Customer Churn Prediction in Telecom Industry using Machine Learning Techniques

MSc Research Project
MSc Data Analytics

Mandar Vaidya
Student ID: X17153409

School of Computing
National College of Ireland

Supervisor: Dr. Catherine Mulwa

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Mandar Vaidya
Student ID:	X17153409
Programme:	MSc. In Data Analytics **Year:** 2018
Module:	MSc. Research Project
Supervisor:	Dr. Catherine Mulwa
Submission Due Date:	18th April, 2019
Project Title:	Customer Churn Prediction in Telecom Industry using Machine Learning Techniques
Word Count:	XXXX **Page Count:** XX

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: ……

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 Prediction in Telecom Industry using Machine Learning Techniques

Mandar Vaidya
x17153409

Abstract

Customer retention has been one of the most important responsibilities in organisations, especially in telecom sector. Customer leaving or about to leave a service is known as churn and process is known as churning, with almost 30-35% rate of churn, telecom sector takes the first place. Due to immense competition between telecom companies, churn prediction has become of the hot topics of research. Although, a lot of algorithms have been implemented, there is always still room for improvement in performance. To retain customers, telecom companies needs to study: (i) why they churn, (ii) which will churn and (iii) return potential churners through various patterns. So, to study, the patterns of customer churn, factors contributing, data mining techniques are used, as retention cost has become expensive than acquisition. To study prediction mechanism, updating strategies from time-to-time has become vital. Prediction techniques like, Logistic Regression, K-Nearest Neighbor, Decision Trees, Random Forest, Naïve Bayes, AdaBoost and Artificial Neural Networks based on different combinations of variables are implemented in this research and results were evaluated on performance metrics like accuracy, precision and recall.

Keywords: Machine Learning, Customer Churn, Prediction, Telecom.

1 Introduction

In introduction section, the idea of project background and motivation and how customer churn is one of the trickiest issues in today's ever-growing world and how it is affecting the telecom sector was discussed in detail. Furthermore, the motivation to choose research on this topic and research question and some sub-objectives were discussed.

1.1 Motivation and Background

Churn is defined as a customer abandoning a service. In today's global world, Customer Churn has become a highly exchangeable topic of argumentation mainly in ever-growing line of business firms like; telecommunications, banking, banking, legal, email, social-networking etc. These days, the area of research with highly debate is prediction/analyzation of Churn. Analysts, researchers, decision makers from various fields are interpreting new data mining, machine learning techniques or modifying the old ones for recommended solutions to know more churners. Due to ever-increasing competition that too on global-level, telecom industries have been affected by churners. That's why, telecom companies have shifted their core from customer acquisition to customer retention. As, customer retention is way more expensive than customer acquisition. In last 15-20 years, churn rate has increased rapidly, added to that, telecom industry is most affected the most. Additional to that issue, variations in numerous

important segments like, less restrictions in market uplifting while entering in the industry, new schemes, new easily accessible laws, new technologies etc., all these factors with added competition, churn rate and customer churn has increased which inevitably causes massive losses which is a concern on the financial aspect of the organizations.
That's why, the focus has been shifted to defensive from aggressive marketing strategies. Although one of the major hurdles is to pinpoint customers who are about to churn or likely to churn in the future. To tackle this hurdle, prediction, classification is done on selected factors with the help of various data mining techniques and machine/deep learning algorithms.

Churn techniques (Fig. 1) can be classified in two types: (i) Traditional and (ii) Conventional. Traditional techniques are fast and has a good interpretation of churning results where as conventional techniques are kind of robust and has a low level of interpretation of churning results. Churners can be classified into two types: (i) Voluntary and (ii) Involuntary.
Both have a significant difference; Voluntary churners are those churners who switch to new service without even thinking of scenarios. They're prompt to change, on the other hand, Involuntary churners are those who telecom organizations remove from the subscribers list due to obvious reasons.
Furthermore, Voluntary churners can be sub-divided into further two sections, (i) Deliberate and (ii) Incidental.

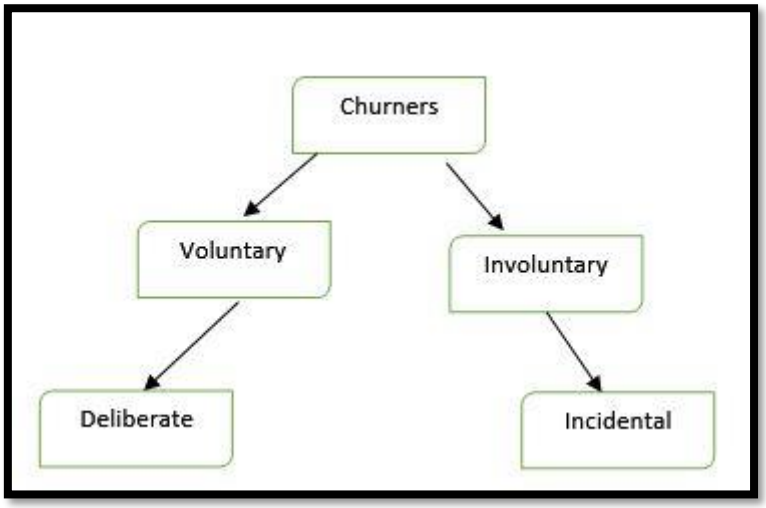

Fig.1 Churn Classification

Deliberate churn is the most important area of focus as it is involved in loads of important aspects and it is one of the most widely researched section for analyzing various patterns and entities required for churn management. Deliberate churn takes place due to various reasons like; other options of new and better technology, current service provider is bad, high priced, various social factors etc.
Incidental churn depends on external factors like; change of location, money related issues etc. It is unplanned by the customer.

1.2 Problem Statement

Various problems are faced in recent times within the domain churning and specially in telecom sector. These problems must be considered as challenges to not only overcome but also to

improve the ever-growing problem of churning and hence, improve the revenue of sectors involved in it by customer retention.
Telecom Industry is extremely dynamic with market constantly altering with new carriers, advertisements, technologies, schemes etc. To lure new customers, organizations are trying anything, as competition is constantly on rise, it has become important to control churn as, domestic churn is roughly 2-3% of customer base. (Mozer et al.; 2000)
In telecom sector, for prepaid customers there are not many options for modelling the data except call detail record (CDR). This is the belief that, same community customers are inclined to have similar behaviors due to various factors. So, to prevent churn, machine learning algorithms plays a decisive role in prediction of churners. (Yan et al.; 2005)
For a traditional model, prediction of customer churn is very difficult, because the class is imbalanced whereas the target task is less. Many Data Mining algorithms are used, while new ones are invented every-now-and-then. To tackle this, multiple classifiers, ensemble models are created for tackling the churn problem, and thus by improving prediction accuracy. (Xie et al.; 2011)
Customer churn prediction is not only critical but also very important for the company's revenue. Although, there are millions of telecommunication customers, reducing even 1% churn rate can yield good profit to those firms. (Xia et al.; 2018)

To address the problem, the following research question has been specified.

RQ: *"How can we enhance/increase churn predictions (i.e. accuracy, precision and recall) in telecom industries using classification techniques (Logistic Regression, K-Nearest Neighbor, Decision Trees, Random Forest, Naïve Bayes, AdaBoost and Artificial Neural Networks) to reduce customer churning?"*

1.3 Research Objectives

The main objectives surrounding this research are:

- A critical review of literature on churn prediction (2000 - 2018).
- Establishing important variables, determining the correlation between dependent variables, checking class imbalance and over-sampling.
- Exploratory data analysis, data preprocessing for implementation of Machine Learning models.
- Implementation, evaluation and result of Logistic Regression model.
- Implementation, evaluation and result of KNN model.
- Implementation, evaluation and result of Decision Tree model.
- Implementation, evaluation and result of Random Forest model.
- Implementation, evaluation and result of Naïve Bayes model.
- Implementation, evaluation and result of AdaBoost model.
- Implementation, evaluation and result of Artificial Neural Networks model.
- Comparison of above-mentioned developed models.

2 Related Work

In literature review section, previously studied and worked conducted on churn by various technologies in telecom sector have being discussed thoroughly.

