Analyzing Telecom Customer Churn using Machine Learning

Stefan Shipinkoski, Hristijan Gjoreski
Faculty of Electrical Engineering and Information Technologies,
University of Ss. Cyril and Methodius in Skopje, Macedonia
stefanshipinkoski@gmail.com, hristijang@feit.ukim.edu.mk

***Abstract*—Predicting customer churn is a critical challenge in the telecommunications industry, with significant implications for business strategy and revenue. This study presents an analysis of various machine learning models to identify at-risk customers. The research evaluates the predictive performance of Decision Tree, Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Random Forest, and XGBoost algorithms. The models are assessed using a dataset preprocessed through feature encoding, scaling, and selection, identifying the top 15 most informative features. Performance metrics such as accuracy, balanced accuracy, precision, F1 score, recall, and ROC_AUC are calculated to compare the effectiveness of each model under default parameters, reduced feature set, and after hyperparameter tuning via Random Search Cross-Validation. The results indicate that ensemble methods, particularly the XGBoost classifier with hyperparameter optimization, outperforms individual classifiers, offering a robust solution for churn prediction. The study also highlights the efficiency of using a reduced feature set without significantly compromising model performance, suggesting a streamlined approach for practical application. These insights provide a valuable contribution to the predictive analytics field and offer a strategic asset for telecom companies in their customer retention efforts.**

Keywords— Customer churn prediction, Ensemble Machine Learning, Feature Selection Optimiziation.

I. INTRODUCTION

In the competitive landscape of the telecommunications industry, customer attrition, or churn, poses a significant threat to the sustainability and profitability of service providers. Churn, which occurs when subscribers discontinue their services, can result in considerable revenue loss and increased marketing costs as companies strive to replace lost customers. Consequently, the ability to predict and preemptively address customer churn has become a strategic imperative for telecom operators [1].

Recent advancements in machine learning (ML) offer promising solutions to this challenge by enabling sophisticated analytical capabilities that can uncover complex patterns in customer behavior and predict potential churn. This paper explores the application of supervised ML algorithms to predict churn within a telecom company. By leveraging a dataset encompassing a range of customer attributes and service details, we aim to develop a predictive model that forecasts churn with high accuracy and provides insights into key determinants of customer retention.

Our research is driven by the premise that predictive analytics, powered by ML, can transform the traditional approach to churn management. Through this study, we aim to demonstrate the predictive power of ML algorithms and provide actionable insights that telecom companies can leverage to devise targeted retention strategies and enhance customer loyalty.

II. DATASET AND ANALYSIS

A. Dataset

We used a publicly available dataset for this research obtained from the Telco Customer Churn data collection, which is hosted on Kaggle platform [2], and originates from a sample provided from IBM [3]. This dataset encapsulates the interactions and service usage of 7043 customers from a hypothetical telecommunications company based in California in the third fiscal quarter. The dataset comprises various customer attributes and service subscription details, which are important for analyzing the churn phenomenon. The dataset is composed of multiple features, each representing a specific aspect of customer information given in Table 1.

The dependent variable of interest, ‘Churn’, denotes whether a customer has terminated their subscription within the last month, with possible responses being Yes or No. The dataset not only provides a granular view of the service to which the customers are subscribed, such as phone lines, internet, and various subsidiary services, but also encompasses critical account information, including tenure, contractual agreements, payment modalities, and financial charges. Furthermore, it includes demographic data points, such as gender, age bracket, and familial status, which are essential for a nuanced demographic analysis in the context of churn.

B. Analysis

The goal of this analysis is to harness predictive modeling to anticipate customer churn, thereby enabling the formulation of targeted customer retention strategies. By leveraging the comprehensive data provided, this study seeks to generate actionable insights that can guide telecom companies in their efforts to mitigate churn and enhance customer engagement.

The exploratory data analysis conducted in this study serves as a preliminary step to understand the underlying structure and patterns within the Telco Customer Churn dataset. To identify key factors that contribute to customer churn and inform the subsequent data processing and modeling phase.

