XGBoost for Customer Churn Prediction: An Iranian Telecom Case Study

DECEMBER 3, 2024
AZRA PARK
[redacted]

1 Executive Summary

Customer churn, the rate at which individuals stop using a service, is a critical issue for subscription-based industries, including telecommunications. This study focuses on predicting customer churn for an Iranian telecom provider, using machine learning to identify which customers are likely to leave and identifying the variables that contribute to this decision.

The research employs a cutting-edge algorithm, XGBoost or eXtreme Gradient Boosting, known for its speed and accuracy in predictive tasks. Using a dataset of over 3,000 customers, the study analyzes behaviours, such as call frequency, text usage, complaints, and engagement duration, to understand their relationship with churn. The goal is to identify the factors that make a customer more likely to churn, enabling targeted actions like personalized incentives or service improvements to enhance retention.

To ensure the reliability of my predictions, I implemented rigourous testing through five-fold cross-validation and extensive hyperparameter tuning. This process involves, systematically divided the data into five groups and validated the model's performance across each subset, confirming consistent accuracy across different customer segments. I then optimized the model by testing hundreds of parameter combinations, fine-tuning aspects like tree depth, learning rate, and sampling methods until I found the precise configuration that maximized prediction accuracy for this dataset.

Key findings reveal that customer inactivity and dissatisfaction, measured by engagement metrics and complaints, are strong predictors of churn. Customers with fewer calls and texts or unresolved complaints are significantly more likely to leave. Conversely, I find that high usage and longer subscriptions correlate with better retention, highlighting the importance of keeping customers engaged with the telecom services.

The XGBoost model demonstrated exceptional performance, achieving an accuracy of nearly 98% and AUC of 99.55%, correctly identifying most at-risk customers. Beyond accuracy, the model provides actionable insights by ranking the most influential factors behind churn. For instance, customer activity status, complaints, price, and service usage emerged as top predictors, giving telecom companies clear areas to focus their retention efforts.

2 Introduction

Customer churn, or the rate at which customers stop using a product or service, is a critical metric in industries that rely on subscription-based business models. Subscription models, commonly used by platforms, streaming services, and telecom companies, are characterized by their recurring revenue streams in which they rely on to sustain profitability, making long-term customer retention a vital element of the subscription-based business model. High churn rates result in significant revenue losses, disruptions in cash flow stability, and increased customer acquisition costs as companies must continually replace lost customers, many of whom have defected to direct competitors.

In subscription-based businesses, the customer lifetime value (CLV) of a customer is directly tied to their time spent as a subscriber, with payments typically being made on a monthly or annual basis. Companies, therefore, focus on reducing churn to maximize CLV and achieve sustainable growth. Understanding the drivers of churn and accurately predicting which customers are likely to leave can help businesses proactively address retention challenges, whether through personalized engagement, targeted incentives, or product improvements.

The telecom industry, built on subscription-based services, highlights the critical need for effective churn management. In this competitive sector, where customer switching costs are often minimal, predicting churn is essential for maintaining a stable and loyal customer base. By identifying at-risk customers in advance, telecom companies can implement targeted interventions,

such as personalized incentives or improved service offerings, to enhance retention and reduce revenue losses.

This study employs the XGBoost, or eXtreme Gradient Boosting, machine learning algorithm to predict customer churn using a dataset from an Iranian telecom provider. To enhance the robustness and performance of my findings, I conduct hyper-parameter tuning using a grid search alongside 5-fold cross-validation, ensuring the model is optimally configured for the given dataset.

The evaluation phase involves a thorough assessment of the model's effectiveness using common classification problem key performance metrics such as AUC-ROC, accuracy, precision, F1 score, and a confusion matrix. Beyond these metrics, an analysis of feature importance is carried out to identify the most influential factors driving customer churn.

This structured methodology ensures both the accuracy of predictions and the interpretability of results, addressing the critical business challenge of customer churn.

3 Literature Review

Research into customer churn prediction in telecommunications has an extensive literature given the immense demand for increasingly accurate and efficient methods to predict and prevent customer attrition.

Keramati and Ardabili (2011), analyzing the same Iranian telecommunications churn dataset used in this research, provided crucial insights by testing multiple hypotheses about factors influencing churn probability using binomial logistic regression. Their work established key relationships between churn and variables including service failures, customer tenure, complaints, credit amounts, and usage patterns. Laying the groundwork for the methodology of this paper, Chen and Guestrin (2016) introduced a breakthrough in tree boosting methods with the advent of eXtreme Gradient Boosting algorithm, more commonly known as XGBoost. This novel method innovated on traditional tree boosting methods by integrated regularization, handling of sparse data, parallelization and much more to achieve an efficient tree-based method for large-scale regression and classification problems in machine learning with impressive accuracy.

Several other significant contributions have shaped the field's development. Almana, Aksoy, and Alzahrani (2014) and Huang, Kechadi, and Buckley (2012) comprehensively analyzed techniques including neural networks, decision trees, k-means clustering, k-nearest neighbour, logistic regression, naive bayes, and support vector machines. Lu, Lin, Lu, and Zhang (2012) demonstrated improved prediction accuracy through AdaBoost combined with logistic regression. Vafeiadis, Diamantaras, Sarigiannidis, and Chatzisavvas (2015) further showed how combining traditional methods with boosting techniques enhanced performance.

With the introduction of the XGBoost algorithm by Chen and Guestrin (2016), its effectiveness quickly gained recognition, leading many researchers to apply it to telecom churn prediction problems. Both Ahmad, Jafar, and Aljoumaa (2019) and Lalwani, Mishra, Chadha, and Sethi (2022) conducted extensive comparisons evaluating multiple algorithms and found XGBoost to be the most effective for churn prediction. While, Fujo, Subramanian, Khder, et al. (2022) found that deep back-propagation neural networks could outperform several techniques, including XGBoost.

