Customer Churn Prediction

Jinyu Wang[1]

[1]Data Science Initiative, Brown University
*Github link: https://github.com/JinyuWang123/DATA1030-MID.git

10/20/2022

1 Introduction

Customer churn is defined as when customers or subscribers discontinue doing business with a firm or service. The telecommunications business has an annual churn rate of 15-25 percent in this highly competitive market. Therefore, As a result, by addressing churn, these businesses may not only preserve their market position, but also grow and thrive [1]. This study analyzes data obtained from kaggle to predict which customers are at high risk of churn.

1.1 Data[2]

The customer churn data was collected from a telco company in California. It is a binary classification problem. About the dataset, All variable are defined in Table 1. This dataset has 7043 rows and 21 columns (1 target variable and 20 features). Among the 20 features, only 3 features, Tenure, MonthlyCharge, and TotalCharges are continuous variables, while the other 17 features are categorical variables.

1.2 Overview of previous work

Because this dataset is from Kaggle, several public projects have studied this dataset before this study. Nilan trained on this dataset to obtain the VotingClassifier model with an accuracy score of 0.81 [3]. Aguiar used the random forest classifier to obtain some preliminary feature importances and she/he found that the three variables Tenure Contract and TotalCharges had the highest feature importances and the final random forest classifier had an accuracy of 0.76 on the test set [4].

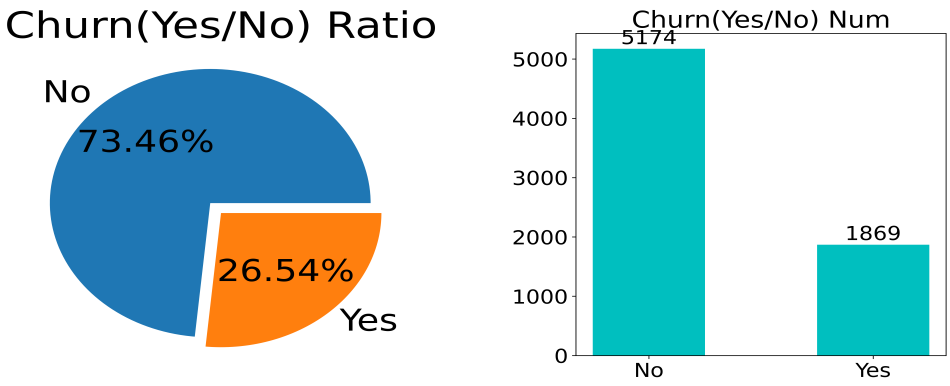
2 Exploratory data Analysis

2.1 Target variable

The target variable is a binary categorical variable and from the Figure 1 we can find that we have a binary classification problem with the imbalanced dataset.

Table 1: Features description

Features	**Description**
Churn (Target Variable)	Whether the customer churned or not
customerID	ID that identifies each customer
Partner	Indicates if the customer is married
gender	The customer's gender
SeniorCitizen	Whether the customer is a senior citizen
Tenure	months the customer has been with the company
Dependents	Indicates if the customer lives with any dependents
Phone Service	Indicates if the customer subscribes to phone service
Multiple Lines	Indicates if the customer subscribes to multiple telephone lines
Internet Service	Indicates if the customer subscribes to an additional online security service
Online Security	Indicates if the customer subscribes to an additional online security service
Online Backup	Indicates if the customer subscribes to an additional online backup service
DeviceProtection	Whether the customer has device protection or not
TechSupport	Whether the customer has tech support or not
StreamingTV	Whether the customer has streaming TV or not
StreamingMovies	Whether the customer has streaming movies or not
Contract	The contract term of the customer
PaperlessBilling	Whether the customer has paperless billing or not
PaymentMethod	The customer's payment method
MonthlyCharges	The amount charged to the customer monthly
TotalCharges	The total amount charged to the customer



(a) (b)

Figure 1: **Visualization for target variable** (a) **Pie Chart** (b) **Bar Chart**

2.2 Categorical features

Among the 17 categorical features we picked the two features, gender and seniorcitizen, and target variable for the graph. from Figure 2, we can see that:

1. Gender is not an indicative of churn.
2. SeniorCitizens are only 16% of customers, but they have a much higher churn ratio: 42% against 23% for non-senior customers.

