Machine learning approach to support ticket forecasting from software logs

Matti Haukilintu

School of Electrical Engineering

Thesis submitted for examination for the degree of Master of Science in Technology.

Espoo 10.08.2022

Supervisor

Prof. Arto Visala

Advisor

MSc Petri Pyöriä

Aalto University
School of Electrical
Engineering

Aalto University
School of Electrical Engineering

Aalto University, P.O. BOX 11000, 00076 AALTO
www.aalto.fi
Abstract of the master's thesis

Author Matti Haukilintu

Title Machine learning approach to support ticket forecasting from software logs

Degree programme Automation and electrical engineering

Major Control, Robotics and Autonomous Systems **Code of major** ELEC3025

Supervisor Prof. Arto Visala

Advisor MSc Petri Pyöriä

Date 10.08.2022 **Number of pages** 42+1 **Language** English

Abstract

Your abstract in English. Keep the abstract short. The abstract explains your research topic, the methods you have used, and the results you obtained. In the PDF/A format of this thesis, in addition to the abstract page, the abstract text is written into the pdf file's metadata. Write here the text that goes into the metadata. The metadata cannot contain special characters, linebreak or paragraph break characters, so these must not be used here. If your abstract does not contain special characters and it does not require paragraphs, you may take advantage of the abstracttext macro (see the comment below). Otherwise, the metadata abstract text must be identical to the text on the abstract page.

Keywords Machine learning algorithms, Robotic process automation, log analyzing, random delay, Microsoft Azure ML Studio

Aalto University
School of Electrical Engineering

Aalto-yliopisto, PL 11000, 00076 AALTO
www.aalto.fi
Diplomityön tiivistelmä

Tekijä Matti Haukilintu

Työn nimi Sovelluslokien ja vikatikettien yhteyden löytäminen koneoppimista hyödyntäen

Koulutusohjelma Automaatio- ja sähkötekniikka

Pääaine Ohjaus, robotiikka ja autonomiset järjestelmät **Pääaineen koodi** ELEC3025

Työn valvoja Prof. Arto Visala

Työn ohjaaja FM Petri Pyöriä

Päivämäärä 10.08.2022 **Sivumäärä** 42+1 **Kieli** Englanti

Tiivistelmä

Tiivistelmässä on lyhyt selvitys kirjoituksen tärkeimmästä sisällöstä: mitä ja miten on tutkittu, sekä mitä tuloksia on saatu.

Avainsanat Koneoppiminen, koneoppimisalgoritmit, ohjelmistorobotiikka, loki, lokin analysointi, satunnainen viive, satunnaisviive, Microsoft Azure ML Studio

Preface

Terve,
ja kiitos kaloista.

Otaniemi, July 15, 2022

Matti Haukilintu

Contents

Symbols and abbreviations

Symbols

P	placeholder-symbol
A	another placeholder-symbol

Operators

$\nabla \times \mathbf{A}$	curl of vectoring **A**
$+$	yep, a plus

Abbreviations

AI	Artificial Intelligence
ML	Machine Learning
HML	Hybrid Machine Learning
RPA	Robotic Process Automation
SQL	Structured Query Language
JSON	JavaScript Object Notation
CSV	Comma-Separated Values
GDPR	General Data Protection Regulation
ADA	Anomaly Detection Algorithm
IoT	Internet of Things
UI	User Interface

1 Introduction

Artificial intelligence (AI) and machine learning (ML) has found their way into more and more fields of business. In banking business they are already used in fraud detection, risk management and service recommendations.[1] Even though these modern technologies utilizing big data are widely used abroad, AI and ML are not that popular in Finnish banking field. Instead, many self-acting solutions are being used to streamline manual labor which could be called intelligent, but are merely highly automated processes and thus cannot be included in the AI category. One of these technologies used in Finnish banking systems is Robotic Process Automation (RPA).

RPA operates "on the user interface of other computer systems in the way a human would do",[2] but is strictly bounded by predefined operations thus being prone to unforeseen situations such as faulty input. RPA, like generally all other software, produces log to "register the automatically produced and time-stamped documentation of events, behaviors, and conditions relevant to a particular system"[3]. Logs don't have any standards or form guidelines to follow which tends to make log analysis and log based problem-solving troublesome. This is also the case with RPAs developed by Oy Samlink Ab.

Oy Samlink Ab (Samlink from now on) was founded in 1994 and is now owned by Kyndryl. From the early years, while going by the name of Samcom, the company was owned by several Finnish banks developing all sorts of IT solutions for them. Nowadays, Samlink offers a wide variety of banking solutions from basic banking system to end user targeted software such as Codeapp mobile application.

Besides banking, Samlink develops multiple other IT solutions to extensive range of customers, for example entertainment platform solutions for DNA. Even though Samlink can be considered a modern technology company, the most modern AI technologies has not yet been adopted in the variety of tools used in development. However, RPA has been actively used in some banking solutions to reduce the amount of manual labor required from banking clerks.

In addition to continuous development as well as product maintenance services, Samlink also offers a technical help desk regarding the software solutions produced. As no IT solutions comes without bugs or misbehaviour, Samlink service desk has to use considerable amount of labor to resolve the possible reason behind the technical support request tickets received. In many cases, the problem-solving starts by reading the log and analyzing the data written by processes in question.

In this study, we aim to find if it is possible to utilize machine learning methods in analyzing logs created by Samlink RPA's. Ultimately, we intend to train ML which is able to predict the arrival of a technical support ticket thus giving a warning for developers about possible issues in the production.

1.1 Background and motivation

In the field of information technology logging is one of the most important methods in problem-solving, be it software or operating system related.[3] Typically, at least

in Samlink processes, logging is a bit more verbose than it needs to be. This is usually because when the problem occurs it is easier to already have the verbose logs available than trying to replicate the issue after setting logging to more verbose mode. Too verbose logging, however, leads into two problematic issues for developers.

First of, the size of log is huge and finding the critical parts related to the problem in hand takes more time. Of course, with more strict logging pinpointing the issue from within the logs would be faster, but then again, solving the problem with only critical error messages could be more time-consuming if crucial context is missing.

Secondly, a well-designed software is able to retry the process after first failure, but logging is done in real time, not after final results of the process has been determined. This means, that each process failure is logged even though said process eventually succeeds. Thus, logs may include dozens of rows of information about a problem encountered, which are not critical information after all. These issues make log analyzing considerably laborious.

Production logs are usually not viewed if everything is presumably working as intended. Technical support tickets are both last and most visible indicator that something is wrong. When a technical support ticket is received from banking clerks it means that something is wrong in a very visible way. Roughly speaking, technical tickets that are due to clear misbehavior of the RPA systems and not, for example, user errors, can be divided into two categories. First are the tickets that uncover an unknown bug in the system which can be either fixed or instructed to user how to avoid. Second type of tickets are somewhat pre-known issues that occur from time to time and are either fixed with updating parts of the system or by rebooting the process.

Typically, in software systems, if issue is known and can be fixed by rebooting something, developers can create log monitors that search for certain keywords and raise an alert if encountered. Developers can either run a reboot manually after a log alert has been received or set up an automated script to do it immediately when such keyword has been found. However, when it comes to RPA's and technical tickets considering them, it is hard to say what type of issue is in question by reading the RPA logs word by word without context. New kind of issues can be more frequent than already confronted ones, and clear keyword linked to a certain problem may not exist without considerable amount of false positive matches.

Machine learning algorithms are widely used to find patterns from massive amount of data making it an ideal tool for log analyzing. Patterns, however, need a connection to a visible issue to be useful. If RPA system has encountered an error but is able to retry successfully, then no issue has practically happened that needs immediate concern. Hence, RPA log analyzing with ML can find meaningful patterns only if they relate to actual technical tickets.

If Samlink support has received a help request the issue behind the request is not fresh anymore. In the event of RPA job failing, it takes some time for the clerk to notice the issue, write a help request to first support level, which then redirects the ticket to the corresponding team. Furthermore, if the issue is noticed during friday, it takes few more days to be handled by RPA developers due to weekend. This leads to noticeable delay in processes that were supposed to be dealt by RPA but which

now have to be manually taken care of by said clerks.

If a correlation between logs and tickets received exists and an ML algorithm is able to find it, it could be possible to create an ML-based log analyzer that can send an alert to developers about an ongoing issue before banking clerks encounter it. With automated scripts set up to receive such alerts, some issues could even be fixed in the production automatically without human interaction. This would reduce significantly the time and labor needed from developers and bank clerks alike.

1.2 Research objectives

This research aims to pave the way for machine learning application developers inside Samlink. Multiple obstacles need to be tackled as most of the phases in this study has not yet been encountered inside the company.

First and foremost, it is crucial to construct some basic rules considering the format of log data to make it usable by ML algorithms. Log data formatting is one of the key elements in automatic log analyzing applications as it is not for just machines but also for people to read.

As today more and more concern is set on anonymization not only due to GDPR, the data used for machine learning must be sanitized. Because of this, one major objective is to create a clean dataset that is safe to use in cloud environment without raising concern around security and privacy issues. In addition to this, data must also be clean enough so that ML algorithms are able to process it.

As mentioned, Samlink has not yet developed ML applications. In order facilitate deployment of applications for future ML application developers, this study aims to document the process of ML deployment well enough to create a simple guide to follow in the possible future ML projects in Samlink.

Finally, the main question this research is aiming to answer is: *is there such a correlation between RPA run logs and technical support tickets that ML algorithm is able to find it, and can this correlation be used to forecast a ticket arrival*?

1.3 Scope

In order to limit the study to feasible length and content, it is necessary to define the scope for the thesis. Before diving in to the scope of research objectives, we must first make one assumption regarding the data that ML is going to find some meaning from. In order to find a connection between log data and support tickets, we make a hypothesis that errors leading to tickets are visible to or parseable by ML algorithm. In addition, as we are going to utilize anomaly detection algorithm in log analyzing, we must also assume that these errors in log are, in fact, anomalies. We will, however, compare the results against these assumptions to test this hypothesis.

Data anonymization

Anonymization in the context of this thesis refers to data sanitization process purposed to edit the data into more secure form in the privacy point of view. In this study

we aim to create a dataset usable in ML training. In this respect, anonymization is not in the main focus of the study but only treated as a sub-phase of the data preprocessing in whole. Nevertheless, anonymization is from the privacy perspective the most important phase of data preprocessing.

