**Churn Prediction Using Big Data in Telecommunication Industry**

Saad Azhar

*Department of Computer Science, George Mason University (*[*sazhar2@masonlive.gmu.edu*](mailto:sazhar2@masonlive.gmu.edu)*)*

**Introduction**

Major problem these days is the loss of revenue of up to around 5% year-over-year for telecom companies due to customers leaving a particular telecom operator because of unsatisfaction with services or packages. This issue is known as churn which is creating a big gap of finances in telecommunication sector and loss of revenue for telecom organizations due to which it is ultimately affecting the economy of that specific country because telecom sector forms a major part of economy these days. Moreover another problem is that researchers are not able to find concrete solution for churn problem. So the purpose of my research is to get this churn problem solved in telecommunication sector using big data and to find out research gap in this domain which is not enabling researchers to get this problem solved. I am trying to solve this problem by predicting churn ratio beforehand (before when it actually happens) so that remedial strategies could be introduced in time so stop customer form leaving.

**Using Big Data to Predict Churn Ratio**

**Big Data**

Big Data is a very ubiquitous term, which has no specific definition, on which a lot of work is going on these days due to the market demand. It means the collection of data from various sources and when that data continuous to grow and the size is in terabytes and petabytes (PB) then that data is known as Big Data. Big Data has four key themes which are Information, Technologies, Methods and Impact[1]. Information means the data itself, Technologies means the devices or ways which are used to collect data, Methods means ways by which data is to be manipulated and impact means the effects and uses of big data in society which in my case is solely focused on prediction of churn ratio. According to authors:

*“*Big Data represents the Information assets characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value”[1].

**Churn Ratio**

Customers switching from one particular telecom operator to another due to unsatisfaction with services and packages is known as churn. The ratio with which customers switch from one operator to another is known as churn ratio[2]. Churn ratio is required by telecom operators so that they may be able to know the present state of their customers and in case of high churn rate, operators then introduce strategies and better services to stop or reduce the customers from leaving that particular operator[3].

Big data plays a vital role in churn prediction. These days’ telecom data is very big, normally in terabytes (TB) and petabytes (PB). It is too much of a size which enables us to analyze huge data sets to produce more accurate results as compared to using small data sets and coming up with inconsistent results. Small data sets can only be used for basic analysis. Algorithms cannot be applied on them to extract churn trends because small data sets contain very less information. So for accurate churn prediction by using specific prediction algorithms huge size of data is very important.

Big data can be manipulated in thousands of ways to come up with useful information such as fraud detection in businesses, stock trends, diseases in health sector and churn prediction in telecom sector[1]. I am using big data in this research to predict churn ratio so that churn problem could be reduced from telecom sector. Big Data can be manipulated in different ways to predict churn ratio. It can be used for classification and clustering of different types of telecom data in separate classes and clusters using Hadoop, which is further divided into training and test data set with a ratio of 7:3. On training data set prediction algorithms are applied and deduced results are then compared with test set at the end to predict churn ratio[4].

Big data can potentially lead us to the most accurate solution for churn problem. It can be very helpful if there is a need to manipulate or extraction of trends using technologies such as clustering using Hadoop to predict churn ratio. Hadoop gives reliable results only if the size of data is very big[4].

**Methods/Algorithms used to Predict Churn Ratio**

Data Scientists can potentially use different algorithms such as Artificial Neural Network, Decision Trees, Covering algorithms, Statistical based techniques, Fuzzy correlation techniques and logistic regression to predict churn ratio.Using different algorithms to predict churn ratio and to come up with most accurate one is a very good approach. It enables us to compare results deduced from different algorithms and then helps in picking the most accurate algorithm. Every algorithm has its own cause and effect. Decision Tree process is a very time taking and an expensive process which is not even very reliable due to inconsistency in the results. Testing different algorithms proved that all results deduced were different from each other. There was no consistency which can lead data scientists to a decision that there was a big research gap in algorithms because data scientists were not using best algorithm to predict churn[5]. Statistical based techniques gave the most accurate results due to its consistency and further future prediction could be based because very limited research is being done on this algorithm[5].

I deduced another method for churn prediction which was based on deep learning convolutional neural networks. Authors mentioned deep learning algorithms can act as a catalyst in filling up the research gap of finalizing the most accurate churn prediction algorithm [6]. Future of big data and Prediction Analysis is based on deep learning algorithms and initial theoretical research results in this domain were very encouraging which were proposed in the form of improved accuracy. Authors suggested prediction using deep learning algorithms combined with correct data sets could be the most accurate solution of this problem[6].

All the sources which I researched for algorithm deduction were interlinked because they were trying to lead me to the algorithm which could produce the most accurate churn results and which could ultimately fill up the research gap in this domain. All the sources were interlinked in such way that they gave me a direction to proceed further and they ultimately lead me to the research gap. Now future research could be in the form of how to fill that gap by using these deduced algorithms.

