Erasmus School of Economics

Erasmus University Rotterdam

Erasmus School of Economics

Master Thesis

MSc Data Science and Marketing Analytics

Experimental Study on the Robustness of Marketing Models to Adversarial Attacks

Student Name: Ridhima Shrivastava

Student ID: 558085

Date Submitted: 05/06/2024

Supervisor: Philip Hans Franses

Second Assessor: Dr. Nuno Almeida Camacho

Table of Contents

Chapter 1: Introduction

1.1 Research Objectives

The advancement of Machine Learning (ML) and Artificial Intelligence (AI) has notably transformed the landscape of marketing strategies, ushering in an era where decisions are increasingly driven by data. The reliance on data-centric approaches, however, brings forth a unique set of challenges, particularly concerning the integrity and reliability of these systems. Among these challenges, *data poisoning* stands out as a critical threat that demands attention. Data poisoning involves the deliberate manipulation of training data with the intent to manipulate the performance of ML models. This manipulation can lead to erroneous model outcomes, significantly corrupting decision-making processes and, by extension, reducing the overall effectiveness of marketing strategies.

This phenomenon's significance is amplified in marketing due to the high stakes involved in decision accuracy and the potential for substantial financial and reputational damage. This impact has been observed in cases where manipulated customer data has led to flawed audience targeting and subsequently ineffective marketing campaigns, causing financial losses and brand damage (Smith, 2021). For instance, in a study by Johnson and Lee (2020), a company experienced a 30% drop in campaign effectiveness after its predictive analytics model was compromised by poisoned data, highlighting the severe implications of such breaches on marketing outcomes. Adversarial attacks through data poisoning not only undermine the reliability and trustworthiness of AI systems but also pose a severe security risk. These risks are not confined to marketing; they extend to other critical applications, including but not limited to autonomous vehicles, medical diagnosis, and financial systems, where the implications of compromised data can be far-reaching (Fox, 2023).

1.2 Central Research Question

Given this context, this thesis aims to address the primary research question: *How does data poisoning affect the performance metrics of marketing models at different levels of attack severity?* This question delves into the core of the issue by investigating the relationship between the extent of data poisoning and the resultant performance degradation of marketing models. Understanding this relationship is crucial for developing robust models capable of withstanding such adversarial threats.

1.3 Subquestions

A thorough analysis of the central research question can be achieved with a series of smaller investigations, separated below for convenience.

1. Which models demonstrate greater resilience to label-flipping attacks and what are the underlying mechanisms contributing to their robustness?
2. What methods can be employed to enhance the robustness of models against data poisoning attacks?
3. What defensive strategies can be implemented to shield marketing models from the adverse effects of data poisoning?
4. At which level of data contamination does the integrity of marketing models begin to significantly deteriorate, impacting key performance metrics such as accuracy, precision, recall, and F1 score?
5. What techniques are effective in detecting and sanitising poisoned data?

Subquestion 1: *Which models demonstrate greater resilience to label-flipping attacks and what are the underlying mechanisms contributing to their robustness?*

This question seeks to classify various models based on their resistance to label-flipping attacks, focusing on identifying and explaining the factors that contribute to their robustness. According to Yerlikaya and Bahtiyar (2022), the model expected to perform the best is Random Forest.

Subquestion 2: *What methods can be employed to enhance the robustness of models against data poisoning attacks?*

This question explores the different strategies and techniques that can be implemented during the model training phase to mitigate the risk and impact of data poisoning, enhancing the model's overall security. The method to be tested is adversarial training, as seen in a paper by Goodfellow et al. (2014).

Subquestion 3: *What defensive strategies can be implemented to shield marketing models from the adverse effects of data poisoning?*

This examines protective measures that can be adopted to secure datasets against corruption, focusing on both preventative care. Goodfellow et al. (2014), Steinhardt et al. (2017), Qiu (2022), and Li et al. (2024) all suggest defensive measures in their papers which will be considered for our dataset.

Subquestion 4: *At which level of data contamination does the integrity of marketing models begin to significantly deteriorate, impacting key performance metrics?*
This question aims to identify the critical threshold of data poisoning at which the performance of marketing models starts to degrade, providing a quantitative basis (with accuracy, precision, recall, and F1 score) for the development of threshold-based defense mechanisms. It is expected that even small amounts (1%) of data poisoning will have an impact on the performance of the chosen models.

Subquestion 5: *What techniques are effective in detecting and sanitising poisoned data?*
This focuses on identifying and evaluating methods for spotting and cleaning corrupted data entries in marketing databases to ensure data integrity and model accuracy. The method that will be explored is detecting discrepancies between a clean version of the dataset and the poisoned version, which is a cheap and effective.

The importance of addressing data poisoning in the context of marketing models cannot be overstated. As businesses increasingly rely on AI and ML to formulate their marketing strategies, the integrity of the data feeding these models becomes paramount. Data poisoning attacks pose a significant threat to this integrity, with the potential to mislead decision-making processes, waste resources, and perhaps most importantly erode consumer trust on the brand. Moreover, the exploration of protective measures against such attacks is essential for maintaining the security and efficacy of marketing models. This research will not only contribute to the academic understanding of adversarial attacks on ML models but will also offer practical insights for businesses seeking to fortify their marketing strategies against these emerging threats.

1.4 Thesis Structure

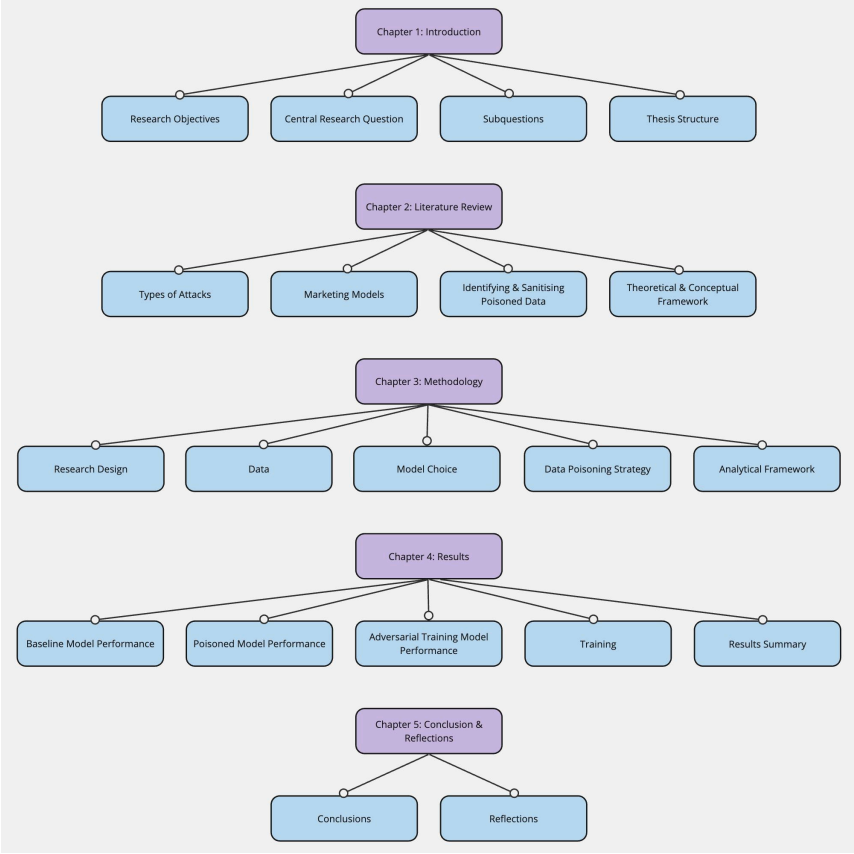
The study conducted in this thesis is presented in five chapters: Introduction, Literature Review, Research Design and Methodology, Results, and Conclusion. Further details are explained below. The first two chapters function as investigating the theoretical answers to the central research question and subquestions as well as providing more background information and context to our inspection. Chapter One starts with a broad introduction to data poisoning and its relevance in various contexts. It continues to clarify the research objectives that are to be achieved with the analysis to be conducted in this paper. Further, it explores the problem at hand

with a central research question along with five sub-questions. Lastly, it mentions the thesis structure followed for ease of readability. Further, Chapter Two includes a literature review of prior investigations into data poisoning across various marketing models, types of attacks, datasets, and defensive strategies. This section concludes with a theoretical and conceptual framework which guides the rest of the analysis.

The following three chapters examine methods, results, and conclusions. Chapter Three focuses on the design as well as the methodology used in this paper. The methodology will focus on the justification behind the chosen analyses of the dataset. Hence, the limitations of the methods used will also be discussed. The following chapter (Four) will exhibit the processing, inspection, and interpretation of results. There will be an in-depth discussion of the general conclusions of the conducted analysis in context of the research objectives aforementioned. Chapter Five will include a conclusion of the thesis outcomes to provide recommendations for future research and closing remarks. A visual representation of the thesis structure is included below for convenience.

Figure 1

Visual overview of thesis structure



Chapter 2: Literature Review

This literature review examines data poisoning by assessing its distinct types, the underlying mechanisms of these attacks, and their negative impact on marketing model performance metrics. Additionally, it scrutinises a variety of defensive strategies that researchers have created to combat such attacks. The review also explores techniques for identifying and purging poisoned data, thereby ensuring the reliability of data-driven decision-making processes in marketing.

Through a detailed examination of both foundational and recent research, this review aims to consolidate knowledge on the strategies to mitigate the impact of data poisoning, identify persisting research gaps, and lay the groundwork for fortifying marketing models against these threats. This comprehensive review not only informs but also inspires the development of robust mechanisms to protect marketing models, ensuring that they continue to operate effectively even in the face of sophisticated data manipulation tactics.

2.1 Types of Attacks

The literature reviewed explores various types of data poisoning attacks such as label-flipping, distance-based label-flipping, watermarking, and clean-label. These attacks are specifically designed to undermine the integrity of machine learning models in distinct and disparate ways. Label-flipping attacks simply alter the labels of training data points, misleading the model's learning process by providing incorrect class associations (Rosenfeld, Winston, Ravikumar, & Kolter, 2020). Distance-based label flipping refines this approach by targeting data points near the decision boundary, flipping their labels to maximise disruption on the model's ability to classify similar data points accurately (Yerlikaya & Bahtiyar, 2022). Watermarking attacks embed irrelevant patterns or signals into the data, which confuses the model during training and diminishes its performance in real-world applications (Qiu, 2022). Lastly, clean-label attacks involve subtle modifications to data features while keeping the labels unchanged, making these poisoned data points appear normal and bypassing straightforward detection methods, thus posing a significant challenge to data integrity and model reliability (Gupta & Krishna, 2023).