Keeping this in mind, anonymization is covered rather superficially, only enough to explain the reasons behind actions taken during anonymization process.

Azure setup

The ML training and result scoring is done in Azure ML environment. Azure is used for ML processes because Samlink already had licenses in Azure Cloud that are used for RPA process control. Integrating existing Azure resources with ML pipelines and endpoints constructed during this study was seen as a big advantage. Thus, no other ML cloud provider was considered. However, other Azure competitors are mentioned to the extent that their existence is recognised.

As one of the objectives is to create an initial guideline for ML process commissioning the Azure setup phase is documented in such detail reflecting the importance of this information for future developers starting Azure ML projects in Samlink.

Data requirements

Data purity in a sense of how easy it is to be used by ML algorithms creates challenges at the beginning of ML training. If data is not consistent, has lots of missing values or is formed in unanticipated way, it requires considerable amount of preprocessing slowing the training process and causing errors in pipeline runs.

In order to create a baseline for Samlink ML projects this study aims to give basic criterion what is required from the data, so it is easily analyzable by ML algorithms.

Machine learning methods

Several different machine learning algorithms exist that are aimed for different applications in mind. For example, to make an algorithm that can predict the price of an apartment listed[4] we could be using linear regression, and in order to detect possible cyber threats from network traffic[5] a two-class support vector machine could be utilized. These two methods are very different in usage and has their pros and cons in different applications.

As different methods can be used in creative ways in very different applications depending on how the data is presented and how the ML problem is formed, this study focuses on just a few easily approachable training methods that were seen suitable to answer the study objectives.

When it comes to anomaly detection algorithms (ADA), only principal component analysis (PCA) is considered because PCA-based Anomaly Detection component is the only one usable from existing two anomaly detection algorithms in Azure ML Studio. As purely Azure ML Studio is used during this study, no other anomaly detection algorithms are debated. The other ADA-component, One-Class Support Vector Machine, is discussed briefly to explain its unsuitableness for current case.

1.4 Structure

In the **Introduction** section we explained the research motivation, main objectives and study scope. The next section, **Background**, clarifies the general machine learning concepts and terms relevant to this study case. We also discuss these topics from the perspective of existing studies.

The third section, **Research material and methods**, explains in detail the data format and contents as well as the steps used to sanitize and preformat it for ML algorithm training. In addition, resources needed to set up the ML designing and training environment are discussed. At the end of the section, the ML pipelines used in the study are unfolded as well as their contents explained in detail.

Results section reveals how the selected algorithms performed and how well the research questions could be answered.

Finally, in the section **Summary**, we summarize the research outcomes, evaluate the results, and discuss what could have been done better.

2 Background

Machine learning, or ML, is a subcategory of the AI field and data science. Typically, ML refers to a set of technologies used to "build computers that improve automatically through experience". [6] This is generally considered a machine way to simulate human learning process. ML usage has become more common and is nowadays widely used in many fields, not just in general information technology and computer science. This is because data can be gathered from anywhere, and where there is data to be processed, ML can be there to process it. Computer algorithms are able to find statistical correlation and patterns from places overlooked by human mind, or where amount of data is just too much for people to process. This is why ML has proved its power in various empirical science fields, such as biology, cosmology or social science. [6]

In this section, key concepts of ML are explained briefly and several ML features are explored that are most relevant to this study. We also discuss shortly about data sensitivity and how it had to be addressed during this study.

2.1 Machine learning algorithms and training

Algorithm means a finite sequence of (typically) mathematical operations that are used to solve a specific problem, generally by repetition of some steps until the problem resolves. [7] Algorithms are the main component inside machine learning By iterating through all the data points algorithm is able to, for example, find repeating patterns, mathematical or logical connections, or unusual anomalies that would be seemingly normal for human eye.

Algorithms operate on set of rules and parameters In order to utilize an algorithm to solve a problem, algorithm is first trained by tuning these parameters to fit the current case. Usually, ML algorithms can be trained in three ways: supervised, unsupervised, and reinforced learning. [6] Even more training methods exist that usually combine those mentioned. [8, 9] For the sake of simplicity, we focus on those three main methods.

In **supervised learning**, algorithm is given data with ready answers on how the data needs to be interpreted. Algorithm then tries to figure out the rules behind how given data and the correct answers are related. [8] In **unsupervised learning**, on the other hand, algorithm does not get model data from which to train itself, but instead it tries to find clusters or groups inside the data that are linked together more closely than to other data points. [4] **Reinforced learning** refers to a method where a computer program is given a goal and provided feedback as a reward. This reward is what program aims to maximize by adjusting given parameters. [8]

In ML, there are multiple algorithms to solve different problems and no jack-of-all-trades algorithm exists. Each algorithm is suitable for certain type of problem. To simplify, algorithms are usually divided into three or four categories based on the problem type. [10]

Regression algorithms predict values and are typically used with supervised learning. Usual example of regression problem is house price prediction using typical

house features such as building year,location, number of rooms etc. . With these varying features the algorithm then gives each feature a weight value which determine the final price of the house. [10]

Classification algorithms predict categories and are also used most commonly with supervised learning. Depending on the algorithm, they can predict between two or several categories. Examples of classification problems could be spam mail identification with two class classification, or flower species recognition from images with multiclass classification. [10]

Clustering algorithms use unsupervised learning to find structures inside data. This is done, for instance, by first providing the amount of clusters to search to algorithm, which then calculates a center point for each cluster so that they are as far away from each other as possible while data points surrounding each center are as close to each other as possible. [9] This could be used, for example, to find meaningful customer segments from transaction data in order to improve targeted advertising. [11]

Dimension reduction algorithms are a separate type of algorithms used with unsupervised learning, but they are usually combined with other algorithms to solve the main problem. With dimension reduction, main algorithm calculations are streamlined by first reducing the amount of feature dimensions. [12]

These four ML problem types and most known algorithms of each type are shown in the graphic 1.

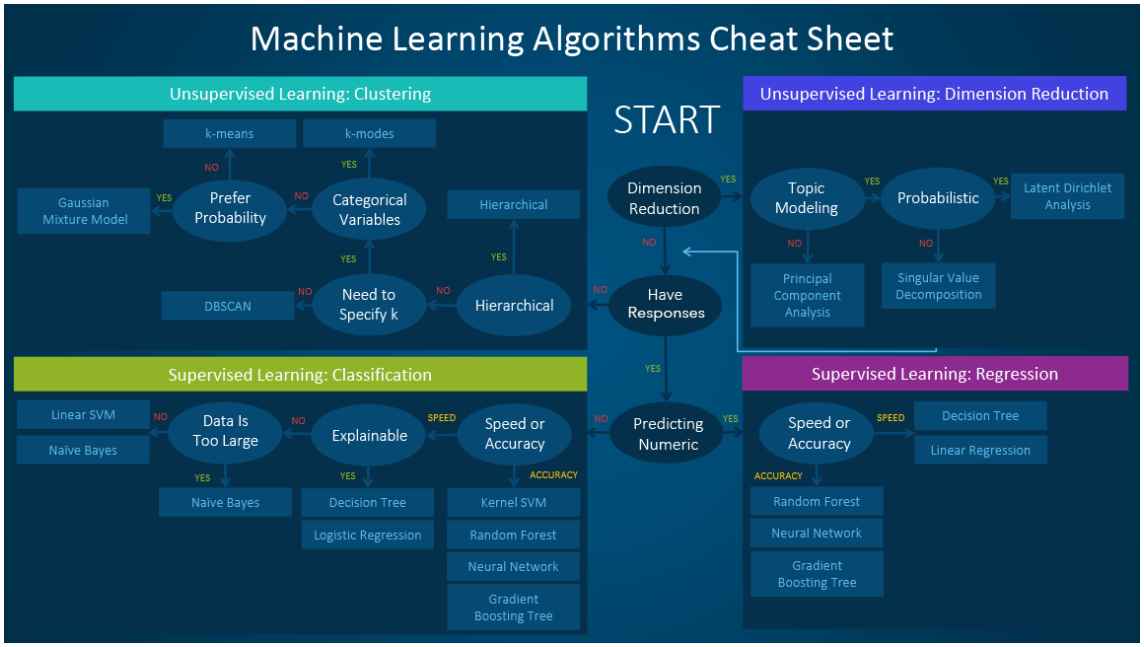

Figure 1: Machine learning cheatsheet for algorithm choosing[12]

This study focuses on anomaly detection, which, roughly simplified, is a clustering problem where anomalies are rare incidents outside common clusters. However, in this study we utilize PCA-based anomaly detection algorithm, where PCA refers to Principal Component Analysis, and which is a dimension reduction algorithm. [12]

More about PCA is discussed later in this section. In addition, we aim to find a connection between anomalies and incident tickets by their amount in a timeframe, which makes the topic in the end a regression problem.

Typically, when training an algorithm, some predefined portion of the data is used as training data. The rest is used to validate the results so that validation data and training data do not overlap. Instead, trained algorithm is given data it has not seen before and the result it produces with it is then validated. [13] For example, in supervised learning the key values the algorithm is trained to find out are hidden from the validation data. The resulting values produced by the algorithm are compared to those hidden values and the difference between the estimate and the real value can be used to determine how well the current trained algorithm compares to others. However, in this study, we are going to break that rule about non-overlapping training and validation data. The reason for this is explained further in section 4.2.

2.2 Cloud ML platforms

Machine learning algorithms are not light to operate. ML is at its best with big data where amount of data points makes it easier for algorithms to find repeating patterns more reliably. [14] Data amount, however, requires huge resources in terms of memory and computing power. Especially with online applications where real time analysis of new input data is required with small latency, cloud computing can make a big difference in terms of processing speed.

Online market offers several solutions for ML computing in cloud. Most notable service providers for MLaaS (Machine Learning as a Service) are Google, Amazon, IBM, and Microsoft. Differences of each service provider are listed in a table 2.

