Detecting fraud with distinctive features: Using website data to analyze transport companies

Data Systems Final Report

Marit Beerepoot
University of Amsterdam (DS)
10983430

Jessy Bosman
University of Amsterdam (DS)
11056045

Ralph Horn
University of Amsterdam (DS)
6069738

Rogier Knoester
University of Amsterdam (IS)
12397660

Guillaume Toubi
University of Amsterdam (IS)
12305405

ABSTRACT

The police and justice departments have a hard time keeping up with the steady increase of fraudulent activities. An automated fraud detection system, based on machine learning, can support current cuts in the police their budget for manpower and could result in other suspicious companies than found with previous methods. This paper explores a unsupervised learning approach, so detect facade/front companies, based on their website data. A user interface was created to visualize the model. The created model is generalisable, which means that it can handle large and dynamic feature sets, support data from different domain as long as it is delivered in the right format and adapt when the dataset is updated.

KEYWORDS

Predictability, machine learning, fraud detection, money laundering, unsupervised learning, Dutch police, feature distinctiveness

1 INTRODUCTION

Fraud can be described as a way of cheating in which things are portrayed differently than they are. It is a conscious attempt at giving a wrong image of reality [1]. While crime rate, in general, lowers each year, fraud increases. Money laundering is an example of fraud. It often happens through drugs trafficking and fraudulent activities. Currently, money laundering is a 16 billion euro problem in the Netherlands [3].

The police and justice departments have a hard time keeping up with the steady increase of fraudulent activities, due to the many varieties of fraud. This paper focuses on facade(front) companies, to limit the scope of this research. The police uses a variety of ways to prevent and apprehend fraudulent companies. The Dutch police has a fraud detection system in place that analyses communication channels and financial data. Information on communication and financial data is gathered from various databases the police has access to. This data is then analysed and an assessment is made whether or not the company should be investigated. However, the police can not investigate every company due to legal and time constrains. The time constraints are a result of limited manpower due to budget cuts [8]. Therefore, a more automated solution that can support current analysts is an approach that could help mitigate the impact of these budget cuts [8]. Furthermore, a more automated solution, based on machine learning, could result in other suspicious companies than found with previous methods.

The police has legal limits on what they are allowed to use in an investigating. For example they are required to have good cause to use privacy sensitive documents. This semi-automated solution could use information that is publicly available and therefore soften the burden these restrictions can impose. A facade company's website is for instance constructed in such a way that it looks legitimate and passes the smell-test. As such, it is possible to automatically check these websites for the presence and absence of website features. These features include aspects such as the presence of an about section, working links, proper Dutch or English, and correct contact info. For example, the absence of correct contact info could be an indicator for a facade website. However, even in 2019 there still exists a plethora of bad website. Therefore, the solution should account for these websites.

In this research paper, there will be a focus on publicly available website data. The website data collected should be structured clearly and help identify fraudulent companies the police might have missed via other methods of analysis. Due to the absence of data about legitimate and fraudulent observations, unsupervised learning methods will be explored to determine the distinctiveness [6]. The research aims to asses if a feature, or combination of features, are important by analyzing their distinctiveness within a dataset. This exploration of the features is dependent on the data itself, therefore in this paper it is assumed that most companies are not a facade/front company.At the request of the stakeholder, the paper will focus on companies in the transportation sector. However, this paper aims to construct a model that could be used in different types of businesses and domains.

The goal of this paper is to answer the following research question: ***How can website data be used to create a fraud prediction model, based on distinctive features*?**

To structure this problem, four sub-questions were created

(1) What are distinctive features of facade websites?
(2) What model can be used to decide the distinctiveness of facade websites?
(3) How can the distinctiveness of facade websites be presented visually?
(4) Can the created model be generalized to other fields?

To answer to this question, first research has to be done into what distinctive features are. The first sub-question thereby is: *"What are distinctive features of facade websites?"* Once these features are found, the facade websites have to be distinguished from real websites, to verify if the chosen approach was correct. The corresponding sub-question is: *”What model can be used to decide the distinctiveness of facade websites?”* After the model is created, it is also important that people with different skill levels and people working in different domains can use the model. Therefore the following sub-question has been added: *”How can the distinctiveness of facade websites be presented visually?”* Finally, it might be interesting to add other data than website data to the model, and look for distinctive patterns in that data. The last sub-question is: *”Can the created model be generalized to other fields?”*

This paper is structured as follows: First, a theoretical framework is constructed based on the possible approach on how to detect fraud. Following this framework, the method is presented. The method section will outline how the research questions will be answered. Next, the results are outlined in the result section. This all accumulates in a conclusion and a discussion. These sections will discuss the pitfalls and problems encountered in this research. Finally this paper will be concluded with a section describing the possibilities for future research.

2 THEORETICAL FRAMEWORK

2.1 Fraud detection financial data

The police and other branches of the government already use specialised tools to find fraudulent companies. Earlier tools mostly used generic statistics, whereas newer tools start to implement forms of machine learning. However, there are some issues with developing new tools. The development becomes difficult due to the limited amount of exchangeable ideas. The reason behind this is due to the fact that describing fraud detection techniques in the public domain gives criminals the necessary information to evade law enforcement. Datasets are not made available and results are censored. The tools that are available use varied methods but they all provide the function of comparing observed data with expected values. Some tools use simple behaviour aspects that focus on graphical summaries, others use multivariate behavioural profiles. Those can be based on the studied system or extrapolation of other systems. The complexity is created due to the fact that an actor is partly fraudulent; some correct behaviour mixed together with some illicit behaviour.

