Artificial Intelligence for Cybersecurity Incidents Prediction

Askar Dyussekeyev
School of Engineering and Digital Science
Nazarbayev University
Nur-Sultan, Kazakhstan
askar.dyussekeyev@nu.edu.kz

Yerzhan Erikuly
School of Engineering and Digital Science
Nazarbayev University
Nur-Sultan, Kazakhstan
yerzhan.erikuly@nu.edu.kz

***Abstract*—This report describes the introduction of artificial intelligence for security incidents prediction for cybersecurity response effectiveness increase that achieved by increasing company's profitability. The project was performed using CRISP-DM methodology for the company that provides security services to government agencies according to its business goals. The introduction of the approach will allow to transform own company's business strategy and reduce expenses for insurance payments. This paper's iteration describes all phases of the CRISP-DM.**

***Index Terms*—CRISP-DM, data mining, cybersecurity, security incidents prediction**

Introduction

Cybersecurity incidents is a threat not only to the security of the state, but also to the stability of world society as a whole. By the end of 2021, cybercrime is expected to cost the world $6 trillion [1]. It is essential to predict information security incidents based on advanced artificial intelligence technologies, such as machine learning and deep learning, to respond to information security incidents quickly.

The Joint Stock Company "Special Technical Service" (STS) is a state-owned enterprise that provides cybersecurity services for government agencies. It operates as a Computer Emergency Response Team (CERT) and Security Operation Center (SOC) for detection, prevention and neutralizing cybersecurity incidents, including attacks on critical national infrastructure. The company's most important task is a timely response to information security incidents to prevent damage to the interests of protected public and private organizations. Responding to these incidents occur only after analyzing the logs of information security tools, which does not satisfy the level of timeliness. At the same time, since STS often cannot completely prevent cyberattacks, it is forced to pay the money within the framework of cyber insurance. It means that to reduce costs, STS needs to predict possible security incidents to prevent and neutralize it promptly.

Since the company in the framework of cyber insurance loses about $3,000,000 yearly, the Board of Directors asked two highly qualified data scientists from the Nazarbayev University to assess the company's business processes and try to introduce data-driven innovations that will allow increasing profitability. They proposed inventing the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology to introduce artificial intelligence technologies.

I. Business Understanding

A. Business Understanding Overview

1) Available Resources: The organizational structure of the company is shown in Figure 1.

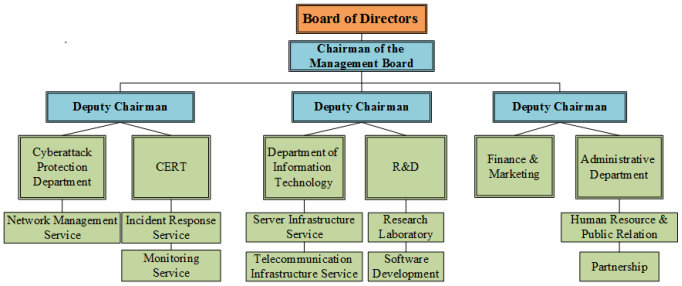

Fig. 1. Organizational structure of the JSC "Special Technical Service"

Since the company exists, it has all the necessary departments to solve a business problem. For a general understanding of the company's business processes, it is essential to describe the functional activities of each division.

The company is headed by the Board of Directors, which usually consists of the current head and his deputies and may also include board members from other organizations who have extensive work experience. The company is managed by three deputy Chairmen, who control the company's departments. The company is divided into six main structural divisions. Each Deputy Chairman of the Management Board control two structural divisions.

The First deputy commands a Cyberattack Protection Department (Network Management Service) and Computer Emergency Response Team (Incident Response Service and Monitoring Service. These structural units monitor critical infrastructures, identify vulnerabilities, analyze information security incidents, counter cyberattacks, block unsafe sources, etc. The second deputy controls the Department of Information Technology and R&D. These structural divisions are engaged in the administration of the internal and external infrastructure of the company, automation of internal business processes,

research of malicious objects, etc. Further, the third deputy oversees such administrative divisions as Finance&Marketing and Administrative Department. The presented divisions deal with financial and economic issues, develop marketing, HR management, etc.

Based on the proposal of the First Deputy of the Management Board, the Management Board allocated funds in the amount of $300,000 for conducting research and introducing the project. The project will involve the structural division of Finance&Marketing, which will be responsible for financing, and it is also possible, if necessary, qualified data analysts, Human Resources & Public Relations will be responsible for recruitment. The creation of copies of records and organizing access to the equipment will be handled by the Telecommunication Infrastructure Service division, where the Monitoring Service department will be engaged in parallel monitoring of incoming data. The Research Laboratory will be responsible for the preparation and analysis of the dataset.

Our project needs to include computer and network engineers from the Department of Information Technology and data scientists from the Research Laboratory. Moreover, the Head of CERT will allocate two security analysts from the Incident Response Service for informational support of the project.

Since the project affects various structural divisions, we identify all Deputy Chairmen of the Board in charge of these divisions as key individuals of the successful implementation. At the same time, the first deputy will be the primary stakeholder because he is interested in improving the efficiency of the CERT.

2) Problems: It should be noted that the company is trying to transform itself up to operate on international markets. Therefore, at present, it is primarily trying to increase its rating by preventing possible incidents and minimizing the negative consequences that arise. Currently, the structural unit of the incident response service CERT has a group of information security officers who go to the facility for each emerging information security incident identified by the monitoring service. However, incidents often cause significant damage for protected organizations since officers already react only after the incident occurs.

At the same time, for each successful attack, STS provides cyber insurance payments for a victim to cover costs on incident recovery that caused serious economic damage. Since the average amount of successful incidents is around 6,000 in year and the average payment is $500, the total amount of annual cyber insurance payments equals $3,000,000 ($500 x 6,000 = $3,000,000). Moreover, each call of a security officer for cyber incident response costs around $50 (one hour pay and trip spendings).

After problem analysis, data scientists decided that prediction of occurrence of information security incidents for engaging information security officers in advance will decrease the number of security incidents. Therefore, it will decrease the total number of cyber insurance payments and partially transform its business model from recovering to incident prevention.

In other words, to increase incomes, it is required to create a prediction model that utilizes historical data about previous incidents. At the same time, this data should be processed in a secure environment to ensure data security because it affects its reputation and customers' safety and security.

3) Goals: It is required to provide a solution for the problem of late response to information security incidents. As a result, the company loses an significant amount of financial resources. The business goal is to analyze the possibility of creating a system that predicts the occurrence of information security incidents based on the available historical data.

