Machine Learning Approaches to Predict Customer Churn

*Note: Model evaluation and selecting the best model

R. A. N. Sankalana
dept. of Computer Science and Engineering
University of Moratuwa
Sri Lanka
nirmalsankalana.19@cse.mrt.ac.lk

***Abstract*—This paper presents a performance comparison of 5 machine learning models for getting predictions about customer churn of a telco company called Chatterbox Pvt Ltd in Banana republic. Five prediction techniques(Support Vector Machine, Logistic Regression, K-Nearest Neighbours, Random Forest Classifier, XGBoost Classifier) are applied in the customer churn as predictors based on 18 predictor variables. The experimental results show that tree-based techniques like Random Forest Classifier and Boosting techniques like XGBooster are more effective than the existing ones for customer churn prediction.**

***Index Terms*—Customer Churn, Logistic Regression, Random Forest Classifier, K Nearest Neighbors (KNN), Support Vector Machine (SVM), XGBoost**

I. Introduction

The telecommunications sector has become one of the main industries in the world. The level of competition increased as a result of technical development and an increase in operators. Businesses are putting a lot of effort into surviving in this competitive market by utilising a variety of techniques. To enhance revenue, three major tactics have been suggested [1]: (1) bringing in new customers, (2) upselling to present customers, and (3) lengthening customer retention. The third strategy proves that keeping an existing customer costs much less than acquiring a new one and is also thought to be much easier than the upselling strategy[1]. To enforce the third strategy, businesses must decrease the risk of customer churn. Numerous studies have shown that machine learning technology is very effective at predicting this circumstance.

The data used in this research contains customer information including their account length, location, activated plans, charges, usage, number of service calls and whether they are charmed or not. We focused on evaluating and analyzing the performance of a set of machine learning methods and algorithms for predicting churn in Chatterbox Telco Pvt Ltd. We have experimented with several algorithms such as SVM, Random Forest, KNN, Logistic regression and XGBoost classifier to build the predictive model of customer Churn after developing our data preparation, feature engineering, and feature selection methods.

Column	Data Type	unique values	null values
account length	float64	204	2
location code	int64	3	0
intertiol plan	object	2	3
voice mail plan	object	2	6
number vm messages	float64	44	3
total day min	float64	1410	1
total day calls	float64	118	3
total day charge	float64	1145	5
total eve min	float64	1357	3
total eve calls	float64	117	4
total eve charge	float64	1206	8
total night minutes	float64	1245	2
total night calls	float64	65	6
total night charge	float64	802	5
total intl minutes	float64	153	2
total intl calls	float64	19	3
total intl charge	float64	150	5
customer service calls	float64	10	1
churn	object	2	5

TABLE I
SUMMERY OF THE TRAINING DATA SET

II. Feature Engineering

In feature engineering, raw observations are converted into features using statistical or machine learning approaches. To make machine learning work well on new tasks, it might be necessary to design and train better features.

A. Data Exploration

It is important to have a quick overview(TABLE 1) of the dataset before going to create new features. In this dataset, there are more numerical data than categorical data and most of them have missing values.

B. Missing Value Imputation

Missing values cause issues not only because the data can be non-representative of the actual population's information, but also because many machine learning algorithms do not work with missing data. Deleting the whole row and inputting the data point are the methods which can be used to treat missing values. In this scenario, the data set is a bit small (2321

rows) and Deleting rows may lose a certain percentage of data. Therefore imputation has been used to deal with missing data. Considering imputation methods there are various ways to impute missing values such as Mean Imputation, Mode imputation, Median Imputation, Regression Imputation etc. Using mean imputation or any other imputation that consists of filling the data with a fixed value is not very accurate because it does not take into consideration the correlation across features. For this data set, I used the Multivariate Imputation by Chained Equation(MICE) method for imputation purposes. MICE works by iterative regressing each feature, referring to missing values using the rest of the features, and repeating this process multiple times.

C. Outlier Imputation

Handling outliers is another important task in the data pre-possessing process because outliers increase the error variance and reduce the accuracy of the Machine Learning model. For outlier detection, I used Inter Quartile Range(IQR) method. Any value, which is beyond the range of -1.5 x IQR to 1.5 x IQR is treated as an outlier. As previously mentioned deleting outliers is not a good practice. Therefore For all numerical features, I used Median imputation because It is less affected by the screwiness of the data set

