BT4101 B.Sc. (Business Analytics) Dissertation

How much is this cookie?

By:

Lim Kai Le (A0136254J)

Department of Information Systems and Analytics

School of Computing

National University of Singapore

2020/2021 Semester 2

BT4101 B.Sc. (Business Analytics) Dissertation

How much is this cookie?

By:

Lim Kai Le (A0136254J)

Department of Information Systems and Analytics

School of Computing

National University of Singapore

2020/2021 Semester 2

Project No: H175410

Advisor: Assoc Prof HAHN Jungpil

Deliverables: Final report

Abstract

The increasing usage of the Internet over the past few decades has become an opportunity for firms to collect user information to generate. One of the many rising concerns from this trend is the potential for serious invasion of user privacy, where data breaches from firms, unauthorised usage of the data collected and/or even illegal data collection without consent and permission can compromise user privacy. To reduce such threats of personal information being misused, users are starting to adopt End-User Privacy Enhancing Technologies (PETs) to guard their personal information. PETs protect user information by introducing impurities in the data being collected in terms of measurement errors and/or missing values in the data. The management of cookies is one of the most familiar privacy-related issues affecting computer users since Web browsing has become such an essential activity, and many information sites, services, and advertising companies on the Web use cookies to track user behaviour and collect personal information. This study aims to systematically find out how does measurement errors induced by PETs such as cookies erasers affects firms' analytical performance in a context of predictive analytics such as predicting purchases. Our simulation experiments find that adoption of cookies erasers can impact the predictive capability in the context of purchase classification significantly by up to a 50% decrease in performance. From these results, we are able to identify which group of consumers firms should put more emphasis on to mitigate the negative impact of the consumer adoption of PETs. The results provide a quantitative explanation for the extent of the impact of the adoption of PETs on firms' analytical performance in the context of a purchase classifier and can provide some generalisation in other areas of analytics conducted by firms.

Keywords:

Privacy-enhancing technologies, purchase classifier, simulation

Implementation of Software and Hardware:

Python 3.7, Google Cloud Platform

Acknowledgment

I would like to express my deepest gratitude to my supervisor Associate Professor Hahn Jungpil for his valuable guidance, patience, kind words and encouragement throughout the whole project. Without him, this project would not be possible and completed.

I would also like to thank Prof Hahn for allowing me to seek advice from his Ph.D. candidate, Mr. Chen Dawei for his valuable feedback during our meeting which provided plenty of insights for the project.

I would also like to thank my main evaluator Dr. Anand Ramchand for his valuable feedback after the midterm presentation that helped me frame the research problem appropriately.

List of Figures & Tables

Table of Contents

1. **Introduction**

Firms are aggressively collecting and storing consumer data to perform consumer analytics, generating behavioural insights for market advantage. This form of analytics allows firms to understand customers' preferences and tailor services to provide better browsing experiences and/or more relevant advertisements resulting in long-term customer relationships (Erevelles et al., 2016). This clearly shows the importance of consumer data to firms and what possible values they can generate justifying the need to collect and store it. With the increasing ease of data collection driven by new technologies, firms that tap on such technologies to collect personality information are more likely to become targets for hackers leading to privacy breaches creating privacy concerns among consumers.

In recent years, privacy concerns among consumers have steadily increased. For instance, 58% of consumers believe that credit-reporting companies and privacy laws do not protect their privacy, and 71% of consumers agreed with the statement that "... consumers have lost all control over how personal information about them is circulated and used by companies." (Wang & Petrison 1993, p. 8). At the same time, privacy laws and regulations are being put in motion to combat the threat of privacy invasion and data breaches. For instance, the General Data Protection Regulation (2018) released by the European Union most notably establishes a consumer's "right to be forgotten" and mandates the removal of all data related to the consumer if requested by the consumer. However, the mitigating effect of such regulations are severely handicapped due to the challenges when it comes to organisations creating guidelines and enforcing regulations. With the increasing need for firms to stay competitive, collection and utilisation of high volumes of personal information are popular ways for firms to stay relevant. Thus, many websites may resort to collecting personal information without consent by exploiting the lack of protection or through illegal means. This results in heightened privacy concerns among users and naturally, users will start to adopt various end-user privacy enhancement technologies (PETs) to protect their personal information. PETs refer to IT artifacts protecting informational privacy by eliminating or minimizing personal data by individuals (Van Blarkom et al., 2003).

Among various PETs, cookies erasers and management tools can overcome one of the most prevalent privacy-related problems concerning users. Web browsing has become an essential activity, and many information sites, services, and advertising companies on the Web use cookies to track user behaviour and collect consumer information. Through the usage of Internet cookies, web browsing may end up being pseudonymous, and users may be identified when personal information is associated with the cookies that are silently accepted by the browsers which leads to concerns arising from the misuse of such personal information. Therefore, the adoption of cookies erasers or cookies management software by consumers can help ensure that only minimal cookies are accepted to allow the website to function properly, with no additional cookies are accepted. For instance, many websites have the option to opt-out of non-compulsory cookies; "Targeting cookies" that are set via advertising partners that aims to build a profile of user's interest and show relevant ads, "Performance cookies" which tracks visits counts and traffic volume and "Functional cookies" which provides personalisation for the hosting domain and partner sites, etc. All the 3 mentioned cookies are non-compulsory and typically opted out, whereas "Strictly Necessary cookies" are needed for the website to function properly and cannot be opted out.

These cookies can be opted out by cookies management and may negatively impact firms' analytical performance as well as generation of value because some data that may be essential for firms' analytics may not be collected accurately. For instance, an average of 2.5 distinct cookies was observed per computer for Yahoo, indicating that there is an overstatement of the number of visitors by a factor of 2.5 times which may affect the true reach of online advertisement campaigns (Abraham et al., 2007). If this were to be applied to China's e-commerce market, it can lead to a tremendous loss due to the sheer volume of sales from being the largest e-commerce market representing over 50% of global online sales (Li, 2017) as accurately predicting sales volume is dependent on web search data (Wei et al., 2014) which can be affected by users' cookie management.

Most prior works have revolved around online advertisers regarding the impact of deleting cookies on online advertisement, users' behaviour on deleting cookies and predicting sales volume of e-commerce. For instance, deletion of cookies can lead to an overstatement of unique visitor counts, understatement of repeated visitor counts and understatement of the conversion rate for site-centric measurements, while for ad server measurements, overstatement of reach and understatement of frequency is also reported (Abraham et al., 2007, p. 13). Since the prediction of sales volume is dependent on search data (Wei et al., 2014), which is a site-centric measure that is affected by the deletion of cookies, it is possible that the prediction of sales volume can be affected by cookies settings as well.

In this study, we aim to uncover how the prediction of online purchase, an important analytics task for firms, can be affected by the adoption of PETs which in this case, cookies deletion technology. We further ask what can firms do to mitigate such negative impacts?

To answer the questions above, we first review the existing frameworks of how cookies work and how PETs can affect the data collected by firms to understand the nature of the data problems induced by cookies deletion technology. Second, we conduct a systematic simulation study to investigate how such a PET may influence firms' analytical performance depending on users' adoption behaviour (e.g., adoption rates & adoption patterns) and the intensity of cookie deletion (interval of deletion). Such results will better position firms to recognise the impact of such degradation in the quality of data collected and strategise about what can be done on their end to mitigate the potential adverse effects.

We aim to contribute to the literature on privacy enhancement technologies (PETs) more specifically the impact of cookies deletion on firms' analytical performance in two ways. First, we provide a summary of how firms use cookies to collect information and how does deletion of cookies affects that information collected from a theoretical standpoint. Second, with the simulation study, the practical impact of the adoption of PETs can be derived and measured at the different extents of the user's adoption behaviour (e.g., adoption rates & adoption pattern) along with the intensity of the cookie deletion (interval of deletion) in the context of a purchase classifier. With these results, insights can be derived for firms' mitigation against the potential negative impact of PETs.

2. **Literature Review**

Privacy-Enhancing Technologies are “a system of information and communication technology (ICT) measures protecting informational privacy by eliminating or minimising personal data thereby preventing unnecessary or unwanted processing of personal data, without the loss of the functionality of the information system.” (Van Blarkom et al., 2003, p. 33).

For this literature review, we attempt to understand how cookies work from the perspective of how firms use them to collect information, the type of information being collected and the degree of privacy issues. We also try to understand how the deletion of cookies affects the data collection and understand what kind of data problem it may induce based on a generalisable framework.

“A "cookie" is a small text file that a Web site sends to be stored on the hard drives of visitors to the site. Cookies contain information on varying topics; some relating to the number of visits a user makes to a particular Web site, others keeping track of a user's passwords
and preferences.” (Zimmerman, R. K. 2000, Vol 4:439 pp 442-443). These text files are formatted strings made up of semi-colon separated key-value pairs. For instance, “Name=Value; Host=example.com; Path=/account; Expires=Tue, 1 Dec 2018 10:12:05 UTC; Secure;” (Cahn et al., 2016). These cookies can be accessed later and display information on each Website that has been visited along with passwords, e-mail addresses and other information that are keyed in, within a period. When revisiting Websites, all the information associated with the cookie is also available to the Website (Zimmerman, R. K. 2000).

Another study regarding information leakage from cookies has shown that 62-73% of browsing history can be reconstructed (Englehardt, S. et al., 2015). It also hypothesised that different cookies can be linked to each other with little (two to three) visits and concluded that cookies enable trafficking. These studies have shown that firms can use cookies to obtain information that can lead to privacy issues especially when there is little control over the usage of data from the users once it has been collected by the websites, thus motivating the adoption of PETs.

From Abraham M., Lipsman A. & Meierhoefer C. (2007), they stated that cookies are prevalent with over 100 million Internet users monthly and studied the impact of cookies deletion on site-severs through the deposit of new cookies and potentially overstate the estimates of new users for cookie-based site severs and/or overstate the reach of online ad campaigns. However, the extent of the overstatement may be dependent on the frequency of visitation to sites or exposure to the campaign. This literature summarised the impact of cookies deletion qualitatively but did not quantify the extent of which it may affect firms' analytical performance, for instance, with such errors in the data collected, how are firms affected which is a critical interest in this paper.

Another study on the classification of end-user PETs stated that they can be classified into six categories, namely; communication anonymizers, privacy setting, transparency enhancing technologies, trackers and evidence erasers, filters and blockers, and personal data stores, all of which protects the users' data and information through leaving out data values resulting in missing data and/or adjusting the data values leading to deviation from the true values (Chen. & Hahn, 2020, pp. 5-7). Among those mentioned, we will be focusing on those that will affect firms' collection of data such as cookies deletion which is categorised under trackers and evidence erasers to uncover what data problem (missing values/measurement error) it causes and the extent of it which will allow us to quantify the impact on firms' analytical performance.

