Assessing and Implementing Supervised Machine Learning for Effective Churn Prediction

Khanh Van Tran
Master of Science in Data Science
Faculty of Science
Thompson Rivers University

Supervisors:

Dr. Jabed Tomal
Department of Mathematics and Statistics
Thompson Rivers University

Dr. Md Erfanul Hoque
Department of Mathematics and Statistics
Thompson Rivers University

1 Introduction

1.1 Overview of Customer Churn

Customer churn is also known as customer attrition. It is a situation wherein a customer stops purchasing products and services from a company. Customers are hard to acquire and might be very expensive for an organization. If a customer does not use your firm's products and services in the long run, it will inevitably lead to the bad fate of an organization. This is why customer retention is also as important as customer acquisition. Without appropriate attention to customer churn management, it can result in heavy monetary losses, reduced profitability, and loss of visibility in the market (Jain et al., 2020). This is why organizations started relying on predictive analytics or machine learning approaches to predict or avoid customer churn.

Since corporations now have access to enormous amounts of customer data through large-scale databases and digital platforms, such information as consumer profiles, transaction histories, notes from communication with customer service personnel, and their interactions with the internet. Companies can now use this vast amount of data to create effective machine learning models for predicting customer attrition. Business entities can lower their customer attrition rate by using previously logged cases to identify trends and markers that predict their churn rates. (Khodabandehlou and Zivari Rahman, 2017).

As a result, an impressive number of studies tell us that machine learning models can effectively predict churn for different industries such as telecom, subscription-based services, e-commerce, and financial services (see Geiler et al., 2022, Jain et al., 2020, Lalwani et al., 2022, Qureshi et al., 2013, Rahman and Kumar, 2020, Ullah et al., 2019). These models can provide valuable customer insights and are particularly useful for companies to personalize their marketing efforts, increase customer experience, and execute focused retention strategies. Some of the impacts on outcomes can be quite dramatic, and using machine learning on churn can yield great returns on investment.

1.2 Project Objectives

As highlighted above, it is essential to understand why churn occurs and make plans to reduce it. Therefore, it increases customer loyalty and achieves corporate objectives. The following are the main goals of this study:

1. Identifying the primary factors contributing to customer churn: This research aims to figure out the main factors that lead to consumers terminating the service by analyzing customer data and identifying trends. Being able to understand these characteristics facilitates the design of targeted interventions aimed at addressing the factors behind churn.

2. Investigating Feature Correlations and Customer Loyalty Tactics: Studying the correlation between different characteristics and turnover status will offer valuable insights into consumer behavior and preferences. The research will provide methods to boost client retention by concentrating on individualized products and enhanced service quality.

3. Create a dependable machine learning model for predicting customer churn. The project attempts to develop a prediction model using sophisticated machine learning techniques to reliably anticipate client attrition. The project aims to provide a strong tool for proactive churn control by evaluating different algorithms and enhancing model performance.

To address these research goals, I will implement a thorough data science approach to understand customer turnover and create successful retention tactics. The full data science life cycle workflow consists of seven steps and follows a comprehensive approach, guaranteeing a systematic and data-driven analysis of the issue.

2 Literature Review

Churn prediction study has gained popularity in recent years due to its impact on corporations and businesses. The following sections summarize the recently published papers in this field of study.

A case study by Qureshi et al. (2013) on churn prediction in the mobile communication market, where people change from one company to another company. It describes the use of regression analysis, decision trees, artificial neural networks, and logistic regression to predict potential churners, looking into the historical data to get an idea of the pattern. The study used a dataset from the Customer DNA website, where usage data for 106,000 customers was provided over three months, along with total usage by the customers. Dealing with the problem of class imbalance in the dataset, the authors check how the different machine learning algorithms performed regarding real usage by the customer, while demonstrating that the decision trees were the best classifiers for identifying potential churners.

Sisodia et al. (2017) built a model for predicting employee churn rate based on HR analytics dataset, using five different machine learning algorithms, namely, linear support vector machine, decision tree classifier, random forest, k-nearest neighbor, and naïve bayes classifier. The authors evaluated the correlation between attributes, generated a histogram to contrast left employees with various factors, and proposed strategies to optimize employee attrition in organizations.

