Data Mining and Predictive Analytics of Customer Churn Using Machine Learning: A Case Study of Telecommunication Industry

Wenzhou Kean University

Instructor: Chun Tee Lee

Abstract

Predicting customer churn is an effective strategy for modern enterprises to retain existing customers. This paper systematically analyzes telecommunications company customer data from IBM. First, through EDA, we analyze and process the customer churn data including processing the outliers, missing values, duplicate values, and other screening of effective data. The second part, through data visualization and data cleaning, on the one hand, analyzes the impact of different data features on the distribution of customer churn, on the other hand, a novel classifies algorithm, which is a combination of XGB algorithm and one-hot encoding, Label encoding, is used for feature selection. Finally, a variety of customer churn prediction models are constructed from the selected feature sub-dataset, and the model performance is compared through the confusion matrix. Get Feature importance priority by XGBoost algorithm. The experimental results show that using XGBoost as the base classifier, combined with the novel feature method combined with one-hot encoding and Label encoding, the features with higher feature importance are screened out and the redundant features are deleted, which can significantly enhance The prediction performance for the commonly-used logit model. The accuracy rate, recall rate, F-score value, and evaluation metrics such as Recall have been greatly improved, and it has high application value.

Keywords: Customer churn prediction; Predictive analytics; Feature Selection; Logistics Regression; Random Forest: XGBoost; Data mining;

Contents

1.Introduction

Customer churn is one of the most substantial problems faced by modern enterprises, due to fierce competition resulting from dynamic market competition, and continuous introductions of new competitive offerings. Previous studies have found that the loss of customer churn to enterprises is huge. First, the lost customers can no longer create value for the company. Every 1% of the lost customers will cause a 5%-16% profit loss to the company [1]. Second, the churn customers may spread some negative word of mouth, affect the image of the company will even the churn of other customers. Third, the cost of developing one new customer for an enterprise is 4-5 times the cost of retaining one old customer. Even a 5% increase in user retention rate can bring an 85% increase in profits for the enterprise [2]. In response, many enterprises have switched from an offer-centric strategy, designed to sell as many offerings as possible, to a customer-oriented retention approach that explicitly seeks to reduce churn [3]. The hypothesis of customer churn prediction is to rank which features of customers have a deeper impact on the customer churn.

Using big data to carry out customer relationship management has the advantages of comprehensiveness, timeliness, and veracity [3][4]. Through machine learning and data mining, we can carry out customer churn prediction by modeling. At present, there are many research directions for the prediction of customer churn: based on traditional statistical methods, based on artificial intelligence, and prediction algorithms based on integrated classifiers [5]. However, the particularity of customer data is manifested in high-dimensional, large indexes, churn and non-churn data imbalance, unclear classification requirements, etc. [4][6]. This particularity affects the performance of the prediction model. The goal of this article is to compare the logistic, decision tree, random forest, XGBoost, and other integrated classifiers used to feature selection, retain important features, and ignore irrelevant attributes. Finally, improve classification performance, save model training time, and improve prediction model accuracy. The following structure of this article is as follows: The second part introduces the research status of related fields of customer churn prediction. The third part describes data sources and performs data sorting and data cleaning. In order to eliminate the outliers, complete data preprocessing. The fourth part, through data visualization and feature selection, analyzes different data characteristics for customers. Then, construct a valid feature sub-dataset preparing for modeling. The fifth part is to build customer churn prediction models and compare the performance of different classifiers. The sixth part is the results and data analysis. Finally, summarize this article and point out the deficiencies in the research and future work directions.

2. Related Works

2.1 Customer churn and machine learning

Customer churn refers to the termination of the current telecom package service and no longer renewal of the contract, which is also referred to as the user leaving the network. According to existing research, the main factors of customer churn can be summarized as price, network quality, service quality, fault response speed, value-added services, etc., of course, also include some own reasons, such as work-related number changes, residential cities, other networks (the same Type of service providers) competition, etc[5]. These potential factors have become an important input for us to conduct churn prediction modeling, in order to achieve the goal of churn warning and maintenance.

Machine learning [5] is a multi-disciplinary interdisciplinary, involving probability theory, statistics, approximation theory, convex analysis, algorithm complexity theory and other disciplines. It uses computers to simulate human behavior to obtain new knowledge and laws. Big data technology provides a broader application stage for machine learning.

The research of telecommunications customer churn based on machine learning algorithms in foreign countries is relatively early. Louis used the decision tree TreeNet not only to achieve high accuracy prediction of Verizon customer churn, but also to find important factors affecting customer churn[7]; Nath et al.[7] People use the Bayes classifier to accurately calculate the probability of customers leaving the network in the next 3 months; Khan et al. [7] used a small number of variables through recursion

The neural network RNN algorithm performs churn prediction analysis on British Telecom customer data. In recent years, the multi-algorithm fusion machine learning method has become an important branch for predicting customer churn and has attracted more and more scholars' attention[6].

The current research on telecommunications customer churn is mostly based on data mining software such as Clementine and SPSS. There are problems such as inconvenience of single data mining algorithm or combined mining algorithm parameter adjustment, limited use of data dimensions, etc., and it is difficult to comprehensively analyze customer churn [6].

3.Data Collection and Preprocessing

3.1 Data Description

The telecommunications customer churn data set implemented in this article comes from IBM Sample Dataset, 2020 update in IBM Cognos Analytics Website. The data set includes 7043 samples, including 52 customer attributes. Such as Tenure, Contract, Cost, etc. some properties are shown in *Table 1*:

Attribute	Description

Number of Referrals	Indicates the number of referrals to date that the customer has made.
Tenure	Indicates the total amount of months that the customer has been with the company by the end of the quarter specified above.
Contract	Indicates the customer’s current contract type: [Month-to-Month, One Year, Two Year].
Payment Method	Indicates how the customer pays their bill: Bank Withdrawal, Credit Card, Mailed Check
Monthly Charge	Indicates the customer’s current total monthly charge for all their services from the company.
Total Charges	Indicates the customer’s total charges, calculated to the end of the quarter specified above.
Internet Service	Indicates if the customer subscribes to Internet service with the company: No, DSL, Fiber Optic, Cable.

Table 1: part attributes of dataset

3.2 Data Preprocessing

In this article, all the code is written in Python3.8 language, running in the *jupyter notebook* environment, using *sklearn, tensorflow, seabor, XGBoost, NumPy, pandas*, and other public toolkits packages. First of all, use the *read_csv()* function to import the initial data set into *.ipynb* file, start data preprocessing. The preprocessing of the data set includes filling up missing values, deleting duplicate values, and binarizing categorical variables. The processed sample set has a total of 7038 rows and 52 columns. Starting to start data preprocessing, there are the following issues that need to be addressed:

First, among them, there are 1865 samples of lost customers, accounting for about 26.5% of the data set, and the remaining customer samples accounted for 73.5%, as shown in *Figure 1*. When there is less than one category in the two-category data set When 40% of the total data set is [4], the data set is a class imbalanced data set. Therefore, SMOTE is used to oversample the churn customer samples in the training set.