2.1 Review of Customer Churn

A theoretical and practical method is needed to explore more in prediction of churning. A feature selection pruning technique called Orientation Ordering Pruning Method (OOPM) is suggested. In this technique, with the help of pruning method, classifiers are used instead of selectors. Later, Random Forest and Transduction (FE_RF&T) is used for feature extraction and results, suggests that OOPM has more advantages than Principal Component Analysis (PCA) and FE_RF&T methods. Feature selection method includes two things i.e., filter and wrapper model. Both methods maintain the integrity of original essential features. Feature extraction transforms the data into high/low dimensional data depending on the need with the help of projection matrix. After, data sampling and outlier detection, feature extraction method was applied with Logistic, C4.5, Logistic Regression and Random Forest models and, results showed Random Forest gave more accuracy than anything else. Thus, by showing, OOPM has more advantages when used with Random Forest thus by improving performance in prediction of churn. (Yihui and Chiyu; 2016)

Author implements a comparison of various machine learning techniques in prediction of customer churning. Initially cross-validation was performed, then boosting was done, Monte Carlo simulations was performed on various parameters. CRM tool was used before applying Data Mining techniques like Artificial Neural Network, Support Vector Machine, Decision Tree, Naive Bayes, Regression Analysis with parameters like Precision, Recall, Accuracy and F-measure was used on boosting algorithm. After performing, Monte-Carlo simulations with two-layer Back-propagation Network classifier was used on AdaBoost.M1 algorithm. After the final phase, SVM-POLY with AdaBoost performed the best. Future work would be performing on different/weak classifiers and trying some different boosting method. (Vafeiadis et al.; 2015)

In this author, examines various data mining techniques in customers who are likely to churn and selection of churning. Several classifiers were used on two datasets. Predictive modelling techniques are often effective in accurate prediction of churn. Various classifiers like: Gradient Boosting, Decision Tree, Support Vector Machine, Random Forest, K-Nearest Neighbor, Ridge Regression Classifier and Logistic Regression was analyzed on metrics like Confusion Matrix, Accuracy, Precession, Recall and F1-score. Feature selection and cross validation was performed before implementing Data Mining techniques. Several state-of-the-art classifiers are used because of binary classification problem and result shows Random Forest and Gradient Boosting performed the best in all metrics. Future work would be to perform techniques on different attributes with the predictor variables. (Chouiekh; 2017)

2.2 Review of Methods and Techniques

In this paper, author, suggests, the most crucial Business Intelligence (BI) application is churn prediction. Data Mining techniques like; Decision Trees, K-means Algorithm, Artificial Neural Networks and Regression Analysis was used. Secondly, while presenting churn prediction

model, various re-sampling methods were performed to improve class imbalance. Measures such as; Recall, Precision and F-measure was performed. For correlation, Spearman's Correlation was performed to check the dependence between two variables. Class imbalance was done with the re-sampling, over-sample and under-sample was done on majority of the class. In Decision Tree, through SPSS, Chi-squared Automatic Interaction Detector (CHAID and Exhaustive CHAID), Quick, Unbiased and Efficient Statistical Tree (QUEST) and Classification and Regression Trees (CART) was performed. Exhaustive CHAID performed the most accurate variant of Decision Trees. New variables were added to check more models. In future work, different variables and different metrics can be performed in SPSS. (Quershi et al.; 2013)

As, various processes like; machine learning, statistics, pattern recognition and visualization techniques are performed. Standard neural network architecture was used for prediction of churn. As, the data was inseparable, multilayer perceptron (MLP) was used instead of McCulloch-Pitts neuron. Here, MLP uses the backpropagation (BP) algorithm because of continuous non-linear activation function. For outer layer of MLP, sigmoid function known as SoftMax was introduced. The weights were implemented on variables using simulated annealing algorithm. For acceleration of process, convergence speed and local minima was done. Based on sensitivity, pruning strategy was done. The overall accuracy of this model was close 99%, but to test on a large-scale dataset, few requirements was needed on variables like cross-sell and up-sell respectively. (Brandusoiu et al.; 2016)

In this paper, author, predicts telecommunication sector is highly affected by churning issue. The churn rate has almost touched 30%. To tackle this ever-increasing issue, predictive modelling was implemented in this paper. Principal Component Analysis (PCA) was used for dimensionality reduction and after that on selected discrete variables Machine Learning algorithms like; Neural Networks, Support Vector Machine and Bayesian Networks were applied. The McCulloch-Pitts neuron which represents the simplest neural network of a weighted neuron was used for activation. By evaluating results for both churners and non-churners, measure such as ROC and accuracy was calculated, and the result showed Bayesian Network has the highest accuracy. As, the dataset was not large, performance of all the models was above 95% thus decision makers can use this approach to retain churners based on predictors and thereby used this approach in customer acquisition and CRM tools too. Future work would be too used a large dataset. (Brandusoiu et al.; 2016)

Sequential cellular data were used for churn prediction. Several state-of-the-art algorithms were analysis and compared for better performance, algorithms like Gradient Boosting Trees, Random Forest, Support Vector Machine (SVM), Basic Long Short-Term Memory (LSTM). Sequential data was integrated in the above-mentioned algorithms, thus by making classifier more rigorous and more accurate for results. Vanilla RNN and LSTM was given more sequential data than others for more boosting, but both showed less accuracy, thus by eliminating them. Various parameters like numerical, categorial, sequential was used for analysis but sequential showed more accuracy. Performance was analyzed on parameter accuracy with the help of ROC and AUC curves. Gradient Boosting Trees showed an accuracy on 97%, the highest among all, thus by making it more suitable for sequential dataset. Future work would be the prediction the rate of churners after 5-6 months and using real-time and adaptive dataset. (Khan and Kozat; 2017)

Churn management strategies have been the focus. A large-scale dataset with eliminating dimensionally reduction technique and an ensemble-Decision Tree was implemented. Several

Key Performance Indicator (KPIs) were analyzed with feature selection method. And, lastly, an Apache Oozie workflow engine was used for production-zing the practical approach in the real-time dataset. Several KPIs were extracted in Hadoop Cluster using PIG scripts and CDR data techniques. Feature selection and dimensionality reduction was done on several KPIs before applying Machine Learning algorithms. Linear and non-linear classifiers like Logistic Regression, Random Forest, Support Vector Machine, Ensemble-based Random sup space was analyzed. Furthermore, KPIs was divided into 2 parts: (i) Snapshot and (ii) Trending. Snapshot contain single whereas trending contains multiple values which are known as KPIs respectively. Wrapper method was used for feature selection because some KPIs have unique feature of patterns. For future work, Principal Component Analysis (PCA) can be used for dimensionality reduction when model is trained and tested for more robust patterns. (Meher et al.; 2017)

2.3 Comparison of Different Models and Techniques

Implementation predictive model for prediction of churners or is about to churn. Four kernels functions with the use of Support Vector Machine (SVM) was analyzed with the help of gain measure as a parameter for the prediction. For training, SVM, sequential minimal optimization (SMO) algorithm was implemented on kernel matrix. As, kernel matrix is large enough to hold all support vectors, but the memory constraint might come into analysis which in turn might give some redundancy issue during prediction, thus by, only 4 kernel functions were selected namely; Radial Basis Function Kernel (RBF), Linear Kernel (LIN), Polynomial Kernel (POL) and Sigmoid Kernel (SIG) respectively. By comparing, gain measure performance, it was evident that models that use RBF and POL performed 80% better than LIN 60% and SIG 50% respectively. Secondly, several multiple configurations of SVM like C, Gamma, Bias and Degree were analyzed, and result showed POL performed best for all the measure. Future work would be to implement this approach on other variables and applying strategies by management people for further acquisition of customers. (Brandusoiu et al; 2013)

As many firms are adopting to customer retention programs for prediction of churners. It has become more and more difficult to develop a specific model. In this research, a Multilayer Perceptron (MLP) has been proposed and Machine Learning techniques like Multiple Regression Analysis (MLA), Logistic Regression Analysis (LRA) and Neural Network were compared and analyzed. Sensitivity, Specificity and Accuracy were the three main parameters that were used in analysis and comparison between various selected models. For, regression, MINTAB was used. Various Multilayer Perceptron's with neural network were implemented for training and testing purpose. And, the results, were compared with LRA and MLA simultaneously. Result showed, Neural Network have good overall performance among all the 3 parameters. Furthermore, as Neural Network has loads of nodes, out of which there is one hidden and output node are the main ones. On output node, Levenberg Marquardt (ML) learning algorithm was applied for cross-checking the analysis. And, the result, showed, Neural Network has an overall accuracy of almost 92%. Future work would be to implement more models on different parameters and mostly on larger dataset respectively. (Ismail et al.; 2015)

