Original Research Article

Telco Customer Churn Prediction Using Machine Learning Method

.

ABSTRACT

Customer churn is an important issue in businesses and service sectors including telecommunication industries. Prediction of potential customer churn can be very useful in these fields and it can help to improve customer retention significantly by providing personalized offering to reduce potential churn. This paper mainly focused on customer churn prediction using machine learning models and Iranian Telco customer churn dataset. Different reasons or variables are involved in customer dissatisfaction or indication of customer's churn. If ML models are trained with such essential and crucial data, they can provide better prediction with high accuracy. This study experimented 16 ML models (single as well as ensemble) and the model performances were evaluated in two ways - 80:20 train-test split and 10-fold cross-validation. In 80:20 split, categorical boosting CatBoost classifier outperformed other models with 97.54% accuracy. However, Light Gradient Boosting Machine classifier LGBMC performed best in 10-fold cross-validation by achieving average accuracy of 96.39% with 1.08% standard deviation while CatBoost classifier was second best performing model. Thus, machine learning techniques can very effectively serve Telco industry as well as other customer-oriented competitive businesses.

Keywords: Customer Churn Prediction, Telco Industry, Machine Learning, Artificial Intelligence, LightGBM, CatBoost, Histogram-based Gradient Boosting

1. INTRODUCTION

Customer churn is a problem for many industries and for the highly competitive businesses, it is a crucial issue. In a highly saturated telecom market where acquiring a new customer it is really difficult, it is much cheaper to retain a customer in that scenario. Therefore, customer churn prediction has become the key component in business strategy for telecom companies. Taking action proactively through targeted retention strategies, personalized offers and service improvements, companies can predict those customers who are most likely to leave, so that they can insure customer loyalty and low revenue loss.

With the advancement of computing technology and availability of large data, currently ML techniques have become powerful tools for data analytics and prediction like customer churn prediction. Data science as well as machine learning can potentially provide powerful insights and better prediction by handling complex patterns of data than normal manual review. These approaches may quickly uncover customers' behavior. Given historical data such as customers' demographics, usage patterns, service interactions, and payments history, ML models can make more accurate and contingent predictions about churn risk.

The emergence of big data and AI in the telecom is growing. The analytic insights and accuracy of churn prediction can help telecom companies understand more about customer behavior, minimize churn and achieve a competitive edge to help them withstand challenges from other telecom companies.

Many research studies investigated customer churn insights and predictive models including ML techniques. The review work of Manzoor *et al.* (2024) scrutinized 212 articles (published during 2015-2023) on predicting customer churn and also included recommendations for Practitioners. Also, De & Prabu (2022) explored literature involving 420 papers (published during 2018-2021). The work of Huang, Kechadi, & Buckley (2012) focused on churn prediction of land-line customers and applied 07 ML models where C4.5 model and support vector machine were more effective than other models. A Swish-RNN-based recurrent neural network was employed in the study of Sudharsan & Ganesh (2022) for predicting customer churn in the telecom industry. Other recent deep learning based works encompass (De Caigny *et al.*, 2020; Pondel *et al.*, 2021; Fujo *et al.*, 2022; Khattak *et al.*, 2023). In predicting customer churn in B2C E-commerce, Xiahou & Harada (2022) found that support vector machine performed better than logistic regression. The study of Al-Najjar, Al-Rousan, & Al-Najjar, (2022) focused ML model applications in churn prediction for credit card customers. Many works on customer churn prediction are applied in various fields, e.g., web browsers (Wu *et al.*, 2022), influencer commerce (Kim & Lee, 2022), telecom (Vo *et al.*, 2021), broadcast industry (Li *et al.*, 2021), banking (Rahman & Kumar, 2020), rental business of home appliance (Suh, 2023). For predicting churn of banking customers, Agarwal *et al.* (2022) applied ML models where Naive Bayes performed better than logistic regression. Ensemble models were mainly focused on the study of Liu *et al.*, (2023) for churn prediction of telecom customer. In predicting customer churn, Lalwani *et al.* (2022) employed popular predictive ML models and two best performing models were AdaBoost (with 81.71% accuracy) and XGBoost (with 80.8% accuracy). XGBOOST was found as best performing algorithm in study of Ahmad, Jafar, & Aljoumaa (2019). Random Forest was found best performing method (with 96.25% accuracy) in the study by Gurung *et al.* (2024). For prediction of customer churn, Mouli *et al.* (2024) also experimented 10 ML classification models of which random forest was best performing.

As said by Andrew Ng "AI is the New Electricity", artificial intelligence (AI) has started to revolutionize the world. Also, Machine learning is the essential part of AI. Since machine learning (ML) models are very useful data-driven techniques and applied in different fields, they might be applied effectively to customer churn prediction. Starting with this hypothesis, study plan was to test the effectiveness of various ML models in Telco industries and for dataset Iranian Telco customer churn dataset was chosen. The remaining sections encompass descriptions of dataset and experimented ML models, presentation of results, discussion and finally conclusion which reflect the applicability of ML models in this area.

2. MATERIAL AND METHODS

This section contains description of datasets, experimented machine learning models, the disease prediction system and evaluation metrics.

2.1Dataset Description

The Iranian Telco customer churn dataset experimented in this study was collected from UCI ML repository (‘Iranian Churn Data’, n.d.) which was randomly collected for over one year period. It has 2850 distinct samples with 13 feature variables excluding target feature. The features are presented with description and values in Table 1.