Ahmad et al. (2019) developed a churn prediction model using machine learning techniques on a big data platform to assist telecom operators in predicting customers who are likely to churn. The model incorporates features from social network analysis and achieves an AUC value of 93.3%. The dataset used for training and testing the model is obtained from SyriaTel telecom company and includes customer information over a period of 9 months. The model experiments with four algorithms, with the XGBoost algorithm yielding the best results for classification in churn prediction.

Ullah et al. (2019) proposed a churn prediction model using classification and clustering techniques to identify churn customers and determine the factors behind customer churn in the telecom sector. The model utilizes the Random Forest algorithm for churn classification and cosine similarity for grouping churn customers. It also employs feature selection techniques and attribute-selected classifier algorithm to identify significant churn factors. The evaluation of the proposed model shows improved churn classification and customer profiling. The results obtained from

the model can help improve customer retention strategies, recommend relevant promotions, and enhance marketing campaigns.

Rahman and Kumar (2020)'s paper focuses on predicting customer churn in a commercial bank using efficient data mining methods. It discusses data transformation techniques and the use of various classification techniques such as k-nearest neighbor, support vector machine, decision tree, and random forest. The paper results show that oversampling improves the accuracy of decision tree and random forest classifiers, while support vector machine is not suitable for large amounts of data. The study also analyzes customer behavior to explore the likelihood of churn and compares the performance of different models, finding that the Random Forest model after oversampling achieves higher accuracy.

Lalwani et al. (2022) predicted customer churn prediction in the telecom industry by using machine learning techniques. The authors applied various predictive models such as logistic regression, naive bayes, support vector machine, random forest, decision trees, boosting, and ensemble techniques. The paper also discusses the use of K-fold cross-validation for hyperparameter tuning and preventing overfitting. The results show that Adaboost and XGBoost classifiers achieve the highest accuracy of 81.71% and 80.8%, respectively.

Geiler et al. (2022) focused on churn prediction in businesses and explored the performance of various supervised and semi-supervised learning methods and sampling approaches on publicly available datasets. The study suggests an ensemble approach should be used for churn prediction.

There are numerous studies on churn prediction using machine learning that have been published, with a predominant focus on the theoretical aspects of the field, such as algorithm comparison and enhancement. There is, however, a notable gap in the literature regarding the practical application of machine learning techniques in this churn prediction area. This research aims to bridge this gap by providing a comprehensive analysis of the practical implementation of machine learning in churn prediction, thereby contributing to the advancement of the field in a more applied manner.

3 Data Description

This study plans to use several public churn datasets to compare and evaluate the performance of various machine learning algorithms.

The first data that I have done preliminary analysis is the telecom customer churn. This data can be accessed on the Maven Analytics website platform. This dataset comprises two tables, presented in CSV format.

The Customer Churn table has data from a total of 7,043 customers who are associated with a telecommunications firm located in the state of California during the second quarter of the year 2022. Each entry in the dataset corresponds to an individual customer and has information about their demographic characteristics, geographical location, length of tenure, subscription services availed, and quarterly status (i.e., whether they joined, remained, or churned), among other relevant factors.

The Zip Code Population table provides further data on the estimated populations of the California zip codes mentioned in the Customer Churn table.

4 Methods

4.1 Machine Learning Algorithms Use in Churn Prediction

This section provides a brief description of all the machine learning algorithms that will be used for this study.

4.1.1 Logistic Classification

Logistic regression is the simplest machine learning algorithm for binary as well as multiclass classification. It estimates the probability that an instance belongs to one of the classes as a function of input features using the logistic (sigmoid) function, which produces outputs in the range of 0-1. During training, the weights for input features are optimized using techniques such as

gradient descent and maximum likelihood estimation. It is very efficient to compute and highly interpretable.

4.1.2 K-Nearest Neighbors (KNN)

KNN is a machine learning method used for both regression and classification tasks. KNN utilizes distance measurements to predict the most probable value of the target feature. KNN employs several metrics to measure distance, including Euclidean, Manhattan, Chebyshev, Minkowski, and Mahalanobis. The optimal k nearest neighbors are identified for the given case based on the cross validation. The predicted class in categorization is the most popular class among the neighbors. This method is non-parametric and does not rely on any parametric distribution. It is a categorization method based solely on examples, utilizing existing data without generalization. It is called a lazy learning algorithm since its stages and operations are executed during the query. The algorithm does not require any preparation. It is effective with low-dimensional data but less so with high-dimensional data.

