Customer Churn Prediction in Telecommunication Sector

Ilia Bukin

February 2024

Abstract

The main goal of this study is to analyze the effectiveness of classification algorithms for binary churn prediction in the telecommunication sector. This paper details the process of preparing the data for analysis and examines the training and assessment of four different machine learning models: Logistic Regression, K-Nearest Neighbors, Decision Trees, and Support Vector Machines. This research outlines the impact of imbalanced data on model accuracy and offers suggestions for future research directions and improvements. These findings offer valuable insights for telecommunication providers seeking to leverage machine learning to improve their customer retention efforts, ultimately contributing to the broader field of customer relationship management in the digital age.

1 Introduction

The market for telecommunications services like home internet broadband and landline phones has one of the highest industry churn rates, with over 20% of US customers leaving their current provider in 2020, according to Statista[1]. Customer churn, also known as attrition, occurs when a customer ends their subscription [2]. This poses a major problem for telecommunications businesses, leading to revenue losses, diminished market share, and increased marketing expenses associated with attracting new customers.

Telecommunications providers generate a lot of data about their customers, encompassing demographics, usage patterns, billing history, service preferences, and more. Using this information, machine learning models can be trained to anticipate behaviours, enabling more efficient targeting of retention strategies on customers with high churn risk. This paper investigates the feasibility of using base machine learning models to predict customer churn in telecommunications. The goal is to identify customers likely to end their services, allowing for targeted retention efforts that improve customer satisfaction and loyalty.

2 Literature review

Extensive literature exists on the application of machine learning techniques for churn prediction in a wide range of industries, including telecommunications.

The first source for the literature review is "Customer Churn Prediction in Telecommunications" by Bingquan Huang [3]. This paper tests seven predictive modelling techniques, including but not limited to Logistic Regression, Decision Trees, and Support Vector Machines. It also stipulates the importance of features that enhance predictive accuracy, such as customer demographics, bills and payments, and possible complaints about the service. This study highlights the importance of appropriate feature selection and demonstrates the effectiveness of selected modelling techniques in the telecommunications sector.

Similarly, "A comparison of machine learning techniques for customer churn prediction" [4] provides a comparative analysis of the performance of Support Vector Machines, Decision Trees and Logistic Regression, both with and without the application of boosting. The results reveal a clear superiority of the boosted models over their non-boosted counterparts, with the boosted SVM model with a polynomial kernel achieving the highest accuracy of nearly 97%. This work underscores the effectiveness of machine learning models for churn prediction.

The "Handling class imbalance in customer churn prediction" by Burez et al[5] explores approaches for handling the class imbalance in churn datasets. This is a prevalent issue, as the number of churned customers tends to be relatively small compared to un-churned ones. This increases the risk of overfitting the model on the majority class and introduces significant bias. The paper discusses more appropriate evaluation metrics and sapling techniques to mitigate this. The study finds

that under-sampling of the majority class is preferable over regular over-sampling, and it introduces a more advanced sampling method such as SMOTE.

3 Data management

3.1 *Datasource*

The dataset chosen for this analysis is called "Telco Customer Churn," and it was sourced from Kaggle[6]. IBM originally created this synthetic dataset to develop focused customer retention programs in the telecommunication sector. The scarcity of publicly available real churn datasets primarily stems from ethical, legal, and commercial restrictions.

3.2 *Feature descriptions*

Column Name	Data Type	Description
customerID	object	Customer ID
gender	object	Whether the customer is male or female
SeniorCitizen	int64	Whether the customer is a senior citizen or not (1: Yes, 0: No)
Partner	object	Whether the customer has a partner or not (Yes, No)
Dependents	object	Whether the customer has dependents or not (Yes, No)
tenure	int64	Number of months the customer has stayed with the company
PhoneService	object	Whether the customer has phone service or not (Yes, No)
MultipleLines	object	Whether the customer has multiple lines (Yes, No, No phone service)
InternetService	object	Customer's internet service type (DSL, Fiber optic, No)
OnlineSecurity	object	Whether the customer has online security (Yes, No, No internet service)
OnlineBackup	object	Whether the customer has online backup (Yes, No, No internet service)
DeviceProtection	object	Whether the customer has device protection (Yes, No, No internet service)
TechSupport	object	Whether the customer has tech support (Yes, No, No internet service)
StreamingTV	object	Whether the customer has streaming TV (Yes, No, No internet service)
StreamingMovies	object	Whether the customer has streaming movies (Yes, No, No internet service)
Contract	object	The contract term of the customer (Month-to-month, One year, Two year)
PaperlessBilling	object	Whether the customer has paperless billing or not (Yes, No)
PaymentMethod	object	The customer's payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))
MonthlyCharges	float64	The amount charged to the customer monthly
TotalCharges	object	The total amount charged to the customer
Churn	object	Whether the customer churned or not (Yes or No)

Figure 1: Features

The "Telco Customer Churn" dataset comprises 21 columns detailing various aspects of customer accounts and their churn status labels and contains 7043 unique customer entities (Fig.1). These features include demographic information, account details, service usage, and billing information, which are instrumental in understanding the factors that may lead to customer churn.

