Implementing Differential Privacy in Machine Learning to Predict Customer Churn

Naramol Pipoppinyo, [redacted]

MScBA Business Analytics & Management
Master Thesis

Coach - [redacted]
Co-reader – [redacted]

June 15, 2023

Preface

The copyright of the Master Thesis rests with the author. The author is responsible for its contents. RSM is only responsible for the educational coaching and cannot be held liable for the content

Acknowledgements

I would like to give a huge thank you to Dr. Jason Roos for his support throughout the thesis trajectory with my constant messages. I would also like to thank Dr. Olga Slivko who was always keen on helping go through my thesis. A special appreciation to Mr. Spencer Giddens, for being willing to help when there was never an obligaton to. I really appreciated the guidance from a topic expert because whenever I was lost, you were able to guide me. Lastly, a shoutout to Dylan, Dat, Lisa, Abdu, Yuchi, Eran, and everyone else who listened to my struggles and/or was always checking in on how I was doing during the trajectory.

Abstract

Using machine learning for customer churn prediction relies heavily on detailed customer data to create accurate predictions. Although, as more consumers are aware of their data being used, more and more consumers are concerned about their privacy. Marketers are set with the task of creating churn models which use data in a legal and ethical privacy-preserving way whilst still having high performance. This research aims to leverage differential privacy (DP) to balance the trade-off between privacy and the performance utility of the churn models. Techniques such as output perturbation and objective perturbation are applied under empirical risk minimization to Regularized Logistic Regression (RLR) and Support Vector Machine (SVM). The churn data borrowed from Telecommunications demonstrate the trade-off exists, however, heavily depends on the combination of the model and DP technique chosen. RLR best demonstrates the trade-off where when privacy grows, the model performance worsens. SVM demonstrates large random variation making it difficult to draw a concrete analysis. This research concludes the practicality of the models in industrial applications is heavily dependent on the use case as well as the nature of the data in terms of data imbalance, dimensionality, and complexity.

Table of Contents

1. Introduction

1.1. Problem background

1.1.1 Privacy Concerns in Today's World

Over the last decade, marketers heavily rely on statistics and customer data to personalize campaigns and communication with consumers (Lin, 2022). As more consumers are aware of their data being used, more consumers are concerned that their privacy is at risk. As a result, policy makers are pushing to enact new laws and regulations that protect individual privacy. For example, in 2016, the General Data Protection Regulation (GDPR) was proposed by the European Union (EU) to protect the fundamental rights and freedoms of individuals' personal data (Marketing Evolution, 2020; IMY., 2021). After the EU acted, many countries around the world registered the gravity of the issue and came up with regulations of their own. For example, in 2018, the California Consumer Privacy Act (CCPA) was passed and in 2020, the "Lei Gerel de Proteçao de Dados" of Brazil was modeled directly after the GDPR.

1.1.2 Impact on Marketers

The rise of privacy regulations has impacted marketers greatly. As reported by Gartner, approximately one out of five marketers reported privacy compliance as a major concern (Marketing Evolution, 2020). The laws and regulations have evolved the way organizations are allowed to gather, manage, and use data. A marketer's ability to garner useful insights is diminishing as less information is available than before. Hence, marketing campaigns may not be as impactful as before "privacy concerns" became a fad. There is also added pressure from the consumers' side as it has been published that 8 out of 10 customers will abandon a brand if their data is being used without their consent (Marketing Evolution, 2020). Using non-consensual data that breach regulations can negatively impact a brand's reputation, the opposite effect of what marketers hope for.

1.1.3. Machine Learning Privacy Considerations in Customer Churn

Customer churn rate is the percentage of customers that quit using a company's product service. Marketing teams may focus on customer churn to understand why their organization is losing customers and how they can create retention campaigns to entice customers to stay. Customer retention is important for marketers as it has been shown that trying to gain more customers is more costly than retaining existing customers (Landis, 2022). Therefore, predicting what makes someone churn is beneficial in creating marketing that entices customers to stay. This can be done through machine learning, a practice that develops algorithms that learn patterns and make decisions based on data. Churn prediction through machine learning requires the use of customer information to understand the variables that will impact churn. For example, variables such as demographic information (e.g., age, job title, location, civil status), customer account information (e.g., payment method, billing, how long they have been with the company), or services (e.g., subscription plan, features indicating customer preferences), can be used to train the algorithm that will predict patterns and behaviors of a "churner" (Smolic, 2022). With machine learning, models are created to learn and improve their predictive accuracy, practically applied through frameworks such as Empirical Risk Minimization (ERM). ERM is used in machine learning as a method of minimizing error so that the data used to create the algorithm is generalizable to other unseen data.

The data used for the machine learning models become less and less informative as a consequence of the privacy regulations in place. Consumers can opt out of data collection which can mean the database is not representative of the full clientele that organizations have (Marketing Evolution, 2020). Additionally, marketers are just not allowed to collect and use the same amount of detailed information as before. This can negatively impact the validity and performance of the model results. Less data about the customer means more privacy but it can also mean less accuracy of the predictive models. Consequently, there is a trade-off between privacy and model utility that marketers need to consider. A method suggested is known as Differential Privacy (DP), a mathematical criterion that adds noise to the data to increase privacy (Dilmegani, 2021). Although privacy is the main goal, specific DP implementation also considers data utility by controlling the amount of noise added to the model.

1.1.4. Differential Privacy

Differential privacy adds noise to the data so that the value is not exactly accurate, but enough to achieve the goal of a query to ensure privacy on an individual level. Famous use cases of DP are within technological organizations like Apple. Apple uses DP to protect the privacy of user activity during a given period when giving Emoji suggestions. Apple ranks the most popular Emojis used based on user data. By implementing DP, Apple protects an individual's personal data while still being able to suggest the most popular Emojis to its users

(Verger, 2017). The use of DP is wide and can be placed in all different industries that use machine learning modeling for their predictions. Thus, it would be of interest to marketers to understand how DP can be implemented with their machine learning models such as for customer churn. This would demonstrate how marketers can ensure the utmost privacy of an individual while still providing accurate insights to drive marketing decisions.

1.2. Problem Statement & Research Question

As a result of the lack of understanding of how effective DP is in the context of marketing and customer analytics, the aim is as follows: understanding how differentially private techniques are effectively implemented in predicting customer churn. To achieve this aim, the research question is proposed:

To what extent does implementing differential privacy via Empirical Risk Minimization balance the trade-off between the privacy and the utility of a model in predicting customer churn?

To answer the research question, *"implementing differentially privacy"* will be done by using DP "data perturbation" technique. Data perturbation adds noise to the data and can happen at different stages of the model training process. The research will focus on two differentially private techniques: output perturbation and objective perturbation, which will be applied during model training. Model training will be done "*Via empirical risk minimization"* as the principle attempts to minimize error and create accurate models. Popular models used in both churn and DP research will be considered including Regularized Logistic Regression (RLR) and Support Vector Machines (SVM). In an ideal world, both privacy and utility can be optimized. However, having both very high privacy and utility is not always realistic and so to consider *"balancing the trade-off",* two things are important to keep in mind: how the perturbation techniques can retain privacy yet still allow marketers to create high-performance churn models that increase the utility of the model. To answer the research question, the churn models and perturbation methods applied will be compared to understand which is better in balancing the trade-off. This entails that the research will focus on analyzing how the statistical model of choice, privacy technique of choice, as well as the nature of the data affect the trade-off between privacy and model performance. Model performance is used to measure the utility of the model: higher model accuracy means higher utility. The research will borrow data from the industry of Telecommunications (Telco) where churn prediction is highly relevant (more details will be covered in the theoretical background).

1.3. Managerial Relevance

There is a lack of research that applies DP Perturbation and does a clear comparative analysis with a specific industrial use case (Chaudhuri et al., 2011; Rubeinstein et al., 2009). This makes it hard for marketers to see the practicality of research. By understanding how to effectively implement DP in customer churn modeling, marketers can analyze and share private data without having to reveal an individual's sensitive information (Dilmegani, 2022). Such a method may be enticing to organizations as it tries to comply with data privacy regulations like the GDPR, while trying to maintain accurate analysis of customer behavior. Companies would not have to give up the predictive power of models for privacy. Or more realistically, companies can understand and find the amount of utility they are willing to sacrifice for privacy. The trade-off also implies the different ways marketers should make organizational decisions. For example, implementing stronger privacy preserving values means marketers would need to rely on less accurate data and more aggregated predictions to create more generalized targeting approaches. Not only does it impact marketing decisions, but managers also need to understand how to communicate their privacy policies and manage stakeholder relations. The choice of balance will cause different reactions depending on stakeholders and their relation to the organization (i.e., hypothetically, a CEO would prioritize profit whereas a customer would prioritize privacy considerations). DP helps marketers find ways to ensure accuracy and utility while still being socially responsible with data.

1.4. Academic Relevance

Currently, DP is widely researched, and machine learning is a complimentary process that can help understand the direction of DP research and its key issues. That is, DP uses private data publishing and data analysis, focusing on the nature of the input/output data, and the different mechanisms applied in machine learning (Zhu et al., 2017). This research will specifically focus on DP data analysis which, aims to publish an approximately accurate analysis model rather than focusing on query answering or DP-synthetic datasets which, also fall under the DP methodologies. Not only is DP widely researched, but predicting customer churn with machine learning is also a commonly studied topic. However, it seems like there is a gap between using DP in customer churn and industry use in general, and thus, academic research on the practicality of DP is a gap in the literature. Only the paper by Hu et al. (2015) has explored DP in real-world industrial data mining systems such as in Telco to predict customer churn. Although the research of Hu et al., (2015) is a specific real-world focus on churn in DP, their focus is limited to DP architectures and not exactly machine learning. The comparison of private algorithms with ERM to understand the trade-off between utility (measured by model performance) and the privacy parameter has not been explicitly researched within the industry of Telco customer analytics and specifically, customer churn. Researchers have presented research on trade-offs, however, looking more into the idea of utility and "fairness" in DP thus, differs in the aim and objective of the research (Xu et al., 2019). This research aims to contribute to the existing body of knowledge by using referencing relevant techniques proposed by well-cited researchers in the field (e.g., Chaudhuri and Dwork) to generalize to real-life cases.

2. Theoretical background

The main objective of this section is to draw the narrative and tie the theory of customer churn and DP to understand the main research approach. Previous and relevant literature will be raised to demonstrate the evolution of privacy preserving techniques and how the decision of the research focus came to be.

2.1. Customer Churn in Telecommunications

Churn is the rate at which a companies' customer stops doing business with the company (Mehta & Steinman, 2016). Specifically in Telco, customers unsubscribe or switch providers which heavily affects the companies' revenue (Modani et. al, 2013). Towards the end of 2021, the industry had an average churn rate of 31% which, can have negative financial consequences for the firm including loss of revenue, loss of opportunity, and increased cost of acquisition (Tessitore, 2023). As a result, one of the most important obstacles in Telco customer retention is in fact, churn management (Duchemin & Matheus, 2021). This is precisely why Telco was chosen an exemplary industry for this research, as it represents a highly relevant use-case scenario. Churn management is a process for organizations to retain their profitable subscribers (Mahajan et al. 2015). A method within the process can include predictive machine learning models. Predicting future trends and behaviors of customers has increased in popularity within customer analytics and even specifically for customer churn in Telco (Mahajan et al., 2015). Such models can have a positive impact on customer retention due to the results of the algorithm's prediction. It allows organizations to remain proactive and make decisions before a loss of profit occurs. For example, models can predict which customer segment is most likely to discontinue their mobile contract and so marketers are able to target those segments before they do churn for good. Churn research has even found ways to incorporate profit-based loss functions to help firms increase profitability of retention campaigns (Lemmens & Gupta, 2017).

2.2. Modeling Customer Churn

A variety of different models have been implemented in predicting churn. RLR is simple and easily interpretable which is advantageous for organizations with limited proficiency in analysis. RLR is a supervised machine learning model for binary classification which has a regularized term. The regularized term helps reduce variance and overfitting, which, is when the model trains too well with the training data, making it difficult to generalize to unseen data (Chang et al., 2022). Due to its advantages, RLR is the most frequently used model for churn forecasting (Ha et al., 2005). By using RLR, the organization can still manage churn but do not have to dedicate high levels of resources. Focusing the research on RLR ensures the research applicability is expansive, reaching out to organizations with all skill levels.

SVM is promising as it is known to be highly robust in churn modeling and seen to have high precision (Xia & Jin, 2008). Other robust models for predicting customer churn exist such as random forest, bagging and boosting (Hu et al., 2022). However, these models are out of the scope of the research because they are less researched within the intersection of customer churn and DP. SVM is a supervised model used for both regression and classification (Rodan et al., 2014). The model proposes an optimal classification decision function when accurately separated by the hyperplane (see Figure 4a). The goal of SVM is to find this hyperplane in the N-dimensional space that can classify the points. The advantages of using SVM is that it has high generalization (less risk of overfitting), scales well to high dimensional data, and solves non-linearity and local minimization problems which traditional customer churn prediction methods do not consider (Xia & Jin, 2008).