4.1.3 Support Vector Machines (SVM)

SVM is a supervised machine learning technique that has been widely used for classification purposes. It searches for an optimal hyperplane that maximizes the margin of separation between two classes when mapped to a feature space. By transforming the data into a higher-dimensional space, SVM can find a nonlinear separator, also known as a functional, making it simpler to model nonlinearly separable patterns. The algorithm then uses linear classification by selecting the support vectors to define the decision boundary or hyperplane. The SVM algorithm can utilize kernel functions, such as linear, polynomial, and radial basis function (RBF) to handle both linearly and nonlinearly separable data.

4.1.4 Naive Bayes

Naive Bayes is a type of probabilistic classifier model, which suggests that it may predict data from several classes at once. This model is based on the Bayes theorem. Multiple class predictions are made feasible by probabilistic classifiers. Based on conditional probability, the decision is made. Instead of using a single method, this approach employs a number of algorithms, but they all share a similar principle. It is assumed in this model that each variable contributes equally to the output. Since it only needs a limited quantity of training data, this model has an advantage over others.

4.1.5 Decision Trees

The decision tree algorithm is a popular supervised machine learning technique used for both classification and regression tasks. It operates by recursively partitioning the dataset into subsets based on the most informative features. Each internal node of the tree represents a decision point where a feature is evaluated, and each leaf node corresponds to a class label or a regression value. Decision trees are constructed using various criteria such as Gini impurity or information gain to optimize the splitting process. They are interpretable models that facilitate human understanding of decision-making processes in complex datasets.

4.1.6 Random Forrest

The random forest algorithm is an ensemble machine learning learning method that uses many decision trees to increase predictive power by combining their contributions. It constructs a forest of decision trees in training, for each tree a random sample of the training data is subsampled with a random sample of the features. Random forests can be fairly insensitive to outliers and more robust to overfitting the training data than single decision trees. It can perform both classification and regression tasks. Random forest has high accuracy and resilience to noisy data. Therefore, it is widely used in practice.

4.1.7 XGBoost

XGBoost or eXtreme Gradient Boosting is a type of gradient boosting algorithm. It is faster than any other algorithm while achieving high accuracy and interpretability. Gradient boosting is an iterative process of building an ensemble of weak decision trees. At each iteration, a weak learner can perform slightly better than random guessing. However, when many weak learners are combined, they can collectively perform very well. In gradient-boosting, each subsequent tree tries to correct the errors made by the preceding ones. To avoid overfitting, XGBoost implements regularization techniques that impose an additional penalty on unnecessarily complicated trees. It uses the gradient boosting framework to optimize a specified objective function, for instance, log-loss for classification. XGBoost has shown the best performance in the most competitive machine learning competitions. XGBoost is scalable and flexible. It can also handle different types of features, missing values, and a variety of objective functions.

4.1.8 LGBM

LGBM is a gradient-boosting framework that excels at handling large scale datasets and is widely used for classification and regression tasks. Developed by Microsoft, it employs a histogram-based learning approach that bins feature values to speed up the training process. LightGBM uses a tree-based model where decision trees are grown leaf-wise instead of level-wise, optimizing for the best splits and reducing computation time. This algorithm supports efficient handling of categorical features and implements a gradient-based optimization strategy to enhance model accuracy. LightGBM's superior performance, scalability, and flexibility make it a popular choice in various applications, particularly in scenarios where large datasets and high-dimensional feature spaces are involved.

4.2 Imbalanced Datasets Handling

Several methods can be employed to handle imbalanced datasets in machine learning classification:

- Resampling Techniques.
  - Oversampling: Increase the number of instances in the minority class by duplicating or generating synthetic samples.
  - Undersampling: Reduce the number of majority class instances by randomly removing samples.
- Synthetic Minority Over-sampling Technique (SMOTE): Create artificial minority samples by adding new data points between corresponding minority instances using interpolation.

This study will apply the above techniques to lessen the impact of the unbalanced dataset. It also uses stratified k-fold cross-validation and correct metrics to fully capture the impact of the unbalanced dataset on the performance of machine learning algorithms.

