Customer Churn Prediction

A report for M.tech project

Submitted by

22MA60R

under the supervision of

Dr. Buddhananda Banerjee

in partial fulfillment for the award of the degree

of

MASTER OF TECHNOLOGY

in

COMPUTER SCIENCE AND DATA PROCESSING

at

DEPARTMENT OF MATHEMATICS
INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR
KHARAGPUR - 721302
WEST BENGAL, INDIA

Contents

Customer Churn Prediction

April 8, 2023

1 Abstract

This project aimed to develop a model for classifying customer churn using four different machine learning algorithms: random forest, SVM, logistic regression, and KNN. However, the initial accuracy of these models was not satisfactory due to the imbalanced dataset. Therefore, a resampling technique called SMOTE-ENN was used to address this issue. The results showed that the model's accuracy improved significantly after applying this technique. The SMOTE-ENN resampling technique effectively balanced the dataset and reduced the impact of the class imbalance, resulting in better performance. The proposed model can be used by companies to predict customer churn and take proactive measures to retain their customers. This project highlights the importance of using appropriate techniques to address imbalanced data and their impact on model performance.

2 Introduction

The customer who ceases a product or service for a given period is referred to as a churner. In a telecommunication company, the individual who has opted for service from a firm is referred to as Churn. The individual who probably intends to depart from the firm in near future was predicted by the churn model. Many industries build a model like churn as a common application for data mining techniques. Mobile telephone organizations present across the globe are almost on the verge of building their own churn model. Furthermore, to retain customers, churn results can be efficiently utilized for various other goals. The Churn Management approach is actually the first step in building a model. In general, the project needs a churn model in the best way instead of taking a single method with the best lift. So here we have built a model for the long run. In this digital era, the client of one company may also be a consumer of one or more telecommunication firms. Some of us may use different carriers based on the distance and some others may use different carriers based on the different plans they offer. While performing the analysis using machine learning customer experience tends to provide valuable insights. Some people will change their service providers from time to time. An increase or decrease in the calling

rate will also depend on different job responsibilities. Based on the availability of the data various situations may reflect.

3 Problem Statement

The telecom service churn problem refers to predicting which customers are likely to switch to a competitor's service or terminate their subscription with a telecom service provider. Churn is a common problem for telecom service providers, as losing customers can have a significant impact on their revenue and profitability. Telecom service providers can use historical customer data such as demographics, usage patterns, call and text logs, and customer service interactions to address this problem to build predictive models. These models can then be used to identify customers who are at a high risk of churning and take proactive measures to retain them, such as offering personalized incentives or improving customer service.

The telecom service churn problem is a classification problem, where the goal is to classify customers as either churned or retained. The predictive model's performance can be evaluated using metrics such as accuracy, precision, recall, and F1 score. Feature engineering, hyperparameter tuning, and ensemble learning are some techniques that can be used to improve the performance of the model.

Solving the telecom service churn problem can help telecom service providers improve customer retention, reduce customer acquisition costs, and ultimately increase revenue and profitability.

4 Data Cleaning and Prepossessing

For the purpose of customer churn prediction, we used a publicly available data set of telco communication data, which contained 7031 rows and 21 columns. The data set included features such as tenure, monthly charges, internet service, online streaming, and others. In order to fit this data into a machine-learning model, we needed to preprocess and clean the data.

One issue we encountered during data cleaning was that the dataset did not contain total charges in four rows. We chose to remove these four rows from the dataset, as it is generally better to remove incomplete rows rather than attempting to impute missing values. Next, we encountered categorical features such as Gender, device protection, online security, and tech support, which contained two different string values. To handle this, we used mapping to replace these values with binary numerical values of 0 and 1. For features such as payment process and contract, which contained more than two non-unique string values, we used one-hot encoding to replace the feature entries. One-hot encoding is a common technique for handling categorical data, which involves creating a new binary column for each unique category in the feature.

Data preprocessing and cleaning are essential steps for preparing a dataset

for machine learning. This process involves identifying and handling missing data, converting categorical data into numerical values, and encoding features in a way that can be understood by a machine learning model. In our telco communication dataset, we removed incomplete rows and used mapping and one-hot encoding to handle categorical features. These preprocessing steps enabled us to fit the data into a machine-learning model for customer churn prediction.

5 Brief Introduction of Models

5.1 Logistic Regression

Logistic regression is a statistical method used to model the relationship between a binary dependent variable and one or more independent variables. The logistic regression model is based on the logistic function, which takes the form:

$$p = \frac{1}{1 + e^{-z}} \tag{1}$$

where p is the probability of the binary outcome, z is the linear combination of the independent variables. The logistic regression model estimates the coefficients of the independent variables that maximize the likelihood of the observed data. The logistic regression model can be used to make predictions on new data by plugging in the values of the independent variables and calculating the probability of the binary outcome using the logistic function. including biomedical research, social sciences, marketing, and finance.

