National College of Ireland

Customer Churn Analysis in Telecom Using Machine Learning Techniques

MSc Research Project
Msc in Data Analytics

Manish Kumar Mittal
Student ID: x20185596

School of Computing
National College of Ireland

Supervisor: Qurrat Ul Ain

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Manish Kumar Mittal
Student ID:	x20185596
Programme:	Msc in Data Analytics
Year:	2022
Module:	MSc Research Project
Supervisor:	Qurrat Ul Ain
Submission Due Date:	15/08/2022
Project Title:	Customer Churn Analysis in Telecom Using Machine Learning Techniques
Word Count:	7103
Page Count:	22

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	19th September 2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Customer Churn Analysis in Telecom Using Machine Learning Techniques

Manish Kumar Mittal
x20185596

Abstract

Telecommunication industry is among few industries with high technological dependency. Companies are struggling to retain their performance. For this, customer churn prediction becomes crucial way to predict customer's information for decisions. Therefore, the study has focused on customer churn prediction the efficient role of machine learning and hybrid modelling techniques. Gradient boosting, random forest, decision tree and logistic regression has been used as machine learning techniques along with hybrid modelling. RandomizedsearchCV was used to improve gradient boost performance. Synthetic Minority Oversampling Technique was used to improve Knowledge Discovery in Databases for better results. The study results are evaluated based on confusion matrix and compared based on accuracy, precision, recall and f1 score. Gradient boosting outperformed all other models by achieving 96.81% of accuracy.

Keywords— Customer Churn, Classification, Decision Tree, Random Forest, Logistic Regression, Gradient Boosting

1 Introduction

The telecommunication industry has recorded massive growth in modern times Mishra and Rani (2017). The increased use of telecommunication channels has increased telecommunication value in society. Globally, the telecom industry has provided a plethora of assistance to both the public and private sectors. In the telecom industry, market competition has significantly increased. Telecom companies are finding it difficult to retain their position in the market Amin et al. (2017). The changing market dynamics and uncertain market condition has been one major challenge for telecom companies. To overcome the market issues, companies are engaged to compensate for the market impact through customer satisfaction. The possibility of customer satisfaction is aligned with customer attraction and trust in the company's services Huang et al. (2012).

Moreover, the increased value of the telecom industry is aligned with customer satisfaction. Telecom companies focus on customer retention due to increased competition in the market. More and more companies mean more competition. Therefore, companies are bound to make effective customer retention strategies for better performance within the industry. Increased technology has integrated customers within the management. To ensure an effective customer management system (CMS) customer churn analysis are crucial for the companies Lu et al. (2012). This indicates that customer churn analysis plays a significant role in good customer relationships in the telecom industry. Telecom companies are facing rigorous problems of churn and customer relationship management Amin et al. (2020).

Additionally, customer behaviour is uncertain Prashanth et al. (2017). The changing nature of customers' behaviour, the unpredictability of product retention and the changes in customer purchasing intentions are key challenges for telecom companies. To overcome these issues, telecom companies use different techniques to analyze customer churn. The induction of technology in the industry has massively contributed to customer churn analysis Li et al. (2014). Companies prefer using advanced technologies such as machine learning techniques for customer churn analysis to predict customers and make strategies accordingly. The survival approach in the competitive environment has forced companies to retain their existing customers Rai et al. (2020). Customer retention has become a major objective for telecom companies in modern times Dahiya and Bhatia (2015).

Keeping the significance of customer churn analysis and machine learning techniques, the study has focused on identifying the role of machine learning techniques in customer churn prediction. The study has used different machine learning techniques to identify the role of machine learning and hybrid modelling techniques in customer churn evaluation or prediction in the telecom industry. The entire study has used a scientific approach to achieve study objectives.

1.1 Background and Motivation

Different studies have been conducted on customer churn prediction so far. Huang et al. (2012) worked on customer churn prediction to land-line customer retention. They used Henley segmentation to identify customer-related issues and provide land-line companies with a chance to retain its customer in the future. Their study used machine learning algorithms for landline customer churn prediction. They concluded that customer churn prediction results through machine learning algorithms are highly effective.

Hereafter, Lu et al. (2012) worked on identifying the impact of digital systems and technological information in the telecom industry. They highlighted the importance of digital transformation to form a digital customer relationship management system (CMS) for the telecom industry. The overall study was based on CMS formation to provide an effective platform for customer retention in the telecom industry. The link between digital systems and digital customer relationship management systems (CMS) has been an important factor for the company's HR management to form effective strategies related to customer retention. Further, Dahiya and Bhatia (2015) used WEKA software to determine the role of Decision tree and Logistic regression in customer churn prediction in the telecom industry.

Vafeiadis et al. (2015) worked on customer churn prediction using a cross-validation approach. Their study worked on datasets that best suites customer churning datasets. They also concluded that machine learning techniques have a high accuracy rate in customer churn prediction in the telecom industry. Dalvi et al. (2016) also conducted a similar study on customer churn prediction in the telecommunication industry. They also highlighted the importance of customer retention and its impact on company performance. They used different machine learning algorithms such as decision trees and logistic regression to identify their role in customer churn evaluation and assessment.

Coussement et al. (2017) study used machine learning techniques in data preparation for customer churn in the telecom industry. Their study used the KDD data mining method for data preparation. Zhao et al. (2021) conducted a similar study in China telecom industry, they focused on customer churn using china telecom datasets. Similarly, Kisioglu and Topcu (2011) used the Bayesian neural technique for customer churn datasets in Turkey. De et al. (2021) used different machine learning techniques to predict customer churn with altered features. Yu et al. (2018) conducted a similar study, however, it used an optimization-based bp network to predict customer churn. Sharma et al. (2020) used the gradient boosting technique for customer churn prediction on telecommunication industrial datasets to find results. Mishra and Reddy

(2017) conducted a comparative study on different classifiers in machine learning techniques. The comparison focused on customer churn predictive datasets.

Amin et al. (2017) worked on customer churn prediction using rough set theory (RST). Their study adopted a rough set approach to extract data and provide decision making information on customer churn and non-churn. Their study used exhaustive, Genetic and covering algorithms to bring efficiency to the data extraction process. effective dimensionality reduction method is crucial in the telecom industry for customer churn prediction Yihui and Chiyu (2016).

Keeping this into consideration, machine learning techniques and their innovative approach to predict customer churn has gained wide attention. Customer churn prediction in the telecommunication industry needs to have machine learning techniques to predict customer churn data and make effective decisions. Therefore, this study is focusing on the use of machine learning techniques to test its prediction role in customer churn. For this, the study has included machine learning algorithms such as Logistic Regression, Random Forest, Decision Tree Classifier, and Gradient Boosting Classifier and hybrid modelling to analyze customer churn.

1.2 Research Question

The research question of the study is "How machine learning and hybrid modelling can predict and analyze customer churn and develops customer retention program in the telecommunication industry"?

1.3 Research Objectives

Following are key research objectives.

- To analyze customer churn behaviour in the telecom industry using machine learning algorithms such as (Logistic Regression, Random Forest, Decision Tree Classifier, Gradient Boosting Classifier).
- To predict customer behaviour in the telecom industry using hybrid model of (Random Forest, Decision Tree Classifier and Gradient Boosting Classifier).
- To develop a focused customer retention program in the telecommunication industry using machine learning algorithms.

1.4 Research Outlines

The overall study outline is based on seven different sections. Each section covers one part of the entire study. In section 1, the study has discussed a basic introduction related to customer churn and machine learning techniques. Through the introduction, a logical connection has been formed between customer churn analysis and machine learning. The section has also discussed study objectives and study questions. The background and motivation sub-section provides an overall view to the reader.

