ISSN 2278-3091
Volume 8, No.3, May - June 2019
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse60832019.pdf
https://doi.org/10.30534/ijatcse/2019/60832019

Churn Prediction using Attention Based Autoencoder Network

Swetha Allam[1], Dr. Srinivas Kudipudi[2]
[1]Velagapudi Ramakrishna Siddhartha Engineering College, India, swethaallam1996@gmail.com
[2]Velagapudi Ramakrishna Siddhartha Engineering College, India, vrdrks@gmail.com

ABSTRACT

User Behaviour Analysis gives valuable insights for customer management. Especially in Telecom sector it can help find the customer churn rate. In Telecom, customer churn is to find whether the customer is going to leave the service of the current operator or not. Effective churn rate prediction is a critical task. Classification models are often used for churn rate prediction. But most of these models have several shortcomings. They require manual feature extraction, or the model cannot balance skewed datasets. To overcome these problems, in this paper, we are proposing a feature extraction model based on Autoencoder with attention mechanism. The Autoencoder network represents the data in latent space representation. Attention mechanism makes the network focus on the features that highly contributes to the target prediction. The network contains a separate classification module for churn prediction. This technique results in using fewer training set yet producing superior results. The proposed model is compared with other classification algorithms and the results show that our proposed model outperforms other models in terms of accuracy and area under the curve.

Key words: Autoencoder, Attention, Churn prediction, Classification.

1. INTRODUCTION

The main focus of customer churn prediction is to determinate the customers who are at risk of leaving the current operator and checking if it is worth retaining. [1]. Accurate churn prediction will help in effective customer management. The Telecom operators can leverage this to offer more customer centric services, which will lead to retaining satisfied old customers and also acquiring new customers.

Most of the literature is focused on Data Mining techniques such as building classification models to predict the customer churn as they are proved to be more effective. Classification models for churn rate can be broadly divided in to several categories, regression based models such as Logistic regression, tree based models like C5.0 decision trees, Bayesian models such as Naive Bayes. Ensemble methods such as AdaBoost and RandomForests are also leveraged to boost the predictive performance of the models. Traditional classification models require the manual extraction of the features form the dataset, which has to done by the experts as there can be many features with no apparent value. These models cannot effectively find the hidden relationships between those features.

Recent advantages in neural network architectures have paved way for new methods that can effectively find the hidden relationships between the features and combine those features to more accurately predict the churn. The many features are given as input to the input layer and at each hidden layer takes the output from the previous layer as input and performs an activation function to produce the output, which is given to the output layer.

The work presented in this paper uses autoencoder network with in built attention mechanism. The Encoder part of the network compresses the data into a latent space with the help of the decoder part. Separate classification module attached to the encoder network is used to predict the customer churn. Both the autoencoder module and the classification module are trained simultaneously. Attention mechanism makes the data in the latent space to be highly effective for robust churn prediction. Telcom churn dataset is used to evaluate the proposed model. The proposed model is evaluated for its accuracy and area under the curve measures and compared against other classification models namely Decision Trees, Logistic Regression, XGBoost, Random Forest, Naïve Bayes and SVM. The results show that the proposed model produces superior results than the compared models.

2. RELATED WORK

Mohanty [2] leveraged Extreme Learning Method(ELM) based model for customer churn prediction in banking sector data. Portuguese Banking Sector dataset was used to train the model. ELMs are a type of a feed forward neural networks. They are faster than other neural networks like probabilistic neural network and decision trees. ELM is a learning algorithm that consists of single-layer feed forward neural network, which contains an input layer, hidden layer and output nodes. The main advantages of ELM are that their

parameters, hidden nodes, input weights and biases are randomly allocated and need not have to be tuned. ELM provides smallest training error and the smallest output weight as well to predict accurately. [3] used the effect of social networks among the bank customers to predict the churn rate. Deep ensemble of classifiers was built that uses the power of individual classifiers in a meta-level by using convolutional neural networks. CNN can extract useful features by stacking of multiple predictions. The data used in this contains a database of all transactions, accounts, and snapshots of a customer database of a retail financial institution in North America over a period of 3.5 years. Ensemble of CNN classifiers can improve the performance of churn predictive classification with automatically capturing and extracting relevant features, especially after adding network features into the model.

