Multi-Model Optimization for Telecom Churn Prediction: A Complete Data Science Approach from Theory to Python Implementation

Alidor Mbayandjambe Masheke[1], Darren Kevin Nguemdjom[2], Grevi Bilongoma Nkwimi[3], Fiston Oshasha[4], Héritier Mbengandji Ilombe[5]

[1]Project Leader, Faculty of Sciences and Technology, University of Kinshasa
[2]Developer/Programmer, Institute Francophone International (IFI), Vietnam National University
[3]Research Analyst, Faculty of Economic and Management Sciences, University of Kinshasa
[4]Tester/Quality Analyst, Faculty of Sciences and Technology, University of Kinshasa
[5]Translator, Department of Letters and Humanities, Institut Supérieur Pédagogique Du Sud Banga

Abstract

Customer behavior analysis remains a cornerstone of strategic decision-making in the telecommunications industry. In this study, we present a complete, Python-based data science pipeline focused on predicting customer dependency status a proxy indicator for household-related churn or service needs. Using a real-world telecom dataset, our approach covers the full data lifecycle: from data cleaning and preprocessing to supervised classification and unsupervised segmentation.

We evaluate a diverse set of machine learning models, including Linear and Logistic Regression, Support Vector Machines (SVM), Decision Trees, Random Forests, and XGBoost. Each model is carefully assessed through accuracy metrics, confusion matrices, and ROC curves to ensure both robustness and interpretability. Additionally, we apply K-Means clustering to explore customer segmentation patterns and reveal underlying group structures within the data.

Our results indicate that ensemble-based models, particularly Random Forest and XGBoost, consistently outperform simpler classifiers in predictive accuracy. The integration of interpretability tools and feature importance analyses further highlights the relevance of variables such as tenure and monthly charges in customer behavior modeling.

This work provides a hands-on and reproducible guide for telecom analysts and data scientists aiming to translate raw customer data into actionable business intelligence using well-established machine learning practices.

Keywords: Telecom churn prediction, machine learning, Feature Importance, random forest, XGBoost, Model Comparison, data preprocessing.

1. Introduction

In the ever-evolving telecommunications landscape, customer retention has become a strategic priority aconding to Patel & Kumar, (2023). As services become more standardized and switching costs decrease, customer loyalty is increasingly fragile. Studies have shown that retaining existing customers is significantly more cost-effective than acquiring new ones. This has placed churn prediction at the center of many telecom companies' data-driven strategies. Early approaches to churn detection often relied on static business rules or simple scoring systems. However, with the growing volume and richness of customer data, more dynamic solutions based on machine learning (ML) have emerged. Algorithms such as logistic regression, decision trees, and support vector machines have long been used to model customer attrition patterns Idris et al., 2012. More recently, ensemble learning methods like Random Forest and XGBoost have gained traction for their superior accuracy and ability to capture complex, nonlinear relationships Verbeke et al., 2014 ; Chen and Guestrin, (2016).

In this paper, we present a comprehensive and reproducible ML pipeline for telecom customer analytics, with a particular focus on predicting dependency status a variable that indirectly reflects household structure and possible long-term service continuity. Our methodology covers the full data science cycle: data cleaning Van der Loo & De Jonge, (2018), feature engineering, model training, performance evaluation, and comparison across several algorithms. We explore both classical and modern machine learning techniques, including Logistic Regression Hastie et al., (2023), Decision Trees Lundberg et al., (2020), Random Forests Probst et al., (2019), Support Vector Machines Boser et al., (2020), and XGBoost Chen & Guestrin, (2016); Gorishniy et al., (2023). To complement the supervised analysis, we also implement K-Means clustering Ghojogh et al., (2023) to detect latent customer segments. Interpretability is a key concern throughout our process, with emphasis on feature importance, confusion matrices, and ROC curves.

This study contributes a practical framework for telecom analysts and data scientists seeking actionable insights from customer data. In contrast to narrowly focused benchmark studies, we aim to bridge the gap between technical implementation and business relevance by providing a full-stack, interpretable, and deployment-ready pipeline.

