Telecom Churn Prediction Using Feature Selection and Hyperparameter Tuning

Nagarathna Sali

Dublin Business School

Applied project is submitted for the degree of

MSc in Business Analytics

January 2021

Declaration

I Nagarathna, hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements.

Nagarathna Sali

January 2021

Acknowledgements

I would like to thank my research supervisor Mr. Philip Hickey for his constant guidance and support through out the research. His expert guidance and constant encouragement helped me to make progress in my thesis.I would like to thank all the faculties and staff of Dublin Business School for this amazing learning experience.Finally I would like to thank my parents , siblings and friends for their love and support which kept me focused in my journey.

Abstract

Telecommunication industry has become very saturated with a lot of service providers in the market. With growing competition, the focus has shifted from acquiring new customers to retaining existing customers. Having an efficient prediction model has become crucial to identify potential churners. In this research different classifiers namely Random Forest, KNN, SVC, XGBoost, and ANN are used to predict potential churners. Research compares the efficiency of classifiers using evaluation metrics to find an efficient model. It also emphasizes improving the performance of the models by implementing feature selection and hyperparameter tuning. Feature selection is applied through Random Forest feature importance method.Hyperparameter tuning is implemented with the help of the grid search algorithm. IBM Watson dataset is utilized to carry out the analysis. Results reveal that the XGBoost algorithm outperformed other classifiers with a recall value of 0.75 initially and 0.77 after applying feature selection which shows 2% improvement.

Keywords: Churn prediction, Telecom, Classification, Neural Network, Feature Selection

Table of contents

List of figures vii

List of tables ix

1 Introduction 1

1.1 Context 1

1.2 Motivation 3

1.3 Research Objective and Research Question 4

1.3.1 Research Question: 4

1.3.2 Research Objective 4

1.4 Research Roadmap: 5

2 Literature Review 6

2.1 Literature Based on the Application of Different Machine Learning Algorithms on Telecom Churn Prediction 7

2.2 Literature Based on the Application of Neural Networks for Telecom Churn Prediction 15

2.3 Literature Based on Application of Feature Selection and Hyperparameter Tuning to Improve the Efficiency of the Model 17

2.4 Literature Review Summary 19

3 Methodology **20**
3.1 CRISP-DM Methodology 20
3.1.1 Business Understanding: 21
3.1.2 Data Understanding: 22
3.1.3 Data Preparation: 22
3.1.4 Modeling: 22
3.1.5 Model Evaluation: 28

4 Implementation **31**
4.1 Random Forest: 33
4.2 Support Vector Machine Classifier(SVC): 34
4.3 K-Nearest Neighbor (KNN): 34
4.4 Extreme Gradient Boosting(XGBoost): 35
4.5 Artificial Neural Network(ANN): 36

5 Evaluation **37**
5.1 Random Forest: 37
5.2 Support Vector Machine Classifier(SVC): 40
5.3 K-Nearest Neighbor(KNN): 42
5.4 Extreme Gradient Boosting(XG-Boost): 44
5.5 Artificial Neural Network(ANN): 46
5.6 Results Discussion 47

6 Conclusions and Future Work **50**
6.1 Conclusion: 50
6.2 Future Work 51

References **52**

List of figures

3.1 CRISP-DM 21
3.2 KNN Algorithm 24
3.3 ANN with 2 hidden layers 27
3.4 ANN Architecture 27

4.1 Correlation heatmap 32

5.1 Random Forest results before feature selection 38
5.2 RF Confusion Matrix before feature selection 38
5.3 Feature Importance 38
5.4 Random Forest results after feature selection 39
5.5 RF Confusion Matrix after feature selection 39
5.6 SVC results before feature selection 40
5.7 SVC Confusion Matrix before feature selection 40
5.8 SVC results after feature selection 41
5.9 SVC Confusion Matrix after feature selection 41
5.10 KNN results before feature selection 42
5.11 KNN Confusion Matrix before feature selection 42
5.12 KNN results after feature selection 43
5.13 KNN Confusion Matrix after feature selection 43

5.14 XGBoost results before feature selection 44
5.15 XGBoost Confusion Matrix before feature selection 44
5.16 XGBoost results after feature selection 45
5.17 XGBoost Confusion Matrix after feature selection 45
5.18 ANN results 46
5.19 ANN Confusion Matrix 46

List of tables

5.1 Summary of False Negatives and Recall values for different classifiers . . . 47

5.2 Model Results Before Feature Selection 48

5.3 Model Results After Feature Selection 48

5.4 Final Recall Values Achieved 48

Chapter 1

Introduction

1.1 Context

Communication is the key in this era of social networking, with this telecommunication industry has grown enormously and had gained huge importance. With the increase in the number of service providers, there is huge competition to retain the existing customer base Pamina *et al.* (2019). Customer churn is a vital issue in almost all industries like telecommunication, insurance, banking, etc. Customer churn in telecommunication is nothing but customers leaving the service provider or disconnecting the services provided. Customers often leave service providers due to lack of satisfaction with the existing services or better offers provided with the new service provider at the value customer are ready to pay Tsai and Lu (2009). Hence losing a loyal or long-term customer is potential loss to organisation and acquiring cost of new customer is 5-6 times higher compared to retaining the existing customer Sivasankar and Vijaya (2019). Having an efficient churn prediction model becomes very significant in order to save telecom service providers from huge revenue loss Idris and Khan (2014). To tackle this issue, Customer Relationship Management (CRM) is interested in having prediction model that will predict the customer who is about to leave in advance Chu *et al.* (2007). These prediction models will help in understanding the behavior pattern of potential churners so CRM can create strategies to retain the customers. This process of identifying the customer who is about to leave and causes them to leave is called churn prediction Umayaparvathi and Iyakutti (2016b). Various prediction methods have been proposed by researchers however theses methods work well with specific domains and are unable to show the desired efficiency on telecom datasets due to their complex nature Hadden *et al.* (2007)

In predicting customer churn, data mining techniques are found to be more efficient from past studies carried out. Building an efficient churn prediction model is a very difficult task because it requires knowing the underlying data set to pick a suitable prediction model, pre-processing of data such as handling missing values, handling outliers, and removing any obsolete and unnecessary features, and selecting important features which contribute churn Umayaparvathi and Iyakutti (2016a).

This research is focused on addressing the complex nature of Telecommunication datasets like class imbalance, large volume of data, and high dimensionality, which are hurdles in getting accurate predictions. The class imbalance of the dataset will be handled with SMOTE oversampling technique. Feature selection is performed by applying Embedded methods with the help of Random Forest Feature importance. To address the nature of Large volume of data, Models like Artificial Neural network is used to classify which tend to perform well with large volume datasets. For the purpose of analysis, the IBM sample dataset is considered with 7043 records. Multiple classifiers implemented in this research are 1) K-nearest Neighbor 2) Support Vector Machine Classifier 3) Random Forest 4) Extreme Gradient Boosting 5) Artificial Neural Network to predict the customers who are about to leave. These models' performance will be evaluated using the evaluation metrics like 1) Accuracy score 2) Precision 3) Recall and 4)F1-score and Confusion Matrix.

1.2 Motivation

Telecommunication market is highly competitive with so many service provider trying to sustain in the saturated market. Customer churn is a vital issue faced by almost all telecommunication service providers all over the world. Customer churn is nothing but customers leaving the service provider due to dissatisfaction with services provided or for better offers provided by another service provider at a price customer is ready to pay. This issue is very crucial since acquiring a new customer is much costlier than retaining the existing ones.

Hence Customer Relationship management is keen on having a churn prediction model that can predict possible churners well in advance so that service providers can develop strategies to retain customers. There has been a lot of research in this area. However, the models have proven to be inefficient in real-world applications due to the telecommunication dataset's complex nature. Hence, while building the prediction model, it is essential to consider imbalance, high dimensionality, and large volumes of data.

There have been many churn prediction models with standard classifiers like Logistic regression, Support Vector Machine, K-nearest neighbor, Random Forest. However, these classifier's efficiency depends on the selection of important features that can improve efficiency. Researches came up with an application of classification along with multiple feature selection methods to find the best combination. Extreme gradient boosting was also efficient in overcoming overfitting and used in many researches. There are also Hybrid models introduced combining classification algorithm with Boosting methods, Hybrid methods with feature reduction techniques, and classification so on. Researches done in past were successful at improving the model efficiency but could not come up with a standard model. Researches conducted based on application of Neural networks considering the large volume of telecom datasets indicate that neural networks are proven to be efficient. Researchers also trying to explore hyperparameter tuning method to improve the efficiency of the model.

After considering all the above points, This research aims to answer the research question mentioned in 1.3

1.3 Research Objective and Research Question

Due to enormous competition, the telecommunication market has saturated. Hence, the service provider's focus has shifted from acquiring a new customer base to retaining the existing customer base. With so many service providers, customers are not hesitant to switch from one service provider to another if they are not satisfied with the services provided or find better offers elsewhere. Retaining customers has become a vital issue for telecommunication service providers. Having an efficient predictive model that can predict potential churners well in advance will help telecom service providers to focus on the key areas in retaining customers.

1.3.1 Research Question:

How can different machine learning classifiers be applied to predict customer churn in telecom use case? Sub Question: How feature selection and hyperparameter tuning help improve the efficiency of different machine learning classifiers?

1.3.2 Research Objective

The primary objectives of this thesis are as follows:

1. To predict potential churners using different machine learning classifiers

2. To carry out a comparative analysis of different machine learning classifiers to identify best performing model

3. To implement feature selection and hyperparameter tuning to improve the performance of models

1.4 Research Roadmap:

The rest of the paper is organized as follows. In chapter 2, the literature review of churn prediction will be discussed. In chapter 3, the methodology of the proposed churn prediction model. In chapter 4, Implementation details will be discussed. In chapter 5 details about Evaluation of the models. In chapter 6, the conclusion and future scope of the work will be discussed.

Chapter 2

Literature Review

Customer churn is one of the trending topics as it is very important to retain the customer in a competitive telecommunication industry with so many service providers out there in the market. Prepaid customers want competitive pricing, value for the money provided, and high-quality service else they will not hesitate to switch service providers without notice as they are not bound by any contracts. Acquiring new customers costs 5-6 times more than retaining old customers MOUNIKA REDDY (2016), Qureshi *et al.* (2013).

In order to sustain in the competitive telecommunication market, Customer Relationship Management (CRM) focus on understanding the behavior pattern of the potential churners, and hence there is a lot of focus on having an efficient predictive model that will predict the customers who are about to churn so telecommunication service providers can act upon it accordingly Idris and Khan (2014).

churn prediction models help in analyzing the behavior pattern by analyzing historical data to understand the potential churners. This will help telecommunication companies to have a better retention strategy by focusing on the customer group who are about to leave instead of investing in retention strategies for the entire customer group as they have a huge customer base. Having efficient prediction models will help to recognize such customers and to build strategies to retain them Dalvi *et al.* (2016).

