# Predicting Customer Churn in the Telecommunications Industry Using Predictive Analytics

BSAN 775: Introduction to Business Analytics

Submitted to: Professor Zahra Ziaei

Manohar Babu Bunga (F984V647)
Minhal Abbas (R349N456)
Renuka Mohanta (F996E733)
Venkata Naga Leela Adithya Ramisetty (P7627268)

## Introduction

Customer churn refers to the phenomenon in which customers discontinue their subscriptions or transition to a competing service provider. It directly reduces revenue and market share, making customer retention far more cost-effective than acquiring new users. Studies show that obtaining new customers could be up to five times as costly as retaining existing customers, and a 5% reduction in customer loss can boost a company's profit by up to 95% (Kumar & Shah, 2004). The telecommunications sector routinely faces substantial revenue losses from churn. For example, Sana et al. (2022) emphasize that correctly identifying likely churners can "reduce dissatisfaction, increase engagement and thus potentially retain [customers]," directly boosting revenue. Similarly, Sikri et al. (2024) note that telecom companies are shifting strategies toward customer retention because gaining a new customer costs an order of magnitude more than preventing an existing customer from leaving.

Given this context, churn prediction using machine learning (ML) has become vital for targeted retention strategies and long-term profitability. This study develops and compares two ML models using the Telco Customer Churn dataset from Kaggle (7,043 customers, 21 features), which includes demographics, service usage, and billing data. After standard preprocessing (handling missing values, encoding categories, scaling features), we train logistic regression for interpretability and random forest for predictive performance.

Our approach combines logistic regression (for interpretability) and random forest (for predictive power) to balance accuracy and actionable insights into churn drivers. By comparing results and analyzing feature importance, we provide actionable, data-driven recommendations for telecom operators. This work demonstrates that even basic ML pipelines can yield actionable insights, offering a cost-effective alternative to complex industry solutions while maintaining relevance.

## Literature Review

Research has revealed several important drivers of customer churn. Neslin et al. (2006) noted service quality, price, and the offerings of competitor companies as the primary drivers. Similarly, Ahn, Han, and Lee (2006) reached the conclusion that billing complaints and service quality dissatisfaction are the main causes of customer attrition. While there has been growth in predictive models, the problem of finding real drivers of churn and translating insight into action strategy remains. Ascarza (2018) noted that traditional retention efforts are ineffective because they are reactive and not data-driven. To drive improvement in retention, companies must utilize customer information including demographics, usage of service, and payment history to build data-driven interventions (Huang et al., 2012).

Sikri et al. (2024) systematically analyze churn prediction classifiers, emphasizing the superior performance of ensemble methods like Gradient Boosting and XGBoost in handling imbalanced data. They propose a novel ratio-based data balancing technique to address class imbalance (common in churn datasets) and show that when class balance is improved, ensemble models significantly outperform simpler classifiers like logistic regression or decision trees. This emphasizes the importance of both data preparation and model selection in improving predictive accuracy.

In contrast, Ana Nurtriana et al. (2024) demonstrate that simpler models can still perform competitively. Using a Customer Churn dataset, they compared logistic regression, random forest, and SVM models. Surprisingly, logistic regression achieved the highest accuracy (≈79%), slightly outperforming the others. This finding highlights the practical value of interpretable models, especially in real-world settings where transparency is critical. The authors also stress