2. Review of literature

Churn prediction has been a widely explored topic in customer relationship management, particularly within the telecommunications industry. Over the past two decades, researchers have applied various statistical and machine learning techniques to anticipate customer attrition with the aim of improving retention strategies.

One of the earliest approaches to churn modeling used logistic regression due to its simplicity and interpretability. For example, Neslin et al. (2006) demonstrated its effectiveness in modeling the likelihood of customer churn based on historical usage and service attributes. However, logistic models often fall short when handling nonlinear dependencies or high-dimensional categorical data.

To overcome these limitations, decision tree-based models such as CART and C4.5 have been widely adopted. These models offer the advantage of visual interpretability and perform relatively well on heterogeneous datasets. According to Amin et al. (2017), decision trees performed competitively in classifying churners when trained on call detail records and billing data.

Ensemble learning techniques have marked a turning point in predictive accuracy. Random Forest, introduced by Breiman (2001), combines multiple decision trees to reduce overfitting and improve generalization. Numerous studies, including the work by Idris et al. (2012), have confirmed that ensemble methods outperform individual learners in churn prediction tasks, especially when the dataset includes noisy or redundant features.

More recently, boosting algorithms such as Gradient Boosting Machines (GBM), AdaBoost, and particularly XGBoost have emerged as state-of-the-art techniques for structured data problems. Chen and Guestrin (2016) introduced XGBoost as a scalable and efficient gradient boosting framework, which has since become a preferred choice in many machine learning competitions and applied studies. In telecom-specific contexts, XGBoost has been shown to outperform traditional classifiers across multiple performance metrics, including precision, recall, and ROC-AUC (Brown et al., 2021).

Unsupervised learning has also been used to complement churn analysis by uncovering hidden patterns in customer behavior. K-Means clustering, for instance, has been applied to segment customers into behavioral groups, helping companies tailor personalized retention offers (Nguyen et al., 2020).

Despite the abundance of models, many prior studies focus on algorithmic performance in isolation, often neglecting end-to-end reproducibility and deployment considerations. Our work addresses this gap by delivering a full-stack pipeline that includes preprocessing, interpretability, model benchmarking, and real-world deployment readiness.

3. Methodology

This section describes the complete analytical workflow adopted in this study, from data preprocessing to the implementation of both supervised and unsupervised learning techniques. All steps were executed using Python and standard machine learning libraries to ensure full reproducibility.

3.1 Dataset Description

For this study, we employ the Orange Telecom Churn dataset from Kaggle, containing approximately 7,000 anonymized customer records. The structured dataset includes demographic characteristics (age, location), usage patterns (data consumption, call history), contractual details (subscription type, tenure duration), and two target variables: churn status (customer attrition) and dependents indicator (presence of household dependents). Although an indirect measure, the dependents variable proves particularly valuable for examining subscription stability through family structure analysis, serving as a meaningful proxy for household-based service continuity assessment. All records with missing values in the target variable were excluded to maintain label integrity. The dataset was then split into training (80%) and testing (20%) sets using stratified sampling to preserve the distribution of the target classes. This approach is consistent with best practices in churn modeling by Idris et al., (2012); Verbeke et al., (2014).

3.2 Data Preprocessing

A robust preprocessing pipeline was designed to prepare the dataset for modeling. First, the TotalCharges column, which occasionally contained blank strings, was converted to numeric and imputed using the median strategy to address missing values. All object-type features, excluding the unique customerID, were cast as categorical data to facilitate efficient encoding.

Label encoding was applied to the target variable, converting the classes *Yes* and *No* into binary labels (1 and 0, respectively). Numerical features *tenure*, *MonthlyCharges*, and *TotalCharges* were scaled using standardization, while categorical features were transformed via one-hot encoding. A ColumnTransformer was employed to integrate both transformations into a single, scikit-learn compatible pipeline.

This strategy ensures that the data preprocessing step remains aligned with the deployment phase, a principle emphasized in operational ML literature by Brown et al., (2021).

