Evaluating Customer Churn Prediction with Machine Learning and Deep Learning Models

Lavanya Peteti[1*], Harsha Vardhini Kandivalasa[2], Jahnavi Sajja[3] and Eva Patel[4]
{ lavanyapeteti7416@gmail.com[1], harshavardhinikandivalasa@gmail.com[2], ahnavi.sajja04@gmail.com[3], evapatel08@gmail.com[4] }

Department of Advanced Computer Science and Engineering, VFSTR Deemed to be University, Vadlamudi, Guntur, 522213, Andhra Pradesh, India[1, 2, 3, 4]

Abstract. Churn is one of the largest growth threats in the telecommunication industry. To manage it effectively, predictive modeling has become the strategy of choice. The following section of this paper compares some machine learning and deep learning techniques. The next part of this paper discusses a comparison of some machine learning and deep learning algorithms. In churn predictions the authors examined deep learning methods against other machine learning methods (Long Short-Term Memory - LSTM, Bi-directional LSTM - Bi-LSTM and Multi-Layer Perceptron - MLP, Support Vector Machine - SVM, Random Forest, Gradient Boosting, AdaBoost and Logistic Regression). The authors developed and evaluated the performances of deep learning and machine learning models with 667 instances in test set and 4568 instances in training set containing 20 features. The machine learning algorithms had the best accuracy from SVM and MLP, with AUC values of 0.8990 and 87.56% and 87.71% respectively. However, the deep learning methods surpassed the conventional algorithm accuracy with LSTM and Bi-LSTM having an accuracy of 91% and 90% respectively. Deep learning methods' great preprocessing, feature engineering and tuning capabilities allowed it to learn the behaviour patterns of its customers better and show higher ability to manage churn and be a part of customer retention programs.

Keywords: Customer Churn Prediction, Long Short-Term Memory, Bidirectional Long Short-Term Memory, Support Vector Machine, Multi-Layer Perceptron.

1 Introduction

Customer retention for the telco business is now as important than acquiring new customers. So many providers are providing essentially the same services at the same price, that loyalty cannot be taken for granted. Small irritations like network issues, billing faults, or failure to personalize offers easily prompt users to move to another provider. Knowing why customers depart and being able to forecast churn before it happens is now a top business imperative. Conventional analytics techniques can recognize some trends, but they are usually not adequate when it comes to detecting very subtle, time-based patterns that result in a customer's decision to depart. It enables firms to know which customers are at risk of departure and act ahead of time. There are typically two forms of churn prediction:

1. **Voluntary Churn Prediction:** The customer voluntarily chooses to exit, usually because of price dissatisfaction or superior offers.
2. **Involuntary Churn Prediction:** The customer is lost as a result of events like payment failures, inactivity accounts, or bad credit.

ICITSM-Part I 2025, April 28-29, Tiruchengode, India

DOI 10.4108/eai.28-4-2025.2357829

3. In this paper, we investigate the performance of state-of-the-art machine learning algorithms, i.e., MLP, SVM, Random Forest, Gradient Boosting, AdaBoost, and Logistic Regression for predicting customer churn. Besides these, we also study two deep learning models - LSTM and Bi-LSTM, which are specifically capable of learning temporal patterns in customer behavior and thus are great predictors of churn with higher accuracy.

2 Related Works

Various investigations have examined the use of machine learning models like Logistic Regression, Random Forest, and Support Vector Machines for customer churn prediction. Deep learning methods, especially those that are able to handle sequential behavior like LSTM and Bi-LSTM, have more recently become prominent as well. The review offered below is a comparative summary of these methods. Bin et al. (2007) [1] Employed decision tree models to forecast Personal Handyphone System (PHS) service churn. The analysis highlights decision trees' utility with telecom data. It demonstrates an early groundwork of churn modelling within telecom systems. Hadden et al. (2007) [2] Discussed contemporary approaches and trends in churn management using computers. It offers a comparative study of statistical and AI methods. The paper also discusses the directions of future research in churn analytics. Fujo et al. (2022) [3] Used deep learning for predicting churn in the telecommunication sector. The study illustrates that deep learning models demonstrate better accuracy compared to classical machine learning models. Discusses the importance of neural networks in understanding intricate customer behavior patterns. Bermejo et al. (2011) [4] Enhanced Naive Bayes Multinomial accuracy in email classification through balanced dataset distribution. While not churn-related per se, it is useful for dealing with imbalanced datasets. Applicable to preprocessing techniques in churn forecasting.

