

Customer Churn Prediction Modelling Based on Behavioural Patterns Analysis using Deep Learning

Sanket Agrawal
Department of Computer Engineering
Sardar Patel Institute of Technology
Mumbai, India
sanket.agrawal@spit.ac.in

Aditya Das
Department of Computer Engineering
Sardar Patel Institute of Technology
Mumbai, India
aditya.das@spit.ac.in

Amit Gaikwad
Department of Computer Engineering
Sardar Patel Institute of Technology
Mumbai, India
amit.gaikwad@spit.ac.in

Sudhir Dhage
Department of Computer Engineering
Sardar Patel Institute of Technology
Mumbai, India
sudhir_dhage@spit.ac.in

Abstract— Customer churn refers to when a customer ceases their relationship with a company. A churn rate, used to estimate growth, is now considered as important a metric as financial profit. With growing competition in the market, companies are desperate to keep the churn rate as low as possible. Thus, churn prediction has gained critical importance, not just for existing customers, but also for predicting trends of future customers. This paper demonstrates prediction of churn on a Telco dataset using a Deep Learning Approach. A multi-layered Neural Network was designed to build a non-linear classification model. The churn prediction model works on customer features, support features, usage features and contextual features. The possibility of churn as well as the determining factors are predicted. The trained model then applies the final weights on these features and predict the possibility of churn for that customer. An accuracy of 80.03% was achieved. Since the model also provides the churn factors, it can be used by companies to analyze the reasons for these factors and take steps to eliminate them.

Keywords— Churn Prediction, Deep Learning, ANN, Correlation Analysis.

I. INTRODUCTION

Customer churn is a concept that is quickly gaining importance and significance, especially in the world's leading and competing organizations. Customer churn [1] can be defined as the group customers or subscribers that discontinue a service over a specific period. It is one of the of the central factors that determines the steady-state of the customer pool that a business can support.

Although used in many scenarios [2], it is a term primarily used in the business aspects. The proportion of customer base that discontinues a service is a potential indication of customer discontent, better deals, sales success or marketing on the part of the competitors.

Customer churn is fast becoming a pressing issue in the industry. This is because customer churn directly affects the profitability of a company. If companies assume that their profits are directly proportional to their customer base, then the easiest way to maintain profit is to ensure the rate of customer growth is always higher than the churn rate.

One of the defining issues in building customer churn prediction models is that in practice we find the percentage of customer churn that occurs in proportion to the entire pool of customer over a fiscal year is characteristically very small [3].

Machine learning models work well with enough effort put on engineering the features of the model. But each model in this case needs to have the proper features defined depending on the problem at hand. Because of the sheer volume and unstructured format of the data, finding behavioral pattern by traditional means fails.

The approach that has been taken is using Deep Learning, whose primary advantage is that it learns multiple levels of representation on its own and comes up with its own features. In a use case like customer churn, this is highly relevant because a large number of factors might surprisingly not be relevant at all. Thus the weights assigned to these would be negligible. Deep learning models, while creating their blueprints of the data, would thus choose to ignore such irrelevant features. This is because deep learning makes use of multiple layers, so certain traits ignored by one layer can easily be picked up by another layer.

Since we are working on a dataset pertaining to the mobile telecommunications industry we decide to analyse the parameters [4] most significantly affecting customer churn rates. This helps contextualise the parameters of the datasets strictly in terms of behaviour analysis of the customers. We believe this would help in making predictions based on customer behaviour and help us draw conclusions that are more effective.

The paper is organised as follows. Section II gives a summarised overview of the literature survey conducted and the related works in the domain. Section III gives a detailed flow of the proposed methodology, with III-A explaining the fundamentals of the Deep Learning models. Section IV and V talk about the dataset that is used and the various techniques that were applied to clean and normalise the data. Section VI deals with the implementation, which involves building a neural network. Section VII evaluates the results that were generated and find the weightage of different parameters in the churn process. Finally, section VIII and IX describe the conclusion and future scope.

II. RELATED WORKS

Research and implementations carried out previously by other authors was studied in order to gain insights to tackle this problem and finding a suitable approach. In this section we will present our findings related to the same.

