Investigation of 3-D Secure's Model for Fraud Detection

Mohammed Aamir Ali
Newcastle University
Newcastle upon Tyne, UK

Thomas Groß
Newcastle University
Newcastle upon Tyne, UK

Aad van Moorsel
Newcastle University
Newcastle upon Tyne, UK

ABSTRACT

Background. 3-D Secure 2.0 (3DS 2.0) is an identity federation protocol authenticating the payment initiator for credit card transactions on the Web.
Aim. We aim to quantify the impact of factors used by 3DS 2.0 in its fraud-detection decision making process.
Method. We ran credit card transactions with two Web sites systematically manipulating the nominal IVs machine_data, value, region, and website. We measured whether the user was challenged with an authentication, whether the transaction was declined, and whether the card was blocked as nominal DVs.
Results. While website and card largely did not show a significant impact on any outcome, machine_data, value and region did.

A change in machine_data, region or value made it 5-7 times as likely to be challenged with password authentication. However, even in a foreign region with another factor being changed, the overall likelihood of being challenged only reached 60%.

When in the card's home region, a transaction will be rarely declined (< 5% in control, 40% with one factor changed). However, in a region foreign to the card the system will more likely decline transactions anyway (about 60%) and any change in machine_data or value will lead to a near-certain declined transaction.

The region was the only significant predictor for a card being blocked (OR = 3).
Conclusions. We found that the decisions to challenge the user with a password authentication, to decline a transaction and to block a card are governed by different weightings. 3DS 2.0 is most likely to decline transactions, especially in a foreign region. It is less likely to challenge users with password authentication, even if machine_data or value are changed.

CCS CONCEPTS

• **Security and privacy** → **Authentication**; **Human and societal aspects of security and privacy**; • **Human-centered computing** → Empirical studies in HCI;

KEYWORDS

3-D Secure, Authentication, Fraud Detection

STAST 2018, December 4, 2018, San Juan, PR, USA

ACM 978-1-4503-7285-5/18/12…$15.00
https://doi.org/10.1145/3361331.3361334

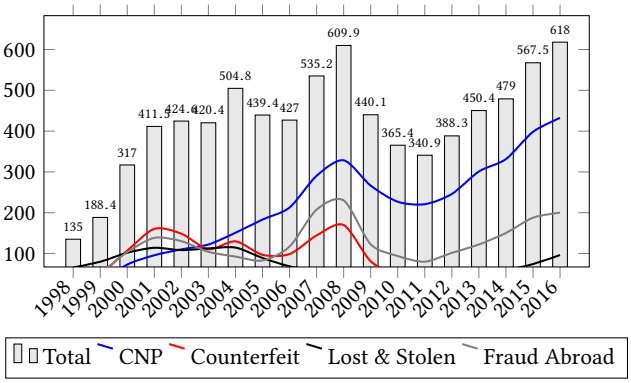


Figure 1: UK Card Fraud by Type from year 1998 to 2016.

ACM Reference format:
Mohammed Aamir Ali, Thomas Groß, and Aad van Moorsel. 2018. Investigation of 3-D Secure's Model for Fraud Detection. In *Proceedings of 8th International Workshop on Socio-Technical Aspects in Security and Trust, San Juan, Puerto Rico, USA, December 4, 2018 (STAST2018),* 11 pages.
https://doi.org/10.1145/3361331.3361334

1 INTRODUCTION

Electronic commerce (e-commerce) is a mainstay of today's Internet, allowing users to buy or sell goods online. In payment systems terminology, e-commerce payments are known as Card Not Present (CNP) payments because the cardholder is not physically present at the merchant. CNP payment sales have shown a significant growth year-by-year. For example, in the UK, sales has been recorded a total of £154 billion for 2017 [3]. This is 18% of increase in the online spending by customers when compared to year 2014 [3].

The convenience of enabling purchases online comes at a price. The system is also prone to attract cyber offenders. Shown in 1, are the UK card payment fraud statistics from year 1998 to 2016. It can be seen from the figure that the payment industry is effective in mitigating card present types of payment card frauds. However, CNP payments fraud has reached its highest mark accounting for 70% of the total card fraud, causing £432.3 million loss exclusively on the UK issued cards [8]. This development has called for a complex fraud detection system to be integrated with the protocol flows of the CNP payment system.

In general, the CNP payment system requires the payment initiator (customer) to enter their payment card information on the checkout page provided by the merchant website. The merchant collects the card information, combines it with the transactions information and forwards it to the card issuing bank for authorization. During the authorization process, the card issuer decides whether to approve or decline the transaction.