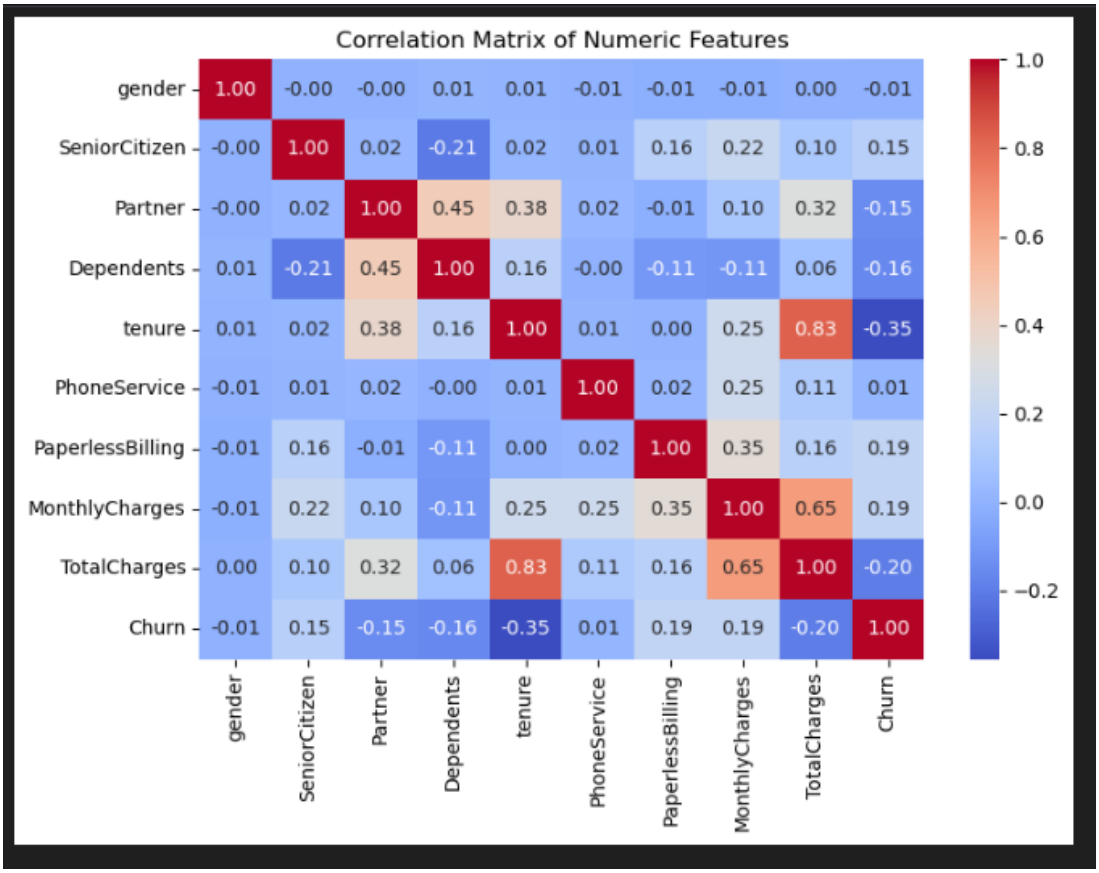
the challenge of class imbalance and reinforce the strategic importance of retention over costly customer acquisition.

Further research by Ijaz, Asghar, and Gul (2019) explores advanced modeling techniques, integrating penalized logistic regression (ridge, lasso, elastic-net) and random forests into ensemble pipelines. Their models, optimized using techniques like regularization and weighted accuracy, show significant gains in metrics such as AUC and sensitivity. Similarly, Abbasimehr et al. (2014) demonstrate that combining classifiers generally boosts performance, reinforcing the value of ensemble learning for churn prediction.

Other studies prioritize feature innovation to enhance predictions. Ahmad et al. (2019) incorporate social network analysis (SNA) into churn modeling by analyzing call-detail records. By building a customer interaction graph and extracting features like PageRank and centrality, they raise AUC scores from 0.84 to 0.93, a significant improvement that underscores how behavioral and relational data can reveal churn risk beyond standard billing or service features. While social network analysis requires robust infrastructure, it underscores limitations in traditional feature sets.

Together, these studies provide strong evidence that churn prediction benefits from balanced data, ensemble models, and thoughtful feature engineering, while also showing that interpretable models like logistic regression remain viable, especially when simplicity and clarity are priorities. Unlike approaches that depend on proprietary or complex network data, our study uses a publicly available dataset and applies a reproducible ML pipeline that balances accuracy and interpretability. These insights directly inform our research hypothesis: that combining interpretable and ensemble models on a well-preprocessed public dataset can yield churn predictions that are both accurate and actionable for telecom providers.

**Correlation Matrix of Numeric Features**




Given the results, logistic regression demonstrated a modest advantage in key metrics, particularly its higher recall for identifying churn cases, which is crucial for customer retention. Despite the random forest's slightly higher accuracy for the majority class, logistic regression's simplicity and better performance on the churn class made it the preferred choice for this analysis. Its higher recall for churn detection, which is critical for proactive retention efforts, solidified its selection as the final model.

