Predicting customers churning in telecommunications industry: A machine learning approach

Chandana Katta
Computer Science
University of Central Missouri
Lee's Summit
cxk04781@ucmo.edu

Lakshmi Vaishali Batchu
Computer Science
University of Central Missouri
Lee's Summit
lxb47320@ucmo.edu

Kandukuri Vijaya Lakshmi
Computer Science
University of Central Missouri
Lee's Summit
vxk23230@uncom.edu

***Abstract*—Churn prediction is one of the commonly used case in machine learning field. It is very important for the telco businesses to be aware about why and when their customers are about to churn. Well accurate and robust churn prediction model helps businesses to take precautions and necessary actions to prevent their customers from churn. Churn prediction is the concept of predicting which customers are likely to stop a subscription to a service based on how they use the service. This work will help various companies to understand the factors behind customer churn and the actual customer churn rate using machine learning which the company can use to reduce the customer churn and also make strategies to retain back the churned customer.**

***Index Terms*—Machine Learning, Customer Churn, Prediction**

I. Introduction

Customer churn is the measure of customers that halted using a particular firm's product or service during a certain time period. One can estimate churn rate by dividing the number of customers that company lost during that time period by the number of customers that company had earlier before that time period. Though predicting customer churn is a challenging task but it is crucial too because business problem especially in industries where the cost of customer acquisition is high and difficult to manage such as in technology, telecom, finance, etc. sectors. Churn prediction consists of detecting which customers are about to cancel a subscription to a service or a product. Predicting customer churn is coined a binary classification problem. Customers either churn or retain in a given period. The reasons behind the customer churn are divided into two types: "accidental and intentional". Accidental churn happens when the plans are updating so as to keep the clients from using the services later on, for example financial terms that make benefits that are not useful and costly for the client. Intentional churn happens when the clients switch to another organization that gives same set of services, with upgraded ideas from rivalry, further developed services and lesser cost for a similar service. In recent years, churn prediction has become an important method for telecommunication industry. To tackle customer churn rate, the telecom operators must find out these customers before they churn. Therefore, developing a unique and accurate classifier that will predict future churns is vital. This classifier must be able to recognize users who might churn in the near future, so the operator can take any effective measures to stop these customers from stop using their services maybe by using discount and promotions or another strategy. When it comes to useful business applications of machine learning, it doesn't get much better than customer churn prediction. This problem, will have lots of data and real time stream of data to work on. And working with this data will, increase the profits of a particular organization. Churn rate is a critical metric of customer satisfaction. Low churn rates mean happy customers; high churn rates mean customers are leaving you. A small rate of monthly/quarterly churn compounds over time. For example, 1monthly churn quickly translates to almost 12In this project, we're using "Telecom Customer Churn" dataset which is available on Kaggle. There are 22 features or independent variables and 1 dependent variable for 7043 clients. Dependent or labeled variable indicates if a customer has left the company (churn=yes) within the last month. Since the dependent variable has two states (yes/no or 1/0), this is a binary classification problem.

The features are: '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'. Customers are very important assets in any industry since they are considered as the main profit source. These days, companies have become advanced that they put much effort not only to convince or attract the customers, but also to retain their existing customers. Churned customers are persons who move to other company for various factors. To decrease the customer churn rate, the company should be able to predict the behavior of customers correctly and establish connections between customer attrition and keep elements in check. This is

https://github.com/LakshmiVaishali/finalproject

a binary classification task, which differentiates churners from non-churners. Machine learning is a data analytical model which allows model analytical building. Using algorithms that repeatedly learn from data, machine learning allows systems to sensitive hidden patterns without being explicitly think where to look for the problems. Machine Learning consists of three types of learning: unsupervised machine learning, semi-supervised machine learning, and supervised machine learning. Supervised learning is the machine learning task of finding the patterns from dataset which already has the output within it. Unsupervised learning is the machine learning task of finding the patterns from dataset which have no output. Semi-supervised learning is a class of supervised learning tasks and techniques which also make use of unlabeled data for training – typically a small set of labelled data with a large set of unlabeled data. Semi supervised learning falls between unsupervised learning and supervised learning. Customer Churn prediction involves these three types of analysis.

II. MOTIVATION

It takes large amount of time and investment to attract new customers in the substitution of the left over loyal customers that churn. This includes much effort to advertise and market the product and services in various ways and make the new customers understand about the services and products that we provide.

The biggest loss we can observe3e due to churn is the credibility, market share and revenue decreases. This in turn effect the company in various ways such as recession in the industry and over production and ends up with marginal income or sometimes no income.