In addition, our company, like other companies, strives to gain experience in artificial intelligence. Our project is a small start and, if successfully implemented, can become a clear incentive for the further development of this industry.

4) Currently Used Solutions to Solve the Addressed Problem: The problem of security incidents prediction is similar to the crimes prediction (predictive policing). It should be noted that the former is devoid of ethical issues that are inherent in the latter. The crime prediction is a well-known problem that resolved in such papers as [2] and [3], which might be used to build an initial model to assess the possibility of the deployment.

Thus, it is proposed to build a prediction model that will provide information about predicted incidents based on historical data and machine learning. After receiving this information, information security officers have the opportunity to strengthen security measures in advance.

B. Business Objectives Understanding

1) Problem You Want to Solve Using Data Mining: STS within the framework of cyber insurance obligations loses money due to the consequences of cyber security incidents. Currently, the company cannot prevent security incidents on the customers' side due to late response.

2) All Business Questions as Precisely as Possible: STS wants to conduct this data mining to get answers to the following questions. Is it possible to correctly and accurately predict information security incidents? How can a prediction model affect effective response? How much money can we save using this prediction model? What additional spending the project will bring? How beneficial using data mining technologies? Answering these questions can help solve business problems, such as decreasing expenses, increasing the effectiveness of the response to information security incidents, and increasing reputation.

3) Any Other Business Requirements: The additional business requirement for implementing this project consists in keeping in secret the introduction of artificial intelligence for cyber security incidents prediction. This will allow avoiding a lack of distrust from customers.

Another possible issue concerns data security because STS provides services for government agencies. It means that the company must ensure the confidentiality and security of

the stored data. Failing to ensure this condition can lead to criminal liability for the company's top management and project team, especially data scientists because this issue relates to national security. Due to this, the project team will use synthetically generated and partially anonymized data to build and validate the model.

4) Expected Benefits in Business Term: It is expected that the total amount of cyber insurance payments will be reduced by not less than to $500,000 annually through timely cyber security incident response, which will be due to the prediction model. In case of successful implementation of this project, other business processes will be considered.

C. Assessing the Situation

1) Resource Inventory: As part of a project with actual data, we cannot use cloud services like Google Drive or Colab. We will use them only with synthetic datasets for the testing purposes. To implement the project, we need to purchase two high-performance servers with minimum RAM characteristics of 32 GB, SSD 1 TB, CPU 2X 2.50GHz 12 Cores, three high-performance workstations of 32 GB RAM, SSD 1 TB, CPU 2.50GHz Dual Core i7. It is also necessary to purchase information security tools (Antivirus, anti-APT, EDR) and DLP systems.

Since obtained datasets contain historical records about security incidents, which are critical for the company, it is required to ensure the security of the information because its leakage might affect the company's reputation.

As previously mentioned, to provide access to datasets, we need data engineers and network administrators of the Information Technology Department.

2) Requirements, Assumptions, and Constraints: There are restrictions on the distribution of project details because some clients might be biased toward artificial intelligence. In other words, a client may doubt the objectivity of such decision-making. It is also necessary to take all the measures required to prevent both dataset leaks and the model itself. The Company's management is very interested in the model being developed. It plans to use similar technologies in other directions in the future.

3) Risks and Contingencies: Contingency Plan for Each Risk: With a successful implementation, the issue of reducing the staff of security officers will most likely be raised because the prediction of emerging incidents will allow us to optimize our activities in this area and attract fewer employees to repel cyber-attacks. After this, there may be distrust on the part of other employees interested in stable work. There is also a risk of an "Italian strike", as a result of which all initiatives to introduce artificial intelligence will be ignored.

In this regard, it is recommended to keep the project strictly secret and gradually switch to these technologies. It is also necessary to conduct training on team building and improve the corporate culture of the company.

There is a possibility obtaining unreliable results, which can be considered as distrust and artificially can undermine the company's reputation. Although incorrect results do not affect our customers in any way. Despite this, it is necessary to test all possible methods and forecasting on the most significant possible amount of data. It is required to raise all the stored information for processing. Next, after simple machine learning methods, we need to train various types of convolutional neural networks. All these methods will be documented, and the end result will be compared.

4) Terminology:

a) IRP system: Incident Response System is a software solution that aggregates and analyzes information about incident response. A minimal set of fields of the record includes date and time, type of the incident and place of incident.

There are no other terms according to this project that would have a different meaning. Communication between members of the company's project team does not require the allocation of specific slang words.

5) Cost/Benefit Analysis: It is expected to collect data for the last four weeks, which may require up to two months of working time. Further, it will likely take about two months to train the model. In parallel, it is necessary to start work on checking our model for half a year. If any shortcomings are identified, the model is returned for correction and tested again.

In the project, to accelerate work and due to the small number of data engineers and data scientists, it will be necessary to work overtime, which implies additional employee payments and bonuses are implemented. $200,000 are allocated for these expenses. The benefits of this project are visible in the short term. It is planned not only to optimize the business process of data analysis but also to gain experience in data mining.

It is planned that data analysis will save $500,000 annually. Considering that the project costs $300,000, the project will pay off after one year. Optimization of other business processes will also be initiated. However, it is out of the scope of this project and should be implemented separately.

D. Data Mining Goals

This project's primary Data Mining goal is to build a model that predicts the occurrence of cyber security incidents using historical data. Meeting this Data Mining goal will prevent the security incident before it occurs and reduce damage to customers and, therefore, the number of cyber insurance payments. Currently, the yearly amount of insurance payments is around $3,000,000. After project introduction, it is expected to decrease by not less than $500,000 annually.

1) Type of Data Mining Problem: To achieve business goals, we need to build a model that predicts a customer who will be a victim of a cyber security incident. This prediction should be based on historical data that reflect previous incidents that occurred in various places.

2) Technical Goals Using Specific Units of Time: The prediction of cyber security incidents should be provided for a given day. Moreover, the model should provide specific hours when an incident might occur.

3) Methods for Model Assessment: We will divide the initial dataset into training and testing subsets in an 80% and 20% ratio, respectively. After the model's training, a training subset will be verified by comparing predicted and actual values to calculate the accuracy and build a confusion matrix.

Several Machine Learning algorithms will be applied, and those, which achieved a 0,25 threshold will be selected as appropriate for the project's purposes.