Since there isn't a generalisable approach in analysing the impact of PETs on firms' ability to make decisions, we build on prior work that analyses the impact of the adoption of end-user PETs on firms' ability to generate recommendations (Chen & Hahn 2020) and adopt a similar approach to analyse the impact of PETs on purchase predictions.

3. **Research Methodology**

3.1 Research Approach

In order to answer the main research (we want to know how the adoption of PETs affects the firm's analytical performance), we need to know the factors that are involved in the adoption of PETs by users and also the intensity of the PETs adopted. User's behaviour will affect their likelihood of adoption of PETs along with the overall proportion of the users adopting PETs. Both factors involving users' behaviour are influential factors in determining the impact for firms' analytical performance as the proportion of users adopting PETs increases the harder it is for firms to make accurate analysis due to the data problem incurred either through missing values and/or measurement errors. The intensity of the PETs' protection can also affect the extent of the data problem induced, higher intensity may lead to a larger measurement error or more missing values and vice versa.

Therefore, the main factors for this simulation are 'Intensity of protection', 'Adoption pattern' and 'Adoption Rates' which will be defined as follows:

Intensity of protection refers to the duration x, which represents the interval in which cookies will be deleted. With a shorter duration of x reflecting a higher intensity, it will result in more frequent deletion of cookies leading to the collection of information being incomplete, resulting in measurement errors as prior sessions will become untraceable after cookies are being deleted.

Adoption Pattern reflects the likelihood of each user to adopt PETs. Different users may have different opinions of their privacy concerns. For instance, users with higher frequency of usage may be more aware of privacy concerns and may choose to increase their level of privacy protection (Kevin et al., 2008). Therefore, such users can be associated with a higher likelihood to adopt PETs. On the contrary, users with a lower frequency of usage may be deemed as being more concerned about their privacy. This will result in such users having a higher probability of adopting PETs. Thus, we will factor these into the simulation to understand the impact of different adoption patterns based on the frequency of usage.

Adoption Rates refer to the proportion of users (identified through unique session-id) adopting the usage of PETs taking on values ranging from 0 to 90%. At higher adoption rates, the firm's analytical performance will be worse as the degradation of data will be more severe through the introduction of more measurement errors of the aggregated values as sessions get broken down.

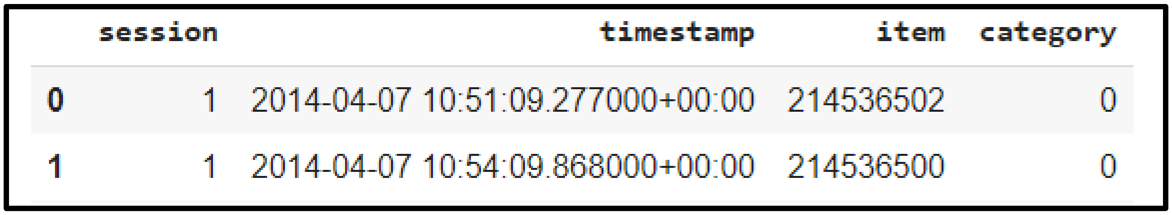
Since we’ve decided to experiment in the context of a purchase classifier which is a binary classification task, the evaluation metric selected is F1-score. F1-score is defined as the harmonic mean between precision and recall, where precision is the fraction of true positive examples among the examples that the model classified as positive while recall is the fraction of examples classified as positive, among the total number of positive examples. The reason for not using accuracy as the evaluation metric is due to the dataset being heavily skewed with approximately 95 – 5% distribution of the two binary classes, thus accuracy will not be a representative metric of the analytical performance i.e., the dummy prediction of the majority class will give 95% accuracy. Hence, we will be using F1-score as the main metric of evaluation by looking at the percentage decrease in performance in each setting against the model without any simulation.

3.2 Data specifications

3.2.1 Data source

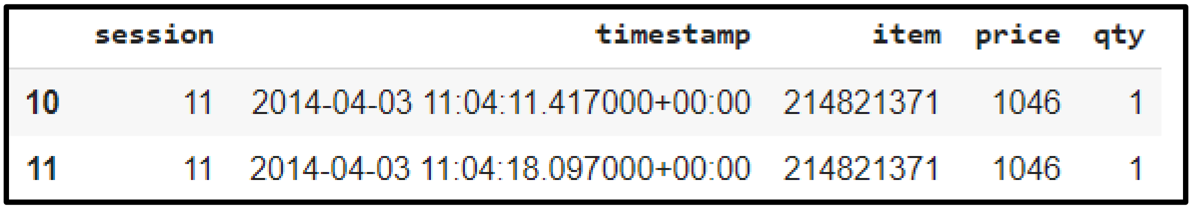
We will use the dataset provided by the RecSys Challenge in 2015 from YooChoose which collected six months of clickstream data in an e-commerce setting selling various goods and products. In this challenge, two main log files were provided: click events and buy events. Some of these click events are associated with buying events based on the session ID. The goal is to develop a model that can predict whether a user (a session) will end up buying something and if the user buys, what items would the user buys. However, for this research, the emphasis will be on the ability of the model to predict if a purchase is being made – generating a purchase classifier.

The format of the 2 files are as follows:



	session	timestamp	item	category
0	1	2014-04-07 10:51:09.277000+00:00	214536502	0
1	1	2014-04-07 10:54:09.868000+00:00	214536500	0

Figure 1: Example of Clicks dataset



	session	timestamp	item	price	qty
10	11	2014-04-03 11:04:11.417000+00:00	214821371	1046	1
11	11	2014-04-03 11:04:18.097000+00:00	214821371	1046	1

Figure 2: Example of Buy dataset

Below are the tables indicating the descriptions of the columns in both datasets:

Column	**Description**
Session ID	The id of the session. In one session there are one or many clicks. Could be represented as an integer number.
Timestamp	The time when the click occurred. Format of YYYY-MM-DDThh:mm:ss.SSSZ
Item ID (item)	The unique identifier of the item that has been clicked. Could be represented as an integer number.
Category	The context of the click. The value "S" indicates a special offer, "0" indicates a missing value, a number between 1 to 12 indicates a real category identifier, any other number indicates a brand. For example, if an item has been clicked in the context of a promotion or special offer then the value will be "S", if the context was a brand e.g, BOSCH, then the value will be an 8-10 digits number. If the item has been clicked under regular category, e.g, sport, then the value will be a number between 1 and 12.

Table 1: Description of clicks dataset

Column	**Description**
Session ID	The id of the session. In one session there are one or many clicks. Could be represented as an integer number.
Timestamp	The time when the click occurred. Format of YYYY-MM-DDThh:mm:ss.SSSZ

Item ID (item)	The unique identifier of the item that has been clicked. Could be represented as an integer number.
Price (price)	The price of the item. Could be represented as an integer number.
Quantity (qty)	The quantity in this buying. Could be represented as an integer number.

Table 2: Description of buy dataset

3.2.2 Exploratory data analysis

The entire click event log consists of 33,003,876 entries while the entire buy event log consists of 1,150,607 entries. Both datasets are from 2014-04-01 to 2014-09-30. As can be seen from the ratio of clicks to buy activities, on average every 28.68 clicks generates 1 buy activity which offers an opportunity for firms to gain a competitive edge by improving customer acquisition rates and sales. (Qiu et al, 2015).

From Figure 3, we can see that the distribution of clicks and buys have a similar shape, which is aligned with the intuition that more clicks are associated with more buys.

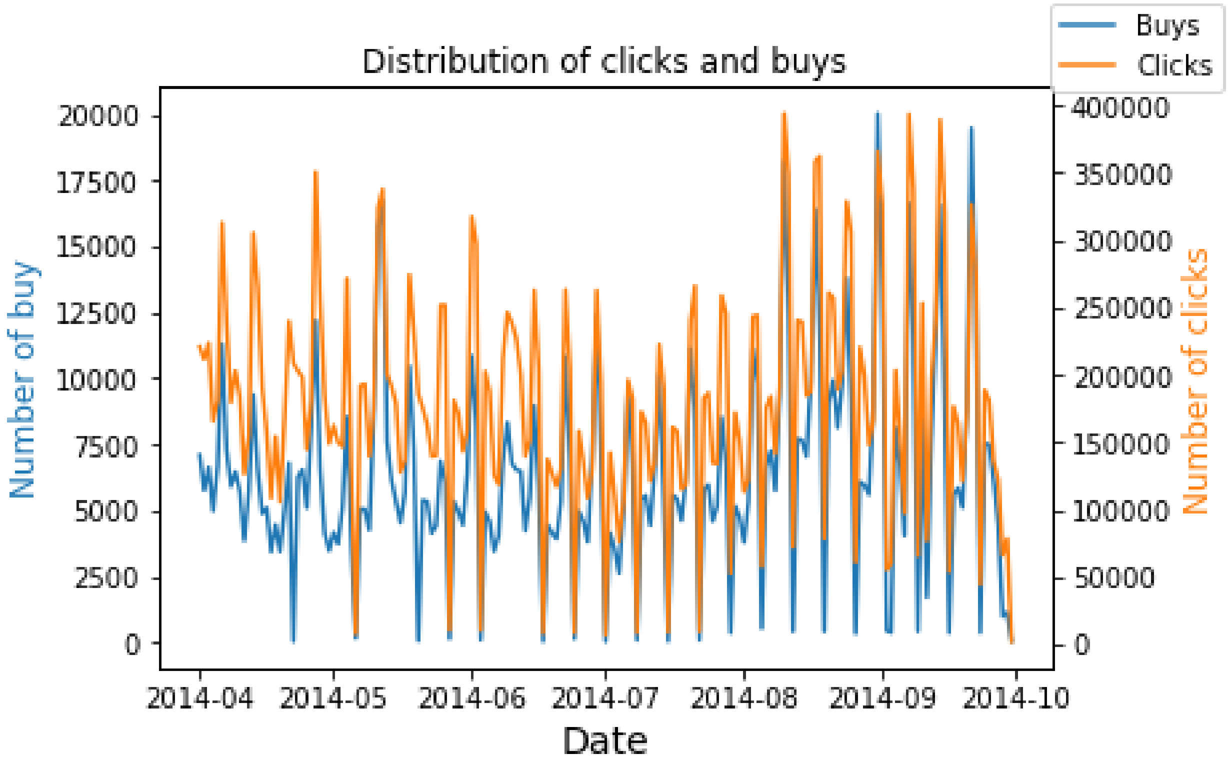


Figure 3: Distribution of buys and clicks across time

From Figure 4, we can see both the distribution and log-scaled distribution of the categories for the clicks. -1 indicates those with special offers, 0 represents missing values, 13 represents brand affiliated clicks while 1-12 represents a real category identifier. It can be observed that most of the distributions are of missing values or special offers.