Huang et al. (2012) [5] Used data mining for telecom churn prediction. Combined various algorithms to improve prediction accuracy. Emphasizes the power of integrating multiple models and data sources. Ahmad et al. (2019) [6] Built a big data platform-based framework for telecom churn prediction via machine learning. Demonstrates scalable solutions for big data. Emphasizes practical deployment considerations of churn models. Yang et al. (2018) [7] Suggested interpretable user clustering and mobile social app churn prediction. The research integrated clustering techniques with predictive modeling to provide enhanced insights into customer churn behavior. Provides explainable AI techniques for new user behavior analysis. Miguéis et al. (2012) [8] Examined partial churn through the study of early product purchase sequences. Applies sequence modeling for forecasting customer retention. Places focus on early-stage customer behavior for forecasting future churn. Ascarza (2018) [9] Contended that serving high-risk customers could prove ineffective owing to retention futility. Proposes reassessing strategies for preventing churn. Provides insights into cost-effectiveness in intervening churn. Umayaparvathi & Iyakutti (2017) [10] Employed deep learning and automatic feature selection to predict churn. Wanted to minimize human intervention in feature engineering. Showed effective performance with less human intervention. Seymen et al. (2020) [11] Applied deep learning models for customer churn classification. Showcased DL's superiority over conventional techniques. Tested the model on telecom datasets with impressive accuracy. Momin et al. (2020) [12] Worried about churn prediction through a number of ML algorithms. Compared model performances. Highlighted early detection of churners from data that is feature-rich.

Lundberg & Lee (2017) [13] Presented SHAP (Shapley Additive explanations) for explaining ML models. Useful in churn prediction for explaining model output. A seminal paper in explainable AI. Poudel et al. (2024) [14] Utilized interpretable ML models for churn prediction in telecom to close the gap between model explainability and predictive accuracy. Employed SHAP to reveal feature contributions.

3 Methodology

The procedures employed in this paper to predict customer churn are outlined in this section. These procedures are dataset preparation, preprocessing, model architecture, training process, and evaluation metrics for the performance of the model.

3.1 Dataset Overview

The dataset for customer retention prediction [15] has training and test subsets of 4,568 and 667 samples, respectively. Each record has 20 attributes, which are a combination of continuous and discrete variables. Examples of variables are total day charge (continuous) and area code (discrete). Categorical features such as state, international plan and churn are also included. This heterogeneous dataset is ideal for predictive modeling.
The dataset summary is presented in Table 1. This summary assists in comprehending the structure of the dataset and makes research on predicting customer churn for a telecom easier. Proper data preprocessing guarantees quality data and correct model training.

Table 1. Dataset Details.

Category	Description
Training Dataset Size	4568 samples
Features	20 input features and 1 target variable
Target Variable	Churn status(categorical)

3.2 Data Preprocessing

Before training the model, there are a number of preprocessing operations that have to be done so as to make your dataset ready for learning. As the original dataset has both categorical and numerical variables, a simple approach can be to perform label encoding on categorical ones. This is so because deep learning models only accept numerical inputs. To mitigate difference in scale for numerical variables we deploy standardization and normalization techniques. All these scale conversions are required to convert all features on the same scale for better model convergence and training time performance. The importance of the preprocessing pipeline is second only to feature selection. Recursive Feature Elimination (RFE): the RFE method uses model insights to order input features by importance and then systematically removes the next least important feature for each iteration. 15 features were chosen. That helped to reduce noise and increase model performance. The dataset had class imbalance problem, one of its major issues. We can overcome this issue via SMOTE(Synthetic Minority Oversampling Technique). SMOTE oversamples the minority class by creating synthetic samples so as to encourage the

model to train with a better balance of the two classes and counteracts bias when it comes to prediction. Everything what we did all those preprocessing steps that got our data in a best condition for training LSTMs and Bi-LSTMs.