Statistical fraud detection methods can be supervised or unsupervised. When supervised, the data of both fraudulent and non-fraudulent actors are used to create a model which enables profiling. The data that is provided should contain truthful class information to begin with. Fraud detection this way will only detect the same type of previously occurred fraud.

Unsupervised methods of statistical fraud detection consist of finding actors that are out of the norm. These outliers have to be further examined to determine their illicitness. It is not possible to indicate if a company is fraudulent or not with this method but the analysis indicates a gradation of abnormal behaviour that can be evaluated more closely. Different methods can be created to grade an actor based on different features that may increase or decrease the credibility that Bolton et al. described as a suspicion score [6].

2.2 Indicators of fraudulent websites

Machine learning approaches to detect fake websites have been researched in the past. The main idea behind these approaches is that a website is scraped, then features are extracted from the websites, a learning algorithm (mostly classification algorithms) are used to classify the data, and then all websites get a prediction label. Most of the resulting approaches base their features on link specific information, like the amount of inlinks and outlinks, or specific fraud cues, such as the appearance of certain features on the website itself. Currently no formal evaluation has been performed to conclude whether the fraud cues found for fake websites also work to detect facade websites. Some of the interesting features extracted from expert interviews are [9]:

- The website content, for example the absence of contact information.
- The linkage, for example the absence of inlinks.
- Images, for example using duplicate images.
- The source code, for example broken links.
- External feature collection, for example when the Kamer van Koophandel (KvK) data does not match up with the contact information on the website.

Some other interesting features from the literature are [9]:

- The use of word phrases, for example outdated copyright.
- Presence of specific genre information, for example having a career section.
- Lexical measures, for example the use of proper English.
- Errors in the source code
- Missing images
- Missing/broken links
- Website registration information

There are, however, three important characteristics for designing a system that can accurately detect fraudulent websites. The first important characteristic is that the system must be able to generalise patterns in existing data, as do all machine learning approaches. The second important characteristic is that the systems must be able to handle large and dynamic feature sets; a website is a complex organ and can change over time. Finally, there should be a layer of dynamic learning included, meaning that the system can adapt when the dataset gets updated [9].

The main conclusion is that there are several characteristics for both the so called facade websites and the system that analyses them. Website data could be retrieved based on those characteristics to help determine fraud. Analysis could then be able to generalise patterns, handle rich and dynamic sets of fraud cues, and have the ability to perform recursive trust labelling (RTL) via a dynamic learning layer to relearn newer training datasets. Finally, a quantitative machine learning approach should consist of enough information to be able to perform correct classifications.

2.3 Possible solutions

Most of the machine learning approaches for detecting fake websites in the past are based on labeled data, which means that supervised learning approaches can be used. Currently the most successful classification techniques for fake websites seem to be neural networks and Support Vectors Machine, but the results seem to vary per type of fake website. For example, different methods work better for spam websites than for phishing websites [5].

Unfortunately enough, most of the time supervised learning approaches cannot be used to classify fraudulent websites due to a lack of labeled data and research into fraudulent website features. Semi-supervised learning and unsupervised learning are thus preferred. With semi-supervised learning, the parameters that indicate facade websites will still not be known, but they can be guessed by an expert. It is, however, hard to not bias the algorithm with this approach. Unsupervised learning thus seems to be the most promising approach [9].

3 METHOD

This section will explain what methodology will be used to answer the research question. The different methodologies all add a variety of knowledge to help explain the subject to the fullest.

3.1 Research Design

3.1.1 *Stakeholder contact.*

To fully understand the problems and develop a solution that fits the needs of the Dutch police it is important to gain knowledge about their way of working. Two data analysts from the Dutch police will give a presentation with a semi-structured Q&A. The questions are of a multitude of different subjects. One of which is in regard to tools that are currently used. The different tools will give insight in what they are already working with so that a possible solution envisions an IT infrastructure that works together with the design of solution. Other questions are in regard to their workflow and scope of investigation. Knowing the workflow will enable the solution to fit to current processes. Finally, questions regarding the data they are already using will clarify which analysis's are already possible. Both knowing the workflow and data the police already uses will benefit the adoption of the solution and a transition can be envisioned towards a future situation where website data analysis benefits investigations.

3.1.2 *KvK data.*

A selection of the Dutch Chamber of Commerce, the Kamer van Koophandel (KvK) [2], data was supplied by the stakeholders. This data consists of a data entries of businesses, with variables such as id, validation number, name, address, business summaries and website URLs. These URLs are used as a starting point for the scraper and data analysis.

3.1.3 *Data collection.*

To further enrich the provided KvK data, it was decided to use the information contained in websites found in the dataset. At first it was decided to manually check each website for extra information. However, this is very time intensive. To reduce this time cost, the different research groups researching this problem distributed the websites between the groups. At first a smaller subsection of the websites was distributed. This however resulted in more biased data as each researcher can have different opinions about a website. As such the teams opted to gather data from these websites through the use of a python scraper and crawler. The development of this tool was distributed between the teams. It should be noted that the combining of the code, the deployment and management, and coordination of the development efforts was done by this research group [4].

