Telecom Churn Prediction Using Seven Machine Learning Experiments integrating Features engineering and Normalization

Hemlata Jain (mailhemajain@gmail.com)
Poornima University https://orcid.org/0000-0003-0679-0835
Ajay Khunteta
PU: Poornima University
Sumit Private Shrivastav
Manipal University - Jaipur Campus

Research Article

Keywords: Machine Learning, Churn Prediction, Random Forest, Feature Importance, PCA, CNN, XGBoost;

Posted Date: April 5th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-239201/v1

Telecom Churn Prediction Using Seven Machine Learning Experiments integrating Features engineering and normalisation

Hemlata Jain a*, Ajay Khunteta, Sumit Srivastava

Abstract – Machine Learning and Deep learning classification has become an important topic in the area of Telecom Churn Prediction. Researchers have come out with very efficient experiments for Churn Prediction and have given a new direction to the telecommunication Industry to save their customers. Companies are eagerly developing the models for predicting churn and putting their efforts to save the potential churners. Therefore, for a better churn prediction model, finding the factors of churn is very important. This study is aiming to find the factors of user's churn by evaluating their past service usage details. For this purpose, study is taking the advantage of feature importance, feature normalisation, feature correlation and feature extraction. After feature selection and extraction this study performing seven different experiments on the dataset to bring out the best results and compared the techniques. First Experiment includes a hybrid model of Decision tree and Logistic Regression, second experiment include PCA with Logistic Regression and Logit Boost, third experiment using a Deep Learning Technique that is CNN-VAE (Convolutional Neural Network with Variational Autoencoder), Fourth, fifth, sixth and seventh experiments was done on Logistic Regression, Logit Boost, XGBoost and Random Forest respectively. First four experiments are hybrid models and rest are using standalone techniques. The Orange dataset was used in this technique which has 3333 subscriber's entries and 21 features. On the other hand, these experiments are compared with already existing models that have been developed in literature studies. The performance was evaluated using Accuracy, Precision, Recall rate, F-measure, Confusion Matrix, Marco Average and Weighted Average. This study proved to get better results as compared to old models. Random Forest outperformed in this study by achieving 95% Accuracy and all other experiments also produced very good results. The study states the importance of data mining techniques for a churn prediction model and proposes a very good comparison model where all machine Learning Standalone techniques, Deep Learning Technique and hybrid models with Feature Extraction tasks are being used and compared on the same dataset to evaluate the techniques performance better.

Keywords: **Machine Learning, Churn Prediction, Random Forest, Feature Importance, PCA, CNN, XGBoost;**

I. Introduction

In the present world, digital media has become a powerful tool for managing large data especially in the telecom industry where there is an essential need to store large dataset. A huge volume of data is being generated by telecom companies at an exceedingly fast rate [**8**]. The large data generated in these companies are bulky and managing and accessing the information out of this data is a main challenging task.

Hemlata Jain
2013PUSBAPHDE02477, Computer Science School of Basic and Applied Sciences, Poornima University Jaipur-303905, India
E-mail: mailhemajain@gmail.com,

Ajay Khunteta
Professor at Poornima Group of Colleges, Jaipur-303905, India
E-mail: khutetaajay@poornima.org,

Sumit Srivastava
Professor, Department of Information Technology, Manipal University Jaipur, Jaipur-303007, India E-mail: sumit.310879@gmail.com

Data mining resolves this issue. Data mining is the process of analysing data from various aspects and summarizes it into valuable information [**4**]. Since the early 1960 Data Mining techniques have been considered to be an area of applied artificial intelligence [**3**]. A large number of data mining techniques available to find out the hidden knowledge about the customer data. Some of them are clustering, classification, attribute selection, Association etc. A churn prediction model is purely based on the customers past service usages behaviour data. Telecom companies develop churn prediction models to increase their client share, maximise profit and stay active in a competitive environment. A consumer churn is switching from one service provider to another. In today's competitive environment customers have multiple options for better services and prices.

There are multiple reasons for customer churn. Unlike post-paid customers, prepaid customers are not bound to a service provider and may churn at any time [**8**]. Customer churn normally happens due to lack of engagement, lack of promotions or new offers, lack of customer service support, high call rate or SMS charges, non-payment bills, fraud or miss usages of services and change of location. When the number of customers dropping below it causes major revenue loss. Churn Prediction model uses a telecom database for prediction. It analyses customer's behaviour and predicts the future churners. Telecom databases are running into terabytes and petabytes having large numbers of attributes and hence to model these complex datasets it needs advanced data sciences models to be developed.

There is a huge advancement in the field of big data and machine learning. Due to that many models have been developed widely. Researchers have developed and compared different machine learning techniques in their models.

Research [**3**] contributed to develop a churn prediction model to assist telecom companies for predicting customers who are near to churn. This research compared the machine learning techniques that are XGBoost, Decision Tree, Random Forest, Gradient Boosted Machine Tree. This research analysed the factors which played an important role in customer churn by feature engineering and selection. [**8]** also identified the factors WHICH LEAD TO CUSTOMER CHURN by selecting features using correlation attributes, ranking attributes and information gain. These researches proved factor identification is useful for churn prediction models. Research [**10**] proved that data preparation techniques they choose affects the churn prediction model performance and enhances Logistic Regression is competitive with advanced single ensemble data mining techniques. [**12**] have shown that customer misclassification, the amount of service they used and some demographic attributes plays an important role in customer churn. This research used binomial Logistic Regression for the prediction.

Research [**9]** used different data mining techniques for churn prediction and compared them. For comparing the different techniques this research used different evaluation metrics and also worked on extracting datasets features. DT handles interaction effects between variables very well but has difficulties to handle linear relations between variables. For LR the opposite is true: it handles linear relations between variables very well but it does not detect and accommodate interaction effects between variables [**1**].

This study added multiple functionalities of feature engineering and selection at one place and worked on improving the model performance. this study used literature models and identified some new work and applied multiple feature analysis tasks to improve performance and at last compare them with each other and with literature works. This study using correlation matrix, feature engineering, feature importance, handling categorical feature, handling continuous features, normalising features and giving this altered and informative dataset to four different hybrid models and to five standalone techniques. some hybrid models already have some feature extraction functionality in it therefore it added the double feature extraction capability to the model. The idea of comparing hybrid techniques and standalone techniques is very helpful for future research and the double feature selection process really worked on improving performance.

This paper is organized in following sections: section II: Literature review highlighting work already done by researchers; Section III briefly describes methodologies leveraged in this study. In Section IV Proposed work and database are detailed while in section V Results and discussion are discussed, section VI is the conclusion of this paper detailing what the author has accomplished and what is planned in future.

II. Literature Review

Telecom Churn prediction is a crucial factor for companies to be concerned about. Many works have been done on the same. In literature Many techniques and methods have been used in prediction models. Machine learning and data mining were the most used approaches in literature. Most researchers have added one technique for knowledge gain and one technique for prediction and many of researches included factor indemnification at most.

Some of the literature are compared and discussed in this section. Table1 shows a very good state of art comparison of the literature work that has been used in this study and tried to get better results.

Table 1: A State Art Comparison of Literature.

Author	Title and Journal	Year	What?	Techniques	Dataset	Results
[2]	An Ensemble Approach for Efficient Churn Prediction in Telecom Industry International Journal of Database Theory and Application)	2016	Churn Prediction by using customer usage history	*decision trees, ensembles, Random Forest and Gradient Boosted*	Orange Dataset, 3333 Subscribers	DT Acc Sec Spe 86 .21 .96 RF 91 .47 .98 GB 91 .49 .98
[1]	A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees (Elsevier)	2018	In the first stage customer segments are identified using decision rules and in the second stage a model is created for every leaf of this tree.	Logistic Regression and Decision Tree	Fourteen Datasets were used	AUC0.63 TDL 1.561
[3]	Customer churn prediction in telecom using machine learning in big data platform (Journal of Big data)	2019	machine learning techniques on big data platform and builds a new way of features' engineering and selection.	Decision Tree, Random Forest, Gradient Boosted Machine Tree "GBM" and Extreme Gradient Boosting"XGBOOST	SyriaTel dataset	AUC After FE (%) XGB 93.3 GSM 90.89 RF 87.76 DT 83
[5]	Echo State Network with SVM-readout for Customer Churn Prediction (*IEEE Conference)*	2015	use of an Echo State Network (ESN) with a Support Vector Machine (SVM) read out	Eco state network, support vector machine	Orange dataset 3333 subscribers, Jordanian cellular with 5000 subscribers	ESN with SVM ACC 99 Churn Rate .97 Hit RATE .99
[8]	A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector (IEEE Access)	2019	Classification as well as clustering for churn factor identification	Random Forest, K-means	South Asia GSM Tekleco with 64107 subscribers, orange dataset with 3333 subscribers	RM Results on SAG Telecom TP - .89 FP - .24 Pre - .89 Rec - .89 FM - .88 ROC - 94 RM Results on Orange Telecom TP - .89 FP - .57 Pre - .89 Rec - .89 FM - .88 ROC - .84

[9]	Customer Churn Prediction in Telecommunication Sector using Rough Set Approach (NeroComputing)	2016	ule-based decision-making technique, based on rough set theory (RST),	Exhaustive Algorithm (EA), Genetic Algorithm (GA), Covering Algorithm (CA) and the LEM2 algorithm	Orange wirh 3333 subscribers	EA GA CA LEM2 Coverage 1 1 0.64 0.668 Precision 0.72 0.86 0.47 0.67 Recall 0.74 1.00 0.50 0.73 MisErr 0.074 0.019 0.122 0.067 Accuracy 0.926 0.981 0.878 0.993 Specification 0.96 0.98 0.93 0.96 F-measure 0.726 0.925 0.487 0.698
1	Automated Feature Selection and Churn Prediction using Deep Learning Models (IRJET)	2017	Feature extraction technique with prediction	ANN, CNN	Cell2cell with 70831 Subscribers, CrowdAnalitix with 3333 subscribers	accuracy of 92.77%.

As Table 1 depicts literature work have been used different techniques with different proposed methods. There are multiple techniques used in literature with very useful insights. Some of them used in this research with some additional functionality to improve the model performance. Researchers used many machine learning techniques like Decision Tree, Logistic Regression, SVM, Random Forest, XGBoost and some hybrid techniques too. feature selection was the most focused content in the literature. All techniques outperformed as shown in table 1. Research [5] and [1] both used a hybrid technique. Both techniques outperformed. Research [5] outperformed with the accuracy 99%. Both hybrid techniques used one technique for clustering similar data and one technique for prediction on the clustered data. All other researchers used standalone technique for prediction with one kind of feature selection methods where some researchers used Deep Learning techniques. all researches used different datasets for different techniques therefore it is very difficult or it will not be a good idea to come to the conclusion that the one technique is best based on the literature data. Techniques should be compared on the same dataset for comparison. This study proposes a study where every kind of technique is being used for churn prediction and comparing on the same dataset to evaluate the technique performance. This study uses hybrid techniques, standalone techniques and Deep Learning techniques too with Feature extraction tasks.

III. Proposed Work

This study is proposing a good example of KDD. KDD (Knowledge Discovery in Databases) is defined as the “non trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns of in data”[**17**]. In churn prediction the customer’s service usage data is very

useful for prediction. But telecom companies have bulk data and there is much need to select important features. Running a churn prediction model on the selected features makes prediction easy for the model and also saves time. Figure 1 shows this study using Feature Importance and Co- Relation Matrix, handling Categorical and continuous features for feature extraction. This study used multiple experiments some of them are hybrid methods and some are single techniques. Later in this study the experiments performance is being compared before and after features selection and compared with similar literature work. Every prediction model starts the process with Dataset acquisition and processing.

