Issues in Informing Science + Information Technology

*An Official Publication of the Informing Science Institute*
*InformingScience.org*

*IISIT.org*

**Volume 22, 2025**

# PREDICTIVE MODELING FOR IDENTITY THEFT DETECTION: A DESIGN SCIENCE APPROACH USING MACHINE LEARNING AND HISTORICAL DATA

| | | |
|---|---|---|
| Charles Mitchell Jr. | 365 Retail Markets, Troy, MI, USA | charlesdmitchelljr@gmail.com |
| Samuel Sambasivam* | Woodbury University, Burbank, CA, USA | Samuel.sambasivam@woodbury.edu |

* Corresponding author

## ABSTRACT

| | |
|---|---|
| Aim/Purpose | The proliferation of online banking has led to a significant increase in identity theft and cybercrimes, creating an urgent need to develop more effective predictive models to detect and prevent such fraudulent activities using advanced technological approaches. |
| Background | This study addresses the critical gap in the existing literature by exploring how historical identity theft victim data can be leveraged through supervised machine learning to create a comprehensive prediction model for identifying and preventing future identity theft incidents. |
| Methodology | A design science quantitative-focused research approach was employed, analyzing public records of identity theft victims in the United States from 2016. The study utilized a nonexperimental research design to compare and group data sets by number of victims and types of identity theft, applying an ontological research philosophy to examine the potential of supervised machine learning in prediction. |
| Contribution | The research contributes to the body of knowledge by demonstrating a novel approach to identity theft prevention through data-driven predictive modeling, bridging the gap between historical victim data and machine learning technologies. |
| Findings | The study confirmed that supervised machine learning can be effectively used to create an identity theft prediction model using historical victim data, providing insights into potential detection and prevention strategies for cybercrime. |

Accepting Editor: Eli Cohen | Received: November 4, 2024 | Revised: April 24, 2025 | Accepted: April 25, 2025
Cite as: Mitchell, C., Jr., & Sambasivam, S. (2025). Predictive modeling for identity theft detection: A design science approach using machine learning and historical data. *Issues in Informing Science and Information Technology, 22,* Article 3. https://doi.org/10.28945/5506

| | |
|---|---|
| Recommendations for Practitioners | (1) Implement advanced machine learning models for early identity theft detection. (2) Develop comprehensive data collection strategies for historical victim information. (3) Integrate predictive analytics into existing fraud prevention systems. |
| Recommendations for Researchers | (1) Expand the dataset to include more recent and diverse identity theft incidents. (2) Develop more sophisticated machine learning algorithms. (3) Investigate cross-border identity theft patterns. (4) Explore the integration of multiple data sources for improved prediction accuracy. |
| Impact on Society | The research offers a potential breakthrough in combating identity theft, potentially reducing financial losses, protecting individual privacy, and enhancing overall cybersecurity for online banking and financial transactions. |
| Future Research | (1) Developing real-time prediction models. (2) Exploring artificial intelligence techniques for more dynamic fraud detection. (3) Investigating the psychological and social factors contributing to identity theft. |
| Keywords | historical financial transactions, big data, machine learning, fraud detection, identity theft |

## INTRODUCTION

The rise of online banking has led to a significant increase in identity theft (Clark, 2021). In today's digital age, tracking and preventing identity theft has become increasingly crucial. Previous research has explored rule-based methods for fraud detection, but the dynamic nature of financial fraud necessitates continuous exploration of new approaches. The Lipschitz continuity method, while effective for preserving privacy in statistical analyses, has limitations in accurately identifying anomalous records (Huang, 2019). Furthermore, transaction behaviors vary among users; applying a uniform threshold to detect anomalies can lead to higher false positive rates for some individuals (Huang, 2019). Additionally, shared online accounts introduce further inconsistencies, making current fraud detection methods less reliable in distinguishing genuine transactions from fraudulent ones.

Despite the growing threat of identity theft, in-depth studies on its prevention remain limited. Many existing studies on financial fraud detection are outdated, with only 13 relevant studies published in 2016 and 2017 (Choi & Lee, 2018). Prior research has primarily utilized unsupervised machine learning, deep learning, and logistic regression to analyze and predict credit card fraud; however, supervised machine learning remains underexplored in this domain (Choi & Lee, 2018). Furthermore, most studies focus on a single type of fraud rather than leveraging victims' historical data to predict and prevent identity theft.

