

Master's Programme in Computer, Communication and Information Sciences

# Predictive Modelling of User Engagement for Subscription Retention

---

**Angeline Oktaviana Eka Pratiwi Jayanegara**

© 2025

This work is licensed under a [Creative Commons](https://creativecommons.org/licenses/by-nc-sa/4.0/) “Attribution-NonCommercial-ShareAlike 4.0 International” license.



---

**Author** Angeline Oktaviana Eka Pratiwi Jayanegara

---

**Title** Predictive Modelling of User Engagement for Subscription Retention

---

**Degree programme** Computer, Communication and Information Sciences

---

**Major** Machine Learning, Data Science, and Artificial Intelligence

---

**Supervisor** Prof. Aki Vehtari

---

**Advisor** Dr. Sini Rautio

---

**Collaborative partner** Sanoma

---

**Date** 31 July 2025

**Number of pages** 41+1

**Language** English

---

**Abstract**

The rise of digital news and media subscriptions has made subscription based business models increasingly important for news and media organisations such as Sanoma. In this context, understanding and predicting subscription churn of Helsingin Sanomat (main product of Sanoma Media Finland and one of the largest news provider in Finland), as well as identifying the key factors behind it, is essential for ensuring the long term sustainability of the digital media.

This thesis investigates subscription retention and churn through predictive modelling, comparing a classical logistic regression model with a more advanced ensemble method, XGBoost. The main objective was not only to improve the accuracy of subscription churn prediction but also to interpret the driving factors behind the churn, thus providing insights for retention strategies.

The analysis presents that XGBoost outperformed logistic regression by effectively capturing nonlinear relationships and interactions within the data. For model interpretation, feature importance analysis, gain and SHAP values were applied to the XGBoost model. The SHAP analysis indicated that customers with longer subscription histories and higher engagement activities were less likely to churn. However, these results should be interpreted as insights rather than causal conclusions.

---

**Keywords** machine learning, data science, logistic regression, XGBoost, subscription retention, feature importance analysis, predictive modelling

---

## **Preface**

I would like to thank Professor Aki Vehtari for his insightful suggestions and guidance which were essential to the completion of this thesis.

I am deeply grateful to my advisor, Dr. Sini Rautio, for the opportunity to work on this intriguing topic in collaboration with Sanoma, and for the valuable support throughout the process. My thanks also go to Juha Manninen and the rest of the Sanoma team for helping me familiarise myself with the data, as well as thoughtful suggestions and discussions.

Finally, I want to thank my husband, my parents, my brother, and my close friends for their endless support. Their love and encouragement have kept me going and have brightened my thesis journey. For these, I am forever grateful to have them in my life.

Vantaa, 31 July 2025

Angeline Jayanegara

# Contents

<b>Abstract</b>	<b>3</b>
<b>Preface</b>	<b>4</b>
<b>Contents</b>	<b>5</b>
<b>Abbreviations</b>	<b>7</b>
<b>1 Introduction</b>	<b>8</b>
1.1 Context and Motivation . . . . .	8
1.2 Problem Statement . . . . .	8
1.3 Research Questions . . . . .	9
1.4 Objectives . . . . .	10
<b>2 Background</b>	<b>11</b>
2.1 Factors Influencing Subscription Retention . . . . .	11
2.2 Machine Learning for Churn Prediction . . . . .	11
2.2.1 Logistic Regression . . . . .	11
2.2.2 Gradient Boosted Trees - XGBoost . . . . .	12
2.2.3 Time-Series and Unsupervised Learning . . . . .	13
2.2.4 Related Work on Churn Prediction and Data-Driven Decision Making . . . . .	13
<b>3 Data &amp; Methodology</b>	<b>16</b>
3.1 Data Overview . . . . .	16
3.2 Data Preprocessing and Feature Engineering . . . . .	16
3.3 Exploratory Data Analysis . . . . .	18
3.3.1 Correlations . . . . .	18
3.4 Model Development, Validation & Evaluation . . . . .	22
3.4.1 Model Development of Logistic Regression . . . . .	22
3.4.2 Model Development of XGBoost . . . . .	23
3.4.3 Evaluation Metrics . . . . .	24
<b>4 Results</b>	<b>26</b>
4.1 Model Performance Comparison . . . . .	26
4.1.1 Logistic Regression . . . . .	27
4.1.2 XGBoost . . . . .	27
4.1.3 Comparison . . . . .	28
4.1.4 Model Sensitivity to Input Features . . . . .	28
4.2 Insights from Predictive Modelling . . . . .	29
4.2.1 Feature Importance Analysis . . . . .	29
<b>5 Summary &amp; Future Work</b>	<b>35</b>



## Abbreviations

AI	Artificial Intelligence
AUC	Area Under Curve
CART	Classification and Regression Trees
CHAID	Chi-squared Automatic Interaction Detection
EDA	Exploratory Data Analysis
FIAM	Finnish Internet Audience Measurement
FN	False Negative
FP	False Positive
GBT	Gradient Boosted Trees
GDP	Gross Domestic Product
INMA	International News Media Association
IoT	Internet of Things
JOMBS	Journal of Media Business Studies
ML	Machine Learning
NaN	Not a Number
ROC	Receiver Operating Characteristic
RNNs	Recurrent Neural Networks
SHAP	SHapley Additive exPlanations
SOM	Self Organizing Map
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
XGBoost	Extreme Gradient Boosting

# 1 Introduction

## 1.1 Context and Motivation

In this current digital era, most people obtain their news information online. The number of print newspaper consumers is declining, as digital news offers greater accessibility, real-time updates, and personalised content.

The decline in print newspaper subscriptions has encouraged news providers to focus more on digital subscriptions, which have become the primary source of subscription revenue from customers. In addition, news organisations worldwide have faced a substantial decline in advertising revenues, further increasing the importance of subscription revenue for their sustainability. For instance, in the United States, newspaper advertising revenues shrank from nearly \$50 billion in 2000 to less than \$20 billion in recent years, with the share of advertising in total revenues dropping from 82% to 65% [1]. While the share of digital advertising has increased over time, reaching 48% of newspaper advertising revenue in 2022, it has not fully compensated for the overall decline in total advertising income [2]. This shift has placed additional emphasis on subscription revenue as a critical source of income for news organisations, making subscription retention a central concern. Therefore, understanding what drives customer engagement and churn is paramount to the sustainability of digital news organisations.

Helsingin Sanomat [3] is one of the largest news providers in Finland. It is owned by Sanoma Media Finland, part of the Sanoma Corporation. According to Finnish Internet Audience Measure (FIAM) in June 2025, Helsingin Sanomat reached more than 1.5 million people and was ranked fourth overall in toplist media of Finland [4]. Like many other news organisations, Helsingin Sanomat has experienced a growing shift from print to digital platforms as readers increasingly consume news online. This shift underscores the importance of understanding customer behaviour and retaining subscribers in the digital environment.

Motivated by these developments, this thesis, in collaboration with Helsingin Sanomat, studies the relationship between user engagement and subscription retention on the digital platform by applying predictive modelling. By analysing engagement patterns and identifying factors associated with retention, this research aims to provide valuable insights to Helsingin Sanomat in developing data-driven strategies to improve subscription retention in the competitive digital news landscape.

## 1.2 Problem Statement

Despite the opportunities of digitalisation, traditional news organisations are introduced to several challenges. One major issue is news avoidance, often driven by political instability and societal crises. Distressing content, particularly related to political crises and armed conflicts, can cause emotional fatigue among readers, causing them to disengage from the news. [5]

In addition, the rapid growth of content creators, streaming platforms, and AI-generated media has intensified competition for user attention. As Piechota [6] stated, media organisations are no longer competing only with each other, but also with entertainment platforms, such as Spotify and Netflix. This shift increases the cost of acquiring and maintaining audience attention, requiring greater investments in marketing and user retention strategies.

Customer churn, typically used to describe customers who terminate a relationship with a business during a specific period or at the end of a contract, poses a critical challenge for subscription-based news organisations [7, 8]. According to previous studies, customer retention returns tend to yield higher long-term margins compared to targeting new customers [9]. Acquiring new customers could cost five to 25 times more than retaining existing customers [8, 10]. Therefore, customer retention strategy has been widely encouraged in both academic literature and business practice [11, 12, 13, 14]. These studies emphasised the importance of retention not only to improve customer lifetime value but also to its direct impact on the bottom line of the company.

Due to the increasing challenges of news avoidance and competition, understanding subscription retention and predicting customer churn have become crucial for digital news organisations. Thus, it is essential to understand the factors that contribute to subscription retention. Analysing feature importance in churn prediction models enables the identification of key behavioural, usage, and possible external factors that influence customer decisions to continue or cancel their subscriptions. These insights can guide digital news organisations in designing early, targeted retention strategies and personalised interventions that encourage long-term subscriptions, thereby supporting the sustainability of media companies in an increasingly competitive digital environment. Thus, this research not only aims to develop models that identify customers who are more likely to churn, but also uncovers the most relevant factors driving subscription retention, enabling actionable strategies for improving customer lifetime value.

### **1.3 Research Questions**

To address the challenge of improving subscription retention in digital news platforms using data-driven approaches, this thesis investigates the following research questions:

1. What factors contribute to long-term subscription retention in Helsingin Sanomat's digital platform?
2. How can machine learning models be applied to predict at-risk subscribers using engagement and user data?
3. Which machine learning model offers better performance for subscription churn prediction?
4. Which features are most important in predicting subscription retention, and can they provide insights for retention strategies?

## **1.4 Objectives**

The objective of this thesis is to develop predictive models for forecasting subscription retention and churn on Helsingin Sanomat's digital platform, with the aim of deriving actionable insights to support customer retention strategies. To achieve this, the following specific objectives are set:

- Develop supervised machine learning models, including logistic regression and gradient boosting with XGBoost, to predict subscription retention.
- Identify and analyse key features influencing retention using digital usage data and historical factors.
- Evaluate the performance of the predictive models to assess their applicability for practical use.

