Prediction on the Churn through Classification

Ding, Ma
Boston University

Hank, Zhong
Boston University

Yifan, Zhang
Boston University

Jiaru, Li
Boston University

June 28, 2021

Abstract

Any business wants to maximize the number of customers. In this paper, we approach a model, based on a dataset from a telecommunications company, to predict the customer churn in order to create a chance for company to retain old clients. The accuracy of our model is close related to cost of the retaining, so in order to get higher accuracy we have to deal with two challenges: missing data and imbalanced data. For missing data, we use either replacing missing data with mean or directly deleting it, and the result demonstrate higher accuracy for deleting method. The major challenge comes from the imbalance, and in order to increase the accuracy, we implement four methods: Penalization on the target class, Decreasing the threshold, Oversampling, and Undersampling. According to the result, Undersampling has highest accuracy for the prediction of the customer churn, and it leads to more effective cost of companies.

1 Data Set Analysis

1.1 Missing Dataset

Before building a model, we analyze the dataset in order to find out the potential cause of bias. After factorizing categorical data and converting string-format real value to float format, we figure out that 10 users has missing data. To minimize the influence of missing data, we implement two method. First one is replacing missing data with mean, and second one is deleting all the missing values.

Table 1: Accuracy of the method for the missing data

Method	Churn	Precision	Recall	F1-score
Replacing	False (Not churn)	0.85	0.98	0.87
	True (Churn)	0.61	0.52	0.56
	average	0.79	0.79	0.79
Deleting	False (Not churn)	0.85	0.91	0.88
	True (Churn)	0.67	0.53	0.59
	average	0.80	0.81	0.80

Shown as Table 1, we use three values, precision, recall and F1-score to determine the accuracy for each level of two methods. Precision gives the ratio of correctly predicted positive observations

to the total predicted positive observations. Recall shows the ratio of correctly predicted positive observations to the all observations in actual class - churn. F1-score is the weighted average of Precision and Recall. From the Table 1, the result clearly demonstrates that when the missing value is not extremely large, deletion is the better way to deal with the missing data.

1.2 Imbalanced Data

The other challenge in the dataset is the imbalanced data. Shown as Figure 1, we have more data on stable customers than leaving customers. This may lead our model to classify not leaving customers more frequently since there are more data to be trained than leaving customers, and it might cause the extra loss of company. Assuming a customer has high probability to leave the service(churn) ,but if our model's prediction excludes him or her from the customer churn which is False Negative. The company will not focus on that customer's services, then this customer will definitely leave the services(churn), which results a customer loss of the company. In the other situation, if a customer is not willing to leave the services and is satisfied, but our model gives the opposite result, then the company will provide this customer extra benefits. This causes the wasted investment of the company. Therefore, other than focusing on the increment of recall or precision, we will use F1-score as the evaluation of the model.

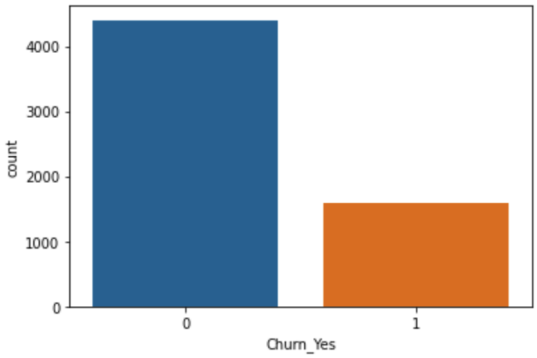

Figure 1: Imbalanced of the dataset

2 Models

2.1 Normal Logistic Regression

Since our numerical data have various range and unknown distribution, before training the model, we implement normalization with the function:

$$x = \frac{x - \bar{x}}{\sigma},$$

in which $\bar{x}$ is the mean and σ is the standard deviation.

Then we apply the dataset in to logistic regression, and the function return the result as Table 2. The accuracy of the model is 0.81, which seems like to a satisfied result. However, when we check the accuracy and recall of different classes, churn and not churn, the result shows that both values of determining the leaving customer are lower than 0.7. Therefore, we need to increase the model's False Negative rate while maintaining the overall performance to prevent the loss of companies mentioned in previous section.