Yerlikaya and Bahtiyar (2022) specifically tested label-flipping and distance-based label-flipping attacks across multiple algorithms and datasets, showing varying levels of impact on model performance metrics like accuracy and F1-score. Qiu (2022) adds to this by categorising various

poisoning strategies, including more complex methods like watermarking and clean-label attacks, which embed hidden patterns or preserve original labels to evade detection, respectively. His work helps frame the breadth of possible attacks that can target machine learning systems, including those employed in marketing models.

This comprehensive exploration of different attack types guided this paper's focus towards label-flipping attacks as it is most appropriate in the context of the chosen dataset (customer churn).

2.2 Marketing Models

The reviewed research papers utilise various machine learning models to explore the impact of data poisoning attacks, offering a comprehensive range of insights into model vulnerabilities and defensive measures. The leading paper for this section will be from Yerlikaya and Bahtiyar (2022), who conducted extensive experiments utilising five different machine learning algorithms: Support Vector Machine (SVM), Stochastic Gradient Descent (SGD), Logistic Regression (LR), Random Forest (RF), Gaussian Naive Bayes (GNB), and K-Nearest Neighbor (KNN). These models are assessed under conditions of random and distance-based label flipping to evaluate their robustness across different attack scenarios, using metrics such as F1 Score, Accuracy rate, and AUC Score. In their paper, the KNN and RF algorithms outperformed the others consistently across datasets and levels of severity and this is what is expected for this research paper as well.

For a customer churn dataset, where the primary objective is to predict binary outcomes effectively (churn or not), several machine learning models stand out for their robustness, accuracy, and applicability. Each of these models offers unique advantages that can be leveraged depending on the specifics of the dataset and the business context. The LR model is highly effective for binary classification tasks like churn prediction due to its simplicity and the interpretability of its output, which represents the probability of churn (Cole, 2020). Known for its high accuracy and robustness, Random Forest can handle large datasets with complex feature interactions without the risk of overfitting, thanks to its ensemble approach that averages multiple deep decision trees (Sharma, 2021). SVM is suitable for high-dimensional spaces and is effective in cases where there is a clear margin of separation between classes (Indriati, 2023). GBM is another ensemble technique that builds trees sequentially, with each new tree helping to

correct errors made by previously built trees (AlShourbaji et al., 2023). This model is known for delivering high accuracy and can handle various types of data, making it a strong candidate for churn prediction. For datasets with complex patterns and interactions that might be difficult to capture with other algorithms, Neural Networks offer a flexible architecture that can learn these patterns directly from the data (Badole, 2023). They require more data and computational power but can significantly outperform simpler models if tuned properly.

Each of these models presents different strengths and trade-offs. For instance, while Logistic Regression and SVM provide clarity and simplicity, Random Forest, and GBM offer greater accuracy through more complex ensemble methods. Neural Networks provide unparalleled flexibility and learning capacity, which can be particularly beneficial in dynamic environments where customer behaviours are non-linear and evolving. Depending on the specific characteristics of the churn dataset, such as the number of features, the volume of data, and the need for model interpretability, one can choose the most appropriate model or a combination of models to maximise predictive performance and operational efficiency.

2.3 Identifying & Sanitising Poisoned Data

Identifying and cleansing contaminated data is pivotal for preserving the dependability and precision of machine-learning models in marketing. Research has introduced various defensive tactics to counteract data poisoning attacks, concentrating on detection techniques and remedial actions to uphold data integrity and model efficacy.

Goodfellow et al. (2014) introduce adversarial training as a defensive strategy. This method incorporates adversarial examples into the training process to prepare the model for potential attacks, effectively enhancing the model's resilience. By training with adversarial examples, the model learns to identify and disregard misleading inputs, a method that has proven effective in reducing vulnerabilities across various types of machine learning models.

Another critical approach discussed by Steinhardt et al. (2017) involves constructing statistical bounds to anticipate the maximum potential loss from data poisoning. This method not only aids in recognising the extent of an attack's impact but also helps in developing strategies that minimise risk by adjusting the model's sensitivity to anomalies in training data.

Furthermore, Qiu (2022) explores comprehensive defensive mechanisms including data sanitisation, which involves cleaning the data set of any anomalies that could potentially skew the model's learning. Techniques such as anomaly detection are vital in this process, allowing for the systematic removal of outliers that do not conform to expected patterns. Comparatively, Li et al. (2024) propose a sampling-based method for detecting anomalies in data sets, specifically through techniques like the Rating Matrix Sampling Method and pinpointing malicious data via the Distance of Rating Vectors. These methods provide practical tools for identifying poisoned data segments by comparing deviations from typical data patterns, offering a robust framework for ensuring data cleanliness before model training.

Each of these studies contributes to a layered defensive strategy against data poisoning. For marketing models, particularly those used in dynamic environments like customer churn prediction, employing a combination of adapted defensive strategies for the dataset considered can significantly enhance security. Adversarial training and data sanitisation processes ensure that the models are not only trained on clean, reliable data but are also resilient to sophisticated attacks aimed at compromising their performance. Implementing these methodologies provides a comprehensive shield, safeguarding the predictive accuracy and reliability of marketing models against the evolving threat of data poisoning.

2.4 Theoretical & Conceptual Framework

2.4.1 Theoretical Underpinnings

This thesis is underpinned by the theory of adversarial machine learning, which scrutinises the vulnerabilities of machine learning systems to manipulative attacks that deliberately alter training data to degrade model performance. Central to this theory is the premise that ML models' dependence on the quality and integrity of their input data can be exploited through adversarial attacks such as label flipping. This specific attack method has been chosen for its relevance and potential impact on marketing models, particularly within the context of customer churn prediction. Theoretical insights from Yerlikaya and Bahtiyar (2022) provide a foundational understanding of how various algorithms withstand such manipulations, emphasising the need for robust defensive strategies.

2.4.2 Conceptual Framework

The conceptual framework of this research is focused on the label-flipping attack applied to a customer churn dataset. This approach involves a comparative analysis of five machine learning models: Logistic Regression (LR), Support Vector Machine (SVM), Gradient Boosting Machines (GBM), Random Forest (RF), and Neural Networks (NN). These models were selected for their diverse capabilities in handling binary classification tasks and their varying levels of susceptibility to adversarial attacks.

This study aims to evaluate and compare the resilience of these models against label-flipping attacks, employing two main defensive strategies:

1. Adversarial Training: Following the methodology suggested by Goodfellow et al. (2014), this technique involves incorporating adversarial examples into the training process to prepare the model for potential adversarial conditions, enhancing its ability to identify and mitigate misleading inputs.
2. Data Sanitisation Methods: Building on the categorisations by Qiu (2022), this research will explore various data cleaning techniques aimed at removing or correcting poisoned data entries, thereby preserving the integrity of the training dataset.

The interactions between these defensive strategies and the five selected models will be critically analysed to determine their effectiveness in maintaining or improving performance metrics such as accuracy, precision, recall, and F1 score under the influence of a label-flipping attack. This framework not only guides the empirical evaluation of model robustness but also contributes to the practical understanding of how to safeguard marketing models against data poisoning.

By integrating these theoretical and conceptual elements, this framework underpins a comprehensive investigation into the dynamics of adversarial threats and defense mechanisms in marketing analytics. This structured approach ensures a thorough examination of the protective measures necessary to enhance the security and reliability of machine learning models used in high-stakes marketing decisions.

Chapter 3: Methodology

3.1 Research Design

This study employs a quantitative research design, leveraging statistical and computational techniques to investigate the impact of data poisoning on marketing models. The quantitative approach is particularly suitable for this research because it allows for precise measurement and analysis of model performance metrics. By using a structured dataset (Telco Customer Churn) and applying various machine learning models, this research quantifies the effects of adversarial attacks at different levels of severity (1%, 3%, and 5%). The primary objective is to produce statistically significant results that can be generalised to similar datasets and contexts, providing a robust understanding of how data poisoning affects predictive performance in marketing applications.

While the study is primarily quantitative, it also integrates elements of experimental research. The experimental setup involves controlled manipulation of the dataset through label-flipping attacks, allowing for a systematic examination of the impact on model performance. This approach not only quantifies the degradation in model metrics but also explores the efficacy of different defensive strategies and adversarial training techniques. By comparing models' performance before and after the introduction of poisoned data, the research design ensures a rigorous evaluation of model robustness and the effectiveness of mitigation methods. This mixed-methods approach, combining quantitative measurement with experimental manipulation, provides a comprehensive framework for understanding and addressing data poisoning in machine learning models.

3.2 Data

The usage of the artificial dataset Telco Customer Churn from Kaggle is ideal for analysis due to its rich and varied feature set, real-world relevance, and high quality. This dataset contains 7043 observations/customers and 21 features. The variables are as follows: Customer ID, gender, senior citizen, partner, dependents, tenure with the company (months), phone service, multiple lines, internet service, online security, online backup, device protection, tech support, streaming TV, streaming movies, contract, paperless billing, payment method, monthly charges, total charges, and churn. This variety allows for development of effective predictive models.

Moreover, the dataset's practical business context in the telecommunications industry makes the findings highly applicable and valuable for real-world customer retention strategies. Its clean and well-structured format reduces the time needed for data preparation. The dataset's size and complexity also make it suitable for testing the scalability and performance of the chosen algorithms, ensuring that the solutions are both efficient and realistic.