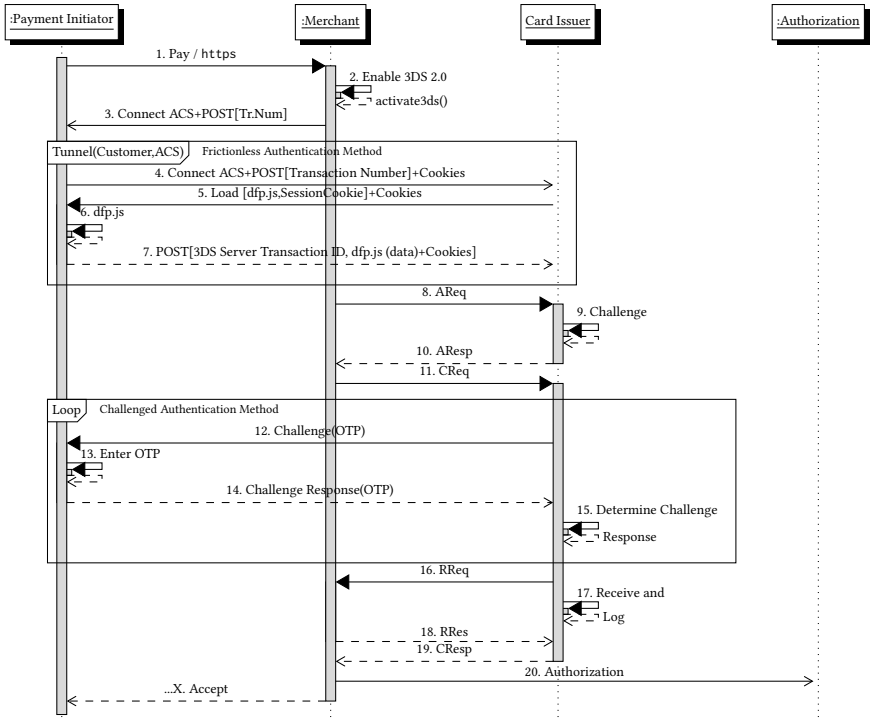


Figure 2: Actions and parties involved in a 3DS 2.0 transaction process

Given that the payment card details are static and are shared with every online merchant, there is a significant risk of the card data leaking and being used in fraudulent transactions. Once the payment card details leaves the payment initiator's device, there is no guarantee that the card details are handled securely by the merchants. This is also reflected by recent attacks on Ticketmaster [2] and British Airways [1, 5] and hundreds of websites where millions of card details were compromised [11]. Such systems lacks in the verification of the payment initiator as the valid owner of the card. Hence, the CNP payment system in itself is based on static card information and, as such, inherently insecure.

To protect the CNP payment system from fraud, VisaInc [10] introduced 3-D Secure 1.0 (3DS 1.0) [9] in 2001. 3-D Secure introduced the concept of payment initiator authentication for CNP payments. 3DS 1.0 redirects CNP transactions from each merchant website to the card issuer so that the payment initiator can be authenticated as the valid owner of the card.

With criticisms voiced on 3DS 1.0's registration and password authentication [6], frictions in the checkout [6], and the steady increase in CNP payment fraud, especially through phishing [6, 7], there was a need for a payment protocol upgrade. In 2016, EMVCo—a consortium of card payment networks—developed the 3D Secure 2.0 (3DS 2.0) [4] to address the requirements of stronger customer authentication yet maintaining the convenience requirements on a merchant checkout page. With 3DS 2.0, the card issuer performs fraud risk assessment for each transaction and authenticates the payment initiator with either of the two schemes: challenged and frictionless. Challenged authentication is designed for higher risks transactions and requires the payment initiator to authenticate him/herself with one-time pass codes sent by the card issuer to the payment initiator's registered device [4]. Frictionless authentication is for purchases with lower risk of fraud and relies on the browser configuration details (hereafter referred to as browser fingerprint) extracted for the payment initiator device during the checkout process [4]. At the same time, the decision making process of 3DS 2.0 to perform fraud risk assessment and the decision for challenged or frictionless authentication is shrouded from and often obscure to the consumers.

In this paper, we quantify the impact of factors used by 3DS 2.0 in its fraud-detection decision making process. That is, we aim at establishing to what extent a change in a factor changes the likelihood of a 3DS 2.0 decision outcome, e.g., whether the authentication is made challenged or frictionless. We run transactions with two Web sites manipulating Independent Variables (IV's) which includes machine_data captured from a user Web browser (WB), transaction value, region and websites. To manipulate machine_data, we set-up an HTTP proxy in the machine used to initiate transactions on 3DS 2.0 website. We measure whether the payment initiator was challenged with an authentication, whether the transaction was accepted with frictionless authentication or declined, and whether the card was blocked. We employ logistic regressions to quantify the change of likelihood observed from changes in the variables we have manipulated and, thereby, shine a light on the 3DS 2.0 fraud decision making process in the backend.

This paper is organized as follows. Section 2 presents an overview of the 3DS 2.0 transaction process and provides an introduction into how the transaction risk assessment decisions are made by the card issuer. The paper follows an empirical-methods standard structure thereafter, describing the method first (Section 4), establishing the core results of the analysis without further interpretation (Section 5), and finally analyzing the results in a discussion (Section 6). We draw attention to the logistic regression plots on pp. 10 and 11 as main tools to interpret the results.

2 OVERVIEW OF A 3DS 2.0 TRANSACTION PROCESS

Figure 2 shows actions and parties involved in a 3DS 2.0 transaction process. The process starts with the payment initiator filling their payment card details on the checkout page provided by the merchant web site. When the "Pay" button is clicked, the merchant web server hosting the 3DS 2.0 plugin generates a unique tranasaction ID and connects the payment initiator's session to the card issuer. As shown in step 4, the card issuer connects to the payment initiator's Web Browser (WB) and sends device fingerprinting JavaScript (dfp.js) programmed to fetch browser and operating system details. The JavaScript mainly includes the following methods:

- deviceprint_browser(): This method extracts information about payment initiator's (WB) and operating system including: browser name, major and minor version, languages supported, languages installed, operating system name, operating system version, and operating system platform (Win32 or Win64).
- deviceprint_display(): This method captures detailed screen information including colour depth, screen width, height, available height, buffer depth, and pixel depth.
- deviceprint_software(): captures (WB's) plugins and their types. The method also has logic to extract browser's tracking and advertisement preferences as provided by DoNotTrack and Useofadblock.
- deviceprint_java(): is used to test if the payment initiator browser supports Java or not.
- cookies(): is used to test if cookies are enabled by the user WB.

The information collected from the above methods is combined into a single string and is encoded into base-64 plain text (as defined by the 3DS 2.0 protocol specifications [4]) before being sent as a form element to the card issuer. It is likely that the card issuer uses IP address as an indicator to extract payment initiator machine location but it is captured differently.

In step 8, the merchant frames an Authentication Request (AReq) which is forwarded to the appropriate card issuer Access Control Server (ACS). The ACS manages 3DS 2.0 authentication request/response messages. The AReq contains card data provided by the payment initiator, merchant account information and other transaction related information. The card issuer collates the transaction information from the merchant and WB details provided by device fingerprinting scrips and performs fraud risk assessment (FRA) on the given transaction. Based on the outcome of FRA, the card issuer decides whether to challenge the payment initiator with a one-time pass codes or to authenticate the payment initiator with frictionless authentication. For the transaction shown in 2, the card issuer decides to have challenged authentication.