This study builds upon these foundations in several ways. While utilizing the same dataset as Keramati and Ardabili (2011), I apply more advanced methods to predict customer churn rather than test hypotheses. Unlike Ahmad et al. (2019), Fujo et al. (2022), and Lalwani et al. (2022) who focused primarily on algorithm comparison, this research develops an optimized approach specifically for the Iranian telecom churn dataset.

4 Data

4.1 Data source

The dataset used for this analysis is the *Iranian Churn* (2020) dataset, sourced from the UC Irvine Machine Learning Repository. This dataset was collected from an Iranian telecom company and contains customer-level data aggregated over a 12-month period.

The dataset comprises 3,150 observations, each representing a unique customer, with 14 variables. These variables capture a wide range of customer behaviours, such as call failures, subscription length, call and SMS usage, and customer complaints. The outcome variable of interest in this dataset is the binary variable *Churn*, which denotes whether a customer had churned or remained an active customer by the end of the observation period.

The dataset's attributes are aggregated over the first 9 months, with the churn labels reflecting customer status at the end of the 12th month. This introduces a three-month gap during which customer behaviour data is not observed. This planning gap reflects a common practice in telecom churn studies, where predictions must be made based on earlier behaviours to allow time for intervention strategies.

4.2 Summary statistics

The summary statistics presented in Table 1 provide an overview of the variables used in the analysis, showcasing their central tendencies and variation alongside supplemental descriptions. As shown in Figure 1, on average, 15.71% of customers churn, demonstrating that a small proportion of customers discontinue their service, while a majority of customers remain active, which is expected.

Customer engagement metrics vary considerably across the sample. For instance, the mean *Seconds of Use* is 4472.46 seconds, with a high standard deviation of 4197.91, capturing the wide variation in call usage among customers. Similarly, *Frequency of Use*, which tracks the number of calls made, averages 69.46 but spans from 0 to 255, indicating diverse calling habits across users. In terms of messaging, the *Frequency of SMS* has a mean of 73.17 but demonstrates significant variation, with a standard deviation of 112.24 and a maximum of 522.

The *Tariff Plan* binary variable, where 1 represents pay-as-you-go customers, highlights that the vast majority (92%) of customers opt for this plan as opposed to fixed contracts. *Subscription Length*, which captures the number of months a customer has been subscribed, has a mean of 32.54 months, with a moderate level of variation shown by the standard deviation of 8.57 months. These metrics suggest a fairly stable customer base with both new and long-term subscribers.

Analyzing customer demographics and behaviour also offers valuable insights. The *Age* of customers has a mean of 31 years, ranging from 15 to 55, reflecting a wide age range catered to by the service. Though, this is somewhat dishonest given *Age* corresponds to *Age Group*, where *Age Group* is an ordinal variable with a mean of 2.83 on a scale from 1 (youngest) to 5 (oldest), representing age groups where: 1 corresponds to 15 years old, 2 to 25 years old, 3 to 30 years old, 4 to 45 years old, and 5 to 55 years old. Despite this, it still holds that most customers fall within the younger to middle-aged range. Furthermore, the *Distinct Called Numbers*, which quantifies the variety of contacts a customer interacts with, averages 23.51, demonstrating active engagement with a diverse set of connections for many users.

Finally, *Customer Value* represents the calculated customer lifetime value reflecting the financial profit derived from a particular customer. *Customer Value* has a mean of 470.97 but has a large variance, as demonstrated in its high standard deviation of 517.02 and a maximum value exceeding 2000. This highlights the heterogeneity in customer profitability, an important factor

for assessing churn risk.

These summary statistics provide a comprehensive view of the dataset, offering critical context for analyzing the drivers of churn. Understanding these metrics is essential for interpreting the relationships between customer behaviour, demographics, and subscription characteristics in predicting churn outcomes.

Table 1: Summary Statistics

Variable	Description	Mean	Std Dev	Min	Median	Max
Call Failure	Number of call failures	7.63	7.26	0.00	6.00	36.00
Complains	One if a complaint was made, else zero	0.08	0.27	0.00	0.00	1.00
Subscription Length	Total months of subscription	32.54	8.57	3.00	35.00	47.00
Charge Amount	Ordinal variable indicating the customer's monthly charge tier	0.94	1.52	0.00	0.00	10.00
Seconds of Use	Total seconds of calls	4472.46	4197.91	0.00	2990.00	17090.00
Frequency of Use	Total number of calls	69.46	57.41	0.00	54.00	255.00
Frequency of SMS	Total number of text messages	73.17	112.24	0.00	21.00	522.00
Distinct Called Numbers	Total number of distinct phone numbers called	23.51	17.22	0.00	21.00	97.00
Age Group	Ordinal variable categorizing the customer's age group	2.83	0.89	1.00	3.00	5.00
Tariff Plan	One if pay-as-you-go, else zero (contract)	0.92	0.27	0.00	1.00	1.00
Status	One if met usage criteria, else zero	0.75	0.43	0.00	1.00	1.00
Age	Age of customer in years	31.00	8.83	15.00	30.00	55.00
Customer Value	Lifetime value of the customer	470.97	517.02	0.00	228.48	2165.28
Churn	One if customer churns, else zero	0.16	0.36	0.00	0.00	1.00

Note: The table summarizes the full sample (N = 3150).