Day by day every business is becoming saturated, mainly telecommunication due to its highly competitive service providers all over the world, thus by making customer retention the main priority among the companies. With the help of predictive modelling, a methodology was implemented to predict the outcome of churners. Two datasets were used for cross-validation. Various classifiers like; Logistic Regression, K-Nearest Neighbor, Random Forest, Support

Vector Machine, Ridge Classifier, Decision Tree and Gradient Boosting were implemented. Parameters like Accuracy, F1 score, Recall, Precision and Confusion Matrix was analyzed for comparison. Python language was used with various libraries like Pandas, NumPy, Sciklit-Learn etc. the main ones. 10-fold cross validation was done on the above classifiers. And, results showed, Gradient Boosting performed the best in all the metrics closely followed by Random Forest and only 5-6 variables contributed the most for prediction of churn outcome. So, in future work, would be to use more call records, more variables, maybe, customer demographic for more accuracy. (Umayaparvathi and Iyakutti; 2016)

Predictive telecommunication industries have grown at a rapid pace with a fierce competition among rivals and thus by going into saturation and churning among customers. And, giving more focus of keeping customers instead of shifting/acquiring. In this paper, techniques like Decision Trees and Logistic Regression were studied thoroughly. R programming was used for implementation and for web interface Shiny package was used. The implementation is divided into 3 parts: (i) View Performance Analysis, (ii) Testing and (iii) Training. The latter two parts were done in R with various attributes while View Performance Analysis was implemented on Machine Learning algorithms like Decision Trees and Logistic Regression respectively. In this paper, just a statistical survival analysis was conducted, and comparison was done between selected above algorithms. Future work would be the selection of right attributes that will help in Customer Relationship Management (CRM) strategies for the retention of customers. Further-more, developers and decision makers can find suitable algorithms for their data using comparative analysis and more accuracy. (Dalvi et al.; 2016)

For customer retention, prediction of churn is important. The effects of network attributes are the prime focus in this paper, which in turned are linked with various prediction models that can be used in implementing and analyzing various Machine Learning algorithms. Correlation is studied between customer churn and network attributes. Machine Learning algorithms like Logistic Regression, Decision Trees and Neural Network were implemented while SAS Enterprise Miner was used for training models. Various combinations were chosen in SAS Enterprise Miner with the support of Gini Reduction technique, which splits the Decision Tree to run Multi-layer Perceptron for better accuracy. After thorough analysis on various attributes, Neural Network attributes provided more accuracy than any other models. Therefore, result predicted, traditional attribute models, network attribute models and various other combination of models can greatly improve prediction accuracy, thus by, implementing, network attributes play an important role in churn prediction. Future work, would be analyzing more variables like, calling behaviors, network signals, location, topologies, that are important network attributes in telecom field. (Zhang et al.; 2010)

If predicting is done of customers before churning, it can help the decision makers through Customer Relationship Management (CRM) in modern-day-era. In this paper, Deep Learning approach was used known as Research Neural Network (RNN), along with Long Short-Term Memory (LSTM) to learn sequential patterns was established, along with a product-based Recurrent Neural Network (pRNN) approach was initiated. After that, the new model was compared with Random Forest, Logistic Regression, PLSTM and LSTMNN. PLSTM includes product layer where-as LSTMNN doesn't. Results were evaluated based on metrics, Area under roc curve, F-1 score with two new measure, Maximum Profit Measure (MPC) and Expected Maximum Profit Measure (EMPC) for customer churn respectively. Over-sampling was done through SMOTE with random-sampling and Disguise Adversarial Networks (DAN). And results, showed, pRNN scored the most in all metrics with values for AUC, F1-score, MPC and EMPC of 0.817, 0.295, 0.467 and 0.584 respectively. So, to conclude, pRNN plays

a crucial role if integrated with CRM. In future, work can be done on real-world dataset with long-term perspective of churning in mind, with accuracy for both layers needs to be evaluated. (Hu et al.; 2018)

2.4 Conclusion

Based on the results reviewed and gaps identified in the literature review, there was a clear evidence that there is a necessity to develop churn prediction in telecom industries as it is increasing at a faster rate and thus by it answers research question (sec 1.2) and research objectives (sec 1.3) respectively. The next chapter represents scientific methodology approach used to develop churn prediction models.

3 Research Methodology

In Methodology section, the process flow of the project and a detailed explanation of methodology applied with the slight overview of data selection, cleaning and pre-processed data with models and techniques implemented was discussed.

3.1 Methodology

Data Mining and Knowledge Discovery are the pillars of this research work. The term KDD or Knowledge Discovery in Database is a well-known methodology followed in various industries. It is an elaborative process of finding knowledge that has a process flow of various aspects of Data Mining steps like gathering the data, cleaning, evaluating till it emphasized the various high-level Data Mining methods are evaluated from the required research. Furthermore, it is one of the most widely used approached by researchers in Machine Learning, Data Visualization, Statistics, Artificial Intelligence and many more. The main goal of KDD process is the extraction of data from large datasets. (Frawley et al.; 1992)
KDD process is sub-divided mainly into following 5 phases (Fig. 2)- Selection, Pre-processing, Transformation, Data Mining and Evaluation.

a) **Data Selection**: Data Selection Phase involves understanding the data and its respective domain. Does the selected data meet the requirements of the domain, prior knowledge of the domain and does it meet the main goal of the research and organization?! It is also focused proper selection of the data, focusing on the variables/measures and various data samples required for the further study.
b) **Data Pre-processing:** Data pre-processing phase is the one the vital steps of KDD process. It is basically known as Data Cleaning phase. Various tools like Microsoft Excel, R, Python etc. are used for cleaning the data, removing the outliers, refilling/removing the missing fields, rescaling the data as per the requirements. Carefully checking the time sequence information like name, address, location and the steps needed for the changes as per the privacy law depending on the country.
c) **Data Transformation:** In this phase, proper projection of variables in the dataset is the main task. Depending on the target variable, various activities are done like combining the columns, selecting and deleting some variables and correct attributes are searched. After that, dimensionality reduction, class imbalance is performed for the effective use and proper functioning of the dataset.

d) **Data Mining:** One of the most important task of KDD process. Its main goal is to train-test the data and applying various Data Mining models over the split target. Data Mining techniques like; classification, regression, clustering etc. Searching for right patterns, deciding models and parameters are checked for models for the overall KDD process.

e) **Data Evaluation:** The last step of KDD process is Evaluation phase. Interpretation of the models run, overall summarization of the results, analyzing patterns and performance metrics of the models applied were evaluated, thus by integrating KDD process and Knowledge and satisfying the research question.