In the section 2, related work with customer churn analysis and machine learning has been discussed. All the related work has been critically reviewed on justifiable grounds for this study. Further, the study has focused on the literature gap to provide reasons for conducting this study. The research gap has critically identified all major loopholes in the previous work.

In the section 3, the research methodology has been discussed with proper justification about why the methodology has opted for customer churn analysis. The section 4 of the study covers design specifications for customer churn analysis using machine learning techniques. The section 5 covers, the implementation and evaluation of the entire design. The section is crucial for reliable results. The section 6 is based on a comparison of results. Lastly, section

7 provides a conclusion and discussion of the role of machine learning techniques in customer churn prediction. Future work has also been discussed in this section.

2 Related Work

The relevant work section contains a critical review of all related work with machine learning and hybrid modeling in customer churn prediction.As part of the critical review, previous research has been analyzed to justify the validity of this study.A hybrid model and machine learning are the two sections of related work.

2.1 Machine Learning Techniques & Customer Churn Prediction

Huang et al. (2010) focused on the multi-objective feature selection method for customer churn prediction in the telecommunication industry. The study mainly used NSGA-II as a key optimization factor. To critically analyze, it is pertinent to mention that the basic idea behind the use of the NSGA-II optimization method is to focus on all the customer features with different sizes and values. The issue with NSGA-II algorithms or approach is that it requires elitist principles or features to optimize the process. In customer churn prediction, the elitist principle can be population for different aspects. On the other hand, using NSGA-II can have an affirmative role in diverse feature selection. Another study by Huang et al. (2012) focused on landline customer churn prediction through machine learning techniques. The study used different landline features to identify customer churns such as monthly billing details of the customers, line information, bill and payment channels, account-related information services information, complaints information and overall demographic information. To evaluate, their study has inducted 6 different features for customer churn prediction. The use of this feature allowed them to use multiple machine learning and deep learning techniques for valid results. However, the issue with the use of multiple machine learning techniques for customer churn prediction is that possibility of error can be high. Also, multiple techniques can make the prediction procedure more complex. The sequence of procedures can require expertise to ensure valid results. For instance, using multiple machine learning and deep learning techniques requires higher expertise in algorithm selection. To further analyze, data acquisition becomes another complex issue in the process. Nevertheless, their study concluded all machine learning and deep learning techniques have provided effective results in customer churn prediction. To critically review, the probability of accurate results is high due to the induction of new customer features as a dataset. Being inquisitive, these machine learning techniques' performance can be questionable with minimal customer features. In their study, SVM performed better than other techniques due to the higher classifying rate and a large dataset of features. The performance of the rest was moderate and acceptable.

Further, Lu et al. (2012) argued that digital customer relationship management systems (CMS) have emerged as a global trend. The use of digital CMS systems has provided leverage to the telecom companies to make effective customer retention strategies. They added that the formation of a digital CMS system requires customer retention which is a major challenge for companies. Therefore, their study focused on customer churn prediction. To critically evaluate, their study focused on the formation of boosting model for customer churn. Unlike other studies, they focused on boosting the accuracy rate of churn prediction in the industry. Further, their studies focused on high-risk customers which were further clustered for differentiation. Even though their study used logistic regression for building a boosting churn prediction model. However, their study ignored the fact for logistic regression, several observations must be higher than customer churn features, if observations are less than customer churn features, the chances

of overfitting are high. To further evaluate, their study showed good results but it was based on single regression modelling. The only issue with this modelling is that it relies on discrete numbers or features. In logistic regression modelling, the assumption of linear formation between the dependent and independent variable is crucial, in this study the linearity was not highlighted. Therefore, it creates confusion for the companies to rely on the results. Secondly, the issue is that the results on customer churn prediction in this study have used logistic regression, however, any powerful algorithm such as a neural network can overcome the results of this modelling. Therefore, relying only on a single technique can increase the chances of invalidity in the results.

Dahiya and Bhatia (2015) argued that increased competition in the telecommunication industry has indirectly stressed companies retain their customers by making advanced technology-based decisions. To critically analyze, their study is based on churn prediction using an advanced novel framework called WEKA to determine the effectiveness of decision trees and logistic regression. Also, data mining used for machine learning techniques provides an accurate and systematic channel to manage data but the issue with WEKA data mining is that it only works on small datasets, large datasets create complications within the process. Below is the proposed data mining framework of the study.

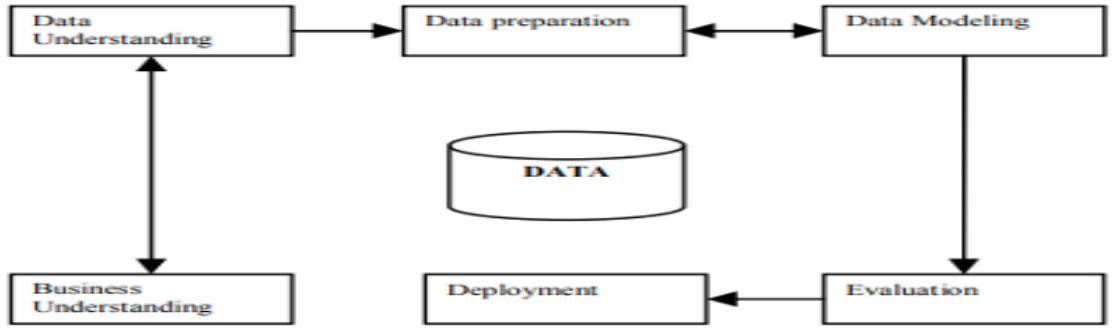

Figure 1: Data Mining Model Dahiya and Bhatia (2015)

Their study mainly focused classification of customer who leaves and who stays as a subscriber with the companies. The issue with his approach is that customers are rational in terms of choices, the uncertainty with customers cannot be calculated directly with machine learning. However, it can classify and provide clusters. Also, the study has used KDD as a key method to determine its effectiveness but it is a complex process. One issue with KDD is that data acquisition is less reliable. It is a mega task during the process, and the probability of data misuse is always high. Data preparation is another major task in KDD. To monitor the effectiveness of decision trees and logistic regression, data preparation must become an essential factor of consideration. The above data mining method is a standard data mining approach for WEKA. To further analyze, their study concluded that decision trees have shown high accuracy rate than logistic regression but the study did not mention the terms or factors that contribute to the accuracy process.

Further, Vafeiadis et al. (2015) conducted a comparative study in customer churn prediction using machine learning algorithms. To critically evaluate, the study has used all machine learning models on the public dataset to determine the overall efficiency of these ML algorithms. The study also focused on cross-validation in all models. This study supports the cross-validation on all ML models, however, the issue is that cross-validation on public datasets can be expensive and risky. For cross-validation, specific customer feature is crucial. Secondly, the study used Monte Carlo simulations as a key method to boost the process. They applied the simulations to different parameters. In analysis, using Monte Carlo Simulation can have certain implications. The use of MC simulation is computationally inefficient in large datasets. In cross-validation,

the process becomes hectic, it requires excessive time for the computation of all the parameters. Secondly, in the absence of the KDD method, MC simulation on poor parameters i.e. poor customer features can have poor results. Therefore, this study has certain objections to the use of Monte Carlo Simulation as a booster in the process.

Furthermore, Dalvi et al. (2016) focused on the importance of customer churn prediction. Their study also used decision trees and logistic regression as key machine learning techniques to check their efficiency in result generation. To analyze, the use of decision trees and logistic regression are commonly adopted techniques in customer churn prediction. However, using decision trees and logistic regression requires datasets that contain different features. The study has used public datasets with communal customer features i.e. customer demographic information, services information and information related to customer complaints, call data records and overall expenses. This indicates that decision tree and machine learning techniques are common approaches, the study did not add any new method to provide a valuable contribution to the topic.