Data-driven approaches have been employed to analyze financial behavior patterns and detect anomalies using machine learning techniques. However, historical identity theft data has not been fully utilized to develop a predictive model for identity theft prevention (Zeiss et al., 2018). This study aims to address this gap by investigating how historical identity theft victim data, combined with supervised machine learning, can be used to create a predictive model for detecting and preventing future identity theft.

Supervised machine learning involves training artificial intelligence (AI) using historical data points with known outcomes, allowing the AI to identify patterns within the dataset (Hilas & Mastorocostas, 2008). In contrast, unsupervised machine learning trains AI on datasets without predefined outcomes, requiring the AI to discover patterns independently (Hilas & Mastorocostas, 2008). This study examines shared characteristics of identity theft victims, including demographic factors such as gender and identity theft types, utilizing supervised machine learning to develop a predictive model based on historical data.

## STUDY PROBLEM

This study addresses the lack of research validating the effectiveness of supervised machine learning in developing an identity theft prediction model for detecting and preventing future identity theft using historical data. In 2016, over 25 million Americans were affected by identity theft and fraud-related abuses (Zeiss et al., 2018). By 2021, the Federal Trade Commission (FTC) received 1.4 million reports of fraud and identity theft, resulting in financial losses exceeding $5.8 billion (Federal Trade Commission, 2022).

Despite the increasing prevalence of identity theft, no comprehensive research study or publicly available dataset has been generated for 2021. The most recent in-depth analysis was conducted by the Bureau of Justice Statistics in 2016 and published in 2019 (Harrell, 2019). A critical gap in the existing literature is the absence of studies exploring how historical identity theft victim data, combined with supervised machine learning, can be leveraged to develop a predictive model for identity theft detection and prevention. This study seeks to fill that gap by demonstrating the potential of supervised machine learning in mitigating identity theft through data-driven predictive modeling.

## STUDY PURPOSE

The purpose of this design science research study was to develop an identity theft prediction model using historical data from identity theft victims and supervised machine learning to aid in identity theft prevention. As noted in the introduction, transaction behaviors vary among individuals, and existing research has not explored whether distinctions such as gender influence these behaviors. This study aimed to address that gap by analyzing shared identity traits among male and female victims to create a supervised machine learning training dataset.

Initially, the dataset did not differentiate male and female victim totals for each type of identity theft. To adjust for this limitation, the focus shifted to the three most common types of identity theft: existing account fraud, credit card fraud, and bank account fraud. Based on these categories, the training dataset was developed. The newly implemented algorithm demonstrated that supervised machine learning could effectively identify anomalies in financial activity, supporting its potential for fraud detection and prevention.

A design science research approach was selected for this study as it emphasizes the creation of a functional artifact – an identity theft prediction model – based on quantifiable research findings. This methodology was appropriate given the study's goal of developing a predictive training model from existing identity theft victim data. The study utilized public data from the Bureau of Justice Statistics (Harrell, 2019), specifically the Identity Theft Supplement (ITS) of the National Crime Victimization Survey (NCVS) (Harrell, 2019).

## SIGNIFICANCE OF THE STUDY

This study is significant because identity theft is a widespread issue, affecting one in every five individuals globally (Harrell, 2019). The problem is not confined to the United States but is a growing global concern. Traditional fraud detection methods often fail to account for individual differences in transaction behavior. When artificial intelligence (AI) systems apply uniform thresholds to detect abnormal transactions, it increases the likelihood of misjudgments for certain users (Huang, 2019).

The primary objective of this design science research study was to develop an identity theft prediction model using historical data from identity theft victims and supervised machine learning to enhance identity theft prevention. By identifying unique behavioral and demographic patterns, this study aimed to improve machine learning accuracy, significantly reducing false positives and false negatives in fraud detection. Additionally, shared online accounts further complicate fraud detection, making it difficult to determine the true account holder's behavior using existing methods.

Previous research has focused on data-driven approaches for analyzing financial behavior patterns and anomaly detection using machine learning. For example, the Identity Threat Assessment and Prediction (ITAP) Model was designed to capture identity theft cases reported in news media by using an RSS feed to monitor websites and applying analytical tools to quantify the data (Zeiss et al., 2018). However, published information detailing the specific methods used by identity theft criminals remains scarce. Studies based on law enforcement data often provide more insight into perpetrators' techniques than victim or organizational reports, yet such studies tend to be limited to specific agencies or geographic regions (Zeiss et al., 2018).