In the past data mining techniques have proven to be effective in churn prediction. however to build an efficient churn prediction model requires a lot of effort right from understanding the underlying dataset, performing data preprocessing, and picking the right models Umayaparvathi and Iyakutti (2016a). In this section, we will go through the past researches to understand the suitable models for telecommunication datasets considering their high volume and dimensionality in nature. These past researches will give a holistic picture of the work done in the past, models used, and their efficiency. By reviewing past research will help in understanding the gap in the researches and scope for future work to be done.

In this section, customer churn key concepts pertaining to this thesis, a review of past work, and research done related to telecom customer churn will be carried out. The literature review is divided into the following subsections mentioned below:

- Literature Based on the Application of Machine Learning Algorithms on Telecom Churn Prediction
- Literature Based on the Application of Neural Networks on Telecom Churn Prediction
- Literature Based on Application of Feature Selection and Hyperparameter Tuning to Improve the Efficiency of the Model

2.1 Literature Based on the Application of Different Machine Learning Algorithms on Telecom Churn Prediction

Research carried out by Brânduşoiu *et al.* (2016) points out that due to globalization customer churn has become a huge problem in many businesses specifically in telecommunication where customer churn is more than 30. The research was carried using call detail dataset containing 3333 customers and 21 attributes and tried to come up with an effective churn prediction model using techniques dimensionality reduction technique with the help of PCA and then machine learning algorithms support vector machines, Bayesian networks and neural networks are applied.

Dahiya and Bhatia (2015) Research explains the competitiveness of the telecommunication industry and with the subscription-based model retaining customers has become very crucial as customers will change the service provider easily if they are unhappy with the existing service provider. As the acquiring cost of a new customer is huge compared to retaining existing ones it is very important to have an efficient churn prediction model. In this research WEKA, data mining software and three datasets with varying sizes and attributes were used. the efficiency of the decision tree and logistic regression was compared. Results indicate the decision tree technique found to be more efficient and research points out possible feature work based on usage of hybrid classification for churn prediction.

Research conducted by Khan *et al.* (2015) stresses how the cost of acquiring a new customer is five times more than retaining the existing customer. Hence it is very crucial to identify the early warning signs of possible customer churn and take preventive action by assigning the churn score to each customer. Brute force feature engineering has been implemented to find features that are significant in customer churn and supervised machine learning algorithms like Random Forest, Support Vector Machine, k-nearest neighbors, Ada Boost, Logistic Regression, and Baseline Model are applied on these features and results are promising. South Asian mobile subscriber dataset was used in prediction. Research points out feature scope in terms of a more systematic approach needed for feature selection as at present lot of manual intervention needed from an analyst.

In the paper presented by Idris *et al.* (2012a) explains due to intensive competition in telecommunication markets, it is very important to have timely information about the customers who are about to quit. It is very crucial to have an efficient churn prediction to retain the customer to avoid the extra cost of acquiring new customers. the different tree-based ensemble classifiers have been widely used for churn prediction. These classifier's efficiency has been affected due to enormous data size, imbalanced datasets, and high dimensionality. In this paper, genetic programming based ADA boosting with ten-fold cross-validation was uniquely employed and an AUC of 0.89 was achieved with this model. Results show that GP with Adaboost to be efficient in Telecom churn prediction.

The study conducted by Zhang *et al.* (2007) proposed a hybrid model with k-nearest neighbor algorithm and Logistic Regression which provided an alternative solution for binary classification. This model increased the efficiency of logistic regression in cases where the complex non-linear relationship between predictor and target variables are present. The results based on four datasets indicate that the hybrid model works well in comparison with well-known classifiers.

Idris and Khan (2014) Stress on having a churn prediction model to retain customers in the highly competitive telecommunication market. The high dimensionality of the telecom dataset is one of the hurdles in achieving high accuracy. The proposed model explores the feature selection method through minimum redundancy and maximum relevance(mRMR) as the first step in prediction and diverse ensemble classifiers Random Forest, Rotation Forest, and KNN are applied in the second step and the majority of the vote is considered for prediction. The datasets which are considered to analyze are telecom datasets Orange and Cell2Cell datasets. The evaluation metrics considered are specificity, sensitivity, AUC, and Q-statistic based measures. Results suggest that with suggested approach research was able to come up with an efficient prediction model with mRMR feature selection and majority voting based ensemble method.

Research carried out by Qureshi *et al.* (2013) applied different machine learning algorithms to classify churners from non-churners.Algorithms like linear regression, artificial neural networks, k-means clustering, decision trees including CHAID, Exhaustive CHAID, CART, and QUEST. The dataset contained 106,000 rows of data containing usage behavior data of 3 months. research also addressed the class imbalance method using various re-sampling methods. The dataset was divided into train and test datasets of 70% and 30% respectively. Results indicate that Exhaustive CHAID algorithm was more efficient in predicting potential churners with an overall accuracy of 75.4%.

Vafeiadis *et al.* (2015) Conducted comparative study of different machine learning techniques for churn prediction in telecommunication using public dataset. Machine learning algorithms ANN, SVM, Decision trees, Logistic regression and Naïve Bayes efficiency were evaluated. except for Naïve Bayes and Logistic Regression on which boosting cannot be applied other algorithms were applied boosting using AdaBoost M1 algorithm to evaluate the impact of boosting on these algorithms. Results revealed that application of boosting improved accuracy between 1% and 4% and F-measure increased between 4.5% and 15%.SVM with AdaBoost performed well overall with an accuracy of 97% and F-measure over 84%. Results revealed the improvement in the efficiency of models with boosting. Feature scope can be extended to see additional boosting algorithm efficiency on telecom datasets.

Research conducted by Subramanya and Somani (2017) for customer churn for e-retailer states that customer churn is an important research topic. with this research, they aim to come up with an efficient prediction model by considering feature selection algorithms and to find out important features for customer churn. Research states that machine learning classifiers (both linear or non-linear) like Support Vector Machine, Artificial Neural Network, and Tree based ensemble methods are primary choices in predicting customer churn. Results indicate that Regularised Linear Regression, Support Vector Machine, Gradient Boost, and Random Forest are efficient.

Azeem and Usman (2018) Points out the limitations of existing prediction models like feature selection did not take into consideration the important features of call details record, Telecom large data sets, and categorization of potential churners to create retention strategy.

research came up with Fuzzy classifiers to overcome the above-mentioned limitation. a dataset from a telecom company in South Asia was considered for analysis. Fuzzy classifiers FuzzyNN, VQNN, FuzzyRoughNN, and OWANN were compared with standard classifiers like Decision Tree, SVM, Linear Regression, Neural Network, AdaBoost. Results showed fuzzy classifiers performed better than standard classifiers with an AUC score of 0.68 with a true positive rate up to 98%.

Lu *et al.* (2012) Study indicates the importance of customer relationship management (CRM) with the globalization in information technologies especially in the telecommunication industry where the market is becoming digitalized and with that, there is an emphasis on the churn prediction model. This study makes use of Boosting algorithms not as a source to increase the efficiency of base learners but the boosting algorithm is used for clustering. The clustering is implemented and customers are classified into high risk and low risk with the weights assigned by boosting algorithms. the churn prediction model is applied to each cluster upon forming the cluster, logistic regression is used as the basis learner. The results are evaluated with results from single logistic regression and the Boosting algorithm proves to be efficient in classifying the churners.

Study conducted by Hung *et al.* (2006) emphasize the importance of customer retention strategy as part of customer management. It states it is very important not only to have an effective churn prediction model but also to act on customer retention strategy. In this study different technique was proposed by including customer service and customer complaint log for building prediction model which were proposed by the previous study **?**. The data which was analyzed for this research was wireless telecommunication data from Taiwan providers. As part of the empirical evaluation, this study compared various data mining techniques. Results show both Decision Tree and Neural Networks techniques were successful in accurately predicting churn.

Literature Based on the Application of Classifiers on Telecom Churn Prediction

In the research carried out by Yabas and Cankaya (2013) about subscription-based churn in mobile and wireless service providers. Public dataset orange telecom from Kaggle was used and aimed to efficiently predict the subscriber who is about to churn. With the previous research, it was evident that preprocessing steps increase the efficiency of classifiers however every classifier works best with a particular preprocessing method. So in this research, they came up with a pair of ensemble classifiers and best suited pre-processing method pair and compared result with individual classifiers, and the new pair of classifier and preprocessing method showed the better result. But, this leads to a subset searching problem to find the best suited preprocessing method for ensemble classifier.

Idris *et al.* (2012b) Study indicates the importance of having an effective churn prediction model to handle the fierce competition in the telecommunication industry. There are a lot of hurdles in achieving high accuracy with a churn prediction model for telecom datasets because of large datasets, imbalanced nature, and high dimensionality. The study explored Particle Swarm Optimization (PSO) undersampling method to overcome the class imbalance. Multiple feature reduction techniques like Principal Component Analysis (PCA), Fscore Fisher's ratio, and Minimum Redundancy and Maximum Relevance (mRMR). After achieving balance and optimal feature set K - Nearest Neighbor (KNN) and Random Forest (RF) classifiers are evaluated using evaluation metrics. Results indicate that PSO undersampling technique, mRMR feature selection, and Random forest classifier combination prove to be effective for large telecom datasets.

Ullah *et al.* (2019) Research to investigate the efficiency of different classifiers and clustering methods to predict customer churn and the factors which are affecting customer churn. feature engineering is done with information gain and correlation attribute ranking filter. different classifiers were implemented to classify the customer which are churning and nonchurners. classifiers implemented include Random Forest, Decision Stump, J48, and Random

Tree with 10-Fold cross-validation and hybrid classifiers like AdaboostM1+DecisionStump and Bagging + Random Tree algorithms. This paper also explores clustering to segment the customers who are churning to group them on different features to which will help in providing group offers to retain customers. Research result revealed that Random Forest performed well with 88.6% accuracy compared to other classifiers. The research stated the study can be further extended to explore artificial intelligence for predicting as they perform well with large tabular datasets.

Idris and Khan (2012) Research uses ensemble classifiers which are effective in predicting churners along with feature extraction method to address the high dimensionality of telecom datasets. implements different feature extraction methods in collaboration with ensemble classifier to find out the best combination to predict the churners in the telecommunication industry effectively. The ensemble classifier methods used are Random Forest, RotBoost, and Rotation Forest. The Feature extraction methods used are mRMR, Fishers' Ratio, and FScore.The results indicate that mRMR was very effective in reducing the feature set from 76 to 31 which helped in dealing with the high dimensionality of the data set which in turn allowed the RotBoost method to learn about the possible churn behavior in predicting.mRMR was also effective in extracting discriminating features which helped RotBoost attain the highest accuracy. hence mRMR in combination with RotBoost proves to be very effective in dealing with the high dimensionality of the telecommunication dataset.