Likewise, Coussement et al. (2017) argued that data preparation is crucial for data analysis in the telecom industry. Their study used a data mining technique and logit model to predict all the data alternatives that are essential for customer churn prediction. The study used European telecom datasets for cross-sectional analysis. Their study concluded that logistic regression is highly compatible with single data mining algorithms. The study also added the key contribution of data preparation in managerial decisions on customer retention. To evaluate, the study has focused on the importance of data preparation for machine learning. The study does not include machine learning techniques as a direct agent of contribution, however, it focused on the support of data preparation. Using the KDD method in the process is an accurate approach but the study ignored data extraction and data duplication steps in data preparation. For machine learning techniques such as logistic regression, data errors need to be highlighted before the process. The chances of data error in logistic regression are high due to poor data preparation techniques. All the missing values in data preparation need to be identified.

Brandusoiu and Toderean (2013) worked on customer churn prediction using call details, they used a support vector machine (SVM) on call detail using datasets of 3333 records. To evaluate, customer churn prediction on call details can provide inaccurate data, the inclusion of multiple customer features is essential to determine machine learning techniques and their accuracy rate.

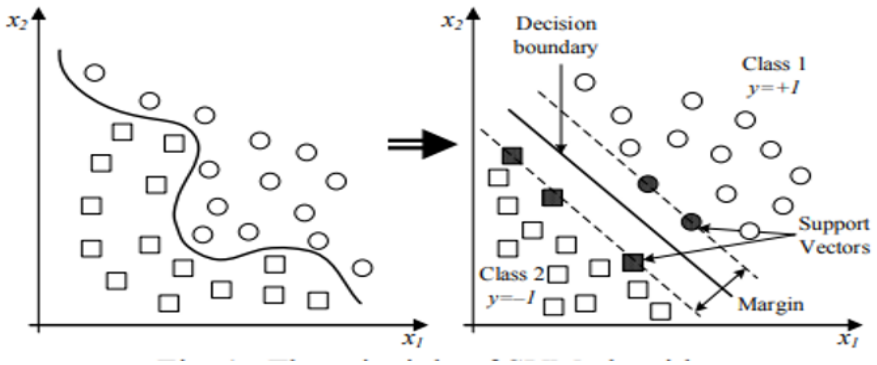

Figure 2: The Principle of SVM algorithm Brandusoiu and Toderean (2013)

Amin et al. (2014) worked on customer churn prediction using machine learning techniques to achieve the objective of one or multi-classifier. To analyze their study, using both one and multi-classifier can have effective results in churn prediction. The key aspect of a multi-classifier is that it covers all aspects of the custom features.

2.2 Hybrid Modelling & Customer Churn Prediction

De et al. (2021) argued that customer churn has been the major challenge for the telecom companies to expand its growth in the market. They added that customer churn is paramount to making effective strategies regarding customers. The use of machine learning techniques and hybrid modelling is playing a vital role in customer churn. Their study analysis of customer churn has used hybrid modelling which provides high accuracy in churn prediction. The hybrid modelling algorithms such as random forest, decision tree and SVM as classifier has been the key factor on focusing rich customer content such as emails, phone calls, and other records. They further stressed that hybrid modelling algorithms as classifier have classified information related with customers. Similarly, Bayrak et al. (2022) study argued that competition and high productive environment has increased the use of technology. Service provider companies are directly and indirectly using advanced technologies to impact customer churn. Customer retention becomes very important because of high competition. To manage customer churn in the fast-food industry, the use of hybrid modelling on deep learning techniques and machine learning techniques such as recurrent neural networks and short-term memory has been used to make sequential data about customer churn. They added that the use of recurrent neural networks helped companies to predict customer churn in the fast-food industry. Likewise, Choudhari and Potey (2018) used two kinds of hybrid classification algorithms to form customer information cluster in the telecom industry. The use of hybrid modelling has effectively predicted customer churn in the telecom industry than single algorithms. The use of a Hybrid Decision Tree and Logistic Regression has increased results accuracy.

2.3 Research Gap

The literature gap of the study is based on a critical review of all the previous work on customer prediction in the telecom industry. Previous studies have failed to compare customer churn datasets of two countries using machine learning techniques. Countries with different datasets and similar features can have unique outcomes to determine the effectiveness of machine learning techniques in customer churn forecasting. None of any studies has shown any objective work on these datasets.

Additionally, previous studies have used machine learning techniques for the automation of the customer churn prediction process. However, none of any studies has discussed the pros and cons and its limitations in the telecommunication industry. Previous studies have not highlighted the crucial role of the data mining process i.e. data extraction from different datasets. Also, previous studies have not shown any effective technique for the handling of data despite KDD data mining. None of any studies has provided a method to counter inconsistency in datasets during data acquisition.

Further, previous studies have not focused on the use of hybrid modelling to target large datasets. This study has used Telco company churn data from the Kaggle website to analyze customer churn. Previously, the studies have not used Kaggle datasets as they are widely accepted in hybrid modelling. To achieve study objectives, the use of machine learning techniques and hybrid modelling algorithms on Kaggle datasets on customer churn makes the study unique.

3 Methodology

The used dataset in the study is comprised of large data generated from multiple sources. The large data in the dataset is a major challenge for the study to extract useful information to increase prediction and analyze the value of customer churn in the telecommunication industry. The use of effective tools and techniques to extract data from the dataset can increase the chances of effective decision making. This indicates the importance of data extraction tools

in the study. Keeping this into consideration, the study has used Knowledge Discovery in Databases (KDD) as a data mining tool to extract data from the Telco customer churn dataset focusing on different variables related to customer information for churn analysis. The accurate use of Knowledge Discovery in Databases (KDD) services helps in data extraction and data transformation from large datasets.