3.3 *Data Pre-processing*

The Telco Customer Churn dataset underwent several pre-processing steps:

- The *TotalCharges* column was converted to float from its original object type using `.to_numeric` function, and 11 missing values were removed using `.dropna` function.
- The dataset has 22 duplicated rows, which were removed using `.drop_duplicates(inplace=True)`.
- The column *CustomerID* contains an equal number of unique values to the number of rows and was dropped because it is irrelevant for training.
- The values in the *SeniorCitizen* column were converted to Yes/No for the sake of consistency using the `.map()` command.
- Some features in the dataset contain redundant values such as *'No phone service', 'No internet service'*. They were replaced with *No* label using `.replace()`.

4 Data Exploration

Synthetic datasets may exhibit a lack of diversity in feature values and generalizations to actual customer proportions, which can lead to bias amplification in the models. Given this fact, detailed data exploration is was conducted to ensure the dataset's validity and reliability.

4.1 *Continuous features*

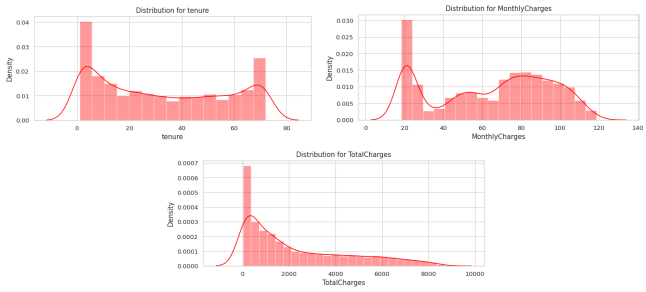

Figure 2: Distributions of Continuous Features

The continuous variables present in this dataset are *Tenure, MonthlyCharges* and *TotalCharges*. Tenure has a bimodal distribution with two peaks at the 0 and 70-month mark, indicating two prominent groups: new and long-time customers. Monthly charges also depict multi-modal distribution with a peak at the beginning of a spectrum. This suggests a high number of customers on basic contracts with fewer charges. This distribution of total charges is skewed to the left with a peak close to 0, indicating that most customers have very low total charges, which correlates with a low peak in tenure for new customers.

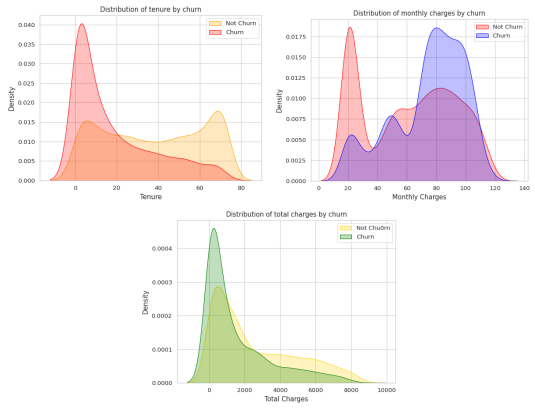

Figure 3: Distributions of Continuous features

Plotting Kernel Density Estimate (KDE) plots provide additional insights into patterns that might differentiate customers who leave from those who stay. The tenure distribution indicates that new customers are more prone to churn. The distribution for monthly charges reveals that customers with higher monthly fees are also more likely to discontinue their services. As for total charges, the distribution is broader, yet there is a noticeable peak at the lower end, indicating that customers who have accrued fewer total charges, typically newer customers, tend to churn at a higher rate.

4.2 *Discrete features*

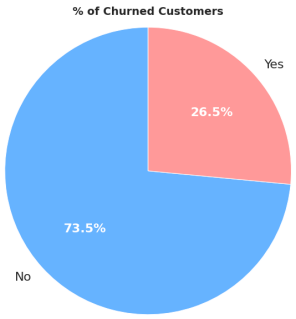

Figure 4: Percentage of Churned Customers.