4 Model Architectures

4.1 Machine Learning Models

We choose a set of machine learning models, each chosen for their own strengths in addressing classification problems like churn prediction. Models utilized are Random Forest, MLP (Multi-Layer Perceptron), AdaBoost, Support Vector Machine (SVM), Gradient Boosting, and Logistic Regression. Random Forest was utilized as it is powerful and can learn non-linear and linear relationships by averaging a set of decision trees, and prevent overfitting. MLP Classifier is multi-hidden-layered deep neural network capable of learning complex patterns from data based on non-linear activation units and depth. AdaBoost as ensemble learner is more interested in step-by-step refining of the mistakes of the weak learner and is well adapted to data where imbalances or patterns lie beneath. SVM with an RBF kernel is added as it is useful with data of hard boundaries, especially in high feature space.

Gradient Boosting, another ensemble algorithm, constructs models sequentially through minimizing errors of earlier models and provides robust performance on structured data. Finally, Logistic Regression is added as a baseline; it's a simple and understandable model that generally works well on binary classification problems when augmented by good feature selection and scaling. By using the combination of these models, the function provides an exhaustive comparison of different learning methods to determine the best method to predict customer churn. Fig 1 shows the Architecture of Machine Learning Models.

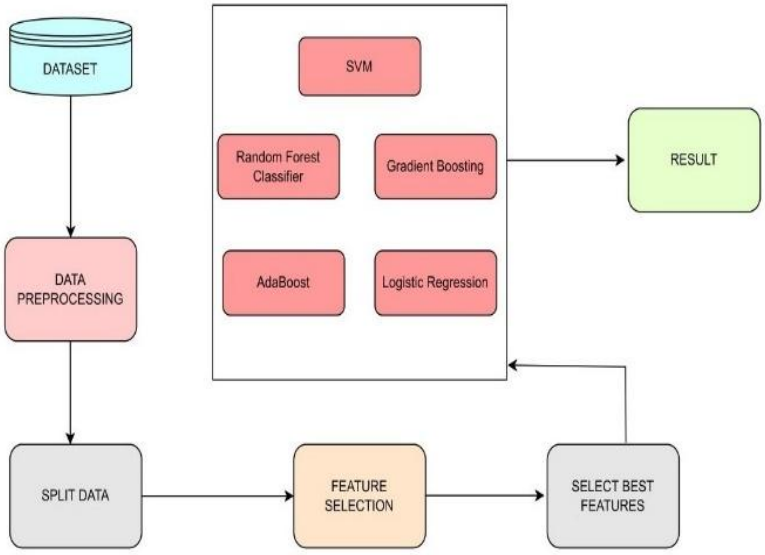

Fig. 1. Architecture of Machine Learning Models.

4.2 Bidirectional Long Short-Term Memory

We used accuracy to gauge its level of performance. We are performing a Bi-LSTM model test for churn prediction because it's actually very capable of dealing with sequential data. The configuration begins with a Bi-LSTM layer considering data forward and backward first, then followed by a Dropout to prevent overfitting. We have another Bi-LSTM and Dropout layer on top of that later on. Then the output is flattened and fed into a Dense layer with ReLU to detect intricate patterns and finished off with an output layer that outputs the probability of churn. We have quantified how well it was doing as far as accuracy goes. Fig. 2 shows the Architecture of Bi-LSTM Model. Fig. 3 shows the Architecture of LSTM Model.

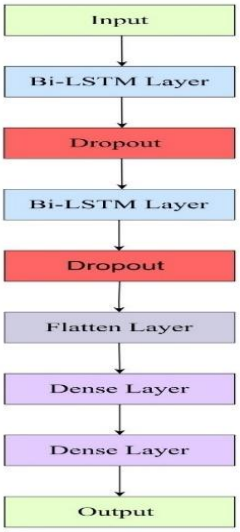

Fig. 2. Architecture of Bi-LSTM Model.

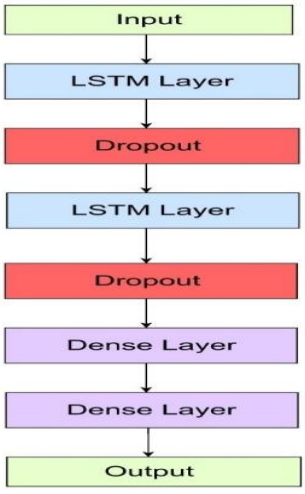

Fig. 3. Architecture of LSTM Model.