The scraper follows these guidelines:

For each website the scraper checks if a feature is present. If said feature is present, set the variable to true. If the feature is not present or the scraper was unable to scrape the feature, set it to false. Most features are stored as booleans. However, there are some features that could be grouped together and as such these features are stored as a dictionary of booleans. An example of the scraped features for a website can be found in table 1.

Table 1: Example scraped features for a website.

Variable	Value
scraper_index	0
about_section	False
career_section	True
contact_info	True
copyright	0
emails	[]
google_analytics	True
https	False
is_scrapable	True
links	[http://www.pax.nl/de/, http://www.pax.nl/en/]
price_calculator	False
redirected	True
social_media	{'linkedin': False, 'facebook': False}
url	http://www.paxbis.nl
wordpress	False

Due to limited development time, the scraper was not fully optimized. Some websites took over a minute to scrape. To counter this problem, the scrapers were deployed into a cluster of six AWS-EC2 instances to scrape the KvK websites in a timely manner.

3.1.4 *Validating the model and user interface.*

To ensure that the model as well as the user interface corresponds to what is needed and to indicate that the research question has been answered a validation should be performed. The model and the user interface will be designed to be able to help identify distinctiveness within a dataset. There are multiple approaches that can be used to evaluate whether or not the user interface requirements are met. Two are considered, one whether the user interface is intuitive to use and the second whether the user interface is satisfying to use. It is possible to evaluate the intuitiveness through usability testing. This can be done by creating a task list with important tasks that the system allows the user to do. Users will then have to perform these tasks. By analysing the outcome, and how long the user took to complete the task, it is possible to decide whether or not the user interface satisfies the intuitiveness requirements. Analysing the user satisfaction can be done by having users use the system without any prior knowledge of it. They can ask questions about the system if they do not understand aspects of it. Through observation and an interview afterwards it becomes possible to conclude whether or not the system is satisfying to use. An open interview with the topics intuitivity and usability is preferred so that there is the ability to further ask questions and receive new ideas, which can be implemented in future versions of the model and user interface. Intuitiveness is based on previous experiences and can be difficult to asses. Since the model and user interface will be dynamic the user validation should be performable by any participant. The sole applicable condition is that the user should have at least some technical know how on how to use similar systems. The technical ability of the user assures that observations and measurements made are generalisable to the targeted audience.

3.1.5 Preprocessing.

After the data was collected, some preprocessing had to be done. The data was first al loaded into python, and with help of the pandas library a dataframe was created. The following actions were taken to create the final dataframe:

- True/False was converted into 1/0 respectively.
- The copyright feature, that extracted the copyright year, was converted in to a 0 when the year of copyright was below 2018 and to a 1 when the copyright is from the year 2018 or 2019.
- The different social media platforms, that were stored in a dict, were spread out to all have their own column in the dataframe (Linkedin, Facebook, Instagram, Twitter, Pinterest). There was also a column added that consisted of 1s when companies have one or more of the social media accounts and a 0 when a company has no social media accounts.
- All possible combinations of features were added to the dataframe. This can be illustrated by the following example: if there are 3 features, X, Y and Z, all their possible combinations are XY, XZ, YZ, XYZ. These are then added as a column to the dataframe and scored. This means that when a website for example has a career section and an about section, it gets the value 1 for the career section and about section column, while another website that only has a career section and no about section, gets a 0.
- All the companies that had a website that was not scrapable, were deleted from the dataframe.
- The data of the KvK was combined with the scraper data, by using the url and the scraper index.

3.1.6 The model.

The formatted data can be used to shape a model. Different possibilities were explored to create the final model. Given the format of the data, with a lack of labeled data, supervised machine learning was not feasible, meaning an unsupervised learning approach was necessary. With these unsupervised methods, a clustering tool was created using built-in scikit learn clustering tools. However, the problem with the discovered clusters was that it was impractical to lead back to features. A more explainable model was required to add meaning to the findings, as this is one of the stakeholder's demands for the model. A more explainable model was created based on a simple but effective construct: creating a penalty score for every website. The main idea behind this score is that a company gets a penalty for not having certain features on its website, since Verhees states that a fraudulent website is more likely to miss certain features [9]. Every feature can be categorized as either True (1) or False (0), which can be used to create a score function. It is, however, important to notice that some features can have more importance (or weight) than others. To solve this, weights for each variable were added to the score function. The weight is based on the proportion of websites that contain a certain feature. This can be illustrated with the help of table 2. This table contains 5 companies, A-E and 1 feature, X. The weight of the feature is therefore 3/5 = 0.6.

Since a penalty score is calculated, the weight is then multiplied by 1 when a website does not contain the feature corresponding to the weight, and by 0 when it does contain the feature. This is not only done for the main features, but also for their combinations/interactions (explained in the previous section). A weighted score function is thereby created, which can be used for ranking. This can be illustrated with the help of table 3 and table 4. Table 3 contains the weight of each feature and 4 the presence or absence of the features in company A. The penalty score of company A will thereby be: 0.1 + 0.07 + 0.03 + 0.002 = 0.202, since it will get a penalty for not having feature Y, XY, YZ and XYZ (see italic number in both the tables). To make the score easier to understand, the scores of all the samples are afterwards scaled from 0 to 10, where 10 is the website with the most penalties (and the smallest amount of features) and 0 the one with the lowest penalties (thus the most features).