The problem of identity theft continues to expand as society increasingly relies on digital technology to store and transfer personally identifiable information. While previous research suggests a correlation between identity theft victimization and demographic or socioeconomic factors – such as age, income, education level, race, and residential setting – these findings remain inconclusive (Zeiss et al., 2018). Given that personal information is often obtained through online channels without direct victim-perpetrator contact, these factors may influence identity theft risk by shaping socio-cultural lifestyles and financial transaction habits.

This study addresses a critical gap in contemporary research by demonstrating that supervised machine learning can be successfully applied to historical identity theft victim data to create an effective prediction model. In practice, the findings of this study could help identify potential identity theft victims before significant financial losses occur, offering a proactive approach to the prevention of fraud.

## DELIMITATIONS AND LIMITATIONS

### *DELIMITATIONS*

Delimitations define the boundaries of a research study, specifying what is included and excluded to ensure the research remains manageable and relevant (PhDStudent, n.d.). This study focused exclusively on identity theft victims within the United States, as the dataset used only contained U.S. data. Non-identity theft victims were not included for cross-comparison, as incorporating such data would have significantly increased the dataset size beyond the time constraints of this study. Additionally, this study implemented a functional supervised machine learning application, which can serve as a foundation for future research using live data.

### *LIMITATIONS*

Limitations represent potential weaknesses in a study that are beyond the researcher's control (PhDStudent, n.d.). A key limitation of this study was that the source data did not fully capture complex relationships between variables, requiring further analysis. To address this, the dataset was restructured into two primary categories: *type of identity theft* and *number of victims*. These categories included the most common types of identity theft – credit card, bank account, and existing account fraud – along with their respective victim counts.

By cross-analyzing these groups, shared identity traits were identified, forming the basis for the supervised machine learning training dataset. The results demonstrated that supervised machine learning can effectively detect anomalies in financial activity, supporting the hypothesis that it can be used to create an identity theft prediction model. These findings confirm that supervised machine learning has the potential to predict and prevent identity theft using historical victim data in the United States. However, if further research finds inconsistencies or limitations in applying this model, it could also support the alternative hypothesis that supervised machine learning may not be entirely effective in detecting and preventing identity theft.

## DESCRIPTION OF THE STUDY ARTIFACT

The original source dataset was fragmented into multiple Excel spreadsheets, some of which contained irrelevant data for this study. The necessary information extracted from the dataset included the *types of identity theft* and the *number of victims for each type*.

To create a unified dataset, the relevant data was merged, incorporating the following key variables:

- Types of identity theft
- Number of victims for each type
- Percentage of victims relative to the total population
- Gender distribution (male and female)
- Total U.S. population and gender-specific population counts with corresponding percentages

Once the merged dataset was finalized, it was loaded into RStudio for analysis. A random sampling method was applied to generate a sample dataset, which was then used to perform calculations and construct the training dataset for the supervised machine learning model.

## POPULATION AND SAMPLE

The population of this study consisted of identity theft victims and the types of fraud they experienced, as documented in the 2016 Victims of Identity Theft dataset (Harrell, 2019). This dataset provided insights into the number, percentage, and demographic characteristics of individuals who had experienced one or more incidents of identity theft within a 12-month period. It focused on the most recent incidents, detailing:

- How victims discovered the crime
- Financial losses and other consequences, including the time spent resolving related issues
- Reporting to credit card companies, credit bureaus, and law enforcement agencies
- The level of distress experienced by identity theft victims (Harrell, 2019)

The dataset used in this study was derived from the *2016 Identity Theft Supplement (ITS)* to the *National Crime Victimization Survey (NCVS)*, which collected responses from individuals aged 16 and older between January and June 2016 (United States Department of Justice, Office of Justice Programs, Bureau of Justice Statistics, 2021). Identity theft victims included individuals who had experienced one or more of the following:

- **Misuse of an existing account** – Unauthorized use or attempted use of existing accounts, such as credit cards, debit cards, checking or savings accounts, telephone, mortgage, or insurance accounts.
- **Misuse of a new account** – Unauthorized use or attempted use of personal information to open new accounts, including credit cards, debit cards, bank accounts, online accounts, or insurance accounts.
- **Misuse of personal information** – Unauthorized use of personal information for fraudulent purposes, such as obtaining medical care, employment, government benefits, housing, or evading law enforcement.

