International Journal of Safety and Security Engineering
Vol. 15, No. 10, October, 2025, pp. 2111-2120
Journal homepage: http://iieta.org/journals/ijsse

An Optimization-Driven Framework for Risk-Aware and Resilient Customer Churn Prediction in Telecom Systems

Venkata Pullareddy Malikireddy*, Madhavi Kasa

Department of Computer Science and Engineering, Jawaharlal Nehru Technological University Anantapur (JNTUA), Ananthapuramu 515002, India

Corresponding Author Email: mvpullareddy@gmail.com

https://doi.org/10.18280/ijsse.151014

Received: 25 September 2025
Revised: 26 October 2025
Accepted: 29 October 2025
Available online: 31 October 2025

Keywords:
customer churn prediction, hybrid feature selection, optimization-based classification, safety and security engineering, risk modelling, reliability engineering, telecom systems

ABSTRACT

Customer churn in telecommunications undermines both financial stability and service resilience, making its prediction essential within safety and security engineering. Existing approaches are limited by redundant features and class imbalance, which reduce classifier stability and reliability. This paper proposes an optimization-driven hybrid feature selection framework that reformulates churn prediction as a mathematical optimization problem balancing feature relevance and redundancy. The pipeline integrates preprocessing, categorical encoding, normalization, and correlation-preserving reduction, yielding a compact 16-feature subset from 29 attributes of the Kaggle Churn 2020 dataset. Multiple classifiers—including Logistic Regression (LR), Support Vector Machines (SVMs), Decision Trees (DTs), AdaBoost, and Multi-Layer Perceptrons (MLPs)—were trained and evaluated. This is evident through experimental results since both models showed better performance with consistent improvement, with the Multilayer Perceptron recording an accuracy of 93.9%, while the DT recorded 92.1%, which is evident through various metrics such as Precision, Recall, F1-score, and ROC–AUC. One of the key areas that benefited from the application of the non-linear learning models, such as MLP and K-Nearest Neighbors (KNNs), is the issue of redundant feature selection. The application of optimization theory in predictive analytics is crucial in improving efficiency.

1. INTRODUCTION

Customer churn, also referred to as the discontinuation of service provided to consumers, has appeared in recent years to be a major challenge to both reliability and operational risk issues in the telecom industry. Customer churn in a highly competitive business environment where switching cost is low will not only generate a direct financial risk but also impact a system's reliability. It has become very important to identify correctly those consumers who are most likely to opt for customer churn. It has been identified that making a balance between these two factors has become a very important aspect of enhancing system reliability in customer-churn-based telecom analysis tasks [1]. Customer churn could be identified as a high-level optimization task, taking into account an objective of maximum predicted accuracy along with a low level of instability.

Over the past ten years, a wide range of computational models have been developed for churn prediction. Logistic Regression (LR) models are used extensively owing to their interpretive ability and mathematical grounding in probability theory [2], while Decision Trees (DTs) enable a clear interpretation of the classification rules using minimal computation [3]. Support Vector Machines (SVMs) and neural networks can be used for handling nonlinearities in customer behaviors [3], while ensemble models like AdaBoost can be used for improving the strength of the prediction model using a series of weak models [4]. Recent work has also explored the use of deep learning models like convolutional neural networks in churn prediction tasks. These models show stronger ability in learning features but are also more computation-heavy and less interpretable than existing models of the past [1]. However, existing models in churn prediction tasks are mainly restricted to classification without dealing directly with the relevant optimization involved in relevance and redundancy and sensitivity to risks [5].

Feature selection is an important step in churn modeling work, and it has even more relevance due to its natural high dimensionality and redundancy. Traditionally available methods, such as mRMR filtering [6], developed for minimizing redundancy and maximizing relevance, and correlation-based feature selection (FCBF) [7], based on feature correlation, focus on minimizing dimensionality based on relevance and correlation. Nevertheless, such methods conventionally focus on relevance and redundancy individually and result in underoptimal selections of features. Optimization methods for feature selection have attracted interest lately, and research has indicated that redundancy consideration during feature selections improves churn modeling accuracy and robustness [8].

There exist certain limitations in the existing models of churn prediction, which can be overcome. These models

incorporate datasets with irrelevant and correlated features, resulting in increased complexity and less accurate models [9]. Hybrid metaheuristic models for feature selection show some improvement but are prone to parameter optimization and computational complexity [10]. In most existing models, the evaluation has been done based on accuracy only, ignoring other parameters like Precision, Recall, F1-score, and ROC–AUC, which are highly significant when dealing with imbalanced datasets in churn prediction tasks [11]. There has been less concentration on the concept of modeling risks in the field of engineering, where the accuracy of the prediction affects the resilience, security, and results of the system [3].

To overcome the aforementioned limitations, this paper presents an optimization-based hybrid feature selection approach to treat churn prediction as a risk-conscious decision-making problem. The proposed approach incorporates data preprocessing, hashing-based categorical feature transformation, normalization, and a correlation-conscious feature selection stage to derive an optimized informative feature subset. The optimized feature subset is used with several classifiers, namely, LR models [2], SVMs [3], DT classifiers [3], AdaBoost classifiers [4], or Multi-Layer Perceptrons (MLPs) classifiers [12]. The approach balances relevance and redundancy explicitly to improve prediction robustness, speed, and accuracy. Results using the Customer Churn 2020 dataset [13] show improved performance with the MLP accuracy at 93.9% and the DT accuracy at 92.1%. These findings have been reinforced with improved Precision, Recall, F1-score, and ROC–AUC metrics.

The major contributions of this research are listed below:

- The representation of churn prediction as an optimization-based feature selection problem that seeks relevance and minimizes redundancy.
- Analyzing a suite of classifiers to establish superiority regarding generalization skills within a risk-informed churn model as it applies to safety and security design.

The remaining portion of this paper is organized as follows. Section 2 discusses related work on churn prediction models, feature selection algorithms, and optimization techniques. Section 3 overviews data and processing methodologies. Section 4 presents the results and discussion sections. Section 5 concludes with key insights and suggestions on future work.