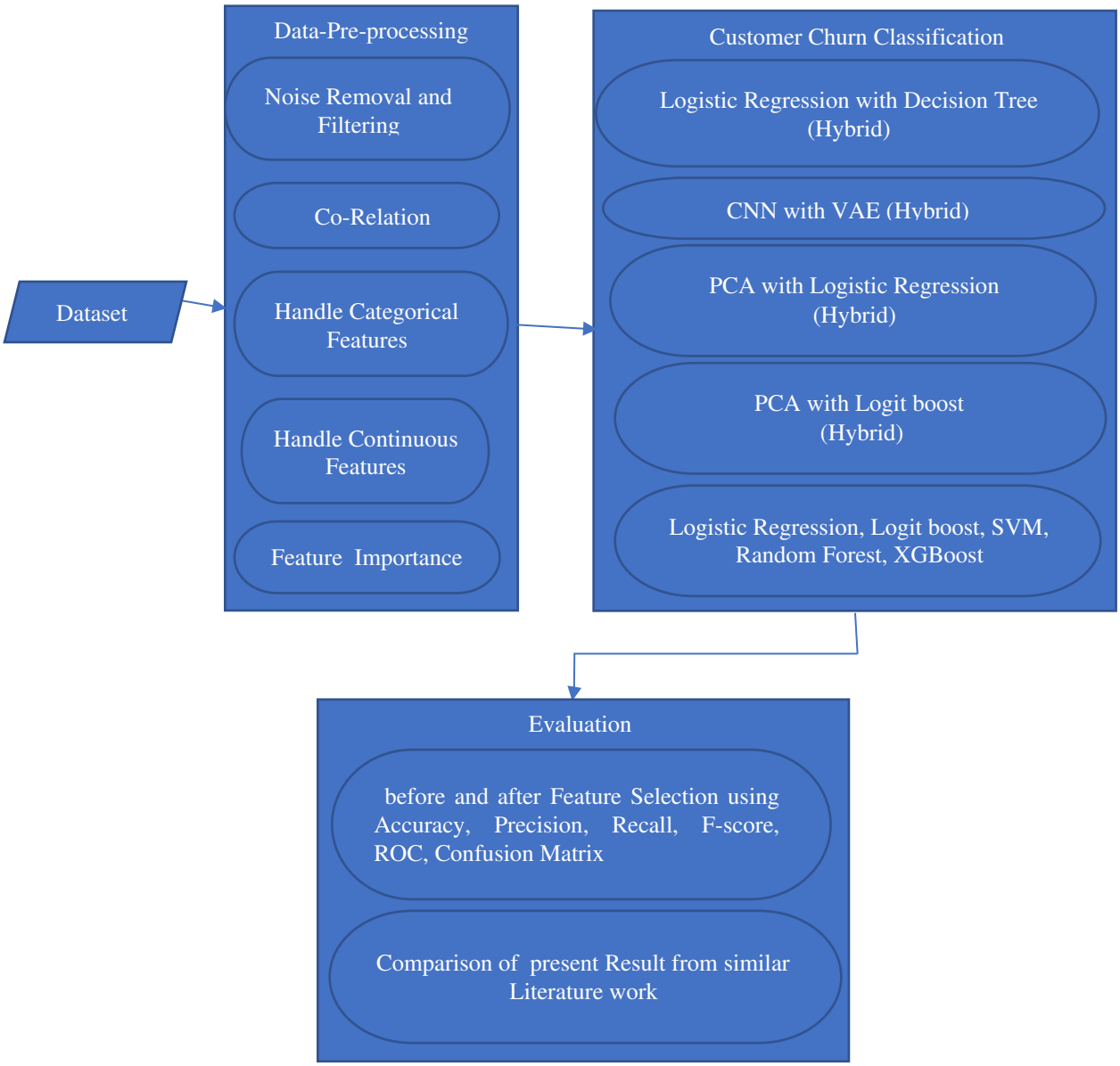


Figure 1: Proposed Model for Customer Churn Prediction.

- **Dataset**

This study used the Orange dataset which is publicly available on a data repository website. Dataset contains 3333 subscriber's entries and 21 attributes. Dataset has one target attribute "Churn" and 20 independent attributes and 483 churners entry. All the independent fields are predictor attributes that will be used to find target attributes. It is required to train the model with the predictors attributes which possess more information for the target attribute [7]. Table 2 depicts all the details about the dataset.

Table 2: Orange Dataset Description

Sr. No.	Variable	Description	Type
1	Account Length	The duration of the account with company	Relational, continuous Attribute
2	VMailMessages	Voice mail messages by customer during the period	Usage, Continues Variable
3	Day Mins	Minutes used by customer in day calls	Usage, Continues Variable
4	Eve Mins	Minutes used by customer in evening calls	Usage, Continues Variable

5	Night Mins	Minutes used by customer in night calls	Usage, Variable	Continues
6	Intl Mins	Minutes used by customer in international calls	Usage, Variable	Continues
7	CustServ Calls	Calls made by customer-to-customer service	Relational, Attribute	Continuous
8	Int'l Plan	Plan recharged by customer for international calls	Recharge, attribute	continuous
9	VMail Plan	Plan recharged by customer for voice mails messages	Recharge, attribute	continuous
10	Day Calls	Call made by customer in day during the period	Usage, Variable	Continues
11	Day Charge	Customer's day call charges	Usage, Variable	Continues
12	Eve Calls	Call made by customer in day during the period	Usage, Variable	Continues
13	Eve Charge	Customer's evening call charges	Usage, Variable	Continues
14	Night Calls	Call made by customer in day during the period	Usage, Variable	Continues
15	Night Charge	Customer's evening call charges	Usage, Variable	Continues
16	Intl Calls	International Call made by customer during the period	Usage, Variable	Continues
17	Intl Charge	Customer's international call charges	Usage, Variable	Continues
18	State	State customer belongs to	Demographic, Attribute	Categorical
19	Area Code	Area in the state customer belongs to	Demographic, Attribute	Categorical
20	Churn	Dependent variable, shows customer has churned or not	Binary attribute	
21	Phone	Customer's mobile number	Unique attribute	

Table 2: Orange Dataset Description

- **Noise Removal and Filtering:**

Noise removal is an important phase in a prediction model. Initially the dataset is filled with some missing values, outliers and null values that does not let machine learning to execute the process. In case of larger missing or unknown values of an attribute, that attribute is removed from the dataset. In this dataset there are no missing values therefore for missing values no steps are taken. Outlier detection method is applied on the dataset. Outliers are unusual values that are typically defined as being more than three standard deviations away from a variable's mean value [**1**]. The outlier values that are in the three slandered divisions are transformed in the accepted values. The next step taken for noise removal is handling NA values. All NA values are placed with a particular column mean. So that values having NA remain in the average range. Having more 0 values might degrade the value performance. There is one more step in data-pre-processing that is essential to take that removing Unique attributes. Orange dataset has one Unique attribute that is "Phone". They are not used in the training process because they have a direct correlation with the target output (specific to the customer itself [**3**].

- **Correlation**

Correlation Attributes Ranking Filter techniques is used for selecting a subset of relevant features [8]. Correlation is used to find out the variable or attributes which have co-related in the sense of dependency or association. It is a very good way to impute missing values by predicting one attribute from another. This study used correlation matrix and reduced the data dimensionality. Figure 2 depicting the correlation presents in the dataset.

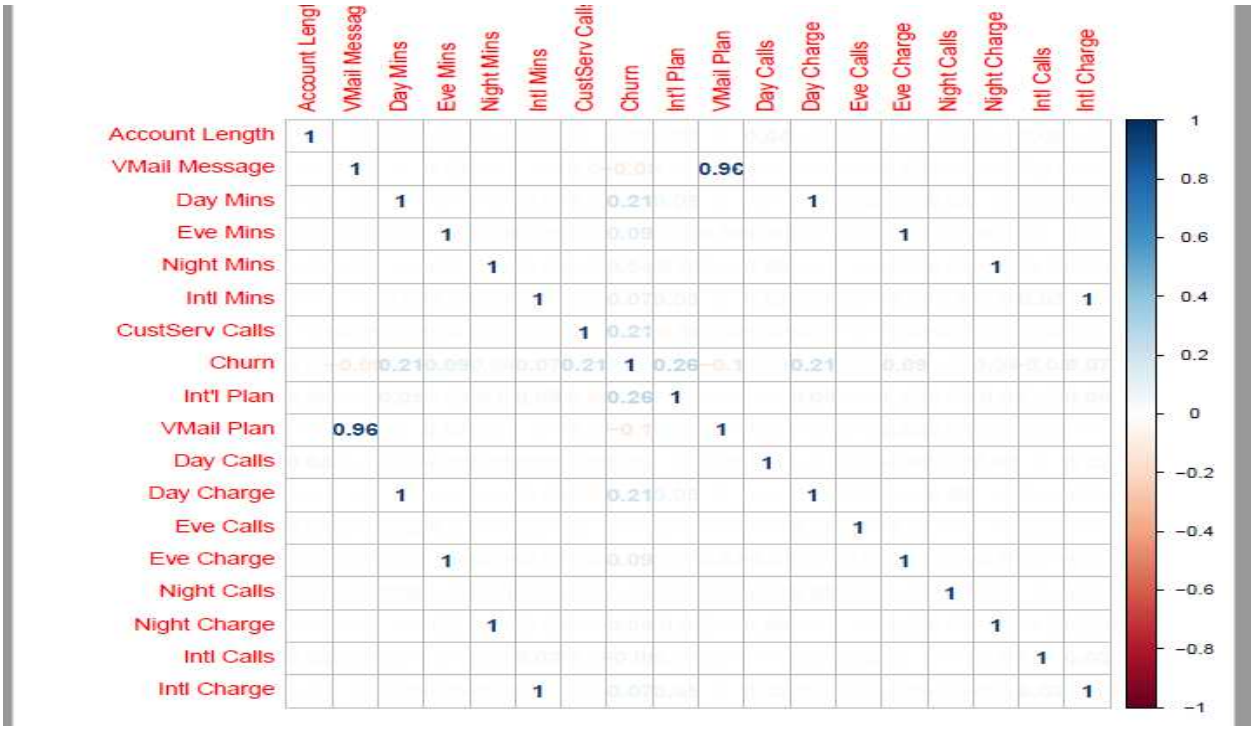


Figure 2: Correlation representation of the dataset Orange.

Figure 2 clearly shows the relationships between the features. "Vmail Plan", Day Charges", "Night Charges" and "International Charges" are co-related with 'Vmail Messages", "Day Mins", "Eve Mins", "Night Mins" and "Intl Mins" respectively. These attributes are dependent on each other therefore keeping all in the dataset is not worth it. In this study one attribute is removed from each pair based on the feature importance that will be discussed later in this study.

