Customer Churn Prediction- Telco Customer Analysis

Project Group P14
Akshada G Malpure agmalpur@ncsu.edu
Rucha M Kulkarni rkulkar5@ncsu.edu
Rishi A Dange radange@ncsu.edu

1 Background

1.1 Problem

Customer churn, the situation of customers discontinuing their relationship with a company, presents a significant challenge in subscription-based industries. A depletion in retention may be caused by a lack of customer engagement and support, lower prices offered by competitive companies or simply marketing a product to the wrong audience.

Losing customers directly impacts profitability. In this highly competitive sector, where customer acquisition costs are substantial, predicting and preventing churn has become crucial for maintaining profitability and sustaining growth. Predicting churn can help companies take proactive steps to retain customers by identifying patterns in their behavior.

The central problem addressed in this project is predicting customer churn in the telecommunications industry. This complex challenge involves analyzing various factors such as demographics, service providers, contract types and other relevant customer attributes to identify those at risk of leaving. By accurately predicting potential churners, companies can implement targeted retention strategies, ultimately improving customer loyalty and reducing financial losses.

1.2 Literature Survey

Recent studies have employed various machine learning and deep learning techniques to tackle the customer churn prediction problem in the telecommunications sector. While these approaches have shown promise, they also reveal areas for improvement and further research.

Wagh et al. (2024) emphasized the importance of demographic and usage features in their analysis, utilizing methods such as logistic regression and random forests. They highlighted how feature engineering enhances model performance. However, their approach may be limited by its reliance on traditional machine learning techniques, which might not capture complex, non-linear relationships in the data as effectively as more advanced methods. [1]

Hu et al. (2020) proposed a hybrid customer churn prediction model combining Decision Tree (C5.0) and Neural Network (BP) techniques to address the limitations of single-model approaches. Using supermarket data, the Decision Tree provided interpretability and confidence-based predictions, while the Neural Network captured complex non-linear relationships. The combined model leveraged weighted confidence from both methods to calculate churn probabilities. It achieved a higher accuracy (98.87%) compared to the Decision Tree (93.47%) and Neural Network (96.42%) individually, demonstrating improved predictive performance and actionable insights for addressing customer churn effectively. However, the study's reliance on a single dataset limits the generalizability of its findings, suggesting a need for validation across diverse datasets and industries to ensure broader applicability. [2]

AlShourbaji et al. (2023) demonstrated the effectiveness of combining gradient-boosting machines with optimization techniques for more accurate churn predictions. Their work emphasized the importance of feature selection and hyperparameter tuning in improving model accuracy. While this approach shows promise, it may be computationally intensive and potentially overfitted to specific datasets, raising questions about its generalizability across different telecom markets. The inclusion of a modified Particle Swarm Optimization (mPSO) with Artificial Ecosystem Optimization features enhanced hyper-parameter tuning and model robustness. This approach bridges gaps in ensemble learning and hybrid modeling for churn prediction, offering a scalable solution for industry-specific challenges. While the study introduces a robust CP-EGBM model, a notable limitation is its reliance on publicly available datasets, which may not capture the unique churn patterns of specific industries. Additionally, the computational expense of metaheuristic optimization methods like mPSO, particularly for high-dimensional data, could limit the model's scalability and real-time applicability in environments with resource constraints. [3]

Jain et al. (2021) explored ensemble methods like XGBoost and highlighted the need for addressing imbalanced datasets and extracting domain-specific features to improve results. Their work provides valuable insights into handling class imbalance, a common issue in churn prediction. It highlights the critical role of feature extraction and advanced methods like CNN in handling large datasets effectively. Its detailed comparison of methodologies offers valuable guidance for researchers entering this field. However, a notable limitation is its reliance on secondary data sources and the lack of implementation insights for practical applications, which could hinder the reproducibility and direct applicability of the findings in real-world telecom environments. Similarly, their focus on ensemble methods may come at the cost of model interpretability, which is crucial for business stakeholders to understand and act upon predictions. [4]

2 Method

2.1 Approach

We aim to address the problem and improve predictive accuracy through the following key components:

1. **Data Preprocessing:** This is an important step that involves understanding the nuances of the dataset structure and futuristic characteristics. In this step, through proper imputation techniques, we handled missing values by replacing numerical feature gaps with the mean and categorical gaps with the mode and standardized numerical features using *StandardScaler* and applied *One-Hot Encoding* to categorical variables for consistency. Features like *'tenure', 'TotalCharges' and 'MonthlyCharges'* were imputed for their missing value. The data was partitioned into a 70% training set and a 30% testing set.
2. **Data Visualization:** The project mainly focused on the analysis of conclusive attributes that help unveil the churn pattern. For this, we employed various visualization techniques to uncover underlying patterns in the data and analyze relationships between features and churn. Ultimately, we identify potential areas for feature engineering or selection. Multivariate analysis helps us understand the impact of several features on churn as a complex entity.