Although, an issue raised with applying these machine learning models is that it requires the use of customer information that can be highly sensitive. The more detailed the information of the customer is, the more accurate these models predict on an individual customer level. This approach considers that everyone is unique and tries to focus on interventions on a personal level. This is advantageous for organizations as they would based their business decisions on higher certainty and less risk. However simultaneously, as customers are increasingly aware of how their data is being used for targeting, they have become more concerned over their privacy. Research has shown a customers' perceived privacy risk, as in an individual's evaluation of how their privacy is being violated or misused will affect how the customer trusts the organization handling their data (Torrao & Teixeira, 2023). Trust is important for the Telco industry as it affects how customers interact with the organization (by either continuing to use its services or not). This means understanding how to maintain privacy while using customer data for predictions is high priority for organizations.

2.3. Threat Analysis

Before understanding privacy-preserving techniques, it is important to know what type of threats the techniques are trying to protect against. This will aid in understanding the reasoning behind why and how techniques protect privacy. Privacy is under threat when the adversary finds ways to extract sensitive information from the data used in the machine learning models. Different methods of attacks have been researched including minimality attacks, attribute disclosure attacks, and reconstruction attacks to name a few.

A minimality attack is when the auxiliary uses background information to discover the techniques of anonymization the data owner used (Zhu et al., 2017). For example, if the data is published that an average Thai woman is 165 CM tall, and the adversary holds information that Jane Doe is 3 CM shorter than the average Thai woman, privacy is breached (Dwork, 2006). By linking information from other sources, anonymous data can be used to deanonymize people (Rathi, 2021). An attribute disclosure attack happens when an individual's information from the training data can be inferred from the output of the machine learning model (Wang & Qu, 2011). In Telco, sensitive information could be attacked such as location (e.g., zip code), billing information, or call logs and texts. Reconstruction attacks happen when the attacker tries to recreate the training samples to recover partial information or the full dataset of sensitive features (Rigaki & Garcia, 2021). These attacks are mainly directed at an individual customer level and thus, a privacy concern for companies. Machine learning models are vulnerable to, but not limited to, the attacks listed. Therefore, researchers have investigated ways of preserving privacy so that these attacks and many others are more difficult for the adversary to achieve (Dwork 2006; Machanavajjhala et al., 2007 Sweeney 2002; Zhu et al., 2017).

2.4. Privacy-Preserving Data Analysis & Machine Learning

To try and prevent these attacks, methods of privacy-preserving in data analysis are prevalent in literature. This section will cover different methods and the type of privacy attacks it defends against as well as what it might not. Firstly, the most popular privacy-preserving technique is k-anonymity (Zhu et al, 2017). K-anonymity requires that individuals are not identifiable from a group size smaller than *k* (Sweeney 2002). l-diversity is an extension of *k*-anonymity*,* and it requires that each group of similar datasets have enough variation and well-represented values within the sensitives attribute (Machanavajjhala et al., 2007). Future researchers have investigated extending *k*-anonymity and *l*-diversity such as *t*-closeness to help strengthen the prevention of attribute disclosure attacks (Zhu et al., 2017). Although the methods are strong in preventing attribute disclosure, there are many other attacks they are still prone to. Such as how *k*-anonymity is still prone to complementary release attack: when data is masked but subsequently released, information can be linked together (Sweeney, 2002). Researchers have investigated a multitude of other privacy-preserving methods to try to tackle these issues including synthetic data generation, data mining techniques, and privacy-preserving data publishing (Zhu et al., 2017).

2.5. Differential Privacy

A promising and prominent method of privacy-preserving data analysis in place today fall under DP research. Though, this research paper focuses on specifically DP in machine learning under ERM, it is important to understand the overarching umbrella in which the topic is from to fully grasp the concept. DP was formerly conceptualized as a definition to ensure that removing or adding an item from a database does not largely affect the analytical outcome (Dwork, 2006). The general definition of differential privacy is that when there are two databases (*D* and *D')* that only differ by one single datapoint, the output distribution of a randomized mechanism *M* applied to *D* will result in a similar output when applied to *D'*. Thus preserves the privacy of the individual to the extent that the (lack of) presence of one individual data point should not affect the outcome. To formulate DP, an algorithmic randomized mechanism is ε-differentially private if for all possible outputs *S* that we want private in range *M* we have the equation:

$$\Pr(\mathrm{M}[D] \in S\ ) \leq \exp(\varepsilon) \Pr(\mathrm{M}[D'] \in S\ ) \tag{1}$$

The equation demonstrates that the probability of M(*D)* within *S* is at most exp(ε) times the probability of M(*D')* within *S.* The privacy budget ε controls the amount of noise inserted which, will control how similar or different the databases are (Zhu et al., 2017). When ε is smaller, there is a stronger privacy guarantee where at 0, *D* and *D'* are indistinguishable. The larger ε gets, the weaker the guarantee of privacy. Providing a smaller ε strengthens the privacy however, there is the trade-off where it results in lower accuracy. This trade-off is an advantage of using DP as it focuses on protecting individual privacy, and sharing sensitive data, while still trying to balance utility. This is important as it is possible for the auxiliary to extract individual records that were used to train a model, also known as a membership inference attack (Hu et al., 2022). Using DP is effective in fighting against machine learning-specific attacks as well. Research shows models are prone to memorizing information from the training data, which can be used to infer private information (Hu et al., 2022). Therefore, protecting individual privacy in machine learning models should be a priority. DP can aid in concealing personal sensitive information as the noise added to the machine learning model distorts the data making it more difficult to get a clear result for the adversary.

While training the model, it is up to the researcher to choose a high or low privacy budget in which balances the opposing goals of privacy and model performance. It is expected that when applying different models to real-world data, the requirements of privacy and utility will vary, ultimately affecting the methodological choice. That is why more specialized DP research has been done in the field. For examples, researchers have tried finding the optimal DP model in a high dimensional setting, or finding how utility of minority groups is largely impacted by DP when using imbalanced data (Farrand et al., 2020; Liu et al., 2021). For this reason, it is important for the marketing and Telco to have research that applies to their industry, such as the focus of this research.

2.6. Privacy vs Utility Trade-off in Differential Privacy

When it comes to preserving privacy techniques, there is a focus on balancing privacy with the accurate performance of models for utility. For example, the trade-off can mean choosing between better customer service via customization or ensuring privacy protection to customers concerned about proper information storage (Rust & Chung, 2006). Such a trade-off exists as neither one nor the other is more preferred by customers (Rust & Chung, 2006). Customers demonstrate their desire for privacy protection, yet they still engage in non-private behaviors or attitudes due to many reasons based on social theory, psychology, behavioral economics and more (Kokolakis, 2017). In fact, people are more willing to give a certain amount of privacy up than they indicated, also referred to as the privacy paradox (Norberg et al., 2007). A survey that was done by Cisco in 2019 on consumer privacy demonstrates consumers understand some extent of personal information must be shared to receive benefits of products, services, and business relationships. Yet simultaneously, consumers are worried about how their personal private information is being handled (Cisco Cybersecurity, 2019).

2.7. Empirical Risk Minimization and Perturbation techniques

Generally, DP ensures that appropriately chosen random noise is added to the algorithm f(D) to achieve ε-DP (Dwork, 2006). In this specific use case, random noise addition methods known as output perturbation and objective perturbation are applied under the framework of ERM. ERM is a classic optimization process to try and minimize the prediction error of the training model so that it can accurately predict unseen data. Such a process is advantageous in the context of DP as ERM ensures high model accuracy whilst implementing the privacy budget (Chaudhuri et al., 2011).

Output perturbation was originally introduced by Dwork (2006) and has flourished in the research field for different use cases. Demonstrated in Figure 1, output perturbation adds noise directly to the output of an algorithm. To implement output perturbation, the models must satisfy certain requirements to meet privacy guarantees. Foremost, implementing ε-DP requires that the noise mechanism M is differentiable and l-strongly convex (Chaudhuri et al., 2011). Strong convexity requires that the absolute value of the derivative of any given point must not exceed the point itself (for proof see Chaudhuri et al., 2011). Objective perturbation is another technique which adds noise to the objective loss function, which, means directly to the learner (Chaudhuri et al., 2011) (see Figure 1). The requirements of objective perturbation are stricter because 1) the loss function and the objective gradient must be differentiable with continuous derivatives and 2) M is l-strongly convex and doubly differentiable (Chaudhuri et al., 2011). Although some researchers have defied these strict requirements[1], the methodology of this research limits itself to the constraints as introduced by the original researchers. This is because the requirements have strong theoretical foundations to guarantee privacy. Complying to more established practices will increase the credibility and reliability of the research.

[1] Kifer et al., (2012) extended the model so that less noise is necessary and applying their models to problems with hard constraints and non-differentiable regularizers.

Although it is an understanding that the privacy budget ε controls the trade-off, the choice of privacy mechanism can also have an impact on the results. Chaudhuri et al (2011) found in their case, objective perturbation was more effective. By considering both output and objective perturbation, it allows thorough deep dive into the way the mechanisms behave in the context of the churn models. Additionally, the model of choice (e.g., SVM and RLR) can also have an impact on the trade-off and how DP performs. This in part, is influenced by the nature of the data in use. Therefore, it is important to combine different models and methods to see which combination leads to the best optimal model for this specific Telco data. Researchers have focused on how the data characteristics impact the effectiveness as well such that in high dimensional cases or imbalanced cases, DP implemented with SVM can effectively balance the trade-off (Rubenstein et al, 2009). Therefore, applying different models and methods per industrial use case is fruitful as the optimal model may differ depending on the scenario.

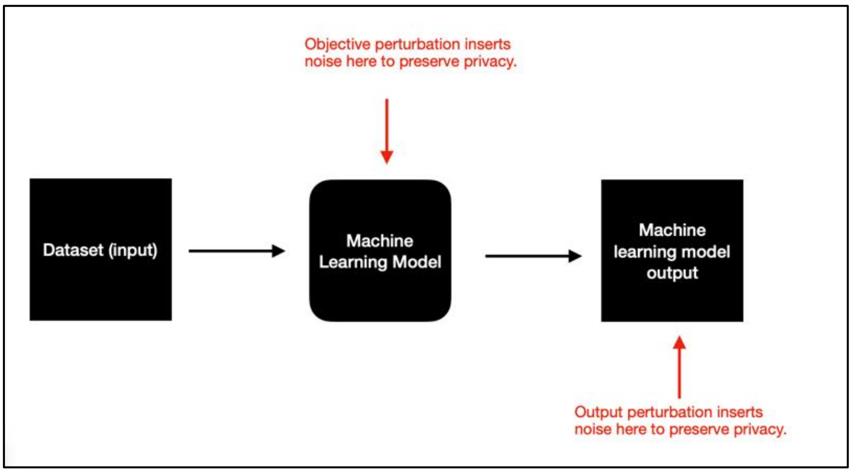
2.8. Trade-off Measurements

The privacy budget ε is commonly used as a measurement of privacy that quantifies the level of privacy protection. However, researchers vary in how they measure model performance for utility. For example, Rubinstein et al. (2009), focused specifically on SVM and used techniques from algorithmic stability to measure utility. Chaudhuri et al. (2011) focused on how the regularization parameter impacted the level of privacy whilst evaluating utility. Song et al. (2013) and Yu et al (2020) measured the utility based on the convergence rate as well as generalization. As demonstrated, the different models lead to different ways of measuring the utility of the DP models. The variety of metrics makes it difficult to compare models from the different literature in the field. Henceforth, implementing common churn metrics for this research is most sensible so that is comparable to industry standards, allowing for meaningful insights and evaluations.

This research focuses on utility analysis based on common industry practices with predictive modeling. Churn prediction in Telco most commonly uses accuracy as a measurement which is calculated by the ratio of total right predictions to the total number of samples (Jain et al., 2020). Oftentimes, accuracy is not enough of a good measure since with small datasets, the value is often more predictable and the same (Jain et al., 2020). Therefore, other classification metrics are often considered. Hu et al (2012) studied differential privacy in Telco big data platforms and used the area under the curve (AUC) to measure the accuracy of their model. Similarly, this research paper will consider the area under the ROC-curve (ROC-

AUC interchangeably referred to as AUC), accuracy, specificity, and sensitivity as utility measurements. Reasons as to why these metrics are specifically chosen is covered in section 3.3.5 Analysis.

Figure 1 - *Simple machine learning algorithm*


Objective perturbation inserts noise here to preserve privacy.
Dataset (input)
Machine Learning Model
Machine learning model output
Output perturbation inserts noise here to preserve privacy.

