TEESSIDEUNIVERSITY

MIDDLESBOROUGH - TS1 3BA
SCHOOL OF COMPUTING, ENGINEERING AND DIGITAL TECHNOLOGIES
MACHINE LEARNING - CIS4035-N

Enhancing Customer Retention in Telecommunications through Predictive Churn Modelling

NAME: SAMSON IJOGHO
ID: C2441395
WORD COUNT: 2,282 (Excluding References)

Abstract

With the rapid development of telecommunications industries, service providers are more eager to expand their subscriber base. To meet the demands of survival in a competitive market, retaining existing clients has become a major concern. An industry survey in the telecom sector indicates that acquiring new customers is substantially more expensive than keeping current ones. Therefore, obtaining data from the telecom sector might help predict customer behaviour, such whether they would leave the company. To maintain their market share, the telecom industries must take the required actions to start gaining their affiliated clientele. value stagnant. Our article proposed a new framework for the churn prediction model, the efficiency and performance of Random Forest, SVM, KNN, Decision tree, and Logistic regression algorithms were tested and compared.

Keywords: Churn prediction, Machine Learning, machine learning, Logistic Regression, Decision Tree Classifier, Random Forest Classifier, Gaussian NB, SVM, KNeighbors Classifier, and XGBoost.

Introduction

The field of communication technologies is competitive today. The main problem that practically all the global telecommunication industries are currently facing is customer churn. In the context of telecommunications, churn refers to the behaviour of customers quitting a business and abandoning its services because they are unhappy with them or because they can get better deals from other network providers within their budget. This could result in a loss of income or profit for the business. Keeping customers has also grown to be a difficult task. Consequently, businesses are putting a lot of effort into launching cutting edge apps and technology to provide their clients with the best services available and keep them around. Being aware of the clients who are most likely to go from the business soon is essential as losing them would mean a large loss of revenue for the enterprise. Churn Prediction is the term for this procedure.

Related work

Several strategies have been used to anticipate customer churn, including as data mining, machine learning, and hybrid technologies. These strategies help firms find, anticipate, and keep churn consumers. They also support industries with CRM and decision-making.

(He, Y., He, Z., & Zhang, D, 2009) suggested a prediction model based on the Neural Network algorithm to handle the customer turnover problem in a large Chinese telecom firm with approximately 5.23 million users. The prediction accuracy standard was the overall accuracy rate, which was 91.1%.

(Idris A, Khan A, Lee YS, 2012) suggested a method for modelling the telecom churn issue using AdaBoost and genetic programming. They tested the model using two sets of common data. One by Orange Telecom and the other by cell2cell, with the latter having an accuracy of 63% and the former having an accuracy of 89%.

(Huang, F., Zhu, M., Yuan, K., & Deng, E. O., 2015) examined the big data platform's issue of customer attrition. The researchers wanted to show that, depending on the volume, diversity, and velocity of the data, big data significantly improves the process of churn prediction. The largest telecom business in China needs a big data platform to handle data from its Operation Support and Business Support departments and engineer the fractures. Utilising the Random Forest technique, AUC was used to evaluate it.

Methodology

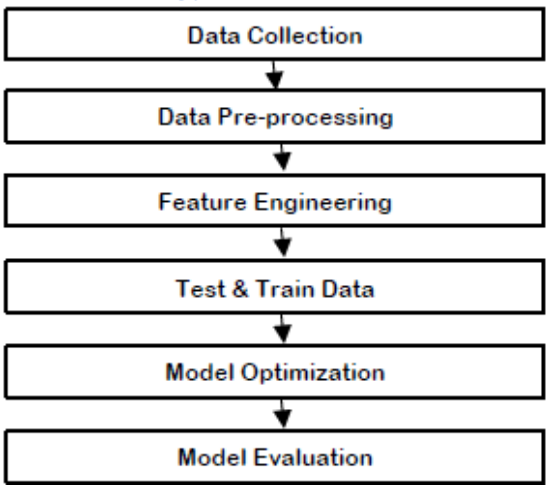

Figure 1. method overview

This study analyses the performance of various machine learning models (Logistic Regression, Gaussian Naïve Bayes, Random Forest Classifier, SVM, KNN, XGboost and Decision Tree) for loan default prediction using the approach indicated below.

Pre-processing and Data Cleaning

The dataset for this research was obtained from Kaggle. It originally had 7032 rows and 21 columns. The dataset has 22 duplicate values. Brief description of the data set by publishing the dataset's size and form, as well as presenting the top five and final five rows using the describe, tail, and head functions to allow for quick analysis of the dataset's structure.

Handling Duplicates and Missing Values: For our machine learning models to maintain our dataset and avoid mistakes, The dataset had 22 duplicate entries, which were eliminated, but no missing values.