Table 1

Descriptive statistics for the Telco customer dataset

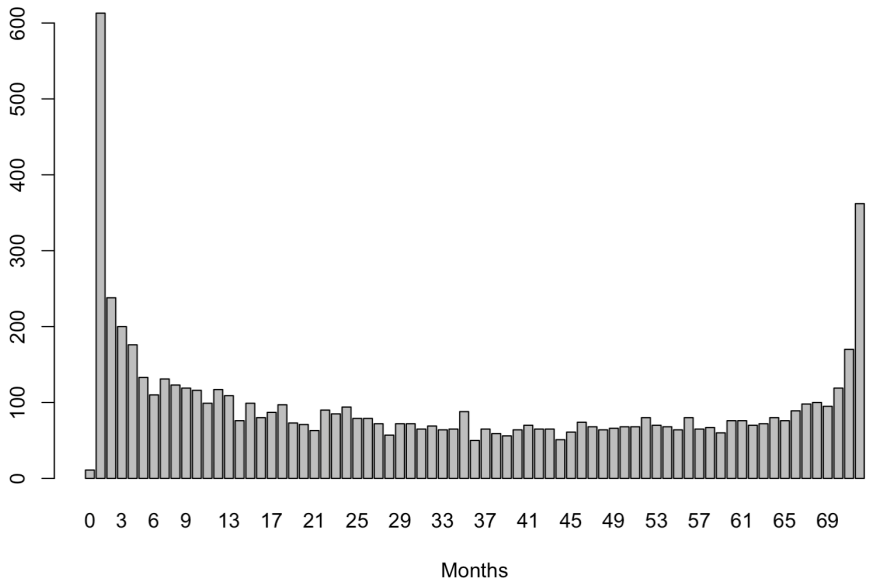
Variable	**Basic Descriptive Statistics**
Gender	Male - 3555 (51%) Female - 3488 (49%)
Senior citizen	No - 5901 (80%) Yes - 1142 (20%)
Partner (whether they have a partner or not)	No - 3641 (49%) Yes - 3402 (51%)
Dependents	No - 4933 (70%) Yes - 2110 (30%)
Tenure	Min - 0 Median - 29 Mean - 32.37 Max - 72
Phone service	No - 682 (10%) Yes - 6361 (90%)
Multiple lines	No - 3390 (48%) No phone service - 682 (10%) Yes - 2971 (42%)
Internet service	No - 1526 (21%) DSL - 2421 (34%) Fiber optic - 3096 (44%)
Online security	No - 3498 (49%) No internet service - 1526 (22%) Yes - 2019 (29%)
Online backup	No - 3088 (44%) No internet service - 1526 (22%) Yes - 2429 (34%)

Device protection	No - 3095 (44%) No internet service - 1526 (22%) Yes - 2422 (34%)
Tech support	No - 3473 (49%) No internet service - 1526 (22%) Yes - 2044 (29%)
Streaming TV	No - 3088 (44%) No internet service - 1526 (22%) Yes - 2429 (34%)
Streaming movies	No - 2810 (40%) No internet service - 1526 (22%) Yes - 2707 (38%)
Contract	Month-to-month - 3875 (55%) One year - 1473 (21%) Two year - 1695 (24%)
Paperless billing	No - 2872 (41%) Yes - 4171 (59%)
Payment method	Bank transfer (automatic) - 1544 (22%) Credit card (automatic) - 1522 (22%) Electronic check - 2365 (33%) Mailed check - 1612 (23%)
Monthly charges	Min - 18.25 Median - 70.35 Mean - 64.76 Max - 118.75
Total charges	Min - 18.8 Median - 1397.5 Mean - 2283.3 Max - 8684.8
Churn	No - 5174 (73%) Yes - 1869 (27%)

Provided below are two visualisations (Figures 2 and 3) - one for the variable tenure and one for monthly charges.

Figure 2

Barplot of distribution of the variable 'Tenure' in months



The distribution follows a U-shaped curve which is common for tenureship in a telecommunications company, with the peak being at the lowest number of months. This measures the customers that used the service for 1-2 months and then either cancelled their service or switched to another service provider.

Figure 3

Barplot of distribution of the variable 'Monthly Charges'

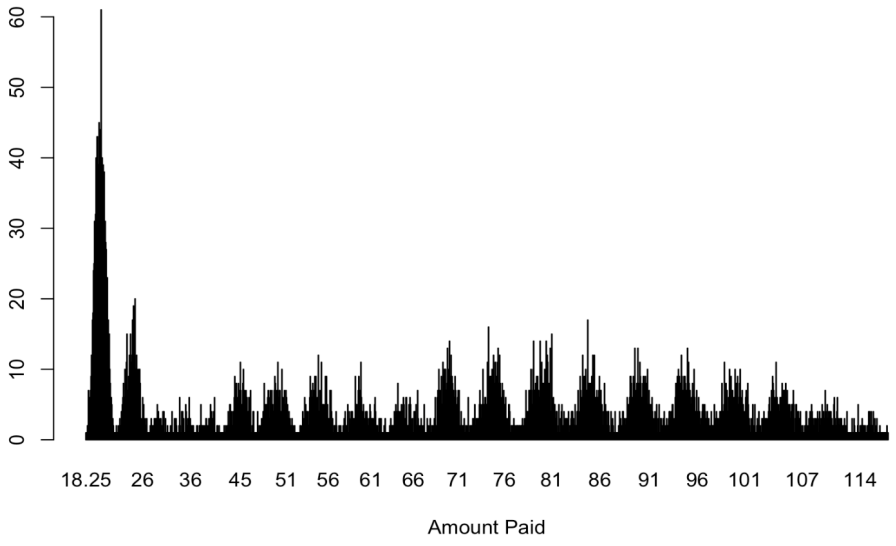


Figure 3 above shows the amount paid monthly by customers, with a notable peak at the lowest amount and short peaks throughout - likely alluding to different plans available.

3.3 Model Choice

In this study, a diverse set of machine learning models is selected to investigate their resilience against data poisoning attacks in the context of churn prediction. The chosen models represent a spectrum of complexity and approaches, from simple, interpretable models to complex, high-performance algorithms based on the research from the literature review. Each model was selected based on its unique strengths and applicability to the task of binary classification, particularly in scenarios involving customer churn. The selection of these models aims to provide a comprehensive understanding of how different types of algorithms respond to adversarial attacks, thereby identifying the most robust and effective models for maintaining data integrity and reliability in marketing applications.

Logistic Regression is a binary classification algorithm that estimates the probability of a binary outcome based on one or more predictor variables. It is widely used in churn prediction due to its simplicity and interpretability as it provides clear and actionable insights by outputting probabilities, which makes it particularly useful for understanding the likelihood of customer churn (Cole, 2020). Despite its simplicity, it is highly effective for binary classification tasks and serves as the baseline model in this study due to its computational efficiency and ease of implementation.

The logistic regression model predicts the probability P of the binary outcome using the following formula:

$$P(Y = 1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_nX_n)}}$$

where:

- $P(Y = 1|X)$ is the probability that the outcome Y is 1 given the predictors X.
- β_0 is the intercept term.
- $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients for the predictor variables $X_1, X_2, \ldots, X_n$.
- e is the base of the natural logarithm.

The formula uses the logistic function, also known as the sigmoid function, which maps any real-valued number into a value between 0 and 1. This mapping is crucial for binary classification, as it ensures that the predicted probabilities are valid probabilities. The exponent term $\beta_0 + \beta_1X_1 + \ldots + \beta_nX_n$ represents a linear combination of the input features, which is then

transformed by the logistic function to produce a probability. This probability helps in making a decision threshold, commonly at 0.5, to classify the outcome (0 or 1).

Support Vector Machine is a powerful classification method that works well in high-dimensional spaces, especially when there is a clear margin of separation between classes. SVM is effective for churn prediction as it aims to find the optimal hyperplane that maximises the margin between different classes (Indriati, 2023). This characteristic makes SVM particularly robust in cases where the data is not linearly separable.

Mathematically, the objective of the SVM algorithm is to find the hyperplane that best divides a dataset into classes. For linearly separable data, the decision boundary can be expressed as:

$w \cdot x - b = 0$

where:

- w is the weight vector perpendicular to the hyperplane.
- x represents the feature vector.
- b is the bias term.

The goal is to maximise the margin, which is the distance between the hyperplane and the nearest data points from either class, known as support vectors. The margin can be maximised by minimising w, subject to the constraint that all data points are correctly classified:

$y_i(w \cdot x - b) \geq 1$

for $i = 1, 2, \ldots n$, where y_i are the class labels (either +1 or -1) and x_i are the feature vectors of the training data.

In cases where the data is not linearly separable, SVM employs kernel functions to map the input features into higher-dimensional spaces, making it possible to find a hyperplane that can separate the data. Commonly used kernels include the polynomial kernel, the radial basis function (RBF) kernel, and the sigmoid kernel. By using these kernel tricks, SVM can handle complex relationships within the data, making it a versatile and robust method for classification tasks, including churn prediction.

Gradient Boosting Machine is an ensemble learning technique that builds models sequentially. Each new model attempts to correct the errors made by the previously built models. This

iterative approach helps in achieving high predictive accuracy and is particularly useful in handling various types of data, including categorical and continuous variables (AlShourbaji et al., 2023). GBM's ability to improve accuracy through successive iterations makes it a strong candidate for churn prediction, as it effectively captures complex patterns in the data.

Mathematically, GBM aims to minimise a loss function $L(y, F_m(x))$where y is the actual value and $F_m(x)$ is the prediction from the mth model. The GBM algorithm can be described as follows:

1. Initialise the model with a constant value:

   $$F_0(x) = arg\ min_\gamma \sum_{i=1}^{n} L(y, \gamma)$$

2. For $m = 1$ to M (total number of iterations):

   Compute the pseudo-residuals:

   $$r_{im} - [\frac{\partial L(y_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)}]$$

   Fit a base learner (e.g., a decision tree) to the pseudo-residuals:

   $$h_m(x) \approx r_{im}$$

   Update the model:

   $$F_m(x) = F_{m-1}(x) + vh_m(x)$$

   where v is the learning rate, a small positive number that controls the step size of each iteration.

The GBM algorithm continues to add new base learners to correct the errors of the ensemble. By iteratively fitting new models to the residual errors of the combined model, GBM is able to improve its accuracy over successive iterations. This process allows GBM to effectively capture complex patterns and interactions within the data, making it a powerful tool for predictive tasks such as churn prediction.

