

PREDICTIVE MODELS FOR CUSTOMER CHURN IN RIDE-HAILING: A MACHINE LEARNING APPROACH

*** Mohamed Yacine SMAIL**

National Higher School of Statistics and Applied Economics, Algeria

Email : smail.mohamedyacine@learning.enssea.edu.dz

Ayoub ASRI

National Higher School of Statistics and Applied Economics, Algeria

Email : asri.ayoub@enssea.edu.dz

Reçu le: 30/07/2025 **Accepté le:** 09/11/2025 **Publication en ligne le:** 09/12/2025

ABSTRACT:

This study investigates the application of advanced machine learning techniques for customer churn prediction within the ride-hailing industry, leveraging real-world transactional data from a regional service in Algiers. Customer churn is explicitly defined as the absence of rides within a six-month period following the evaluation period. The comprehensive methodology encompasses robust data preprocessing, which benefited from the naturally balanced distribution of the target variable, thereby simplifying the modeling pipeline. A systematic evaluation of multiple classification algorithms, including XGBoost, Random Forest, Logistic Regression, LightGBM, and Decision Tree, was conducted. A key aspect of this approach was the rigorous hyperparameter optimization using Grid Search, Bayesian Optimization, and Simulated Annealing, all implemented in R. The findings demonstrate that LightGBM, particularly when tuned with Simulated Annealing, achieves superior predictive performance (Accuracy: 0.762, Recall: 0.781, F1-Score: 0.779), effectively identifying potential churners. This research provides actionable information for ride-hailing companies to enhance customer retention strategies by focusing on critical behavioral patterns identified through feature importance analysis. The technical rigor, evidenced by the meticulous preprocessing and advanced optimization, culminates in a solution that offers direct business utility by enabling targeted interventions for customer retention.

Keywords : Customer Churn Prediction, Ride Hailing, Machine Learning, Hyperparameter tuning, Lift Analysis.

* Mohamed Yacine SMAIL

I. INTRODUCTION

The ride-hailing sector has undergone a decade of rapid global expansion, fundamentally transforming urban transportation and logistics (Olayode et al., 2023). This dynamic environment is characterized by intense competition, making customer retention a paramount concern for companies like Uber and Lyft, which have faced substantial financial losses due to rider attrition. The economic imperative for effective customer retention strategies is underscored by the widely accepted principle that the cost of acquiring new customers is generally five times higher than retaining existing ones, highlighting its critical importance for sustained business growth and profitability (Maan & Maan, 2023).

In this highly competitive landscape, businesses increasingly rely on data-driven methodologies to forecast future revenues and losses, enabling timely strategic interventions and risk mitigation (Chang et al., 2024). Customer Churn Prediction (CP) has emerged as a vital methodology, allowing organizations to construct predictive models that analyze consumer behaviors and anticipate customer attrition or a decline in sales (Mohaimin et al., 2025). The application of machine learning and data science has become instrumental in formulating methods to understand and mitigate customer attrition, with significant research conducted in various sectors, including telecommunications, retail, banking, healthcare, and insurance (Chang et al., 2024).

The present study addresses the challenge of predicting customer churn within the ride-hailing sector, a domain that, while expanding, often operates with limitations in customer data. Smaller ride-hailing companies, in particular, frequently rely primarily on transactional records from their Enterprise Resource Planning (ERP) systems due to budget constraints that limit investment in extensive market research or access to alternative data sources (Jawad & Balázs, 2024). This study's focus on transactional data directly addresses a common constraint for smaller ride-hailing companies, positioning this research as highly practical and relevant for businesses seeking deployable solutions without extensive demographic or social data. This approach transforms what might be perceived as a data limitation into a strength, demonstrating that significant business value can be extracted even under common data constraints.

This research aims to develop and evaluate robust machine learning models for customer churn prediction in such a context, specifically leveraging transactional data from a regional ride-hailing service and employing advanced optimization techniques. The contribution of this study is multifaceted. It defines a customer retention framework within the ride-hailing industry where research related to customer churn is not as rich as in other industries. Furthermore, it presents a comprehensive methodology that includes detailed feature engineering, rigorous preprocessing, and a comparative analysis of multiple state-of-the-art machine learning models. Furthermore, the study emphasizes the systematic evaluation of hyperparameter tuning strategies, including Grid Search, Bayesian Optimization, and Simulated Annealing, to achieve optimal model performance. The entire

analytical workflow is implemented in R, ensuring a structured, consistent, and reproducible approach to predictive modeling.

II. LITTERATURE REVIEW

The field of customer churn prediction has undergone a significant evolution, transitioning from traditional statistical methods to sophisticated Machine Learning (ML) and Deep Learning (DL) approaches (Imani, Joudaki, et al., 2025). This shift has been driven by the ability of ML and DL algorithms to effectively analyze high-dimensional, dynamic, and complex customer datasets, often outperforming conventional statistical techniques (Imani, Joudaki, et al., 2025). Recent advancements in ML and DL for churn prediction, particularly between 2020 and 2025, have been extensively documented across diverse domains such as telecommunications, retail, banking, healthcare, and insurance (Imani, Joudaki, et al., 2025). Reviews of peer-reviewed studies highlight a marked increase in publications, with a notable peak in 2023, underscoring the growing research interest in this area. These studies consistently demonstrate the power of data-driven approaches to enhance customer retention strategies (Manzoor et al., 2024).

A significant trend observed is the emergence of ensemble deep learning models, which have shown superior performance in terms of accuracy and F1-score across various sectors compared to single models (AbdelAziz et al., 2025). Advanced deep learning architectures, such as Convolutional Neural Networks (CNNs) for sequential data, Recurrent Neural Networks (RNNs), and transformer-based models, are also gaining traction in churn prediction due to their ability to capture complex temporal patterns in customer behavior (Shihua et al., 2025a). Concurrently, there is a growing emphasis on the incorporation of profit-driven modeling to align predictive efforts directly with business profitability (Manzoor et al., 2024).