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) helps us comprehend the features of data and identify trends and patterns. We imported relevant libraries such as NumPy, Pandas, and Seaborn, which allows us to do exploratory analysis such as value counts to obtain the frequency of categorical variables. Our EDA allows us to gain insight into the dataset's imbalance, as seen in fig 2 below. We may also conclude from the distribution of customer churn rate by monthly charges in fig 3. We also used charts like histograms and bar chats to visualise the distribution of our numerical and categorical data in order to understand their frequency and investigate the link between our target variables.

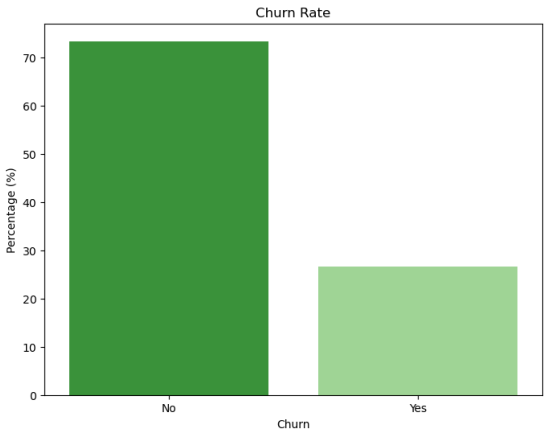

Figure 2. churn rate.

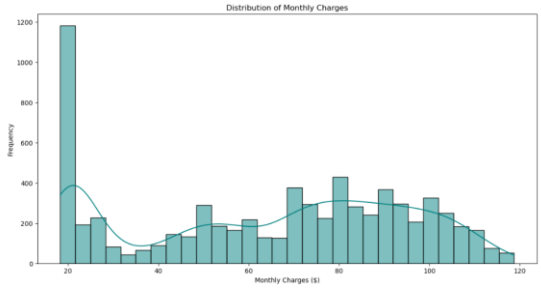

Figure 3. distribution of monthly charges.

As seen in fig. 4 below, we also use correlation matrices to determine the type of correlation that exists between our numerical variables. The relationships between variables are shown visually in the correlation matrix. Strong correlations are indicated by high positive (blue) or negative (red) colours. There is a strong positive association between tenure and total charges, indicating that they rise simultaneously.

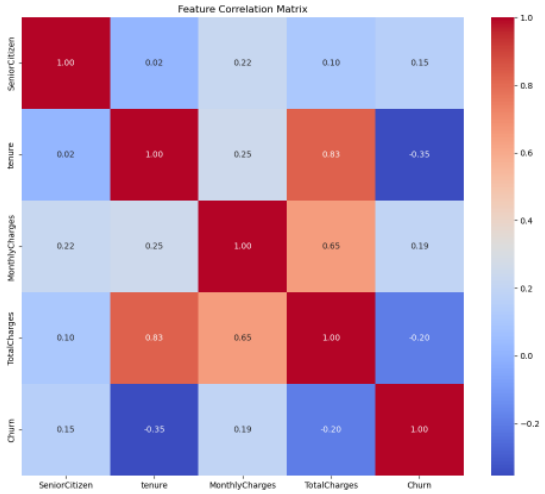

Figure 4. correlation matrix.

Feature Engineering

We employed a particular method in this research to make our data suitable for the deployment of machine learning models.

Encoding: Data in numerical form is used by machine learning. We used the label encoding approach in this project to convert our categorical variables into numerical features.

Splitting Data: We divided our data into training and test datasets in order to make sure our model performs well overall on fresh or unexplored data. With the test's size parameters set at 0.2, 80% of the dataset was used for training and 20% was used to evaluate our model. We set our random state to 42 so that every time the code is executed, the same random split is produced.

Feature selection: The most pertinent features in the dataset are found using this technique. To ascertain whether there are any significant variations between the means of two or more groups, we employed the ANOVA F-value test. The top 10 features with the best statistical relevance to our target variable were eliminated using this selection approach. Of the features that were chosen for the project, the credit type has the greatest F-value**.**

Normalization:

Using StandardScaler, important features such as tenure and numerous service-related factors were standardised. The normalisation improves algorithm performance by equal the scales of data inputs, which is important for our subsequent predictive modelling.

Model Implementation

For this research, we employed seven different machine learning algorithms Logistic Regression, Decision Tree Classifier, Random Forest Classifier, Gaussian NB, SVM, KNeighbors Classifier, and XGBoost. The algorithms used are all well-established and have been regularly utilised in binary classification for many years.

Random Forest: Random Forest is a machine learning technique that creates several decision trees and then combines them to get a more accurate and consistent forecast. It is well-known for its accuracy and robustness, and it is especially useful for classification and regression tasks, where it can deal with unbalanced datasets and avoid overfitting (Breiman, 2001). This was chosen because it provides safeguard for overfitting.