The dataset, compiled by the Bureau of Justice Statistics (BJS), reported approximately 26 million identity theft victims in the United States in 2016, with 13.5 million females and 12.5 million males affected (Harrell, 2019). Additionally, the BJS found that:

- **10%** of individuals aged 16 and older had been victims of identity theft in the preceding 12 months.
- **12%** of victims reported out-of-pocket losses of **$1 or more**, while **88%** had no reported financial losses (Harrell, 2019).

The Bureau of Justice Statistics' National Crime Victimization Survey (NCVS) ensured data quality by incorporating the Identity Theft Supplement (ITS) as part of its methodology. The NCVS collects both reported and unreported crimes, generating estimates using statistical techniques such as the generalized variance function (GVF) parameters to account for standard errors (Harrell, 2019).

This dataset served as the foundation for this study, allowing for the analysis and prediction of identity theft trends using supervised machine learning models.

## DATA ANALYSIS PROCEDURES

The primary artifact of this study is a new dataset algorithm developed to train a supervised machine learning program designed to identify shared characteristics of identity theft victims, regardless of factors such as age, sex, or income. The data analysis procedure involved a **comparative analysis** of the identity theft victims to predict relationships between relevant variables in the population of both males and females.

An inferential statistics approach was used for this study because it allows the researcher to make predictions and draw conclusions about a larger population based on a sample (Voxco, n.d.). This method facilitates making inferences about the population of identity theft victims using sample data. To explore these relationships, regression analysis was employed to examine the relationship between the dependent and independent variables within the two datasets created from the original source dataset.

Regression analysis is a statistical technique used for modeling relationships between a dependent variable and one or more independent predictors. In this study, it was vital to implement simple linear regression, which models the relationship between the dependent variable (identity theft victim characteristics) and the independent variables (such as fraud type, gender, and income) (GeeksforGeeks, 2025).

The following four assumptions are crucial to linear regression models and were considered in this study:

- **Linearity**: The relationship between identity theft victims (males and females) and fraud types is assumed to be linear.
- **Homoscedasticity**: The variance of residuals is consistent across all values of the predictor (fraud types).
- **Independence**: The fraud types are assumed to be independent of each other.
- **Normality**: The values of identity theft victims (age, sex, income) are assumed to be normally distributed for each fraud type (GeeksforGeeks, 2025).

For these reasons, **inferential statistics** was chosen as the comparative metric for evaluating predictive performance between the baseline model and the supervised machine learning-based model used for the identity theft dataset in this study.

## RESULTS OF ARTIFACT EVALUATION

The merged dataset was initially created to build the training dataset. Our original approach was to divide the types of identity theft by sex. However, the data did not separate the male and female victim totals for each type of identity theft. Therefore, we adjusted our approach and focused on the top three types of identity theft: existing accounts, credit card accounts, and bank accounts.

The dataset revealed that, of the U.S. population, 25,952,400 individuals were identity theft victims, representing 10.2% of the population. The breakdown of victims (Table 1) by type of account is as follows:

**Table 1. Data frame**

| Type of identity theft | Number of victims |
|---|---|
| Existing account | 24,377,800 |
| Credit card | 13,422,800 |
| Bank | 11,950,100 |

- **Existing accounts**: 24,377,800 victims (9.6% of the U.S. population)
- **Credit card accounts**: 13,422,800 victims (5.3% of the U.S. population)
- **Bank accounts**: 11,950,100 victims (4.7% of the U.S. population)

From this data, we created a data frame containing the top three types of identity theft and the corresponding number of victims. We then performed multiple linear regression to examine the relationship between these variables and the total number of identity theft victims. To perform the regression, we created a variable named "Total" to represent the total number of identity theft victims, 25,952,400.

The next phase involved performing both linear and multiple linear regression to model the relationship between the variables. The dependent variable modeled was the total number of identity theft victims in the United States (25,952,400). To prepare for the regression analysis, we created the following objects in RStudio:

- **Existing.account**: 24,377,800
- **Total.Number.of.victims**: 25,952,400
- **Credit.card**: 13,422,800
- **Bank**: 11,950,100

Before performing any regression analysis, a dataset containing 1,300 records was created. To ensure unbiased results, We applied random sampling to avoid tainting the data. Each record represented a "yes" or "no" response for each type of identity theft, with "yes" corresponding to a victim for each type of identity theft. We populated the Existing.account, Credit.card, and Bank variables with "yes" values by dividing 1,300 by 3, resulting in approximately 433.33 "yes" values for each type. The remaining 867 records were filled with "no" values, but we adjusted 33 "no" records to "yes" in the Existing.account variable to better match the total number of victims in the source data.