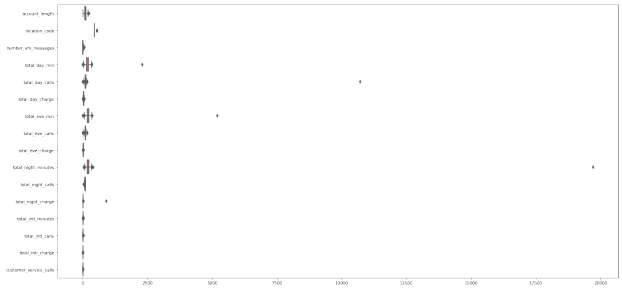

Fig. 1. Outliers of train data set before imputation

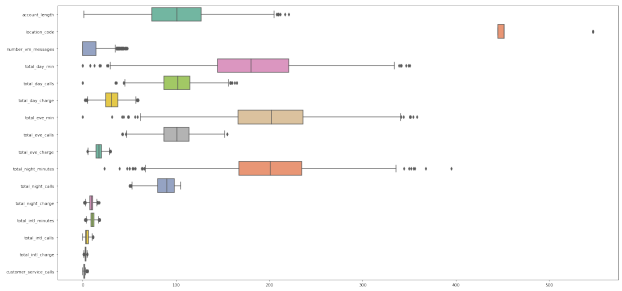

Fig. 2. Outliers of test data set after imputation

D. Data Encoding

Encoding categorical data is a process of converting categorical data into integer format. One-Hot encoding is used to encode categorical data in this data set. One-Hot Encoding is the process of creating dummy variables. This technique is used for categorical variables where order does not matter. One-Hot encoding technique is used when the features are nominal(do not have any order). In one hot encoding, for every categorical feature, a new variable is created. For example location code has three categorical values. After one hot encoding three distinct variables(location code 445, location code 452, location code 547)(Fig:3) are created.

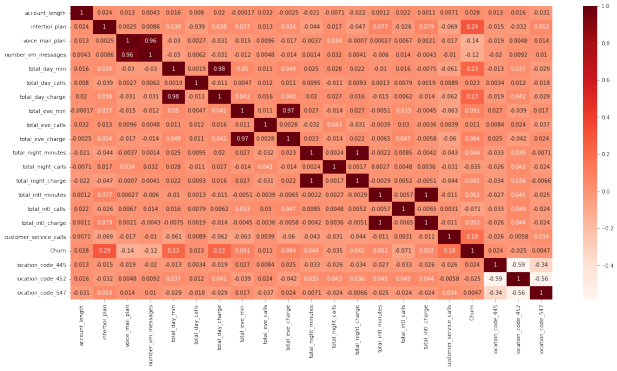

Fig. 3. Correlations matrix of train data set

E. Feature Scaling

Feature scaling is essential for machine learning algorithms that calculate distances between data. If not scale, the feature with a higher value range starts dominating when calculating distances.

F. Principle Component Analysis

Principle component analysis(PCA) is one of the linear dimensionality reduction techniques. It transforms a set of correlated variables (p) into a smaller k (k¡p) several uncorrelated variables called principal components while retaining as much of the variation in the original data set as possible. By reducing the dimensionality of the data, PCA will reduce the size of the data improving the performance of machine learning algorithms.

Fig. 4. Correlations matrix after principle component analysis

After calculating the correlation matrix of features in the churn data set, it is obvious that 10 features are highly correlated with each other(Fig: 3). PCA can be used to eliminate those highly correlated features to 5 uncorrelated features. The

data set has 20 features. When I plot the variance over the number of principal components(PCs), the number of principal components goes to 15, the variance reaches 1 and the change of variance reaching to 0(Fig: 5).

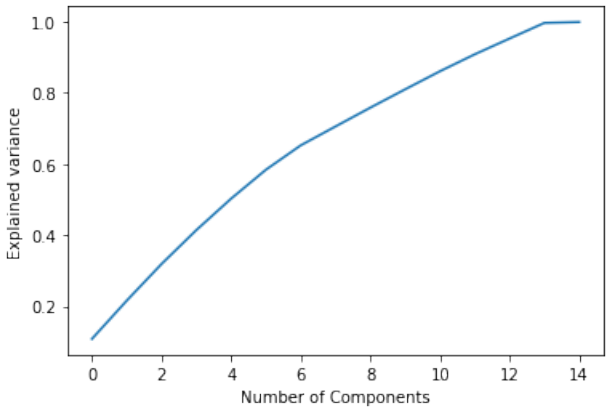

Fig. 5. Number of principle components over variance

III. Modelling

A. Logistic Regression

Logistic regression is a supervised ML algorithm used for binary classification problems. It follows the Bernoulli distribution to model a binary output variable. Logistic regression's range is bounded between 0 and 1 and does not require a linear relationship between inputs and target variables because of applying a nonlinear log transformation to the odds ratio.