A new hybrid algorithm named logit leaf model (LLM) which overcomes the shortcomings of decision trees and logistic regression was proposed by [4]. The algorithm is a combination of decision trees and logistic regression. It reduces the shortcomings of DT and LR while maintaining their strength. In Wangperawong's [5] work, data was represented as images and churn prediction is performed on those images using Convolutional Neural Networks(CNN) and Autoencoders. CNN is a supervised learning model primarily used for image classification. In order to use CNN, customer data was represented as images. A 2D array of normalized pixels was constructed. To understand customer behavioural patterns and to explain the reasons for churning, unsupervised learning approach autoencoders was used. Autoencoders are neural networks where the inputs and outputs are identical. They can be used for dimensionality reduction on data and have performed better than principal components analysis.

3. METHODOLOGY

The proposed methodology is shown in the Figure 1. It contains Telco Dataset Acquisition, Preprocessing the dataset to make it compatible for the training model, simultaneously training the autoencoder and classification modules, and finally evaluating the comparing the results with state of the art classification models.

3.1 Dataset

Telcom customer churn Dataset was taken IBM Watson Analytics Datasets. It contains 7043 samples across 21 attributes. Out of the 21 attributes 20 are independent variables and the last attribute is dependent variable. Each sample contains data about customer's attributes. The complete list of attributes is given in Table 1. Around 73% of the samples belong to non-churn class and 27% of the samples belong to churn class.

Table 1. Attributes in Dataset

Attribute	Description
customerID	ID of the customer
gender	Specifies if customer is male or female
SeniorCitizen	Specifies if Customer is senior citizen or not
Partner	Specifies whether customer has partner
Dependents	Specifies customers has dependents or not
tenure	Number of months the customer is with that particular company
PhoneService	Specifies the customers phone service
MultipleLines	Specifies if customer has multiple lines
InternetService	Specifies the customers type of internet connection
OnlineSecurity	Specifies if customer has online security
OnlineBackup	Specifies if customer has online backup
DeviceProtection	Specifies if customers device has protection
TechSupport	Specifies whether customer has used Tech support
StreamingTV	Specifies if customer has streaming TV
StreamingMovies	Specifies if customer has streamed movies
Contract	Type of contract term
PaperlessBilling	Specifies whether customer supports paperless billing
PaymentMethod	Type of customers payment method
MonthlyCharges	Amount charged to the customer every month
TotalCharges	Total amount charged to the customer
Churn	Specifies if customer has churned or not

3.2 pre-processing

Initially the samples that contains null values are removed from the dataset. Then the categorical variable with two categories such as gender are Label encoded in to 0 and 1. Remaining attributes that contains more than two categories are encoded using one-hot encoder into corresponding columns. The dataset is then normalized using standard scaling which is given as:

$$y = \frac{(x - \mu)}{\sigma} \quad (1)$$

Where x is the sample attribute, μ is the mean of the attribute and σ is the standard deviation