Table 2: Accuracy of the normal Logistic Regression

Churn	Precision	Recall	F1-score
False (Not churn)	0.85	0.91	0.88
True (Churn)	0.67	0.53	0.59
average	0.80	0.81	0.80

2.2 L2 Penalization

Due to the imbalanced property of our dataset, the training model appears to have higher frequency on classifying 0 (not churn). Therefore, we try to add L2 penalization on target class so that the classifier are forced to classify more 1s (churn) than 0, in order to increase the recall of classifying 1. The error function with L2 penalization is

$$E(w) = -[y * ln(\hat{y}) + (l-y) * ln(1-\hat{y})] + \lambda * (1-y)^2,$$

where λ represents for penalty. According to this error function, we derive new back propagation equations for weights and bias, such as

$$\begin{aligned}
\frac{\partial E(w)}{\partial \hat{y}} &= -[\frac{y}{\hat{y}} - \frac{1-y}{\text{ffl}1-\hat{y}}] - 2\lambda * (1-\hat{y}) \\
\frac{\partial E(w)}{\partial w} &= \frac{\partial E(w)}{\partial \hat{y}} * \frac{\partial \hat{y}}{\partial w} \\
&= [-[\frac{y}{\hat{y}} - \frac{1-y}{1-\hat{y}}] - 2\lambda * (1-\hat{y} * [\frac{\partial}{\partial w}(\sigma(w*x))] \\
&= [\hat{y} - y - 2\lambda * \hat{y} * (1-\hat{y})^2] * x^T
\end{aligned}$$

After we use this back propagation formula to update the original logistic regression model, the result changes as we want. In order to improve the model, we try different punishment intensity and form a graph according to the corresponding values of model's True recall and True F1-score.

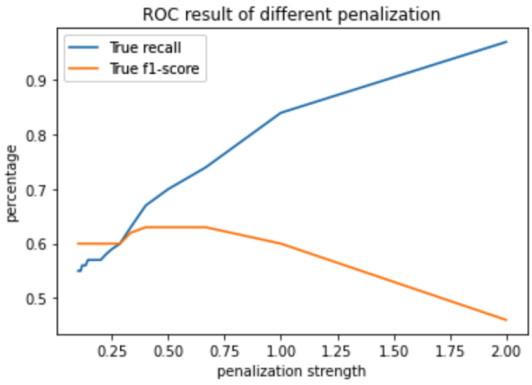

Figure 2: ROC of different penalty strength

Shown by Figure 2, the best choice of penalty strength should be 0.666, so we determine λ value in our previous error function to be 0.666, and recalculate the accuracy of the model, which is shown in Table 3.

Table 3: Accuracy of L2 penalization on the target class

Churn	Precision	Recall	F1-score
False (Not churn)	0.90	0.91	0.88
True (Churn)	0.54	0.74	0.63
average	0.80	0.77	0.78

2.3 Oversampling and Undersampling

2.3.1 Oversampling

The other idea of minimize the bias caused by the imbalanced dataset is to manually change the proportion of different classes through either Oversampling or Undersampling.

Oversampling is usually achieved by repeatedly sampling the minority class, which is the churn class in our case, but simple repetition is likely to lead to overfitting of the model. Therefore, we use the SMOTE, Synthetic Minority Oversampling Techniques, as an improvement, which can effectively reduce overfitting by synthesizing new samples with fewer categories through k-nearest neighbors in the local area. After we implement SMOTE, sizes of two class, churn and not churn, become the same, which is shown as Figure 3.

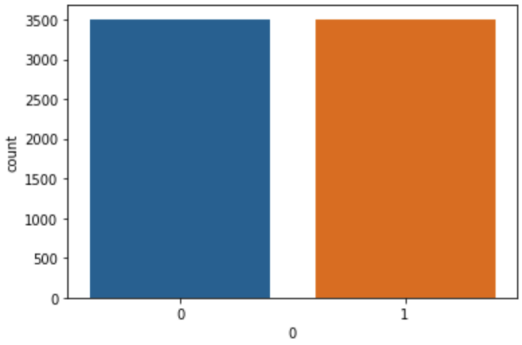

Figure 3: Oversampling

2.3.2 Undersampling

Undersampling is to balance the sample size of different categories by randomly deleting samples from majority class. However, the number of samples in this dataset is not very large. If too many samples are discarded, the recognition accuracy of the model for the categories with more samples will decline. So we plan to use multiple under sampling, then train multiple models, and finally use voting to determine the final prediction results. After implementing Undersampling, shown as Figure 4, the size of majority class decrease to the same level as the minority class.