3. Data & Methodology

3.1. Data Description

Data was retrieved from Kaggle, "an online community platform for data scientists and machine learning enthusiasts" (Uslu, 2022). Specifically, the dataset "Telco customer churn" by BlastChar (2018) on Kaggle will be used as it has the highest number of upvotes at 2282 (upvote reflects the positive sentiment of the community towards the dataset). Additionally, the usability is rated relatively high on the scale at 8.82 out of 10. The data contains information about a fictional Telecommunications company in a contractual setting and is used to represent a standard telco customer churn dataset. The company provides services for internet and the home phone to 7043 customers in California.

3.2. Data Preprocessing

Several data preprocessing procedures were done to ensure reliable data and a smooth modeling process. There were 11 “NA” values that were removed from the dataset. There were no observations that were duplicated. Certain variables were formatted to be compatible for the machine learning algorithm. For example, categorical variables were encoded into dummy variables. “PaymentMethod” was split into automatic methods (“bank transfer” and “credit card” = 1), and checks (“mailed check” and “electronic check” = 0). “Contract” was categorized as “month-to-month” (=1) and those longer than one year was coded to 0. Character variables that are binary in nature were recoded to 1 and 0. The remaining categorical variables (e.g., “MultipleLines”, “InternetService”, “OnlineSecurity”, “OnlineBackup”, “DeviceProtection”, “TechSupport”, “StreamingTV”, “StreamingMovies”) were factored (see Appendix A for the result of data*)*. Table 1 is a descriptive data table that was created to demonstrate the continuous and binary variables.

The categorical variables were also described. In total there are 5163 non-churners (0 = 73.4%) and 1869 churners (1 = 26.6%). Approximately 44% of the customers preferred automatic payment methods (bank transfer or creditcard) while the remaining preferred electronic or mailed check. Additionally, 55% of the customers had a month-to-month contract while 20% had a one-year contract while the other 20% had a two-year contract.

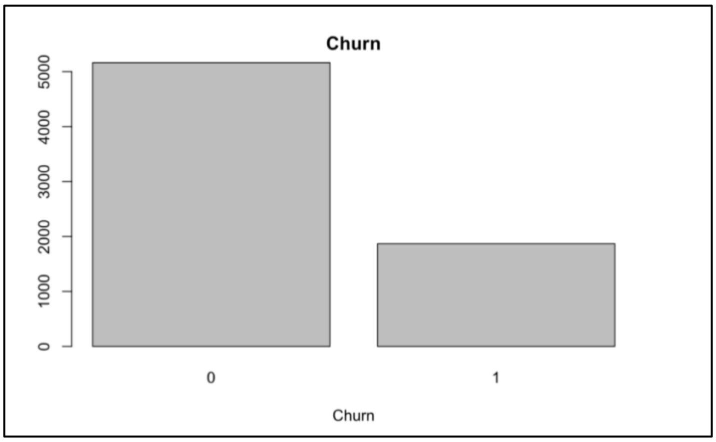
Lastly, 44% of the customers have fiber-optic internet service, 30% use DSL, while the remaining do not have internet service.

Table 1 - *Continuous and dummy variables*

Variable	N	Mean	St.dev	Min	Max
gender	7032	0.50	0.50	0	1
SeniorCitizen	7032	0.16	0.37	0	1
Partner	7032	0.48	0.50	0	1
Dependents	7032	0.30	0.46	0	1
tenure	7032	32.4	24.55	1	72
PhoneService	7032	0.90	0.30	0	1
PaperlessBilling	7032	0.59	0.49	0	1
MonthlyCharges	7032	64.80	30.09	18.25	118.75
TotalCharges	7032	2283.30	2266.77	18.80	8684.80
Churn	7032	0.27	0.44	0	1

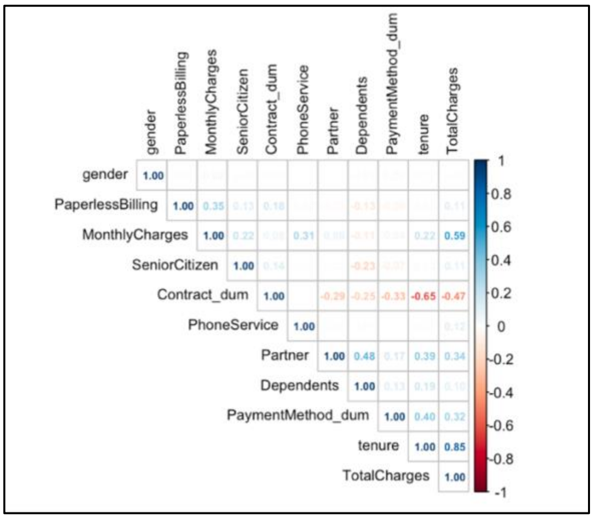
To understand the relationship between the variables, a Pearson correlation plot (see Figure 3) was created. Highly correlated can be problematic when it creates issues like multicollinearity. Multicollinearity creates high variance and makes it more difficult to interpret each individual independent variable effect on the output variable, undermining its significance. Additionally, considering highly correlated variables can help remove redundancy in the model, simplifying it and making it more interpretable.

Figure 2 - *Churners vs non-churners in the dataset*



A threshold of above 0.5 or below – 0.5 was used as values that had a high correlation. “TotalCharges” had a 0.82 correlation with “tenure” and 0.65 with “MonthlyCharges”. Additionally, “Contract_dum” had a -0.65 correlation with “tenure”. Therefore, the variables were categorized as highly correlated and considered when training the model. Whilst training the model, these variables were removed to see if it would improve the model performance as well.

Figure 3 - *Correlation Plot*



3.3. Methdology

The research approach follows a standard approach to predicting customer churn within supervised machine learning. The baseline model used will be the non-private RLR and non-private SVM as they are two statistical models widely used in popular research papers of both the fields of DP and churn (Chaudhuri et al., 2011; Chaudhuri & Monteleoni, 2008; Ha et al., 2005). Hence, using popular models common in both fields will encourage synergy between the research of DP and customer churn analytics. Commonality strengthens the proof of concept in bringing the topics together.

3.3.1. Empirical Risk Minimization

Given the Telco dataset with D $\{x_i, y_i\}$ for all $i \in \{ 1, 2, \ldots, n\}$ from the universe *S* and a closed convex set C, ERM is used to minimize the empirical loss function *L* $(f,$ *D)* over $f \in$ C . The loss function is an important function in the field of machine learning so that the model output is as close as possible to the actual churn values of the Telco customer data. In supervised machine learning under ERM, the problem can be generalized to the following formula:

$$L(f, D) = \frac{1}{n}\sum_{i=1}^{N} l(f(x_i), y_i) + \lambda R(f) \tag{2}$$

To prevent overfitting, the regularization term is added to the objective function. Where λ is the regularization parameter and *R* is the regularizer which helps prevent overfitting by penalizing, imposing a cost on the optimization function. The regularizer penalizes the classifier for predicting each training data point (Chauduri et al, 2011). Tuning λ to effectively optimize the model was done with a grid search with a range of values for λ. The regularizer in this case will be denoted as the l_2 norm (ridge regression) denoted as $\| (.) \|_2$ (Chaudhuri et al., 2011). Its strong convexity is what leads to meeting the generalization and privacy requirements necessary for the model (Chaudhuri et al., 2011).

3.3.2. Privacy Preserving via Differential Privacy

Output perturbation is applied based on previous works such as Dwork (2006) and Chaudhuri (2011). As demonstrated in Figure 1, noise is implemented in the output of the machine learning model. To balance privacy and model performance, the magnitude of the noise is added proportionallly to the sensitivity. The sensitivity quantifies an upper bound that specifies the maximum amount of noise that can be introduced to the algorithm. The sensitivity is the maximum distance between all possible pairs within a defined space. The sensitivity is used to balance the model so that the noise does not significantly compromise the model performance. For the vector-valued function, the l_2 sensitivity is used:

$$S(Lpriv) = \Delta f = \max_{D,D'} \parallel f(D) - f(D') \parallel_2 \qquad (3)$$

The l_2 sensitivity employs the Euclidean distance ($\sqrt{(x_2 + x_1)^2 + (y_2 - y_1)^2}$), as the maximum distance to derive an upper bound for the output perturbation to balance the privacy and utility of the model. Additionally, an assumption that will be used for modeling is that for $x_1, x_2, x_3, \ldots . x_n$ of the database, $\parallel x_i \parallel_2 < 1$. These restricting bounds allow for tighter privacy guarantees, by controlling the sensitivity and essentially bounding the magnitude of the input data to a limited number of possible values that the individual data points can take. This helps control the random noise that is used so that the model also minimizes its negative impact on the utility of the output as well.

To sum up, when *D* and *D'* differs by at most one element and where *b* is the noise vector bounded by the sensitivity, the probability density from Chaudhuri & Monteleoni (2008) used as noise for output perturbation is as follows:

$$p(b) \propto e^{-\frac{n\mathcal{E}\lambda}{2}\parallel b \parallel_2} \qquad (4)$$

Thus, the algorithm and its goal to minimize function the loss function and the noise vector *b* to balance the privacy and model performance at different values of the privacy budget.

$$Lprivout(f, D) = L(f, D) + b \qquad (5)$$

Objective perturbation does not depend on perturbing the data such as output perturbation and thus, does not rely on the sensitivity. Instead, objective perturbation directly adds noise to the objective function where k is noise from the density function:

$$p(k) \propto e^{-\frac{\mathcal{E}}{2}\| k \|_2} \tag{6}$$

The loss function incorporates the regularization term in which $\| b \|_2^2$ quantifies the magnitude of the noise that is added to objective function. An added term $< k, f >$ represents the inner product of the noise term and the output of the algorithm:

$$Lprivobj(f, D) = \frac{1}{n}\sum_{i=1}^{R} l(f(x_i), y_i) + \frac{\lambda}{2} \| b \|_2^2 + < k, f > \tag{7}$$

Minimizing $Lprivobj(f, D)$ will optimize the learner so that it balances the highest utility (model performance) with varying degress of noise input. The assumption that for $x_1, x_2, x_3, \ldots. x_n, \| x_i \|_2 < 1$ will also be used for objective perturbation for the same reason.

3.3.3. Logistic Regression

Logistic regression estimates the probability that an instance will belong to a particular class. The binary classifier will identify a probably above 50% as positive ('1' = churn), and anything below as negative ('0' = no churn) given that the features are plugged into the logistic function:

$$\hat{y} = \sigma(x) = \frac{e^{-\beta 0 + \beta 1 x}}{1 + e^{-\beta 0 + \beta 1 x}} \tag{8}$$

When implementing the algorithm for the logistic regression, the cross-entropy loss function is used for the optimization objective to minimize the error of the model. The equation quantifies the error that exists between the predicted probabilities $\hat{y}$ and the actual labels from the dataset y:

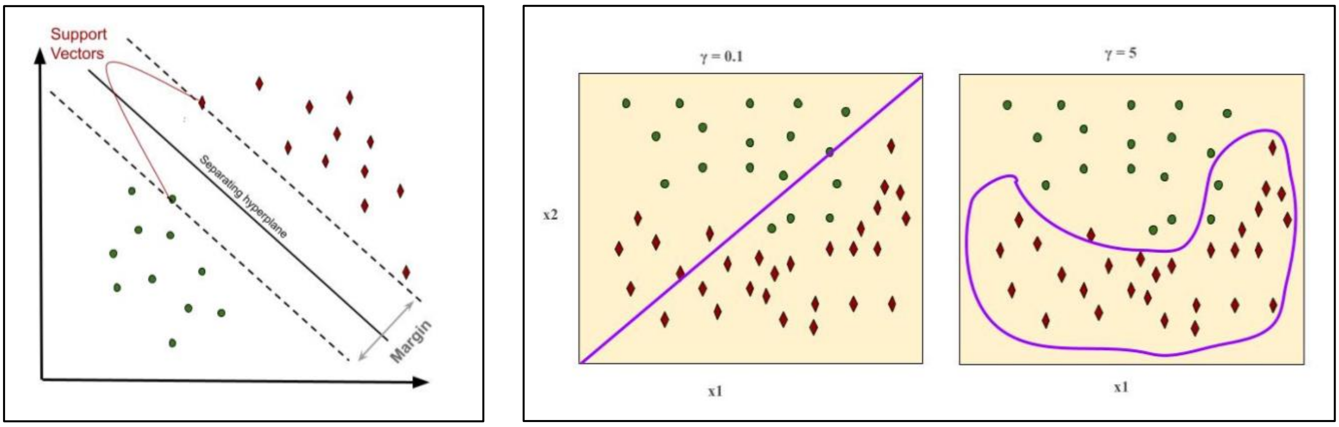
$$l(\hat{y}, y) = l_{logreg}(f(x_i), y_i) = -y * \log(\hat{y}) + (1 - y) * \log(1 - \hat{y}) \tag{9}$$

The equation satifies the requirements that the loss function is continuous, differentiable, and doubly differentiable and so will be used alongside output and objective perturbation. Since there is a regularization term in RLR, the model introduces the l_2 regularization term to meet the privacy guarantee. The term is controlled by the parameter lambda (λ) which will balance the performance and generalizability of the model. A higher term increases the penalty which helps constrain the model complexity. However, too high may lead to overfitting, making it difficult to generalize to unseen data. Therefore, hyperparameter tuning will help optimize the value.