While the telecommunications sector has historically dominated churn research, the ride-hailing industry presents unique challenges and opportunities for churn prediction (Tebu & Izang, 2025). Its dynamic nature, characterized by real-time transactions and evolving customer behaviors, necessitates robust predictive models. The reliance on transactional data, as opposed to richer demographic or social data often available in other sectors, poses a specific constraint that this study seeks to address. Existing literature often draws parallels from related service industries, but specific research on ride-hailing churn, particularly utilizing real-world transactional data and advanced optimization techniques, remains an area for further exploration (Tebu & Izang, 2025).

The effectiveness of churn prediction models also varies significantly based on the choice of methodology, dataset characteristics, and industry-specific factors, highlighting the need for careful model selection and tuning (Imani, Joudaki, et al., 2025).

III. DATA AND METHODOLOGY

This section provides a detailed analysis of the dataset utilized, the comprehensive preprocessing steps undertaken, the machine learning models employed, and the framework

for their evaluation. The entire machine learning workflow was systematically implemented in R.

III.1. DATA COLLECTION AND CHARACTERISTICS

The dataset central to this study comprises transactional records obtained from the Yassir ride-hailing application, which operates within Algiers city. Data collection was facilitated directly through an SDK integrated into the ride-hailing application, enabling real-time tracking of both rider and driver activities. Key metrics, including trip details, pickup and drop-off locations, timestamps, estimated costs, and various user interactions, were meticulously recorded. This raw data was then transmitted to Google Cloud Platform (GCP) for secure storage and efficient processing, with BigQuery employed for structuring and querying to extract actionable insights.

For the purpose of this study, customer churn was explicitly defined as the absence of any rides within a six-month period following the evaluation period. This definition was inspired by numerous literature reviews, such as the one described by (Zheng, 2025). This clear definition is fundamental for framing the binary classification problem and ensuring consistency in churn identification. The dataset encompasses a rich set of variables, each capturing different facets of rider behavior and trip characteristics. These variables are detailed in (Table 1).

Table 01 : Variable Definitions

Variable	Description
number_finished_trips_by_rider	The total number of trips successfully completed by a rider.
avg_finished_trips_monthly	The average number of trips completed by the rider per month, calculated over the available period.
number_of_DCC_DCRC_and_RC_trips	The count of trips categorized as DCC (Driver Coming Cancelled), DCRC (Driver Coming Rider Cancelled) and RC (Rider Cancelled).
passivity_days	The number of days the rider was inactive or did not perform any trips.
commune_of_pickup	The name of the commune (local administrative unit) where the rider picks up most of their trips.
wilaya_of_destination	The destination wilaya (province) if the trips extend outside of Algiers.
average_estimated_cost_eta_distance	The average values for trip cost (estimated), estimated time of arrival (ETA), and distance covered for completed trips.
number_of_used_coupons	The total number of discount coupons used by the rider during their trips.
number_of_trips	The total number of trips initiated, including completed and non-completed trips.
does_the_rider_travels_mostly_on_weekend_or_weekdays	An indication of whether the rider predominantly performs trips during weekends or weekdays.
success_rate_of_trips	The ratio of completed trips to the total number of trips initiated, expressed as a percentage.
average_value_of_discount	The average monetary value of the discounts applied to the rider's trips.
maximum_minimum_mean_and_median_time_difference_between_trips	The statistical measures (maximum, minimum, mean, and median) for the time intervals between consecutive trips performed by the rider.

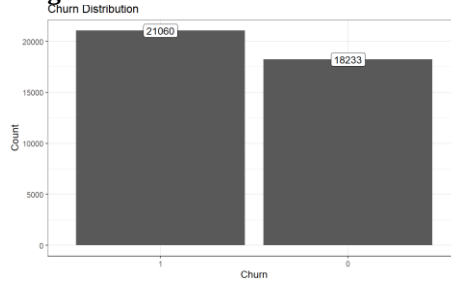
Source : Authors' own work

III. 2. EXPLORATORY DATA ANALYSIS (EDA)

Exploratory Data Analysis (EDA) was conducted to summarize key characteristics of the dataset, identify underlying patterns, and uncover initial relationships, often leveraging

visualization techniques (Deming et al., 2018). A critical observation from the EDA was the balanced distribution of the target variable, 'churn'. Specifically, the dataset exhibited nearly equivalent representation between churners and non-churners, with churners (churn = 1) comprising 53.6% of observations and non-churners (churn = 0) accounting for the complementary 46.4%. For instance, the raw data showed a marginal predominance of churners, confirming a fair class balance. This equitable distribution is visually confirmed in (Figure 1).

Figure 01 : Churn Class Distribution



Source : Authors' own work

Univariate analysis provided summary statistics for each numerical feature, offering a quick overview of their distributions, central tendencies, spread, and the presence of missing values. These statistics are presented in (Table 2).

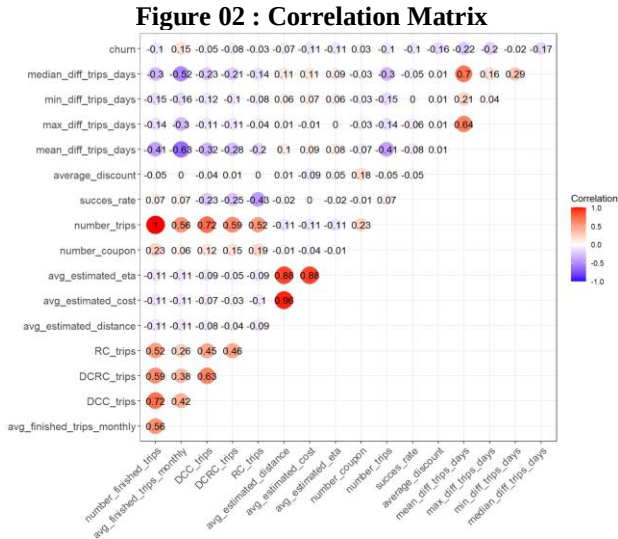
This visual and quantitative confirmation of imbalance highlighted the necessity for specific handling during the preprocessing phase to prevent model bias towards the majority class. Univariate analysis provided summary statistics for each numerical feature, offering a quick overview of their distributions, central tendencies, spread, and the presence of missing values. These statistics are presented in (Table 2).

