Machine Learning-based prediction on customer churn

By Violet Yan

AUTHOR BIO

Violet Yan is a rising sophomore at Langley High School in McLean, Virginia. She has been interested in applied math, computer science, business and economics. She takes courses in advanced Python, artificial intelligence and machine learning outside of school and hopes to study multidisciplinary fields in the future.

ABSTRACT

Machine learning is a very powerful tool for studying data and creating mathematical models to solve real-world problems. This study seeks to make customer churn prediction in telecommunication business using machine learning modeling methods. Customer churn prediction plays a critical role in business management and development in the highly competitive telecommunication business. A telecommunication customer dataset is used to train a sequential machine learning model with several stacks of layers, which predicts customer churn. A 4-fold cross validation method, instead of the traditional data validation method, has been conducted to fully utilize the dataset. It is found that the prediction accuracy for customer retention is around 90%, which is very high, while the prediction accuracy for customer churn is around 60%, which is not ideal. The overall accuracy of the model is 0.80. Several factors, including technical support, are deemed important to predict customer churn, which could be used by telecommunication business executives to develop better customer retention strategies.

Keywords: *machine learning, customer churn, telecommunication, logistic regression, customer retention, mathematical modeling, statistics, churn prediction*

INTRODUCTION

According to an online survey by the German company Statistica that was conducted in August 2020 among 1,100 customers in the United States, the industry in the United States where customers were most likely to leave their current provider due to poor customer service was cable television, with a 25% churn in 2020, while Telecom/Wireless companies has a 21% of churn (Bohne, 2024). Churn, sometimes called churn rate, or attrition rate, is the percentage of customers that stop utilizing a service within a given time period. It is often used to measure businesses which have a contractual customer base, especially subscriber-based service models. A 21% churn means there's one in five customers leaving annually.

In almost any business, it is more cost-effective to retain an existing customer than acquiring a new one. Acquisition of a new customer usually requires substantial spending on marketing, promotion and incentives. According to HubSpot's 2024 Annual State of Customer Service Report, customer acquisition is 5 times more expensive than customer retention (El-Abidin, 2023). Therefore, telecom companies nowadays pay lots of attention to customer retention or reducing churn. Retention focuses on nurturing established relationships. By reducing churn, the company can allocate resources more efficiently, improving its overall financial health (Camp, 2024).

As a result, it is very crucial for the telecom business to predict whether a customer will leave in a given period, which will help the company develop better customer retention strategies and improve services to meet customer expectations. There could be many reasons for customer churn, including competitors offering a lower price, bad customer service, disappointing experience, etc.

The purpose of this paper is to use a machine learning model to predict telecom business customer churn to provide insights into the company's customer retention strategies.

METHOD

Dataset

The dataset contains basic Telco customer information, as well as an indicator, showing whether the customer left Telco in the past months (Kaggle, 2018). The goal is to build a machine learning model to predict, given the basic customer information, whether this customer will switch in the next month. Table 1 below shows a sample of the Telco dataset.

Table 1

Dataset Sample - the first 6 rows.

customer ID	gender	Senior Citizen	Partner	Dependents	tenure	Phone Service	Multiple Lines	Internet Service	Online Security	Online Backup	Device Protection	Tech Support	Streaming TV	Streaming Movies	Contract	Paperless Billing	Payment Method	Monthly Charges	Total Charges	Churn
7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No	Yes	No	No	No	No	Month-to-month	Yes	Electronic check	29.85	29.85	No
5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	No	Yes	No	No	No	One year	No	Mailed check	56.95	1889.5	No
3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	Yes	No	No	No	No	Month-to-month	Yes	Mailed check	53.85	108.15	Yes
7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes	No	Yes	Yes	No	No	One year	No	Bank transfer (automatic)	42.3	1840.75	No
9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No	No	No	No	No	No	Month-to-month	Yes	Electronic check	70.7	151.65	Yes
9305-CDSKC	Female	0	No	No	8	Yes	Yes	Fiber optic	No	No	Yes	No	Yes	Yes	Month-to-month	Yes	Electronic check	99.65	820.5	Yes

