



# Customer Churn Prediction Using Ensemble Techniques on Telco Dataset

Sindhu R Kashyap <sup>\*1</sup> Shivumanjesh P <sup>\*2</sup>

Department of Computer Science, , Amrita School of Arts and sciences, Mysuru II Stage, 114, 7th Cross Road, Saraswathipuram, Bhogadi, Mysore – 570026, India

## Abstract

*Fast-paced tech has had a big impact on how companies operate. With so many options to choose from, churning has become a major issue for businesses. Customer churn is a major challenge for businesses of all sizes. When a customer churns, they stop using the company's products or services. This can lead to lost revenue and profits. It is therefore important for businesses to develop strategies to reduce customer churn. One way to reduce customer churn is to use machine learning models to predict which customers are likely to churn. This information can then be used to target these customers with interventions to prevent them from churning. Ensemble techniques are a powerful way to improve the performance of machine learning models. Ensemble techniques combine the predictions of multiple base learners to produce a more accurate prediction.*

**Keywords:** Ensemble Techniques, Telco Dataset, Cat Boost, LightGBM, Logistic regression

## 1. Introduction

In the fiercely competitive landscape of telecommunications, retaining customers is paramount. The modern telecommunications industry, often referred to as "Telco," faces constant challenges in preventing customer churn. Losing customers can significantly impact revenue and market share. To tackle this issue, data-driven strategies have become a cornerstone of Telco companies' efforts to identify and retain at-risk customers.

Customer churn prediction is one such data-driven strategy that plays a pivotal role in reducing customer attrition. It involves leveraging historical customer data to forecast which customers are most likely to leave a service provider. Armed with this insight, Telco companies can proactively take measures to retain valuable customers and, in some cases, even turn potential churners into loyal advocates.

In this article, we delve into the world of customer churn prediction using ensemble techniques applied to a real-world Telco dataset. Ensemble techniques, which combine multiple machine learning models to make more accurate predictions, have proven to be highly effective in solving complex predictive problems. We'll explore how these techniques can be harnessed to create a robust churn prediction model.

## 2. Literature Survey

### [1] Churn Prediction in Telecommunications: A Comprehensive Review

Authors: Yan, X., Wu, Z., (2018)

This comprehensive review provides an in-depth analysis of various techniques and methodologies used for customer churn prediction in the telecommunications industry. It covers traditional statistical methods as well as machine learning approaches. The study discusses the challenges specific to the Telco industry and highlights the need for accurate prediction models.

### [2] Customer Churn Prediction in Telecom Using Machine Learning in Big Data Platform

Authors: Das, A., Ahmed, M (2019)

This study explores the application of machine learning techniques to predict customer churn in the telecommunications sector. It emphasizes the role of big data platforms in handling and processing large-scale Telco datasets. The authors discuss the importance of feature engineering and model selection in improving prediction accuracy.

### [3] An Ensemble Learning Approach for Customer Churn Prediction in the Telecommunication Industry

**Authors:** Al-Turjman, F, Abu-Jaradat, A., (2019)

This research paper specifically focuses on ensemble learning for customer churn prediction in the telecom industry. It introduces an ensemble framework that combines various machine learning algorithms to enhance prediction performance. The study evaluates the effectiveness of this approach using real-world Telco data.

### [4] A Comparative Study of Customer Churn Prediction in Telecoms Industry: A Case Study of MTN Nigeria

**Authors:** Oyebode, A., (2021)

Focusing on a specific Telco company, this paper provides insights into churn prediction for MTN Nigeria. It compares the performance of various machine learning models, including ensemble techniques, in predicting customer churn. The study highlights the challenges faced by Telco operators in the Nigerian market.