Amazon's new SageMaker service has replaced the old Amazon Machine Learning service. It is very much like Azure Machine Learning service produced by Microsoft. Azure, however, has one major advantage which is the ML Studio environment. Most of the MLaaS provider's solutions have some sort of no-code to low-code design features which makes pipeline designing easy. Azure ML Studio lets the developer design and deploy full ML pipelines with drag-and-drop user interface. Compared to SageMaker and Azure, Google AI Platform is missing anomaly detection and ranking abilities. IBM Watson has even less features, as demonstrated in the table 2. [15]

2.3 Azure ML Studio

Microsoft Azure offers a Machine Learning Studio environment for easy ML pipeline designing. ML Studio gives ML designer a possibility to train algorithms and publish cloud endpoints utilizing all Azure resources connecting the power of ML to all other Azure features like data storages, IoT-services, and cloud computing. [15, 16]

Each component in pipeline can be tuned to a certain extent. ML Studio has a predefined set of ready algorithms to use. Example of Azure ML Studio interface is shown in figure 3. Data to the ML Studio environment can be imported from local storage, but also from various other Azure services such as storage accounts with table and blob data. Trained ML pipeline can be inserted into wider operation chain

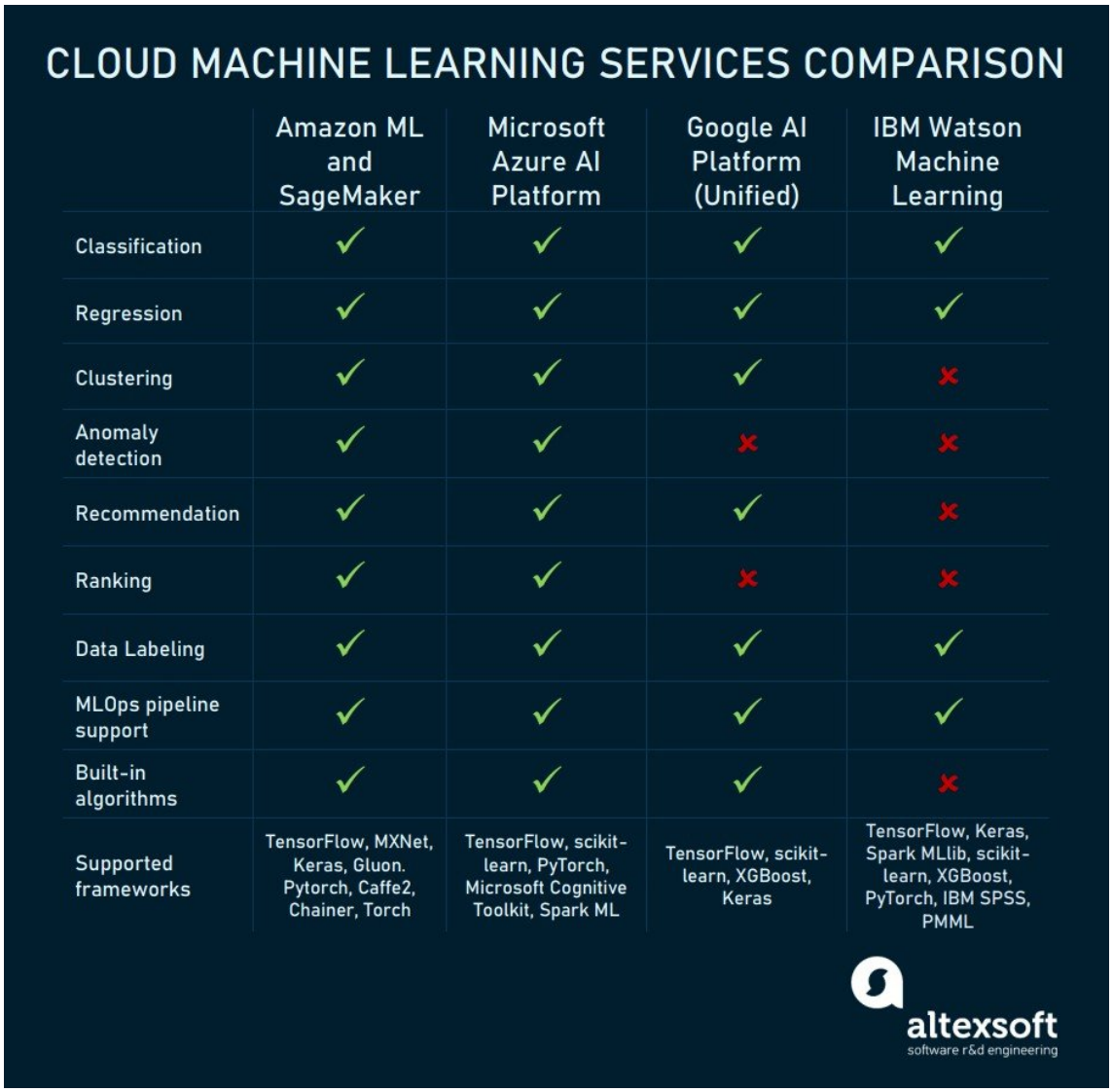

Figure 2: Machine learning as a Service comparison. [15]

combining other Azure services to it. This allows designer to use ML computing capabilities with existing production environments utilizing services like IoT, API, or Kubernetes.

2.4 Regression algorithm

`TODO: science and math behind regression algorithm`

Regression analysis is typical approach in statistical science. It is used to find relationships with a set of variables.

2.5 PCA-based anomaly detection

`TODO: Explain PCA and mention other ADA algorithms`

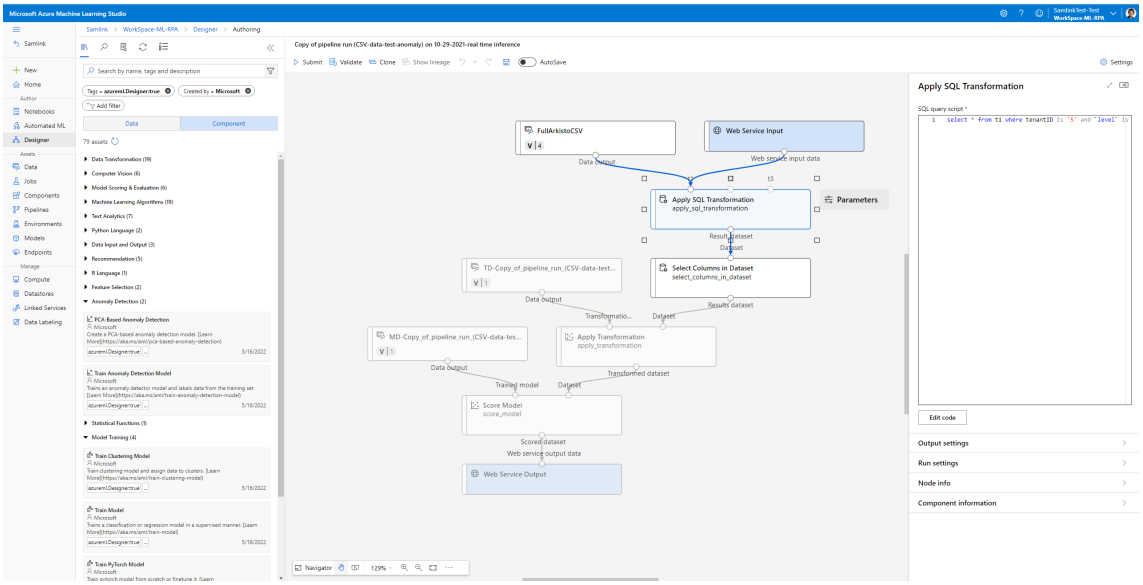

Figure 3: With drag-and-drop pipeline designer it is easy to get started with ML programming in Azure ML Studio, and visualizing the process helps understand all pipeline components and their relations to each other.

Principal Component Analysis, or PCA, is a machine learning technique used to analyze data and explain the variance inside it. [17]

Other anomaly detection methods exist, but they are not supported by ML Studio in a ready component level.

`TODO: Something about Anomaly and Novelty detection differences?`

`TODO: placeholder picture. replace with mathematical explanation`

One-Class support vector machine

Azure ML Studio has also another anomaly detection algorithm to use. This module is called One-Class Support Vector Machine. In our case, however, this module was not deemed suitable as the documentation mentioned that "The dataset that you use for training can contain all or mostly normal cases." Because the content of the data used did not meet this requirement, the usage of this component was decided to skip.

2.6 N-gram features and feature hashing

`TODO: cover basic n-gram features and ML connection`

Comment: unorganized text below:

As stated before, features are the key elements in ML algorithm training. As textual input does not have any meaning to machines as itself, it is necessary to create a

