CSE475 Project Report

2018-1-60-265, 2018-1-60-164, 2018-1-60-192

September 2021

1 Introduction

Customer churn refers to the phenomenon where customers related to a business no longer purchase from or interact with the business. Customer churning mainly relates to dissatisfaction with customer service and support systems. Customer churning may happen from several misunderstandings or even from one bad incident/ service. In this project, we will analyze the data of any company to predict the nature of customer churning. The reason and the most crucial time for customer relation improvement is the work which will be done in this project. We will be using ANN (Artificial Neural Network) to analyze the whole customer dataset by backpropagating the method and some activation functions accordingly to accurately find the test results and predict customer churn.

1.1 Objectives

Our main objective of the project is to identify customer churning from a dataset based on the customer behavior of a particular company so that they can be notified about the customer churn before it happens and accomplish the below objectives to retain their customers.

- Find the solution for customer retention.
- To improve the relationship with the customer according to their need and expectation.
- To find out the nature of customer churning.
- Using previous data, we will be able to see if any customer is going to leave the company.
- Increase the chance of holding into the existing customer with less investment.

1.2 Motivation

It takes more effort and investment to target and gets a new customer. When things don't go right existing customers may leave the company which can affect the company in long run. To improve the business strategy predicting the customer churn is our ultimate reason. After a thorough analysis of the customer behavior, if we can predict the customer have an intention to leave the company, the company can provide them with some offers so that they can retain the customer.

1.3 Existing Works

Numerous neural network algorithm has been used in this sector to find pretty much accurate result or to gain high accuracy. All the work is available on the internet. We have also added some of the related paperwork in our project proposal slide. The neural network solution for this problem is added again in this section.

- **Customer Churn Prediction using Naïve Bayes:** The implementation of data processing techniques using K-Means and Equal-Width Discretization (EWD) combined with Naïve Bayes are performed respectively to conduct a comparison of techniques to identify probable churn activities. Usually, the data generated are of massive size and with high dimensionality. To accommodate fast processing, casual heuristics is a preferred deployment [1]

- **Customer Churn Prediction using ANN:** Artificial neural network (ANN) approach for prediction of customers intending to switch over to other operators. This model works on multiple attributes like demographic data, billing information and usage patterns from a telecom company's data set [2]

- **Customer Churn Prediction using RNN:** Customer Lifetime Value (CLV) is the monetary value of a customer relationship. No change in CLV for a given customer indicates a decrease in loyalty. Recurrent Neural Network identifies churners based on Customer Lifetime Value time-series regression. The use of the K-means algorithm as a replacement to a rule-extraction algorithm. The K-means algorithm contributed to a more comprehensive analytical context regarding churn prediction.[3]

- **Customer Churn Prediction using CNN:** A layered architecture that comprises two different phases that are data load and preprocessing layer and a 2-D CNN layer is used to predict customer churn. In addition, the Apache Spark parallel and distributed framework are used to process the data in a parallel environment[4]

- **Customer Churn Prediction using Random Forest:**

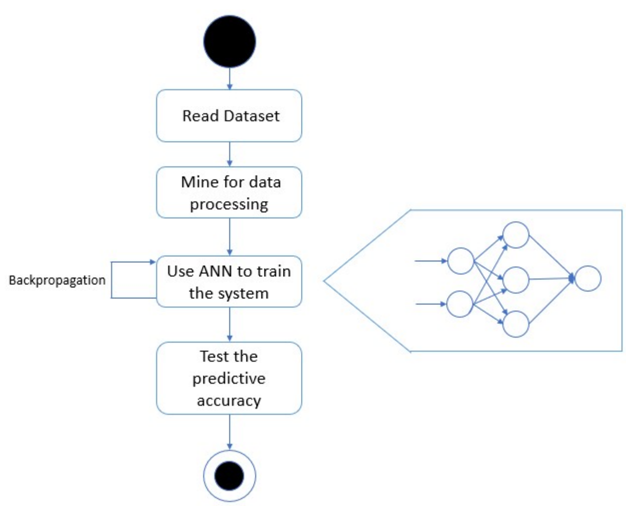
  Business analysts and customer relationship management (CRM) analyzers need to know the reasons for churn customers, as well as behavior patterns from the existing churn customers' data. A churn prediction model that uses classification, as well as clustering techniques to identify the churn customers and provides the factors behind the churning of customers in the telecom sector. Feature selection is performed by using information gain and correlation attribute ranking filter [5].