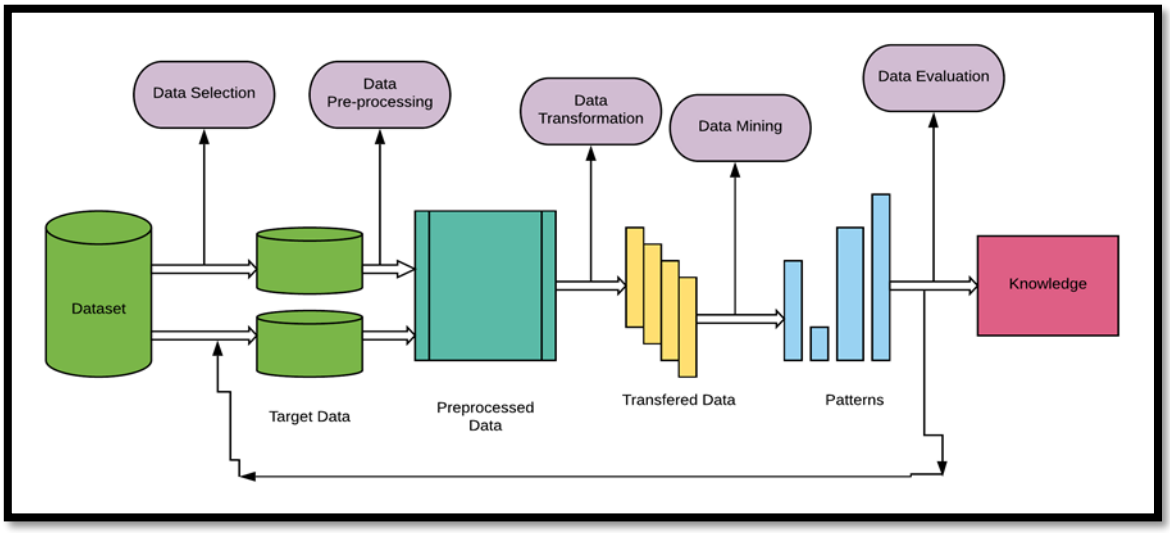

Fig. 2 KDD Methodology

3.2 Data Selection

For the following research Telecom dataset was needed. The main aim of this research is to predict rate of Churn in Telecom Industries using various attributes and parameters. Any information containing name, personal details, location have been avoided for privacy issues while selection of the data.
Churn dataset was accessed from website BigML. BigML dataset with target variable churn was used for prediction on various Data Mining algorithms.

3.3 Data Pre-processing

One of the most important phases in any Data Mining project is Data Pre-processing. The dataset is been thoroughly checked (Table. 3) here for any outliners or class imbalance. It is basically known as data cleaning process. Missing values from cells are either filled or dropped as per the requirement of algorithms. Special characters, signs, symbols etc. were dropped. Conversion of formats were done like; string to categorial or numerical or vice versa as per needed.

Churn Dataset Attributes:

- State, Area Code, Phone Integer - Integer, String.
- Account Length, International Plan, Voice Mail Plan, Number Vmail Messages, Total Day Minutes, Total Day Calls, Total Day Charge, Total Evening Minutes, Total Evening Calls, Total Evening Charge, Total Night Minutes, Total Night Calls, Total Night Charge, Total International Minutes, Total International Calls, Total International Charge, Customer Service Calls - Inter, Categorical.
- Churn - Inter, Categorical, Target.

3.4 Data Cleaning

After dataset is done with cleaning process, selecting appropriate variable and parameter is done to perform for over-fitting and under-fitting. Low-level data is generally replaced with high-level data for data climbing. If required, normalization is performed. Once correct attributes are selected for target variable, feature selection is performed, if the dataset is large, which reduces the training and testing times significantly.

3.5 Data Implementation

As the research is based on classification, it was best to utilize various techniques of Data Mining which are used for classification. Techniques like, Random Forest, K-Nearest Neighbor, Logistic Regression and Decision Trees were performed and for results, parameters like accuracy, recall and precision were checked.

3.6 Data Evaluation

After applying classification models, parameters like accuracy, precision and recall were calculated.

3.7 Models and Techniques

a) Logistic Regression:

One of the basic and important algorithms in Machine Learning is Logistic Regression. For the level of x=1, we can interpret h(x) as the probability. The hypothesis class associated is known as Sigmoid Function.

Sigmoid Function is defined as: $\emptyset_{sig}(z) = \frac{1}{1 + exp(-z)}$

As the plot is of s-shaped that's why it is known as Sigmoid Function. Its hypothesis class is defined as:

$$H_{sig} = \emptyset_{sig} \circ L_d = \{x \to \emptyset_{sig}((w, x)): w \in R^d\}$$

If (w, x) is very large then sig ((w, x)) is close to 1 whereas if (w, x) is very small then sig ((w, x)) is close to 0.

Generally, in Logistic Regression, Empirical Risk Minimization (ERM) which is how bad is prediction if the target variable, which is known as loss function. It can be solved by standard methods with the help of convex functions. A Maximum Likelihood Estimator is a standard statistical approach for maximizing probabilities in dataset which improves the prediction of the target variable with the help of other parameters. (Shalev-Shwartz and Ben-David; 2017)

b) K Nearest Neighbor:

KNN is one of the simplest algorithms of all the Machine Learning algorithms. It can be used in both classification and Regression problems. It is very easy to interpret; calculation time is very less, and predictive power is the most. Firstly, the training set is studied thoroughly, then prediction of new instance based on closet neighbors is taken into consideration. In this, real and predicted values are analyzed for accuracy. The main aim of KNN is, domain points features are relevant to their libeling's as compared to closely-points on the same label.

For example, X is having a metric function p. Then p: X x X $\rightarrow$R is a function that returns the distance between the 2 points, which is known as Euclidean distance. So, if X = R^d , then p (x, x') = ||x – x'|| = $\sqrt{\sum di = 1 (xi - xi')2}$
The factor of K in KNN is where exactly K influence in the algorithm. So, even if the model is simple, it can give highly competitive results. (Shalev-Shwartz and Ben-David; 2017)

c) Decision Trees:

Decision Tree is a predictor, which are initiated with base parameters with different levels and branches for every node. For example, h: x $\rightarrow$ y, predicts the label associated with x by traveling to a root node as a Decision Tree.

At each node, a successor child is chosen by splitting. One of the main problems with Decision Trees is where to split the data exactly. Training and Testing is done through k-fold cross validation, as it doesn't violate accuracy and avoids overfitting. The thresholding value of a single feature is a good technique of splitting the data at various internal nodes.
For overfitting issue, Minimum Description Length (MDL) is used as it has two options; where one hand is not too large while the other hand fits the data. The effect of splitting a single node is called iteration and it is measure by 'gain' metric with Gini Index. For further iteration check, Iterative Dichotomizer 3 (ID 3) is used. (Shalev-Shwartz and Ben-David; 2017)

d) Random Forest:

A mixture of trees that are used for prediction algorithms is known as Random Forest. It is the ensemble version of Decision Trees as the class of Decision Trees has infinite dimensions, so to reduce the danger of over-fitting, Random Forest is constructed. Random vectors are generated and sampled into small trees as per the required model. Several Decision Trees are executed, and the most popular class is voted for further consideration. Random Forest classifier is defined as the combination of multiple classifiers which are in a tree-type structure. Its equation is: (h (x, θ_k), k = 1…)

Herein, (θ_k) i.e. random vectors are independent and disturbed identically.
x= class input

The prediction of Random Forest is done by majority vote over the predictions of individual trees simultaneously. (Breiman; 2001)

e) Naïve Bayes:

Naïve Bayes algorithm is based on Bayes Theorem. It is a classical algorithm on generative assumptions and parameters used while estimation of the model process.
For example; consider a predicting label $y \in \{0,1\}$

with feature vectors $x = (x_1 \ldots. x_d)$ where x_i is in $\{0,1\}$

Then, Bayes classifier is defined as: $h_{bayes}(x) = \text{argmax}_{y \in \{0,1\}} P|Y = y|X = x|$
Where probability function is defined as 2^{nd} parameters for a value in the range of $x \in \{0, 1\}^d$.

Thus, in Naïve Bayes, naïve assumptions are given to the label with features independent of each other with the resulting classifier known as Naïve Bayes Classifier. These classifiers are extremely fast as compared to other classifiers, because the curse of dimensionality is reduced because of decoupling of class. For classification, GaussianNB is used for implementation of Naïve Bayes algorithm. (Shalev-Shwartz and Ben-David; 2017)

f) AdaBoost Algorithm:

AdaBoost is known as boosting algorithm. It is an adaptive algorithm which adapts to error rates of individual weak hypothesis. It has no parameters and it is fast and easy to implement. It focuses on weak learning algorithms, instead of trying to design a weak one. In each weak hypothesis, it gets us confidence scores along with predicted classifications.

For example: - $(x_1, y_1)\ldots(x_m, y_m)$ where,
x_i = domain space X and y_1 = each label in set Y= {-1, +1}
and algorithm is t = 1…T.

The weak learner's job at weak hypothesis $h_t = X \rightarrow \mathbb{R}$, at distribution D_t
The hypothesis error is measured by $\in_t = Pr_{i \sim Dt}[h_t(x_i) \neq y_i] = \Sigma_{iht(xt) \neq yt} D_t(i)$
Thus, to minimize error and over-fitting issues, boosting is preferred with AdaBoost algorithm as one of the vital boosting algorithms involved in Machine Learning. (Jinbo et al; 2007)

g) Artificial Neural Network

ANN is a computation model of the structure Neural Networks. ANN are formal computated models which are constructed and remodified into computated paradigm. It consists of nodes which are interconnected.