4.3 Stratified K-Fold Cross Validation

Given that the dataset is highly imbalanced, we use stratified K-fold cross validation for the evaluation of the ML models. In K-fold cross validation, the initial dataset is first divided into K folds. Each fold contains an equal number of training cases and an equal number of test cases. During the training process, the model is built using K-1 folds as a training set and tested on the remaining test fold. This process is repeated K times, where each fold takes a turn to be in the test set, resulting in a comprehensive evaluation of the modeled performance. The main issue with the K-fold cross validation is when there exists a large class imbalance between the classes in the initial dataset, such that the size of one class in the training fold is much less than the number of cases in the other class of the dataset. In such cases, the results of the validation evaluation are too optimistic and biased in favor of the model. Therefore, stratified K-fold cross validation is recommended in these cases, where each fold should be stratified in the same way as the dataset. This ensures that the class distribution of each fold is an exact mirror image of the class distribution of the initial dataset. Figure 1 illustrates the stratified K-fold cross validation.

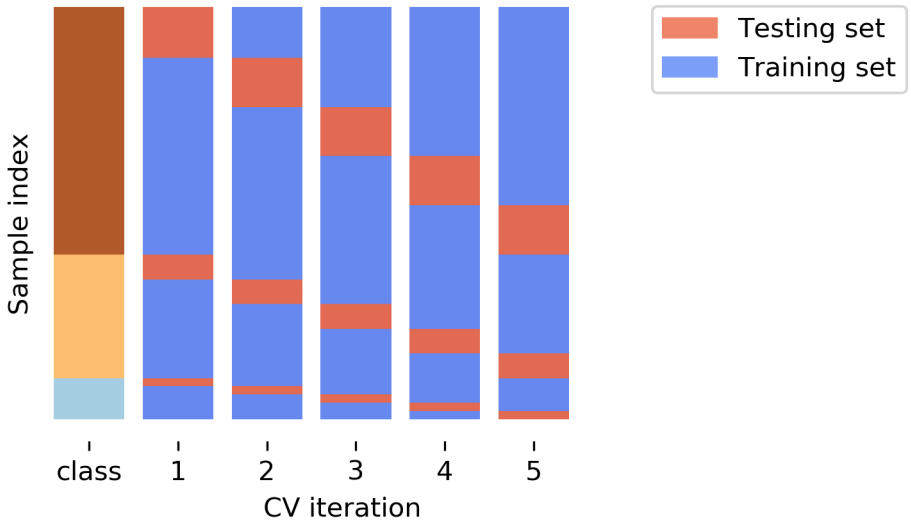

Figure 1: Stratified K-Fold Cross Validation.

4.4 Model Performance Metrics

The performance of the models developed in this study is assessed using the confusion matrix and its derived metrics, namely accuracy, precision, recall, specificity, and F1 score. Figure 2, below, demonstrates the confusion matrix represented with actual and predicted outcome categories plotted against its axis.

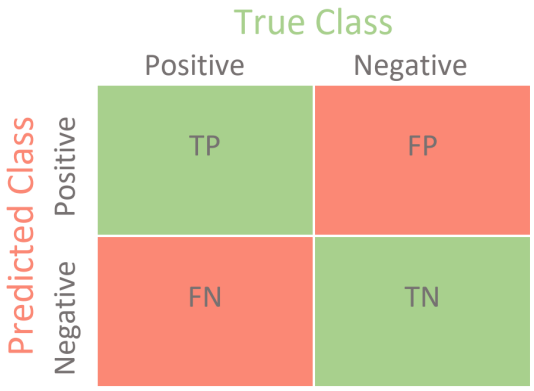

Figure 2: Confusion matrix.

4.4.1 Accuracy

Accuracy is the ratio of correctly predicted classes over the total number of instances in the dataset and is calculated by the equation 1 (Witten & James, 2013). Accuracy is used to measure the overall performance of a model but can be misleading in imbalanced datasets.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

4.4.2 Precision

Precision shows the proportion of positive predictions that are truly positive (Witten & James, 2013). It indicates the reliability of a model in predicting a class of interest. It is basically a ratio of correctly positively labeled to all positively labeled and can be calculated as follows:

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

4.4.3 Recall

Recall has other names Sensitivity or True Positive Rate. It is a measure of the percentage of actual positive classes that are predicted as positive (Witten & James, 2013). It can be calculated as follows:

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

4.4.4 Specificity

Specificity or True Negative Rate is used to measure the portion of the negative class that has been correctly classified (Witten & James, 2013). It is calculated by the following equation:

$$Specificity = \frac{TN}{TN + FP} \tag{4}$$

4.4.5 F1 Score

The F1 Score is a measure that is calculated from precision and recall (Equation 5). It is often a preferred metric over accuracy when data is unbalanced.