- **Handling Categorical and continuous Variable**

In telecom dataset there are a number of categorical features that exist, they may store useful information about customers therefore that features are essential for the churn prediction model. Some categorical features may have null importance for prediction. Features having null less importance can be removed from the dataset but removing categorical features having much importance will decrease model performance. Machine learning cannot handle these variables. Therefore, they need to be handled in such a way so that model performance may increase. Encoding a categorical variable is a good idea for handling categorical variables. Dataset has two categorical variables "State" and ``Area". "Area" feature is converted to 3 dummy features and assigned with 0 or 1. "state" feature converted into 52 dummy features and assigned with 0 or 1 values. After Encoding categorical features now, the dataset has 73 total features. categorical variables dummy variables are created based on the values of categorical variables. The number of dummy variables depends on the number of values a feature has. Later in this study feature importance is checked for all features that will be discussed later, based on feature importance, dummy features are removed and kept for the prediction.

as stated in Table 2 there are 17 continuous features and all features have different value ranges. Therefore, there is a big need to normalise all Continuous fields. Having values in different ranges makes problems for machine learning algorithms. Normalisation basically is a scaling technique which scales feature values in a range of 0 and 1. That works on a min-max scalar. It takes the maximum and minimum value of feature and scale value accordingly. Normalisation works in the following way:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

X_{max} *and* X_{min} are the maximum and minimum value of the feature respectively. If the value of X is minimum value of the feature then it will 0 in numerator and the value of X' will be 0 or else if the value is maximum in the feature column then the numerator will be equal to the denominator where the value of field will be 1 else the value will lie in range between 0 to 1.

After creating new dummy features and new normalised feature original features are removed from the dataset.

- **Feature Importance**

One very important key purpose of the churn prediction model is to find out the factor why customers are churning. Feature importance is a very good technique to visualize the importance of each feature in the dataset. Random forest provides a very good ranking method that is "feature importance". Feature Importance is calculated as the decrease in the node impurity weighted by the probability of reaching that node. The probability of a node is calculated by the number of samples that reach the node divided by total number of samples. The higher the value calculated the importance will be that high. It can be understood in following way:

$$nj_i = w_i c_i - w_{left(i)} c_{left(i)} - w_{right(i)} c_{right(i)} \quad (2)$$

Here nj_i is the node importance, w_i are weighted number of samples reached to node, c_i the impurity value of node where $right(i)$ and $left(i)$ are the right split child node and left split child node respectively. The feature importance for each feature is calculated by this way

$$fi_j = \frac{\sum \quad i:node\ i\ splits\ on\ feature\ j\ nj_i}{\sum \quad n \in all\ nodes\ nj_k} \quad (3)$$

Where fi_j is the importance of feature j and nj_i is the importance of node i. after calculating importance of all feature values are normalize in the range 0 to 1 in the following way

$$normfi_j = \frac{fi_j}{\sum \quad i \in all\ features fj_j} \quad (4)$$

At the end the final importance of a feature at the Random Forest level is calculated by calculating its average over all the trees. The sum of the features importance on each tree is divided by the total number of trees.

$$RFfi_j = \frac{\sum \quad i \in all\ trees\ normsfi_{ji}}{K} \quad (5)$$

Here $RFfi_j$ is final feature importance, $normsfi_{ji}$ is the normalized feature importance of feature f in tree I and K is the total number of trees. The Feature importance in this study is visualized below:

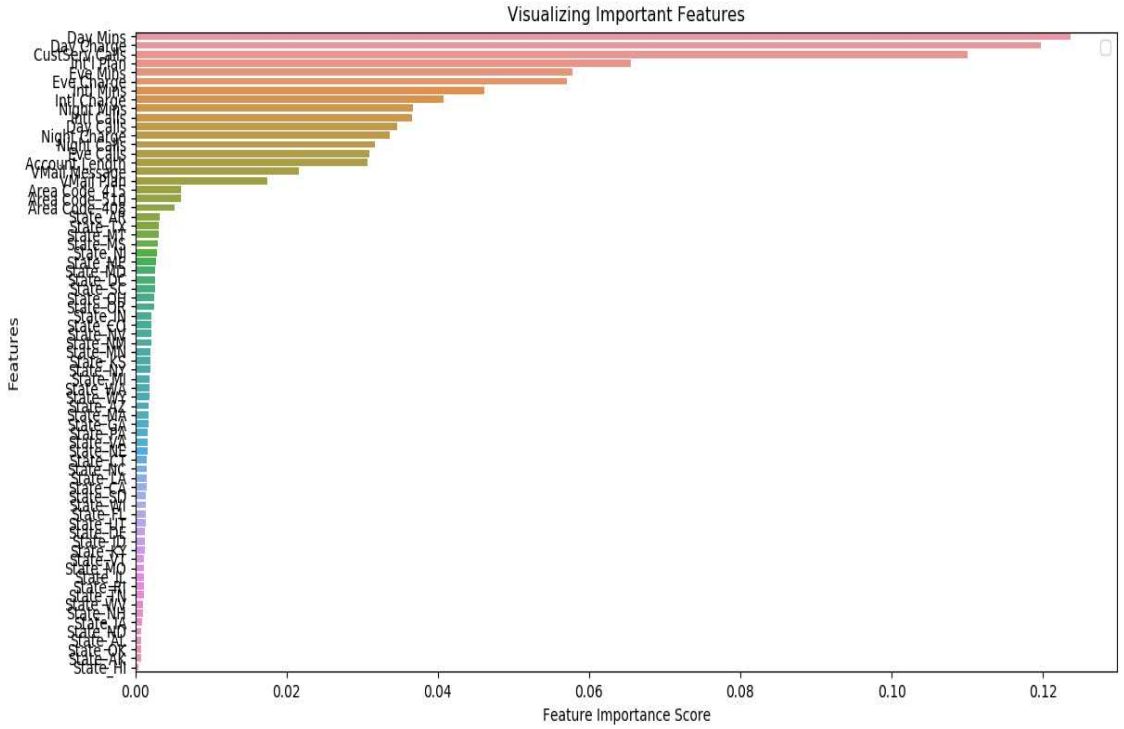


Figure. 3: Feature Importance Visualisation on Orange dataset.

Figure 3 shows that dummy features created after encoding "State" feature increased the dataset size to 52 new attributes and all dummy state features do not have much importance for the churn model. Therefore the "state" feature is removed from the dataset before category Encoding. Where feature "Area Code " gave three new features to the dataset with good importance therefore 'Area Code" is not removed. On the other hand, after correlation it was examined that features "Vmail Plan", Day Charges", "Night Charges" and "International Charges" are co-related with 'Vmail Messages", "Day Mins", "Eve Mins", "Night Mins" and "Intl Mins" respectively. Therefore, to effectively remove features from the dataset, features are removed based on their importance.

Table 3 : Data Pre-processing handled based on Feature Importance.

Sr. No.	Feature	Importance	Removed/Kept
1	Intl Mins	0.133903	Kept
2	Intl Charge	0.044322	Removed
3	Day Mins	0.133903	Removed
4	Day Charge	0.134758	Kept
5	Eve Mins	0.069142	Kept
6	Eve Charge	0.067976	Removed
7	Night Mins	0.038991	Kept
8	Night Charge	0.038902	Removed
9	VMail Plan	0.018475	Removed
10	VMail Message	0.025275	Kept

2. Experiments

- **Decision Tree with Logistic Regression (Hybrid LLM)**

Logistic Regression and Decision Tree both are very popular techniques and well known for good prediction and comprehensibility. Logistic Regression and Decision Tree both are most used Techniques in the literature. Having very good strength both techniques have some flaws as well. DT handles interaction effects between variables very well but has difficulties to handle linear relations between variables. For LR the opposite is true: it handles linear relations be- tween variables very well but it does not detect and accommodate interaction effects between variables [**1**]. In this study both techniques are combined and using the strength of both techniques. Decision tree first split the dataset into subsets based on their similarity. On each subset Logistic Regression is fit for classification.

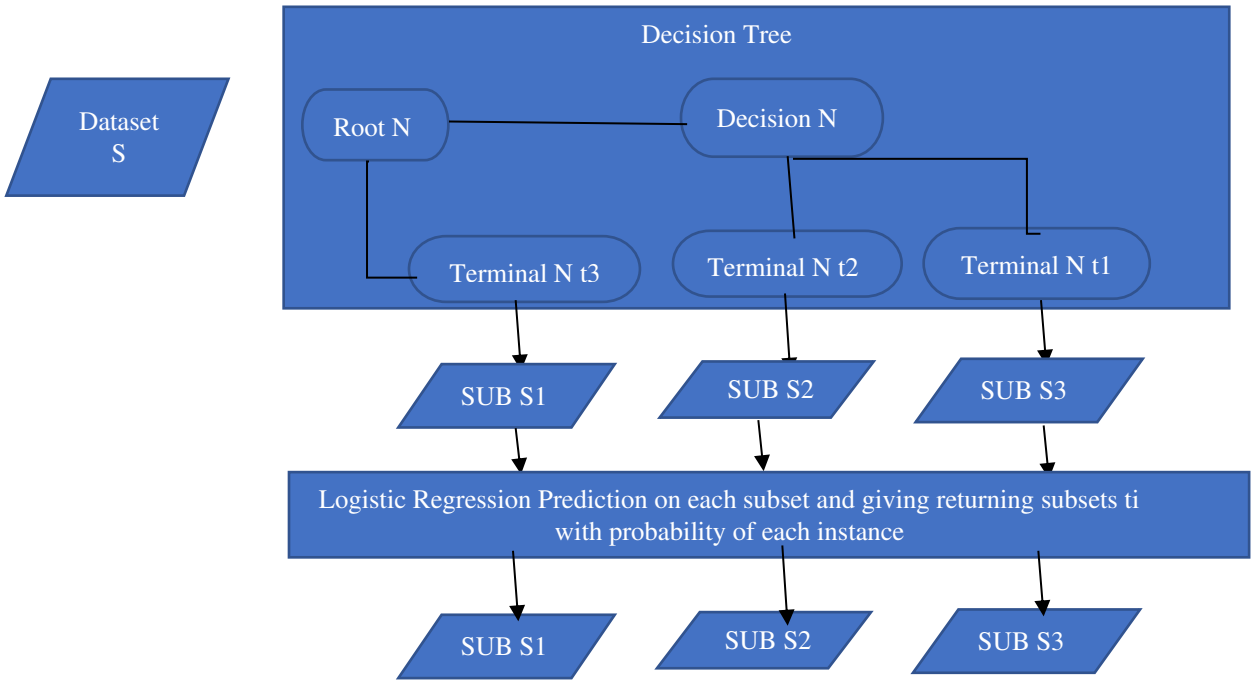


Figure 4: Structure of experiment one using logistic Regression with Decision tree.

The first experiment consists of two phases, the first phase using a decision tree and returning homogeneous customer segments. Second phase using Logistic regression for prediction on each customer segment. Decision Tree uses a process where the data is recursively split into smaller and purer subsets by repeatedly applying a greedy search through the space of possible decision trees branches and choosing optimal splits based upon a splitting criterion [**1**]. in a disjoint subdivision of *S* into customer subsets *S t* where every subset is represented by a leaf *t* in the tree:

$$S = \bigcup_{t \in T} S_t;\ \forall\, t \neq t' : S_t \cap S_{t'} = \emptyset \tag{6}$$

Decision tree using pruning to overcome overfitting. Overfitting occurs due to repeatedly splitting the tree and makes the model more complex. Logistic regression is a single classification technique mostly used for churn prediction. Logistic Regression proved to provide very good prediction results in the field of churn prediction. This experiment was done on R platform using Orange dataset. Dataset is divided into training and test sets in a 75:25 ratio.