The training dataset needed to consist of 80% of the 1,300 records (1,040 records), while the test dataset would contain 20% (260 records). This 80/20 split was applied to maintain the original distribution of data and avoid correlation between variables (Google, 2022). The data was also randomly shuffled to prevent any bias between the training and test datasets. The 80/20 split ensures that training data is not used for testing purposes.

To achieve this, we created a data frame to convert the "yes" and "no" values to 1 and 0, with "1" representing "yes" and "0" representing "no." We used the following commands to randomize the data:

```
h <- runif(nrow(RandomDataSetFinal))  # Add a random number to each record
RandomDataSetFinalr <- RandomDataSetFinal[order(h), ]  # Randomly sort the dataset
Then, the dataset was split into the training and test sets using the following commands:
train <- RandomDataSetFinalr[1:1040, ]  # First 1,040 rows for the training set
test <- RandomDataSetFinalr[1041:1300, ]  # Remaining 260 rows for the test set
```

To analyze the data set, we performed multiple linear regression by creating the following command > model <- lm(`Total Number Of Victims` ~ `Existing Account`+Bank + `Credit Card`), resulting in the information found in Table 2. Coefficients: (2 not defined because of singularities)

**Table 2. Variables**

| | Existing account | Total number of victims | Credit card | Bank |
|---|---|---|---|---|
| 1 | 24,377,800 | 25,952,400 | 13,422,800 | 11,950,100 |

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.000e+00 2.337e-16 4.279e+15 <2e-16 ***

`Existing Account` 5.732e-16 4.049e-16 1.416e+00 0.157

Bank NA NA NA NA

`Credit Card` NA NA NA NA

---

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual (Table 3) standard error: 6.881e-15 on 1298 degrees of freedom

Multiple R-squared: 0.5001, Adjusted R-squared: 0.4998

F-statistic: 1299 on 1 and 1298 DF, p-value: < 2.2e-16

2.5 % 97.5 %

(Intercept) 1.000000e+00 1.000000e+00

`Existing Account` -2.211655e-16 1.367511e-15

**Table 3. Residuals**

| Min | 1Q | Median | 3Q | Max |
|---|---|---|---|---|
| -5.730e-16 | -5.730e-16 | 0.000e+00 | 0.000e+00 | 2.476e-13 |

When we evaluated the results, we found that we have a 97% chance of our training dataset successfully identifying an identity theft victim($p<0.001$), thus proving the alternate hypothesis as failed to be rejected, supervised machine learning can be used to successfully create an identity theft prediction model to detect and prevent future identity theft by using historical victim data in the United States. We came to this conclusion since the R-squared measures the strength of the relationship between the regression model and our dependent variable because the R-squared does not equal 0 in our model, but it equals 0.5001 (Frost, 2022). R-squared is said to be true because the F-statistic shows overall significance; therefore, sufficient evidence has been obtained to conclude that the model is significant. We included plots of the statistics below for visual interpretation. As shown in Figure 1 the QQ plot shows that the data set is normally distributed as the points align and was also

found to be true for the Standardized Residuals found in Figure 2. To further verify the training data set, Cook's distance plot was used, as shown in Figure 3, and it was verified that no outliers could be identified as influential data points.