1.4 Necessity

For a business acquiring customers is the main motto but gaining customers is a really hard task for a company. A company can be successful by either acquiring customers every day or keeping the existing customers with the business for a long time. As mentioned before acquiring a customer is really hard work but keeping them satisfied to stay with the company is the best way of achieving success. People may leave the company or don't want to interact with the company anymore due to some incident. So, to prevent them from leaving the company and save some investment to gain new customers, it is necessary to learn from previous data. Only if we can predict if any potential customer will think about leaving the company, the outcome of the profit or face of the company may differ.

2 Methodology

We have used ANN (Artificial Neural Network) to predict customer churning because according to our knowledge ANN can self-learn itself to reduce the errors using backpropagation. So, our system can be well trained using this method.



3 Implementation

We have collected the data from a website and then data has been processed accordingly to use for predicting customer churn. All the process is described below.

3.1 Data collection

We have executed our project based on the telco customer churn dataset which was acquired from Kaggle.

http://www.kaggle.com/blastchar/telco-customer-churn?fbclid=IwAR3XZ1bqW3MATgwfnKAwUtU2SA
f5tfnODyqK5F9MZsLQ5tDJNLKfb8/

Properties of the dataset are described in this section:

- **CustomerID :** The unique id is provided to the customer to identify them uniquely.
- **Gender :** Gender of the customer.
- **SeniorCitizen:** If the customer is a senior citizen or not.
- **Partner:** If defines that, if the customer has a partner or not
- **Dependents:** It defines if the customer is dependent or not
- **tenure :**Tenure defines how long the customer is connected to the organization
- **PhoneService:**It defines if the customer has phone services or not.
- **MultipleLines:**It defines if the customer has multiple lines for phone services
- **InternetService:**It defines what type of internet connectivity service the user is currently taking.
- **OnlineSecurity :**It defines if the customer is under online security service or not.
- **OnlineBackup :**It defines if the customer is under online backup service or not.
- **DeviceProtection:**It defines if the customer is under device protection service or not.
- **TechSupport:**It defines if the customer is under tech support service or not.

- **StreamingTV :**It defines if the customer is under Streaming TV service or not.
- **StreamingMovies:**It defines if the customer is under Streaming Movie service or not.
- **Contracts:** It defines what type of contract the company has with the customer.
- **PaperlessBilling:**It defines that if the customer pays using an online transaction or not.
- **PaymentMethod:**It defines what type of payment process the customer uses to pay his/her bills
- **MonthlyCharges:**it defines how much the company charges the customer every month for their services
- **TotalCharges:**It indicates how much the customer has been charged by the company after joining.
- **Churn:**It defines if the customer has fled or not.

3.2 Data processing

- As customerID is not important for our project, we have dropped this column.
- In TotalCharges there were some null values as the number of columns was significantly less, we decided to drop them.
- We have found out that the TotalCharges of our dataset was a string. So, we have changed it to a numeric type that is float.
- In MultipleLines we have found three attributes (yes, no, no phone service). No phone service is similar to "no" for this column. So, we have replaced no phone service with "no".
- In OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies we can find three attributes (yes, no, no internet service). No internet service is similar to "no" for this column. So, we have replaced no phone service with "no".
- In Partner, Dependents, PhoneService, MultipleLines, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, PaperlessBilling, Churn we have two attributes either yes or no. so, we have replaced yes with 1 and no with 0.

- In InternetService, Contract, PaymentMethod we have different types of attributes that cannot be defined using numeric values. So, we have replaced each of the columns with new columns containing the different types. Example – For InternetService now we have InternetService-DSL,InternetService-Fiber optic, InternetService-No.

- In tenure, MonthlyCharges, TotalCharges we have a large number of numerical values which are continuous and have no limits. So, we have converted them between 0 to 1.