Table 1. Feature and their description

Feature Name	Description
customerID	A distinct identifier for each customer.
gender	The customer's gender, categorized as Female or Male.
SeniorCitizen	A binary indicator reflecting whether the customer is aged 65 or above.
Partner	A binary indicator of whether the customer has a life partner.
Dependents	A binary indicator of whether the customer has dependents.
tenure	The duration, in months, of the customers relationship with the company.
PhoneService	A binary indicator of the customer's subscription to phone service.
MultipleLines	The customer's subscription status to multiple telephone lines.
InternetService	The type of internet service availed by the customer.
OnlineSecurity	The customer's subscription status to online security services.
OnlineBackup	The customer's subscription status to online backup services.
DeviceProtection	The customer's subscription status to device protection plans.
TechSupport	The customer's subscription status to technical support services.
StreamingTV	The customer's subscription status to streaming TV services.
StreamingMovies	The customer's subscription status to streaming movie services.
Contract	The term of the customer's service contract.
Paperless billing	A binary indicator of the customer's preference for paperless billing.
PaymentMethod	The method by which the customer completes payments.
MonthlyCharges	The monetary cost billed to the customer monthly.
TotalCharges	The aggregate monetary cost billed to the customer.

The target variable, 'Churn', exhibits a class imbalance with approximately 26.6% of customers having canceled their subscriptions as shown on Figure 1 [5]. This imbalance underscores the necessity for careful consideration of evaluation metrics beyond accuracy, such as balanced accuracy, precision, recall, F1 score, and the ROC-AUC score, which provide a more nuanced assessment of model performance across imbalanced classes. The gender distribution within the dataset is nearly balanced, with a slight majority of male customers (50.5%). In contrast, the proportion of senior citizens is relatively low at 16.2%, suggesting that the majority of the customer base is younger. The presence of a partner and dependents among customers is almost evenly split, with a slight majority not having a partner (51.7%) and a significant majority not having dependents (70.2%).

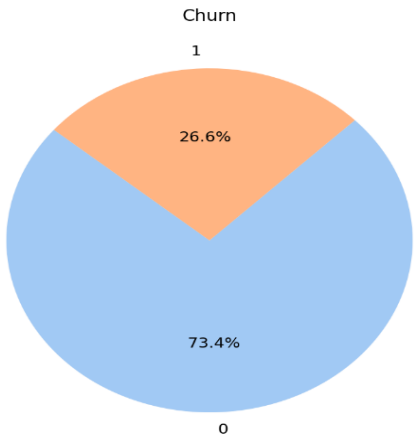

Figure 1 Churn Distribution.

Many customers have subscribed to phone services (90.3%). However, the presence of multiple lines among these customers does not show a significant difference in churn rates. Internet service type distribution reveals that customers with fiber optic service are more likely to churn than those with DSL or no-internet service, highlighting the potential impact of service quality or pricing on customer retention.

The analysis of additional services such as online security, online backup, device protection, and tech support indicates that customers lacking these services exhibit higher churn rates. This pattern extends to streaming services, where customers without streaming TV or movies tend to churn more frequently. Notably, the type of contract is a strong indicator of churn, with customers on month-to-month contracts showing a higher propensity to churn compared to those on longer-term contracts. Customers who have adopted paperless billing and those who use electronic checks as their payment method are more prone to churn. These findings may reflect a segment of customers who are more comfortable with digital services and potentially more sensitive to service dissatisfaction or competitive offers.

This analysis has revealed several features with significant differences in churn rates, as seen on the Figure 2, including the type of internet service, contract length, and billing practices. These features warrant closer examination in the predictive modeling phase and suggest potential areas for targeted customer retention strategies.