4.3 Long Short-Term Memory

The LSTM model is designed to learn patterns over time in the data. It begins with an LSTM layer that feeds its entire sequence to a second LSTM layer, allowing it to learn more complex relationships. Dropout layers are inserted in between to avoid overfitting. Following the LSTMs, it passes through a Dense layer to learn higher-level patterns, and concludes with a single output that provides the churn probability. Accuracy is employed to quantify how well it does.

4.4 Training Procedure

For ensuring best performance on all models, each algorithm was finely tuned using certain hyper parameters. The Random Forest classifier was set to have 18 estimators and a depth of 15 to manage overfitting, with a constant random state for reproducibility. The Multilayer Perceptron (MLP) model had a deep structure with five hidden layers, each having 64 neurons, and used the ReLU activation function. It was optimized with the Adam optimizer up to 1000 iterations. The AdaBoost model was initialized with 250 weak learners and a 0.1 learning rate, finding a trade-off between performance and generalization. Support Vector Machine (SVM) utilized a radial basis function (RBF) kernel with regularization constant C=10 and probability estimate to allow measures like AUC for evaluation. Gradient Boosting used tree depth as 10, min samples per leaf as 7, max leaf nodes as 5, and square root feature selection for generalization during training. Lastly, Logistic Regression was run with a strong regularization of C=10000 and was trained for 1000 iterations until convergence. Table 2 shows the Hyper Parameters of Bi-LSTM Model. Table 3 shows the Hyper Parameters of LSTM Model.

Table 2. Hyper Parameters of Bi-LSTM Model.

Layer Type	Parameters
Bidirectional LSTM	100 units, return_sequences=True
Dropout	rate=0.3
Bidirectional LSTM	50 units
Dropout	rate=0.3
Dense	32 units, activation= ReLU
Output Dense	1 unit, activation=Sigmoid
Optimizer	Adam, learning rate=0.001
Loss Function	Binary Cross entropy
Metrics	Accuracy

Table 3. Hyper Parameters of LSTM Model.

Layer Type	Parameters
LSTM	64 units, return_sequences=True
Dropout	rate=0.3
Bidirectional LSTM	64 units, return_sequences=False
Dropout	rate=0.3
Dense	32 units, activation= ReLU
Output Dense	1 unit, activation=Sigmoid
Optimizer	Adam, learning rate=0.001
Loss Function	Binary Cross entropy
Metrics	Accuracy

4.5 Performance Metrics

We employ the below measures to compare the performance of the models presented in sections 4.1 to 4.3.

Accuracy: The degree to which a measured, calculated, or predicted value corresponds to the true or accepted value. It reflects how close a result is to the correct or target value.

$$Accuracy = \frac{\mathrm{T\,P + T\,N}}{\mathrm{T\,P + T\,N + F\,P + F\,N}} \quad (1)$$

where TP, TN, FP, and FN are the true positives, true negatives, false positives, and false negatives, respectively.
Precision: Precision comprises the proportion of true positive results to all the positive results acquired in the sample, including the false detections.

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP + FP}} \quad (2)$$

Recall: Recall or sensitivity is the proportion of real positives detected correctly.

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP + FN}} \quad (3)$$

F1-Score: F1 score is a harmonic mean between recall and accuracy, preferring balances by penalizing models harder for poor performance in their measure. This makes it a crucial measure to gauge classification models, especially when both accuracy and recall are significant.

$$F1\ Score = 2 \times \frac{\mathrm{PrecisionRecall}}{\mathrm{Precision + Recall}} \quad (4)$$

5 Results and Evaluation

In this section we discuss the performance of the evaluated models in Table 4.

Table 4. Performance Comparison of Machine Learning Models.