3.3 Model Development

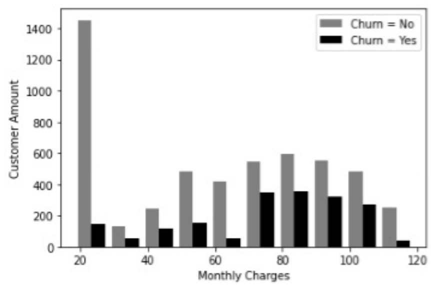
While developing our model we have used anaconda software to use Jupyter notebook for our python coding environment. We have used some library functions such as 'pandas', 'TensorFlow, 'NumPy', 'SKLearn', 'seaborn'.

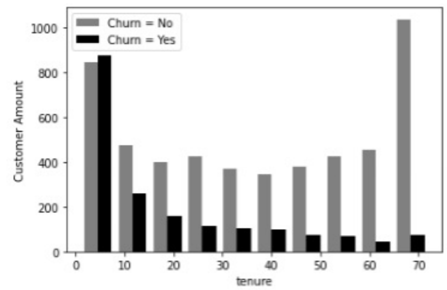
3.4 Result

```
customerID          object
gender              object
SeniorCitizen        int64
Partner             object
Dependents          object
tenure               int64
PhoneService        object
MultipleLines       object
InternetService     object
OnlineSecurity      object
OnlineBackup        object
DeviceProtection    object
TechSupport         object
StreamingTV         object
StreamingMovies     object
Contract            object
PaperlessBilling    object
PaymentMethod       object
MonthlyCharges     float64
TotalCharges        object
Churn               object
dtype: object
```

```
gender              object
SeniorCitizen        int64
Partner             object
Dependents          object
tenure               int64
PhoneService        object
MultipleLines       object
InternetService     object
OnlineSecurity      object
OnlineBackup        object
DeviceProtection    object
TechSupport         object
StreamingTV         object
StreamingMovies     object
Contract            object
PaperlessBilling    object
PaymentMethod       object
MonthlyCharges     float64
TotalCharges       float64
Churn               object
dtype: object
```

In the above images, we can see the conversion of TotalCharges from object or string to float.





Here, we have identified the relationship between tenure and customer amount and between MonthlyCharges and customer amount.

```
gender: ['Female' 'Male']
Partner: ['Yes' 'No']
Dependents: ['No' 'Yes']
PhoneService: ['No' 'Yes']
MultipleLines: ['No phone service' 'No' 'Yes']
InternetService: ['DSL' 'Fiber optic' 'No']
OnlineSecurity: ['No' 'Yes' 'No internet service']
OnlineBackup: ['Yes' 'No' 'No internet service']
DeviceProtection: ['No' 'Yes' 'No internet service']
TechSupport: ['No' 'Yes' 'No internet service']
StreamingTV: ['No' 'Yes' 'No internet service']
StreamingMovies: ['No' 'Yes' 'No internet service']
Contract: ['Month-to-month' 'One year' 'Two year']
PaperlessBilling: ['Yes' 'No']
PaymentMethod: ['Electronic check' 'Mailed check' 'Bank transfer (automatic)'
 'Credit card (automatic)']
Churn: ['No' 'Yes']
```

```
gender: ['Female' 'Male']
Partner: ['Yes' 'No']
Dependents: ['No' 'Yes']
PhoneService: ['No' 'Yes']
MultipleLines: ['No' 'Yes']
InternetService: ['DSL' 'Fiber optic' 'No']
OnlineSecurity: ['No' 'Yes']
OnlineBackup: ['Yes' 'No']
DeviceProtection: ['No' 'Yes']
TechSupport: ['No' 'Yes']
StreamingTV: ['No' 'Yes']
StreamingMovies: ['No' 'Yes']
Contract: ['Month-to-month' 'One year' 'Two year']
PaperlessBilling: ['Yes' 'No']
PaymentMethod: ['Electronic check' 'Mailed check' 'Bank transfer (automatic)'
 'Credit card (automatic)']
Churn: ['No' 'Yes']
```

Here, we have replaced some attributes with no class so that binary attributes can be sustained.

