CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter would be discussing the overview of customer churn prediction, followed by the background of e-commerce sector. The comparisons between Linear Regression, Logistic Regression, and Random Forest Model are made to find out the strengths and limitations within the model. At the end of the chapter, the best model that is stronger in overall would be concluded.

2.2 Theoretical Background

Customer churn is a kind of scenario where the customers are leaving the platform without getting noticed. This is a loss to the organization, because organizations unable to make money from the churned customers, especially if the churned customers contribute majority of the sales towards the organization, leading the organization to have more funding for future development.

The customer churn can be categorized into 2 types, that is voluntarily churn and involuntary churn (Fredrik Fagerholm, 2022). Voluntarily churn means the customer decided to leave the platform by stopping the support initiatively without including personal frustration while involuntary churn means the customer leave the platform due to the factors that is beyond their control, causing them to accept the decision forcefully. The studies related to voluntary churn and involuntary churn would be further discussed in the next paragraph.

A study from Arun Velu (2021) saying that the voluntarily churn happens when the customer quits the service or switches to the service provided by the competitor. Voluntary churn could be caused by various factors, such as product or service dissatisfaction, the availability of superior alternatives or change of personal situations (Long, 2023). It is hard to be identified because the churners would not directly tell the reason they leave with initiatives. They would use action to express out the finalized decision without any reasoning. The reason can be in variety, such as cheaper alternative exist in the competitor platform and service does not meet the minimum requirements of customer needs (Daniel and William, 2023).

From the study done by Nayema Taskin (2023), the involuntary churn happens when the companies resolve the issue beyond the standard timeframe, which they could not control the timeframe for companies' resolution time. The customers who are dissatisfied with anger would be spreading the bad experience towards the third parties, including friends, families, or co-workers. The motive is to instill the negative impacts of keep supporting the thing that they keep support with, so that they would have a mentality to prepare the negative consequences that will be happening on them on time.

In the behavior analysis done by Emílio José Montero Arruda Filho and Alexis de Araújo Barcelos (2020), consumers tried to resolve the complaints through private and public channels, but the result is less satisfying. The expected result from the effort is not equal level with the actual result. Hence, equity instantiation is done to the company, causing the company have detrimental effect in the company reputation. The revenge caused by the resolution alert the companies to take serious action to minimize the financial impact as low as possible.

2.3 Dataset Review

Dataset is the key contributor to the model training. The dataset found in these previous researches would be analyzing its strength, limitations, and the application in the real life case study. The details are articulated in the next paragraphs.

E-commerce platforms generate vast behavioral datasets that capture customer interactions, purchases, and churn signals (Abdulrahman Alshamsi, 2022). These datasets enable researchers to identify patterns in customer attrition using machine learning (ML). For instance, Alshamsi (2022) applied CRISP-DM methodology to a Kaggle dataset of 5,000+

customers, testing Decision Trees, Logistic Regression, and Random Forest models. Strengths included six months of temporal data (June–November 2021) and granular behavioral attributes (e.g., login devices, satisfaction scores), which revealed that mobile app users had 23% higher retention than web users. However, limitations such as class imbalance (few churned customers) and moderate sample size restricted model generalizability, necessitating synthetic oversampling techniques.

The study conducted by Rehka Yadav (2024) stated that the dataset found in Kaggle contain 250,000 entries and 13 columns. Out of the 13 columns, the key attribute churn is the result predicted after summing up the all the attributes into it. These attributes are customer ID, Product Price, Quantity, Total Purchase Amount, Customer Age, Returns, and Churn. Given that the Age and Customer Age share the same characteristics, either Age or Customer Age could be removed. This benefits in model training, consuming less computational resources to process the dataset. At the same time, removal of duplicated attribute reduces the chance to let the model memorize the attribute rather than training it to recognize.

Mukun Chang (2023) analyzed niche live-streaming e-commerce data using clustering techniques. The dataset's strength was novel engagement metrics (e.g., average watch time, gift purchases), which identified four user segments. Surprisingly, "high-spending casual viewers" exhibited 62% churn rates despite high satisfaction scores. A key limitation was platform-specific data collection, limiting cross-segment comparisons. The study demonstrated how unconventional behavioral indicators could uncover non-linear relationships between engagement and churn.

Table 2.3: Comparison between the strengths and limitations of dataset

References	Experiment	Strength	Limitation
Abdulrahman Alshamsi, 2022	Customer Churn in E-Commerce Sector (EIT)	• Include preferred login device and satisfaction metrics • Include churn flag for labelling	• Imbalanced dataset.