Random Forest is another ensemble method that constructs multiple decision trees during training and outputs the mode of the classes as the final prediction. Known for its robustness and ability to handle large datasets with high dimensionality, RF reduces the risk of overfitting by averaging multiple trees (Sharma, 2021). This ensemble approach makes it highly effective for complex datasets, providing reliable and stable predictions for churn analysis.

The Random Forest algorithm can be described as follows:

1. Bootstrap Sampling: Create multiple subsets of the original training data by randomly sampling with replacement. Each subset is called a bootstrap sample.
2. Tree Construction: For each bootstrap sample, grow a decision tree:
    - At each node, select a random subset of features.
    - Split the node using the feature that provides the best split according to a specific criterion (e.g., Gini impurity or information gain).
    - Continue splitting until the stopping criterion is met (e.g., maximum depth or minimum number of samples at a leaf node).
3. Aggregation: Combine the predictions of all the decision trees to make a final prediction:
    - For classification tasks, use majority voting to determine the class label.
    - For regression tasks, average the predictions of the individual trees.

Mathematically, let $\{T_1(x), T_2(x),..., T_B(x)\}$ be the set of B decision trees trained on different bootstrap samples. For a given input x, the Random Forest prediction for classification is:

$$\widehat{y} = mode\{T_1(x), T_2(x),..., T_B(x)\}$$

For regression, the prediction is:

$$\widehat{y} = \frac{1}{B}\sum_{b=1}^{B} T_B(x)$$

The Random Forest method leverages the diversity of multiple trees to reduce variance and improve generalisation performance. This ensemble technique is particularly effective for churn analysis, as it provides robust and stable predictions even in the presence of complex interactions and noisy data.

Neural Networks offer a flexible architecture capable of learning complex patterns directly from the data. They consist of layers of interconnected nodes that simulate the workings of the human brain, enabling the model to capture non-linear relationships within the data (Badole, 2023). Although they require more data and computational power, Neural Networks can significantly outperform simpler models if tuned properly. Their ability to adapt to evolving patterns makes them particularly useful in dynamic environments where customer behaviors are non-linear and constantly changing.

Neural Networks are trained using a process called backpropagation, which involves the following steps:

1. Forward Propagation: Compute the output $\hat{y}$ by passing the input x through the network.
2. Loss Calculation: Compute the loss $L(y, \hat{y})$ using a suitable loss function (e.g., cross-entropy loss for classification).
3. Backward Propagation: Compute the gradients of the loss with respect to the weights and biases using the chain rule.
4. Weight Update: Update the weights and biases using an optimisation algorithm such as gradient descent:

$$W \leftarrow W - \eta \frac{\partial L}{\partial W}$$

$$b \leftarrow b - \eta \frac{\partial L}{\partial b}$$

where η is the learning rate.

Neural Networks' ability to adapt to evolving patterns and capture complex non-linear relationships makes them particularly useful in dynamic environments where customer behaviors are non-linear and constantly changing.

Each of these models presents unique strengths and trade-offs. While Logistic Regression and SVM provide clarity and simplicity, Random Forest and GBM offer greater accuracy through more complex ensemble methods. Neural Networks provide unparalleled flexibility and learning capacity, which can be particularly beneficial in dynamic environments where customer behaviors are evolving. Depending on the specific characteristics of the churn dataset, such as the number of features, volume of data, and need for model interpretability, the appropriate model or a combination of models can be selected to maximise predictive performance and operational efficiency.

3.4 Data Poisoning Strategy

Data poisoning, particularly label-flipping attacks, is chosen as the primary adversarial strategy for this study due to its relevance and impact on binary classification tasks such as churn prediction. Label-flipping involves altering a certain percentage of the dataset labels, transforming 'churn' labels to 'non-churn' and vice versa. The choice of label-flipping as the attack method is driven by its straightforward implementation and significant impact on model performance. Label-flipping attacks are well-documented in literature as an effective means to

degrade the performance of classification models (Goodfellow et al., 2014). For churn prediction, where the primary objective is to accurately distinguish between customers who will churn and those who will not, altering the labels undermines the model's learning process, leading to erroneous predictions.

Three levels of attack severity are selected for this study: 1%, 3%, and 5%. These specific levels are chosen based on the precedent set by Goodfellow et al. (2014), who demonstrated the effectiveness of these proportions in simulating realistic adversarial scenarios. The chosen levels provide a gradient of attack intensity, allowing for a comprehensive analysis of how varying degrees of label-flipping impact model performance.

- 1% Label-Flipping: Represents a minimal level of attack severity, providing insight into the model's robustness against small-scale adversarial manipulations.
- 3% Label-Flipping: Represents a moderate level of attack severity, offering a balanced view of model resilience under more pronounced adversarial conditions.
- 5% Label-Flipping: Represents a significant level of attack severity, testing the model's ability to withstand substantial data corruption.

The introduction of label-flipping at varying levels of severity allows for a detailed evaluation of model robustness. By retraining and evaluating each model on the poisoned datasets, this study assesses how well each algorithm can maintain its predictive performance in the face of adversarial attacks. The performance metrics used for evaluation include accuracy, precision, recall, and F1 score, providing a comprehensive view of how data poisoning affects key aspects of model effectiveness.

3.5 Analytical Framework

The analysis in this study is meticulously structured to evaluate the impact of data poisoning on the performance metrics of various marketing models. The steps of analysis included data preprocessing, initial model training, implementation of data poisoning attacks, retraining, and performance evaluation. Each step is designed to ensure a comprehensive assessment of model robustness against adversarial attacks.

3.5.1 Initial Model Training

Before training a model, it is essential to have a clean and preprocessed dataset. This is a critical step in preparing the Telco Customer Churn dataset for analysis. The preprocessing steps generally included (if applicable to the respective model's dataset preparation requirements) transforming categorical variables into dummy variables, and scaling. The dummy variables are created using one-hot encoding. This is achieved by applying the model.matrix function, which generates a binary column for each category, allowing for effective inclusion in the models.

Each machine learning model is then initially trained on the clean dataset to establish baseline performance metrics. This step includes:

- Data Partitioning: The dataset is split into training (70%) and testing (30%) subsets using the createDataPartition function from the caret package.
- Scaling: Continuous variables are scaled to standardise the range of the data, enhancing model performance and convergence. The preProcess function with the center and scale methods is used for this purpose.
- Model Training: Five machine learning models (Logistic Regression, SVM, GBM, Random Forest, and Neural Networks) are trained using the training subset. Each model is tuned to optimise its performance.
- Performance Evaluation: The models' baseline performance without poisoning is measured using accuracy, precision, recall, and F1 score.

3.5.2 Implementation of Label-flipping

To evaluate the impact of data poisoning, label-flipping attacks are introduced at three levels of severity: 1%, 3%, and 5%. To begin, a specified percentage of the labels (1%, 3%, or 5%) are randomly selected for flipping. Next, the selected labels are flipped, changing 'churn' labels to 'non-churn' and vice versa. This simulated the effect of a data poisoning attack on the test dataset.

The models are retrained on the poisoned datasets to assess the impact of the attacks. The steps included:

- Data preprocessing as aforementioned
- Retraining Models: Each of the five models is trained on unpoisoned training subsets.
- Performance Evaluation: The models are evaluated on the poisoned testing subsets using the same (accuracy, precision, recall, and F1 score) performance metrics, allowing for

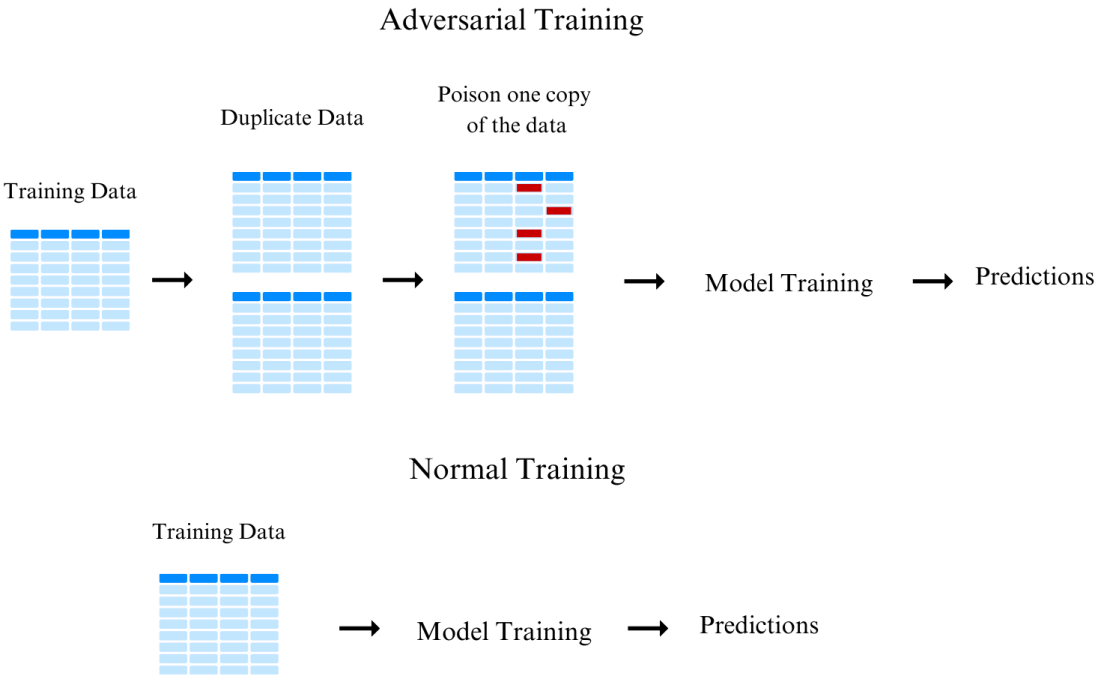
clear comparisons. These metrics provide a detailed understanding of how data poisoning affected each model's predictive capabilities.

3.5.3 Adversarial Training and Defensive Strategy

To enhance model robustness, adversarial training will be implemented. This involves creating additional training data with flipped labels to simulate adversarial conditions. The original and adversarial training data are combined to form a comprehensive training set. Models are then retrained on the combined dataset, incorporating both clean and adversarial examples to improve resilience against data poisoning. Provided below in Figure 4 is this process visualised and compared to regular model training.