To show distributions of categorical features `.value_counts(column, normalize=True)` command was used. It was discovered that churn distribution in Fig. 4 exhibits a significant class imbalance with a majority of 73.5% of customers not having churned, compared to a minority of 26.5% who have. This raises concerns about bias when training models and will be addressed further.

The continuous variables in this dataset can be generally classified into the following groups: customer demographics, contracts, service usage, and payment methods. A bi-variate analysis was conducted to assess how these categories of features impact churn.

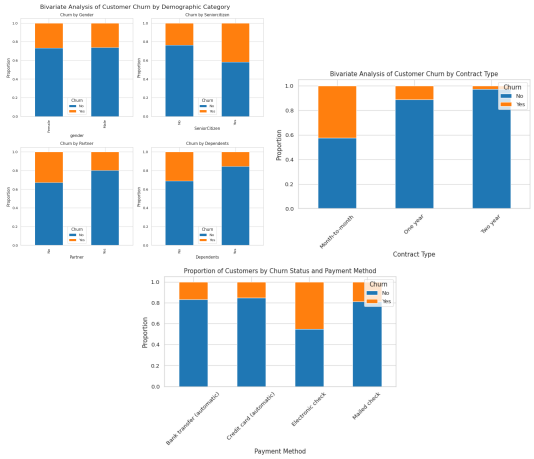

Figure 5: Bi-variate analysis

The gender influence on churn was found to be insignificant as their distributions relative to churn are fairly equal. Senior citizens have a distinctly higher churn rate. Customers with partners and dependents exhibited lower overall churn rates. There is a substantially higher churn rate among customers with month-to-month contracts than those on one and two-year contracts. With regard to payment methods, customers using electronic checks have the highest churn rate.

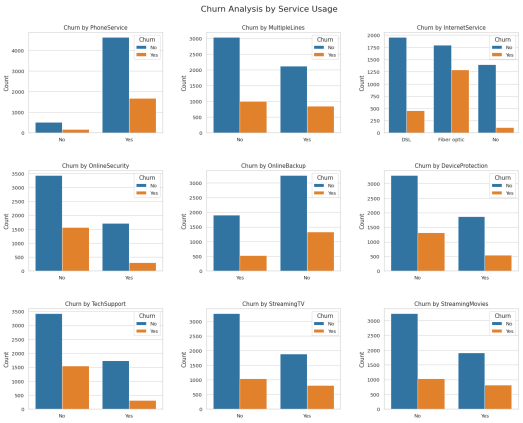

Figure 6: Bi-variate analysis of Services

Customers with phone service have a higher churn count compared to those without. DSL Internet users show a lower churn rate than fibre optic users, and customers without internet service have the lowest churn.

For *Online Security, Online Backup*, and *Device Protection*, customers without these services tend to churn more than those with them. The availability of tech support seems to play a significant role in customer retention, as seen by the higher churn rate among those who do not have tech support. Customers who do not use *Streaming TV* and *Streaming Movies* services have a higher churn rate than those who do. Customers who use electronic checks have the highest churn rate among all payment methods.

5 Preparation for Modelling

5.1 Data Scaling

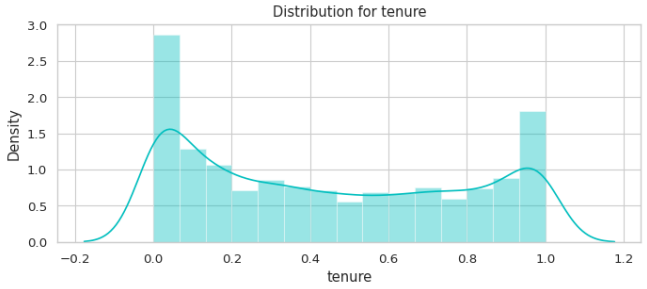

Figure 7: Scaled Tenure

The continuous features in the dataset are spread across different ranges, which can introduce bias to a model because higher-scaled values are treated as having greater significance. Therefore, `.MinMaxScaler()` was used to fit numeric values to a consistent range between 0 and 1 without distorting the original range (Fig.7).

5.2 Encoding categorical features

Two primary encoding strategies are used to prepare categorical data for machine learning, which requires numerical input: Label Encoding and One-Hot Encoding. Label Encoding assigns each category of a feature a unique integer, but this can unintentionally introduce a false sense of order in features with more than two categories, potentially misleading the model. One-Hot Encoding circumvents this by creating distinct binary columns for each category, ensuring no artificial ordinal relationship is implied. However, it greatly expands the number of features in the dataset. To find balance, Label Encoding is applied to binary and One-Hot Encoding to multiple-category features. This approach has expanded the feature space from 20 to 27 features, providing the required format for modelling.