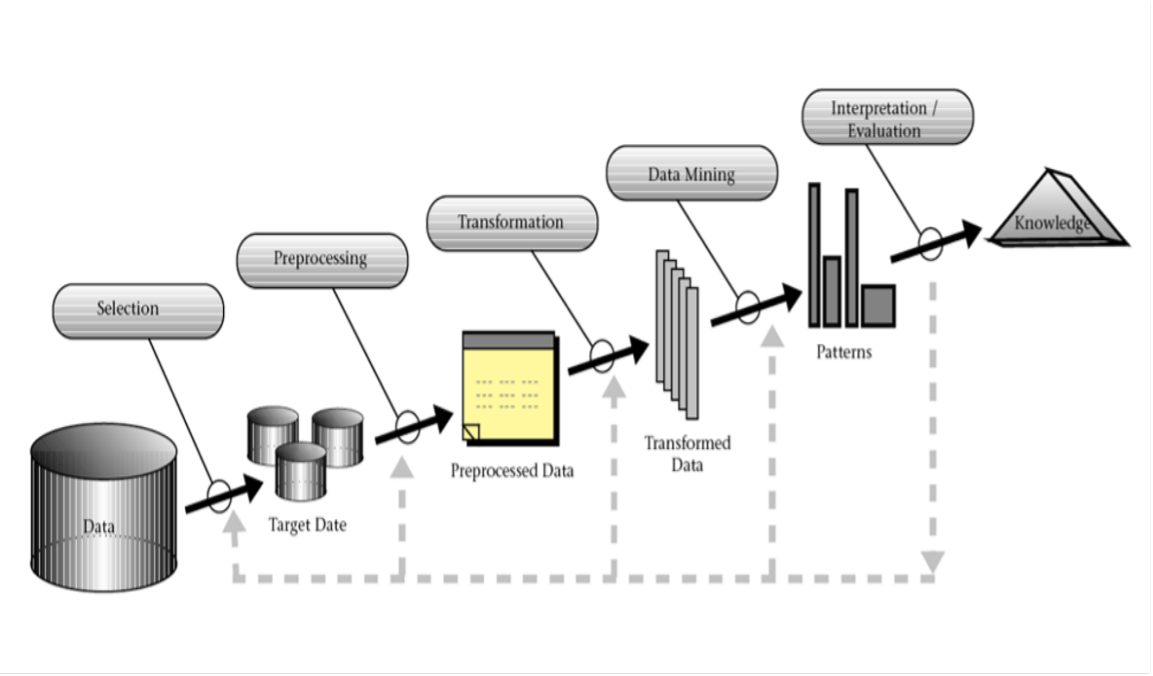

Figure 3: Knowledge Discovery in Databases

3.1 Data Collection

Data collected in this study has been taken from the Kaggle dataset Bojer and Meldgaard (2021). In the Kaggle dataset, Telco customer churn data has been selected to predict customer churn. Using telco customer churn data in the study is based on key aspects that make the data sets relevant. The Telco customer churn data is comprised of information related to fictional telco companies. The company provides all kinds of information related to customer phone and other internet services usage Rahman et al. (2022). A total of 7043 customers' information in California has been used in Telco company churn data. The 7043 customer data indicates the large size of customer data which requires KDD as major data extracting tool with high accuracy. Also, Telco Company customer churn data provides information related to the customer to terminate or shifted from the network. Likewise, information related to customer retention or staying in the network is another key factor of consideration in this dataset. Additional signed-up customers are another point of consideration. Each section in the Telco company dataset determines different attributes of customers. Hence, below are key factors that determine the use of the Kaggle dataset i.e. (Telco Company customer churn). The dataset provides information related to customers who quit or left company services within the last month. This column is named churn in the dataset. This column "churn" is highly pivotal because it indirectly determines customer churn quantity and company performance in the previous month.

1. The dataset provides information related to customer services such as phone numbers, multiple lines, internet usage, online security, online backup, TV, streaming, device protection information and tech support. Information related to customers signed-in into

these services indicates customer overall behaviour.

2. Information related to the customer account is another key point in this study. Information related to customer contract time, payment methods, e-services, monthly expenses and other charges and accumulated charges (both paid and due) is provided in this Telco customer churn data.

3. Demographic information related to customers is another key point of consideration. The dataset provides demographic information i.e., customer age, income range, customer financial condition i.e., dependent or partner.

3.2 Exploratory Data Analysis

In the data exploration phase, different aspects of the customer churn datasets are explored. Among the total dataset of 7043 in 21 columns features, different attributes of datasets were explored. In the exploration phase, counting of customer churns features were analyzed.

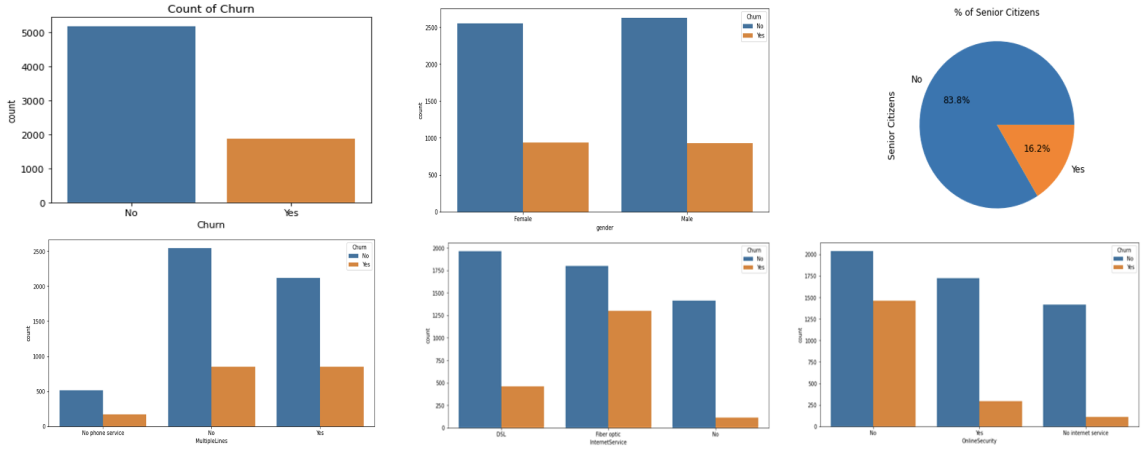

Figure 4: Exploratory Data Analysis

The dataset was explored by questioning the gender of both males and females. In the dataset, 50% female and 50% male were found. Questioning on senior citizen, 16.2% were found among all 7043 customers. Similarly, in dependency, none of any related attributes was found. Upon questioning on the number on whether customers stayed with the company, a total of 759 relevant information were found. Similarly, on the identification of customer multiple lines in both Yes or No, among the total of 682 customers, 48% were none while 32% matched were found. Customer internet service in both fibre optic and DSL, in a total of 7043, 44% customers used fibre optic while 34% were using DSL. The remaining 10% which were 1526 of 7043 customers used other ways. Further, in the context of online security, 50% showed no interest while 29% were shown interest in company services. The other 22% which is 1526 of the totals showed no interest.

3.3 Data Pre-Processing & Transformation

Data prepossessing and transformation focused to increase the validity and quality of Telco Company churn data. In the data preprocessing phase, the data mining technique mainly focused on the transformation of raw data into a standard, useful and efficacious format Cil et al. (2018). All the data from Telco customer churn was labelled according to the customer attributes in different columns. The data preprocessing phase was based on two major considerations.

- Outlier and null value removal
- The use of the Synthetic Minority Oversampling Technique (SMOTE) for balancing the overall Telco customer churn data.

Initially, outliers and null values were removed from Telco customer churn data to decrease variability. The outlier's presence increases variability in the datasets. This can result in a reduction in the statistical power of datasets. Extraction and removal of outliers from the dataset can increase the result's significance and validity Zimek and Filzmoser (2018). In this regard, the study used focused on outlier and null value removal to avoid duplication and provide standard data for the processing and data mining phase. Further, the study applied Synthetic Minority Oversampling Technique (SMOTE) to manage all unbalanced data for machine learning techniques such as Logistic Regression, Random Forest classifier, Decision Tree Classifier, Gradient Boosting (GB) Fernández et al. (2018) Elreedy and Atiya (2019).

While applying the SMOTE technique, all minority classes in 7043 customers and 21 features were identified. The identification helped to determine the nature and impact of unbalanced data on customer churn prediction. The nearest number (k) was decided. All the lines between the minority data and its neighbours were adjusted at the balanced point. 0 to 1 class were used during the data balancing on all features. During the process, 0 classes were used which indicated that the customer has not churned from Telco company while 1 class indicated that the customer has churned. This model was applied to all the 21 features of Telco company churn data.

3.4 Data Mining

By eliminating data noise and using only relevant data, feature selection reduces the input variable to your model.In this study, all relevant data were extracted from Decision Tree, Random Forest, Logistic Regression, and Gradient Boosting. SelectKBest class was used for Feature Selection. A feature is selected according to the k highest score using the SelectKBest method. Both classification and regression data can be analyzed using the 'score_func' parameter.

Further, in the Gradient boosting (GB), RandomizedsearchCV is used to improve the gradient boosting performance during the implementation process. On the Gradient Boosting (GB), RandomizedsearchCV will focus on different sets and their parameters in the Telco Company data, all the scores on each parameter will be calculated which can give hyperparameters with a higher value as output. RandomizedsearchCV provides better output results by transforming parameters into hyper-parameters. The use of RandomizedsearchCV can improve machine learning techniques' performance Kelsingazin et al. (2021).