Figure 4

Simple visualisation of adversarial training compared to normal model training



To identify poisoned data, it is essential to maintain data integrity and keep a sample or copy of the original dataset. Techniques for cleaning corrupted data are implemented (finding discrepancies between the original and poisoned dataset) to restore the dataset to its original state. This defensive strategy is simple, yet very effective and requires minimal effort.

Chapter 4: Results

Prior to exploring the results, it is essential to understand the metrics being used to evaluate the models: accuracy, precision, recall, and F1 score. Accuracy measures the overall correctness of the model by calculating the proportion of true results (both true positives and true negatives) among the total number of cases examined. Precision, or positive predictive value, assesses the accuracy of the positive predictions by calculating the proportion of true positives among all positive predictions made by the model. Recall, or sensitivity, measures the model's ability to correctly identify all relevant instances, calculating the proportion of true positives detected among all actual positives. The F1 score is the harmonic mean of precision and recall, providing a single metric that balances the two. In the context of predicting customer churn, accuracy and recall should be prioritised as it is crucial to identify as many potential churn cases as possible. Failing to identify a customer who is likely to churn could result in missed opportunities for retention efforts, which can be costly for businesses. However, precision and F1 score are also important to ensure the quality of the predictions and avoid unnecessary retention efforts on customers who are not actually at risk of churning.

4.1 Baseline Model Performance Metrics

The table below (Table 2) presents the baseline performance metrics for the chosen machine learning models providing a benchmark for evaluating the impact of subsequent data poisoning attacks.

Table 2

Performance metrics for all models before poisoning

Model	Accuracy	Precision	Recall	F1_Score
Logistic Regression	0.814	0.670	0.588	0.626
SVM	0.811	0.684	0.538	0.602
GBM	0.817	0.720	0.505	0.594
Random Forest	0.803	0.655	0.548	0.597
Neural Network	0.792	0.658	0.454	0.537

Note. All analyses were performed in R with the Telco Customer Dataset from Kaggle.

As seen in the table above, the accuracy values seem to oscillate around 0.8, with the Neural Network being the worst-performing model with an accuracy of 79%. The GBM has the highest

accuracy, followed by LR, and SVM having very similar results around 81%. Next, the precision peaks again with GBM at 72%.

The next best performing model is GBM with 72%, and the lowest performing being the Neural Network again. The only notable values from Recall and F1 score are those of XGBooost which are markedly higher than the rest of the models. It is worth considering that the remaining models performed near 50-60%, which makes them not advisable to use.

In summary, Logistic Regression is fairly accurate, it may miss a significant number of true churn cases, as reflected in its lower recall. SVM is better at avoiding false positives but may fail to identify many actual churn cases. The higher precision indicates that GBM makes fewer false positive errors, but its recall suggests it still misses a considerable number of actual churn cases.

The Random Forest model provides a reasonable balance between identifying churn cases and minimising false positives but is not the best performing model in any measure. The lower recall indicates that the Neural Network is less effective at identifying true churn cases, which could limit its practical utility for churn prediction.

4.2 Poisoned Model Performance Metrics

Presented below are the model performance metrics after the datasets have been poisoned at 1%, 3%, and 5% (Tables 3, 4, and 5). Additionally, Table 6 provides the percentage changes in metrics from before poisoning (Table 2) to the respective level of attack severity.

Table 3

Performance metrics for all models with 1% level of attack severity

Model	Accuracy	Precision	Recall	F1_Score
Logistic Regression	0.788	0.632	0.524	0.573
SVM	0.811	0.684	0.538	0.602
GBM	0.808	0.679	0.517	0.587
Random Forest	0.787	0.635	0.502	0.561
Neural Network	0.778	0.649	0.395	0.491

Note. All analyses were performed in R with the Telco Customer Dataset from Kaggle.

Table 4

Performance metrics for all models with 3% level of attack severity

Model	Accuracy	Precision	Recall	F1_Score
Logistic Regression	0.784	0.655	0.505	0.570
SVM	0.782	0.673	0.450	0.539
GBM	0.791	0.667	0.594	0.628
Random Forest	0.780	0.655	0.477	0.552
Neural Network	0.772	0.657	0.413	0.507

Note. All analyses were performed in R with the Telco Customer Dataset from Kaggle.

Table 5

Performance metrics for all models with 5% level of attack severity

Model	Accuracy	Precision	Recall	F1_Score
Logistic Regression	0.766	0.634	0.476	0.543
SVM	0.768	0.658	0.430	0.521
GBM	0.765	0.619	0.612	0.615
Random Forest	0.766	0.633	0.481	0.546
Neural Network	0.761	0.632	0.442	0.52

Note. All analyses were performed in R with the Telco Customer Dataset from Kaggle.

Table 6

Percentage change from before poisoning at each level of attack severity

1% Level				
Model	**Accuracy**	**Precision**	**Recall**	**F1_Score**
Logistic Regression	-3.14%	-5.74%	-10.73%	-8.47%
SVM	0%	0%	0%	0%
GBM	-1.06%	-5.67%	2.31%	-1.14%
Random Forest	-2.05%	-3.00%	-8.48%	-6.06%
Neural Network	-1.78%	-1.31%	-12.89%	-8.51%

3% Level				
Model	**Accuracy**	**Precision**	**Recall**	**F1_Score**
Logistic Regression	-3.62%	-2.23%	-14.04%	-8.90%
SVM	-3.63%	-1.69%	-16.31%	-10.45%
GBM	-3.14%	-7.42%	17.60%	5.81%
Random Forest	-2.84%	0.09%	-13.07%	-7.53%
Neural Network	-2.51%	-0.17%	-8.94%	-5.55%

5% Level				
Model	**Accuracy**	**Precision**	**Recall**	**F1_Score**
Logistic Regression	-5.88%	-5.44%	-19.02%	-13.20%
SVM	-5.37%	-3.75%	-19.92%	-13.53%
GBM	-6.29%	-14.06%	21.09%	3.62%
Random Forest	-4.59%	-3.26%	-12.34%	-8.42%
Neural Network	-3.94%	-3.96%	-2.61%	-3.17%

4.2.1 One Percent Level of Attack Severity

After consulting Table 3, one result that stands out is that SVM performs the same for all metrics at the 1% level. The accuracy values for the other models show a slight decline compared to their baseline performance (Table 2), generally in the range of 0.7-08. The Neural Network remains the worst-performing model with an accuracy of 77.8%, experiencing a 1.8% decrease from its baseline. Logistic Regression and Random Forest also show notable declines in accuracy, dropping by 3% and 2%, respectively.

In terms of precision, the best-performing model is SVM with a precision of 68.4%, though it suffering no decrease from its baseline. GBM follows with a precision of 67.9%, showing a 5.6% decline. The Neural Network again shows the poorest performance with a similar percentage decrease from its baseline.

Almost all models experience substantial declines in recall, with Neural Network showing the largest drop of 12.8%, resulting in a recall of only 39.5%. Logistic Regression and Random Forest both see their recall drop by around 10% and 9%, respectively, indicating a significant impact from the data poisoning. We do see that GBM has a percentage increase of 2.3% when compared to its baseline.

The F1 scores reflect similar trends, with SVM performing the best. The Neural Network experiences the most significant decline in F1 score, dropping by 9% to 49.1%. Logistic Regression, GBM, and Random Forest also show notable declines in F1 score, with decreases of around 8%, 1% and 6% respectively.

In summary, the results indicate that SVM does not suffer under small levels of poisoning. Logistic Regression, while fairly accurate, suffers from a significant drop in recall, missing a

considerable number of true churn cases. SVM, although good at avoiding false positives, also shows a decline in recall. GBM's high precision is undermined by a drop in recall, while Random Forest, despite maintaining a balance, shows reduced performance across all metrics. The Neural Network, with the largest drops in recall and F1 score, proves to be the least robust model against this level of data poisoning, limiting its utility for churn prediction in adversarial scenarios.

4.2.2 Three Percent Level of Attack Severity

In Table 4, the performance metrics for the models continue to decline as the severity of label-flipping increases to 3%. The Neural Network remains the worst-performing model with an accuracy of 77.2%, which is a 2.5% decrease from its baseline and a further 0.6% drop from the 1% level. Logistic Regression, Random Forest, and SVM also show similar declines in accuracy, dropping by around 3% from their baseline.

Regarding precision, the best-performing models are Neural Network and Logistic Regression, both around 65.5%. However, both models show declines compared to their baseline (0.2% and 2.2%, respectively). It is worth noting that the logistic regression suffered a larger drop in performance from the 1% level at approximately 2.3% while the Neural Network only dropped by around 0.5%. GBM's precision drops to 66.7%, a 7.4% decrease from its baseline and a further 1.2% drop from the 1% level.

When examining recall, GBM while still relatively high, drops by 1.7% from its baseline but notably shows an improvement from the 1% level (from 0.517 to 0.594). Other models experience more substantial declines, with the Neural Network showing a drop of 9% from its baseline and a further 1.9% from the 1% level, resulting in the lowest recall of only 41.3%. SVM sees its recall drop by 1.6% from its baseline and an additional 0.88% from the 1% level, indicating a significant impact from the increased severity of data poisoning.

The F1 scores follow the same pattern. The Neural Network experiences the most significant decline in F1 score, dropping by 5.5% from its baseline. Logistic Regression, GBM, and Random Forest also show notable declines in F1 score, with decreases of 8.9%, 5.8%, and 7.5% from their baseline, respectively.

In summary, the results indicate that GBM is the best performing model under the 3% label-flipping attack, achieving higher scores than the rest in most of the four metrics. Logistic Regression, while maintaining a reasonable level of accuracy, continues to suffer from significant drops in recall, missing a considerable number of true churn cases. SVM, although good at avoiding false positives, also shows a substantial decline in recall. GBM's performance degradation is counterbalanced by a notable increase in recall from the 1% level; while Random Forest, despite maintaining a balance, shows reduced performance across all metrics.

The Neural Network, with the largest drops in recall and F1 score, remains the least robust model against data poisoning. The degradation in performance from the 1% to 3% levels underscores the increased vulnerability of these models to more severe data poisoning attacks.

