Reproducible Prediction Framework of Customer Churn Using Machine Learning, Advanced Data Science and Business Intelligence Techniques

Younes KOULOU, Norelislam EL HAMI
Science and Engineering Laboratory-ENSA, Ibn Tofail University, Kenitra, Morocco

***Abstract*—The telecommunications sector has evolved in recent years, resulting in intense competition and high customer acquisition costs. As a result, retaining customers has become a key concern for telecom operators. In this work, we propose the design and implementation of a complete customer churn prediction system that combines data science, machine learning and business intelligence approaches. The methodology is structured into five main steps: exploratory data analysis, development of an ETL pipeline, feature engineering, predictive modeling using a Random Forest algorithm, and the creation of decision-support dashboards in Power BI. Random Forest demonstrated higher performance with AUC-ROC of 0,85 and the results demonstrated that the main predictors of churn are monthly charges, contract type, and customer tenure. Our approach, which validated by a confusion matrix, offers decision-makers an operational tool to anticipate departures and implement targeted loyalty actions. This study proposes a reproducible methodological framework for companies facing the problem of churn and contributes to the use of machine learning in relationship marketing.**

***Keywords*—Churn prediction; telecommunications; machine learning; random forest; business intelligence; ETL; Power BI; feature engineering; decision support**

I. Introduction

In today's digital economy, the telecommunications sector is characterized by intense competition, saturated mature markets, and particularly high customer acquisition costs. Acquiring a new customer can cost up to 25 times more than retaining an existing one [1]. In this context, customer churn poses a direct threat to the profitability and long-term viability of businesses.

For this reason, the phenomenon of churn takes on particular strategic importance in telecommunications where the sudden interruption of contractual revenues significantly affects financial forecasts, loyal customers tend to subscribe to more additional services, and the quality of customer relationship represents a decisive competitive advantage in such a market where differentiation by supply becomes difficult [2].

Churn prediction has been addressed from various methodological perspectives. There are ensemble learning approaches, which currently dominate, utilize algorithms such as Decision Trees, Random Forest, XGBoost, LightGBM, and CatBoost. These algorithms demonstrate superior performance for churn prediction in the telecommunications sector [3] [4]. In this context, [4] explored the effectiveness of ensemble learning techniques by combining several basic learners, such as decision trees and gradient boosting in an ensemble framework in order to improve predictive accuracy to that of individual learners [5]. In parallel, the advent of Deep Learning has opened up new perspectives. While other researchers have focused on the interpretability of models and the identification of critical factors affecting churn [6].

A recent systematic review analyzed 240 studies, published between 2020 and 2024, on churn prediction. The authors identified several persistent challenges: class imbalance, model interpretability, the concept of drift such as the temporal evolution of customer behavior, and the limited use of profit-oriented metrics [4].

Unlike most existing studies, which focus on algorithmic performance, we propose a holistic approach integrating the entire data value chain within a comprehensive methodological framework that includes: a rigorous ETL (Extract, Transform, Load) process containing business rules; a feature engineering phase to create interpretable variables; Optimized modeling with cross-validation and hyperparameter search; an interactive application for Operational deployment; and integration of business intelligence via BI dashboards.

This approach addresses the need to bring predictive models closer to the operational needs of businesses.

The structure of this article is as follows: Section 2 details our methodology. Section 3 presents the results. And Section 4 discusses the results and suggests some perspectives.

II. Related Work

Numerous studies have been conducted to predict customer churn. Despite their interpretability, early approaches lacked the ability to capture complex non-linear relationships through traditional statistical methods like logistic regression. The introduction of machine learning has resulted in a significant improvement in predictive performance.

Churn prediction is now dominated by ensemble methods. According to [3], the authors demonstrated that combining multiple base learners (decision trees, gradient boosting) in an ensemble framework can result in superior predictive accuracy than individual learners. Telecommunications data [3] [4] has demonstrated robust performance for Random Forest, XGBoost, LightGBM, and CatBoost.

New possibilities have been opened up by deep learning architectures simultaneously. The author in [4] conducted a

systematic review examining 240 publications from 2020 to 2024 and found that the use of temporal dependencies-captured LSTM networks is gradually being applied to churn prediction. Data volume requirements and interpretability concerns limit their implementation.