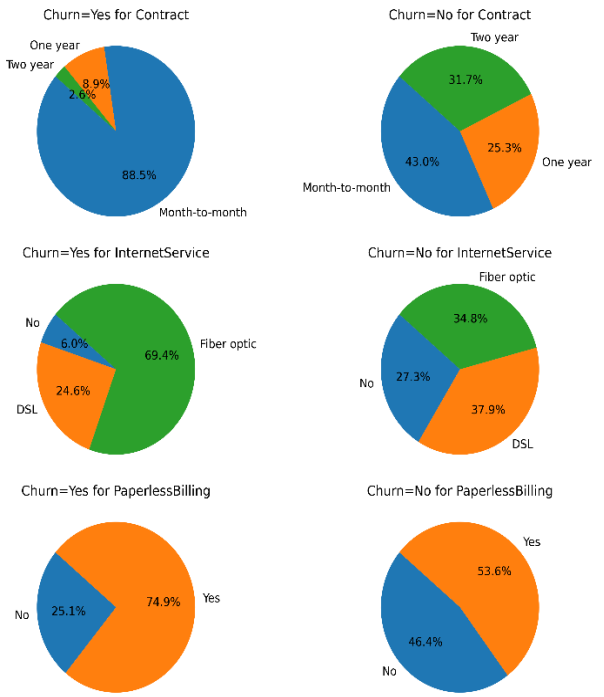

Figure 2 Additional Services and Contract Distribution

To develop our predictive models, we employed both individual classifier approaches and ensemble techniques. The individual classifier approach involved the use of several machine learning algorithms, each with its own theoretical underpinnings and assumptions:

- **Decision Tree Classifier**: A non-parametric supervised learning method ideal for classification and regression tasks. It operates by recursively partitioning data into subsets based on the feature that results in the maximum information gain at each decision node. The end result is a tree-like model of decisions as shown on Figure 3, where each path from root to leaf represents a classification rule. One of the key advantages of decision trees is their interpretability, as they can mimic human decision-making processes. However, they can be prone to overfitting, especially when they grow too complex [4].

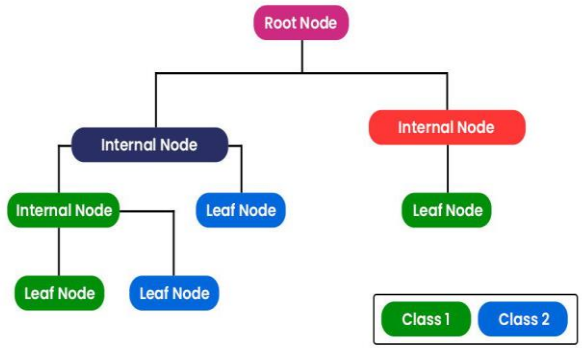

Figure 3 Decision Tree basic structure.

- **Logistic Regression**: Logistic regression is used for binary classification rather than regression. It predicts the probability of the target variable belonging to a particular class using a logistic function, which is an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1. Logistic Regression is robust to noise and provides probabilities that can be used to interpret model predictions in terms of odds ratios.
- **Support Vector Machine Classifier (SVM)**: SVM is a powerful algorithm that works well for both linear and non-linear classification. It aims to find the optimal separating hyperplane that maximizes the margin between different classes, seen on the Figure 4. In cases where the data is linearly inseparable SVM uses kernel functions to project the data into higher-dimensional spaces where a hyperplane can be used to separate the classes. SVMs are effective in high-dimensional spaces and are versatile due to the different kernels that can be applied [4].

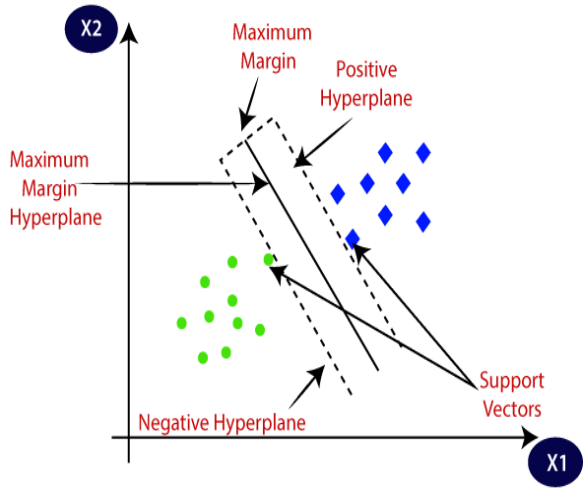

Figure 4 Support Vector Machine.