Table 02 : Univariate Statistics

	Min	1st Q	Median	Mean	3rd Q	Max	NAs
number_finished_trips	20	25	36	52.55	58	807	
avg_finished_trips_monthly	1.111	3.092	4.4	5.685	6.818	70.6	
DCC_trips	0	3	6	7.941	10	213	34
DCRC_trips	0	2	4	6.998	9	500	34
RC_trips	0	4	9	15.55	19	972	34
first_trip_month	1	4	7	6.632	10	12	1
last_trip_month	1	2	2	4.558	8	12	1
passivity_days	0	74.42	102.24	175.79	225.24	922.26	1
lifetime_days	0	217	379	382.6	531.2	869.8	1
avg_estimated_distance	970.8	8355.3	10912.9	11749.2	14107.8	102112.8	47
avg_estimated_cost	16.76	440.71	519.16	540.99	614.46	3229.17	47
avg_estimated_eta	224.7	1011	1214.1	1249.7	1445.5	5639.4	47
number_coupon	0	0	0	0.4572	0	36	
number_trips	20	25	36	52.55	58	807	
success_rate	0.01944	0.54098	0.63636	0.63038	0.72549	1	
average_discount	0	0	0.001538	0.010829	0.010132	0.957381	
mean_diff_trips_days	0.2236	4.6855	8.4791	10.2746	14.2405	80.7233	2
max_diff_trips_days	0.7734	40.3641	74.3481	95.1919	126.8967	696.9336	2
min_diff_trips_days	0.000005	0.019389	0.035437	0.080659	0.068154	7.947334	2
median_diff_trips_days	0.00045	1.21687	2.59933	3.65683	4.90739	51.98191	2

Source : Authors' own work

Multivariate analysis, typically visualized through a correlation matrix heatmap presented in (Figure 02), provided insights into the relationships between features. The correlation matrix reveals several key observations about customer behavior and churn drivers.



Source : Authors' own work

First, trip frequency and consistency emerge as critical factors influencing retention. The strong negative correlations between churn and metrics like number_trips, RC_trips, and DCRC_trips suggest that customers who take fewer trips are more likely to leave. Additionally, irregular trip schedules—captured by mean_diff_trips_days and median_diff_trips_days—show significant negative relationships with trip volume, indicating that habitual users are more likely to stay engaged. This pattern indicates that customer engagement (frequent, regular trips) and a positive service experience (implied by success_rate) are primary drivers of retention.

Interestingly, while discounts (average_discount) have a slight positive link to success_rate, their overall impact appears minimal, implying that blanket discounts may not be as effective as targeted incentives in driving long-term loyalty. Another notable finding is the disconnect between operational metrics and churn.

Although avg_estimated_cost and avg_estimated_distance are nearly perfectly correlated (0.99), their weak ties to churn suggest that pricing and distance are secondary concerns for customers. Instead, service reliability—reflected in success_rate—may play a more pivotal role. Meanwhile, the strong negative correlation between

DCC_trips (Driver Coming Cancelled trips) and avg_finished_trips_monthly (-0.72) hints that over-reliance on promotions could undermine sustainable engagement. This observation challenges a common business assumption that price reductions or blanket discounts are the most effective ways to retain customers. Instead, it shifts the strategic focus

towards operational excellence, fostering habitual usage through loyalty programs, and ensuring a seamless user experience, which can lead to more sustainable customer relationships and a higher return on investment than purely price-driven strategies.

In summary, the data underscores the importance of habitual usage and service quality in reducing churn. Companies may benefit from incentivizing regular trips (e.g., through subscription models or loyalty rewards) while ensuring operational reliability. Discounts, while useful in specific contexts, should be deployed strategically rather than as a one-size-fits-all solution. By focusing on these behavioral and experiential levers, businesses can build stronger, more enduring customer relationships.

III. 3. FEATURE ENGINEERING

Feature engineering involved the creation of new, more informative features from the raw transactional data. Examples include `number_finished_trips_by_rider`, `avg_finished_trips_monthly`, `passivity_days`, and various time difference metrics between trips (Khedekar et al., 2025).

The inclusion of features such as `passivity_days` and `time_difference_between_trips` is particularly noteworthy, as these directly capture the disengagement aspect of customer behavior. In service industries, a change in customer behavior, specifically a cessation of service usage, is a primary indicator of churn. Longer periods of inactivity or larger time differences between consecutive trips serve as direct signals of disengagement, which are often more predictive than static demographic data. This approach moves beyond simple data points to extract their underlying meaning in the context of churn prediction.

Additionally, a filtering step was applied to exclude riders with `passivity_days` less than 350 days (Yap & Cats, 2025). This step aimed to focus the analysis on a cohort of users with a sufficient observation history, and to exclude newly acquired users who have not yet had the opportunity to exhibit churn behavior, thereby refining the active customer base for more accurate churn analysis. This methodological choice implicitly defines the active customer base for churn analysis, influencing the generalizability of the model's findings to different customer segments. While it improves accuracy for established customers, it might limit insights into very early-stage churn.

III. 4. DATA PREPROCESSING

The raw data underwent several critical preprocessing steps to ensure its quality, handle inconsistencies, and optimize its suitability for machine learning model training. These steps are fundamental for achieving reliable and accurate predictive performance.