Interpretability of models has become a crucial concern. According to the authors of [6], XAI-Churn TriBoost is a model that is interpretable and integrates XGBoost, CatBoost, and LightGBM into a weighted voting ensemble, and it combines LIME and SHAP techniques to explain individual predictions and uncover critical factors impacting churn.

According to the systematic review by [4], there are many challenges that persist: class imbalance, model interpretability, concept drift (temporal evolution of customer behavior), and the limited use of profit-oriented metrics. The primary focus of existing studies is on algorithmic performance, leaving out the operational integration of predictive models into business processes. Few works address the complete data value chain from extraction to strategic visualization, which constitutes the main contribution of our research.

III. Methodology

The proposed approach is broken down into five main phases, as illustrated in Fig. 1.

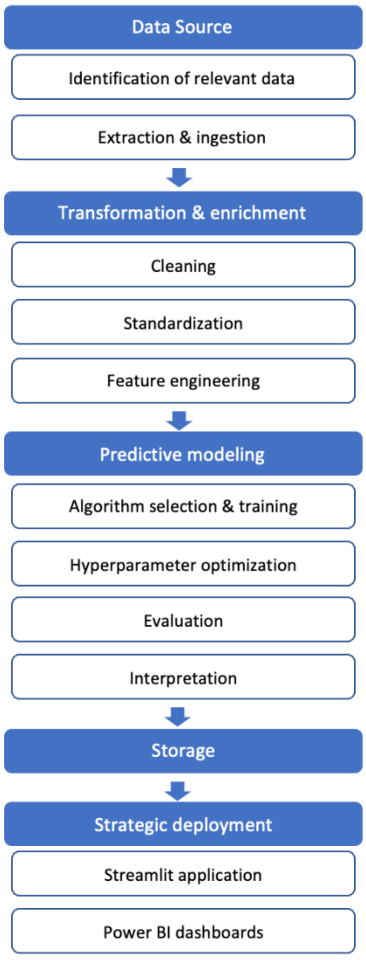

Fig. 1. Methodological diagram.

A. Data Used

The dataset used in this study comes from a telecommunications company that provides phone and internet services. It comprises 7043 observations and 21 variables. Table I presents the main characteristics of this Dataset.

TABLE I. Dataset Characteristics

Characteristic	Value
Total observations	7 043
Number of features	21
Non-churn customers	5 174 (73,5%)
Churn customers	1 869 (26,5%)
Class ratio (non-churn:churn)	2.8:1
Memory size	7.79 MB

B. Data Extraction and Ingestion

This initial phase consists of loading the data from its source and verifying its structural integrity. The dataset's informative diversity covers four categories of variables (see Table II):

TABLE II. Typology of Dataset Variables

Category	Variables	Description
Demographics	Gender, Senior Citizen, Partner, Dependents	Sociodemographic profile
Services	Phone Service, Multiple Lines, Internet Service, Online Security, Online Backup, Device Protection, Tech Support, StreamingTV, Streaming Movies	Package of subscribed services
Contractual	Contract, Paperless Billing, Payment Method	Contractual terms
Financial	Monthly Charges, Total Charges, Tenure	Financial indicators and loyalty
Target	Churn	Predictable variable (Yes/No)

C. Pipeline ETL and Feature Engineering

The reliability of predictive models depends largely on the quality of the input data. In this step, we implemented an ETL pipeline to progressively clean, structure, and transform the data as follows:

1) Data cleaning: This step aims to detect and correct data quality issues [7]. During the initial analysis, the totalcharges variable was found to contain 11 observations with spaces in place of numerical values. After examination, these cases corresponded to customers with zero tenure. To resolve this issue, the variable was converted to a numeric type, where missing values are automatically treated as null and then imputed as 0. This solution preserved observations while maintaining data integrity.

2) Handling of categorical variables: This step aims to reduce inconsistencies in categorical variables by grouping similar categories under a unified representation [8]. For example, We found that the service variables had three categories: "yes", "no", and variants such as "no internet

service" and "no telephone service". To harmonize the data, we grouped these variants into a single category "no".