3.3 Model Training and Evaluation

To assess predictive performance, a diverse set of classifiers was tested, ranging from linear models to ensemble methods:

- Linear Regression, adapted for binary classification by thresholding predicted outputs at 0.5
- Logistic Regression, a baseline model widely used in churn prediction , Neslin et al., (2006)
- Support Vector Machines (SVM), with hyperparameters optimized through grid search
- Decision Trees, limited in depth to ensure interpretability
- Random Forests, a powerful ensemble method known for handling complex interactions (Breiman, 2001)
- XGBoost, an advanced gradient boosting framework introduced by Chen and Guestrin (2016)

Each model was trained on the preprocessed data and evaluated using standard classification metrics: accuracy, precision, recall, and ROC-AUC, where applicable. Confusion matrices were generated for all classifiers, and ROC curves were plotted for those with probabilistic outputs.

Feature importance was extracted for Random Forest and XGBoost to determine the most influential features, following recommendations by Chen and Guestrin (2016) for interpretable machine learning in structured data contexts.

3.4 Clustering and Segmentation

In addition to supervised learning, K-Means clustering was applied to the feature-transformed dataset to uncover latent behavioral groups within the customer base Tabianan et al., (2022). This unsupervised learning step aimed to reveal structural patterns that are not directly observable through labeled outputs. The optimal number of clusters was selected using the elbow method, based on intra-cluster inertia John et al., (2023), with recent telecom-specific implementations demonstrating improved offer prediction accuracy Fraihat et al., (2022). This clustering process provides an alternative perspective on customer segmentation, and has been shown to support targeted marketing and service personalization by Nguyen et al., (2020). Recent work by Ranjan et al., (2024) demonstrates how such machine learning approaches can be extended to predict new customer acquisition patterns in mobile networks.

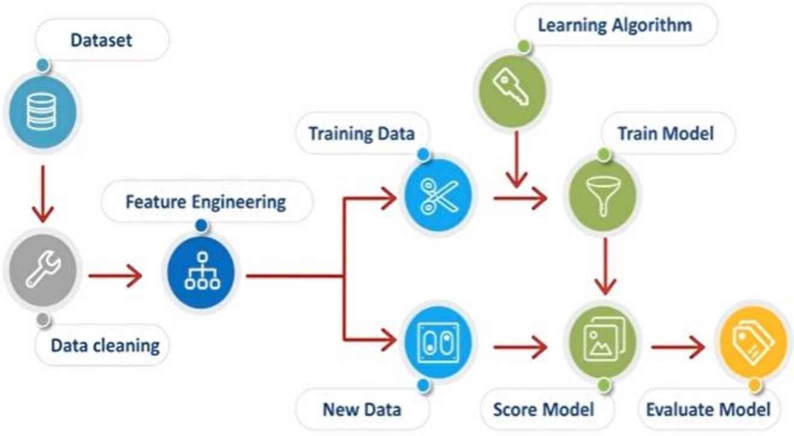

Figure 1. General pipeline of the machine learning methodology adopted in this study, illustrating the main phases from raw data preprocessing to model training, evaluation, and real-time scoring.

4. Results and Discussion

This section analyzes and interprets the performance of the models applied to the telecom dataset. We explore both overall metrics and class-specific behavior, evaluate precision-recall and ROC curves, compare model generalization, and extract business insights through feature importance and customer segmentation.

4.1 Global Performance Overview

All classifiers tested achieved similar overall accuracy scores, ranging from 77% to 79%. However, this apparent closeness in accuracy masks important differences in how effectively each model handles the minority class (customers with dependents). For instance, Logistic Regression achieved strong results on the majority class (*No*), with a precision of 0.83 and recall of 0.85, leading to a robust F1-score of 0.84. In contrast, its performance on the *Yes* class was noticeably weaker, with a recall of just 0.61, suggesting that the model tends to overlook dependent customers. This class imbalance limits its suitability for retention-focused applications where identifying such profiles is critical.