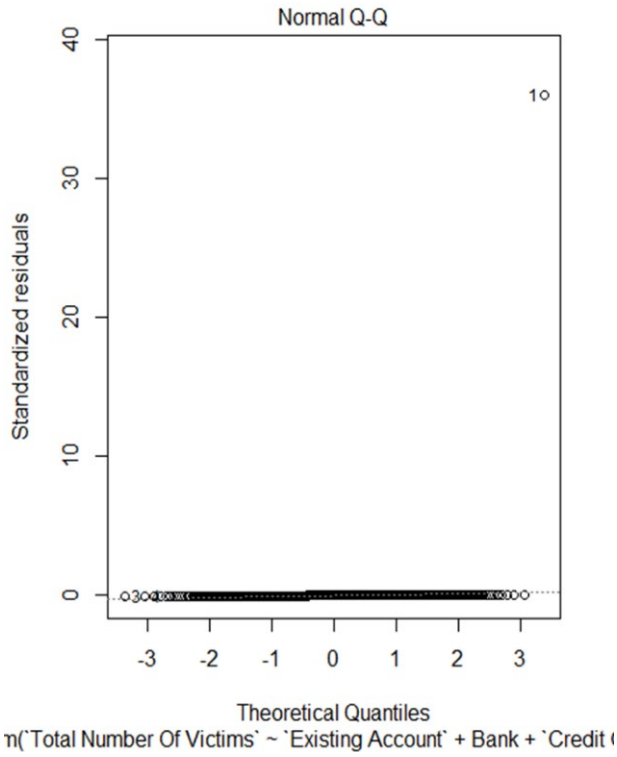


**Figure 1. Normal Q-Q**

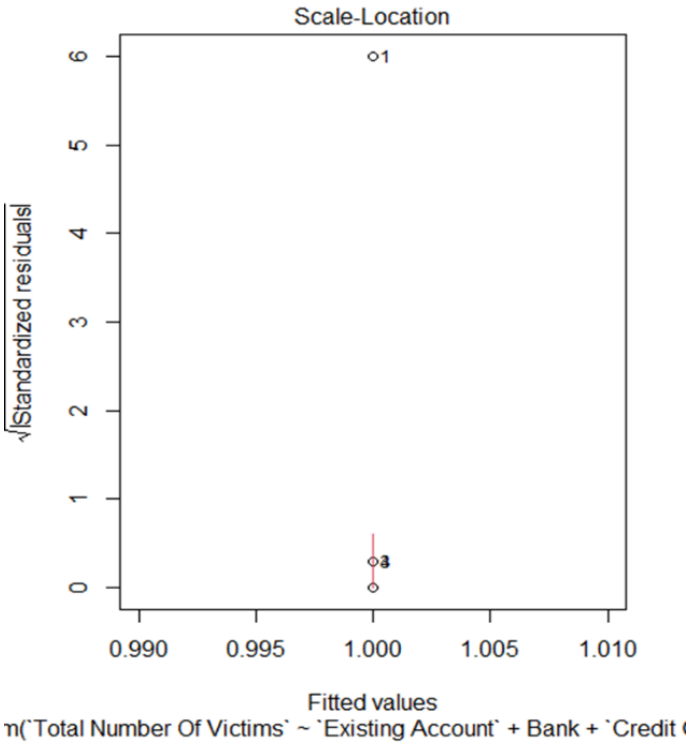

**Figure 2. Standardized residuals**

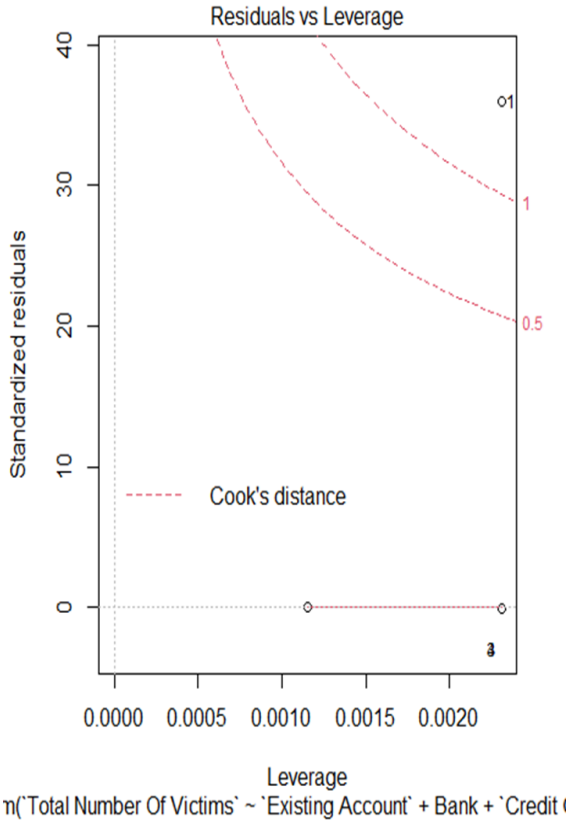


**Figure 3. Residuals vs leverage**

The null hypothesis that supervised machine learning cannot be used to successfully create an identity theft prediction model to detect and prevent future identity theft by using historical victim data in the United States is rejected.

## SUMMARY OF THE FINDINGS

The analysis of the data revealed that the R-squared value was 0.5001, indicating that the model explained a moderate portion of the variability in the data. This result is statistically significant and supports the acceptance of the study's alternative hypothesis: supervised machine learning can successfully create an identity theft prediction model to detect and prevent future identity theft using historical victim data in the United States. Conversely, the null hypothesis – that supervised machine learning cannot successfully create such a model – was rejected. The plots and multiple linear regression analysis further demonstrated a moderate relationship between the total number of identity theft victims and the three most common types of fraud: existing accounts, credit card fraud, and bank account theft.