5.2 K-nearest neighbor

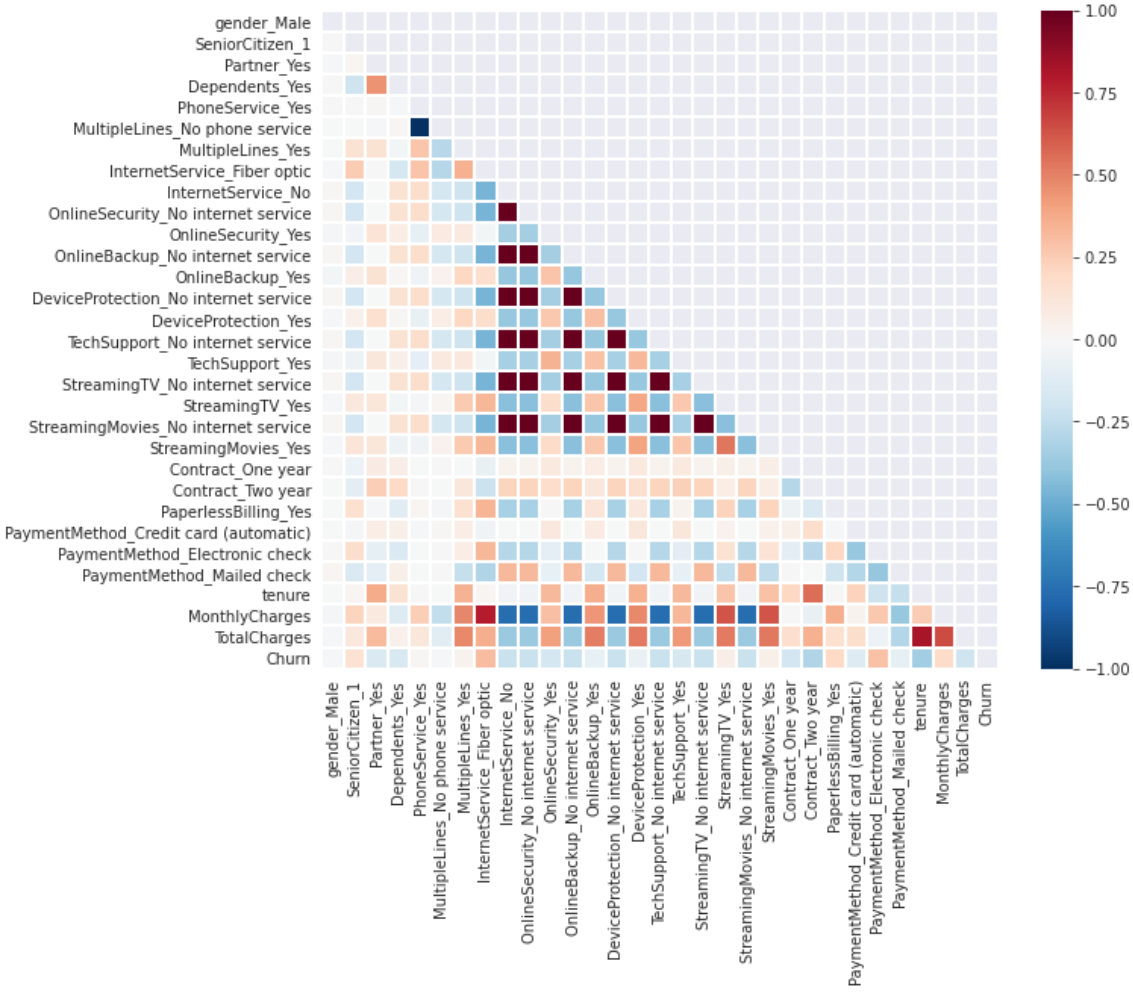
K-Nearest Neighbors (KNN) is a classification algorithm used to predict the class of a given sample based on the class of its k-nearest neighbor in a training dataset. KNN works by measuring the distance between the test data and all training data points and then selecting the k data points that are closest in distance. The distance measure used is typically Euclidean distance. Once the k nearest neighbors have been identified, the algorithm assigns the class label that is most common among those neighbors to the test data point. The value of k is a hyperparameter that can be tuned to improve the algorithm's performance.

5.3 Support Vector Machine

Support vector machine is a supervised machine learning problem where we try to find a hyperplane that best separates the two classes. SVM works best when the dataset is small and complex. It is usually advisable to first use logistic regression and see how does it performs, if it fails to give a good accuracy we can go for SVM without any kernel. SVM uses a kernel trick to transform the data into a higher dimensional space where the classes can be separated by a hyperplane. The choice of kernel function can affect the performance of the model. During training, SVM tries to find the optimal hyperplane by solving an optimization problem. The goal is to minimize the margin and simultaneously minimize the classification error.