Table 2: Example data of 5 companies and feature X.

	A	B	C	D	E
Feature X	0	1	0	1	1

Table 3: Example data of features and there weight.

	X	Y	Z	XY	XZ	YZ	XYZ
Fraction/weight	0.6	*0.1*	0.05	*0.07*	0.04	*0.03*	*0.002*

Table 4: Example data of company A and the presence of features.

	X	Y	Z	XY	XZ	YZ	XYZ
Company A	1	*0*	1	*0*	1	*0*	*0*

This score function creates a general idea of the uniqueness of a website, or in principle the deviation of websites in comparison to the whole set of websites available.

Besides the website penalties, the website are clustered via a choice tree, supplied with all different combinations of variables. For each combination, the choice tree is created based on the True and False entries. Each leaf (endpoint) is then labeled as a cluster. Therefore, each cluster contains the websites with exactly the same composition of True and False variables. After all the cluster were created, a uniqueness score was calculated in the same way the weight was calculated in the model: dividing the total number of companies by the amount of companies that use the (combination of) features they were clustered by. These clusters can be used to explore connections, cluster sizes and unique occurring combinations. An insight is created in the way the variables are distributed and key indicators might be derived from the cluster data. Finally, a method was created to link the found scores back to the Kamer van Koophandel data. A selection (for example the most deviating 100 websites) can be supplied to a function. This function then links the websites back to the KvK data. Together with all the available KvK data, a proportion test is created for a user given variable. This creates insight in the distribution of the variable, inside and outside of the supplied selection. For example, for the variable City, it might show that from the selection there are significantly more businesses located in Amsterdam in comparison to the complete data set. If the user supplies a suspicious selection of websites with a explorative variable for the KvK data, the output can show the difference in the distribution of values from the proportions. This is shown in figure 2.

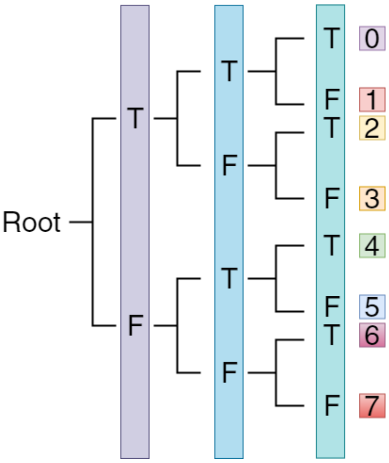
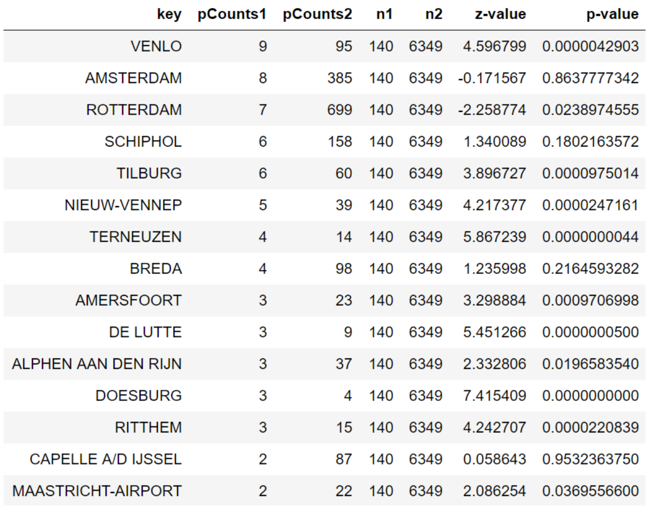


Figure 1: Clustering algorithm visualized



key	pCounts1	pCounts2	n1	n2	z-value	p-value
VENLO	9	95	140	6349	4.596799	0.0000042903
AMSTERDAM	8	385	140	6349	-0.171567	0.8637777342
ROTTERDAM	7	699	140	6349	-2.258774	0.0238974555
SCHIPHOL	6	158	140	6349	1.340089	0.1802163572
TILBURG	6	60	140	6349	3.896727	0.0000975014
NIEUW-VENNEP	5	39	140	6349	4.217377	0.0000247161
TERNEUZEN	4	14	140	6349	5.867239	0.0000000044
BREDA	4	98	140	6349	1.235998	0.2164593282
AMERSFOORT	3	23	140	6349	3.298884	0.0009706998
DE LUTTE	3	9	140	6349	5.451266	0.0000000500
ALPHEN AAN DEN RIJN	3	37	140	6349	2.332806	0.0196583540
DOESBURG	3	4	140	6349	7.415409	0.0000000000
RITTHEM	3	15	140	6349	4.242707	0.0000220839
CAPELLE A/D IJSSEL	2	87	140	6349	0.058643	0.9532363750
MAASTRICHT-AIRPORT	2	22	140	6349	2.086254	0.0369556600

Figure 2: Selection proportion test with selected KvK variable

3.1.7 *Back-end.*

To present the scores and general information of the dataset it is necessary to request it from a server. Flask is used for this. Flask is a micro web framework written in Python. Because the web server is also written in Python it becomes easier to integrate the model. Through this it becomes possible to generate new scores whenever a new csv file is uploaded. To allow the front-end to communicate with the back-end several HTTP endpoints were defined. These endpoints effectively open several csv files and based on which endpoint is called returns the requested data. The data is presented in the JSON format. By having the data available in JSON the front-end can easily interpret the data without any extra actions.