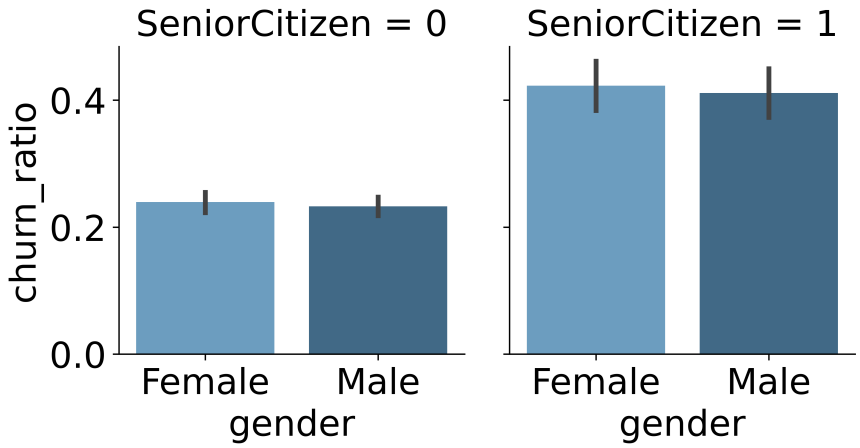


Figure 2: **Visualization for Categorical features (gender, SeniorCitizen)**

2.3 Continuous features

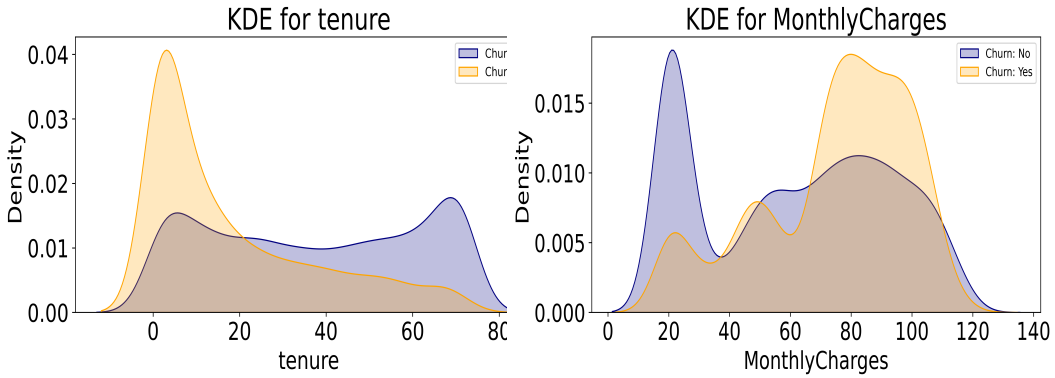
Among the 3 continuous features we picked the two features, MonthlyCharges and Tenure, and target variable for the graph. from Figure 3, we can see that:

1. Recent clients are more likely to churn.
2. Clients with higher MonthlyCharges are also more likely to churn.
3. Tenure and MonthlyCharges are probably important features.

3 Methods

Neither group structure nor time-series data are in the dataset. And all data are independent and identically distributed (IID) in this in imbalanced dataset. Thus we utilize the stratified method (training set 60%, validation set 20%, test set 20%) to split the dataset. As a result, the number of data points in the train set was 4,225, the number of data points in the validation set was 1,409, and the test set had the same number of instances as the validation set.

In the dataset, the customerID feature has no real meaning, Phone Service and Multiple Lines these two features have the same meaning, so we drop customerID and Phone Service these two features. we apply standardscaler on the continuous features as they have tailed distribution and apply onehotencoder on categorical



(a) (b)

Figure 3: **Visualization for Continuous features (MonthlyCharges, Tenure)**

features because their categories are unordered. During the preprocessing, we found that TotalCharges for 11 new users (tenure = 0) are missing values. However, according to the telecommunication company policy, new users need to pay the first month's charges, so we choose their MonthlyCharges to impute missing values. And the preprocessed data has 7043 rows and 37 columns, including 1 target variable and 36 features.