3) Feature engineering: The goal of this step is to enrich the dataset by creating derived variables that offer greater analytical value and contribute to improved predictive performance. Eight new variables were created:

- ChurnFlag: conversion of the Churn variable into a binary format.
- Tenure Group: Segmentation of tenure into four classes: 0-1 year, 1-2 years, 2-4 years, >4 years.
- Total Services: This represents the number of subscribed services by each customer. Possible values 0 to 7.
- Customer Type: Behavioral typology ("Phone Only", "Internet Only", "Bundle").
- Is Premium Customer: Binary indicator for Total Services $\geqslant 5$.
- Avg Monthly Charges: Ratio of Total Charges/tenure to compare between new and existing customers.
- Family Support: Indicator of family stability (Partner=Yes AND Dependents=Yes).
- Tenure Years: Conversion to years for improved business readability.

D. Predictive Modeling

1) Data preparation for learning: 17 categorical variables were transformed using label encoding. To prevent data leakage and ensure unbiased evaluation, we split the dataset into training (80%) and test (20%) sets before any transformation. The separation is stratified, ensuring the same proportion of churn in both sets (26,5%).

Standardization (centering and scaling) was applied to the numerical variables. An important step for algorithms sensitive to scale, although Random Forest is theoretically less sensitive in this regard.

2) Choice of algorithm: The choice of the random forest classifier is based on several theoretical and practical considerations, confirmed by recent literature. The authors of [9] showed that random forest and XGBoost offer better performance with 96% accuracy, while SVM and KNN offer lower performance. Also, in [10], random forest demonstrates higher performance on e-commerce data. Among the advantages of random forest are the following [11]:

- Robustness: Aggregating multiple trees reduces sensitivity to noise and outliers.
- Non-linearity: The model allows for the capture of complex relationships without relying on parametric assumptions.
- Interpretability: Provides a measure of variable importance.
- Stability: The average of sets produces a lower variance than that of single decision trees.
- Scalability: Good performance on moderately sized datasets.

3) Hyperparameter optimization: To identify the optimal configuration, an exhaustive search for hyperparameters was performed using gridsearchcv [12], with cross-validation (5 folds) and optimization based on the roc-auc metric:

- Number of estimators: [100, 200, 300]
- Maximum depth: [10, 20, 30, None]
- Minimum number of samples per leaf: [1, 2, 4]
- Split quality criterion: ['gini', 'entropy']

4) Evaluation metrics: To evaluate the model's performance, we used a set of complementary metrics:

- Accuracy: Gives the overall proportion of correct classifications.
- Precision: Measures the number of churners predicted who actually churn.
- Recall: Proportion of churn detected among all actual churn.
- F1-Score: Provides a balanced view by combining precision and recall.
- ROC-AUC: Model's ability to discriminate between classes.
- Confusion matrix: Offers visualization of prediction errors.

E. Deployment and Decision-Making Integration

1) Deployment architecture: After optimization, the final model was saved with all necessary components, e.g., scaler, encoders, and feature list, to ensure reproducibility for production deployment.

2) Streamlit application: For deployment, we propose using Streamlit, a Python framework that transforms scripts into interactive web applications without needing certain front-end skills. This step aims to make the models accessible to business teams [13].

The interactive web application offers:

- A form for inputting customer characteristics.
- Real-time calculation of churn probability.
- Risk level classification (Low, Medium, High) based on thresholds.
- Automated recommendations personalized to the risk profile.

3) Business intelligence integration: The goal of this phase is to structure the data and model predictions in a MySQL database and then connect them to Power BI. This architecture

combines descriptive analytics (BI) and predictive analytics (ML) within the same decision-making environment.

IV. Results and Analysis

A. Exploratory Analysis Results

1) Data structure and quality: The loading of the dataset exposed a structure of 7,043 rows and 21 columns, which is manageable for iterative processing. Analysis of variable types identified 18 categorical variables (object) and 3 numerical variables.

A variable "Total Charges" represents quantitative financial data, and we observed that it was stored in object format. This indicated that a conversion would be necessary.

2) Distribution of the target variable: Analysis of the churn variable indicates an imbalanced class configuration (see Fig. 2):

- Non-churn customers: 5,174 (73,5%)
- Churn customers: 1,869 (26,5%)
- Overall churn rate: 26,5%
- Class ratio: 2,8 (non-churn) to 1 (churn)

Moderate imbalances require careful attention during modeling, especially when selecting evaluation metrics.