3.3.4. Support Vector Machine

SVM can be applied as a linear classification or non-linear classification model. Both will be investigated to find the optimized performance. As a linear model, the SVM will create a hyperplane (i.e., the decision boundary) that splits the data into the available classes (churners and non-churners) (see Figure 4a). The aim of SVM is to fit the widest possible parallel lines around the linear discriminant, also known as maximizing the margin. Support vectors refer to the points closest to the hyperplane and thus, have an influence on the position and orientation of the line. SVM introduces the parameter often referred to as *C* in SVM literature which represents the soft margin when classifying the training data. Smaller *C* is more open to outliers and allows for larger misclassification whereas a larger *C* leads to a decision boundary that more fits the training data, potentially leading to overfitting. Hyperparameter tuning will be used to optimize the parameter. However, the parameter for the remainder of the research will be referred to as lambda (λ).

Figure 4a -4b - *Left (4a): Visual Representation of SVM; Right (4b): Demonstrating how as γ grows, the more bell shaped the decision boundary becomes*



A non-linear kernel is considered as well since the dataset may be complex and not linearly separable. SVM uses a “kernel method” which defines a function that directly computes the dot product between the data points in the original input space to transform to an output space in its required form. This makes SVM more computationally efficient for more complex results. The linear kernel will be used for the linear classification:

$$K(x, y) = \mathrm{x} \cdot \mathrm{y} \tag{10}$$

Gaussian kernel (Radial Basis Function) will be used to handle the data “non-linearly” and the equation is as follows:

$$K(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right) \tag{11}$$

The radial kernel is often preferred because it can handle complex data (Keerthi & Lin, 2003). Gaussian (radial) kernel captures more complex non-linear relationships between the predictors. The radial kernel introduces a new parameter often referred to as gamma (γ) in literature. γ controls the bell-shaped curve of the decision boundary (see Figure 4b) . As γ increases, the more irregular and non-linear the decision boundary becomes. γ acts just like a regularization parameter so as the value becomes higher, the more likely it is to overfit. Hyperparameter tuning will help choose the optimal γ. For SVM, the hinge loss function is typically used as the error penalty, and it is equal to:

$$l_{SVM}(z) = \max(\,0\,, 1 - z) \tag{12}$$

and where $R(f) = \frac{1}{2} \| f \|^2$. The typical hinge loss function for the SVM satisfies the requirements for output perturbation in that it is differentiable, yet unfortunately, it is not doubly differentiable, a requirement of objective perturbation. Therefore, as demonstrated and suggested in Chaudhuri et al., (2011) and Chapelle (2007), the Huber loss function is implemented as a smooth approximation to the hinge loss function. The Huber loss function is a differentiable approximation to the linear penalization in which can guarantee privacy when used for the DP models. *Z* represents the loss between the predict and true values, *h* determines at which point the loss switches from a quadratic function to not:

$$l_{Huber}(z) = \begin{cases} 0 & if\ z > 1 + h \\ \frac{1}{4h}(1 + h - z)^2 & if\ |1 - z| \leq h \\ 1 - z & if\ z < 1 - h. \end{cases}$$

(13)

3.3.5. Model Training

Several points were considered for model training:

- The data was split so that 80% of the observations were in the training set and 20% of the observations were in the test set. The training data was imbalanced since 1033 observations were non-churners and 374 observations were churners, which, can lead to biased and inaccurate results. Using the ROSE method generates synthetic-balanced samples that can strengthen the estimation of the binary classifier (Menardi and Torelli, 2014). The ROSE method is commonly used for binary classification problems in the presence of imbalanced classes (Lunardon et al., 2014). Balanced samples were created by undersampling the majority group (no churn) and oversampling the minority group (churn). The choice of using both was to ensure a large enough of observations while not relying on a large amount of synthetic data that may be unreliable. Accordingly, there was a total of 2864 observations of non-churners and 2761 observations of churners in the training set. Both the imbalanced dataset and balanced dataset will be used to see the improved accuracy.
- In total, 8 different algorithms are trained: three non-private models including RLR and SVM (radial kernel), SVM (linear kernel), and 6 private models where output and objective perturbation were applied to SVM (radial kernel), SVM (linear kernel) and RLR via the DPpack on R (Giddens & Liu, 2023). The non-private models were used as the comparative baseline model for the algorithms.
- Ridge-regression incorporates the l_2 norm to act as the regularization term for all models since a privacy guarantee requires the l_2 norm. Desirably, l_2 also provides added benefits to the model such as preventing multicollinearity and reducing the impact of highly correlated variables. This helps balance the overall model, having the advantage of preventing overfitting.
- *K*-fold cross-validation is a resampling procedure used in machine learning to evaluate, compare, and select the algorithm for a given situation. The parameter *K* is the number of splits in the training set that will be used. This research uses a 5-fold cross-validation where each fold is used as a validation set while the rest is used for training. The process is repeated 5 times as each fold is used as the validation set once. Not only does cross-

validation prevent overfitting and bias, but it also gives a better understanding on how the model will generalize to unseen data. In combination with hyperparameter tuning, cross-validation will choose the parameter values that give the optimal performance.

- Choosing the optimal regularization function (λ) with hyperparameter tuning is done with varying grid searches depending on the model. Hyperparameter tuning allows model training by testing out a variety of predefined regularization terms. A grid of a range of hyperparameter values was run for each $\boldsymbol{\varepsilon}$ to find the regularization term (λ) that would result in the highest performance. When there are multiple parameters such as with SVM, the grid will run through all possible combinations to find the desired output. The grid search was adapted depending on the model and thus, further explained in the results section.
- When running preliminary models, it was discovered that the performance at different $\boldsymbol{\varepsilon}$ above 0.5 were very similar. The implication is that for higher values of the privacy budget, the model performance was not noticeably affected. For a more diverse and comprehensive research, $\boldsymbol{\varepsilon}$ between 0.001 and 0.5 was chosen. Focusing on smaller values also meant the research could focus on stronger privacy protection.
- Additional terms in SVM include the huber (h) (Equation 12) which is a positive real number that gives an indication to the degree of which the Huber loss approximates the hinge loss (Chapelle, 2007; Gidden & Liu, 2023).
- With radial kernel SVM model, an added parameter is D, a parameter approximating the kernel as it identifies the number of dimensions of the transform space. Higher values of D provide better approximations. The choice of values will be further explained in the results section.

3.3.5. Analysis

The test set is used to assess the performance of the models with the chosen metrics. This is because the test consists of unseen data which provides a more realistic evaluation of the models' generalizability and their ability to perform well on unseen instances. Privacy is measured based on ε, in which the higher the value the less noise is added, and the less private it is. The classification accuracy is measured by the number of correct predictions over the total number of predictions. Additionally, the ROC-AUC is used as a metric for binary classification. The ROC (Receiver Operator Characteristic Curve) plots the test sets true positive rate (TPR) as the y-axis against the false positive rate (FPR) as the x-axis. The values plotted represent the value at different classification thresholds to create a ROC curve. An ideal curve represents a high TPR and a low FPR. The AUC is used to summarize the classification model performance from the ROC, at different classification thresholds, in one value. It demonstrates the degree to which the classifier can distinguish between the classes. The AUC is a percentage that ranges between 0 – 1 where a value above 0.5 demonstrates that the model does better than if a random classification was done. Lastly, the sensitivity and specificity were included as metrics for further analysis of how well the model classified each class. Sensitivity focuses solely on the TPR whereas specificity focuses on the true negative rate (TNR). Thereby, they provide valuable insights into how well the model classifies each class specifically. Note that the sensitivity metric is different to the sensitivity used output perturbation. Over the course of the research, the AUC was preferred over any other metric because it demonstrates a more holistic model performance, calibrating the trade-off between the true positive rate and the false positive rate as different thresholds. Although accuracy is commonly used for churn prediction, it is often not enough to capture the model performance (Jain et al., 2020).

4. Results

This section will discuss the process behind training and testing the model in which lead to the optimal model performance. This section will do a comparative analysis on the model performance to determine the optimal model. With all models, the numerical variables were scaled, and the categorical variables were turned into dummy variables. Scaling the numerical variables to have a mean of 0 and a standard deviation of 1 ensures all variables are on a similar scale so that the magnitudes are comparable. Feature scaling is highly recommended for RLR because, with regularization, coefficients with large and small variables are penalized. Variables that have inherently large or small values will be penalized more, creating bias in the model. Similarly, with SVM, the hyperplane will be heavily influenced by variables with large values, leading to less-than-optimal results. Further scaling was done for private models so that the assumption $\| x_i \|_2 < 1$ to hold. Therefore, when practically applying the training model, stating an upper and lower bound for each x_i was essential to keep the values in the meaningful range before scaling the values. The minimum and maximum value of each independent variable was set as the lower and upper bound respectively. Keeping the bounds as tight as possible avoids adding unnecessarily large amounts of noise that will hinder the model's performance (Giddens & Liu, 2023). Additionally, 5-fold cross-validation and hyperparameter tuning was used to choose the best combinations of parameters that lead to the highest model performance.

4.1. Baseline Models

Non-private Regularized Logistic Regression. As a result of the 5-fold cross-validation and hyperparameter tuning, the model with the highest AUC was when $(\lambda) = 10^{-5}$. The penalty term was used to predict the test set which led to an AUC of 84.19%, the accuracy of 80.67%, the sensitivity of 89.93%, and specificity of 55.08% (see Table 2). When considering the highly correlated variables, removing them did not improve the model and the variables were kept in for training and testing. Modifying the current decision boundary (observations predicted above 50% are considered more likely to churn and those below are considered unlikely to churn) can improve certain metric values but can also simultaneously decrease the performance of other metrics. Therefore, the threshold at 50% was kept as the choice of improving a metric in this scenario is entirely up to the use case of the churn model. More about the practical business implications will be covered in Section 5.3. Business Contributions and Recommendations. This model will be used as the baseline to compare with the performance of the private models. With the private models, the goal is to try and perform just as well as the non-private model, or as close as possible.

Non-private Support Vector Machine. Both linear and radial kernel SVM were considered for the baseline SVM model as it is still uncertain which one will perform better. Similarly, to the RLR model, SVM requires hyperparameter tuning to find the optimal penalty term for the model. When performing the radial SVM model the grid search adds the extra parameter γ (refer to Section 3.3.4. Support Vector Machine). Higher values of γ may lead to overfitting; however, it is better at capturing complex and non-linear relationships. Therefore, hyperparameter tuning will capture the right balance alongside using 5-fold cross-validation which will help generalize the findings better. Again, the AUC was preferred over any other metric in determining the best values for γ and the regularization term. The SVM linear model results in an AUC of 83.32%, the accuracy of 66.38%, the sensitivity of 88.5% and a specificity of 58.37% (see Table 2). The radial kernel SVM model results in an AUC of 53.11%, accuracy of 77.35%, sensitivity of 9.89%, and specificity of 96.32%. Here, the highly correlated variables were also considered and did not seem to improve the model. Additionally, these models were based on the balanced data as they performed better than the imbalanced data. As a result, the private models will also be using balanced data as the same outcome is expected and also to maintain comparability. This model will be used as the baseline to compare with the performance of the private models. This is also with the goal of trying and creating private models that perform close to or even better, just as high of performance.

Between the SVM models, the linear kernel SVM model is preferred as it has a higher AUC. The model also has a high sensitivity although performs worse in accuracy and specificity. It seems that the radial kernel SVM model has extremely high specificity but extremely low sensitivity. Adjusting the probability threshold can help achieve a better balance if necessary for specific use cases. Overall comparing the three models, RLR outperforms with the highest performance in AUC, accuracy, and sensitivity.

Table 2 - *Non-private Optimal Model Results*

Model	ROC-AUC	Accuracy	Sensitivity	Specificity
RLR	84.19%	80.67%	89.93%	55.08%
SVM Linear	83.32%	66.38%	88.5%	58.37%
SVM Radial	53.11%	73.35%	9.89%	96.32%

4.2. Private Models

To assess the private models alongside the non-private models, a stepwise explanation will be used. Initially, similar family models will be compared side-by-side (i.e., RLR objective perturbation, RLR output perturbation and non-private RLR). The smaller-scale analysis allows the research to focus on the strengths and weaknesses of each method and technique, and how the different combinations affect the trade-off. Then, the analysis will cross-compare models such as analyzing the different RLR models with the different SVM models. The broader comparison aims to choose the best model overall on the given dataset. The best model is one that can keep the error lowest at different values of ε. These comparisons will be used to understand how well the different types of model features and DP methods balance the privacy budget ε and the performance metrics to represent the trade-off between privacy and performance. Like the baseline model, the AUC will be the prioritized metric as it captures the holistic performance of the model.