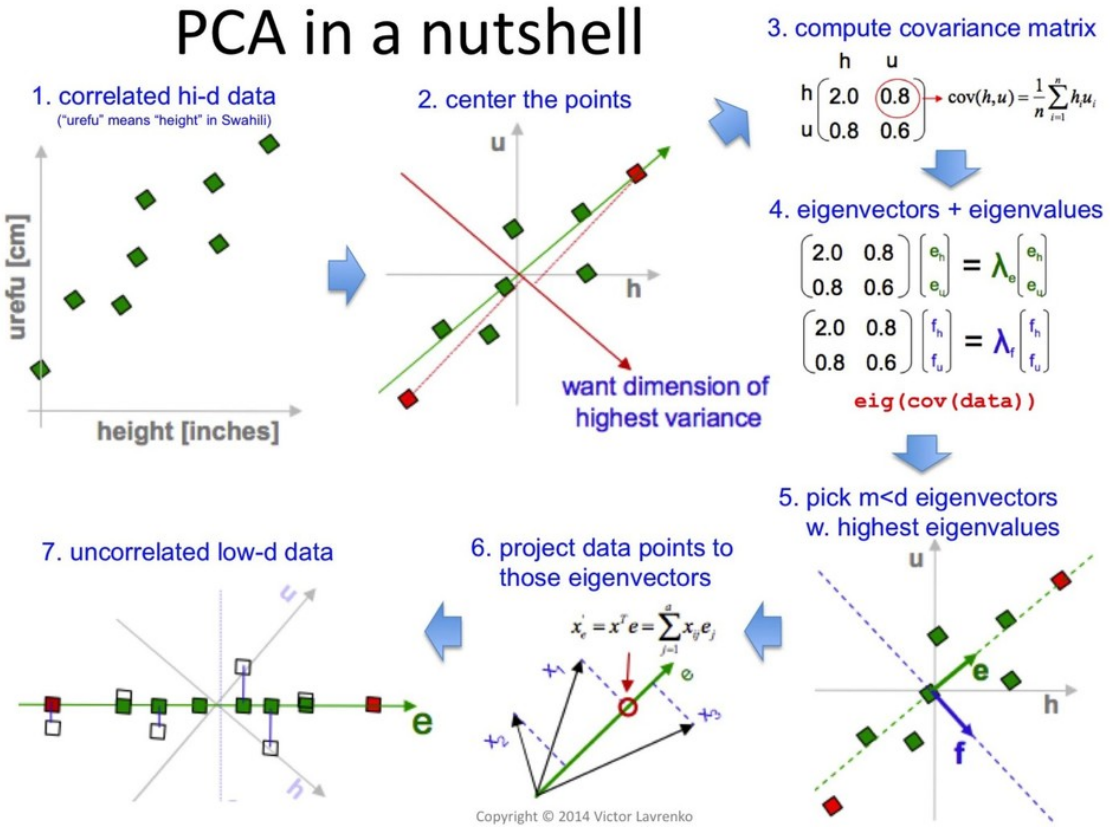

Figure 4: PCA in a nutshell

connection between words and features for algorithm. In ML training, one typical approach is to convert textual input to numerical features. For example, by creating a dictionary of words used in the input and assigning each word an identification number, we can express words as a count of certain words used. In addition, as words include meanings not only individually but also with relation to each other and in their order, we can add more information for the algorithm by creating word pairs and groups in the dictionary. These groups are referred as word grams, where **n** in n-gram refers to the maximum number of words in a group of consecutive words in the input sentence.

`TODO: explain feature hashing component functionality briefly`

As the number of word grams in a dictionary can increase significantly in complex input cases, it is necessary to limit the resource usage by decreasing the features analyzed. One way to do this is use feature hashing. This means that instead of pure n-gram count we use hashed value of several n-grams thus reducing the amount of features. As a drawback, the amount of information might also get reduced as the data is "compressed" but this way we can include more features for algorithm training without significant resource demands.

2.7 Robotic process automation in Samlink

TODO: Short explanation of RPA in Samlink, mostly to clear out the terms used in this study

Robotic process automation, or RPA, is used to automate mechanical tasks. Usually it operates on the UI level and can be used to repeat meaningful functions instead of mechanical actions. For example, with screen recording macros only position at the screen and mechanical key pressing is recorded and repeated. RPA automation, however, is able to repeat the functionalities those actions trigger, such as inputting text to a certain named field on the UI, or logging in with given username and password regardless of the location of those fields on the layout.

TODO: references!

In Samlink RPA operations, a central coordinating system called "Orchestrator" supervises the RPA processes.

TODO: Add reference to hierarchy picture

TODO: Dummy picture, replace with better

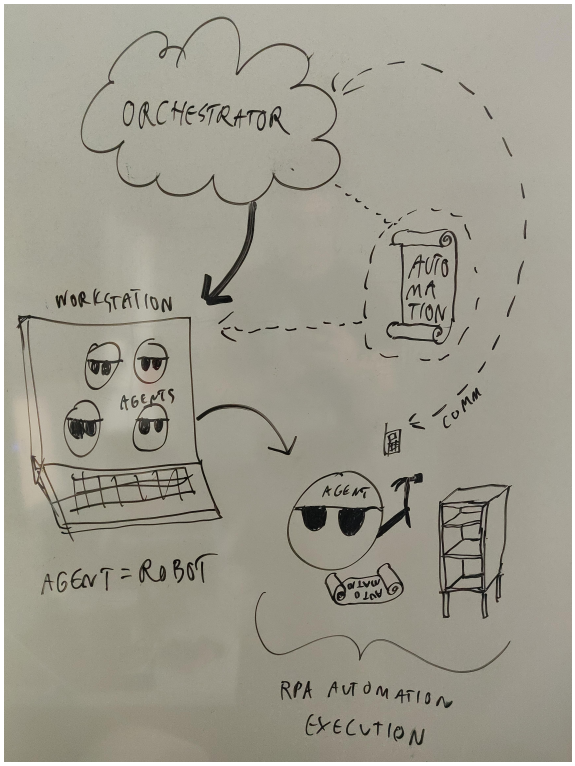

Figure 5: TODO:! Dummy. To be replaced with proper picture. Hierarchy of RPA components explaining the terms and their relations

2.8 Data sensitivity

During this study, it was necessary to make sure no sensitive data was moved out of the production environment. This was mostly due to restrictions imposed by GDPR. In order to maintain the data security, data had to be anonymized before it could be exported to the cloud environment. After anonymization, data would not include any information that can be connected to real individuals. Three different anonymization methods were considered, which were pseudonymization, k-anonymization and full anonymization.

Pseudonymization refers to a method where sensitive information is de-identified. This means, that each sensitive piece of information is replaced with a decrypted value so that no information is lost but human cannot identify individuals when reading the data. Decryption and de-identification could be reversed, i. e. data could be re-identified, with decryption key which tells computer how to convert the replaced value back to the original form. As machine learning algorithms do not care about the meanings behind personal identification information, such as phone numbers or addresses, pseudonymization would preserve the information in the data unchanged for ML algorithms to use so that no information would be lost. [18]

As pseudonymization is reversible operation with encryption key, it is not the safest way to anonymize the data because encryption key leaking is always a risk. K-anonymization is the next step to secure the data sensitivity. Excluding all unique identifiers such as full name or social security number, information like home street, age, workplace or last name are not on their own enough to identify certain individual, but combined they can single out a person. K-anonymization is unreversable anonymization approach where identifying information is generalized to mask individuals into crowd. With k-anonymization, algorithm replaces single informative details with more general variants, for instance, address to hometown or age to age range. K-anonymization loses information as it cannot be reversed. If personal information is essential for the use case of ML algorithm, this method weakens the algorithm results. [19]

However, because it was determined that individual information in the log data used in this study was not relevant for connecting the log events to technical support ticket timestamps, full anonymization was decided to execute on the log data. This way, each personal information was replaced with a general token disclosing only what type of information (phone number, email address etc.) was anonymized.

Data anonymization was executed in production environment with PowerShell script. Several predefined identification features were searched with regular expression (or regex) patterns and replaced with default keys.

2.9 Log data analyzing and anomaly detection with ML

Using machine learning for log data analyzing is not a new field of study. [20, 21, 22, 23] Key issue tends to be the format of the log data which shifts the dilemma to natural language processing. Some studies also combine anomaly detection using machine learning to log data analysis. [24, 25] Comparing existing studies to our case raises

Info type	**Example**	**Replaced value**	**Comment**
Social security number	010190-0123	10105051470101	Includes '-','+' and 'a' format
Email	author@thesis.fi	EmailAddress0101	
IBAN number	FI8612345600000123	1010IBANnumber0101	Only Finnish format
BBAN number with dash	123456-123	1010BBANnumber0101	
Phonenumber, international	+358501234567	1010PhoneNumberInt0101	With or without whitespaces
Phonenumber, local	050-1234567	1010980230101	With or without whitespaces or dashes
Business ID	1234567-8	1010BusinessID0101	Finnish format
Business ID, international	FI12345678	101086512350101	
Business ID, int. zero form	0012345678	1010865123500101	
Credit card number	4920191061682346	1010664900101	
Windows Identity	K123456	1010WinID0101	Used in company processes
Address, common	Teekkarikuja 1 a 42	1010AddressCommon0101	Common street name endings
Address, ZIP	Bulevardi 2 B 69, 00100	1010AddressZip0101	Disregards city name after ZIP
Bank ID	12345678	10108426100101	Bank user ID
BBAN without dash	12345600000123	101088420101	
Artificial business ID	8123456789	101086512354970101	Used in RPA processes

Table 1: Information replaced with Regex search from log data. Data values are replaced with patterns with numbers or numbers and letters depending on the original format in the data. Patterns are formatted uniquely so that they can be recognized amongst the anonymized data, each starting with 1010 and ending with 0101, and having a typewise identifier in the middle. With numeric patters, numbers are selected as letter representations, like business ID = 8651235 (BUSINES)

Info type	**Regex**
SSN	`?<![a-zA-Z0-9])[\d ]{6}[-a+]?[\d ]{3}[\w ]{1}(?:0{0}|0{3})(?![a-zA-Z0-9])`
Email	`[^\‘"\s ]+@[\.\w -]*[\w ]`
IBAN	`(?:(?<![a-zA-Z0-9])|(?<=\\\D ))(?:FI|fi)(?: ?\d ){16}(?![a-zA-Z0-9])`
BBANwith-Dash	`(?<![a-zA-Z0-9])[\d ]{6}-[\d ]{2,8}(?![a-zA-Z0-9])`
PhoneInt	`(?<![a-zA-Z0-9])\+358(?: ?\d ){8,10}(?![a-zA-Z0-9])`
PhoneLoc	`(?<![a-zA-Z0-9-])[0][\d ]{2,3}[ -]?(?: ?\d ){6,8}(?![a-zA-Z0-9-])`
BusinessId	`(?<![a-zA-Z0-9])[\d ]{7}-[\d ]{1}(?![a-zA-Z0-9])`
BusinessIdInt	`(?<![a-zA-Z0-9])[a-zA-Z]{2}[\d ]{8}(?![a-zA-Z0-9])`
BusinessId-IntZero	`(?<![a-zA-Z0-9])[0]{2}[\d ]{8}(?![a-zA-Z0-9])`
CreditCard	`(?<![a-zA-Z0-9-.])[\d ]{1}(?: ?\d ){14,15}(?![a-zA-Z0-9-])`
WinId	`(?<![a-zA-Z0-9])[a-zA-Z]{1,2}[\d ]{6}(?![a-zA-Z0-9])`
AddressCom	`[^\s ""’,.]* ?(katu|tie|kuja|polku|kaari|linja|raitti|rinne|penger|ranta|väylä|taival|tanhua|portti |veräjä|laita|reuna|syrjä|aukio|tori|laituri|tunneli)[\d ]{1,3}( ?[a-zA-Z.]{1,4} ?[\d ]{0,3})?(?!\w )`
AddressZip	`(?<=\s )[\S ]* [\d ]{1,3}( ?[a-zA-Z.]{1,4} ?[\d ]{0,3})?(\s |,\s )[\d ]{5}(?!\w )`
BankId	`(?<![a-zA-Z0-9-])[\d ]{8}(?![a-zA-Z0-9-])`
BBANnoDash	`(?<![a-zA-Z0-9])[\d ]{14}(?![a-zA-Z0-9])`
ArtifBusines-sId	`(?<![a-zA-Z0-9])[89]{1}[\d ]{9}(?![a-zA-Z0-9])`

Table 2: Regex search patterns for sensitive info finding. Most of the regex patterns start with negative lookbehind and end with negative lookahead so that found pattern is not part of another string. Order of the regex patterns is as shown on the table as some patterns gives overlapping matches. By searching shorter patterns before longer it is possible to recognize the type of anonymized information with higher reliability.

at least one major suggestion for improvement: log data refining.