5.4 Random Forest

Random forest is a powerful machine learning algorithm that can be used for both classification and regression tasks. It belongs to the family of ensemble methods, which means it combines multiple decision trees to make predictions. The algorithm works by creating a set of decision trees, each trained on a random sample of the data and features. During the training process, at each split, a subset of features is randomly selected and the best feature is chosen as the split point. In the case of making a classification, each tree in the forest independently produces a prediction, and the final prediction is obtained by taking the majority vote of all the trees. The key advantages of the random forest include its ability to handle high-dimensional data, nonlinear relationships between variables, and missing data. Additionally, it provides a measure of feature importance, which can help in feature selection.



correlation between different features

6 Pre-Sampling Analysis:

Table 1: Analysis

Model	Outputs	Precision	Recall	F1-Score
Logistic Regression	0	0.83	0.90	0.87
	1	0.65	0.51	0.57
KNN	0	0.78	0.91	0.84
	1	0.54	0.30	0.38
Random Forest	0	0.83	0.90	0.86
	1	0.64	0.48	0.55
SVM	0	0.83	0.90	0.86
	1	0.64	0.48	0.54

Based on the information provided, it seems that the logistic regression model has the highest accuracy on the test set, but this accuracy may be inflated due to the class imbalance in the dataset. The fact that 66% of the data is labeled as nonchurner suggests that a model that simply predicts nonchurner for every data point would achieve 66*per* accuracy. Therefore, it is important to look at other metrics like precision, recall, and F1-score, which provide a more detailed picture of the model's performance. From the information given, it seems that the models have poor F1 scores for the churner class (i.e., a class with label 1), indicating that the models are not doing well in correctly identifying churners.

7 Re-sampling

Resampling is a statistical method used to estimate the properties of a population by analyzing a sample of data. In essence, it involves repeatedly drawing samples from a given dataset and analyzing the statistics of these samples to estimate the population parameters.

There are different types of resampling techniques, including bootstrap and cross-validation. Bootstrap resampling involves randomly drawing samples from the original dataset with replacement, and calculating the desired statistic on each of the resampled datasets. This method is useful for estimating the variability of a statistic and constructing confidence intervals.

Cross-validation, on the other hand, involves splitting the dataset into training and testing subsets and using the training set to build a model, which is then tested on the testing set. This process is repeated several times, with different subsets of the data used for training and testing, to estimate the performance of the model on new data.

Overall, resampling is a powerful technique for understanding the properties of a dataset and assessing the performance of statistical models.

SMOTE, which stands for Synthetic Minority Over-sampling Technique, is an oversampling technique used in machine learning to handle imbalanced datasets. The algorithm works by generating synthetic examples in the minority class using the existing data. **General algorithm of SMOTE:**

Input: A dataset with minority and majority classes.

- Determine the number of synthetic examples to generate using SMOTE.
- For each example in the minority class, select k nearest neighbors from the same class.
- Randomly choose one of the k nearest neighbors and create a synthetic example by combining the features of the selected example and the original example.
- Add the synthetic example to the minority class.
- Repeat steps 3-5 until the desired number of synthetic examples have been generated.
- Return the new dataset with the balanced classes.

In step 3, the value of k can be chosen based on the size of the minority class. Typically, a value of 5 is used, but this can be adjusted based on the specific problem.

The SMOTE algorithm is a useful technique to handle imbalanced datasets in machine learning and has been shown to improve the performance of models on such datasets.

8 Post-Resampling Analysis

Table 2: Analysis

Model	Outputs	Precision	Recall	F1-Score
Logistic Regression	0	0.91	0.87	0.89
	1	0.90	0.93	0.91
KNN	0	0.98	0.98	0.98
	1	0.99	0.99	0.99
Random Forest	0	0.97	0.96	0.96
	1	0.97	0.98	0.97
SVM	0	0.95	0.94	0.94
	1	0.95	0.96	0.96

9 Conclusion

In conclusion, this project aimed to classify customer churn using four different machine learning algorithms: random forest, SVM, logistic regression, and KNN. However, due to the imbalanced dataset, the accuracy of the models did not meet our expectations. To address this issue, we employed a resampling technique called SMOTE-ENN, which resulted in a significant improvement in the performance of our model. As a result of this technique, our model was able to provide accurate classifications on the test data. Thus, it can be concluded that SMOTE-ENN resampling is an effective approach to improve the accuracy of models when dealing with imbalanced datasets. This project highlights the importance of using appropriate techniques to address imbalanced data to obtain better results.

10 Reference

- Kim H S, Yoon C H. Determinants of subscriber churn and customer loyalty in the Korean mobile telephony market. Telecommunications Policy, 2004, 28(9): 751–765.
- Rosset S, Neumann E. Integrating customer value considerations into predictive modeling. Third IEEE International Conference on Data Mining, 2003: 1–8.
- model for managers in mobile telecomcommunication. Networks and Services Research Conference, Proceedings of the 3rd Annual16-18 May 2005: 48–53.