4) Benchmarks for Evaluating Success: Cyber security experts from the Incident Response Service extracted the historical log of incidents that occurred in the government sector from January 2010 to April 2021. This dataset was cleaned, anonymized and mixed with synthetic data for security purposes to ensure the confidentiality and security of the company's customers.

5) Subjective Measurement: By the opinion of the project group, the probability of successful deployment of the project and the likelihood of meeting the requirements assessed is very high. The arbiter of the success will be a commission that consists of management and technical experts, who will assess the fact of the successful invention of the project.

6) Successful Deployment of Model Results is Part of Data Mining Success: Since the fine-tuning of the model is a part of the project's lifecycle, the Data Mining success depends on the successful deployment of model results.

E. Project Plan

1) Project Tasks and Proposed Plan: The project plan consists of the following main stages: formation of the project team, purchasing servers and computers, collecting data, data structuring, training the model, testing models and evaluating their accuracy, modification of the model, and deployment. See figure 2. The specified work plan was drawn up after a long analysis and obtaining the consent of all participants. In addition, the plan describes all previously mentioned attracted resources, risks and goals.

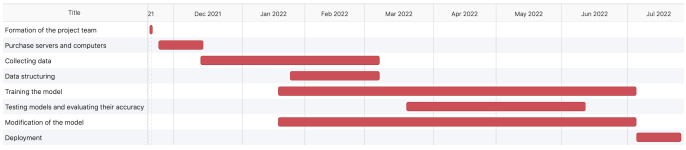

Fig. 2. Proposed plan

2) Time Estimates for All Phases or Task: The time estimate for all stages or tasks is carried out taking into account the opinions of all interested parties. In general, approximately 1 year is allocated for the project.

3) Effort and Resources Needed to Deploy: In the event that existing employees will not cope with the tasks of data mining, the H&R division should prepare specialized companies in advance to attract consultants.

4) Decision Points and Review Request: As a result of the business understanding review, the main goal that the business requires from this DM model was formulated. The Management Board is is highly interested in creating an incident prediction system, as a significant amount of money is spent on the late response. In case of successful implementation, the Board is ready to solve other problematic issues with the help of Data Mining. In addition, it is also interested in developing skills in the application of artificial intelligence technologies. Data mining will be the first step for mastering this industry.

5) Multiple Iteration: It should be noted that the model test will take place through multiple iterations. It will be tested after partial training, complete training, and even after implementing the system. Because the character of incidents is always changing, in this regard, we need to update our model.

In addition, the repetition of the model training stages involves the use of various methods. It is expected that in order to obtain the best result, several options will be developed, from which the most optimal one will then be selected.

II. Data Understanding

A. Collect Initial Data

It is assumed that the dataset was obtained from the Incident Response Platform in cooperation with a computer engineer from Telecommunication Infrastructure Service, who helped us get access to data. After securing procedure, the dataset was uploaded to a publicly available storage [4]. This dataset contains 35,280 records from 05.01.2010 to 29.04.2021 for 10 various customers and 11 different cyber security incidents.

```
In [23]: df.describe()
```

Out[23]:

	Dates	Category	Customer
count	35280	35280	35280
unique	30535	11	10
top	17.04.2018 13:00	SPAM	HOUSE OF MINISTRIES
freq	6	6556	8102

Fig. 3. Described dataset

Using the special function of a Pandas framework, we analyzed the initial dataset. We concluded that the dataset is complete for all attributes, which is demonstrated in Figure 4.

```
In [14]: df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 35280 entries, 0 to 35279
Data columns (total 3 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   Dates     35280 non-null  object
 1   Category  35280 non-null  object
 2   Customer  35280 non-null  object
dtypes: object(3)
memory usage: 827.0+ KB
```

Fig. 4. Completeness of dataset

1) 3vs of data: The dataset contains only 35,280 records (rows), which might not be identified as "big data". However, this data contains valuable information about previous security incidents.

The obtained dataset contains three attributes. While the first represents the date and time of the security incident, the second and thirds relate to its type and place, respectively. The second attribute, 'Category', may take one of 11 values, and the last one, 'Customer,' have ten various values.

Since the total number of rows is 35,280, the average number of incidents per day is around 16. It means that the data are not rapid in terms of velocity.

2) Outline data requirements: We need to obtain historical information from the IRP for each customer to address the data mining goals. The type of data should be any, but the raw CSV file is preferable.

3) Verify data availability: We confirm that the required data exists and can be used. Data sources are represented by the IRP system and DBMS, where records about previous incidents are stored.

4) Import the data into the data-mining platform: As data mining platform we use open source frameworks for data analysis and machine learning such as Pandas, scikit-learn and Jupyter Notebook.

5) List of datasets, methods of acquisition, problems: Dataset was downloaded via link https://github.com/dyussekeyev/security-incidents-prediction. Dataset was acquired from the IRP system and preprocessed as a CSV file. The obtained data is compatible with our data-mining platform.

As it was mentioned before, ensuring of the confidentiality of customers is company's priority. Due to that, the dataset was anonymized and mixed with synthetical data.

B. Data Description

The dataset consists of 3 columns and 35,280 rows that describe the date of incidents, their category, and to which customer it belongs. Using the Jupiter Notebook tool, the following attributes of columns with data types are obtained Dates(object), Category (object), and Customer (object). In addition, for convenience, using the describe() function, we looked at the statistical data of the dataset. See Figure 3 [4].

As we mentioned in the "Collect Initial Data" section, our dataset is in CSV format, consisting of 35280 rows and three columns. See Figures 3 and 4.

1) Dataset prioritization: The priority of the dataset might be recognized as highly valued to the business because historical records in case of a successful introduction of the project might significantly reduce the company's expenses on cyber insurance payments.

2) Accessibility and availability of attributes/features: The dataset contains three attributes: Dates, Category and Customer. All these attributes are accessible and available for the project team.

3) Attributes/features types, value ranges, understanding of attributes, relevance analysis, consistency of attribute usage, domain experts opinions on attribute relevance, data balancedness: As is mentioned before, the dataset contains three attributes. While the first attribute is interval, the type of second and third is nominal. The dataset contains 35,280 records between 05.01.2010 and 29.04.2021. The average amount of records during one year is around 2,500.