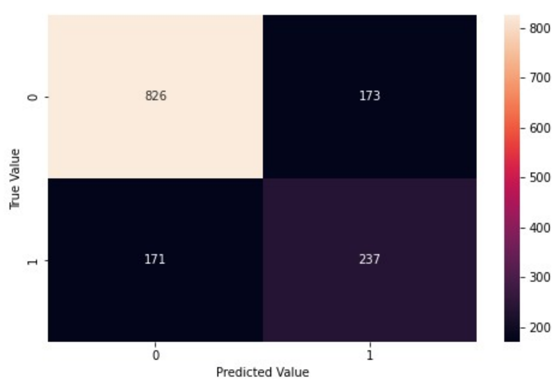
```
gender: ['Female' 'Male']
SeniorCitizen: [0 1]
Partner: [1 0]
Dependents: [0 1]
tenure: [ 1 34  2 45  8 22 10 28 62 13 16 58 49 25 69 52 71 21 12 30 47 72 17 27
  5 46 11 70 63 43 15 60 18 66  9  3 31 50 64 56  7 42 35 48 29 65 38 68
 32 55 37 36 41  6  4 33 67 23 57 61 14 20 53 40 59 24 44 19 54 51 26 39]
PhoneService: [0 1]
MultipleLines: [0 1]
InternetService: ['DSL' 'Fiber optic' 'No']
OnlineSecurity: [0 1]
OnlineBackup: [1 0]
DeviceProtection: [0 1]
TechSupport: [0 1]
StreamingTV: [0 1]
StreamingMovies: [0 1]
Contract: ['Month-to-month' 'One year' 'Two year']
PaperlessBilling: [1 0]
PaymentMethod: ['Electronic check' 'Mailed check' 'Bank transfer (automatic)'
 'Credit card (automatic)']
MonthlyCharges: [29.85 56.95 53.85 ... 63.1  44.2  78.7 ]
TotalCharges: [  29.85 1889.5   108.15 ...  346.45  306.6  6844.5 ]
Churn: [0 1]
```

```
gender: [1 0]
SeniorCitizen: [0 1]
Partner: [1 0]
Dependents: [0 1]
tenure: [0.         0.46478873 0.01408451 0.61971831 0.09859155 0.29577465
 0.12676056 0.38028169 0.85915493 0.16901408 0.21126761 0.8028169
 0.67605634 0.33802817 0.95774648 0.71830986 0.98591549 0.28169014
 0.15492958 0.4084507  0.64788732 1.         0.22535211 0.36619718
 0.05633803 0.63380282 0.14084507 0.97183099 0.87323944 0.5915493
 0.1971831  0.83098592 0.23943662 0.91549296 0.11267606 0.02816901
 0.42253521 0.69014085 0.88732394 0.77464789 0.08450704 0.57746479
 0.47887324 0.66197183 0.3943662  0.90140845 0.52112676 0.94366197
 0.43661972 0.76056338 0.50704225 0.49295775 0.56338028 0.07042254
 0.04225352 0.45070423 0.92957746 0.30985915 0.78873239 0.84507042
 0.18309859 0.26760563 0.73239437 0.54929577 0.81690141 0.32394366
 0.6056338  0.25352113 0.74647887 0.70422535 0.35211268 0.53521127]
PhoneService: [0 1]
MultipleLines: [0 1]
OnlineSecurity: [0 1]
OnlineBackup: [1 0]
DeviceProtection: [0 1]
TechSupport: [0 1]
StreamingTV: [0 1]
StreamingMovies: [0 1]
PaperlessBilling: [1 0]
MonthlyCharges: [0.11542289 0.38507463 0.35422886 ... 0.44626866 0.25820896 0.60149254]
TotalCharges: [0.0012751  0.21586661 0.01031041 ... 0.03780868 0.03321025 0.78764136]
Churn: [0 1]
InternetService_DSL: [1 0]
InternetService_Fiber optic: [0 1]
InternetService_No: [0 1]
Contract_Month-to-month: [1 0]
Contract_One year: [0 1]
Contract_Two year: [0 1]
PaymentMethod_Bank transfer (automatic): [0 1]
PaymentMethod_Credit card (automatic): [0 1]
PaymentMethod_Electronic check: [1 0]
PaymentMethod_Mailed check: [0 1]
```