Logistic Regression: Logistic Regression is a statistical method for analysing a dataset in which one or more independent factors influence the outcome. The outcome is quantified using a dichotomous variable. It is widely used in medicine, social sciences, and machine learning, providing a simple and fast method for modelling the likelihood of a specific class or event occurring, such as pass/fail, win/lose, alive/dead, or healthy/sick. This is accomplished using a logistic function that simulates the probability of the default class (Peng, C.-Y. J., Lee, K. L., & Ingersoll, G. M., 2002). Logistic Regression was chosen for its interpretability and efficiency in binary classification tasks using our dataset.

Decision Tree: A Decision Tree is a flowchart-like structure in which each internal node represents a test on an attribute, each branch indicates the test's result, and each leaf node represents the class label. Decision trees are easy to grasp and analyse, and they are useful for a variety of predictive modelling applications (Quinlan, 1986).

The Decision Tree was chosen for our project because of its ease of comprehension, which enhances understanding of the decision-making process, as well as its capacity to handle both numerical and categorical data.

Gaussian Naive Bayes (NB): Gaussian Naive Bayes (NB) is a version of the Naive Bayes algorithm in which the likelihood of the characteristics is assumed to be Gaussian or regularly distributed. It computes the probability and conditional probability of each class for each input value. These probabilities are used to anticipate fresh data (Murphy, Kevin P., 2006). Gaussian Naive Bayes was used for this project because it is efficient and effective with continuous data. It works effectively in cases where the premise of feature independence holds true, making it appropriate for high-dimensional datasets.

Support-vector networks: SVM is a strong supervised machine learning technique used for

classification and regression tasks. It creates a hyperplane or set of hyperplanes in a high-dimensional space to distinguish between classes with as large a margin as possible. The ideal hyperplane is chosen to maximise the margin between nearest data points in any class, also known as support vectors (Cortes, C., Vapnik, V. Support-vector networks, 1995). SVM was chosen for this project due to its effectiveness in handling non-linear boundaries.

K-Nearest Neighbors (KNN): KNN is a simple instance-based learning method that may be used for classification as well as regression. In KNN, the output is selected by a majority vote of each point's nearest neighbours: a query point is allocated to the data class having the most representatives among its nearest neighbours (T. Cover, P.Hart, 1967). KNN was chosen for this project because of its simple implementation and high performance on tiny datasets.

XGBoost (Extreme Gradient Boosting): XGBoost is an improved version of gradient boosting algorithms. This method uses gradient descent to reduce errors in sequential models that forecast residuals from preceding models, gradually increasing prediction accuracy. XGBoost is optimised for efficiency, scalability, and performance (XGBoost: A Scalable Tree Boosting System, 2016). XGBoost was chosen for this project because of its ability to handle a wide range of data formats, its computational efficiency, and its success in enhancing model accuracy, particularly in classification challenges.

Model optimization:

Resampling Technique: During data preparation, we used the Synthetic Minority Over-sampling Technique (SMOTE) to resolve class imbalance in our target variable 'Churn'. This strategy artificially generates new examples of the minority class, resulting in a balanced dataset with an equal number of instances for both groups. After applying SMOTE, the dataset showed a uniform distribution, with 5153 occurrences in each class (churn and no churn).

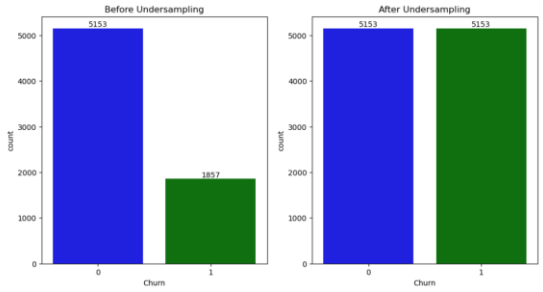

Figure 5: Target variable before and after under sampling.

To increase model performance, we use grid search to tune hyperparameters. This involves looking over a range of values for each hyperparameter and selecting the optimum combination based on the dataset. K-fold cross-validation improves performance by separating the training dataset into five folds, with one-fold for training and one-fold for testing, the method is performed five times each fold. This improves model performance estimates and minimises evaluation metric variance. After optimisation, the algorithm was reapplied, and the increased performance of the models was shown by bar charts, confusion matrix, and ROC curve.

In conclusion, the XGBoost and Random Forest models outperformed the others in terms of model performance, as demonstrated by the ROC curves, with AUC scores of 0.85 and 0.84. On the other hand, the Decision Tree and Naive Bayes models lag behind with AUC scores of 0.79 and 0.81, respectively, showing that their predictive performance can be improved.