- **Convolutional Neural Network with Variational Autoencoder (Hybrid)**

The second experiment was performed using deep learning techniques. This is also a hybrid model implemented using CNN(Convolutional Neural Network) and VAE (Variational Autoencoder). Convolutional Neural Network also called ConvNet is a Deep Learning Algorithm of a type Artificial Neural Network. The general idea behind Convolutional Neural Network (CNN) has been explained in four steps i.e. first convolution, second non-linearity, third pooling and finally classification [**7**]. This algorithm takes input (images/features) and assigns the weights or importance for prediction to the input attributes so that they can be distinguished from other attributes. Churn prediction related models contain one input layer that transfers all extracted features from the training set. Sigmoid function is used to calculate weight and that weight is assigned to the input features. This weighted sum is sent to the activation function in the hidden layers and output layers which generates output. It is important to increase the number of hidden layers to increase the performance. The mathematical understanding can be seen below.

$$f(x; w, a) = x_1 w_1 + x_2 w_2 + \cdots + x_n w_n + a \tag{7}$$

Here in the input layer, Linear function is used. Output of this linear function is dependent on the value of the weight w, a represent the bias factor or co-officiant and x is the input vector. Sigmoid function maps the input values from 0 to 1 that is more useful and is given by:

$$f(x; w, a) = g(w^T x + a) \ \text{, where} \tag{8}$$

$$g(v) = \frac{1}{1+exp^{-v}} \tag{9}$$

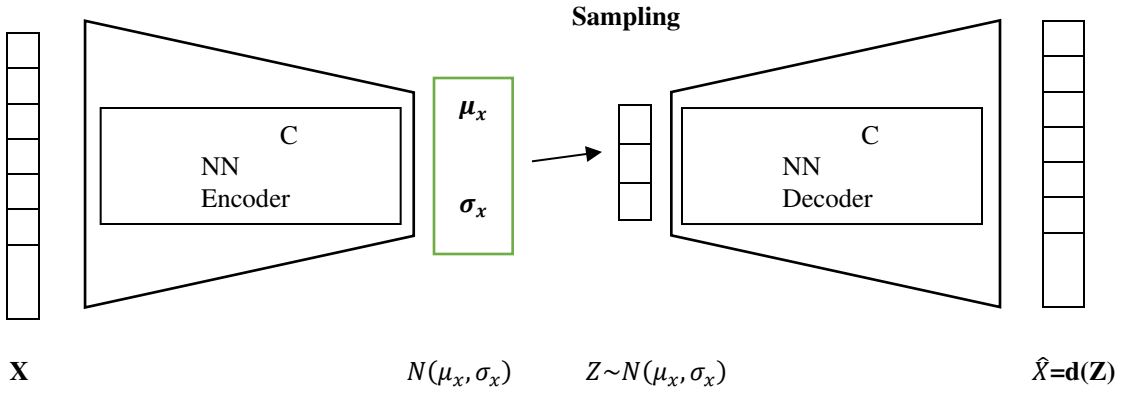
In the last few years **Deep Learning** based generative models have become the centre of interest. VAE allows us to create a complex generate model and also make them fit to large data. Variational autoencoders encoding distribution regularised at the time of training and it ensures that its **latent space**'s properties are good and allow us to generate new data. VAE automatically helps in **Dimensionality Reduction**. Dimensionality Reduction is the process of reducing the number of features that describes some data. In machine learning PCA is a very good technique for dimensionality reduction. in autoencoder, every data point is encoded as the real value that leads to no reconstruction loss in decoding it. In this case there is a high degree of freedom for autoencoders that ensures no reconstruction loss and low dimension latent space. However, there is a major issue in this process which is **overfitting**. Some data points come in the decoded data that are meaningless and to resolve this problem, model need to ensure the latent space to be regular enough. Regularising of the training process is done to avoid overfitting and making latent space properties more meaningful in Variational Autoencoder. VAE provides implementation in keras. VAE models are trained using loss function and compare the original data which is reconstructed. For optimisation purposes VAE is trained using a variational lower bound £ using stochastic gradient ascent method. The negative value of stochastic gradient descent is used for loss function L_{vae}. This loss function is calculated by summation of reconstruction loss L_{rec} and kullback-Leibler divergence loss L_{kl}.

$$L_{vae} = L_{rec} + L_{kl} \tag{10}$$

$$L_{rex} = E_{q\langle t|x\rangle}[log_{p\langle x|t\rangle}] \tag{11}$$

$$L_{kl} = KL_{(q\langle t|x\rangle \| p(t))\|} \tag{12}$$

Here x is the data to be reconstructed and t is a latent space vector**.** $p\langle x|t\rangle$ is the probabilistic decoder of VAE.



$$Loss = \|X - \hat{X}\|^2 + KL[N(\mu_x, \sigma_x), N(0,1)] = \|X - d(Z)\|^2 + KL[N(\mu_x, \sigma_x), N(0,1)]$$

Figure 5: Representation of Second Hybrid Model using CNN with VAE.

CNN-VAE implementation was done using keras libraries. This implementation includes phases:
○ Input layers were defined by giving the dimension of the dataset.
○ Encoded 2 dense layers were created. This dense layer is itself a **CNN** (convolutional Neural Network).
This convolutional neural network filters all the features by giving them weights and generating the output layer that have filtered feature set. This dense layer will change the dimension of the input vector. In this dense layer RelU activation function was used and in further hidden layers linear activation function was used. After this the model was encoded by converting inputs into latent variables. The output at any layer was calculated as mean i.e. centre point. From the hidden layers decoder dense layers were created that is again a **CNN (Convolutional Neural Network)** and decoded the output in the output layer.

- **PCA with Logistic Regression and Logitboost**

In the third experiment PCA was used for more dimensionality reduction on the orange dataset and for prediction Logistic Regression and logitBoost is used. PCA (Principal Component Analysis) mainly used for dimensionality reduction. It is used to transform large sizes of features into smaller one with containing more information with respect to churn prediction. Smaller dataset is easy to explore and visualise, it makes the machine learning process easy. In PCA dimensionality reduction is to trade a little accuracy for simplifying the process. PCA does the process in a few steps First Standardization, that standardizes the range of continuous features so that each feature contributes equally in the process. Mathematically this process works by subtracting the mean and dividing with standard deviation for each value of each variable.

$$x' = \frac{x-\mu}{\sigma} \tag{13}$$

Second Covariance Matrix Computation, is used to identify the relationship between variables so that variable having similar information and effect can be handled and redundancy can be removed. Third Identifying the Principal components, that is constructed by computing the Eigenvectors and Eigenvalues of covariance matrix. PCA give a new angle from there data can be evaluated and seen effectively.

The dataset given by PCA then given for the prediction and two powerful techniques are used logistic Regression and Logitboost. Logit Boost is an Additive Logistic Regression Model. The LogitBoost model is like the AdaBoost model. The main idea behind Logit Boost is to apply boosting in building logit model. The Logit Boost is classified as a “weak” or “base” learning algorithm, Logit Boost takes different training examples repeatedly due to that the base learning algorithm generates a new weak prediction rule, that causes so many rounds and later boosting algorithm must convert these weak rules into one strong prediction rule that, normally, become much more accurate than a weak rules. The difference between AdaBoost and logit boost is to use a weak classifier. Logit Boost is is Additive Logistic Model. An additive Logistic model forms the equation:

$$log\ log\ \frac{\langle x \rangle}{\{x\}} = \sum_{m=1}^{M} \quad f_m(x). \tag{14}$$

The monotonic logit transformation o the left side says that for any value of F(x)= $\sum_{m=1}^{M} \quad f_m(x). \in R$, the estimated probability will lie in the range 0 and 1. Inverting we get:

$$p(x) = P\{x\} = \frac{e^{F(x)}}{1+e^{F(x)}} \tag{15}$$

Here is the fitting of additive logistic regression by stagewise optimisation of log likelihood. Here probability y=1 by p(x) where

$$p(x) = \frac{e^{F(x)}}{e^{F(x)}+e^{-F(x)}} \tag{16}$$

Logistic regression is the one machine-learning algorithm that is not black box model. Normally black box models are complex but the logistic regression we understand what it does actually. Logistic regression can be binary, multinomial or ordinal. In our case, it is binary logistic regression. The logistic regression takes the real valued inputs and makes the prediction like input class belonging to the class 0. If the prediction is >0.5 then it takes the output as class 0 otherwise it takes output as class 1(here class 0 refers to non-churners and 1 refers to churners).

Logistic regression is achieved by taking the log odds of $\frac{P_i}{1-P_i}$ where P is the probability of being churn or not churn. P always will come in rage 0 to 1.

$$Z_i = ln\ ln\ \left(\frac{P_i}{1-P_i}\right) = \beta_0 + \beta_1 X_1 + .. + \beta_n X_n \tag{17}$$

Here **β** is the coefficients to be learnt and X_1 ... X_n are the independent variables. Here churn is the dependent variable and rest features are independent variables.

Taking the exponent of both side:

$$P_i = (\langle X_i \rangle) = \frac{e_z}{1+e_z} = \frac{e^{\alpha+\beta_i x_i}}{1+e^{\alpha+\beta_i x_i}} \tag{18}$$

In machine learning algorithms, we estimate the value of the coefficient by using stochastic gradient descent. It just calculates the value of prediction for each instance in the training set and calculates error for each prediction. In addition, this process continues until the model is accurate enough. In addition, the coefficient keeps updated in the process. For updating the coefficient, the following equation is used:

$$b = b - \alpha * (y - p) * p * (1 - p) * X \tag{19}$$

This experiment is performed on Weka. in this experiment orange dataset is evaluated twice, once in the earlier feature selection phase and second in PCA.

- **Experiments on Singles Techniques:**

All the experiments discussed above are hybrid experiments. In this experiment few standalone techniques were applied for prediction on Orange evaluated dataset. Those techniques are Logistic Regression, Logit boost, SVM, XGBoost, Random Forest. **Logistic Regression and Logit Boost** are already discussed above and used with PCA, now both techniques are being used alone. Both techniques individually performed well as well as with PCA. Next classification technique used is SVM. **SVM** is a powerful algorithm. SVM is based on mapping training data points into a higher dimensional space. This mapping is accomplished using a nonlinear function and then SVM performs linear regression in that space [**5**]. SVM does not minimise the training error but works on the generalisation error by minimising the upper bound. SVM plot datapoint in n dimensional space here n is taken for the number of feature dataset have. Then SVM separates data in two classes by defining a HyperPlane.

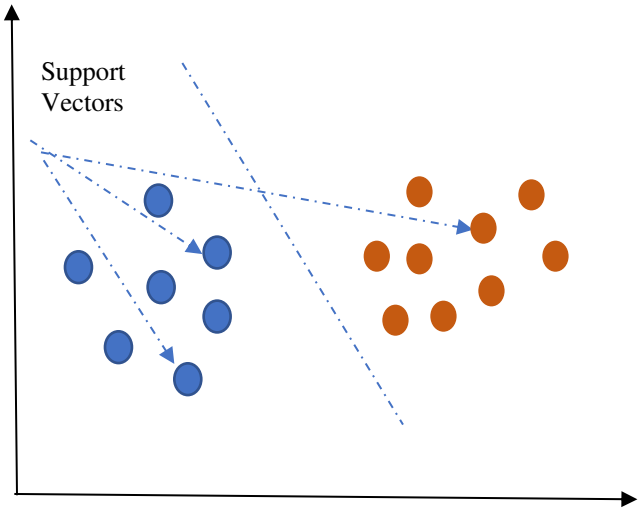


Figure 6: SVM Representation.