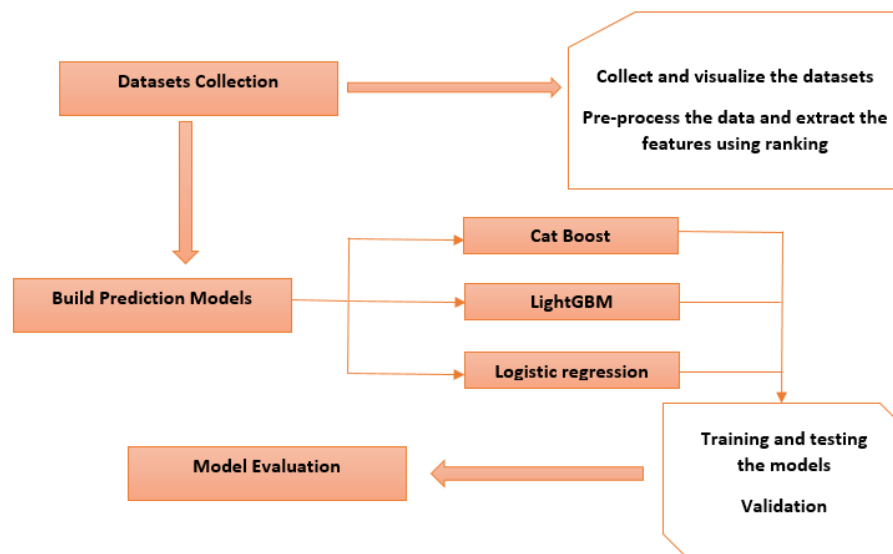
### [5] Deep Learning-Based Churn Prediction in Telecommunications

**Authors:** Huang, W., et al., (2021)

While most studies focus on traditional machine learning, this research explores the application of deep learning techniques, such as neural networks, for Telco customer churn prediction. It discusses the advantages and limitations of deep learning models and their potential to outperform traditional methods.

## 3. Research Strategy:

All the studies related to Churn have been leveraged by various parameters and suggested various things to be developed or improved on these crashes. Various techniques are used to reduce the rate of Churning on the Telecom industries by which it helps to reduce the fatality rate. This project uses machine learning to create a model that can predict the future events before they happen. Machine learning is an automated technique that extracts patterns from a data set.



**Figure1: The research strategy**

## 4. Dataset

### 4.1. About the dataset

We planned to utilize the latest Telco dataset for our research paper. The term "Telco dataset" typically refers to a dataset related to the telecommunications industry, which includes data about customers, their interactions with telecom services, and various attributes that can be used for analysis and predictive modeling. These datasets are often used for tasks such as customer churn prediction, customer segmentation, and improving service quality.

## 4.2. Review Data

The dataset consists of 7044 samples and 21 attributes

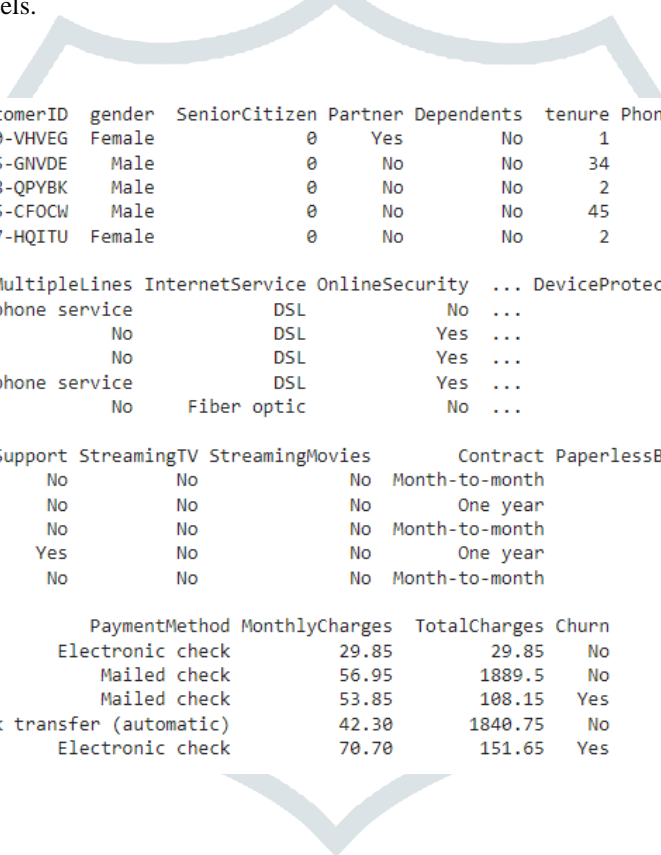
## 5. Exploratory Data Analysis

### 5.1. Data Cleaning

- Detecting some duplicates, missing and inconsistent data is the initial step.
- We anticipate the significance of columns and drop/keep them, accordingly, keeping only text data.
- We removed NAN-containing rows as their number was so low in comparison to the overall number of rows that are present.