Feature Importance

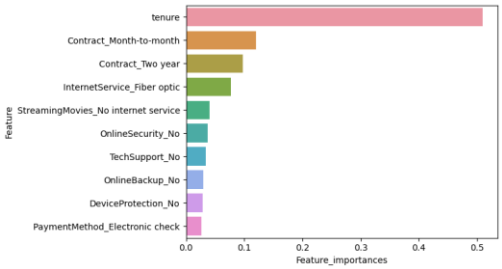

Figure 6. feature importance.

After concluding that the XGBoost and random forest classifier are our top performing models, we investigated the significance of the project's attributes and how much they contributed to model performance. We visualised these features in order of relevance to gain insight

into how much they contributed to the model's performance.

Result and Findings

To validate our assertion that XGBoost is the most performing model, followed by the random forest classifier, we used accuracy, recall, F1 score, and roc curve as evaluation metrics to measure the performance of the seven models used in this research. The ROC curves of the seven algorithms below show that XGBoost has the biggest area under the curve, indicating the best performing model.

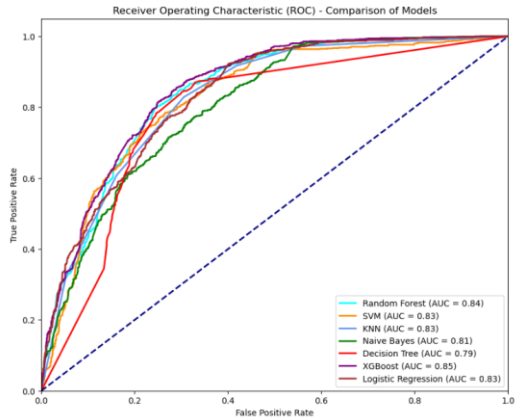

Figure 7. ROC of all the models.

Analysis shows that Random Forest (RF) has the highest test accuracy at just over 79%. XGBoost (Xgb) isn't far behind (78%), with the best precision and F1 scores, indicating a balanced and exact model. On the other hand, Gaussian Naive Bayes (GNB) has the lowest F1 score, indicating possible areas for improvement. Decision Trees (DT) provide respectable overall performance.

models	Test Accuracy	Trained Accuracy	Precision	Recall	F1 Score
Xgb	0.7696	0.7718	79	78	78
RF	0.7905	0.7774	78	77	77
LG	0.7435	0.7506	75	74	74
KNN	0.7614	0.7572	77	76	76
SVN	0.7599	0.7629	76	75	75
DT	0.77158	0.7581	77	77	77
GNB	0.7182	0.7165	73	73	71

Figure 8. Evaluation matrix of the seven models.

Conclusion and Discussion

The model developed for this project yields a good outcome; however, there is still space for development in subsequent studies, such as obtaining additional data with more features integrated and investigating additional features that would be pertinent to the task. Given that we employed Logistic Regression, Decision Tree Classifier, Random Forest Classifier, Gaussian NB, SVM, K Nearest Neighbours Classifier, and XGBoost, it might be worthwhile to attempt different algorithms, such as deep neural networks or Linear Regression, for this project Future research that will be worthwhile to do will include experimenting with various ensemble learning techniques to raise the performance of our underperforming model.

In summary, the goal of this research project is to evaluate the effectiveness of the chosen algorithm for predicting loan default. It was determined that banks and other lending institutions can rely on machine learning models to help them control risk and make wise decisions.

References

Breiman, L. (2001). Ramdon Forest. *Machine Learning*, 5-32.

Cortes, C., Vapnik, V. Support-vector networks. (1995). Support-vector networks. *Mach Learn 20*, 273-297.

He, Y., He, Z., & Zhang, D. (2009). A study on prediction of customer churn in fixed communication network based on data mining. *In Sixth International Conference on Fuzzy Systems and Knowledge Discovery (Vol.1)*, (pp. 92-94).

Huang, F., Zhu, M., Yuan, K., & Deng, E. O. (2015). Telco churn prediction with big data. *ACM SIGMOD International Conference on Management of Data*, (pp. 607-618).

Idris A, Khan A, Lee YS. (2012). Genetic programming and adaboosting based churn prediction for telecom. *IEEE international conference on systems, man, and cybernetics*, (pp. 1328–32).

Murphy, Kevin P. (2006). Naive Bayes Classifirs. *University of British Columbia*, 1-8.

Peng, C.-Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *The Journal of Educational Research*, 3-14.

Quinlan, J. (1986). Induction of decision trees. *Mach Learn 1*, 81-106.

T. Cover, P.Hart. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 21-27.

XGBoost: A Scalable Tree Boosting System. (2016). *International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). Association for Computing Machinery.