Support vectors are the coordinates of the observations. These data points help in building SVM. While building SVM, the margin between the coordinates and hyperplane tries to be maximised. The loss function is calculated by:

$$c(x, y, f(x)) = \{0,\ \ if\ y * f(x) \geq 1\ 1 - y * f(x),\ \ else\ \}$$
(20)

Loss is calculated 0 in case both actual and predicted are in the same range and is not the loss is calculated. For balancing margin maximisation and loss the regularisation parameter is added to the cost function. Cost function is given by:

$$min_w\ \lambda \|w\|^2 + \sum_{i=1}^{n} \quad (1 - y_i(x_i, w)) + \tag{21}$$

After loss function partial derivatives are taken to find gradients with respect to weights. Gradient is used to update weights.

$$\frac{\delta}{\delta w_k} \lambda\ \|w\|^2 = 2\lambda w_k \tag{22}$$

$$\frac{\delta}{\delta w_k} (1 - y_i(x_i, w))_+ = \{0,\ \ if\ y_i\ (x_i, w) \geq 1\ - y_i x_{ik},\ \ else\ \} \tag{23}$$

In order to train SVM, two main parameters are required: C and Sigma. The C parameter affects the prediction. It indicates the cost of penalty. Large value For C means high accuracy in training and low accuracy in testing. While small value for C indicates unsatisfactory accuracy [**14**]. in SVM Sigma values influence the hyper parameter partitioning more. The large value of sigma and small value of sigma leads to overfitting and underfitting respectively.

- **XGBoost**

Both AdaBoost (Adaptive Boost) and Stochastic Gradient Boosting algorithms are ensemble-based algorithms that are based on the idea of boosting. They try to convert a set of weak learners into a stronger learner [**14**]. The boosting methods are different from Random Forests and follow a constructive ensemble formation strategy. The idea behind boosting is to add new learning models in a continuous manner while building ensembles [**2**]. XGBoost is an optimised distributed gradient boosting library designed to be highly efficient, flexible and portable. XGBoost is the abbreviation for extreme Gradient Boosting. The primary purpose of using XGBoost is due to its execution speed [**16**]. Gradient descent helps in minimising the differentiable function but in gradient boosting the average gradient components will be computed. For each node in the tree there is a factor y with which each learner hm(x) is multiplied. This function adds the difference on the impact of splitting of each branch. Gradient boosting helps in predicting the optimal gradient for the additive model unlike classical gradient descent techniques which reduce error in the output at each iteration. The gradient boosting works in the following steps:

$$f0(x) = argmin_y \sum_{i=1}^{n} \quad l(y_i, \gamma) \tag{24}$$

The gradient is the loss function that is computed iteratively.

$$R_{im} = -\alpha \left[\frac{\partial(l(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{m-1}(x)}, where\ a\ is\ the\ learning\ rate \quad (25)$$

Each hm(x) is fit on the gradient obtained at each step. The factor ym a multiplicative factor for each terminal node is derived and then boosted model is given as:

$$f_m(x) = f_{m-1}(x) + y_m h_m(x) \quad (26)$$

XGBoost is also known as Regularised Boosting. It helps in reducing overfitting and perform parallel tasks that make it faster. The boosting methods are different from Random Forests and follow a constructive ensemble formation strategy [**2I**]. XGBoost adds new learning models continuously at the time of building ensembles. After every iteration a cumulative error is considered and based on that error a basic new weak learner is trained. This experiment is performed on python.

- **Random Forest**

Random Forest is a learning that is operated by multiple decision trees. The final decision is made based on the majority vote of the tree and chosen by Random Forest.

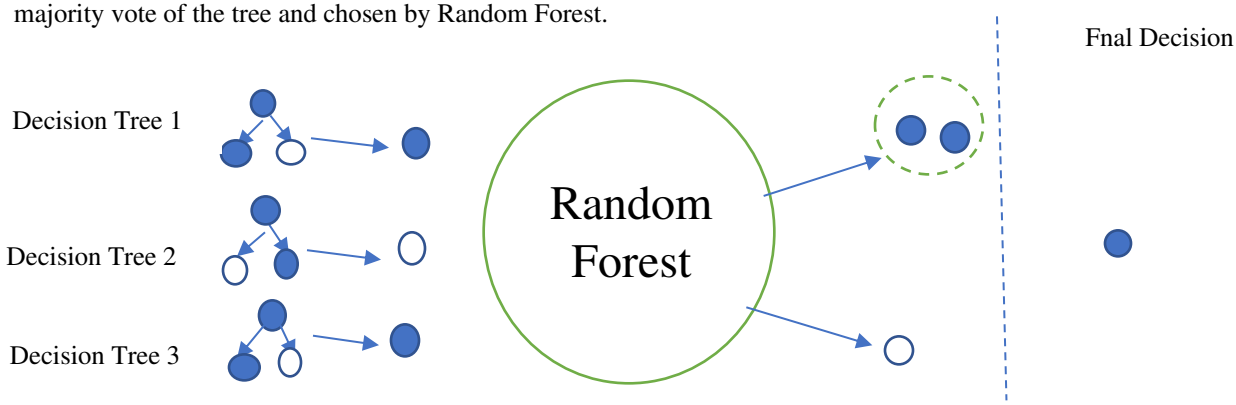


Figure. 7: Random Forest Structure

Random Forest belongs to the family of classifiers which populate the forest of Decision Tree. In Random Forest bootstrap aggregation or bagging is applied on the tree learning of training algorithm, where the training set is given by $X = x_1 \dots x_n$ and responses as $y = y_1 \dots y_n$. Bagging repeatedly calls random samples from and fit trees to that samples. Let's take random samples as R
$r = 1 \dots R$.

Samples for replace with training samples calls training samples X and Y as $X_r\ and\ Y_r$, then train the classification tree $f_r\ on\ X_r Y_r$. for prediction on unseen samples X' are made by averaging of the prediction of all individual trees on X' .

$$f' = \frac{1}{R}\sum_{r=1}^{R} \quad f_r X' \quad (27)$$

Prediction on X' unseen training samples also can be made by majority votes. Bootstrap procedure of random forest is good for better model performance. It does not increase the bias.

- **Parameters:** While fitting Random Forest some parameters are needed to be given. Tuning this parameter is very essential for improving model efficiency. These parameters are listed below:
    a) **Bootstrap:** Bootstrap tells the Random Forest about sampling data point methods. Sample will be sent with replacement or without replacement. In this study it is set to be False for using all samples.
    b) **n_estomators:** It is set to tell the Random Forest about the number of trees. The larger the number of trees leads good performance. But sometimes very large values come out with overfitting. In this study 100 default values were used only. Using a value greater than 100 is not affecting the performance and using a lesser value decreases the performance.

c) **max_depth:** This parameter tells Random forest about the max number of levels every decision tree. In this study it is set to 20.
d) **max_features:** This is used to tell the max number of features considered for splitting the node. It is set to 'auto'.
e) **minimum_sample_split:** This parameter tells the minimum number of datapoints to be placed in the node before splitting it. This is set to 2.
f) **minimum_sample_leaf;** This parameter gives the value of minimum data points that are allowed in leaf nodes. This is set to 1.

Random forests can be easily deployed in a distributive manner because of parallel execution while Gradient boosted trees cannot as it executes trial after trial[**2**].

IV. Result and Discussion

In this section the result visualisation is presented. The results are analysed to compare the performance of 8 experiments done on Orange dataset. all experiments are performed on the same dataset to analyse the result better. this study dealing with unbalanced dataset and comparing results in two scenarios. The first scenario was when the feature engineering and feature selection task is not performed using feature engineering. In the second scenario feature extraction task is performed using feature engineering In both parts, the performance is evaluated using the accuracy score of all models and it is compared with already existing similar models from the literature review. in the second part all the models are compared based on other performance measures. In the third part all models are compared using a confusion matrix.

- **Accuracy**

Accuracy indicates the ability to differentiate the credible and non-credible cases correctly [**14**]. It is the true positive and true negative portion from all the predicted instances. Figure 8 shows the accuracy comparison of all developed models before feature engineering (FE) and extraction tasks and after FE and extraction tasks and also compared with the accuracy achieved in literature on the same model. It can be seen that the maximum accuracy achieved by ensembled model Random Forest (RF) is 93 before FE boost, and 95 after FE in standalone technique where in literature it was 80%. Logistic Regression, Logit Boost, XGBoost got 85% accuracy before FE tasks After FE tasks these techniques got 86%, 89% and 88% accuracy respectively where in literature accuracy achieved was 79%, 87% and 78% respectively. SVM got 85% accuracy before FE and 89% after FE where in literature SVM got 86%. Standalone technique Random forest outperformed in terms of accuracy.

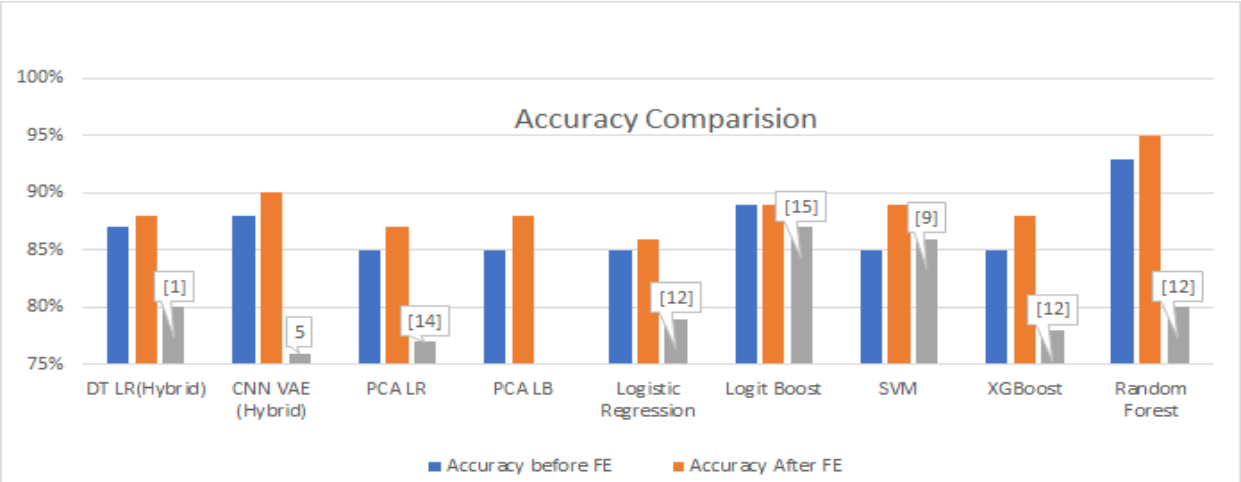


Figure 8: Accuracy Comparison if all models before and after FE with Literature work

In hybrid techniques maximum accuracy achieved by CNN with VAE (Convolutional Neural Network with Variational Autoencoder) is 88% before FE and it is 90% after FE. in literature the accuracy achieved on the same hybrid model is 76%. Where the Decision Tree with Logistic Regression (DTLR) got 87% Accuracy before FE and 88% After FE. the literature model of DTLR got 80% accuracy. PCALR and PCALB got 85% accuracy before FE and 87% and 88% after FE respectively. Feature engineering plays an important roll right prediction that can be seen by the results. Accuracy is a good performance measure but not enough to tell a model performance is good. so more

performance measures are essential to add. In the next part models are being compared according to some performance measures.