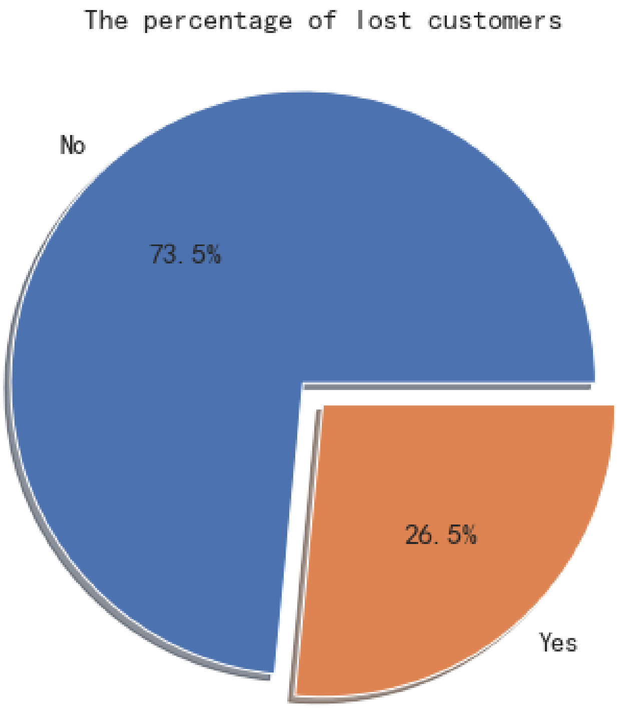
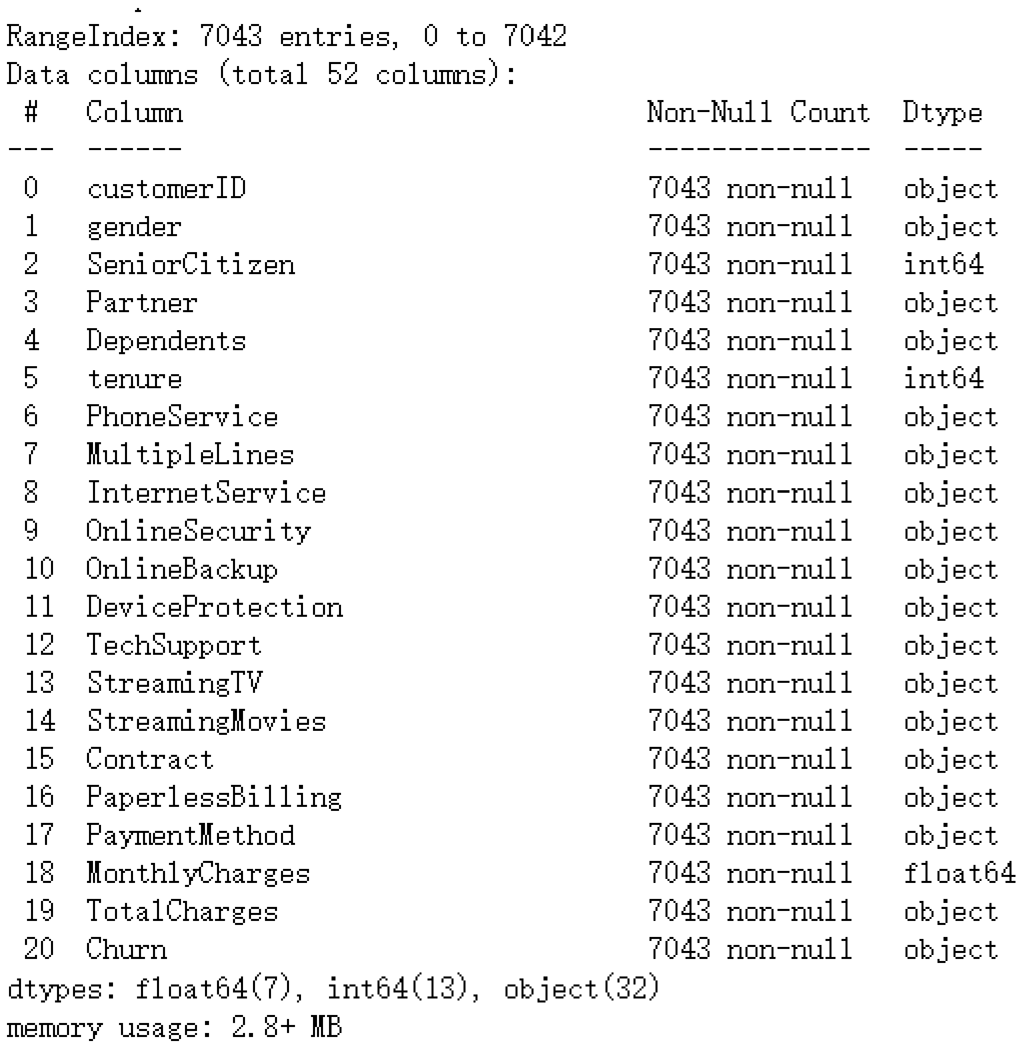


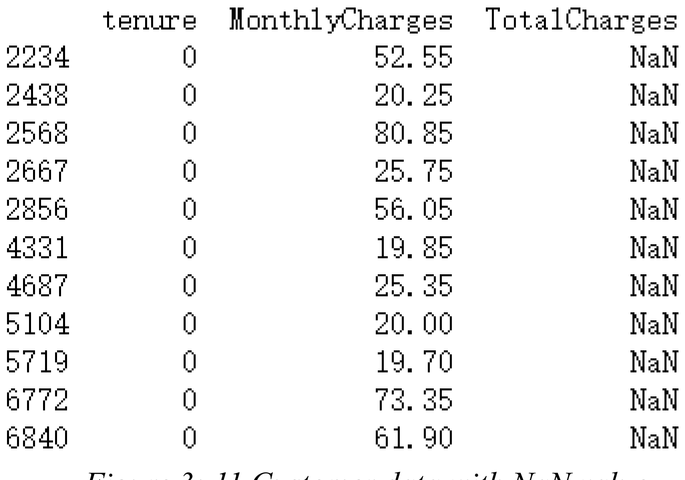
Figure 1: Percentage of Lost(churn) Customers Data

Second, the data types are diverse and there is no direct comparison, such as *Figure 2.* Among the dataset, we found that, first, the total cost of the data "*Total Charges*" should be the same type as "*Monthly Charges*". Therefore, it is necessary to convert *"Total Charges"* from *"object"* to *"float64"*. Second, it is found that 11 users' *tenure* (frequently connected to the network) is 0, as shown in *Figure 3.* It is speculated that they are new users of the current month. However, these 11 users have produced "*Monthly Charges*" (monthly fees). We fill in the monthly fees of the current month into *"Total Charges"*. The candidate method is to directly delete the sample where the missing value is located, but this will cause the training set to shrink, so it is not used.



```
RangeIndex: 7043 entries, 0 to 7042
Data columns (total 52 columns):
 #   Column                              Non-Null Count  Dtype
---  ------                              --------------  -----
 0   customerID                          7043 non-null   object
 1   gender                              7043 non-null   object
 2   SeniorCitizen                       7043 non-null   int64
 3   Partner                             7043 non-null   object
 4   Dependents                          7043 non-null   object
 5   tenure                              7043 non-null   int64
 6   PhoneService                        7043 non-null   object
 7   MultipleLines                       7043 non-null   object
 8   InternetService                     7043 non-null   object
 9   OnlineSecurity                      7043 non-null   object
 10  OnlineBackup                        7043 non-null   object
 11  DeviceProtection                    7043 non-null   object
 12  TechSupport                         7043 non-null   object
 13  StreamingTV                         7043 non-null   object
 14  StreamingMovies                     7043 non-null   object
 15  Contract                            7043 non-null   object
 16  PaperlessBilling                    7043 non-null   object
 17  PaymentMethod                       7043 non-null   object
 18  MonthlyCharges                      7043 non-null   float64
 19  TotalCharges                        7043 non-null   object
 20  Churn                               7043 non-null   object
dtypes: float64(7), int64(13), object(32)
memory usage: 2.8+ MB
```