4.2.3 Five Percent Level of Attack Severity

Accuracy values decline further as the label-flipping severity increases to 5%, as presented in Table 5. The Neural Network remains the least accurate model with an accuracy of 76.1%, representing a 3.9% decrease from its baseline. Logistic Regression and Random Forest both show similar decreases in accuracy, dropping by 5.9% and 4.6% from their baseline respectively. The biggest drop was from the GBM, falling by 6.3%

Regarding precision, SVM and Neural Network have scores of around 63%, both experiencing slight declines from their baseline (3.8% and 3.9%) and minimal changes from the 3% level. GBM's precision drops to 61.9%, reflecting a 14.1% decrease from its baseline. Neural Network's precision also decreases slightly by 3.96%.

When examining recall, Logistic Regression shows a largedrop, decreasing by 19% from its baseline and further declining from the 3% level to 47.6%. SVM's recall falls by 19.9% from its baseline, highlighting its increased vulnerability to severe data poisoning. GBM, while still relatively high in recall at 61.2%, experiences a large increase of 21.06%, indicating some resilience. Neural Network's recall drops by 2.6% from its baseline but remains relatively stable from the 3% level.

The F1 scores show that Logistic Regression, SVM, and Random Forest suffer declines in F1 score, dropping by 13.2%, 13.5%, and 8.4% from their baseline, respectively, and further

decreasing from the 3%. GBM shows a slight improvement of 3.6% from the 1% level, indicating some recovery in performance.

4.3 Adversarial Training Model Performance Metrics

When adversarial training is applied, the performance metrics of the models indicate an overall improvement in robustness against data poisoning attacks, as shown in Table 7.

Table 7

Performance metrics for all models with adversarial training, tested against 5% level of attack severity

Model	Accuracy	Precision	Recall	F1_Score
Logistic Regression	0.781	0.651	0.508	0.571
SVM	0.798	0.688	0.510	0.585
GBM	0.801	0.656	0.502	0.569
Random Forest	0.897	0.910	0.950	0.569
Neural Network	0.772	0.653	0.407	0.503

Note. All analyses were performed in R with the Telco Customer Dataset from Kaggle.

Logistic Regression shows a slight recovery, achieving an accuracy of 78.1%, which is a 1.52% increase from its accuracy at the 5% attack level. Its precision improves to 65.1%, representing a 2.72% increase from the 5% level, while recall improves to 50.8%, up by 3.29%. The F1 score also increases slightly to 57.1%, a 2.8% improvement.

SVM demonstrates a notable improvement in its metrics, with accuracy increasing to 79.8%, up by 3.9% from the 5% attack level. Its precision rises to 68.8%, an increase of 3%, and recall improves to 51.0%, up by 8.05%.

The F1 score improves to 58.5%, marking a 6.2% increase. GBM also benefits from adversarial training, with an accuracy of 80.1%, a 4.72% increase from the 5% level. Precision rises to 65.6%, a 3.7% improvement, and recall remains relatively stable at 50.2%, a slight decrease of 1.9% from the 5% level but still an overall improvement in robustness. The F1 score is 56.9%, showing a minor increase of 1.1%.

Random Forest shows the most significant improvement among the models, with its accuracy rising to 89.7%, a remarkable 17.1% increase from the 5% attack level. Precision improves significantly to 91%, up by 27.8%, and recall increases to 95%, an impressive 47.4% rise. However, the F1 score remains at 56.9%, showing that while individual metrics have improved dramatically, the overall balance between precision and recall still needs attention.

The Neural Network shows some recovery with an accuracy of 77.2%, a 1.11% increase from the 5% attack level. Precision improves to 65.3%, up by 3.4%, and recall increases slightly to 40.7%, a 1.4% improvement. The F1 score also rises to 50.3%, showing a 2.9% increase.

Overall, the application of adversarial training significantly enhances the robustness of the models, particularly Random Forest and SVM, which show substantial improvements across most metrics. While Neural Network and Logistic Regression show some recovery, their lower recall and F1 scores suggest they are still less reliable under adversarial conditions compared to the other models.

These results highlight the effectiveness of adversarial training in bolstering model resilience, making it a valuable strategy for maintaining the integrity of marketing models against data poisoning.

Table 8

Performance metrics for all models with adversarial training, tested on unpoisoned data

Model	Accuracy	Precision	Recall	F1_Score
Logistic Regression	0.780	0.651	0.508	0.570
SVM	0.822	0.714	0.550	0.621
GBM	0.755	0.816	0.102	0.182
Random Forest	0.927	0.928	0.974	0.950
Neural Network	0.728	0.592	0.322	0.491

Note. All analyses were performed in R with the Telco Customer Dataset from Kaggle.

Table 8 above summarises the performance metrics for all models with adversarial training, now tested on unpoisoned data. Logistic Regression shows consistent performance with an accuracy of 78%, which is a slight decrease of 0.13% from its accuracy with adversarial training against

5% attack severity. Precision remains stable at 65.1%, with no change from the 5% attack level, while recall also stays steady at 50.8%. The F1 score shows a minor decrease.

SVM demonstrates significant improvement when tested on unpoisoned data, with accuracy increasing to 82.2%, up by 2.9% from the 5% attack level. Precision rises to 71.4%, an increase of 3.8%, and recall improves to 55%, up by 8%. The F1 score improves to 62.1%, marking a 6.2% increase.
GBM exhibits mixed performance, with accuracy decreasing to 75.5%, a 5.8% drop from the 5% attack level. Precision jumps significantly to 81.6%, representing a 24.4% increase, while recall drops sharply to 10.2%, a decrease of 79.5%. The F1 score falls to 18.2%, showing a 67.9% reduction, indicating a trade-off between precision and recall.

Random Forest shows substantial improvement, with accuracy rising to 92.7%, a notable 3.4% increase from the 5% attack level. Precision improves to 92.8%, up by 2.9%, and recall increases to 97.4%, an impressive 2.5% rise. The F1 score jumps to 95%, a significant improvement of 66.9%, highlighting the model's enhanced performance on unpoisoned data.

The Neural Network shows reduced performance on unpoisoned data, with accuracy decreasing to 65.8%, a 14.7% drop from the 5% attack level. Precision falls to 59.2%, down by 9.4%, and recall drops to 32.2%, a 20.6% decrease. The F1 score slightly decreases to 49.1%, showing a 2.4% reduction.

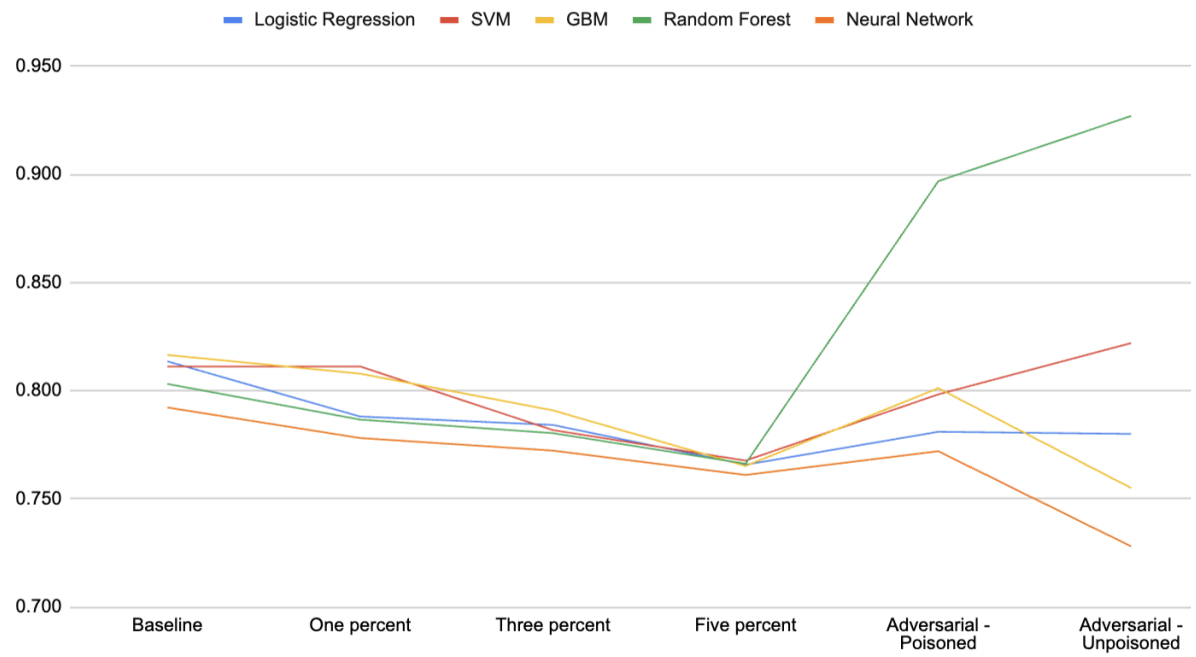
Overall, the application of adversarial training significantly enhances the robustness of models, particularly SVM and Random Forest, which show substantial improvements across most metrics when tested against unpoisoned data. The Random Forest model, in particular, achieves remarkable precision and recall, making it highly reliable under both adversarial and normal conditions.

However, models like GBM and Neural Network exhibit mixed results, indicating that while adversarial training can improve precision, it might not necessarily enhance recall or F1 score uniformly across all models. These results highlight the effectiveness of adversarial training in bolstering model resilience, making it a valuable strategy for maintaining the integrity of marketing models against data poisoning.

4.4 Trends

Figure 5

Accuracy of models across different levels of data poisoning and adversarial training



In Figure 5 above, we see that all models experience a decline in accuracy as the severity of data poisoning increases. Specifically, Neural Network consistently shows the lowest performance, indicating it is the most vulnerable to label-flipping attacks.