- **Other Performance Measures**

Table 4 shows all performance measures achieved by all models before and after FE. in hybrid models CNN with VAE out performed with .92 precision, .98 recall and .94 f-score. where after feature engineering CNN model got 88 precision, .99 recall and .93 f-score. **Precision** is how many are actual positive out of what we have predicted correctly. When a model achieves a low precision rate with high **recall** then it becomes difficult to measure the model performance and vice-versa. In this case Performance is calculated by **f-score.** F-score uses the Harmonic Mean at the place of Arithmetic Mean by punishing the extreme values more. CNN model got a .93 f-score that is a very good score. Other hybrid models also got very good scores of precision, recall and f-score but less than CNN models. where in single techniques Random Forest outperformed with .93 precision, 1 Recall and .96 f-score before feature engineering process. After feature engineering Random Forest got .95 precision, .99 recall and .97 f-score. Other models also performed well with good scores that are given in Table 4.

Table 4: Comparison based on other performance measures.

Model	Precision	PrecisionFI	Recall	RecallFI	F-score	F-scoreFI	Support	SupportFI	Macro Avg	Macro AvgFI	Weighted Avg	Weighted AvgFI
DT LR(Hybrid)	0.85	0.86	0.99	1	0.9	0.92	709	709	0.9	0.92	0.4	0.88
CNN VAE (Hybrid)	0.92	0.88	0.98	0.99	0.94	0.93	709	516	0.82	0.79	0.89	0.85
PCA LR (Hybrid)	0.87	0.85	0.97	0.99	0.92	0.9	0.92	709	0.71	0.9	0.83	0.4
PCA LB (Hybrid)	0.87	0.86	0.97	1	0.92	0.92	0.92	709	0.71	0.92	0.83	0.88
Logistic Regression	0.87	0.87	0.97	0.98	0.92	0.92	0.92	709	0.71	0.73	0.83	0.83
Logit Boost	0.91	0.91	0.97	0.97	0.93	0.93	709	709	0.81	0.81	0.88	0.81
SVM	0.85	0.91	1	0.97	0.92	0.93	709	709	0.43	0.81	0.72	0.81
XGBoost	0.85	0.86	1	1	0.92	0.93	709	709	0.43	0.93	0.72	0.88
Random Forest	0.93	0.95	1	0.99	0.96	0.97	709	709	0.96	0.95	0.94	0.95

The other performance measure added in this study is **Marco-Average**. Marco-Average performance measure is used when there is a need to check the overall performance on all classes. Marco average is calculated simply by taking an average of precision and recall achieved on all classes. F-score Marco average is the harmonic mean of both calculated Marco average of precision and recall. **Weighted Average** is also an important performance matrix for a machine learning model. Weighted average is also used to tell the overall performance of the model. It is also calculated for precision, Recall, f-score and support. To calculate the weighted average of all, get the precision, recall, f-score and support of each class and weight by the number of instances of each class. This study is working on two class classification problems so the weighted average will be calculated in this way.

$$weighted_p = (p_{c1} * |c1|) + (p_{c2} * |c2|)/|c1| + |c2| \quad (28)$$

where PC1 and PC2 are the precision of class one and class 2 respectively. and c1 and c2 are the number of instances of class 1and 2. According to the given formula, recall will be calculated in the same way. Here in this study, we are focusing on precision and recall that is why f-score $2*p*r/p+r$ is important to calculate. Weighting by class frequency gives better calculation of overall performance of the model. The **Figure 9** shows the macro-average and Weighted average matrices of all models.

Marco Average and Weighted Average Results for all models After FE

Logit Boost

	Precision	Recall	F-score
macro avg	0.81	0.7	0.74
weighted avg	0.88	0.89	0.88

CNN VAE

	Precision	Recall	F-score
macro avg	0.79	58	0.6
weighted avg	0.85	0.87	0.84

Logistic Regression

	Precision	Recall	F-score
macro avg	0.73	58	0.61
weighted avg	0.83	0.86	0.83

Random Forest

	Precision	Recall	F-score
macro avg	0.96	0.78	0.84
weighted avg	0.94	0.93	0.93

SVM

	Precision	Recall	F-score
macro avg	0.43	0.5	0.46
weighted avg	0.72	0.85	0.78

XGBoost

	Precision	Recall	F-score
macro avg	0.93	0.56	0.56
weighted avg	0.88	0.87	0.82

Figure 9: Marco Average ang Weighted Average result Comparison of all models.

Overall Random forest proved to be the best standalone technique for churn prediction model. Random Forest (RF) is a useful algorithm that suits classification and can handle nonlinear data very efficiently. RF produces better results and better accuracy and performance compared to the other techniques [**12**]. On the other hand, CNN with VAE proved to be the best Hybrid technique for churn prediction models.

- **Confusion Matrix**

Given the number of categories C, Confusion matrix represents the results of a machine learning model in CXC tabular format. that display the records count by their actual and predicted class. Confusion matrix is used to evaluate a classification model but not for a regression model.it categorise the outcome into two or more categories. Confusion matrix is used to calculate some performance measure like precision, recall, f-score, error rate etc. **Table** shows the confusion matrix achieved in different models after feature engineering.

Table 5: Confusion Matrix comparison.

Confusion Mtrix Comparison of all models After FE												
RF			LR			LB			PCALR		
	1	0		1	0		1	0		1	0
1	705	4	1	692	17	1	685	24	1	2793	57
0	38	87	0	101	24	0	73	52	0	435	48
CNN			SVM			XGBoost			PCALB		
	1	0		1	0		1	0		1	0
1	510	6	1	709	0	1	709	0	1	2800	50
0	70	14	0	125	0	0	111	14	0	444	39

V. Conclusion

This model is presenting a very good comparison model for Customer churn prediction in Telecommunications using a wide variety of machine learning and deep learning techniques. Additionally, this study set a very good example for feature engineering and feature extraction for a churn prediction model. On the other hand, this study showed a start art comparison of all the similar literature work that has been used in this study. Later in this study all literature works are compared with the similar models developed in this study using feature engineering based on Accuracy, precision, Recall, f-score, support, Marco-average, weighted average and confusion matrix. For feature engineering this study used correlation matrix, handled continuous features, handled Categorical features and used the feature importance function of random forest. These feature engineering tasks helped at best for improving the accuracy of churn prediction models. This study used two types of models: first hybrid models and second standalone techniques. all the techniques and models are compared later.

Standalone techniques used in this study are Logistic regression, Logit Boost, SVM, Random Forest, XGBoost. out of all techniques Random forest outperformed with 95% prediction accuracy where without Feature Engineering tasks Random forest got 93% accuracy. SVM and Logit Boost got 89% prediction accuracy After Feature Engineering that is second highest accuracy. On the other hand, Random Forest got the highest value as .93 precision, 1 Recall and .96 f-score before feature engineering process. After feature engineering Random Forest got .95 precision, .99 recall and .97 f-score. In this study Marco Average and Weighted average also explained and listed the achieved value. Other standalone techniques also performed well and got very good value of accuracy and other performance measures but Random Forest proved to be the best standalone technique.

In this study four hybrid models LLM (Decision Tree with Logistic Regression), CNN with VAE, PCA with Logistic Regression and PCA with Logit Boost are used for churn prediction. Out of all models CNN with VAE outperformed with 88% accuracy before Feature Engineering and 90% accuracy after Feature Engineering. where CNN with VAE got .92 precision, .98 recall and .94 f-score before Feature Engineering and 88 precision, .99 recall and .93 f-score after Feature Engineering. all hybrid models also performed well after feature engineering tasks. This Study we

compared predictive accuracy and comprehensibility of explicit, implicit, and hybrid machine learning models for telecom churn prediction on Orange Dataset. for machine learning models Weka, R and Python platforms are used.

In this paper several promising machine learning models have been identified which are suitable for learning knowledge and decision support. These models produced very good and understandable results. This study also used several feature engineering tasks like correlation matrix, feature normalisation, feature extraction, feature engineering, feature importance, handling categorical variables and continuous variables that also set an example of feature engineering and proved encouraging for future research. Random Forest and CNN with VAE have achieved good prediction results but all other models got similar results that need to be improved. Future research may include two or more big datasets.

Declaration of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Conflict of Interest Statement

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

[1] Arno De Caigny a , Kristof Coussement a , Koen W. De Bock b, "A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees", European Journal of Operational Research, 269 (2018) 760–772

[2] Pretam Jayaswal+, Bakshi Rohit Prasad*, Divya Tomar!, and Sonali Agarwal, "An Ensemble Approach for Efficient Churn Prediction in Telecom Industry", International Journal of Database Theory and Application, Vol.9, No.8 (2016), pp.211-232, http://dx.doi.org/10.14257/ijdta.2016.9.8.21.

[3] Abdelrahim Kasem Ahmad* , Assef Jafar and Kadan Aljoumaa, "Customer churn prediction in telecom using machine learning in big data platform", journal of BigData, https://doi.org/10.1186/s40537-019-0191-6.

[4] Nadeem Ahmad Naz, Umar Shoaib and M. Shahzad Sarfraz, "A Review on Customer Churn Prediction Data Mining Modeling Techniques", Indian Journal of Science and Technology, Vol 11(27), DOI: 10.17485/ijst/2018/v11i27/121478, July 2018.

[5] Ali Rodan and Hossam Faris, "Echo State Network with SVM-readout for Customer Churn Prediction", 2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), 2015.

[6] Preeti K. Dalvi, Siddhi K. Khandge, Ashish Deomore, Aditya Bankar, Prof. V. A. Kanade, "Analysis of Customer Churn Prediction in Telecom Industry using Decision Trees and Logistic Regression", 2016 Symposium on Colossal Data Analysis and Networking (CDAN), 2016.

[7] Abinash Mishra, U. Srinivasulu Reddy, "A Novel Approach for Churn Prediction Using Deep Learning", DOI: 10.1109/ICCIC.2017.8524551, Conference Paper · December 2017.

[8] Irfan Ullah1, Basit Raza 1, Ahmad Kamran Malik 1, Muhammad Imran1, Saif Ul Islam 2, And Sung Won Kim 3," A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector", IEEE Access, May 6, 2019.

[9] Adnan Amin, Sajid Anwar, Awais Adnan, Muhammad Nawaz, Khalid Alawfi, Amir Hussain, Kaizhu Huang,"Customer Churn Prediction in Telecommunication Sector using Rough Set Approach", http://dx.doi.org/10.1016/j.neucom.2016.12.009.

[10] Kristof Coussement, Stefan Lessmann, Geert Verstraeten, "A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry", doi: 10.1016/j.dss.2016.11.007.

[11] V. Umayaparvathi1, K. Iyakutti2, "Automated Feature Selection and Churn Prediction using Deep Learning Models", International Research Journal of Engineering and Technology (IRJET), Volume: 04 Issue: 03 | Mar -2017.

[12] Abbas Keramati a,n, SeyedM.S.Ardabili b,1, "Churn analysis for an Iranian mobile operator", Telecommunications Policy35(2011)344–356.

[13] A.Keramatia,∗, R. Jafari-Marandia, M. Aliannejadib, I. Ahmadianc, M. Mozaffaria,U. Abbasi," Improved churn prediction in telecommunication industry using datamining techniques", Applied Soft Computing 24 (2014) 994–1012.

[14] Sahar F. Sabbeh, "Machine-Learning Techniques for Customer Retention: A Comparative Study", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 9, No. 2, 2018.

[15] Wai-Ho Au, Keith C. C. Chan, and Xin Yao, "A Novel Evolutionary Data Mining Algorithm With Applications to Churn Prediction", IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, VOL. 7, NO. 6, DECEMBER 2003.

[16] V. Kavitha, S. V Mohan Kumar, G. Hemanth Kumar, M. Harish, "Churn Prediction of Customer in Telecom Industry using Machine Learning Algorithms", International Journal of Engineering Research & Technology (IJERT), Vol. 9 Issue 05, May-2020.