Currently the generated data in the back-end does not adhere to a schema. If, in the future, the storage of the data moves from generic csv files to a relational database, it becomes possible to create a more RESTful API. The presentation of the data could be more efficient. For example, currently several endpoints have been limited to a subset of the available data because the front-end would be unable to handle such amounts of data at once. Another feature that then becomes available is the ability to version the data periodically. This would allow the presentation of historical data and visualise changes in that data.

3.1.8 *Front-end.*

The front-end is a single page application. The decision for this is based upon the expectation that the end users will be navigating between different screens a lot. By creating a single page application it becomes possible to store the state (e.g. company details, scores) locally and immediately show them on subsequent navigation actions. Furthermore, through a single page application filtering the shown data becomes easier as well. Filtering the data is an important aspect of the front-end because, as detailed earlier, the model can handle any binary encoded set of features. This means that it is possible to load a csv file with many features which will result in many possible clusters which can be navigated through. Besides listing data in formats such as tables, it is also required that a user can zoom in on a specific company that interests them. This should be possible from every location the company is mentioned, e.g. the list of companies in a certain cluster. The technical side of the front-end is completed through the help of React, a JavaScript library that enables developers to easily create interactive user interfaces. The interactivity comes forth from observing the state of the application. Whenever the state changes React renders the changed elements. The state of the application is filled by calling several endpoints provided by the back-end. For example, the companies in the csv are loaded by calling the "/companies" endpoint. The visual aspect of the front-end is built with the help of Semantic UI and its React component set. Semantic UI is a CSS framework that provides a large collection of pre-designed components. This makes it quicker to design a functioning user interface, as not every element has to be handcrafted by the developer. Instead, already proven elements can be used.

With the above a first version of the front-end and user interface has been created to effectively present the results of the model.

4 RESULTS

4.1 Police workflow

Currently police investigations are initiated due to various reasons. Some companies have financial behaviours that are out of the ordinary. Others have communication channels that are being used to exchange non business related information. To increase effective policing the solution should be applicable as a starting point for an investigation but also for an ongoing investigation where additional information helps to identify the likeliness of a company's illicitness.

4.2 What are the distinctive features of facade websites?

Currently there is not enough research done into facade websites to know what features are distinctive. There are, however, lists of features that are distinctive for fake websites. Since the time span of this research was not very long and the lack of labeled data (as further detailed in the discussion), it was impossible to find and validate which features are distinctive for facade websites. It is, however, possible to determine the distinctiveness of the features of the fake websites within the data set. This can be done by making the assumption that most of the websites in the data set are not facades. Then by using the score/fractions calculated for the model, it is possible to distinguish features that are used often in website, and features that are not used often on the websites. Some of the features that are used often on the websites are https (56,8%), social media buttons (40,9%), contact information (39,9%), an about section (35,0%) and a copyright statement from 2018 or 2019 (28,7%). When assuming that real websites are in the majority, it could be possible that the lack of these features on a website are an indication of a facade website. There are also some really rare combinations on features that only a few websites use, like using https, having an about section, having a career section, using Wordpress and using Google Analytics, that only 0.2% of the website uses. It could be that these unique combinations of features are also an indication of facade websites. However, this is not confirmed, since there was no labeled data available.

4.3 What model can be used to decide the distinctiveness of facade websites?

The website score and the website clustering seem to work well. When using the website score, the websites that do not use many of the features are distinguished from the websites that do use those features. Below some example sites are shown. The highest scoring websites are really incomplete websites or websites that only contain plain text or a 404 error. The lowest scoring websites are complete and really decent looking websites.

The decision tree clustering part of the model also did a remarkably good job in distinguishing different kind of websites. A total of 712 clusters were created. The websites inside the clusters all show a close resemblance to each other, and seems to group websites accurately by features. Figure 3 shows an example of grouped websites in a cluster. Both of the approaches have the same problem: it was not possible to validate whether this score or the clusters did actually distinguish the real and facade websites with the help of the distinctive features on the websites, due the lack of labeled data. The main thing what the model did was cluster similar websites and give a penalty to websites without many features. There features were based on features of websites that can be used to distinguish fake websites from the real ones, that were chosen to use in this research due to lack of time and lack of data to validate whether these features also work on actual facade websites.