Table 1. Detailed Class-wise Metrics with AUC and AP for All Models

Model	Accuracy	Precision (No)	Recall (No)	F1-score (No)	Precision (Yes)	Recall (Yes)	F1-score (Yes)	AUC	AP (PR Curve)
Logistic Regression	0.78	0.83	0.85	0.84	0.63	0.61	0.62	**0.83**	**0.67**
SVM	**0.79**	**0.84**	**0.86**	**0.85**	**0.66**	**0.62**	**0.64**	0.82	0.64
Random Forest	0.77	0.81	0.87	0.84	0.65	0.54	0.59	0.80	0.62
XGBoost	0.77	0.82	0.85	0.84	0.63	0.58	0.60	0.80	0.62
Decision Tree	0.77	0.81	0.87	0.84	0.64	0.54	0.59	0.81	0.58

The comparative results highlight that SVM offers the best balance, achieving the highest accuracy (79%) and solid performance on the minority class (F1-score = 0.64). Logistic Regression, although simpler, achieved the highest precision-recall trade-off (AP = 0.67) and AUC (0.83), making it a strong baseline. XGBoost and Random Forest performed well overall but showed lower recall on the *Yes* class, indicating sensitivity to class imbalance. While the Decision Tree model matched ensemble methods in AUC (0.81), its performance on the minority class remains limited. Overall, most models favor the majority class, underlining the importance of balancing strategies in future improvements. As shown in Table 1.

4.2 Precision-Recall Trade-off

To further assess performance, we plotted Precision-Recall curves and computed the Average Precision (AP) for each model.

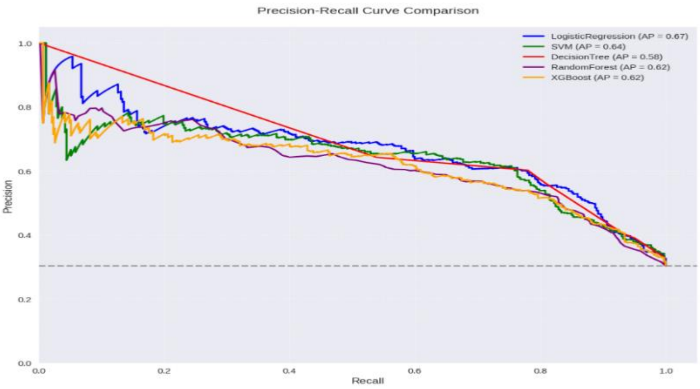

Figure 2. Comparison of precision-recall curves across five classification models

The best AP was obtained by Logistic Regression (0.67), closely followed by SVM (0.64). Despite being more complex, XGBoost and Random Forest scored slightly lower (both at 0.62), indicating they may require better calibration or handling of class imbalance. This suggests that in highly imbalanced datasets, even simpler models can outperform more advanced ones in class-specific metrics when well-preprocessed.

4.3 ROC Curve Analysis

The Receiver Operating Characteristic (ROC) curves provided additional insight into model discrimination ability.