[17] DR. M.BALASUBRAMANIAN *, M.SELVARANI **, "CHURN PREDICTION IN MOBILE TELECOM SYSTEM USING DATA MINING TECHNIQUES", International Journal of Scientific and Research Publications, Volume 4, Issue 4, April 2014.

[18] Mr. Nand Kumar1, Mr. Chetankumar Naik2, "Comparative Analysis of Machine Learning Algorithms for their Effectiveness in Churn Prediction in the Telecom Industry", International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056, Volume: 04 Issue: 08 | Aug -2017.

[19] Praveen Asthana, “A comparison of machine learning techniques for customer churn prediction”, International Journal of Pure and Applied Mathematics Volume 119 No. 10 2018, 1149-1169.

Figures

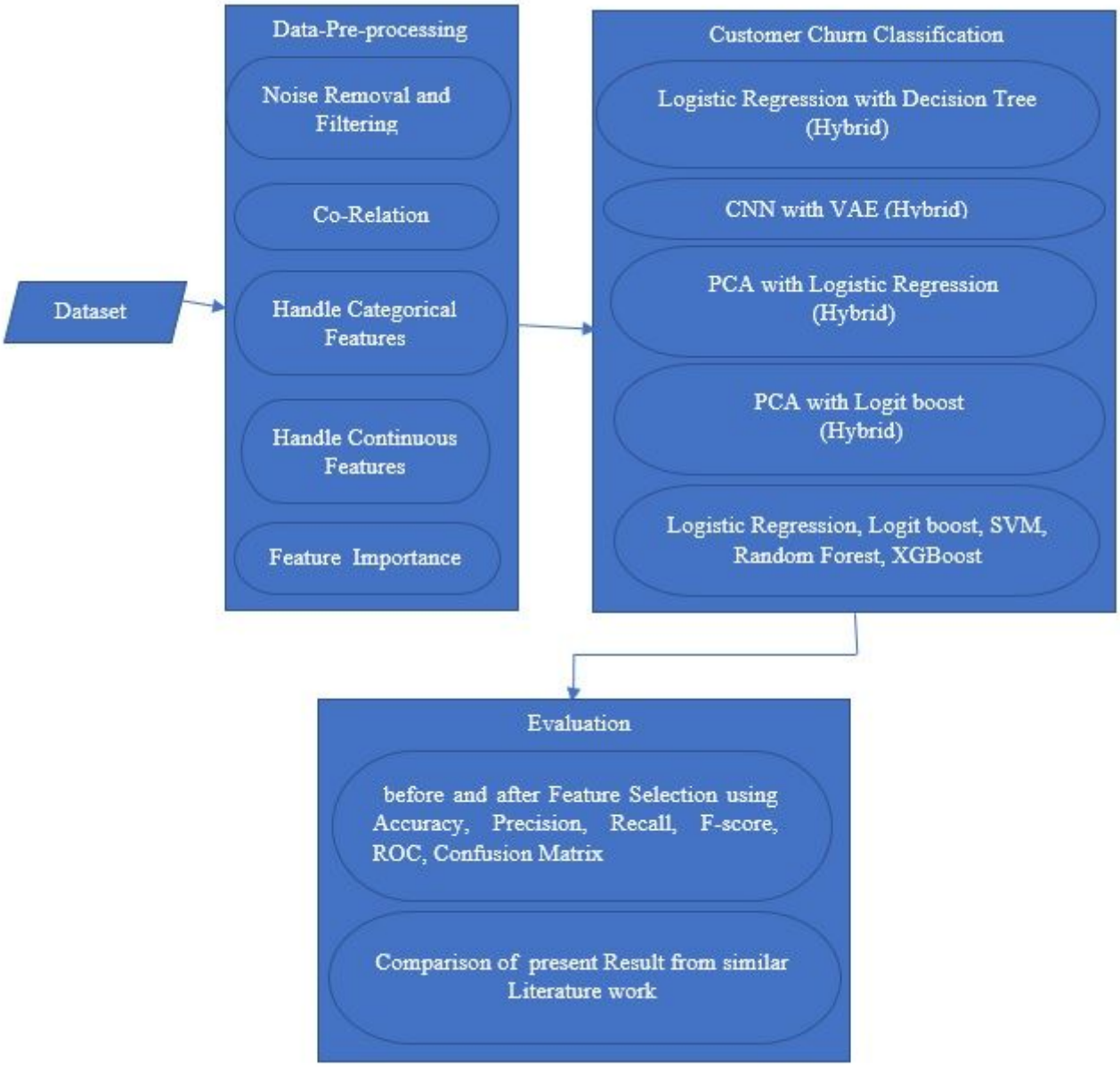


Figure 1

Proposed Model for Customer Churn Prediction.

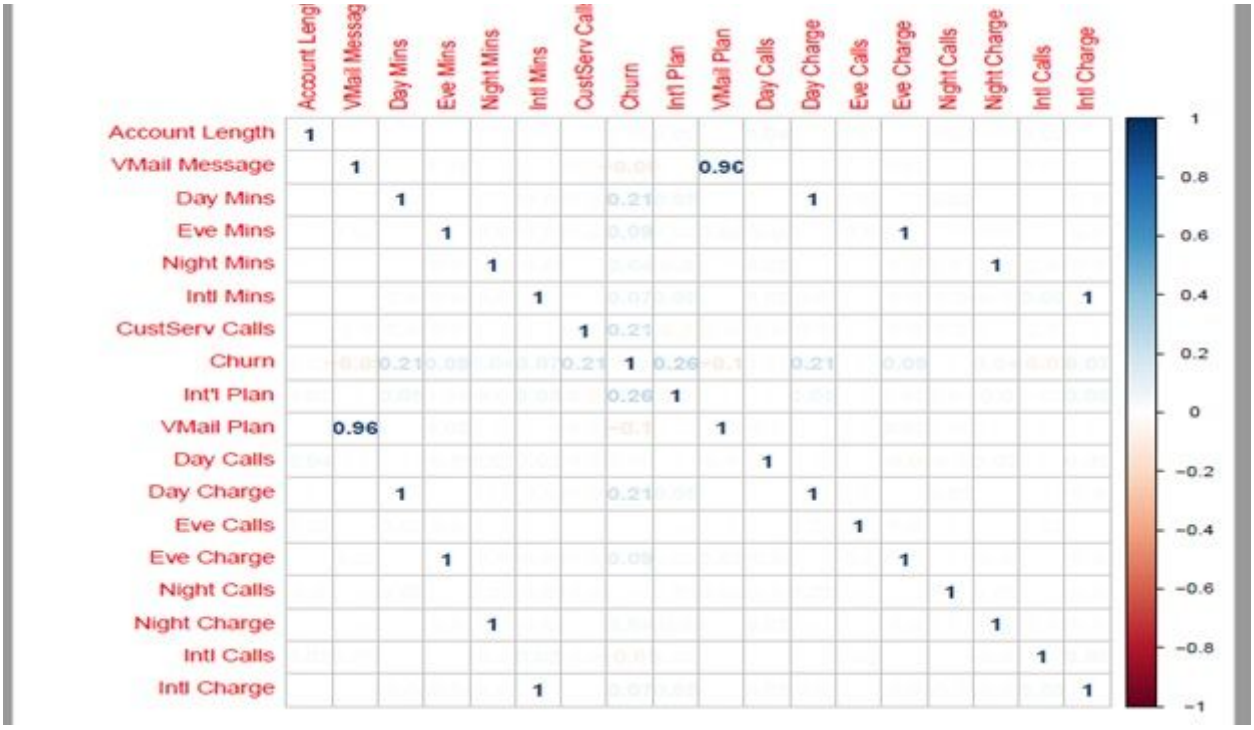


Figure 2

Correlation representation of the dataset Orange.

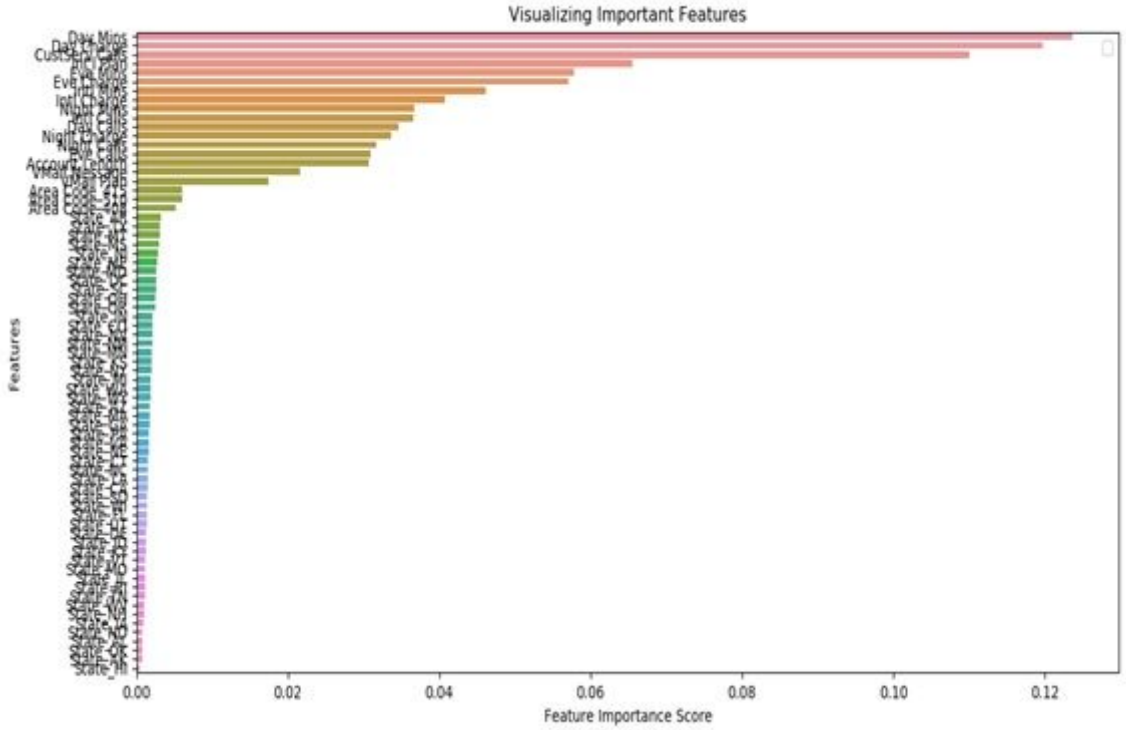


Figure 3

Feature Importance Visualisation on Orange dataset.