3.5 Evaluation

The evaluation phase is considered to be the essential phase to understand the overall data mining process before data specification. The use of performance evaluation interprets all kinds of crucial steps during the machine learning and data mining process Hussain et al. (2019). Performance evaluation can use different metrics to evaluate machine learning and deep learning techniques results Xu et al. (2021). In this regard, the study has used performance metrics such as confusion metrics based on Accuracy, Precision, Recall and f1-score to evaluate the results and ML and hybrid modelling performance. Both precision, and accuracy in results provide authenticity of customer churn prediction. Therefore, the evaluation is done after concluding customer churn assessment using Logistic Regression, Random Forest classifier, Decision Tree Classifier, Gradient Boosting (GB) and hybrid modelling which are presented in section 6.

4 Design Specification

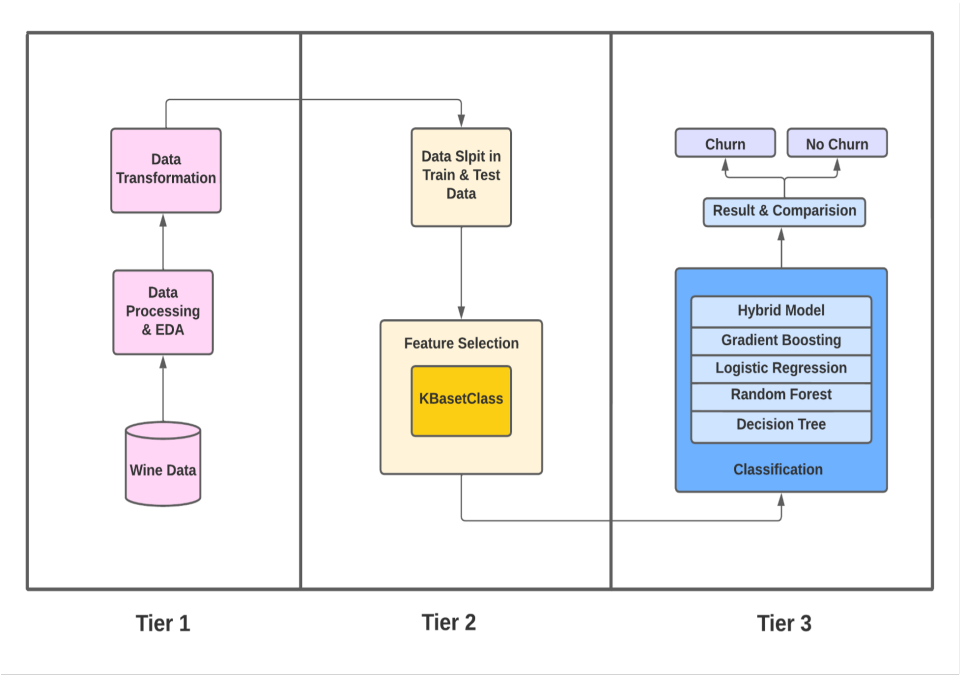

Figure 5: Process Flow Diagram

The completion of the Knowledge Discovery in Database (KDD) method for machine learning (ML) techniques and hybrid modelling allows the study for design specification. In the design specification section, all the study workflow is explained systematically. The design specification provides information about how the entire project has been carried out till the end. In this regard, all the work from initiation to the hand has been graphically presented. The design specification and workflow are crucial for the study Faris (2018).

Above is the project design for Customer churn Prediction in the Telecom Industry.

For the references, Kaggle datasets are widely accepted for their authenticity and availability of large data. Using Kaggle provides data that associates with the study objectives. The use of Kaggle Datasets to predict customer have a high result accuracy rate. Testing over the Kaggle dataset increases efficiency for results. Therefore, the 1st step in workflow was to import the Kaggle dataset to customer churn analysis. The study focused on Telco Company churn data imported from the Kaggle dataset. Finalization of all probable factors before applying machine learning and hybrid modelling algorithms has been done using the KDD method to avoid all kinds of anomalies and outliers from the dataset. The use of outlier removal through SMOTE techniques helped the study avoid duplication and ensure accurate customer data. The probability at 0.80 was 0.20 and with 0 and 1 to determine customer not churn and customer churn provides a parallel line to classify customers into groups using Machine learning classifiers. And all the result are evaluated based on accuracy, precision, recall and f1 score.

5 Implementation

5.1 Decision Tree Implementation

An essential component of machine learning is the decision tree, a technique for both categorizing and regressing problems.Classification of problems is the most common application of

decision trees Ma et al. (2009).The DecisionTreeClassifier function in the scikit learn library was used to implement this approach.Figure 6 displays the decision tree's hyper-parameters. A measure of impurity is gini. Gini impurity is an indicator of how frequently a randomly selected element from the set would be mislabeled if it were randomly classified in accordance with the distribution of labels in the subset. max_depth is another hyperparameter that can be used to control a tree's depth. It does not do any sample proportion or impurity calculations. When max_depth is achieved, the model halts splitting. The minimal number of samples necessary to be at a leaf node is indicated by the variable min_samples_leaf.

```
[ ] # decisionTree Classifier
    Dtc_sampling = DecisionTreeClassifier(criterion = "gini",random_state = 100,max_depth=7, min_samples_leaf=15)
    Dtc_sampling.fit(X_train_sap, y_train_sap)
    dtc_sampling_pred = Dtc_sampling.predict(X_test_sap)
```

Figure 6: Hyper-parameter used for Decision Tree

5.2 Random Forest Implementation

A highly common machine learning algorithm known as Random Forest is a supervised machine learning method. In the domain of machine learning, this approach treats both regression and classification problems Geetha et al. (2020). This strategy was put into practice using the RandomFoestClassidier function from the scikit learn library. Figure 7 displays the hyper-parameters that were used in the decision tree. We can learn about the quantity of trees in the forest from the hyper-parameter n_estimators. Each decision tree's maximum number of levels is indicated by the variable max_depth. The minimum amount of data points permitted in a leaf node is indicated by the variable min_sample_leaf. When a node is divided, min_sample_split informs us of the bare minimum of data points that must be present.

```
[ ] # Random forest classifier
    Rfc_sampling = RandomForestClassifier(n_estimators=150,criterion='gini', max_depth=15, min_samples_leaf=10, min_samples_split=6)
    Rfc_sampling.fit(X_train_sap, y_train_sap)
    rfc_sampling_pred = Rfc_sampling.predict(X_test_sap)
```

Figure 7: Hyper-parameter used for Random Forest

5.3 Logistic Regression Implementation

A classification method that uses a linear model is logistic regression. Maximum entropy classification, logit regression, and occasionally log-linear classifier are other names for logistic regression. The likelihood of the likely results of a single experiment is modeled using a logistic function. This strategy was put into practice using the LogisticRegression function of the scikit learn library.Figure 8 displays the hyper-parameters used in logistic regression.

```
[ ] # logistic regression
    Log_reg_sampling = LogisticRegression(C=10, max_iter=150)
    Log_reg_sampling.fit(X_train_sap, y_train_sap)
    Log_sampling_pred = Log_reg_sampling.predict(X_test_sap)
```

Figure 8: Hyper-parameter used for Logistic Regression

5.4 Gradient Boosting Implementation

A class of machine learning techniques known as gradient boosting classifiers combines a number of weak learning models to produce a powerful predicting model. This approach allows for the optimization of any differentiable loss function and constructs an additive model in a forward stage-wise manner. The optimal parameters for the Gradient Boosting RandomSearchCV method were chosen for the hyper-parameters selection, as shown in Figure 9. This strategy was put into practice using the GradientBoostingClassifier function from the scikit learn library.

```
[ ] # GradientBoostingClassifier
    gbc_tunning = GradientBoostingClassifier(criterion='squared_error', learning_rate=0.5,
                                max_depth=7, max_leaf_nodes=12, min_samples_leaf=11,
                                min_samples_split=5, n_estimators=300)
    gbc_tunning.fit(X_train_sap, y_train_sap)
    pred = gbc_tunning.predict(X_test_sap)
```

Figure 9: Hyper-parameter used for Gradient Boosting

5.5 Hybrid Model Implementation

Hybrid models are often used where it is believed that traditional algorithms are not as efficient. Therefore, multiple algorithms are combined to get better results. This has been proven by previous research works as well, however, the results tend to vary. Other machine learning algorithms are single algorithms and are not a combination of two or more algorithms. Using a hybrid model algorithm for dataset testing to predict customer churn is another key part of the study design. The use of Decision Tree, Random Forest, and Gradient Boosting classifier to analyze customer churn provides combine algorithms. DT and RF work as classifiers to form groups by featuring different points of similarities and differences. In hybrid modelling, the algorithms connect with the nearest scale or point. All the unnecessary data at the points are side-lined for accuracy purposes.