We utilized five ML algorithms on this dataset: Ridge, Lasso, Support Vector Classifier,Random Forest Classifier, and XGBoost Classifier. Lasso and Ridge are linear models, while others are non-linear models. The details about parameter tuning for ML models are summarized in Table 2. Besides, Accuracy and F1-score were determined as the evaluation metric because the project is a classification problem. The model evaluation metric of Accuracy has the advantages of computational simplicity and low time complexity, but it is not sufficient for evaluating models trained on unbalanced dataset. But the F1-score combines the precision and recall of a classifier into a single metric by taking their harmonic mean, which could be used to evaluate models trained on unbalanced dataset. Two kinds of uncertainties could be generated due to splitting and due to non-deterministic ML methods . We utilized five random states to measure the uncertainties due to splitting. As the result in Table3, 0.8176 holds the lowest uncertainty among all models. Since random forest is the only non-deterministic model, we trained the random forest model of random state = 1 five times, I obtained that uncertainty of accuracy of RF was 0.0205 and uncertainty of F1-score of RF was 0.0196.

The Cross validation pipeline in this research was describe in five steps: (I) Select 5 random states: 0, 1, 2, 3, 4, 5. (II) Stratified split dataset (training set 60%, validation set 20%, test set 20%). (III) Apply standardscaler on the continuous features and onehotencoder on categorical features. (IV) Loop through all combinations of hyperparameter combos. (V) Output the best model and best scores of each state.

4 Results

With the formula of accuracy and F1-score, the mean of baseline accuracy and the mean of F1-score could be calculated. The mean of baseline accuracy is 0.7346 and the mean of baseline F1-score is 0.4194. And from Table 3, we can find detailed performances for each Machine Learning model. Among five models, XGBoost Classifier was the most predictive model on this dataset.

Table 2: Basic hyperparameter tuning (each ML Algorithm)

	ML Algorithms	Parameter	Values tuned
1	Ridge	C	1e-3,1e-2,1e-1,1,1e+3,1e+2,1e+1
		max_iter	10,100
2	Lasso	C	1e-3,1e-2,1e-1,1,1e+3,1e+2,1e+1
		max_iter	10,100
3	SVC	C	1e-3,1e-2,1e-1,1,1e+3,1e+2,1e+1
		max_iter	10,100
4	Random forest	max_depth	7,8,9,10,11,12,13,14
		max_features	0.25, 0.5,0.75,1.0
5	XGBoost	max_depth	3,5,7,9,11,13,15
		min_child_weight	1,5,7,9
		n_estimators	100,1000

Table 3: Performances of ML algorithms

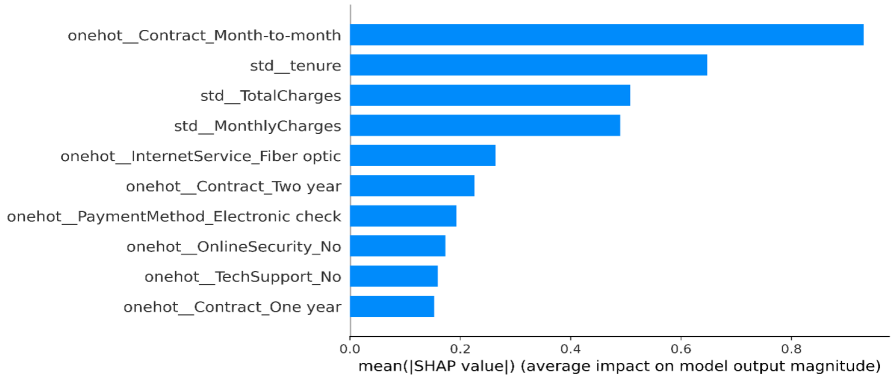
Model	Accuracy (mean + SD)	F1-score (mean +SD)	SD above (accuracy)	SD above (F1-score)	Baseline Accuracy	Baseline F1-score
Ridge	0.7728+0.0180	0.4598+0.2282	2.122	0.177	0.7346	0.4194
Lasso	0.7615+0.0113	0.4223+0.2597	2.381	0.011		
SVC	0.7991+0.0194	0.5117+0.0858	3.325	1.076		
RF	0.8126+0.0149	0.5864+0.0102	5.234	16.37		
XGBoost	**0.8176+0.0079**	**0.5871+0.0144**	**10.51**	**11.65**		

Since the XGBoost Classifier was the most predictive model, SHAP method was utilized to calculate the global feature importance derived from XGBoost Classifier model. From Figure 4(a), the top 3 most important features in the SHAP method are Contract Month to Month, tenure and TotalCharges and the least important feature is Contract one year. Additionally, From Figure 4(b), we concluded that (I) The contract is on a Month to Month is generally a bad thing (This means that the customer may discontinue purchasing telecommunication company's services), but the horizontal dispersion also implies that it depends on other factors. (II) The larger the customer's tenure, the more likely they will continue to buy the company's services.