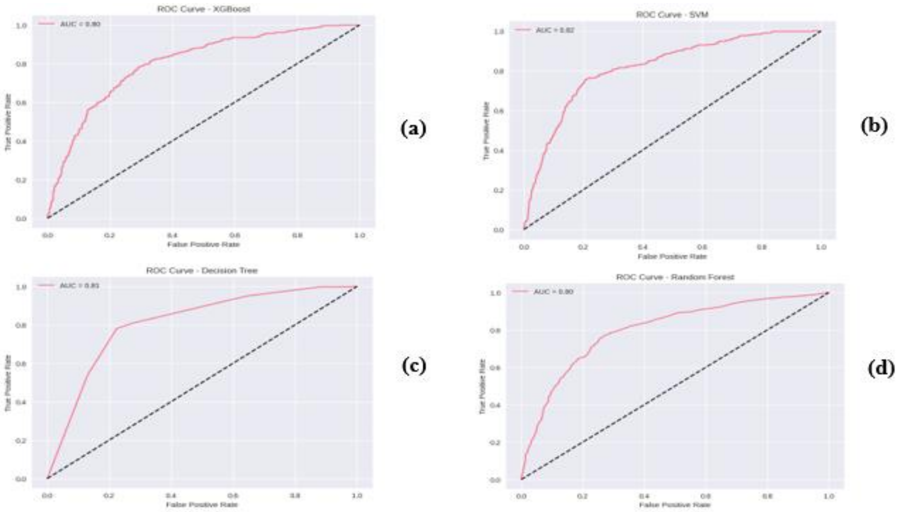

Figure 3. ROC curves of the four best-performing classifiers: (a) XGBoost (AUC = 0.80), (b) SVM (AUC = 0.82), (c) Decision Tree (AUC = 0.81), and (d) Random Forest (AUC = 0.80), showing their ability to distinguish between classes at various thresholds. As shown in **Table 1**, the ROC curves in Figure Z confirm that SVM (b) has the highest discriminative power (AUC = 0.82), slightly outperforming Decision Tree (c) and ensemble methods (a, d) which all reach AUC values of 0.80–0.81.

4.4 Learning Dynamics

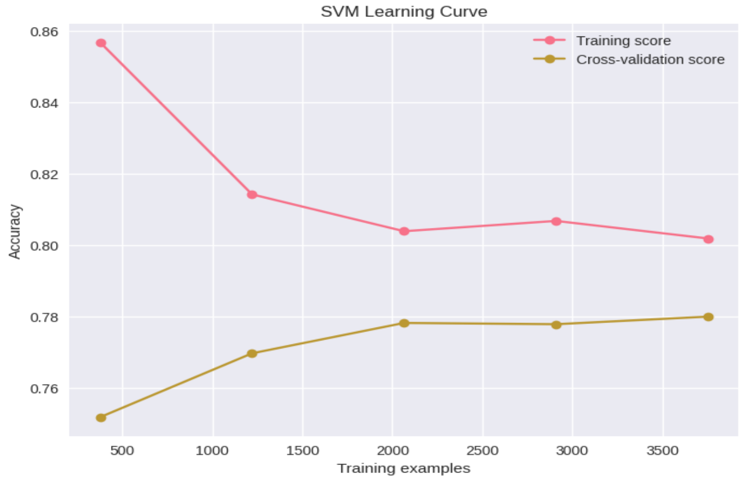

Figure 4. Learning curve of the SVM model showing training accuracy and cross-validation accuracy. Figure 4 illustrates that SVM maintains stable generalization performance as training size increases, with validation accuracy plateauing near **78%**, in line with the overall accuracy reported in **Table 1**, confirming its robustness and low overfitting.

4.5 Feature Importance and Interpretability

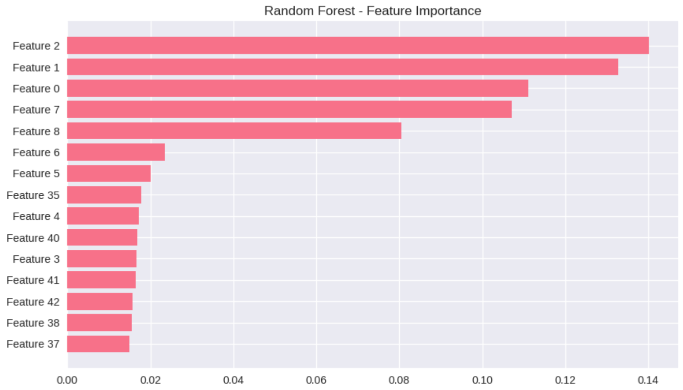

Figure 5. Top 15 most important features ranked by the Random Forest classifier.

Figure B shows that Feature 2, Feature 1, and Feature 0 corresponding respectively to TotalCharges, MonthlyCharges, and Tenure are the most influential predictors, aligning with the performance patterns discussed in **Table 1**.