First, variables exhibiting near-zero variance were systematically removed from the dataset. Variables with very little or no variation across observations provide minimal to no discriminatory power for a predictive model, as they are almost constant. Their removal contributes to simpler and faster models, and it also prevents potential errors or numerical instability that can arise in certain machine learning algorithms, particularly those sensitive to division by zero (e.g., normalization methods) or non-invertible matrices (e.g., PCA).

Second, missing values within the dataset were imputed. Numerical features with missing values were imputed with their respective medians, while missing categorical values were imputed with their modes. The statistical rationale behind imputation is to create a dataset that best approximates the "true" data as if no information were lost (Wu et al., 2023). Median imputation is a robust choice for numerical features, especially when data distributions are skewed or contain outliers, as it is less sensitive to extreme values compared to mean imputation (Wu et al., 2023). For categorical variables, mode imputation is a straightforward approach, though careful consideration is advised if the distribution of categories is highly uneven, as it could potentially introduce bias.

Finally, all categorical features were transformed into numerical dummy (one-hot encoded) variables. This conversion is a standard practice because most machine learning algorithms are designed to operate on numerical input. Dummy encoding effectively allows categorical information to be incorporated into the models without implying an ordinal relationship where none exists.

III. 5. MACHINE LEARNING MODELS

A diverse set of machine learning classification algorithms was employed to predict customer churn, ranging from simpler linear models to complex ensemble methods. The entire modeling process, encompassing model specification, fitting, and evaluation, was meticulously implemented in R. The deliberate selection of a diverse set of models, spanning different algorithmic paradigms, ensures a comprehensive and robust evaluation. This approach provides deeper understanding into which types of models are inherently better suited for the specific churn prediction problem in the ride-hailing industry, rather than just identifying a single best performer.

The following models were initially evaluated in their default configurations to establish a baseline performance benchmark:

- **XGBoost (eXtreme Gradient Boosting)** (Chen & Guestrin, 2016).
- **Random Forest** (Breiman, 2001).
- **Logistic Regression.**
- **LightGBM (Light Gradient Boosting Machine)** (Ke et al., 2017).
- **Decision Tree.**

This comprehensive evaluation across linear (Logistic Regression), single tree (Decision Tree), bagging ensemble (Random Forest), and boosting ensembles (XGBoost, LightGBM) allowed for a robust assessment of how different algorithmic characteristics perform with ride-hailing transactional data. The aim was to uncover not just the best-performing algorithm, but also to gain deeper understanding into why certain model types, particularly gradient boosting ensembles, are inherently well-suited for the complexities of churn prediction in this domain. This approach provided valuable insights into the underlying structure of the data and the most effective strategies for modeling customer behavior in the ride-hailing industry.

III. 6. HYPERPARAMETER TUNING STRATEGIES

The raw data underwent several critical preprocessing steps to ensure its quality, handle inconsistencies, and optimize its suitability for machine learning model training. These steps are fundamental for achieving reliable and accurate predictive performance.

Hyperparameter tuning is a critical phase in the machine learning workflow, as the selection of optimal hyperparameter values significantly influences a model's efficacy, its capacity to generalize to unseen data, and its ability to avoid overfitting. The hyperparameters are crucial parameters that govern the learning process of a model and are typically established prior to the commencement of training. While often initialized with default values, their refinement through experimentation is essential to optimize the model's learning efficacy and overall performance. This study systematically compared three advanced tuning techniques applied to LightGBM, which emerged as the best-performing baseline model.

The following hyperparameter tuning techniques were employed:

- **Grid Search** (Liashchynskiy & Liashchynskiy, 2019).
 - **Bayesian Hyperparameter Tuning** (Nguyen, 2019) (Shihua et al., 2025b) (Wang et al., 2023).
- **Simulated Annealing (SA)** (“The Theory and Practice of Simulated Annealing,” 2019).

The comparative evaluation of Grid Search, Bayesian Optimization, and Simulated Annealing provides valuable insights into the efficiency and effectiveness trade-offs of different optimization paradigms. While Bayesian Optimization is often praised for its efficiency in black-box optimization, Simulated Annealing, despite potentially being more computationally intensive, can sometimes achieve marginally better results. This is attributable to its inherent ability to escape local optima more effectively due to its stochastic exploration, which allows it to jump out of suboptimal regions that a more guided search might get stuck in.

III. 7. MODEL TRAINING AND EVALUATION METRICS

The dataset was systematically partitioned into two distinct subsets: training (70% of total data) and testing (30%). The training set served as the foundation for model development, enabling algorithms to learn underlying patterns, while hyperparameter tuning was conducted internally through 5-fold cross-validation—a rigorous process that cyclically partitions the training data to optimize model configurations and prevent overfitting without requiring a separate validation set. The testing subset remained strictly isolated from all training and tuning activities, providing an independent benchmark to evaluate the model's generalizability and deliver an unbiased performance estimate on unseen data.

For evaluating the performance of the classification models, especially given the imbalanced nature of the churn dataset, the selection of appropriate metrics was paramount. Simple accuracy can be misleading in such scenarios, as a model might achieve high accuracy by merely predicting the majority class. Therefore, a confusion matrix was employed:

Figure 03 : Tuning the neighbors hyperparameter

		Predicted class		
		Classified positive	Classified negative	
Actual class	Actual positive	TP	FN	TPR: $\frac{TP}{TP + FN}$
	Actual negative	FP	TN	FPR: $\frac{TN}{TN + FP}$
		Precision: $\frac{TP}{TP + FP}$	Accuracy: $\frac{TP + TN}{TP + TN + FP + FN}$	

Source : (Richter-Laskowska et al., 2022)

Based on this matrix, the following metrics were calculated:

- **Accuracy:** The proportion of correctly classified instances over both positive and negative classes, calculated as $\frac{TP+TN}{total}$. While reported, its limitations for imbalanced data were acknowledged (Thölke et al., 2023).
- **Recall (Sensitivity or True Positive Rate):** Defined as $\frac{TP}{TP + FN}$, this metric measures the ability of the model to identify all actual positive cases (churners). In the context of churn prediction, recall is particularly critical, as failing to identify a customer who will churn (a False Negative) can result in a lost customer and revenue that could have been saved, often at a higher cost than incorrectly predicting churn for a non-churner (a False Positive).
- **Precision:** Calculated as $\frac{TP}{TP + FP}$, precision represents the proportion of predicted positives that are, in fact, truly positive.
- **F1-Score:** This metric represents the harmonic mean of precision and recall and is defined as $2 \times \frac{Precision \times Recall}{Precision + Recall}$. The F1-score provides a balanced measure of a model's performance, making it especially valuable for evaluating models on imbalanced datasets where both false positives and false negatives are important considerations (Imani, Beikmohammadi, et al., 2025).
- **Area Under the Receiver Operating Characteristic (ROC) Curve (AUC):** The ROC curve graphically depicts the trade-off between the false positive rate (plotted on the horizontal axis) and the true positive rate (recall, plotted on the vertical axis) across various classification thresholds. The Area Under the Curve (AUC) provides a single scalar measure of overall discriminatory power, where a higher AUC value (closer to 1) indicates better performance in distinguishing between positive and negative classes.

IV. RESULTS

This section presents the empirical findings from the evaluation of various machine learning models, both in their default configurations and following hyperparameter optimization, along with insights derived from feature importance analysis.

IV.1. BASELINE MODEL PERFORMANCE

The initial evaluation involved assessing the performance of several machine learning models in their default configurations to establish a baseline for comparison. Table 3 provides a comparative analysis of these models across key predictive metrics: accuracy, recall, precision, and F1-score.

Table 03 : Default Models' Performance

Model Name	Accuracy	Recall	Precision	F1 Score	ROC AUC
XGBoost	0.754432	0.8077227	0.7523585	0.779058	0.84557
Random Forest	0.728561	0.8064567	0.7204864	0.761051	0.812295
Logistic Regression	0.638392	0.7520177	0.6380236	0.690346	0.689176
LightGBM	0.783951	0.7797120	0.8100954	0.794613	0.866686
Decision Tree	0.713547	0.7583478	0.7214694	0.739449	0.780992

Source : Authors' own work

As indicated in (Table 3), LightGBM emerged as the most robust baseline model, achieving the highest accuracy (0.7839), recall (0.7797), and F1-score (0.7946). This performance highlights its inherent capacity to accurately predict churners while maintaining a strong balance between precision and recall. Random Forest and XGBoost also demonstrated competitive performance, exhibiting strong predictive power with F1-scores of 0.7610 and 0.7790, respectively. Logistic Regression, despite its simplicity, achieved the worst performance overall, suggesting its struggles in identifying all actual churners. The Decision Tree model also performed the weakest among the evaluated algorithms, with the lowest accuracy (0.7135) and F1-score (0.7394), likely due to its susceptibility to overfitting or its inability to capture complex patterns in the data without further optimization.

LightGBM's superior overall performance is particularly significant for churn prediction in a business context. Minimizing False Negatives—that is, failing to identify a customer who will ultimately churn—is often more critical than minimizing False Positives (incorrectly identifying a non-churner as a churner). A high recall value ensures that the company can proactively target a larger proportion of actual at-risk customers, even if it means some non-churners are also included in retention campaigns. This directly enhances the effectiveness and potential return on investment of customer retention efforts.

IV. 2. IMPACT OF HYPERPARAMETER TUNING

Following the baseline evaluation, hyperparameter tuning was performed on the LightGBM model, which demonstrated the most promising initial performance. The objective was to further optimize its predictive capabilities. Three distinct hyperparameter

tuning techniques—Grid Search, Bayesian Search, and Simulated Annealing—were applied and their results are presented in (Table 4).

The results indicate that hyperparameter tuning significantly enhanced the performance of the LightGBM model across all metrics. Grid Search, while providing reliable performance, yielded an accuracy of 0.7544 and a recall of 0.78. However, it was outperformed by both Bayesian Search and Simulated Annealing, indicating its limitations in efficiently exploring complex parameter spaces.

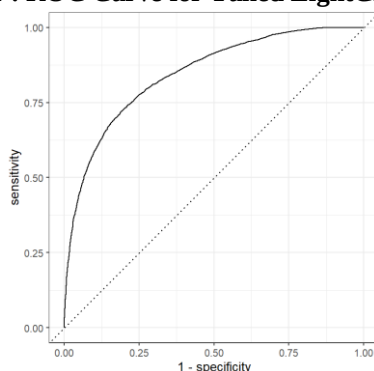
Table N° 4 : Tuned LightGBM Performance

Tuning Technique	Accuracy	Recall	Precision	F1 Score	ROC AUC
Grid Search	0.754432	0.7800285	0.7539003	0.779058	0.84557
Bayesian Search	0.761388	0.7896819	0.7707754	0.780114	0.844237
Simulated Annealing	0.762406	0.7812945	0.7767464	0.779014	0.844987

Source : Authors’ own work

Bayesian Search demonstrated superior performance over Grid Search, achieving an accuracy of 0.7613, a recall of 0.7896, and an F1-score of 0.7707. This improvement can be attributed to its probabilistic approach, which intelligently models the objective function and focuses on more promising regions of the hyperparameter space, leading to greater efficiency without compromising performance (Shihua et al., 2025b).