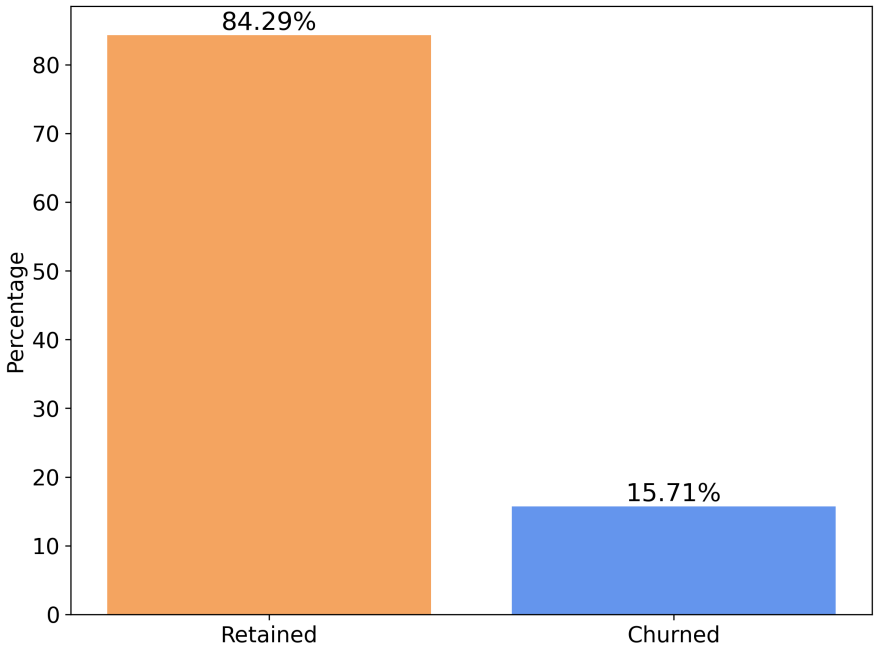

Figure 1: Churn Distribution

The correlation heatmap presented in Figure 2 provides insights into the relationships between the variables, highlighting key associations that are relevant to understanding customer churn dynamics. The variable *Churn*, as the outcome of interest, exhibits its strongest positive correlation with *Complains* (0.53), indicating that customers who file complaints are significantly more likely to churn. This emphasizes the importance of addressing customer dissatisfaction proactively to mitigate the risk of churn.

Other variables show weaker, yet noteworthy, negative correlations with *Churn*. For instance, *Seconds of Use* (-0.30), *Frequency of Use* (-0.30), and *Frequency of SMS* (-0.22) are negatively correlated with churn, suggesting that, higher service engagement levels, as determined by calls or texting, are associated with a reduction in churn. These findings, though somewhat expected, reinforce the importance of customer's engagement with the service in promoting retention.

The *Status* variable, which distinguishes active accounts from inactive ones based on usage metrics, exhibits a strong negative correlation with *Churn* (-0.50). This indicates that inactive customers are significantly more likely to churn, making *Status* a critical predictor in churn modeling. Furthermore, *Status* shows notable correlations with other key variables, including *Seconds of Use* (0.46), *Frequency of Use* (0.45), *Frequency of SMS* (0.30), *Distinct Called Numbers* (0.41), *Charge Amount* (0.36), and *Complaints* (-0.27). These relationships suggest that *Status* serves as a mediator for *Churn*, consistent with findings by Keramati and Ardabili (2011).

Similarly, *Distinct Called Numbers*, which captures the diversity of a customer's interactions, demonstrates a negative correlation with *Churn* (-0.28), suggesting that customers with broader communication patterns are less likely to churn.

Other relationships of interest emerge within the predictor variables themselves. For example, *Seconds of Use*, *Frequency of Use*, and *Frequency of SMS* exhibit strong positive correlations with one another (greater than 0.90), indicating that these variables capture complementary aspects of customer engagement. Similarly, *Age Group* and *Age* are almost perfectly correlated (0.96), reflecting that they are effectively interchangeable for analysis purposes.

Variables such as *Customer Value* (-0.29) also show a moderate negative correlation with churn, indicating that high-value customers are less likely to churn. Meanwhile, *Tariff Plan*, while weakly correlated with *Churn* (0.11), shows a stronger negative correlation with *Customer Value* (-0.25), which may imply indirect effects on churn.

Overall, the heatmap reveals clear patterns that can guide churn analysis. Complaints, engagement metrics, and customer status emerge as primary factors influencing churn, while relationships among the predictor variables suggest opportunities to simplify or optimize modeling approaches. These insights provide a solid foundation for predictive modeling and targeted interventions to reduce churn.