Research conducted by Mishra and Reddy (2017) states the importance of the churn prediction model in domains like banking, insurance, and telecommunication industry. It is very crucial to predict the possible churners in telecommunication to retain the customers. The dataset analyzed for the research is UCI provided educational purpose machine learning dataset with 3333 records. This study conducted an evaluation of churn prediction model using ensemble-based classifiers like Bagging, Boosting and Random Forest also compared the ensemble methods with existing classifiers like Support Vector Machine, Naïve Bayes classifiers, Decision Tree to find the efficient model. Results indicate that Random Forest proves to be an efficient model in predicting customer churn in terms of accuracy, sensitivity, and error rate. Random Forest Classifiers achieved 91.66% Accuracy.

Hanif (2019) Conducted research to find the efficiency of Extreme Gradient Boosting algorithm to retain customers from churning by predicting possible churners in advance. Extreme gradient boosting proves to be efficient on telecommunication datasets that are imbalanced in nature. research conducted an evaluation of prediction efficiency of XGBoost with Logistic Regression algorithm. Results indicate that XGBoost algorithm was efficient in classifying possible churners from nonchurners with higher accuracy, sensitivity, and ROC curve compared to Logistic regression.

Ahmad *et al.* (2019) Research implemented different machine learning techniques on a big data platform to build a prediction model based on SyriaTel telecom company dataset of customer data for 9 months. The research implemented classification algorithms Random Forest,Decision Tree, Gradient Boosting machine Tree, and XGBoost. The dataset was divided into train and test datasets of 70% and 30% respectively with 10 fold cross-validation. Hyperparameter optimization and feature engineering was implemented to increase the efficiency of the models and the class imbalance was handled by the undersampling technique. Results indicate XGBoost algorithm was very efficient in predicting the possible churners with an AUC value of 93.301%.

Pamina *et al.* (2019) Conducted comparative study of machine learning classifiers Random Forest, K-Nearest Neighbor, and Extreme gradient Boosting to predict the customers who are about to churn.IBM Watson dataset was considered in this research for analysis purposes. The dataset has 7043 rows and 21 attributes and it was analyzed to predict potential customers who are about to leave. A comparative study of these well-known classifiers was performed with evaluation metrics like Accuracy and F Score. Results indicate that Extreme gradient Boost proves to be efficient with an Accuracy of 0.798 and F Score of 0.582.

Based on the literature review done on different machine learning classifiers like Support Vector Machine, K-Nearest Neighbour, and Tree-based ensemble classifiers like Random Forest and Extreme Gradient Boosting are primarily used to predict customer churn.

2.2 Literature Based on the Application of Neural Networks for Telecom Churn Prediction

Research carried out by Sharma *et al.* (2013) emphasize on how it is cost-effective to retain loyal customers than acquiring a new customer base and stress having a churn prediction model to retain the customers. This research analyzed the cellular wireless service provider churn dataset provided by California Irvine university machine learning database from UCI repository. This research proposes a Neural Network-based churn prediction model and tested churn prediction with different neural network topologies. The result indicates the medium-sized Neural network that is Neural Network with single hidden layer performed well in terms of efficiency. The accuracy achieved was more than 92%. The future work indicates the data preprocessing steps like feature reduction or feature selection methods to achieve the efficiency of the model.

Umayaparvathi and Iyakutti (2017) Conducted research to find out the efficiency of deep neural networks in predicting customer churn in telecommunication. Since telecommunication market has reached saturation it is very important to have an efficient prediction model in place to find out potential churners. The efficiency of these prediction models highly dependent on feature selection. There are a lot of feature selection methods available but it becomes highly difficult to manually select features for each dataset and it is a time-consuming process. Hence this research explores the efficiency of deep neural networks which does not require feature engineering as it is automatically taken care by Deep Neural networks. The two datasets CrowdAnalytix and Cell2Cell were considered for analysis. Deep Neural Networks used are Baseline Neural Network, Small Feedforward Neural Network (SFNN), Large Feedforward Neural Network (LFNN), Convolution Neural Network (CNN). Results show that Deep Neural Network performed well in comparison to traditional classifiers SVM or Random forest. Using Baseline Neural Network, an accuracy of 92.77% for CrowdAnalytix dataset and accuracy of 71.27% for Cell2Cell dataset was achieved.

Study conducted by Keramati *et al.* (2014) discusses the importance of classifying customer who is not about to churn to sustain in the competitive telecommunication market. The study focus on having a churn prediction system and the same has been employed by using different data mining techniques with Decision Tree, ANN, SVM, and KNN, evaluation of their performances. The dataset analyzed was Iranian mobile company data. Results show that the Artificial Neural Network (ANN) significantly outperformed the other three models. Also, the research focused on extracting important features by a new dimensionality reduction method using a Decision Tree classifier. The study not only compared each model but studied the behavior of each model to come up with a hybrid model that showed improvement to values of evaluation metrics.

Based on the literature review done based on the application of Neural Networks for churn prediction indicate that Neural networks are proven to be efficient in handling large volume of telecom datasets.

2.3 Literature Based on Application of Feature Selection and Hyperparameter Tuning to Improve the Efficiency of the Model

Hyperparameter Tuning to Improve the Efficiency of the Model

Customer churn model efficiency can be improved by hyperparameter settings. These hyperparameter settings can be different for different machine learning algorithms even though the same dataset is being used. So it is very important to find the optimal value of hyperparamerter settings to increase the efficiency of the model. In this research, the different hyperparameter settings are explored to find the optimal values for each model.

To increase the efficiency of the machine learning models it is very important to set the parameters of the model to the optimal value according to the dataset which is being analyzed and this process is called hyperparameter tuning. Often most of the machine learning algorithm's efficiency will be increased by hyperparameter settings. the study conducted by Ahmad *et al.* (2019) changed the hyperparameter values of Random Forest algorithm in each run by changing the values of n_estimators that is the number of trees to be grown in the forest. The results indicated that the best value of n_estimators was found at 200. The research also changed hyperparameter of other machine learning algorithms used. XGBoost was performed well with n_estimators value of 180. The study states that hyperparameter optimization was performed to achieve better outcomes from the model.

Research carried out by Li and Marikannan (2019) focus on Hyperparameter tuning to improve the efficiency of prediction models for telecom customer churn. The public dataset used to analyze was obtained from Kaggle containing 3333 records with 21 attributes. Research mainly focused to determine the ideal hyperparameter settings for Naive Bayes, Decision Tree, and configuration settings for Artificial Neural Networks to optimize the performance of these models. Ten-fold cross-validation method and Grid search method was utilized to find out the hyperparameter for each model which will improve the efficiency. Results indicated that each model has obtained an accuracy of more than 80% with hyperparameter tuning and configuration settings. ANN outperformed the other two algorithms with an accuracy of 86.82%.

Based on the literature review, hyperparameter tuning can help improve the efficiency of the model if they are set at an optimal value suitable for the model and dataset utilized.

Feature Selection to help Improve the Efficiency of the Model

In this research, Feature selection is performed to increase the efficiency of the models. Feature selection is applied through Random Forest feature importance method. Feature selection through Random Forest comes under the category of embedded methods. These embedded methods combine both filter and wrapper methods. These methods are more efficient in terms of their accuracy and they are easily interpretable Dubey (2018).

Saraswat and Arya (2014) Research conducted for the classification of leukocytes using Random Forest state that the selection of important features plays a very important role in improving the efficiency of the model. This research introduces a Novel binary Random Forest feature selection method based on Gini importance. Results indicate that Random Forest based feature selection was efficient in reducing irrelevant features and increasing the efficiency of the Random Forest classifier and proved to be more efficient than other feature reduction techniques. However, the efficiency of this method depends on the performance of RF and values chosen in Random Forest

Research conducted by Lee *et al.* (2010) for spam detection used Random Forest which took feature importance and the optimal number of variables into consideration which was missing in earlier studies. The research utilized spam base dataset for analysis. Random Forest classifier generates important features and generates numeric values for each feature. These features are ranked in descending order of feature importance. property feature_importances_ denotes feature importance based on impurity, higher the value higher the importance of the feature. once the Random Forest is run, features with lower values of importance can be removed and the Random Forest classifier can be run again to improve the efficiency. Results denote that Random Forest proved to be efficient in detecting spam by lowering processing overhead and increased high detection rates.

There are few studies that focused on the ensemble based feature selection process Saeys *et al.* (2008). There are studies that used Random Forest classifier for choosing important features to improve the efficiency of the model Kursa and Rudnicki (2011). After reviewing all these research papers, Random Forest based feature selection which is based on feature importance will be utilized in this research to increase the efficiency of the models.

2.4 Literature Review Summary

From the overall literature review conducted, The research will be implementing classifiers Support Vector Machine, K-Nearest Neighbor, Random Forest, Extreme Gradient Boosting, and Artificial Neural Network to come up with a standard prediction model which will address the complex nature of telecom datasets such as class imbalance, high dimensionality and large volume of data. Feature selection and Hyperparameter tuning will be applied to help improve the efficiency of the model further.

Chapter 3

Methodology

3.1 CRISP-DM Methodology

In this research project, The methodology used is Cross Industry Process for Data Mining methodology(CRISP-DM). This method provides a framework that is flexible with industry and technology used. This methodology provides a structured approach to follow during research. This follows a sequence of steps or phases that can be performed and reiterated if needed as shown in figure3.1 Wirth and Hipp (2000).

These steps are as follows:

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Evaluation and Results
6. Deployment

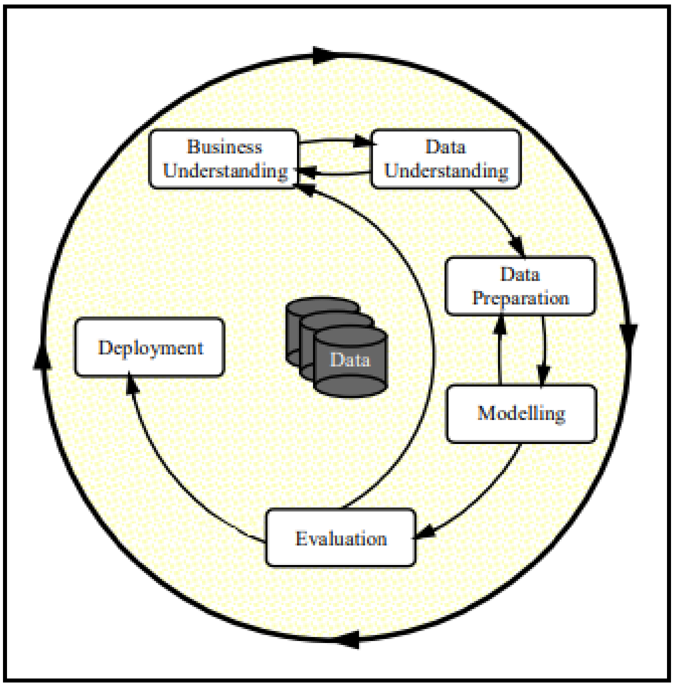


Fig. 3.1 Phases of CRISP-DM Process Model Wirth and Hipp (2000)

3.1.1 Business Understanding:

The first part of the methodology is understanding the Business aspect of the research. Customer churn is one of the crucial issues in telecommunication market all over the world. According to telecommunication terminology churn is nothing but customers leaving a particular service provider or unsubscribing from services due to dissatisfaction in terms of service provided or better offers provided by competitors for the price customer can afford. This will be a huge loss to service providers since the acquiring cost of a new customer is 5-6 times higher than retaining the existing ones. Hence Customer Relationship Management(CRM) is focused on making customers happy with new offers or better services. However, it is very crucial for CRM to know the customer who is about to leave so that strategies to retain those customers can be formedIdris and Khan (2014). The process of knowing customers who are about to leave through prediction models is called churn prediction model.