It is defined as: $y_i = f_i(\sum^n_{j=1} w_{ij}x_j - \theta_i)$ where;
i = node, y_i = output of the node,
x_j = j^{th} output of the node,
w_{ij} = connection between nodes,
θ_i = threshold, f_i = non-linear function

Whereas summation is defined as $y_i = f_i(\sum^n_{j,\, k=1} w_{ijk}x_jx_k - \theta_i)$ where all symbols have equal definitions at all nodes. That's why, ANN is known a topological structure. (Yao, 1999)

3.8 Conclusion

Hence, for this project a KDD methodology approach is used as it fits best for this research project. Also, the process flow diagram can help decision makers, analysts etc. on how to approach this problem.

4 Implementation of Customer Churn Prediction Models

In this section, the project implementation of the research was conducted with steps like data selection, cleaning, transformation, mining and implementation of various machine learning algorithms was discussed (Fig. 3).

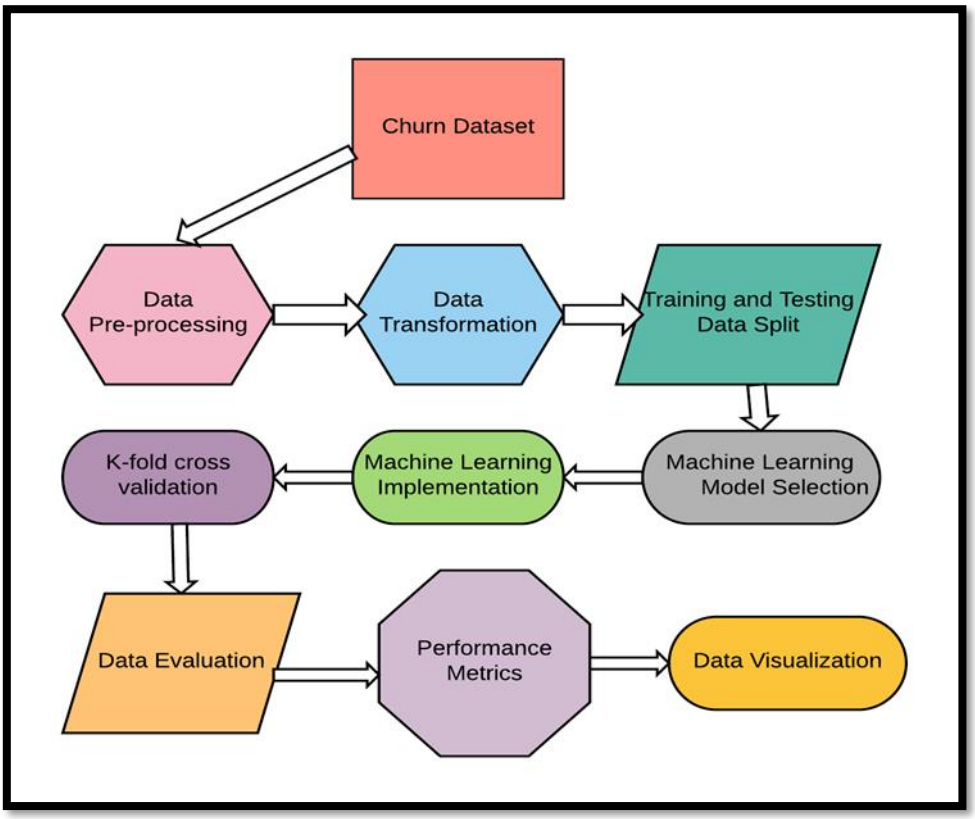

Fig. 3 Process Flow Diagram of the Research

4.1 Data Exploration

One of the essential aspects of any research is understanding the dataset and exploring through its variables. In this research, it was observed, almost all variables expect (state and phone number) are valuable for further analysis.

4.2 Data Cleaning

One of the most important phases in implementation of Machine Learning Models. In this research, Python programming language was used for cleaning process. Firstly, levels of target variable, Churn [True, False] was converted to [1,0] for analysis. String columns like State and Phone Number was removed as they were irrelevant for the research. Further-more, columns like International Plan and Voice Mail Plan both had levels [Yes, no] which were converted to [1,0] respectively for better analysis. As, dataset was mostly cleaned, no null values, no empty cells, no missing values were ensured. Outliers were not removed as a smaller number of outliers won't affect the quality of model training as dataset contains less rows and columns.

4.3 Data Transformation

One of the vital steps in the process flow is Data Transformation.

4.3.1 Determining the Important Variables

Before applying Machine Learning algorithms, one of the most important steps is to determine the right set of variables, which are known as predictors. To determine that its p-value is with target variable here which is 'churn' is calculated (Fig. 4). Its defined as, the probability of the sample data observed has the null hypothesis true. If p-value is less than 0.05, then null hypothesis is rejected, thus p-value is above 0.05 is rejected.

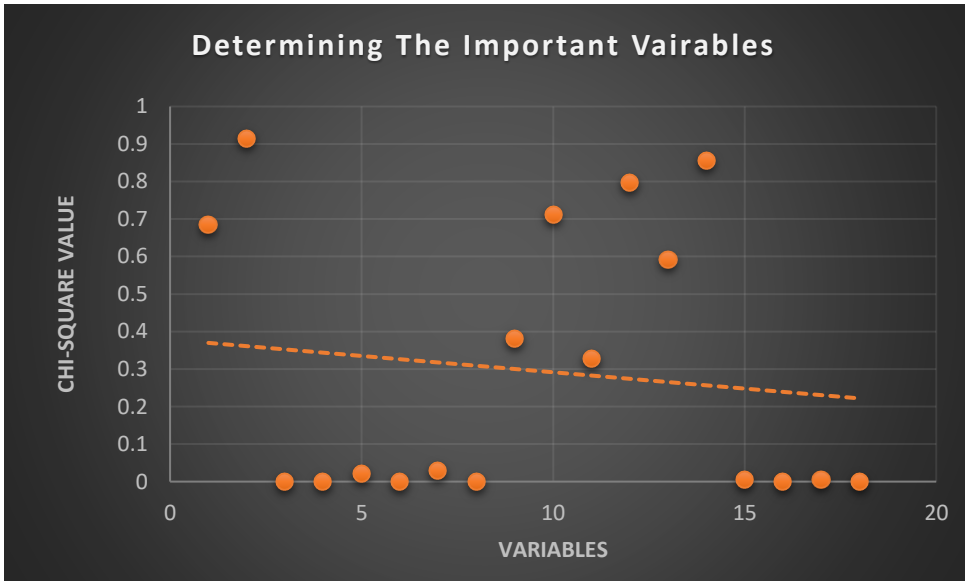

Fig. 4 Chi-Square Values vs Variables

4.3.2 Determining the Correlation

Next step is to determine the correlation the correlation between variables (Fig. 5). It determines the determines the dependence among the variables. As seen, by the heatmap below, except few variables, many are correlated with target variable, churn.