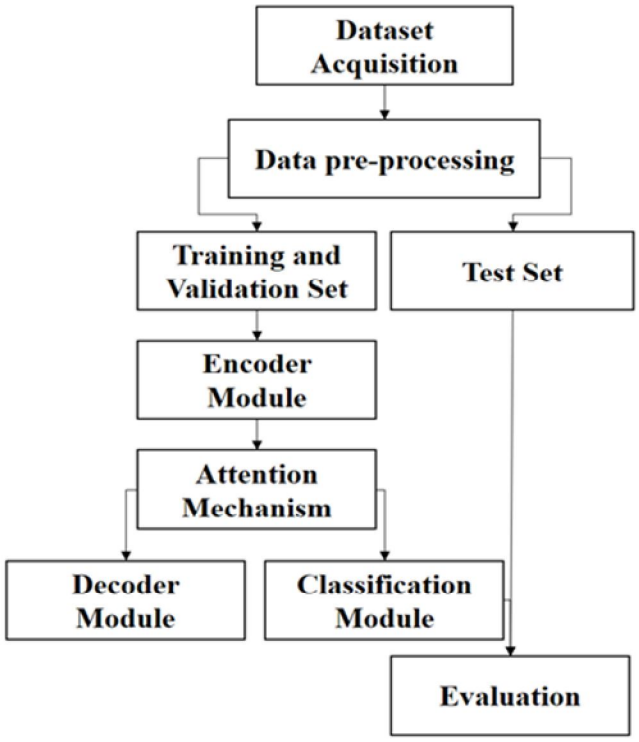

Figure 1. Proposed Methodology

3.3 Autoencoder with Attention mechanism

The proposed network contains an encoder decoder module and a classification module. Autoencoders are unsupervised learning neural networks that tries to reconstruct data that is given as input to the model [9]. It contains two modules 1) an encoder module and 2) a decoder module. The encoder module represents the data in minimal latent representation which the decoder module uses to reconstruct the original data. We are mainly focused on the latent representation of data as it gradually decreases the dimension of the data and represents the data in a compressed form. By having to reconstruct the data it learns the useful properties that are hard to find in the data.

Data representation in the autoencoders are often lossy and may still contain redundant features. Even though the features are important for prediction their importance is not equal across the dataset. Attention layer contains a vector of values that focuses on the features that contribute the most. Each value in the vector contains the importance value of the corresponding feature. This mechanism eliminates the unnecessary features and gradually increases the performance and generalization of the network.

As shown in the Figure 2. The encoder part contains normal hidden layers with neuros in a decreasing order, which the decoder tries to reconstruct. At the end of the encoder the data is in compressed form. This is given as input to the attention layer which computes the attention vector for each of the reduced features. The attention vector and the features are multiplied to produce the data for input to the classification module. When the attention vector finds that a particular feature does not contribute to the prediction it makes that particular value in the vector to 0 and that makes the network to forget about that feature. The classification module finds the hidden representations and outputs the final predicted value.

For this particular dataset we have taken an encoder module with three hidden layers each containing 150, 100 and 50. After the attention layer the total number of neurons will be 20. The decoder module resembles the encoder module in increasing order of neurons. The classification module has one hidden layer and an output layer. The output layer in the classification module has a single neuron with sigmoid activation function. Except this layer all the layers in the network has ReLU activation. Dropout technique is used to make the network generalize well.

4. EXPERIMENTS

4.1 Implementation Details

The proposed network was implemented using keras framework [8]. Keras is a deep learning framework that works on top of TensorFlow library. It contains functional API that is suitable for implementing networks with multiple outputs efficiently. Adam [6] optimizer is used to train the network with an initial learning rate of 0.0001. The autoencoder module uses Mean Squared Error as the loss function and the classification module uses binary cross entropy as the loss function. For training Nvidia GTX 1070 GPU with a batch size of 100 is used.

4.2 Evaluation Metrics

The performance of the model is evaluated using accuracy and area under the curve (AUC). Accuracy is a general evaluation metric which can be calculated as follows.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

Where TP is True Positives, TN is True Negatives, FP is False Positives and FN is False Negatives.

Area under the curve (AUC) is a common metric used for classification models [7]. It can be accounted for both true positives and false positives.