Figure 2:part Datatype information of Dataset

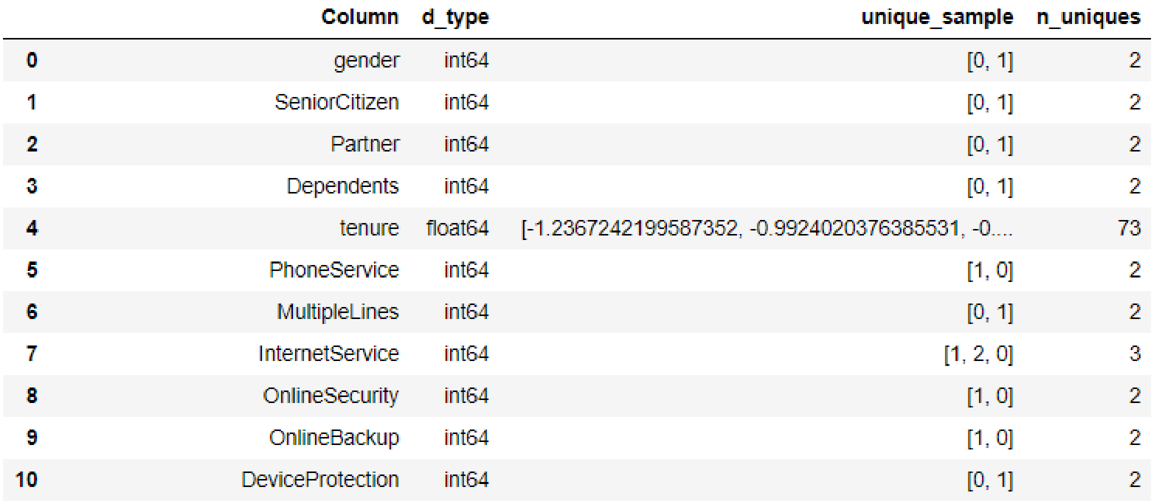


```
      tenure  MonthlyCharges  TotalCharges
2234       0           52.55           NaN
2438       0           20.25           NaN
2568       0           80.85           NaN
2667       0           25.75           NaN
2856       0           56.05           NaN
4331       0           19.85           NaN
4687       0           25.35           NaN
5104       0           20.00           NaN
5719       0           19.70           NaN
6772       0           73.35           NaN
6840       0           61.90           NaN
```

Figure 3: 11 Customer data with NaN value

Third, as shown in *Figure 2*, the data set contains *float64(7), int64(13), object(32)*, and the inconsistency of data types will make it difficult to form an effective comparison in the modeling step. Therefore, the basic idea of this article is to talk about objects as much as possible Type data is converted into comparable numeric data. Data with "large dimensional difference" in numeric type data is processed by standard scaler, that is, numeric conversion. Among them, for *[Male, Female], [Yes, No, No internet service],* such kind of object data, does not have an obvious numerical type relationship, we use integer encoding to convert it to *[0,1,2]*. This process will be described in the fourth part, feature selection.

Finally, we finish data prepossessing. The data set currently has 7043 samples, with a total of 7043 rows and 52 columns. As shown as *Figure 4:*



	Column	d_type	unique_sample	n_uniques
0	gender	int64	[0, 1]	2
1	SeniorCitizen	int64	[0, 1]	2
2	Partner	int64	[0, 1]	2
3	Dependents	int64	[0, 1]	2
4	tenure	float64	[-1.2367242199587352, -0.9924020376385531, -0....	73
5	PhoneService	int64	[1, 0]	2
6	MultipleLines	int64	[0, 1]	2
7	InternetService	int64	[1, 2, 0]	3
8	OnlineSecurity	int64	[1, 0]	2
9	OnlineBackup	int64	[1, 0]	2
10	DeviceProtection	int64	[0, 1]	2

Figure 4: Partially Preprocessed Dataset

4.Methodology

4.1 Data Visualization

In this step, we use the drawing skills of machine learning, try to analyze the relationship between different customer's features, and further prepare for feature selection. First, we perform correlation analysis and use heat map to draw the Pearson matrix, as shown in *Figure 5 and Figure 6:*