The possible values of the "Category" attribute are:

- SPAM
- PHISHING
- MALWARE
- DDOS
- INTERNAL OFFENER
- BOTNET
- SECURITY SYSTEM FAULT
- UNAUTHORIZED ACCESS
- NETWORK ATTACK
- WEBSITE UNAVAILABILITY
- RANSOMWARE

The possible values of the "Customer" attribute are:

- E-GOV DATA CENTER
- OFFICE OF THE PRIME-MINISTER
- HOUSE OF MINISTRIES
- AKIMAT OF THE NUR-SULTAN
- MINISTRY OF DEFENCE
- SUPREME COURT
- OFFICE OF THE PRESIDENT
- POLICE DEPARTMENT
- PARLIAMENT
- MINISTRY OF DIGITAL DEVELOPMENT

We discovered that old records might be irrelevant due to rapidly changing trends of cyber security attacks. It means that we should analyze data to decide what records might be removed/cleaned.

We asked domain experts about attribute relevance and data balancedness, and they confirmed that data inside the dataset represent regular cyber-attack activity, i.e. data balanced. However, the attribute 'Category' might be irrelevant and useless because, in a real-case scenario, it is difficult to predict what type of attack will occur. Since we are focused only on the prediction of the customer, it was recommended to exclude this attribute from further analysis.

4) Problems with data: It should be noted that the main problem with the considered dataset is missing records for specific dates. This data might not be recovered using data augmentation procedures.

C. Data Exploration

1) Attributes, properties, statistics, sub-groups: The gathered dataset consists of three attributes: "Dates", "Category", and "Customer", which are also essential attributes. The result of simple grouping demonstrates that all three features have 35,280 records. The attributes "Category" and "Customer" have 11 and 10 unique records, respectively.

Figure 5 demonstrates the number of incidents by each customer. We might observe that most incidents are about 8000 arise in the house of ministries. Figure 6 shows the number of incidents by each category. ”Spam” and ”Website unavailability” categories include more than 6000 incidents. In Figure 7, we can conveniently look at each customer’s number of incidents in various categories.

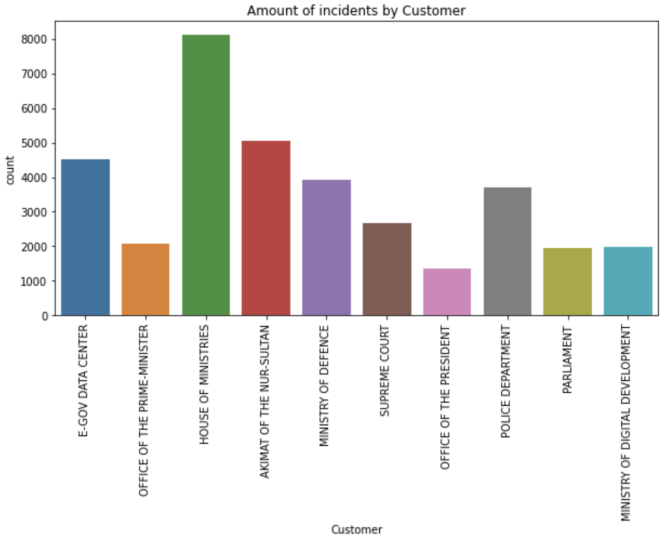

Fig. 5. Amount of incidents by customer

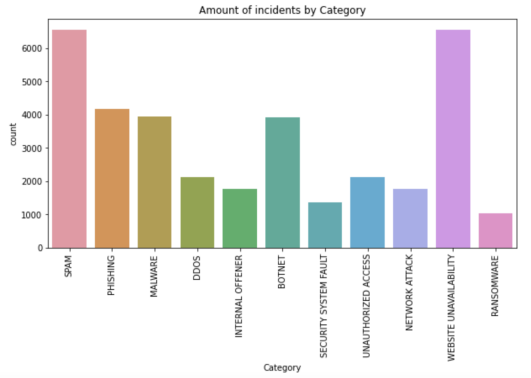

Fig. 6. Amount of incidents by categories

2) Quality problems and inconsistencies: While exploring the dataset, we found that the records are not present for all days. However, it is not a severe problem because the prediction of security incidents is still possible. At the same time, we did not find any duplicate data, inconsistent units or default values.

3) Check for errors in data: We conducted a manual check and did not find any errors in data.

4) Variables summary: We have only three attributes: ”Date”, ”Category” and ”Customer”. Since the first attribute is interval and others are nominal, it is impossible to provide a five-number summary (min, max, Q1, Q3, median/mean).

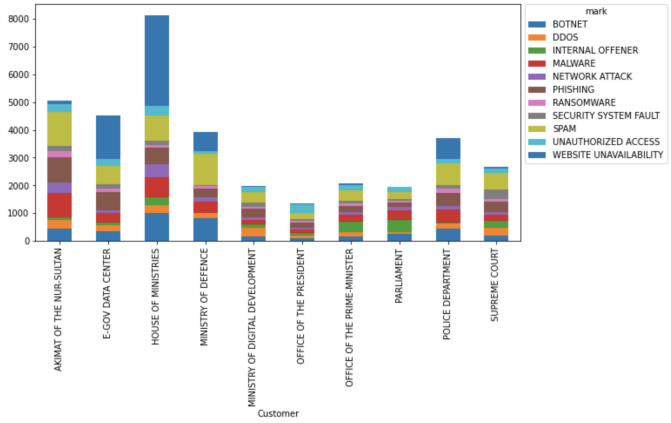

Fig. 7. Amount of incidents in each customer

5) Visualization: In Figure 9, the gaps in a dataset are shown. We can observe the alternation of weeks for which data is available and weeks for which data is not available.

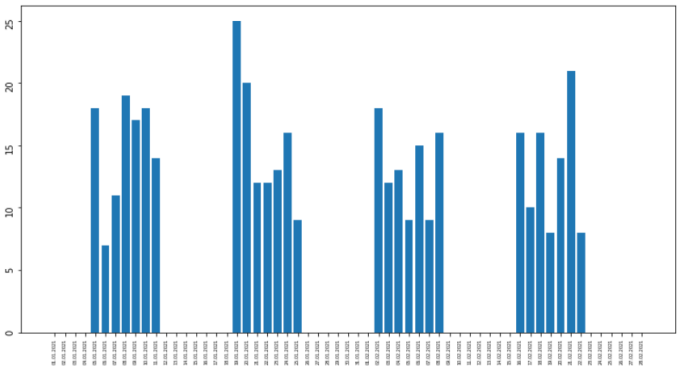

Fig. 8. Data gaps

6) Data provenance: The only problem with data is that the records are not present for all days. We believe that it is related to source provenance due to technical problems in a database. However, these issues should not lead to a data bias.