- **k - Nearest Neighbor Classifier (KNN)**: KNN is an instance-based learning algorithm that classifies instances based on the majority vote learning of the k-nearest neighbors in the feature space. It is a type of a lazy learning, where the generalization of the training data is delayed until a query is made to the system. This model is simple and effective but can become computationally expensive as the size of the data grows, and it may suffer from the curse of dimensionality in high-dimensional spaces.
- **Random Forest Classifier**: Random Forest is an ensemble learning method that builds upon the simplicity of decision trees and enhances their performance. It constructs a multitude of decision trees during training and outputs the mode of the classes (for classification) as the final prediction. By aggregating the predictions of individual trees, random forest reduces the risk of overfitting and is able to capture complex relationships in the data [7].
- **XGBoost Classifier**: XGBoost stands for Extreme Gradient Boosting and is an advanced implementation of gradient-boosted decision trees. It is designed for speed and performance, utilizing both hardware optimization and software algorithmic enhancements. XGBoost provides regularizations to prevent overfitting and is capable of handling missing data. It works by sequentially adding

predictors that correct the predecessors' errors, with each new model being fitted on the residual errors of the last prediction.

Prior to applying these algorithms, the data was processed through feature encoding, feature scaling, and data splitting. Categorical values were encoded to numerical values to facilitate computation, while numerical features were standardized to have a mean of zero and standard deviation of one. The dataset was then split into 80%-20% training and testing set, respectively.

Feature selection was performed using the SelectKBest technique to identify the top 15 features that are most predictive of churn. This step is crucial to improve model efficiency and effectiveness by reducing dimensionality and focusing on the most relevant predictors [8].

III. Experimental Results

A. Evaluation Metrics

The following section will detail the results obtained from applying these methods, including the performance metrics and comparative analysis of the six different machine learning models applied to the task of predicting customer churn. The models were evaluated using a comprehensive set of metrics [9] to provide a thorough understanding of their predictive capabilities. The evaluation metrics include:

- Accuracy: The proportion of true results (both true positives and true negatives) among the total number of cases examined. It is calculated as:

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

- Balanced Accuracy: The average recall obtained on each class. It is used to deal with imbalanced datasets and is calculated as:

$$\text{Balanced Accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2}$$

  where, sensitivity is the "true positive rate" – the percentage of positive cases the model is able to detect; and specificity is the "true negative rate" – the percentage of negative cases the model is able to detect.

- Precision: The ration of true positives to all positive results predicted tby the classifer. It is calcualted as:

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

- F1 Score (F1): The harmonic mean of precision and recall, providing a balance between the two in cases where one may be more imporatnt then the other. It is calculated as:

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

- Recall: The ratio of true positives to the sum of true positives and false negatives. It is calculated as:

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

- ROC-AUC: The area under the receiver operating characteristic curve, a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. It is a value between 0 and 1, where 1 represents a perfect model and 0.5 represents a model with no discriminative ability.

The models were trained and evaluated twice: first using all available features, and then using a subset of the top 15 features selected through the SelectKBest technique. Additionally, hyperparameter tuning was performed on the models (Logistic Regression, Random Forest, and XGBoost) using Random Search Cross-Validation to optimize their parameters.

B. Results with Default Parameters (All Features)

The initial set of experiments was conducted using all the features with default model parameters. The results are summarized in Table 2.

Table 2: Model Performance with Default Parameters (All Features)

Metrics	Decision Tree	Logistic Reg	SVM	KNN	Random Forest	XGBoost
Accuracy	0.7292	0.7789	0.7341	0.6837	0.7860	0.7796
Bal accuracy	0.6586	0.6831	0.5000	0.5244	0.6811	0.6844
Precision	0.4909	0.6067	0.0000	0.3301	0.635688	0.6081
F1	0.4993	0.5351	0.0000	0.2367	0.531882	0.5373
Recall	0.5080	0.4786	0.0000	0.1844	0.4572	0.4812
ROC-AUC	0.6586	0.6831	0.5000	0.5244	0.6811	0.6844

C. Results with Default Parameters (Top 15 Features)

The models were then trained using only the top 15 features, that were selected using the SelectKBest technique. The results are summarized in Table 3.