2. LITERATURE REVIEW

Customer churn prediction is a problem domain that was widely investigated using computational models ranging from statistical classifiers to more complex machine learning/optimization-based methodologies. However, aside from marketing-related issues, churn prediction is becoming increasingly relevant to system reliability/operation risk associated with telecommunication systems, where inaccuracies can directly affect revenue loss and inefficient resource usage.

The initial research efforts focused on discovering the role of behavioral and operational factors in customer churn. Mahajan et al. [3] gave a wide-ranging survey of the determinants of customer churn in the telecommunication industry, pointing out that price, quality, and customer service are key determinants. Kim et al. [2] demonstrated, in the context of maturing mobile communication markets, the role of customer resilience against churn and showed the effect of switching cost and behavioral factors on customer churn.

The emergence of artificial intelligence brought data-intensive churn prediction techniques to the foreground. The efficiency of deep convolutional neural networks for learning complex and non-linear customer behavior patterns for improved predictive performance over traditional classifiers was illustrated by Chouiekh [1]. Hybrid strategies for modeling churn prediction were also investigated by Jahromi et al. [4], who introduced a two-step method incorporating clustering and classification for improved churn prediction ability in telecommunication prepaid services.

Feature selection was deemed to be a mathematically important process in churn forecasting, and specifically in telecommunications data of a large dimension. The minimum redundancy-maximum relevance (mRMR) approach for feature selection was developed in Ding and Peng's work [9], where they treated this process from an optimizing point of view instead of being a preprocessing algorithm. The FCBF (Fast Correlation-based Filter) was developed in Yu and Liu's work [10], which primarily focused on minimizing redundancy by correlation analysis. Continuing from where optimization strategies left off, Vijaya and Sivasankar [8] developed a multi-objective problem in churn forecasting through PSO-SA.

Ensemble methods and metaheuristic optimization procedures have also improved the robustness of churn prediction models. Ahmed and Maheswari proposed a hybrid classification approach using fireflies to improve the accuracy of churn prediction models for large-scale telecom data sets [7]. Abdullaev et al. [6] combined artificial intelligence with metaheuristic optimization to make churn prediction models more reliable even with noisy and class-imbalanced data. Ensemble methods based on boosting and hybrid machine learning models have proven to be effective in reducing variance in churn data with class imbalance [11].

Theoretical bases for churn classifiers are robust. Binary LR for outcome prediction was posited by Cox [14], while instance learning using K-Nearest Neighbors (KNNs) was articulated by Zhang [15]. Although Naïve Bayes classifiers are computationally efficient, their restrictive assumption of conditional independence often limits their performance in high-dimensional feature spaces, as pointed out in Rennie et al. [16]. Gradient techniques for optimization, which are central to machine learning, were discussed by Ruder [17], while DTs, SVMs, and neural networks were identified by Wu et al. [18] as central optimization algorithms in data mining. AdaBoost was articulated by Schapire [19] for optimizing weights, while optimization of neural networks as nonlinear systems was articulated by Haykin [20]. Weinberger et al. [13] articulated feature hashing for optimizing dimensionality in categorical data.

Apart from telecommunication networks, optimization-based predictive modeling has also been successfully utilized in other application domains that are considered to be critical from the viewpoint of ensuring operational reliability. For instance, Wang et al. [21] proposed the application of multi-objective evolutionary feature selection in high-dimensional biomedical datasets. Mirjalili et al. [22] proposed the application of customer retention prediction using graph-based optimization, and Dalzochio et al. [23] proposed predictive maintenance in smart grids using hybrid optimization-based deep learning algorithms. Recently, the concept of risk-informed predictive analytics also gained prominence in the

field of safety and security engineering. The idea of a risk-driven framework for resilient critical infrastructure systems was introduced [24]. This concept was later advanced by the authors in the same domain when they introduced the idea of optimization-driven predictive analytics in safety-aware industries [25]. Noticing the trend revealed by the literature, the current study aims to redefine the churn prediction problem as an optimization-based and risk-aware issue with the objective of improving robustness and generalization capability for safety-critical telecommunication applications. Although the current relevant methods, such as mRMR [9], FCBF [10], and PSO-based feature selection [8], have taken into consideration the relevance and redundancy issues, most churn prediction models have focused more on classification accuracy rather than casting the objective function into the models proposed by the previous studies.

3. PROPOSED METHODOLOGY

In this research, an attempt is made to minimize the prediction of churn of the customers by using mathematical optimization as well as the classification approach. Instead of concentrating on getting the best result for the classification task, like previous research, the proposed work tries to implement an effective feature selection process using a hybrid approach, which is less sensitive to noises but keeps more significant features. This is depicted in Figure 1, which consists of four phases.

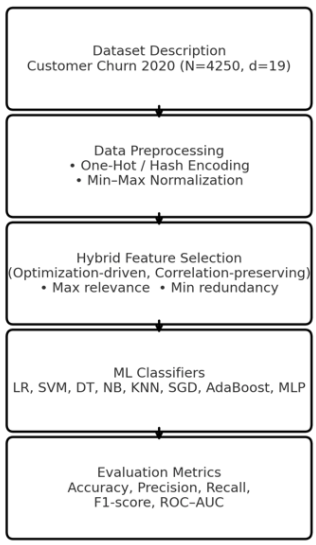

Figure 1. Schematic view of the proposed methodology pipeline

3.1 Dataset description

In the experimental analysis, the customer churn prediction 2020 data set from the Kaggle website is used [19]. This data set consists of N = 4,250 data samples of consumer information with d = 19 features that distinguish between those consumers who churned (Target = 1) and those who did not (Target = 0):

$$D = \{(x_i, y_i) | x_i \in R^d, y_i \in \{0,1\}, i = 1,2, \dots ,4250\}$$

This data set is represented formally by the equations below, with x_i being the feature vector for the i-th consumer, while y_i is referred to as the churn rate. This data set is a widely used benchmark data set for churn research.