Table 1. Feature variables with description and values of Iranian telco customer churn dataset

Feature	Description	Value
Call Failures	Number of failed calls	Numeric
Complains	Customer complains of any form	Binary (0 for No complaint or1 forcomplaint)
Subscription Length	Total months of subscription	Numeric
Charge Amount	Type of charge amount	Ordinal (0-9, i.e., lowest-highest amount)
Seconds of Use	Total time of calls (in seconds)	Numeric
Frequency of use	Total number of calls	Numeric
Frequency of SMS	Total number of short text messages	Numeric
DistinctCalled Numbers	Total number of distinct phone calls	Numeric
Age Group	Customer’s Age Group	Ordinal (1-5, i.e., younger-older age)
Tariff Plan	Customer’s chosen tariff plan	Binary (1 for Pay-as-you-go; 2 for contractual)
Status	Status of use	Binary (1 for activeor 2for non-active)
Customer Value	The calculated value of customer	Numeric
Churn	Customer’s churn status	Binary (1 for churnor 0for non-churn)

The statistical summary of the features are presented in Table 2 which includes mean, standard deviation (std), minimum (min), first quartile (25%), second quartile or median (50%), third quartile (75%) and maximum (max).

Table 2. Descriptive statistics of Iranian Telco customer churn dataset

Feature	count	mean	std	min	25%	50%	75%	max
Call Failure	2850	7.80	7.33	0	1	6	12	36
Complains	2850	0.08	0.27	0	0	0	0	1
Subscription Length	2850	32.45	8.72	3	29	35	38	47
Charge Amount	2850	0.97	1.55	0	0	0	2	10
Seconds of Use	2850	4534.24	4199.71	0	1458.8	3041	6500.0	17090
Frequency of use	2850	70.48	57.40	0	28	54.5	96	255
Frequency of SMS	2850	73.79	112.06	0	7	22	88	522
Distinct Called Numbers	2850	23.87	17.19	0	11	21	34	97

Age Group	2850	2.84	0.89	1	2	3	3	5
Tariff Plan	2850	1.08	0.27	1	1	1	1	2
Status	2850	1.24	0.43	1	1	1	1	2
Age	2850	31.08	8.86	15	25	30	30	55
Customer Value	2850	474.99	514.44	0	117.5	232.5	790.1	2165.3
Churn	2850	0.16	0.36	0	0	0	0	1

Correlations among variables in the Iranian Telco churn data are presented in the following correlation matrix in Fig. 1. It shows that high positive correlations of churn with other variables are correlation between churn and complains is 0.55, correlation between churn and status is 0.49. The high negative correlations of churn with other variables are correlation between churn and seconds of use is -0.3, correlation between churn and frequency of use is -0.3, correlation between churn and customer value is -0.29, correlation between churn and distinct call numbers is -0.27, correlation between churn and frequency of SMS is -0.22, correlation between churn and charge amount is -0.2, correlation between churn and tariff plan is -0.11.

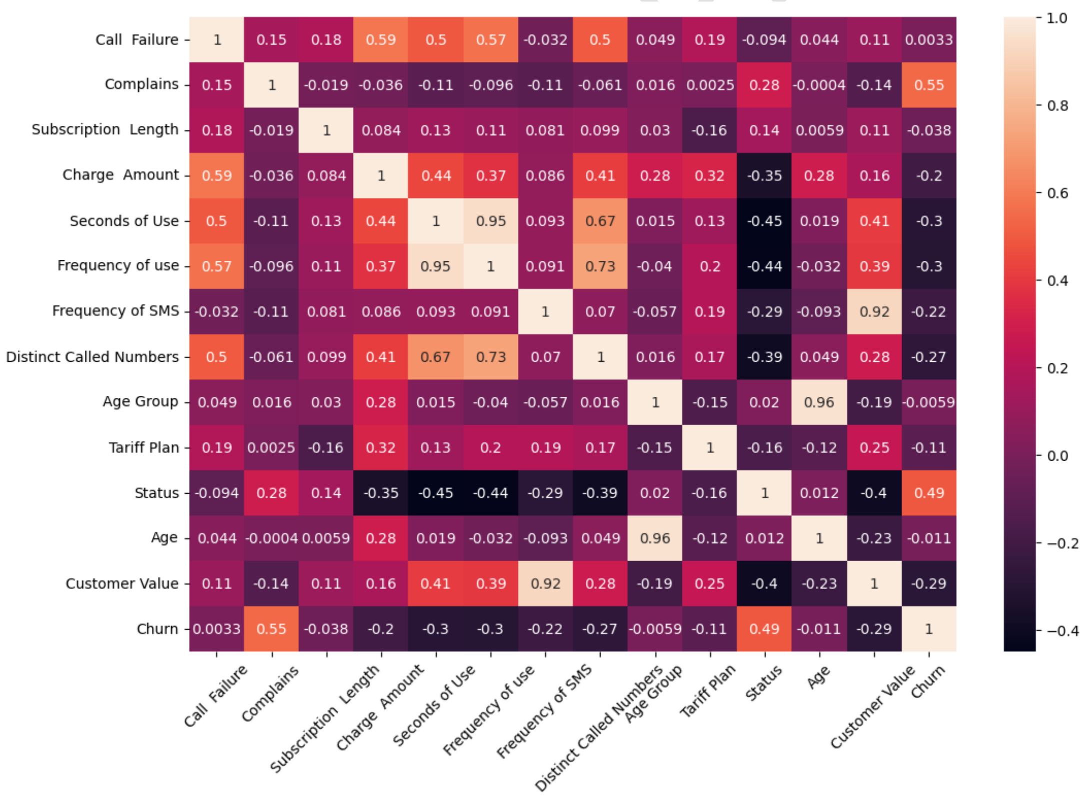

Fig. 1.Correlation matrix of input features and target feature/class of Iranian Telco customer churn.