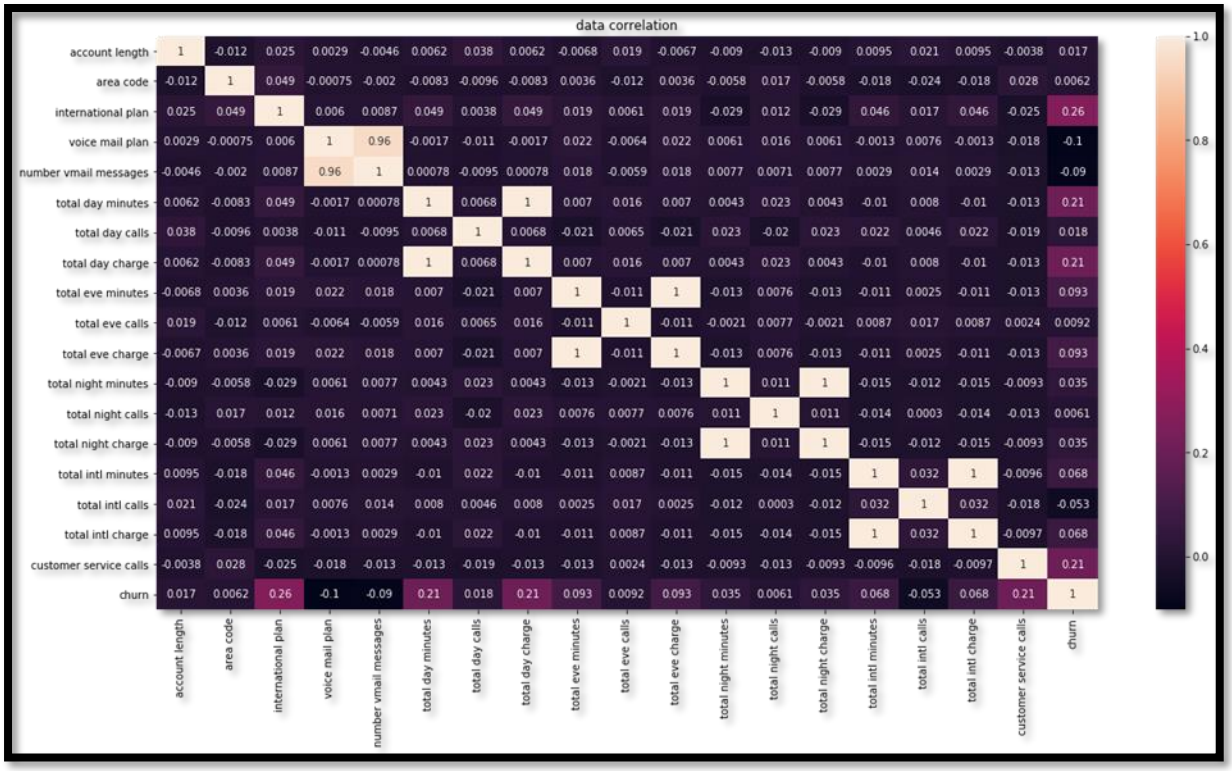

Fig. 5 Correlation Matrix

4.3.3 Checking about the Class Imbalance

Data features plays a decisive role in data implementation as different datasets have different features, which might pose some issues while implementation, that's why to check whether is imbalanced or not is necessary. As, imbalanced dataset can cause accuracy issues. As, seen from the bar-chart, the target variable 'churn' is imbalanced (Fig. 6). Outlier analysis is skipped as there is an impact over data due to class imbalance which can solved further as Random Forest is implemented which can deal with outliers.

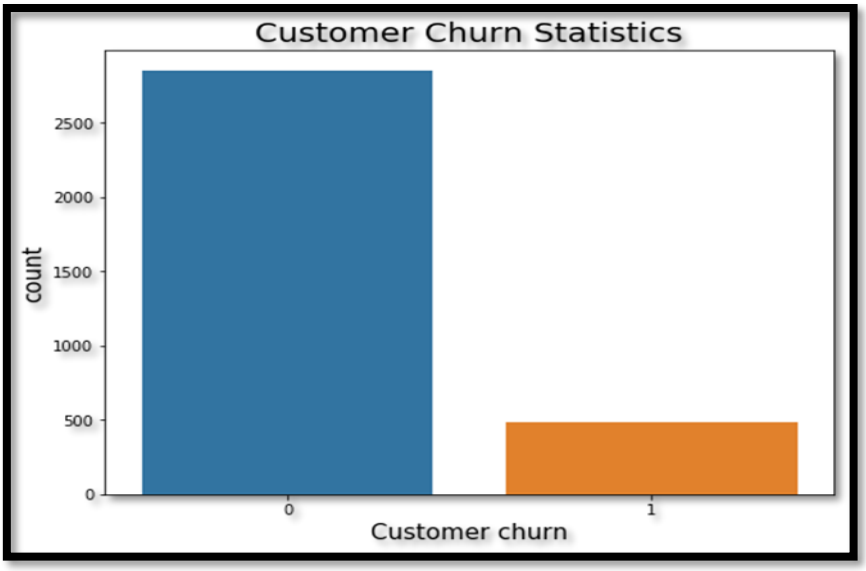

Fig.6 Customer Churn Statistics

4.3.4 Checking Over-Sampling

Over-sampling is done about class imbalance. There are 2 ways it is done: (i) Over-sample and (ii) Under-sample. To increase the strength of minority class, random over-sampling is preferred. In this research, Synthetic Minority Over-Sampling Technique (SMOTE) technique is used which can implemented with imblearn package in Python.

4.4 Data Mining

To start with training models, the data was split into training and testing. Data was divided into 90% and 10% respectively. For analysis, over-fitting and under-fitting was considered. The interpretation of performance of the model usually depends on factors like; problem solution, metrics used, prediction of model. All the models were evaluated in two ways of validation: (i) solo run and (ii) k-fold validation. In this research, confusion matrix was considered as it is used for prediction and original classification. Results were evaluated based on performance metrics i.e. accuracy, precision and recall.

4.5 Implementation of Logistic Algorithm

As the target variable was dichotomous, there was a slight problem in implementing because the target variable was in string format which needs to be converted to integer [0,1]. For better running of the model, the inputs were converted to contiguous variable and output were categorical variable respectively.

4.6 Implementation of KNN Algorithm

Categorial independent variable plays a crucial role in KNN algorithm. It forms classes of all the attributes which form circles around each class. Loop iteration was done on k-Neighbor Classifier to get an efficient value of k. The optimum value of k was 5 which was later used

for training again for cross-validation for better performance (Fig. 7). KNeighborsClassifier package was used.

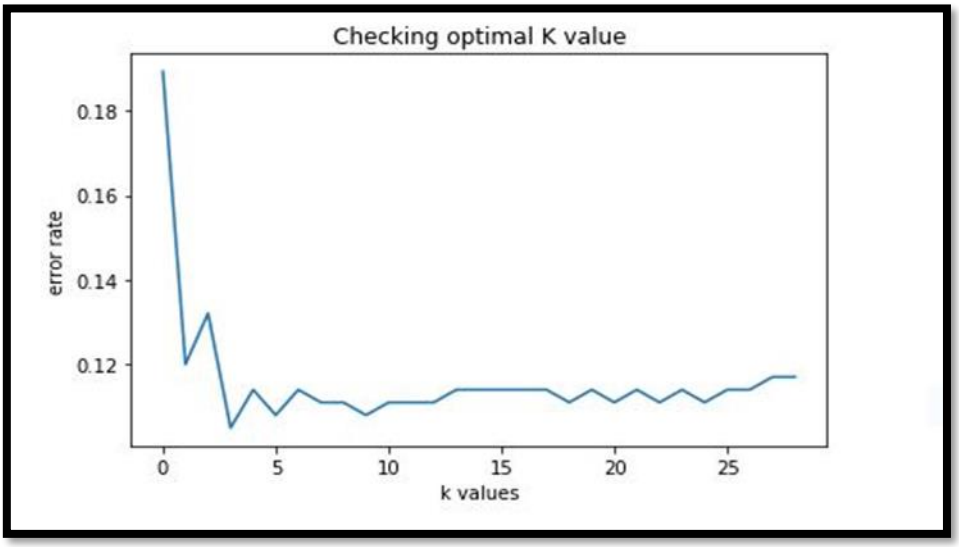

Fig. 7 Checking Optimal K Value

4.7 Implementation of Decision Tree Algorithm

Decision Tree was implemented with the base parameters. As, number of levels and branches are different in every node, everyone was given a base parameter. Training and testing with 90% and 10% was done with k-fold cross validation respectively. K-fold cross validation was done because it makes use of every data subset and avoid over-fitting and under-fitting problem. DecisionTreeClassifier package was used.

4.8 Implementation of Random Forest Algorithm

The ensemble version of Decision Tree is known as Random Forest. It is like a voting-based algorithm, as, the highest voter(tree) is selected. After several Decision Tree were executed, the tree with maximum votes is considered. Default values are taken at every split, with default values for number of levels too. Random Forest is generally used after Decision Tree as it avoids all over-fitting issues. RandomForestClassifier package was used.