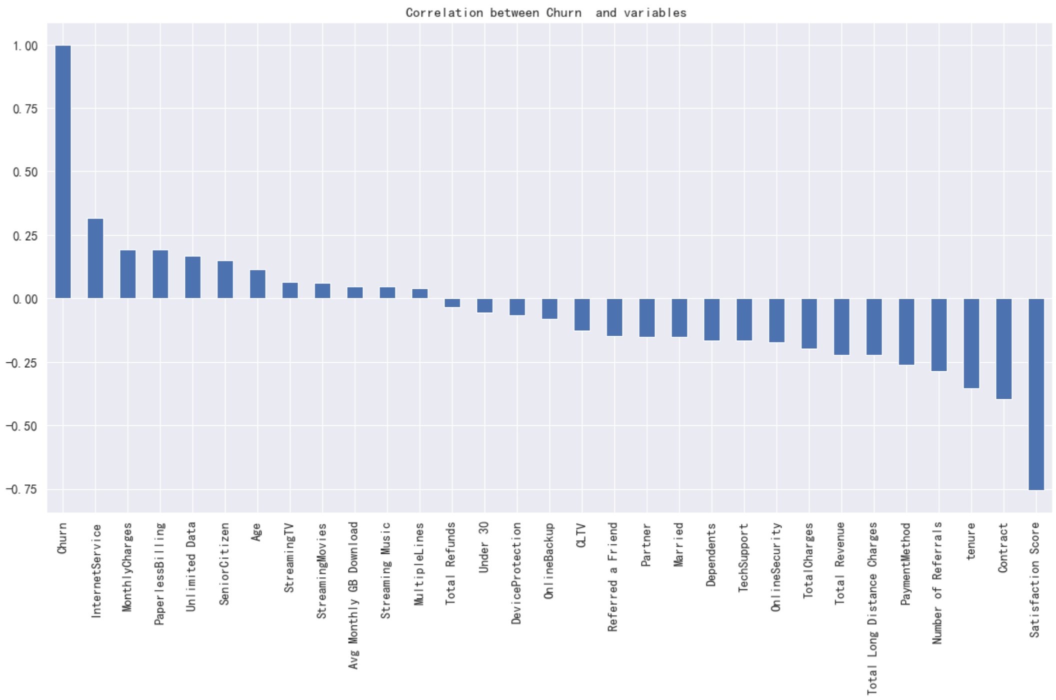


Figure 5: Correlation between Churn and Features

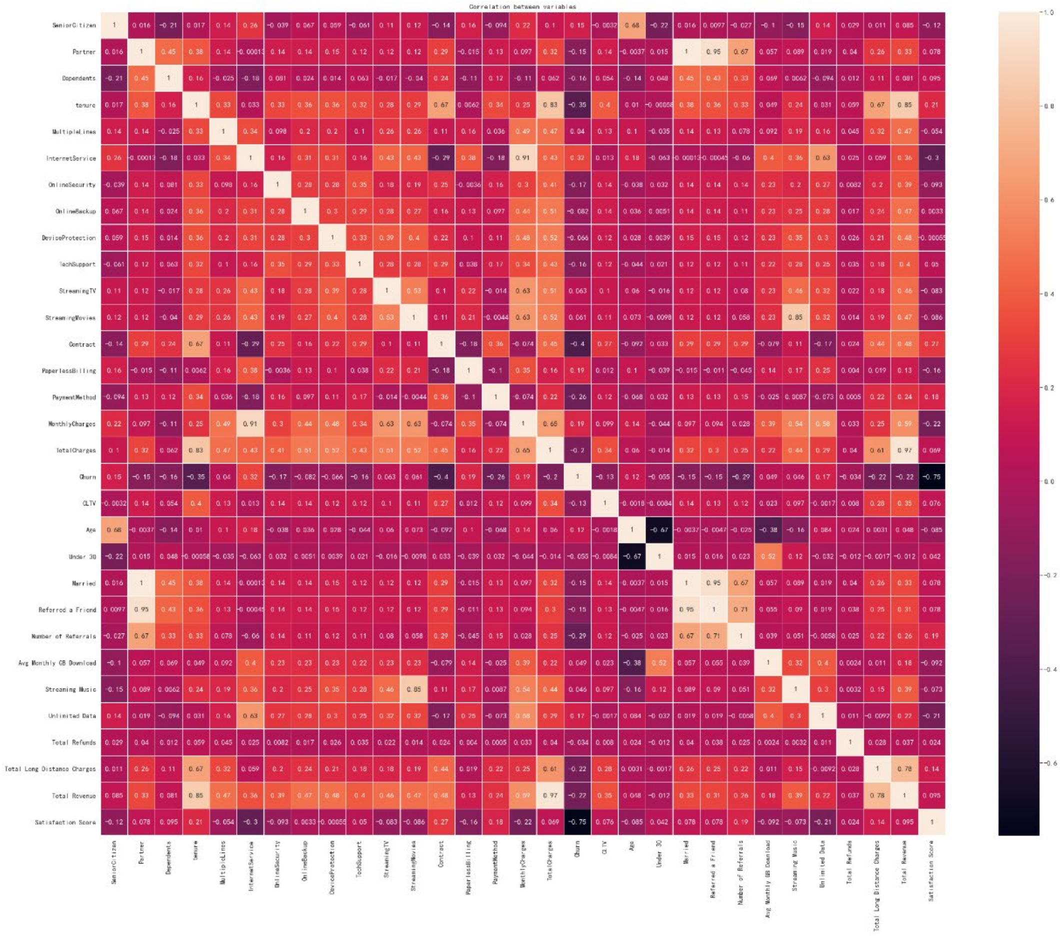


Figure 6: Correlation Heat Map between Churn and Features

It can be seen intuitively from Figure 5 and Figure 6 that the two features *"PhoneService"* and *"gender"* have the weakest correlation with the *"churn"* target variable. In addition, *"customerID"* cannot determine the probability that someone will churn. *"Count", "Country", "State"* and other variables have unique values, so they will not affect the target variable, so they can be ignored. Similarly, it represents the *"ZIPcode", "Longitude",* and *"Latitude"* of the customer's personal information "Will also be deleted. Finally, the variables *"Churn Rate", "Churn Score"* and *"Churn Reason"* are too highly correlated with the target variable, and the final result can even be derived directly. This is not necessary for the predictive model and may even lead to overfitting Problem, so it can also be ignored.

The above secondary data processing is based on subjective inference. We found 22 redundant data features, but before deleting them, I used data visualization techniques for cross-validation, as shown in *Figure 7, Figure 8:*