2.1.1 Train-Test Split

The Iranian Telco customer churn data has 2850 distinct instances. The train-test split experimented in this study was 80:20 (i.e., 2280 training samples and 570 test samples).

2.2 Machine Learning Models

This study employed 16 ML classification for predicting Iranian Telco customer churn dataset. Seven of these are single benchmark models; nine others are ensemble models. The single methods encompass Decision Tree Classifier (DTC), Gaussian Naïve Bayes Classifier (GNBC), k-Nearest Neighbors Classifier (kNNC), Linear Discriminant Analysis (LDA), Logistic Regression (LR), Quadratic Discriminant Analysis (QDA) and Support Vector Classifier (SVC). The ensemble models are AdaBoost Classifier (ABC), Bagging Classifier (BC), Categorical Boosting Classifier (CBC), ExtraTrees Classifier (ETC), Gradient Boosting Machine Classifier (GBMC), Histogram-based Gradient Boosting Classifier (HGBC), Light Gradient Boosting Machine Classifier (LGBMC), Random Forest Classifier (RFC) and Extreme Gradient Boosting Classifier (XGBC). These models are briefly described in the following sections.

2.2.1 Logistic Regression (LR)

LR is a very simple and useful algorithm for ML classification task. In particular, it is used to predict the outcome as the use of a logistic (sigmoid) function. It works for both binary and multi-class classification. As the class with the highest probability, it models the relationship between the input features and the class labels and hence predicts the output of which class the input belongs. Assuming the data is linearly separable, LR is efficient and effective. It is used widely in medical diagnosis, spam detection and risk assessment because the interpretation of results is simple.

2.2.2 Support Vector Classifier

Support Vector Classifier (SVC) is a powerful ML algorithm used to do classification tasks. This algorithm works by finding boundary (or hyperplane) based on which data points belonging to different classes should be separated. A common goal of SVC is to maximize the margin, which is the difference between the hyperplane and nearest data points, to better generalize.SVC works well for situations where classifications are linear or nonlinear, with the use of kernel functions to deal with complexities in data. With high dimensional data, it works well and has applications for image recognition, text classification, and bioinformatics. Accuracy and also robustness are its strengths.

2.2.3 k-Nearest Neighbors Classifier

k-nearest neighbor classifier (k-NNC) is a simple and intuitive algorithm for classification tasks. Given an input, it picks up k-nearest data points and determines the most common label of k number of them by calculating how far apart using distance measures (i.e. Euclidean distance). k-NN implementation is simple and no training phase is required. However, kNNC is very sensitive to both the choice of k and the distance metric, and is slow. Its applications encompass recommendation systems, pattern recognition and anomaly detection.

2.2.4 Gaussian Naïve Bayes Classifier

Gaussian Naïve Bayes Classifier (GNBC) is a simple and efficient classification algorithm based on Bayes' theorem. GNBC assumes that features are independent and normally distributed. With these assumptions, GNBC computes probability of each class and outputs input to the class with the highest probability of it. It is simple to implement, can take small data, and is very flexible to problems which hold some (approximate) independence assumption. Its diverse applications range from spam detection to medical diagnosis and text classifications.

2.2.5 Decision Tree Classifier

Decision Tree Classifier (DTC) is a simple and an intuitive ML algorithm used in classification tasks. It works by creating a tree like structure, splitting data into branches by feature value. A decision is represented by a node and a class label by a leaf. One of the reasons why DTCs are used is because they are easy to understand, interpret, and visualize. Both numerical and categorical data are handled DTCs and they work for small to medium sized datasets. However, if not managed well, they can overfit the data. Examples of application include medical diagnosis, credit scoring and customer segmentation.

2.2.6 Random Forest Classifier

The Random Forest Classifier (RFC) is a very powerful and versatile ML algorithm for classification. In an RFC algorithm, majority voting on several decision trees (built with the random subsets of data and features) is used. There are several important advantages of this method. Generally, RFC does not overfit, it can handle huge datasets and lots of features. Often, it is useful in applications like fraud detection, data-mining, recommendation system and healthcare. Also, it possesses robustness, scalability and can provide feature importance rankings.

2.2.7 Gradient Boosting Machine Classifier

Gradient Boosting Machine Classifier (GBMC) is one of the very useful algorithms used for ML classification problems. It makes series of decision trees one at a time and each tree tries to correct errors that are not fixed by their previous trees. This improves over time making prediction accuracy better. GBMC is very flexible and can handle complex dataset. However, it is computationally expensive and requires tuning the parameters too. This is used widely in applications such as fraud detection, customer segmentation and predictive modeling. Its strengths are precision and the ability to capture complex patterns in data.

2.2.8 Light Gradient Boosting Machine Classifier

Light Gradient Boosting Machine Classifier (LGBMC) is one of the fast and efficient ML algorithm for classification. It allows building decision trees sequentially using gradient boosting to learn and each tree corrects the previous ones' errors. LGBMC tries to be fast and performant by using histogram-based learning and leaf-wise tree growth.