4.2.1. RLR Output Perturbation

The mean AUC score was calculated from the hyperparameter combination chosen from the training set and then used on test set to produce the performance metrics with the unseen data. Table 3-6 demonstrate the optimal model performance for all metrics for each value of ***ε***. It seems that the model demonstrates an extent of the privacy and utility trade-off because as ***ε*** increases, privacy decreases and the better the model performs. Particularly there is a sharp increase in performance from when ***ε*** is at 0.001 to 0.01 (from 54.05% AUC to 78.64% AUC) and another sharp increase from when ***ε*** is at 0.01 to 0.05 (from 78.64% AUC to 82.01% AUC). However, from there, ***ε*** does not seem to affect the AUC performance largely as the value maintains relatively stable around between. 82 – 83%.

There are anomalies in the data where the performance goes up and down due to the random nature of the noise; the trade-off is not a direct continuous linear relationship. For example, the AUC when privacy is higher (AUC is 78.64% when ***ε*** = 0.01) can be higher than when privacy is lower (AUC is 76.9% when ***ε*** is 0.03). The line of best fit in Figure 5a still demonstrates an incline in performance as ***ε*** increases (privacy decreases). Using the AUC as an indicator, models in which ***ε*** is above 0.05 performs almost as well as the non-private RLR-model, with approximately only 2% lower AUC.

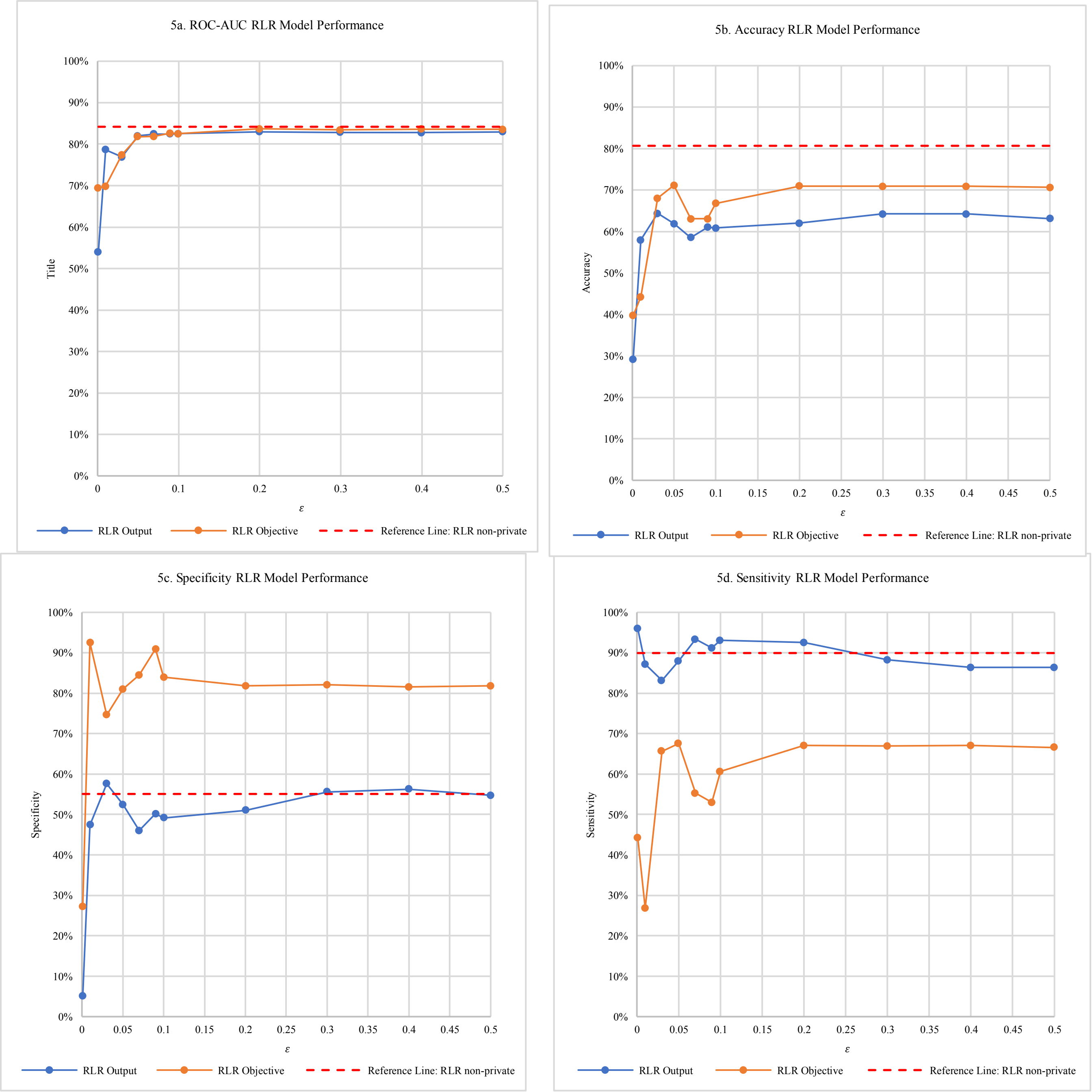
The model has high sensitivity and at certain values of ***ε***, performs even higher than the non-private model. This demonstrates that the added noise shifts the classification to better classify the positive instances. Just like the non-private RLR-model, RLR output perturbation demonstrates low specificity. It seems the model closely follows the baseline model, demonstrating that introducing privacy does not compromise the model's performance. All these findings can be accounted to how the random noise changes the way the model behaves and how it allocates the true positives and true negatives.

4.2.2. RLR Objective Perturbation

It was more critical to strictly define the criteria for the hyperparameter in objective perturbation since failure to converge was more common if proper combinations of parameters were not chosen carefully. Convergence in machine learning is important as it indicates when the model finds an optimal stable point after the iterative training process. The stable point indicates when the algorithm cannot improve the objective function anymore and achieves the predefined model requirements. However, failure to converge happens when the model was unable to find a stable solution that meets the criteria set in the period it was given. Failure to converge was mainly seen when there were small values of the regularization parameter with small values of ε. This can be linked to the fact that the regularization term implies a weaker penalty which was not strong enough to constrain the model's parameters. More noise is inserted with lower values of ε which can create higher instability. Therefore, the more important it is for a strong regularization term to penalize the model and reduce complexity. Thus, varying grids with varying parameter values were used for objective perturbation to meet the strict constraints imposed. As the smaller the ε got, the stronger penalty values were imposed on the grid search. The model with the highest AUC was selected and evaluated with various metrics on the unseen test set with the results presented in Table 3.

The RLR objective perturbation model demonstrates an extent of the privacy and utility trade-off where as ε increases, the less private and thus, the more accurately the model performs. Particularly there is a clear increase in performance from when ε is at 0.001 to 0.03 (from 69.42% AUC to 77.4% AUC) and another increase from when ε is 0.01 to 0.05 (from 77.4% AUC to 81.79% AUC). However, when ε is between 0.05 and 0.5, the privacy parameter does not seem to affect the model performance as the AUC maintains relatively stable around 81.9 – 83.7%. Object RLR performs similarly to output RLR in terms of AUC but outperforms in terms of accuracy (see Figure 5a – b). Analyzing the specificity and sensitivity can demonstrate the underlying reason. It seems that objective perturbation cause the model to be worse at classifying the individuals who are more likely to churn (lower sensitivity) and even better than the non-private model at classifying the non-churners (higher specificity). It seems that adding noise in the output results in better classification of the individuals who are more likely to churn (higher sensitivity and lower specificity) whereas adding noise in the objective results in better classification of those who are not likely to churn (higher specificity and lower sensitivity) (see Figure 5c - d). The results demonstrate a trade-off between sensitivity and specificity which can be mitigated by choosing a cut-off level for classifying the positive class (Chu, 2002).

Figure 5a – 5d – *Private regularized logistic regression models compared to the baseline non-private model*



5a. ROC-AUC RLR Model Performance

5b. Accuracy RLR Model Performance

5c. Specificity RLR Model Performance

5d. Sensitivity RLR Model Performance

4.2.3. SVM Output and Objective Perturbation

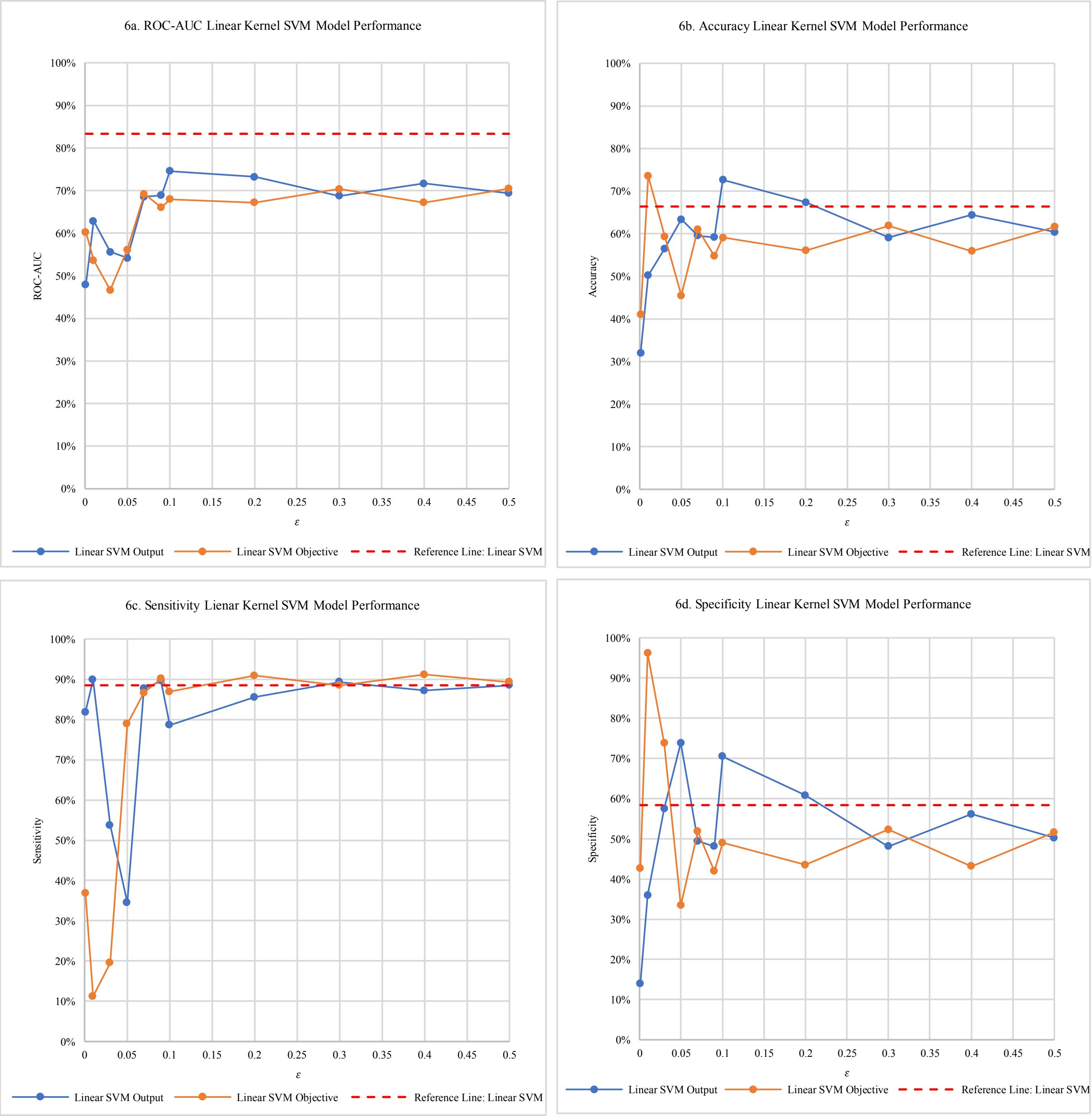
Linear Kernel SVM Models. Firstly, the linear kernel SVM models were trained with output and objective perturbation. The regularization term with the highest AUC was used for the entire training set and used on the unseen test set to measure the performance of the model per ***ε*** presented in Tables 3 -6.

To train the model, SVM has an additional parameter to set called *h.* The term *h* (in the equation) approximates the Huber loss function where the value 0 is equivalent to the hinge loss, and thus, the model behaves like traditional SVM. Increasing *h* increases the robustness of the Huber loss, providing a smoother approximation and possibly losing the discriminative power of the hinge loss. The Huber loss function 0.01 and 0.5 are considered typical values. Since *h* should not largely affect the model performance as long as it is not too large or small, 0.3 was chosen as an approximate middle point (Chapelle, 2007).