Table 3: Model Performance with Default Parameters (Top 15 Features)

Metrics	Decision Tree	Logistic Reg	SVM	KNN	Random Forest	XGBoost
Accuracy	0.7086	0.7818	0.7896	0.7512	0.7810	0.7661
Bal accuracy	0.6471	0.6927	0.6835	0.6583	0.6820	0.6727
Precision	0.4573	0.6084	0.6477	0.5375	0.6153	0.5728
F1	0.4849	0.5505	0.5360	0.4956	0.5333	0.5183
Recall	0.5160	0.5026	0.4572	0.4598	0.4705	0.4732
ROC-AUC	0.6471	0.6927	0.6835	0.6583	0.6820	0.6727

D. Results after Hyperparameter Tuning (All Features)

Hyperparameters tuning was performed on Logistic regression, Random Forest, and XGBoost models using all features. The results are summarized in Table 4.

Table 4: Model Performance after Hyperparameter Tuning (All Features)

Metrics	Logistic Reg	Random Forest	XGBoost
Accuracy	0.7867	0.7938	0.7988
Bal accuracy	0.6918	0.6856	0.7095
Precision	0.6267	0.6640	0.6531
F1	0.5495	0.5396	0.5782
Recall	0.4893	0.4545	0.5187
ROC-AUC	0.6918	0.6856	0.7095

The comparison of results seen on Figure 5 indicates that hyperparameter tuning generally improved the performance of the models. Notably XGBoost with hyperparameter tuning achieved the highest accuracy and balanced accuracy among all configurations. The use of only 15 features did not significantly diminish model performance, suggesting that these features capture the most relevant information for predicting churn.

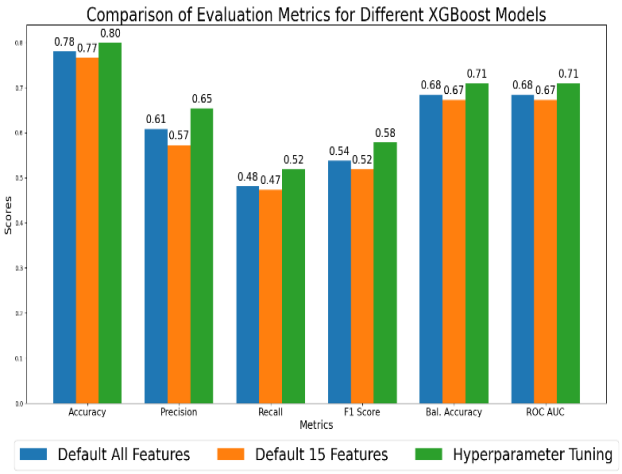

Figure 5: Comparison of the XGBoost model with all metrics.

On the Figure 6, the confusion matrix offers a detailed evaluation of the XGBoost model's performance in predicting customer churn post-hyperparameter tuning. It accurately identifies 930 non-churning and 194 churning customers but misclassifies 103 non-churning as churners and 180 churning as non-churners. While proficient at identifying non-churners, the model struggles with false negatives indicating in recognizing all churning customers. This assessment underscores the model's strengths and areas requiring improvement in discriminating between churning and non-churning customers.

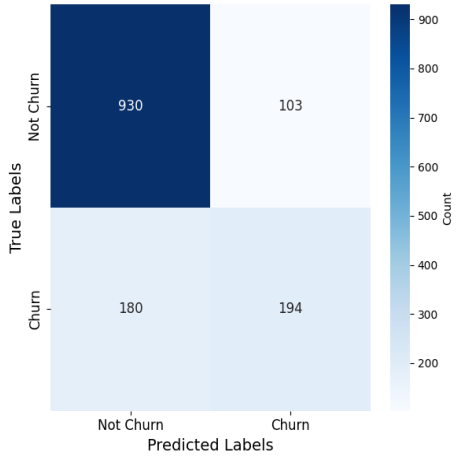

Figure 6: Confusion Matrix XGBoost - Hyperparameter tuned.