As a result, this method is both effective for treating large data and high dimensional datasets while attaining high accuracy. On financial modeling, recommendation systems and predictive analytics, LGBMC is widely used. The strength or good sides of this algorithm are that it is very fast, scalable and it works well on complex datasets.

2.2.9 Extreme Gradient Boosting Classifier (XGBoost)

Extreme Gradient Boosting Classifier (XGBC) is very efficient and robust ML algorithm used to solve classification problems. Generally, it attempts to improve results by trading optimizing speed and performance and in the process it utilizes parallel processing, regularization, and tree pruning. In XGBC, decision trees are built off sequentially and each tree tries to correct the errors done by previous trees. It is very adaptive and it works well with large datasets. XGBC is used in various classification tasks such as fraud detection, recommendation systems and predictive analytics.

2.2.10 Categorical Boosting Classifier

Categorical Boosting or CatBoost Classifier (CBC) is great and fast ML classifier for working on categorical data. It is a kind of gradient boosting method and its remarkable function is that it processes categorical feature automatically without need for massive preprocessing. CatBoost is highly accurate, fast and very robust to overfitting. Large datasets are easily processed by this algorithm and it has been used in customer segmentation, recommendation system, fraud detection and credit scoring.

2.2.11 ExtraTrees Classifier

Like Random Forest, the ExtraTrees classifier (ETC) is an ML algorithm which creates a large number of decision trees to learn for classification tasks. Unlike other tree-based methods, it it makes random splits while making its splits at each node of the tree to increase the diversity of the trees and subsequently reduce overfitting. In general, ETC is fast, robust and very accurate even for large datasets with many features. It is a very useful algorithm to handle numerical as well as categorical data and hence it is effective in solving diverse problems. Also, speed, simplicity and outstanding performance, especially when dealing with large data are the key advantages of ETC.

2.2.12 AdaBoost Classifier

Adaptive Boosting or AdaBoost classifier is an ML algorithm which boosts the accuracy of weak classifiers by taking them and combining them into a more accurate strong model. It trains a series of classifiers in sequence and the next classifier concentrates on a mistake that the previous one made. All the classifiers predict and the final prediction is made by combining the predictions of all the classifiers. AdaBoost is both simple, and effective, and has the ability to boost the performance of weak models. For binary classification tasks it works well and is sensitive to noisy data. AdaBoost is used very widely in face detection, text classification and image recognition. This approach provides for boosting accuracy while keeping computational efficiency.

2.2.13 Linear Discriminant Analysis

Linear discriminant analysis (LDA) classifier is used for classification tasks including binary as well as multiple classes. LDA mainly finds a linear combination of features that could best separate classes in the dataset. The assumption of LDA is that data samples from each class are Gaussian distributed and have the same covariance. LDA is simple, efficient and it works well when the training classes are linearly separable. In face recognition, medical diagnosis and spam detection tasks, LDA is often used. A special characteristic of this algorithm is dimensionality reduction technique, i.e., it reduces dimensionality preserving some or all of the class separability.

2.2.14 Quadratic Discriminant Analysis

Quadratic Discriminant Analysis (QDA) is an ML classifier algorithm particularly suitable for the case when classes are not separable in linear fashion. QDA is better than Linear Discriminant Analysis (LDA) in a sense because QDA can hold different covariance to each class, which enables it to deal with more complex defined feature spaces. QDA fits the quadratic function to each class of the data. When LDA assumptions are not met, it is a good approach to use QDA. However, using more data would help evade overfitting. Common QDA applications include speech recognition, medical diagnostics, and pattern recognition. Advantages of QDA are that it can handle non-linear decision boundaries well and it is flexible with class specific covariances.

2.2.15 Histogram-based Gradient Boosting Classifier

The Histogram based Gradient Boosting Classifier (HGBC) is an ML algorithm which increases the performance of gradient boosting by utilizing histogram to speed up the learning process. Instead of using the exact values of the data, it bins its values into bins decreasing the computational complexity which is very good for big data sets. HGBC builds decision trees sequentially and each new tree solves the problems made by previous tree. The histogram based approach showed high accuracy and higher efficiency. Speed, efficiency and high performance on large datasets, are its key strengths.

2.2.16 Bagging Classifier

Bagging Classifier (BC) is a popular ML classifier used to enhance the accuracy of a model by training many copies of the same base model on many subsets of the dataset. Finally, prediction is done by combining each model's predictions (usually through majority voting when the task is classification). Like bagging creates multiple training sets. The good thing about this technique is that since it tries to reduce variance it prevents from overfitting especially with unstable models like decision trees. BC is widely used algorithm like Random Forest. It is known for improved accuracy, robustness and ability to handle noisy data.