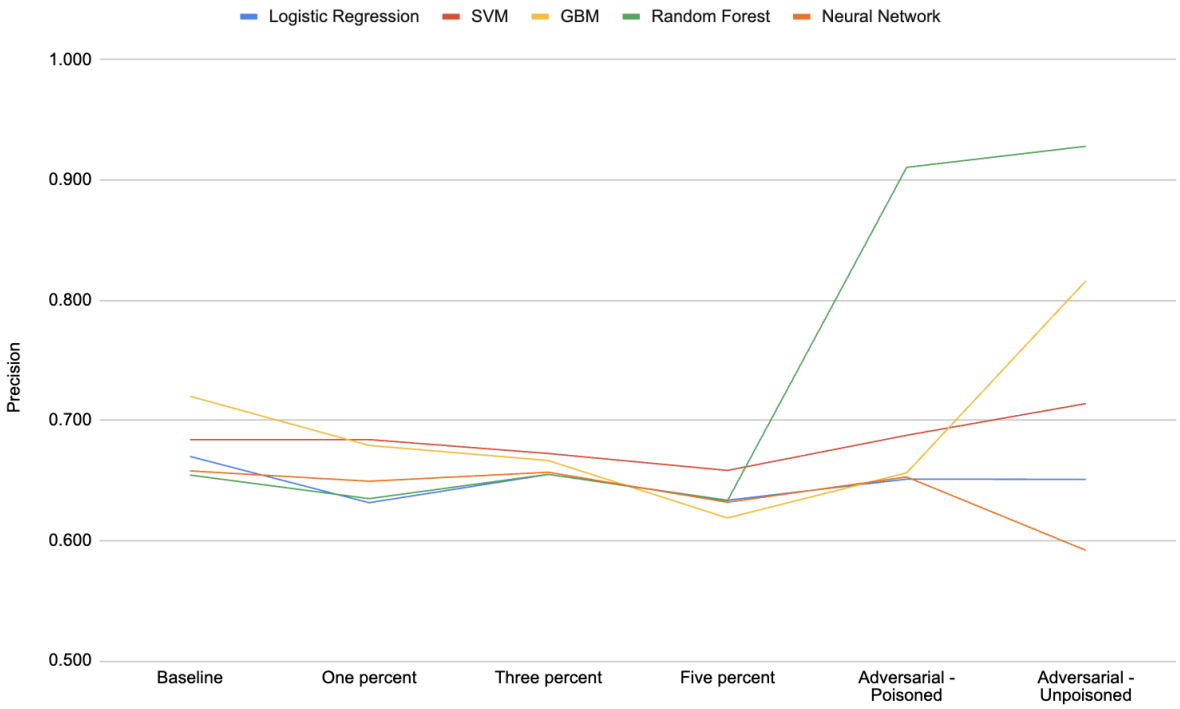
Adversarial training significantly improves the performance of all models when tested on poisoned data. Random Forest shows a remarkable recovery, with accuracy increasing dramatically after adversarial training. SVM and GBM also show substantial improvements, while Logistic Regression and Neural Network exhibit moderate gains but still lag behind the other models.

However, when testing on unpoisoned data, only RF and SVM see an improvement. The rest of the models suffer small or large drops in accuracy. GBM and NN's performance degrades the most.

Overall, Random Forest after adversarial training stands out as the best performer. The Neural Network consistently underperforms, particularly in poisoned scenarios.

Figure 6

Precision of models across different levels of data poisoning and adversarial training



In Figure 6, the models' precision over the phases of analysis is visualised. The similarity in the trends to accuracy is that the best-performing model is Random Forest after adversarial training. However, it is important to recognise that Random Forest is one of the worst-performing models until adversarial training has been employed; as with accuracy, all models benefit from adversarial training when tested on poisoned data. On unpoisoned data, we see that RF and GBM benefit the most, seeing stark increases in this metric. NN drops the most in precision when adversarial training is tested on unpoisoned data, as is seen in Figure 6 clearly.

One notable difference from accuracy is that GBM is the lowest-performing model at the 5% level of poisoning, despite the eventual recovery after adversarial training. Another clearly visible distinction is that the metrics increase when moving from poisoning levels of 1% to 3% and reduce slightly again for 5%. Lastly, as expected from the results aforementioned in earlier sections, the Neural Network and Logistic Regression models are consistently the worst performing over most phases.

Overall, the best-performing model for precision remains the Random Forest.

Figure 7

Recall of models across different levels of data poisoning and adversarial training

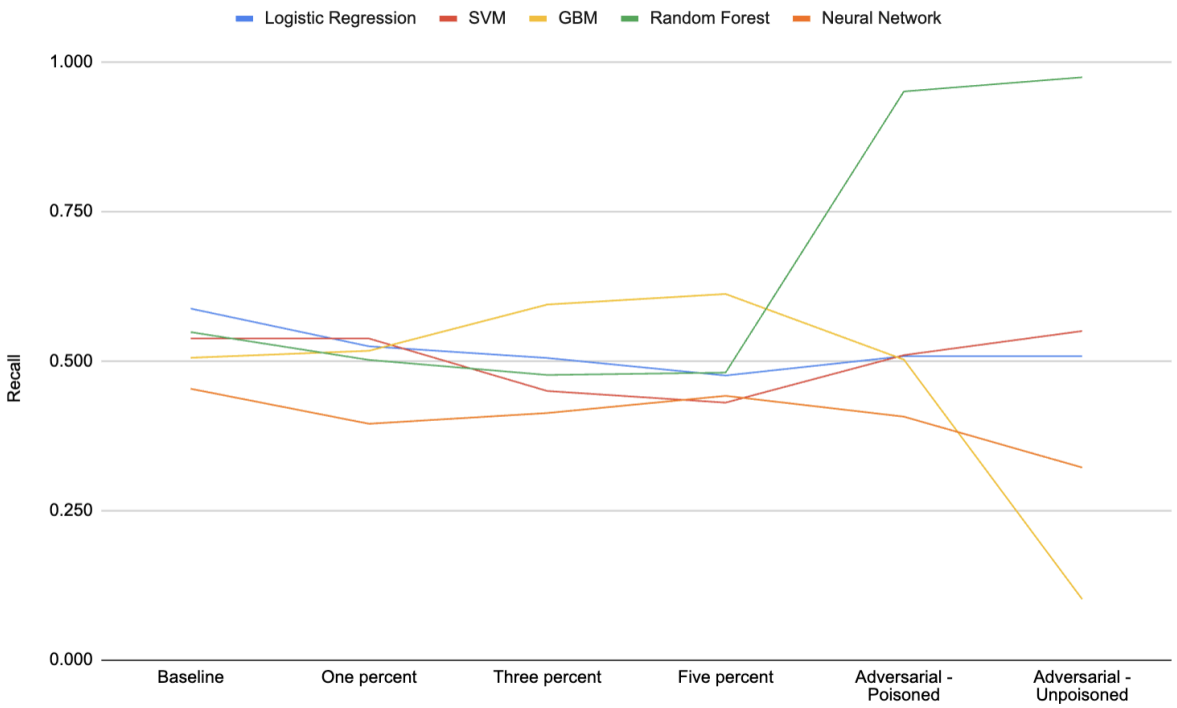


Figure 7 examines Recall, which is one of the more important measures in our research. In this metric, each model experiences different impacts depending on the phase it is in.

The logistic regression model starts as one of the best-performing models, but quickly drops, and ends in fourth. The SVM is stable until the level of poisoning increases, becoming the worst model at the 5% level. This model does recover significantly after adversarial training and performs slightly better than the logistic regression. GBM does exceptionally well throughout the phases but suffers from a substantial decline in the last phase. The Neural Network model trails around 0.4 and drops even further when the adversarially trained model is tested on unpoisoned data. GBM suffers a large drop in recall in the last two phases, making it the least preferable model for our research. Most other models do not see drastic changes.

Lastly, Random Forest is the most preferable model for real-world usage as it would be fruitful in predicting actual churn cases. This is especially considering that after being trained adversarially, and tested on poisoned and unpoisoned data, this model does exceptionally. In addition, one could do supplementary analyses to discover the cause behind customer churn by computing variable importance or similar measures.

Figure 8

F1 score of models across different levels of data poisoning and adversarial training

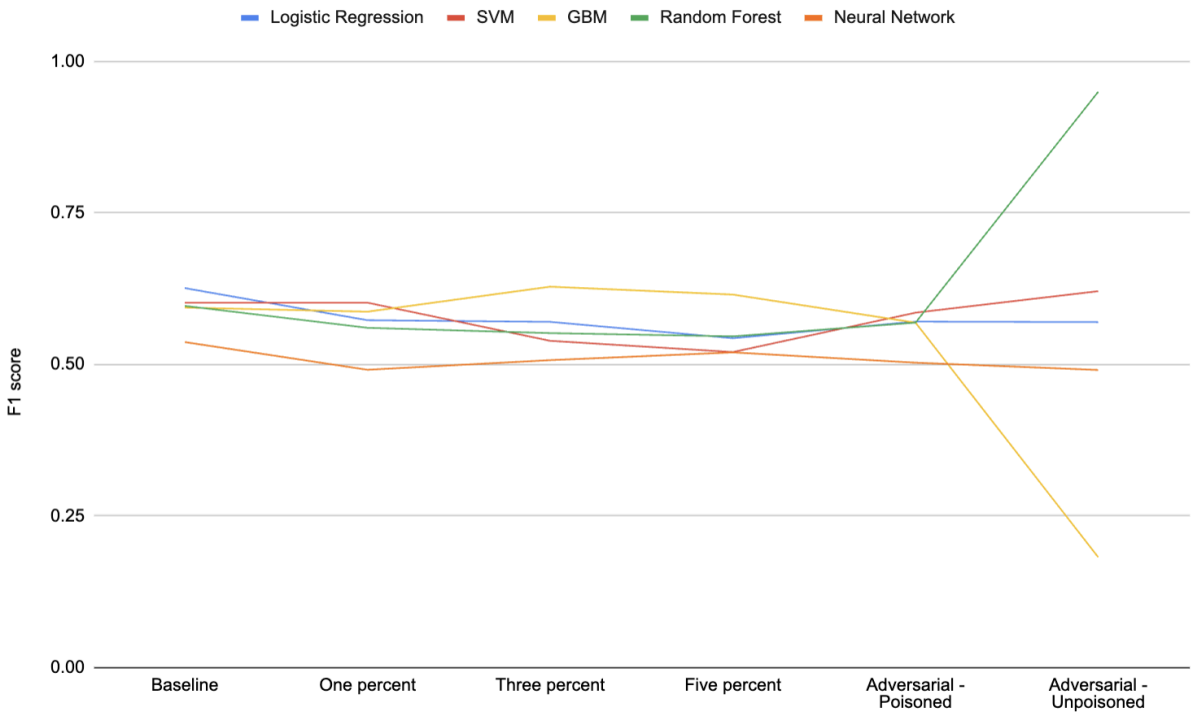


Figure 8 displays the F1 score of the models which perform quite differently to each other depending on the phase. GBM performs far better than the rest of the models at the 3% and 5% levels of poisoning, making it resilient. It is worth noting here that GBM and Neural Network both decrease after adversarial training and increase when exposed to additional levels of poisoning.