D. Verify Data Quality

1) Correctness: Since required security incident prediction cases are covered with minor insufficiency, the dataset might be recognized as partially correct. The dataset’s structure is constructed in such a way that the weeks in which the data are present and the weeks in which the data are absent alternate. However, since the primary statistical characteristics of the events are preserved, the dataset can be utilized.

2) Completeness: The dataset in its current form cannot be used for prediction purposes because we might derive new attributes. For example, we might consider to derive ’Hour’, ’Day of week’, ’Day of month’ and ’Month’ columns. Moreover, nominal values of ’Category’ attribute and ’Customer’ class should be replaced by numerical values.

3) Dirtiness and understanding: Since the date and time values of the dataset are recorded using distributed NTP

protocol, its dirtiness is unlikely. The 'Category' attribute and 'Customer' class are nominal and derived from human operators. It means that assessing its dirtiness is impossible, and therefore they should be considered as truthful. Moreover, this data is also used for insurance payments, which additionally increases its confidence.

4) Degrees of detectability, data issues types, data entry error mitigation: The dataset does not contain issues related to detectability or invalid data types (Transcription, Insertion, Deletion, Transposition). It means that a mitigation procedure of data entry error is not required.

E. Quality Report

According to quality report there is no quality issues.

1) Accessibility: Analysts have access to the data in a useful form. It exposed in a running database to analyze the data.

2) Accuracy: All values represent the true value of the entity.

3) Coherency: Data can be accurately combined with other relevant data.

4) Completeness: There is a minor issue related to missing data, which is related to missing records for certain days. We discovered that the weeks in which the data are present and the weeks in which the data are absent alternate. Since we still have records for other weeks, it is possible to build an effective model that predicts security incidents occurrence.

5) Consistency: Data is in agreement. All values of the attributes matches to related data.

6) Definition: All data fields each have a well-defined, unambiguous meaning.

7) Relevancy: Data has bearing on the data analysis being conducted.

8) Reliability: There is minor issue related to missing data, which might be derived from existed (required additional attributes: 'Hour', 'Day of week', 'Day of month', 'Month').

9) Timely: The dataset contains historical records of security incidents for more than ten years. Thus, the data is useful and can be used for planning and forecasting.

III. Data Preparation

A. Selecting Data

1) Samples, attributes: We obtained a dataset that involves information about security incidents for more than ten years. It was discovered that the dataset contains only three attributes that represent the time, type and place of the security incidents.

Since we are interested in the prediction of the customer by the time, we might exclude the 'Category' attribute. Moreover, we discovered changes in security incidents trends for various years. It means that historical data from 10 years ago are irrelevant to predicting contemporary security incidents. Thus, we limited the dataset by only 9544 records representing security incidents that occurred in 2018, 2019, 2020 and 2021.

2) Relevancy, quality, validity, salvageability: Based on the results of the dataset analysis, it was established that the relevancy and quality of data are high. The data is reliable and salvageable.

B. Data Cleaning

1) Row/sample exclusion, logic Analysis, imputation: After manually examining all fields, it was found that there is no missing or insufficient data. It means that the dataset does not require filling some fields by blank/NaN values. Since the dataset hasn't data errors or coding inconsistency, the data cleaning procedure is not required.

However, as it was mentioned before, we excluded the 'Category' column because, in a real-world scenario, we did not know what type of incident would occur. At the same time, we limited rows (records) by 2018-2021 years because older historical data are not relevant for predicting contemporary security incidents.

We do not need to exclude any other rows or columns because the dataset is complete. Also, there are no data errors. To assess the dataset, we firstly used logic analysis, then implemented Google Colab.

2) Cleaning report: After removing the 'Category' column and rows, we have a dataset that consists of only two columns ('Dates' and 'Customer') and 9543 rows (records) for 2018, 2019, 2020 and 2021 years. We do have no other issues that should be reported.

a) Types of noise in the data: There is no any noise in the data.

b) Approaches to remove the noise: We do not have approaches to remove the noise, because of absence any noise.

c) Cases or attributes that could not be salvage: There is no cases or attributes that could not be salvaged.

C. Constructing New Data

1) Deriving attributes: We have constructed four additional attributes from a date value to ensure the flexibility of working with data. New columns "Hour", "Day of Week", "Day of Month", and "Month" are shown in Figure 9. We returned shortly to the Data Exploration stage to analyze various distributions for new attributes that are demonstrated in Figures 10, 11, 12 and 13.

D. Integrating New Data

There is no integration of data.

1) Merging or appending: There is no merging or appending of data.

We do not need merging or aggregating datasets because datasets are gathered from the Incident Response Platform as a single file. At the moment, the available attributes and records are sufficient to build an effective mode. We will split the data into training and testing datasets.

IV. Modeling

A. Modeling Techniques

Our modelling is conducted in multiple iterations to define the most accurate and appropriate model for the data mining goal. In the period of modelling, we have experimented with various parameters of the model.

According to available data types and data mining goals, we experimented with five modelling techniques. These are Naive

	Hour	Period	DoW	DoM	Month	Customer
0	22	7	3	29	4	E-GOV DATA CENTER
1	21	7	3	29	4	OFFICE OF THE PRIME-MINISTER
2	18	6	3	29	4	HOUSE OF MINISTRIES
3	18	6	3	29	4	HOUSE OF MINISTRIES
4	18	6	3	29	4	HOUSE OF MINISTRIES
...	...	...	...	...	...	...
9539	0	0	0	1	1	AKIMAT OF THE NUR-SULTAN
9540	0	0	0	1	1	AKIMAT OF THE NUR-SULTAN
9541	0	0	0	1	1	MINISTRY OF DEFENCE
9542	19	6	6	31	12	E-GOV DATA CENTER
9543	18	6	6	31	12	MINISTRY OF DIGITAL DEVELOPMENT