When log data has consistent format, multiple different algorithms can be utilized for anomaly detection and log analyzing. Log events can be clustered and different types of events can be counted if the amount of types is finite and known. [24]

If training data already have information we wish to teach the algorithm to forecast (i. e. data is labeled), combining the results of the log data analysis to external features is more feasible. With labeled data, supervised learning methods can improve the results of the algorithm forecast abilities. [20]

As explained in the section 3.1, data features connected to anomaly detection results are pure datetime values. With more insight to ticket data properties, ML algorithms could be able to extract more valuable information from the log data.

2.10 Random delay in log event analyzing

`TODO: Anything about the topic?`

Random delay in input data features is not unusual aspect in time-series forecasting. Time-series in ML context refers to data features that varies over time and is usually affected by past values. As an example, ML algorithm could try to predict future weather based on measured temperature and air pressure. Both these features change over time and also affect to their own future values.

Random delay in such an example could be due to some other local or global features like

This study, however, is not time-series because the majority of the log rows are not affected by previously logged events. Random delay in this case is caused by banking clerks finding the issue and writing a technical support request. This delay can span from hours to several days depending on the weekday and time the issue occurred.

As random delay of such does not seem to be trivial to take into account with ML algorithms, a simple method to solve this was used which we call “time frame compression method”. More about this approach is discussed in the section 4.1

2.11 Hybrid machine learning approach in anomaly detection

Hybrid machine learning (HML) refers to an ML technique where two or more ML methods are combined to overcome the limitations of or to boost the estimation capabilities of a single method alone. [26] In this study, we combine PCA-based anomaly detection algorithm with regression algorithm in order to amplify the prediction powers of our ML algorithm when trying to determine the possible ticket count based on log events.

Hybrid machine learning is not rare technique in ML field. [27, 28, 29, 30, 31, 32, 33, 34]

`TODO: Something about the HML in ADA`

In order to clarify whether hybrid approach is suitable for the current study problem we will compare the results of hybrid ML technique with a single ML algorithm usage.

`TODO: Include wireframe model about hybrid model`

With ML algorithm utilizing n-gram features combined with time frame compression it is possible to get estimates about the support tickets based on the log events. It is not feasible to use anomaly detection on its own to do this as plain sum of anomalies detected is not correlating with tickets received.

We can, however, amplify our ticket estimating algorithm with anomaly value features. As we first count the anomaly numbers with anomaly detection algorithm and their calculated statistical features with another algorithm, like regression algorithm, we get more relative information to use when creating the final ticket number estimations.

Comment:

Hybrid ML as a term is used here to explain that we use two different ML algorithms in two separate phases. In first phase we try to give an anomaly certainty value for each log row using PCA-based anomaly detection component. In second phase we use this value as a feature to estimate ticket amount in time range by utilizing regression algorithm.

Dual algorithm approach should not be unusual in ML field, but how existing studies or case examples relate to our way is uncertain.

If nothing exists about the topic (at least nothing easily to be found) it should be worth to mention. But if there is a lot of case examples about this, it feels unnecessary to discus about it in more detail.

3 Research material and methods

In the next section we explain in more detail what the data used in the study consists of and what methods were used in attempt to answer the research goals. The content of the section is briefly described below, and the steps of the research are explained.

The data in the research is mainly made of two parts. The most important part is, obviously, the log data produced by the numerous RPA processes. The second data part, complementing the study, is the support ticket data written by clerks of customer banks. In order to use the data safely in the cloud environment it was necessary to sanitize the data from any sensitive information. This was done by anonymizing the log data and using only timestamps from the support tickets.

After confirming the results of anonymization, the data was preprocessed into a better form to make it more usable by algorithms. More processing was done inside the pipeline as ML Studio offered several usable components for this but main cleaning was easier and to execute in local environment. This was also done with PowerShell scripting.

The actual ML pipeline structure is discussed in the next section.

3.1 Support ticket data

Like all other software, RPA components fail from time to time. As described before, RPA logs are verbose making possible error identification from among them hard. Due to that, it is not feasible to create log parsers that would be able to identify critical errors from within thousands of lines of log. When critical error happens causing the RPA process to fail, the banking clerks need to finish manually the job left by the RPA robot. Every time this happens, these clerks then send a support request ticket to Samlink technical help desk and ask to fix the issue.

When clerks send the ticket to technical support a verbose description of the situation is written to help developers to identify the problem. This description often contains sensitive end customer information like bank account details and social security numbers. To avoid privacy issues when processing this data, it was decided to use only timestamps of the tickets. The resulting data was practically a list of date and time values. More about the issue from privacy point of view is described in section 3.3.

3.2 RPA log data

Robotic process algorithms used in Samlink are designed to ease the workload of bank clerks. RPA robots work in behalf of bank clerks executing routine tasks that require mostly manual labor.

Like other software, RPA also produces log data during runtime. As dozens of RPA automations are running in several bank environments the amount of log entries produced is also significant, up to over a million lines per week. This log data is not in consistent structure as it is formed out of typical CSV data and injected with even more inconsistent JSON data that varies in contents vastly.

TODO: refer to appendices, include examples of data

RPA log data is stored in SQL database. The database is split in live production log that is gathered for two weeks and then moved to archive that has several years worth of log. In this study we used archived data as it was easier to acquire in one run without the need to merge different parts together. Archive also had data that was considered as sufficient amount for machine learning algorithm training, with data entries spanning almost two and a half years and rowcount exceeding 80 million.

Samlink RPA logs have few standard fields. These are listed in a table below 3. The most notable aspect of these fields are the *message* and *rawmessage* fields. *Message* holds the short log message written by the RPA agent during automation execution. This includes details about the issue, what part of the process failed, and possible stack trace of the error.

Rawmessage is JSON-formatted representation of all the default features, including the message and multiple other additional features that RPA agent is able to output regarding the execution. These additional fields are what vary from log entry to log entry. Some of the possible fields are shown in a table 3, but several other field types exist, and the JSON data in them can be nested in multiple layers.

Without rawmessage, the data was in pure CSV-format. However, it was not certain that rawmessage would not hold usable data for ML . In the end, the usage of rawmessage was not studied in satisfying extent as it demanded too much resources in ML Studio to process.

TODO: Maybe some more info about the log data? Open up the table?

3.3 Data anonymization

Support ticket data privacy

Samlink handles highly sensitive banking customer data in its processes, such as personal identification numbers, home addresses, email addresses and bank account numbers. All possibly sensitive data had to be removed before data could be transferred out from production environment to cloud. Due to bureaucratic reasons, technical support tickets were under more strict policies. Because of this, they were allowed to be used in the research on condition that no business critical nor customer sensitive information was processed in the first place. Only way to assure this was to select solely timestamp fields from ticket data. Thus, no sanitation for ticket data was needed as ticket data consisted of only list of datetime values.

RPA log data sanitization

Information privacy is one of the key values in Samlink business promise as company develops high security banking applications and processes sensitive customer data. Thus, several aspects were needed to take into consideration before log data could be authorized for thesis study usage. To improve privacy, it was decided to assume that personal customer details are not critical information for ML algorithm training if

Field	**Contents**	**Examples**
organizationUnitId	Samlink organization unit	5
level	Log level	Information | Warning | Error | etc.
logType	Type of log entry	Default | User
timeStamp	Date and time for log entry	2019-09-10T03:00:01.6278373+03:00
fingerprint	Unique identifier for log entry	bcd51984-agdc-40a6-b571-y6a97f98a4e3
machineName	Workstation name of the RPA agent	T2490A1011
processName	Name of the RPA automation	RPA-bank-hakemusten-tietojen-siirto_Samlink Production
jobId	Identifier for current RPA automation execution	24a84531-010b-457f-90t1-5ayc98d7b557
robotName	Name of the agent executing automation	RPA-BANK-1-1234
machineId	Workstation ID number	5
message	Short message regarding the entry	Throw exception: Saldo ei riitä | Tarkistettiin A:n nimi
rawmessage	JSON formatted message including most of the above columns and several more fields regarding the log entry	{ **“message”**: “Siirryttiin Varallisuus-sivulle.”, **“level”**: “Information”, **“logType”**: “User”, **“timeStamp”**: “2019-09-10T03:00:50.6103121+03:00”, **“fingerprint”**: “70f44345-22bb-46df-885e-75f180fc4d48”, **“windowsIdentity”**: “LOCAL\\T123456”, **“machineName”**: “T2490A1011”, **“processName”**: “RPA-bank-hakemusten-tietojen-siirto_Samlink Production”, **“processVersion”**: “1.0.7111.31245”, **“jobId”**: “24a84531-010b-457f-90t1-5ayc98d7b557”, **“robotName”**: “RPA-BANK-1-1234”, **“machineId”**: 11, **“fileName”**: “LisaaTiedot_Talous_Varat”, **“logF_BusinessProcessName”**: “rpa-sp-011-OTT-valmistelu” }

Table 3: Log fields in RPA log data

goal is to find possible problems in RPA runtime and not detect individual customer related problems. This way it was not necessary to achieve adequate security by less secure and more effort consuming ways such as pseudonymization or k-anonymization (explained in the section 2.8), which would have also required strict inspections before data could have been approved for cloud processing.