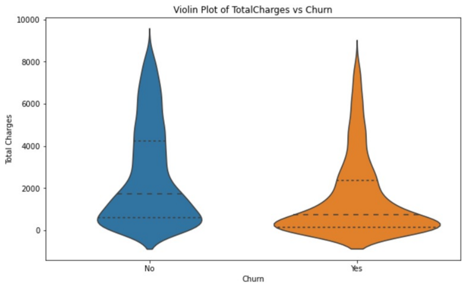


Figure 1: *Violin Plot showing Density of Churn values for Total Charges*

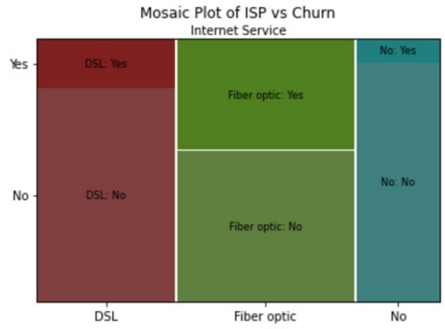


Figure 2: *Mosaic Plot showing spread values of ISP for Target Variables*

3. **Feature Selection:** We employed correlation analysis and feature importance derived from tree-based models to identify underlying patterns and significant predictors like *'Tenure', 'ContractType', and 'PaymentMethod'* to fine-tune them to further strengthen the chosen model's accuracy. We also excluded low-variance features like 'CustomerID' that did not correlate with Churn to reduce noise and improve model performance. The below heatmap demonstrates how each feature correlates with the other.

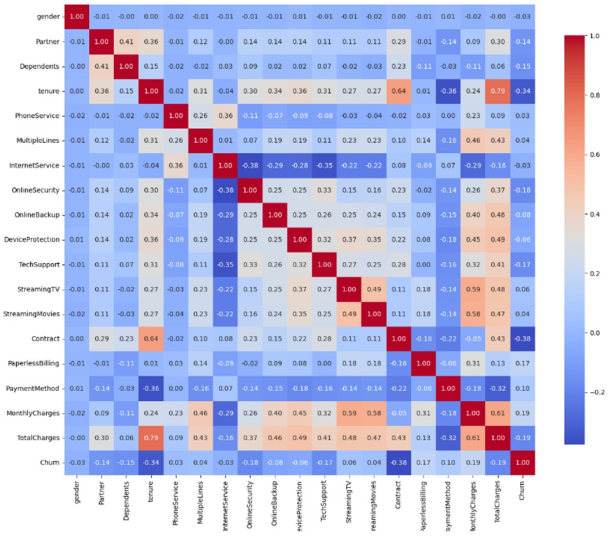


Figure 3: *Visualizing the Correlation between different attributes*

4. **Handling Data Imbalance:** To address the class imbalance inherent in churn datasets, we implemented Synthetic Minority Over-sampling Technique (SMOTE) to balance the distribution of churn and non-churn instances. We consequently leverage the impact of balancing techniques on model performance.

5. **Model selection and training:** We explored multiple models including Logistic Regression, Gradient Boosting, SVMs, Gaussian Naive Bayes, K-NN, Decision Tree, XGBoost, and Random Forest. We began by exploring a range of machine learning algorithms to evaluate their performance in predicting customer churn. These models were chosen to balance simplicity and interpretability with the ability to handle non-linear relationships and complex interactions. During model training, we paid particular attention to hyperparameter tuning to optimize performance. For the Random Forest model, we fine-tuned parameters such as the number of estimators *(n_estimators)*, maximum tree depth *(max_depth)*, minimum samples per split *(min_samples_split)*, and minimum samples per leaf *(min_samples_leaf)*. Similarly, for Gradient Boosting and LightGBM, we adjusted parameters like learning rate *(learning_rate)*, number of estimators, and maximum depth. Then, we trained and evaluated multiple predictive models and fine-tuned hyperparameters using cross-validation to enhance performance.