## Methods

The analysis used the Kaggle Telco Customer Churn dataset, which contains 7,043 customer records and 21 variables, including demographic and service-related features such as customerID and the target variable Churn. These features include information like gender, senior citizenship, service subscriptions, contract type, billing method, monthly charges, and total charges. An initial inspection of the dataset confirmed that there were no missing values or duplicates, though the TotalCharges column was read as an object due to blank entries in some rows. To address this, the column was converted to numeric format using the pd.to_numeric(..., errors='coerce') function, which turned blank values into NaN. Subsequently, any rows containing NaN values were removed, resulting in a final dataset of 7,032 records.

After cleaning the data, the customerID column, which was non-predictive, was dropped. Categorical features, including the target variable Churn, were encoded: binary categories (e.g., gender) were label-encoded, while multi-class variables (e.g., contract type) underwent one-hot encoding. The target variable Churn was label-encoded to binary values (0 for "No" and 1 for "Yes"). For other categorical variables, binary features were label-encoded, while features with more than two categories were one-hot encoded using the pd.get_dummies() function. The numeric features, including tenure, MonthlyCharges, and the newly converted TotalCharges, were standardized using StandardScaler to ensure that all numeric inputs were on the same scale, improving model performance. After encoding and scaling, the feature set was expanded to 40 predictors, along with the target variable, resulting in a final dataset of 7,032 rows and 41 columns.

We split the dataset into features (X) and the target variable (y), reserving 80% for training and 20% for testing, with a fixed random state (42) to ensure reproducibility. Two classification

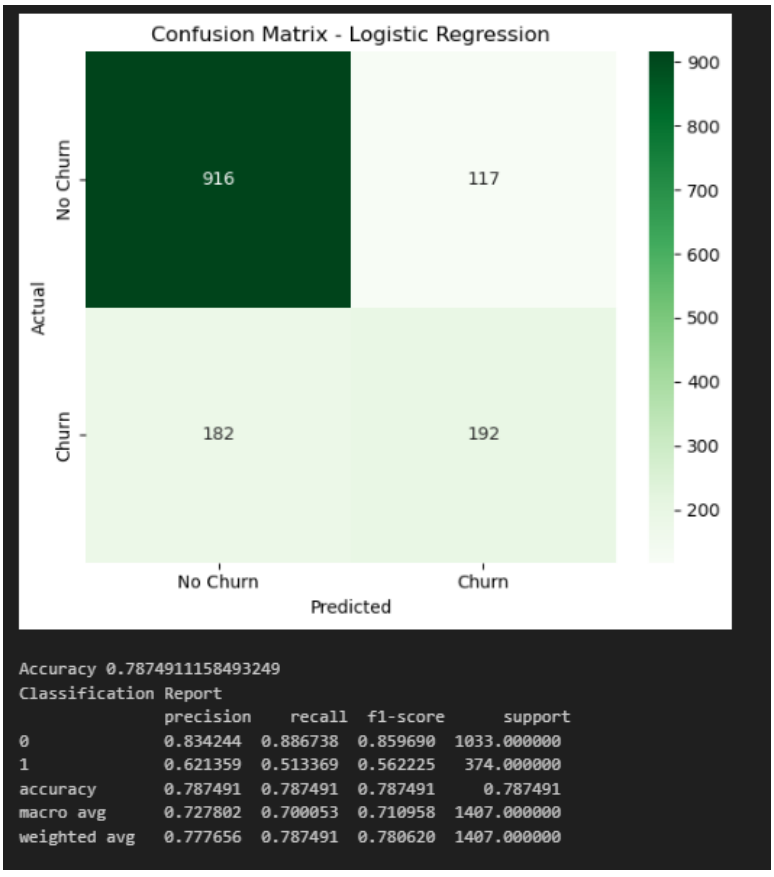
models were implemented using scikit-learn: logistic regression and random forest. Logistic regression was chosen as a baseline linear model due to its simplicity and interpretability, allowing us to understand the relationship between the features and the target. The random forest model, an ensemble technique, was selected to capture more complex, nonlinear relationships in the data, offering greater flexibility. The logistic regression model was initialized with a higher iteration limit (max_iter=1000) to ensure that the model converged during training. The random forest classifier was set to use 100 trees (n_estimators=100) to ensure stability and robustness in its predictions. Although the dataset exhibited some class imbalance, no explicit class balancing or resampling techniques were applied, as the models were still expected to perform adequately without these adjustments.