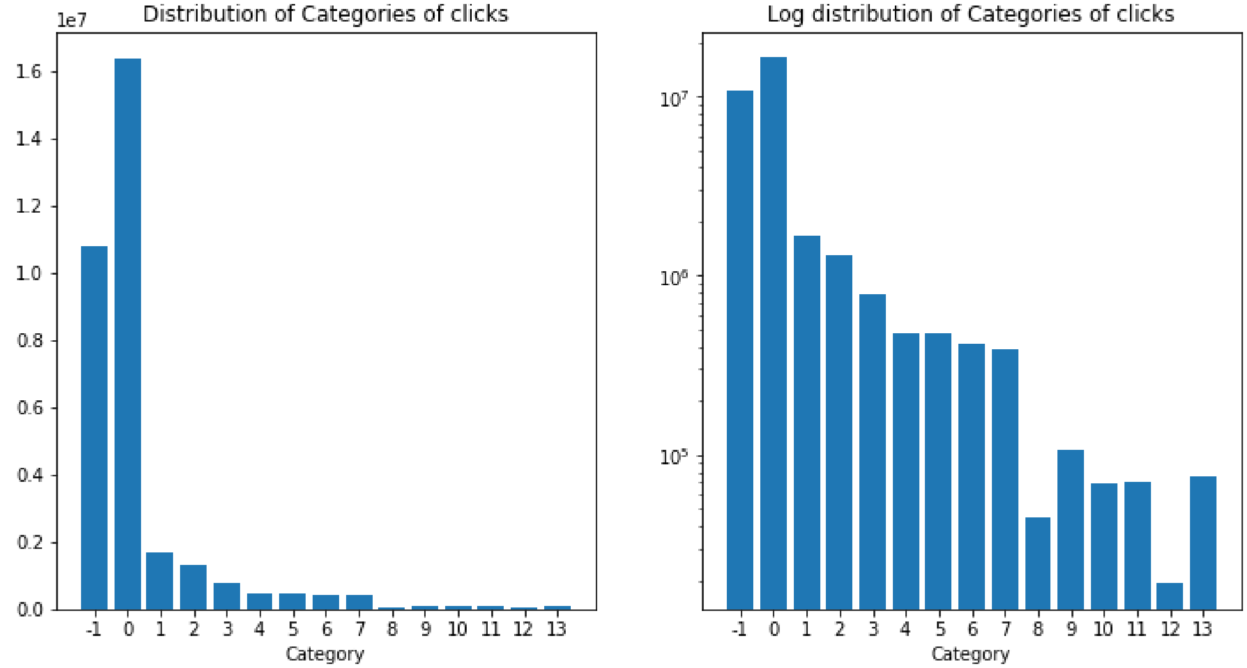


Figure 4: Distribution of categories of click

Since we are more interested in session-level information, Table 3 below summarises some statistics of the clicks and buy event logs aggregated at the session-level. There are 9249729 unique sessions among which 509696 of the sessions are buy sessions which consist of only 5.51%.

Percentage of buying session	Average clicks per session	Number of unique items clicked	Number of unique items bought
5.51%	3.57	52739	19949

Table 3: Statistic of sessions

From Figure 5, it is observed that the distribution of clicks per session is skewed to the right with the majority of the sessions having only 2 clicks be it a session with or without purchase made. It is reasonable to observe that the sessions without any purchase made are shorter than those with at least one purchase made. Hence this may be an indicator for what features to be included in the model to predict purchases made within a session. Similarly, an observation can be made that the percentage of buying sessions with 1 click is significantly smaller than those without comparing between the 2 proportions (~4% for purchase sessions, ~14% for non-purchase sessions). The percentage decrease in proportion is larger for non-purchase sessions as the number of clicks increase as compared to those with at least one purchase made, which can be due to users wanting to compare between different options to find the best deal before a purchase is being made.

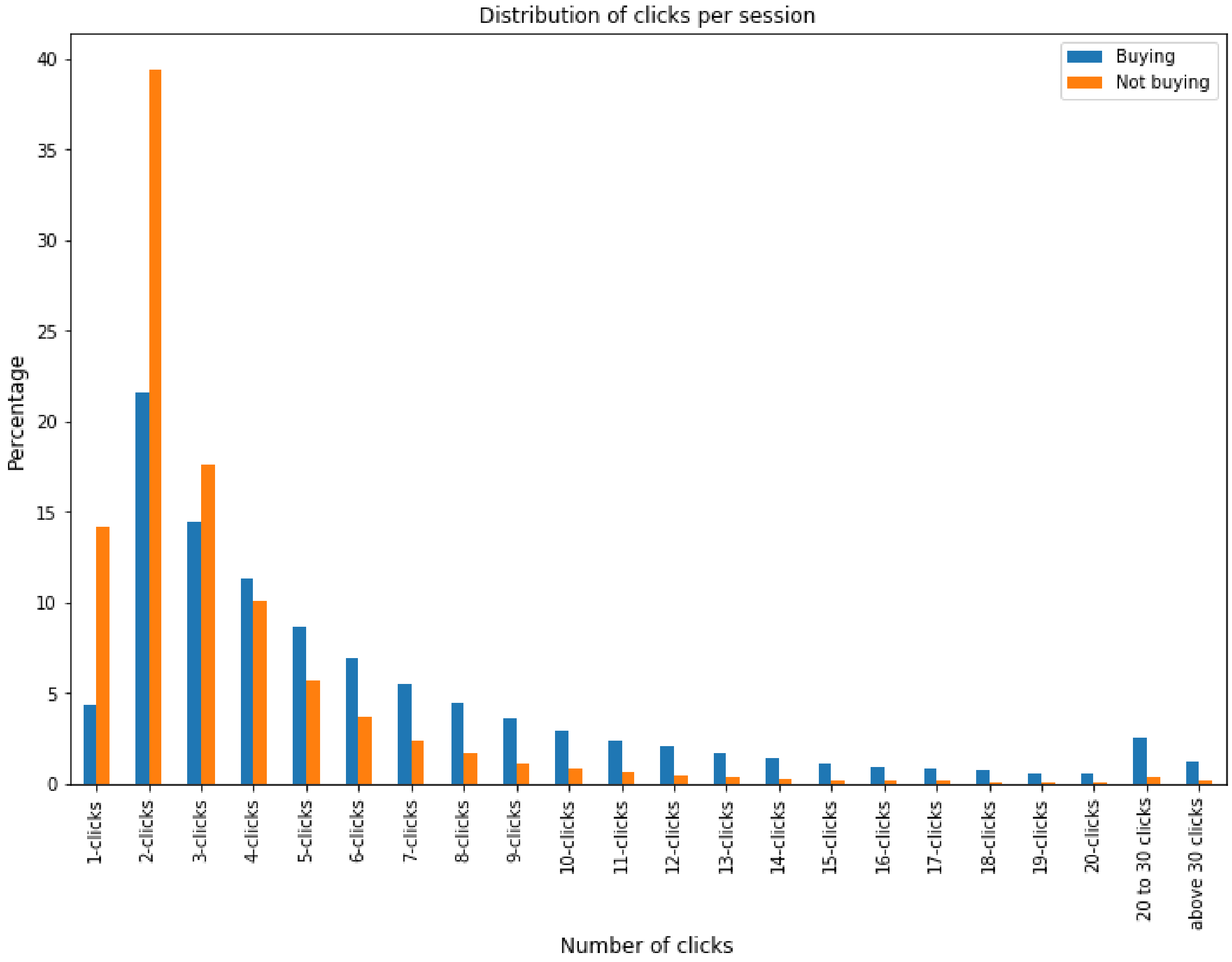


Figure 5: Distribution of clicks per session

From Figure 6, we can see that the average and median dwell time across the number of clicks per session seems to be decreasing, which is somewhat aligned with the intuition that clicks are usually clustered together especially for those who are planning to purchase as users will look through the different options. It can be observed that for those sessions with 1 click there isn't any dwell time as there is no subsequent click for comparison. This may provide an idea of how the sessions may be broken up based if cookies are deleted at specific intervals.

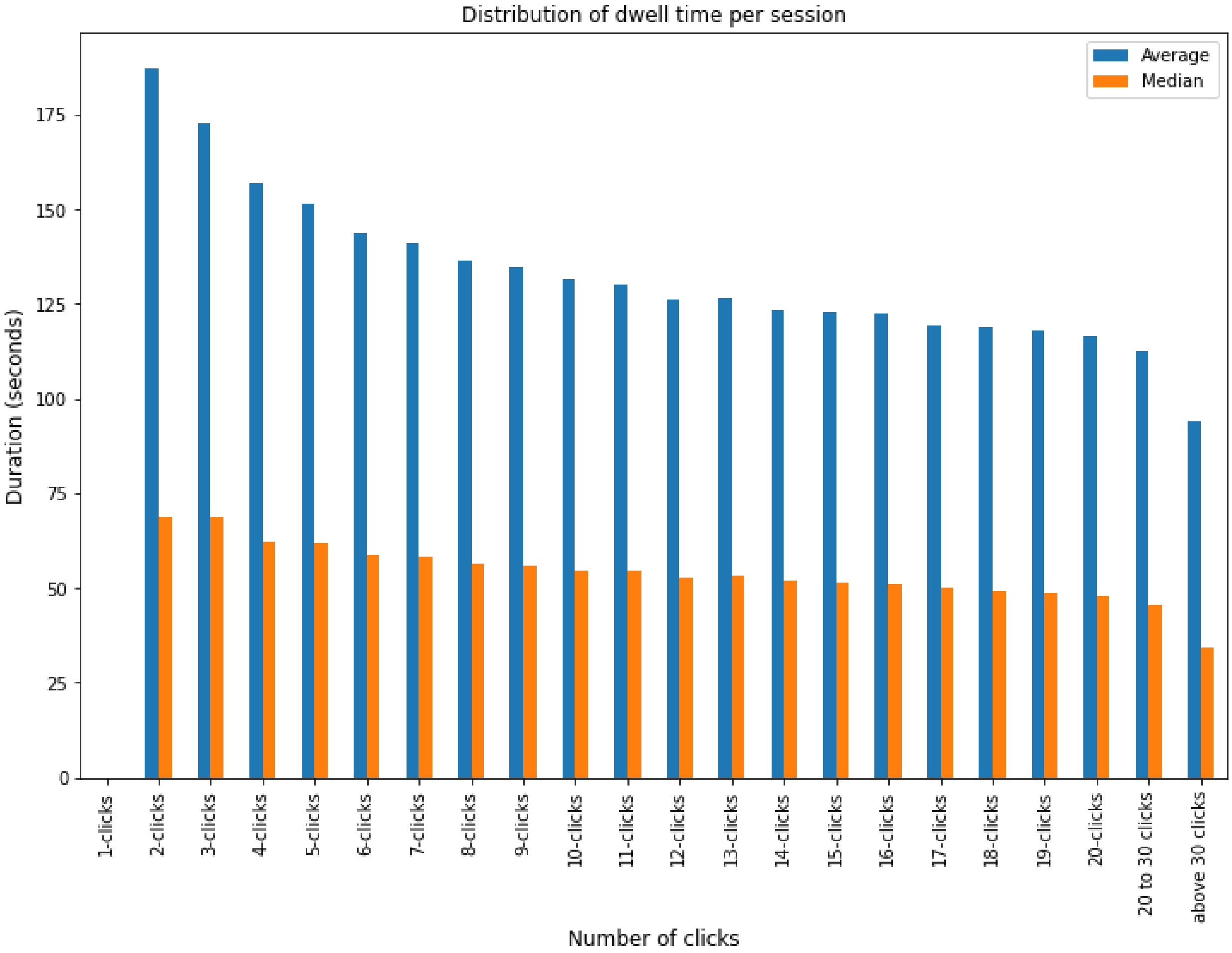


Figure 6: Dwell time across number of clicks

3.3 Detailed Methodologies

3.3.1 Feature engineering

Features generated include time-based features, categorical features with regards to the item as well as numerical features aiming to capture the trends of items.