6. **Performance Metrics:**The effectiveness of the models was assessed using accuracy, precision, recall, F1 score, and AUC-ROC. There was a slight improvement in the accuracies after hyperparameter tuning.

2.2 Rationale

We selected imputation techniques (*mean* for numerical and *mode* for categorical data) to handle missing values because they preserved the dataset's original distribution while being computationally efficient. This choice was particularly suitable as our dataset exhibited a relatively small number of missing values, where more complex techniques like Multiple Imputation or KNN-based imputation could have introduced unnecessary complexity and increased processing time without substantial gains in accuracy.

Standardization using *StandardScaler* was crucial for ensuring numerical features like MonthlyCharges and TotalCharges were on comparable scales, especially for distance-based models like K-Nearest Neighbors (KNN). We avoided normalization or Min-Max scaling because they could disproportionately affect models sensitive to outliers.

SMOTE was chosen to address the imbalance between churn and non-churn instances because it creates synthetic samples for the minority class, preserving the original distribution. Alternative methods like undersampling would have reduced the dataset size, potentially discarding valuable information from the majority class. More complex techniques, such as Generative Adversarial Networks (GANs) for data synthesis, were not pursued due to their computational intensity and the adequacy of SMOTE for our dataset size and imbalance ratio.

Correlation analysis and tree-based feature importance were chosen because they provide interpretable insights into the relationships between features and churn, enabling actionable insights. While dimensionality reduction techniques like PCA could have simplified the dataset, they were avoided as they obscure feature interpretability, which is essential for business applications. Excluding low-variance features like *CustomerID* reduced noise and computational complexity, ensuring the models focused on relevant predictors.

The models were chosen based on their ability to handle non-linear relationships, imbalanced data and diverse feature types: *Random Forest* was selected for its robustness and interpretability, ideal for datasets like ours with mixed feature types and class imbalance. *Gradient Boosting* and *LightGBM* were included due to their scalability and efficiency, particularly for datasets with complex relationships between features. Simpler models like *Logistic Regression* and *Naive Bayes* were used as baselines for comparison. These models were less computationally expensive and offered quick insights but lacked the predictive power of ensemble methods for non-linear data. Deep learning models, such as CNNs or RNNs, were excluded because the dataset size and feature complexity did not justify their computational cost. Similarly, Support Vector Machines (SVMs) with non-linear kernels were not prioritized due to their scalability issues with larger datasets and longer training times compared to ensemble methods.

Grid search was employed for hyperparameter tuning to systematically explore and identify the best parameter combinations for each model. This approach ensured all potential configurations were evaluated, providing a controlled framework for optimization. While Bayesian optimization or random search could have been applied, grid search was computationally feasible given the size of our parameter space and dataset.

3 Experiment setup

3.1 Dataset description

The *Telco Customer Churn* dataset, available on Kaggle, contains *7,043* customer records and *21* features. These features fall into three categories:

1. **Personal information** including details like customer gender, age and dependents.
2. **Services subscribed** detailing services such as phone lines, internet and security features.
3. **Billing information** which includes monthly charges, payment methods, tenure and the overall cost of services.
4. The primary target feature in this dataset is **Churn**, a binary variable indicating whether or not a customer has discontinued their services.

The dataset contains both numerical and categorical variables, which will require pre-processing before model training. The dataset is split into a training set (70%) and a test set (30%). Out of the (70%) of train data, it was again split into an (80%) model training set and a (20%) model testing set to calculate the performance metrics. The original dataset on Kaggle had no missing values. To increase the model's reliability and robustness we added random noise in the data which resulted in nearly 5% missing values in the overall dataset.

Dataset: **Telco Customer Churn Dataset(Source: Kaggle)**

3.1.1 Hypothesis

This section reports some of the hypotheses of our project. In our study, we have formulated two research hypotheses that will be tested and validated using the Telco customer churn prediction dataset.

- **Hypothesis 1**: There is no significant relationship between the customer's internet service provider and their likelihood of churn.
- **Hypothesis 2**: It is anticipated that more sophisticated models, such as Random Forest and Gradient Boosting, will yield better results compared to simpler models like Logistic Regression.

Hypothesis 1 was validated through a chi-square test, where we anticipated the results to indicate no significant relationship. Moreover, we aimed to understand whether the customer's internet service provider could be considered a critical factor in churn prediction models.