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

4.5 Methodology

To address these research goals, I will implement a thorough data science approach to understand customer turnover and create successful retention tactics. The data science workflow consists of seven steps and follows a comprehensive approach, guaranteeing a systematic and data-driven analysis of the issue (Figure 3).

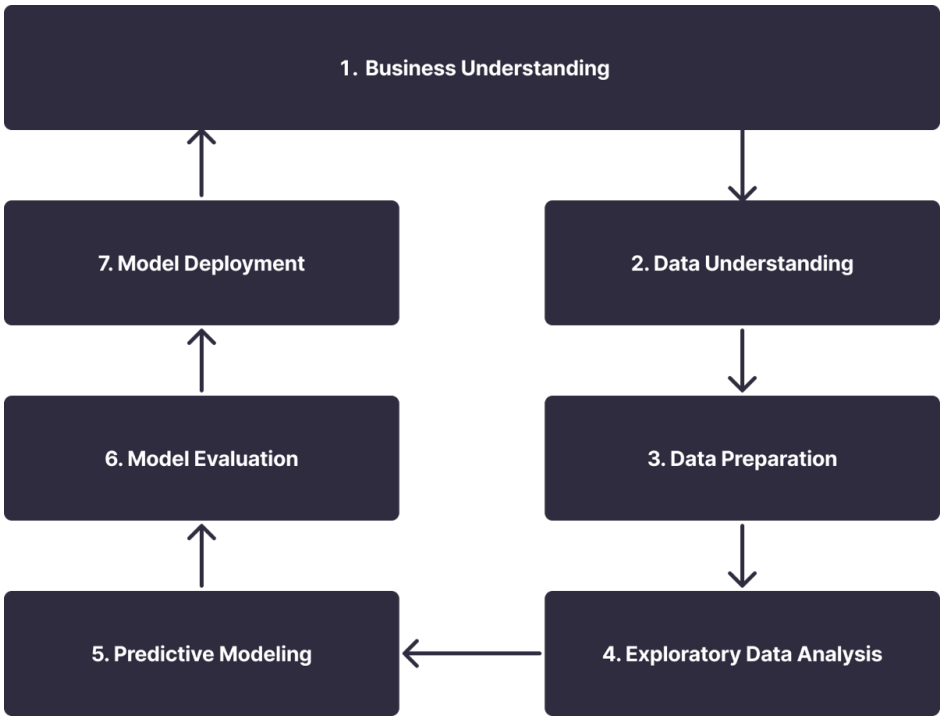

Figure 3: Data Science Life Cycle.

4.5.1 Business Understanding

The first phase will concentrate on clarifying the corporate objectives and matching them with the analytical aims. The main point is to comprehend the financial and operational consequences of client turnover and establish a definitive standard for effective retention efforts.

4.5.2 Data Understanding

During this stage, the focus is on analyzing the existing data to understand its organization, information, and reliability. It is the foundation for finding the necessary data to answer research questions and sets the stage for thorough data examination.

4.5.3 Data Preparation

Data preparation is a process of performing all the necessary tasks to generate the final dataset from a raw dataset, which will include data cleaning, feature selection, data transformation, and splitting the dataset into training set and test set for model building and validation.

4.5.4 Exploratory Data Analysis

Through exploratory data analysis (EDA), it aims to discover the underlying patterns, relationships, and conclusions found from the telecom customer churn dataset, which can help guide predictive modeling and further strategic direction. The dataset comprises information on 7043 customers, encoded across 38 features that span demographic details, service usage, billing information, and churn specifics. This comprehensive analysis forms the foundation for identifying key factors influencing customer behavior and retention in the telecom sector.

4.5.5 Predictive Model Building

Insights from the EDA step will be used to develop and refine predictive models for forecasting customer attrition. All the popular machine learning models will be implemented and evaluated.

The top 3 models will be selected and fine-tuned to improve their capacity to forecast outcomes, focusing on being easily understood and providing practical insights.