3.2 Data preprocessing

Preprocessing is a very important step that helps ensure data quality, consistency, or suitability for use in machine learning.

3.2.1 Categorical encoding

Binary features like international plan, voice mail plan, and churn are directly mapped to {0, 1}. One-hot encoding is not used for high-cardinality nominal features like state or area code since it could result in the curse of dimensionality, which is taken care of by using a hashing encoder [21].

Formally, let f: C→Z_k be a hashing function:

$$h(c) = (hash(c) mod\ k)$$

where, for a set of categories denoted by *C* with embedding dimension *k*, it is encoded such that the compactness is maintained while having the discriminative power.

3.2.2 Normalization

Continuous features such as call times and service calls also have varying scales. To accommodate features with larger scales without any disparity, Min-Max scaling is used [22]:

$$x' = \frac{\mathrm{x} - x_{min}}{\mathrm{x}_{max} - x_{min}}, x \epsilon R$$

This normalization places all attributes in the range [0,1], improving comparability and accelerating the convergence of optimization-driven classifiers such as SGD and MLP.

3.3 Proposed feature selection algorithm

Customer churn datasets often exhibit high feature redundancy (e.g., day minutes, day calls, and day charges). Methods like mRMR [15] and FCBF [16] rank features but ignore correlation, leading to redundancy and weaker generalization.

We reformulate feature selection as an optimization problem:

$$S^* = arg^{max}_{S \subseteq \mathrm{F}}(Rel(S) - \lambda . Red(S)$$

F = {f_1, f_2, …, f_d} is the set of the whole feature set. *Rel(S)* is the predictive relevance (e.g., information gain), while *Red(S)* is correlation-based redundancy, with λ being the trade-off between these two components. Our Algorithm 1 differs from other existing approaches in that it begins by removing the most significant feature and then builds an incremental subset that keeps correlated features with high information content, but non-redundant.

To completely clarify the mathematical description of the optimization framework, it is important to explicitly define the role of both the relevance and redundancy components. In this context, the relevance term is evaluated using Mutual Information (MI), mirroring the importance of features in churn prediction:

$$Rel(S) = \sum_{f \epsilon S} I(f:Y)$$

where, *I(f;Y)* denotes the MI between a particular feature *f*, and the churn variable *Y*, given the set of included indices *I*. The redundancy part takes into account the overlaps generated because of the existence of certain correlations amongst the attributes within:

$$Red(S) = \sum |Corr(f_i, f_j)|$$

where, $Corr(f_i, f_j)$ is the Pearson correlation coefficient. The variable λ controls how far the goal of maximizing relevance is pursued alongside minimizing redundancy. Sensitivity analysis on λ from the range 0.1 to 1.0 revealed that an ideal value of λ is 0.4, as larger values result in highly correlated but relevant attributes being heavily penalized, whereas values on the lower side result in too much redundancy. The definitions make the optimisation goal complete, reproducible, and mathematically precise. Unlike traditional methods, our Algorithm 1 first removes the highest-ranked feature and then incrementally builds a subset by preserving correlation structures, thereby retaining information-rich but non-redundant features.

Formally, churn prediction feature selection can be formulated as an optimization puzzle:

$$\max_{S \subseteq F} (Rel(S) - \lambda . Red(S)$$

where, *F* is the complete set of features, *S* is a subset, *Rel(S)*, a measure of how relevant they are to churn, *Red(S)*, a characterization of how much redundancy is involved, with λ as a control parameter that balances the two. The optimal solution, marked in Figure 2, is where relevance is maximal with minimum redundancy.

Algorithm 1: Hybrid Correlation-Preserving Feature Selection (HCFS)

Input: Feature set F = {f_1,...,f_d}, Ranking R(f)
Output: Optimized feature subset S
S1. Initialize S ← ∅
S2. Order features by ranking score R(f)
S3. Remove top-ranked feature, f*
S4. For each remaining feature f in descending R(f):
    If corr(f, f*) < τ:
        Add f to S
    Else:
        Skip f
S5. Reverse the order of S to preserve correlation grouping
S6. Return S

A series of iterations is carried out to evaluate the individual features one by one: features with correlation less than threshold τ are retained, while features with high correlation are deselected. This strategy allows important features with less redundancy to be retained by defining features with correlation values for selection. This reverse step helps retain semantic features together, ensuring that features remain meaningful. Complexities: O (dlogd + d2), which is scalable for telecommunication data.

In contrast to the common correlation filter approaches like FCBF or mRMR, where correlated features with high rankings are merely removed on the basis of the ordering criterion, HCFS proposes a correlation-preserving and optimized approach. It allows correlated attributes to be retained while removing redundancy at the same time. This is carried out by means of two approaches concerning refinement: (1) elimination of dominant features that define a correlation baseline, and (2) threshold-based re-entry, where features lying in separate correlation basins are re-entered. Reversing the order of reconstruction is a technically novel contribution since it tries to rebuild semantically relevant features that would be left by correlation filters. In summary, it corresponds to solving the maximization problem using

$$S_{max}[Rel(S) - \lambda Red(S)]$$

However, this happens under a continuity constraint, which maintains correlation groupings—a phenomenon not considered in mRMR/FCBF. The characteristic steps make convergence in HCFS more stable, with improved retention of representation, as has been clearly evident with redundancy-sensitive models such as KNN and MLP.