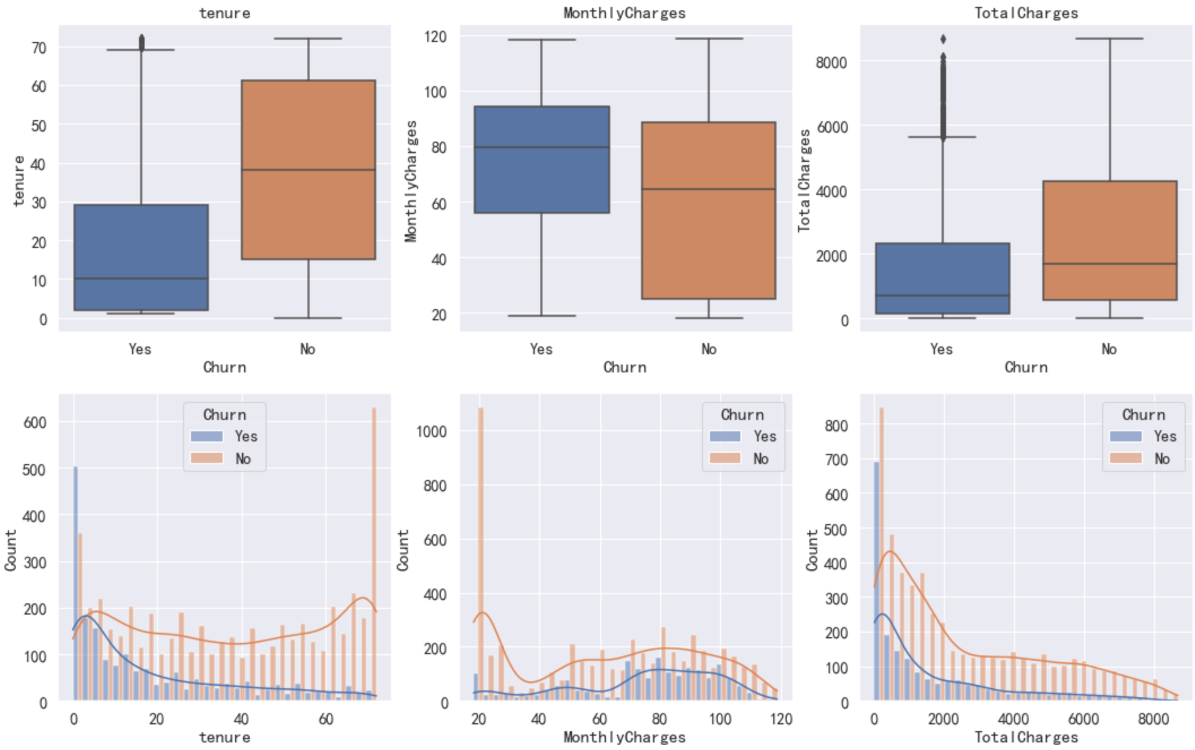


Figure 7.1: Distributions of Numerical Data Features

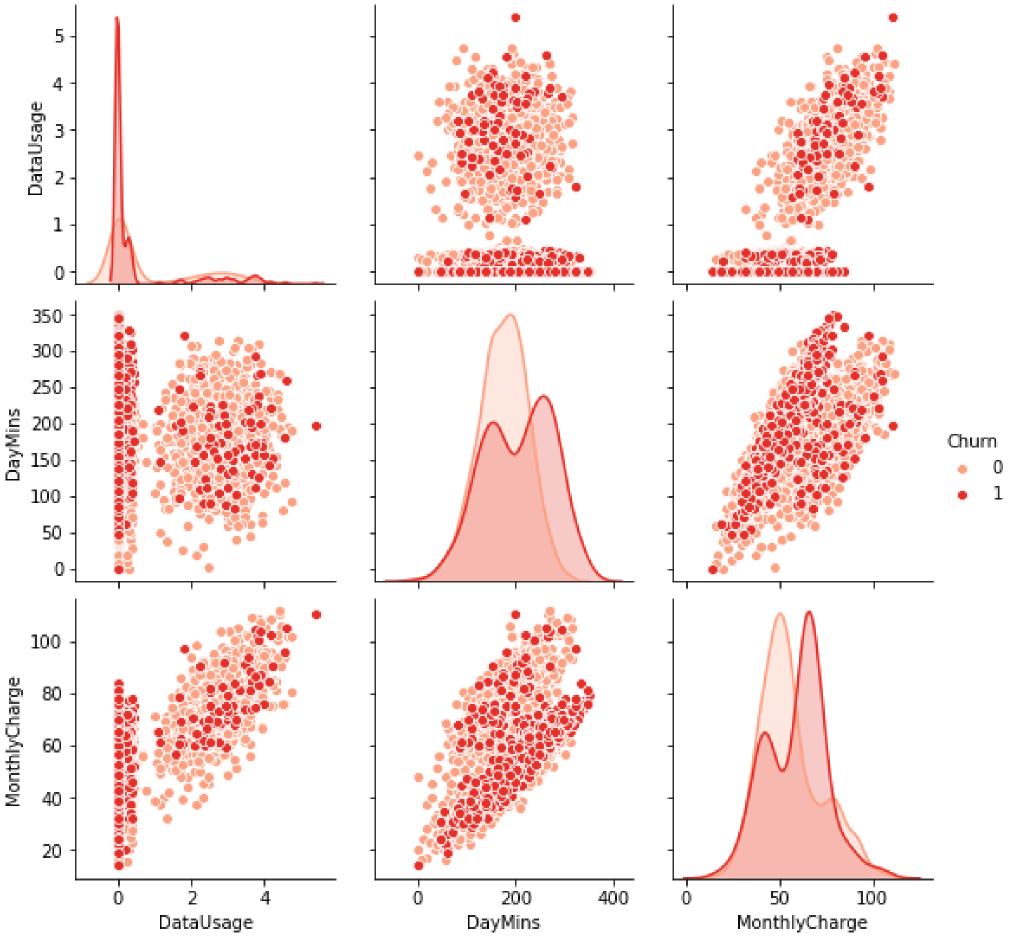


Figure 7.2: Distributions of Numerical Data Features

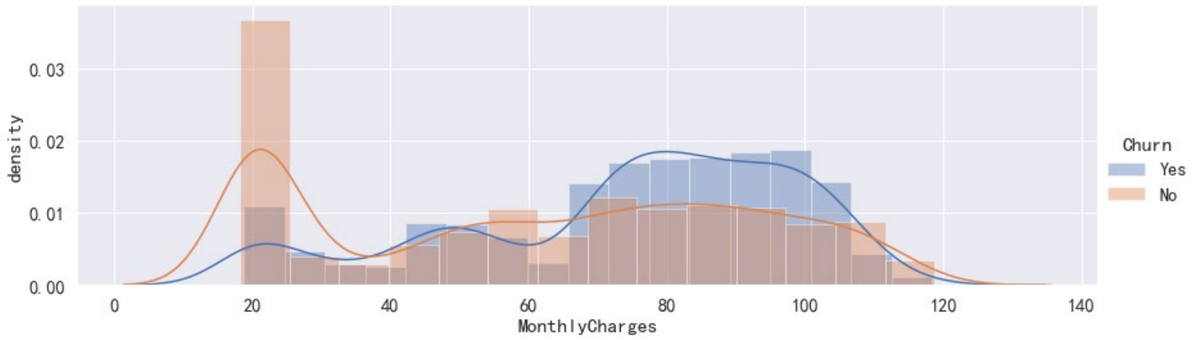


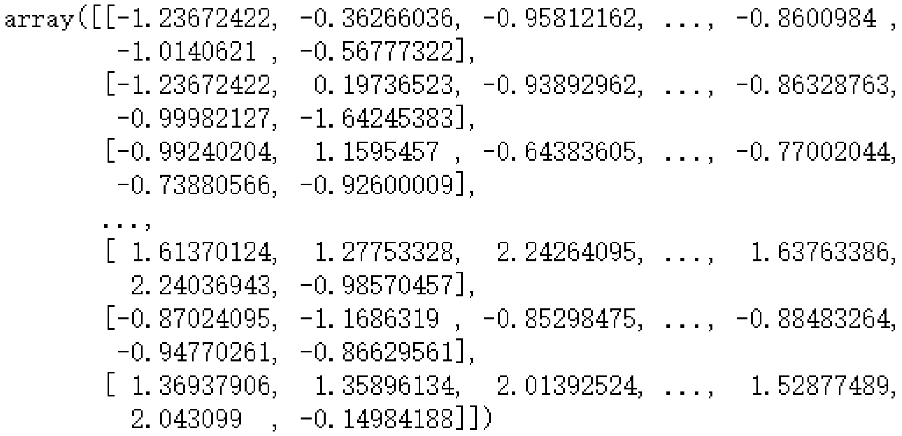
Figure 8: Kernel Plot of Monthly Charge Feature

From *Figure 7 and Figure 8*, it can be seen that the user churn rate in the range of about 70-110 monthly consumption is relatively high. In addition, *"monthly charge", "total charge", "Tenure"* and other features are highly correlated with the objective feature *"churn"*, and the curve also shows regularity.

4.2Feature Selection

After completing the data conversion processing for numeric type data and non-data type data, it is now necessary to continue processing according to the modeling requirements. We have already described the feature selection operation in the data prepossessing part before, and then we will explain in detail how to use the Encoding technique.

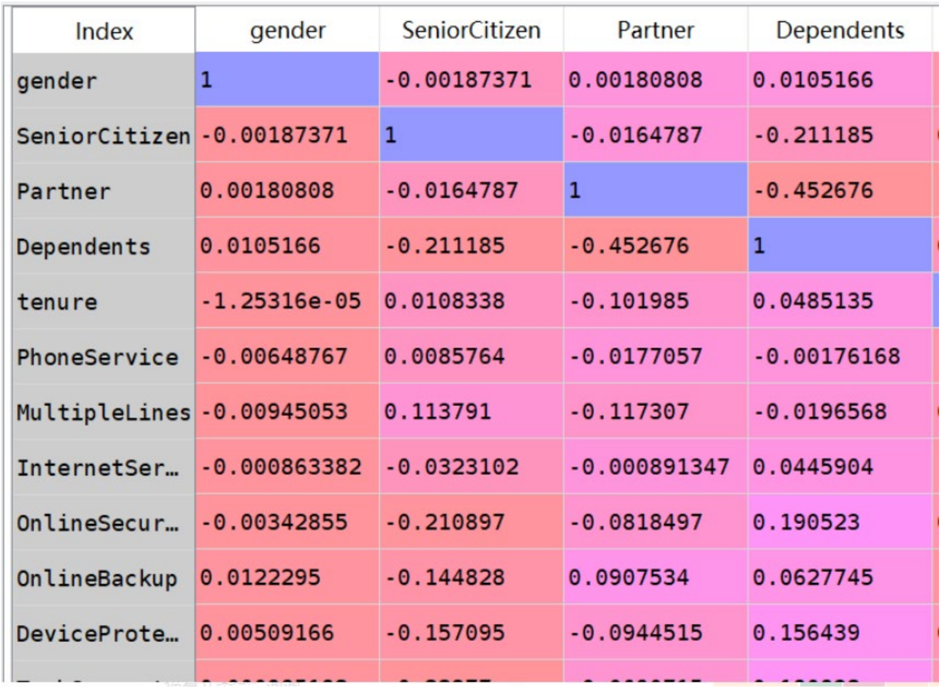
In the whole data set, features are divided into continuous features and discrete features, among which discrete features are subdivided into two categories according to whether there is a numeric relationship between the features. For example, continuous features: *"tenure", "MonthlyCharges", and "TotalCharges"* are generally processed by normalization. After *StandardScaler* processing, the variance of the feature data is 1, and the mean is 0, which reduces the impact of excessive numerical features on the prediction results, as shown as *Figure 9:*



```
array([[-1.23672422, -0.36266036, -0.95812162, ..., -0.8600984 ,
        -1.0140621 , -0.56777322],
       [-1.23672422,  0.19736523, -0.93892962, ..., -0.86328763,
        -0.99982127, -1.64245383],
       [-0.99240204,  1.1595457 , -0.64383605, ..., -0.77002044,
        -0.73880566, -0.92600009],
       ...,
       [ 1.61370124,  1.27753328,  2.24264095, ...,  1.63763386,
         2.24036943, -0.98570457],
       [-0.87024095, -1.1686319 , -0.85298475, ..., -0.88483264,
        -0.94770261, -0.86629561],
       [ 1.36937906,  1.35896134,  2.01392524, ...,  1.52877489,
         2.043099  , -0.14984188]])
```

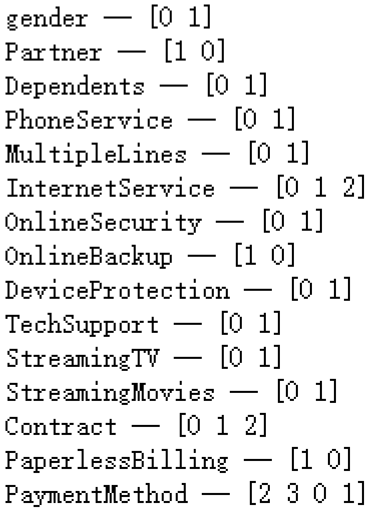
Figure 9: StandardScaler Process of Continue Features

Deal with discrete features, for example, *PaymentMethod: [bank transfer, credit card, electronic check, mailed check],* there is no numeric relationship between payment methods, generally use *one-hot encoding*, auxiliary use *integer encoding, and Label Encoding*. For *[Male, Female], [Yes, No, No internet service],* such kind of object data, does not have an obvious numerical type relationship, we use integer encoding to convert it to *[0,1,2]*. For other discrete features: if there is a correlation between features, then the numerical mapping is used, as shown in *Figure 10 and Figure 11*:



Index	gender	SeniorCitizen	Partner	Dependents
gender	1	-0.00187371	0.00180808	0.0105166
SeniorCitizen	-0.00187371	1	-0.0164787	-0.211185
Partner	0.00180808	-0.0164787	1	-0.452676
Dependents	0.0105166	-0.211185	-0.452676	1
tenure	-1.25316e-05	0.0108338	-0.101985	0.0485135
PhoneService	-0.00648767	0.0085764	-0.0177057	-0.00176168
MultipleLines	-0.00945053	0.113791	-0.117307	-0.0196568
InternetSer…	-0.000863382	-0.0323102	-0.000891347	0.0445904
OnlineSecur…	-0.00342855	-0.210897	-0.0818497	0.190523
OnlineBackup	0.0122295	-0.144828	0.0907534	0.0627745
DeviceProte…	0.00509166	-0.157095	-0.0944515	0.156439

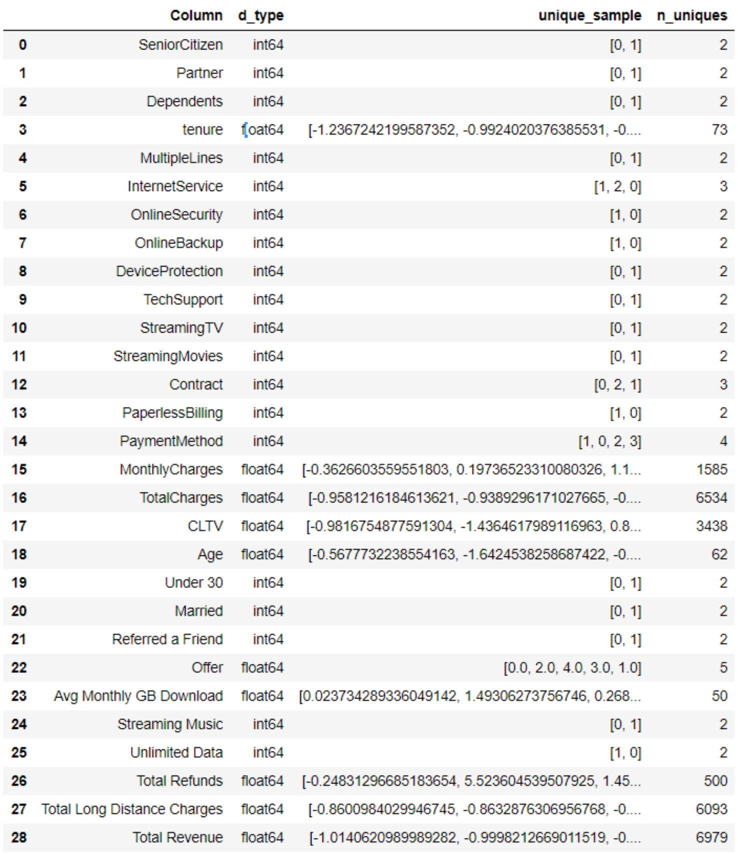
Figure 10: One-hot Encoding of Discrete Features



```
gender — [0 1]
Partner — [1 0]
Dependents — [0 1]
PhoneService — [0 1]
MultipleLines — [0 1]
InternetService — [0 1 2]
OnlineSecurity — [0 1]
OnlineBackup — [1 0]
DeviceProtection — [0 1]
TechSupport — [0 1]
StreamingTV — [0 1]
StreamingMovies — [0 1]
Contract — [0 1 2]
PaperlessBilling — [1 0]
PaymentMethod — [2 3 0 1]
```

Figure 11: Label Encoding of Discrete Features

After plotting each two 52 features, depend on the value of evaluation metrics. I decided to keep 29 customer feature is used for modeling, then completed data dimensionality reduction processing, as shown as *Figure 12:*



	Column	d_type	unique_sample	n_uniques
0	SeniorCitizen	int64	[0, 1]	2
1	Partner	int64	[0, 1]	2
2	Dependents	int64	[0, 1]	2
3	tenure	float64	[-1.2367242199587352, -0.9924020376385531, -0....	73
4	MultipleLines	int64	[0, 1]	2
5	InternetService	int64	[1, 2, 0]	3
6	OnlineSecurity	int64	[1, 0]	2
7	OnlineBackup	int64	[1, 0]	2
8	DeviceProtection	int64	[0, 1]	2
9	TechSupport	int64	[0, 1]	2
10	StreamingTV	int64	[0, 1]	2
11	StreamingMovies	int64	[0, 1]	2
12	Contract	int64	[0, 2, 1]	3
13	PaperlessBilling	int64	[1, 0]	2
14	PaymentMethod	int64	[1, 0, 2, 3]	4
15	MonthlyCharges	float64	[-0.3626603559551803, 0.19736523310080326, 1.1...	1585
16	TotalCharges	float64	[-0.9581216184613621, -0.9389296171027665, -0....	6534
17	CLTV	float64	[-0.9816754877591304, -1.4364617989116963, 0.8...	3438
18	Age	float64	[-0.5677732238554163, -1.6424538258687422, -0....	62
19	Under 30	int64	[0, 1]	2
20	Married	int64	[0, 1]	2
21	Referred a Friend	int64	[0, 1]	2
22	Offer	float64	[0.0, 2.0, 4.0, 3.0, 1.0]	5
23	Avg Monthly GB Download	float64	[0.023734289336049142, 1.49306273756746, 0.268...	50
24	Streaming Music	int64	[0, 1]	2
25	Unlimited Data	int64	[1, 0]	2
26	Total Refunds	float64	[-0.24831296685183654, 5.523604539507925, 1.45...	500
27	Total Long Distance Charges	float64	[-0.8600984029946745, -0.8632876306956768, -0....	6093
28	Total Revenue	float64	[-1.0140620989989282, -0.9998212669011519, -0....	6979

Figure 12: Sub-Dataset Used for Modeling

5.Prediction Model

5.1 Model Parameter Setting and Evaluation Metrics

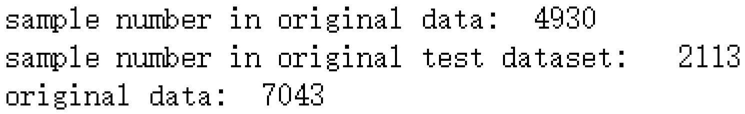
In this experiment, all the code is written in Python3.8 language, running in the jupyter notebook environment, using *sklearn, tensorflow, seabor, XGBoost, numpy, pandas,* and other public toolkits packages. The data set used in the experiment is the sub-dataset after the above operation and processing. It contains 7043 samples, with a total of 7043 rows and 29 columns. 23 features are deleted and 29 features are retained.

In order to eliminate the impact of unbalanced samples, as shown in *Figure 1*. Using the SMOTE method for oversampling [5], The basic idea of SMOTE is to randomly select a sample $x_{i(nn)}$ from its k neighbors for each minority sample x_i, ($x_{i(nn)}$ is a sample in the minority class), and then choose between x_i and $x_{i(nn)}$ Randomly select a point on the connection line as the newly synthesized minority sample, that is, synthesize a new sample $\hat{x}_i$ according to the following *Formula 1:*:

$$\hat{x}_i = x_i + \text{rand}(0,1) * (x_{i(nn)} - x_i)$$

Formula 1: Definition of SMOTE Method

Then, we divide the data set into a training set and a test set to achieve the purpose of cross-validation, where *Test_sise = 0.3*, which means 70% data is used for training the model, and 30% data is used for testing. As shown in *Figure 13:*



```
sample number in original data:  4930
sample number in original test dataset:  2113
original data:  7043
```

Figure 13: Partition Training and Testing data

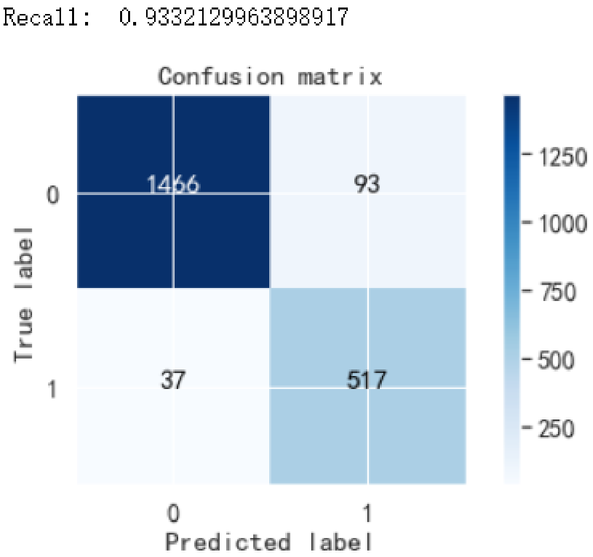
During the experiment, setting Logistic Regression regular punishment parameter *c = 1.05*, Result threshold (trained by the Sigmoid function, the probability of the final result is greater than how much we treat it as a positive or negative sample) *h = 0.1*

In terms of evaluation, choose Accuracy, Recall, Precision, F1-score, and confusion Matrix as the evaluation indicators of the model.