4.9 Implementation of Naïve Bayes

The supervised learning algorithm is known as Naïve Bayes. Training and Testing was done with 90% and 10% respectively. Real and predicted values are checked as the algorithm is focused on assumptions also k-fold cross validation after train-test-split. GaussianNBClassifier package was used.

4.10 Implementation of AdaBoost Algorithm

Next algorithm implemented is boosting algorithm known as AdaBoost. Training and Testing was done with the same ratio as for the other algorithms. Real and predicted values were checked for consistency after that k-fold cross validation was applied. AdaBoost Classifier package was used.

4.11 Implementation of Artificial Neural Networks

Lastly, Artificial Neural Network was implemented. Training and testing were done with same ratio. Sequential package was used with Relu and Sigmoid as activation functions, while compilation loss of binary classification [0,1] is calculated with metrics accuracy and validated data accuracy. Selected number of iterations was trained to fit the entire dataset, which are known epochs. The predicted data gets validated across the real data, for the model prediction, on which accuracy was measured. Sklearn package with keras and TensorFlow backend which was used and then, algorithm parameters were evaluated with mean training and testing loss and accuracy for both respectively.

4.11 Conclusion

After pre-processing, implementation of all the machine learning models were conducted using Python. The next section represents the evaluation of models.

5 Data Evaluation and Results

In this section, Data Mining models for classification are evaluated and compared based on performance metrics.

5.1 Performance Metrics

Depending upon the churn problem model, prediction was interpreted with performance metrics like accuracy, recall and precision. Confusion matrix is a table that is used to describe the performance of a model in which true values are known.
There are 4 types of parameters (Table. 1): -

- True Positive: The value of both classes i.e., actual and predicted are yes.
- True Negative: The value of both classes i.e., actual and predicted are no.
- False Positive: When actual class is no but predicted class is yes.
- False Negative: When actual class is yes but predicted class is no.

Table. 1 Confusion Matrix

	Predicted Class		
Actual Class		Class = Yes	Class = No
	Class = Yes	True Positive	False Negative
	Class = No	False Negative	True Negative

Following parameters were used in this research:

- **Accuracy:** Accuracy is the most logical performance measure. It can be calculated from the confusion matrix. It is simply the ratio of correctly predicted observation to total observations. When datasets are symmetric, accuracy is a great measure, but if they are not, other performance metrics are evaluated.

$$\text{Accuracy} = \frac{True\ Positive + True\ Negative}{True\ Positive + False\ Positive + False\ Negative + True\ Negative}$$

- **Precision:** The ratio of correctly predicted positive observations to total predicted positive observations.

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

- **Recall:** The ratio of correctly predicted positive observations to all observations in actual class – yes.

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

5.2 Evaluation of Logistic Regression

Initially, Logistic Regression was applied using sklearn package in Python. The linear model was trained using controlled parameters; cross-validation and linear model. Result showed, the percentage of people who are likely to churn yielded an accuracy of 86% with precision and recall of 21% and 55% respectively.

```
Logistic Regression metrics calculated

Accuracy =  86.48648648648648  %
Precision =  21.27659574468085  %
Recall =  55.55555555555556  %
```

5.3 Evaluation of k-Nearest Neighbor

Secondly, KNN also was applied, sklearn package was used. To determine the optional k-value, K-Neighbor classifier was implemented with the range of 1 to 30. From the plot, it was evident that k value seems to be stable after 22 and at k=5, it seems to be lowest. After, real

values and predicted values were checked for performance metrics, result showed accuracy of 88% with precision and recall of 29% and 73% respectively.

```
KNN metrics calculated

Accuracy =  88.58858858858859  %
Precision =  29.78723404255319  %
Recall =  73.68421052631578  %
```

5.4 Evaluation of Decision Trees

Next was, Decision Trees, as there was no problem in overfitting, 100 number of trees was selected for implementation. sklearn package was used. Decision Tree experienced high accuracy of 95%, with high precision and recall values of 85% for both respectively.

```
Decision Trees metrics calculated

Accuracy =  95.7957957957958  %
Precision =  85.1063829787234  %
Recall =  85.1063829787234  %
```

5.5 Evaluation of Random Forest

Next algorithm was the ensemble-version of Decision Tree, named, Random Forest. The package used was sklearn. And, results showed, performance metrics was highest in everything with accuracy, precision and recall showing 97%, 87% and 97.6% respectively.

```
Random forest metrics calculated

Accuracy =  97.8978978978979  %
Precision =  87.2340425531915  %
Recall =  97.61904761904762  %
```

5.6 Evaluation of Naïve Bayes

One of the supervised algorithms, Naïve Bayes was implemented next. Gaussian Naïve Bayes classifier with sklearn package was used for performance metrices. Results were in the mid-range showing 85.88% accuracy, with precision and recall of 55.31% and 50.00% respectively.

```
Naive Bayes metrics calculated

Accuracy =  85.88588588588588  %
Precision =  55.319148936170215  %
Recall =  50.0  %
```

5.7 Evaluation of AdaBoost

One of the boosting algorithms, AdaBoost was implemented next. With the help of sklearn package and AdaBoost Classifier, performance metrics were calculated with accuracy of 89.78 %, precision of 44.68% and high recall value of 72.41% respectively.

```
Ada Boost metrics calculated

Accuracy =  89.7897897897898  %
Precision =  44.680851063829785  %
Recall =  72.41379310344827  %
```

5.7 Evaluation of Artificial Neural Network

One of the sequential algorithms, Artificial Neural Networks was implemented lastly. Relu and Sigmoid functions were used with epochs to train entire dataset and then ANN metrics at every layer was evaluated with results of training loss at 234.89 with training accuracy at 60 epochs of 85.41% and testing loss at 227.49 and testing accuracy at 85.88 % respectively.

```
Total params: 411
Trainable params: 411
Non-trainable params: 0
_________________________________________________________________
Training Loss =  234.89790767980585
Training Accuracy =  85.417223073033
Testing Loss =  227.49264059473032
Testing Accuracy =  85.88588672715267
```

As, the various algorithms implemented to predict customer churn performed differently at different computated time. Also, k-fold cross validation avoided over-fitting issues and provided better results too. The table (Table. 2), the values of performance metrics for accuracy, precision and recall of each model. Judging by the table below shows, Random Forest performed better in all the three-performance metrics with values of accuracy (97.89%), precision (87.23%) and recall (97.61%) followed by Decision Trees.

Table. 2 Results of Implemented Models

Models	Accuracy	Precision	Recall
Logistic Regression	86.48	21.27	55.55
K-Nearest Neighbor	88.58	29.78	73.68
Decision Tree	95.19	85.10	81.62
Random Forest	97.89	87.23	97.61
Naïve Bayes	85.88	55.13	50.00
AdaBoost	89.78	44.68	72.41
Artificial Neural Network	85.46	-	-