Model	Accuracy	Precision	Recall	F1 Score
MLP	0.8771	0.5736	0.7327	0.6435
SVM	0.8756	0.5652	0.7723	0.6527
Random Forest	0.8231	0.4525	0.8020	0.5786
Gradient Boosting	0.7706	0.3889	0.9010	0.5433
AdaBoost	0.7436	0.3577	0.8713	0.5072
Logistic Regression	0.6852	0.3088	0.8713	0.4560

5.1 Receiver Operating Characteristic curve for Bi-LSTM and LSTM

Receiver Operator Curves (ROC Curves) demonstrate good performance by both models, with Bi-LSTM having a better AUC of 0.90 compared to 0.88 for LSTM. The Bi-LSTM does have an upper hand, as it learns from both the past and future samples, whereas LSTM merely observes past patterns alone. But both models are capable of identifying churners from non-churners quite well. Fig. 4. Shows the ROC curve of Bi-LSTM Model. Fig. 5 shows the ROC curve of LSTM Model.

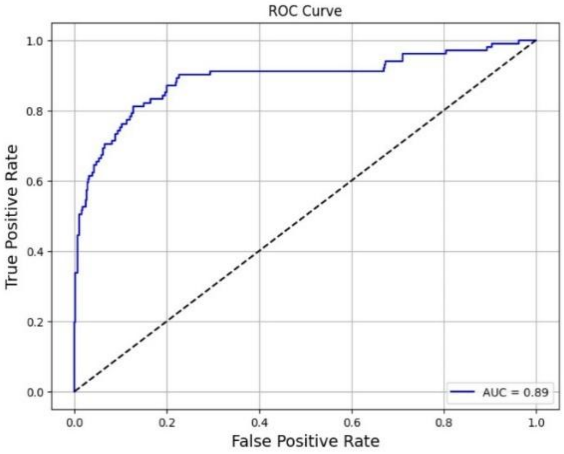

Fig. 4. ROC curve of Bi-LSTM Model.

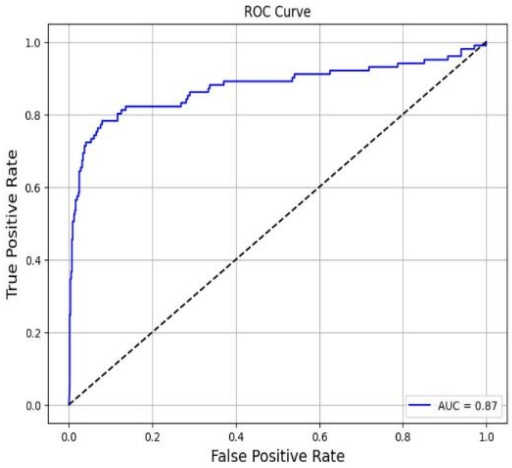

Fig. 5. ROC curve of LSTM Model.

5.2 Learning Curves of Bi-LSTM and LSTM

The first plot shows that the training loss of the Bi-LSTM model slowly decreases over 300 epochs, and validation loss also decreases though with some fluctuations reflecting good learning and generalization. The second plot, for the LSTM model over 250 epochs, is the same

too: both training and validation loss reduce, and the small gap between them reflects that the model is learning well without overfitting. Fig. 6 shows the Learning Curve of Bi- LSTM Model. Fig 7 shows the Learning Curve of LSTM Model.

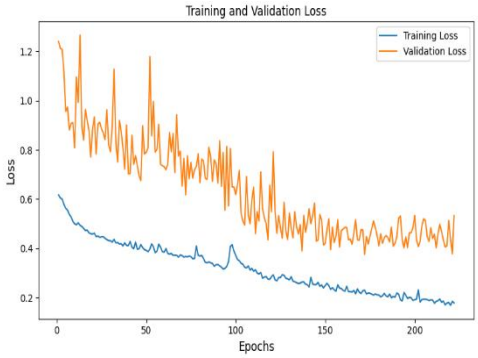

Fig. 6. Learning Curve of Bi- LSTM Model.

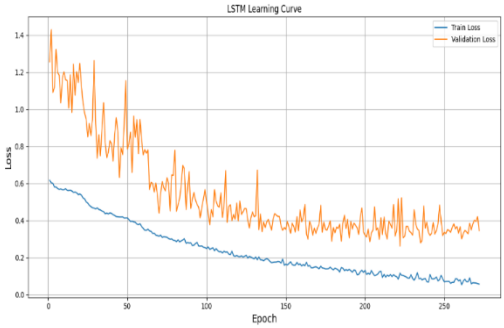

Fig. 7. Learning Curve of LSTM Model.