B. Support Vector Machine(SVM)

Support Vector Machine(SVM) is a supervised ML algorithm which is used for both classification and regression problems. The objective of the SVM is to find a hyperplane in N-dimensional space(N is the number of features) that distinctly classifies the data points.

C. K-Nearest Neighbours(KNN)

The k-nearest neighbours (KNN) is a simple supervised machine learning algorithm that can be used for both classification and regression problems. It assumes the similarity between the new data and available data and put the new data into the category that is most similar to the available categories.

D. Random Forest Classifier

The Random forest classifier is a supervised, tree-based ML algorithm used for classification problems. It consists of a large number of individual decision trees that operate as an ensemble(Bagging). Each tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction. In Random Forest A large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models. Those uncorrelated models can produce ensemble predictions that are more accurate than any of the individual predictions.

E. Extreme Gradient Boosting Classifier(XGBoost)

XGBoost Classifier is a very powerful tree-based, supervised machine learning algorithm based on the Greedy Function Approximation, In Boosting models are built sequentially by minimizing the errors from previous models while increasing(boosting) the influence of high-performing models. Gradient boosting employs a gradient descent algorithm to minimize errors in sequential models. XGBoost is one of the fastest implementations of gradient boosted trees. It does this by tackling one of the major inefficiencies of gradient-boosted trees.

IV. Evaluation Critaria

Evaluation metrics are used to measure the accuracy of the model. Evaluation metrics can help to assess the model's performance, monitor the ML system in production, and control the model to fit business needs.

A. Confusion Matrix

A confusion matrix is a table that shows the number of correct and incorrect predictions made by the model compared with the actual classifications in the test set. This matrix describes the performance of a classification model on test data for which true values are known. It is an n*n matrix, where n is the number of target variables. This matrix can be generated after making predictions on the test data.

B. Accuracy

Accuracy is the proportion of true results among the total number of cases examined. It is a good evaluation method for classification problems which are well balanced and not skewed.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

C. Precision

The precision proportion of true positive over the sum of True positive and False positive.

$$Precision = \frac{TP}{TP + FP}$$

D. Recall

The recall is the proportion of actual Positives is correctly classified.

$$Recall = \frac{TP}{TP + FN}$$

E. F1 score

The F1 score is a number between 0 and 1 and is the harmonic mean of precision and recall. F1 score sort of maintains a balance between the precision and recall for your classifier. It is a good evaluation method to evaluate binary classification problems.

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN}$$

Model	Accuracy	F1	AUC
Logistic Regression	0.8039	0.4966	0.6763
KNN	0.8484	0.6159	0.7471
SVM	0.9150	0.8024	0.8767
Random Forest Classifier	0.9268	0.8024	0.8819
XGB Classifier	0.9329	0.8403	0.8882

TABLE II
ACCURACY OF EACH ALGORITHMS

F. AUC

AUC is the area under the ROC curve. AUC ROC indicates how well the probabilities from the positive classes are separated from the negative classes. AUC is scale-invariant. It measures how well predictions are ranked, rather than their absolute values.

V. MODEL PERFORMANCE COMPARISON

The three main metrics used to evaluate the model are accuracy, F1 score, and AUC. According to table Logistic regression has lowest accuracy and XGB classifier has highest accuracy.

In this context Logistic regression gives less accurate results because Logistic regression is trained by minimizing logistic loss (or maximizing likelihood) Therefore it will be guaranteed to minimize this loss, rather then something else. K-NN, is a lazy learner, has no training process, and in consequence, it does not try to optimize any effectiveness measure. SVM provides a decent accurate results for customer churn because it classifies data by finding the best hyperplane that separates all data points of one class from those of the other class. The best hyperplane for an SVM means the one with the largest margin between the two classes.

Random forest classifier and XGBoost classifier give better accuracy in customer churn prediction. Both of those algorithms are tree-based algorithms. In a random Forest, parallel decision trees are built independently using different features and ensembling them(bagging). In XGBoost models are built sequentially by minimizing the errors from previous models while increasing(boosting) the influence of high-performing models. Therefore XGBoost produces more accurate predictions than Random Forest.

CONCLUSION

In this context, I tried 5 different classification approaches to predict customer churn. According to the result in table 2 It is obvious that tree-based algorithms produce more accurate predictions in binary classification problems. For Customer churn analysis optimised tree-based algorithms(like random forest classifier and Extreme Gradient, Booster) perform more accurate. Extremely Gradient Boosting Algorithm is more accurate for customer churn prediction because it use both boosting to produce its model.

ACKNOWLEDGMENT

I thank every one who helped me for doing this project.

REFERENCES

[1] Wei CP, Chiu IT. Turning telecommunications call details to churn prediction: a data mining approach. Expert Syst Appl. 2002;23(2):103–12.