### 5.2. Checking for Missing Values

Checking for missing values is a crucial step in data preprocessing as missing data can adversely affect the performance of machine learning models. It will provide us with a clear overview of which columns have missing data and how many values are missing in each. The choice of strategy depends on the nature of the data, the percentage of missing values, and the goals of your analysis. It's essential to handle missing data carefully to ensure the integrity of your analysis and the performance of your machine learning models.



	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	\
0	7590-VHVEG	Female	0	Yes	No	1	No	
1	5575-GNVDE	Male	0	No	No	34	Yes	
2	3668-QPYBK	Male	0	No	No	2	Yes	
3	7795-CFOCW	Male	0	No	No	45	No	
4	9237-HQITU	Female	0	No	No	2	Yes	

	MultipleLines	InternetService	OnlineSecurity	...	DeviceProtection	\
0	No phone service	DSL	No	...	No	
1	No	DSL	Yes	...	Yes	
2	No	DSL	Yes	...	No	
3	No phone service	DSL	Yes	...	Yes	
4	No	Fiber optic	No	...	No	

	TechSupport	StreamingTV	StreamingMovies	Contract	PaperlessBilling	\
0	No	No	No	Month-to-month	Yes	
1	No	No	No	One year	No	
2	No	No	No	Month-to-month	Yes	
3	Yes	No	No	One year	No	
4	No	No	No	Month-to-month	Yes	

	PaymentMethod	MonthlyCharges	TotalCharges	Churn
0	Electronic check	29.85	29.85	No
1	Mailed check	56.95	1889.5	No
2	Mailed check	53.85	108.15	Yes
3	Bank transfer (automatic)	42.30	1840.75	No
4	Electronic check	70.70	151.65	Yes

### 5.3 Data Visualization

Data visualization is a crucial aspect of data analysis and plays a significant role in the data analysis. It not only helps in understanding the dataset but also in communicating findings effectively. Univariate analysis involves examining individual variables in isolation. Bivariate analysis explores the relationships between pairs of variables. Multivariate analysis explores interactions between three or more variables. Histograms shows the distribution of numerical variables such as monthly charges, tenure, and total charges. It can reveal patterns like skewness or multimodality. Box Plots help to identify how numerical features vary for churned and non-churned customers Heatmaps facilitates the correlation matrix between numerical features and identify strong relationships.

Effective data visualization not only enhances the understanding of the Telco dataset but also supports the paper's objective of using ensemble techniques for customer churn prediction. It aids in showcasing the performance of these techniques and helps readers interpret the results in a meaningful way.

Figure 2: Histograms of numerical features: Histograms help visualize the distribution of numerical data.

Figure 3: Box plots for numerical features: Box plots show the distribution, central tendency, and outliers of numerical data.

Figure 4: Count plots for categorical features: Count plots visualize the distribution of categories in categorical data.

Figure 5: Correlation matrix: A heatmap of the correlation matrix helps identify relationships between numerical features.