9544 rows × 6 columns

Fig. 9. Derived new attributes

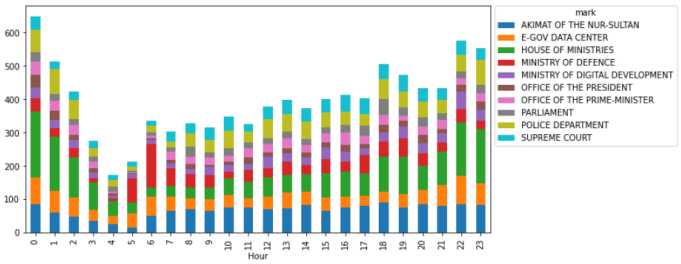

Fig. 10. Distribution of incidents by hour and customer

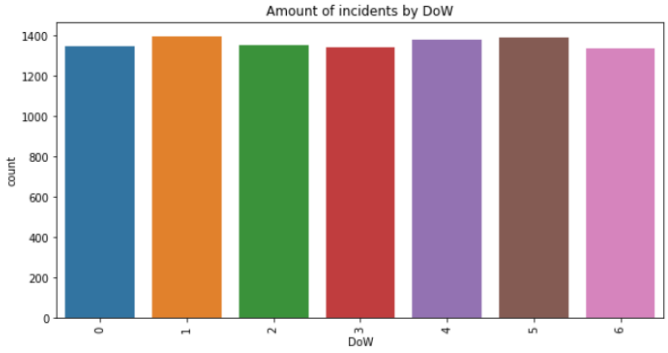

Fig. 11. Distribution of incidents by week

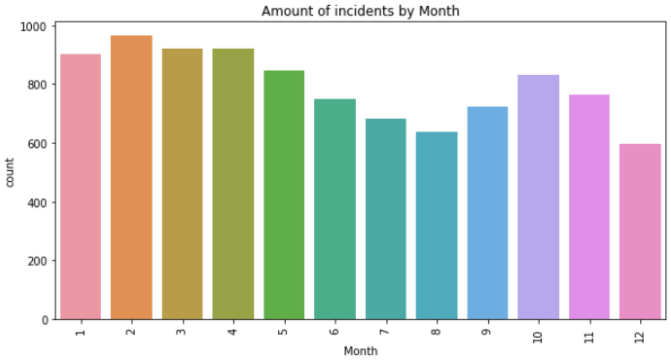

Fig. 12. Distribution of incidents by month

Bayes, KNN (k=3), Decision tree, Random Forest, Multi-layer perceptron.

1) Data types: After data preprocessing, we derived the following data types: 'Hour', 'DoW', 'DoM' and 'Month' are intervals, and 'Customer' attribute that we predcit (i.e. it is a class) is nominal. To ensure correctness, we need to convert nominal data to numerical values.

2) Data mining goals: The data mining goal of this project is to create a model that predicts the customer where cyber security incidents will occur by time data. In case of successful prediction, it will provide a timely response and prevention of damage. In general, the annual amount of cyber insurance payments will be reduced, so the company's expenses will decrease.

3) Requirements: Our model requires a certain level of data quality, and we can meet this level with the current dataset.

4) Data splitting, data size, data quality: The prepared dataset that consists of 9543 rows (records) have been split into 80% for training and 20% for testing subsets. Such decision-making can be explained that data until 2018 is irrelevant to predict today security incidents. We believe that amount of data is large enough to build an adequate model. There are some data quality problems related to the absence of some records due to technical issues. Then we will analyze it on cost/benefit to ensure that our expenses are less than potential benefits.

B. Modeling Assumptions

1) Generating test design: Criteria for goodness and assessment: Since the prediction of place by time is a complex task, we have very light criteria for the model's goodness that are presented by error rate or accuracy. We will declare that the designed model is good when its accuracy is more than random guesses.

However, relating to an economic benefit and business goals, the model might be deployed if its accuracy is not less than 0.25. We will build several models based on various Machine Learning algorithms. Then, the one with the highest accuracy rate will be chosen.

2) What data?: In the data preparation stage, we have split our dataset into training and testings subsets. We will assess models on a testing subset, which is 20% of the total amount of records.

3) Iterations?: We will provide several iterations of model building to ensure that all model parameters are picked correctly and provide adequate results.

C. Building Models

1) Model description: We utilized five various models with default parameters of the classical Machine Learning approach. They are Naive Bayes, KNN (k=3), Decision Tree, Random Forest and Multi-Layer Perception. We sequentially apply these models to our data and assess the accuracy. After

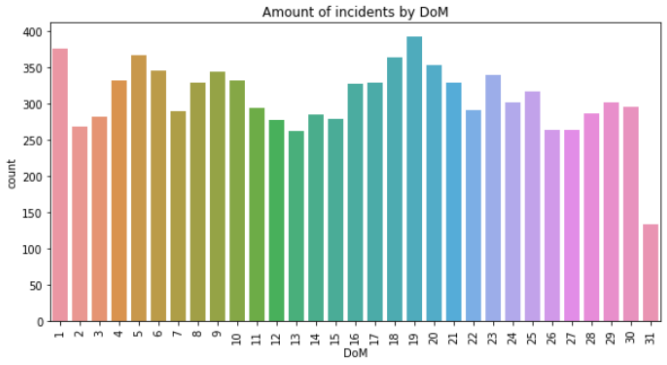

Fig. 13. Distribution of incidents by day of month

the first iteration, we fine-tuned these models to ensure that the models provided the best accuracy.

The meaningful conclusion from the built model is that various types of Network Attack, Ransomware, Security System Fault have a low prediction rate. We believe that this fact relates to its rarity. The confusions matrix demonstrates several patterns: "Unauthorized access" (class 8) is often mistakenly predicted as "Internal offender" (class 3). The correlation between these events might explain this.

During the first iteration of the project, we selected all available records to build a prediction model. However, we receive a very low accuracy rate. After analysis, we discovered that we should exclude very old historical data due to its inadequate prediction of current security incidents. We returned to a Data preparation stage and limited the dataset to a records for 2018, 2019, 2020 and 2021 years.

D. Assessing the Model

1) Evaluate the results of your model: After fine-tuning, we received accuracy for various models provided in Table I. We might observe that the Random Forest model provides the best results, which represents the ensemble algorithm. The Confusion Matrix of this algorithm is shown in Figure 14. Since one more model, based on a Decision Tree algorithm, satisfies success criteria, it might be reserved as a "Plan B" variant. The confusion matrix for this model is demonstrated in Figure 15.

TABLE I
MODEL RESULTS

#	Method	Accuracy
1	Gaussian Naïve Bayes	0.2095
2	k-Nearest Neighbors	0.2478
3	Decision Tree	0.2551
4	Random Forest	0.2645
5	Multi-Layer Perceptron	0.2362

2) Review of the results based on your understanding of the business problem: The modelling results mean that we might predict the occurrence of the security incident for a specific customer using only current date and time attributes.