Mainly the company market shares may drop down gradually because of the lack of sales and revenue decrease. This could in turn influence the other financial funding that the company might acquire from the external sources.

It is highly difficult task to find out he reason why customers or a group of customers stops using the product or service of a particular organization or company. This needs much data analysis and also the accurate prediction as it will effect the revenue of the company at the end.

III. RELATED WORK

Currently the competition between different companies had increased globally that provide similar kind of services due to the evolution of digital marketing and 4 technology over the internet. Tracking customer retention rate, customer requirements and fulfil them accordingly is of great importance while marketing.

A. Abbreviations and Acronyms

Along with the growth of industries, the need for prediction of customer attrition or churn rate has also grown because it's very difficult to bring in new customers in place of the existing, loyal customers when they stop using your company services or products. Learning the reasons for the customers churn will help the companies to reduce the customer churn rate.

Kriti [1] in their paper Customer churn: A study of factors affecting customer churn using machine learning has used various factors affecting customer churn (price sensitivity, technology, customer service, tenure, security) to predict the customer chrun rate. Also comparing various algorithms to analyze customer churn and prescribe solutions to avoid this churn. She has given the future work as the predictions from the ML model can help in understanding the customers who might leave their services. Also suggested various solutions based on those predictions. Essam Abou El Kassem and Shereen Ali Hussein [2] in their paper clearly described that Customer churn is a problem for most companies because it affects the revenues of the company when a customer switches from a service provider company to another. Social media sentiment analysis is used to predict the factors behind customer churn. Praveen Lalwani and Manas Kumar Mishra [3] in their paper has Compared the time taken to train the model and accuracy of various ML algorithms. Their paper concludes that ensemble learning techniques such as XGBoost classifier gives maximum accuracy when compared to other models. Saran Kumar A. [4] In his paper had conducted a survey on various ML algorithms and techniques to predict customer attrition or churn rate. Proposed to use various boosting classification techniques for better accuracy. Pradeep B and Sushmitha Vishwanath Rao [5] in their paper have explained how to use various ML algorithms to analyze customer attrition or churn rate in the logistics industry.

This research work concludes that the purpose analysis of customer churn rate is to identify valuable customers that potentially contribute to the profitability of the company.

Problem Definition

To build an effective customer churn rate prediction system using Machine Learning algorithms to predict whether the customer(s) may churn or not, also factors that lead to customer churn. Advantages of Logistic Regression algorithm:

1. Easy implementation, accurate and efficient to interpret.
2. IT allows multi feature regression (multinomial regression).
3. Unknown records can be classified swiftly.
4. High amount of accuracy for many small and large data sets and it performs well for linearly separated data-sets.

Disadvantages of Logistic Regression algorithm:

1. Overfitting is not self managed by the Logistic Regression.
2. Logistic regression creates an assumption of linearity between the dependent and independent variables.
3. Logistic regression cannot solve the non-linerarity problems
4. Logistic Regression requires average or no multicollinearity between independent variables.

Need of New System:

Limitations of Existing Systems:

1. Time-consuming and not accurate.
2. Cannot predict customer attrition or churn rate based on huge dataset.
3. Feedback Forms are not sufficient.

4. Competitive Market.

IV. PROPOSED FRAMEWORK

- Loading Data.
- Data Cleaning, Feature engineering, Data Visualization.
- Encoding Categorical Data.
- Folow the train-test split procedure to estimate the per formance of machine learning algorithms when they ar used to make predictions on data.
- Displaying the results using Bar plot.

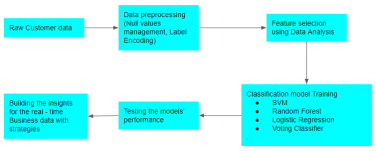

Fig. 1. Example of a figure caption.

Initially, we imported the python modules such as NumPy, pandas, sklearn, matplotlib., seaborn, plotly to use all the ml algorithms to perform various operations on the dataset and to predict customer churn rate and the factors lead tocustomer churn.

In order to visualize and compare the attributes in the dataset WE used matplotlib and seaborn modules. Along with these we imported the modules like Label Encoder, One Hot Encoder which are used to correct the data discrepancies by label encoding the categorical data, and accuracy score to check the accuracy of various algorithms to predict customer churn rate.

Then loaded the telco dataset which we got from Kaggle. Using the head function and pandas library we've checked the top 5 rows and attributes or the services on the basis of which we'll be able to calculate customer churn and find out the important factors behind customer attrition or churn rate in telecom company. Dependent variable contained imbalanced class distribution. Positive class (Churn=Yes) is much less than negative class (churn=No). Imbalanced class distributions influence the performance of a machine learning model negatively. Next, we used describe function to figure out the count of the rows, mean value, standard deviation, minimum and maximum value, values for numerical attributes (like Senior citizen, Tenure, Monthly Charges).