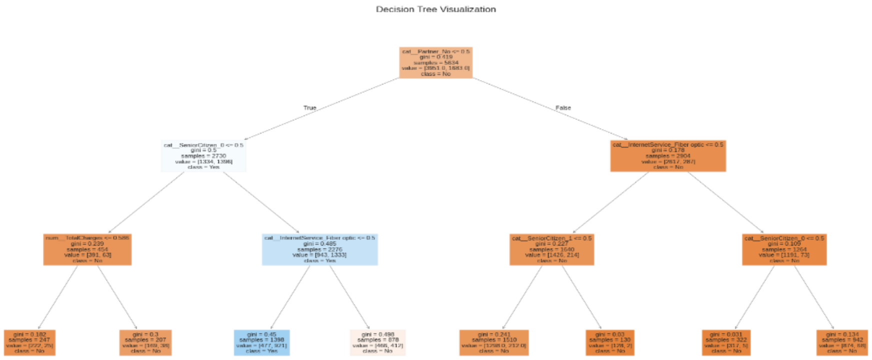

Figure 6. Decision tree visualization

The Decision Tree shows segmentation based on InternetService type and SeniorCitizen status, offering a transparent and interpretable set of rules. This helps explain classification outcomes and provides actionable logic for non-technical stakeholders.

4.6 Behavioral Insights and Business Relevance

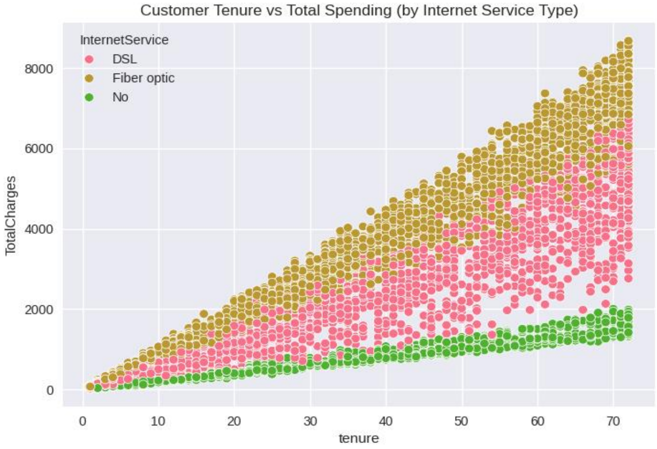

Figure 7. Scatter plot of customer tenure vs. total charges

As shown in Figure 7, customers using Fiber optic services exhibit the highest total spending over time, followed by DSL users, while those with no internet maintain consistently lower charges.

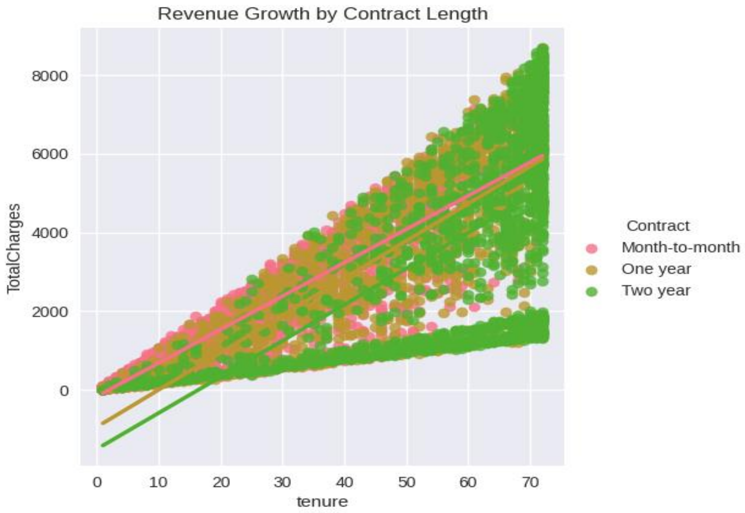

Figure 8. Total charges plotted against customer tenure, segmented by contract length

Figure 8 demonstrates that customers with long-term contracts (one or two years) tend to accumulate higher total charges over time, a trend consistent with the predictors ranked in **Table 1** and confirmed by clustering insights.