5. RESULTS

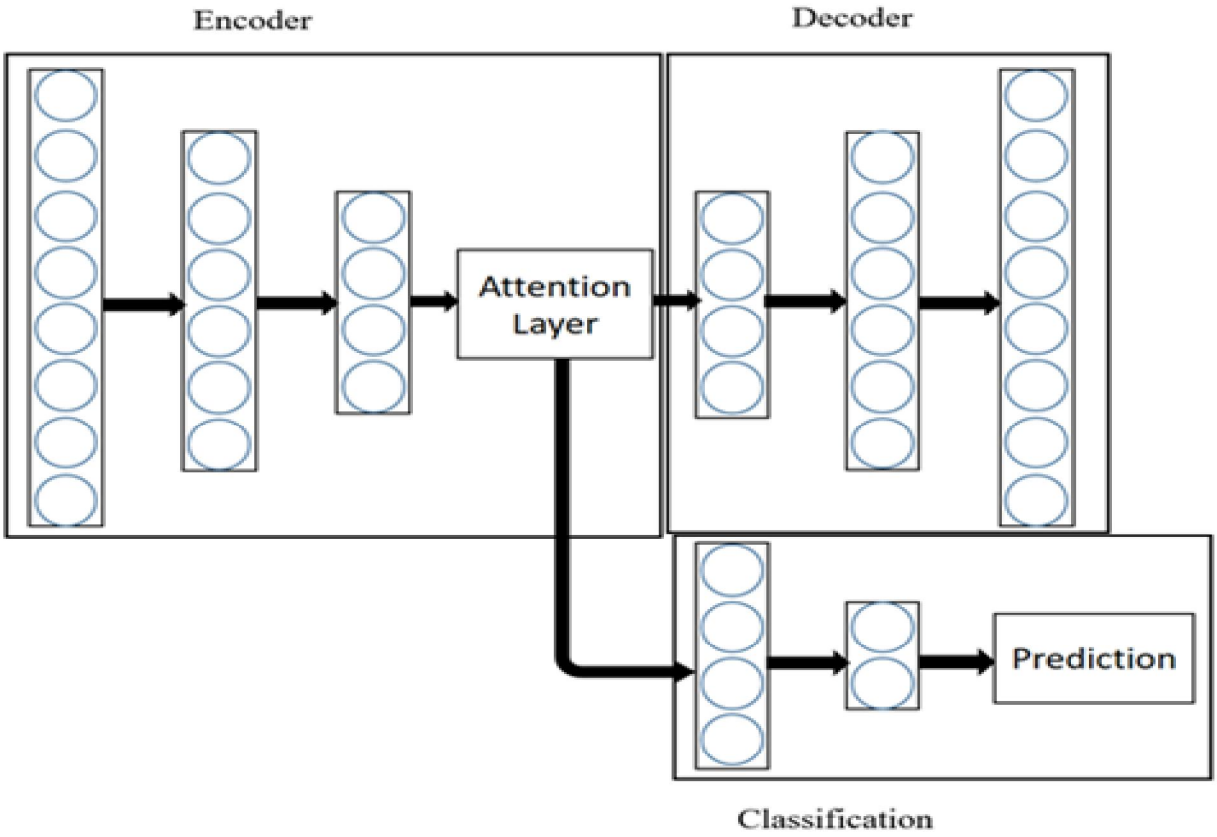

Figure 2. Proposed Autoencoder with Attention Mechanism

The proposed model was compared against other state of the art models generally used in classification problems. Their performance was evaluated using accuracy and area under the curve (AUC). As shown in the Table 2. Our model achieves better results in both the accuracy and AUC. After our proposed model Random Forest model gave good results. Although Random Forest classifier was able to give good accuracy it falls behind in AUC. AUC is very important for achieving high true positives which are critical in application like churn prediction.

Receiver operating characteristic or ROC curve is widely used to visualize the sensitivity and false alarm rate of the model. It is shown in Figure 3. Figure 4. Shows the training and testing accuracy of the proposed model. It can be seen that the model was able to learn faster due to the attention mechanism and achieves good accuracy in just 20 epochs. It also generalizes well for the testing data. The combined loss of the decoder module and classification module in training and testing is given in Figure 5. Figure 6. Shows the individual losses of both modules. It is to be noted that both these modules have different loss functions and different data representations as output. The network was able to maintain a steady decrease in the losses throughout the network.