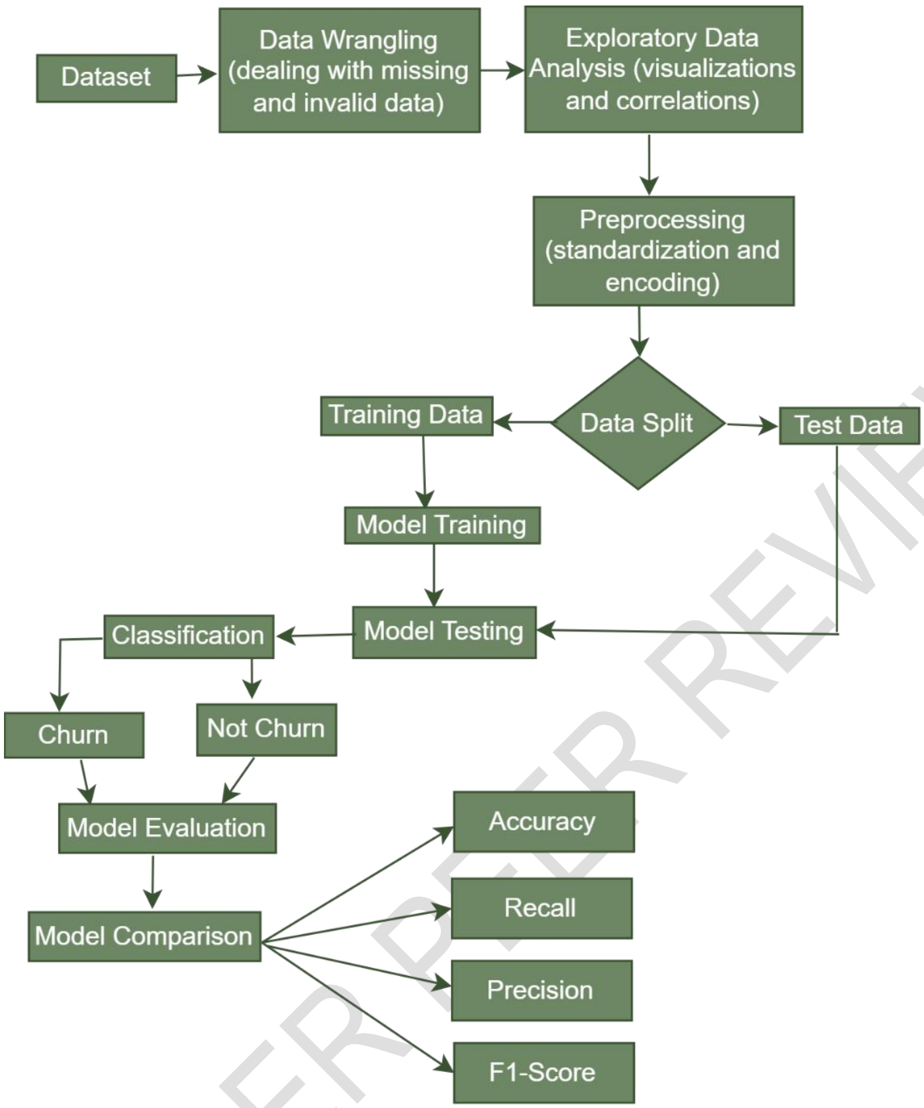

Fig. 2. Flowchart of machine learning techniques implementation for Iranian Telco customer churn prediction

2.4 Performance Metrics

The most commonly used evaluation metrics of ML classification problems are accuracy, Recall, precision and F1-score. These four metrics used in this study are defined below: $\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}$, $\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}$, $\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}$, $\text{F}_1 = \frac{\text{2TP}}{\text{2TP+FP+FN}}$
where True Positive (TP) is a positive value which is predicted correctly; while False Positive (FP) is actually a negative value which is incorrectly predicted as positive. Also, False Negative (FN) is actually a positive value which is incorrectly predicted as a negative value; while True Negative (TN) is a negative value which is predicted correctly.

Also, ROC curve is widely used visual performance tool for binary classification. A receiver operating characteristic (ROC) curve is the graph of true positive rate ($\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP+FN}}$) against false positive rate ($\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP+TN}}$) at each threshold.

2.5 Cross-validation (k-fold)

To validate performance of ML models against data bias, cross-validation method plays important role to find robust and stable model. Frequently used such method is k-fold cross validation. In this process the original dataset is divided into "k" equal folds. The model is trained on (k-1) folds and tested on the remaining fold. This approach is iterated (k) times, with each fold becoming a test set once. Finally, average performance is computed by averaging performance metrics of each result found for all folds. Advantages of cross-validation are that it reduces overfitting, uses data efficiently and gives a stable performance estimate. However, it is computationally expensive specially for large datasets. A k-fold cross-validation is shown visually in Fig. 3.

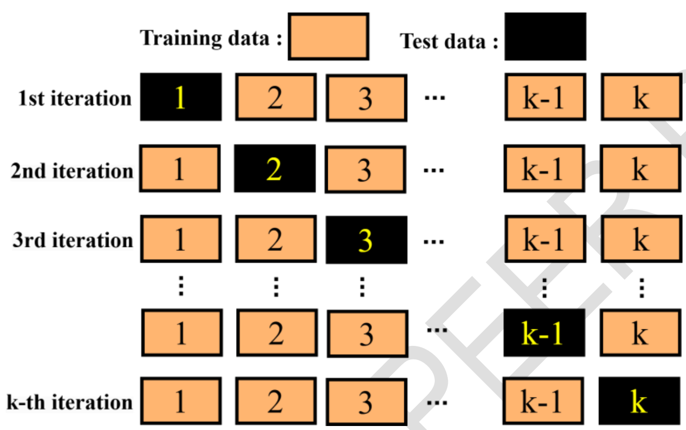

Fig. 3. K-fold cross-validation.