Figure 04 : ROC Curve for Tuned LightGBM Model



Source : Authors’ own work

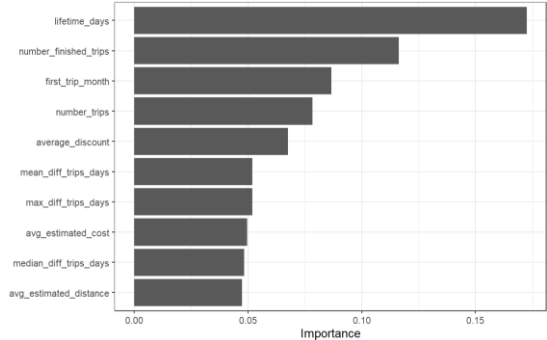
Simulated Annealing emerged as the most effective tuning technique, achieving the highest accuracy (0.7624), recall (0.7812), and F1-score (0.779) for the LightGBM model. Its stochastic nature allows for a broader exploration of the parameter space and a greater ability to escape local optima by occasionally accepting less optimal solutions, particularly in the initial stages of the search (“The Theory and Practice of Simulated Annealing,” 2019). This leads to marginally better results compared to Bayesian Search, as it can navigate complex, multi-modal hyperparameter landscapes more effectively.

The overall discriminatory power of the best-performing model, LightGBM tuned with Simulated Annealing, is further illustrated by its Receiver Operating Characteristic (ROC) curve, presented in (Figure 4) ROC Curve for Tuned LightGBM Model. This curve visually depicts the model's ability to distinguish between churners and non-churners across various classification thresholds, providing a comprehensive understanding of its performance trade-offs.

IV. 3. IMPACT OF HYPERPARAMETER TUNING

Understanding which features contribute most to the model's predictions is crucial for deriving actionable strategies. A Variable Importance Plot (Figure 5), typically a bar chart, would visually rank features by their importance scores (e.g., gain, split count) as determined by the trained model (LightGBM in our case). This plot would clearly identify which aspects of rider behavior or trip characteristics are most influential in predicting churn.

Figure N° 5 : Variable Importance Plot



Source : Authors' own work

This plot illustrates the feature importance for a churn prediction model, highlighting the most influential variables. Temporal activity features, such as `lifetime_days` and `first_trip_month`, are highly influential. This suggests that recency of engagement and tenure duration play significant roles in customer retention. Similarly, trip completion metrics like `number_finished_trips` and `number_trips` demonstrate critical importance. This indicates that consistent service usage and successful transaction volume are strong indicators of engagement. Furthermore, trip interval patterns reflected in `mean_diff_trips_days`, `max_diff_trips_days`, and `median_diff_trips_days` show considerable predictive power. Longer gaps between trips or irregular usage patterns can signal declining engagement and increased churn likelihood. These findings provide a direct link between the model's predictive power and practical business understanding.

IV. 4. LIFT ANALYSIS

While traditional metrics such as Accuracy, Recall, and F1-Score assess the overall statistical performance of a churn prediction model, business stakeholders often require

metrics that quantify the model's ability to concentrate churners within a targeted group. Lift table analysis or decile analysis is an alternative way to assess the model performance especially in a marketing context to gauge the efficiency of marketing campaign. The customers are scored using the obtained predictive model then they are sorted decreasingly by these scores and segmented into equal deciles (or percentiles depending on the positive event distribution). The company can use this table to design its customer retention campaigns (Mena et al., 2019).

Table N° 5 : Lift Analysis results

decile	base	churners	Cumulative churners	Min prob	Max prob	base %	churners % (recall)	lift	cumulative churners	precision	Cumulative precision
1	1,179	1,127	1,127	0.946	1	10%	18%	1.78	18%	96%	96%
2	1,179	1,064	2,191	0.866	0.946	10%	17%	1.68	35%	90%	93%
3	1,179	966	3,157	0.767	0.866	10%	15%	1.53	50%	82%	89%
4	1,179	854	4,011	0.656	0.767	10%	14%	1.35	63%	72%	85%
5	1,179	703	4,714	0.541	0.656	10%	11%	1.11	75%	60%	80%
6	1,179	542	5,256	0.436	0.541	10%	9%	0.86	83%	46%	74%
7	1,179	429	5,685	0.335	0.436	10%	7%	0.68	90%	36%	69%
8	1,179	347	6,032	0.224	0.335	10%	5%	0.55	95%	29%	64%
9	1,179	230	6,262	0.084	0.223	10%	4%	0.36	99%	20%	59%
10	1,178	57	6,319	0	0.084	10%	1%	0.09	100%	5%	54%

Source : Authors' own work

The lift analysis, presented in Table 5, demonstrates the model's effectiveness in prioritizing high-risk churners. The top decile exhibits exceptional performance, capturing 18% of all churners (recall) with a precision of 96% and lift factor of 1.78 (maximum lift that can be achieved here is 1.86). This indicates that targeting the top 10% of users ranked by the model would yield 1.78 times more churners than random selection. The model maintains strong predictive capability through subsequent deciles, with the top three deciles collectively identifying 50% of all churners at 89% cumulative precision. Performance gradually declines in lower deciles, as expected, with the seventh decile marking the threshold where recall falls below the expected baseline proportion (7% vs. 10%).

V. DISCUSSIONS AND MANAGARIAL IMPLICATIONS

The current study successfully developed and evaluated robust machine learning models for customer churn prediction within a ride-hailing service, leveraging real-world transactional data and an R-based machine learning approach. The comprehensive methodology, which included meticulous data preprocessing and advanced hyperparameter optimization techniques, culminated in a highly effective LightGBM model. The superior performance of the tuned LightGBM model, particularly its high recall, underscores its suitability for this complex, imbalanced classification task and offers significant practical value for identifying at-risk customers.

The identified important features, such as `passivity_days`, `number_finished_trips_by_rider`, and `average_discount`, align with intuitive understandings

of customer engagement and loyalty in the ride-hailing sector. For instance, prolonged periods of inactivity or a decrease in the number of completed trips are strong indicators of disengagement, suggesting that customers are gradually reducing or ceasing their use of the service. Conversely, features related to discount usage highlight the role of incentives in fostering retention.