## 2 Background

### 2.1 Factors Influencing Subscription Retention

According to Journal of Media Business Studies (JOMBS), subscription purchase predictions are often driven by variables such as recency, frequency, and monetary value of customers. In contrast, churn prediction tends to be influenced more by behavioural variables, such as time spent on the platform, frequency of visits, and number of actions taken per visit. These behavioural variables are parts of the users' engagement patterns and can serve as important predictors for identifying at-risk subscribers. [15]

*News media companies leverage games, bundles to encourage habits* article from the International News Media Association (INMA) highlights the role of habit-building features in improving retention on the digital news platform. Such features include newsletters, push notifications, horoscopes, games, cooking recipes, stock market updates, and sports coverage. These features encourage regular usage of the platform, creating user habits that can contribute to long-term retention. [16]

A notable example is *The New York Times*, a leading news provider in the U.S., which has successfully leveraged habit-forming content to enhance retention. Incorporation of games such as Wordle [17] and other entertainment-based content has significantly increased user visits and subscriptions to the platform [18]. Inspired by these developments, this thesis incorporates features such as playing games and reading comics as part of the analysis of churn prediction and feature importance evaluation. By including these variables alongside traditional behavioural and engagement metrics, this research aims to assess their potential influence on subscription retention within Helsingin Sanomat's digital platform.

### 2.2 Machine Learning for Churn Prediction

Machine learning (ML) methods offer powerful approaches for modelling churn prediction and customer retention. In this research, the ML method used for predictive modelling is supervised learning, which utilises training data that includes correctly labelled data points [19]. Supervised learning is also the most applied ML method in practice [20], particularly for churn prediction tasks where labelled historical data on customer retention and churn is available. Algorithms such as logistic regression, decision trees, gradient boosting, and other ensemble methods are widely employed in this context to identify customers at risk of churn and to uncover patterns within user behaviour and subscription data [21, 22].

#### 2.2.1 Logistic Regression

Logistic regression, also known as *logit regression*, is a supervised machine learning method designed for binary classification. It classifies data into two categories with binary labels and returns a conditional probability distribution of the label  $y$  given the input features  $x$ , using the logistic (sigmoid) function: [19, 23]

$$p_\alpha[y|x] = \frac{1}{Z(x)} \exp(y \alpha \cdot \Phi(x)) \quad (1)$$

In equation 1,  $\alpha$  represents the weight vector, and  $\Phi(x)$  the feature transformation. Despite its name, logistic regression is a *classification algorithm* and is implemented as a linear model in libraries such as `scikit-learn`. It is also known in the literature as *maximum-entropy classification (MaxEnt)* or the *log-linear classifier*.

Logistic regression is widely utilised as a baseline due to its simplicity and interpretability. However, its performance can be limited for a more complex task such as churn prediction. For instance, while it performs comparably to ensemble methods in purchase prediction, it often underperforms in churn prediction, where models such as gradient boosting tend to yield better results. [24]

## 2.2.2 Gradient Boosted Trees - XGBoost

Gradient Boosted Trees (GBT) is an ensemble method that fits regression trees as weak learners to the gradients of a loss function and improves predictions iteratively. [20]

The following equation 2 is the form of each individual regression tree  $f_m$  with  $J_m$  as the number of leaf regions in the tree  $m$ , and  $w_{jm}$  ((In general,  $w_{jm}$  is a vector) as the predicted output for region  $R_{jm}$ . [20]:

$$F_m(x) = \sum_{j=1}^{J_m} w_{jm} \mathbb{I}(x \in R_{jm}), \quad (2)$$

Among various supervised learning techniques, GBT has demonstrated exceptional performance in numerous real-world applications and standard classification benchmarks. These models work in a sequential manner, allowing the algorithm to focus on difficult-to-predict instances, resulting in it being particularly robust and accurate. [25] It is a commonly used method that achieves state-of-the-art performance on many classification and regression tasks [26, 25].

XGBoost (Extreme Gradient Boosting) is a widely adapted and scalable implementation of GBT. It is highly efficient since it includes several improvements such as regularisation on tree complexity to control overfitting, second-order (Hessian) approximation of the loss to improve optimisation, samples features at split points similar to random forests to enhance diversity among trees, and utilises various computer science methods to ensure scalability [20]. Additionally, the model includes various hyperparameters that allow fine-tuning of the model behaviour [27].

Hyperparameters can be tuned to control how machine learning models behave. Hyperparameters that start with `max_` tend to increase the complexity of the model when their values are increased, whereas those that start with `min_` simplify the model when their values are increased. Regularisation parameters such as `learning_rate`, `gamma`, `reg_alpha`, and `reg_lambda` control the overall complexity of the model. Finally, hyperparameters such as `scale_pos_weight` and `max_delta_step` are tuned to handle imbalanced data. [27]

An effective applied technique to also decrease the possibility of overfitting in XGBoost is early stopping. Early stopping halts the training process when the performance on a validation set no longer improves, thereby preventing overfitting to the training data and enhancing the general performance of the model. [27]

XGBoost is often considered a black box model due to its complexity and the difficulty of interpreting its ensemble of potentially hundreds of decision trees. Although it is technically possible to trace individual decision tree paths within the model, doing so would be impractical in most real-world scenarios. Consequently, in many business settings where interpretability is paramount, simpler “white box” models are sometimes preferred despite their typically lower predictive performance. [27]

Nevertheless, XGBoost is a scalable implementation of gradient boosting, known for its speed and high predictive performance. It has been successfully applied in numerous industrial and academic settings, consistently achieving state-of-the-art results in machine learning competitions and practical tasks. Due to its effectiveness, scalability, and ability to deliver state-of-the-art results, XGBoost is selected as one of the core models in this thesis to study customer churn prediction in the context of subscription data, providing a strong benchmark for predictive modelling in comparison to other models.

### 2.2.3 Time-Series and Unsupervised Learning

Some ML methods do not require knowing the label value of any data point and are therefore referred to as unsupervised ML methods. In the customer segmentation, unsupervised learning methods such as K-means clustering can be applied to group users with similar behaviours based on feature vectors  $\mathbf{x}^{(i)} \in \mathbb{R}^n$ . Hard clustering methods assign each user to one and only one cluster  $C^{(c)}$ , helping to identify distinct user personas or churn risk profiles.[19]

In the context of churn analysis, time-series analysis could be utilised to examine temporal engagement trends, helping to detect churn patterns over time.

However, these methods are not applied in this thesis. Unsupervised learning and time-series analysis, while valuable for exploratory analysis and segmentation, are less suitable for directly predicting churn probabilities or for quantifying feature importance, which are primary to this study’s objectives. Instead, this research focuses on supervised learning methods that explicitly utilise churn labels to predict customer churn and analyse feature importance to understand the factors influencing subscription retention.

### 2.2.4 Related Work on Churn Prediction and Data-Driven Decision Making

Subscription retention and churn prediction using machine learning models is a widely researched topic in today’s highly competitive digital business environment. By understanding subscription retention with accurate churn prediction models, businesses can identify factors contributing to the churn, and thus enable companies to implement preventive measures and reduce revenue loss.

Previous studies have investigated the use of various machine learning models, such as logistic regression, random forests, and gradient boosting, for subscription retention and churn prediction. Boozary et al. [28] conducted a comparative analysis of ensemble models, demonstrating that these methods increase the accuracy of churn prediction. Mach-Król and Hadasik [29] and Hassouna et al. [30] applied decision trees and logistic regression using real-world data from an English mobile service provider. The study included multiple decision tree algorithms (CART, C5.0, and CHAID) and evaluated their performance with logistic regression using AUC, top decile performance, and accuracy. The C5.0 algorithm outperformed logistic regression with an AUC of 0.763, challenging earlier findings by Owczarczuk [31].

Several studies have highlighted the trade off between model performance and interpretability. Onari et al. [32] integrated SHAP (SHapley Additive exPlanations) values with decision trees and logistic regression to improve the interpretability of credit scoring, while Chen and Guestrin [25] introduced XGBoost, a scalable tree boosting system that has demonstrated high accuracy in churn prediction but is less interpretable due to its complexity.

Table 1 summarises a selection of related works, their objectives, methods used, and key findings.

The review of related work reveals a lack of comprehensive comparisons between classical machine learning models (e.g., Logistic Regression, K-nearest neighbors) and ensemble techniques (e.g., XGBoost, random forest) within the context of churn prediction. This thesis addresses this gap by applying logistic regression as the classical model and XGBoost as an ensemble technique with a detailed performance assessment using multiple metrics such as confusion matrix, AUC score and ROC curve to compare their effectiveness in subscription churn prediction.

**Table 1:** Summary of Related Work in Churn Prediction with Machine Learning

Reference	Methods	Objective	Key Findings
Boozary et al. [28]	Ensemble Models	Churn prediction in customer retention	Comparative analysis showed ensemble models enhance churn prediction accuracy
Onari et al. [32]	Decision Trees, Logistic Regression	Explainable credit scoring	Improvement of SHAP interpretability and guided rejected customers
Eslami et al. [33]	SOM, Decision Tree	IoT customer segmentation	Identified satisfaction and loyalty factors
Rudd et al. [8]	Deep Learning, Causal Analysis	Churn prediction with causality	Combined deep learning with causal inference to improve churn understanding
Mach-Król & Hadasik [29]	CART, C5.0, CHAID, Logistic Regression	Churn prediction	With AUC 0.763, C5.0 outperformed logistic regression
Hassouna et al. [30]	Decision Trees, Random Forest	Telecom churn	Random Forest managed large data well
Chen & Guestrin [25]	Tree Boosting (XG-Boost)	Scalable ML system	Provided high-performance, scalable boosting widely used in churn prediction
Owczarczuk [31]	Logistic Regression, SVM	Comparing churn models	Highlighted logistic regression limitations
Chandar & Krishna [34]	Predictive Data Mining	Bank customer churn	Modeled churn behaviour using predictive data mining in banking

## 3 Data & Methodology

### 3.1 Data Overview

This research utilises dataset containing 35 features, including variables such as user demographics, reading volume, game activity, and comic consumption. The study aims to capture user engagement with these features to understand and predict churn behaviour.

The dataset comprises six months of subscriptions data, covering the period from 1 August 2024 to 31 January 2025, and is limited to subscribers with longer-term subscriptions, excluding customers on trial subscriptions. Digital usage data were collected for 30-day window before the billing period ends, meaning the actual observed digital activity spans from 1 July 2024 to 31 January 2025.

The target variable in this study is subscription churn, defined as whether a subscription continues to the next billing period. The data includes various feature groups that capture different aspects of user interaction and customer profile. These features are grouped as follows:

- Digital usage data: reading frequency, session duration, number of articles read, and content preferences.
- Subscription history data: subscriptions renewal/cancellation timestamps, duration of active subscriptions.
- Customer data: demographic (municipality), sales channel

### 3.2 Data Preprocessing and Feature Engineering

Data preprocessing and feature engineering are essential steps in preparing the dataset for modelling. In this thesis, preprocessing and feature engineering are to address the characteristics and challenges of the subscription dataset, including missing values, potential anomalies, and categorical variables.