Although there are some random variation (expected as the noise is random) both output and objective perturbation for linear kernel SVM have similar patterns of behavior and are therefore analyzed together (see Figure 6a - d). Overall, considering output and objective perturbation of the linear kernel SVM, it is clear that adding noise leads to a decrease in the model's AUC value compared to the baseline model (see Figure 6a). The trade-off demonstrates that as ***ε*** increases, the AUC also increases to an extent. Specifically considering output perturbation, when ***ε*** is at 0.001, the model performs its lowest at 48% and jumps to an AUC of 63% at 0.01. Subsequently, there are some anomalies in which the AUC decreases in performance when ***ε*** is between 0.03 – 0.05 (see Table 3) however, the AUC continues to grow and stabilizes between 69% - 75% when ***ε*** is above 0.07.

In general, the line of best fit (see Figure 8) demonstrates a positive incline in performance as ***ε*** increases for both output and objective perturbation. Overall, linear kernel SVM output perturbation and objective perturbation demonstrate similar performance in terms of AUC. However, due to the difference in random noise perturbation, there are slight random variation. In terms of accuracy and specificity, the model performs slightly below the baseline model for both output and objective perturbation. This is not taking into account some large variation where the specificity for the linear kernel SVM objective perturbation spikes to a performance of 96.13% when ***ε*** is at 0.01, much higher than the baseline model. It seems that compared to the RLR models, linear kernel SVM has larger variation anomalies in the optimal model performance. Meaning as ***ε*** increases, the performance seems to vary largely, making it difficult to draw a more concrete conclusion on the trade-off. In terms of the sensitivity, when ε is below 0.1, it is also difficult to draw a conclusive statement about the trade-off due to the large spikes in performance other than saying that there is a trade-off in how the model classifies the true positive and true negative rates. This is demonstrated by objective perturbation when ε is at 0.01, sensitivity is at its lowest (11.23%), whilst specificity is at its highest (96.13%). Although, once ε is above 0.2, both objective and output perturbation specificity performs close to the baseline model (see Figure 6c).

Figure 6a – 6d - *Private Linear Kernel SVM models compared to the baseline non-private mode*



Radial Kernel SVM Models. Although the non-private radial kernel performed lower in AUC than the non-private linear kernel, the models may behave differently once the noise is considered. Like all models, hyperparameter tuning was done where the radial kernel had an additional term *D*. *D* was defined to approximate the dimensionality of the transformed space. Whilst training, he models would reach the maximum duration for running time due to computational capacity when $D > 10$. Therefore, a value of 10 was chosen as it was the highest value able to properly run without running into errors and issues with long running time (e.g., the model could not converge properly). Just like the other models, the average AUC of the 5-fold cross-validation was used to compare to the models in hyper parameter tuning. Demonstrated in Table 3-6, the model with the highest AUC was used on the unseen test set to derive the performance per $\boldsymbol{\varepsilon}$ value.

It seems that both output perturbation and objective perturbation with radial kernel SVM demonstrates very similar results and so are analyzed together. In terms of the AUC, adding noise for both objective and output perturbation actually resulted in models that perform better than the baseline model (see Figure 7a). Due to the technical limitation and time constraints, it could be that training the non-private radial SVM model only led to suboptimal results. There are a total of 5625 observations in the training sample and thus, a total of 5625 x 5625 training samples are compared for similarity. This makes it computationally inefficient especially on top of hyperparameter tuning and cross-validation.

In terms of balancing the trade-off, it is difficult to draw a concrete conclusion for both perturbation methods. Overall, the trendline for objective perturbation seems static while output perturbation has a slight incline (see Figure 8a). Looking deeper into the models, it seems when $\boldsymbol{\varepsilon}$ is between 0 and 0.2, the variation in AUC-performance remains high, jumping back and forth around 50% - 70% AUC (see Figure 7a). However, starting at when $\boldsymbol{\varepsilon}$ is above 0.2, the AUC begins to decline as ε increases, which is the opposite of the expected trade-off. For example, output perturbation presents an AUC of 69.2% when $\boldsymbol{\varepsilon}$ is 0.2 and gradually declines to 60.04% when $\boldsymbol{\varepsilon}$ is at 0.5. A similar pattern exists for objective perturbation. It could be that the parameters used for tuning were not exhaustive enough where different privacy budgets required different parameter values for optimal results. Hence, the grid search may not have covered the necessary parameters for the optimal results when $\boldsymbol{\varepsilon}$ is 0.2 and thus, it performs worse than when less noise is inserted. Overall, the radial kernel does not balance the trade-off between utility and privacy as well as the linear kernel. The radial kernel takes a longer duration of running time than the linear kernel which this research was not able to meet its model needs. It could also be that the complexity of the radial kernel was not fit for the nature of the data. Additionally, both SVM models seem to vary more in the trade-off performance than RLR which makes it also hard to distinguish the difference in performance of output perturbation and objective perturbation in SVM.

Figure 7a – 7d - *Private Radial Kernel SVM models compared baseline non-private model*

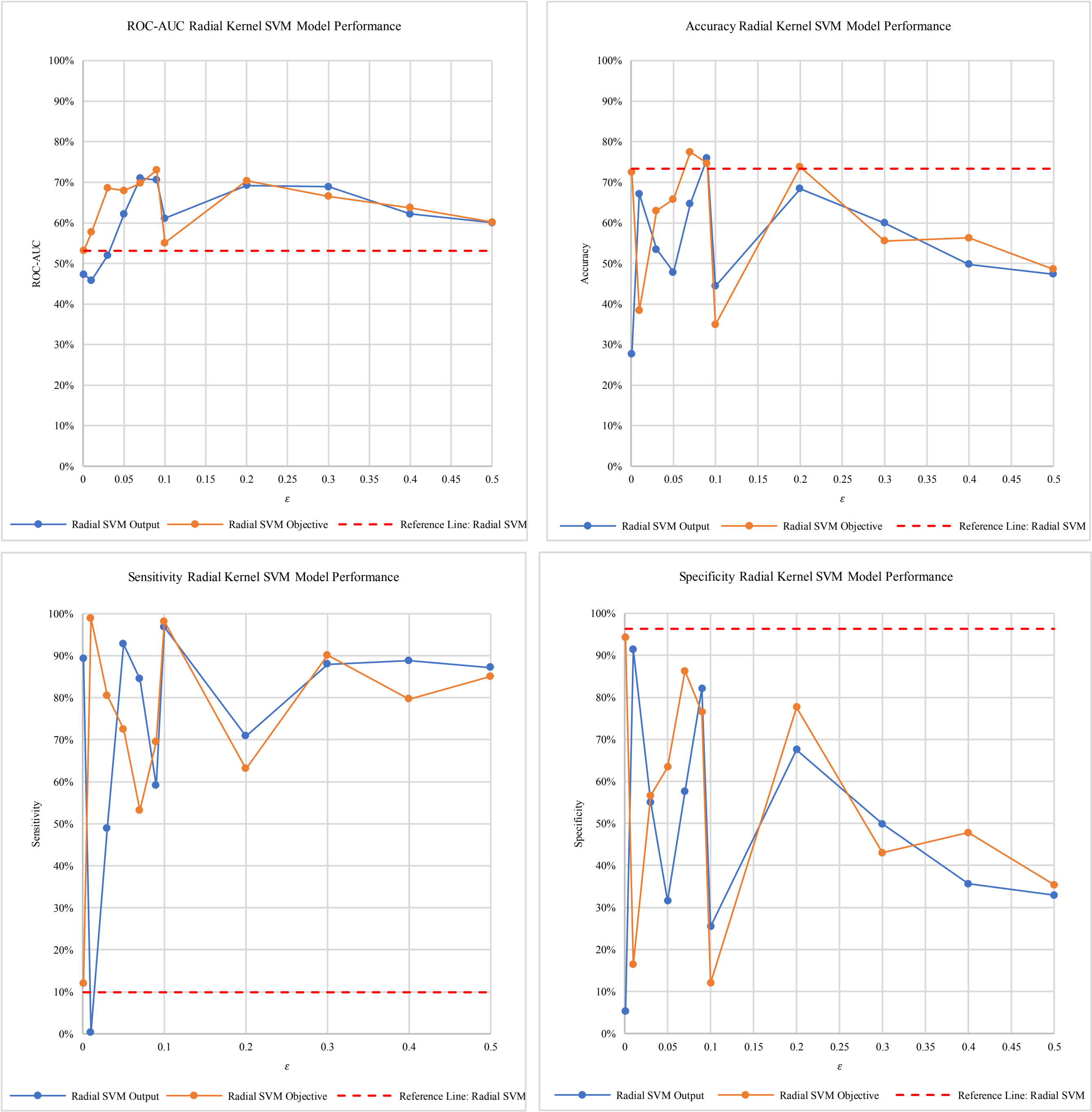


Table 3 - *ROC-AUC of all optimal models at different values of ε*

ε	0.001	0.01	0.03	0.05	0.07	0.09	0.1	0.2	0.3	0.4	0.5
RLR Output	54.05%	78.64%	76.90%	82.01%	82.45%	82.47%	82.51%	82.96%	82.84%	82.80%	82.95%
RLR Objective	69.42%	69.81%	77.40%	81.79%	81.87%	82.63%	82.51%	83.72%	83.45%	83.59%	83.54%
SVM Linear Output	47.88%	62.88%	55.62%	54.18%	68.54%	68.84%	74.54%	73.18%	68.71%	71.66%	69.37%
SVM Linear Objective	60.25%	53.68%	46.64%	56.14%	69.21%	66.01%	67.94%	67.19%	70.39%	67.18%	70.45%
SVM Radial Output	47.31%	45.83%	52.01%	62.17%	71.05%	70.59%	61.13%	69.21%	68.91%	62.20%	60.04%
SVM Radial Objective	53.16%	57.69%	68.56%	67.93%	69.73%	73.00%	55.07%	70.37%	66.54%	63.75%	60.18%

Table 4 - *Accuracy of all optimal models at different values of ε*

ε	0.001	0.01	0.03	0.05	0.07	0.09	0.1	0.2	0.3	0.4	0.5
RLR Output	29.21%	58.00%	64.39%	61.90%	58.56%	61.05%	60.84%	62.05%	64.25%	64.25%	63.11%
RLR Objective	39.73%	44.28%	68.02%	71.14%	63.04%	63.04%	66.81%	71.00%	70.93%	70.93%	70.65%
SVM Linear Output	31.98%	50.25%	56.50%	63.40%	59.56%	59.13%	72.64%	67.38%	59.06%	64.39%	60.41%
SVM Linear Objective	41.08%	73.56%	59.35%	45.49%	61.05%	54.73%	59.06%	56.08%	61.90%	55.93%	61.62%
SVM Radial Output	27.65%	67.16%	53.45%	47.83%	64.75%	75.98%	44.42%	68.44%	59.99%	49.75%	47.33%
SVM Radial Objective	72.42%	38.38%	62.97%	65.81%	77.47%	74.63%	34.90%	73.77%	55.51%	56.29%	48.54%