Logistic regression remains relatively stable across poisoning levels, similar to the Random Forest Model. Contrastingly, SVM slopes downwards between each poisoning level (although 1% does not impact its performance), and recovers after training.

We do see a difference in behaviour after the adversarially trained models are tested on unpoisoned data. Random Forest does extremely well, SVM, LR, and NN stay relatively stable, and GBM's performance lacks significantly. Overall, RF emerges yet again as our most reliable model for predictions.

4.5 Results Summary

Provided below is a table that summarises the best performing model for each metric, for all levels and trainings implemented.

Table 8

The results summary of the best-performing models

	Accuracy	Precision	Recall	F1 score
Baseline	GBM	GBM	LR	LR
One percent	SVM	SVM	SVM	SVM
Three percent	GBM	SVM	GBM	GBM
Five percent	RF	SVM	GBM	GBM
Adversarial - Poisoned	RF	RF	RF	GBM/RF
Adversarial - Unpoisoned	RF	RF	RF	RF

Based on these results, prior to adversarial training GBM/SVM seem to be the best performing models. However, after this training has been tested on poisoned and unpoisoned data, the most successful model is unequivocally Random Forest.

Chapter 5: Conclusions and Reflections

5.1 Conclusions

This study set out to investigate the impact of data poisoning on the performance metrics of various marketing models, with a particular focus on label-flipping attacks. The results provide a comprehensive view of how different machine learning models respond to varying levels of adversarial attacks, shedding light on their robustness and effectiveness in maintaining predictive performance.

Central Research Question: *How does data poisoning affect the performance metrics of marketing models at different levels of attack severity?*

Data poisoning, particularly through label-flipping attacks, certainly impacts the performance metrics of marketing models. As the severity of the attack increases, there is a noticeable decline in accuracy, precision, recall, and F1 score for most models. The Neural Network model exhibits

significant performance degradation, particularly in recall and F1 score, under adversarial conditions.

Subquestion 1: *Which models demonstrate greater resilience to label-flipping attacks and what are the underlying mechanisms contributing to their robustness?*

Random Forest after adversarial training demonstrates the greatest resilience to label-flipping attacks (see Table 8). This robustness can be attributed to its ensemble approach, which mitigates the impact of corrupted data by averaging multiple decision trees (Yerlikaya & Bahtiyar, 2022).

Subquestion 2: *What methods can be employed to enhance the robustness of models against data poisoning attacks?*

Adversarial training is identified as an effective method to enhance model robustness against data poisoning attacks (Goodfellow et al., 2014). This technique involves incorporating adversarial examples into the training process, thereby preparing the models for potential attacks and improving their ability to identify and mitigate misleading inputs. This method significantly improved the performance of most models, particularly Random Forest and SVM.

Subquestion 3: *What defensive strategies can be implemented to shield marketing models from the adverse effects of data poisoning?*

The defensive strategy of maintaining a clean copy of the original dataset and correcting discrepancies between the original and new dataset proved effective. Another approach allows for the identification and cleaning of poisoned data by comparing the proportions, expected levels, and ranges within the dataset to spot significant outliers (Qiu, 2022). While this method may be challenging for label-flipping attacks, as they only alter one column, it can be effective against other techniques by highlighting inconsistencies and unusual patterns.

Subquestion 4: *At which level of data contamination does the integrity of marketing models begin to significantly deteriorate, impacting key performance metrics such as accuracy, precision, recall, and F1 score?*

The study indicates that model performance begins to significantly deteriorate at even low levels of data contamination. Notable degradation is observed at the 1% label-flipping level, with further declines at the 3% and 5% levels. Neural Networks, in particular, show marked performance drops across all metrics, while models like Logistic Regression and SVM also exhibit significant declines in recall and F1 score as contamination increases.

Subquestion 5: *What techniques are effective in detecting and sanitising poisoned data?*

Maintaining a clean copy of the original dataset and analysing data proportions and ranges can help detect significant outliers indicative of data poisoning. Effective techniques for detecting and sanitising poisoned data include anomaly detection methods and statistical analysis to identify discrepancies between original and corrupted datasets. Adversarial training also plays a dual role by not only preparing models for attacks but also helping to identify poisoned data during the training process.

5.2 Reflections

Reflecting on the research process, several key insights emerge that could inform future work in this area. The choice of machine learning models proved to be a critical factor in determining the resilience of the system against adversarial attacks. While Random Forest and GBM showed remarkable robustness, both simpler models (like Logistic Regression) and more complex ones (like Neural Networks) exhibited significant vulnerabilities. This suggests that the complexity of a model does not necessarily correlate with its resilience, emphasising the need for careful selection and evaluation of models based on the specific context and threat landscape it will be applied in.

The experimental design, which involved controlled label-flipping attacks at different severity levels, provided a clear and structured approach to understanding the impact of data poisoning. However, real-world scenarios may involve more complex and varied types of adversarial attacks, necessitating broader studies that encompass a wider range of attack methods and conditions.

Adversarial training emerged as a highly effective defensive strategy, significantly bolstering model performance across various metrics. Future research could explore the integration of other

defensive techniques, such as anomaly detection and data sanitisation, to develop a multi-layered defense mechanism that offers comprehensive protection against adversarial attacks (Qiu, 2022).

Maintaining a clean copy of the original dataset and identifying discrepancies proved to be effective and low-effort defensive strategy. However, the challenges posed by label-flipping attacks highlight the need for continuous improvement and adaptation of these strategies. The importance of continuous monitoring and updating of models in response to evolving threats cannot be overstated. As adversarial tactics become more sophisticated, static models and defense strategies may quickly become obsolete. Ongoing research and adaptation are essential to stay ahead of potential threats and ensure the long-term effectiveness of marketing models.

In conclusion, this research provides valuable insights into the impact of data poisoning on marketing models and highlights the critical need for robust and adaptive defense mechanisms. By advancing our understanding of these threats and the efficacy of various defensive strategies, this study contributes to the development of more resilient machine learning applications in marketing and beyond.

Bibliography

AlShourbaji, I., Helian, N., Sun, Y., Hussien, A. G., Abualigah, L., & Elnaim, B. (2023). An efficient churn prediction model using gradient boosting machine and metaheuristic optimization. Scientific Reports, 13, 14441. https://doi.org/10.1038/s41598-023-41093-6

Badole, M. (2023, March 31). Customer churn prediction using artificial neural network. Analytics Vidhya. Retrieved from https://www.analyticsvidhya.com/blog/2021/10/customer-churn-prediction-using-artificial-neural-network/

Blastchar. (2018). Telco Customer Churn. Kaggle. Retrieved from https://www.kaggle.com/datasets/blastchar/telco-customer-churn

Cole, A. (2020, May 13). Predicting customer churn using logistic regression: Part 2: Building the model. Towards Data Science. Retrieved from https://towardsdatascience.com/predicting-customer-churn-using-logistic-regression-c6076f37eaca

Davi, L., et al. (2004). Game-Theoretic Approaches to Attack and Defense in Cybersecurity. Retrieved from https://homes.cs.washington.edu/~pedrod/papers/kdd04.pdf and further explained at https://cseweb.ucsd.edu/~akmenon/AdversarialTalk.pdf.

Fox, J. (2023, July 26). Cobalt.io Data Poisoning Attacks: A New Attack Vector Within AI. Retrieved from https://www.cobalt.io/blog/data-poisoning-attacks-a-new-attack-vector-within-ai

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and Harnessing Adversarial Examples. Retrieved from https://arxiv.org/abs/1412.6572

Gupta, A., & Krishna, A. (2023). Adversarial clean label backdoor attacks and defenses on text classification systems. In Proceedings of the 8th Workshop on Representation Learning for NLP (RepL4NLP 2023) (pp. 1-12). Toronto, Canada: Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.repl4nlp-1.1

Indriati, F. (2023, January 4). Customer churn prediction in the banking sector using support vector machine (SVM) in Python. Medium. Retrieved from https://fiqey.medium.com/bank-customer-churn-prediction-using-support-vector-machine-svm-82d206cf7206

Johnson, D., & Lee, A. (2020). The impact of data integrity on marketing analytics: A case study. Journal of Marketing Analytics, 8(3), 145-159. https://doi.org/10.1057/s41270-020-00073-3

Qiu, W. (2022). A Survey on Poisoning Attacks Against Supervised Machine Learning. Electrical and Computer Engineering, University of Toronto. Retrieved from https://ar5iv.labs.arxiv.org/html/2202.02510

Rosenfeld, E., Winston, E., Ravikumar, P., & Kolter, J. Z. (2020). Certified robustness to label-flipping attacks via randomized smoothing. arXiv. https://arxiv.org/pdf/2002.03018

Sharma, A. (2021, May 3). Applying random forest on customer churn data. Data Science on Customer Churn Data. Retrieved from https://medium.com/data-science-on-customer-churn-data/applying-random-forest-on-customer-churn-data-53883efb25bf

Smith, R. (2021). Adversarial attacks in digital marketing: Understanding and mitigating data poisoning. Digital Marketing Review, 15(2), 67-83.

Steinhardt, J., Koh, P. W., & Liang, P. (2017). Certified Defenses for Data Poisoning Attacks. Retrieved from https://proceedings.neurips.cc/paper/2017/hash/9d7311ba459f9e45ed746755a32dcd11-Abstract.html

Yerlikaya, F. A., & Bahtiyar, Ş. (2022). Data poisoning attacks against machine learning algorithms. Department of Computer Engineering, Istanbul Technical University. Retrieved from https://doi.org/10.1016/j.eswa.2022.118101