Table 5: Top 10 highest scoring websites, high score means lack of features, low score means lots of features on the website

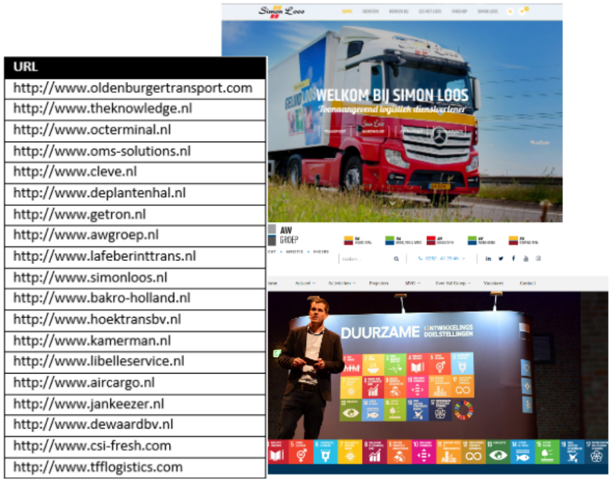
Website	Score
www.zoologistics.nl	10
www.supermaritime-vanreems.nl	10
www.westsidetrading.eu	10
www.documentexpresse.nl	10
www.friendly-service.nl	10
venemaantiques.nl	9.9
www.buchele.nl	9.8
www.milestonelogistics.nl	9.8
www.all4youlogistics.com	9.7
www.royalstars.nl	9.7

Table 6: Top 10 lowest scoring websites, high score means lack of features, low score means lots of features on the website

Website	Score
www.cleve.nl	0
www.hoektransbv.nl	1.3
www.dr-logistics.nl	2.2
www.gjpersoneelsdiensten.nl	2.4
www.cargonaut.nl	2.8
baselogistics.com	2.8
www.alpi.nl	3.2
www.cargoshipping-international.com	3.2
www.ter-linden.nl	3.3
www.a2b-online.com	3.3

4.4 How can the distinctiveness of facade websites be presented visually?

One of the goals was to create a generalisable model. The user interface has to incorporate this dynamic behavior and be flexible with how it presents the results of the model. The dashboard is an example of UI flexibility. It features the companies that deviate the most from other companies. The attached score is derived from the features which is dynamic by nature. Besides the dashboard, the company detail and cluster overview pages are other examples. The company detail page, seen in figure 4, includes whether or not the features in the feature set are present on the website of the company, and the distribution of features of its website score. The cluster overview page is entirely dynamic, all the data presented on the page is gathered from the feature set that is provided to the back-end. When this feature set changes the possible presented



URL
http://www.oldenburgertransport.com
http://www.theknowledge.nl
http://www.octerminal.nl
http://www.oms-solutions.nl
http://www.cleve.nl
http://www.deplantenhal.nl
http://www.getron.nl
http://www.awgroep.nl
http://www.lafeberinttrans.nl
http://www.simonloos.nl
http://www.bakro-holland.nl
http://www.hoektransbv.nl
http://www.kamerman.nl
http://www.libelleservice.nl
http://www.aircargo.nl
http://www.jankeezer.nl
http://www.dewaardbv.nl
http://www.csi-fresh.com
http://www.tfflogistics.com

Simon Loos
WELKOM BIJ SIMON LOOS
Toonaangevend logistiek dienstverlener
AW GROEP
DUURZAME ONTWIKKELINGSDOELSTELLINGEN

Figure 3: Cluster that consists of websites that are made with Wordpress, do have a career section and have an up to data copyright

clusters also changes. Therefore, the user is able to quickly swap the csv file on the back-end to analyse another set of data.

The explainability of the model is one of the most important UI aspects according to the stakeholder. The requirement has to do with the legal perspective of the Dutch police force. An investigation cannot be initiated unless an explanation is provided as to why an investigation is necessary. In other words, a black box model is unusable due to legal constraints. The model is made explainable in two separate, standalone ways. First, the website score. The score is made explainable by accompanying the score with the value of each feature. This value is either true or false, which is represented by either green or red, shown in the top right of figure 4.

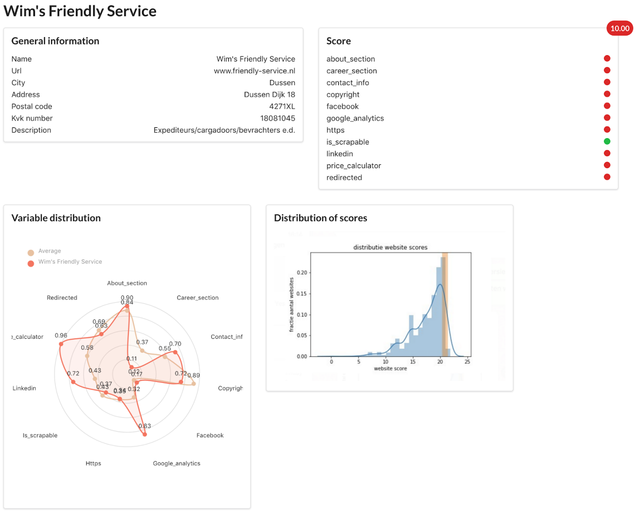
When navigating through clusters the explainability is done through showing which features the cluster has and which other websites are in the same cluster. The user interface of this is shown in figure 5.

4.4.1 *User interface validation.*

As described in the methodology section it is important to validate the user interface. When a user interface is not validated it is impossible to know whether or not the user interface is effective at doing what it should do. Two questions were asked:

- Is the user interface intuitive to use?
- Is the user interface satisfying to use?