In step 10, the issuer through the Authentication Response (ARes) message responds back to the merchant indicating that the challenge is required to further process the transaction. For frictionless authentication, ARes will indicate a successful authentication.

The merchant initiates a Challenge Request (CReq) message and posts it to the card issuer. The issuer sends a challenge user interface (UI) to the payment initiator's WB. The UI is an interaction platform where the card issuer can interact with the payment initiator to obtain challenge response. At the point, the card issuer prompts challenge or OTP on the payment initiator's registered device (mobile phone for example).

The payment initiator enters the OTP on the 3DS 2.0 interface and uponn successful authentication, the issuer determines the payment initiator as appropriate owner of the card and formats the Results Request (RReq) message with a cryptographic hash which is forwarded to the merchant. The RReq and the hash is later used by the Authorization network to verify the integrity of authentication messages. To acknowledge the receipt of the RReq, the merchant prepares the Results Response (RRes) and forwards it to the issuer. Finally, the issuer formats the Challenge Response (CRes) message and shuttles it back to the merchant. The CRes indicates the completion of challenged authentication. It is to be noted that the CReq and CRes messages are only applicable to challenged 3DS2.0 transaction.

3 AIMS

RQ 1 (Impact of predictors on authentication outcomes). *Which factors impact the fraud-detection decisions with what magnitude of change in acceptance likelihood?*

Table 1 gives an overview of the operationalization of this research question. As nominal independent variables (IV) we have machine_data, value, region, and website.

As nominal dependent variables we consider whether the user was challenged with a password authentication, whether the transaction was declined[1] and whether the card was blocked.

Iterating over the independent variables $X \in \{$ machine_data, value, region, website $\}$ and the dependent variables $Y \in \{$ challenged, declined, blocked $\}$, we consider the following statistical hypotheses:

Alternative Hypotheses. $H_{1,X,Y}$: The independent variable X systematically impacts the likelihood of a change in the dependent variable Y.

Null Hypotheses. $H_{0,X,Y}$: The independent variable X does not yield an impact on the likelihood of change in the dependent variable Y.

Note that we, thereby, investigate 5×3 relations with corresponding alternative and null hypotheses. Our main interest lays in the IVs { machine_data, value, region }.

Logistic Regression Classifier. We use logistic regressions to establish the magnitude of impact on the likelihood on change in the response variable.

[1] In the pre-registration of the experiment, the IV declined was called transaction_status.

Table 1: Operationalization.

Variable	Levels
IV: machine_data	0 := intact
	1 := overwritten
IV: value	0 := low ($13)
	1 := high ($406)
IV: region	0 := credit card home region (UK)
	1 := foreign region (Germany)
DV: challenged	0 := passed (User passed without password authentication)
	1 := challenged (User was challenged with password authentication)
DV: declined	0 := accepted (Transaction was accepted)
	1 := declined (Transaction was declined)
DV: blocked	0 := continued (Credit card continued to be active)
	1 := blocked (Credit card was blocked by the bank)

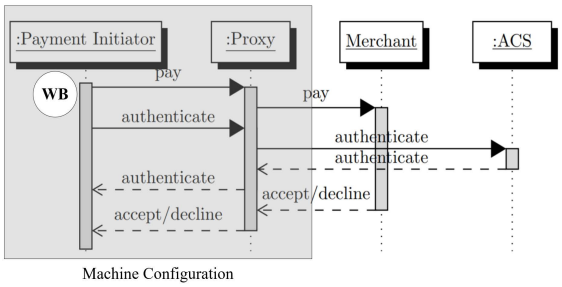


Figure 3: Reverse engineering set-up, intercepting 3DS 2.0 transactions through a proxy.

4 METHOD

The study—its statistical hypotheses and analysis plan—have been preregistered at the Open Science Framework (OSF)[2] prior to any statistical analysis. Analyses, graphs and statistical reporting in this paper were computed directly from the data using the R package knitr. The OSF repository includes the dataset and its Datacite 4.0 meta-data description.

4.1 Sampling

In a repeated-measures experiment, four different payment cards (three Visa and a MasterCard) were used to make Card Not Present (CNP) payment transactions. We, thereby, sampled CNP payment transactions with 3DS-2.0-enabled home appliance Web sites (specifically: argos.co.uk and bmstores.co.uk). The sampling frame was created by enumerating combinations of cards and Web sites, which were then exposed to different conditions.

4.2 Procedure

We manipulated the three IVs (machine_data, value, region). To manipulate the machine_data, we intercepted the communication between the payment initiator's browser WB, the merchant and the card issuer. We achieve this by placing Fiddler on the payment initiator's device (i.e., our own machine). Fiddler allows us to add breakpoint to alter the data before it is forwarded from WB to the communicating server. Using this platform, we are able (1) to sniff the communication, (2) to control the input to WB and (3) to control the output from WB.

4.2.1 *Manipulation.*

Machine Data. To alter the machine_data, we add two breakpoints. First, when the payment initiator click the 'Pay' button on the merchant website. This is to modify the HTTP headers flowing from WB to the merchant. The second breakpoint we add is when card issuer connects to WB to fetch the browser fingerprint, as shown in Figure 3. This is to change the machine_data. We alter the HTTP headers and the base-64 string of WB device fingerprint with that of recorded by Fiddler from a machine with different browser fingerprint.

Value. To change the value of the transaction, we selected and purchased items that either cost $13 or $406.

Region. We kept the transaction either in a region local to the country of where credit card is issued (UK) or a region foreign to the credit card where the transactions were made from Germany.

4.2.2 *Measurement.*

We coded the outcomes of transactions on a nominal scale, either as '0' or '1', depending on whether the user could proceed with the transaction or was interrupted. This outcome was obtained from the response of the 3DS protocol to the browser.

We classified interruptions manually, based on (1) whether the user was challenged with a password authentication (challenged), (2) whether the card transaction was declined (declined), and (3) whether the payment card was blocked altogether (blocked).

4.3 Ethics

The experiment was run in accordance with the requirements of the institution's ethical review board and an ethics case signed off.