6 Feature Selection

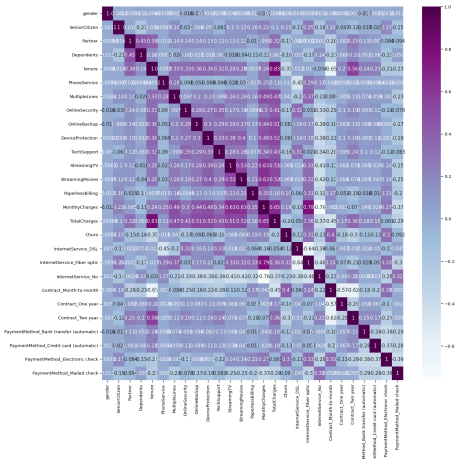

Figure 8: Correlation Matrix

Features like *Tenure* and *TotalCharges* have a high correlation coefficient of 0.83, and *MonthlyCharges* and Fibre optic internet have 0.79 indicating a strong linear relationship. However, without further analysis, it is not worth removing them.

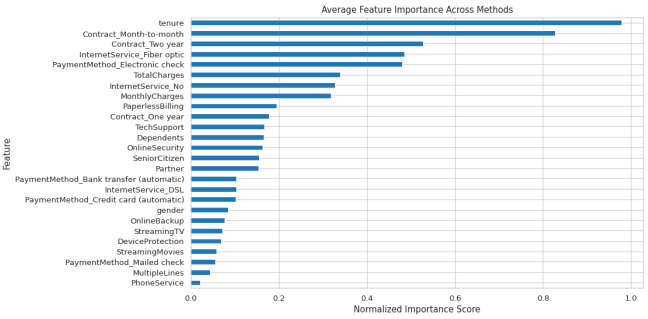

Figure 9: Average Feature Importance

A shortlisting process was conducted to narrow down a set of features and remove complexity. *Chi-squared* and *Mutual Information (MI)* scores were utilised for discrete columns, and for continuous features, *ANOVA F-test* and MI scores were applied. Lastly, the whole dataset was analysed using the *Extra Tree Classifier*. The results of all tests were normalized using the `.MinMaxScaler()` and then averaged, presenting the final score (Fig.10). Features scoring below a 0.15 thresh-

old were considered insufficiently critical and subsequently eliminated, reducing the dataset to 16 features.

```
((7010, 27),
 (7010, 16),
 Index(['DeviceProtection', 'InternetService_DSL', 'MultipleLines',
        'OnlineBackup', 'PaymentMethod_Bank transfer (automatic)',
        'PaymentMethod_Credit card (automatic)', 'PaymentMethod_Mailed check',
        'PhoneService', 'StreamingMovies', 'StreamingTV', 'gender'],
       dtype='object'))
```

Figure 10: Removed Features.

7 Data Analytics Methods

Customer churn prediction represents a binary classification problem (1 for churned, 0 for not churned). Four models renowned for their efficacy in binary classification tasks were selected: Logistic Regression, K-nearest Neighbours (KNN), Decision Trees, and Support Vector Machines (SVMs).

7.1 Logistic Regression

Logistic regression is a popular method for binary classification, often used as a baseline for more complex models. The model estimates the likelihood of features belonging to a class by using the sigmoid function, which maps any real-valued number to a value between 0 and 1, as shown in the equation below:

$$\sigma(z) = \frac{1}{1+e^{-z}} \tag{1}$$

where z represents the linear combination of features.

The model then assigns a class to a given instance using a decision threshold:[7]

$$\text{if } p \geq 0.5, \text{ then class} = 1$$

$$\text{if } p < 0.5, \text{ then class} = 0$$

7.2 K-Nearest Neighbors

K-Nearest Neighbors (KNN) is a non-parametric, lazy learning algorithm that can be used for classification and regression. In classification tasks, KNN assigns a class to a new instance based on the values of its nearest neighbours. It is an easy algorithm to overfit as too low a K captures lots of noise in the training data. It also suffers from the ”curse of dimensionality”, so its performance degrades with larger datasets.[8]

7.3 Decision Trees

Decision Trees are a non-linear predictive modelling tool that can be used for both classification and regression tasks. In the context of churn prediction, it splits the data into subsets based on the values of input features. Each internal node of the tree represents a ”test” on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label. Decision Trees are intuitive and easy to interpret, but they can also be prone to overfitting, especially with complex trees. Setting hyperparameters like pruning and setting a maximum depth can mitigate this issue.