**Data used to Predict Churn Ratio**

Mare used In previous researches churn was predicted using real time data sets but previous churn data could also be used for comparisons in the form of training data set and test data set. Training data set is the data on which algorithms are applied and we tend to learn some trends from them and test data sets are the ones with which training results are compared to deduce a result. Here previous churn data sets (old) could be used as test data sets for comparisons[3], [5], [7], [8]. But in majority of researches these days Demographics Data and CDRs are used to predict churn ratio[2], [9]. All the algorithms were applied on this data set such as decision trees, regression and Neural Network etc, to come up with results but the problem with these data sets are that just the physical location and call history of a customer is used to predict churn and this don’t gives a true impression of a customer’s feeling about the operator. So the effects of these data sets are that we can never know what the customer feels about the services or what could be the reasons for customers churning and we can never know which strategies to introduce to stop customers from leaving[9]. These strategies and data sets have become very old now which have no scope for further improvement. So new areas of further research are required in this domain.

Other data sets that could be used in future for churn prediction are Complaints and Repairs data. All the same prediction algorithms such as decision tress, regression and Artificial Neural Networks can be applied on these data sets too and which could produce more accurate churn results than CDRs and Demographics data. The effects of these data sets are that they can enable data scientists what the customer feel about the services or what could be the reasons for customers churning and they can help us in introducing strategies to stop customers from leaving. These data sets can enable data scientists to know the real problems faced by customers which could be a very important Key Performance Indicator in churn prediction and reasons of churn. These data sets can help us in filling up the research gap which is in the form of use of wrong types of data sets [9]. It also makes sense to predict churn using the data sets in which problems faced by customers are involved.

All these types of researches on such data sets are interlinked because they were trying to lead me to the correct and best data sets which can produce the most accurate churn results. These data sets helped me in giving further insight of research gap in churn prediction domain and the ways to solve them.

**Conclusion**

My findings from this research were in the form of most accurate algorithms for churn prediction, research gap and problems due to which companies were not able to reach to a concrete solution for churn problem. Companies were not using the best data sets and algorithms for churn prediction. They were just trying to re-invent the wheel by using old strategies which became outdated as data increased. Future areas of research could be the use of complaints and repair data which could produce more accurate results, with the use of Deep Learning Convolution Neural Network algorithms, as compared to other algorithms and demographics data. These algorithms have the capability to deal with huge and inconsistent complaints data sets which could be a strong basis for future research in churn prediction domain based on initial perceptions and because of all this, evolution of and improvementcould be In conclusion, this research helped me in filling up the research gap in a small capacity. May be building up on this gap in the form of its solution could give us a concrete solution in future for Churn Prediction using Big Data in Telecommunication Sector.

.

**References:**

[1] A. De Mauro, M. Greco, and M. Grimaldi, “What is big data? A consensual definition and a review of key research topics,” 2015, pp. 97–104.

[2] M. Kaur and P. Mahajan, “Churn Prediction in Telecom Industry Using R,” *Int. J. Eng. Tech. Res. IJETR ISSN*, pp. 2321–869.

[3] R. J. Jadhav and U. T. Pawar, “Churn prediction in telecommunication using data mining technology,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 2, no. 2, 2011.

[4] N. Kamalraj and A. Malathi, “Hadoop Operations Management for Big Data Clusters in Telecommunication Industry,” *Int. J. Comput. Appl.*, vol. 105, no. 12, 2014.

[5] A. M. Almana, M. S. Aksoy, and R. Alzahrani, “A survey on data mining techniques in customer churn analysis for telecom industry,” *Int. J. Eng. Res. Appl.*, vol. 45, pp. 165–171, 2014.

[6] A. Wangperawong, C. Brun, O. Laudy, and R. Pavasuthipaisit, “Churn analysis using deep convolutional neural networks and autoencoders,” *ArXiv Prepr. ArXiv160405377*, 2016.

[7] D. BALASUBRAMANIAN and M. Selvarani, “Churn Prediction In Mobile Telecom System Using Data Mining Techniques,” *Int. J. Sci. Res. Publ.*, vol. 4, 2014.

[8] S.-Y. Hung, D. C. Yen, and H.-Y. Wang, “Applying data mining to telecom churn management,” *Expert Syst. Appl.*, vol. 31, no. 3, pp. 515–524, Oct. 2006.

[9] J. Hadden, A. Tiwari, R. Roy, and D. Ruta, “Churn prediction using complaints data,” in *Proceedings Of World Academy Of Science, Engineering and Technology*, 2006.

[10] Y. Richter, E. Yom-Tov, and N. Slonim, “Predicting customer churn in mobile networks through analysis of social groups,” in *Proceedings of the 2010 SIAM International Conference on Data Mining*, 0 vols., Society for Industrial and Applied Mathematics, 2010, pp. 732–741.

[11] M. R. Khan, J. Manoj, A. Singh, and J. Blumenstock, “Behavioral modeling for churn prediction: Early indicators and accurate predictors of custom defection and loyalty,” in *2015 IEEE International Congress on Big Data*, 2015, pp. 677–680.