The payment cards used belonged to one of the experimenters, who exercised informed consent in volunteering the cards for the experiment. The card holder was aware that repeated transactions as done in this experiment may impact future payment behavior of the card.

The card transactions were made by the card holder, and the relevant personal identifiable information not stored outside of the holder's control.

[2]DOI 10.17605/OSF.IO/X6YFH; https://osf.io/x6yfh/

The transactions were done manually, over a longer timeframe, and restricted to at most 100 transactions, thereby, limiting the impact on the fraud detection efforts of providers, banks and 3DS.

5 RESULTS

The statistics were computed with a significance level of $\alpha = .05$. As a common approach, we conducted binomial logistic regressions with the dependent variables as response and a target model including all independent variables.

5.1 Common Analysis Approach

For each dependent variable, we created a logistic regression that is to quantify the change in likelihood caused by the different predictors.

Model Significance and Fit. The first question is, whether a selected model is a valid and well-fitting model, at all. We checked overall model significance with the Wald test, checked for significant higher-level interactions, and selected the final model using the Akaike Information Criterion corrected for small samples (AIC_c). Thereby, we ascertained that the selected model contains substantial evidence vis-à-vis of the minimal-AIC model and substantiated its suitability with a goodness-of-fit check.

While we aimed for a full model with all predictors as specified in the pre-registration, we checked that the model is actually defensible. While this was the case for the models on the user being challenged and on the transaction being declined, we found that for the card being blocked, there was not enough evidence to vouch for the full model. Here, we have selected a model only with the core predictors, as an alternative.

Impact of Predictors. We computed odds ratios and 95% confidence intervals thereon for effects of significant predictors. Odds ratios are an effect size of choice for logistic regressions. They quantify the *multiplicative* change in likelihood of the of the outcome, given a change in one predictor and everything else being equal. We may say, "Other predictors held constant, a change in transaction value makes a rejection five times as likely."

It is important to note that this multiplicative change is with respect to a baseline specific for each model. Hence, similar odds ratios in different models might lead to different absolute likelihoods of outcomes, given selected interventions.

Scenario Probabilities. We also discuss the absolute likelihoods of outcomes for particular scenarios. Hence, then we factor in the odds ratios of active predictors and obtain the overall likelihood in that situation. In this case, we consider combinations of predictors being manipulated and offer a likelihood estimate for the outcome. Here we may say that "In a foreign region, the transaction is 99% likely to be declined if the machine data is faulty."

Model Evaluation. Finally, we evaluated each model with regression diagnostics as well as accuracy (prediction vs. observation). We computed repeated 10-fold cross-validations with the R package caret (with 10 repetitions).

We report the results of the model evaluation in Appendix A.

Table 2: Logistic Regression: User Challenged

	Estimate	SE	z-value	p-Value
(Intercept)	-2.981	1.021	-2.921	.003**
Machine.Data	1.893	0.727	2.602	.009**
Value	1.498	0.705	2.125	.034*
Region	1.893	0.727	2.602	.009**
Website	-0.219	0.663	-0.330	.741
Card	-0.563	0.312	-1.805	.071

Note: Overall Model: Wald $\chi^2(5) = 21.593$, $p < .001$
R^2= .28 (Hosmer & Lemeshow), .29 (Cox & Snell), .41 (Nagelkerke)

5.2 Logistic Regression: User Challenged

We computed a binomial logistic regression to test whether the independent variables impact the likelihood of challenged as response. Appendix A.1 contains details of the model selection and evaluation.

5.2.1 *Fitted Model.*

Predictors and Odds Ratios. Table 2 offers an overview estimates of the selected model.

There was a statistically significant positive impact of overwriting the machine data on the user being challenged with a password, $z = 2.6$, $p = .009$, OR = 6.64, 95% CI [1.7, 31.7]. Everything else being equal, a user whose machine data is corrupted is 6.6 times as likely to be challenged with a password authentication.

A change of the value of a transaction from low to high had a statistically significant effect on the user being challenged, $z = 2.12$, $p = .034$, OR = 4.47, 95% CI [1.2, 19.9]. The increase in value from \$13 to \$406 made it 4.5 times as likely to be challenged.

The region being changed to a foreign country had a statistically significant impact on being challenged with a password authentication, $z = 2.6$, $p = .009$, OR = 6.64, 95% CI [1.7, 31.7]. A change in region to Germany made it 6.6 times as likely.

Given these results we reject the null hypotheses $H_{0,X,\text{challenged}}$ for $X \in \{$ machine_data, value, region $\}$.

Scenario Probabilities. We display the response plots of the different predictor variables in Fig. 4 on p. 10. Note that given the similar odds ratios of the predictors, we expect the response and overlay plots below to look rather similar to one another.

In Fig. 5, we overlay by region the likelihoods of changing machine data or value, respectively. Overall, we observe that the probability of being challenged while being in home region of the card is less than 5% when neither machine data nor value are manipulated. Should either of the two predictors be changed (machine data overwritten or the value high), the probability of being challenged is less than 20%.

If the card does a transaction from the foreign region, the situation is quite different. Here, the card will challenge the user at probability of 20% or 25% even if machine data are intact and the value low. Should either of the two variables be manipulated, the user will be challenged at a probability of around 60%.

5.2.2 *Model Evaluation.*

Table 3: Logistic Regression: Transaction Declined

	Estimate	SE	z-value	p-Value
(Intercept)	-6.333	1.709	-3.706	<.001***
Machine.Data	2.944	0.944	3.119	.002**
Value	3.397	0.996	3.410	<.001***
Region	3.397	0.996	3.410	<.001***
Website	0.285	0.757	0.376	.707
Card	0.975	0.398	2.449	.014*

Note: Overall Model: Wald $\chi^2(5) = 44.409$, $p < .001$
R^2= .50 (Hosmer & Lemeshow), .50 (Cox & Snell), .67 (Nagelkerke)

Performance. Having computed a observation-vs.-prediction classification, we found that the regression had an accuracy of 73%, Hosmer-Lemeshow not rejecting the fit, HL_C $\chi^2(8) = 10.77$, $p = .215$.