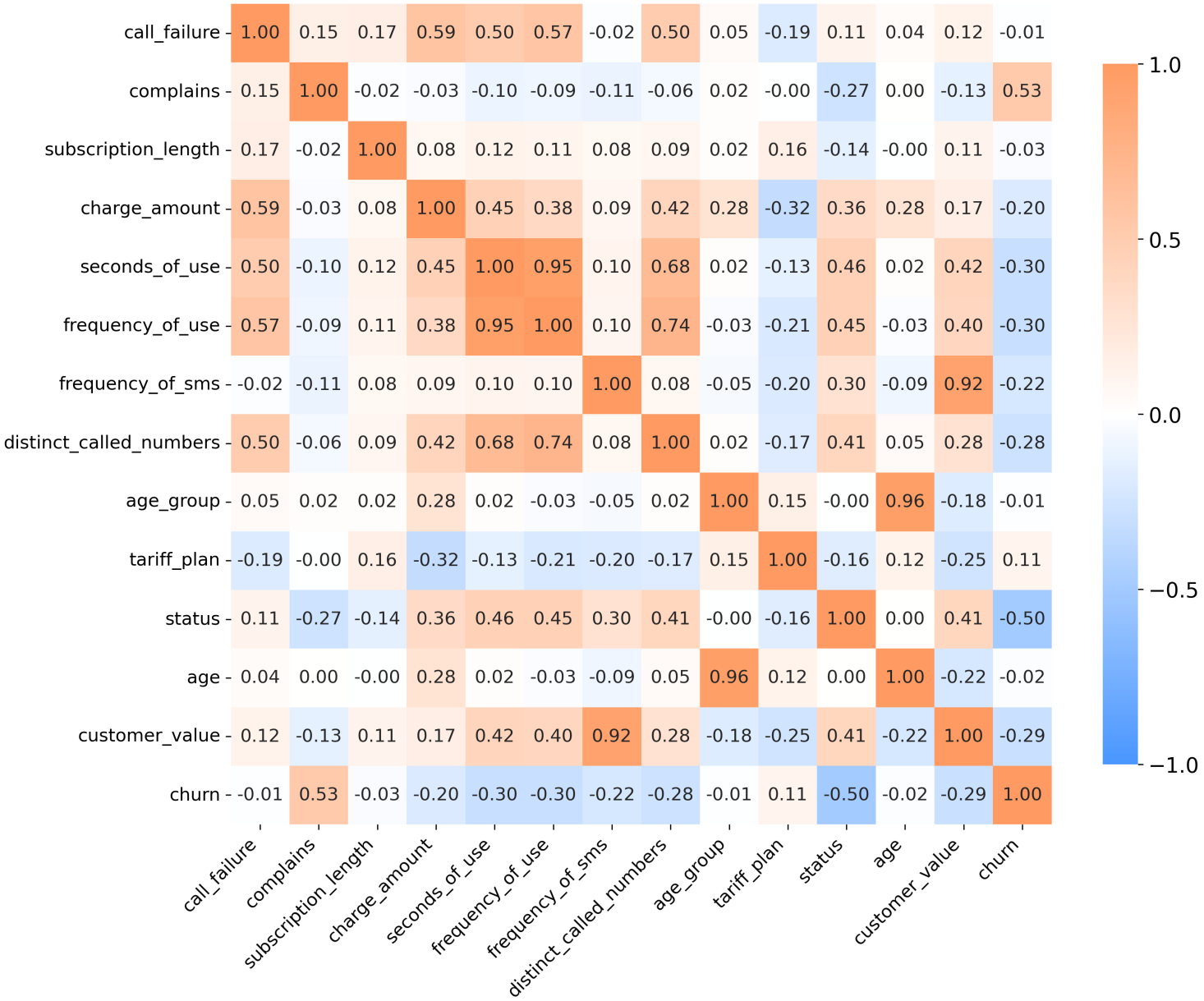

Figure 2: Correlation Heatmap

The visualizations in Figure 3 illustrate the distributions of key variables segmented by *Churn* status, offering valuable insights into the differences between customers who churn and those who remain active. For *Call Failures*, the distribution for churned customers leans slightly towards a higher frequency of failures compared to non-churned customers, suggesting that poor call quality may contribute to dissatisfaction and churn. Similarly, *Complaints* exhibit a marked distinction; churned customers are more likely to have expressed complaints, as evidenced by a higher density near *Complains* $= 1$.

The length of subscription, captured by *Subscription Length*, also reveals notable differences. Non-churned customers tend to have longer subscriptions, while churned customers are distributed more broadly, including shorter durations. This suggests that telecom companies should focus more effort on retaining newer subscribers compared to long-standing customers. Another important variable, *Charge Amount*, shows that churned customers are more prevalent in higher charge tiers, potentially highlighting cost sensitivity as a driver of churn, potentially driving customers to competitors.

Usage metrics such as *Seconds of Use*, *Frequency of Use*, and *Frequency of SMS* further illustrate the behavioural disparities. Non-churned customers consistently demonstrate higher engagement across all three metrics, with distributions skewed towards higher values. These patterns suggest that more engaged customers are less likely to churn, making engagement a critical factor in retention efforts. A similar trend is observed in *Distinct Called Numbers*, where non-churned customers tend to interact with a greater variety of contacts, pointing to the importance of social connectivity in customer retention.

Demographic variables such as *Age Group* and *Age* exhibit subtle variations suggesting that young adults and middle-aged customers may be marginally more likely to churn. For *Tariff Plan*, while the majority of customers in both groups are on pay-as-you-go plans, the churned group appears more concentrated, indicating potential associations with certain plans.

The starkest distinction is observed in *Status*, where churned customers are overwhelmingly represented in the inactive category (*Status* $= 0$). This variable directly correlates with the churn outcome and serves as a clear indicator of disengagement. Lastly, *Customer Value* shows that non-churned customers tend to have higher customer lifetime values. The broader spread and longer tail in the non-churned distribution shows the correlation between higher customer value and retention.

These patterns collectively highlight key drivers of churn, including dissatisfaction, lower engagement, and cost sensitivity. Understanding these relationships is essential for designing targeted interventions to mitigate churn and retain valuable customers.

4.3 Data Limitations

The primary limitation of the dataset lies in the absence of temporal context. All attributes, except for the churn label, represent aggregated data from the first nine months of the year. The churn labels capture the customers' status at the end of 12 months, leaving a three-month planning gap where no specific behavioural or transactional data is available. This lack of temporal granularity complicates efforts to analyze the timing and progression of events leading to churn.

Another limitation somewhat related to the lack of temporal context is that many of the variables in the dataset are binary, limiting the richness of the information provided. For example, the *Complains* attribute is a binary indicator that only signals whether a customer has made any complaints, but does not provide details on how many complaints were made or when they occurred during the nine-month period. Complaints may accumulate over time and be indicative of escalating dissatisfaction, but the binary nature of the variable masks such trends. Similarly, the *Status* variable reflects whether a customer is active or inactive based on usage metrics, but does not capture transitions between these states or how status evolved over time.