The research work carried out by Ning Lu et al.[5] discusses the highly skewed class distribution and the lack of churn data which is generally typical in churn analysis. They present a churn prediction model using a boosting algorithm which is believed to be very robust and has demonstrated success in churn prediction in the banking industry. They have used logistic regression as the base learner and it enhanced using the boosting algorithm. Their training set includes 7190 customers drawn randomly, with 678 churners and 6512 nonchurners. A churn prediction was made based on each three-month period, in order to simulate the real-world scenario. Experimental evaluation shows that in customer churn data, which is highly skewed, the weight given by Gentle AdaBoost algorithm also suggests a good separation, and provides an opportunity to define a high risk customer group. This paper has also raised multiple questions including increasing performance by a hybrid of different classifiers.

The paper by Xiaojun Wu et al.[6] describes the process of E-commerce customer churn prediction. It begins by discussing the existing work done in the churn prediction on the real bank datasets in China and then goes on to state the problem of data imbalance in the e-commerce domain. Their proposed prediction model was based on Improved SMOTE and AdaBoost algorithm. SMOTE is an oversampling technique that generates synthetic samples from the minority class, which generates new synthetic samples along the line between the minority examples and their selected nearest neighbors. This combines oversampling and undersampling methods to balance the datasets and generate a certain number of positive and negative samples by setting sampling ratio. The processed data was then passed into the AdaBoost algorithm, which is based on the resampling technique where misclassified data is more favored to be sampled. This improves the classifying ability after a number of iterations. Confusion matrix was then created to estimate the number of false positives and false negatives. Finally, B2C e-commerce data is used to show that this model achieves a better score and less time than earlier implementations and so can be applied to other domains.

Philip Spanoudes et al.[7] describe the various applications of Deep Learning in the context of churn prediction in their research. This paper has chosen neural network in order to allow for unsupervised learning, which in turn helps to skip the feature engineering step, and at the same time increasing the overall prediction accuracy of the system. A novel data representation model was generated that can be applied to various datasets regardless of the company feature set. For this they have generated an event vector for every user at a given timeframe, with each vector having a hundred dimensions. They have then used a 4-layered feed-forward discriminative architecture which classifies users as churners or non-churners based on their call patterns. This model was finally trained on an implementation of SGD and early stopping to adjust for over-fitting. 390 days of data of three companies was taken for the evaluation purpose and was split into a set of 30 days. The paper concludes by stating that the results have largely remained constant at 70-80% in all the cases.

The paper written by V. Umayaparvathi et al. [8] is a survey on various datasets and implementations to carry out the process of predicting the churn rate. It starts off with explaining the problem of customers getting churned and then reviewing the existing work in this domain. It tries to define

metrics to define the churn rate and a high level view of the entire process to be followed. The paper emphasizes mainly on the telecom industry and categorises the parameters into six broad categories. It also enlists various datasets available to use, pertinent to the task at hand, namely - PAKDD 2006 Data Mining Competition Dataset, ACM KDD Cup 2009 Orange Labs Dataset, Cell2Cell Dataset and CrowdAnalytix Community. It then conducts a comparative study of the datasets used, features extracted and the methods employed by various authors. Various classification techniques like SVM, Decision Trees, Neural Networks and simple K-Means clustering were studied for analysis of models. The paper also discusses the different performance metrics that are generally employed to evaluate the models designed, some of which are - Confusion Matrix, Accuracy, Precision and Recall, F1-Score, AUC curve, and Correlation Graph.

The work by Federico Castanedo et al. [9] goes about investigating the use of autoencoders, deep belief networks and multi-layered feedforward networks of different configurations in the field of mobile network churn prediction. However, it gives special importance to Deep Learning because of its inherent ability to come up with good features. A dataset of call records from an enterprise BI system of 1.2 million customers spanning over 16 months was used for this purpose. They have used 4 states - new, active, inactive and churned to classify events. They have eventually made use of the multi-layer architecture and have encoded the input data into a hierarchical representation. A total of 12 models were trained and predictions were generated with each of the models on all 12 months. SGD is used to reduce the loss function and dropout layers have been used to prevent the situation of overfitting. They have proved that the model is stable for different months and have predicted improved results with the inclusion of location data. One of the major issues identified in this paper is that it does not scale properly for long-term user interactions.