Cross-Validation. We computed a repeated 10-fold cross-validation on the same dataset. This means, that the dataset was partitioned into 10 parts, that the model was then re-computed using 9 parts as training data (with $N_T = 57 \pm 1$), and used to predict the observations of the 10-th part.

The cross-validation yielded an accuracy of 70%, 95% CI [66%, 74%]. With a Cohen's $\kappa = .23$, we consider the cross-validation accuracy as low.

5.3 Logistic Regression: Transaction Declined

We computed a binomial logistic regression with the independent variables as predictors and the transaction being declined as response variable. We report on the model evaluation in Appendix A.2.

5.3.1 *Fitted Model.*

Predictors and Odds Ratios. We are offering an overview of all estimates in the regression Table 3.

Overwriting the machine data has a statistically significant impact on the transaction being declined, $z = 3.12$, $p = .002$, OR = 18.99, 95% CI [3.7, 162.2]. Everything else being equal, overwriting the machine data made it 19 times as likely to get the transaction declined.

There was a statistically significant effect of the value of the transaction on it being declined, $z = 3.41, p < .001$, OR = 29.88, 95% CI [5.4, 292.9]. Other predictors held constant, increasing the value to $406 made it 29.9 times as likely to have transaction declined.

The region had a statistically significant effent on the transaction being declined, $z = 3.41$, $p < .001$, OR = 29.88, 95% CI [5.4, 292.9]. Other predictors constant, a change to the foreign region (Germany) made it 29.9 times as likely to have transaction declined.

We thereby reject the null hypotheses $H_{0,X,\text{declined}}$ for $X \in \{$ machine_data, value, region $\}$.

In addition to these predictors, the card used also had a statistically significant effect on the transaction being declined.

Scenario Probabilities. We are giving an overview of regression (response and overlay) graphs for the transaction-declined regression in Figures 6 and 7 on p. 11.

Table 4: Logistic Regression: Card Blocked

	Estimate	SE	z-value	p-Value
(Intercept)	-22.798	2855.831	-0.008	.994
Value	20.194	2855.830	0.007	.994
Region	2.327	0.951	2.447	.014*
Machine.Data	1.096	0.888	1.234	.217

Note: Overall Model: Wald $\chi^2(3) = 26.358$, $p < .001$
R^2= .45 (Hosmer & Lemeshow), .34 (Cox & Snell), .56 (Nagelkerke)

As expected, the response graphs shown in Fig. 6 are similar due to the similar odds ratios of the predictors in question.

We consider the overlay of [machine data or transaction value] by region in Fig. 7.

In the home region, we observe that the probability to get a transaction declined in below 5%, if machine data and value stay at control level. If either of them are changed to machine data being overwritten or the value increased, we expect a probability of about 50% to get the transaction declined.

Once the user requests a transaction from the foreign region, the probabilities are not in the user's favor. Everything else at control level, we expect a probability of 50% to 60% of the transaction being declined. If either the machine data is overwritten or value is increased, it is almost certain for the transaction to be declined.

5.3.2 *Model Evaluation.*

Performance. We have a classification accuracy of 83%, Hosmer-Lemeshow not rejecting the fit, HL_C $\chi^2(8) = 6.96$, $p = .540$.

Cross-Validation. The repeated 10-fold cross-validation showed an accuracy of 79%, 95% CI [75%, 82%]. The model offers reasonably accuracte predictions (Cohen's $\kappa = .57$), with accuracy statistically significantly greater than the no-information rate, $p < .001$.

5.4 Logistic Regression: Card Blocked

We established a binomial logistic regression on the impact of predictors machine_data, value and region on the credit card being blocked. We offer details on model selection and evaluation in Appendix A.3.

5.4.1 *Fitted Model.*

Predictors and Odds Ratios. We offer an overview of the predictor estimates and p-values in Table 4.

The only predictor statistically significantly impacting the likelihood of the card being blocked was the region, $z = 1.23$, $p = .217$, OR = 2.99, 95% CI [0.6, 19.8]. A change from the card's home region (UK) to the foreign region (Germany) made it 3 times more likely to get the card blocked.

We thereby failed to reject the null hypotheses $H_{0,X,\text{blocked}}$ for $X \in \{$ machine_data, value, region $\}$.

5.4.2 *Model Evaluation.*

Performance. The classification accuracy was 73%, Hosmer-Lemeshow not rejecting the fit, HL_C $\chi^2(8) = 1.78$, $p = .987$.

Cross-Validation. The repeated 10-fold cross-validation yielded an accuracy of 84%, 95% CI [81%, 87%], Cohen's $\kappa = .41$.

5.5 Overall Model Properties

The three selected models stay valid with each $p < .001$ under Bonferroni-Holm correction for multiple comparisons. The selected models show a classification accuracy of $73\% - 83\%$, with a passable fit. At the same time, the cross-validation accuracy was low to medium, which makes the selected models less useful as predictive classifiers for other datasets.

6 DISCUSSION

The discussion is best seen in context of the likelihood plots of Figures 4 and 6 on pp. 10 and 11.

6.1 The overall likelihoods of outcomes differ characteristically.

From the quantification on likelihoods obtained from the logistic regressions, we can observe a consistent order of likelihoods. It was most likely for a transaction to be rejected especially in a foreign region. It seems that 3DS is taking no chances in the case of either the machine data being corrupt or the value being too high: the that transactions are declined is all but certain.

It is noteworthy that the likelihood to decline transactions was consistently higher in foreign and home regions alike than the likelihood to challenge the user with a password authentication. There seems to be a prioritization of user convenience in the sense creating less interruptions in payment flow overall.

Of the three outcomes considered, the card being blocked had the lowest effect size (odds ratio), that is, 3DS seems least likely to have a card blocked as *ultima ratio*. Of course, this makes sense given the hassle for consumers and banks alike to get a card unblocked or a new card issued.

6.2 The three independent variables have an effect in the same order of magnitude.

For being challenged and the transaction declined, we find that the three interventions investigated (overwriting machine data, changing to a foreign region, or increasing the transaction value) all yielded an impact on the respective outcome with a change in likelihood roughly in the same order of magnitude. Hence, we conclude that 3DS takes into account all three variables in its decision making process and that the variables are roughly equally weighted.