This Telco customer dataset contains 7,043 rows and 21 columns in total. Each row contains information about each customer, including a unique customer id, gender, etc. The numerical data provided for each customer includes the monthly charges, total paid amount, and the number of months that the customer has been with Telco. The rest of the data columns have all categorical information, such as payment method, family plan, etc. There are a total of 20 customer features. The last column indicates whether each customer has churned or not in the past month, and the goal of this research project is to predict the possibility of customer churn.

When examining the dataset, 11 new customers were discovered. Since none of them had churn value in the last column, those data points of new customers were removed. The rest of the dataset was used for model training and validation which will be elaborated in the Machine Learning Model section below. The numerical variables in the dataset include monthly charges (Figure 1 below), total charges (Figure 2 below), and tenure, which denotes the number of months with Telco. The total charges don't always equal the product of the tenure and the monthly charges, especially for customers with a higher tenure. This could have been caused by the changing monthly rates the customer experienced with Telco.

Figure 1
Histogram of Monthly Charges

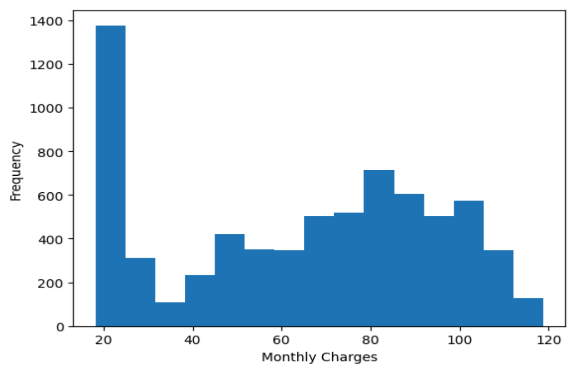
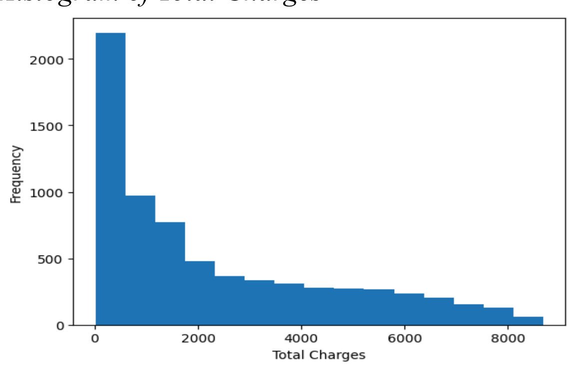


Figure 2
Histogram of Total Charges



All three numerical variables are scaled and normalized into new variables with mean 0 and variance 1, before feeding them into the machine learning model. The following two figures show the histogram of the normalized monthly charges and the normalized total charges.

Figure 3
Normalized Numerical Variable - Monthly Charges

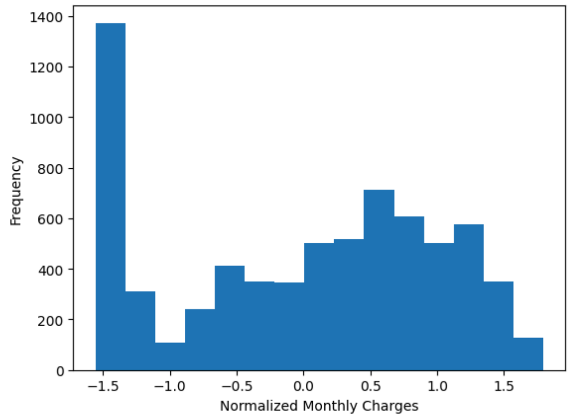
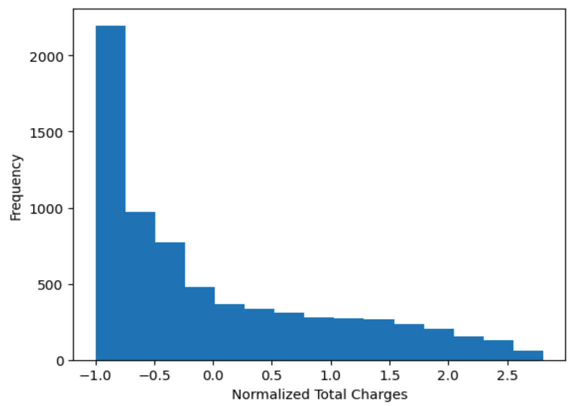