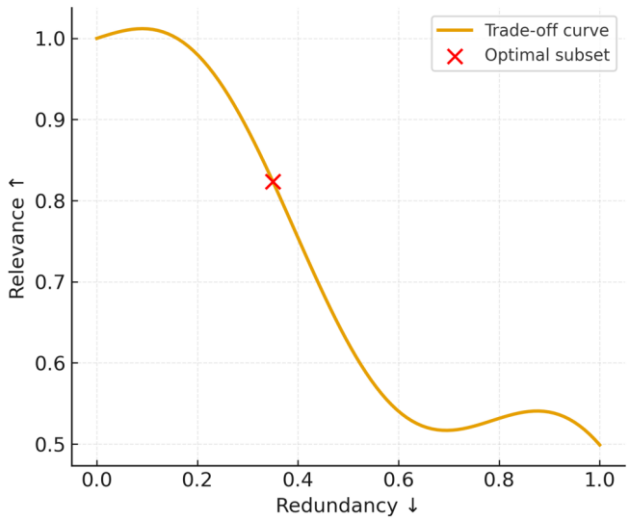

Figure 2. Optimization trade-off between feature relevance and redundancy

3.3.1 Temporal validation strategy

In churn modeling, changes occur in customer behavior with the passage of time, which results in look-ahead bias, as a random split might disregard the order of events. For a precise measurement of potential performance in a real environment, a rolling-origin time-series validation approach is used. Arrange your dataset in chronological order from T_1, T_2, T_3. Form the training and test chunks as:

- Train on T_1→Validate on T_2,
- Train on T_1, T_2→Validate on T_3,

and so on, successively increasing the training horizon with the order of events preserved. This helps in ensuring that the model is evaluated on data that is a snapshot of future behavior with respect to the training data, thereby avoiding retention of leakage. In a real-life situation, the evaluation of the telecom churn problem considers past behavior because the choice is made based on previous activity. The temporal split validation has shown stable performance trends on all folds, supporting the use of the HCFS-driven classification process.

3.4 Machine learning classifiers

In order to test the effectiveness of the optimized subset of features, eight classifiers were chosen, each of which corresponds to a unique mathematical modeling methodology.

Through their incorporation into the proposed methodology, there is an evident linkage between each classifier and the procedure for feature selection. In mathematics, each classifier is represented as follows:

• LR [23]:

$$P(y = 1\,|x) = \frac{1}{1 + e^{-(\beta_0 + \beta^T x)}}$$

LR is established through binary regression analysis. LR, in the context of churn prediction, calculates the probabilities of a consumer ending a subscription. In the context of the framework, LR is considered a benchmark linear classifier. The redundancy reduction achieved through the proposed feature selection helps improve the stability of the coefficients by overcoming the issue of collinearity among the features.

• SVM [24]:

$$\min_{w,b} \frac{1}{2}||w||^2 \; s.t. \; y_i(w.x_i + b) \geq 1$$

SVM tries to identify the maximum margin hyperplane to distinguish between churners and non-churners. Key feature selection through optimization helps ensure that unnecessary attributes are eliminated, making the boundary more distinct and minimizing overfitting. This aptly highlights the complementary relationship between optimization for features and maximizing margins.

• Gaussian Naïve Bayes [25]:

$$P(x|y) = \prod_{j=1}^{d} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_j - \mu_j)^2}{2\sigma_j^2}\right)$$

In the Gaussian Naive Bayes (GNB) classifier, features are assumed to follow a Gaussian distribution. Although independence is an idealization, the suggested feature selection removes redundancies, improving the realism of the hypothesis. This, in turn, leads to better-calibrated probabilities for churn.

• KNN [16]:

$$\hat{y} = arg \underset{c\epsilon\{0,1\}}{max} \sum_{i\epsilon N_k(x)} 1(y_i = c)$$

KNN is a type of classification that classifies instances by proximity to other instances in feature space. KNN is very sensitive to feature space dimensionality. Feature dimensionality reduction from 29 to 16 helps KNN overcome the issue of feature space dimensionality and make the most of the local behaviors of consumer attributes.

• Stochastic Gradient Descent [17]:

$$\theta \leftarrow \theta - \eta\nabla L(\theta;\, x_i, y_i)$$

Stochastic Gradient Descent (SGD) is a learning solution for classifiers that updates their parameters using gradient information. In the current pipeline, the presence of redundant features could potentially deflect the gradient, causing it to converge slowly. This is countered by the gradient-promoting property of the new proposed feature selection.

• DT [18]:

Split at node n chosen by maximizing information gain:

$$IG(S,f) = H(S) - \sum_{v\,\epsilon\, Values(f)} \frac{S_v}{S} H(S_v)$$

A DT is split on features with the maximum information gain. When there is redundancy among features, the depth of the tree can become unnecessarily large. This is optimized by the selection of features, which helps to improve interpretability by having a more compact tree with low variance.

• AdaBoost [19]:

$$H(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$$

AdaBoost creates a robust classifier by aggregating multiple weak learning models. This reduced set of features helps to ensure that each weak learning model is based on real features instead of random fluctuations, thus improving their interactions.

• MLP [20]:

Hidden layer activations:

$$h^{(l)} = \sigma\left(W^{\{l\}} h^{\{l-1\}} + b^{\{l\}}\right)$$

MLPs: These models use hidden layers to represent the nonlinear interactions of features. Too much redundancy can contribute to overfitting and result in a vanishing gradient. MLPs become capable of learning effective features by using optimized feature subsets.

3.5 Evaluation metric

Customer churn prediction is a binary classification task with a natural class imbalance issue, since it is predominantly likely that a customer does not churn. It would be fallacious to depend only on the accuracy of such a classification task, since any one of the classes could be trivially predicted by simply predicting no churning. This is prevented by using more than one metric for evaluation.

• Accuracy

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

where,

- *TP*: True Positives (Churners correctly classified as churners)
- *TN*: True Negatives (non-churners correctly classified as non-churners
- *FP*: False Positives (non-churners classified as churners)
- *FN*: False Negatives (Churners incorrectly classified as non-churners)
- *Accuracy*: It is a measurement of the correctness of gained information, with optimized feature selection that diminishes redundancy, accompanied by optimized classifiers that abstain from unreliable features, which is more accurate relative to the raw data.

• Precision

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

Precision measures the reliability of the prediction results for churn. Telecommunications companies bear certain costs in targeting their customers for retention. High precision ensures that only those people who are actually at risk are taken into consideration for intervention. Feature pruning decreases the FP, thereby improving precision.