Time-based features aim to capture information regarding the purchase period such as the time of the day, week and even year are indicative of the purchase intent (Mokryn et al., 2019). Categorical features indicate the type of item being clicked, whether it has a special offer or belongs to a branded category or other categories as well as the item itself. Numerical features aim to capture the user's browsing and purchase behaviour with regards to the dwell time defined as the duration customer viewed a particular product or page, where it has been correlated to the interest that the browser has on the product the user is currently looking at (Yi et al., 2014), number of unique items, number of clicks and number of items purchased as repeated browsing of an item represents a strong intention of purchase. Statistical information on the items concerning the entire training dataset such as the top N items purchased and categories clicked are also encoded as numerical features capturing the seasonal trend of items within the training window.

After pre-processing, there are a total of 194 features, of which 28 are temporal features, 18 are categorical features, 147 are numerical features which include statistical details of the top N items and 1 dependent variable - whether a purchase is made in the respective session. The temporal features are encoded into numerical features by using a package named fastai and applying the transformation on the date and time of the first and last click of the session respectively. While the information on the dwell time is generated by the time difference between consecutive clicks in seconds. The subsequent features are done by aggregating the counts of the sessions with items, categories or a combination of the two.

Feature Description	**Number / Type**
Numerical time features of the first and last click of the session (Encoded into cyclic temporal features)	2 * 14 **Numerical**
Total dwell time, average dwell time, maximum dwell time	3 **Numerical**

Number of clicks, unique items, categories and item-category pairs	4 **Numerical**
Top 5 items and top 3 categories by the number of clicks in the session	8 * **Categorical**
The item which is first/last click at least k = 1,2, …, 5 times in the session	10 * **Categorical**
Sparse matrix of click counts and total duration for the top 50 items and top 20 categories that were most popular in the whole training set	70 * 2 **Numerical**

Table 4: List of features used in the model

3.3.2 Classification method

The most difficult task is to find a classifier that allows the training of categorical features with many levels as there are over 50,000 unique items. Popular libraries that use ensembling such as XGBoost, LightGBM, etc. do not support categorical features directly and require them to be one-hot encoded into real-valued features. This will result in a huge number of training features that are hard to operate with. Thus, CatBoost is the natural choice for this task as it is a gradient boosting library developed recently that is capable of handling categorical features and has a better performance than the 2 other libraries mentioned (Dorogush et al., 2018). Since this is a binary classification task, binary log-likelihood is selected as the loss function to train the training set and further optimisation will be made to set the prediction threshold to optimise for the F1-score of the overall predictions with the validation dataset.

3.3.3 Preliminary results

A sample baseline model has been trained using a subset of the data from 2014-04-01 to 2014-04-15 using a CatBoost classifier. The training set has roughly 6.5% of the sessions having a purchase made and the model has achieved a training accuracy score of 0.93837 and AUC of 0.940278.

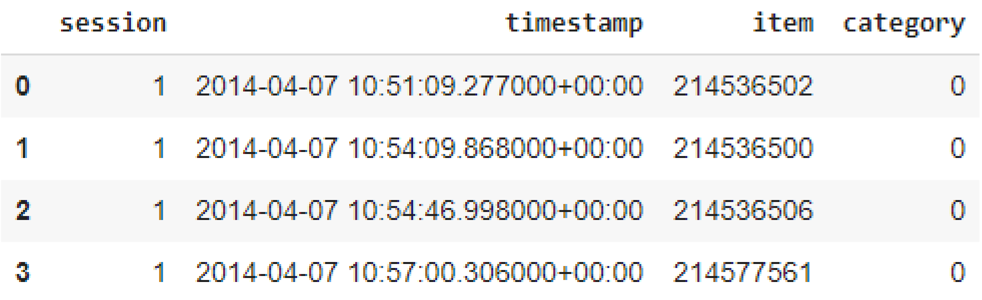
Testing has been done for the period between 2014-04-20 to 2014-04-25 with the validation set between 2014-04-15 to 2014-04-20 before any simulation of PETs for preliminary analysis. The test set has roughly 4.7% of the sessions having a purchase made and the model has achieved a testing accuracy score of 0.8926 and AUC of 0.8589. This shows that before any adoption of PETs, firms can have a fairly accurate prediction of customers' purchase decisions.

From Table 5, we've identified the top 10 most important features to be the specific orders of the item being clicked, the frequency of item being clicked, average dwell time of the sessions and the starting time of the first click.

Feature	Importance score
3rd last item clicked	32.868
3rd item clicked	15.760
2nd last item clicked	9.217
Most clicked item	7.267
4th last item clicked	5.492
Average dwell time	3.847
2nd most clicked item	3.775
2nd item clicked	2.792
Total duration	2.494
Starting hour of the first click	2.325

Table 5: Top 10 features of preliminary results

From Figure 7, you can see that session 1 has 4 clicks activities and the adoption of PETs may lead to the session being broken up based on timestamp, for example, clicks before 10:55 are grouped together while clicks after 10:55 within session 1 will be isolated leading to errors in subsequent calculations.



	session	timestamp	item	category
0	1	2014-04-07 10:51:09.277000+00:00	214536502	0
1	1	2014-04-07 10:54:09.868000+00:00	214536500	0
2	1	2014-04-07 10:54:46.998000+00:00	214536506	0
3	1	2014-04-07 10:57:00.306000+00:00	214577561	0

Figure 7: Clicks of session 1

Thus, we can see that, with the adoption of PETs, error values will be introduced to these variables leading to a decrease in the predictive ability of this model as the top 10 features generated are based on timestamps of the click activities.

3.3.4 Simulation

To simulate the impact of the adoption of PETs, the dataset is split randomly into 3 sets of training, validation and testing sets in an approximately 60-20-20% ratio. This split is done in a time-sensitive manner, where the training set is from sessions where the dates are before the validation set and the testing set is from sessions where dates are after the validation set to retain the temporal properties of the dataset. This simulation will mainly introduce measurement errors on the training and validation sets by identifying those who will adopt the usage of PETs based on adoption patterns and adoption rates. Those sessions identified to adopt PETs may be broken into smaller, unique and untraceable sessions dependent on the dwell duration of the clicks within the session and the intensity of protection. After the data problem is introduced, we will go through the same training algorithm with the simulated dataset that contains the measurement errors.

As mentioned previously, the main factors for this simulation are ‘Intensity of protection’, ‘Adoption Rates’ and ‘Adoption pattern’ which can take on different values at each setting.

Intensity of protection refers to the duration x, that after a session elapsed by will be broken down into another session. For instance, clicks that occur at time n, will be grouped together with those clicks that occur between n and $n + x$ inclusive. Subsequent clicks that occur within the session but not within time n and $n + x$ will be separated from the initial session. This is to reflect the usage of PETs via cookie erasers where sessions will be untraceable after cookies are erased. 3 ordinal intensities of protection (Low, Moderate, High) are adopted for this simulation by extracting the quantile values of the dwell times excluding 0 seconds – 75%, 50%, 25% which are 130.0565, 58.427, 26.76 seconds respectively reflecting the value x takes on at each intensity (Low, Moderate, High). Intuitively, at higher intensity, 1 session will more likely be broken down into more sessions due to the shorter interval time which is reflected by using the 25% percentile dwell time.

Adoption Rates will be set at intervals of 10%, ranging from 10% to 90% representing the proportion of PETs adopters. Each session selected as the adopters will be broken down during the simulation based on the intensity of the setting as mentioned above.

Adoption Pattern reflects the likelihood of each user to adopt PETs. 3 different approaches have been explored to determine the likelihood of adoption by each user – Uniform, Light-sensitive and Heavy-sensitive. For the Uniform pattern, each user has an equal probability of adopting PETs. For the case of Heavy-Sensitive, users that have a higher number of clicks will be associated with a higher frequency of usage thus, having a higher probability to adopt PETs. Lastly, for the case of Light-sensitive, users with a lower frequency of clicks will have a higher probability of adopting PETs. We will use the number of clicks in the session divided by the total clicks to generate the probability of the Heavy-sensitive setting and use the inverse to generate the probability of the Light-sensitive setting, while for Uniform setting, all sessions will have the same probability.

For all session *i, adoption pattern*:

$$P_{i,heavy} = \frac{M_i}{\text{Total number of clicks}}$$

$$P_{i,light} = \frac{1/P_{i,heavy}}{\sum_{i=1}^{N}\left(1/P_{i,heavy}\right)}$$

$$P_{i,uniform} = \frac{1}{\mathrm{N}}$$

Where $P_{i,adoption\,pattern}$ is the probability of session *i* adopting PETs under the given *adoption pattern*, M_i refers to the number of clicks in session *i* and N refer to the number of unique sessions.

With the given probability of adoption based on one of the 3 adoption patterns and the specified adoption rates, we can generate the sessions that will be selected as PETs adopters and perform the simulation of breaking down the sessions into possible smaller sessions based on the intensity at the given setting. The performance of the model will be recorded and we will be able to quantitatively look at the decrease in performance across different settings.

Figure 8 shows the overall steps to be taken for the whole simulation process where each setting will run on 20 different seeds and be compared to the performance prior to the simulations for analysis.