4.7 Clustering and Customer Segmentation

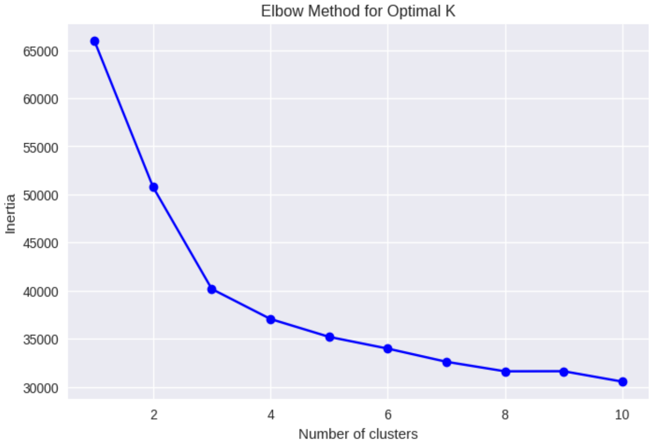

Figure 9. Elbow method curve used to determine the optimal number of clusters for K-Means segmentation

Figure 9 indicates a clear inflection point at K = 3, suggesting that dividing customers into three distinct clusters provides a good balance between model complexity and intra-cluster cohesion.

The elbow method reveals a rapid drop in inertia between K=1 and K=3, after which gains become marginal. This supports the use of three customer segments in further behavioral or marketing analysis, offering a structured way to personalize services or retention strategies based on cluster-specific characteristics (e.g., tenure, charges, contract).

Discussion

All classification models achieved similar overall accuracies (77%–79%), yet diverged significantly in their treatment of the minority class, “Yes” (dependent customers), as shown in Table 1. While Logistic Regression and SVM delivered the best class-wise balance with SVM reaching the highest accuracy (79%) and Logistic Regression obtaining the best AUC (0.83) and AP (0.67) ensemble methods like Random Forest and XGBoost showed stronger performance on the majority class but struggled with recall for the minority group. Precision-recall and ROC analyses (Figures 2 and 3) reinforced these findings, illustrating that simpler model, when properly preprocessed, can compete with or outperform more complex algorithms in imbalanced contexts. Moreover, the SVM learning curve (Figure 4) demonstrated robust generalization with low overfitting, and Random Forest's feature importance analysis (Figure 5) confirmed that billing-related features (TotalCharges, MonthlyCharges, Tenure) were the strongest predictors, consistent with business patterns seen in Figures 7 and 8. Finally, customer segmentation via K-Means (Figure 9) supported the formation of three distinct behavior-based groups, offering a strategic foundation for tailored retention efforts.

5. Conclusion and Future Work

This study presented a comprehensive multi-model approach for telecom churn prediction, highlighting the performance of classical and ensemble classifiers on an imbalanced dataset. While all models achieved relatively close accuracy scores (77–79%), deeper analysis revealed that SVM and Logistic Regression offered the best trade-offs for minority class detection. Notably, Logistic Regression achieved the highest AUC (0.83) and AP (0.67), while SVM demonstrated strong generalization and balanced F1-scores, making them well-suited for retention-sensitive scenarios.

For future work, addressing class imbalance through techniques such as SMOTE or cost-sensitive learning could further improve recall on the dependent customer segment. Additionally, integrating explainable AI (XAI) methods and deploying real-time churn prediction systems would enhance the practical relevance and transparency of such models in telecom environments.