`TODO: References?`

As production environment is built on Microsoft Server based solution, and because it was highly unrecommended to install additional software to the production server, data acquiring and anonymization tools were chosen based on what was already usable in the RPA production environment. Microsoft PowerShell offers sufficient tools for database SQL querying and stream editing. The amount of data was significant which made straight file editing impossible due to the memory limitations. Thus, stream editing was necessary for finding and replacing sensitive information from the data.

`TODO: maybe references?`

Anonymization took good proportion of the time in workdays as processes were slow, the amount of data was huge and multiple re-runs were needed before the results were seemed adequate.

`TODO: Appendix of the script used. Does this need more explaining?`

3.4 Data formatting

At the beginning of the research, the log data from RPA was in SQL database. However, the database used was not "pure" in a way that typical relational databases are, but some columns included JSON-formatted data in them. For ML algorithms to be able to read the given data with ease this kind of impurities needed to be cleared from the data.

When feeding the log data to anomaly detection algorithm, it was necessary that all the rows were as minimally unique as possible in order to use the pattern finding abilities of the algorithm. Too unique data points would have made all of them anomalies compared to each other. Thus, all unique features were stripped from the data, such as the fingerprint value that was unique for each data point. As rawmessage-field included data in textual JSON form that identified unique automation executions at some extent, it was processed with additional script which removed manually such field values. These fields were timestamp, fingerprint, and job ID values. When training ML algorithm with rawmessage, timestamp and job ID values were included from their corresponding fields outside rawmessage.

`TODO: Add reference to appendices`

3.5 Azure cloud resources

TODO: This section is under construction! Here are some things we are discussing here:

Comment:

Azure resources, such as virtual networks, storage spaces, connections to ML studio etc.

What kind of resources were usable based on the issue at hand and limitations by the company (cost, basically). We discuss some things about Azure ML Studio but only in extent of what parts needed configuring.

Azure provides a vast set of tools and resources for different kinds of cloud projects. Resources needed for Azure ML Studio usage depends on the subscription used and security restrictions set by subscription administrators. When starting this study, due to these restrictions, all resources used for ML training in this project needed to be inside the same virtual network. Azure ML Studio environment could be opened from any network, but most of the features were unavailable if computer browsing studio UI was outside this virtual network. Thus, a virtual machine had to be acquired as Azure resource from within the same network as other resources and ML Studio UI had to be opened with the browser on this machine. Later on it was found that Azure Machine Learning Workspace networking feature could be configured to allow public access making possible to access ML Studio UI from all networks.

TODO: Explain picture

3.6 Azure ML Studio

During the initial pipeline runs the execution came to an abrupt stop and Azure notified about memory issues. These problems were linked to the data amount which had to be reduced to 600 megabytes before any pipeline could be finished using the data. This reduction was against the initial goal where preferably all the data could have been used.

Considerable amount of time was used to fix or avoid this issue but nothing clear was found that would explain the error received. While working with the issue it was also noted that data needed more cleaning in order to ease the preprocessing phase as described with more detail in section 3.3 Thus, the data had to be imported from log archive and anonymized once more.

To advance the study more efficiently it was decided to trim info-type log messages from data hence reducing the data amount considerably. Final data included 8.6 million log rows which was about 10% of the original data size. Before final cleaning operations the data took 8.1GB od disk space, and after cleaning the rawmessage-field the final disk size was 6.6GB. Even with this data size, some Azure ML components

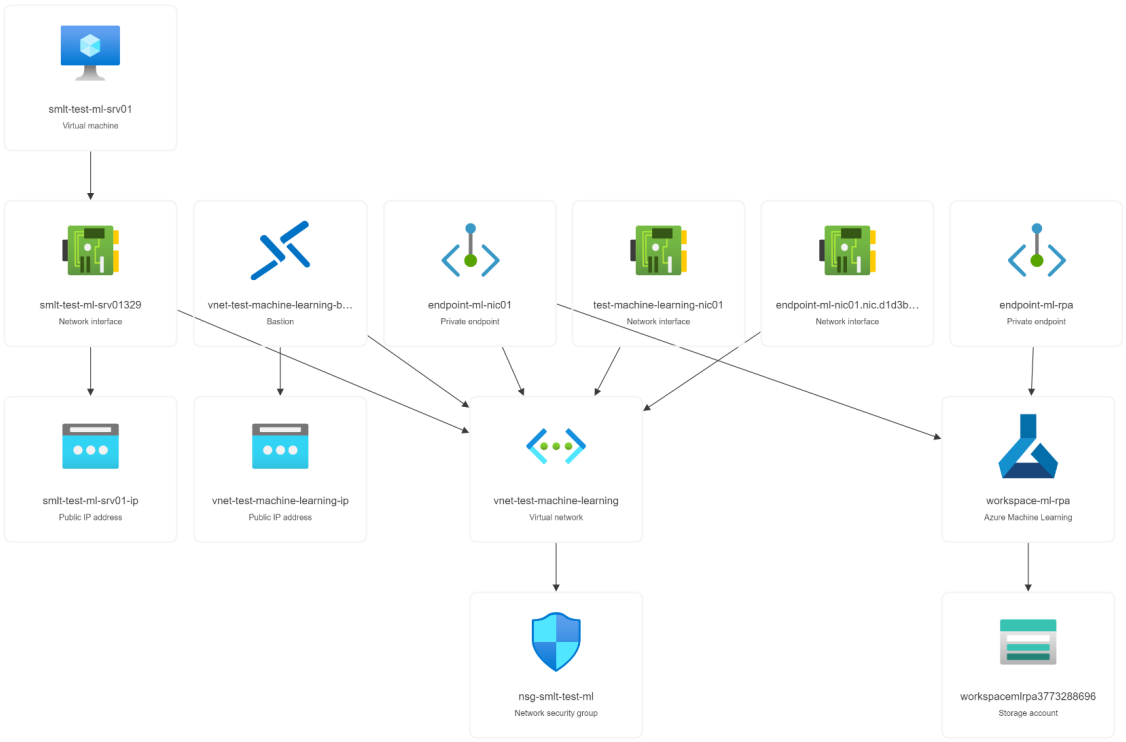

Figure 6: Azure cloud resources needed and their relations

faced this memory issue and forced us to choose such components that were able to handle these data amounts.

In addition to the data we used to train the algorithm, we needed to set up computing instance in Azure ML studio. Some predefined resource limitations affected the computing instance choosing and the memory issue encouraged us to pick memory prioritized instances. Single computing instance did not work, but we needed to choose a computing cluster instead to be able to run ML training pipeline.

`TODO: Some more info about the resources needed`

4 Machine learning pipeline structure

The full component chain from input to output with algorithm training and result validating is called a machine learning pipeline. In this section, we discuss how ML pipeline was created in Azure ML Studio. Several ML algorithms were compared in order to find the most feasible for our goal in mind. ML training was organized in two different phases in order to find the relation between log anomalies and technical tickets. Although the Azure environment and ML Studio requirements were the objectives of the study and therefore part of the outcome of the results, these results were also a prerequisite for solving the final objective considering the ML algorithm possibilities. Therefore, the resulting Azure resources and ML Studio pipeline components are demonstrated in this section.

Results of the trained algorithms were validated against newly acquired production data in order to estimate how well the initial goals of the study were fulfilled. These results are presented in the next section 5.

Azure ML Studio makes ML pipeline creation easy and comparing different methods and algorithms effortless. Nevertheless, with hybrid approach having two different phases, and result comparison being done against anomaly hypothesis, the pipeline drafts started to accumulate in content.

When starting the ML pipeline testing the initial plan was to feed the log data to anomaly detection algorithm and try to get some sort of estimate of possible anomaly count. This plan had several problems. First, as stated, logging is very abundant and several thousands of rows is logged during a single day. Some errors encountered are not critical and RPA agent is able to recover from them finalizing the initial task. This means that errors that could be deemed anomalous may not result to a ticket in the end.

In addition, one single error case noticed by bank clerks may be linked to several problems in runtime, meaning that one ticket received is, in fact, linked to multiple, dozens or even hundreds of log rows.

Two different algorithms were needed. In phase 1, algorithm defines how likely one datapoint, or log row, is to be considered an anomaly. In phase 2, another algorithm aims to predict how many tickets are to be expected to receive within a time frame. This dual algorithm approach is referred as a hybrid machine learning approach.

`TODO: Explain a bit the contents of these phases. Intro to next subsections.`

4.1 Time frame compression and statistical features

To avoid the problem with random delay between log rows and technical ticket timestamps, as stated in the section 2.10, log rows were grouped by time stamp into certain time frame groups with a method we call “time frame compression”,

This means that in order to eliminate the effects of random delay we compress some features in certain time frame which is at least as long as the longest estimated

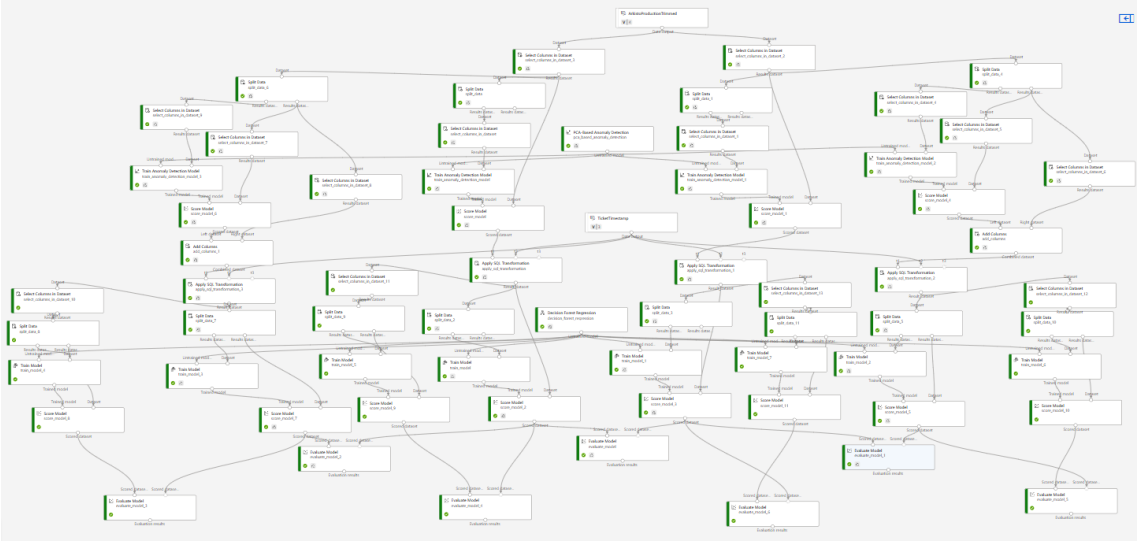

Figure 7: TODO: Pipeline in all its glory. Find a better place and edit caption.

delay. Simply put, if we count possible anomalies during one hour of log, we cannot compare this number to actual tickets received at the same hour or the next. What we can do, with time frame compression, is that we count some statistical values of anomaly estimates, for example, the mean and median values of a week, and then compare these numbers with the tickets received during the same week.