7.4 Support Vector Machines

Support Vector Machines are a set of supervised learning methods used for classification, regression, and outliers detection. SVMs aim to find the hyperplane that best separates the classes in the feature space. The best separation is achieved by the hyperplane that has the largest distance to the nearest training data points of any class (margin), ensuring that the classes are as far apart as possible. SVMs are effective in high-dimensional spaces. However, SVMs require careful tuning of parameters and can be computationally intensive for large datasets.

8 Results and Evaluations

Each model was trained on 80% of the dataset and evaluated on the remaining 20% using train_test_split with `stratify=y` parameter to ensure the model is exposed to the same proportion of values as in the full dataset. The model was then tested on the unseen 20% and metrics gathered. To evaluate further, StratifiedKFold validation was performed by splitting the entire dataset into 5 distinct parts while preserving the proportion of target variable classes in each of them. Then, the mean accuracy across all folds was calculated. Preserving the distribution of target variables is necessary to capture accurate model performance on unbalanced datasets[9].

	Model	Accuracy	Precision	Recall	F1 Score	CV Accuracy
0	LogisticRegression	0.804565	0.660066	0.539084	0.593472	0.800999
3	SVC	0.798859	0.656140	0.504043	0.570122	0.791155
1	KNeighborsClassifier	0.773894	0.582822	0.512129	0.545194	0.778174
2	DecisionTreeClassifier	0.722539	0.477387	0.512129	0.494148	0.721969

Figure 11: Testing and Evaluation Results

The best-performing model for the churn prediction task was Logistic Regression.

Accuracy compute quantifies the overall correctness of the model considering both true positives and true negatives among the total number of instances with Logistic Regression achieving 80%. In cases with class imbalance like churn, the model can still get a high score without truly capturing the minority class. Therefore, metrics like Precision and Recall take precedence.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{2}$$

Precision measures the percentage of correctly predicted positives. High precision is critical in scenarios where the consequences of false positives are substantial. The Logistic Regression model achieved a precision rate of 66%, signifying that when it predicts churn, it is correct about two-thirds of the time.

$$\text{Precision} = \frac{TP}{TP+FP} \tag{3}$$

Recall (sensitivity) assesses the model's ability to identify all relevant instances of churn. Specifically, Logistic Regression was able to detect churn accurately in 53% of the actual churn cases.

$$\text{Recall} = \frac{TP}{TP+FN} \tag{4}$$

The F1 score is the harmonic mean of precision and recall. In this case, it is 59%, which is to be expected.

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

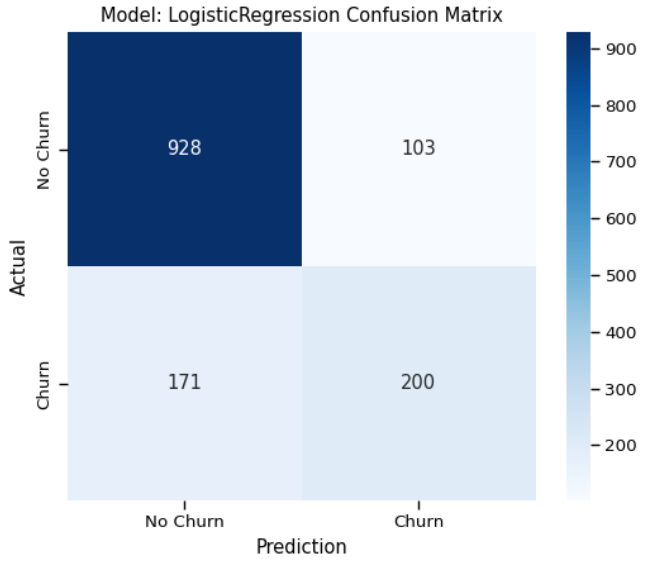

Figure 12: Logistics Regression Conf. Matrix

The second-best model was the Support Vector Machine Classifier (SVC), with metrics only marginally lower than Logistic Regression.