3. RESULTS AND DISCUSSION

The prediction performances of 16 experimented ML models on Iranian Telco customer dataset are presented in the subsequent tables and graphs. The results on hold-out test data for 80:20 train-test split are shown in Table 3. (with accuracy, recall, precision and F1-score and), Fig. 4. (accuracy comparison plot) and Fig. 5. (ROC curve). Based on these results ensemble model CBC was found to be the best performing model with accuracy 97.54%, precision 95.72%, F1-score 95.3% and recall 94.88%; while second best model was HGBC. Also, single model GNBC performed comparatively very poorly among all models with accuracy 72.98%, precision 66.88%, F1-score 66.42% and recall 80.79%.

In addition to prediction with 80:20 train-test split, 10-fold cross-validation was implemented on the ML models with the Iranian Telco customer churn prediction. The results (Table 4.) of 10-fold cross-validation reflect that LGBMC outperforming among the models with average accuracy 96.39% and standard deviation 1.08%. Also, CBC was found to be second best performing model with average accuracy 96.38% and standard deviation 1.15%. However, with least standard deviation (0.81%) and comparative better accuracy, ETC was winner with average accuracy of 95.83%. Also, based on 10-fold cross-validation, the worst performing model was GNBC with average accuracy 73.58% and standard deviation 2.58%.

Table 3. Prediction performance of 16 ML models on Iranian telco customer churn dataset.

Model	Precision	Recall	F1	Accuracy
CBC	95.72	94.88	95.3	97.54
HGBC	95.19	94.78	94.98	97.37
RFC	94.9	93.65	94.26	97.02
LGBMC	94.01	94.01	94.01	96.84
XGBC	94.75	93.09	93.9	96.84
ETC	94.21	92.99	93.59	96.67
kNNC	93.04	90.64	91.78	95.79
BC	92.67	91.1	91.86	95.79
GBMC	92.54	88.95	90.62	95.26
ABC	89.97	85.83	87.72	93.86
DTC	91.72	82.42	86.19	93.51
SVC	87.87	71.74	76.62	90.18
LR	85.36	73.01	77.21	90
LDA	84.21	73.37	77.22	89.82
QDA	71.41	85.82	73.71	80.7
GNBC	66.88	80.79	66.42	72.98

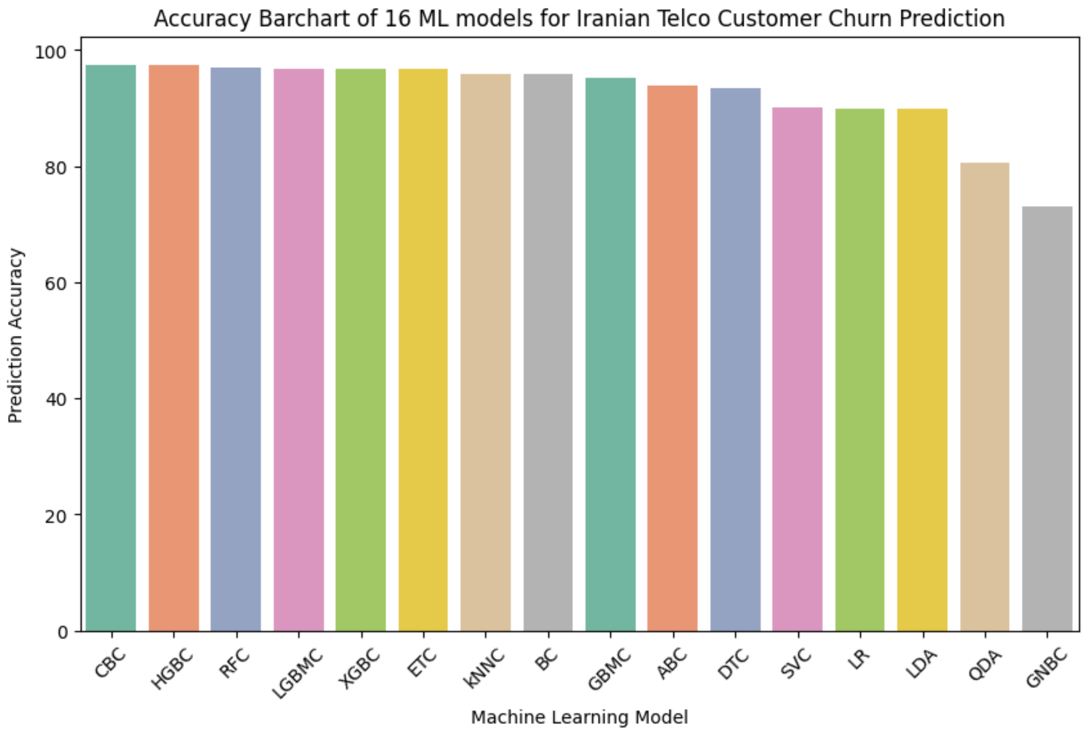

Fig. 4. Barchart of 16 ML Models Test Accuracy with Iranian Telco Customer Churn Prediction.