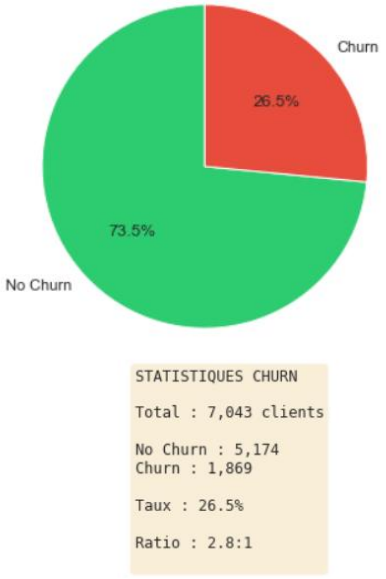

Fig. 2. Churn variable Analysis.

3) Discriminant factors of churn: Analysis identified a number of variables that were strongly associated with churn: (See Fig. 3, Fig. 4, Fig. 5, and Fig. 6).

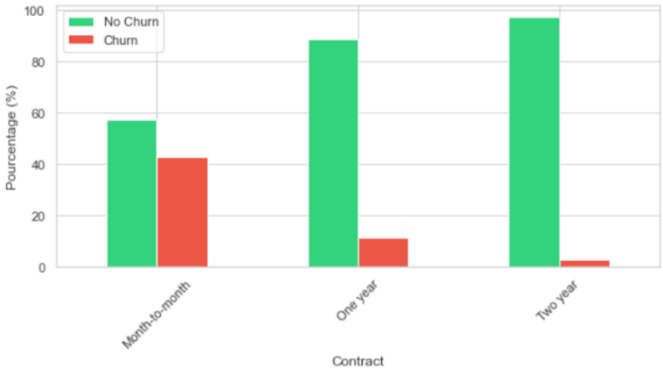

Fig. 3. Churn rate by contract.

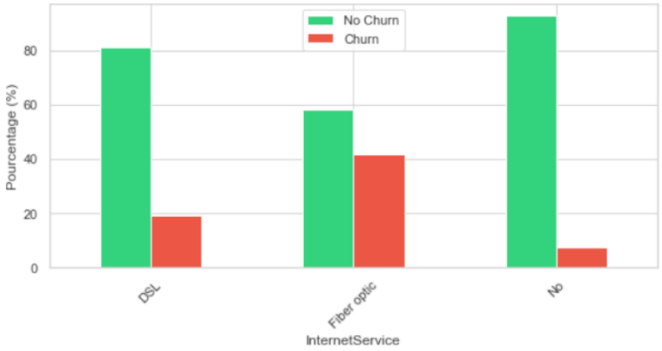

Fig. 4. Churn rate by internet service.

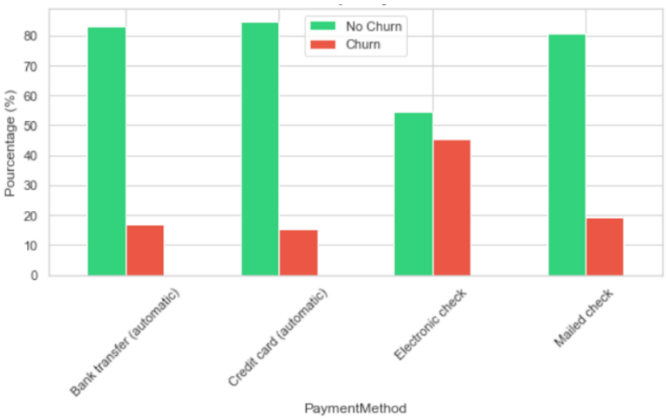

Fig. 5. Churn rate by payment method.

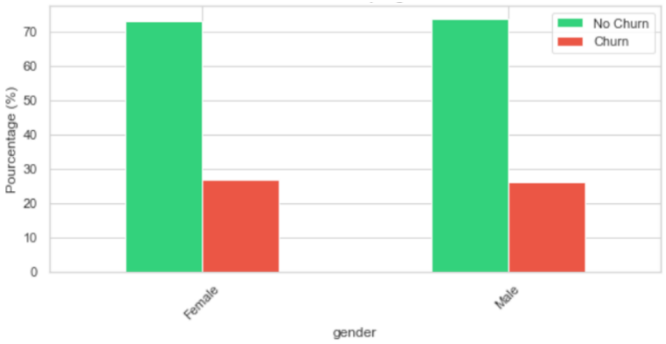

Fig. 6. Churn rate by gender.