Rehka Yadav, 2024	Machine Learning Insights into E-Commerce Churn	• Comprehensive customer behaviour and purchase history. • Large dataset (250,000 entries with 13 columns)	• Duplicated features exist. (Customer Age and Age columns)
Mukun Chang, 2023	Customer Churn Prediction based on E-Commerce Live Streaming.	• Revealed non-intuitive insights, such as satisfaction, that is not always correlated to churn.	• Focus on live-streaming e-commerce segment

2.4 Supervised Learning in Customer Churn Prediction

Supervised Learning is a type of Machine Learning approach that learns the training data with specified information. The training data is the food that needs to feed the model so that model able to learn the way to classify the data. It can be divided into 3 stages, that is training, testing and validation. (Shreyas Rajesh Labhsetwar, 2020) Firstly, the training data act as a sample to feed into the model. Then, the model is tested using testing dataset, determining the accuracy of the output based on the last training approach. In case the model output is less satisfying, the validation step would be conducted to fine tune the hyper-parameters in the model. The process would be iterating over the model development, until the model fits the expected output in training dataset. In this case, the 3 supervised learning approach, that is Linear Regression, Logistic Regression, and Random Forest would be described in detail.

2.4.1 Linear Regression

Linear Regression is a method that is used to find out the correlation between independent variable X and dependent variable Y. The independent variable in the dataset is suitable in Linear Regression because it is the only factor that influences the other dependent variable. Hence, the graph is linear line. The formula for single Linear Regression is as below.

$$y = \beta 0 + \beta 1x + \epsilon \qquad (2.1)$$

Where:

y=Predicted value
β=The coefficient of the parameter
x=The nth feature value
ϵ=Error term

According to Enlin Deng (2025), linear regression is used to evaluate the relationship between independent variables and dependent variables. When there are multiple independent variables affect dependent variables, it indicates that the formula that is used in Single Linear Regression becomes incompatible. In such case, Multiple Linear Regression formula is introduced to cope the scenario where multiple dependent variables affect the same variable. Multiple Linear Regression Formula is shown below.

$$y = \beta 0 + \beta 1x1 + \beta 2x2 + \cdots + \beta kxk + \epsilon \qquad (2.2)$$

Where:

y=Predicted value
β=The coefficient of the parameter
x=The nth feature value
ϵ=Error term

2.4.2 Logistic Regression

Logistic Regression is an algorithm that is used to classify the problems using binary approach (Aurélien Géron, 2023). It is used to estimate the probability of the instance belonging to a particular class. In this case, churn response is the key to define the class. There are two types of churn response, that is churn and not churn. Churn is the case where the situation is true in the class identification whereas not churn is the case where the situation is false in the class identification. Churn is defined as 1 while not churn is defined as 0.

According to Ahmed (2024), Logistic Regression is a method that links the binary dependent variable together with one or more independent variables. The linkage uses fundamental statistical method to discover the relationship between the attributes found in the dataset. At the end of the interpretation, the key variables that affect the customer churn rate could be addressed.

2.4.3 Random Forest

Random Forest is the complex model that is built in the foundation of Decision Tree. Multiple Decision Trees combine together to form a Random Forest model, which is more complex than single Decision Tree. According to Swetha P, Dayananda R B (2020), Decision Tree able to make decision using divide and conquer method, meaning that it break the problem into smaller problem first. Then, it would answer the smaller problem accordingly. At the end of the answering process is completed, all the answered questions combine together to form the final predicted result.

The following divide and conquer concept could be addressed more technically from the analysis done by Xinyu Miao and Haoran Wang (2025). The complex relationship between the first variable and second variable is the advantage of Random Forest. The origin of the Random Forest, which is Decision Tree, only contain a tree. A tree can only contain one set of data. The limitation in the single tree would cause the tree to be overfitting, due to the tree is sensitive towards the pretrained dataset. To reduce the overfitting level, Ensemble Learning is introduced to use the same algorithm to train the original dataset in multiple batches. Each batch of dataset is split randomly. Then, the batch of dataset is dumped into the training process, iterating till the last batch of dataset. The train result of each batch is the aggregated into generalized result, improving the generalization in prediction.

2.4.4 Random Forest Attached with SMOTE

Synthetic Minority Over-Sampling Technique (SMOTE) is a type of technique that improves accuracy based on the foundation of Random Forest Algorithm (Hafiz Ma'ruf and Rodiah, 2021). SMOTE tackles class imbalance by creating new synthetic samples to balance the dataset. The synthetic samples are created in diverse, to ensure that the model have more opportunities to learn the minority class as well, not only limit to learning the majority class.

