access*, vol. 8, pp. 220 816–220 839, 2020.

[4] S. F. Bischof, T. M. Boettger, and T. Rudolph, "Curated subscription commerce: A theoretical conceptualization," *Journal of Retailing and Consumer Services*, vol. 54, p. 101822, 2020.

[5] I. Brânduşoiu, G. Toderean, and H. Beleiu, "Methods for churn prediction in the pre-paid mobile telecommunications industry," in *2016 International Conference on Communications (COMM)*, 2016, pp. 97–100.

[6] O. Caelen, "A bayesian interpretation of the confusion matrix," *Annals of Mathematics and Artificial Intelligence*, vol. 81, no. 3, pp. 429–450, 2017.

[7] M. Chen, "Music streaming service prediction with mapreduce-based artificial neural network," in *2019 IEEE 10th Annual Ubiquitous Computing, Electronics Mobile Communication Conference (UEMCON)*, 2019, pp. 0924–0928.

[8] ——, "Music streaming service prediction with mapreduce-based artificial neural network," in *2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*. IEEE, 2019, pp. 0924–0928.

[9] K. Dahiya and S. Bhatia, "Customer churn analysis in telecom industry," in *2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions)*, 2015, pp. 1–6.

[10] S. De, P. Prabu, and J. Paulose, "Effective ml techniques to predict customer churn," in *2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA)*. IEEE, 2021, pp. 895–902.

[11] A. Deligiannis and C. Argyriou, "Designing a real-time data-driven customer churn risk indicator for subscription commerce." *International Journal of Information Engineering & Electronic Business*, vol. 12, no. 4, 2020.

[12] B. Gregory, "Predicting customer churn: Extreme gradient boosting with temporal data," *arXiv preprint arXiv:1802.03396*, 2018.

[13] J. Hadden, A. Tiwari, R. Roy, and D. Ruta, "Churn prediction: Does technology matter," *International Journal of Intelligent Technology*, vol. 1, no. 2, pp. 104–110, 2006.

[14] H. Jain, G. Yadav, and R. Manoov, "Churn prediction and retention in banking, telecom and it sectors using machine learning techniques," in *Advances in Machine Learning and Computational Intelligence*. Springer, 2021, pp. 137–156.

[15] Z. Jianyu and F. Soulie Fogelman, "Kkbox's music recommendation challenge solution with feature engineering. 11th acm international conference on web search and data mining wsdm 2018. february 5-9 2018, los angeles, california, usa. wsdm cup 2018 workshop. 2018." 02 2018.

[16] C. Junior and G. Dinis, "Churn analysis in a music streaming service: Predicting and understanding retention," 2017.

[17] 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.

[18] J. S. Lee and J. C. Lee, "Customer churn prediction by hybrid model," in *International Conference on Advanced Data Mining and Applications*. Springer, 2006, pp. 959–966.

[19] A. Maasø and H. S. Spilker, "The streaming paradox: Untangling the hybrid gatekeeping mechanisms of music streaming," *Popular Music and Society*, pp. 1–17, 2022.

[20] H. Martins, "Predicting user churn on streaming services using recurrent neural networks," 2017.

[21] D. M. McCarthy and E. S. Oblander, "Scalable data fusion with selection correction: An application to customer base analysis," *Marketing Science*, vol. 40, no. 3, pp. 459–480, 2021.

[22] G. Nie, W. Rowe, L. Zhang, Y. Tian, and Y. Shi, "Credit card churn forecasting by logistic regression and decision tree," *Expert Systems with Applications*, vol. 38, no. 12, pp. 15 273–15 285, 2011.

[23] S. Nimmagadda, A. Subramaniam, and M. L. Wong, "Churn prediction of subscription user for a music streaming service," *CS229 Lecture, Standford Univ., Standford, CA, USA, Project Rep., Fall*, 2017.

[24] F. Osisanwo, J. Akinsola, O. Awodele, J. Hinmikaiye, O. Olakanmi, and J. Akinjobi, "Supervised machine learning algorithms: classification and comparison," *International Journal of Computer Trends and Technology (IJCTT)*, vol. 48, no. 3, pp. 128–138, 2017.

[25] A. Ozgur, "Supervised and unsupervised machine learning techniques for text document categorization," *Unpublished Master's Thesis, İstanbul: Boğaziçi University*, 2004.

[26] R. Prey, "Nothing personal: algorithmic individuation on music streaming platforms," *Media, Culture & Society*, vol. 40, no. 7, pp. 1086–1100, 2018.

[27] M. Radja and A. W. R. Emanuel, "Performance evaluation of supervised machine learning algorithms using different data set sizes for diabetes prediction," in *2019 5th International Conference on Science in Information Technology (ICSITech)*, 2019, pp. 252–258.

[28] A. Rodan, H. Faris, J. Al-sakran, and O. Al-Kadi, "A support vector machine approach for churn prediction in telecom industry," *International journal on information*, vol. 17, 08 2014.

[29] A. Santharam and S. B. Krishnan, "Survey on customer churn prediction techniques," *International Research Journal of Engineering and Technology*, vol. 5, no. 11, p. 3, 2018.

[30] A. Saran Kumar and D. Chandrakala, "A survey on customer churn prediction using machine learning techniques," *International Journal of Computer Applications*, vol. 975, p. 8887, 2016.

[31] N. Seliya, T. M. Khoshgoftaar, and J. Van Hulse, "A study on the relationships of classifier performance metrics," in *2009 21st IEEE international conference on tools with artificial intelligence*. IEEE, 2009, pp. 59–66.

[32] F. Stojanovski, "Churn prediction using sequential activity patterns in an on-demand music streaming service," 2017.

[33] 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.

[34] Y. Xie, X. Li, E. Ngai, and W. Ying, "Customer churn prediction using improved balanced random forests," *Expert Systems with Applications*, vol. 36, no. 3, pp. 5445–5449, 2009.

[35] B. Zhang, G. Kreitz, M. Isaksson, J. Ubillos, G. Urdaneta, J. A. Pouwelse, and D. Epema, "Understanding user behavior in spotify," in *2013 Proceedings IEEE INFOCOM*. IEEE, 2013, pp. 220–224.

Appendix A

Supplemental Information

Faculty of Computing, Blekinge Institute of Technology, 371 79 Karlskrona, Sweden