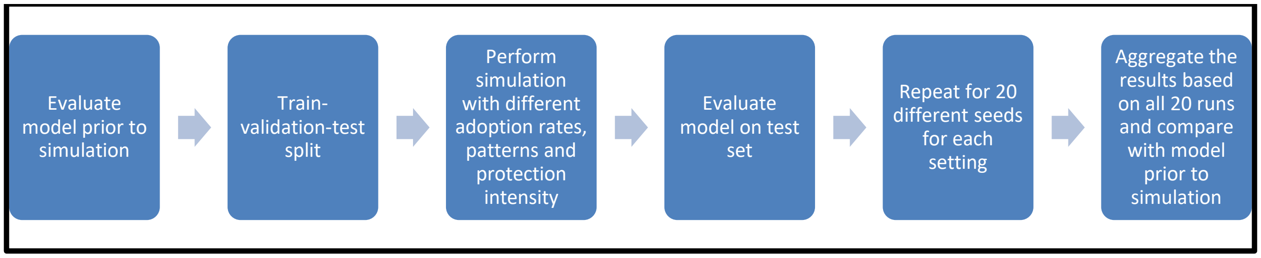

Evaluate model prior to simulation
Train-validation-test split
Perform simulation with different adoption rates, patterns and protection intensity
Evaluate model on test set
Repeat for 20 different seeds for each setting
Aggregate the results based on all 20 runs and compare with model prior to simulation

Figure 8: Flow of simulation

4. **Results Analysis**

The result for the simulation at each setting is aggregated by taking the average decrease in performance across all seeds.

4.1 Main Effect

4.1.1 Adoption Rates

Figure 9 shows the results of the average percentage decrease in F1-score across the adoption rates. As the adoption rate increases, it leads to a greater decrease in F1-score which is aligned with intuition as a higher adoption rate will result in a larger data problem where more measurement errors are introduced.

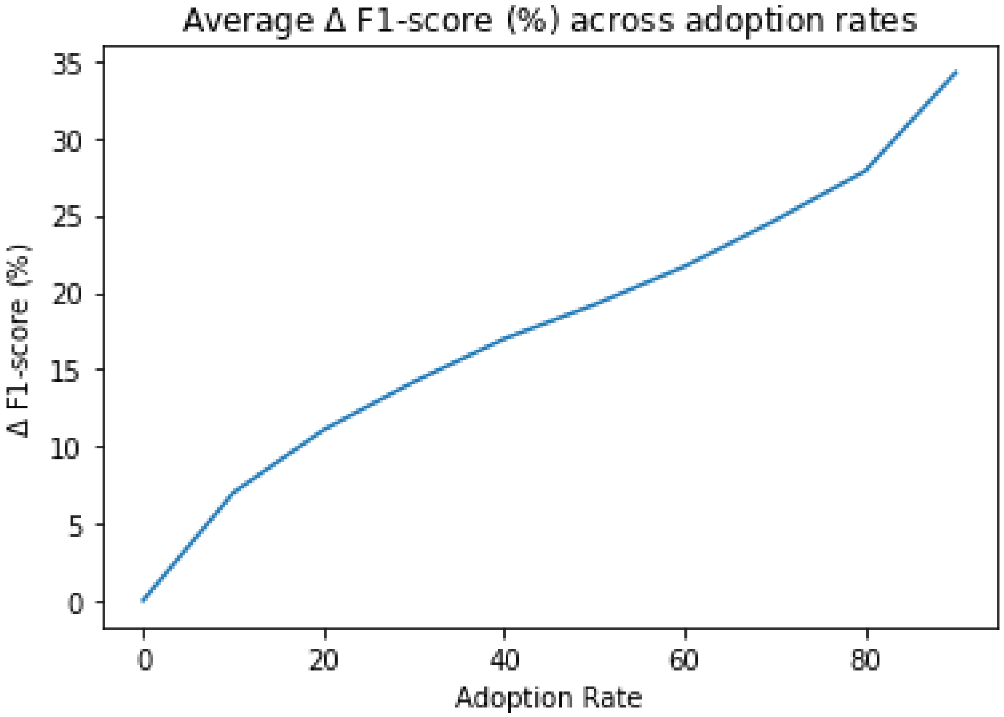


Figure 9: Average decrease in F1-score across adoption rates

4.1.2 Adoption Pattern

Figure 10 shows the results of the average percentage decrease in F1-score across the different adoption patterns. For the Heavy-sensitive adoption pattern, it resulted in the largest percentage decrease in average F1-score while the Light-sensitive adoption pattern experiences the lowest average decrease. This is reasonable as the Heavy-sensitive adoption pattern will result in sessions with more clicks being the more likely adopters and leading sessions being broken down into a larger volume of smaller sessions, which represents a bigger measurement error. The converse is also true, where Light sensitive adoption pattern will result in sessions with fewer clicks to be adopters and since there are lesser clicks to begin with, there is a limit to the extent of measurement error it can cause.

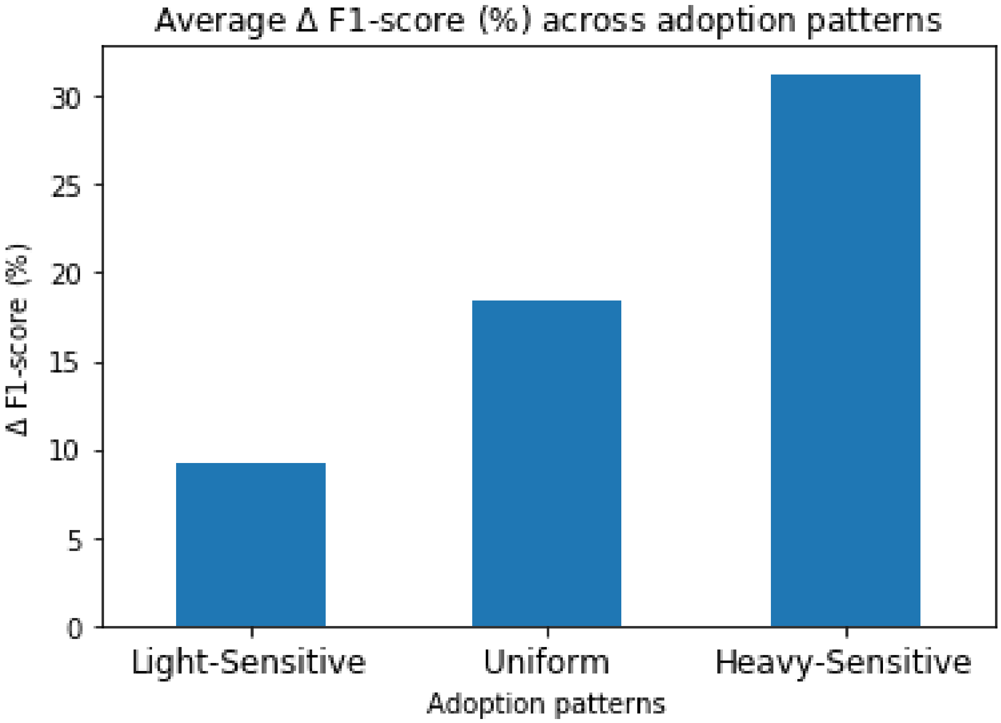


Figure 10: Average decrease in F1-score across adoption patterns

4.1.3 Protection Intensity

Figure 11 shows the results of the average percentage decrease in F1-score across the different protection intensities. We observed that the greatest decrease is from the Moderate intensity followed by Low intensity and High intensity resulting in the least decrease in performance. This may be due to a confounding effect by aggregating by the average among the 2 other factors (all 9 different adoption rates – 10 to 90% and the three different adoption patterns).

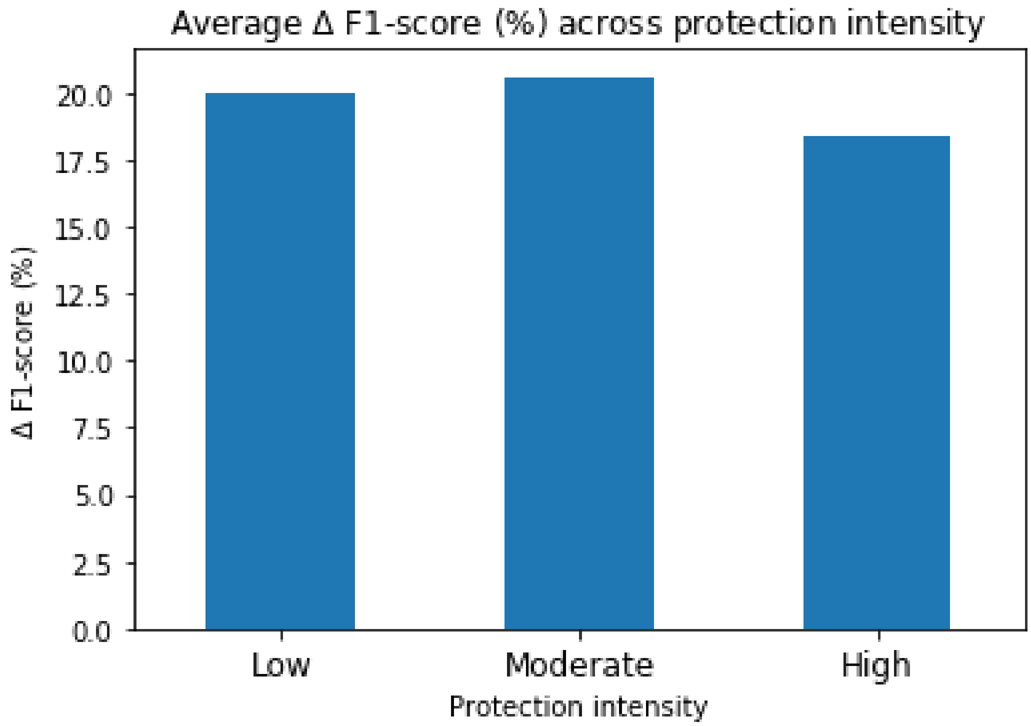


Figure 11: Average decrease in F1-score across protection intensity

4.2 Interaction effect

4.2.1 Adoption rate and adoption pattern

Figure 12 shows the results of the average percentage decrease in F1-score across the 3 different adoption patterns along with the adoption rates. The results are aligned with what we observed from the 2 factors individually above. The Heavy-sensitive adoption pattern resulted in the largest average decrease in F1-score followed by the Uniform adoption pattern then Light-sensitive. Higher adoption rates also result in a greater average decrease in F1-score as well.