Missing values, represented as NaN, were imputed with 0 to ensure compatibility with machine learning algorithms. In addition, a threshold-based approach was applied by removing some extreme outliers. These extreme outliers are potentially originated from internal testing accounts created by Helsingin Sanomat employees. They did not reflect typical user behaviour and, if included, could distort model training.

Categorical variables were transformed using one-hot encoding, which represents each category as a separate binary feature. This approach was applied to variables such as the outbound channel, invoice method, and municipality (e.g., Uusimaa), allowing the models to effectively process non-numeric inputs.

Furthermore, the numerical data were standardised to have zero mean and unit variance. This step is especially essential for models such as logistic regression that are sensitive to differences in feature scale.

Table 2 displays the summary of the data preprocessing steps.

The next sections provide a more elaborate and detailed explanation of one-hot encoding and feature scaling, including the application in the context of this study.

**Table 2:** Summary of Data Preprocessing Steps

Step	Description
Handling Missing Values	Replaced NaN with 0
Outlier Removal	Removed unrealistic values based on domain knowledge (e.g., test accounts with extreme feature values)
Categorical Encoding	Applied one-hot encoding to variables like out-bound channel, invoice method, and municipality
Feature Scaling	Applied standardisation for logistic regression data to ensure zero mean and unit variance

### Encoding Categorical Features: One-Hot Encoding

In predictive modelling, many machine learning methods, particularly linear models such as logistic regression, require numerical input features. However, real-world data, including the data in this study, contain categorical variables that should be transformed into a numerical format without introducing unintended biases. One-hot encoding is a commonly applied feature engineering technique for handling such variables.

One-hot encoding method transforms categorical data into multiple binary features, each representing a distinct category. Each of the variables cannot belong to multiple categories at once. For example, if the feature `invoice_method` includes three categories: *Credit Card*, *E-mail*, and *Bank Invoice*, one-hot encoding will generate three new binary columns—one for each payment method. Table 3 displays this example. Each data point will have a value of 1 in the column corresponding to its category and 0 in the others [35].

**Table 3:** One-hot encoding of the `invoice_method` column

	Credit Card	Bank Transfer	E-mail
Subscriber 1	1	0	0
Subscriber 2	0	1	0
Subscriber 3	0	0	1

This approach avoids assigning arbitrary numerical labels to categories (e.g., *Credit Card* = 0, *Bank Invoice* = 1, *E-mail* = 2), which could mislead models like logistic regression into interpreting an ordinal or metric relationship where none exists. By encoding each category as an independent dimension, one-hot encoding ensures that the model treats categorical values without implying any hierarchy or distance between them.

Thus, one-hot encoding is essential in preparing categorical variables for models that assume numeric input, preserving the values and interpretability of the data.

### Feature Scaling: Standardisation

In many real-world datasets, including the dataset in this study, input features are often on different scales: age in years, income in euros, or usage in minutes. This variation can affect the performance of machine learning models such as logistic regression that is sensitive to the relative magnitudes of the input features. To address this potential bias in the learning process, feature standardisation was utilised using `StandardScaler` from the `scikit-learn` library. This is done to prevent overfitting and to ensure that the model treats all features on the same scale.

Standardisation transforms each feature to have a mean of zero and a variance of one. For a given feature  $x_d$ , this is done by subtracting its sample mean  $\hat{\mu}_d$  and dividing by its sample standard deviation  $\hat{\sigma}_d$  [20]:

$$\text{standardise}(x_{nd}) = \frac{x_{nd} - \hat{\mu}_d}{\hat{\sigma}_d} \quad (3)$$

where the sample mean and variance are computed as:

$$\hat{\mu}_d = \frac{1}{N} \sum_{n=1}^N x_{nd} \quad (4)$$

$$\hat{\sigma}_d^2 = \frac{1}{N} \sum_{n=1}^N (x_{nd} - \hat{\mu}_d)^2 \quad (5)$$

## 3.3 Exploratory Data Analysis

Exploratory Data Analysis (EDA) aims at exploring and understanding the data through summary statistics and visualisations. This process is essential to gain deeper insight into the data as well as to understand the relationships between the features and the target label. A solid understanding of the data would enable more informed modelling decisions.

### 3.3.1 Correlations

Correlation coefficients are commonly applied to examine the relationships between numerical features. The most widely applied method is the Pearson correlation coefficient, which measures the strength of linear relationships. Alternatively, the Spearman correlation coefficient can be applied to assess monotonic relationships, especially when data are nonlinear or ordinal in nature. [27]

The Pearson correlation coefficient is defined as:

$$\rho_{X,Y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \quad (6)$$

where  $x_i$  and  $y_i$  are paired observations, and  $\bar{x}$ ,  $\bar{y}$  their respective sample means [36].

The feature correlation matrix shown in figure 1 shows the Pearson correlation values. As observed, several features in the middle of the matrix show strong positive correlations, highlighted in warmer red tones. These features are primarily associated with the volume of reading articles, the reading frequency, and the session duration, which logically tend to be interrelated.

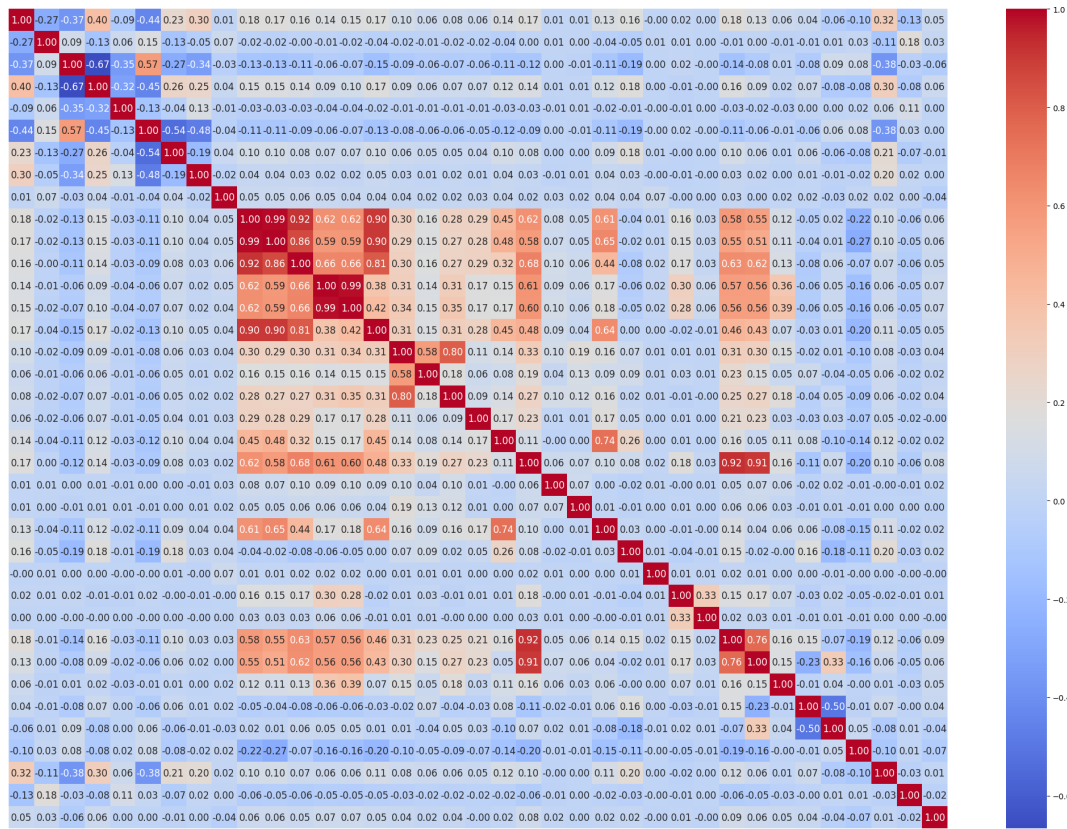
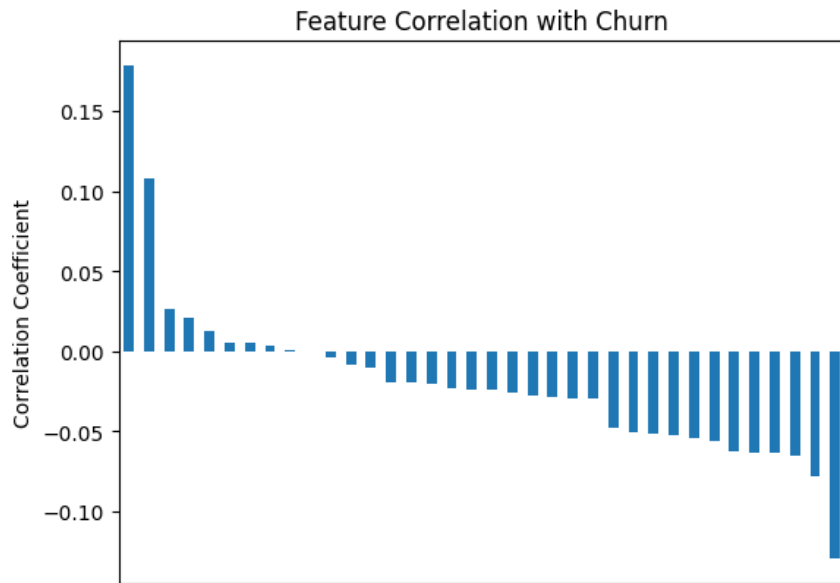