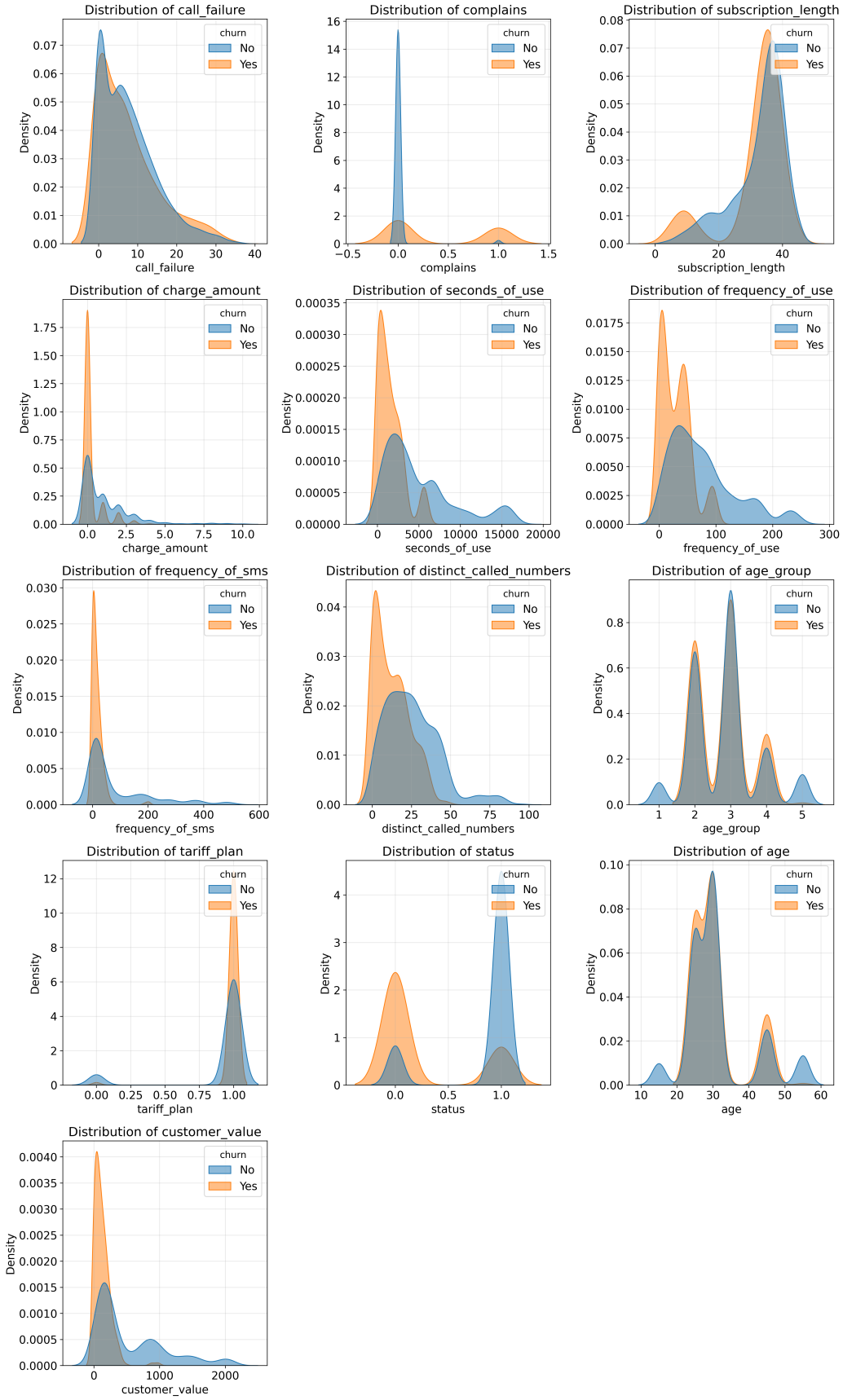

Figure 3: Variable Distributions Based on Churn Classification

5 Methodology

5.1 XGBoost

XGBoost, or eXtreme Gradient Boosting, is a robust and scalable tree-based ensemble learning method designed for classification and regression problems. XGBoost builds models sequentially, with each tree correcting the errors of the previous ones while incorporating regularization techniques to prevent overfitting. For the churn prediction task, XGBoost optimizes a binary cross-entropy loss function to handle the binary nature of the target variable *Churn*.

The binary cross-entropy loss, denoted as $\ell(y, \hat{y})$, is computed as:

$$\ell(y, \hat{y}) = -\frac{1}{n} \sum_{i=1}^{n} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$$

Here, y_i represents the true labels, $\hat{y}_i$ represents the predicted probabilities, and n is the total number of samples.
The overall objective function for XGBoost consists of two components: the loss function and a regularization term:

$$\mathcal{L} = \sum_{i=1}^{n} \ell(y_i, \hat{y}_i) + \sum_{t=1}^{T} \Omega(f_t)$$

The regularization term, $\Omega(f_t)$, penalizes the complexity of the trees to reduce overfitting:

$$\Omega(f_t) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} w_j^2$$

where T represents the number of leaves in a tree, w_j is the weight of each leaf, γ controls the penalty for the number of leaves, and λ regularizes the leaf weights.
XGBoost generates predictions iteratively by adding the output of new trees to previous predictions, scaled by a learning rate η:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta f_t(X)$$

The final prediction is computed by summing the outputs of all trees and applying a sigmoid function to convert these values into probabilities:

$$\hat{y}_i = \sigma \left(\sum_{t=1}^{T} f_t(X) \right), \quad \sigma(z) = \frac{1}{1 + e^{-z}}$$

To address class imbalance, XGBoost incorporates the `scale_pos_weight` parameter, defined as:

$$\texttt{scale\_pos\_weight} = \frac{\text{Number of non-churn instances}}{\text{Number of churn instances}}$$

This parameter assigns a higher weight to the minority class (churn instances), improving the model's ability to handle imbalanced datasets effectively.