The data frame(df) and info function we check for the null value in our dataset and also the datatype of the attributes /services of telco company provided in the dataset.

Data Cleaning is an important step in machine learning or data mining as it helps to clean the data or rectify the mistakes which is there in the dataset. Completion of this we have visualize or display a bar chart to compare the customer attrition or churn rate precentage as "Yes" and not churned as "No". Since our project is based on supervised learning it already has the output to train and test the model.

```
In [239]: df.columns.values

Out[239]: array(['customerID', 'gender', 'SeniorCitizen', 'Partner', 'Dependents',
                 'tenure', 'PhoneService', 'MultipleLines', 'InternetService',
                 'OnlineSecurity', 'OnlineBackup', 'DeviceProtection',
                 'TechSupport', 'StreamingTV', 'StreamingMovies', 'Contract',
                 'PaperlessBilling', 'PaymentMethod', 'MonthlyCharges',
                 'TotalCharges', 'Churn'], dtype=object)

In [240]: df.dtypes

Out[240]: 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
```

Fig. 2. Attributes of the Telco dataset.

V. ALGORITHMS USED

A. Logistic Regression

Logistic regression is a classification model rather than a regression model. It is a classification algorithm used to assign observations to a discrete set of classes. Unlike linear regression which outputs continuous numeric values, logistic regression transforms its output using the logistic sigmoid function to return a probability value which can then be mapped to two or more discrete classes. It is a classification model, which is very easy to realize and achieves very good performance with linearly separable classes. It is an extensively employed algorithm for classification in industry. The logistic regression model is a method for binary classification that can be generalized to multiclass or multi-attributes classification.

B. Random Forest

Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.Most important features of the Random Forest Algorithm are that it can handle the data set containing continuous variables as in the regression case and categorical variables as in the case of classification. It performs better results for classification problems.

C. K-nearest neighbors

The k-nearest neighbors (KNN) algorithm is a very simple, easy-to-implement machine learning supervised type algorithm that can be used to solve both regression problems and classification problems. As the name (K Nearest Neighbor) suggests it considers K Nearest Neighbors (Data points) to predict the class or continuous value for the new Datapoint.

D. Adaboost Classififer

We might play around with the parameters for a bit or augment the data, but in the end, we are still using a single model. Even if we build an ensemble, all if the trained models and applied to our data separately. Boosting, takes a more

iterative approach. It's still technically an ensemble technique in that any models are combined together to perform the final one, but takes a cleverer approach.

E. Support Vector Machine(SVM)

Support vector machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problems. After giving in SVM model sets of labeled training data for each category, they're able to categorize new text.

VI. CHURN RATE DISTRIBUTION

Here in Churn Rate Distribution we check the customer churn rate based on the attributes given in the dataset.

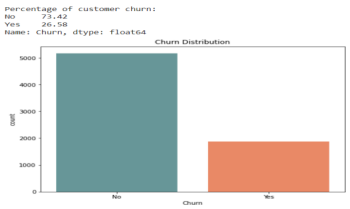

Fig. 3. Distribution of the Churn Rate.

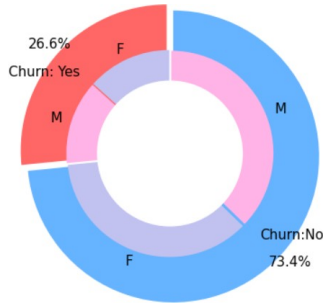

Fig. 4. Distribution of the Churn Rate with respect to gender.

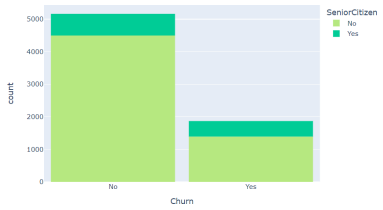

Fig. 5. Distribution of the Churn Rate with respect to senior citizen.