These findings translate directly into actionable strategies for ride-hailing companies. Based on the prominence of `passivity_days` and `number_finished_trips`, companies can implement proactive engagement campaigns targeting riders who show early signs of disengagement, such as increasing inactivity or declining trip frequency. Personalized offers, reminders, or push notifications could be deployed to re-engage these at-risk segments. Insights from `number_coupon` and `average_discount` suggest that well-designed loyalty programs or targeted re-engagement campaigns leveraging discounts could be effective in retaining customers. While not explicitly detailed in the feature set, the `success_rate_of_trips` implies a connection to service quality; improving trip completion rates and overall rider experience could further enhance retention. The predictive model's ability to identify potential churners allows companies to allocate their retention resources more efficiently, focusing efforts on high-value, at-risk customers, thereby maximizing the return on their marketing and operational investments (Manzoor et al., 2024).

Comparing the achieved model performance (accuracy, recall, F1-score) with benchmarks in recent churn prediction studies, especially those in service industries, reveals that the results are competitive and align with broader trends. The study's findings support the general observation that ensemble methods, such as LightGBM, and advanced deep learning models often outperform simpler, single solutions in complex classification tasks (AbdelAziz et al., 2025). The incremental performance gains achieved through advanced hyperparameter tuning techniques like Bayesian Optimization and Simulated Annealing further validate the necessity of sophisticated optimization strategies to extract maximum predictive power from machine learning models, even for those already demonstrating strong baseline performance.

Despite the robust methodology and promising results, this study has certain limitations. Primarily, the analysis relied predominantly on transactional data. While this approach addresses a common data constraint for smaller ride-hailing companies, integrating additional feature sets, such as demographic information, direct customer feedback, or social interaction graphs (as recommended in recent literature), could potentially enhance model accuracy and provide a richer understanding of the underlying drivers of churn. Furthermore, while the models demonstrated high predictive performance, the evaluation did not explicitly incorporate profit-based metrics, such as the Maximum Profit Criterion (MPC) or Expected Maximum Profit (EMP). These metrics are increasingly advocated in churn prediction research to align model optimization directly with business profitability and the economic impact of retention efforts.(Manzoor et al., 2024)

Directions for future research include exploring the integration of additional data sources (e.g., demographic, social network data, customer service interactions, and direct feedback) to construct more holistic churn prediction models. Investigating the application of advanced deep learning architectures, such as Recurrent Neural Networks (RNNs) for

sequential transactional data or transformer-based models, which are gaining traction in churn prediction, could also yield further improvements (Lachi et al., 2025). Implementing and evaluating profit-driven evaluation metrics would allow for direct optimization of models for business value. Conducting controlled A/B testing of different retention strategies guided by model predictions could quantify their real-world impact and refine intervention effectiveness. Finally, considering the use of automated machine learning (AutoML) frameworks for end-to-end pipeline optimization, including automated feature engineering and model selection, could streamline the development and deployment of future churn prediction systems (Seo et al., 2025).

VI. CONCLUSION

This study successfully developed and evaluated robust machine learning models for customer churn prediction in a ride-hailing service, utilizing real-world transactional data and an R-based machine learning approach. The comprehensive methodology, which included meticulous data preprocessing and advanced hyperparameter optimization techniques (Grid Search, Bayesian Optimization, and Simulated Annealing), yielded a highly effective LightGBM model. The superior performance of the tuned LightGBM model, particularly its high recall, offers significant practical value for identifying at-risk customers within the ride-hailing industry. The insights derived from feature importance analysis provide actionable information for the ride-hailing company to design targeted retention campaigns, ultimately contributing to sustained customer loyalty and business profitability. This research underscores the potential of advanced machine learning to transform customer retention strategies, even when facing data limitations, by focusing on behavioral indicators and optimizing model performance through sophisticated techniques.

VII. BIBLIOGRAPHIE :

AbdelAziz, N. M., Bekheet, M., Salah, A., El-Saber, N., & AbdelMoneim, W. T. (2025). A Comprehensive Evaluation of Machine Learning and Deep Learning Models for Churn Prediction. *Information*, 16(7), 537. <https://doi.org/10.3390/info16070537>

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/a:1010933404324>

Chang, V., Hall, K., Xu, Q., Amao, F., Ganatra, M., & Benson, V. (2024). Prediction of Customer Churn Behavior in the Telecommunication Industry Using Machine Learning Models. *Algorithms*, 17(6), 231. <https://doi.org/10.3390/a17060231>

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://doi.org/10.1145/2939672.2939785>

Deming, C., Dekkati, S., & Desamsetti, H. (2018). Exploratory Data Analysis and Visualization for Business Analytics. *Asian Journal of Applied Science and Engineering*, 7(1), 93–100. <https://doi.org/10.18034/ajase.v7i1.53>

Imani, M., Beikmohammadi, A., & Arabnia, H. R. (2025). Comprehensive Analysis of Random Forest and XGBoost Performance with SMOTE, ADASYN, and GNUS Under Varying Imbalance Levels. *Technologies*, 13(3), 88. <https://doi.org/10.3390/technologies13030088>

Imani, M., Joudaki, M., Beikmohamadi, A., & Arabnia, H. R. (2025). *Customer Churn Prediction: A Review of Recent Advances, Trends, and Challenges in Conventional Machine Learning and Deep Learning*. Computer Science and Mathematics. <https://doi.org/10.20944/preprints202503.1969.v1>

Jawad, Z. N., & Balázs, V. (2024). Machine learning-driven optimization of enterprise resource planning (ERP) systems: A comprehensive review. *Beni-Suef University Journal of Basic and Applied Sciences*, 13(1), 4. <https://doi.org/10.1186/s43088-023-00460-y>