III.

PROPOSED METHODOLOGY

A. Churn Neural Architecture

The methodology that will be followed in this paper will work on customer user records data provided by the company. This raw data will have to be cleaned in order to work with it. A high level analysis is required to identify the variance present in the data and accordingly applying techniques to normalize the data. After this is done, parameters will be selected which will become the part of the feature set used for the training of a model. This processed data is then fed into a multi-layered Artificial Neural Network which is designed for this problem. Once the model gets trained on the training data, it is verified by using validation data and then tested on the testing set. This would give a percent accuracy metric to quantify the model in question. To determine the major factors that were responsible for the churn, a correlation graph will be generated. Micro analysis will also be done on numeric parameters to check for their effectiveness in the overall churn trends and try to predict patterns. This is outlined in Fig. 1.

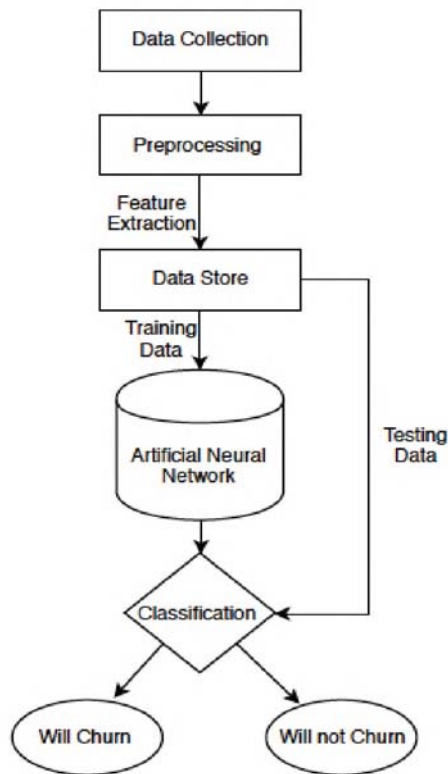


Fig. 1 Proposed Methodology Flow Diagram

B. Deep Learning

Deep learning is a subset of machine learning algorithms based on the human brain structure called artificial neural networks [10]. It helps systems learn by example and teaches them what humans naturally do. In deep learning, the system learns to perform classification tasks directly from images, text, sound or videos. Deep learning models can achieve very high accuracy, sometimes exceeding the performance of humans. Models are trained using neural network architectures which contain many layers on a large set of labeled dataset. The term “deep” usually refers to the magnitude of hidden layers in the neural network, as shown in Fig. 2.

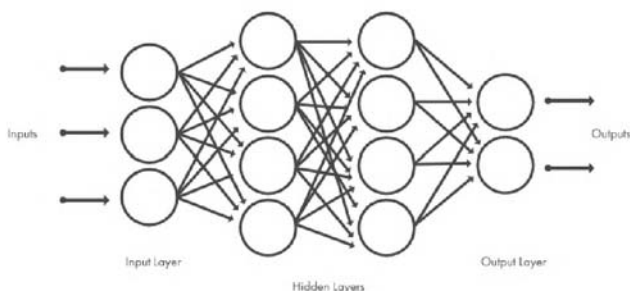


Fig. 2 A standard Neural Network Architecture

Deep learning is one of the only methods which can be exploited to overcome the challenges of feature extraction because deep learning models are capable of learning to focus on the right features by themselves thus requiring little guidance from the programmer. This makes deep learning an extremely powerful tool for modern machine learning.