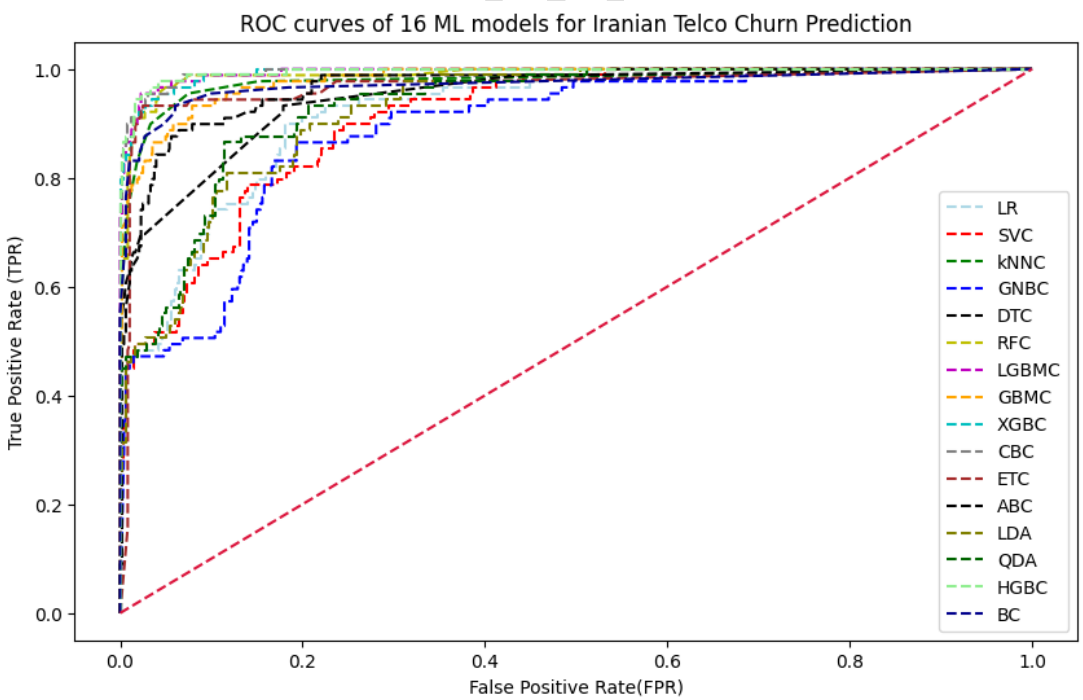

Fig. 5. ROC curves of 16 ML Models with Iranian Telco Customer Churn Prediction.

Table 4. Average accuracy with standard deviation of 10-fold cross-validation achieved by 16 ML models on Iranian Telco customer churn dataset.

Model	**Average Accuracy**	**Std**
LGBMC	96.39	1.08
CBC	96.38	1.15
HGBC	96.29	1.08
XGBC	96.01	1.01
ETC	95.83	0.81
RFC	95.66	1.00
kNNC	95.06	1.26
BC	94.98	1.41
GBMC	94.33	1.01
ABC	92.93	1.52
DTC	92.56	1.41
LR	89.94	1.71
SVC	89.86	1.50
LDA	89.70	1.82
QDA	81.47	1.92
GNBC	73.58	2.58

All models are not same. Each ML model has its strength and weakness. Also, data from different fields and subfields are diverse and of different pattern. Learning of 16 different models from the same training data of same dataset (i.e., Iranian Telco customer churn data in this study) was distinct and hence they achieved distinct results for same test data. This is due to data pattern and varied model capacity of handle provided data. As found in results of 80:20 split, most better-performing models were ensemble models with an exception of kNNC. Gradient boosting algorithms and tree-based models achieved greater than 95% accuracy except ABC. Also, CBC, HGBC and RFC achieved greater than 97% accuracy which indicates that they were highly capable of reducing overfitting. Similar performance was found in 10-fold cross-validation. The models which achieved greater than 95% average accuracy were ensemble models (both tree-based and gradient boosting) except kNNC. Also, these high performing ensemble models showed low variation with standard deviation less than 1.15%. However, considering both assessment of 80:20 split and 10-fold cross-validation, QDA and GNBC performing comparatively very poorly. In addition, their predictive variations were comparative high (i.e., 1.92% and 2.58% respectively) among all models.

Limitation of this study is that only 16 specific ML models were experimented on one medium-sized dataset (with 2850 samples and 14 features including target feature). Hence, for practical implementation especially for deployment further researches are required.

This research study can be useful to ML practitioners, researchers and business people especially from Telco industries who are interested on the applicability and effectiveness of ML models. It can also inspire other areas of research and application involving data-driven prediction using ML models.

4. CONCLUSION

To predict customer churn using Iranian Telco customer churn dataset, 16 machine learning models (both single and ensemble) were experimented in this paper. Both 80:20 train-test split and 10-fold cross-validation were implemented. Based on the results of 80:20 split, CatBoost classifier (CBC) outperformed all other models with 97.54% accuracy and 95.3% F1-score. In 10-fold cross-validation, LightGBM classifier (LGBMC) and CBC performed very close with average accuracy 96.39% and 96.38% respectively. However, LGBMC was comparatively better in terms of lower standard deviation (i.e., 1.08% for LGBMC and 1.15% for CBC). Both 80:20 split and 10-fold cross-validation show that outperforming ML models achieved high prediction accuracy for Telco customer churn. This reflects that ML techniques (with analytics and predictive system) can assist Telco industry by predicting potential churn in and take steps to prevent churn by providing improved customer care and personalized services. If automation is implemented in each segment of the ML process (using error-correcting pipeline), it will be great comfort at the user-level. Thus quality data and stable ML models can serve Telco industry as well as other sectors especially where adequate data availability can be ensured. Future work plan includes data-driven predictive modeling with machine learning and deep learning in business, economics and healthcare sector.