a) Contract type: The number of customers who canceled their contracts was 42,7% for monthly customers, whereas those with one year and two years contracts have rates of only 11,3% and 2,8%. This difference highlights the effect of long-term commitments.

b) Internet service: With a churn rate of 41,9%, fiber optic customers depart more frequently than DSL users 19% and those without internet connection. This may reflect higher expectations among fiber users or more intense competition in this segment.

c) Payment method: Customers who pay by electronic check are highest chance of unsubscribing, with a rate of 45,3%. This rate is too higher than those paying by bank transfer at 26,5%, credit card at 21,9%, or by mailed check at 21%.

d) Gender: The difference is minimal, 26,1% of men churn and 26,9% of women, meaning that gender does not play a significant role.

4) Analysis of numerical variables: This analysis of continuous variables identified different profiles between churned and non-churned customers:

a) Tenure: Median duration for churned customers is 10 months compared to 38 months for non-churned customers.

b) Total Charges: Average among churned customers (1531$) is lower than the non-churned customers (2555$).

The correlation matrix (see Fig. 7) indicates a negative correlation between tenure and churn (-0,352), a positive correlation with Monthly Charges (0,193), and a weak correlation with Senior Citizen (0,151).

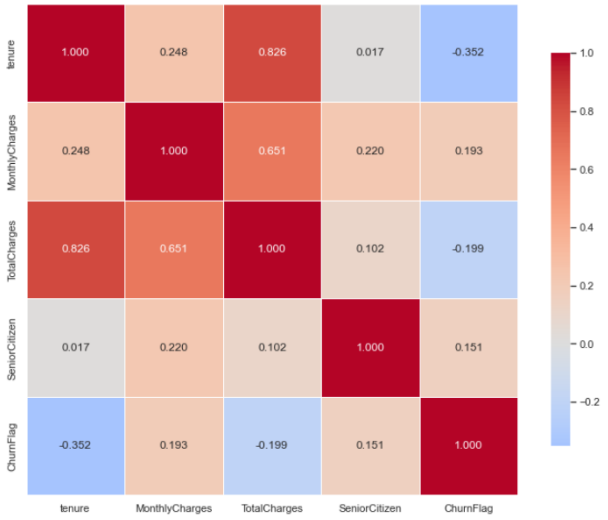

Fig. 7. Correlation matrix.

B. Modeling Results

1) Baseline model performance: The baseline random forest model, without optimization, attained the following performance on the test set:

- Accuracy: 79%
- Precision: 63%
- Recall: 48%
- F1-Score: 55%

These results provide a functional performance, but recall at 48% indicates a limitation. In other words, more than half of churned customers are not detected.

2) Optimized model performance: In order to improve the model performance, an optimization using gridsearchcv has been applied (see Table III):

TABLE III. PERFORMANCE COMPARISON BEFORE/AFTER OPTIMIZATION

Metric	Initial model	Optimized model	Improvement
Accuracy	79%	80%	+1,3%
Precision	63%	65%	+3,2%
Recall	48%	49%	+2,1%
F1-Score	55%	56%	+1,8%
ROC-AUC	0,83	0,85	+2,4%

Although these gains may seem modest, each additional percentage point of recall represents dozens of identified customers. Therefore, the improvement can be considered significant.

3) Confusion matrix analysis: The confusion matrix of the optimized model is illustrated in Fig. 8:

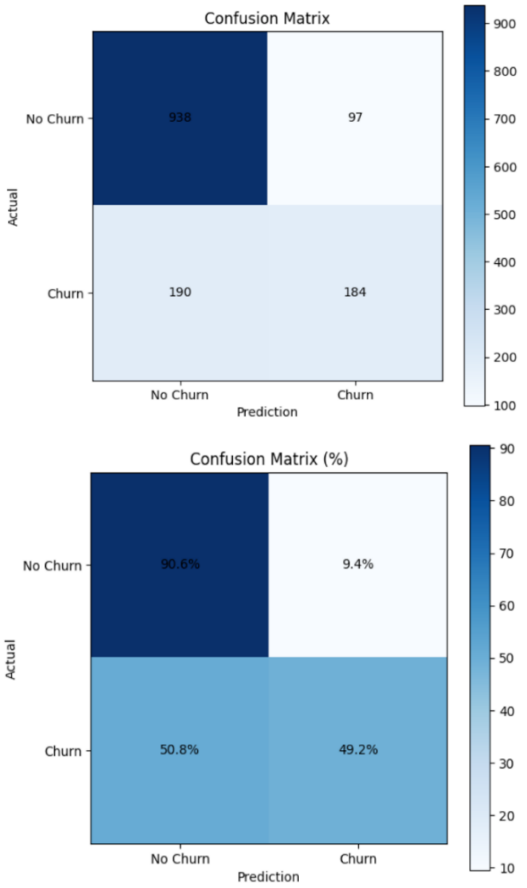

Fig. 8. Confusion matrix

The model prioritizes reducing false positives (97 loyal customers identified) at the expense of a higher recall rate (190 false negatives).

In an operational context, mobilizing retention resources to customers who will not leave (false positives) means spending without return, while failing to identify those who will leave (false negatives) means lost revenue.

4) ROC curve: The ROC curve presents an area under the curve greater than 0.84 (see Fig. 9). This indicates an excellent distinction capability of the model, and the value, above 0,8, is generally considered acceptable in real contexts.

The deviation from the random diagonal confirms that the model provides real added value compared to a random prediction.

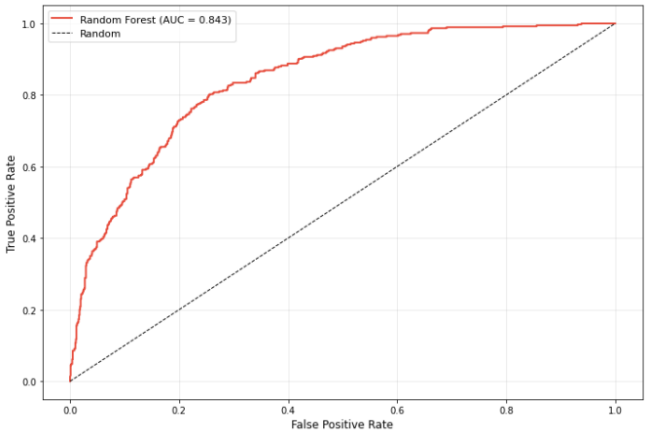

Fig. 9. Courbe ROC

5) Features importance: Fig. 10 illustrates the hierarchy of the most influential predictors detected during the analysis of variable importance.

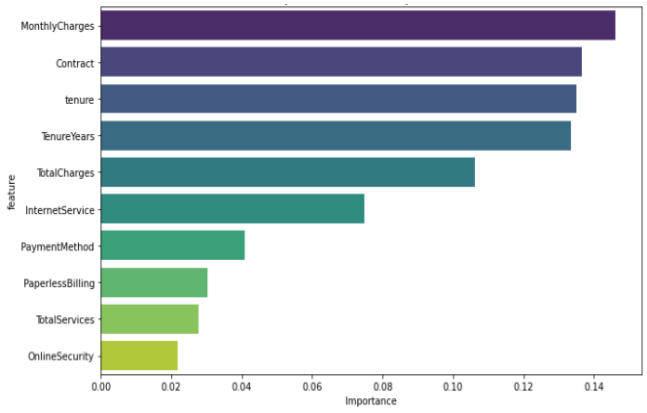

Fig. 10. Features importantes.

C. Operational Deployment Results

1) Streamlit application: Real-time interaction with the model is made possible through the application that we developed and usability tests have demonstrated:

- Less than 2 minutes on average for each customer's input.
- Intuitive comprehension of the results, such as probability and risk level.

The application was structured to support two different ways of using:

- Individual prediction for a specific customer (customer service).
- Batch prediction via file upload (marketing campaigns).