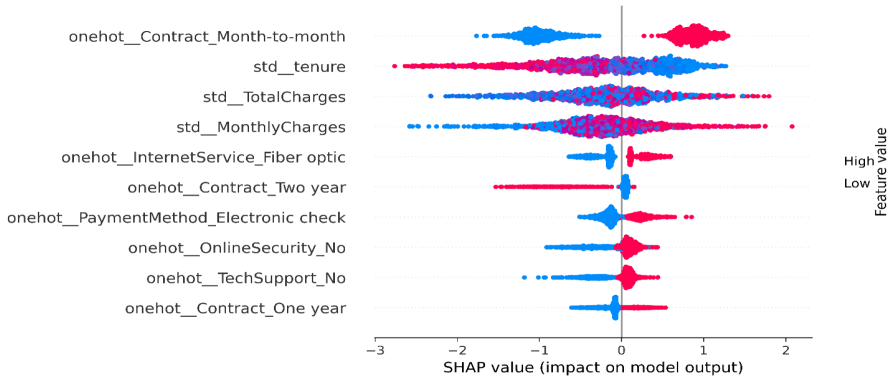
Regarding the local importance, data points with index 0 was analyzed. The main factors that affect the model output results were Contract One Year = 1, Payment Method Electronic check = 1, Contract Month to Month = 0 and tenure, which were also essential features in global feature importance. In the final model, this customer was Predicted to continue purchasing the telecommunication company's services (expit(-3.73) = 0.02). The XGBoost model predicted intuitively that a customer who signs a long-term contract using electronic check has a high level of telecommunication brand loyalty. And in conclusion, the features about contract, tenure and charges could significantly influence the results of XGBoost Classifier model.

5 Outlook

Although this study has preliminarily achieved the preliminary results on telecommunication customer churn, several limitations also apply. First, the sample size was 7043. In the subsequent study, we can obtain more data by reviewing consumer surveys published by different telecommunication companies. Second, Some hyperparameters of five ML algorithms in this study were set with default values. After we obtain



(a)



(b)

Figure 4: **Visualization for global importance (SHAP)**

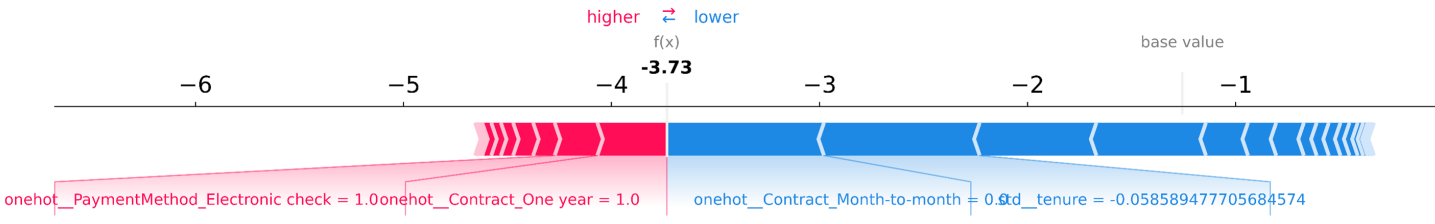


Figure 5: **Visualization for local importance (SHAP)**

a larger dataset, we can try more advanced ML algorithms (such as Gradient boosting) and tune more hyperparameters.

References

[1] P. Sulikowski and T. Zdziebko, "Churn factors identification from real-world data in the telecommunications industry: Case study," *Procedia Computer Science*, vol. 192, pp. 4800–4809, 2021.

[2] "Data link," [Online]. Available: `https://www.kaggle.com/datasets/blastchar/telco-customer-churn`.

[3] "Customer churn prediction," *n.p.*, 2021. [Online]. Available: `https://www.kaggle.com/code/bhartiprasad17/customer-churn-prediction/notebook`.

[4] "Exploratory analysis with seaborn," *n.p.*, 2018. [Online]. Available: `https://www.kaggle.com/code/jsaguiar/exploratory-analysis-with-seaborn`.