3.1.2 Data Understanding:

The dataset considered for evaluating the proposed model is IBM Watson Dataset. IBM Watson dataset is obtained from IBM community datasets released in 2015. The dataset has 21 attributes and 7043 instances including both numerical and categorical data. The last column is churn column, out of 7043 instances 5174 instances of churn "No" values and 1869 instances of churn "Yes" values are present IBMCommunity (2015), The dataset is in CSV format.

3.1.3 Data Preparation:

Data Preprocessing is one of the essential step in model building. After the dataset is collected, Data may contain noise as well as irrelevant data so it is important to clean the data before proceeding further. This process of preparing the data by applying preprocessing steps is called data cleaning. This is very important to make the data suitable for the model so models can generate accurate predictions. This process may involve the following steps 1) Checking for any missing values and handling them appropriately 2) converting categorical variables into numerical values 3) Removing redundant values and features to avoid overfitting 4) checking for any outliers in the data and handling them accordingly. In this research, The dataset which is in CSV format is uploaded to the Google Colab environment and is converted to a data frame using python pandas. The preprocessing steps are applied so that the dataset is made ready for model building. The dataset is split into training and Testset of 70% and 30% respectively

3.1.4 Modeling:

Multiple machine learning Classifiers are applied to the cleaned pre-processed data to predict the churn in the dataset used. The models used in this research project are:

Random Forest(RF):

Random Forest is a popular supervised machine learning technique used for both classification and regression problems. Random Forest classifier is an ensemble learning algorithm. It is one of the ensemble of decision tree which works on majority voting techniques. Random Forest classifier grows set of decision trees based on randomly selected subsamples of training data, N decision trees are constructed based on these N bootstrapped subsamples and then it uses majority voting to decide the final class of the test data. In Random forest classifier bootstrap set to 'True' by default which means the dataset will be divided into subsamples to grow trees else, entire dataset is considered.

In Random Forest trees are allowed to grow deeply without pruning until the forest is grown with multiple trees. To make sure one tree differs from other, only a small random sample of predictors is considered at each split this random sample size is the square root of the total number of predictors. Random Forest method uses multiple trees to train data which makes it a more accurate model. Random Forest is a popular approach because of its accuracy. It is also a preferred model due to its high tolerance for abnormal values and noiseLiaw *et al.* (2002)

Support Vector Machine Classifier(SVC):

It is one of the most popular supervised machine learning algorithms used for both classification and regression problems. SVC model is efficient in higher dimensions. It separates data with decision boundary or hyperplane into two different non-overlapping classes. SVM uses kernel trick to transform the data given to make it suitable for separating it into different classes. Feature of SVM is that it creates a hyperplane or line in such a way that the margin between hyperplane and data points is maximum to reduce the error in classification. Hence SVM has superior generalization performance in churn prediction.SVM performance mainly depends on the kernel selected. By default, SVM uses linear kernel function other kernel is Gaussian, polynomial, and binomial.

The important hyperparameter in the SVC classifier is C which is nothing but the regularization parameter which tells the model how much the SVM should avoid misclassifying. The larger value of C tells optimization to have a hyperplane with a smaller margin and a small value of C means the larger margin of hyperplane to classify the training points. The higher value of C will tend to overfit the model.

K-Nearest Neighbor(KNN):

The K-Nearest Neighbor is a supervised machine learning algorithm that is used for both classification and regression. KNN classifier works based on the k value, it classifies data points based on its neighbor so it is always best to consider a k value greater than one to make correct classification. The value of k will decide the number of data points from a specific point and it classifies every new data point based on the majority of votes from its neighbors.

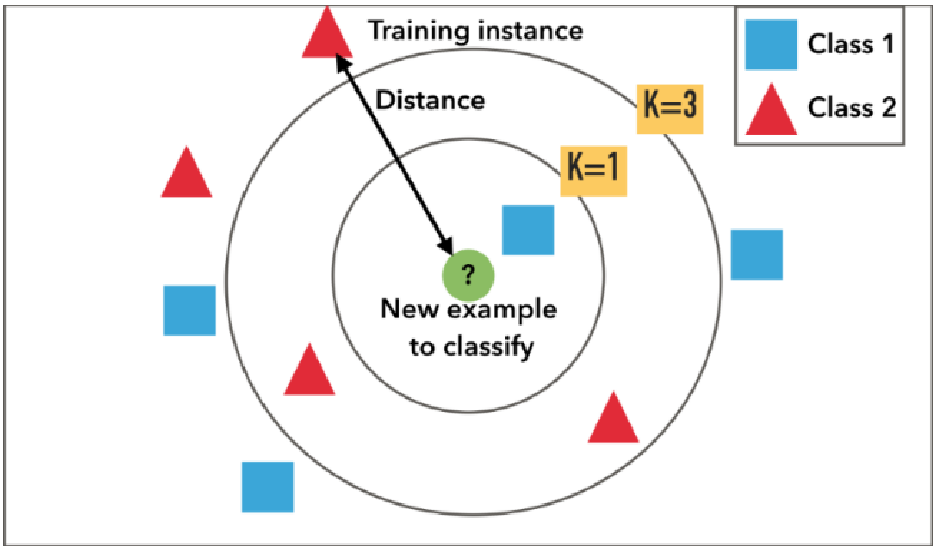


Fig. 3.2 k-nearest neighbor algorithm Bronshtein (2017)

KNN classification works in two steps, at first, the neighboring data points are taken into consideration based on the k value that is the number of neighbors, and in the next step the classification is done based on the majority of the class of these neighbors belong to. Neighbour is decided based on the distance measure parameter and the default value of this is ‘minkowski’ which is nothing but Euclidean distance Ajit (2016),Euclidean function is

$$D = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2} \tag{3.1}$$

where D is the distance x and y are data points

Extreme Gradient Boosting(XGBoost):

XGBoost stands for Extreme Gradient Boosting. XGBoost is an implementation of gradient boosted decision trees designed for speed and performance. It combines traditional gradient boosting with regularization. It is a supervised and ensemble learning method that combines trees to give a more generalizable machine learning model. XGBoost classifier is generally built by many numbers of trees and averaging the results for a higher prediction.

XGBoosting algorithm is an advanced version of the gradient boosting method which literally means extreme gradient boosting that is designed to focus on computational speed and model efficiency. It is one supervised ensemble machine learning method. XGBoost Algorithm was designed as a research project at the University of Washington. It can be used for regression, classification, ranking, and user-defined prediction problem. It is best suited for small to medium tabular datasets. It is a decision tree based ensemble machine learning algorithm that uses gradient boosting framework. XG has features that improve speed and efficiency

- Regularisation – helps in preventing overfitting
- Handling sparse data

- Block structure for parallel learning - XGBoost can make use of multiple cores of CPU. Data is stored in the memory unit and called blocks and this enables to re-utilize in iterations instead of computing again.
- Out of core computing -optimizes the available disk space and maximizes its usage while handling large datasets
- It implements distributed computation method to evaluate any complex modules
- Cache optimization to make best use of hardware

when compared to Gradient Boosting, XGBoost avoids overfitting by using regularisation methodology and hence enhances its efficiency

Artificial Neural Network(ANN):

Artificial Neural Networks (ANN) often known as Neural Network, is a machine learning algorithm that was inspired by a biological neural system. ANN mimics the behavior of the human brain to solve the complex data-driven problems. The ANN has three layers an input layer, a hidden layer (ideally one or two), and an output layer, each layer is fully connected with the other. figure 3.3 shows a typical example of a fully connected Artificial Neural Network with two hidden layers

Each layer consists of neurons or nodes, the number of input neurons depends on the number of target features in the output layer. Each connection has a numeric number associated with it which is called weight. Each node adds up input weights and bias. This will be sent as input to the next node finally output layer will generate output and this will be compared with the actual result. If the output has too much error then the backpropagation algorithm will be applied. The weights values will be changed to reduce the error.

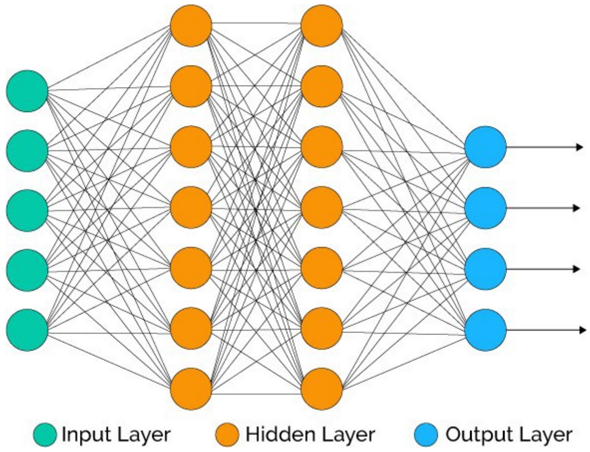


Fig. 3.3 Artificial Neural Network with two hidden layers, Source : McDonald (2017)

There are number of activation functions, they are ReLU(rectified linear unit) , sigmoid, TanH / Hyperbolic tangent, Leaky ReLU, Parametric ReLU, Softmax, Swish The architecture of ANN is defined as shown in the below diagram

$$h_i = \sigma\Big(\sum_{j=1}^{N} V_{ij}x_j + T_i^{hid}\Big)$$

Fig. 3.4 ANN Architecture

Hyperparameter tuning and Feature selection

Hyperparameter tuning is applied to all methods to find the optimal values of the hyperparameter to increase the efficiency of the models. Feature selection will be performed to get rid of the features which are irrelevant for prediction to avoid overfitting. Feature selection is applied on all the models except ANN, as ANN can learn important features on its own.

Hyperparameter Tuning: During the training phase model learns its parameter but there are hyperparameters that cannot be learned by the model instead they have to be tuned by analysts. These hyperparameter settings may vary for different machine learning algorithms even though the underlying dataset is the same. Tuning these hyperparameters will improve the efficiency of the model. Hyperparameter tuning will be performed to find the optimal values hyperparameter to increase the efficiency of the model. In this research, Grid search algorithm is used for hyperparameter tuning.