IV. DATASET DESCRIPTION

The dataset that we will use for the remaining part of discussion in this paper will be IBM Watson Telco Customer Churn Data Set [11]. This consists of user data of 7043 customers along with their final label of either “Churned” or “Not Churned”. It provides information about the customer subscription along with their basic demographic details, which will help in training the base model. Main features that can be extracted from this dataset are :

- Demographic Data - CustomerID, Gender, and whether they have partners and dependents or not
- Subscriptions availed by the customer - This includes Phone, internet security, online backup, multiple lines, device protection, tech support and streaming TV and Movies
- User billing profile - Payment method, contract, paperless billing, Monthly Charges and Total charges
- Duration for which the customer was with the company(Tenure)

Along with this data, each customer is also classified as churned or non-churned, which in this context is those who have left the company within the previous month.

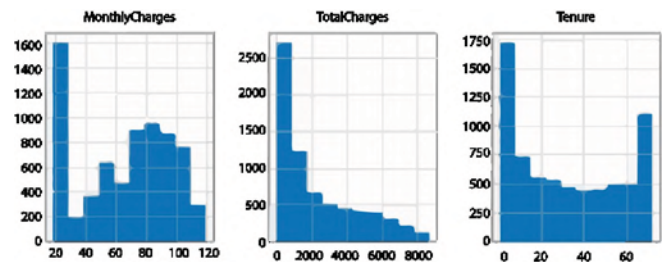


Fig. 3 Ranges of Numeric Parameters

The graphs in Fig. 3 show us that the numeric parameters in this dataset, although spread over a wide range, are concentrated around the lower margins. This will result in oversampling for a particular range if not normalized using discretization or logarithmic transformations. Due to this, data processing is highly required.

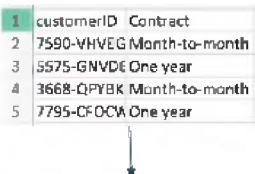
V. DATA PREPROCESSING

The raw data acquired has to be processed for application of any algorithm. To do this, a number of algorithms were applied to clean the data to make it suitable for the Neural Network. Firstly, columns which were completely irrelevant to the training of the model were removed. This included columns like CustomerID which was unique for each data entry. Also, incomplete entries were removed from the dataset to remove any ambiguity or source of error.

A major issue with the data is that it contains a number of textual parameters, and Machine Learning, especially Deep Learning Algorithms falter when fed with textual data, and so it is necessary to convert them into numerical entries. The technique of Label Encoding [12] is used to tackle this issue. It is used to transform object labels to numerical labels which are always between 0 and $n_classes-1$. This is used for a number of columns where the entries are categories in either a yes or a no. A problem with this technique is that it assumes that categories with higher values are better in nature, which

is in fact not true with a number of columns. This drawback renders this technique ineffective for multiclass attributes.

The next major processing algorithm applied was One Hot Categorical Encoding [13]. This is binary vector representation of categorical parameters. The indexed columns thus is given the value of 1 while all other indices become 0. This technique allows the data to be more expressive. This is used more in the Deep Learning models since binary categorisation results in better training as compared to a larger numerical range. Another reason for this being a preferable choice for multi-valued columns was that the categories in a column were largely unrelated. Using this technique avoid assigning any higher preference to a given technique.



customerID	Contract
7590-VHVEG	Month-to-month
5575-GNVDE	One year
3668-QPYBK	Month-to-month
7795-CFOCA	One year

customerID	Contract_Month-to-month	Contract_One year	Contract_Two year
7590-VHVEG	1	0	0
5575-GNVDE	0	1	0
3668-QPYBK	1	0	0
7795-CFOCA	0	1	0

Fig. 4 Categorical Encoding on the data

For columns like tenure, the entries are spread over a large number of months, with many months having few entries. A technique of Discretization [14] was used, which is basically a reduction mechanism to convert a large domain of numerical values to a small range. This data was normalised by compartmentalizing them into bins, which were decided based on the number of years. This ensured that there were sudden spikes in the graph of tenures.

VI. CUSTOMER CHURN PREDICTION MODEL

The implemented model is based on artificial feed-forward neural networks. A sequential model has been used which can be viewed as a linear stack of neural layers and mainly comprises of dense layers. A Dense [15] or a fully connected layer is a linear operation on the layer's input vector in which every input is connected to every output by a weight so there are n inputs * n outputs weights, generally followed by a non linear activation function. Finally, an m -dimensional vector is received as an output. Thus, a dense layer can be used to change the dimensions of the vector as we can apply various transformations like rotation, scaling and translation to the vector.