5.2 Performance of Logistic Regression

Predicting customer churn is a classic two-category model. We judge "customer churn" as negative and "customer retain" as positive. This article uses *logistic regression, KNN, ANN, random forest, decision tree, naïve Bayes, XGBoost, Catboost,* etc., such classification algorithms are used to establish prediction models. After the model training is completed, the features priority ranking can be output. Based on evaluation metrics, the performance of different customer churn prediction models is compared. The comparison results will be discussed in detail in the sixth part, *result and analysis*. In the below part, we explain the basic algorithm: the establishment and improvement of logistics regression.

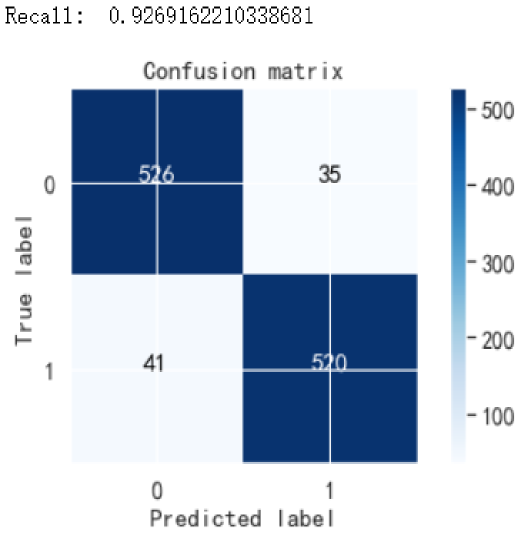
Implementation the data set and parameter settings of 5.1 above to train the model, call the logistic regression model, and get the test results as shown in *Figure 14:*



```
Recall:  0.9332129963898917
```

Figure 14: LR Confusion Matrix of Test Dataset

Use confusion matrix for visual display analysis. In the test dataset *Recall=0.9321*. Due to the model will eventually return to reality, it is necessary to evaluate the effect on the test set of the original dataset (30% test data set without SMOTE oversampling and downsampling), and get the result as shown in *Figure 15:*



```
Recall:  0.9269162210338681
```

Figure 15: LR Confusion Matrix of Original Dataset

Test on the original data, the *Recall*=0.9269, the results of the two datasets are relatively close, it can be determined that the data balancing algorithm based on SMOTE is effective, we can use test dataset to replace the original dataset in other classification algorithms, in order to save model training time.

When debugging the threshold *h*, we found the following rules, as shown in *Figure 16:*

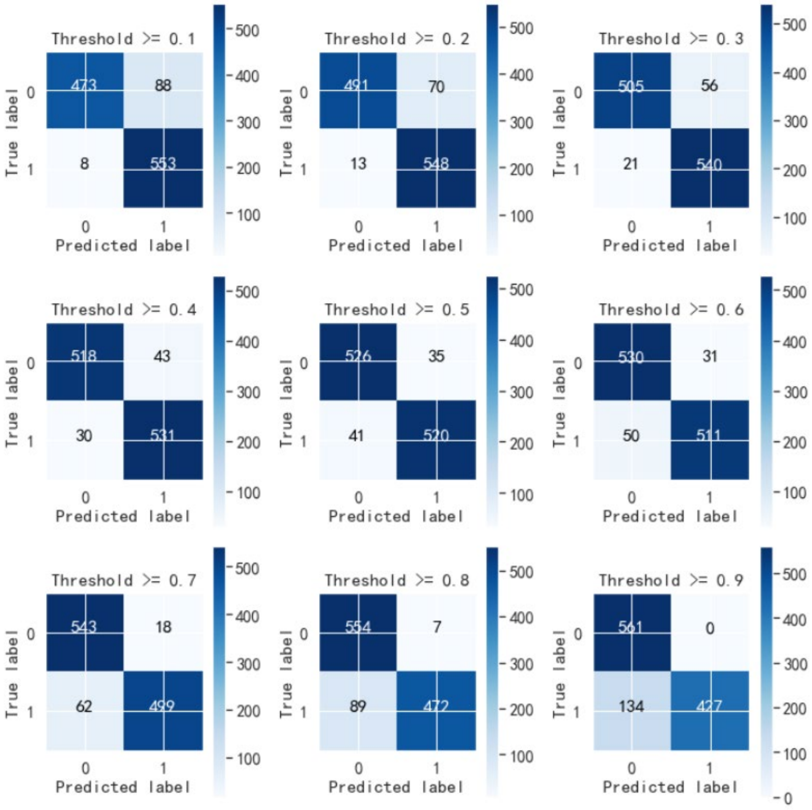


Figure 16: LR Confusion Matrix of Threshold Influence

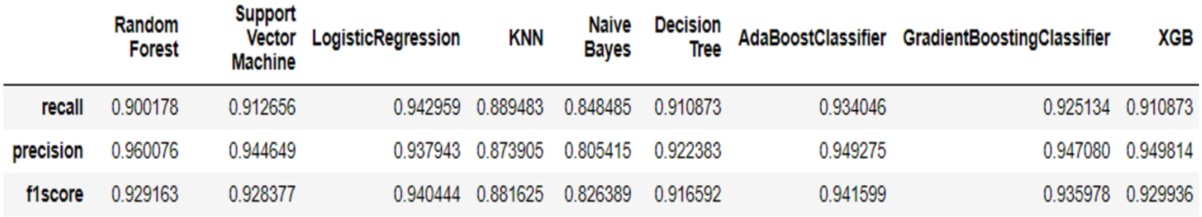
By comparing nine different thresholds, it is found that when the threshold is smaller, the recall value is higher, but the FP probability (False Predict) is also higher, which has little practical meaning. As the threshold increases, the recall value decreases, and the NP probability also decreases.

For example, when the threshold is equal to 0.1, 553 churn customers are correctly predicted (TN), 30 churn customers are not predicted (FP), but 88 normal users are mistaken as churn users (FN), the Recall= 0.9821, but the *Accuracy=0.54;* When the threshold is equal to 0.5, 520 churn customers are correctly predicted (TN), 41 churn customers are not predicted (FP), 35 normal users are mistaken as churn users (FN), the *Recall= 0.7950*, the *Accuracy= 0.78*, and the accuracy values increase.

Therefore, it can be concluded that in an ideal state, all churn users are predicted, so that we can perform refined operations on these customers in advance, effectively prevent customer churn and increase customer retention; at the same time, there is no customers are misjudged as churn one, reducing operating costs and improving user experience. However, it is difficult to have both. The specific threshold adjustment needs to be combined with the actual business, starting from the actual business, and selecting the optimal threshold. The Recall value does not completely represent the model performance.