The advantage of SMOTE is that the original data is preserved during training. Higher accuracy could be achieved and recalled for churn class. The disadvantage is that the noise may be introduced if the dataset itself contain outliers and noise. Moreover, introducing the scattered data have chance to cause the dataset to scatter, making the dataset itself become more unrealistic to actual instances.

2.4.5 Random Forest Attached with XGBoost

Extreme Gradient Boosting (XGBoost) is a type of Random Forest algorithm that is strengthened with multiple supplemental characteristics (Sana Fatima et al., 2023). Compared to pure Random Forest algorithm, XGBoost provides more sophisticated result due to more fine tune approach is done based on Random Forest algorithm. XGBoost able to deal with the imbalanced class by giving priority on underrepresented classes, making the underrepresented class be considered in attention during training process.

The advantage of XGBoost is that it combines weak learners a.k.a individual decision trees to form a big forest a.k.a powerful ensemble model. This powerful ensemble model improves the overall accuracy by filling in the strengths of the multiple models into the weakness found in the individual modules. But, the XGBoost challenge and limitations lies on the hyperparameter setting (Yashkumar Burnwal and Dr. R.C. Jaiswal, 2023). If the hyperparameter is not adjusted properly, it would result in overfitting, which indicates that the model always generates the result as True, even the fact is False. When the data is sparse, the model would require more complex parameter tuning to make the model compatible to the dataset, causing the model complexity increase. Model would become difficult to maintain in the future implementation.

Table 2.4: Comparison between the performance in multiple models

Model	Strengths	Limitations	Citation
Linear Regression	• Simple Implementation method	• Only capture linear relationships. • Sensitive to outliers	Enlin Deng (2025)
Logistic Regression	• Effective for binary classification problems. • Provides probability estimates • Less prone to overfitting with regularization	• Limited to linear decision boundaries • Require more data for stable estimation	Ahmed (2024)
Random Forest	• Reduce overfitting via ensemble averaging.	• Increasing trees cause longer prediction time.	Swetha P, Dayananda R B (2020)

		•More computational resources required. • Biased to dominant class if not properly tuned.	
Random Forest + SMOTE	• Improve imbalanced dataset performance. • Achieve higher accuracy and recall for churn class.	• Artificial data noise • More computational resources than pure Random Forest.	(Hafiz Ma’ruf, 2021)
Random Forest + XGBoost	• Combine bagging and boosting for higher accuracy output.	• Extra parameter tuning required • Model become more complex, increasing maintenance difficulty	(Yashkumar Burnwal and Dr. R.C. Jaiswal, 2023).

2.5 Evaluation Metrics and Performance Analysis

The evaluation of customer churn prediction models requires comprehensive metrics that capture both statistical accuracy and business relevance. The literature review reveals inconsistencies in evaluation approaches across different studies, making performance comparison challenging.

Accuracy remains the most commonly reported metric across the reviewed studies. Alshamsi (2022) achieved varying accuracy levels with different models, with Random Forest showing superior performance compared to Decision Trees and Logistic Regression. Similarly, Yadav (2024) reported accuracy improvements when using comprehensive behavioral features from the large-scale dataset. However, accuracy alone can be misleading in imbalanced datasets where churned customers represent a small percentage of the total customer base.

Precision and recall scores give further insight into model performance, specifically for the minority churn class. Ma'ruf and Rodiah (2021) pointed out the significance of recall in churn prediction since failing to detect actual churners (false negatives) is more expensive than mistakenly identifying loyal customers as potential churners (false positives). Their SMOTE-based Random Forest method was tailored to improve recall for the churn class at no loss in overall accuracy.

F1-score, an area between recall and precision, is a more inclusive measure of an assessment metric. Fatima et al. (2023) highlighted how XGBoost optimization aims at improving F1-score rather than accuracy alone, particularly in dealing with imbalanced datasets. This is more in tune with business objectives where both precision and recall are critical for successful churn prevention programs.

Area Under the Curve (AUC) and Receiver Operating Characteristic (ROC) curves offer an additional level of assessment. Burnwal and Jaiswal (2023) indicated that XGBoost models have greater AUC scores, which reflect the models' ability to accurately distinguish churned and non-churned customers at varying threshold levels. This measure is especially useful for companies that must set their intervention thresholds according to available resources and campaign expenditure.