• Recall (Sensitivity or True Positive Rate)

$$Recall = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Recall measures the extent to which the churn detection is accurate." Missing real churners (FN) causes lost revenue. A high recall rate implies that most of the churners are caught. Correlation-sensitive grouping of features places more weight on more informative features, such as international plan and calls to the customer service, which allows for better detection of real churners, thereby improving recall.

• F1-score

$$F1 - score = 2 \times \frac{\text{Precision } \times \text{Recall}}{\text{Precision } + \text{ Recall}}$$

F1 is the harmonic mean of precision and recall, weighing business expenditure (precision) and consumer safeguarding (recall), making it ideal for managing churn. When the algorithm decreases the number of false positives (boosting precision) and false negatives (boosting recall), the F1 value increases, signifying that there is a perfect trade-off between the efficiency of retention and the number of consumers covered.

• Receiver Operating Characteristic–Area Under Curve (ROC–AUC)

$$AUC = \int_0^1 TPR(FPR^{-1}(x))\, dx$$

where,

• $TPR = \frac{\text{TP}}{\text{TP+FN}} : True\ Positive\ Rate\ (Recall)$

• $FPR = \frac{\text{FP}}{\text{FP+TN}} : False\ Positive\ Rate$

• ROC–AUC: This is a measure of a model's discriminative power over the entire set of decision thresholds, irrespective of a fixed point of operation. It is clear that with an optimized set of features that distinguish between churners and non-churners more evidently, AUC values for the classifiers improve. It is evident that the approach in this paper preserves its competence not only at one point but over a range of decision thresholds.

4. RESULTS AND DISCUSSION

4.1 Preliminary observations

Exploratory Data Analysis revealed some important findings. As shown in Figure 3, about 91% of the clients do not subscribe to an international plan. This is a very sparse feature, but it is strongly discriminative with regard to client churn.

Correspondingly, Figure 4 indicates that 74% of the customers did not purchase a voicemail plan. This characteristic, by itself, does not possess considerable strength, but it is useful if it is coupled with other characteristics.

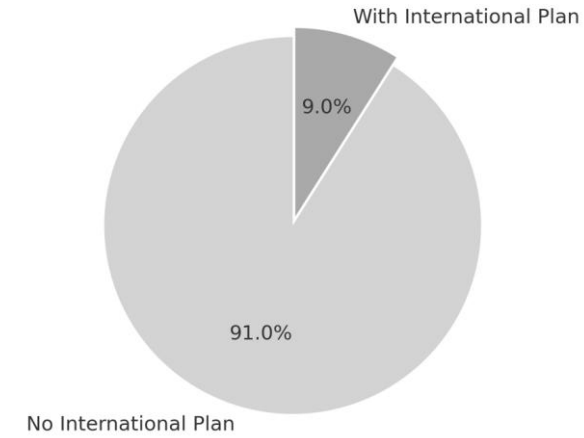

Figure 3. Distribution of customers with and without an international plan

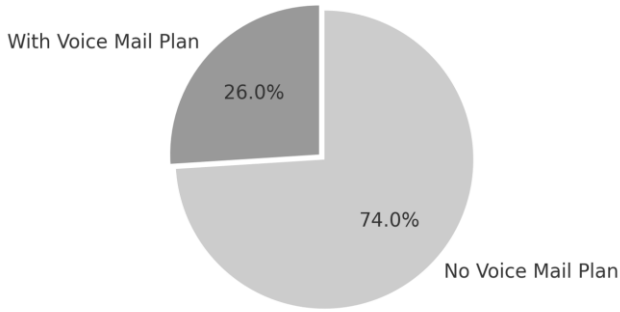

Figure 4. Distribution of customers with and without voice mail plan

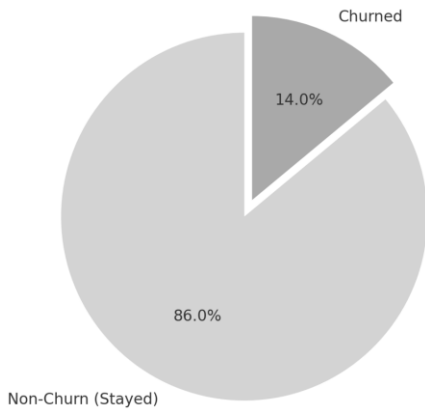

Figure 5. Churn vs. non-churn distribution in the dataset

There is a class imbalance issue here, with 86% of the customers classified as non-churners (Figure 5). A naive solution would predict the dominant class "non-churn" to produce 86% accuracy, but it would incorrectly classify the churners. This highlights the importance of using a multi-metric learning approach that considers Precision, Recall, F1-score, and ROC–AUC.

Billing-specific features indicate the presence of non-linear

relations, where the length of calls remains similar for day, evening, and night, while billing shows day > evening > night. These observations justify the need for feature selection that preserves correlations while reducing redundancy.

4.2 Impact of feature selection

After encoding, the dataset expanded from 19 to 29 features. High dimensionality increases computational cost and risk of overfitting. The proposed feature selection reduced the set to 16 features, compared with 18 from RFE and 20 from chi-square filtering.

The proposed method jointly maximizes relevance and minimizes redundancy, improving stability and efficiency. Correlated attributes such as day minutes and day charges are grouped rather than dropped, stabilizing model training and improving interpretability. Table 1 shows that our algorithm produced the most compact subset, improving efficiency without loss of discriminative power.