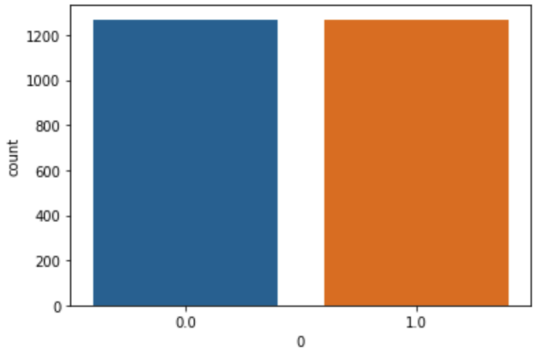

Figure 4: Undersampling

The accuracy table of undersampling and oversampling is shown as Table 4, and it shows that even undersampling uses a smaller proportion of data, the result is better than oversampling. One reason of this is that oversampling more or less causes overfitting which may affect the accuracy of the model when validation uses the test set.

Table 4: Accuracy of the method for the missing data

Method	Churn	Precision	Recall	F1-score
Oversampling	False (Not churn)	0.91	0.76	0.83
	True (Churn)	0.54	0.79	0.64
	average	0.81	0.77	0.78
Undersampling	False (Not churn)	0.91	0.78	0.84
	True (Churn)	0.56	0.78	0.66
	average	0.82	0.78	0.79

2.4 Decreasing Threshold

Another way of minimizing the bias from the imbalanced data is simply changing the threshold when predicting using the output of the ordinary logistic regression, as shown in the following function.

$$class = \begin{cases} 0 & \hat{y} > threshold \\ 1 & \hat{y} \leq threshold \end{cases}$$

Shown as Figure 5, when the threshold value roughly around 0.3 to 0.4, True recall and True F1-score is balanced. After we multiple attempts on changing threshold between this interval, 0.333 gives our best feedback on the accuracy, shown as Table 5.

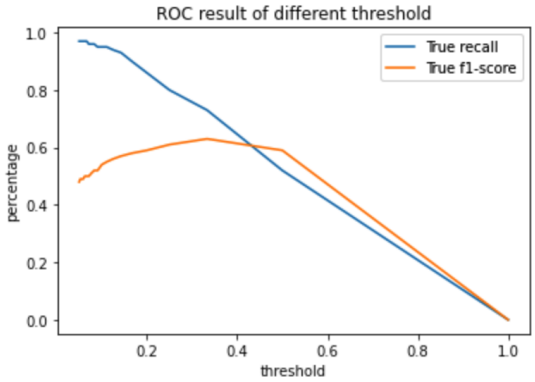

Figure 5: ROC of threshold

Table 5: Accuracy of threshold value 0.333

Churn	Precision	Recall	F1-score
False (Not churn)	0.89	0.79	0.84
True (Churn)	0.55	0.73	0.63
average	0.80	0.77	0.78

3 Conclusion

After we implement four methods to overcome the major challenge, the imbalance of the data, we use value of recall and F1-score as the evaluation standard. From Table 6, the penalization method has the maximum False F1-score, minimum variance, and close values on other measurements. In comparison with changing threshold when predicting, adding penalization is more stable when changing the strength of penalization. Oversampling and undersampling show good performance in predicting minority class but because of manually change the distribution of original data. This method has greatly decrease the accuracy of predicting the majority class from 0.88 down to 0.83. However, as we state above, we do not care much about the majority class, so to sum up, oversampling and undersampling are best methods for this challenge. Undersampling acchieves higher True F1-score than oversampling, it may because the given data is large enough for predicting True label even though it abandon most of the majority class's data. while oversampling has good performance when the number of samples is not enough.

Table 6: The accuracy of all four methods

Method	True recall	True F1-score	False F1-score	Model F1-score	variance
Original	0.53	0.59	**0.88**	**0.80**	
Penalization	0.74	0.63	**0.88**	0.78	**0.03366**
Oversampling	**0.79**	0.64	0.83	0.78	
Undersampling	0.78	**0.66**	0.84	0.79	
Threshold	0.73	0.63	0.83	0.78	0.12487