Additionally, XGBoost optimizes the objective function using a second-order Taylor approximation:

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t)$$

where g_i and h_i are the first- and second-order gradients of the loss function, enabling efficient tree construction.

This methodology, combined with regularization and gradient-based optimization, makes XGBoost a powerful model for predicting churn. Its ability to handle missing data, manage imbalanced datasets, and capture complex patterns in high-dimensional data sets it apart as a suitable choice for this analysis.

5.2 Cross-Validation

Cross-validation is an essential component of the modeling framework in this study, ensuring the robustness and generalizability of the churn prediction model. Specifically, a k-fold cross-validation approach was utilized during hyperparameter tuning to evaluate the model's performance across multiple data subsets and to minimize the risk of overfitting. In this approach, 90% of the dataset (the training set) is divided into k equal-sized folds, and the model is trained on k-1 folds while validated on the remaining fold. This process is repeated k times, allowing each fold to serve as the validation set exactly once. For this study, I implemented a 5-fold cross-validation.

During the cross-validation process, the model's performance was assessed using the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) as the primary metric. For each combination of hyperparameters, the mean AUC score across the five folds was calculated, with the configuration achieving the highest mean score selected as the optimal set of parameters. This iterative process ensures that the model is not only trained effectively, but also validated against a diverse range of distributions, promoting consistent and reliable performance across different subsets of the data.

The use of cross-validation is particularly crucial in this study due to the characteristics of the churn dataset, which exhibit imbalanced classes, where the majority of customers do not churn. Cross-validation helps expose the model to representative distributions of both churn and non-churn cases across all folds, ensuring balanced training and validation. Additionally, it provides a realistic estimate of the model's ability to generalize to unseen data, a critical consideration for real-world deployment. By integrating cross-validation with hyperparameter tuning, I ensure that the XGBoost model is not only optimized for the training dataset but is also robust in its ability to maintain high predictive accuracy on new data.

$$\text{Mean AUC} = \frac{1}{K} \sum_{k=1}^{K} \text{AUC}_k$$

Where:

- K: Number of cross-validation folds
- AUC_k: AUC score for the k-th fold, where AUC is defined below

5.3 Evaluation Criteria

The performance of the XGBoost model is evaluated using our held-out test set comprising 10% of the dataset using several metrics to ensure a comprehensive assessment of its predictive capabilities. The evaluation criteria are detailed as follows:

Accuracy measures the proportion of correct predictions among all predictions made by the model:

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

Where:

- TP: True Positives (correctly predicted churned customers)
- TN: True Negatives (correctly predicted non-churned customers)
- FP: False Positives (non-churned customers incorrectly predicted to churn)
- FN: False Negatives (churned customers incorrectly predicted as non-churned)

Precision quantifies the proportion of positive predictions that are actually correct:

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

Where:

- TP: True Positives
- FP: False Positives

The **True Negative Rate**, also known as specificity, measures the proportion of correctly identified negatives:

$$\text{TNR} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

The **False Positive Rate** captures the proportion of negatives that are incorrectly classified as positives:

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

Where:

- TN: True Negatives
- FP: False Positives

The **False Negative Rate** measures the proportion of positives that are incorrectly classified as negatives:

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

The **True Positive Rate**, also known as recall or sensitivity, measures the proportion of actual positives that are correctly classified:

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

Where:

- FN: False Negatives
- TP: True Positives

The **F1 Score** combines precision and recall into a single metric by taking their harmonic mean:

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

The **Area Under the ROC Curve** (**AUC**) measures the area under the Receiver Operating Characteristic (ROC) curve, which plots the True Positive Rate (TPR) against the False Positive Rate (FPR):

$$\text{AUC} = \int_0^1 \text{TPR}(FPR)\, d(FPR)$$

Where:

- TPR: True Positive Rate
- FPR: False Positive Rate

These evaluation metrics collectively ensure a balanced and thorough assessment of the model's performance.

6 Empirical Results

6.1 Results

The results presented in Table 2 summarize the performance of the XGBoost model in predicting customer churn. The model achieved an exceptional AUC-ROC score of 0.9955, as illustrated in Figure 4, which indicates that the model is highly effective in distinguishing between churned and non-churned customers across various classification thresholds. The nearly perfect AUC value reflects the robustness of the model in handling the class imbalance inherent in the dataset.

Aligning with the high AUC-ROC score, the model achieved an accuracy of 97.78%, and a precision of 90.57%. The F1 score, a harmonic mean of precision and recall, was 93.20%, highlighting the modelâs balanced performance in identifying churned customers while minimizing false positives. The True Positive Rate (TPR) and False Negative Rate (FNR) were 96.00% and 4.00%, respectively, with the True Negative Rate (TNR) and False Positive Rate (FPR) being 98.11% and 1.89%, respectively, exemplifying the model's robust performance in identifying both customer groups.

Figure 5 provides a detailed view of the confusion matrix, highlighting the model's predictive breakdown. The model correctly classified 260 true negatives and 48 true positives while making only 5 false positive and 2 false negative predictions. This breakdown demonstrates the model's effectiveness in both classes, with particular success in minimizing errors for the minority churn class.

Figure 6 showcases the variable importance plot, which reveals the most influential features driving the model's predictions. The variable *Status*, capturing whether a customer is active or inactive based on usage metrics, emerged as the most significant predictor of churn, followed by *Complains*, indicating whether a customer has filed a complaint. Variables such as

Charge Amount, *Frequency of Use*, and *Seconds of Use* also played notable roles, reflecting the importance of both customer behaviour and dissatisfaction in driving churn outcomes.

Overall, the results highlight the exceptional predictive capabilities of the XGBoost model. The combination of strong performance metrics, a well-balanced confusion matrix, and meaningful feature importance insights confirms the robustness and practical utility of the model for churn prediction. These findings provide a solid foundation for implementing targeted interventions to retain at-risk customers and improve overall satisfaction.