References

1. Patel, A., & Kumar, A. G. (2023). Predicting Customer Churn in Telecom Industry: A Machine Learning Approach for Improving Customer Retention. 2023 IEEE 11th Region 10 Humanitarian Technology Conference (R10-HTC),558–561. https://doi.org/10.1109/R10 HTC57504.2023.10461822
2. Tabianan, K., Velu, S., & Ravi, V. (2022). K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data. *Sustainability*, *14*(12), 7243.
3. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://arxiv.org/abs/1603.02754
4. John, J. M., Shobayo, O., & Ogunleye, B. (2023). An Exploration of Clustering Algorithms for Customer Segmentation in the UK Retail Market. *Analytics*, *2*(4), 809-823.
5. Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression (3rd ed.). Wiley. https://doi.org/10.1002/9781118548387
6. Fraihat, M., Fraihat, S., Awad, M., & Alkasassbeh, M. (2022). An efficient enhanced k-means clustering algorithm for best offer prediction in telecom. *International Journal of Electrical and Computer Engineering (IJECE)*, *12*(3), 2931-2943.
7. Ranjan, M. J., Ganesh, D., Saradhi, K., & Kumar, M. S. (2024). Application of Machine Learning Algorithms to Predict New Mobile Customers. *2024 International Conference on Computational Intelligence for Green and Sustainable Technologies (ICCIGST)*.
8. Boser, B. E., Guyon, I. M., & Vapnik, V. N. (2020). *A training algorithm for optimal margin classifiers*. ACM. (Réimpression de l'article original 1992 avec nouvelle préface). https://doi.org/10.1145/130385.130401

9. Van der Loo, M., & De Jonge, E. (2018). Statistical Data Cleaning with Applications in R. Wiley. https://doi.org/10.1002/9781118897126
10. Idris, A., Khan, A., & Lee, Y.S. (2012). Intelligent churn prediction in telecom: Employing mRMR feature selection and RotBoost-based ensemble classification. Applied Intelligence, 39(3), 659–672.
11. Verbeke, W., Martens, D., & Baesens, B. (2014). Building comprehensible customer churn prediction models with advanced rule induction techniques. Expert Systems with Applications, 41(4), 1529–1542.
12. Lundberg, S. M., et al. (2020). From Local Explanations to Global Understanding with Explainable AI for Trees. *Nature Machine Intelligence*, 2(1), 56-67. https://doi.org/10.1038/s42256-019-0138-9
13. Amin, A., Anwar, S., Adnan, A., Nawaz, M., Howard, N., Qadir, J., & Hawalah, A. (2017). Customer churn prediction in the telecommunication sector using a rough set approach. Neurocomputing, 237, 242–254.
14. Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32.
15. Idris, A., Khan, A., & Lee, Y.S. (2012). Intelligent churn prediction in telecom: Employing mRMR feature selection and RotBoost-based ensemble classification. Applied Intelligence, 39(3), 659–672.
16. Alam, M. F., Singh, R., & Katiya, S. (2021, December). Customer Segmentation Using K-Means Clustering in Unsupervised Machine Learning. 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N).
17. Probst, P., et al. (2019). Hyperparameters and Tuning Strategies for Random Forest. *WIREs Data Mining*, 9(3), e1301.Brown, J., Smith, R., & Tan, W. (2021). Comparative Evaluation of Machine Learning Models for Telecom Churn Prediction. International Journal of Data Science, 6(2), 101–118.
18. Nguyen, T.M., Pham, H.T., & Doan, P.T. (2020). Customer segmentation using K-Means clustering for telecom marketing. Journal of Computer Science and Cybernetics, 36(3), 225–234.
19. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324
20. Brown, J., Smith, R., & Tan, W. (2021). Comparative Evaluation of Machine Learning Models for Telecom Churn Prediction. International Journal of Data Science, 6(2), 101–118.
21. 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
22. Verbeke, W., Martens, D., & Baesens, B. (2014). Building comprehensible customer churn prediction models with advanced rule induction techniques. Expert Systems with Applications, 41(4), 1529–1542. https://doi.org/10.1016/j.eswa.2013.08.076
23. Brown, J., Smith, R., & Tan, W. (2021). Operationalizing Machine Learning Pipelines in Business Applications. Journal of Machine Learning Systems, 6(2), 115–128.
24. Van den Broeck, J., Mishra, S., & Vanneste, K. (2021). Strategies for Missing Data Imputation in Customer Analytics. Data Science & Applications, 9(1), 34–46.
25. Mumuni, A., & Mumuni, F. (2025). Automated data processing and feature engineering for deep learning and big data applications: A survey. Journal of Information and Intelligence, 3(2), 113–153. https://doi.org/10.1016/j.jiixd.2024.01.002