Both models were trained on the training set and evaluated on the test set using several performance metrics, including accuracy, precision, recall, F1-score, and confusion matrix. This comprehensive evaluation allowed us to compare the performance of the linear logistic regression model with the more flexible random forest model, assessing their effectiveness in predicting customer churn.

## Results

The logistic regression model achieved a slightly higher overall accuracy (approximately 78.7%) compared to the random forest model (approximately 78.1%). For the majority class ("No churn"), logistic regression demonstrated strong performance, with precision and recall values of 0.83 and 0.88, respectively. The random forest model yielded similar precision (0.82) and slightly higher recall (0.89), indicating robust identification of non-churners by both models

**Logistic Regression - Confusion Matrix and Metrics**



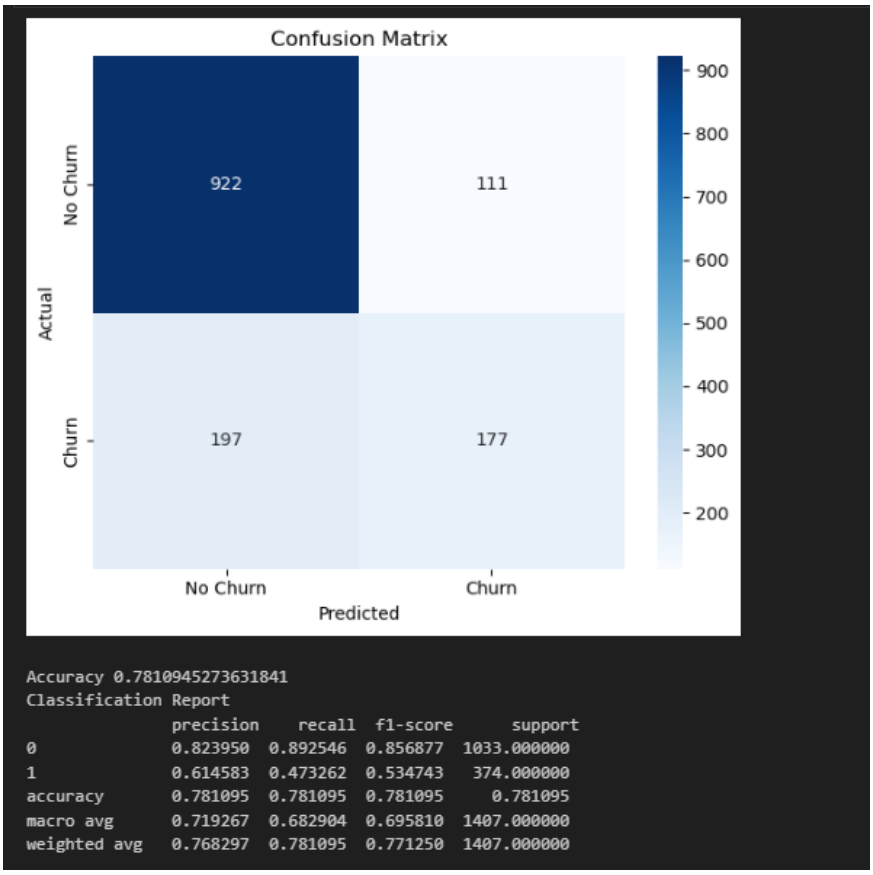

Accuracy 0.7874911158493249
Classification Report

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.834244 | 0.886738 | 0.859690 | 1033.000000 |
| 1 | 0.621359 | 0.513369 | 0.562225 | 374.000000 |
| accuracy | 0.787491 | 0.787491 | 0.787491 | 0.787491 |
| macro avg | 0.727802 | 0.700053 | 0.710958 | 1407.000000 |
| weighted avg | 0.777656 | 0.787491 | 0.780620 | 1407.000000 |

However, when predicting the minority class ("Yes churn"), performance diverged more clearly. Logistic regression achieved a precision of 0.62 and a recall of 0.51, while the random forest model produced slightly lower precision (0.615) and recall (0.473). These differences resulted in F1-scores of 0.562 (logistic regression) and 0.535 (random forest) for churn detection, indicating

logistic regression's slight edge. For the "No churn" class, F1-scores were 0.860 and 0.857, respectively. Overall, logistic regression offered a modest but consistent advantage across key metrics, particularly for the minority churn class.