For *Hypothesis 2*, we anticipate that advanced models like Random Forest and Gradient Boosting will outperform simpler models such as Logistic Regression in predicting churn. Recognizing that churn rate attributes often exhibit non-linear relationships, we implemented these sophisticated models to address the limitations of Logistic Regression. This approach enhances accuracy and improves feature selection, enabling us to uncover complex interactions that influence customer churn and consequently validates our hypothesis is true. Results for this have been included in the later part of this report.

Some of the Research Questions that we plan to answer are listed in the section below.

3.2 Experimental design

Let us address the research questions we came across while working on the project:

1. **What are the primary features affecting the churn rate?**
   We trained the dataset using various models, including Linear Regression, Gradient Boosting, XGBoost, Random Forest and LightGBM, while applying hyperparameter tuning to the top-performing models to enhance their robustness. Among these, the Random Forest model emerged as the most accurate and high-performing. To identify the key features influencing the churn rate, we analyzed feature importance from the Random Forest model and the following graph highlights these features ranked by their relevance.

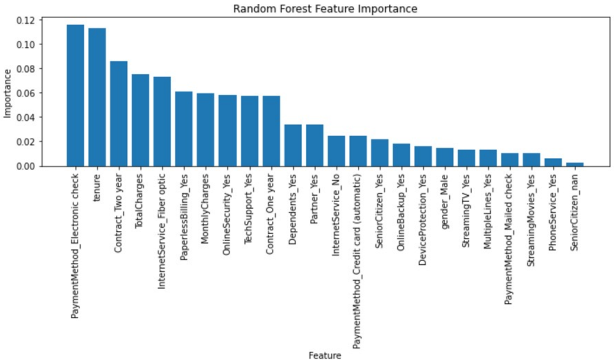


Figure 4: *Feature importance to understand the key features affecting churn rate*

2. **Is there a relationship between the tenure of a customer and their likelihood of churn?**
   Yes, there is a clear relationship between customer tenure and the likelihood of churn. Exploratory Data Analysis (EDA) was used to group customers by tenure and calculate churn rates for each group, revealing that churn rate decreases as tenure increases. This suggests that customers with longer tenures are less likely to churn, indicating stronger loyalty and satisfaction over time. By uncovering this trend through EDA, we gained valuable insights that highlight the importance of retaining customers early in their tenure to foster long-term loyalty, which is a positive sign for the company.
3. **Does the type of internet service influence customer churn?**
   Yes, the type of internet service significantly influences customer churn. Based on the

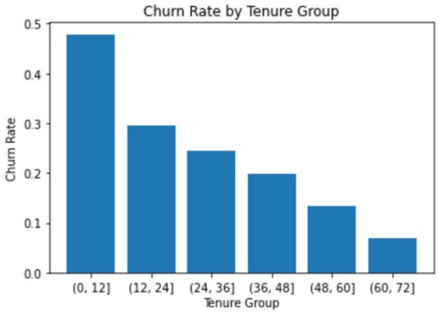


Figure 5: *Relationship between different Tenures (Grouped) and Churn Rate*

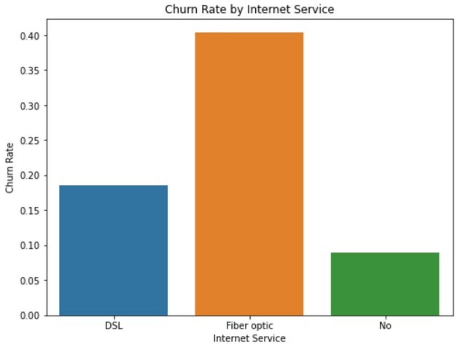


Figure 6: *Interpreting the relation between ISP and Churn Rate*

analysis, customers with fiber optic internet service exhibited the highest churn rate, followed by those using DSL, while customers without internet service had the lowest churn rate. The high churn rate among fiber optic users may be attributed to dissatisfaction with service quality, such as frequent disruptions or unmet expectations of performance. Additionally, the higher cost of fiber optic services may lead customers to switch if they feel the value does not justify the price or if they find cheaper alternatives. On the other hand, customers without internet service may rely on alternative sources like public Wi-Fi or mobile data, leading to reduced dependency on the telecom provider and lower churn rates. This analysis underscores the need to address service quality and pricing strategies to retain customers in specific internet service categories.