The regular approach for these methods is to find a sample population of the target audience to test with. However, our target audience is the police. This target audience is significantly harder to test with. For this reason we made use of convenience sampling; asking peers to test. Making use of convenience sampling does introduce several drawbacks. These drawbacks are that our sample population is relatively small and that there is likely a bias in this population. The bias is introduced by the fact that a large part of the population studies a subject related to computer sciences. However, their (partial) hands-on experience with user interfaces might help finding certain less performing elements in the user interface.



Wim's Friendly Service

General information

Name	Wim's Friendly Service
Url	www.friendly-service.nl
City	Dussen
Address	Dussen Dijk 18
Postal code	4271XL
Kvk number	18081045
Description	Expediteurs/cargadoors/bevrachters e.d.

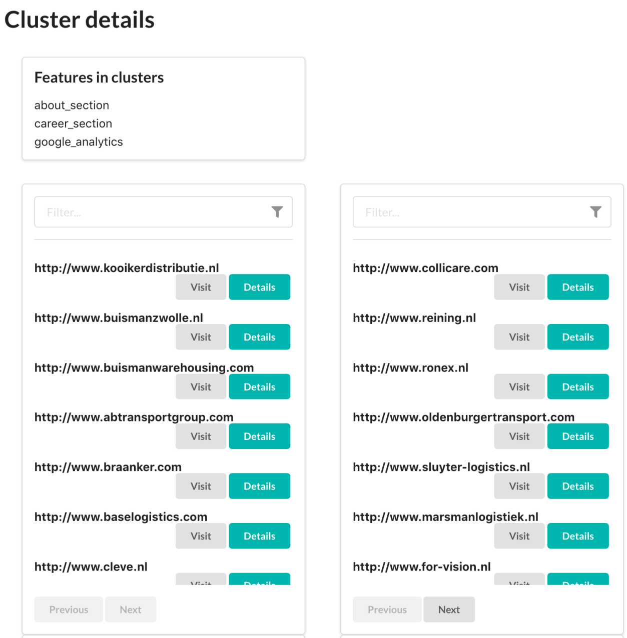
Score 10.00

about_section	🔴
career_section	🔴
contact_info	🔴
copyright	🔴
facebook	🔴
google_analytics	🔴
https	🔴
is_scrapable	🟢
linkedin	🔴
price_calculator	🔴
redirected	🔴

Variable distribution

Distribution of scores

Figure 4: The details of a company showing the explainability of the website score



Cluster details

Features in clusters

about_section
career_section
google_analytics

http://www.kooikerdistributie.nl	http://www.collicare.com
http://www.buismanzwolle.nl	http://www.reining.nl
http://www.buismanwarehousing.com	http://www.ronex.nl
http://www.abtransportgroup.com	http://www.oldenburgertransport.com
http://www.braanker.com	http://www.sluyter-logistics.nl
http://www.baselogistics.com	http://www.marsmanlogistiek.nl
http://www.cleve.nl	http://www.for-vision.nl

Figure 5: The details of clusters showing the features and which other websites are present in the same set

To test the intuitiveness certain tasks were created that the participants had to perform:

- Find a company that performs badly according to the model
- Search for a company
- Filter the list of companies on your home city
- Find a company through feature exploration

The tests have time limits, these are 10, 10, 20, and 30 seconds, respectively. The tests were performed by 6 participants, of which five were male and one female. All of the participants were able to complete the above tasks. Four of the participants were able to complete the tasks within the set time limits. Most of the issues encountered were during the feature exploration tests. Participants had issues understanding what a feature combination represented. The common opinion was that the feature exploration interface lacked details as what could be done and what would happen if a cluster was selected. Furthermore, often the participants did not understand what the different groups of companies within a feature combination meant. This could possibly be improved by adding more textual hints on the pages the user is and by adding actions that can be performed on the visible data.

Whether or not the user interface is satisfying to use is done through an open interview. It is satisfying if the system does what the user interface portrays what it does. This open interview was done immediately after the first test. Several questions were asked to deduct whether or not their use of the system was satisfying. These questions were:

- Did you have any troubles using the application?
- Did you expect different results from certain actions?
- Overall, do you think this application achieves its goals?

As stated with the other test, several participants had issues understanding the feature exploration. It was not immediately clear what it meant and how it could be used by the user. After explaining what the feature was about they understand the usefulness of it, however, a common opinion was that it requires a redesign to make it more effective. The second question was mostly answered positively; the users expected the same results as what the system presented. The final question was also answered mostly positive. Most users voiced that they think that this kind application and user interface can help the police in their work.
With these results, we can assume that this first version of the user interface is intuitive in its use and that the overall experience while using it is also satisfying. However, it should be kept in mind that this is a first version. There are no other versions that the results can be compared with.

4.5 Can the created model be generalized to other fields?

It is possible to use this model in different fields. The model works as long as the data is binary. This means that yes/no features work, but also categorical data that can be transformed to binary data through dummy variables. The model will only have problems with numbers (int or float) that can not be transformed into categorical data. This means that in theory nearly all possible data can be used in this model, since the model just searches unique clusters and scores companies on not using certain features. To apply the model in other fields it is only necessary to provide a csv file with the data. The user interface provides a input field to upload a csv file, however, it is not functioning as it is considered beyond the scope of this project.