VII. ANALYSIS OF THE DATA WITH THE FEATURES AVAILABLE/ DATA DESCRIPTION

The categorical features need to be converted to numbers so that they can be included in calculations done by a machine learning model. The categorical variables in our data set are not ordinal (i.e., there is no order in them). For example, "DSL" internet service is not superior to "Fiber optic" internet service. An example for an ordinal categorical variable would be ratings from 1 to 5 or a variable with categories "bad", "average" and "good". When we encode the categorical variables, a number will be assigned to each category. The category with higher numbers will be considered more important or effect the model more. Therefore, we need to do encode the variables in a way that each category will be represented by a column and the value in that column will be 0 or 1. We also need to scale continuous variables. Otherwise, variables with higher values will be given more importance which effects the accuracy of the model. As we briefly discussed in the beginning, target variables with imbalanced class distribution are not desired for machine learning models. We use up sampling which means increasing the number of samples of the class with less samples by randomly selecting rows from it. The features such as gender and partner are evenly distributed with approximate percentage values. The difference in churn is slightly higher in females. A higher proportion of churn can be observed in younger customers , customers with no partners, and customers with no dependents. The demographic section of data highlights on-senior citizens with no partners and dependents as a particular segment of customers likely to churn. Internet service variable is definitely important in predicting churn rate. Correlation measures the linear relationship between two variables. Features with high correlation are more linearly dependent and have almost the same effect on the dependent variable. So, when two features have a high correlation, we can drop one of them. In our case, we can drop highly correlated features like Multiple Lines, Online Security, Online Backup, Device Protection, Tech Support, Streaming TV, and Streaming Movies. Churn prediction is a binary classification problem, as customers either churn or are retained in a given period. Two questions need answering to guide model building: 1. Which features make customers churn or retain? 2. What are the most important features to train a model with high performance? Monthly Charges and Phone Service columns will not be used to reduce multicollinearity in the data. Here s the Fig. 7. we can observe the correlation between the each attribute that is the dataset that effects the customer churn rate.

VIII. RESULTS ANALYSIS

For the machine learning algorithms we used, here we place the individual graphs and the confusion matrix obtained for

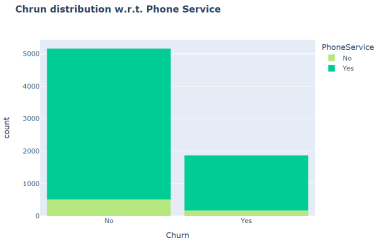

Fig. 6. Distribution of the Churn Rate with respect to Phone Service.

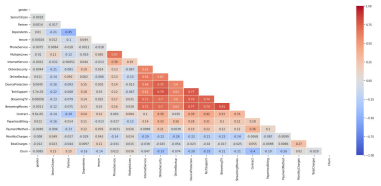

Fig. 7. Correlation co efficient matrix.

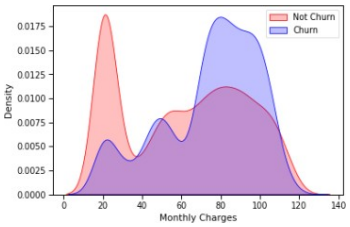

Fig. 8. Distribution of monthly charges by churn.

each algorithm to analyse the best algorithm that has produced or predicted the customer churn rate t the maximum accuracy.

A. Logistic Regression

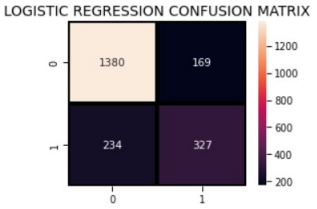

Fig. 9. Logistic Regression Confusion matrix.

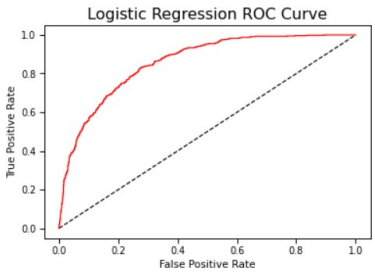

Fig. 10. Logistic Regression ROC curve.

B. Random Forest

C. Final Classification analysis

In the our research we concluded that the Voting Classifier with Adaboost algorithm, Gradient boosting algorithm and Logistic Regression algorithms we are able to build the accurate classfier.

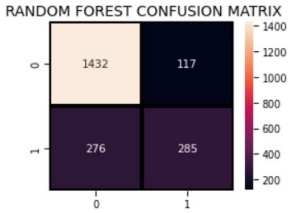

Fig. 11. Random Forest Confusion matrix.

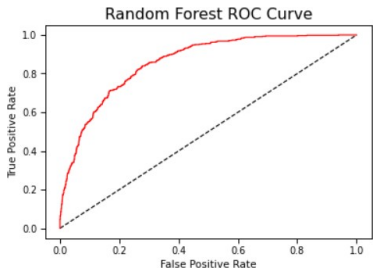

Fig. 12. Random Forest ROC curve.