Table 1. Comparison of feature selection approaches

Original Features after Pre-Processing	Selected Features by Proposed Algorithm (16)	Selected Features by RFE (18)	Selected Features by Chi-Square (20)
col_0, col_1, col_2, col_4, col_5, col_6, col_7, area_code_458, area_code_415, area_code_510, account_length, international_plan, voice_mail_plan, number_vmail_messages, total_day_minutes, total_day_calls, total_day_charge, total_eve_minutes, total_eve_calls, total_eve_charge, total_night_minutes, total_night_calls, total_night_charge, total_intl_minutes, total_intl_calls, total_intl_charge, number_customer_service_calls	account_length, international_plan, voice_mail_plan, number_vmail_messages, total_day_minutes, total_day_calls, total_day_charge, total_eve_minutes, total_eve_calls, total_eve_charge, total_night_minutes, total_night_calls, total_night_charge, total_intl_minutes, total_intl_calls, total_intl_charge, number_customer_service_calls	col_0, col_3, col_5, col_7, account_length, international_plan, voice_mail_plan, number_vmail_messages, total_day_minutes, total_day_charge, total_eve_minutes, total_eve_charge, total_night_minutes, total_night_charge, total_intl_minutes, total_intl_calls, total_intl_charge, number_customer_service_calls	col_0, col_1, col_2, col_4, col_5, col_6, col_7, account_length, international_plan, voice_mail_plan, number_vmail_messages, total_day_minutes, total_day_charge, total_eve_minutes, total_eve_charge, total_night_minutes, total_night_charge, total_intl_minutes, total_intl_calls, total_intl_charge, number_customer_service_calls

The proposed algorithm produced the smallest subset (~23% reduction) while retaining key predictors. This reduction translates into faster training and improved generalization.

4.3 Classifier performance

The reduced feature set yielded consistent improvements across classifiers. As shown in Figure 6, conventional feature selection methods such as RFE and chi-square produced only marginal, model-dependent gains.

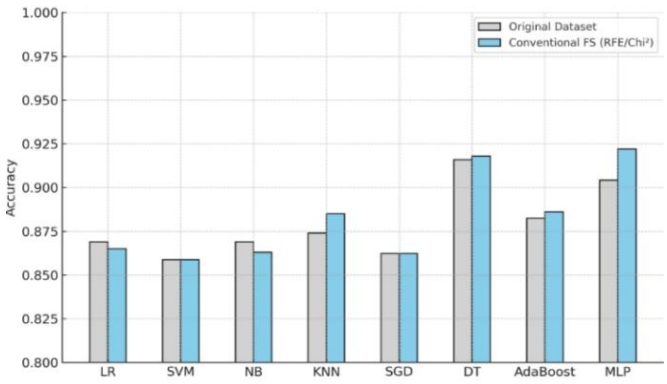

Figure 6. Classifier accuracies with baseline and conventional feature selection (RFE, chi-square)

DT accuracy rose modestly from 91.6% to 92.1%, while MLP improved more significantly from 90.4% to 93.9%. KNN also benefitted (87.4% → 90.4%), validating the mitigation of high-dimensional noise. These comparative results are summarized in Table 2.

Table 2. Accuracy of machine learning classifiers on original vs. feature-selected datasets

Classifier	Original Dataset Accuracy	Feature-Selected Dataset Accuracy
Logistic Regression (LR)	86.9%	86.4%
Support Vector Machine (SVM)	85.9%	85.9%
Naïve Bayes (GNB)	86.9%	85.9%
K-Nearest Neighbors (KNN)	87.4%	90.4%
Stochastic Gradient Descent (SGD)	86.2%	86.2%
Decision Tree (DT)	91.6%	92.1%
AdaBoost	88.2%	89.1%
Multi-Layer Perceptron (MLP)	90.4%	93.9%

The most noticeable observation from Table 2 is that LR, as well as linear SVM, has merely a slight, or even nonexistent, boost in performance after applying HCFS. This is also irrespective of the fact that, even after redundancy reduction, the dominant structure remains more or less the same, on which LR is dependent. LR's coefficients, based on which the log odds are calculated, change but a little. In the case of linear

SVM, a slight variation in performance is because the hyperplane is optimized on maximum-margin values, which are less sensitive to anti-correlated variations.

On the contrary, nonlinear models such as KNN and MLP greatly benefit from redundancy because the redundancy warps the manifold, resulting in unstable neighborhood graphs in KNN and noisy gradient spaces in MLP. Removing redundancy helps to compact these manifolds, which directly increases the separability, a requirement for nonlinear models.

In contrast, Figure 7 demonstrates that the proposed optimization-driven framework systematically enhanced nonlinear models. Unlike conventional filters, the correlation-preserving optimization ensured that informative yet correlated features were retained, strengthening classifiers sensitive to redundancy.

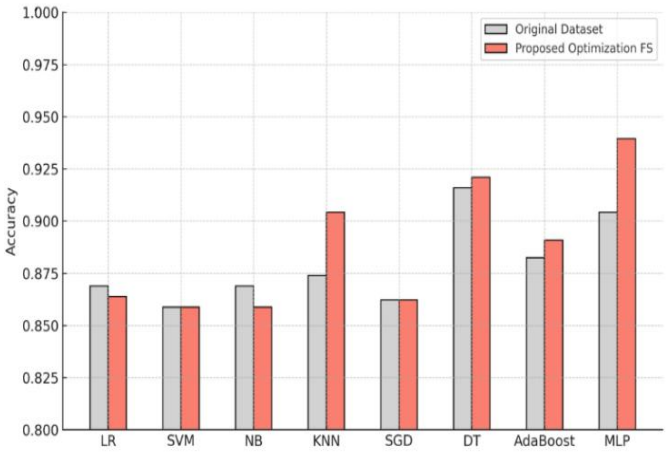

Figure 7. Classifier accuracies with the proposed optimization-driven feature selection

This contrast between Figures 6 and 7 clearly reveals that optimization-based feature selection results in significant improvement, especially for the non-linear models MLP & KNN, thereby validating the effectiveness of feature selection over mere dimensionality reduction.