Table 5. Elaborations of abbreviations

	Abbreviation	Elaboration	Type
1	ABC	AdaBoost Classifier	Ensemble
2	BC	Bagging Classifier	Ensemble
3	CBC	Categorical Boosting Classifier	Ensemble
4	DTC	Decision Tree Classifier	Single
5	ETC	ExtraTrees Classifier	Ensemble
6	GBMC	Gradient Boosting Machine Classifier	Ensemble
7	GNBC	Gaussian Naïve Bayes Classifier	Single
8	HGBC	Histogram-based Gradient Boosting Classifier	Ensemble
9	kNNC	k-Nearest Neighbors Classifier	Single
10	LDA	Linear Discriminant Analysis	Single
11	LGBMC	Light Gradient Boosting Machine Classifier	Ensemble
12	LR	Logistic Regression	Single
13	ML	Machine Learning	-
14	QDA	Quadratic Discriminant Analysis	Single
15	RFC	Random Forest Classifier	Ensemble
16	SVC	Support Vector Classifier	Single
17	XGBC	Extreme Gradient Boosting Classifier	Ensemble

REFERENCES

Huang, B., Kechadi, M. T., & Buckley, B. (2012). Customer churn prediction in telecommunications. Expert Syst. Appl. **39**, 1414–1425.

Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. J. Big Data **6**, 1–24.

De Caigny, A., Coussement, K., De Bock, K. W., & Lessmann, S. (2020). Incorporating textual information in customer churn prediction models based on a convolutional neural network. Int. J. Forecast. **36**, 1563–1578.

Rahman, M., & Kumar, V. (2020). Machine learning based customer churn prediction in banking. In *2020 4th international conference on electronics, communication and aerospace technology (ICECA)*: 1196–1201. IEEE.

Li, Y., Hou, B., Wu, Y., Zhao, D., Xie, A., & Zou, P. (2021). Giant fight: Customer churn prediction in traditional broadcast industry. J. Bus. Res. **131**, 630–639.

Pondel, M., Wuczyński, M., Gryncewicz, W., Łysik, Ł., Hernes, M., Rot, A., & Kozina, A. (2021). Deep learning for customer churn prediction in e-commerce decision support. In *Business Information Systems*: 3–12.

Vo, N. N., Liu, S., Li, X., & Xu, G. (2021). Leveraging unstructured call log data for customer churn prediction. Knowl.-Based Syst. **212**, 106586.

Agarwal, V., Taware, S., Yadav, S. A., Gangodkar, D., Rao, A., & Srivastav, V. (2022). Customer-Churn Prediction Using Machine Learning. In *2022 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS)*: 893–899. IEEE.

Al-Najjar, D., Al-Rousan, N., & Al-Najjar, H. (2022). Machine learning to develop credit card customer churn prediction. J. Theor. Appl. Electron. Commer. Res. **17**, 1529–1542.

De, S., & Prabu, P. (2022). Predicting customer churn: A systematic literature review. J. Discrete Math. Sci. Cryptogr. **25**, 1965–1985.

Fujo, S. W., Subramanian, S., Khder, M. A., & others. (2022). Customer churn prediction in telecommunication industry using deep learning. Inf. Sci. Lett. **11**, 24.

Kim, S., & Lee, H. (2022). Customer churn prediction in influencer commerce: An application of decision trees. Procedia Comput. Sci. **199**, 1332–1339.

Lalwani, P., Mishra, M. K., Chadha, J. S., & Sethi, P. (2022). Customer churn prediction system: a machine learning approach. Computing **104**, 271–294.

Sudharsan, R., & Ganesh, E. (2022). A Swish RNN based customer churn prediction for the telecom industry with a novel feature selection strategy. Connect. Sci. **34**, 1855–1876.

Wu, X., Li, P., Zhao, M., Liu, Y., Crespo, R. G., & Herrera-Viedma, E. (2022). Customer churn prediction for web browsers. Expert Syst. Appl. **209**, 118177.

Xiahou, X., & Harada, Y. (2022). B2C E-commerce customer churn prediction based on K-means and SVM. J. Theor. Appl. Electron. Commer. Res. **17**, 458–475.

Khattak, A., Mehak, Z., Ahmad, H., Asghar, M. U., Asghar, M. Z., & Khan, A. (2023). Customer churn prediction using composite deep learning technique. Sci. Rep. **13**, 17294.

Liu, Y., Fan, J., Zhang, J., Yin, X., & Song, Z. (2023). Research on telecom customer churn prediction based on ensemble learning. J. Intell. Inf. Syst. **60**, 759–775.

Suh, Y. (2023). Machine learning based customer churn prediction in home appliance rental business. J. Big Data **10**, 41.

Gurung, N., Hasan, M. R., Gazi, M. S., & Chowdhury, F. R. (2024). AI-Based Customer Churn Prediction Model for Business Markets in the USA: Exploring the Use of AI and Machine Learning Technologies in Preventing Customer Churn. J. Comput. Sci. Technol. Stud. **6**, 19–29.

Manzoor, A., Qureshi, M. A., Kidney, E., & Longo, L. (2024). A Review on Machine Learning Methods for Customer Churn Prediction and Recommendations for Business Practitioners. IEEE Access.

Mouli, K. C., Raghavendran, C. V., Bharadwaj, V., Vybhavi, G., Sravani, C., Vafaeva, K. M., Deorari, R., & Hussein, L. (2024). An analysis on classification models for customer churn prediction. Cogent Eng. **11**, 2378877.

Iranian Churn Data. (n.d.).