Cross-validation approaches differ widely between studies. Labhsetwar (2020) employed traditional train-test splits, whereas more recent research such as Chang (2023) employed k-fold cross-validation to preserve model stability when applied to various subsets of the data. The chosen validation approach impacts the model's generalizability and even the validity of the metrics of performance.

Temporal validation presents a special challenge for churn prediction evaluation. Taskin (2023) highlighted the importance of time-based validation in which models are trained on past data and evaluated on future time ranges, mimicking real-world deployment conditions. This temporal dimension is overlooked in research employing random sampling for train-test splits.

The testing duration also differs between studies. Whereas some concentrate on short-term churn prediction (next month), others such as Velu (2021) consider longer prediction horizons. The difference in testing durations makes it difficult to develop standardized performance benchmarks for various business contexts and industries.

2.6 Research Gaps and Limitations

The comprehensive review of customer churn prediction literature reveals several critical gaps that limit the practical application and theoretical advancement of current approaches. It can be discussed in 5 different perspectives, that is methodological gap, dataset and data quality limitations, model interpretability and explainability gaps, real world implementation gaps, and external factors and environmental context.

2.6.1 Methodological Gaps

The lack of standardized evaluation frameworks represents a significant methodological gap. Studies use different metrics, validation approaches, and performance criteria, making meaningful comparison impossible. Imani (2024) noted this inconsistency when evaluating classification methods under varying imbalance levels, highlighting the need for standardized evaluation protocols that consider both statistical and business performance measures.

Feature engineering approaches remain largely ad-hoc across studies. While Yadav (2024) identified duplicate features as a data quality issue, the broader challenge of systematic feature selection and engineering lacks comprehensive treatment. Different studies focus on various feature types - transactional, behavioral, demographic - without establishing clear guidelines for feature selection based on business context or data availability.

The temporal aspect of churn prediction receives insufficient attention in most methodological approaches. Xu et al. (2021) attempted to address this through ensemble learning with feature grouping, but the broader challenge of incorporating time-varying factors and seasonal patterns remains underexplored. Most studies treat churn prediction as a static problem rather than a dynamic process that evolves over time.

2.6.2 Dataset and Data Quality Limitations

Class imbalance remains a persistent challenge across all reviewed studies. Despite various technical solutions like SMOTE (Ma'ruf and Rodiah, 2021) and XGBoost optimization (Fatima et al., 2023), the fundamental issue of limited positive examples for training continues to constrain model performance. More importantly, the artificially balanced datasets may not reflect real-world distributions, raising questions about model performance in actual deployment scenarios.

Dataset generalizability presents another significant limitation. Chang (2023) worked with platform-specific live-streaming data, while Alshamsi (2022) used general e-commerce datasets. The lack of cross-platform and cross-industry validation limits the applicability of findings beyond specific contexts. This is particularly problematic for organizations seeking to implement proven solutions in their unique business environments.

Data privacy and availability constraints are inadequately addressed in current literature. Most studies rely on anonymized or synthetic datasets, but real-world implementation requires dealing with privacy regulations, data governance, and integration challenges that are rarely discussed in academic research.

2.6.3 Model Interpretability and Explainability Gaps

The trade-off between model accuracy and interpretability receives limited attention across the reviewed literature. While ensemble methods like XGBoost achieve higher accuracy (Burnwal and Jaiswal, 2023), their complexity makes it difficult for business users to understand and trust the predictions. This interpretability gap becomes critical when models need to support business decision-making and customer retention strategies.

Feature importance and model explainability are mentioned but not thoroughly explored. Business stakeholders need to understand which factors drive churn predictions to design effective intervention strategies. The current literature focuses primarily on predictive accuracy without adequately addressing the explanatory requirements of practical applications.

2.6.4 Real World Implementation Challenges

The gap between research and practice remains substantial. Most studies focus on achieving high accuracy in controlled experimental conditions but fail to address deployment challenges such as real-time prediction requirements, system integration, model maintenance, and continuous learning from new data.

Scalability considerations are largely absent from the reviewed literature. While Yadav (2024) worked with 250,000 records, the computational requirements and scalability challenges of deploying these models to process millions of customer records in real-time are not adequately addressed.

The economic impact and cost-benefit analysis of churn prediction models receive minimal attention. Organizations need to understand not just the accuracy of predictions but also the economic value of different intervention strategies and the optimal allocation of retention resources.

2.6.5 External Factors and Environmental Context

The reviewed literature largely ignores external factors that influence customer churn. Market conditions, competitive actions, economic cycles, and external events like the COVID-19 pandemic can significantly impact customer behavior, but current models are not designed to incorporate these external variables.