Figure 4
Normalized Numerical Variable - Total Charges

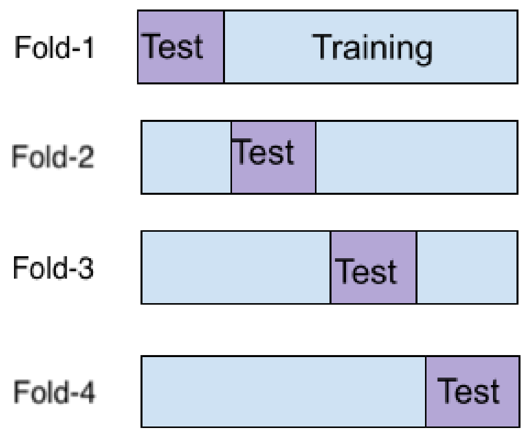


All other categorical variables, assigned as 0 and 1, were also normalized into new variables with mean 0 and variance 1, before feeding them into the machine learning model. For example, the variable "Paperless Billing", assigned 0 for no and 1 for yes, became -1.21 and 0.83 respectively after normalization.

MACHINE LEARNING MODEL

The normalized dataset was then split into 4 subsets, with 1,758 customers (25%) in each subset. 3 folds of the 4 subsets were used as the training dataset while the remaining 1 fold of the 4 subsets was used as the validation dataset. The training dataset was used to train the machine learning algorithm to predict the churn rate of Telco customers, and the validation dataset was to evaluate the effectiveness of the model. In this case, the iteration was repeated for 4 times and each time different subsets of the dataset were used for training and validation. Figure 5 below illustrates the 4-fold cross-validation process:

Figure 5
An Illustration the 4-fold Cross-validation used in the Machine Learning Model



Note. Fold-1 corresponds to using the first fold (25%) of data for validation; Fold-2 corresponds to using the second 25% of data for validation, and so on.

The machine learning model implemented involves several layers of activation functions. The logistic regression model is used as the first layer to process input variables, before feeding them to the subsequent layers. Logistic regression is a statistical method which predicts the likelihood of an event based on the input variables. It is based on the logistic function, also known as the sigmoid function, which is expressed as follows:

$$f(x) = \frac{1}{1+e^{-x}}$$

The rectified linear unit activation function ("ReLu"), a neural network model, is used in the subsequent layers. The binary cross-entropy function is used as a loss function to calibrate parameters in each machine learning layer, which is given by:

$$\text{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[y_i ln(p_i) + (1 - y_i)ln(1 - p_i)]$$

For the current machine learning algorithm, two layers of ReLu activation functions for a total of three linear layers, are implemented. Adding an additional ReLu layer doesn't improve any prediction accuracy but only adds computational time.

RESULTS

Different variables could have different impacts on the machine learning algorithm. Those important features are kept and used in the machine learning algorithm. For example, including or excluding the category variable of "Senior Citizen" did not change the prediction results or speed up the convergence. However, including the category variable, such as "Technical Support", did improve the precision, recall, and the f1-score (discussed in more detail below) for predicting the non-churn rate. Table 2 below lists the data summary on "Technical Support" and "Churn" columns. The table shows those internet customers with technical support are less likely to leave this telco company compared to those internet customers who don't have the technical support.