Figure 1: Feature Correlation Matrix

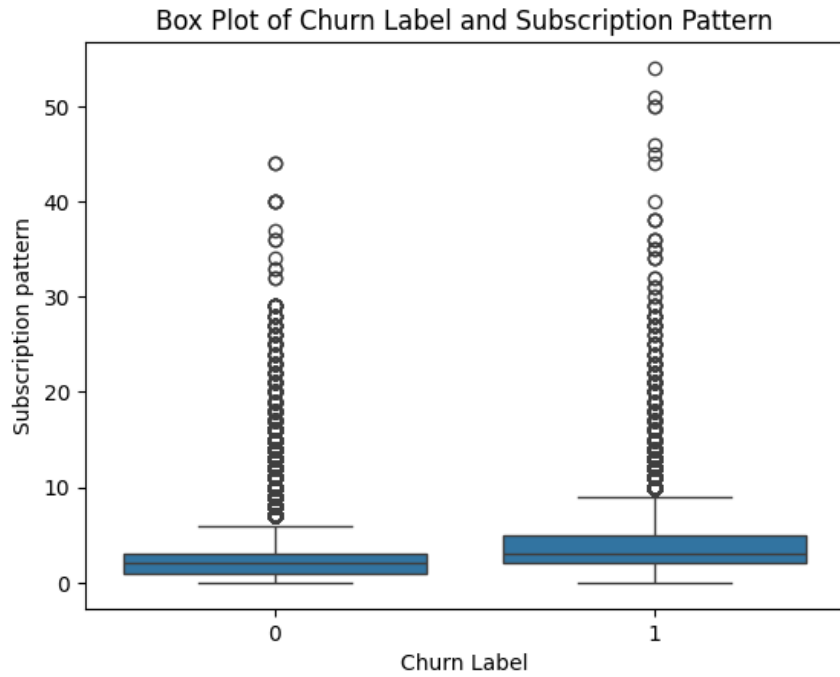


**Figure 2:** Features Correlation with Churn

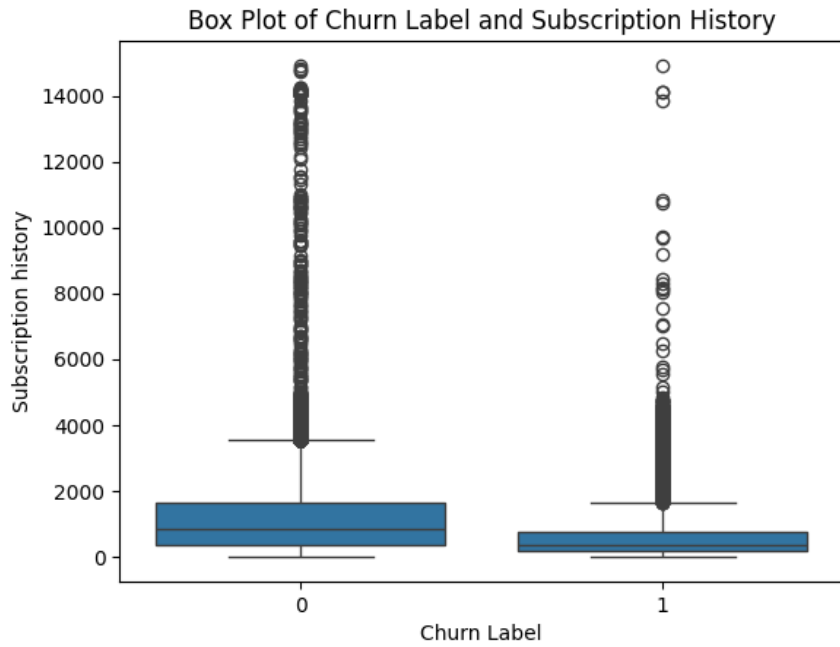
From the correlation calculations, we obtain the highest positively churn-correlated features with  $\rho = 0.178$ ,  $\rho = 0.108$  and  $\rho = 0.026$ , which are also displayed on figure 2. The features are related to user subscription patterns, invoicing method and user engagement.

In contrast, the subscription history feature has a negative correlation ( $\rho = -0.130$ ), indicating that lower values in this feature are associated with a higher correlation with churning.

These relationships are then further examined utilising various plots, such as box plots.



**Figure 3:** Box plot of subscription patterns and churn label



**Figure 4:** Box plot of subscription history and churn label

From the box plots in Figures 3–4, we can observe the relationships between selected features and the churn\_label. Figure 3, *Subscription patterns and churn label*, illustrates that users with lower values in subscription patterns feature generally

have lower average churn rates (label 0) compared to those with higher values in this feature, suggesting that frequent subscription repetitions are associated with an increased likelihood of churn.

In the case of subscription history feature as in figure 4, a longer history in this feature appears to correlate with a reduced likelihood of churn, indicating that customers with longer engagement histories are less prone to churning.

Similarly, higher values of the features that capture user activity are associated with lower churn rates, suggesting that users who engage more frequently with the service tend to remain subscribed.

### 3.4 Model Development, Validation & Evaluation

This section presents the development, validation, and evaluation of the predictive models used in this study. The main models are logistic regression and the XGBoost model implemented in Python using several libraries including Scikit-learn, XGBoost, NumPy, pandas, PySpark, matplotlib, and Optuna.

The data is initially split randomly with predetermined seed into 80% for training and validation, and 20% for testing. The 80% portion is then further divided into 70% training and 30% validation, resulting in 56% for training, 24% for validation, and 20% for testing overall.

#### 3.4.1 Model Development of Logistic Regression

The first model implemented in this thesis is Logistic Regression, utilising the implementation provided by the `scikit-learn` library [37]. In practice, `scikit-learn`'s implementation supports binary, one-vs-rest, and multinomial classification tasks, with optional  $L1$ ,  $L2$ , or Elastic-Net regularisation [37]. Regularisation technique controls the complexity of the model to prevent ML methods from overfitting [38, 19]. Notably, regularisation is applied by default within this library, thus aiding in mitigating overfitting during model training.

The dataset utilised in this thesis is highly imbalanced, with significantly fewer churn labels compared to non-churn. This class imbalance poses challenges for model performance, especially in predicting the minority class. To address this, class balancing is applied using the `class_weight` parameter in `scikit-learn` Logistic Regression implementation:

$$\{\text{class\_label} : \text{weight}\}$$

Prior to balancing, all classes are assigned equal weights, resulting in models that tend to favour the majority class and underpredict churn instances. By setting `class_weight` to “balanced”, the algorithm automatically adjusts the class weights in proportion to the inverse of the class frequencies:

$$\frac{n_{\text{samples}}}{n_{\text{classes}} \times \text{np.bincount}(y)}$$

This adjustment ensures that higher weights are assigned to the minority churn class during model training, promoting a more balanced and unbiased representation of both classes and improving the predictive performance for churn cases.

### 3.4.2 Model Development of XGBoost

The second model implemented in this thesis is gradient boosting model using the XGBoost library [38].

Gradient boosting models are generally insensitive to monotone transformations of input features, as the split points are based on ranking the data points rather than absolute values. As a result, standardisation of input data is unnecessary for gradient boosting methods [20].

To address the significant class imbalance in the dataset, the `scale_pos_weight` [39] parameter is used to adjust the weighting of the minority class:

$$\text{scale\_pos\_weight} = \frac{\text{Number of negative samples}}{\text{Number of positive samples}}$$

The `scale_pos_weight` parameter in XGBoost helps balance the contributions of positive and negative classes for binary classification. This is especially relevant for imbalanced data of this study where the proportion of churns could be as low as 5% compared to the 95% of non-churn class. The `scale_pos_weight` adjusts the weight of churn class during the training by increasing its relative weight accordingly. [39, 40]

The XGBoost model is trained on the training set and validated on the validation set with early stopping employed to prevent overfitting. As with many decision tree-based models, the performance of XGBoost can be influenced by the choices of the hyperparameters.

### Hyperparameter Optimisation

Hyperparameter optimisation is a process of finding the optimal set of hyperparameters to improve the performance of machine learning models, particularly the XGBoost model for this study. Traditional and modern approaches are available, and both were explored.

The first approach is Grid Search, a traditional approach available in `GridSearchCV` class in scikit-learn [41]. In this method, users can manually define the hyperparameter values and try various sets of combinations. The key hyperparameters selected for tuning include `max_depth`, `learning_rate` (also known as `eta`), and `n_estimators`. It is a straightforward and commonly utilised approach, however, sometimes Grid Search can be computationally expensive and inefficient for a complex model such as XGBoost. [27]

Optuna [42], on the other hand, is a more modern and efficient automated approach to hyperparameter optimisation. It is an open-source Python library that finds optimal hyperparameter values with reduced computational cost while maintaining performance.

After exploring different sets of hyperparameters, it was found that all of these settings led to similar results. The final model was then retrained using the optimal hyperparameters and evaluated on the test set. Validation curves were also plotted to visualise the effect of hyperparameter changes on model performance.

In the following chapter, the evaluation metrics used to assess the performance of the model will be introduced and discussed in detail.

### 3.4.3 Evaluation Metrics

The evaluation metrics for the developed classification models include Confusion Matrix, AUC and ROC Curve, Accuracy, Precision, Recall, and F1 Score. These are applied to evaluate and compare the performance of the models.

#### Confusion Matrix

The confusion matrix is visualisation of the model's performance in the form of table. The table includes the number of TP (True Positive), TN (True Negative), FP (False Positive), and FN (False Negative).

- TP (True Positive) = Number of instances where the actual class is 1 and the predicted class is also 1
- TN (True Negative) = Number of instances where the actual class is 0 and the predicted class is also 0
- FP (False Positive) = Number of instances where the actual class is 0 but the predicted class is 1
- FN (False Negative) = Number of instances where the actual class is 1 but the predicted class is 0

#### Accuracy

Accuracy metric measures the proportion of correct predictions out of total number of predictions.

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

#### Precision

Precision metric measures the proportion of correct positive predictions out of the total number of positive predictions. This metric is important to decide whether the results are relevant and when false positive is costly.

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

## Recall

Recall or Sensitivity metric measures the proportion of actual positive predictions that the model correctly predicts. This metric is essential for understanding the number of relevant results that are returned.

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

## F1 Score

F1 Score is the balanced metric of precision and recall. It measures the harmonic mean of precision and recall.

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2TP}{2TP + FP + FN}$$

## ROC Curve and AUC

The Receiver Operating Characteristic (ROC) curve was originally developed during World War II to evaluate the effectiveness of radar signals in detecting enemy aircraft. In the context of Machine Learning, the ROC curve plots the True Positive rate (Recall) against the False Positive rate (fallout) across various classification thresholds, offering a view of the model's sensitivity and its false alarm rate. [27]

The Area Under Curve (AUC) summarises the ROC curve into a single value. AUC score of 0.5 indicates that the model performs similarly as random guessing. An AUC score below 0.5 suggests that the model is performing worse than random guessing.

While an AUC score that is very close to 1.0 may appear ideal, it could signal potential data leakage or an overly simple problem that does not require ML model. Therefore, while AUC is a valuable tool for identifying under-performing models, it should be interpreted with caution and considered alongside other evaluation metrics to form a comprehensive view of model performance.