The model requires the data input to the first layer with a definite shape. This input shape will be the number of initial parameters that has been extracted from preprocessing stage. The number of neurons for the first layer has been set to 16 for optimized results. In this model, the network weights were initialized using the argument kernel initializer to a small random number generated from a uniform distribution between 0 and 0.05 which is the default 'uniform' weight initialization. In the first two layers of the neural network, the rectifier linear unit activation function has been applied, which is defined as the positive part of the following equation:

$$f(x) = x^+ = \max(0, x) \quad (1)$$

A dropout layer has been added between every 2 Dense layers with a probability of 0.1. Dropout [16] is a regularization technique which is used to reduce the complexity of the model and its goal is to prevent overfitting. Using "dropout", a certain amount of neurons in a layer are deactivated with a certain probability 'p' from a Bernoulli distribution which is set to 10%. Thus, this sets one-tenths of the activations of a layer to zero, so that the neural network does not rely on particular activations during training in a feed-forward pass. This ratio is chosen since it does not result in a significant loss of neurons, as well as ensure efficient running of the network. As a consequence, because of the certain neurons being dropped, the neural network will learn differently thus preventing overfitting. Redundant representations are thus avoided as the network cannot rely on the particular neurons and the combination (or interaction) of these to be present. Another pseudo advantage of this is that training speed increases. It is also to be noted that Dropout is only applied during the training phase, and is ignored while validating and testing the model. Finally, if the training has finished, we use the complete network for testing or in other words, we set the dropout probability to 0.

Similarly for the second dense layer of the network the uniform kernel initializer and relu activation function was used. The final output layer has a single neuron, since this classification has just two outcomes, namely yes or no for the churn. A uniform kernel initializer was used in this layer and the activation function was changed to sigmoid.

Sigmoid Activation function is a activation function of form

$$f(x) = \frac{1}{1 + e^{-x}} \quad (2)$$

whose range is between 0 and 1. It forms an S-shaped graph.

This thus formed the 5-layered ANN which was used for the training and prediction purpose. After trying with a few more optimizers in the hyperparameter optimization, best results were obtained using the Nadam[17] optimizer. Nadam is the Nesterov Adam optimizer which is Adam RMSprop with Nesterov momentum. The data was split into 80% training dataset and 20% validation dataset. The model was trained for one hundred epochs and the batch size taken in each epoch was ten.

The implementation of this network was carried out using the Keras neural networks API, which is a Python deep learning library. This library enables us to design a sequential neural network which can be designed manually and provides a set of various types of layers that can be used and configured as per the requirements of the use case.

VII. EVALUATIONS AND RESULTS

Based on the above procedure, we have trained the model and tested it on the validation dataset. The model was designed to give a percentage accuracy on the data and we have achieved an accuracy of 80.03%, which is a significant increase from the numbers that was encountered in our studies.

To evaluate this model and to find out the major affecting parameters for churning, various methods of analysis were carried out. Firstly it has to be noted that the dataset includes 3 numeric parameters, which are important to do a

deterministic evaluation of their impact on churn as the values gradually increase. These are Tenure for which the customer was associated, Monthly Charges and Total Charges incurred by the customer.

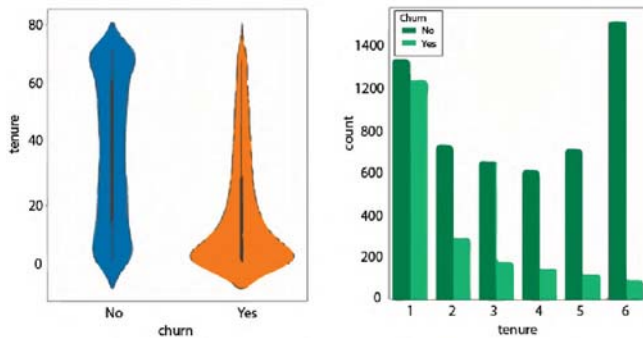


Fig. 5 Effect of Tenure on churning

Tenure is generally a direct indicator of the loyalty of a customer as so it is imperative to check if this is in line with the data available. A violin plot, which is a modification of the box plot to give a more detailed spread of the data, as shown in Fig. 5, reveals that most of the churners are those customers who have been with the company for less than 18 months and this reduces to almost a negligible number as the amount of months crosses beyond 60. This can be seen in a more definitive manner in the histogram where the ratio of non-churners to churners is close to 1 in the people whose tenure is less than a year, while this ratio increases manifold when the customer is with company for a number of years.