Feature Selection: In this research, feature selection is performed through Random Forest feature importance method. Feature selection through Random Forest comes under the category of embedded methods. Feature importance is a list of features with numeric values assigned to them. These features are sorted in descending order of feature importance. In this research context features with higher values of feature importance have a higher influence on the churn, and features with low feature importance have less significance and can be dropped to improve the efficiency of the model.

3.1.5 Model Evaluation:

Data evaluation of the built model is very crucial to analyze model efficiency. In this research, we will be using confusion matrix and different confusion matrix metrics like Accuracy, Recall, and F1 Score for evaluating the models

Confusion Matrix:

Confusion matrix is one of the important results generated in terms of understanding the efficiency of the model on a particular dataset. It is a table or matrix containing the information about predicted and actual values. The entries of the confusion matrix meaning are as follows:

- True Positives (TP): Actual value was 1 and the model predicted it as 1
- True Negative (TN): Actual value was 0 and the model predicted it as 0
- False Positives (FP): Actual value was 0 but the model predicted it as 1

- False Negative (FN): Actual value was 1 but the model predicted it as 0

In the context of this research confusion matrix values can be interpreted as follows:

- TP - Predicted and Actual value is churn
- TN - Predicted and Actual value is nonchurn
- FP - Predicted value is churn but in real it is nonchurn
- FN - Predicted value is nonchurn but in real it is churn

Accuracy:

Accuracy is defined as the number of correct predictions made divided by the total number of predictions made. Accuracy can also be calculated by true positive + true negatives divided by the total number of observations.

Accuracy is an evaluation metric that is mainly considered when the dataset is balanced, it is not suitable for highly imbalanced datasets.

$$Accuracy = \frac{NumberofCorrectPredictions}{TotalNumberofPredictions} \tag{3.2}$$

$$Accuracy = \frac{TruePositives + TrueNegatives}{TotalNumberofobservations} \tag{3.3}$$

Recall:

Recall which is also called sensitivity is calculated by True Positives divided by True Positive + False Negatives. When it comes to predicting customer churn it is very important to reduce the false negatives and to increase true positives as wrongly predicting the customer who is about to leave as nonchurner will be a huge loss to the company. So, lower values False

Negative values result in higher Recall values.

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives} \tag{3.4}$$

Precision:

Precision is calculated by considering the total number of True Positives divided by sum of True Positives and False Positives.

$$Precision = \frac{TruePositives}{TruePositives + FalsePositive} \tag{3.5}$$

F1 Score or Precision Recall Tradeoff:

F1 Score combines both recall and precision methods and this evaluation metric will be considered to evaluate the model. In this, we have two scenarios

- Low recall high precision
- High Recall and Low precision

Since the research aims to increase recall that is to minimize false negatives, high recall, and low precision will be preferred

$$F1 = \frac{2 * Recall * Precision}{Recall + Precision} \tag{3.6}$$

$$F1 = \frac{2 * TruePositive}{2 * TruePositives + FalseNegatives + FalsePositives} \tag{3.7}$$

Chapter 4

Implementation

In this research project, different machine learning classifiers are implemented on IBM Watson dataset to predict the customer churn.Evaluation of these models will be performed to find the efficient classifier in predicting the customer churn.

The design process follows the CRISP-DM Methodology. In the first step of implementation, The dataset, the CSV file is loaded to Google Colab environment, and data is converted to data frames in python programming environment using pandas and python libraries.

In the next step, the data is explored to get deeper insights into the data. A heatmap is plotted to get insights into the degree of correlation between the attributes of the dataset.

In the third step, the data are pre-processed and prepared to be modeled. As part of the data pre-processing, data is cleaned to remove any irrelevant or redundant features, missing values will be treated and outliers will be handled and categorical features will be converted to numerical features. In this research data preprocessing is handled by carrying out the required cleaning steps mentioned above. The dataset considered has 7043 instances and 21 attributes. The dataset is checked for Null values and no null values are found. The Customer Id column is dropped as it is only an indexing column and doesn't have any impact on the churn. a heatmap is plotted as shown in figure 4.1 to get insights into the degree of correlation between the attributes. All the categorical variables are converted to numerical values using a

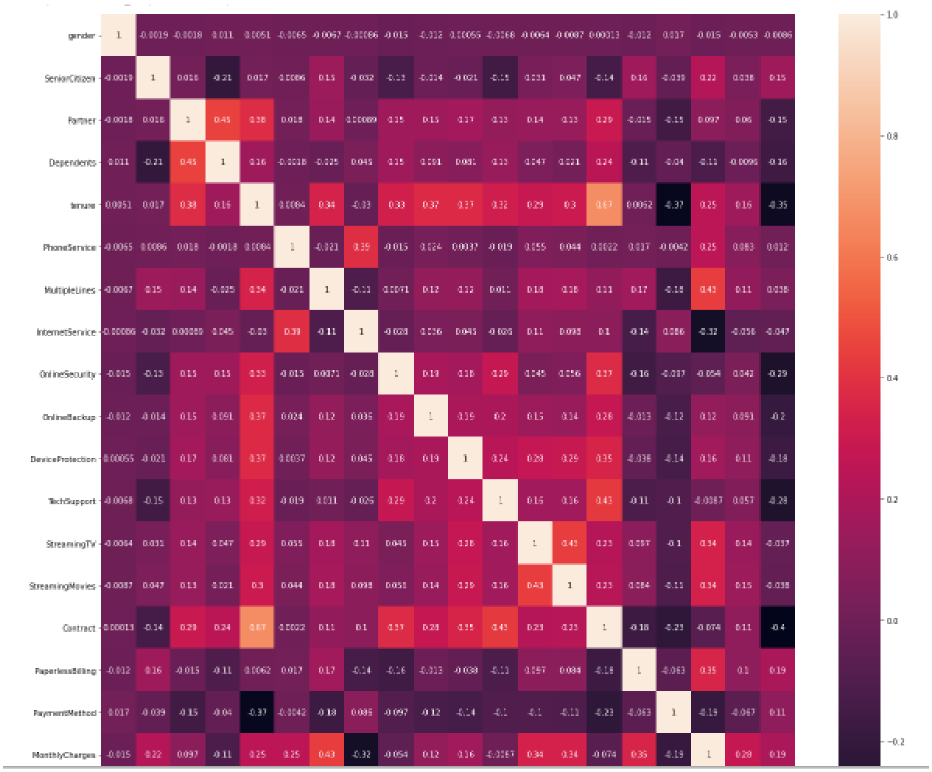


Fig. 4.1 Correlation heatmap

label encoder and the dataset is divided into label and feature sets. Each feature in the feature set is normalized and then the dataset is split into training set and test set. The data is split into 70% train set and 30% test set. As part of the last step in the preprocessing step, the class imbalance in the training set is handled. In this research, the class imbalance is handled by Synthetic Minority Oversampling Technique(SMOTE) technique. This is very important as telecom datasets are imbalanced in nature and this affects the performance of the model. The SMOTE technique oversamples the minority class by duplicating the minority class prior to fitting a model. This balances the dataset however will not add any additional information to the model.

In the fourth step of implementation, different classifiers are applied to preprocessed data of the dataset. The jupyter Notebook in Google colab environment is used as a programming environment. Python Programming is used for modeling the classifiers. All the required python packages were imported in Google colab.

4.1 Random Forest:

The data which was preprocessed and split into train and test are trained using sklearn ensemble package. In this research RandomForestClassifier() model criterion parameter is set to entropy which is for information gain. The hyperparameter tuning is applied and the Random Forest classifier is run for different values of these parameters.GridsearchCV function is used to find the optimal values of these parameters. Fivefold cross-validation is applied. Scoring is set 'recall' value as we want to minimize the number of false negatives. In GridSearchcv param-grid is used to find out the number of optimal trees(n_estimator). The optimal n_estimator value obtained was 800 when no features are dropped from the dataset.

The Random Forest classifier fit function was used to train the model. Prediction of the model is carried out using the predict function. Confusion matrix, Accuracy score and classification report are generated to check model efficiency. The feature importance of the features is generated using pandas series feature_importances_. The feature importance is numerical values printed in descending order of the values. These feature importance carry a lot of significance as it will reveal which features are contributing to churn and which features are irrelevant which can be dropped. The feature importance value, the higher the value higher the significance of the feature. Tenure, Monthly Charges, Contract, Total Charges, and Payment Method are Top5 significant features.

In this research, Random Forest model was rerun after performing feature selection using the backward elimination method based on Random Forest feature importance list. The features with less feature importance were dropped and the model was run again, this process is repeated until the Recall value was increasing. Random Forest model with a subset of features was run to check the increase in efficiency of the model. Grid search was performed to find the optimal values for the model.n_estimator value obtained is 700 after dropping few features.

4.2 Support Vector Machine Classifier(SVC):

Support vector machine classifier(SVC) is imported from the sklearn SVM package. The data which was preprocessed and split into train and test set are trained using SVC classifier.GridSearchcv is used to find the optimal value for the model. Scoring is set to recall as the aim here is to reduce the number of False Negatives.Param_grid is provided with different values of C. C value is regularisation parameter which decides the hyperplane margin distance between training points while classifying. Lower values of C imply a higher margin of hyperplane which is optimal for classification. Grid search is run with five fold cross-validation to find the optimal values. The optimal value of C obtained from the grid search is 25. The svc model is trained with the help of the fit function. SVC model trained for the best parameters obtained that is c=25 from grid search function. prediction function is used to predict the values. Confusion matrix, Accuracy score, and classification report are generated to check model efficiency.

In this research, To improve the efficiency, the SVC model was rerun with features selected based on feature importance result from Random Forest classifier by dropping few features with low feature importance value. Grid search with five fold cross-validation was run and c=25 value was obtained. The model is trained with fit function and predicted test values with prediction function. Confusion matrix, Accuracy score, and classification report are generated to check the efficiency of the model.

4.3 K-Nearest Neighbor (KNN):

KNeighborsClassifier is imported from sklearn neighbors package. The data which was preprocessed and split into training and test set are trained using the model. The basic KNN classifier is implemented in this research, hyperparameters of KNN classifiers are manually altered to check the efficiency of the model.n_neighbors is the main parameter and it denotes the number of nearest neighbors considered to classify. In this research n_neighbors value is set to 4. The metric is set to minkowski by default with the p value of 2. The model is trained with the fit() function and once the model learned from the training set predict function is used to predict test set values. The model efficiency is analyzed by generating an Accuracy score, Confusion matrix, and classification report. Confusion matrix is plotted with matplotlib.pyplot confusionmetrics() and heat map is generated with the seaborn package.

In this research, To check the impact of feature selection the KNN model is trained again with features selected based on feature importance result from Random Forest classifier by dropping few features. The KNN model is trained again by manually modifying the value of n_neighbors=5 to check the impact of feature selection. Accuracy score, Confusion matrix, and classification report are generated. Confusion matrix is plotted.