It is important to note, however, that the variables are not KO criteria: If something is amiss in only one of those variables, the outcome will just be biased towards the user being challenged or the transaction being declined. However, only if two variables come together in a deviation from the norm (machine data intact, home region, value relatively low) then the likelihood of a 3DS intervention is predicted to be more than 50%.

6.3 The impact of the region is consistently strong.

Having said that the three variables seem to have an impact of equal order of magnitude, it seems that the region still ranks first among them, consistently being in the first effect-size rank. This becomes especially apparent when looking at the "by region" plots presented in Figures 5 and 7 on pp. 10 and 11.

Here, we see that the change from home to foreign region impacts the likelihood of a negative outcome more strongly than the combined changes in likelihood caused by problems with machine data or transaction value.

6.4 The impact of the card used seemed consistently weak.

While we used four different payment cards from different providers in the experiment, we found consistently low effect sizes on the impact of the card used.

6.5 Limitations

6.5.1 Generalizability. We are the first to state that the generalizability of this experiment is somewhat limited. In terms of experiment design, this is rooted in a small number of credit cards, card providers and merchant sites being evaluated. Furthermore, the different payment cards used were linked to a single card holder.

While the card being used generally was linked to a comparatively low effect size, the experiment was thereby not prepared to discern whether 3DS and payment card institutions *personalize* their responses to the card holder.

To gain a quantification of the impact of different card holder profiles on the outcomes, one would need a wide range of participants with different credit card histories, which was beyond the scope of this study.

While such future research might yield interesting results, it comes with ethical caveats that participants might expose their credit card accounts with a host of failed transactions, which in turn could impact the future behavior of their payment cards.

6.5.2 Sample Size & Power. Operating on live credit cards owned by real people, we saw a need to exercise restraint how many transaction we would run.

We computed the logistic regressions with a sample of $N = 64$ and a maximal number of predictors $k = 5$.

An *a priori* power analysis with G*Power based on a presumed H_1 probability of 50% and a presumed H_0 probability of 20% for the impact of one predictor (assuming the others to have $R^2 = .3$), highlighted a need of a minimal sample size of $\hat{N} = 55$ to reach 80% power.

We are aware that we are operating below the rule-of-thumbs limits of sample sizes used for binomial logistic regressions. We accepted that we accepted that in terms of sensitivity, we could only detect effect of OR ≥ 4 at 80% power. To be prudent operating at this small a sample size, we used the corrected Akaike Information Criterion (AIC_c) for the model selection and profile-likelihood limits for the interval estimation on odds ratios (both said to be superior for small sample sizes).

7 CONCLUSION

In this paper, we presented the first attempt to quantify back-end decision making process of 3-D Secure (3DS). Considering the 3-D Secure decisions as probabilistic, we have employed an empirical experiment to evaluate to what extent different deviations from the norm (overwriting machine data, leaving a payment card's home region, or increasing the value of the transaction) change the likelihood of a "negative" outcome for the user.

To the best of our knowledge, we are the first to employ logistic regression to quantify the changes in likelihood in the outcomes of the 3DS decision, that is, whether the user is challenged with a password authentication, whether the transaction is declined, or whether the card is blocked altogether.

Overall, we observed that the likelihood of the different outcomes follow different distributions, transactions declined being the most likely, card blocked the least. While all predictors showed the same order of magnitude on the biasing the decision to a "negative" outcome, we found that the impact of the region was consistently in the first rank.

While this study is limited in its scope and the sample size too small to obtain accurate predictive logistic regression classifiers for other datasets, we believe that the result is an interesting first step. By itself, it already offers insights in the characteristics of the 3-D Secure decision making in the back-end, normally shrouded from the user.

7.1 Future Work

So far, we have considered each line of outcomes separately. Of course, these analyses do not take into account the interplay between dependent variables. As future work, we anticipate it to be fruitful to analyze 3-D Secure either with a multinomial logistic regression or hidden model estimation.

To evaluate the impact of *personalized* user profiles on payment card transactions governed by 3-D Secure, future work could include a large-scale experiment with many participants and a diversity of card payment histories.

ACKNOWLEDGMENTS

This work was in parts supported by a grant of the National Cyber Security Centre (NCSC) on "Pathways to Enhancing Evidence-Based Research Methods in Cyber Security" in its approach to research integrity and reproducibility. Pre-registration, dataset and meta-data annotation are independently hosted at the Open Science Framework (OSF)[3] and publicly available under Creative Commons license. Thomas Groß was supported by the European Research Council (ERC) Starting Grant "Confidentiality-Preserving Security Assurance (CASCAde)," GA n°716980.

[3]DOI 10.17605/OSF.IO/X6YFH; https://osf.io/x6yfh/

REFERENCES

[1] BBC. 2018. British Airways boss apologises for 'malicious' data breach - BBC News. (2018). https://www.bbc.co.uk/news/uk-england-london-45440850
[2] BBC. 2018. Ticketmaster admits personal data stolen in hack attack - BBC News. (2018). https://www.bbc.co.uk/news/technology-44628874
[3] E commerce Europe. 2017. European Ecommerce Report 2017 - Ecommerce continues to prosper in Europe, but markets grow at different speeds. (2017). https://goo.gl/AZFqVs.
[4] EMVCo. 2017. 3-D Secure 2.0. (2017). https://goo.gl/d1ksLf.
[5] Financial Times. 2018. The British Airways data breach shows that regulation works | Financial Times. (2018). https://www.ft.com/content/a301f46a-b4df-11e8-bbc3-ccd7de085ffe
[6] S. J. Murdoch and R. Anderson. 2010. Verified by Visa and MasterCard SecureCode: Or, How Not to Design Authentication. In *Proceedings of the 14th international conference on Financial Cryptography and Data Security*. Springer Verlag, 336–342.
[7] K. Richard. 2009. Verified by Visa: bad for security, worse for business - Richard's Kingdom. (2009). https://goo.gl/NgUUvn.
[8] Financial Fraud Action UK. 2017. *Fraud the Facts 2017. The Definitive Overview of Payment Industry Fraud.* Technical Report. https://goo.gl/3cLu8N.
[9] Visa. 2017. 3D Secure. (2017). https://goo.gl/TZSTEc.
[10] VisaCo. 2018. Visa - Leading Global Payment Solutions. (2018). https://usa.visa.com/.
[11] Worldpay. 2018. Card fraud increases as stolen cards used once every 20 seconds - Payments Cards & Mobile. (2018). http://www.paymentscardsandmobile.com/card-fraud-increases-as-stolen-cards-used-once-every-20-seconds/