Here, we have scaled MonthlyCharges, TotalCharges, and tenure from 0 to 1.

```
Epoch 192/200
176/176 [==============================] - 0s 2ms/step - loss: 0.2954 - accuracy: 0.8654
Epoch 193/200
176/176 [==============================] - 0s 2ms/step - loss: 0.2940 - accuracy: 0.8672
Epoch 194/200
176/176 [==============================] - 0s 2ms/step - loss: 0.2990 - accuracy: 0.8638
Epoch 195/200
176/176 [==============================] - 0s 2ms/step - loss: 0.2942 - accuracy: 0.8686
Epoch 196/200
176/176 [==============================] - 0s 2ms/step - loss: 0.2931 - accuracy: 0.8686
Epoch 197/200
176/176 [==============================] - 0s 2ms/step - loss: 0.2948 - accuracy: 0.8649
Epoch 198/200
176/176 [==============================] - 0s 2ms/step - loss: 0.2941 - accuracy: 0.8683
Epoch 199/200
176/176 [==============================] - 0s 2ms/step - loss: 0.2930 - accuracy: 0.8644: 0s - loss: 0.2918 - accuracy:
Epoch 200/200
176/176 [==============================] - 0s 2ms/step - loss: 0.2939 - accuracy: 0.8663

38]: <keras.callbacks.History at 0x1cee24607f0>

39]: model.evaluate(X_test, y_test)

44/44 [==============================] - 0s 2ms/step - loss: 0.6524 - accuracy: 0.7555

39]: [0.652396023273468, 0.7555081844329834]
```

After running 200 epochs our accuracy was 86.63%. After evaluating the model of x_test and y_test the accuracy was 65.2% and 75.5%.



After evaluating the test case with the predicted case, we have found accuracy for 0 (no) 826 and accuracy for 1 (yes) 237.

	precision	recall	f1-score	support
0	0.83	0.83	0.83	999
1	0.58	0.58	0.58	408
accuracy			0.76	1407
macro avg	0.70	0.70	0.70	1407
weighted avg	0.76	0.76	0.76	1407

This classification report shows statistics on precision and recall. This is the plot of the performance of the overall model. The accuracy here is 76% and the two precision values for 0 class and 1 class are 83% and 58% respectively. That means out of the prediction that our model made this is the percentage of correct predictions made for both of them.

4 Conclusions

We have tried to implement a dataset-based customer churn prediction model which we have trained based on the dataset we have collected, and the implementation result was about 76% accuracy. A large portion of the implementation has been occupied by the data processing task and the rest of it was ANN (Artificial Neural Network) based learning and testing.

4.1 Challenges

During the development of the project, we have faced some problems.

- First of all, during the data filtration process, we have faced some complications in that some of the filled were empty. Secondly, some float-type variables were in string.
- Some values could have been included using the "no" case, but they were not. So, we had to change their properties.
- We had to identify what type of activation function should we use, and the most problematic part has not differed between "softmax" and "sigmoid".

4.2 Limitations

- Our model is predicting about 86% for the test case and 76% for the predictive cases.
- Our model is based on values within a range of 0 to 1. So, if a company decides to use our model using the large continuous values the model might provide a biased answer.

4.3 Future Directions

We have used a dataset containing user and company interactive behaviors, but the general stores, banks, and other companies can use streaming data to constantly update and identify the user churn. So that, if a customer comes, they can instantly take measures to retain their customer.

5 Reference:

[1] T. Y. Fei, L. H. Shuan, L. J. Yan, G. Xiaoning, and S. W. King, "Prediction on customer churn in the telecommunications sector using discretization and Naïve Bayes classifier," International Journal of Advances in Soft Computing and its Applications, vol. 9, no. 3, 2017.
[2] Y. Khan, S. Shafiq, A. Naeem, S. Hussain, S. Ahmed, and N. Safwan, "Customers churn prediction using Artificial Neural Networks (ANN) in telecom industry," International Journal of Advanced Computer Science and Applications, vol. 10, no. 9, 2019, doi: 10.14569/ijacsa.2019.0100918.
[3] J. Ljungehed, "Predicting Customer Churn Using Recurrent Neural Networks," 2017.
[4] M. U. Tariq, M. Babar, M. Poulin, and A. S. Khattak, "Distributed model for customer churn prediction using convolutional neural network," Journal of Modelling in Management, 2021, doi: 10.1108/JM2-01-2021-0032.
[5] I. Ullah, B. Raza, A. K. Malik, M. Imran, S. U. Islam, and S. W. Kim, "A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector," IEEE Access, vol. 7, 2019, doi: 10.1109/ACCESS.2019.2914999.