4.4 Extreme Gradient Boosting(XGBoost):

XGBClassifier() is imported from XGBoost library.The preprocessed data is split into training and test set. The XGBClassifier() model is trained using a training dataset. The XGBoost classifier is run for the optimal values of the model with the help of the GridSearchcv() function. Grid search is run to find the optimal values of n_estimators and learning_rate. The scoring parameter is set to Recall value as the aim is to reduce the False Negatives. Grid search is implemented with five fold cross-validation. Optimal value obtained is learning_rate= 0.01, n_estimators=450. The model is trained with fit function to learn from the training dataset. The predict function is used to predict test values. Confusion matrix , accuracy score, and classification report are generated.

In this research, To check the impact of feature selection on the efficiency of XGB-Classifier(), the model is rerun again after eliminating few features from the data frame based on the feature importance list obtained from the Random Forest feature importance list.GridSearchcv function with five-fold cross-validation is run to find the optimal values of n_estimators and learning_rate.Values obtained are learning_rate= 0.01, n_estimators=400. Confusion matrix , accuracy score, and classification report are printed again to check the efficiency of the model.

4.5 Artificial Neural Network(ANN):

Artificial Neural network is implemented using TensorFlow and Keras library. A simple ANN architecture with two hidden layers and reLu activation function is used to train the data. Add function is used to add layers and Dense function is used to have fully connected layers. The output layer is constructed with sigmoid activation function, binary_crossentropy loss function is used, optimizer value is set to adam. The fit function is used to train the model with different batch sizes and epoch values. predictions from the model are obtained using the predict function. Confusion matrix, accuracy score, and classification report are generated. The model is rerun by training the model for different batch and epoch values to find the optimal value where the model classified the churners efficiently.The optimal values obtained are batch_size = 10, epochs = 50.

After all the models are trained and predicted, these models are evaluated based on the evaluation metrics Recall, Accuracy Score, Precision, F1 Score. Additionally, the confusion matrix is generated to find out TP, TN, FP, and FN values. These evaluation metrics will be discussed in detail in chapter 5.

Chapter 5

Evaluation

The models implemented are evaluated using evaluation metrics Accuracy, Recall, Precision, and F1 Score. The confusion matrix is generated to find out True Positive, True Negative, False Positive, and False Negative values.

Research aim is to predict possible churners accurately so that Customer Relationship Management can take corrective actions or make strategies to retain the customer. Hence research aim at minimizing False Negative values. False Negative values are ones where the model predicted the value as nonchurn but the actual value is churn. Wrongly predicting the customer who is about to leave will be a huge loss to the telecom service provider. So the main focus of the research is on minimizing the False Negative(FN) or to increase the Recall value for customer churn.

The following sections illustrate the evaluation for each model:

5.1 Random Forest:

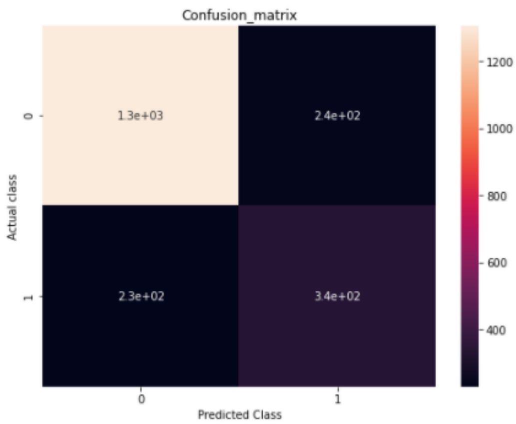
In this research, Random Forest model is implemented for the optimal values using Grid search and five fold cross-validation including all the features in the dataset. The model is evaluated using the above mentioned evaluation metrics. The values obtained for customer churn are FN = 229 and Recall = 0.60 as shown in 5.1.confusion matrix generated as shown in 5.2

```
[[1308  239]
 [ 229  337]]
Accuracy score: 0.7785139611926172
              precision    recall  f1-score   support

           0       0.85      0.85      0.85      1547
           1       0.59      0.60      0.59       566

    accuracy                           0.78      2113
   macro avg       0.72      0.72      0.72      2113
weighted avg       0.78      0.78      0.78      2113
```

Fig. 5.1 Random Forest results before feature selection




```
Confusion matrix:
 [[1308  239]
 [ 229  337]]
TP:  337
TN:  1308
FP:  239
FN:  229
```

Fig. 5.2 RF Confusion Matrix before feature selection

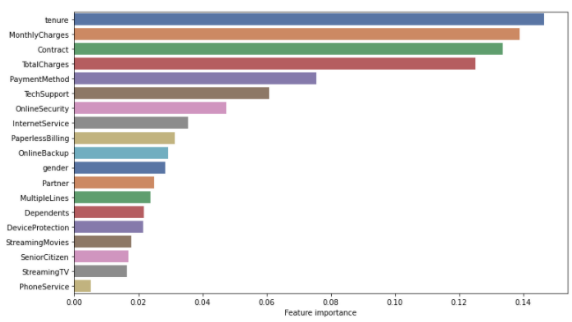


Fig. 5.3 Feature Importance

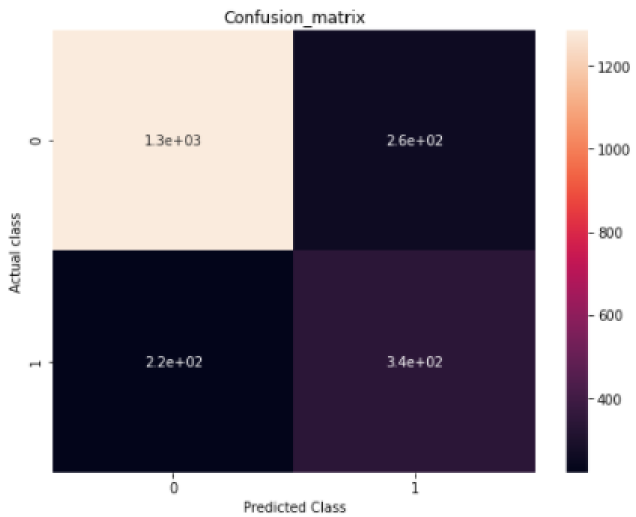
Model is rerun for the optimal values of the model using Grid search and five fold cross-validation after dropping few features with low feature importance based on feature importance list obtained by feature_importances_ as shown in figure5.3. The new results obtained for customer churn are FN = 222 and Recall= 0.61. The results are shown in the figure5.4. The confusion matrix is shown in figure5.5.

```
[[1286  261]
 [ 222  344]]
Accuracy score: 0.7714150496923805
              precision    recall  f1-score   support

           0       0.85      0.83      0.84      1547
           1       0.57      0.61      0.59       566

    accuracy                           0.77      2113
   macro avg       0.71      0.72      0.71      2113
weighted avg       0.78      0.77      0.77      2113
```

Fig. 5.4 Random Forest results after feature selection




```
Confusion matrix:
 [[1286  261]
 [ 222  344]]
TP:  344
TN:  1286
FP:  261
FN:  222
```

Fig. 5.5 RF Confusion Matrix after feature selection

5.2 Support Vector Machine Classifier(SVC):

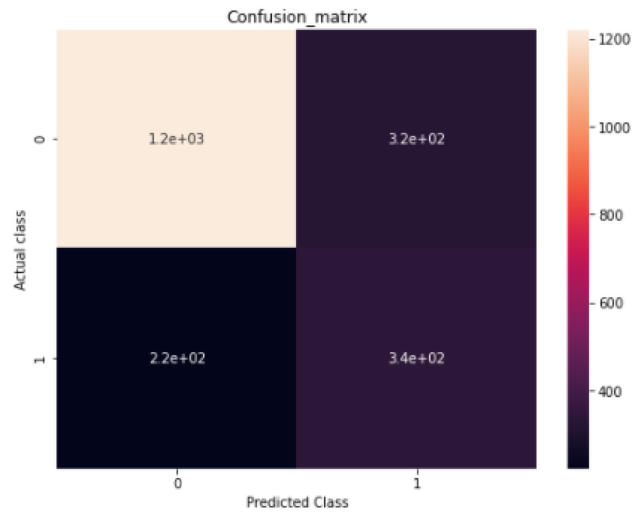
SVC model is run for the optimal values using five fold cross-validation and Grid search including all the features in the dataset. The model is evaluated using the above mentioned evaluation metrics. The results obtained for customer churn are FN =223 and Recall = 0.61 as shown in the figure5.6. confusion matrix generated as shown in 5.7

```
[[1222  325]
 [ 223  343]]
accuracy_score: 0.7406530998580217
              precision    recall  f1-score   support

           0       0.85      0.79      0.82      1547
           1       0.51      0.61      0.56       566

    accuracy                           0.74      2113
   macro avg       0.68      0.70      0.69      2113
weighted avg       0.76      0.74      0.75      2113
```

Fig. 5.6 SVC results before feature selection




```
Confusion matrix:
 [[1222  325]
 [ 223  343]]
TP:  343
TN:  1222
FP:  325
FN:  223
```

Fig. 5.7 SVC Confusion Matrix before feature selection

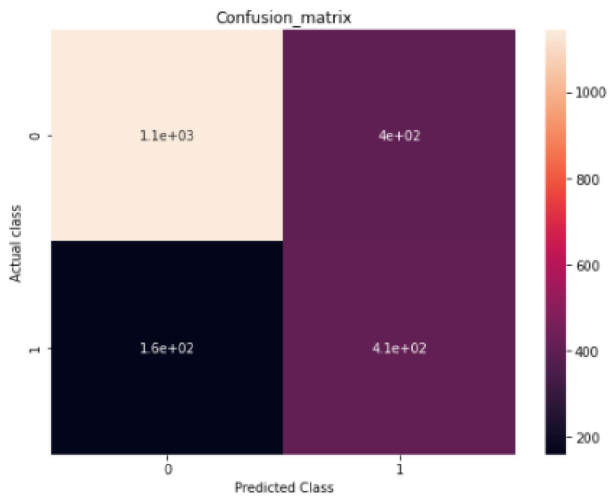
Model is implemented again after dropping features with low feature importance based on feature importance list obtained from Random forest model. The model optimal parameters are found using Grid search and five fold cross-validation. The results obtained for customer churn are FN = 159 and Recall = 0.72 as shown in the figure5.8. The confusion matrix is shown in figure5.9.

```
[[1146  401]
 [ 159  407]]
accuracy_score: 0.7349739706578324
              precision    recall  f1-score   support

           0       0.88      0.74      0.80      1547
           1       0.50      0.72      0.59       566

    accuracy                           0.73      2113
   macro avg       0.69      0.73      0.70      2113
weighted avg       0.78      0.73      0.75      2113
```

Fig. 5.8 SVC results after feature selection




```
Confusion matrix:
 [[1146  401]
 [ 159  407]]
TP:  407
TN:  1146
FP:  401
FN:  159
```