Cross-cultural and geographical variations in customer behavior are not adequately considered. Most studies focus on specific regions or platforms without examining how cultural factors, regulatory environments, and market maturity affect churn patterns and prediction accuracy.

The dynamic nature of customer preferences and market conditions requires adaptive models that can continuously learn and update their predictions. Current approaches largely rely on static models that require manual retraining, limiting their effectiveness in rapidly changing business environments.

2.7 Discussion

The existing literature shows the methodologies used in customer churn prediction, majorly in e-commerce domain. Even though the other domain such as telecommunication and banking sector also be introduced, the underlying concept is still the same, as all the domain addresses the same problem, that is customer churn prediction. The accuracy of the previous research is a challenging issue in respective to practical implementation constraints.

Theory behind demonstrates a fundamental distinction between voluntary and involuntary churn that will impact how we approach the issue of prediction. Voluntary churn, as defined by Velu (2021), happens when customers consciously decide to switch or leave for competitors. This type of churn is more predictable as it follows some patterns of behavior. But involuntary churn is another matter entirely. Taskin (2023) describes how customers churn because of circumstances outside of their control, including lengthy service resolution times.

The behavioral study by Arruda Filho and Barcelos (2020) introduces an additional layer of complexity in that it demonstrates that unhappy customers can actively harm company reputation through negative word of mouth. This indicates that churn prediction models must take into account not only purchase behavior, but also satisfaction levels and external stimuli that impact customer decision-making.

Data analysis in both studies has common challenges that constrain model performance. Alshamsi (2022) dealt with 5,000+ customers with the CRISP-DM approach, whereas Yadav (2024) dealt with a significantly larger dataset containing 250,000 records and 13 columns. Although of varying sizes, both study works had the same ground problem: class imbalance. The majority of the customers do not churn; thus, the model does not have many examples to learn from the minority class that churns. Chang (2023) attempted to solve this by incorporating live-streaming engagement metrics and found something surprising - high-spending casual watchers had 62% churn rates with high satisfaction rates. This contradicts the typical assumption that spending and satisfaction always result in retention.

The machine learning model comparison shows clear progression in capability, but also increasing complexity. Linear Regression, as analyzed by Deng (2025), provides a simple starting point but can only capture linear relationships between variables. Logistic Regression improves on this by handling binary classification problems better, as Ahmed (2024) demonstrates. However, the real advancement comes with Random Forest models that can handle complex, non-linear relationships. The ensemble approach of Random Forest, explained by Swetha P and Dayananda R B (2020), reduces overfitting by combining multiple decision trees, making predictions more stable and accurate.

The advanced Random Forest methods have even better performance but at the cost of more complexity. SMOTE method, as explained by Ma'ruf and Rodiah (2021), assists in dealing with imbalanced data by generating synthetic samples of the minority class. It improves accuracy and recall in churn prediction but may introduce noise if the initial dataset contains outliers. XGBoost, however, uses a different approach by combining bagging and boosting techniques, as explained by Fatima et al. (2023). While this is more accurate, Burnwal and Jaiswal (2023) explain that XGBoost requires careful hyperparameter tuning and ends up being cumbersome to maintain in practical applications.

What emerges from the debate is a trade-off between accuracy and practicality. The more accurate models are also the more complicated ones, demanding greater computer resources and technical know-how to set up and administer. Organizations must weigh the need for high accuracy against the practical limitations of their technical environment and personnel.

The literature also identifies a number of gaps that circumscribe the utility of prevailing approaches. First, churn prediction in the majority of the studies is assumed to occur in a static world, whereas the actual world is in a continuous state of change because of market forces, competition, and externalities. Second, feature engineering in the studies differs considerably, and hence comparison of results or determining best practices is challenging. Third, the evaluation metrics value statistical performance more than business impact - the model can be statistically correct but still produce nonsensical retention campaigns that squander resources.

Another critical lacuna involves the lack of consideration for temporal considerations and spillover effects. Although some research incorporates time-sensitive attributes, none of them provide a satisfactory way of specifying how evolving market conditions or competitors' actions could impact trends in churn. The pandemic from the COVID-19 virus, for instance, drastically altered e-commerce habits, but existing models would be unable to respond to such drastic shifts unless re-trained.

The deployment problems of operations are not well covered in the literature as well. Although most studies concentrate on having high accuracy in controlled experimental settings, deployment entails other factors like real-time processing needs, model explainability to business users, and integration with other systems.