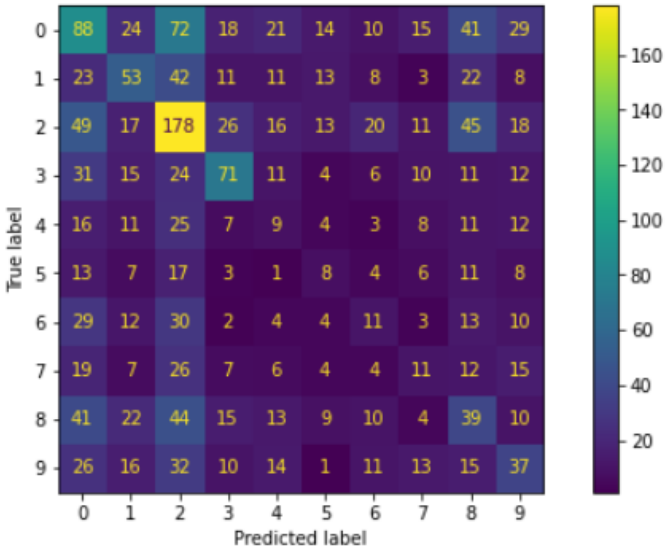

Fig. 14. Convolution Matrix for Random Forest

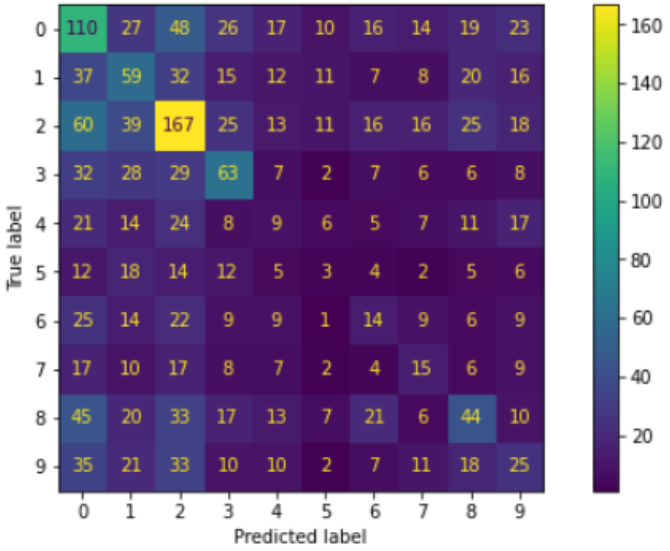

Fig. 15. Convolution Matrix for Decision Tree

We will predict for the following day by enumerating the 'Hours' attribute and will build a schedule of sending security officers to customers to prevent a security incident before it occurs.

Since we have only 0,26 prediction accuracy, around 74% of all calls would be false positive. However, there is economic expediency for the invention of the project. It means that the company will bear expenses.

It should be noted that it is impossible to deploy the permanent place of work at the customer's side because it requires an increase of the security personal up to 40 people (10 customers x 4 people) to provide a 24/7 service. At the same time, we need only 6 people (2 people x 3 eighth-hours period) that will be a mobile team that prevents security

incidents.

3) Model's results are easily deployable: The model's results are easily deployable because it does not require powerful calculations, and assigning of 6 security officers might be done from the current personnel.

4) The impact of results on your success criteria: The accuracy of the Random Forest algorithm is successful because it provided an achievement of the project's goals.

V. Evaluation

A. Results Evaluation

1) Summarise assessment results in terms of business success criteria: The proposed model meets our business objectives only about 20%. Because our chosen model Random Forest give us approximately 26%. Here, the deduction of expenses for the false calls model allows us to save up around $780,000.

However, since we have a false positive prediction, the company incurs additional expenses for mistaken calls. From the business understanding stage, we know that each call of a security officer for cyber incident response costs around $50 (one hour pay and road spendings). If we will calculate annual expenses for false calls the total amount will be around $220,000 ((1-0,26) x 6,000 incidents x $50).

Finally, we receive a total savings of $560,000 ($780,000 - $220,000), which correspond to our business goal.

We discovered that another model satisfies a set baseline. Since the accuracy of the Decision Tree algorithms is 0,25, the total savings for this model is around $525,000. This amount still satisfies our business success criteria.

2) Understand the data mining result: Our data mining goal was to build a model that successfully predicts where the security incident will occur by a time and date. During the Modelling stage, we designed a model that satisfies the data mining goal with an accuracy of 0.26. It means that we might correctly predict 26% of security incidents.

In case of applying of second model the prediction rate decreases to a 25% of total amount of security incidents.

3) Interpret the results in terms of the application: The results of the modelling stage mean that we might successfully predict 1560 of 6000 annual incidents the exact customer, where the security incident will occur. We might predict incidents for the following week and prepare the schedule of incident prevention. At the same time, since we have a 74% false positive rate, a vast amount of calls would be an error. However, since we predict only one incident per hour, we need two security officers minimally (one on-site incident and one arriving at the next). In the case of 24/7 service, the minimal amount of the incidents prevention mobile team is six employees (3 eight-hour periods by two people). Similar results are for second model.

4) Impacts for data mining goal: The data mining goal's successful impact consists of the transformation of a responsive approach to a predictive approach. Since the model's accuracy will increase over time, the project's positive impact will expand.

5) Check the data mining result against the given knowledge base:

6) Evaluate and assess results W.R.T. business success criteria: We have evaluated and assessed results and discovered that they correlate to business success criteria because the savings ($560,000 and $525,000 for two models) exceeded the amount set by the Board ($500,000).

7) Compare evaluation results and interpretation: Evaluation results and interpretation are correlate and have not any inconsistencies.

8) Create ranking of results W.R.T. business success criteria: The best results related to business success criteria are provided by a Random Forest algorithm, saving $560,000. In case of impossible to apply this model, we might utilize another based on the Decision Tree algorithm. The latter can decrease the total amount of cyber insurance up to $525,000.

9) Impacts of result for initial application goal: The impacts of results were checked, and it was discovered that they are correlated to an initial application goal.

10) Approved models: We might approve two models that differ by the amount of potential savings - $560,000 and $525,000 for Random Forest and Decision Tree, respectively.

B. Review Process

1) Review and highlight activities that have been missed and those that should be repeated: Since three attributes of the dataset are presented only by interval and nominal types, it was impossible to provide a five-number summary (min, max, Q1, Q3, median/mean). We also missed the new data integration step during the data preparation stage because we had enough data to build a prediction model. Moreover, since we have no broken data, the data cleaning process is also skipped.