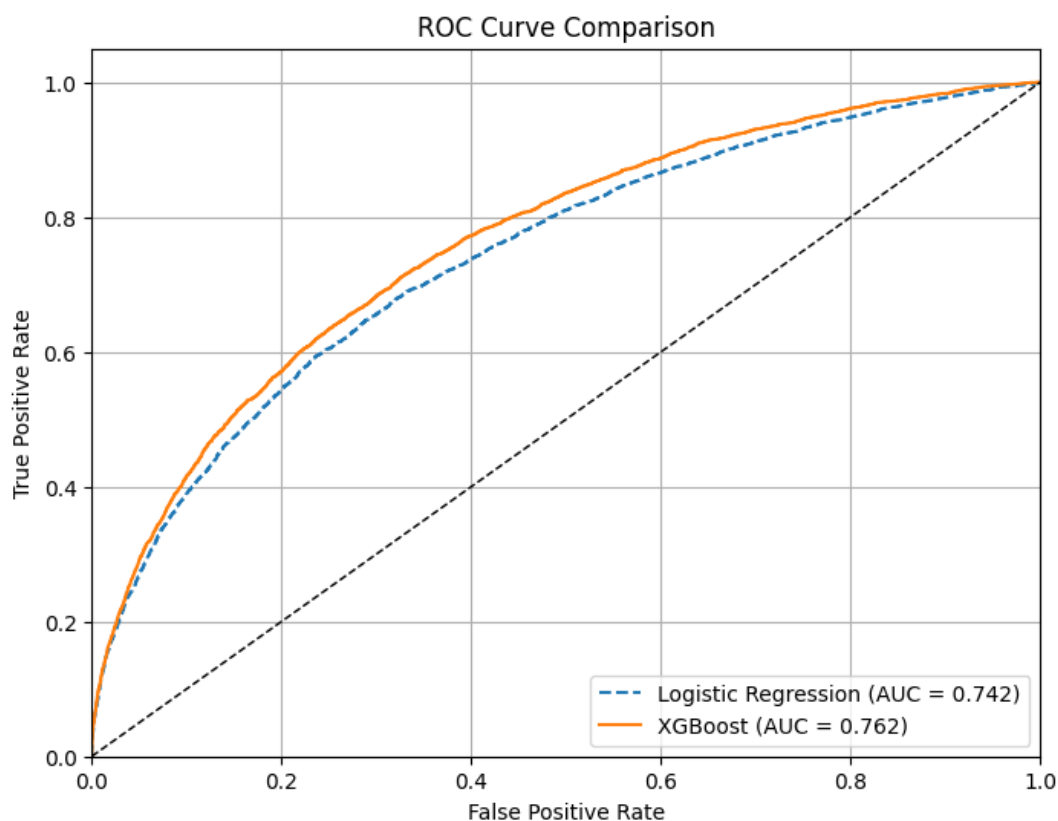
## 4 Results

### 4.1 Model Performance Comparison

This section presents the evaluation results for the Logistic Regression and XGBoost models trained on Helsingin Sanomat subscribers data for a period of six months, from 1 August 2024 to 31 January 2025. Primary performance metrics include accuracy, precision, recall, F1-score, and the area under the ROC curve (AUC). The target variable is the binary `churn_label`, with class 1 indicating churn and class 0 indicating no churn.

**Table 4:** Performance Comparison Between Logistic Regression and XGBoost

Metric	Logistic Regression	XGBoost
Accuracy	0.720	0.724
Precision (Churn)	0.14	0.20
Recall (Churn)	0.39	0.65
F1-score (Churn)	0.21	0.30
AUC	0.742	0.762

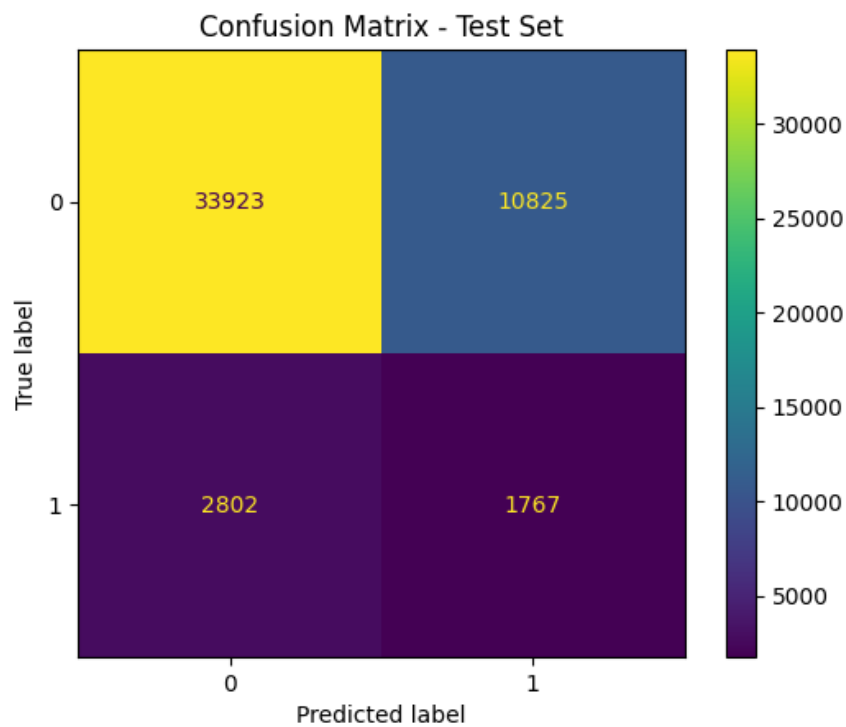


**Figure 5:** ROC Curve Comparison with AUC score

### 4.1.1 Logistic Regression

Although the logistic regression model achieves high precision (0.92) and F1-score (0.83) for class 0 (non-churners), it struggles with identifying churners (class 1), as reflected in its low precision and F1-score for that class on table 4. The moderate AUC score of 0.742 suggests the model performs better than random but has limited discrimination ability between churn and non-churn.

The confusion matrix for Logistic Regression is as shown as in Figure 6.

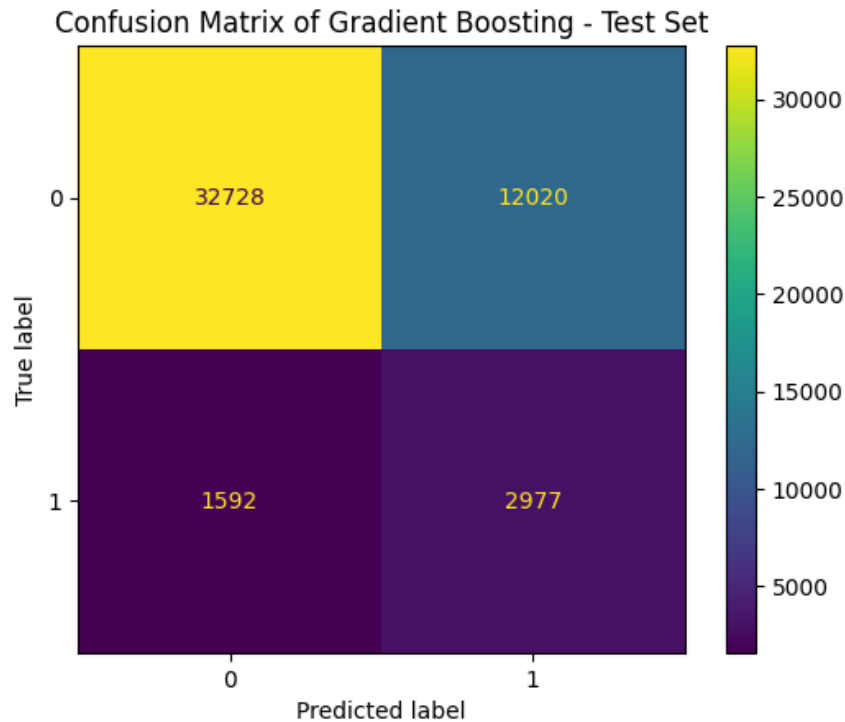


**Figure 6:** Confusion Matrix Logistic Regression

### 4.1.2 XGBoost

The XGBoost model demonstrates improved performance in key metrics. Compared to Logistic Regression, XGBoost improves both recall and F1-score for churners, which is crucial in imbalanced classification problems like churn prediction. The higher AUC of 0.762 also indicates better separation between classes.

The confusion matrix for XGBoost is as shown as in Figure 7.



**Figure 7:** Confusion Matrix XGBoost

### 4.1.3 Comparison

While both models achieve similar overall accuracy, XGBoost significantly outperforms Logistic Regression in identifying churners. It achieves a higher recall, F1-score, and AUC, making it more suitable for this task. The improved ability of XGBoost to correctly classify minority class instances (churners) is particularly important for targeted interventions and reducing customer attrition.

### 4.1.4 Model Sensitivity to Input Features

During studies, the performance of the logistic regression model sometimes changes depending on the inclusion or exclusion of certain features. This indicates that logistic regression is highly sensitive to data noise and feature selection. Minor changes in the feature set can lead to different results in metrics such as precision, recall, and AUC.

In contrast, the XGBoost model has a more stable performance. Gradient boosting methods are more robust to irrelevant or noisy features. Their iterative nature allows them to focus on harder-to-predict (noisy) samples across trees, reducing sensitivity to feature-level disturbances. It exhibits greater resilience to feature perturbations compared to simpler models [25]. This results in XGBoost as a more reliable choice for real-world churn prediction tasks where feature uncertainty is common.

Table 5 presents a comparison of logistic regression and XGBoost models for three different experiments: 1, 3, and 6 months. The metrics include recall, F1-score, and AUC, which are particularly important in churn prediction due to class imbalance.

**Table 5:** Performance Comparison of Logistic Regression and XGBoost Across Time Periods

Model	Period	Accuracy	Precision (1)	Recall (1)	F1-score (1)	AUC
Log Reg	1 mo	0.79	0.15	0.19	0.17	0.67
	3 mo	0.69	0.16	0.44	0.23	0.72
	6 mo	0.72	0.14	0.39	0.21	0.74
XGBoost	1 mo	0.73	0.21	0.54	0.30	0.71
	3 mo	0.73	0.22	0.64	0.33	0.75
	6 mo	0.72	0.20	0.65	0.30	0.76

From the 6-month results, XGBoost outperforms logistic regression in all key metrics, achieving a higher recall (0.65 vs. 0.39), F1-score (0.30 vs. 0.21), and AUC (0.76 vs. 0.74). This indicates that XGBoost is better at identifying churners while maintaining a more balanced performance. The results also show that logistic regression is more sensitive to noise and feature selection, while XGBoost remains relatively stable.