Table 2: Model Performance Metrics

Metric	**Value**
AUC-ROC	0.9955
F1 Score	0.9320
Precision	0.9057
Accuracy	0.9778
False Positive Rate	0.0189
False Negative Rate	0.0400
True Positive Rate	0.9600
True Negative Rate	0.9811

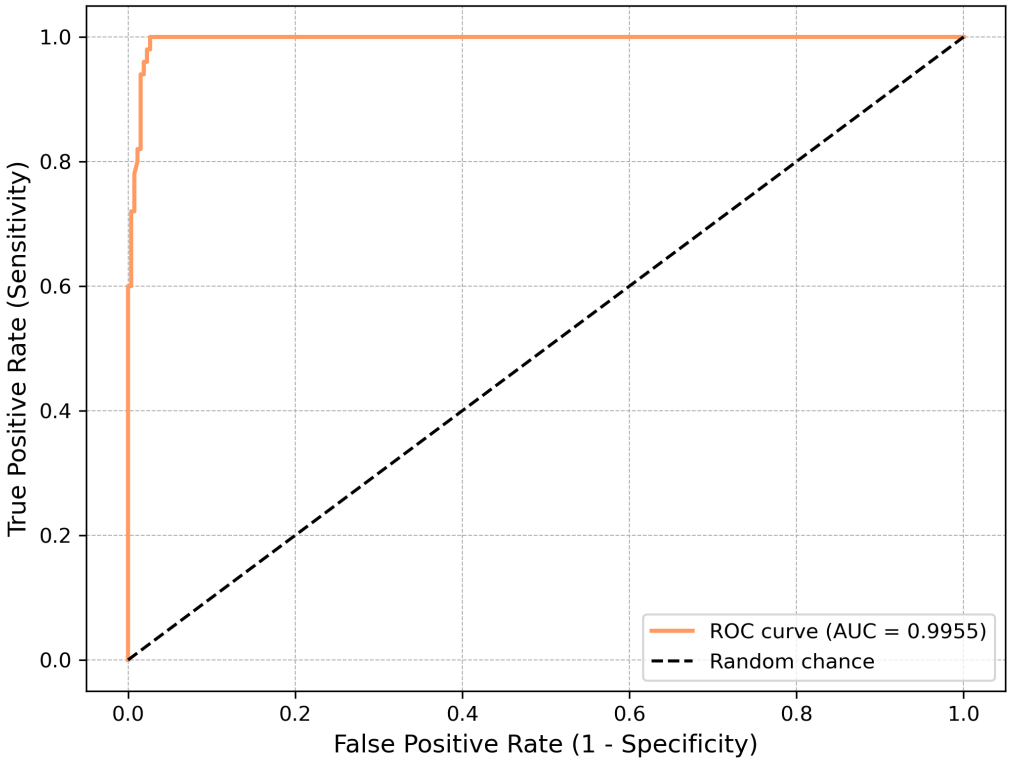

Figure 4: ROC Curve

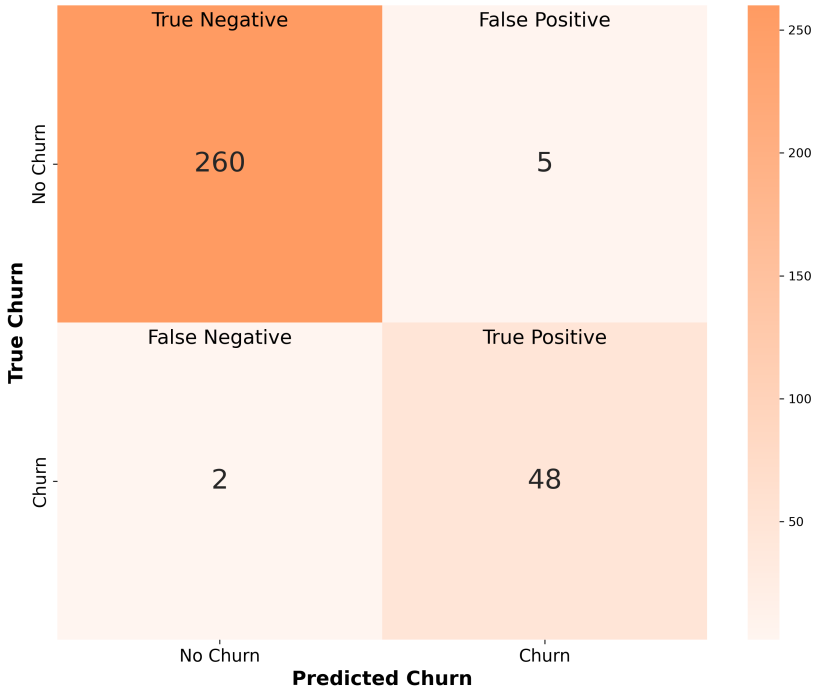

Figure 5: Confusion Matrix

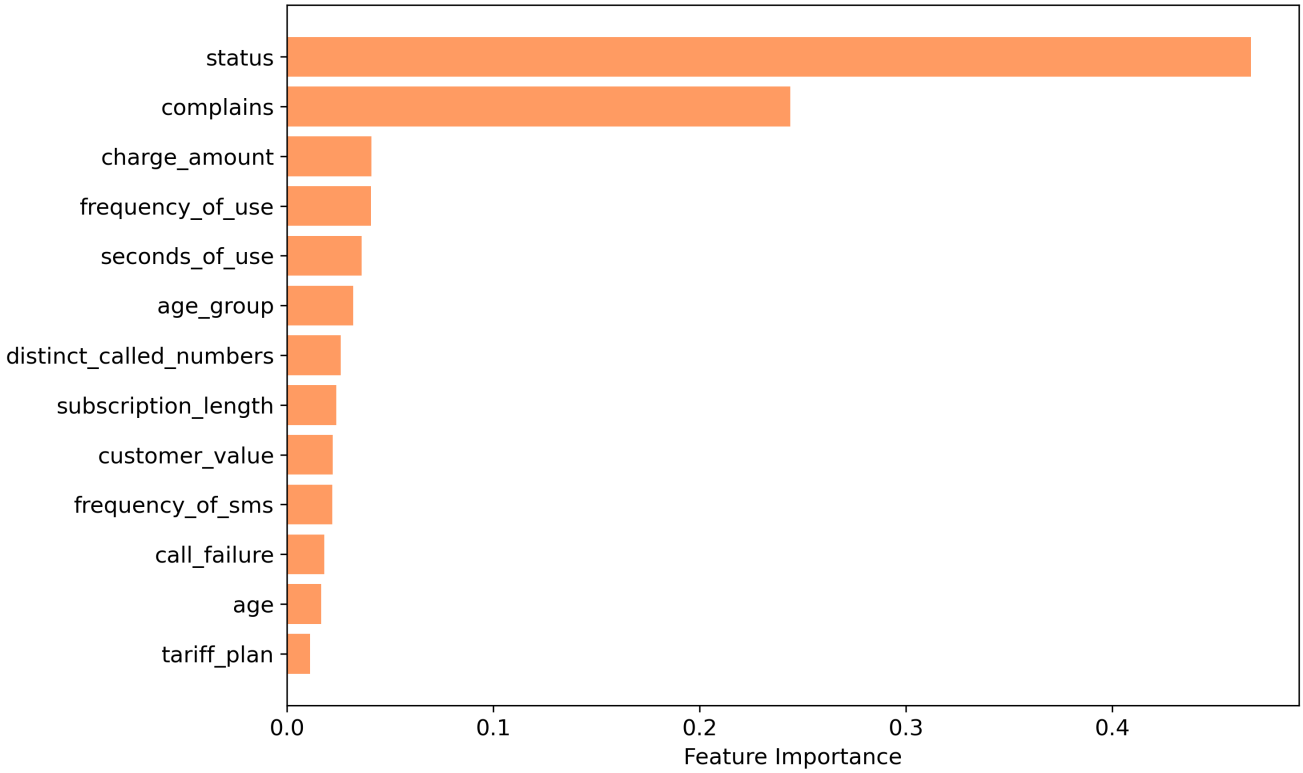

Figure 6: Top 15 Variables Influencing Customer Churn

7 Conclusion

This study provides an in-depth exploration of customer churn prediction using an Iranian telecom dataset using the XGBoost algorithm. By leveraging robust machine learning techniques, the research achieved highly accurate predictions of customer churn, as evidenced by an AUC-ROC score of 0.9955 and other performance metrics, including precision (90.57%), accuracy (97.78%), recall (96.00%), and an F1 score (93.20%). These results highlight the model's ability to effectively distinguish between churned and non-churned customers, offering practical implications for targeted retention strategies.

The analysis emphasized the importance of key features such as customer status, complaints, and behavioural metrics like frequency and duration of service use. These insights align with previous literature and underline the critical role of customer engagement and satisfaction in reducing churn. The findings also demonstrated the value of employing advanced ensemble learning methods like XGBoost to address the challenges of imbalanced datasets and complex feature interactions.

While this study highlights the efficacy of XGBoost for customer churn prediction, it also uncovers avenues for further exploration. With a customer's *Status* being the greatest indicator of *Churn* a causal analysis into what drives customer's usage of telecom services would be a valuable extension, potentially uncovering actionable insights for retention strategies. However, this dataset is fairly limited in this capacity, though this should not limit telecom companies utilizing their more comprehensive proprietary customer data to conduct deeper investigations.