2) Power BI dashboards: Two dashboards were created to meet various decision-making needs (see Fig. 11, Fig. 12):

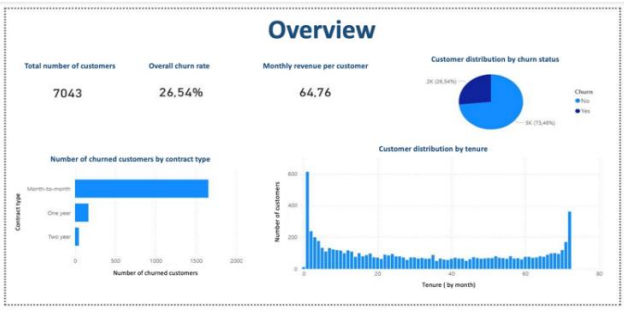

Fig. 11. Dashboard overview.

a) Overview: Designed for strategic management. It presents key performance indicators such as the number of customers, churn rate, average revenue, etc., and the classification of churn by contract and tenure. It provides an immediate overview of the portfolio.

b) Detailed analysis: Focused on marketing analysis. This report explores the relationships between churn and explanatory variables such as internet service, payment method, and cross-analysis of contract and payment. Churn evolution by tenure is also visualized. And a list of the top ten customers with the level of risk they pose.

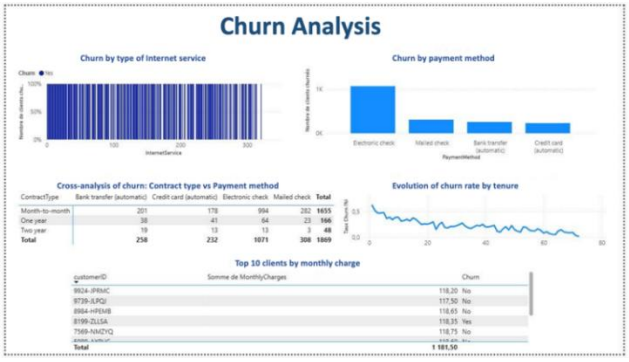

Fig. 12. Dashboard: Detailed Analysis.

V. Discussion

A. Interpretation of Results

The results obtained confirm and enrich the existing literature on churn prediction, while providing important nuances regarding the hierarchy of predictive factors.

Contrary to several studies that place tenure as the top predictor [4], our analysis reveals that MonthlyCharges constitute the most influential factor. This result suggests that price sensitivity is a major determinant of the cancellation decision.

Contract type ranks second, confirming the protective effect of long-term commitments. Month-to-month contract customers have a churn rate of 42.7%, compared to only 11.3% for one-year contracts and 2.8% for two-year contracts.

Tenure as a protective factor, but not preeminent. Although tenure ranks third, it does not dominate the hierarchy. This result nuances the idea of an automatic "lock-in" effect linked to relationship duration.

The presence of TotalServices in the top nine validates the relevance of the feature engineering work. The more services a customer consumes, the more engaged they are and the less likely they are to cancel.

B. Impact of Class Imbalance

The dataset exhibits a moderate class imbalance with a ratio of 2.8:1. While this challenge is identified in the literature [4], we chose not to apply oversampling or under-sampling techniques. Random Forest is inherently robust to moderate imbalance due to its bootstrap sampling and tree aggregation mechanisms. Furthermore, applying balancing techniques would alter the real-world class distribution, potentially leading to overoptimistic estimates that do not reflect operational conditions. The ROC-AUC of 0.85 is robust to imbalance, whereas the accuracy of 80% is partially inflated by the majority class.

C. Methodological Contributions

Our main methodological contribution is the integration of the data value chain from extraction to deployment. The feature engineering phase was particularly productive. A number of variables were created, such as IsPremiumCustomer and TotalServices, which emerged as important predictors of the final model. This confirms the intuition that the creation of relevant variables, guided by business knowledge, can be more decisive than the choice of algorithm.

The dual-channel deployment approach (interactive application for operational staff, dashboards for decision-makers) addresses a real need of organizations: making predictions available to different stakeholders according to their specific working methods.

D. Limitations and Perspectives

Our research, although having achieved its objectives, presents certain limitations that open up promising research directions.

1) Temporal limitation and longitudinal perspective: The study is based on a cross-sectional dataset, capturing churn at a given point in time. A temporal perspective would allow modeling the evolution of risk and identifying weak signals that precede churn. The use of longitudinal data could refine the understanding of the role of tenure and monthly charges, which occupy a central place in our variable hierarchy.

2) Informational limitation and enrichment perspective: The data used, although rich, are limited to the operator's internal information. The integration of external data like sentiment on social networks, macroeconomic indicators, and competitive data would constitute an important enrichment. Significantly, this addition makes it possible to understand why monthly charges become the primary predictive factor in our model by placing them within their competitive environment.