## 4.2 Insights from Predictive Modelling

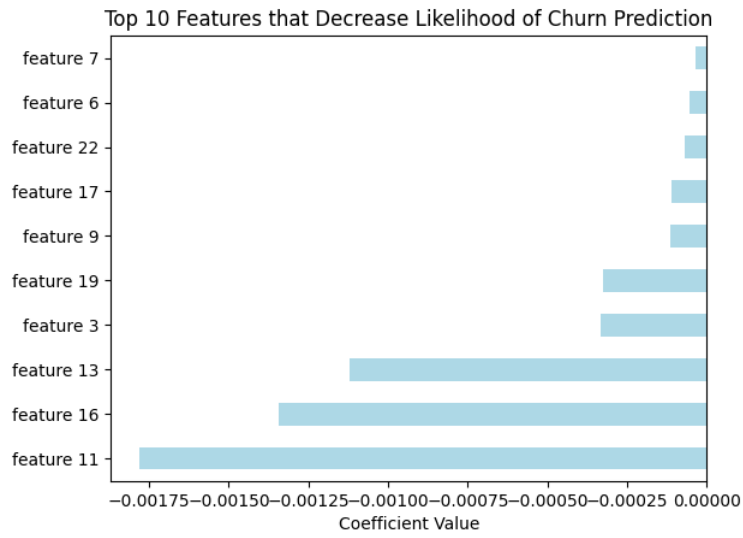
### 4.2.1 Feature Importance Analysis

One of the main objectives of this research is to interpret the key features of subscription churn. Feature importance analysis enables the understanding of which features contribute most significantly to the model's predictions. This section presents feature importance analysis for both logistic regression and XGBoost models.

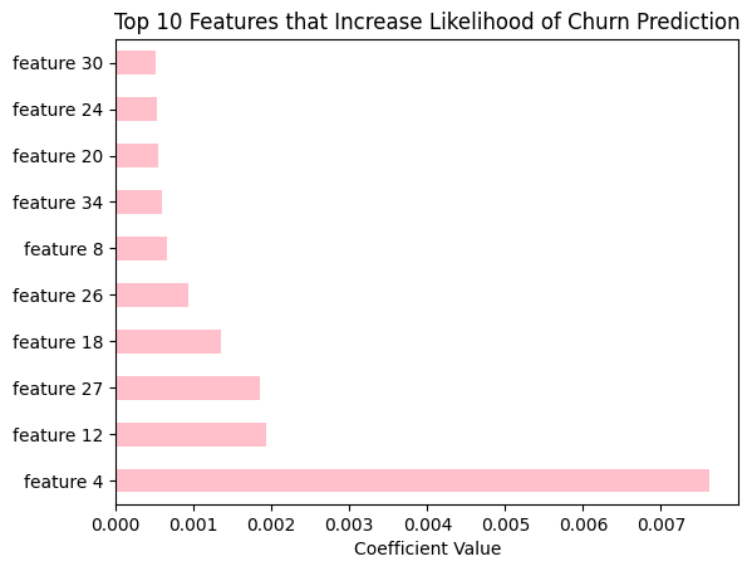
#### Feature Importance Analysis of Logistic Regression

Feature importance analysis of logistic regression is performed by using the coefficients of the trained model. Logistic regression is a linear model, therefore, the magnitude and sign of the learned coefficients directly indicate the direction and strength of each feature's influence on the probability of churn. Positive coefficients increase the likelihood of churn, whereas negative coefficients decrease it.

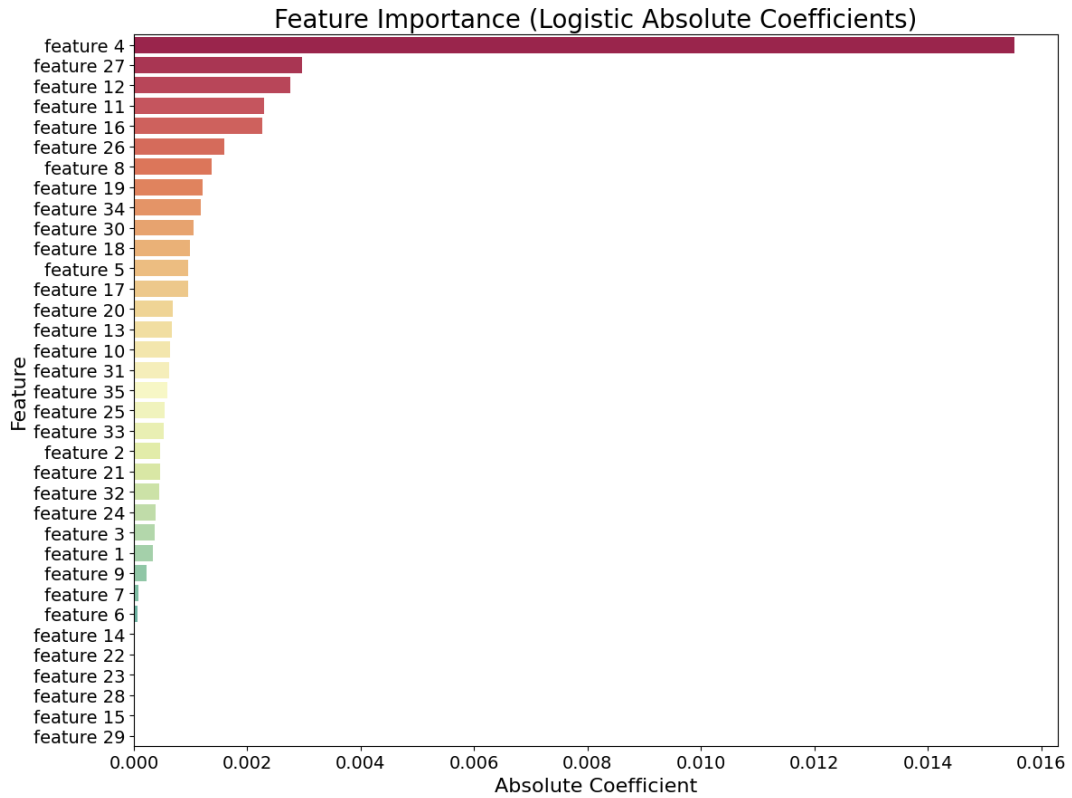
The feature coefficients are visualised by sorting them according to their values. The top ten features with the lowest (most negative) coefficients are plotted to identify the features most associated with retention as displayed in Figure 8. The top ten features with the highest positive coefficients are plotted to identify the features most associated with increase likelihood of churn as shown in Figure 9. Additionally, feature importance is visualised using the absolute values of the coefficients in Figure 10, highlighting the most important features regardless of the direction of their influence. This analysis provides insights of which user behaviours or characteristics are most predictive of churn within the dataset, supporting the interpretability of the predictive model.



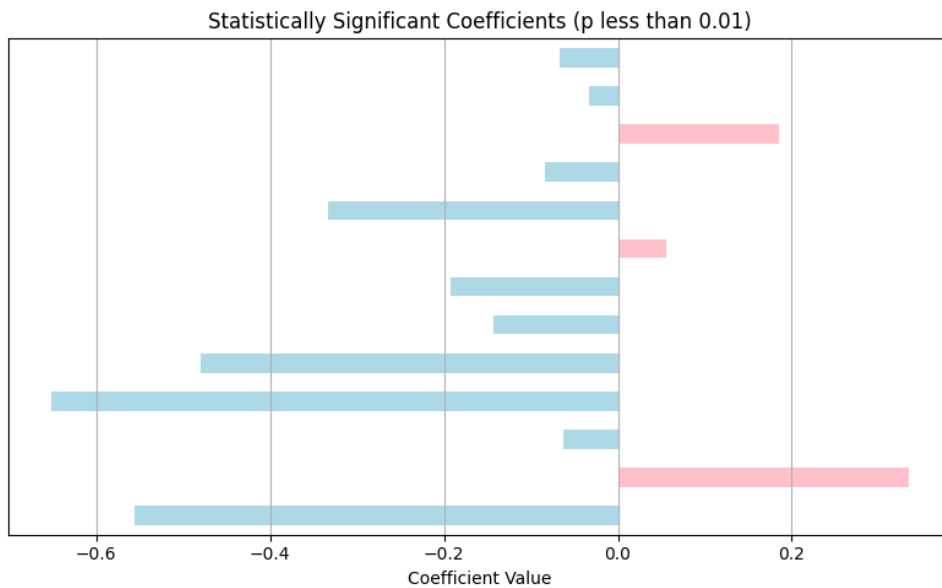
**Figure 8:** Feature importance of logistic regression (retention)



**Figure 9:** Feature importance of logistic regression (churn)



**Figure 10:** Feature importance of logistic regression (absolute coefficients)



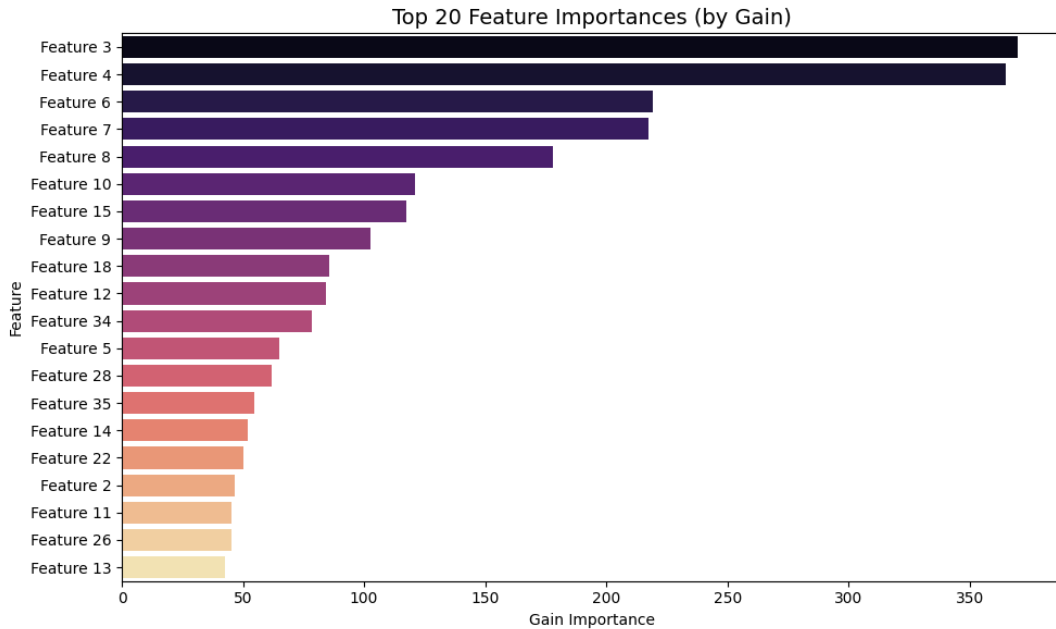
**Figure 11:** Statistically significant coefficients of logistic regression (p less than 0.01)

Figure 11 displays the statistically significant values of the characteristics with p-value coefficients less than 0.01.