6 Conclusion

The research contrasted the performance of the conventional machine learning models and the newer deep learning models in the prediction of telco customer churn on a large scale. Out of the machine learning models, MLP and SVM displayed comparatively better accuracy and F1-scores reflecting their capacity to recognize non-linear data patterns. However, although very good, the models proved to be somehow limited in regard to precision-recall balance, particularly when utilized with imbalanced classes, an issue often inherent in churn prediction tasks. When compared, the deep learning-based models LSTM and Bi-LSTM were significantly improved. With maximum training and test set accuracy as 91% and 90% respectively, these models successfully learned temporal and long-range relations in customer behavioral data. In addition, the ROC curves also confirmed the better discriminative capacity of LSTM-based networks compared to traditional approaches. In general, the results unequivocally demonstrate that deep learning models, i.e., LSTM, provide a more scalable and robust solution for predicting customer churn. Their capacity to learn sequential dependencies and context positions them perfectly placed to deal with dynamic and complex data common

in the telecommunications sector. This comes into focus as the need for embracing deep learning models in constructing more proactive and efficient customer retention strategies.

References

[1] Bin, L., Peiji, S., Juan, L. (2007). Customer churn prediction based on the decision tree in personal handyphone system service. Proceedings of the International Conference on Service Systems and Service Management, 1–5.

[2] Hadden, J., Tiwari, A., Roy, R., Ruta, D. (2007). Computer assisted customer churn management: State-of-the-art and future trends. Computers Operations Research, 34(10), 2902–2917.

[3] Fujo, S. W., Subramanian, S., Khder, M. A., et al. (2022). Customer churn prediction in telecommunication industry using deep learning. Information Sciences Letters, 11(1), 24.

[4] Bermejo, P., G´amez, J. A., Puerta, J. M. (2011). Improving the performance of Naive Bayes multinomial in e-mail foldering by introducing distribution-based balance of datasets. Expert Systems with Applications, 38(3), 2072–2080.

[5] Huang, B., Kechadi, M. T., Buckley, B. (2012). Customer churn prediction in telecommunications. Expert Systems with Applications, 39(1), 1414–1425.

[6] Ahmad, A. K., Jafar, A., Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. Journal of Big Data, 6(1), 1–24.

[7] Yang, C., Shi, X., Jie, L., Han, J. (2018). I know you'll be back: Interpretable new user clustering and churn prediction on a mobile social application. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery Data Mining, 914–922

[8] Migu´ eis, V. L., Van den Poel, D., Camanho, A. S., e Cunha, J. F. (2012). Modeling partial customer churn: On the value of first product category purchase sequences. Expert Systems with Applications, 39(12), 11250–11256.

[9] Ascarza, E. (2018). Retention futility: Targeting high-risk customers might be ineffective. Journal of Marketing Research, 55(1), 80–98.

[10] Umayaparvathi, V., Iyakutti, K. (2017). Automated feature selection and churn prediction using deep learning models. International Research Journal of Engineering and Technology (IRJET), 4(3), 1846–1854.

[11] Seymen, O. F., Dogan, O., Hiziroglu, A. (2020). Customer churn pre-diction using deep learning. International Conference on Soft Computing and Pattern Recognition, 520–529.

[12] Momin, S., Bohra, T., Raut, P. (2020). Prediction of customer churn using machine learning. EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing, 203–212.

[13] Lundberg, S. M., Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.

[14] Poudel, S. S., Pokharel, S., Timilsina, M. (2024). Explaining customer churn prediction in telecom industry using tabular machine learning models. Machine Learning with Applications, 17, Article 100567.

[15] https://www.kaggle.com/datasets/spscientist/telecom-data

[16] M. Ranjith Kumar, P. S, J. Srinivasan Anusha, V. Chatiyode, J. Santiago and D. Chaudhary, "Enhancing Telecommunications Customer Retention: A Deep Learning Approach Using LSTM for Predictive Churn Analysis," *2024 International Conference on Data Science and Network Security (ICDSNS)*, Tiptur, India, 2024, pp. 01-07, doi: 10.1109/ICDSNS62112.2024.10691038.