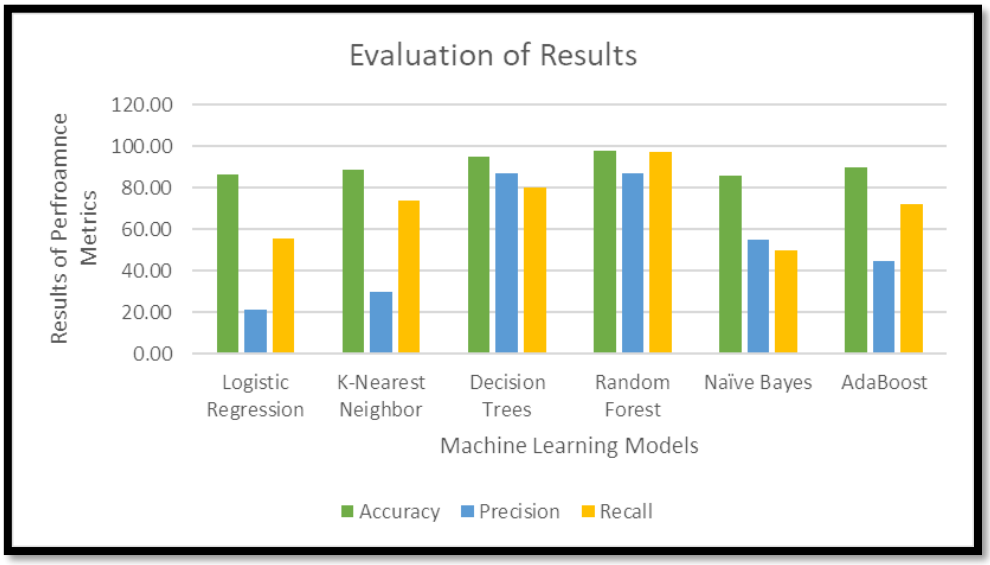

Fig. 8 Evaluation of Results based on Performance Metrics

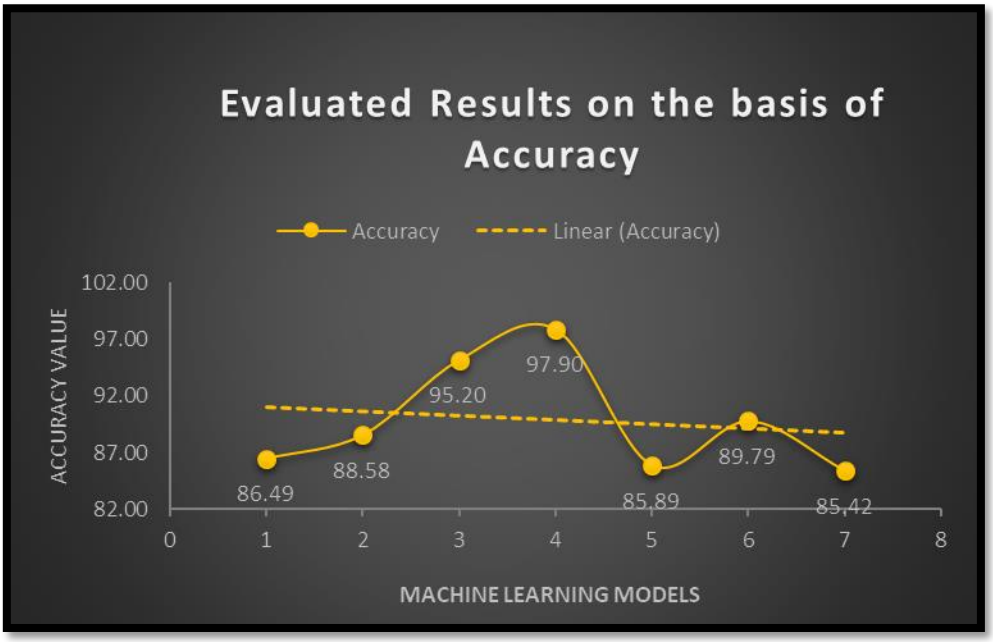

Fig. 9 Evaluated Results based on Accuracy

5 Conclusion and Future work

As several researchers had studied and implemented various classification models in prediction of churn, which was discussed in literature review section. In this research, as the dataset was imbalanced, variables with low predictive power was eliminated and which has strong relationship with target variable was chosen for better results. The results of the tests based on performance metrics was concluded that prediction of churn in telecom industries, machine learning can be the easy and efficient approach.

Amongst, the algorithms applied, Random Forest (RF) got the best results. It gave an accuracy of (97.89%) with high values of precision (87.23 %) and recall (97.61%) too (Fig. 8 and Fig. 9). In today's fast-growing world one of the important jobs in Client Management is Customer Churn Management. If given, proper time for Artificial Neural Networks (ANN), although it can be time consuming and expensive, it may improve the efficiency of prediction. Although for the academic purposes, Random Forest is quite efficient model.

In future work, research must be done with larger dataset, more factors of telecom companies need to be studied for analysis. As, Neural Networks, takes slightly more time for computation, a simple machine learning algorithm with different attributes with different settings may provide good accurate results. Also, with different classifiers, different boosting methods can be performed on real-time and adaptive dataset. Furthermore, if integrated with Customer Relationship Management (CRM) technologies can give insights for decision makers for better retention methods.

Acknowledgment

I would like to thank my supervisor Dr. Catherine Mulwa whose feedback and guidance helped me immensely. She guided and cleared the doubts on time. Furthermore, I thank National College of Ireland for providing proper support and resources needed for the research. I would also like to express my gratitude towards to School of Computing, National College of Ireland.

References

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

Brandusoiu, I.B. and Toderean, G., 2016. Churn prediction in the telecommunications sector using neural networks. Acta Technica Napocensis, 57(1), p.27.

Brândușoiu, I., Toderean, G. and Beleiu, H., 2016. Methods for churn prediction in the pre-paid mobile telecommunication industry.

Breiman, L., 2001. Random forests. Machine learning, 45(1), pp.5-32.

Chouiekh, A., 2017, April. Machine Learning techniques applied to prepaid subscribers: case study on the telecom industry of Morocco. In 2017 Intelligent Systems and Computer Vision (ISCV) (pp. 1-8). IEEE.

Dalvi, P.K., Khandge, S.K., Deomore, A., Bankar, A. and Kanade, V.A., 2016, March. Analysis of customer churn prediction in telecom industry using decision trees and logistic regression. In Colossal Data Analysis and Networking (CDAN), Symposium on (pp. 1-4). IEEE.

Frawley, W.J., Piatetsky-Shapiro, G. and Matheus, C.J., 1992. Knowledge discovery in databases: An overview. AI magazine, 13(3), pp.57-57.

Hu, J., Zhuang, Y., Yang, J., Lei, L., Huang, M., Zhu, R. and Dong, S., 2018, December. pRNN: A Recurrent Neural Network based Approach for Customer Churn Prediction in Telecommunication Sector. In 2018 IEEE International Conference on Big Data (Big Data) (pp. 4081-4085). IEEE.

Ismail, M.R., Awang, M.K., Rahman, M.N.A. and Makhtar, M., 2015. A multi-layer perceptron approach for customer churn prediction. International Journal of Multimedia and Ubiquitous Engineering, 10(7), pp.213-222.

Jinbo, S., Xiu, L. and Wenhuang, L., 2007, June. The Application of AdaBoost in Customer Churn Prediction. In 2007 International Conference on Service Systems and Service Management (pp. 1-6). IEEE.

Khan, F. and Kozat, S.S., 2017, May. Sequential Churn Prediction and Analysis of Cellular Network Users–A multi-class, multi-label perspective. In 2017 25th Signal Processing and Communications Applications Conference, SIU 2017. Institute of Electrical and Electronics Engineers Inc.

Meher, A.K., Wilson, J. and Prashanth, R., 2017, July. Towards a Large-Scale Practical Churn Model for Prepaid Mobile Markets. In Industrial Conference on Data Mining (pp. 93-106). Springer, Cham

Mozer, Michael & Wolniewicz, Richard & B. Grimes, David & Johnson, Eric & Kaushansky, Howard. (2000). Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry. Neural Networks, IEEE Transactions on. 11. 690 - 696. 10.1109/72.846740.

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

Shalev-Shwartz, S. and Ben-David, S., 2014. Understanding machine learning: From theory to algorithms. Cambridge university press.

Umayaparvathi, V. and Iyakutti, K., 2016, March. Attribute selection and Customer Churn Prediction in telecom industry. In Data Mining and Advanced Computing (SAPIENCE), International Conference on (pp. 84-90). IEEE.

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.

Xia, X., Zeng, L. and Yu, R., 2018, October. HMM of Telecommunication Big Data for Consumer Churn Prediction. In 2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI) (pp. 1903-1910). IEEE.

Xie, L., Li, D. and Xiao, J., 2011, October. Feature selection-based transfer ensemble model for customer churn prediction. In 2011 International Conference on System science, Engineering design and Manufacturing informatization (Vol. 2, pp. 134-137). IEEE.

Yan, L., Fassino, M. and Baldasare, P., 2005, July. Predicting customer behavior via calling links. In Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005. (Vol. 4, pp. 2555-2560). IEEE

Yao, X., 1999. Evolving artificial neural networks. Proceedings of the IEEE, 87(9), pp.1423-1447.

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

Zhang, X., Liu, Z., Yang, X., Shi, W. and Wang, Q., 2010, July. Predicting customer churn by integrating the effect of the customer contact network. In Service Operations and Logistics and Informatics (SOLI), 2010 IEEE International Conference on (pp. 392-397). IEEE.