Table 2

Summary of Customers with Technical Support and Churn

Technical Support	**Yes Churn**	**No Churn**
Yes	16.59%	33.51%
No	77.37%	39.24%
No internet	6.05%	27.25%
Total	100.00%	100.00%

Table 3 below lists the data summary on “Paperless Billing” and “Churn” columns. The table shows those customers with paperless billing are more likely to leave this telco company compared to those customers who receive the billings in the mail. This is one important categorical feature that was included in the machine learning model.

Table 3
Summary of Customers with Paperless Billing and Churn

Paperless Billing	**Yes Churn**	**No Churn**
Yes	74.9%	53.6%
No	25.1%	46.4%
Total	100.00%	100.00%

Table 4 below lists the data summary on “Payment Method” and “Churn” columns. The table shows those customers who pay their bills using electronic checks are more likely to leave this telco company compared to those who pay using other methods. A new categorical feature “AutoPay” was created for those customers with the payment methods “Credit Card (automatic)” and “Bank Transfer (automatic)”; and “ElectronicCheck” for those customers with the payment methods “Electronic Check”. Both features were used in the machine learning algorithm.

Table 4
Summary of Payment Methods and Churn

Payment Methods	**Yes Churn**	**No Churn**
Mail Check	16.5%	25.1%
Electronic Check	57.3%	25.1%
Credit Card (automatic)	12.4%	25.0%
Bank Transfer (automatic)	13.8%	24.9%
Total	100.00%	100.00%

Table 5 below lists the data summary on “Internet Service” and “Churn” columns. The table shows those customers with Fiber optic internet are more likely to leave this telco company, and those customers without internet are less likely to leave. Two important categorical features, with/without internet and with/without fiber were used in the machine learning model.

Table 5

Summary of Internet Service and Churn

Internet Service	Yes Churn	No Churn
DSL	24.6%	37.9%
Fiber optic	69.4%	34.8%
No	6.0%	27.3%
Total	100.00%	100.00%

When training the machine learning code on different neural layers and input parameters, one needs to determine both the model error and the prediction accuracy. The model error is calculated by the binary cross-entropy function defined above, and the prediction accuracy can be shown by three numbers: precision, recall and f1-score. The precision, the recall and f1-score are numbers between 0 and 1, and defined as following:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$\text{f1-score} = 2\times\frac{Precision\times Recall}{Precision+Recall}$$

The f1-score, depending on both the precision and recall, is the harmonic mean of both numbers. The higher the f1-score, the higher the accuracy of the machine learning model predicting this classification. The f1-score was calculated in the classification report, so one can test the impact of different categorical variables. A total of 8 categorical variables were used in the machine learning model, including “TechSupport”, “PaperlessBilling”, “AutoPay”, “ElectronicCheck”, “Internet”, “Fiber”, “PhoneService”, and “MultipleLines”. All of them are tested separately and can either improve the prediction accuracy and/or speed up the convergence of the machine learning algorithm.

The training set classification report from fold-1 shows that the precision of the model is 0.84 on predicting correctly non-churn customers (classification = 0) over all predicted non-churn ones, and the recall 0.92 on predicting correctly non-churn customers over all non-churn customers. The f1-score 0.88 is the harmonic mean of precision and recall numbers. However, the f1-score of 0.57 for predicting churn customers (classification = 1) is not ideal.

Table 6

Training Set Classification Report from Fold-1

churn indicator	precision	recall	f1-score	support
0	0.84	0.92	0.88	3900
1	0.68	0.49	0.57	1374
accuracy			0.81	5274
macro avg	0.76	0.71	0.72	5274

weighted avg	0.80	0.81	0.80	5274

Model validation has also been executed by checking the validation dataset in the trained machine learning model. The validation set classification report for fold-1 shows the precision is 0.82 on predicting correctly non-churn customers (classification = 0) over all predicted non-churn ones; the recall 0.92 on predicting correctly non-churn customers over all non-churn customers. The f1-score 0.87 is the harmonic mean of precision and recall numbers. However, the f-1 score of 0.57 for predicting churn customers (classification = 1) is not ideal. The overall accuracy of the model using fold-1 is 0.80.