5 DISCUSSION

The biggest limitation of this research is that it was not possible to validate whether the created model can detect clusters of fraudulent websites or if the website penalty score is a good indicator for fraudulent websites. Unfortunately, the Dutch Police could not provide labeled data due to confidentiality, as some ongoing investigations would be in that data. The focus of the research was thereby not necessary on detecting fraudulent websites but on detecting unique websites. The assumption was, based on literature, that companies with unique websites are more likely to be fraudulent. Furthermore, the stakeholders also provided the information that most suspects in this domain are older males with little knowledge regarding websites. This four-month project was guided by the UvA. The UvA provided the project with a stakeholder. Due to unfortunate circumstances the stakeholder had the inability to provide fraud data on multiple occasions. Finally, the focus of the project changed to still enable a project without fraud data. Due to these setbacks the four months turned into four weeks to complete this project. Therefore, a change in scope where multiple aspects, such as example validation of the model and user interface, will have to be looked into in the future. However, the setbacks were not enough to not be able to create a model and user interface that can be tested in the future. Methods are provided in the future work to help with the validation.

6 CONCLUSION

The goal of this paper was to answer the following research question:*How can website data be used to create a fraud prediction model, based on distinctive features?*

In this research it was not possible to conclude one set of features that make a facade website distinct from other websites. However, it was possible to build a model that can group websites into distinctive groups. This was done through clustering with binary trees. This results in clear non colliding clusters that group websites by similarity. From this similarity the distinctiveness can be calculated based on how unique this cluster is in relation to the other clusters in the dataset.

This distinctiveness was presented visually through the use of normalised scores between zero and ten and an interactive dashboard. As stated in the discussion, it should be noted that a user validation with the end user should be done to fully validate that the interface works as expected.

Finally, the model is build in such a way that it can be applied in other fields. The model accepts a csv file as input and as long as the data is normalized, as described section 4.5, it will be applicable to different fields. This is due to the nature of the chosen clustering algorithm, this algorithm is agnostic to the input data and computes the uniqueness based on the input dataset.

7 FUTURE WORK

7.1 Within the fraudulent website field

First of all, it would be great if this model can be validated in the future, with labeled data. This way it can be determined whether the fake website features also work for fraudulent websites. Furthermore, more data on the companies can be added to the model, such as financial data, past crime data and employee data. It would be interesting to see what clusters this model can find with more features.

7.2 Other fields: The food sector as possible candidate

This research was focused on finding distinctiveness in website data for transportation companies to, in the future, be able to achieve probability analysis of illicitness. The transportation branch has been chosen, as stated earlier, due to the higher fraud chance in that branch. Due to the flexibility of the model and accompanying user interface other domains can be envisioned as said before. According to the stakeholder, another branch of interest could be researched is the food sector [7]. Transporting food provides criminals with the opportunity to perform illicit activities. Combining fraudulent website data of food companies should provide the current model with enough information to asses website distinctiveness of such companies and possibly cluster unidentified illicit companies together with the existing fraudulent companies creating possible leads for future investigations.

7.3 User interface and experience validation

The results described indicate a somewhat positive view on the built user interface. However, as stated, the population taken for the testing is likely biased. The population is familiar with the domain of user interfaces. This could, of course, also help in seeing elements that might be performing less than required. For future testing we suggest broader tests. These tests should include the actual target audience. This was limited for our research due to time and logistical constraints. If user tests with the actual target audience has positive results it is possible to use the user interface in actual applications.

REFERENCES

[1] [n. d.]. fnv uitkeringsfraude. ([n. d.]). https://clientenraad-roerdalen.nl/wp-content/uploads/2018/12/Factsheet-Uitkeringsfraude-FNV.pdf

[2] [n. d.]. KVK. ([n. d.]). http://www.kvk.nl/

[3] 2018. Nog altijd wassen criminelen in Nederland miljarden euro's wit. (Nov 2018). https://www.trouw.nl/samenleving/nog-altijd-wassen-criminelen-in-nederland-miljarden-euro-s-wit~ae172327/

[4] 2019. *Project github repository.* https://github.com/Ralphhorn/datasystems-collaorative-scraper

[5] Ahmed Abbasi, Zhu Zhang, David Zimbra, Hsinchun Chen, and Jay F. Nunamaker. 2010. Detecting Fake Websites: The Contribution of Statistical Learning Theory. *MIS Quarterly* 34 (2010), 435–461.

[6] Richard J. Bolton and David J. Hand. 2002. Statistical Fraud Detection: A Review. *Statist. Sci.* 17, 3 (08 2002), 235–255. https://doi.org/10.1214/ss/1042727940

[7] Ilse de kruif Vincent van megen Nederlandse politie. 2019. Stakeholder meeting roeters eilandcomplex. (Jan 2019).

[8] Remco Meijer. 2017. *Groot tekort aan agenten dreigt: komende jaren 14.000 politiemensen met pensioen.* https://www.volkskrant.nl/nieuws-achtergrond/groot-tekort-aan-agenten-dreigt-komende-jaren-14-000-politiemensen-met-pensioen~bea9ff63/

[9] L.C. Verhees. 2018. The Potential of Machine Learning in Risk Assessment for Policing - Detecting Front Firms based on their Business Website. (2018).