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

Khedekar, L. S., Chhajed, A. S., Bhaganagre, R. C., Kokate, N. S., & Patil, S. (2025). Uber Rides Prediction using Machine Learning. *2025 International Conference on Electronics and Renewable Systems (ICEARS)*, 1731–1736. <https://doi.org/10.1109/icears64219.2025.10940984>

Lachi, V., Longa, A., Bevilacqua, B., Lepri, B., Passerini, A., & Ribeiro, B. (2025). *Boosting Relational Deep Learning with Pretrained Tabular Models* (Version 1). arXiv. <https://doi.org/10.48550/ARXIV.2504.04934>

Liashchynskyi, P., & Liashchynskyi, P. (2019). *Grid Search, Random Search, Genetic Algorithm: A Big Comparison for NAS* (Version 1). arXiv. <https://doi.org/10.48550/ARXIV.1912.06059>

Maan, J., & Maan, H. (2023). *Customer Churn Prediction Model using Explainable Machine Learning* (No. arXiv:2303.00960). arXiv. <https://doi.org/10.48550/arXiv.2303.00960>

Manzoor, A., Atif Qureshi, M., Kidney, E., & Longo, L. (2024). A Review on Machine Learning Methods for Customer Churn Prediction and Recommendations for

Business Practitioners. *IEEE Access*, 12, 70434–70463. <https://doi.org/10.1109/ACCESS.2024.3402092>

Mena, C. G., Caigny, A. D., Coussement, K., Bock, K. W. D., & Lessmann, S. (2019). *Churn Prediction with Sequential Data and Deep Neural Networks. A Comparative Analysis* (No. arXiv:1909.11114). arXiv. <https://doi.org/10.48550/arXiv.1909.11114>

Mohaimin, M. R., Das, B. C., Akter, R., Anonna, F. R., Hasanuzzaman, M., Chowdhury, B. R., & Alam, S. (2025). Predictive Analytics for Telecom Customer Churn: Enhancing Retention Strategies in the US Market. *Journal of Computer Science and Technology Studies*, 7(1), 30–45. <https://doi.org/10.32996/jcsts.2025.7.1.3>

Nguyen, V. (2019). Bayesian Optimization for Accelerating Hyper-Parameter Tuning. *2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, 302–305. <https://doi.org/10.1109/aike.2019.00060>

Olayode, I. O., Severino, A., Justice Alex, F., Macioszek, E., & Tartibu, L. K. (2023). Systematic review on the evaluation of the effects of ride-hailing services on public road transportation. *Transportation Research Interdisciplinary Perspectives*, 22, 100943. <https://doi.org/10.1016/j.trip.2023.100943>

Richter-Laskowska, M., Kurpas, M., & Maśka, M. (2022). *A learning by confusion approach to characterize phase transitions*. <https://doi.org/10.48550/ARXIV.2206.15114>

Seo, W., Lee, J., & Bu, Y. (2025). *SPIO: Ensemble and Selective Strategies via LLM-Based Multi-Agent Planning in Automated Data Science* (Version 1). arXiv. <https://doi.org/10.48550/ARXIV.2503.23314>

Shihua, J. M., Saves, P., Liem, R. P., & Morlier, J. (2025a). *Bayesian Optimization of a Lightweight and Accurate Neural Network for Aerodynamic Performance Prediction* (No. arXiv:2503.19479). arXiv. <https://doi.org/10.48550/arXiv.2503.19479>

Shihua, J. M., Saves, P., Liem, R. P., & Morlier, J. (2025b). *Bayesian Optimization of a Lightweight and Accurate Neural Network for Aerodynamic Performance Prediction* (Version 1). arXiv. <https://doi.org/10.48550/ARXIV.2503.19479>

Tebu, G., & Izang, A. (2025). Customer Churn Prediction in the Telecommunication Industry Over the Last Decade: A Systematic Review. *Asian Journal of Research in Computer Science*, 18(4), 256–271. <https://doi.org/10.9734/ajrcos/2025/v18i4618>

The Theory and Practice of Simulated Annealing. (2019). In D. Henderson, S. H. Jacobson, & A. W. Johnson, *International Series in Operations Research & Management Science* (pp. 287–319). Kluwer Academic Publishers. https://doi.org/10.1007/0-306-48056-5_10

Thölke, P., Mantilla-Ramos, Y.-J., Abdelhedi, H., Maschke, C., Dehgan, A., Harel, Y., Kemtur, A., Mekki Berrada, L., Sahraoui, M., Young, T., Bellemare Pépin, A., El Khantour, C., Landry, M., Pascarella, A., Hadid, V., Combrisson, E., O'Byrne, J., & Jerbi, K. (2023). Class imbalance should not throw you off balance: Choosing the right classifiers and performance metrics for brain decoding with imbalanced data. *NeuroImage*, 277, 120253. <https://doi.org/10.1016/j.neuroimage.2023.120253>

Wang, X., Jin, Y., Schmitt, S., & Olhofer, M. (2023). Recent Advances in Bayesian Optimization. *ACM Computing Surveys*, 55(13s), 1–36. <https://doi.org/10.1145/3582078>

Wu, H., Dong, Z., & Rojas-Cessa, R. (2023). DICE: Data Imputation for Cost Estimates from Multiple Sources to Model User Decision-Making. *2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI)*, 149–154. <https://doi.org/10.1109/ictai59109.2023.00029>

Yap, M., & Cats, O. (2025). Ride-hailing vs. public transport: Comparing travel time perceptions using revealed preference data from Washington DC. *Travel Behaviour and Society*, 41, 101069. <https://doi.org/10.1016/j.tbs.2025.101069>

Zheng, P. (2025). The Impact of Contract Type and Monthly Charges on Customer Churn: Comparative Analysis Across Age Groups. *Advances in Economics, Management and Political Sciences*, 153(1), 22–27. <https://doi.org/10.54254/2754-1169/2024.19469>