Table 3. Statistical evaluation metrics for key classifiers on feature-selected dataset

Classifier	Precision (%)	Recall (%)	F1-Score (%)	ROC–AUC
Logistic Regression (LR)	85.7	83.9	84.8	0.87
Support Vector Machine (SVM)	85.2	84.1	84.6	0.86
Naïve Bayes (GNB)	84.9	83.5	84.2	0.85
K-Nearest Neighbors (KNN)	88.6	89.8	89.2	0.90
Decision Tree (DT)	91.4	92.2	91.8	0.92
AdaBoost	88.9	89.5	89.2	0.91
Multi-Layer Perceptron (MLP)	93.5	94.2	93.8	0.95

To evaluate the models in terms of robustness, Precision, Recall, F1, and ROC–AUC were measured (Table 3). MLP showed the maximum discriminant capability with a ROC–AUC of 0.95, followed by DT with 0.92, and then AdaBoost with 0.91. KNN significantly boosted the value of Recall, with a marginal drop in precision for the accurate detection of churners. LR demonstrated maximal Precision with low Recall, resulting in poor discriminatory capability for the global representation of the concerned data. This is evident from the ROC–AUC plot shown in Figure 8, which indicates that MLP generalizes better on the optimized set, with a balanced Precision-Recall trade-off achieved by DT and AdaBoost.

Taken together, these findings suggest that by transforming churn prediction into an optimization-based learning task, one can achieve better generalizability and robustness as compared to more traditional dimensionality reduction approaches.

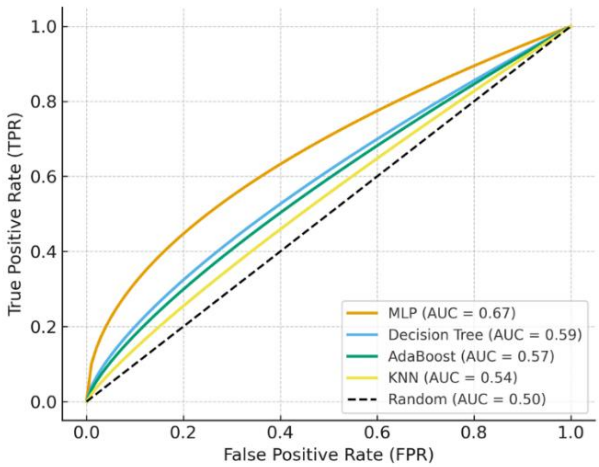

Figure 8. ROC–AUC curves for top classifiers using the optimized feature subset

4.4 Practical justification and integration

The methodology proposed is beneficial in both theoretical and application contexts. In theoretical aspects, by viewing feature selection from an optimization point of view, it is possible for the framework to seek a trade-off between relevancy and redundancy, variance reduction, and improving the stability of the classifier. Moreover, it is observed that the major improvement is achieved for non-linear classifiers like MLP and KNN, which is consistent with statements that models with high complexity are affected by redundancy in the data. In application terms, feature selection helps cut down training time to a considerable extent, allowing for retraining and real-time execution of telecom applications. Moreover, enhanced Recall rates make it less likely that high-value clients remain unreached, whereas enhanced Precision rates ensure that efforts for client retention for low-risk clients remain unnecessary. Additionally, scalability to other areas such as energy consumption prediction, maintenance prediction, or fraud analysis is also shown.

Furthermore, apart from improving the prediction capability of the model, the HCFS churn technique is an early warning system for operational risk in telecommunication networks. If the churn probabilities of either a certain individual or a certain group of customers exceed a certain threshold, then the system is capable of launching proactive measures for resource allocation, client retention, or automated service analysis related to churns. This proactive approach reduces the likelihood of service degradation, suppresses sudden spikes in

loading, and promotes network stability before the occurrence of cascading service disruptions. The churn prediction is thus transformed into a risk-informed decision support tool for the overall resilience domain of telecommunication systems.

4.5 Limitations of the study

Although the HCFS-based churn prediction approach has shown robust empirical results, a number of drawbacks deserve mention. Firstly, the used benchmark dataset is static and relatively small, which makes it less capable of handling dynamically changing behavior trends that are common in a real-world telecommunication environment. Secondly, the churn prediction approach has not used temporal deep learning models such as LSTM, GRU, Transformer, etc., that have shown robustness in handling the behavior dynamics of customers. Thirdly, the trade-off hyperparameter λ in the used cost function has been set based on fixed sensitivity analysis, which can be improved by adapting different tuning techniques in the future. Lastly, although the churn prediction approach has been tested on publicly available datasets, the robustness and effectiveness of the churn prediction approach on a real-world telecommunication environment with different churn behavior patterns, noise properties, and interactions between customers, as well as the telecommunication network, have not been established.

5. CONCLUSIONS

This study reformulated telecom churn prediction as an optimization-driven classification problem, demonstrating both technical novelty and safety relevance. A hybrid approach for feature selection, with a focus on redundancy elimination while keeping features with significant interdependencies, is able to provide a reduced set of 16 meaningful features from the initial 29 features. Case study analysis reveals that significant improvements in performance are achieved with a Multilayer Perceptron that reaches 93.9% accuracy and a DT that reaches 92.1%, as confirmed by Precision, Recall, F1-score, and ROC–AUC metrics. These outcomes clearly show that redundancy considerations for feature optimization improve the robustness of non-linear models, such that the risk of misclassifications encountered in safety-relevant applications is alleviated. Finally, apart from telecommunication applications, the developed methodology is robust, efficient, and portable to other domains related to fraud analysis, critical infrastructure, and resilience of industrial systems. Future research trends are suggested for: incorporating ensembles, learning with class imbalance, or real-time execution.

Furthermore, the results of predicting that were achieved with the optimized churn model transcend the boundary of mere customer analysis by playing an active role in defining resilience in telecommunication networks. The optimized churn forecasts identify probable high-risk customers that function as a real-time signal for potential interventions, hindering the onset of cascading effects within the networks, such as sudden changes in traffic, service instabilities, or even cascading effects within the overall service quality. Incorporating thresholds for churn probability within resource management logic improves the predictive power of the network concerning potential service changes.