This study's findings reaffirm the transformative potential of machine learning for strategic decision-making in customer retention. By identifying at-risk customers with exceptional precision, this methodology equips telecom providers to design proactive interventions that not only reduce churn but also foster long-term customer loyalty. This research underscores the critical role of data-driven approaches in navigating the challenges of a highly competitive subscription-based industry, offering a foundation for both immediate application and future innovation.

References

Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. *Journal of Big Data*, *6*(1), 1–24.

Almana, A. M., Aksoy, M. S., & Alzahrani, R. (2014). A survey on data mining techniques in customer churn analysis for telecom industry. *International Journal of Engineering Research and Applications*, *4*(5), 165–171.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794).

Fujo, S. W., Subramanian, S., Khder, M. A., et al. (2022). Customer churn prediction in telecommunication industry using deep learning. *Information Sciences Letters*, *11*(1), 24.

Huang, B., Kechadi, M. T., & Buckley, B. (2012). Customer churn prediction in telecommunications. *Expert Systems with Applications*, *39*(1), 1414–1425.

Iranian Churn. (2020). UCI Machine Learning Repository. (DOI: https://doi.org/10.24432/C5JW3Z)

Keramati, A., & Ardabili, S. M. (2011). Churn analysis for an iranian mobile operator. *Telecommunications Policy*, *35*(4), 344–356.

Lalwani, P., Mishra, M. K., Chadha, J. S., & Sethi, P. (2022). Customer churn prediction system: a machine learning approach. *Computing*, *104*(2), 271–294.

Lu, N., Lin, H., Lu, J., & Zhang, G. (2012). A customer churn prediction model in telecom industry using boosting. *IEEE Transactions on Industrial Informatics*, *10*(2), 1659–1665.

Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., & Chatzisavvas, K. C. (2015). A comparison of machine learning techniques for customer churn prediction. *Simulation Modelling Practice and Theory*, *55*, 1–9.