6. Result and Analysis

This article uses *logistic regression, KNN, random forest, decision tree, naïve Bayes, XGBoost,* AdaBoostClassifier, GradientBoostingClassifier, etc., such classification algorithms are used to establish prediction models. After the model training is completed, the features priority ranking can be output. Based on evaluation metrics Recall, Precision, F1-score, the performance of different customer churn prediction models is compared, as shown as *Figure 17:*



	Random Forest	Support Vector Machine	LogisticRegression	KNN	Naive Bayes	Decision Tree	AdaBoostClassifier	GradientBoostingClassifier	XGB
recall	0.900178	0.912656	0.942959	0.889483	0.848485	0.910873	0.934046	0.925134	0.910873
precision	0.960076	0.944649	0.937943	0.873905	0.805415	0.922383	0.949275	0.947080	0.949814
f1score	0.929163	0.928377	0.940444	0.881625	0.826389	0.916592	0.941599	0.935978	0.929936

Figure 16: LR Confusion Matrix of Threshold Influence

Through the comparison of evaluation metrics, it can be seen that based on the data processing method of feature selection in this article, Logistic Regression has the highest Recall value 94.29%, and F1-score 94.04%. Random Forest has the highest Precision Value 96.00%. The

Accuracy of all prediction models are above 80.54% (obtained by the precision of naïve Bayes), which means that the above algorithms can basically be applied to the actual telecommunications customer churn prediction What's more? The error rate can be optimized by tuning (Logistic Regression regular punishment parameter c, threshold h). In general, using the feature selection method combined with XGBoost and Encoding in this article can make the customer churn prediction model achieve better performance.

According to the XGBoost algorithm, the special importance priority is shown in *Figure 17:*

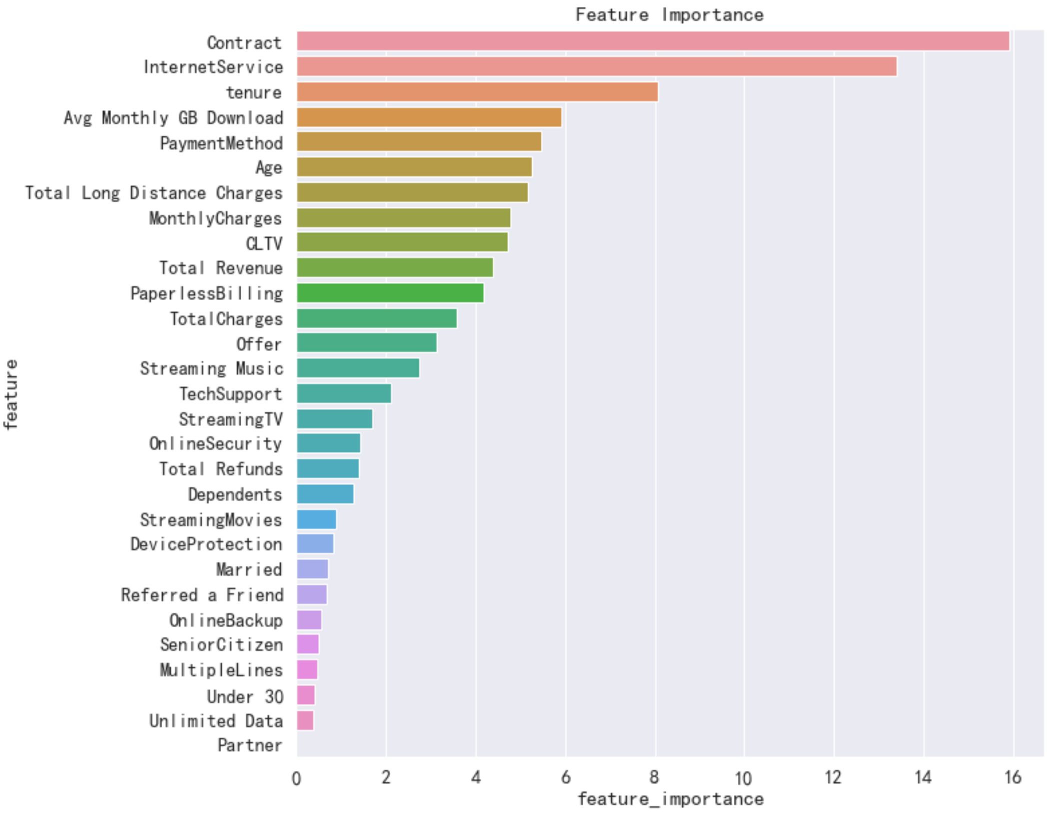


Figure 17: Feature Importance Priority

From the feature priority in the above figure, it can be concluded that Contract, Internet Service, and Tenure are the three most important factors. Among them, *contract: [Month-to-Month, One Year, Two Year]*, the customer who chooses the "two-year" contract is easier to retain, "month-to-month" users are more likely to churn. This is because the longer the contract period of the customer, the easier it is for the customer to have a sense of belonging to the company and form corporate identity, so it is not easy to change telecom. The importance of Internet Sercive

Sexuality is also in line with reality. The development of the Internet era and the popularization of 5G have caused users to have an increasing demand for the network. The quality of network services directly affects the user experience. In addition, homogeneous products in the market increasingly, if users feel that they need to improve their network experience, they will naturally change telecommunications service companies. For enterprises, this has resulted in customer churn. Tenure represents the impact of user income on target variables. Among them, the average monthly income of customers the lower the telecommunications service fee (monthly rent), the more likely it is to lose customers.

In summary, for new customers who join the telecom, the enterprise should give away value-added service experience qualifications at the time of purchase to enhance users’ awareness of value-added services; for existing old customers, according to user behavior characteristics, for potential users who have purchased this part of the service, focus on publicity and promotion to encourage users to make further purchases and strengthen users' stickiness to the product. Enhancing user experience in a variety of ways can achieve the purpose of reducing customer churn.

7. Conclusion

This article takes IBM Telecom’s customer churn data as an example, systematically analyzes the related factors of telecommunications customer churn, and elaborates the process of establishing a customer churn prediction model. The first part is the use of missing value completion, outlier processing, and correlation analysis, complete data preprocessing. After that, by a novel feature selection algorithm, which is a combination of One-hot encoding, label encoding, and XGBoost, based on the evaluation metric to select valid features and drop redundant features. Experimental results on the telecommunications customer churn data set Show that the method has good performance, and can also be applied to other classification fields for feature selection. Next step, bring the test data into 9 classification algorithms such as Logistics Regression, Random Forest, Decision Tree, etc., create prediction models, and compare performance. Then, result shows that both Logistics Regression and Random Forest have good predictive performance. But if the data sets are different, we need to readjust the parameters (such as the punishment function parameter C, the threshold h). Finally, we get the feature importance priority based on the XGBoost function. We can judge which features are more important for customer churn prediction.

The future work will focus on two aspects: First, using deep learning techniques, artificial neural networks (ANN), building models, and then comparing and optimizing existing prediction models. Second, due to the limitations of the sampling data set, the data set reflects Disorder, the next step is to use clustering methods such as KNN, K-Means, etc. to cluster the features with a high degree of correlation, and summarize the correlation rules, so as to achieve a better balance of outliers.

Reference

[1] Coussement, K., Lessmann, S., & Verstraeten, G. (2017). A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry. Decision Support Systems, 95, 27-36.

[2] Li, Y., Hou, B., Wu, Y., Zhao, D., Xie, A., & Zou, P. (2021). Giant fight: Customer churn prediction in traditional broadcast industry. Journal of Business Research, 131, 630-639. [3] Chu, P. C., & Beasley, J. E. (1998). A genetic algorithm for the multidimensional knapsack problem. Journal of heuristics, 4(1), 63-86.

[3] Ji Huijie, Ni Feng, Liu Jiang, Lu Qiling, Zhang Xuyang & Que Zhongli. (2021). Telecom customer churn prediction based on XGB-BFS feature selection algorithm. Computer Technology and Development (05), 21-25. doi:CNKI: SUN: WJFZ.0.2021-05-004.

[4] Jiao Gui'e & Xu Hong.(2021).Analysis and Comparison of Forecasting Algorithms for Telecom Customer Churn. Journal of Physics: Conference Series(3),. doi:10.1088/1742-6596/1881/3/032061.

[5] Cai Nan. (2020). Telecom customer churn prediction and analysis based on improved random forest algorithm (Master's thesis, Nanchang University). https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD202101&filename =1020053360.nh.

[6] Khan F,Kozat S S.Sequential churn prediction and analysis of Cellular network users-A multi-class,multi-label perspective [C].Signal Processing and Communications Applications Conference.IEEE,2019

[7] NATH S.V.Data Warehousing and Mining:Customer Churn Analysis in the Wireless Industry[D].Boca Raton,Florida,Florida Atlantic University,2003.