**Random Forest - Confusion Matrix and Metrics**

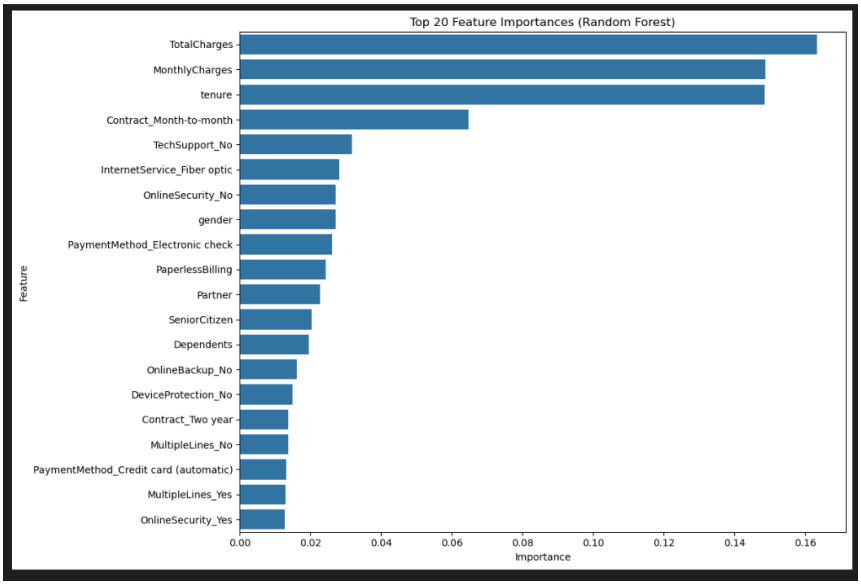



The confusion matrices further support this conclusion. For a test set of 1,407 instances, the logistic regression model correctly identified 916 non-churn cases and 192 churn cases, resulting in 117 false positives and 182 false negatives. The random forest model identified 922 non-churn and 177 churn cases, but with slightly more false negatives (197), suggesting a reduced ability to detect churn cases compared to logistic regression.

Feature importance analysis from the random forest model highlighted 'TotalCharges', 'MonthlyCharges', and 'tenure' as the most significant predictors of churn. Additional

contributing factors included service-related features such as contract type and technical support options.

**Top 20 Feature Importances (Random Forest)**




**Feature Importance Table (Random Forest)**

| | Feature | Importance |
|---|---|---|
| 8 | TotalCharges | 0.163332 |
| 7 | MonthlyCharges | 0.148727 |
| 4 | tenure | 0.148435 |
| 33 | Contract_Month-to-month | 0.064829 |
| 24 | TechSupport_No | 0.031815 |
| 13 | InternetService_Fiber optic | 0.028185 |
| 15 | OnlineSecurity_No | 0.027097 |
| 0 | gender | 0.027089 |
| 38 | PaymentMethod_Electronic check | 0.026120 |
| 6 | PaperlessBilling | 0.024330 |
| 2 | Partner | 0.022843 |
| 1 | SeniorCitizen | 0.020378 |
| 3 | Dependents | 0.019533 |
| 18 | OnlineBackup_No | 0.016170 |
| 21 | DeviceProtection_No | 0.014875 |
| 35 | Contract_Two year | 0.013834 |
| 9 | MultipleLines_No | 0.013683 |
| 37 | PaymentMethod_Credit card (automatic) | 0.013116 |
| 11 | MultipleLines_Yes | 0.012999 |
| 17 | OnlineSecurity_Yes | 0.012688 |

These findings align with observed correlations, where longer tenure was negatively associated with churn (correlation coefficient: -0.35), and higher total charges were associated with more loyal, long-term customers.

## Discussion

This study used logistic regression and random forest models to predict customer churn using a telecommunications dataset. Logistic regression achieved an accuracy of approximately 78.7% on the test set, with a precision of 0.83 and recall of 0.89 for non-churners, and precision of 0.62 and recall of 0.51 for churners. The random forest model produced a similar overall accuracy of 78.1%, but with slightly lower recall for churners (0.49). These results show that although both models perform well in identifying customers who are not likely to churn, they struggle to accurately detect those who are.