4. **Is there a correlation between the customer's payment method and churn rate?** Yes, there is a clear correlation between the customer's payment method and churn rate. Customers using electronic checks showed the highest churn rates followed by those using mailed checks. Conversely, customers with automatic payment methods, such as bank transfers or credit cards, exhibited significantly lower churn rates. These findings are critical for data analysis as they highlight specific groups of customers more likely to churn, enabling the company to investigate potential issues with electronic checks and mailed checks, such as convenience or user experience. Additionally, this analysis emphasizes the value of promoting automatic payment methods to improve customer retention.

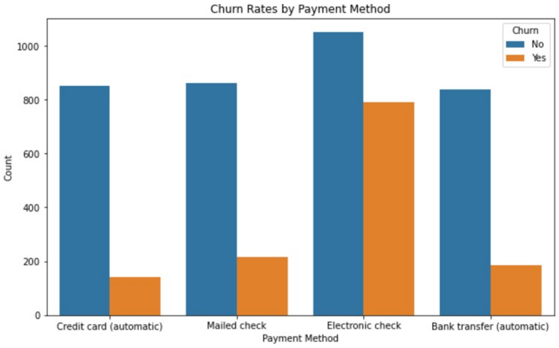


Figure 7: *Visualizing how paperless billing affects Churn Rate*

5. **Does the presence of paperless billing affect customer churn?** Yes, the presence of paperless billing appears to affect customer churn. From the data, customers without paperless billing have a lower churn rate (16.8%) compared to those with paperless billing (33.3%). This suggests that customers opting for paperless billing are more likely to churn. The increased churn rate could be due to potential dissatisfaction with digital billing experiences, such as accessibility issues, unclear charges or reduced customer

engagement. These insights highlight the importance of improving the digital billing process to enhance customer satisfaction and reduce churn among paperless billing users.

6. **How can churn prediction insights support customer loyalty initiatives?**
Churn prediction insights can greatly enhance customer loyalty initiatives by enabling targeted strategies to retain at-risk customers. For instance, early engagement efforts can be directed toward customers with shorter tenures to improve their experiences and foster loyalty. Additionally, providing incentives for customers who use automated payment methods can encourage satisfaction and reduce churn. Finally, targeted marketing and educational campaigns can build loyalty by informing customers about the value of services and features that cater to their needs. Overall, these insights allow companies to develop tailored initiatives that enhance customer satisfaction and promote long-term loyalty.

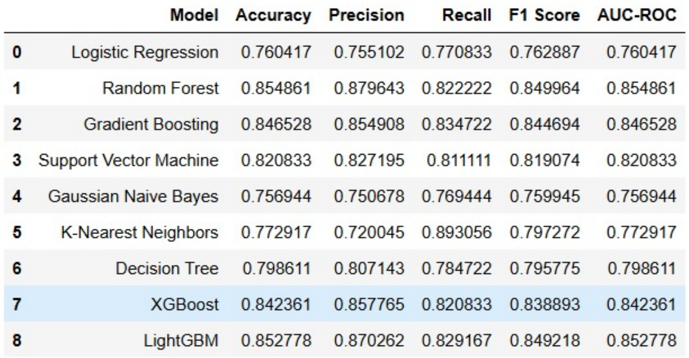
4 Results

The performance of each model is evaluated using metrics like accuracy, precision, recall, F1 score, and AUC-ROC. Based on the F1 score. The table below summarizes the performance metrics of each model tested:

The top two performing models are:

1. **Random Forest**: F1 score of 0.854355
2. **K-Nearest Neighbours**: F1 score of 0.797272

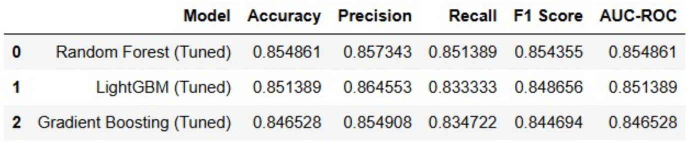
Other models like Logistic Regression and Naive Bayes performed adequately but were not as effective in handling the dataset's complexity and imbalanced distribution.