Table 2. Accuracy and Area Under Curve Metrics for proposed and comparison models on Test set

Method	Accuracy	AUC
Logistic Regression	76.5	0.75
Decision Trees	76.16	0.66
Random Forest	79.74	0.7
Naïve Bayes	75.42	0.74
SVM	75.08	0.75
XGBoost	76.45	0.69
Proposed	80.57	0.841

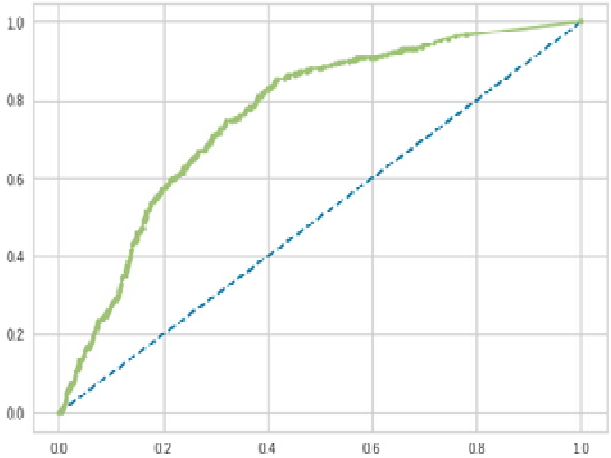

Figure 3. ROC curve

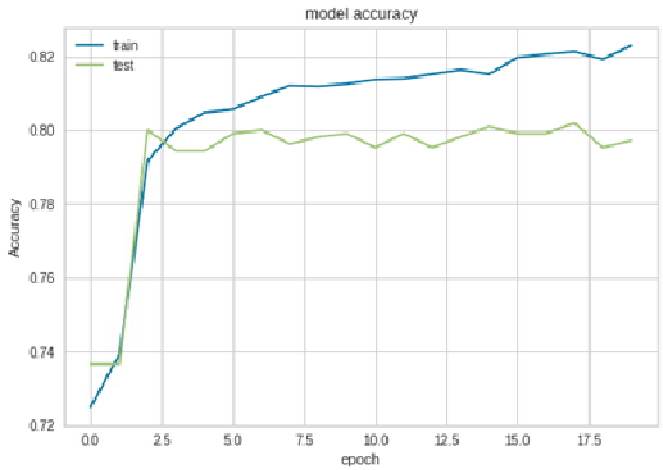

Figure 4. Model Training and Testing Accuracy

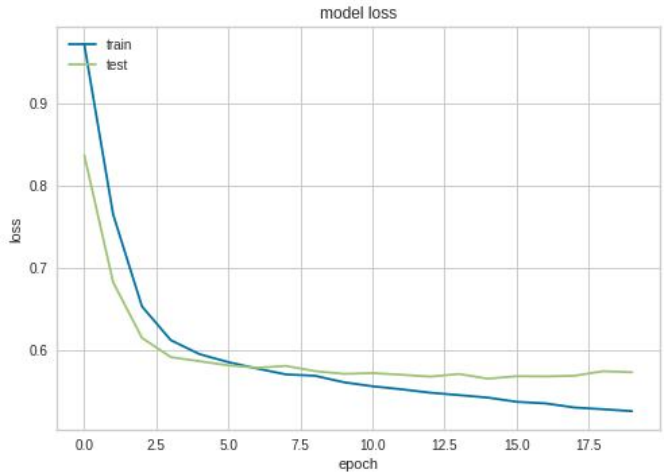

Figure 5. Training and Testing Overall loss

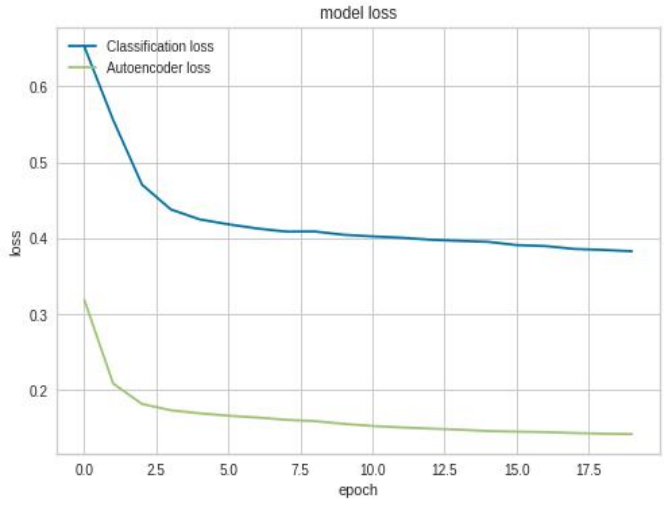

Figure 6. Autoencoder and Classification Loss Plotted Separately

6. CONCLUSION

In this paper, we proposed a autoencoder based network with attention mechanism for churn rate prediction. The autoencoder and the attention module can extract hidden features which are often difficult for other classification algorithms. Representing the user data in latent space compresses the given dataset and applying attention mechanism to the compressed data can identify the features that contribute the most to the output. The proposed network gave better results with limited training dataset, which helped in creating a simple yet powerful neural network model. The dimensionality reduction can be generalized for large amounts of data even though it is trained on limited data. The proposed model outperformed all the other traditional classification model in terms of accuracy and AUC.

REFERENCES

1. Teemu Mutanen, Sami Nousiainen, and Jussi Ahola. 2010**. "Customer churn prediction --a case study in retail banking".** *In Proceedings of the 2010 conference on Data Mining for Business Applications*, Carlos Soares and Rayid Ghani (Eds.). IOS Press, Amsterdam, The Netherlands, The Netherlands, 77-83.