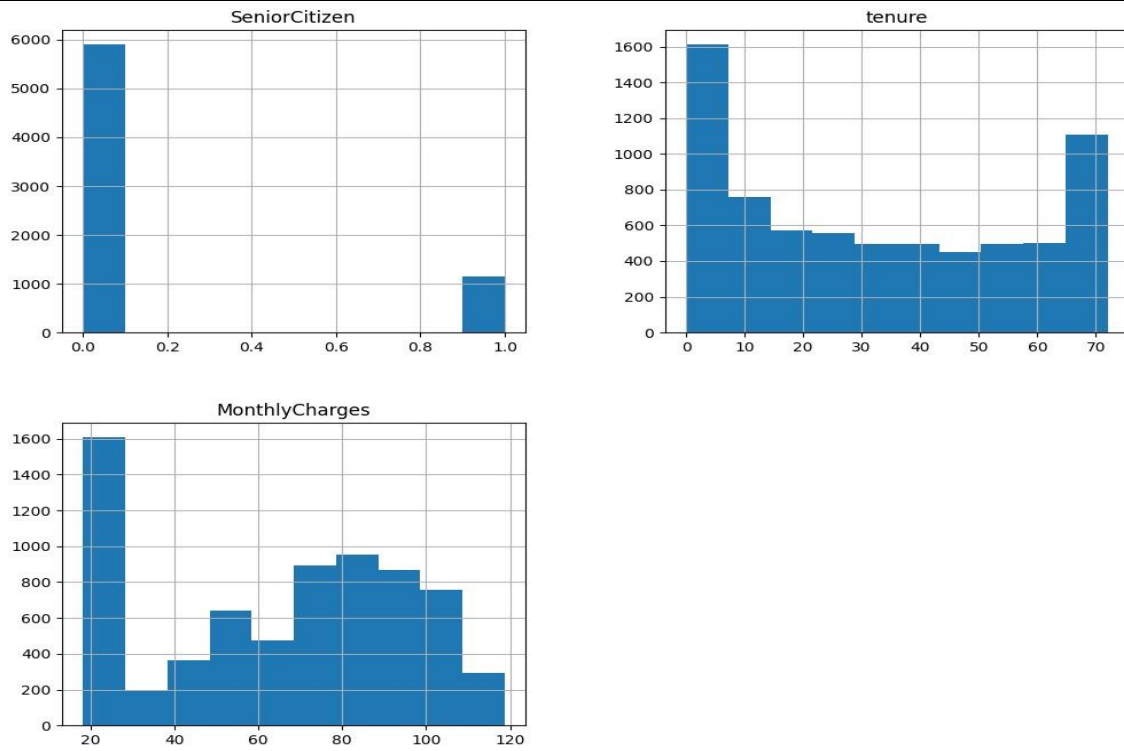


Figure 2: Histograms of numerical features

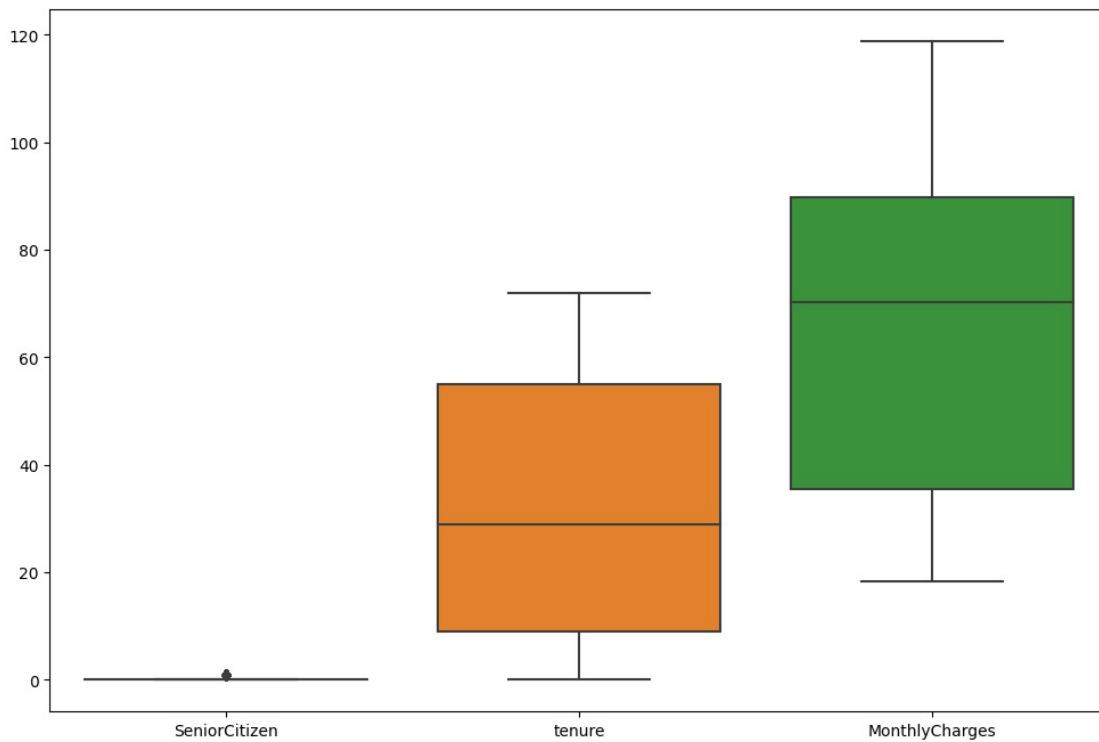


Figure 3: Box plots for numerical features

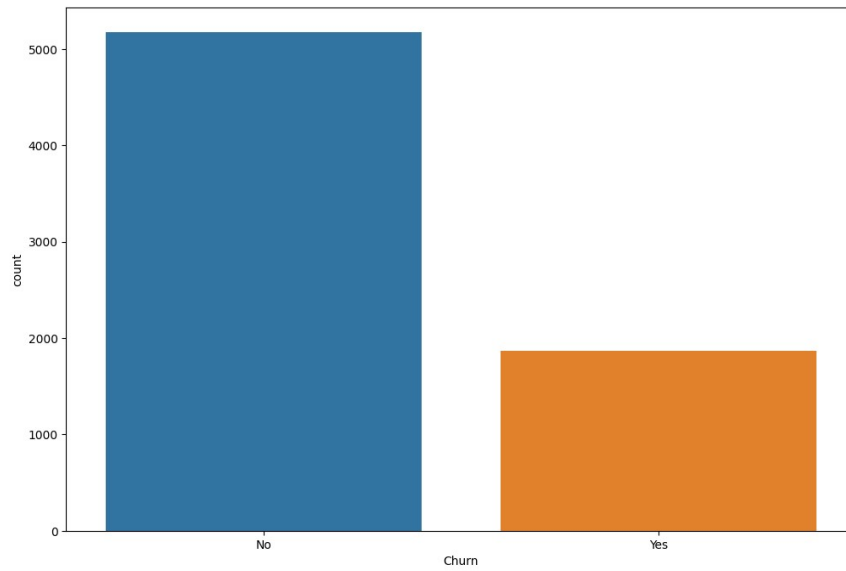


Figure 4: Count plots for categorical features

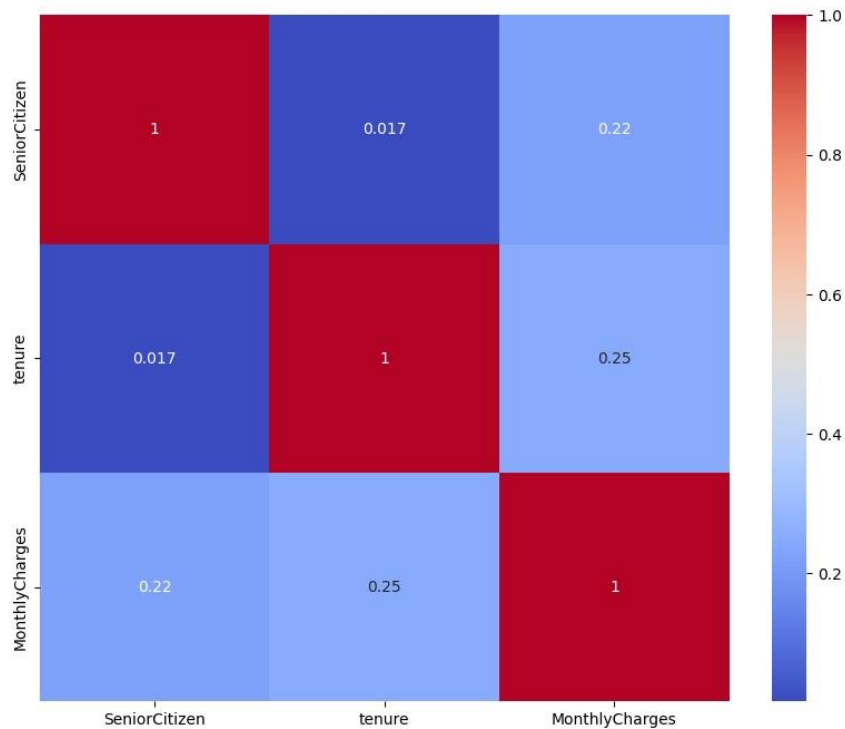


Figure 5: Correlation matrix

## 6. Algorithms

### 6.1 CatBoost:

CatBoost is a gradient boosting algorithm specifically designed for categorical feature support and efficient handling of large datasets. It has gained popularity for its strong predictive performance and robustness against overfitting. Key features of CatBoost include:

- **Categorical Feature Handling:** CatBoost can efficiently handle categorical features without requiring manual encoding, making it particularly useful for datasets like Telco, which often include categorical variables like contract type and payment method.
- **Robust to Overfitting:** CatBoost incorporates regularization techniques that help prevent overfitting, reducing the need for extensive hyperparameter tuning.