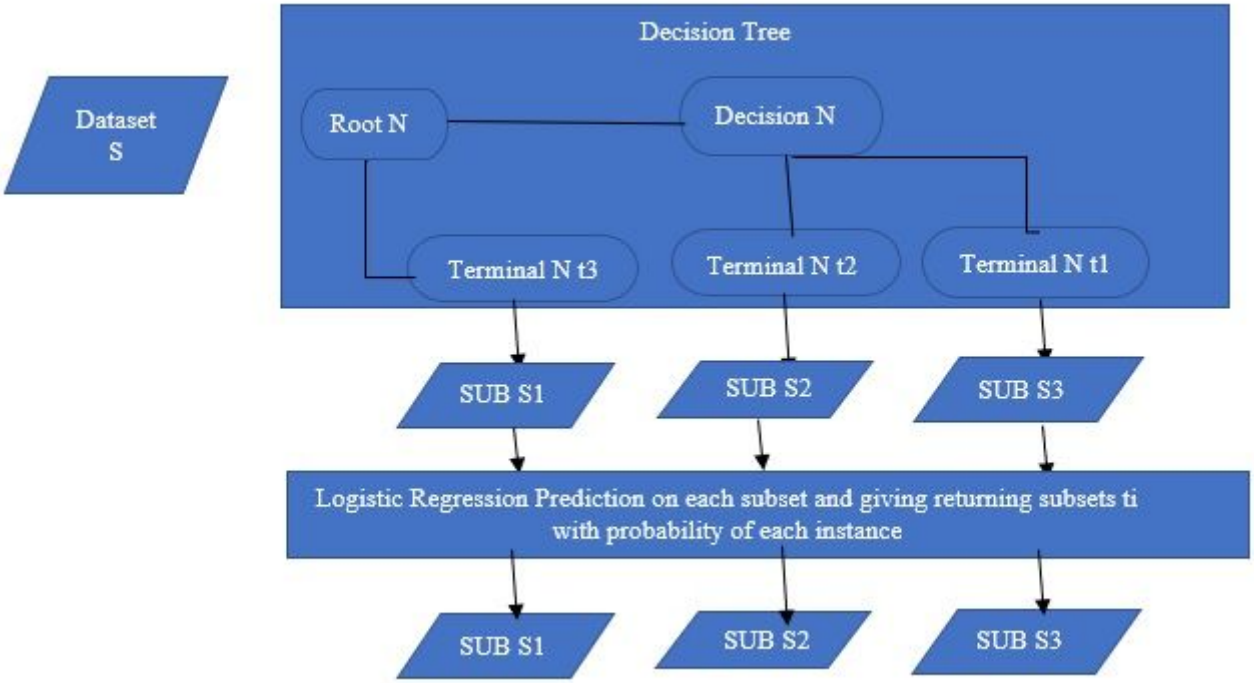


Figure 4

Structure of experiment one using logistic Regression with Decision tree.

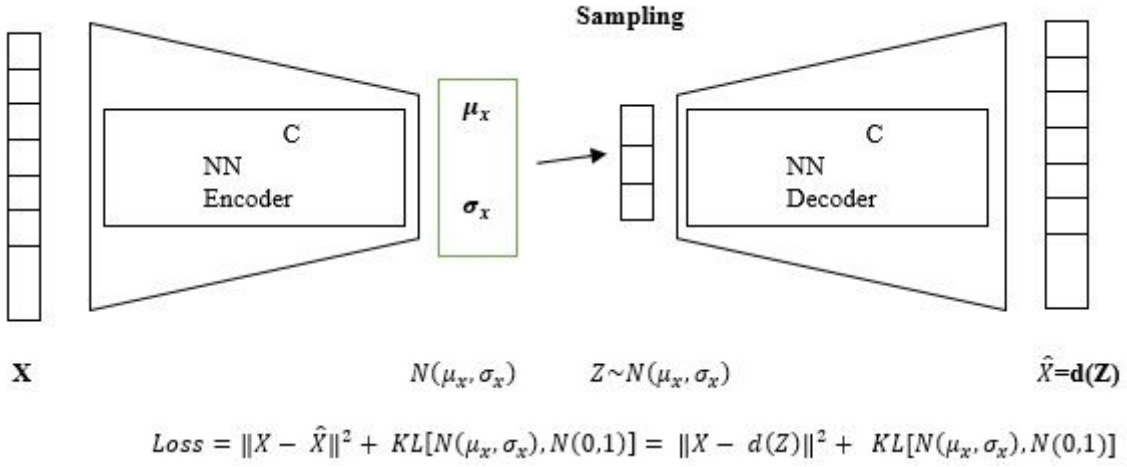


Figure 5

Representation of Second Hybrid Model using CNN with VAE.

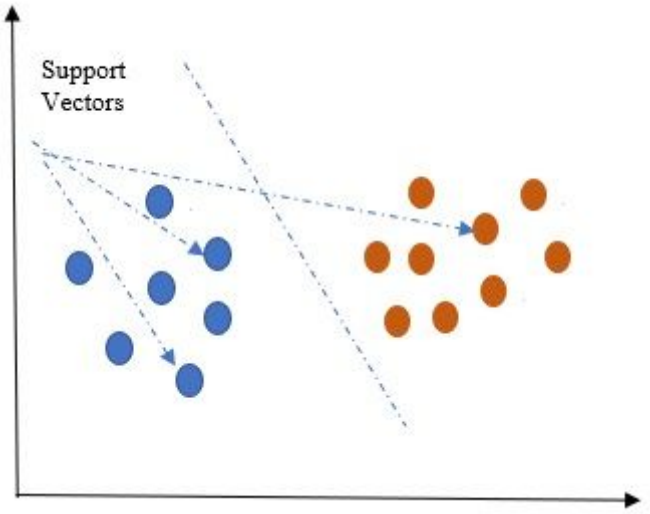


Figure 6

SVM Representation

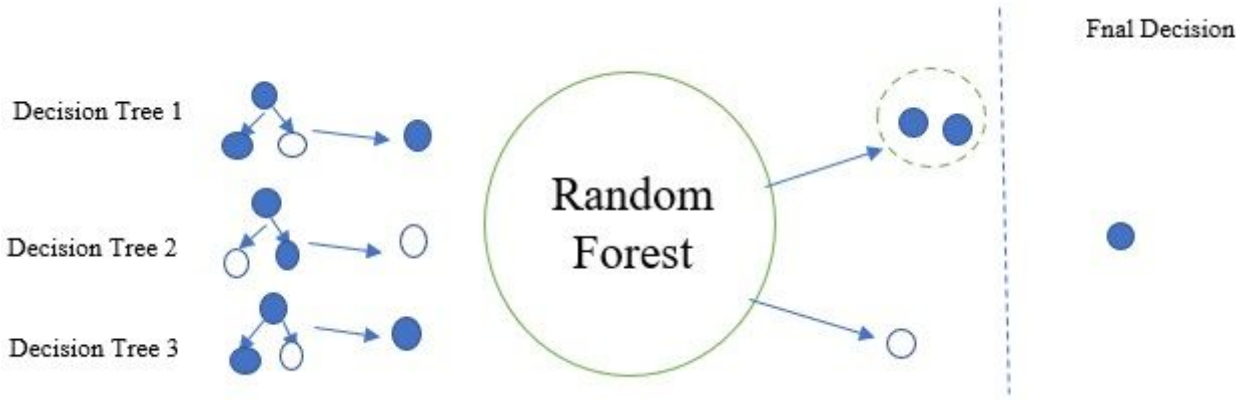


Figure 7

Random Forest Structure

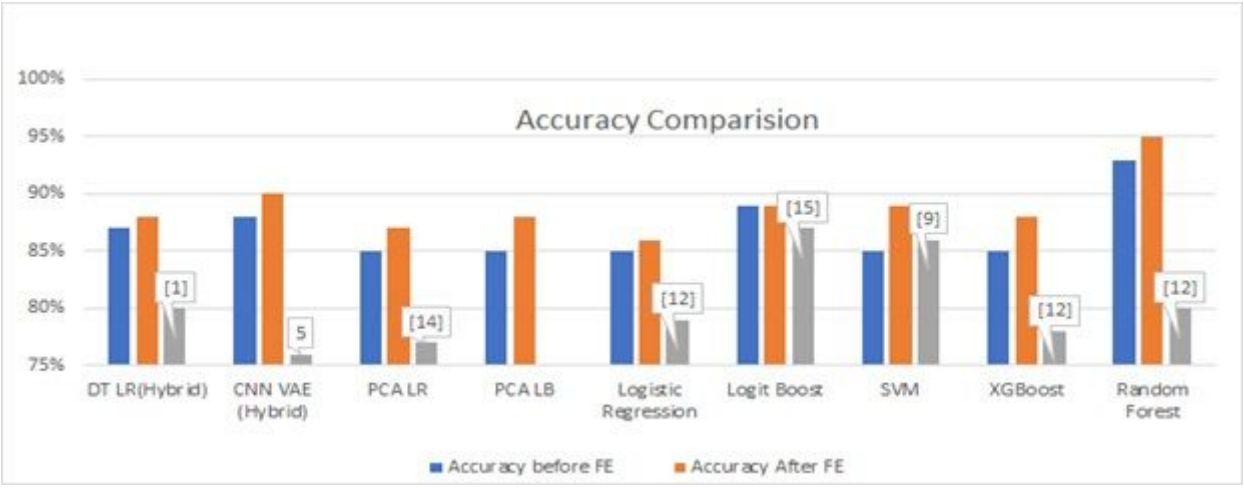
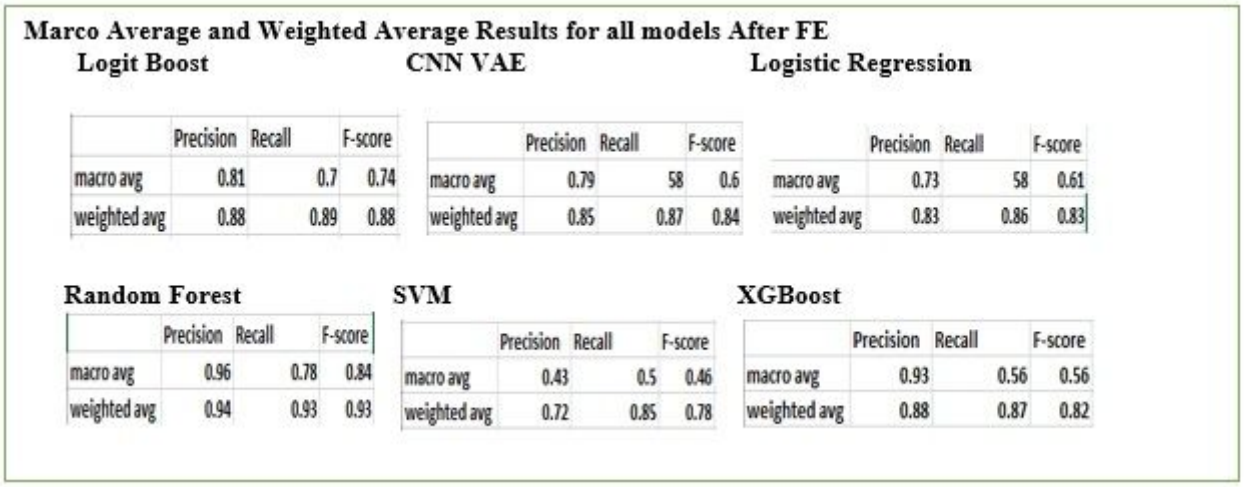


Figure 8

Accuracy Comparison if all models before and after FE with Literature work



Marco Average and Weighted Average Results for all models After FE

Logit Boost

	Precision	Recall	F-score
macro avg	0.81	0.7	0.74
weighted avg	0.88	0.89	0.88

CNN VAE

	Precision	Recall	F-score
macro avg	0.79	58	0.6
weighted avg	0.85	0.87	0.84

Logistic Regression

	Precision	Recall	F-score
macro avg	0.73	58	0.61
weighted avg	0.83	0.86	0.83

Random Forest

	Precision	Recall	F-score
macro avg	0.96	0.78	0.84
weighted avg	0.94	0.93	0.93

SVM

	Precision	Recall	F-score
macro avg	0.43	0.5	0.46
weighted avg	0.72	0.85	0.78

XGBoost

	Precision	Recall	F-score
macro avg	0.93	0.56	0.56
weighted avg	0.88	0.87	0.82

Figure 9

Marco Average ang Weighted Average result Comparison of all models.