2. Mohanty R., Naga Ratna Sree C. (2018), **Churn and Nonchurn of Customers in Banking Sector Using Extreme Learning Machine**. In: Bhateja V., Tavares J., Rani B., Prasad V., Raju K. (eds) *Proceedings of the Second International Conference on Computational Intelligence and Informatics. Advances in Intelligent Systems and Computing*, vol 712. Springer, Singapore. https://doi.org/10.1007/978-981-10-8228-3_6
3. Chen Y., Gel Y.R., Lyubchich V., Winship T. (2018) **Deep Ensemble Classifiers and Peer Effects Analysis for Churn Forecasting in Retail Banking**. In: Phung D., Tseng V., Webb G., Ho B., Ganji M., Rashidi L. (eds) *Advances in Knowledge Discovery and Data Mining. PAKDD 2018. Lecture Notes in Computer Science*, vol 10937. Springer, Cham. https://doi.org/10.1007/978-3-319-93034-3_30
4. De Caigny, Arno Coussement, Kristof De Bock, Koen. (2018). **A New Hybrid Classification Algorithm for Customer Churn Prediction Based on Logistic Regression and Decision Trees.** *European Journal of Operational Research.* 269. 10.1016/j.ejor.2018.02.009.
5. Wangperawong, Artit, Cyrille Brun, Olav Laudy, and Rujikorn Pavasuthipaisit." *Churn analysis using deep convolutional neural networks and autoencoders*." **arXiv preprint arXiv:**1604.05377 (2016).
6. Kingma, Diederik P. and Jimmy Ba. "**Adam: A Method for Stochastic Optimization**." *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, CoRR abs/1412.6980 (2014).

7. Bradley, Andrew. (1996). **The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms.** *Pattern Recognition*. 30. 1145 - 1159. 10.1016/S0031-3203(96)00142-2.
8. F. Chollet, "Keras," https://github.com/fchollet/keras, 2015.
9. Tschannen, Michael & Bachem, Olivier & Lucic, Mario. (2018**). Recent Advances in Autoencoder-Based Representation Learning**, *arXiv:*1812.05069.