A MODEL EVALUATION

A.1 Logistic Regression: User Challenged

Model Selection. The model with all five predictors was statistically significant, $\chi^2(5) = 21.593$, $p < .001$. We checked for second-level and third-level interactions wrt. the manipulated independent variables and found none.

This model yields an Akaike Information Criterion corrected for small sample sizes of $\mathrm{AIC_c} = 69.73$. Compared to the best model only including the predictors with the greatest residual drop (machine_data, value, and region), this fitted model experiences a small enough information loss ($\Delta = 1.14$) to be classified as having substantial support.

Goodness of Fit. We computed a likelihood-ratio test to compare the full model selected against the one with minimal $\mathrm{AIC_c}$ (and fewer predictors). We failed to reject the null hypothesis that the reduced model is true, and, hence, keep the full model. We report the goodness-of-fit in different variants of Pseudo-R^2 in Table 2. McFadden's $R^2 = .28$.

Diagnostics. There were two cases with large residuals, but DFbetas were well below .5. There were no cases with large leverage. Assessing for multicollinearity, we found the Variance Inflation Factors (VIFs) all close to 1, with a mean VIF of 1.09.

A.2 Logistic Regression: Transaction Declined

Model Selection. The model including all predictors was statistically significant, $\chi^2(5) = 44.409$, $p < .001$. Second-level and third-level interactions between manipulated variables were not statistically significant.

This model comes with a corrected Akaike Information Criterion $\mathrm{AIC_c} = 57.72$. Compared to the min-$\mathrm{AIC_c}$ model with the predictors machine_data, value, region, and card, the fitted model has an information loss of $\Delta = 2.3$. This is which past the threshold of substantial support, but still considered good evidence.

Goodness of Fit. Comparing the min-AIC_c model with the chosen model on goodness-of-fit, we find that a likelihood-ratio test does not reject the null hypothesis. We report the goodness-of-fit in Pseudo-R^2 in Table 3. McFadden's $R^2 = .50$.

Diagnostics. There were four cases with large residuals, yet DFbetas shown to be below 1. There were 12 cases with leverage just touching twice the average leverage, however the DFbetas are consistently less than 1 and Cook's distance less than 0.1. The VIFs are smaller than 2, where the mean VIF is 1.49.

A.3 Logistic Regression: Card Blocked

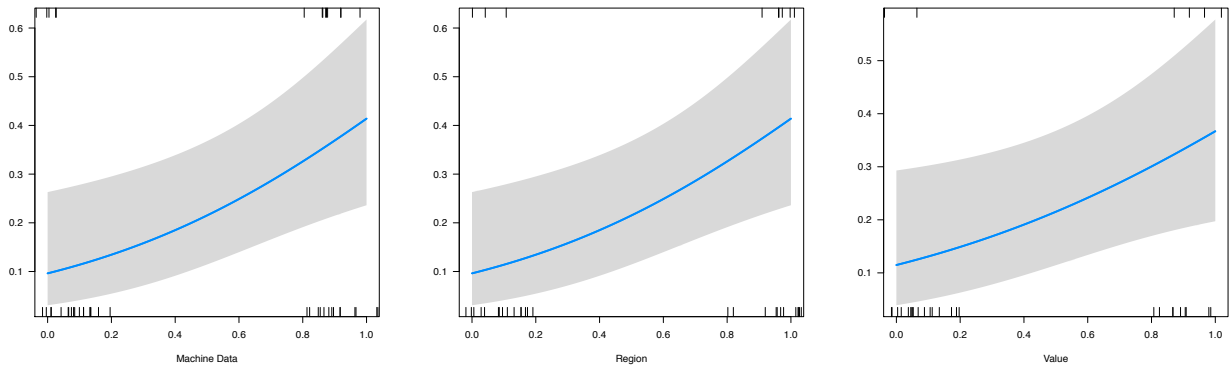
Model Selection. We have a scenario in which the model including only value and region has the minimal $AIC_c = 40.39$.

The full model including the three other predictors ($AIC_c = 44.71$) yields an information loss of $\Delta = 4.32$, having considerably less support. Given the data, this model only carries a likelihood of 1%.

Even though the likelihood-ratio test does not reject the null hypothesis, we consider the model with value, region andmachine_data as robust alternative ($AIC_c = 41.05$). In comparison with the minimal-AIC_c model, we have an information loss of $\Delta = 0.67$, yielding substantial evidence. Hence, we select this model.