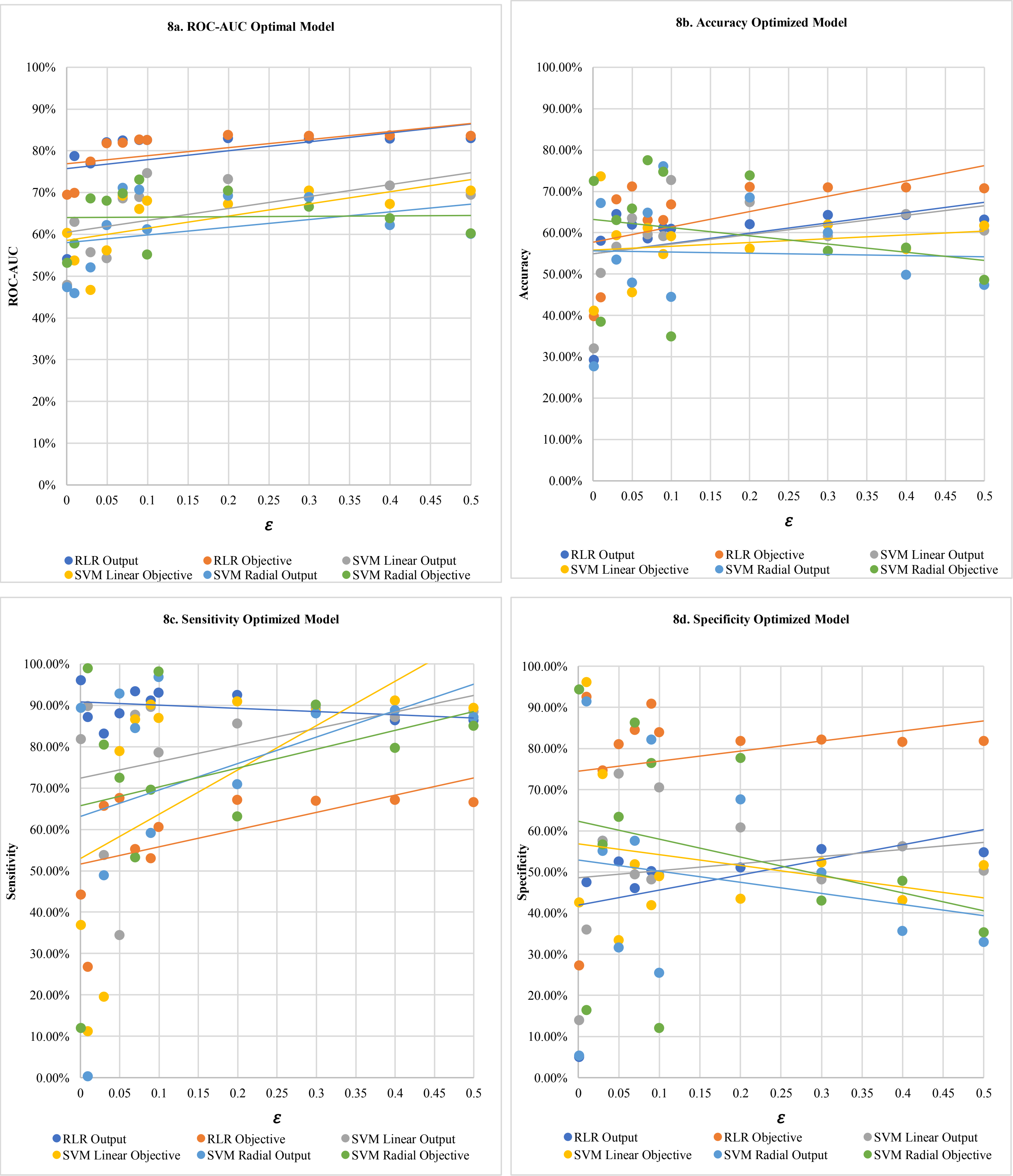
Table 5 - *Sensitivity of all optimal models at different values of ε*

ε	0.001	0.01	0.03	0.05	0.07	0.09	0.1	0.2	0.3	0.4	0.5
RLR Output	95.99%	87.17%	83.16%	87.97%	93.32%	91.18%	93.05%	92.51%	88.24%	86.36%	86.36%
RLR Objective	44.24%	26.82%	65.63%	67.57%	55.28%	52.95%	60.60%	67.09%	66.89%	67.09%	66.60%
SVM Linear Output	81.82%	89.84%	53.74%	34.49%	87.70%	89.57%	78.61%	85.56%	89.30%	87.17%	88.50%
SVM Linear Objective	36.90%	11.23%	19.52%	78.88%	86.63%	90.11%	86.90%	90.91%	88.50%	91.18%	89.30%
SVM Radial Output	89.30%	0.27%	48.93%	92.78%	84.49%	59.09%	96.79%	70.86%	87.97%	88.77%	87.17%
SVM Radial Objective	12.03%	98.93%	80.48%	72.46%	53.21%	69.52%	98.13%	63.10%	90.11%	79.68%	85.03%

Table 6 - *Specificity of all optimal models at different values of ε*

ε	0.001	0.01	0.03	0.05	0.07	0.09	0.1	0.2	0.3	0.4	0.5
RLR Output	5.03%	47.43%	57.60%	52.47%	45.98%	50.15%	49.18%	51.02%	55.57%	56.24%	54.70%
RLR Objective	27.27%	92.51%	74.60%	81.02%	84.49%	90.91%	83.96%	81.82%	82.09%	81.55%	81.82%
SVM Linear Output	13.94%	35.91%	57.50%	73.86%	49.37%	48.11%	70.47%	60.79%	48.11%	56.15%	50.24%
SVM Linear Objective	42.59%	96.13%	73.77%	33.40%	51.79%	41.92%	48.98%	43.47%	52.27%	43.18%	51.60%
SVM Radial Output	5.32%	91.38%	55.08%	31.56%	57.60%	82.09%	25.46%	67.57%	49.85%	35.62%	32.91%
SVM Radial Objective	94.29%	16.46%	56.63%	63.41%	86.25%	76.48%	12.00%	77.64%	42.98%	47.82%	35.33%

Figure 8a – 8d – *Performance metrics of all private models and their line of best fit*



5. Discussion

5.1. Conclusion

Based on the results obtained from the trained models, the research question is answered: the extent to which DP balances the trade-off is dependent on the privacy budget (***ε***), model choice, and DP technique:

- The model choice affects the trade-off because it seems that RLR compared to SVM better demonstrates that as ***ε*** increases, privacy decreases which leads to an overall higher performance in the AUC and accuracy.
- The sensitivity and specificity also demonstrate a trade-off between utility and performance except for the sensitivity value of RLR output perturbation where when ***ε*** increases, the sensitivity decreases (see Figure 8c). This could be attributed to the trade-off that also exists between sensitivity and specificity. Sensitivity and specificity vary greatly between models, sometimes performing better than the baseline models. To better balance the sensitivity and specificity, changing the probability threshold of classifying the classes could increase or decrease the values interchangeably.
- Deciding on a threshold is dependent on the organization and specific use case. High sensitivity means the model does well in correctly identifying individual cases that churn. This might be important for companies who would like a full grasp of all the characteristics of a customer that are more likely to churn so the most optimal communication campaign message is created. If companies were to send out a cost-effective retention campaign companies can prioritize high specificity. High specificity correctly classifies the non-churners, minimizing costs by avoiding unnecessary campaign messages sent out to non-churners. Although a simplified example, it gives an understanding on how the output of the model can be applied in real life.

Additionally, impacting the trade-off was the choice of the model between RLR and SVM as well as the method of perturbation:

- When ***ε*** is the same value, RLR leads to a better AUC performance than SVM, no matter if noise was inserted in the output or objective function. This could be because RLR is more robust to noise and less sensitive to irrelevant features or noisy inputs.

- SVM does not do as well when noise is added and can be considered more complex to tune. This added complexity does not bind well with DP as more points to consider means increased computational challenges.
- When comparing between the perturbation techniques for RLR, the results are comparable to that of Chaudhuri et al., (2011) where on average, objective perturbation performs with higher accuracy and AUC.
- RLR objective perturbation balances the utility and privacy trade-off better than output perturbation. The perturbation techniques when compared within just SVM does not make large differences in performance when measuring accuracy and AUC.

Understanding the anomalies and variations that exist within the models gives a better understanding of how the perturbation techniques behave with the model and data:

- The large variation in performance from one ε to another can be attributed to the random noise insertion of DP that leads to varying random performances of the trained model. The efforts of cross-validation may have not been enough to offset the anomalies.
- Another contributing factor for the anomalies could be the choice of regularization parameter. Prior literature indicates that the regularization parameter can impact the privacy and utility trade-off. For instance, Chaudhuri et al. (2011) proposed that a larger regularization parameter makes the model less sensitive, thereby requiring less noise. This suggests that the range of values chosen for the regularization parameter in the grid plays a part in the optimal trade-off value. There is a possibility that the choice of regularization term in the grid was not exhaustive enough to capture the highest performance for the specific ε.
- Lastly, the large variation in results may be because the models were unable to reach the optimal performance, only producing suboptimal results due to computational limitations. SVM models, especially with the radial kernel, take significant running time. With the technology available, the research was not capable of covering all possible parameters that could lead to higher model performance. Consequently, it is possible that at certain values of ε, the chosen parameters allowed higher performance only by chance.
- Overall, RLR models outperform SVM models due to its stability in performance as well as its better ability to balance the trade-off between privacy and utility.

5.2. Dataset Generalizability of Findings

The research objective was to understand how DP can be implemented in managing churn rate models while preserving privacy. To extend the generalizability of the findings, it is important to consider the nature of the data in terms of data imbalance, dimensionality, and complexity.

- Firstly, considering the choice of model is no different in DP than with non-private models. SVM and RLR are appropriate for the Telco churn data since it included a binary outcome. Just like non-private models, implementing DP with SVM or RLR would not be appropriate if the outcome was continuous. Thus, incorporating DP via other modeling methods (e.g., random forest, neural networks) can be appropriate depending on the objective of the modeling.
- Secondly, the class distribution of the data can impact how well the trade-off is executed when implementing DP. The imbalance can result in biased predictions by favoring the classification of the majority class, leading to a decrease in performance (Farrand et al., 2020). Therefore, how the imbalance is dealt with affects how the model will perform in other use cases.
- Thirdly, higher dimensional data performs worse with DP algorithms (Chaudhuri et al., 2011). The infamous curse of dimensionality provides an extra layer of complexity and noise from the predictors itself that needs to be first dealt with (possibly preprocessing) before feeding into the DP-model. So, when dealing with high-dimensional DP models, there is a set of problems that requires special treatments (Chen et al., 2015).

All-in-all, the generalizability of DP models requires similar model decisions and preprocessing steps just like non-private models. So, when considering the generalizability of the findings, users must consider the nature of the dataset and its resemblance to the one used in this research before implementing the model for other use cases. This could just mean preprocessing results like rebalancing the data or using dimensionality reduction, or it might mean choosing a completely different model that works better. Managers must try what works best within their context while keeping in mind the privacy budget that comes with using DP.

5.3. Business Contributions and Recommendations

The results demonstrate Telco businesses trying to predict churn should most likely apply RLR objective perturbation. Although this research shows slightly better results for objective perturbation, managers should test both techniques in case the better performance is a result of chance and randomness. The research focuses on how managers can balance the performance of their predictive models while still considering the privacy of their customers. This means managers must also decide for themselves the privacy budget that works best for them. For example, data that is highly sensitive (e.g. IP address, geolocation) may require higher ε at the expense of model performance. Otherwise, data that is aggregate in nature and less sensitive (e.g., age, gender) may need less ε allowing for optimal model performance.

Managers should refer to privacy guidelines set by the organization as well as ther local laws (e.g., GDPR). After determining the desired level of noise, managers should proceed to use the models in their usual manner. Referring to Section 2.1 Customer Churn in Telecommunications gives a good indication of what "usual" could mean. Although, it is advisory that if managers opt for higher privacy, they should apply a more skeptical and realistic approach when assessing the effectiveness of their communication, design, and marketing strategies. For example, if the churn model is to understand the customer segmentation of those more likely to stay or more likely to churn, managers that use a lower ε should create a broader and more general retention campaign as the model performance may not be as accurate as when ε is big. Striking the right balance is crucial because if managers play it too safe, they might miss the opportunity to create successful and personalized campaign strategies. Conversely, excessive risks may lead organization to disaster with legal fines and negative sentiment for the organization. Therefore, organizations must find a balance so that they still follow legal guidelines but still stay competitive by capitalizing on their data collection.

5.4. Limitations of the Study

Several limitations were acknowledged during the research. Firstly, computational capacity largely hindered the process. The 8-core CPU was enough for a reasonable amount of modeling but models with high computational needs such as radial kernel SVM would have benefited from higher computational capacity. The SVM model parameters (i.e., parameter *D)* were adjusted to avoid extensively slow model training time which would often lead to errors. This meant that the research had to give up possible optimal SVM models to run models at the possible computation capacity.

Another limitation of the DP model was it required strongly convex functions for privacy guarantee. Therefore, the perturbation techniques are not generalizable to non-convex optimization problems. This limits the applicability to situations with strict requirements. The model is further restricted from using non-differentiable regularizer such as the l_1 norm. The l_1 norm and its solutions of sparsity can be advantageous over l_2 for high-dimensional cases and thus in DP, is not possible to benefit from. Henceforth, the requirements for privacy guarantee limits the generalizability of the findings. Future investigations could try loosening the DP requirements while still guaranteeing privacy.

Lastly, there are points of data leakage to consider such as during model tuning, through the SVM model itself, and when providing the model bounds. The research focused on using perturbation techniques for privacy, yet, there are risks to not incorporating privacy-preserving techniques in other parts of the modeling process. Firstly, with model tuning, each iteration accesses the data which could reveal patterns that expose individuals in the data. Consequently, the models are more prone to attacks than if a fully private preserving process was used. End-to-end privacy guarantees would incorporate preserving techniques during parameter tuning. Such as in Chaudhuri et al., (2011) used disjoint subsets of the data and randomized the classifier to release as a privacy-preserving cross-validation which, could be used to further extend the model in this research as well. One thing to note, however, is this technique does require large datasets.

Another point of privacy leakage is the SVM model itself. The kernel function provides a measure of similarity between training data points for the decision boundary. This could lead to data leakage as the model derives patterns from the training data to determine the optimal classification. Thus, attackers that get information on the model can use reverse engineering to infer sensitive information about the training data from the combination of the kernel functions. The last leakage point of the model considered is when providing the bounds and scales of the variables within the model. Certain methodological choices with the model such as $\| x_i \|_2 < 1$ and scaling the features to properly use RLR and SVM is required to state the upper and lower bounds of the variable features. Unfortunately, stating the maximum and minimum values of the features can create privacy leakage since it makes it easier for attackers to infer range of possible values. A solution would be to the 5% and 95% quantiles with DP, resulting in an output that could be used to specify the upper and lower bound of the original model (Giddens & Liu, 2023).