VI. Conclusion

This work aimed at designing and implementing a churn prediction system adapted to the telecommunications context. Rather than focusing only on modeling, we structured the approach around several successive steps, including exploratory data analysis, the development of an ETL process, feature engineering, and the use of a Random Forest model. The results were then integrated into visualization tools to support decision-making.

The obtained results show that the proposed approach is effective in identifying the main factors related to customer churn. In particular, variables such as monthly charges, contract type, and customer tenure appear to play a significant role. The feature engineering step also brought additional value. For instance, the variables TotalServices and IsPremiumCustomer, which were constructed during the process, turned out to be informative for the model.

One important point that emerges from this study is that predictive performance alone is not sufficient. In practice, the utility of such a model depends on how it is integrated into workflow, from data preparation to the interpretation of results. This perspective makes it possible to better align technical outputs with operational needs, and suggests that similar approaches could be applied in other contexts.

References

[1] F. F. Reichheld and W. E. Sasser, “Zero defections: Quality comes to services,” Harvard Business Review, vol. 68, no. 5, pp. 105–111, 1990.

[2] E. Ascarza, S. A. Neslin, O. Netzer, Z. Anderson, P. S. Fader, S. Gupta, et al., “In pursuit of enhanced customer retention management: Review, key issues, and future directions,” Customer Needs and Solutions, vol. 5, no. 1, pp. 65–81, 2018.

[3] S. Dhariya and H. Bhaidasna, “Ensemble learning approach for customer churn prediction in the telecom industry,” AIP Conference Proceedings, vol. 3288, no. 1, p. 040007, 2025.

[4] M. Imani and H. R. Arabnia, “Customer churn prediction: A systematic review of recent advances, trends, and challenges in machine learning and deep learning,” Machine Learning and Knowledge Extraction, vol. 7, no. 3, p. 105, 2025.

[5] Y. Koulou and N. El Hami, “Machine learning for recommender systems under implicit feedback and class imbalance,” International Journal of Advanced Computer Science and Applications, vol. 16, no. 9, 2025, doi: 10.14569/IJACSA.2025.0160954.

[6] D. Asif, M. S. Arif, and A. Mukheimer, “A data-driven approach with explainable artificial intelligence for customer churn prediction in the telecommunications industry,” Results in Engineering, vol. 26, p. 104629, 2025, doi: 10.1016/j.rineng.2025.104629.

[7] Y. Koulou, N. El Hami, A. El Attaoui, and S. Rhouas, "Application of Interoperability to Intelligent System of Systems," in Methods and Applications of Artificial Intelligence Dynamic Response Learning Random Forest Linear Regression Interoperability Additive Manufacturing and Mechatronics Volume 2. Open source preview, 2025, vol. 2, pp. 193-222.

[8] X. Su and T. M. Khoshgoftaar, "A Survey of Collaborative Filtering Techniques," Advances in Artificial Intelligence, vol. 2009, Art. no. 421425, 2009.

[9] B. Jagadeesh and K. Sashirekha, "Enhancing accuracy of customer churn prediction in e-commerce using random forest algorithm over K-nearest neighbors algorithm," AIP Conference Proceedings, vol. 3270, no. 1, p. 020101, 2025.

[10] M. A. Shaikhsurab and P. Magadum, "Enhancing customer churn prediction in telecommunications: An adaptive ensemble learning approach," arXiv preprint arXiv:2408.16284, 2024, doi: 10.48550/arXiv.2408.16284.

[11] N. Elsayed, S. A. Elaleem, and M. Marie, "Improving prediction accuracy using random forest algorithm," International Journal of Advanced Computer Science and Applications, vol. 15, no. 4, 2024, doi: 10.14569/IJACSA.2024.0150445.

[12] N. A. Alrajeh, B. Elmir, B. Bounabat, and N. E. Hami, "Interoperability optimization in healthcare collaboration networks," Biomedizinische Technik, vol. 57, no. 5, pp. 403-411, 2012.

[13] N. I. Yusof, N. M. M. Zainuddin, N. H. Hassan, N. N. A. Sjarif, S. Yaacob, and W. A. W. Hassan, "A guideline for decision-making on business intelligence and customer relationship management among clinics," International Journal of Advanced Computer Science and Applications, vol. 10, no. 8, 2019, doi: 10.14569/IJACSA.2019.0100865.