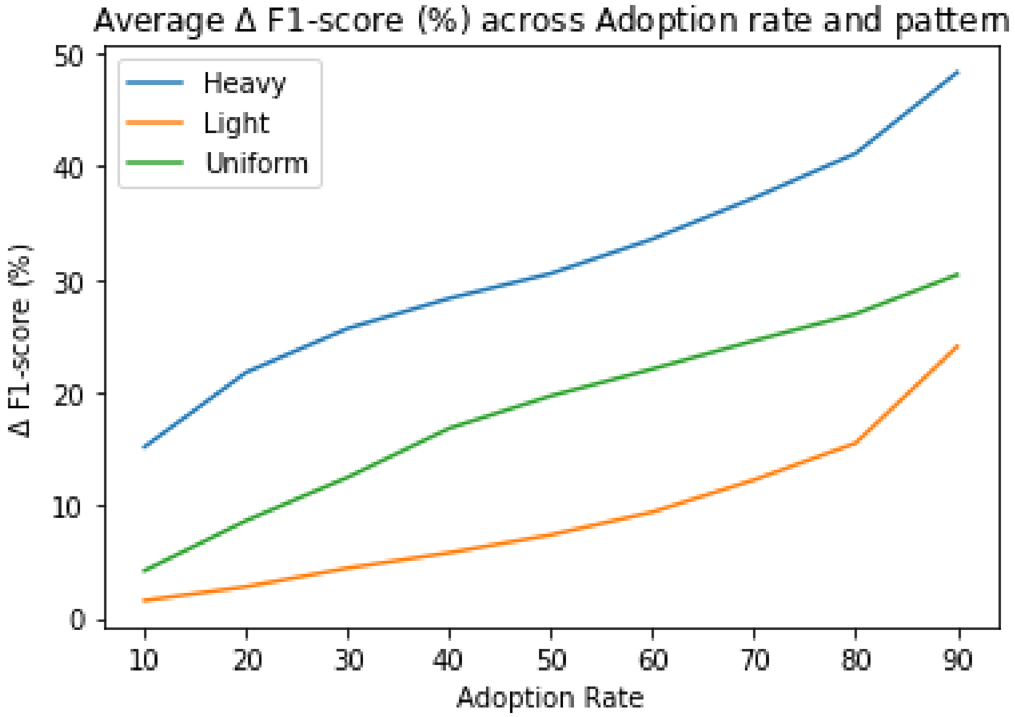


Figure 12: Average decrease in F1-score across adoption rates and pattern

4.2.2 Adoption rate and protection intensity

Figure 13 shows the results of the average percentage decrease in F1-score across the 3 different protection intensities along with the adoption rates. The results are still consistent with what we observed when looking at the factors individually with a larger average decrease in F1-score when the adoption rate increases, while the moderate protection intensity still resulted in the greatest decreases in average F1-score, however, we can observe that the changes in F1-score for the different intensity are not linear across the adoption rates which is an intersection effect that is not present when viewed individually.

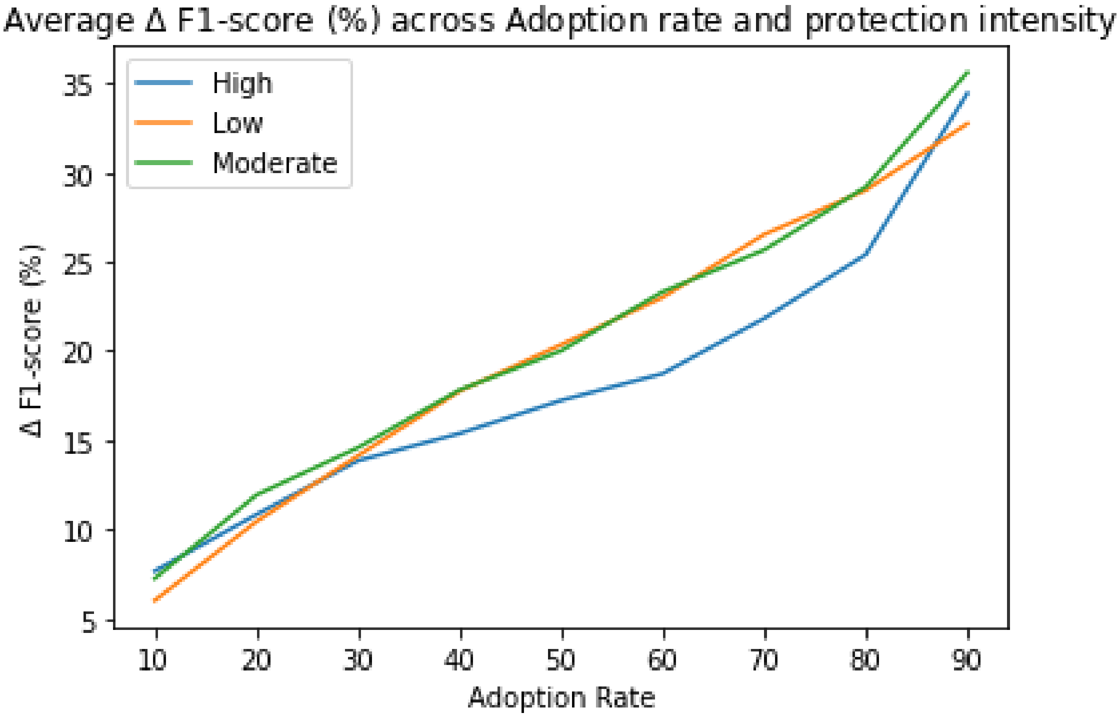


Figure 13: Average decrease in F1-score across adoption rates and protection intensity

4.2.3 Protection intensity and adoption pattern

Figure 14 shows the results of the average percentage decrease in F1-score across the 3 different adoption patterns and the 3 different protection intensities. The results are consistent with what is observed individually where the Heavy sensitive adoption pattern resulted in the greatest average decrease in F1-score followed by Uniform and lastly Light-sensitive. However, with this intersection, we can observe that the intensity of the different adoption patterns has a varying effect on the average decrease in F1-score. For the Heavy-sensitive adoption pattern, the High protection intensity has the smallest decrease while the effect for Low and Moderate intensity is relatively similar. For Light sensitive adoption pattern, the Low protection intensity resulted in the smallest decrease in performance while the moderate intensity still resulted in the greatest decrease. Lastly, for the Uniform adoption pattern, the ascending order of decrease in the performance is High, Low followed by Moderate protection intensity.

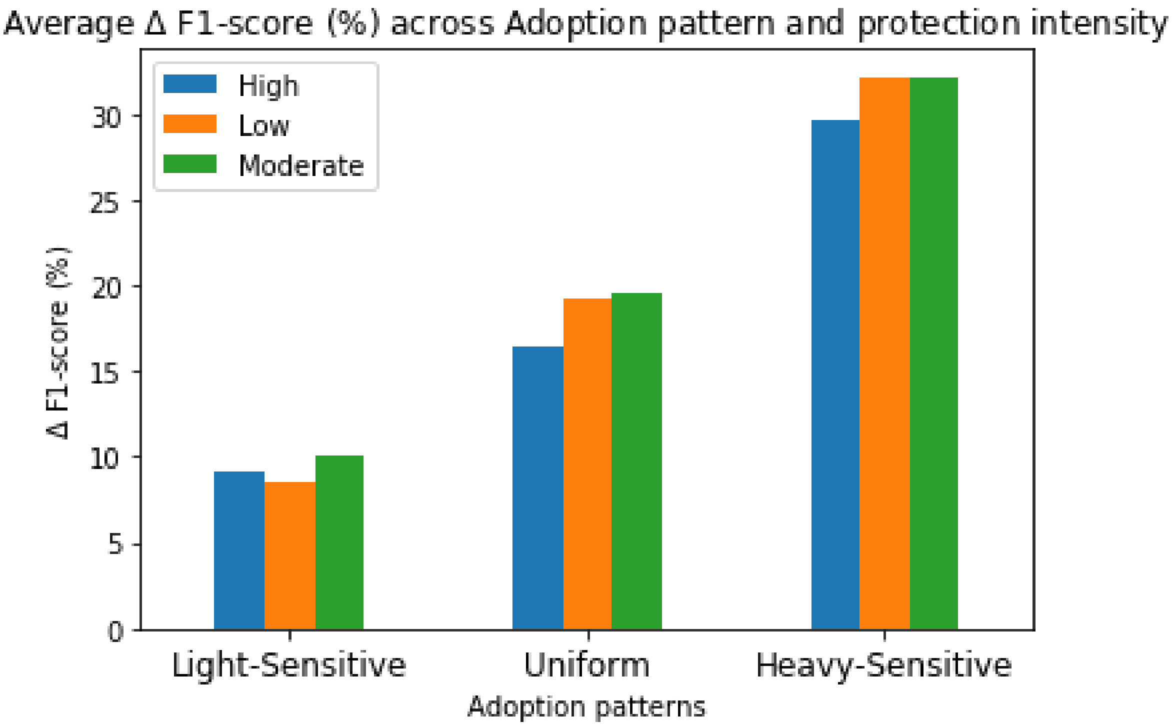


Figure 14: Average decrease in F1-score across adoption pattern and protection intensity

4.2.4 All three factors combined

The simulation results have shown that as the adoption rate increases, the model's performance of predicting customer purchases decreases as well. The adoption pattern has a greater impact than adoption rates. When comparing across the Uniform adoption pattern, the decrease in performance is relatively linear. For the Heavy-sensitive pattern, the decrease in performance is concave for low intensity as the adoption rate increases while for moderate and high intensity it exhibits concave at lower adoption rates and convex at higher adoption rates. For Light-sensitive, the decrease in performance is concave for all 3 types of intensities. Among the 3 different adoption patterns, Heavy sensitive has the largest decrease in performance, while Light sensitive has the lowest decrease in performance.

As seen from the distribution of clicks from Figure 5 above, we know that majority of the users are infrequent users (the most frequent number of clicks is 2), it explains why for Light sensitive adoption pattern (last row of Figure 15 below), sessions selected as PET adopters are less likely to be broken down into further smaller sessions as they have very little sessions to being with, resulting in the smallest decrease in percentage change of F1-score. At higher adoption rates where most of the infrequent users are already adopting, the newer adopters will be those heavy users resulting in their sessions being broken into smaller sessions where each session has fewer clicks. Thus, this explains a concave pattern as seen in all the 3 intensities across the last row of the results (Figure 15) where the decrease in performance spike after 80% adoption rates.

For the Uniform adoption pattern, with every session having the same probability of being selected as an adopter of PET regardless of the number of clicks it has, a linear trend is observed for the deterioration of the model's performance as the adoption rate increases which is aligned with intuition as the proportion of consumers adopting PETs increase, the data degradation will be more severe leading to a decrease in performance in a relatively constant rate. However, such deterioration is lower for High intensity where the cut-off duration for each session is shorter.

For Heavy Sensitive pattern, at Low intensity it exhibits a concave shape which may be attributed to the fact that there are not many heavy users to begin with, thus at the lower adoption rates, most of the adopters will be frequent users leading to larger sessions being broken down into smaller sessions cause a higher degree of data degradation. Subsequently, as the adoption rate increase, the remaining adopters are infrequent users where their sessions are less likely to be broken down into smaller sessions due to the lower intensity as well as lesser number of clicks to begin with, leading to the data degradation problem being less severe thus the performance of the model is not as greatly affected.

However, at moderate or high intensity, when the adoption rates increase, we see a larger decrease in performance which maybe be attributed to those infrequent users' sessions having relatively longer dwell time (From Figure 6) leading to their sessions being broken into small sessions with a higher likelihood which results in the model's performance to decrease much more significantly as reflected in Figure 15 below.