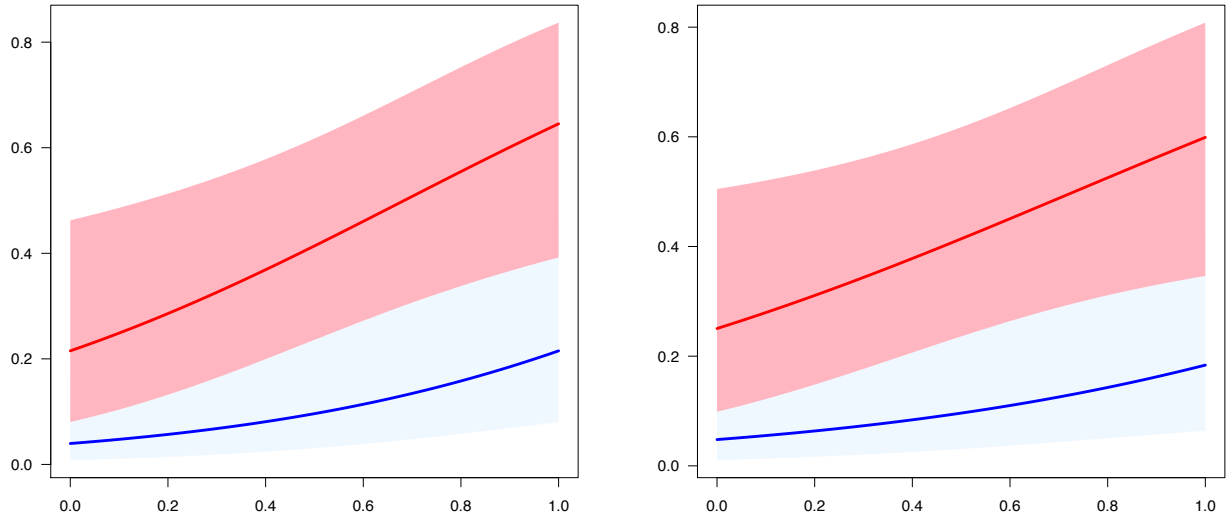
Goodness of Fit. We evaluate a likelihood-ratio test to check the goodness-of-fit of the min-AIC_c model vis-à-vis the chosen model. It did not reject the null hypothesis. Table 4 contains customary Pseudo-R^2 estimates. McFadden's $R^2 = .45$.

Diagnostics. There was one case with high residuals, but DFbetas smaller than 1. There were 9 cases with a leverage past the double-mean-leverage threshold. Inspecting DFbetas, we find them to be below 1, and inspecting the cooks distance, we find it below 0.2 max. The VIF was consistently close to 1, with a mean VIF of 1.04.



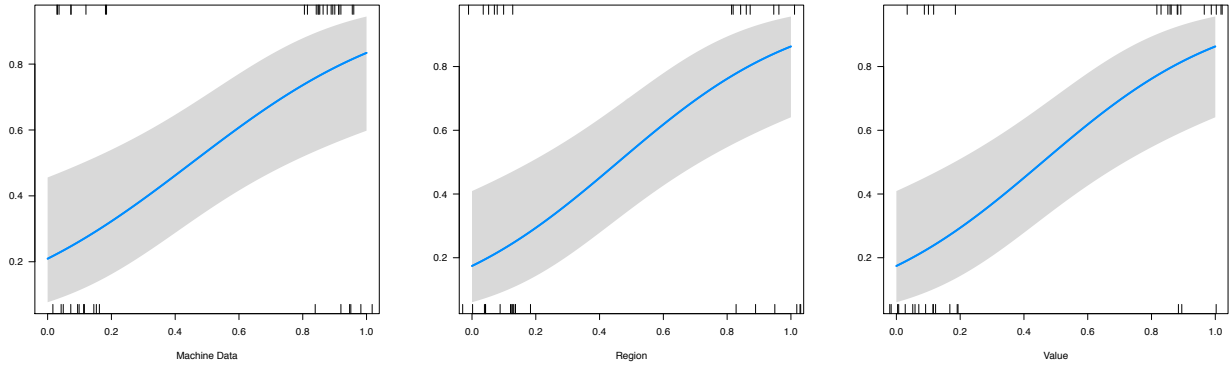
(a) Change of Machine Data (b) Change of Region (c) Change of Value

Figure 4: Probability(challenged) depending on significant predictors.



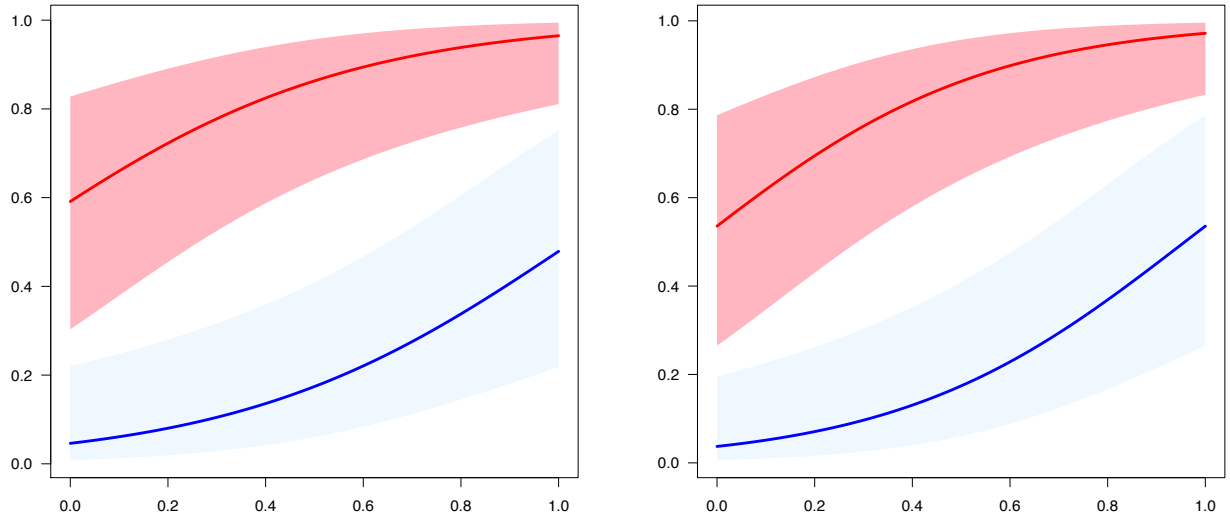
(a) machine.data by region (b) value by region

Figure 5: Probability(challenged) depending on [machine data *or* transaction value] by region. (The blue line on the bottom shows the home region, the red line on the top the foreign region)



(a) Change of Machine Data (b) Change of Region (c) Change of Value

Figure 6: Probability(declined) depending on significant predictors.



(a) machinedata by region (b) value by region

Figure 7: Overlay Plot: Probability(declined) depending on [machine data *or* transaction value] by region. (The blue line on the bottom shows the home region, the red line on the top the foreign region)