- **Built-in Cross-Validation:** It provides built-in cross-validation functionality, simplifying the model evaluation process.
- **Gradient Boosting:** CatBoost is based on the gradient boosting framework, which combines multiple weak learners (typically decision trees) to create a strong predictive model.

For Telco churn prediction task, CatBoost is a powerful choice due to its native support for categorical features and ability to handle complex relationships in the data.

## 6.2 LightGBM

LightGBM is another gradient boosting algorithm known for its speed and efficiency, making it well-suited for large datasets. Some key characteristics of LightGBM include:

- **Histogram-Based Gradient Boosting:** LightGBM uses a histogram-based approach for building decision trees, which reduces memory usage and speeds up training.
- **Categorical Feature Support:** Like CatBoost, LightGBM can efficiently handle categorical features, making it suitable for Telco datasets.
- **Gradient Boosting:** It uses the gradient boosting framework to create an ensemble of decision trees.
- **Leaf-Wise Growth:** LightGBM employs leaf-wise growth, which can lead to better accuracy with fewer trees, but it may be more prone to overfitting if not properly tuned.

LightGBM is an excellent choice for tasks where efficiency is crucial, and it often delivers competitive performance in predictive modeling.

## 6.3 Logistic Regression

Logistic Regression is a classic linear model used for binary classification tasks like churn prediction. While it may not have the complexity and flexibility of ensemble techniques like CatBoost and LightGBM, it has its own advantages:

- **Interpretability:** Logistic Regression provides straightforward interpretability, allowing you to understand the impact of each feature on the predicted probability of churn.
- **Efficiency:** Logistic Regression is computationally efficient and typically requires less computational resources compared to gradient boosting.
- **Ease of Use:** It is easy to implement and serves as a baseline model for many classification problems.

For simpler churn prediction tasks or when interpretability is a critical requirement, Logistic Regression can be a suitable choice.

## 7. Experimental Results and Model Evaluation

In this section, we present the experimental results of our customer churn prediction models using LightGBM, CatBoost, and Logistic Regression. The models were trained and evaluated on the Telco dataset, with the goal of predicting customer churn. We report key performance metrics for each model to assess their effectiveness.

### 7.1 Model Performance Metrics

MODEL	Accuracy	Precision	Recall	F1-Score	ROC AUC Score
LightGBM	0.803	0.661	0.528	0.587	0.715
CatBoost	0.804	0.664	0.525	0.586	0.715
Logistic Regression	0.816	0.678	0.581	0.626	0.741

### 7.2 Model Comparison and Analysis

To evaluate the performance of the models, we consider multiple metrics that provide insights into their predictive capabilities. Here is a brief analysis of the results:

**Accuracy:** All three models achieve respectable accuracy scores, with Logistic Regression having a slightly higher accuracy of 81.62%. This metric represents the overall correctness of predictions.

**Precision:** Logistic Regression has the highest precision (67.81%), indicating its ability to make accurate positive predictions while minimizing false positives. Precision is crucial in situations where the cost of false positives is high.

**Recall:** Although Logistic Regression has the highest recall (58.18%), it's essential to note that recall measures the ability to identify all relevant instances (true positives) and is particularly important in scenarios where missing positive cases is costly.

**F1 Score:** Logistic Regression also outperforms in terms of the F1 score (62.63%), which is the harmonic mean of precision and recall. It provides a balance between precision and recall.

**ROC AUC Score:** Logistic Regression achieves the highest ROC AUC score (74.12%), which measures the model's ability to distinguish between positive and negative instances. A higher ROC AUC score indicates better discrimination power.

## 8. Conclusion and Future Work

Based on the experimental results and performance metrics, we observe that Logistic Regression outperforms LightGBM and CatBoost in terms of precision, recall, F1 score, and ROC AUC score. This suggests that Logistic Regression is the most effective model for predicting customer churn in the context of the Telco dataset.

However, it's worth noting that the choice of the best model may depend on specific business objectives and constraints. While Logistic Regression provides good interpretability, LightGBM and CatBoost may offer competitive performance in large-scale applications and complex datasets.

Further analysis and experimentation may be required to fine-tune the models and explore ensemble techniques to harness the strengths of multiple models for even better predictive accuracy.