Table 7
Validation Set Classification Report from Fold-1

churn indicator	**precision**	**recall**	**f1-score**	**support**
0	0.82	0.92	0.87	1263
1	0.70	0.49	0.57	495
accuracy			0.80	1758
macro avg	0.76	0.70	0.72	1758
weighted avg	0.79	0.80	0.78	1758

Table 8 shows the model accuracy using the 4-fold cross validation method. The accuracy using each fold and the mean overall accuracy are all 0.80. This machine learning model produces consistent accuracy results.

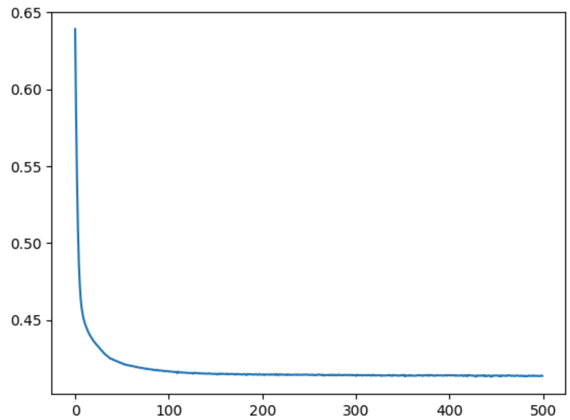
Table 8
4-fold Cross Validation Accuracy Results

Validation Dataset	**Accuracy**
Fold 1	0.80
Fold 2	0.80
Fold 3	0.80
Fold 4	0.80
Mean Accuracy	0.80

The error of the machine learning calibration, calculated by the binary cross-entropy function is shown in Figure 6, which decreases quickly and is stabilized after 100 iterations.

Figure 6

Error of the Machine Learning Model (y-axis) with Number of Iterations (x-axis) for Fold-1



CONCLUSION

Telecommunication companies face fierce competition nowadays to attract and retain customers. Reducing churn and improving customer retention plays an important role in telecommunication business. The machine learning model used in this research has an overall accuracy of 80%. It has an accuracy of about 60% for predicting customer churn, which is not ideal. It is possible that the dataset is not extensive enough. On the other hand, the model shows a high accuracy of about 90% in predicting customer retention. Several categorical factors have been identified as having a big impact on churn rate in this machine learning model. For example, technical support is an important factor in this model, which suggests that telecommunication companies can improve customer retention by investing in the technical support department. However, further investigation is needed to fully understand how these categorical variables could be used to provide insight for company retention strategies.

In order to improve the prediction model for customer churn, a more comprehensive dataset with more customer features may be needed. It may be helpful to include more variables such as customer age or monthly income.

Not only in the telecommunication business, such a machine learning model could be used in other business fields such as the cable television industry, which has the highest churn rate based on the survey mentioned in the introduction. However, collecting the right dataset is the key step to producing a good machine learning model.

REFERENCES

Bohne, R. (2024, November 9). *Customer service: Churn rate by industry U.S.* Statista. https://www.statista.com/statistics/816735/customer-churn-rate-by-industry-us/

Camp, M. (2024, July 16). *Ultimate Guide: Building Customer Retention in the telecoms industry*. https://blog.propellocloud.com/customer-retention-in-the-telecoms-industry

El-Abidin, R. (2023, February 16). *50 customer retention statistics to know*. HubSpot Blog. https://blog.hubspot.com/service/statistics-on-customer-retention

IBM. (2018, February 23). *Telco customer churn*. Kaggle. https://www.kaggle.com/datasets/blastchar/telco-customer-churn