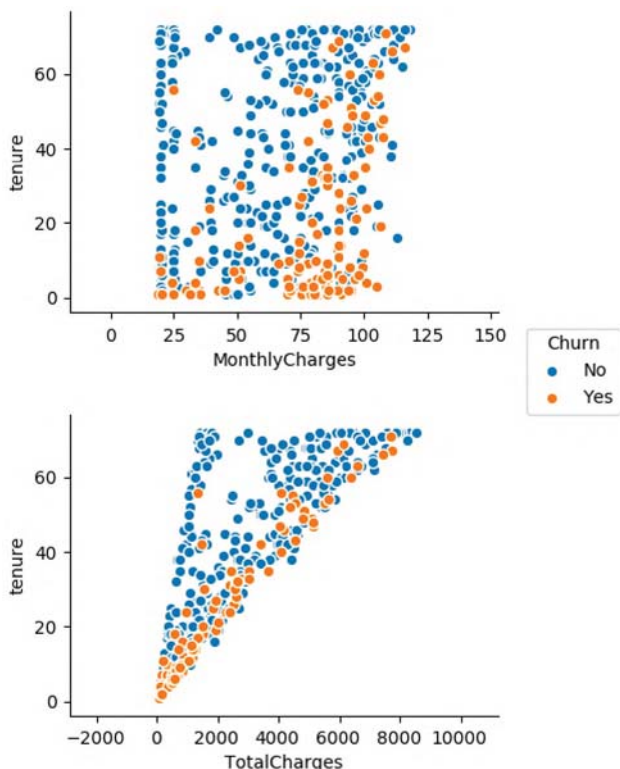


Fig. 6 Effect of tenure on Billing Charges

Next we see the effect of charges on tenure of the customer. In Fig. 6, 500 samples are randomly selected from the dataset and are plotted on the graphs as per their tenure and charge values. It can be inferred from these graphs that even

among those with low tenure, it is mainly those with high monthly charges that are labelled churned. It is also noted that although the tenure touches 60-80 months, if the monthly charges are high, the customer is like to be churned away. As with total charges, we can see the for every tenure value, the concentration of churn is around the region where total charges are maximum, thus indicating that most of the churners are leaving the company due to highly charged bills.

Now to find out the effectiveness of various parameters used in the prediction, we use the concept of correlation analysis [18]. Data correlation is the behaviour of data in presence of another set of data, which would refer to the correspondence of the features with respect to the output label. In this case, each of the extracted parameter is compared with the final churn label of the dataset. This helps in finding the similarity of each parameter with the final result and shows a direct dependence of the weight of the factor contributing to the churn. This correlation coefficient returns a value between +1 and -1; +1 meaning highly similar and -1 being highly dissimilar. Fig. 7 depicts these values. A positive correlation value would mean that the parameter contributes to the churning while a negative value would suggest that the parameter prevents churning. As we can see in the above figure, major factors that are preventing churn are:

- 1-year tenure
- Month-to-month contracts
- Fibre optic internet services

This behaviour is in direct association with the facts that those customers that have long term contracts or those associated with the firm for a longer duration are more likely to be retained than those with short term contracts and small tenures.

Other parameters like paper billing, no device protection and no phone services, as it can be seen, play a negligible role in deciding the churn, since their correlation values are very close to 0.