6 Evaluation

As was covered in the section before, various matrices, including precision, recall, f1-score, and accuracy, are used to evaluate the model. For this implementation, a total of 7043 samples were taken into account, with a train and test split of 80:20. There are 5634 train sets and 1409 test sets, or 80% train sets and 20% test sets, respectively. The identical training set and test set were utilized for all studies.

6.1 Experiment 1: Decision Tree

```
Accuracy score : 0.9423407917383821
Precision score : 0.9487179487179487
Recall score : 0.9441786283891547
F1 score : 0.9464428457234213
Confusion matrix :
 [[503  32]
 [ 35 592]]
Classification report :
              precision    recall  f1-score   support

           0       0.93      0.94      0.94       535
           1       0.95      0.94      0.95       627

    accuracy                           0.94      1162
   macro avg       0.94      0.94      0.94      1162
weighted avg       0.94      0.94      0.94      1162
```

Figure 10: Classification Report of Decision Tree

A decision tree classifier was used in the implementation phase to predict customer churn. The use of a decision tree as a machine learning technique predicted customer churn. The accuracy level of the decision tree in customer churn prediction was recorded at 94.23%. The model precision during the testing was 94.87%. During the implementation process, decision trees showed a recall value of 94.41% and the overall F1 score was 94.64%. The use of the decision tree algorithm in customer churn prediction has recorded high accuracy ratio and overall F1 score. The overall score supports the use of a decision tree to predict customer churn.

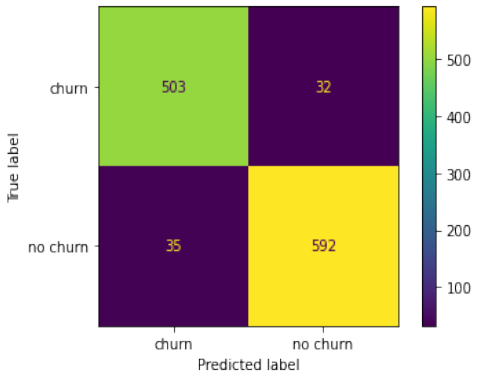

Figure 11: Confusion Matrix of Decision Tree

Figure 11 confusion matrix contrasts the True label with the anticipated label. The Decision Tree model accurately predicted 503 consumers who would not churn and 592 customers who would (TN). 32 churned as anticipated by the churn, but 35 did not.

It is important for the telco companies to retain customers for as long as possible. Decision tree help in achieving the goals of the study. The decision tree helps us to understand that which decision is being taken by the customers for each step. This enables to gather all data and enables to predict the customer behaviour in a successful manner.

6.2 Experiment 2: Random Forest

```
Accuracy score : 0.9526678141135972
Precision score : 0.9615384615384616
Recall score : 0.9508716323296355
F1 score : 0.9561752988047808
Confusion matrix :
 [[507  24]
 [ 31 600]]
Classification report :
              precision    recall  f1-score   support

           0       0.94      0.95      0.95       531
           1       0.96      0.95      0.96       631

    accuracy                           0.95      1162
   macro avg       0.95      0.95      0.95      1162
weighted avg       0.95      0.95      0.95      1162
```

Figure 12: Classification Report of Random Forest

A random forest machine learning technique was used to predict customer churn. During the implementation and evaluation phase, Random Forest recorded an accuracy of 95.26%. The precision of random forest was recorded on 96.15%. The random forest recall prediction value stands at 95.08% while the F1 score is recorded at 95.61%. The use of random forest in customer churn prediction has a higher accuracy rate. The results show that the Random Forest classifier score is higher than the decision tree by 0.97% while the accuracy rate is also high by 1.03%.

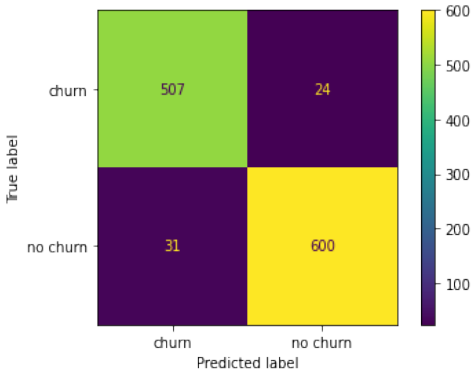

Figure 13: Confusion Matrix of Random Forest

Figure 13 confusion matrix contrasts the True label with the anticipated label. The Random Forest model accurately predicted 507 consumers who would not churn and 600 customers who would (TN). 24 were not anticipated to churn by the churn, while 31 were.

Random forest works on decision trees and makes use of bootstrap aggregation due to which it has a higher efficiency over decision tree. The final outcome is based on the individual trees and their averages. Monthly prediction of customer data can be done easily with better accuracy. Results show it is good model for prediction of customer data.

6.3 Experiment 3: Logistic Regression Classifier

```
Accuracy score : 0.9199655765920827
Precision score : 0.9294871794871795
Recall score : 0.9220985691573926
F1 score : 0.925778132482043
Confusion matrix :
 [[489  44]
 [ 49 580]]
Classification report :
              precision    recall  f1-score   support

           0       0.91      0.92      0.91       533
           1       0.93      0.92      0.93       629

    accuracy                           0.92      1162
   macro avg       0.92      0.92      0.92      1162
weighted avg       0.92      0.92      0.92      1162
```

Figure 14: Classification Report of Logistic Regression

Logistic Regression as a classifier was used to predict customer churn using Telco company data. The model was predicted with an acceptable accuracy rate of 91.99%. The precision value of logistic regression was recorded at 92.94%. The recall value of logistic regression during the implementation phase was 92.20% while the F1 score stands at 92.57%. The overall use of the random forest as a classifier during the implementation phase has shown a 90+ accuracy value.

However, the accuracy rate and precision value are lower than the decision tree and random forest by 2.24% and 2.88% respectively.

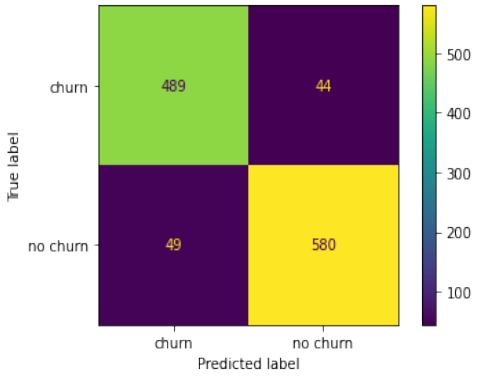

Figure 15: Confusion Matrix of Logistic Regression

Figure 15 confusion matrix contrasts the True label with the anticipated label. The Logistic Regression model successfully predicted 489 consumers who would not churn and 580 who would (TN). 44 churned as anticipated by the churn, but 49 did not.