## Feature Importance in XGBoost

There are several methods of feature importance analysis available for tree based models such as XGBoost. These include Gain, Split Count, and Saabas methods. However, these methods are often inconsistent because they sometimes rely more on a given feature, yet the estimated importance of that feature decreases [43]. To enhance interpretability and reliability, more consistent alternatives, such as SHAP and permutation-based importance are often recommended.

The feature importance for XGBoost in this thesis is initially evaluated using the Gain method. Figure 12 displays the top features based on Gain, identifying which ones most strongly influenced the model’s decisions. Unlike logistic regression, however, this method does not indicate whether a feature increases or decreases the likelihood of churn, only its relative contribution.

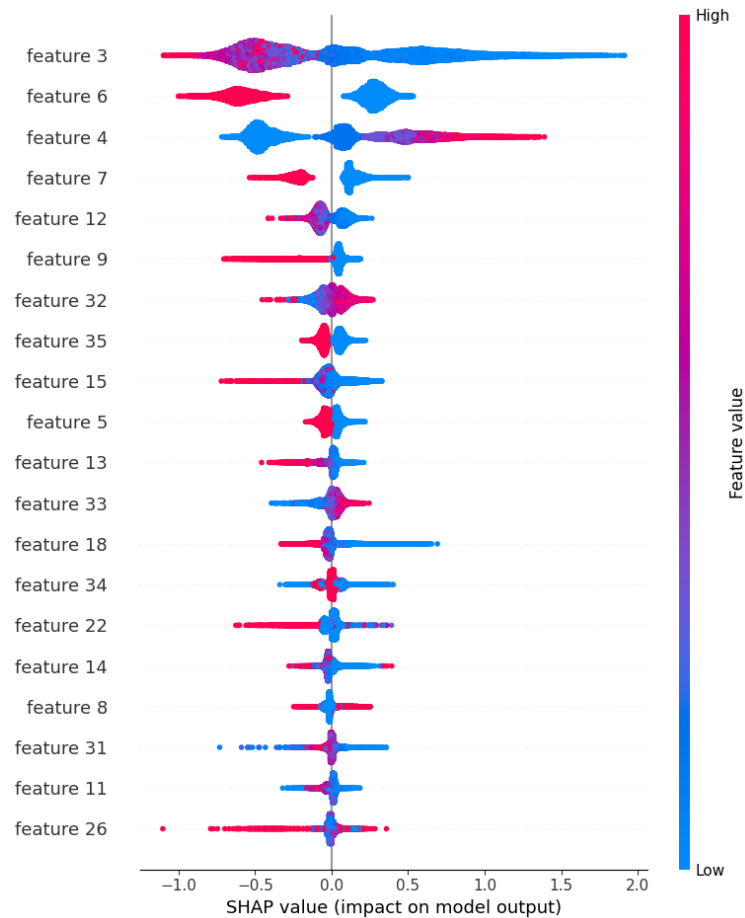


**Figure 12:** Feature Importance by Gain

SHAP (SHapley Additive exPlanations) [44] and permutation-based feature importance methods are known for more consistent performance in feature importance of tree-based models. When applied to tree-based models such as XGBoost, SHAP provides accurate explanations of how each feature contributes to model predictions on average [45]. In addition, SHAP is particularly useful to identify features that are the most important for predicting churn. It is also valuable for diagnosing and improving model performance by highlighting dominant features. Thus, to provide more robust and interpretable explanations, SHAP is also applied for the XGBoost model.

Figure 13 presents the SHAP beeswarm plot, which summarises overall impact, magnitude, and direction of the top features on the model’s output [46]. On each

feature row, each instance in the plot is represented by a single dot. “Pile up” dots along the x-axis show the density distribution. The color represents the value of a feature (`shap_values.data[instance, feature]`), with red indicating high feature values and blue indicating low values. The position of the dots along the x-axis corresponds to the SHAP value of that feature (`shap_values.value[instance, feature]`), indicating how much that feature contributed to the prediction. In this instance, it indicates whether the effect was to increase or decrease the probability of predicting churn.



**Figure 13:** Feature importance SHAP beeswarm plot

Feature 3, representing the subscription history, plays a key role in the SHAP beeswarm plot. For feature 3, the red dots predominantly appear on the left (negative) side of the SHAP value axis. This suggests that higher values of subscription history are associated with a decreased likelihood of churn, meaning that customers with longer subscription histories are more likely to stay.

This pattern is consistent with the earlier boxplot analysis shown in Figure 4, which also suggests that longer engagement durations are correlated with customer retention. Together, these observations reinforce the idea that historical engagement is a strong predictor of customer loyalty in this dataset.

Feature 32 displays a more complex pattern with red and blue dots spread across

the left and right sides of the SHAP value axis, especially the red dots. This might suggest that while higher value of feature 32 (reading and engagement feature) might generally indicate lower risk of churn, in some contexts it could also reflect last-minute activity before cancellation, thus increasing the risk of churn. These variations suggest that the impact may depend on interactions with other features or contextual factors within the data.

## 5 Summary & Future Work

The rise of digital news and media consumption has made subscription-based business models increasingly vital for media organisations. In this context, understanding and predicting subscription churn is essential for ensuring long term sustainability.

This thesis explored the application of machine learning models, particularly logistic regression and XGBoost, to predict subscription churn in digital news media domain, utilising data from Helsingin Sanomat. The main objective was not only to accurately predict the subscription churn but also to interpret the driving factors behind the churn, thus providing insight for retention strategies.

The analysis demonstrated that the ensemble model, XGBoost, generally outperformed the classical model, logistic regression, in multiple evaluation metrics, including precision, recall, F1-score and AUC. The results can be attributed to the ability of XGBoost model to capture complex nonlinear relationships and interactions between features, making it particularly well suited for modelling real world customer behaviour in churn prediction.

Although logistic regression provides transparency and easier interpretability, its linear nature limits the ability to model intricate patterns within the data. In contrast, XGBoost, despite being a more complex and less interpretable “black box” model, proved to be more effective in handling the nuances of the subscription data.

To address the interpretability challenge, SHAP was utilised in the XGBoost model. SHAP displays analysis of feature contributions, helping to uncover which variables most influenced the churn predictions. In particular, the subscription history, invoice method, and engagement features emerged as the most influential predictors of churn.

These insights may support data-driven retention strategies. For instance, SHAP values revealed that customers with longer subscription histories and higher engagement were less likely to churn. This finding aligns with behavioural expectations and is consistent with visual analyses presented in earlier chapters (e.g., boxplot in Figure 4).

While SHAP values enable statements like:

“Customers with fewer logins and lower engagement scores are more likely to churn, according to the model.”

It is important to emphasise that SHAP explanations are predictive rather than causal. As noted in the SHAP documentation:

“SHAP makes transparent the correlations picked up by predictive ML models. But making correlations transparent does not make them causal.”  
[47].

XGBoost in combination with SHAP excels in explaining model behaviour but does not establish causation. SHAP values demonstrate how the model uses input features to make its predictions, capturing the correlations the model has learned from the data. However, they do not show how changing a feature would directly influence outcomes in the real world. Therefore, SHAP explanations should be interpreted as insights and

it is important to avoid drawing causal conclusions from them [43, 47]. Understanding these limitations is crucial when translating model insights into business strategies.

### **Future Work**

Although this thesis provides a strong foundation for predicting subscription churn and features that contribute to churn, there are several possible further developments to be analysed for future research and improvement. These include exploring other advanced predictive models, applying causal inference methods, incorporating external economic indicators, and developing hybrid approaches that combine interpretability with high performing predictive models.

The first potential future development is to explore causal inference methods to understand beyond correlation-based insights. In this study, SHAP and feature importance methods help interpret model predictions. However, these methods do not explain why certain features lead to churn. Causal graphs, do-calculus, or uplift modelling can provide a better understanding of the causal effects of interventions. Causal inference methods can also help predict changes in the outcome of variables, which can be especially essential to support more effective and targeted retention strategies.

In terms of models, other advanced predictive models and deep learning models can be explored. The application of Recurrent Neural Networks (RNNs) or transformer (deep learning architectures) may result in improved performance. These models can capture complex temporal patterns in customer behaviour [48], particularly in sequential datasets or data with timestamp interactions, such as subscription history or reading and engagement features in the Helsingin Sanomat data.

In addition to model complexity and additional methods, including external economic indicator data could enhance the predictive modelling performance of churn models. Macroeconomic indicators such as inflation rates, GDP trends, or unemployment rates can add valuable information that influences customer churn decisions. These additional economic factors can help the model to capture broader economic pressures that may impact subscription retention, particularly during periods of financial uncertainty or global disruption (e.g., recession, pandemic, war).

In conclusion, predicting subscription churn and understanding factors that contribute to retention will continue to be one of the main objectives of news and media services. This study, while comparing a classical logistic regression model with a more advanced ensemble XGBoost model, contributes to this objective by providing these predictive models and interpretive insights with real world news subscription data (Helsingin Sanomat). Future developments in advanced modelling techniques, causal inference methods, hybrid approaches, and the inclusion of external economic data sources can potentially expand our understanding further. Combining high predictive modelling with better interpretability and real world external data can lead to the improvement of subscription retention and ensure long-term sustainability of digital subscription services.