VIII. CONCLUSION

It was seen how the issue of customer churn is becoming increasingly pressing by the day and previous works were analyzed to find the gaps in the implementation of the solution. This was also used to extract a set of parameters which were seen to affect the churn. The multi-layered Artificial Neural Network model designed to tackle this problem resulted in an accuracy of 80.03%. The model also displayed a list of the attributes which are directly and inversely related with the churn rate, which can be used for isolating churn parameters. This analysis becomes a powerful tool for organizations to decide which parameters to focus upon for retaining customers and avoiding losing them to competitors.

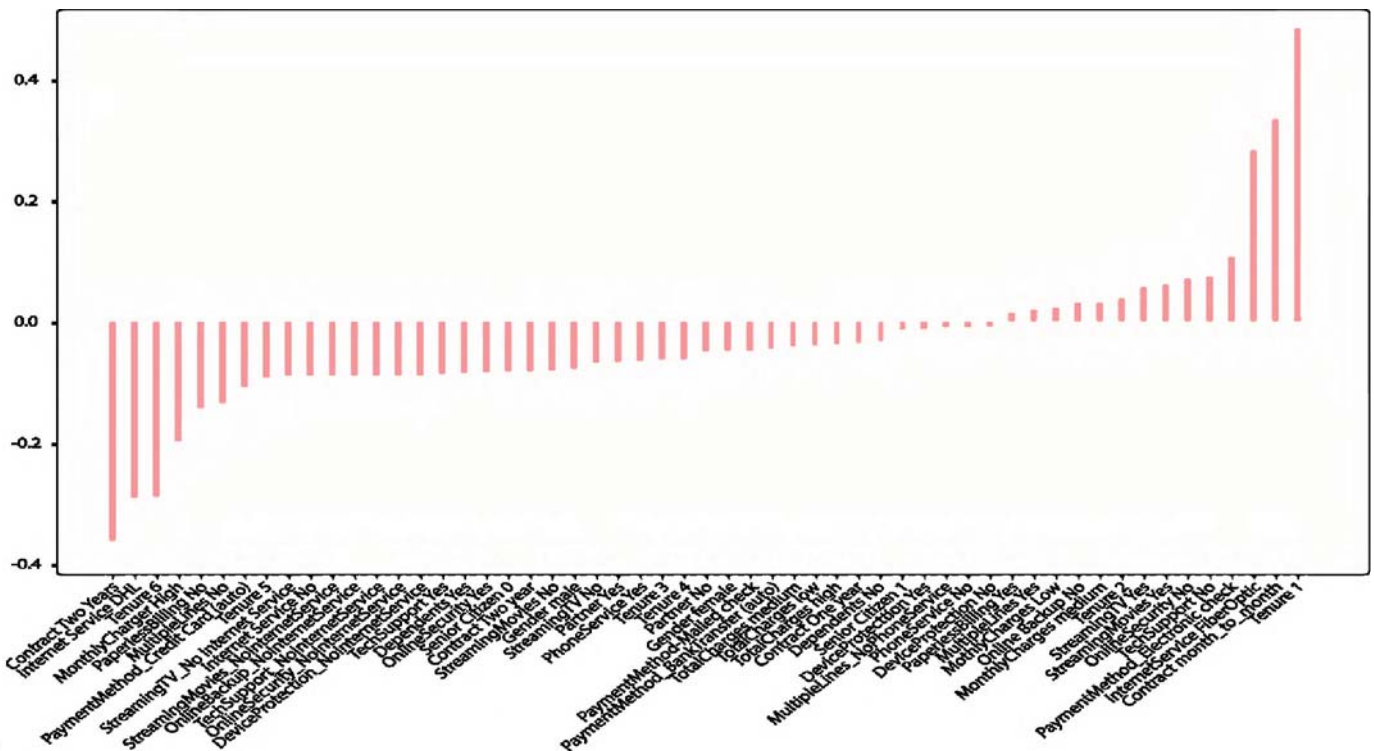


Fig. 7 Correlation Analysis for Churn Parameters

IX. FUTURE SCOPE

Behavioral analysis can be extended to association between prediction of churn and lifetime customer value. This can help shape retention policies as well, which will further boost the revenue of the company. A churn retention model can also be designed to retain customers. These retention models can further be studied to analyse the cost of adding a new customer to the base against the cost of implementing retention policies to retain existing customers. .