Comment:

This part explains some statistical features used when starting regression training in hybrid ML phase 2. These features consist of both log row amounts and PCA-component output values. All these values are grouped together in timescales.

4.2 Unconventional training approach

As stated in section 2.1, the approach used in this study is, if expression is allowed, unorthodox. Typically, data points used in ML algorithm training and validating should always be different. Acting otherwise leads to algorithm processing with same data it was trained with thus creating a situation where algorithm already knows what to do with the current data point. If the results were validated after this the algorithm would get unreliable score as it had the validation data already in the training phase. This could be compared to giving some right answers to students during test and scoring test results as if no help was given. However, due to the nature of the study problem and contents of the data, it was decided to test whether bending this rule would provide better results in algorithm training.

Even though the amount of data was large enough to cause issues with memory, the hybrid approach and the time frame compression discussed <later in this section> lead to significant reduction of data in phase 2. As a general rule of thumb in ML training, only 20–30% of the data is used to validate the algorithm. With the hybrid

ML approach, the validation results in phase 1 are what actually form the data used in the phase 2. This data is further compressed to time frame groups leading to only a few dozen data points in phase 2 ML training compared to millions of rows in phase 1.

Because of the way the anomaly detection algorithms work the over-lapping data points is not as big of an issue as it would be with other types of algorithms like regression algorithms. This is why we could use part of the data for training the anomaly detection algorithm as usual and then use all of the available data for validation without overfitting the algorithm. Also, because the main forecasting functionality comes in the phase 2, overtraining in phase 1 may not cause issues.

To verify if this unconventional training method gives good results without issues, the trained algorithms were tested with new production data that had zero overlapping data points with training and validation data. This way we were able to compare different training approaches to determine the best overall pipeline.

4.3 Memory issues and limitations

Memory is crucial resource in ML training. Algorithms take multiple steps while iterating the data and intermediate results are stored in the RAM rather than on the disk. While building ML pipeline in Azure ML studio, a memory issue emerged that affected several components and caused serious limitations in terms of usable components and data size. Due to the time limits of this study this issue was not resolved and the problem behind it was not found. As several conditions considering the environment costs were already issued by the company, the issue was supposed to be linked with compute instance property limitations and thus could not be resolved without cost overruns. However, this was not certain.

`TODO: Limitations to components. List effects of this problem.`

4.4 Feature format for PCA-ADA

In Azure ML Studio there is only one module selectable for anomaly detecting, the PCA-based anomaly detection module, which is explained in section 2.5. However, with textual input like logs it can be used at least in two ways. First, input data can be fed to the algorithm trainer as is, letting the PCA-based ADA component do the work without further modifying the log rows. This way, the component tries to recognize the anomalies based on all the information included in the row. Practically this means that the component processes data in textual format making each row in the input a feature as a whole to consider.

`TODO: PCA-ADA should be explained in background-section`

Second option is to convert the textual features into numerical n-gram features. Each word or n-gram is now a number of said instances found on the row being processed, and each row can be presented as a sequence of numbers indicating the number of those features.

N-grams can in addition have a weight based on the frequency they appear in the entire data. Different weights usable in Azure ML component are listed below.

1. Binary Weight
2. TF Weight
3. IDF Weight
4. TF-IDF Weight

`TODO: Explain different weights and open up more Azure ML Studio PCA-component`

4.5 Anomaly probability

Comment:

Here we explain what PCA-component outputs and how the result is used in the pipeline.

The output values of the PCA-ADA component are, as explained in the section 2.5, normalized so the values range between 0 and 1. This anomaly probability value is the main output of hybrid ML phase 1. Based on our initial hypothesis that each anomalous event in the log is linked to a real life support ticket received, the bigger a single anomaly probability value is for a log row the more likely that row is related to a ticket inducing event.

4.6 Regression based estimating

Comment:

Here is more information about different regression algorithms used in ML pipeline. Some basic information about all of them is given so the results are understandable by reader in the Result section.

1. Linear regression
2. Decision forest regression
3. etc.
4. etc.

Branching node	Options	Layer count
Input text column	message rawmessage	2
Text preprocess	Yes No	2
Numeric conversion	No N-gram Feature Feature Hashing	3
ADA training	Unconventional Proper	2
Validation without anomaly metrics	Yes No	2
Regression algorithms	Decision forest regression Boosted decision tree regression Neural Network Regression Linear regression	4

Table 4: Pipeline divergent layers

4.7 Pipeline branching

In order to compare different results, some comparable metrics are needed.

Several diverging points caused the pipeline to branch. First, the error message used to calculate the anomaly probability of a log row had two options. We could either use simple *message*, or more verbose *rawmessage*. This textual data could be fed to the ADA-component in several forms. Most straightforward way was using textual data without any preformatting. Text could also be run through "Preprocess text" -component. N-gram features could have been extracted from the text and these features could have been used instead. Instead of n-gram features, textual data could be converted to numeric with "Feature Hashing" -component. After getting the ADA-component results, the anomaly probabilities were compressed with R-code or SQL. <This concludes the phase 1.>

In phase two, branching of the pipeline was due to either different regression algorithms used or comparing results without the anomaly probability values calculated in phase 1. In practice this means that in order to validate the results against our initial hypothesis, we used pure statistical log data such as row count and unique job ID count without anomaly probabilities to determine whether anomaly metrics provided any insight regarding the ticket data.

Each branching step, or layer, multiplies the amount of results used in final comparison that would determine the best possible pipeline combination. These layers are simplified in the table 4.

The divergent count implies the number of branches diverging from the previous component. The total count of branch ends, or leaves, would then be the multiplication of all divergent counts, totaling to 192 comparable pipeline combinations. Moreover, n-gram feature extraction and feature hashing have several tunable parameters that

strongly influence the end results of the algorithm training. To reduce this amount when considering the best possible pipeline, we simplified this by narrowing down the options based on initial test run results of some of the divergent options.

For example, n-gram feature component suffered greatly from the memory problem and the data amount that the *Extract N-Gram Features from Text* -component was able to handle only 2% of the original data amount. This was deemed as too small amount for training as it would be extremely likely with 98% of the data skipped that also possible rows relevant to the ticket anomalies would get trimmed out.

Algorithm	**Mean Absolute Error**	**Root Mean Squared Error**	**Relative Squared Error**	**Relative Absolute Error**	**Coefficient of Determination**
Rawmsg, unconventional training	3.5	4.355313	0.890083	0.924528	0.109917
Rawmsg, unconventional training, anomaly values excluded	3.544643	4.204987	0.8297	0.936321	0.1703
Rawmsg, proper training	2.848214	3.696191	0.641063	0.752358	0.358937
Rawmsg, proper training, anomaly values excluded	3.133929	4.025449	0.760362	0.82783	0.239638

Table 5: Rawmessage with text preprocessed, Decision Forest Regresssion used in phase 2

5 Results

`TODO: Contents of this section has been reorganized to methods and pipeline -sections. Final results are still gathered. Below is some quickly written notes, subjected to change.`

Even though with rawmessage we were forced to reduce the data amount significantly, it was seen as valuable data source as it included more data than pure message had in JSON structure. As data amount was skimmed to less than 1% of the original amount, the results may not be reliable. However, with rawmessage preprocessing used with decision forest regression, the coefficient of determination was over 0.35.

6 Summary

TODO: Very under construction...

<Sum up here what we did and why>

6.1 Discussion

<Here some thinking what should have been improved>

Integrating with real time logging?

Data formatting

The most time-consuming tasks in the study was without a doubt the anonymization and preformatting of the data. Although sensitive information may sometimes be crucial in error fixing as problems may consider just one client, it is necessary that the data sanitation is possible to do in order to use the data in less secure environment. By preformatting the data in such way that all different personal information types do not differ between use cases.

Possible ML methods

Some sort of time delay forecasting?[35] Could we estimate the number of tickets based on some log metrics in time frame?

Memory error fixing would have given more options.

TODO: Something else specifically needed?

If log would have been better organized/preformatted it would have been possible to use One-Class Support Vector Machine. This, however, would have needed a manually constructed dataset including only "normal" log events or such events that could have been certain of that they were not part of the ticket inducing issues.

TODO: Shifting timeframes from mon-sun -> sat-fri

Some more testing would be needed to determine whether splitting data randomly or not in the phase 1 provides better results. By splitting data randomly it is possible to miss important anomalous rows that come in groups which would provide necessary insight to identify anomalous events.

TODO: Show some explaining graphics, here or elsewhere.

The original hypothesis was that anomalous events in the logs were clearly linked to the tickets received. However, as memory errors due to the size of data forced us to skip info-typed rows, it is possible the data anomalies did not reflect to the tickets. As stated before, multiple error lines in the log may be linked to a single issue, which could make the ticket inducing events more common log feature. Thus, by tuning the statistical values used to take into account more common error messages, it could have been possible to get better results.

TODO: Could anomaly probability metrics be calculated per job ID?

References

[1] P. K. Donepudi, "Machine learning and artificial intelligence in banking," *Engineering International*, vol. 5, no. 2, pp. 83–86, 2017.

[2] W. M. Van der Aalst, M. Bichler, and A. Heinzl, "Robotic process automation," pp. 269–272, 2018.