Future research should address these limitations by developing more robust approaches that can handle dynamic environments and provide practical value to organizations. This includes creating standardized evaluation frameworks that consider business impact, developing models that can adapt to changing conditions, and focusing on interpretability alongside accuracy. The field needs to move beyond purely technical improvements to create solutions that work effectively in real-world business contexts.

2.8 Summary

The models used in customer churn prediction is not the best solution, due to the variety of factors changed unexpectedly. The churn prediction result provides a valuable insight to the organization, saying that the. In the previous studies that uses Linear Regression and Logistic Regression model able to perform churn identification with fewer attributes. With the presence of Random Forest model, churn identification becomes more generalized and able to generate more stable and accurate prediction, which is condensed from multiple decision trees. Out of all Random Forest Model, Random Forest attached with XGBoost performs the best, due to the ability in processing imbalanced handling, resulting in higher accuracy at the end, compared to pure Random Forest model and Random Forest model attached with SMOTE. The next chapter would be discussing the gap, followed by Data Collection process. The collected data would be used in Data Preprocessing Steps. Cleaned dataset is the final product. Full Exploratory Data Analysis (EDA) would be conducted to find out all the data attributes. Feature Engineering would be conducted to transform the data into features that are compatible with the machine learning models. A comparison of accuracy between the selected Machine Learning model would be made. A thorough analysis of data quality and model selection would be done.

REFERENCES

Aljifri, A. (2024). Predicting Customer Churn in a Subscription-Based E-Commerce Platform Using Machine Learning Techniques.

Alshamsi, S. A. (2022). *Customer Churn Prediction in E-Commerce Sector.* Dubai: Rochester Institute of Technology.

Barcelos, E. J. (2020). Negative Online Word-of-Mouth: Consumer's Retaliation in the Digital World. *Journal of Global Marketing*, 1-19.

Chang, M. (2023). *Customer Churn Prediction based on E-Commerce Live Streaming Data.* Rotterdam: Erasmus University Rotterdam.

Daniel Dahlén, W. M. (2023). *Machine Learning-based Prediction of Customer Churn in SaaS.* Lund: Lund University.

Deng, E. (2025). Customer Churn Prediction based on Multiple Linear Regression and Random Forest. *5th International Conference on Signal Processing and Machine Learning*, 22-28.

Géron, A. (2023). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow.* San Francisco: O’Reilly Media, Inc.

Hafiz Ma'ruf, R. (2021). Analysis of Random Forest Algorithm on Customer Churn Prediction to Handle Imbalanced Data. *International Research Journal of Advanced Engineering and Science*, 102-106.

Imani, M. (2024). Evaluating Classification and Sampling Methods for Customer Churn Prediction Under Varying Imbalance Levels.

Labhsetwar, S. R. (2020). Predictive Analysis of Customer Churn in Telecom Industry using Supervised Learning. *ICTACT Journal On Soft Computing*, 2054-2060.

Sana Fatima, A. H. (2023). XGBoost and Random Forest Algorithms: An In-Depth Analysis. *Pakistan Journal of Scientific Research, PJOSR*, 26-31.

Swetha P, D. R. (2020). Customer Churn Prediction and Upselling using MRF (Modified Random Forest) Technique. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, 475-482.

Taskin, N. (2023). *Customer Churn Prediction Model in Telecommunication Sector Using Machine Learning Technique*. Uppsala University.

Tianpei Xu, Y. M. (2021). Telecom Churn Prediction System Based on Ensemble Learning Using Feature Grouping. *Applied Sciences*, 1-12.

Tran, L. D. (2023). *Enhancing Telecom Churn Prediction: Adaboost With Oversampling and Recursive Feature Elimination Approach.* San Luis Obispo: California Polytechnic State University.

Velu, A. (2021). Customer Churn Management Using Predictive Modelling - A Machine Learning Approach. *Journal of Emerging Technologies and Innovative Research (JETIR)*, 1410-1421.

Xinyu Miao, H. W. (2022). Customer Churn Prediction on Credit Card Services using Random Forest Model. *2022 7th International COnference on Financial Innovation and Economic Development (ICFIED 2022)*, 649-656.

Yadav, R. (2024). *Machine Learning Insight into E-Commerce Churn: Prediction and Preventing Customer Loss.* Dublin: Dublin Business School.

Yashkumar Burnwal, D. R. (2023). A Comprehensive Survey on Prediction Models and the Impact of XGBoost. *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, 1552-1556.