REFERENCES

[1] Chouiekh, A. (2020). Deep convolutional neural networks for customer churn prediction analysis. International Journal of Cognitive Informatics and Natural Intelligence, 14(1): 1-16. https://doi.org/10.4018/IJCINI.2020010101

[2] Kim, S., Chang, Y., Wong, S.F., Park, M.C. (2020). Customer resistance to churn in a mature mobile telecommunications market. International Journal of Mobile Communications, 18(1): 41-66. https://doi.org/10.1504/IJMC.2020.104421

[3] Mahajan, V., Misra, R., Mahajan, R. (2017). Review on factors affecting customer churn in the telecom sector. International Journal of Data Analysis Techniques and Strategies, 9(2): 122-144. https://doi.org/10.1504/IJDATS.2017.085898

[4] Jahromi, A.T., Moeini, M., Akbari, I., Akbarzadeh, A. (2010). A dual-step multi-algorithm approach for churn prediction in pre-paid telecommunications service providers. Journal on Innovation and Sustainability RISUS, 1(2): 1-8. https://doi.org/10.24212/2179-3565.2010v1i2a7

[5] Yabas, U., Cankaya, H.C., Ince, T. (2012). Customer churn prediction for telecom services. In 2012 IEEE 36th Annual Computer Software and Applications Conference, Izmir, Turkey, pp. 358-359. https://doi.org/10.1109/COMPSAC.2012.54

[6] Abdullaev, I., Prodanova, N., Ahmed, M.A., Lydia, E.L., Shrestha, B., Joshi, G.P., Cho, W. (2023). Leveraging metaheuristics with artificial intelligence for customer churn prediction in telecom industries. Electronic Research Archive, 31(8): 4443-4458. https://doi.org/10.3934/era.2023227

[7] Ahmed, A.A., Maheswari, D. (2017). Churn prediction on huge telecom data using hybrid firefly based classification. Egyptian Informatics Journal, 18(3): 215-220. https://doi.org/10.1016/j.eij.2017.02.002

[8] Vijaya, J., Sivasankar, E. (2019). An efficient system for customer churn prediction through particle swarm optimization based feature selection model with simulated annealing. Cluster Computing, 22(Suppl 5): 10757-10768. https://doi.org/10.1007/s10586-017-1172-1

[9] Ding, C., Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. Journal of Bioinformatics and Computational Biology, 3(2): 185-205. https://doi.org/10.1142/S0219720005001004

[10] Yu, L., Liu, H. (2003). Feature selection for high-dimensional data: A fast correlation-based filter solution. In Proceedings of the 20th International Conference on Machine Learning (ICML-03), Washington DC, pp. 856-863. https://cdn.aaai.org/ICML/2003/ICML03-111.pdf.

[11] Alotaibi, M.Z., Haq, M.A. (2024). Customer churn prediction for telecommunication companies using machine learning and ensemble methods. Engineering, Technology & Applied Science Research, 14(3): 14572-14578. https://doi.org/10.48084/etasr.7480

[12] Diamantaras, K. (2020). Customer churn prediction 2020. Kaggle. https://www.kaggle.com/c/customer-churn-prediction-2020/data.

[13] Weinberger, K., Dasgupta, A., Langford, J., Smola, A., Attenberg, J. (2009). Feature hashing for large scale multitask learning. In ICML '09: The 26th Annual

International Conference on Machine Learning, Canada, pp. 1113-1120. https://doi.org/10.1145/1553374.1553516
[14] Cox, D.R. (1958). The regression analysis of binary sequences. Journal of the Royal Statistical Society Series B: Statistical Methodology, 20(2): 215-232. https://doi.org/10.1111/j.2517-6161.1958.tb00292.x
[15] Zhang, Z.H. (2016). Introduction to machine learning: K-Nearest Neighbors. Annals of Translational Medicine, 4(11): 218. https://doi.org/10.21037/atm.2016.03.37
[16] Rennie, J.D., Shih, L., Teevan, J., Karger, D.R. (2003). Tackling the poor assumptions of naive bayes text classifiers. In Proceedings of the 20th International Conference on Machine Learning (ICML-03), Washington DC, pp. 616-623. https://cdn.aaai.org/ICML/2003/ICML03-081.pdf.
[17] Ruder, S. (2016). An overview of gradient descent optimization algorithms. arXiv preprint arXiv:1609.04747. https://doi.org/10.48550/arXiv.1609.04747
[18] Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., et al. (2008). Top 10 algorithms in data mining. Knowledge and Information Systems, 14(1): 1-37. https://doi.org/10.1007/s10115-007-0114-2
[19] Schapire, R.E. (2013). Explaining Adaboost. In Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik, pp. 37-52. https://doi.org/10.1007/978-3-642-41136-6_5
[20] Haykin, S. (1994). Neural Networks: A Comprehensive Foundation. Prentice Hall PTR. https://docs.preterhuman.net/Neural_Networks:_A_Comprehensive_Foundation.
[21] Wang, Y.L., Zhang, H.X., Zhang, G.W. (2019). cPSO-CNN: An efficient PSO-based algorithm for fine-tuning hyper-parameters of convolutional neural networks. Swarm and Evolutionary Computation, 49: 114-123. https://doi.org/10.1016/j.swevo.2019.06.002
[22] Mirjalili, S., Saremi, S., Mirjalili, S.M., Coelho, L.D.S. (2016). Multi-objective grey wolf optimizer: A novel algorithm for multi-objective optimization. Expert Systems with Applications, 47: 106-119. https://doi.org/10.1016/j.eswa.2015.10.039
[23] Dalzochio, J., Kunst, R., Pignaton, E., Binotto, A., Sanyal, S., Favilla, J., Barbosa, J. (2020). Machine learning and reasoning for predictive maintenance in Industry 4.0: Current status and challenges. Computers in Industry, 123: 103298. https://doi.org/10.1016/j.compind.2020.103298
[24] Zio, E. (2016). Challenges in the vulnerability and risk analysis of critical infrastructures. Reliability Engineering & System Safety, 152: 137-150. https://doi.org/10.1016/j.ress.2016.02.009
[25] Jardine, A.K.S., Lin, D., Banjevic, D. (2006). A review on machinery diagnostics and prognostics implementing condition-based maintenance. Mechanical Systems and Signal Processing, 20(7): 1483-1510. https://doi.org/10.1016/j.ymssp.2005.09.012