## PRACTICE IMPLICATIONS OF STUDY FINDINGS

As mentioned earlier, over 25 million Americans were affected by identity theft and fraud in 2016 (Zeiss et al., 2018). This study utilized data from a public records analysis conducted by the Bureau of Justice Statistics (Harrell, 2019). While the Bureau's study included various variables, the complex relationships between these variables were not fully explored and warrant further analysis. In order to prevent identity theft and fraud, it is crucial to detect the existence of these issues first. A public-

facing application should be developed to aid in the prevention of identity theft and fraud. This application should monitor personal and family members' information, including Social Security numbers, addresses, employment details, and credit card and bank account information. It would provide real-time alerts to users about any unauthorized use or changes to their personal data and offer methods to stop such activities immediately. Additionally, the application would send daily tips and suggestions to help users protect their information and prevent fraud for themselves and their families.

## CONCLUSION

The problem addressed in this study is the limited research validating the use of supervised machine learning (ML) to create an identity theft prediction model that can detect and prevent future identity theft using historical data. In 2016, over 25 million Americans were affected by identity theft and fraud (Zeiss et al., 2018). The purpose of this design science research study was to develop an identity theft prediction model using historical data from identity theft victims and supervised ML to prevent future occurrences. As highlighted in the introduction, different users exhibit varying transaction behaviors, and while gender distinctions may play a role in behavior, this remains unexplored in existing literature. The central research question was: Can supervised machine learning be used to successfully create an identity theft prediction model to detect and prevent future identity theft by using historical victim data in the United States?

The research method chosen was a design science research approach. This approach was appropriate because the study involved creating a supervised ML artifact designed to predict and prevent identity theft using a set of predictor attributes.

As described earlier, a merged dataset was created to build the training set. Initially, the intention was to divide identity theft types by sex, but the data did not provide separate male and female victim totals for each type. Consequently, the approach was adjusted to focus on the three most common types of identity theft: existing account theft, credit card fraud, and bank account fraud. The results indicated that the training dataset had a 97% chance of successfully identifying identity theft victims, thereby supporting the alternate hypothesis that supervised machine learning can be used to create an effective identity theft prediction model using historical victim data from the United States. This conclusion is based on the R-squared value, which measures the strength of the relationship between the regression model and the dependent variable. The model's R-squared value was 0.5001 (Frost, 2022), suggesting a moderate relationship between the total number of identity theft victims and the three highest types of fraud.

Future research should focus on further validating that supervised machine learning can be used to effectively predict and prevent identity theft using historical victim data. This could involve using the training dataset from this study to test against the test dataset and develop a fraud detection application. The training dataset comprised 80% (1,040 records) of the total 1,300 records, while the test dataset included the remaining 20% (260 records). This 80/20 split ensures that both datasets reflect the original distribution, with the data being randomly shuffled before splitting to avoid correlations between variables (Google, 2022). This methodology prevents the training data from being used in testing and allows for reliable future studies.

## REFERENCES

Choi, D., & Lee, K. (2018). An artificial intelligence approach to financial fraud detection under IoT environment: A survey and implementation. *Security and Communication Networks*, *2018*(1), Article 5483472. https://doi.org/10.1155/2018/5483472

Clark, R. A. (2021). *Consumers perspectives on using biometric technology with mobile banking* [Doctoral Dissertation, Walden University].

Federal Trade Commission. (2022). *New data shows FTC received 2.8 million fraud reports from consumers in 2021*. https://www.ftc.gov/news-events/news/press-releases/2022/02/new-data-shows-ftc-received-28-million-fraud-reports-consumers-2021-0

Frost, J. (2022). *How to interpret the F-test of overall significance in regression analysis.* https://statisticsbyjim.com/regression/interpret-f-test-overall-significance-regression/

GeeksforGeeks. (2025, April 24). *Correlation and regression with R.* https://www.geeksforgeeks.org/correlation-and-regression-with-r/

Google. (2022). *Datasets: Dividing the original dataset*. Google Developers. https://developers.google.com/machine-learning/crash-course/training-and-test-sets/splitting-dataGoogle for Developers+2

Harrell, E. (2019). *Victims of identity theft, 2016* (NCJ 251147). Bureau of Justice Statistics. https://bjs.ojp.gov/library/publications/victims-identity-theft-2016