As it is extension of linear regression therefore the accuracy is not as much as decision trees and Random forest. A stable customer base is key for any telco company and therefore more accurate prediction are important for the telecom industry in terms of customer churn. Prediction can be done using this method, however, the more accurate methods are adopted, the better.

6.4 Experiment 4: Gradient Boosting Classifier

```
Accuracy score : 0.96815834767642
Precision score : 0.9711538461538461
Recall score : 0.9696
F1 score : 0.9703763010408327
Confusion matrix :
 [[519  18]
 [ 19 606]]
Classification report :
              precision    recall  f1-score   support

           0       0.96      0.97      0.97       537
           1       0.97      0.97      0.97       625

    accuracy                           0.97      1162
   macro avg       0.97      0.97      0.97      1162
weighted avg       0.97      0.97      0.97      1162
```

Figure 16: Classification Report of Gradient Boosting

Gradient Boosting (GB) classifier was used during the implementation phase for customer churn prediction and analysis. The results of Gradient boosting show a high accuracy value of 96.81% surpassing random forest, Decision tree and Logistic regression algorithms. The precision value

of gradient boosting was higher at 97.11% while the recall value was 96.96%. Therefore, the F1 score during gradient boosting during the implementation was 97.03%. The F1 score of gradient boosting was higher than random forest, Decision tree and Logistic regression algorithms.

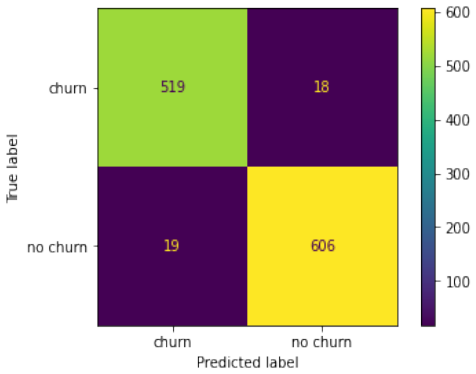

Figure 17: Confusion Matrix of Gradient Boosting

Figure 17 confusion matrix contrasts the True label with the anticipated label. The Gradient Boosting model successfully identified 519 customers who would churn (TP) and 606 customers who would not turnover (TN). 18 churns instead of the 19 suggested by the churn.

There are various internal parameters in gradient boosting which are known as the hyper parameters and it has an impact on the overall model of gradient boosting. Gradient boosting gives us much bigger edge in predicting customer churn and it gives an edge to a company when ensuring the customer retention which is crucial for every business.

6.5 Experiment 5: Hybrid Model

```
Model: HybridModel
Accuracy Score: 0.9569707401032702
Precision Score: 0.9526813880126183
Recall Score: 0.967948717948718
F1 Score: 0.9602543720190779
Confusion Matrix:
 [[508  30]
 [ 20 604]]
Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.94      0.95       538
           1       0.95      0.97      0.96       624

    accuracy                           0.96      1162
   macro avg       0.96      0.96      0.96      1162
weighted avg       0.96      0.96      0.96      1162
```

Figure 18: Classification Report of Hybrid Model

Hybrid modelling algorithms during the testing phase showed an accuracy rate of 95.69%. Similarly, during the testing, the precision value was recorded at 96.26%. The overall recall value of hybrid modelling was 96.79%. In this regard, the F1 score of hybrid modelling stands at 96.02%.

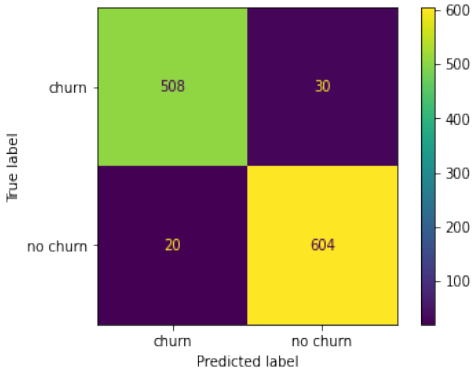

Figure 19: Confusion Matrix of Hybrid Model

Figure 19 confusion matrix contrasts the True label with the anticipated label. The Hybrid model successfully predicted 508 consumers who would not churn and 604 who would (TN). The churn incorrectly anticipated 30 not churning and 20 churning.

Since hybrid model is a combination of models therefore it gives good results. This model is also successful in predicting customer churn the telecommunication industry. However, its results are not as accurate as the gradient boosting.

6.6 Result & Comparison

Model Name	Accuracy	Precision	Recall	F1 Score
Decision Tree	94.23%	94.87%	94.41%	94.64%
Random Forest	95.26%	96.15%	95.08%	95.61%
Logistic Regression	91.99%	92.94%	92.20%	92.57%
Gradient Boosting	96.81%	97.11%	96.96%	97.03%
Hybrid Model	95.69%	95.26%	96.79%	96.02%

Table 1: Comparison of Models

In this section, the result comparison of both machine learning techniques and hybrid modelling is done. The results provide key information to relate all machine learning and hybrid modelling techniques.

- In terms of model accuracy, gradient boosting has recorded a higher value of 96.81% than all techniques including hybrid modelling. Hybrid Modelling accuracy stands at 95.69%. Random Forest, Decision tree and Logistic Regression come at 3rd, 4th and 5th respectively.
- In terms of precision, gradient boosting has a higher value of 97.11% followed by random forest valuing 96.15%.

- In terms of recall value, gradient boosting is higher at 96.96% followed by hybrid modelling at 96.79%.
- F1 score of gradient boosting is higher at 97.03% followed by hybrid model of 96.02%.
- The overall score of gradient boosting is higher due to the additional use of RandomizedsearchCV to improve its performance.

6.7 Result Discussion

The project's main goal was to conduct customer churn analysis using models that produce precise and effective results. A careful analysis of the available literature revealed certain fundamental gaps and limitations. For the analysis of customer turnover in this research project, the machine learning algorithms Decision Tree, Random Forest, Logistic Regression, and Gradient Boosting were used. For the analysis, a hybrid decision tree, random forest, and gradient boosting model was created. With an overall accuracy of 96.81%, the Gradient Boosting algorithm produced the greatest results of all. The same's precision, recall, and f1-score were, respectively, 0.97, 0.96, and 0.97.

7 Conclusion and Future Work

7.1 Conclusions

To conclude, customer churn prediction in the telecom industry has a pivotal role in effective decision making customer retention strategies. The use of AI technologies such as machine learning algorithms and hybrid modelling techniques plays an effective role in customer churn prediction. Companies can use machine learning techniques such as Random Forest, Decision Tree, Logistic regression and Gradient Boosting to predict customer churn and make decisions. The telecommunication industry is among industries with higher customer challenges. Customer switching and rational decisions are massive challenges for companies. Therefore, AI technologies assist companies to make quality decisions by targeting different aspects of customers i.e information. Likewise, hybrid modelling techniques can have an affirmative role in customer churn prediction by scoring above 90%. This indicates that machine learning techniques have revolutionalised the customer churn analysis process for telecommunication companies. Gradient boosting recorded a higher accuracy score of 96.14% due to RandomizedsearchCV for improved performance. Thus, ML techniques in customer churn prediction provide impactful results.

7.2 Future Recommendations

Studies can use machine learning techniques other than Random Forest, Decision Tree, Logistic regression and Gradient Boosting to predict customer churn in the telecom industry. Likewise, hybrid modelling techniques can be customized to predict customer churn in the future. Further, deep learning techniques play a key role in managerial operations. Future studies have a chance to use deep learning techniques and models for customer churn prediction in the telecommunication industry. The use of both ML and DL techniques is not only liable for the telecommunication industry, future studies can test ML and DL techniques by targeting other industries. Future studies have a higher chance to explore factors that effects both machine learning and deep learning techniques in a similar industry.