REFERENCES

- [1] Pradeep B, S. V. Rao, S. M. Puranik and A. Hegde, "Analysis of Customer Churn prediction in Logistic Industry using Machine Learning", *Int. J. of Scientific and Res. Publications*, vol. 7, no. 11, 2017.
- [2] R. Mattison, *The Telco Churn Management Handbook*, 2001.
- [3] Forhad N, Hussain MS, Rahman RM, "Churn analysis: Predicting churners", In *Digital Information Management (ICDIM)*, 2014 Ninth International Conference, 2014.
- [4] Dahiya K., Bhatia S., "Customer churn analysis in telecom industry", *Reliability, Infocom Technologies and Optimization (ICRITO)(Trends and Future Directions)*, 2015 4th International Conference, pp. 16, 2015.
- [5] N. Lu, H. Lin, J. Lu, G. Zhang, "A customer churn prediction model in telecom industry using boosting", *IEEE Trans. Ind. Informat.*, vol. 10, no. 2, pp. 1659-1665, 2014.
- [6] Xiaojun Wu, Sufang Meng, "E-commerce customer churn prediction based on improved SMOTE and AdaBoost", 2016 13th International Conference on Service Systems and Service Management (ICSSSM), pp. 1-5, 2016.
- [7] Spanoudes, P., Nguyen, T, "Deep Learning in Customer Churn Prediction: Unsupervised Feature Learning on Abstract Company Independent Feature Vectors", *arXiv preprint arXiv:1703.03869*, 2017.
- [8] V. Umayaparvathi, K. Iyakutti, "A Survey on Customer Churn Prediction in Telecom Industry: Datasets Methods and Metrics", *International Research Journal of Engineering and Technology (IRJET)*, vol. 04, no. 4, pp. 1065-1070, 2016
- [9] F. Castanedo, G. Valverde, J. Zaratiegui, and A. Vazquez, "Using Deep Learning to Predict Customer Churn in a Mobile Telecommunication Network", http://www.wiseathena.com/pdf/wa_dl.pdf, 2014.
- [10] LeCun, Y., Bengio, Y., Hinton, G. "Deep learning", *Nature* 521, 436444, 2015.
- [11] IBM Analytics Communities, *Telco Customer Churn Dataset*, 2015. [ONLINE]. Available: <https://www.ibm.com/communities/analytics/watson-analytics-blog/predictive-insights-in-the-telco-customer-churn-data-set/>.
- [12] C. Li, Q. Kang, G. Ge, Q. Song, H. Lu, J. Cheng, "Deep BE: Learning Deep Binary Encoding for Multi-label Classification", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 39-46, 2016.
- [13] Allen Chieng Hoon Choong, Nung Kion Lee, "Evaluation of Convolutionary Neural Networks Modeling of Dna Sequences Using Ordinal Versus One-hot Encoding Method", *International Conference on Computer and Drone Applications*, pp. 60-65, 2017.
- [14] Mridu Sahu, Shreya Sharma, Vyom Raj, N.K. Nagwani, Shrish Verma, Impact of discretization on classification of data using divide and conquer paradigm, *International Conference on Electrical Electronics and Optimization Techniques (ICEEOT)*, 2016.
- [15] T. J. OShea, J. Hoydis, "An introduction to Deep Learning for the Physical Layer", *IEEE Trans. Cogn. Commun. Netw.*, vol. 3, no. 4, pp. 563-575, 2017.
- [16] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., "Dropout: a simple way to prevent neural networks from overfitting", *J. Machine Learning Res.* 15, pp. 19291958, 2014.
- [17] Andrychowicz, M., Denil, M., Gomez, S., Hoffman, M. W., Pfau, D., Schaul, T., Shillingford, B., de Freitas, N., "Learning to learn by gradient descent by gradient descent", *Neural Information Processing Systems conference*, pp. 398189, 2016.
- [18] G. Andrew, R. Arora, J. Bilmes, K. Livescu, *Deep Canonical Correlation Analysis*, *International Conference on Machine Learning*, 2013