	Model	Accuracy	Precision	Recall	F1 Score	AUC-ROC
0	Logistic Regression	0.760417	0.755102	0.770833	0.762887	0.760417
1	Random Forest	0.854861	0.879643	0.822222	0.849964	0.854861
2	Gradient Boosting	0.846528	0.854908	0.834722	0.844694	0.846528
3	Support Vector Machine	0.820833	0.827195	0.811111	0.819074	0.820833
4	Gaussian Naive Bayes	0.756944	0.750678	0.769444	0.759945	0.756944
5	K-Nearest Neighbors	0.772917	0.720045	0.893056	0.797272	0.772917
6	Decision Tree	0.798611	0.807143	0.784722	0.795775	0.798611
7	XGBoost	0.842361	0.857765	0.820833	0.838893	0.842361
8	LightGBM	0.852778	0.870262	0.829167	0.849218	0.852778

Figure 8: *The Results of different models before Hyperparameter Tuning*

- **Impact of Hyperparameter Tuning:** Hyperparameter tuning significantly improved the performance of ensemble models. For instance:



	Model	Accuracy	Precision	Recall	F1 Score	AUC-ROC
0	Random Forest (Tuned)	0.854861	0.857343	0.851389	0.854355	0.854861
1	LightGBM (Tuned)	0.851389	0.864553	0.833333	0.848656	0.851389
2	Gradient Boosting (Tuned)	0.846528	0.854908	0.834722	0.844694	0.846528

Figure 9: *The Precision Metrics after Hyperparameter Tuning*

- **Random Forest:** F1 score increased from 84.99% (default parameters) to 85.43% after tuning parameters.
- **Handling Class Imbalance:**The application of SMOTE effectively balanced the dataset, leading to notable improvements in recall across all models. For Random Forest, recall improved from 82% to 85% post-SMOTE, indicating better detection of churned customers.

The following attributes had a major impact on Churn, namely:

1. **Tenure:** Customers with shorter tenure were more likely to churn.
2. **ContractType:** Customers on month-to-month contracts had higher churn rates compared to those on longer-term contracts.
3. **Payment Method:** Payments through Electronic Checks indicated the highest Churn Rate, which could be nullified through automatic payments instead.

5 Conclusion

By applying machine learning techniques, this project provided valuable insights into customer churn prediction in the telecommunications industry. Key lessons learned include addressing data imbalance using SMOTE, which significantly enhanced model performance, particularly for Random Forest, which achieved the highest F1 score of 85.43%, showcasing its robustness in predictive capability. Additionally, the analysis revealed critical churn factors such as:

- *Tenure:* Churn rates decline as customer tenure lengthens, indicating that long-term customers are less likely to leave. This pattern suggests that loyalty and satisfaction tend to grow over time, reflecting a stronger relationship with the company.
- *Payment Method:* The type of payment method correlates with churn rates. Customers using electronic or mailed checks show higher churn rates, while those who use bank transfers or credit cards with automatic payments exhibit lower churn rates. This pattern suggests that convenient, automated payment options may enhance satisfaction and loyalty.
- *Paperless Billing:* Customers who choose paperless billing show a higher churn rate than those preferring traditional billing methods. This suggests that customers who opt for digital processes may have unique expectations or experiences that impact their decision to stay or leave.

These insights can help inform strategies to reduce churn by focusing on factors that influence customer retention and loyalty.

We also discovered that hyperparameter tuning and feature engineering play pivotal roles in improving model outcomes. A limitation of this study was the time constraint, which restricted the exploration of advanced models like XGBoost by tuning different features that could boost its performance. Future work could address these areas, enabling more accurate and scalable churn prediction models. Additionally, combining different outperforming models together or including different data sources to improve robustness and accuracy metrics could also act as a power move. Overall, this project highlights the potential for machine learning to proactively address customer churn and strengthen customer retention strategies in the telecom sector.

References

[1] Ibrahim AlShourbaji et al. *An efficient churn prediction model using gradient boosting machine and metaheuristic optimization.* 2023. DOI: 10.1038/s41598-023-41093-6.

[2] Xin Hu et al. *Research on a Customer Churn Combination Prediction Model Based on Decision Tree and Neural Network*. 2020. DOI: 10.1109/ICCCBDA49378.2020.9095611.

[3] Hemlata Jain, Ajay Khunteta, and Sumit Srivastava. *Telecom churn prediction and used techniques, datasets and performance measures: a review*. 2021. DOI: 10.1007/s11235-020-00727-0.

[4] Sharmila K. Wagh et al. *Customer churn prediction in telecom sector using machine learning techniques*. 2024. DOI: 10.1016/j.rico.2023.100342. URL: https://doi.org/10.1016/j.rico.2023.100342.

GitHub Repository Link: Project Group 14 Repository