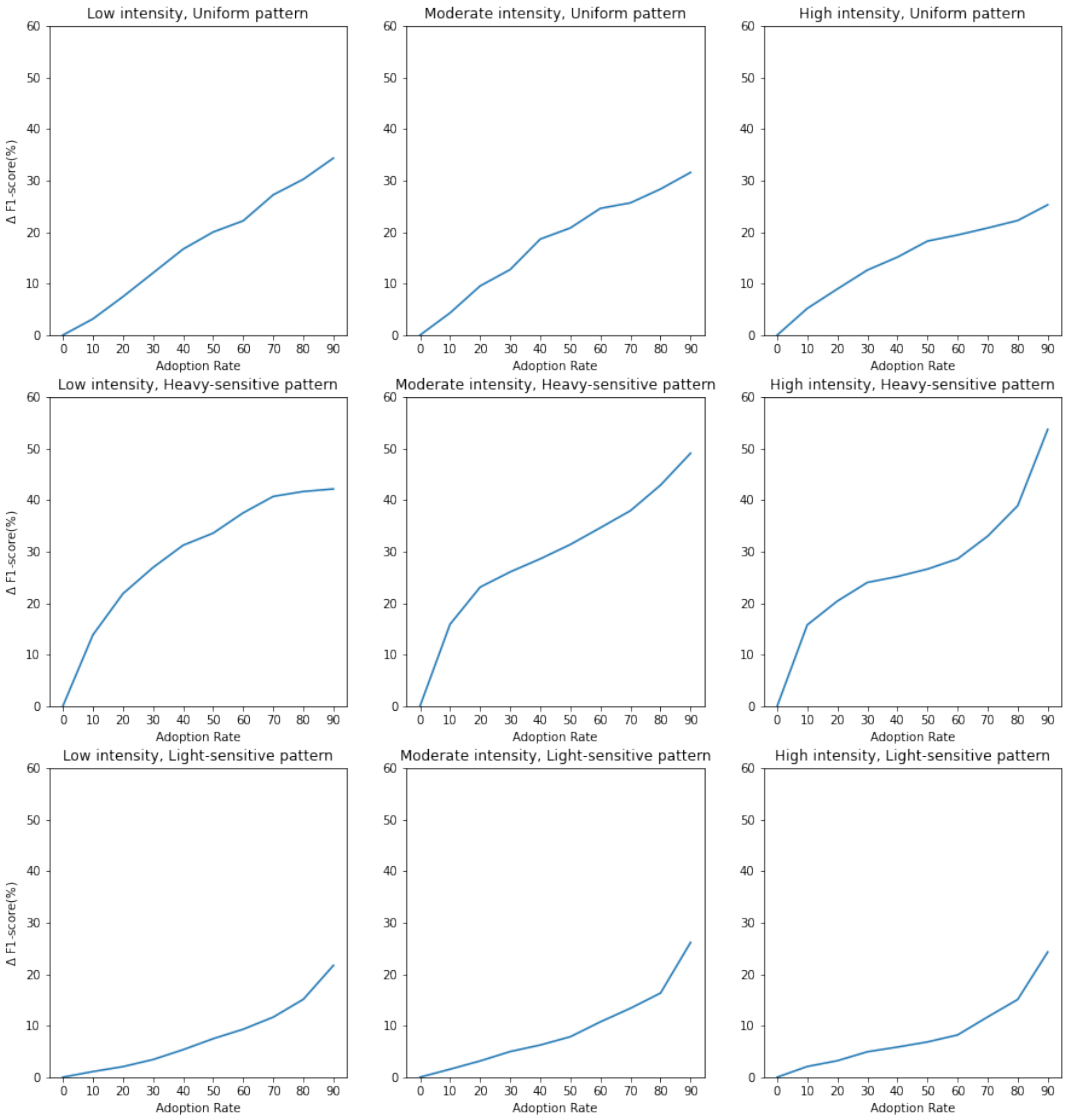


Figure 15: Main results of simulation (Δ F1-score - %)

4.2.5 Regression analysis

The regression result from Table 6 is consistent with the findings above. Adoption rates are scaled from 0 to 1 while dummy variables for High Intensity, Moderate intensity, uniform adoption pattern and heavy-sensitive adoption pattern are generated. The dependent variable for this regression is the percentage decrease in F1-score for consistency with the results from Figure 15. The results have shown that higher protection intensity has a direct impact on the percentage decrease in F1-score, the same can be observed for the Uniform adoption pattern and Heavy-sensitive adoption pattern compared to the Light-sensitive adoption pattern.

Variable	Coefficient	Standard error	P – value
Adoption Rate	35.6959	0.577	0.000
Intensity High	1.0515	0.447	0.021
Intensity Moderate	0.9790	0.439	0.029
Pattern Uniform	0.9790	0.439	0.029
Pattern Heavy Sensitive	1.0515	0.447	0.021

Table 6: Regression results

5. **Conclusion**

5.1 Summary

We have managed to find out the main objective of this research, (how the adoption of PETs affects firms' analytical performance). Fundamentally, the adoption of end-user PETs will adversely affect firms' analytical performance. However, the extent of impact varies depending on the 3 factors (adoption rate, adoption pattern & protection intensities) with some differences when looking at their direct and mixed effects.

From the results shown above, we can see that firms should be concerned with frequent users adopting PETs as it results in the largest decrease in analytical performance in this context of purchase prediction. While at lower adoption rates and lower intensities the decrease in performance is not that large (< 20% decrease for adoption rates below 50%), firms should still try to find ways to reduce the number of adopters through increasing personalisation values such as membership discounts which may scale with usage as firms' main concerns are from those frequent users. This conclusion is consistent with prior work that states that firms should be more worried about frequent users being adopters as they provide most of the data for the firm (Chen. & Hahn, 2020, pp. 13).

This study provided a summary of how cookies function and collect information from both theoretical and practical aspects, along with some qualitative explanations for the impact of erasing cookies for websites. It also generalises the response firms should adopt to mitigate the impact of users adopting PETs by ensuring the frequent users are not part of the adopting population.

5.2 Limitations

In this study, we have only looked at how the data problem induced by measurement errors affects firms' analytical performance through the usage of cookie erasers by splitting up sessions at fixed intervals simulating HTTP cookies being erased. There are also other sophisticated ways that cookies may be erased or to make sessions untraceable that have not been explored in this paper.

Furthermore, this study did not explore much about the data problem introduced by missing values as it is not viable to examine it given the nature of the dataset where minimal information is captured without any demographic details.

5.3 Recommendations for Future Work

From the limitations mentioned above, future work may explore how missing values can affect firms' analytic performance in the context of a classification task when an appropriate dataset is available. At the same time, factors that may affect how the users adopt PETs apart from their behaviour (adoption pattern) can be explored to provide a more comprehensive analysis, for instance, their demographic information (e.g. Education level, age, income, etc.).

6. **References**

Abraham, M., Meierhoefer, C., & Lipsman, A. (2007). The impact of cookie deletion on the accuracy of site-server and ad-server metrics: An empirical comscore study. Retrieved October, 14, 2009.

Cahn, A., Alfeld, S., Barford, P., & Muthukrishnan, S. (2016, April). An empirical study of web cookies. In Proceedings of the 25th international conference on world wide web (pp. 891-901).

Chen, Dawei and Hahn, Jungpil, "Impact of End-User Privacy Enhancing Technologies (PETs) on Firms'Analytics Performance" (2020). ICIS 2020 Proceedings. 3.https://aisel.aisnet.org/icis2020/digital_commerce/digital_commerce/3

Dorogush, A. V., Ershov, V., &amp; Gulin, A. (2018, October 24). CatBoost: Gradient boosting with categorical features support. arXiv.org. https://arxiv.org/abs/1810.11363v1.

Englehardt, S., Reisman, D., Eubank, C., Zimmerman, P., Mayer, J., Narayanan, A., & Felten, E. W. (2015, May). Cookies that give you away: The surveillance implications of web tracking. In Proceedings of the 24th International Conference on World Wide Web (pp. 289-299).

Erevelles, S., Fukawa, N., & Swayne, L. (2016). Big Data consumer analytics and the transformation of marketing. Journal of business research, 69(2), 897-904.

Kevin Lewis, Jason Kaufman, Nicholas Christakis (2008) Journal of Computer-Mediated Communication, Volume 14, Issue 1, 1 October 2008, Pages 79–100, https://doi.org/10.1111/j.1083-6101.2008.01432.x

Li, A. H. F. (2017). E-commerce and Taobao Village. China Perspectives, 03.

Mokryn, O., Bogina, V., & Kuflik, T. (2019). Will this session end with a purchase? Inferring current purchase intent of anonymous visitors. Electronic Commerce Research and Applications, 34, 100836.

Qiu, J., Lin, Z., & Li, Y. (2015). Predicting customer purchase behavior in the e-commerce context. Electronic commerce research, 15(4), 427-452.

Van Blarkom, G. W., Borking, J. J., and Olk, J. E. 2003. “PET,” in Handbook of Privacy and Privacy- Enhancing Technologies - The Case of Intelligent Software Agent, G. W. van Blarkom, J. J. Broking, and J. G. E. Olk (Eds.), The Hague: College bescherming persoonsgegevens, pp. 33-53.

Wang, P., & Petrison, L. A. (1993). Direct marketing activities and personal privacy: A consumer survey. Journal of Direct Marketing, 7(1), 7-19.

Wei, D., Geng, P., Ying, L., & Shuaipeng, L. (2014, May). A prediction study on e-commerce sales based on structure time series model and web search data. In The 26th Chinese Control and Decision Conference (2014 CCDC) (pp. 5346-5351). IEEE.

Yi, X., Hong, L., Zhong, E., Liu, N.N., Rajan, S., (2014) RecSys '14: Proceedings of the 8th ACM Conference on Recommender systems October 2014 Pages 113–120 https://doi.org/10.1145/2645710.2645724

Zimmerman, R. K. (2000). The way the cookies crumble: internet privacy and data protection in the twenty-first century. NYUJ Legis. & Pub. Pol'y, 4, 439.