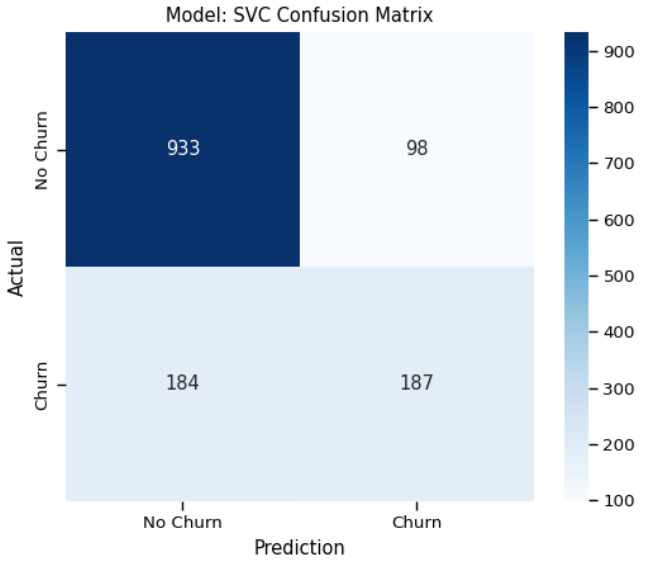

Figure 13: SVC Confusion Matrix

Drawing from the comparison in the confusion matrices, we can identify Logistic Regression as being able to perform slightly better at correctly predicting churn, with a higher number of True Positives (200 vs 187) and fewer False Negatives (171 vs 184). However, it also exhibits a higher number of False Positives (103 vs 98), which means it more often incorrectly predicted customers would churn when they did not (fig. 12,fig.13).

Correctly identifying customers who are likely to leave and minimizing the risk of not identifying them is more important than reducing false positives, as the financial impact of a customer's departure often outweighs the cost of offering retention incentives. With that in mind, Logistic Regression retains its place as a preferred model.

9 Discussion

To conclude, the evaluation of the machine learning model performance on the "Telco Customer Churn" dataset yielded somewhat insightful results. The models exhibited a balanced trade-off between precision and recall, with each recall consuming approximately half of the actual churn cases. Considering that only 26.5% of customers have churned, the models' performance is not catastrophic. Yet, none of the models achieved groundbreaking results due to the nuanced nature of dealing

with unbalanced data and training on the default hyperparameters.

Several improvements can be suggested:

- Hyperparameter tuning can be conducted using methods such as GridSearchCV or more sophisticated approaches like Bayesian Optimization.
- Implementing sampling techniques such as the Synthetic Minority Over-sampling Technique (SMOTE) when training could help models capture nuances of the minority class.
- Integrating features such as gender and dependents to capture nuanced relationships that might otherwise be overlooked by predictive models.
- Utilising advanced ensemble techniques like boosted models like XGBClassifier and CatBoostClassifier, which were shown to have superior performance in churn prediction.
- Acquiring a real-world dataset to ensure the model's performance is grounded in true customer behaviour dynamics.

References

[1] S. R. Department, *Customer service: Churn rate by industry u.s.* Jul. 2022. [Online]. Available: https://www.statista.com/statistics/816735/customer-churn-rate-by-industry-us/.

[2] Wikipedia, *Churn rate*, [Online; accessed 01-February-2024], 2024. [Online]. Available: https://en.wikipedia.org/wiki/Churn_rate.

[3] B. Huang, M. T. Kechadi, and B. Buckley, "Customer churn prediction in telecommunications," *Expert Systems with Applications*, vol. 39, no. 1, pp. 1414–1425, 2012, ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2011.08.024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417411011353.

[4] T. Vafeiadis, K. Diamantaras, G. Sarigiannidis, and K. Chatzisavvas, "A comparison of machine learning techniques for customer churn prediction," *Simulation Modelling Practice and Theory*, vol. 55, pp. 1–9, 2015, ISSN: 1569-190X. DOI: https://doi.org/10.1016/j.simpat.2015.03.003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1569190X15000386.

[5] J. Burez and D. Van den Poel, "Handling class imbalance in customer churn prediction," *Expert Systems with Applications*, vol. 36, no. 3, Part 1, pp. 4626–4636, 2009, ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2008.05.027. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417408002121.

[6] BlastChar, *Telco customer churn*, Feb. 2018. [Online]. Available: https://www.kaggle.com/datasets/blastchar/telco-customer-churn.

[7] [Online]. Available: https://www.ml-cheatsheet.readthedocs.io.

[8] [Online]. Available: https://www.saedsayad.com/k_nearest_neighbors.htm.

[9] T. E. Inc, *Different types of cross-validations in machine learning*. Mar. 2022. [Online]. Available: https://www.turing.com/kb/different-types-of-cross-validations-in-machine-learning-and-their-explanations.