REFERENCES

[1] Kriti, "Customer churn: A study of factors affecting customer churn using machine learning" Iowa State University Capstones, Theses and Dissertations, March 2019

[2] Essam Abou el Kassem, Shereen Ali Hussein, Alaa Mostafa Abdelrahman, Fahad Kamal Alsheref. "Customer Churn Prediction Model and Identifying Features to Increase Customer Retention based on User Generated Content" IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 11, November 5, 2020

[3] I Praveen Lalwani, Manas Kumar Mishra, Jasroop Singh Chadha, Pratyush Sethi. "Customer churn prediction system: a machine learning approach" in Springer

[4] Saran Kumar A., Chandrakala D. "A Survey on Customer Churn Prediction using Machine Learning Techniques" International Journal of Computer Applications (0975 – 8887) Volume 154 – No.10, November 2016.

[5] Pradeep B, Sushmitha Vishwanath Rao, Swati M Puranik, Akshay Hegde "Analysis of Customer Churn prediction in Logistic Industry using

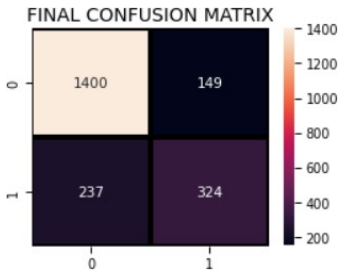

Fig. 13. Final Confusion matrix.

	precision	recall	f1-score	support
0	0.86	0.90	0.88	1549
1	0.68	0.58	0.63	561
accuracy			0.82	2110
macro avg	0.77	0.74	0.75	2110
weighted avg	0.81	0.82	0.81	2110

Fig. 14. Final Classification Report.

Machine Learning" International Journal of Scientific and Research Publications, Volume 7, Issue 11, November 2017 ISSN 2250-3153.
[6] A. Payne and P. Frow, "A Strategic Framework for Customer Relationship Management," Journal of Marketing, vol. 69, no. 4, pp. 167-176, 2005.
[7] S.-Y. Hung, D. C. Yen and H.-Y. Wang, "Applying data mining to telecom churn management," Expert Systems with Applications , vol. 31, no. 3, pp. 515-524, October 2006.
[8] T. Jiang, J. L. Gradus and A. J. Rosellini, "Supervised Machine Learning: A Brief Primer," Behavior Therapy , vol. 51, no. 5, pp. 675-687, 2020.
[9] P. Mehta, M. Bukov, C. H. Wang, A. G. Day, C. Richardson, C. K. Fisher and D. J.Schwab, "A high-bias,l ow-variance introduction to Machine Learning for physicists," Physics Reports, vol. 810, pp. 1-124, 2019.
[10] O. Simeone, "A Very Brief Introduction to Machine Learning With Applications to Communication Systems," IEEE Transactions on Cognitive Communications and Networking, vol. 4, no. 4, pp. 648-664, 2018.
[11] S. Marsland, Machine Learning: An Algorithmic Perspective, 2nd edition ed., Chapman and Hall/CRC, 2014.
[12] M. Mohammed, M. b. Khan and E. B. M. Bashier, Machine Learning: Algorithms and Applications, CRC Press, 2016.
[13] M. Oral, E. L. Oral and A. Aydin, "Supervised vs. unsupervised learning for construction crew productivity prediction," Automation in Construction, vol. 22, pp. 271-276, 2012.
[14] V. Verdhan, publisher logoSupervised Learning with Python: Concepts and Practical Implementation Using Python, Apress, 2020.
[15] Z.-H. Zhou, Ensemble Methods - Foundations and Algorithms, Taylor Francis group, LLC, 2012.
[16] A. Kumar and M. Jain, Ensemble Learning for AI Developers: Learn Bagging, Stacking, and Boosting Methods with Use Cases, Apress, 2020.
[17] M. Van Wezel and R. Potharst, "Improved customer choice predictions using ensemble methods," European Journal of Operational Research, vol. 181, no. 1, pp. 436-452, 2007.
[18] J. Karlberg and M. Axen, "Binary Classification for Predicting Customer Churn," Umeå University, Umeå, 2020.
[19] D. Windridge and R. Nagarajan, "Quantum Bootstrap Aggregation," in International Symposium on Quantum Interaction, 2017.
[20] B. Raja and P. Jeyakumar, "An Effective Classifier for Predicting Churn in Telecommunication," Journal of Advanced Research in Dynamical and Control Systems, vol.11, no. 1, Juni 2019.
[21] V. F. Rodriguez-Galiano, B. Ghimire, J. Rogan, M. Chica-Olmo and J. Rigol-Sanchez, "An assessment of the effectiveness of a random forest classifier for land-cover classification," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 67, pp. 93-104, 2012.