5.5. Directions for Future Research

Based on the results, future research could help better understand DP within machine learning. The anomalies in the performance of the model when the privacy budget increases lead to questions about the relationship between the regularization parameter and DP. Researching any patterns or correlations may lead to researchers quantifying the privacy loss because of the regularization parameter. This will allow future users of DP in machine learning to better understand how the regularization parameter plays a role in their model to better preserve privacy.

Taking a broader step and considering the topic in general, there are other methods of DP or even other privacy-preserving techniques that could be considered. DP was only used within the machine learning process which means the data and model is still vulnerable to attacks. Future research could try to implement the model with synthetic data generation which involves creating artificial data that tries to mimic similar properties of the original dataset. This allows organizations to train and test the model while minimizing exposure to personal and identifiable information. Such a method may be favorable for organizations that want to outsource their analytics team to create a model for them as they are able to share the dataset.

6. Reference

BlastChar. (2018, February 23). *Telco customer churn*. Kaggle. Retrieved February 3, 2023, from https://www.kaggle.com/datasets/blastchar/telco-customer-churn?datasetId=13996&sortBy=dateRun&tab=collaboration

Chapelle, O. (2007). Training a support vector machine in the primal. *Neural Computation*, *19*(5), 1155–1178. https://doi.org/10.1162/neco.2007.19.5.1155

Chen, R., Xiao, Q., Zhang, Y., & Xu, J. (2015). Differentially private high-dimensional data publication via sampling-based inference. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/2783258.2783379

Chaudhuri, K., & Monteleoni, C. (2008). Privacy-preserving logistic regression. *Advances in Neural Information Processing Systems 21* .

Chaudhuri, K., Monteleoni, C., & Sarwate, A. D. (2011). Differentially private empirical risk minimization. *Journal of Machine Learning Research : Jmlr*, *12*, 1069–1109.

Chu, K. (1999). An introduction to sensitivity, specificity, predictive values and likelihood ratios. *Emergency Medicine Australasia*, *11*(3), 175–181. https://doi.org/10.1046/j.1442-2026.1999.00041.x

Cisco Cybersecurity. (2019). Maximizing the value of your data privacy investments [Review of *Maximizing the value of your data privacy investments*]. In *cisco* (pp. 1–14). CISCO. https://www.cisco.com/c/dam/en_us/about/doing_business/trust-center/docs/dpbs-2019.pdf

Dilmegani, C. (2021, June 1). *Differential Privacy: How It Works, Benefits & Use Cases in 2023*. AIMultiple. Retrieved February 3, 2023, from https://research.aimultiple.com/differential-privacy/

Duchemin, R., & Matheus, R. (2021). Forecasting customer churn: Comparing the performance of statistical methods on more than just accuracy. *Journal of Supply Chain Management Science*, *2*, 115–137. https://doi.org/https://doi.org/10.18757/jscms.2021.6125

Dwork, C. (2006). Differential Privacy. *Automata, Languages and Programming*, 1–12. https://doi.org/10.1007/11787006_1

Farrand, T., Mireshghallah, F., Singh, S., & Trask, A. (2020). Neither private nor fair. *Proceedings of the 2020 Workshop on Privacy-Preserving Machine Learning in Practice*, 15–19. https://doi.org/10.1145/3411501.3419419

Xia, G., & Jin , W. (2008). Model of customer churn prediction on support vector machine. *Systems Engineering - Theory & Practice*, *28*(1), 71–77. https://doi.org/10.1016/s1874-8651(09)60003-x

Hu, H., Salcic, Z., Sun, L., Dobbie, G., Yu, P. S., & Zhang, X. (2022). Membership inference attacks on Machine Learning: A Survey. *ACM Computing Surveys*, *54*(11s), 1–37. https://doi.org/10.1145/3523273

Hu, X., Yuan, M., Yao, J., Deng, Y., Chen, L., Yang, Q., Guan, H., & Zeng, J. (2015). Differential privacy in Telco Big Data Platform. *Proceedings of the VLDB Endowment*, *8*(12), 1692–1703. https://doi.org/10.14778/2824032.2824067

IMY. (2021, June 10). *The purposes and scope of GDPR*. IMY. Retrieved February 3, 2023, from https://www.imy.se/en/organisations/data-protection/this-applies-accordning-to-gdpr/the-purposes-and-scope-of-gdpr/#:~:text=One

Jain, H., Khunteta, A., & Srivastava, S. (2020). Telecom churn prediction and used techniques, datasets and performance measures: A Review. *Telecommunication Systems*, *76*(4), 613–630. https://doi.org/10.1007/s11235-020-00727-0

Keerthi, S. S., & Lin, C.-J. (2003). Asymptotic behaviors of support vector machines with Gaussian Kernel. *Neural Computation*, *15*(7), 1667–1689. https://doi.org/10.1162/089976603321891855

Kifer, D., Smith, A., Thakurta, A., & Mannor, S. (2012). *Private Convex Empirical Risk Minimization and High-Dimensional Regression*, *23*, 25.1-25.4.

Kokolakis, S. (2017). Privacy attitudes and privacy behaviour: A review of current research on the Privacy Paradox Phenomenon. *Computers & Security*, *64*, 122–134. https://doi.org/10.1016/j.cose.2015.07.002

Landis, T. (2022, April 12). *Customer retention marketing vs. Customer Acquisition Marketing*. OutboundEngine. Retrieved February 3, 2023, from

Lemmens, A., & Gupta, S. (2017). Managing churn to maximize profits. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2964906

Lin, P. (2022, July 27). *Tackling the issues of data privacy in the Marketing Landscape*. MarTech Cube. Retrieved February 3, 2023, from https://www.martechcube.com/tackling-the-issues-of-data-privacy-in-the-marketing-landscape/

Liu, X., Kong, W., & O2021, S. (n.d.). Differential privacy and robust statistics in high dimensions. *Annual Conference Computational Learning Theory*. https://doi.org/https://doi.org/10.48550/arXiv.2111.06578

Machanavajjhala, A., Gehrke, J., Kifer, D., & Venkitasubramaniam, M. (2006). L-diversity: Privacy beyond K-anonymity. *22nd International Conference on Data Engineering (ICDE'06)*. https://doi.org/10.1109/icde.2006.1

Marketing Evolution. (2020, August 14). *Tackling data privacy and digital marketing concerns*. A New Approach to Marketing Measurement & Optimization. Retrieved February 3, 2023, from https://www.marketingevolution.com/knowledge-center/data-privacy-issues-in-data-driven-marketing

Mehta, N., Steinman, D., & Murphy, L. (2016). *Customer success: How innovative companies are reducing churn and growing recurring revenue*. Wiley.

Modani, N., Dey, K., Gupta, R., & Godbole, S. (2013). CDR analysis based Telco Churn Prediction and Customer Behavior Insights: A Case Study. *Lecture Notes in Computer Science*, 256–269. https://doi.org/10.1007/978-3-642-41154-0_19

Norberg, P. A., Horne, D. R., & Horne, D. A. (2007). The privacy paradox: Personal information disclosure intentions versus behaviors. *Journal of Consumer Affairs*, *41*(1), 100–126. https://doi.org/10.1111/j.1745-6606.2006.00070.x

Rigaki, M., & Garcia, S. (2021). *A Survey of Privacy Attacks in Machine Learning*. https://doi.org/https://doi.org/10.48550/arXiv.2007.07646

Rodan, A., Faris, H., Alsakran, J., & Al-Kadi, O. (2014). Rodan, A., Faris, H., Alsakran, J. and Al-Kadi, O. (2014) A Support Vector Machine Approach for Churn Prediction in Telecom Industry. International Journal on Information, 17, 3961-3970. *International Journal on Information*, *17*, 3961–3970.

Rubinstein, B. I., Bartlett, P. L., Huang, L., & Taft, N. (2012). Learning in a large function space: Privacy-preserving mechanisms for SVM learning. *Journal of Privacy and Confidentiality*, *4*(1). https://doi.org/10.29012/jpc.v4i1.612

Rust, R. T., & Chung, T. S. (2006). Marketing models of service and relationships. *Marketing Science*, *25*(6), 560–580. https://doi.org/10.1287/mksc.1050.0139

Song, S., Chaudhuri, K., & Sarwate, A. D. (2013). Stochastic gradient descent with differentially private updates. *2013 IEEE Global Conference on Signal and Information Processing*. https://doi.org/10.1109/globalsip.2013.6736861

Smolic, H. (2022, July 4). *How to predict churn, and what kind of data is essential?* Graphite Note. Retrieved February 3, 2023, from https://graphite-note.com/how-to-predict-churn-data-you-need

Giddens, S. & Liu, F. (2023). DPpack: Differentially Private Statistical Analysis and Machine Learning (Version 0.1.0). Retrieved from https://CRAN.R-project.org/package=DPpack

Sweeney, L. (2002). K-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*(05), 557–570. https://doi.org/10.1142/s0218488502001648

Tamaddoni, A., Stakhovych, S., & Ewing, M. (2017). The impact of personalised incentives on the profitability of customer retention campaigns. *Journal of Marketing Management*, 1–21. https://doi.org/10.1080/0267257x.2017.1295094

Torrão, J., & Teixeira, S. (2023). The antecedents of customer satisfaction in the Portuguese Telecommunications Sector. *Sustainability*, *15*(3), 2778. https://doi.org/10.3390/su15032778

Tessitore, S. (2023, March 7). *What's the average churn rate by industry?* CustomerGauge. https://customergauge.com/blog/average-churn-rate-by-industry#:~:text=Telecommunications%3A%2031%25&text=However%2C%20towards%20the%20end%20of,prior%20in%202020%2C%20for%20example

Verger, R. (2017, December 8). *Here's how Apple can figure out which emojis are popular*. Popular Science. Retrieved February 3, 2023, from https://www.popsci.com/apple-figure-out-popular-emojis-differential-privacy/

Vishal, M., Mistra, R., & Mahajan, R. (2015). Review of Data Mining Techniques for Churn Prediction in Telecom. *Journal of Information and Organizaitonal Sciences* , *37*, 183–197.

Xu, D., Yuan, S., & Wu, X. (2019). Achieving differential privacy and fairness in logistic regression. *Companion Proceedings of The 2019 World Wide Web Conference*. https://doi.org/10.1145/3308560.3317584

Yu, D., Zhang, H., Chen, W., Yin, J., & Liu, T.-Y. (2020). Gradient perturbation is underrated for differentially private convex optimization. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*. https://doi.org/10.24963/ijcai.2020/431

Zhu, T., Li, G., Zhou, W., & Yu, P. S. (2017). Differential privacy and applications. *Advances in Information Security*. https://doi.org/10.1007/978-3-319-62004-6

7. Appendices

Appendix A: Description of Data

Variable Name	Variable Description	Class
customerID	The identification number of a customer	Character
gender	If customers are “male” = 0 or “female” = 1	Numeric
SeniorCitizen	If customers are senior citizens or not (Yes = 1, No = 0)	Numeric
Partner	If customers have a partner or not (Yes = 1/No = 0)	Numeric
Dependents	If customers have dependents or not (Yes = 1/No =0)	Numeric
tenure	How long a customer has stayed with the telco company in months	Numeric
PhoneService	If customers have a phone service or not (Yes = 1/No =0)	Numeric
MultipleLines	If customers have multiple liens or not (Yes/No/No phone service)	Factor
InternetService	The customer’s internet service provider (DSL, Fiber optic, No)	Factor
OnlineSecurity	If customers have online security or not (Yes/No/no internet service)	Factor
OnlineBackup	If customers have online backup or not (Yes/No/No internet service)	Factor
DeviceProtection	If customers have device protection or not (Yes/No/No internet service)	Factor
TechSupport	If customers have tech support or not (Yes/No/No internet service)	Factor
StreamingTV	If customers have streaming TV or not (Yes/No/No internet service)	Factor
StreamingMovies	If customers have streaming movies or not (Yes/No/No internet service)	Factor
Contract -> “Contract_dum”	If customers have contract terms “month-to-month” = 1/ “one-year” and “two-year” = 0	Numeric

PaperlessBilling	If customers have paperless billing or not (Yes = 1/No = 0)	Numeric
PaymentMethod	If customers paid via “electronic check” and “mailed check” = 0, “bank transfer (automatic) and credit card (automatic) = 1	Numeric
MonthlyCharges	Amount in dollars that customers were charged monthly	Numeric
TotalCharges	Amount in dollars that customers were charged in total	Numeric
Churn	If customers churned or not ("Yes” or “1”/ “No” or “0”)	Numeric