4.5.6 Model Evaluation

Models will be thoroughly evaluated using suitable performance metrics, namely, accuracy, precision, recall, specificity, and F1 score. The goal is to choose a model that properly forecasts customer attrition and also offers insights into the probability of churn, allowing the organization to take proactive actions.

4.5.7 Model Deployment

Once the final model is validated and evaluated for robustness, it will be implemented in a production environment. This would enable real-time prediction of customer churn and the automation of efforts to retain customers, integrating data-driven decision-making into the company.

5 Preliminary Results

The preliminary study was performed with Logistic Regression, K Nearest Neighbors, Support Vector Machine, Naive Bayes, Decision Tree, Random Forest, Lightgbm, Gradient Boosting, AdaBoost Classifiers, and MLP Classifier using Sklearn (Pedregosa et al., 2011) packages in Python on Telcom churn dataset. It has been executed in Python version 3.11.4 on a Windows 11 machine with 8GB RAM. The Accuracy, Precision, Recall, and F1-Score were used to evaluate and compare the algorithms' performance. The preliminary results are given in Table 1.

Table 1: **Model Comparison**

Model	Accuracy	Precision	Recall	F1 Score
LGBM	0.8367	0.8324	0.8367	0.8335
Random Forest	0.8357	0.8298	0.8357	0.8306
XGBoost	0.8306	0.8256	0.8306	0.8269
Logistic Regression	0.7992	0.7996	0.7992	0.7987
Decision Tree	0.7751	0.7778	0.7751	0.7759
KNeighbors	0.7481	0.7283	0.7481	0.7336
Naive Bayes	0.7227	0.7749	0.7227	0.7375
SVC	0.6604	0.7279	0.6604	0.6380

A Research Schedule

The tentative research schedule is proposed in detail in Table 2

Table 2: **Tentative Research Schedule**

Research phase	Objectives	Deadline
Literature review	Literature review	30-Feb-24
Proposal Preparation	Gather data from public domain	15-Mar-24
Proposal Submission	Supervisors review and approval	21-Mar-24
Methodology	Review project methodology with supervisors	01-Apr-24
Model building	Build and compare models performance	15-Apr-24
Preliminary analysis results	Review preliminary results with supervisor	30-Apr-24
Revised analysis results	Review revised results with supervisor	15-May-24
Project writing	Finish draft report writing	30-May-24
Project Revision	Supervisors review and feedback	15-Jun-24
Final review	Final review with supervisors	30-June-24
Project Presentation	Presenting research and outcomes	15-July-24

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.

Geiler, L., Affeldt, S., & Nadif, M. (2022). A survey on machine learning methods for churn prediction. *International Journal of Data Science and Analytics*, *14*(3), 217–242.

Jain, H., Yadav, G., & Manoov, R. (2020). Churn prediction and retention in banking, telecom and it sectors using machine learning techniques. In *Advances in machine learning and computational intelligence: Proceedings of icmlci 2019* (pp. 137–156). Springer.

Khodabandehlou, S., & Zivari Rahman, M. (2017). Comparison of supervised machine learning techniques for customer churn prediction based on analysis of customer behavior. *Journal of Systems and Information Technology*, *19*(1/2), 65–93.

Lalwani, P., Mishra, M. K., Chadha, J. S., & Sethi, P. (2022). Customer churn prediction system: A machine learning approach. *Computing*, 1–24.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Qureshi, S. A., Rehman, A. S., Qamar, A. M., Kamal, A., & Rehman, A. (2013). Telecommunication subscribers' churn prediction model using machine learning. *Eighth international conference on digital information management (ICDIM 2013)*, 131–136.

Rahman, M., & Kumar, V. (2020). Machine learning based customer churn prediction in banking. *2020 4th international conference on electronics, communication and aerospace technology (ICECA)*, 1196–1201.

Sisodia, D. S., Vishwakarma, S., & Pujahari, A. (2017). Evaluation of machine learning models for employee churn prediction. *2017 international conference on inventive computing and informatics (icici)*, 1016–1020.

Ullah, I., Raza, B., Malik, A. K., Imran, M., Islam, S. U., & Kim, S. W. (2019). A churn prediction model using random forest: Analysis of machine learning techniques for churn prediction and factor identification in telecom sector. *IEEE access*, *7*, 60134–60149.

Witten, D., & James, G. (2013). *An introduction to statistical learning with applications in r*. springer publication.