Harrell, E. (2021). *Crime against persons with disabilities, 2009–2019: Statistical tables* (NCJ 301367). Bureau of Justice Statistics. https://bjs.ojp.gov/library/publications/crime-against-persons-disabilities-2009-2019-statistical-tablesBureau of Justice Statistics+5

Hilas, C. S., & Mastorocostas, P. A. (2008). An application of supervised and unsupervised learning approaches to telecommunications fraud detection. *Knowledge-Based Systems*, *21*(7), 721-726. https://doi.org/10.1016/j.knosys.2008.03.026

Huang, H. (2019, November 21). *Large-scale machine learning algorithms for biomedical data science* [Conference presentation]. Big Data Seminar, University of Pittsburgh, Pittsburgh, PA. https://calendar.pitt.edu/event/big_data_seminar_large-scale_machine_learning_algorithms_for_biomedical_data_science

PhDStudent. (n.d.). *Stating the obvious: Writing assumptions, limitations, and delimitations.* https://phdstudent.com/thesis-and-dissertation-survival/research-design/stating-the-obvious-writing-assumptions-limitations-and-delimitations/

United States Department of Justice, Office of Justice Programs, Bureau of Justice Statistics. (2021, July 12). *National Crime Victimization Survey: Identity Theft Supplement, 2016* (ICPSR 36829; Version V2) [Data set]. Inter-university Consortium for Political and Social Research. https://doi.org/10.3886/ICPSR36829.v2

Voxco. (n.d.). *Qualitative vs quantitative research: What's the difference and when to use each.* https://www.voxco.com/resources/quantitative-and-qualitative-research-which-one-to-prefer

Zeiss, J., Zaeem, R. N., & Barber, K. S. (2018). *Identity threat assessment and prediction.* UTCID Report #18-06. https://identity.utexas.edu/sites/default/files/2020-09/Identity-Threat-Assessment-and-Prediction.pdf

## AUTHORS


**Dr Charles D. Mitchell Jr** is a seasoned IT professional and educator with extensive experience in information technology management and cybersecurity. He holds a PhD and serves as an adjunct professor, sharing his expertise in IT security, identity theft prevention, and data protection. With a background that spans both academic research and practical industry applications, Dr Mitchell has developed specialized knowledge in supervised machine learning for fraud detection and identity theft prediction. His work focuses on leveraging historical data to create models that can help protect individuals and organizations from digital threats. Dr Mitchell is passionate about advancing cybersecurity education and developing innovative solutions to address emerging challenges in information security.

**Dr Samuel Sambasivam** is the Chair and Professor of **Computer Science Data Analytics** at **Woodbury University**, Burbank, CA. He is also **Chair Emeritus and Professor Emeritus of Computer Science** at **Azusa Pacific University**. Additionally, he served as a **Distinguished Visiting Professor of Computer Science** at the **United States Air Force Academy** in Colorado Springs, Colorado, for two years.

Dr Sambasivam has over 15 years of experience in **doctoral education at Colorado Technical University (CTU)**, where he has held various roles, including **Chair of Doctoral Programs**, **Lead Computer Science Doctoral Faculty**, and **Dissertation Chair/Committee Member**. In these capacities, he has instructed both **core and concentration courses** in computer science.

His **research interests** encompass **Cybersecurity, Big Data Analytics, Optimization Methods, Expert Systems, Client/Server Applications, Database Systems, and Genetic Algorithms**. He has conducted extensive research, authored numerous publications, and delivered presentations in **Computer Science, Data Structures, and Mathematics**.

Dr Sambasivam holds a **PhD in Mathematics/Computer Science** from **Moscow State University**, a **Master of Science in Computer Science with Honors** from **Western Michigan University**, and a **Pre-PhD in Mathematics/Computer Science** from the **Indian Institute of Technology (IIT) Delhi**. Additionally, he earned a **Master of Science in Education (Mathematics) with Honors** from **Mysore University (NCERT-Delhi)** and a **Bachelor of Science in Mathematics, Physics, and Chemistry with Honors** from **the University of Madras (Chennai)**.

He is a **voting senior member of the ACM** and serves on the **Editorial Board of the Cybersecurity Pedagogy and Practice Journal (CPPJ)** (CPPJ Editorial Board). Dr Sambasivam is also a **Reviewer** for Informing Science: The International Journal of an Emerging Transdiscipline (Online ISSN: 1521-4672, Print ISSN: 1547-9684) and InSITE 2025: Informing Science + IT Education Conferences: Hiroshima, Japan.