## References

- [1] C. Angelucci and N. Cage, “Newspapers in times of low advertising revenues,” *Sciences Po Working Paper*, no. 2019-09, 2019. [Online]. Available: <https://sciencespo.hal.science/hal-03391880> (accessed Jun. 6, 2025).
- [2] Pew Research Center, “Newspapers fact sheet,” Pew Research Center, 2023. [Online]. Available: <https://www.pewresearch.org/journalism/fact-sheet/newspapers/> (accessed Jun. 6, 2025).
- [3] Helsingin Sanomat, “Etusivu – HS.fi,” Helsingin Sanomat, 2025. [Online]. Available: <https://www.hs.fi/> (accessed Jun. 6, 2025).
- [4] FIAM, “Tulokset – Finnish Internet Audience Measurement,” FIAM, 2025. [Online]. Available: <https://fiam.fi/tulokset/> (accessed Jun. 26, 2025).
- [5] N. Newman, “Reuters Institute Digital News Report 2023,” Reuters Institute, 2023. [Online]. Available: [https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2023-06/Digital\\_News\\_Report\\_2023.pdf](https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2023-06/Digital_News_Report_2023.pdf) (accessed Jul. 23, 2025).
- [6] G. Piechota, “News avoidance is a massive threat,” *International News Media Association (INMA)*, Nov. 2023. [Online]. Available: <https://www.inma.org/blogs/value-content/post.cfm/news-avoidance-is-a-massive-threat> (accessed Jul. 23, 2025).
- [7] K. O’Brien and A. Downie, “What is customer churn?,” *IBM Consulting*, Sep. 9, 2024. [Online]. Available: <https://www.ibm.com/think/topics/customer-churn> (accessed May 7, 2025).
- [8] D. H. Rudd, H. Huo, and G. Xu, “Causal analysis of customer churn using deep learning,” *arXiv preprint*, [Online]. Available: <https://arxiv.org/pdf/2304.10604> (accessed Jul. 1, 2025).
- [9] F. F. Reichheld and W. E. Sasser, “Zero defections: Quality comes to services,” *Harvard Business Review*, vol. 68, no. 5, pp. 105–111, 1990.
- [10] A. Gallo, “The value of keeping the right customers,” *Harvard Business Review*, Oct. 2014. [Online]. Available: <https://hbr.org/2014/10/the-value-of-keeping-the-right-customers> (accessed Jul. 2, 2025).
- [11] R. N. Bolton, “A dynamic model of the duration of the customer’s relationship with a continuous service provider: The role of satisfaction,” *Marketing Science*, vol. 17, no. 1, pp. 45–65, 1998.
- [12] R. T. Rust and T. S. Chung, “Marketing models of service and relationships,” *Marketing Science*, vol. 25, no. 6, pp. 560–580, 2006.

- [13] P. C. Verhoef, “Understanding the effect of customer relationship management efforts on customer retention and customer share development,” *Journal of Marketing*, vol. 67, no. 4, pp. 30–45, 2003.
- [14] J. Villanueva and D. M. Hanssens, “Customer equity: Measurement, management and research opportunities,” *Foundations and Trends in Marketing*, vol. 1, no. 1, pp. 1–95, 2007.
- [15] J. Rosada, E. Koch, A. B. Burmester, and M. Clement, “The impact of smart speakers and podcasts on news media consumption,” *J. Media Bus. Stud.*, 2024, doi: 10.1080/16522354.2024.2418718.
- [16] INMA, “News media companies leverage games, bundles to encourage habits,” *INMA*, 2024. [Online]. Available: <https://www.inma.org/blogs/product-initiative/post.cfm/news-media-companies-leverage-games-bundles-to-encourage-habits> (accessed Jun. 4, 2025).
- [17] The New York Times, “Wordle – a daily word game,” [Online]. Available: <https://www.nytimes.com/games/wordle/> (accessed Jul. 23, 2025).
- [18] A. Silberling, “Wordle brought ‘tens of millions’ of new users to the New York Times,” *TechCrunch*, May 4, 2022. [Online]. Available: <https://techcrunch.com/2022/05/04/wordle-new-york-times-user-growth/> (accessed Jul. 23, 2025).
- [19] A. Jung, *Machine Learning: The Basics*. Singapore: Springer, 2022. [Online]. Available: <https://alexjungaalto.github.io/MLBasicsBook.pdf> (accessed May 17, 2025).
- [20] K. P. Murphy, *Probabilistic Machine Learning: An Introduction*. Cambridge, MA: MIT Press, 2022. [Online]. Available: <http://probml.github.io/book1> (accessed Jun. 13, 2025).
- [21] W. Verbeke, K. Dejaeger, D. Martens, J. Hur, and B. Baesens, “New insights into churn prediction in the telecommunication sector: A profit driven data mining approach,” *European Journal of Operational Research*, vol. 218, no. 1, pp. 211–229, 2012.
- [22] A. Lemmens and C. Croux, “Bagging and boosting classification trees to predict churn,” *Journal of Marketing Research*, vol. 43, no. 2, pp. 276–286, 2006.
- [23] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*, 2nd ed. Cambridge, MA: MIT Press, 2018. [Online]. Available: <https://cs.nyu.edu/~mohri/mlbook/> (accessed Jul. 2025).
- [24] N. H. Al-Maqaleh, M. I. Hammouri, and H. I. Jaam, “Ensemble methods in customer churn prediction: A comparative analysis of the state-of-the-art,” *Mathematics*, vol. 11, no. 5, p. 1137, Feb. 2023, doi: 10.3390/math11051137.

- [25] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, San Francisco, 2016, pp. 785–794.
- [26] J. H. Friedman, “Greedy function approximation: A gradient boosting machine,” *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.
- [27] M. Harrison, *Effective XGBoost: Optimizing, Tuning, Understanding, and Deploying Classification Models*, E. Krueger, A. Rook, and R. Legere, Eds. Treading on Python Press, 2023.
- [28] P. Boozary, S. Sheykhan, H. GhorbanTanhaei, and C. Magazzino, “Enhancing customer retention with machine learning: A comparative analysis of ensemble models for accurate churn prediction,” *Int. J. Inf. Manag. Data Insights*, 2025. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2667096825000138> (accessed Jul. 29, 2025).
- [29] M. Mach-Król and B. Hadasik, “On a certain research gap in big data mining for customer insights,” *Appl. Sci.*, vol. 11, no. 15, Art. no. 6993, Jul. 2021, doi: 10.3390/app11156993.
- [30] M. Hassouna, A. Tarhini, T. Elyas, and M. S. AbouTrab, “Customer churn in mobile markets: A comparison of techniques,” *arXiv preprint arXiv:1607.07792*, 2016.
- [31] A. Owczarczuk, “Churn models for prepaid customers in the cellular telecommunication industry using large data marts,” *Expert Syst. Appl.*, vol. 37, no. 6, pp. 4710–4712, Jun. 2010, doi: 10.1016/j.eswa.2009.12.048.
- [32] M. A. Onari, M. J. Rezaee, M. Saberi, and M. S. Nobile, “An explainable data-driven decision support framework for strategic customer development,” *Knowledge-Based Systems*, vol. 297, Art. no. 111617, Sep. 2024, doi: 10.1016/j.knosys.2024.111761
- [33] E. Eslami, N. Razi, M. Lonbani, and J. Rezazadeh, “Unveiling IoT customer behaviour: Segmentation and insights for enhanced IoT-CRM strategies: A real case study,” *Sensors*, vol. 24, no. 4, Art. no. 1050, Feb. 2024, doi:10.3390/s24041050. [Online]. Available: <https://www.mdpi.com/1424-8220/24/4/1050> (accessed Jul. 30, 2025).
- [34] M. Chandar and P. A. L. Krishna, “Modeling churn behavior of bank customers using predictive data mining techniques,” in *Proc. Nat. Conf. Soft Comput. Techn. Eng. Appl. (SCT)*, 2006, pp. 24–26.
- [35] A. Zheng and A. Casari, *Feature Engineering for Machine Learning*, 1st ed. Sebastopol, CA: O’Reilly Media, 2018. [Online]. Available: <http://oreilly.com/safari> (accessed Jun. 14, 2025).

- [36] K. Pearson, “Notes on regression and inheritance in the case of two parents,” *Proc. R. Soc. Lond.*, vol. 58, pp. 240–242, 1895.
- [37] scikit-learn, “LogisticRegression,” [Online]. Available: [https://scikit-learn.org/stable/modules/generated/sklearn.linear\\_model.LogisticRegression.html#sklearn.linear\\_model.LogisticRegression](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html#sklearn.linear_model.LogisticRegression) (accessed May 7, 2025).
- [38] XGBoost, “XGBoost,” [Online]. Available: <https://xgboost.readthedocs.io/en/stable/tutorials/model.html> (accessed May 7, 2025).
- [39] XGBoost, “XGBoost parameters,” [Online]. Available: <https://xgboost.readthedocs.io/en/stable/parameter.html> (accessed Jun. 27, 2025).
- [40] M. Voros, “Handling imbalanced datasets with XGBoost: Optimizing model performance with smart parameter tuning,” *Medium*, 2024. [Online]. Available: <https://medium.com/@mate.voros1998/handling-imbalanced-datasets-with-xgboost-optimizing-model-performance-with-smart-parameter-tuning-18568c7783cf> (accessed Jun. 27, 2025).
- [41] scikit-learn, “GridSearchCV,” 2025. [Online]. Available: [https://scikit-learn.org/stable/modules/generated/sklearn.model\\_selection.GridSearchCV.html](https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html) (accessed Jul. 28, 2025).
- [42] Optuna, “A hyperparameter optimization framework,” 2025. [Online]. Available: <https://optuna.org/> (accessed Jul. 28, 2025).
- [43] S. M. Lundberg, G. Erion, and S.-I. Lee, “Consistent individualized feature attribution for tree ensembles,” *arXiv*, Feb. 2018. [Online]. Available: <https://arxiv.org/abs/1802.03888> (accessed Jun. 2025).
- [44] S. M. Lundberg and S.-I. Lee, “SHAP (SHapley Additive exPlanations) documentation,” 2025. [Online]. Available: <https://shap.readthedocs.io/en/latest/index.html> (accessed Jun. 18, 2025).
- [45] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 4765–4774.
- [46] SHAP, “Beeswarm plot,” *SHAP Documentation*, [Online]. Available: [https://shap.readthedocs.io/en/latest/example\\_notebooks/api\\_examples/plots/beeswarm.html](https://shap.readthedocs.io/en/latest/example_notebooks/api_examples/plots/beeswarm.html) (accessed Jul. 31, 2025).
- [47] E. Dillon *et al.*, “Be careful when interpreting predictive models in search of causal insights,” *SHAP Documentation*. [Online]. Available: [https://shap.readthedocs.io/en/latest/example\\_notebooks/overviews/Be%20careful%20when%20interpreting%20predictive%20models%20in%20search%20of%20causal%20insights.html](https://shap.readthedocs.io/en/latest/example_notebooks/overviews/Be%20careful%20when%20interpreting%20predictive%20models%20in%20search%20of%20causal%20insights.html) (accessed Jun. 25, 2025).

- [48] Y. Feng *et al.*, “Deep learning models for hepatitis E incidence prediction leveraging meteorological factors,” *PLOS ONE*, vol. 18, no. 3, 2023, doi: 10.1371/journal.pone.0282928.
- [49] C. Zhang *et al.*, “Enhancing robustness of gradient-boosted decision trees through feature perturbations,” *arXiv preprint arXiv:2304.13761*, 2023. [Online]. Available: <https://arxiv.org/abs/2304.13761> (accessed Jul. 2025).
- [50] World Association of News Publishers, “How Germany’s Der Spiegel uses games, comments to boost engagement,” [Online]. Available: <https://wan-iffra.org/2025/05/how-germanys-der-spiegel-uses-games-comments-to-boost-engagement/> (accessed May 8, 2025).