Fig. 5.9 SVC Confusion Matrix after feature selection

5.3 K-Nearest Neighbor(KNN):

In this research, a simple KNN Model is implemented by manually changing the hyperparameter values including all the features in the dataset. The model is evaluated using the above mentioned evaluation metrics. The results obtained for customer churn are FN =213 and Recall = 0.62 as shown in the figure5.10. confusion matrix generated as shown in 5.11

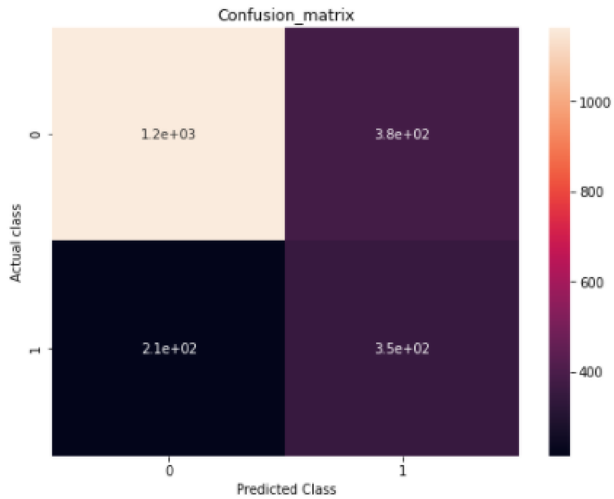
```
[[1166  381]
 [ 213  353]]
              precision    recall  f1-score   support

           0       0.85      0.75      0.80      1547
           1       0.48      0.62      0.54       566

    accuracy                           0.72      2113
   macro avg       0.66      0.69      0.67      2113
weighted avg       0.75      0.72      0.73      2113

0.7188831045906294
```

Fig. 5.10 KNN results before feature selection




```
Confusion matrix:
 [[1166  381]
 [ 213  353]]
TP:  353
TN:  1166
FP:  381
FN:  213
```

Fig. 5.11 KNN Confusion Matrix before feature selection

The model is implemented again after dropping features with low feature importance based on the feature importance list obtained from Random Forest model. The results obtained for customer churn are FN = 176 and Recall = 0.69 as shown in the figure5.12. The confusion matrix is shown in the figure5.13.

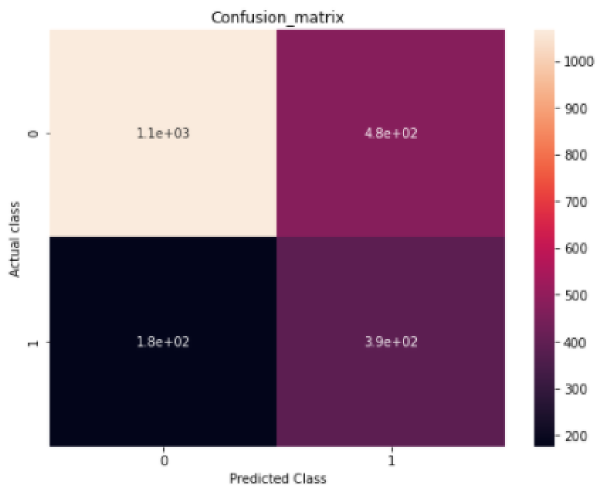
```
[[1069  478]
 [ 176  390]]
              precision    recall  f1-score   support

           0       0.86      0.69      0.77      1547
           1       0.45      0.69      0.54       566

    accuracy                           0.69      2113
   macro avg       0.65      0.69      0.65      2113
weighted avg       0.75      0.69      0.71      2113

0.690487458589683
```

Fig. 5.12 KNN results after feature selection




```
Confusion matrix:
 [[1069  478]
 [ 176  390]]
TP:  390
TN:  1069
FP:  478
FN:  176
```

Fig. 5.13 KNN Confusion Matrix after feature selection

5.4 Extreme Gradient Boosting(XG-Boost):

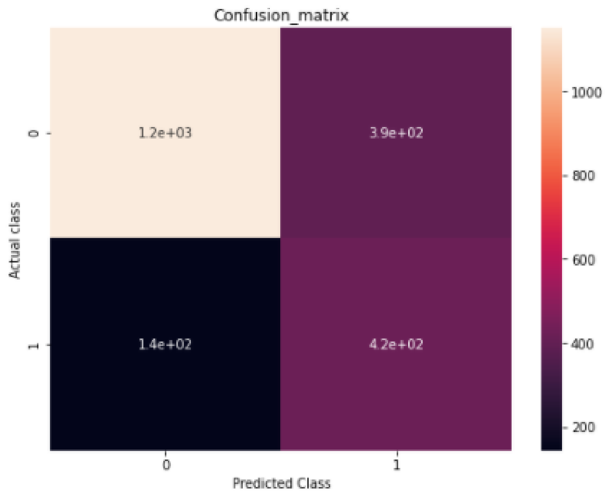
XG-Boost model is run for the optimal values using five fold cross-validation and Grid search including all the features in the dataset. The model is evaluated using the above mentioned evaluation metrics. The results obtained for customer churn are FN =144 and Recall = 0.75 as shown in the figure5.14. confusion matrix generated as shown in 5.15

```
[[1153  394]
 [ 144  422]]
Accuracy score: 0.7453857075248462
              precision    recall  f1-score   support

           0       0.89      0.75      0.81      1547
           1       0.52      0.75      0.61       566

    accuracy                           0.75      2113
   macro avg       0.70      0.75      0.71      2113
weighted avg       0.79      0.75      0.76      2113
```

Fig. 5.14 XGBoost results before feature selection




```
Confusion matrix:
 [[1153  394]
 [ 144  422]]
TP:  422
TN:  1153
FP:  394
FN:  144
```

Fig. 5.15 XGBoost Confusion Matrix before feature selection

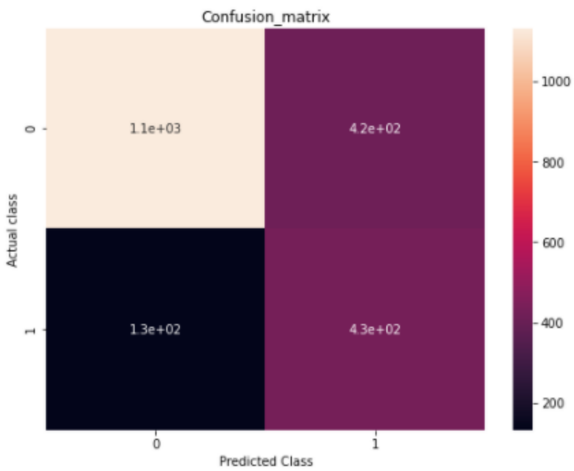
The model is implemented again after dropping features with low feature importance based on the feature importance list obtained from Random Forest model. The model optimal parameters are found using Grid search and five fold cross-validation. The results obtained for customer churn are FN = 133 and Recall = 0.77 as shown in the figure5.16. The confusion matrix is shown in the figure5.17.

```
[[1132  415]
 [ 133  433]]
Accuracy score: 0.7406530998580217
              precision    recall  f1-score   support

           0       0.89      0.73      0.81      1547
           1       0.51      0.77      0.61       566

    accuracy                           0.74      2113
   macro avg       0.70      0.75      0.71      2113
weighted avg       0.79      0.74      0.75      2113
```

Fig. 5.16 XGBoost results after feature selection




```
Confusion matrix:
 [[1132  415]
 [ 133  433]]
TP:  433
TN:  1132
FP:  415
FN:  133
```

Fig. 5.17 XGBoost Confusion Matrix after feature selection

5.5 Artificial Neural Network(ANN):

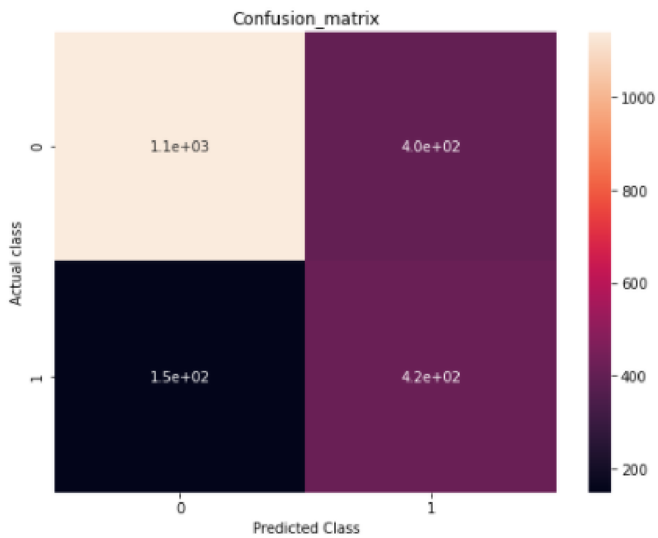
ANN model is evaluated using the above-mentioned evaluation metrics. The evaluation is carried out by using the predicted values against actual test values for churn. The results obtained for customer churn are FN =149 and Recall = 0.74 as shown in the figure5.18. confusion matrix generated as shown in 5.19

```
[[1142  405]
 [ 149  417]]
accuracy: 0.737814
              precision    recall  f1-score   support

           0       0.88      0.74      0.80      1547
           1       0.51      0.74      0.60       566

    accuracy                           0.74      2113
   macro avg       0.70      0.74      0.70      2113
weighted avg       0.78      0.74      0.75      2113
```

Fig. 5.18 ANN results




```
Confusion matrix:
 [[1142  405]
 [ 149  417]]
TP:  417
TN:  1142
FP:  405
FN:  149
```

Fig. 5.19 ANN Confusion Matrix

5.6 Results Discussion

The Table 5.1summarises False Negative and Recall values obtained for all the models before feature selection and after feature selection for optimal values of hyperparameters. The higher value of false negatives shows that the model was not efficient in predicting customer churn whereas lower values of false negatives indicate the model was efficient in predicting customer churn. Column 3 in Table 5.1shows that there is a decrease in false negative values for each model after implementing feature selection. Table 5.1 Column 4 shows the increase in Recall or sensitivity values after feature selection implemented for each model. Table reffn results indicate that XGBoost outperforms other classifiers in terms of minimizing false negative values hence achieving high sensitivity or recall values.XGBoost algorithm was successful at minimizing false negative values from 144 to 133 with an increase in recall value from 0.75 to 0.77 after feature selection was implemented which is about 2% improvement. As Neural Networks does not require feature engineering and it is automatically taken care of by the model itself, feature selection is not implemented on ANN. The table 5.1shows false negative values achieved by ANN with optimal batch and epoch values is 149 with a recall value of 0.74 which is quite satisfactory.