2) Analyse data mining process: The data mining process was performed according to a CRISP-DM approach. The results were obtained adequately.

3) Identify failures: During the first iteration of the Modelling stage, we applied all available dataset's rows (records). However, our model predicts security incidents inadequately. We discovered that this issue is related to using very old historical records that are unsuitable for predicting current security incidents.

4) Identify misleading steps: We had not identified misleading steps that were performed during this project.

5) Identify possible alternative actions, unexpected paths in the process: We might identify a possible alternative way to implement the project. It relates to the usage of the 'Category' column as an addition to the date and time attributes. Since we have no information about what type of incident will occur, we might focus on two groups: a) security incidents with the most amount of damage, and b) most likely (in terms of probability) security incidents. However, we have not any information about relations between the type of the security incident and its damage in costs. It means that we use an average value of damage that correlated to a cyber insurance payment.

C. Next Steps

1) List of possible actions: Our next step of possible action is that the model will not be sent for revision or remodelling. If only there were unforeseen circumstances that will be demonstrated that the model is not working at well, the model would be sent to remodelling. Now, it is planned to deploy the model, and in parallel on an ongoing basis, it will be improving by training with new coming data.

We are against the referral for revision because we have to start testing the model in parallel. Even if the model does not function correctly, it cannot negatively affect the current business process.

2) Decision:

a) Potential for deployment of each result: As we mentioned before, with the proposed model, we receive a total savings of $560,000, which correspond to our business goal. Also, we have the technical capability to deploy the project. Because of this, we should coordinate the implementation with the board of directors and create a plan for deployment and integration, plan to disseminate this information to strategy makers, plan alternative deployment.

b) Potential for improvement of current process: To improve the current process, we should continuously monitor and maintain the deployed project.

c) Remaining resources to determine if they allow additional process iteration: We have enough resources to process the iteration and deploy the project. So we will deploy and rerun the model in parallel for verification.

d) Recommend alternative continuations: If the model's shortcomings are found, it is recommended to return the model even during the deployment process.

e) Refine process plan: The further plan of the process will be defined in the deployment chapter. At the same time, changes to the process plan will be promptly made only under unforeseen circumstances.

VI. Deployment

After modelling and evaluation phases, we have concluded that our model predict information security incidents with 26% accuracy. It is not a high result, but even this result allows us to save money up to $500,000. Also, gradually by extending the size of the dataset, our model's accuracy will be increased. Because of this, we have decided to deploy this project.

A. Planning Deployment

We are planning deployment according to the following steps.

1) Summarize your results—both models and findings: The best accuracy is about 26% we get from model Random Forest. We decided to deploy this model with the possibility of continuous improvement or replacement with another model.

2) Step-by-step plan for deployment and integration with your systems:

- After defending our project on the board of directors primarily, we should use the capabilities of the division R&D, prepare a user-friendly interface for the user.
- Then we must bring the Monitoring Service room all databases regarding incidences.
- These data should be automatically pre-processed, cleaned and converted to CSV files for subsequent continuous use as training and test data.
- The Monitoring Service room must establish those high-performance computers to apply a prediction model that uses the Random Forest algorithm. Also, this computer should be connected computer described in the above paragraph to get datasets for train our model and improve accuracy continuously.
- The prediction information should be online displayed on the big screen.

3) Plan to disseminate this information to strategy makers: To effectively implement this project, we should create a plan to disseminate this information to responsible employees.

- Firstly, there will create a group of employees from each involved structural division. The involved departments are Telecommunication Infrastructure Service, Monitoring Service, Finance&Marketing and Human Resources&Public Relations. The idea is that each involved employee should control the business process of the implementation of the project.
- If any problems or questions emerge regarding performance details, they can promptly discuss with a group of responsible employees and give appropriate directions.
- After implementing the model, data analysts should train monitoring workers to apply the model correctly and extract useful information.
- If the employees of the Monitoring Service discover the prediction of the information security incidents, they must notify employees of the Incident Response Services on time.

4) Alternative deployment plan: We are creating an alternative deployment plan for non-standard cases where can arise threat to disruption of implementation.

- In case of any local unsolvable problems, it is necessary to organize meetings at the management level to decide.
- In case of financial problems, the issue will be raised to the Board of Directors.
- In case of unrealizable technical problems, these aims should be changed to an alternative option.
- In case of a problem with personnel qualifications, the issues of attracting outsourcing specialists or companies should be considered.

5) Deployment will be monitored: The deployment will be monitored by data scientists and Finance&Marketing employees every quarter. Here primary accent will be the correctness of the prediction model and the effects of saving money on cyber insurance.

6) Deployment problems and plan for contingency: It should be noted that the original plan will not change, only changes will be made. Plan for contingency and deployment problems will be closely related to mentioned before alternative deployment plan.

B. Monitoring and Maintenance

The proposed model will constantly be monitored and maintained to make the model relevant for current data and the current situation.

1) Factors or influences to be tracked: Here need to be tracked continuous data flow from ERP. Because if the data do not expand datasets, our data will not be accurate. Also, we must track situations on the customer's side because the incidents may occur due to specific reasons. It is also necessary to monitor the data formats that are received for analysis.

2) Validity and accuracy of each model be measurement and monitoring: Our model's validity and accuracy every month will be measured and monitored by data scientists. We need to ensure that it is being used properly on an ongoing basis, and that any decline in model performance will be detected. Otherwise, we will incur significant financial losses.

VII. Conclusion

In conclusion, we would like to note that much data is required for successful implementation. Despite the low results of the model, it is possible to improve the results in the future. This project was implemented by the methodology of CRISP-DM, which helped reveal all the details of the project and get a general understanding of the implementation of such a nature of the project.

References

[1] "Cybersecurity Statistics for 2021: Packetlabs," Packetlabs, 2021. [Online]. Available: https://www.packetlabs.net/ cybersecurity-statistics-2021/ [Accessed: 03-Aug-2021].

[2] Chandrasekar, Addarsh, Abhilash Sunder Raj, and Poorna Kumar, "Crime prediction and classification in San Francisco City." http://cs229. stanford.edu/proj2015/228 _ report. pdf (2015).

[3] Sardana, Divya, Shruti Marwaha, and Raj Bhatnagar, "Supervised and Unsupervised Machine Learning Methodologies for Crime Pattern Analysis." International Journal of Artificial Intelligence and Applications (IJAIA) 12.1 (2021).

[4] Security Incidents Prediction [Dataset]. https://github.com/dyussekeyev/security-incidents-prediction.