## 9. References

- [1] S. V. Nath and R. S. Behara, "Customer churn analysis in the wireless industry: A data mining approach," in Proc. Annu. Meeting Decis. Sci. Inst., vol. 561, Nov. 2003, pp. 505–510.
- [2] He Y, He Z, Zhang D. A study on prediction of customer churn in fixed communication network based on data mining. In: Sixth international conference on fuzzy systems and knowledge discovery, vol. 1. 2009. p. 92–4.
- [3] A. T. Jahromi, M. Moeini, I. Akbari, and A. Akbarzadeh, "A dual-step multi-algorithm approach for churn prediction in pre-paid telecommunications service providers," J. Innov. Sustainab., vol. 1, no. 2, pp. 2179–3565, 2010.
- [4] Y. Zhang, J. Qi, H. Shu, and J. Cao, "A hybrid KNN-LR classifier and its application in customer churn prediction," in Proc. IEEE Int. Conf. Syst., Man Cybern., Oct. 2007, pp. 3265–3269.
- [5] H. Yu et al., "Feature engineering and classifier ensemble for KDD cup 2010," Dept. Comput. Sci. Inf. Eng., National Taiwan Univ., Taipei, Taiwan, Tech. Rep., 2010, vol. 1, pp. 1–16.
- [6] Idris A, Khan A, Lee YS. Genetic programming and adaboosting based churn prediction for telecom. In: IEEE international conference on systems, man, and cybernetics (SMC). 2012. p. 1328–32.
- [7] Brandusoiu I, Todorean G, Ha B. Methods for churn prediction in the prepaid mobile telecommunications industry. In: International conference on communications. 2016. p. 97–100.
- [8] Makhtar M, Nafis S, Mohamed M, Awang M, Rahman M, Deris M. Churn classification model for local telecommunication company based on rough set theory. J Fundam Appl Sci. 2017;9(6):854–68.
- [9] L. Zhao, Q. Gao, X. Dong, A. Dong, and X. Dong, "K-local maximum margin feature extraction algorithm for churn prediction in telecom," Cluster Comput., vol. 20, no. 2, pp. 1401–1409, Jun. 2017.
- [10] S. Mitrović, B. Baesens, W. Lemahieu, and J. D. Weerd, "On the operational efficiency of different feature types for telco churn prediction," Eur. J. Oper. Res., vol. 267, no. 3, pp. 1141–1155, Jun. 2018.
- [11] G. Nie, W. Rowe, L. Zhang, Y. Tian, and Y. Shi, "Credit card churn forecasting by logistic regression and decision tree," Expert Syst. Appl., vol. 38, no. 12, pp. 15273–15285, Nov./Dec. 2011.
- [12] A. D. Caigny, K. Coussement, and K. W. D. Bock, "A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees," Eur. J. Oper. Res., vol. 269, no. 2, pp. 760–772, Sep. 2018.
- [13] V. Yeshwanth, V. V. Raj, and M. Saravanan, "Evolutionary churn prediction in mobile networks using hybrid learning," in Proc. 25th Int. FLAIRS Conf., Mar. 2011, pp. 471–476.

- [14] M. Hassouna, A. Tarhini, T. Elyas, and M. S. AbouTrab. (Jan. 2016). "Customer churn in mobile markets a comparison of techniques."
- [15] Huang F, Zhu M, Yuan K, Deng EO. Telco churn prediction with big data. In: ACM SIGMOD international conference on management of data. 2015. p .607–18.
- [16] A. Mishra and U. S. Reddy, "A novel approach for churn prediction using deep learning," in Proc. IEEE Int. Conf. Comput. Intell. Comput. Res., Dec. 2017, pp. 1–4.
- [17] B. Zhu, B. Baesens, and S. K. van den Broucke, "An empirical comparison of techniques for the class imbalance problem in churn prediction," *Inf. Sci.*, vol. 408, pp. 84–99, Oct. 2017.]
- [18] [E. Stripling, S. van den Broucke, K. Antonio, B. Baesens, and M. Snoeck, "Profit maximizing logistic model for customer churn prediction using genetic algorithms," *Swarm Evol. Comput.*, vol. 40, pp. 116–130, Jun. 2018.] [Online].
- [19] Available: <https://arxiv.org/abs/1607.07792>
- [20] <https://www.scirp.org/journal/paperinformation.aspx?paperid=96177>