Table 5.1 Summary of False Negatives and Recall values for different classifiers

	Before Feature Selection		After Feature Selection	
Model	FN	Recall	FN	Recall
Random Forest	229	0.60	222	0.61
SVC	223	0.61	**159**	**0.72**
KNN	213	0.62	176	0.69
XGBoost	**144**	**0.75**	**133**	**0.77**
ANN	**149**	**0.74**		

Table 5.2 is summary of results obtained based on different evaluation metrics for all five classifiers by applying hyperparameter tuning before feature selection. Random Forest is able to perform well in terms of Accuracy with 0.78 followed by XGBoost with 0.74, ANN with 0.74, and SVC also with 0.74 respectively. However, in terms of Recall value, XGBoost

Table 5.2 Model Results Before Feature Selection

Evaluation Summary				
Model	Accuracy	Recall	Precision	F1 Score
Random Forest	**0.78**	0.60	0.59	0.59
SVC	0.74	0.61	0.51	0.56
KNN	0.72	0.62	0.48	0.54
XGBoost	0.75	**0.75**	0.52	0.61
ANN	0.74	**0.74**	0.51	0.60

and ANN outperformed with Recall values of 0.75 and 0.74 respectively. Table 5.2results indicate XGBoost is outperforming with an Accuracy value of 0.75 and Recall value of 0.75. ANN also seems to perform quite well after hyperparameter tuning is performed with different Batch size and Epoch values by achieving Accuracy of 0.74 and Recall of 0.74.

Table 5.3 Model Results After Feature Selection

Evaluation Summary				
Model	Accuracy	Recall	Precision	F1 Score
Random Forest	0.77	0.61	0.57	0.59
SVC	0.73	0.72	0.50	0.59
KNN	0.69	0.69	0.45	0.54
XGBoost	0.74	**0.77**	0.51	0.61

Table 5.3provides summary of results obtained for different classifiers by applying hyperparameter tuning and feature selection. The results are obtained for different evaluation metrics. Random Forest achieved the highest accuracy compared to all classifiers with a value of 0.77.XGBoost outperformed with a Recall value of 0.77.

Table 5.4 Final Recall Values Achieved

Recall Summary	
Model	Recall
Random Forest	0.61
SVC	0.72
KNN	0.69
XGBoost	**0.77**
ANN	**0.74**

Table 5.4 Results shows the final Recall values achieved by each classifier. Overall XGBoost Outperformed all the other classifiers with a Recall value of 0.77.

Chapter 6

Conclusions and Future Work

6.1 Conclusion:

Customer Churn is one of the crucial issues faced by all telecom service providers globally due to the saturated telecommunication market. Customers will not hesitate to leave the service provider if they are not satisfied with the services provided or due to better offers provided by competitors at the value customer is ready to pay. Hence it is very important to have an efficient churn prediction model in place. There have been a lot of prediction models proposed however they fail to take into consideration, the complexities of the real telecom dataset. Keeping all this in mind this research presents a churn prediction model for telecommunication customers. The model is built by taking into consideration the complexities of telecom datasets like high dimensionality, class imbalance and large volume of data. Different classifiers that are suitable for predicting telecom churn are taken into consideration with hyperparameter tuning and feature selection to help improve the efficiency of the models further. Hyperparameter tuning is implemented with the help of GridSearchCV algorithm. Feature selection is applied through Random Forest feature importance method which comes under the embedded method. The efficiency of these models is compared with confusion matrix evaluation metrics using IBM Sample telecom dataset. Results indicate that feature selection and hyperparameter tuning improved efficiency of the each model.XGBoost outperformed all the other classifiers with a sensitivity or recall value of 0.77 and false negative value of 133. Feature selection and hyperparameter tuning improved efficiency of XGBoost model by 2%.

6.2 Future Work

The future direction of this work can include exploring other hyperparameter tuning algorithms like Random search, and Bayesian optimization to overcome limitations of the Grid search algorithm. Since in grid search, the number of models to train grows exponentially with the number of hyperparameters provided, which makes the algorithm inefficient in terms of computing power and time. In addition other deep learning models can be explored for churn prediction as Neural Networks prove to be efficient on large volume datasets like telecom datasets. Finally, the current model can be implemented in other sectors like insurance or banking.

References

Ahmad, A.K., Jafar, A. and Aljoumaa, K. (2019) Customer churn prediction in telecom using machine learning in big data platform *Journal of Big Data* **6**(1), p. 28

Ajit, P. (2016) Prediction of employee turnover in organizations using machine learning algorithms *algorithms* **4**(5), p. C5

Azeem, M. and Usman, M. (2018) A fuzzy based churn prediction and retention model for prepaid customers in telecom industry *International Journal of Computational Intelligence Systems* **11**(1), pp. 66–78

Brânduşoiu, I., Toderean, G. and Beleiu, H. (2016) Methods for churn prediction in the pre-paid mobile telecommunications industry in: *2016 International conference on communications (COMM)* pp. 97–100 IEEE

Bronshtein, A. (2017) A quick introduction to k-nearest neighbors algorithm

Chu, B.H., Tsai, M.S. and Ho, C.S. (2007) Toward a hybrid data mining model for customer retention *Knowledge-Based Systems* **20**(8), pp. 703–718

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

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 in: *2016 Symposium on Colossal Data Analysis and Networking (CDAN)* pp. 1–4 IEEE

Dubey, A. (2018) Feature selection using random forest

Hadden, J., Tiwari, A., Roy, R. and Ruta, D. (2007) Computer assisted customer churn management: State-of-the-art and future trends *Computers & Operations Research* **34**(10), pp. 2902–2917

Hanif, I. (2019) Implementing extreme gradient boosting (xgboost) classifier to improve customer churn prediction

Hung, S.Y., Yen, D.C. and Wang, H.Y. (2006) Applying data mining to telecom churn management *Expert Systems with Applications* **31**(3), pp. 515–524

IBMCommunity (2015) ibm community home sample datasets

Idris, A. and Khan, A. (2012) Customer churn prediction for telecommunication: Employing various various features selection techniques and tree based ensemble classifiers in: *2012 15th International Multitopic Conference (INMIC)* pp. 23–27 IEEE

Idris, A. and Khan, A. (2014) Ensemble based efficient churn prediction model for telecom in: *2014 12th International Conference on Frontiers of Information Technology* pp. 238–244 IEEE

Idris, A., Khan, A. and Lee, Y.S. (2012a) Genetic programming and adaboosting based churn prediction for telecom in: *2012 IEEE international conference on Systems, Man, and Cybernetics (SMC)* pp. 1328–1332 IEEE

Idris, A., Rizwan, M. and Khan, A. (2012b) Churn prediction in telecom using random forest and pso based data balancing in combination with various feature selection strategies *Computers & Electrical Engineering* **38**(6), pp. 1808–1819

Keramati, A., Jafari-Marandi, R., Aliannejadi, M., Ahmadian, I., Mozaffari, M. and Abbasi, U. (2014) Improved churn prediction in telecommunication industry using data mining techniques *Applied Soft Computing* **24**, pp. 994–1012

Khan, M.R., Manoj, J., Singh, A. and Blumenstock, J. (2015) Behavioral modeling for churn prediction: Early indicators and accurate predictors of custom defection and loyalty in: *2015 IEEE International Congress on Big Data* pp. 677–680 IEEE

Kursa, M.B. and Rudnicki, W.R. (2011) The all relevant feature selection using random forest *arXiv preprint arXiv:1106.5112*

Lee, S.M., Kim, D.S., Kim, J.H. and Park, J.S. (2010) Spam detection using feature selection and parameters optimization in: *2010 International Conference on Complex, Intelligent and Software Intensive Systems* pp. 883–888 IEEE

Li, K.G. and Marikannan, B.P. (2019) Hyperparameters tuning and model comparison for telecommunication customer churn predictive models

Liaw, A., Wiener, M. *et al.* (2002) Classification and regression by randomforest *R news* **2**(3), pp. 18–22

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), pp. 1659–1665

McDonald, C. (2017) Machine learning fundamentals (ii): Neural networks

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

MOUNIKA REDDY, C. (2016) Customer churn predictive heuristics from operator and users' perspective

Pamina, J., Raja, B., SathyaBama, S., Sruthi, M., VJ, A. *et al.* (2019) An effective classifier for predicting churn in telecommunication *Jour of Adv Research in Dynamical & Control Systems* **11**

Qureshi, S.A., Rehman, A.S., Qamar, A.M., Kamal, A. and Rehman, A. (2013) Telecommunication subscribers' churn prediction model using machine learning in: *Eighth International Conference on Digital Information Management (ICDIM 2013)* pp. 131–136 IEEE

Saeys, Y., Abeel, T. and Van de Peer, Y. (2008) Robust feature selection using ensemble feature selection techniques in: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* pp. 313–325 Springer

Saraswat, M. and Arya, K. (2014) Feature selection and classification of leukocytes using random forest *Medical & biological engineering & computing* **52**(12), pp. 1041–1052

Sharma, A., Panigrahi, D. and Kumar, P. (2013) A neural network based approach for predicting customer churn in cellular network services *arXiv preprint arXiv:1309.3945*

Sivasankar, E. and Vijaya, J. (2019) Hybrid ppfcm-ann model: an efficient system for customer churn prediction through probabilistic possibilistic fuzzy clustering and artificial neural network *Neural Computing and Applications* **31**(11), pp. 7181–7200

Subramanya, K.B. and Somani, A. (2017) Enhanced feature mining and classifier models to predict customer churn for an e-retailer in: *2017 7th International Conference on Cloud Computing, Data Science & Engineering-Confluence* pp. 531–536 IEEE

Tsai, C.F. and Lu, Y.H. (2009) Customer churn prediction by hybrid neural networks *Expert Systems with Applications* **36**(10), pp. 12547–12553

Ullah, I., Raza, B., Malik, A.K., Imran, M., Islam, S.U. and Kim, S.W. (2019) A churn prediction model using random forest: analysis of machine learning techniques for churn prediction and factor identification in telecom sector *IEEE Access* **7**, pp. 60134–60149

Umayaparvathi, V. and Iyakutti, K. (2016a) Attribute selection and customer churn prediction in telecom industry in: *2016 international conference on data mining and advanced computing (sapience)* pp. 84–90 IEEE

Umayaparvathi, V. and Iyakutti, K. (2016b) A survey on customer churn prediction in telecom industry: Datasets, methods and metrics *International Research Journal of Engineering and Technology (IRJET)* **3**(04)

Umayaparvathi, V. and Iyakutti, K. (2017) Automated feature selection and churn prediction using deep learning models *International Research Journal of Engineering and Technology (IRJET)* **4**(3), pp. 1846–1854

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**, pp. 1–9

Wirth, R. and Hipp, J. (2000) Crisp-dm: Towards a standard process model for data mining in: *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining* pp. 29–39 Springer-Verlag London, UK

Yabas, U. and Cankaya, H.C. (2013) Churn prediction in subscriber management for mobile and wireless communications services in: *2013 IEEE Globecom Workshops (GC Wkshps)* pp. 991–995 IEEE

Zhang, Y., Qi, J., Shu, H. and Cao, J. (2007) A hybrid knn-lr classifier and its application in customer churn prediction in: *2007 IEEE International Conference on Systems, Man and Cybernetics* pp. 3265–3269 IEEE