IV. CONCLUSION

The objective of this study was to develop a predictive model for customer churn that could assist telecom companies in identifying at-risk customers and formulating effective retention strategies [10]. Through the application of various machine learning algorithms, both individually and in ensemble, we sought to determine the most effective approach for this classification task.

The experimental results revealed several key findings. Initially, when all features were used, the Random Forest and XGBoost classifiers demonstrated superior performance in terms of accuracy and balanced accuracy compared to other models like Decision Tree, Logistic Regression, SVM, and KNN. This outcome can be attributed to the ensemble methods ability to reduce overfitting and capture complex patterns in the data. Upon reducing the features set to the top 15 most significant features, as determined by the SelectKBest technique, there was a negligible decrease in performance for most models. This suggests that these selected features retain the essential information required for churn prediction, and that a more parsimonious model could be just as effective as one using the full features set. Notably, the SVM classifier showed a marked improvement in balanced accuracy with the reduced feature set, indicating that feature selection can have a positive impact on the performance of certain models.

The application of hyperparameter tuning to the Logistic Regression, Random Forest, and XGBoost models further refined their performance, XGBoost benefited from this optimization, achieving the highest accuracy and balanced accuracy across all models and configuration tested. This underscores the importance of hyperparameter tuning in the development of predictive models and its potential to significantly enhance model accuracy.

In conclusion, the study demonstrates that machine learning can be a powerful tool in predicting customer churn. The XGBoost classifier, with hyperparameter tuning, emerged as the most effective model in this study. However, the relatively strong performance of models with only top 15 features indicates that a simpler model could be nearly effective, which has implications for computational efficiency and ease of model interpretation.

Future work could explore the integration of additional data sources, the application of more sophisticated feature selection methods, and the deployment of these models in a real-world setting to validate their predictive power. Additionally, the development of a cost-sensitive model that takes into account the financial implications of false positives and false negatives could provide a more nuanced approach to customer retention strategies.

This study contributes to the growing body of knowledge on the application of machine learning in the telecommunication industry and provides a foundation for further research and practical applications in the area of churn prediction.

REFERENCES

[1] Ribeiro, Hugo, et al. "Determinants of churn in telecommunication services: a systematic literature review." Management Review Quarterly (2023): 1-38.

[2] Kaggle Data Sample: https://www.kaggle.com/datasets/blastchar/telco-customer-churn [Accessed, March 2024]

[3] IBM Original Data:
https://community.ibm.com/community/user/businessanalytics/blogs/steven-macko/2019/07/11/telco-customer-churn-1113 [Accessed, March 2024]

[4] Customer Churn Prediciton Using Machine Learning: Main Approaches and Models: https://www.kdnuggets.com/2019/05/churn-prediction-machine-learning.html [Accessed, May 2024]

[5] Burez, Jonathan, and Dirk Van den Poel. "Handling class imbalance in customer churn prediction." Expert Systems with Applications 36.3 (2009): 4626-4636

[6] Lalwani, Praveen, et al. "Customer churn prediction system: a machine learning approach." Computing 104.2 (2022): 271-294.

[7] Idris, Adnan, Muhammad Rizwan, and Asifullah Khan. "Churn prediction in telecom using Random Forest and PSO based data balancing in combination with various feature selection strategies." Computers & Electrical Engineering 38.6 (2012): 1808-1819.

[8] Huang, Bingquan, Brian Buckley, and T-M. Kechadi. "Multi-objective feature selection by using NSGA-II for customer churn prediction in telecommunications." Expert Systems with Applications 37.5 (2010): 3638-3646.

[9] Understanding Classification Metrics:: https://www.kdnuggets.com/understanding-classification-metrics-your-guide-to-assessing-model-accuracy [Accesssed, May 2024]

[10] Keramati, Abbas, Hajar Ghaneei, and Seyed Mohammad Mirmohammadi. "Developing a prediction model for customer churn from electronic banking services using data mining." Financial Innovation 2 (2016): 1-13.