Appendix A – Alternative metrics

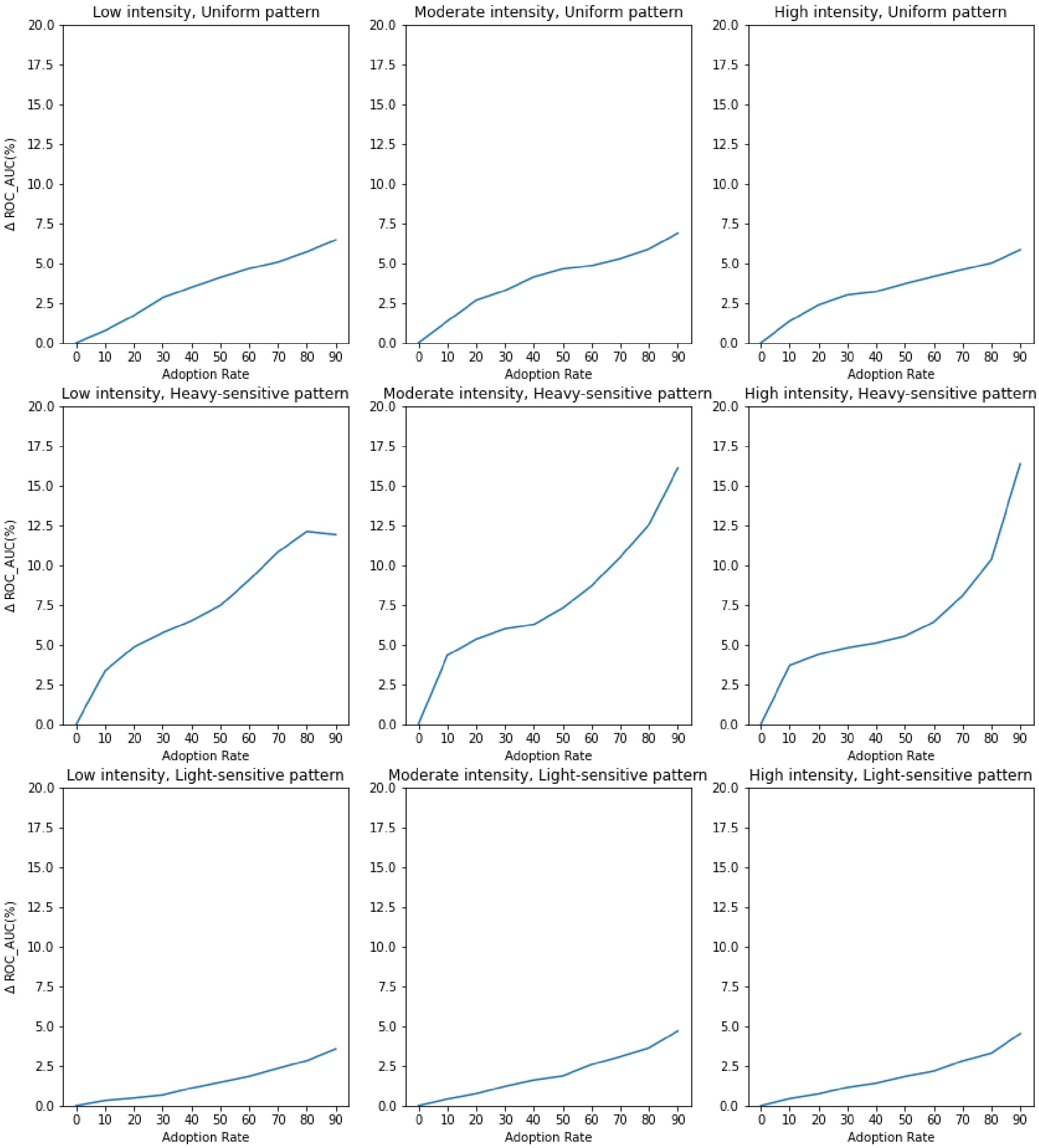


Figure 16: Results of simulation (Δ ROC_AUC - %)

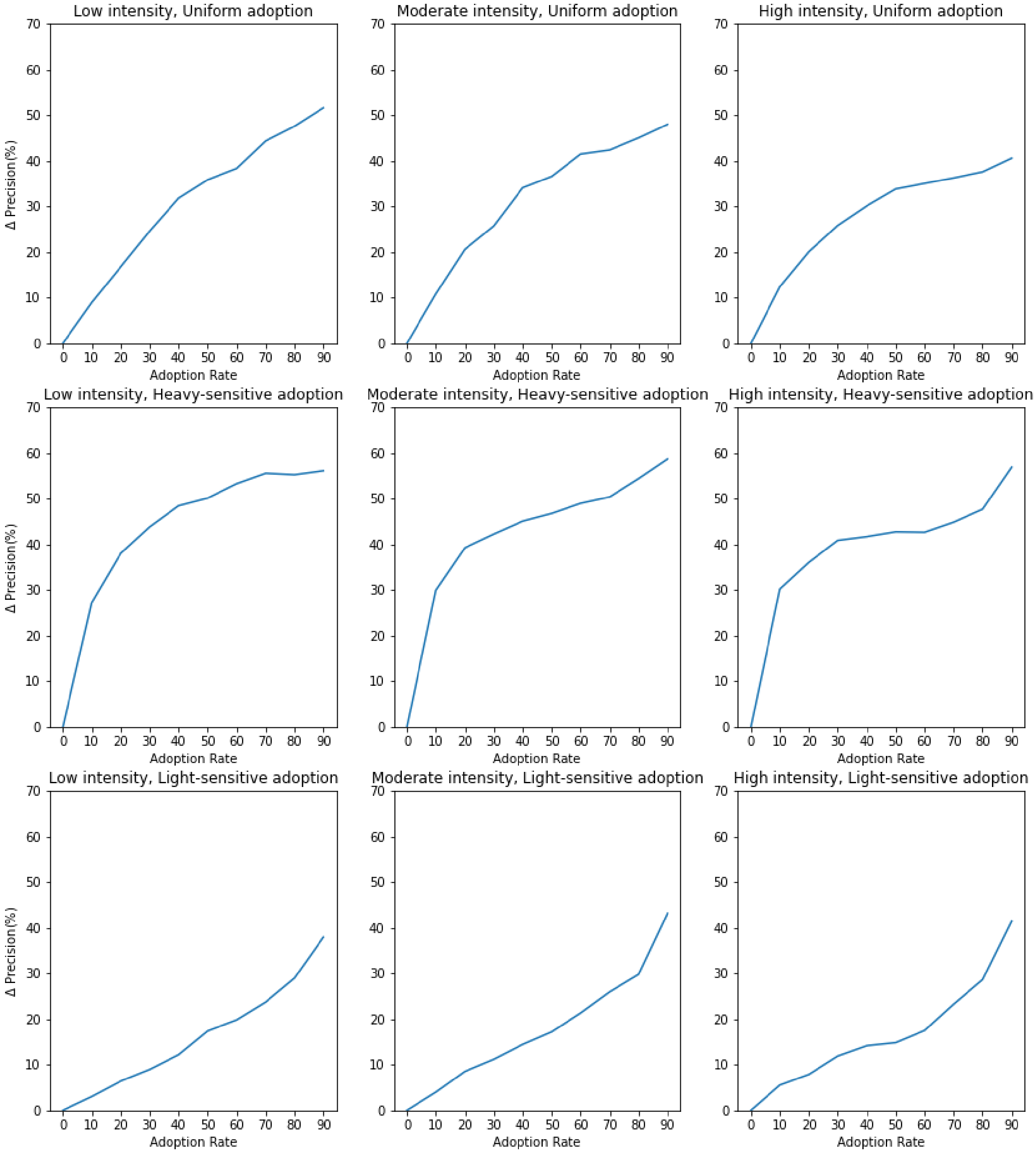


Figure 17: Results of simulation (Δ Precision - %)

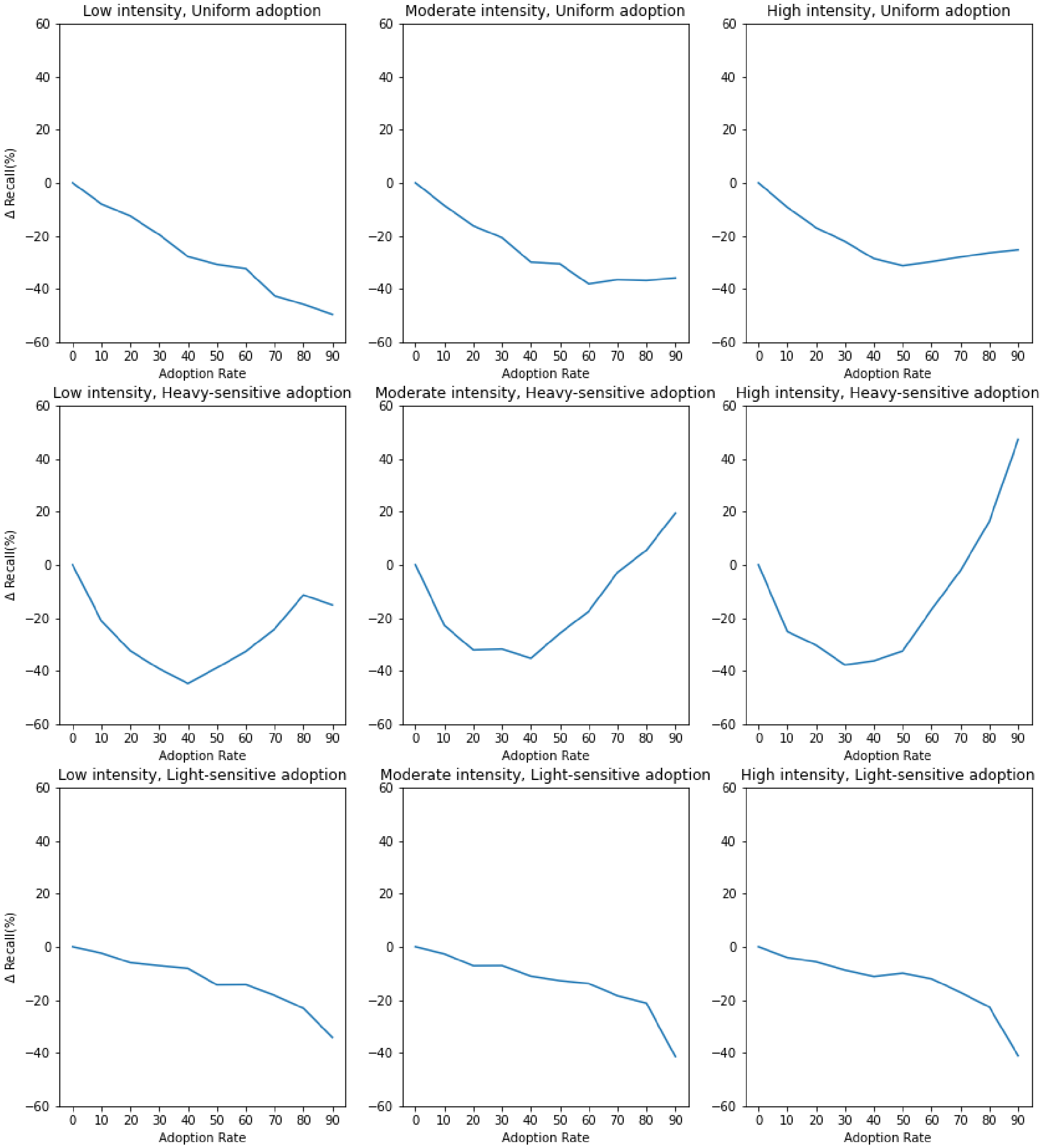


Figure 18: Results of simulation (Δ Recall - %)