8 Acknowledgement

Without Dr. Qurrat Ul Ain's oversight, this study endeavor would not have been able to be finished. I want to thank Dr. Qurrat Ul Ain for always guiding me in the proper route and giving me suggestions on how to structure this research. I also want to express my gratitude to my family for their unwavering support over the course of the endeavor.

References

Amin, A., Al-Obeidat, F., Shah, B., Tae, M. A., Khan, C., Durrani, H. U. R. and Anwar, S. (2020). Just-in-time customer churn prediction in the telecommunication sector, *The Journal of Supercomputing* **76**(6): 3924–3948.

Amin, A., Anwar, S., Adnan, A., Nawaz, M., Alawfi, K., Hussain, A. and Huang, K. (2017). Customer churn prediction in the telecommunication sector using a rough set approach, *Neurocomputing* **237**: 242–254.

Amin, A., Khan, C., Ali, I. and Anwar, S. (2014). Customer churn prediction in telecommunication industry: with and without counter-example, *Mexican international conference on artificial intelligence*, Springer, pp. 206–218.

Bayrak, A. T., Yücetürk, G., Bahadır, M. B., Yalçinkaya, S. M., Demirdağ, M. and Sayan, I. U. (2022). Comparative methods for personalized customer churn prediction with sequential data, *2022 IEEE International Conference on Big Data and Smart Computing (BigComp)*, IEEE, pp. 222–225.

Bojer, C. S. and Meldgaard, J. P. (2021). Kaggle forecasting competitions: An overlooked learning opportunity, *International Journal of Forecasting* **37**(2): 587–603.

Brandusoiu, I. and Toderean, G. (2013). Churn prediction in the telecommunications sector using support vector machines, *Margin* **1**: x1.

Choudhari, A. S. and Potey, M. (2018). Predictive to prescriptive analysis for customer churn in telecom industry using hybrid data mining techniques, *2018 Fourth international conference on computing communication control and automation (ICCUBEA)*, IEEE, pp. 1–6.

Cil, F., Çetinyokuş, T. and Gökçen, H. (2018). Knowledge discovery on investment fund transaction histories and socio-demographic characteristics for customer churn, *International Journal of Intelligent Systems and Applications in Engineering* **6**(4).

Coussement, K., Lessmann, S. and Verstraeten, G. (2017). A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry, *Decision Support Systems* **95**: 27–36.

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

Dalvi, P. K., Khandge, S. K., Deomore, A., Bankar, A. and Kanade, V. (2016). Analysis of customer churn prediction in telecom industry using decision trees and logistic regression, *2016 symposium on colossal data analysis and networking (CDAN)*, IEEE, pp. 1–4.

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

Elreedy, D. and Atiya, A. F. (2019). A comprehensive analysis of synthetic minority oversampling technique (smote) for handling class imbalance, *Information Sciences* **505**: 32–64.

Faris, H. (2018). A hybrid swarm intelligent neural network model for customer churn prediction and identifying the influencing factors, *Information* **9**(11): 288.

Fernández, A., Garcia, S., Herrera, F. and Chawla, N. V. (2018). Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary, *Journal of artificial intelligence research* **61**: 863–905.

Geetha, V., Punitha, A., Nandhini, A., Nandhini, T., Shakila, S. and Sushmitha, R. (2020). Customer churn prediction in telecommunication industry using random forest classifier, *2020 International Conference on System, Computation, Automation and Networking (ICSCAN)*, IEEE, pp. 1–5.

Huang, B., Buckley, B. and Kechadi, T.-M. (2010). Multi-objective feature selection by using nsga-ii for customer churn prediction in telecommunications, *Expert Systems with Applications* **37**(5): 3638–3646.

Huang, B., Kechadi, M. T. and Buckley, B. (2012). Customer churn prediction in telecommunications, *Expert Systems with Applications* **39**(1): 1414–1425.

Hussain, S., Muhammad, L., Ishaq, F., Yakubu, A. and Mohammed, I. (2019). Performance evaluation of various data mining algorithms on road traffic accident dataset, *Information and Communication Technology for Intelligent Systems*, Springer, pp. 67–78.

Kelsingazin, Y., Akhmetov, I. and Pak, A. (2021). Sentiment analysis of kaspi product reviews, *2021 16th International Conference on Electronics Computer and Computation (ICECCO)*, IEEE, pp. 1–5.

Kisioglu, P. and Topcu, Y. I. (2011). Applying bayesian belief network approach to customer churn analysis: A case study on the telecom industry of turkey, *Expert Systems with Applications* **38**(6): 7151–7157.

Li, P., Li, S., Bi, T. and Liu, Y. (2014). Telecom customer churn prediction method based on cluster stratified sampling logistic regression.

Lu, N., Lin, H., Lu, J. and Zhang, G. (2012). A customer churn prediction model in telecom industry using boosting, *IEEE Transactions on Industrial Informatics* **10**(2): 1659–1665.

Ma, H., Qin, M. and Wang, J. (2009). Analysis of the business customer churn based on decision tree method, *2009 9th International Conference on Electronic Measurement & Instruments*, IEEE, pp. 4–818.

Mishra, A. and Reddy, U. S. (2017). A comparative study of customer churn prediction in telecom industry using ensemble based classifiers, *2017 International Conference on Inventive Computing and Informatics (ICICI)*, IEEE, pp. 721–725.

Mishra, K. and Rani, R. (2017). Churn prediction in telecommunication using machine learning, *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*, IEEE, pp. 2252–2257.

Prashanth, R., Deepak, K. and Meher, A. K. (2017). High accuracy predictive modelling for customer churn prediction in telecom industry, *International Conference on Machine Learning and Data Mining in Pattern Recognition*, Springer, pp. 391–402.

Rahman, M. S., Alam, M. S. and Hosen, M. I. (2022). To predict customer churn by using different algorithms, *2022 International Conference on Decision Aid Sciences and Applications (DASA)*, IEEE, pp. 601–604.

Rai, S., Khandelwal, N. and Boghey, R. (2020). Analysis of customer churn prediction in telecom sector using cart algorithm, *First International Conference on Sustainable Technologies for Computational Intelligence*, Springer, pp. 457–466.

Sharma, T., Gupta, P., Nigam, V. and Goel, M. (2020). Customer churn prediction in telecommunications using gradient boosted trees, *International Conference on Innovative Computing and Communications*, Springer, pp. 235–246.

Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G. and Chatzisavvas, K. C. (2015). A comparison of machine learning techniques for customer churn prediction, *Simulation Modelling Practice and Theory* **55**: 1–9.

Xu, W., Jang-Jaccard, J., Singh, A., Wei, Y. and Sabrina, F. (2021). Improving performance of autoencoder-based network anomaly detection on nsl-kdd dataset, *IEEE Access* **9**: 140136–140146.

Yihui, Q. and Chiyu, Z. (2016). Research of indicator system in customer churn prediction for telecom industry, *2016 11th International Conference on Computer Science & Education (ICCSE)*, IEEE, pp. 123–130.

Yu, R., An, X., Jin, B., Shi, J., Move, O. A. and Liu, Y. (2018). Particle classification optimization-based bp network for telecommunication customer churn prediction, *Neural Computing and Applications* **29**(3): 707–720.

Zhao, M., Zeng, Q., Chang, M., Tong, Q. and Su, J. (2021). A prediction model of customer churn considering customer value: an empirical research of telecom industry in china, *Discrete Dynamics in Nature and Society* **2021**.

Zimek, A. and Filzmoser, P. (2018). There and back again: Outlier detection between statistical reasoning and data mining algorithms, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **8**(6): e1280.