This challenge is partly due to the class imbalance in the dataset: only about 27% of customers actually churned. As a result, a naive model that predicts "no churn" for everyone would still achieve about 73% accuracy. This highlights a key limitation—accuracy alone can be misleading in imbalanced datasets. A model may achieve high overall accuracy but fail to identify the smaller, high-impact churn class. For this reason, we focused on other performance metrics like precision, recall, and F1-score, which revealed that both models missed a substantial number of true churners.

Despite this limitation, both models can serve as useful tools for decision support in telecom retention efforts. Because they generate churn probability scores, they can help managers flag customers at high risk of leaving. These predictions can inform targeted interventions, such as special offers or improved customer support. Logistic regression offers the added benefit of interpretability, making it easier for analysts to understand which features contribute most to churn. Random forests, although less transparent, are better at capturing nonlinear interactions and still provide useful information through feature importance scores. In our results, top predictors such as contract type, tenure, and monthly charges aligned with previous research

showing that short-term contracts and high billing amounts are strong indicators of churn. By identifying such drivers, telecom managers can create tailored retention strategies—such as offering discounts or long-term contracts to at-risk customers.

In conclusion, our analysis shows that simple models like logistic regression and random forest can offer meaningful insights into customer churn. However, their performance—particularly on the churn class—remains limited due to data imbalance and modeling constraints. As the literature emphasizes, the goal of churn prediction is not just accuracy, but actionable insight. When paired with business knowledge and further model refinement, these tools can help telecom providers prioritize retention efforts and make smarter customer management decisions.

## References

Abbasimehr, H., Setak, M., & Tarokh, M. J. (2014). A comparative assessment of the performance of ensemble learning in customer churn prediction. *The International Arab Journal of Information Technology*, *11*(6), 599–606. https://www.researchgate.net/publication/333194418_A_Comparative_Assessment_of_the_Performance_of_Ensemble_Learning_in_Customer_Churn_Prediction

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), 28. https://doi.org/10.1186/s40537-019-0191-6

Ahn, J. H., Han, S. P., & Lee, Y. S. (2006). Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry. *Telecommunications Policy*, *30*(10-11), 552 568. https://doi.org/10.1016/j.telpol.2006.09.006

Ascarza, E. (2018). Retention futility: Targeting high-risk customers might be ineffective. *Columbia Business School Research Paper No. 16-28*. https://doi.org/10.1509/jmr.16.0163

Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. *Expert Systems with Applications*, 36(3), 4626-4636. https://doi.org/10.1016/j.eswa.2008.05.027

Huang, B., Kechadi, M. T., & Buckley, B. (2012). Customer churn prediction in telecommunications. *Expert Systems with Applications*, *39*(1), 1414-1425. https://doi.org/10.1016/j.eswa.2011.08.024

Ijaz, M., Asghar, Z., & Gul, A. (2019). Ensemble of penalized logistic models for classification of high-dimensional data. *Communications in Statistics - Simulation and Computation, 50*(7), 2072–2088. https://doi.org/10.1080/03610918.2019.1595647

Kostić, S. M., Simić, M. I., & Kostić, M. V. (2020). Social network analysis and churn prediction in telecommunications using graph theory. *Entropy*, *22*(7), 753. https://doi.org/10.3390/e22070753

Kumar, V., & Shah, D. (2004). Building and sustaining profitable customer loyalty for the 21st century. *Journal of Retailing, 80*(4), 317-329. https://doi.org/10.1016/j.jretai.2004.10.007

Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., & Mason, C. H. (2006). Defection detection: Measuring and understanding the predictive accuracy of customer churn models. *Journal of Marketing Research*, *43*(2), 204-211. https://doi.org/10.1509/jmkr.43.2.204

Sana, J. K., Abedin, M. Z., Rahman, M. S., & Rahman, M. S. (2022). A novel customer churn prediction model for the telecommunication industry using data transformation methods and feature selection. *PLoS ONE, 17*(12), e0278095. https://doi.org/10.1371/journal.pone.0278095

Sikri, A., Jameel, R., Idrees, S. M., & Kaur, H. (2024). Enhancing customer retention in telecom industry with machine learning driven churn prediction. *Scientific Reports, 14*(1), 13097. https://doi.org/10.1038/s41598-024-63750-0