[3] A. DeLaRosa, "Log monitoring: not the ugly sister," *Pandora FMS*, 2018. [Online]. Available: https://web.archive.org/web/20210901031146/https://pandorafms.com/blog/log-monitoring/

[4] W. K. Ho, B.-S. Tang, and S. W. Wong, "Predicting property prices with machine learning algorithms," *Journal of Property Research*, vol. 38, no. 1, pp. 48–70, 2021. [Online]. Available: https://doi.org/10.1080/09599916.2020.1832558

[5] K. Ghanem, F. J. Aparicio-Navarro, K. G. Kyriakopoulos, S. Lambotharan, and J. A. Chambers, "Support vector machine for network intrusion and cyber-attack detection," in *2017 Sensor Signal Processing for Defence Conference (SSPD)*, 2017, pp. 1–5.

[6] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, 2015.

[7] "Definition of algorithm," accessed: 2022-05-19. [Online]. Available: https://web.archive.org/web/20220510183749/https://www.merriam-webster.com/dictionary/algorithm

[8] T. O. Ayodele, "Types of machine learning algorithms," *New advances in machine learning*, vol. 3, pp. 19–48, 2010.

[9] B. Mahesh, "Machine learning algorithms-a review," *International Journal of Science and Research (IJSR).[Internet]*, vol. 9, pp. 381–386, 2020.

[10] R. Vickery, "Beginners guide to the three types of machine learning," 2019, accessed: 2022-07-12. [Online]. Available: https://web.archive.org/web/20220712170551/https://towardsdatascience.com/beginners-guide-to-the-three-types-of-machine-learning-3141730ef45d?gi=9db5aaf56001

[11] X. Chen, Y. Fang, M. Yang, F. Nie, Z. Zhao, and J. Z. Huang, "Purtreeclust: A clustering algorithm for customer segmentation from massive customer transaction data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 3, pp. 559–572, 2017.

[12] H. Li, "Which machine learning algorithm should i use?" 2017, accessed: 2022-07-13. [Online]. Available: https://web.archive.org/web/20220402120055/https://blogs.sas.com/content/subconsciousmusings/2020/12/09/machine-learning-algorithm-use/

[13] P. Baheti, "Train test validation split: How to and best practices 2022," 2022, accessed: 2022-07-14. [Online]. Available: https://web.archive.org/web/20220714125714/https://www.v7labs.com/blog/train-validation-test-set

[14] L. Zhou, S. Pan, J. Wang, and A. V. Vasilakos, "Machine learning on big data: Opportunities and challenges," *Neurocomputing*, vol. 237, pp. 350–361, 2017.

[15] "Comparing machine learning as a service: Amazon, microsoft azure, google cloud ai, ibm watson," accessed: 2022-04-04. [Online]. Available: https://web.archive.org/web/20220714131635/https://www.altexsoft.com/blog/datascience/comparing-machine-learning-as-a-service-amazon-microsoft-azure-google-cloud-ai-ibm-watson/

[16] "Azure machine learning - ml as a service - microsoft azure 2022," 2022, accessed: 2022-07-12. [Online]. Available: https://web.archive.org/web/20220714154625/https://azure.microsoft.com/en-us/services/machine-learning

[17] "Pca-based anomaly detection: component reference - azure machine learning," 2022, accessed: 2022-07-10. [Online]. Available: https://web.archive.org/web/20220710090015/https://docs.microsoft.com/en-us/azure/machine-learning/component-reference/pca-based-anomaly-detection

[18] R. Noumeir, A. Lemay, and J.-M. Lina, "Pseudonymization of radiology data for research purposes," *Journal of digital imaging*, vol. 20, no. 3, pp. 284–295, 2007.

[19] J.-W. Byun, A. Kamra, E. Bertino, and N. Li, "Efficient k-anonymization using clustering techniques," in *International Conference on Database Systems for Advanced Applications*. Springer, 2007, pp. 188–200.

[20] A. Rantala, "Applying machine learning to automatic incident detection from software log output," Master's thesis, Aalto-yliopisto, Sähkötekniikan korkeakoulu, 2019, available: http://urn.fi/URN:NBN:fi:aalto-201906234018.

[21] S. Allagi and R. Rachh, "Analysis of network log data using machine learning," in *2019 IEEE 5th International Conference for Convergence in Technology (I2CT)*. IEEE, 2019, pp. 1–3.

[22] N. Kondo, M. Okubo, and T. Hatanaka, "Early detection of at-risk students using machine learning based on lms log data," in *2017 6th IIAI international congress on advanced applied informatics (IIAI-AAI)*. IEEE, 2017, pp. 198–201.

[23] Q. Cao, Y. Qiao, and Z. Lyu, "Machine learning to detect anomalies in web log analysis," in *2017 3rd IEEE International Conference on Computer and Communications (ICCC)*. IEEE, 2017, pp. 519–523.

[24] D. Liu, "Log analysis for anomaly detection," 2019, accessed: 2022-07-15. [Online]. Available: https://web.archive.org/web/20220715102721/https://davideliu.com/2019/10/26/log-analysis-for-anomaly-detection/

[25] X. Zhang, Y. Xu, Q. Lin, B. Qiao, H. Zhang, Y. Dang, C. Xie, X. Yang, Q. Cheng, Z. Li *et al.*, "Robust log-based anomaly detection on unstable log data," in *Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 2019, pp. 807–817.

[26] F. Anifowose, "Hybrid machine learning explained in nontechnical terms," *JPT*, 2020. [Online]. Available: https://web.archive.org/web/20220612142143/https://jpt.spe.org/hybrid-machine-learning-explained-nontechnical-terms

[27] T. Shon and J. Moon, "A hybrid machine learning approach to network anomaly detection," *Information Sciences*, vol. 177, no. 18, pp. 3799–3821, 2007.

[28] C.-F. Tsai and M.-L. Chen, "Credit rating by hybrid machine learning techniques," *Applied soft computing*, vol. 10, no. 2, pp. 374–380, 2010.

[29] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE access*, vol. 7, pp. 81 542–81 554, 2019.

[30] N.-C. Hsieh, "Hybrid mining approach in the design of credit scoring models," *Expert Systems with Applications*, vol. 28, no. 4, pp. 655–665, 2005.

[31] A. Jain and A. M. Kumar, "Hybrid neural network models for hydrologic time series forecasting," *Applied Soft Computing*, vol. 7, no. 2, pp. 585–592, 2007.

[32] H.-j. Kim and K.-s. Shin, "A hybrid approach based on neural networks and genetic algorithms for detecting temporal patterns in stock markets," *Applied Soft Computing*, vol. 7, no. 2, pp. 569–576, 2007.

[33] T.-S. Lee, C.-C. Chiu, C.-J. Lu, and I.-F. Chen, "Credit scoring using the hybrid neural discriminant technique," *Expert Systems with applications*, vol. 23, no. 3, pp. 245–254, 2002.

[34] R. Malhotra and D. Malhotra, "Differentiating between good credits and bad credits using neuro-fuzzy systems," *European journal of operational research*, vol. 136, no. 1, pp. 190–211, 2002.

[35] G. H. Erharter and T. Marcher, "On the pointlessness of machine learning based time delayed prediction of tbm operational data," *Automation in Construction*, vol. 121, p. 103443, 2021.

A Pipeline draft

TODO: This is dummy appendix

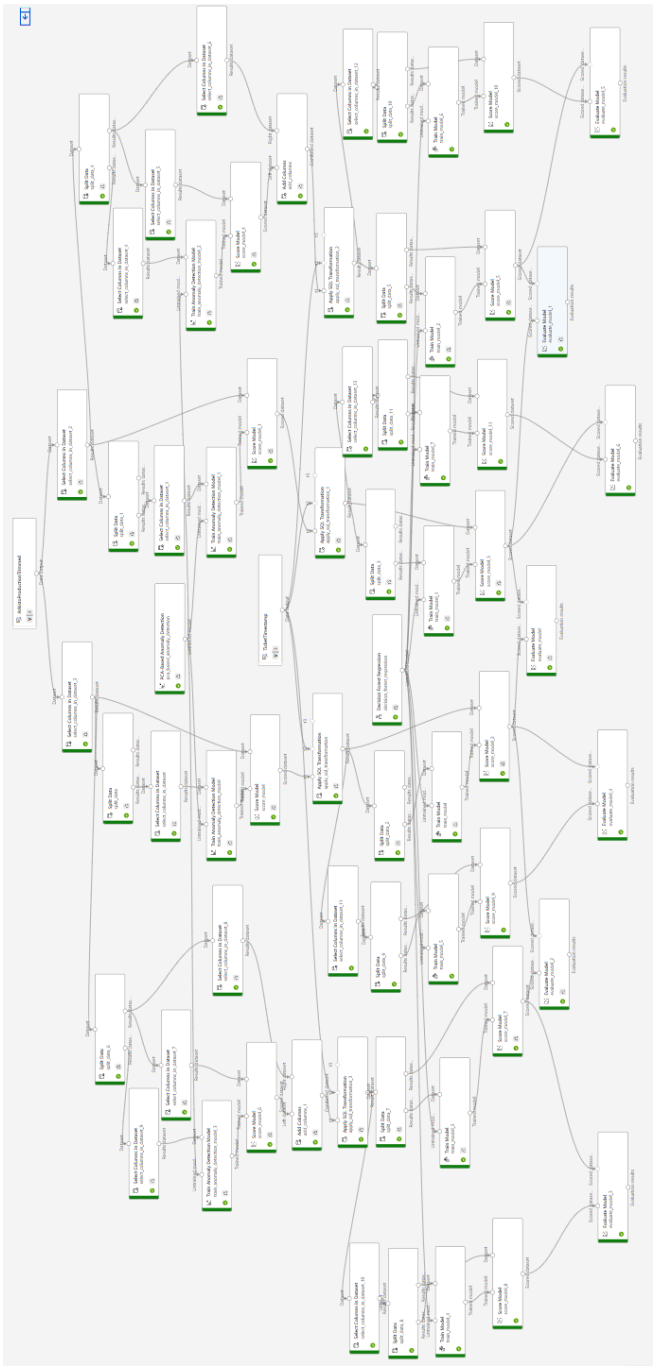

Figure A1: TODO: dummy picture in appendices