Detecting fraud with distictive features

Using website data to analyze transport companies

Groep E1
Marit Beerepoot - Jessy Bosman - Ralph Horn
Rogier Knoester - Guillaume Toubi

Problem

- Front companies use websites to appear as a legitimate company
- The Dutch Police has limited resource to detect these companies
- The detection of front websites is hard
- There is a lack of qualitative data

Challenge: Investigate/detect companies that would otherwise go unnoticed

"How can website data be used to create a fraud prediction model based on distinctive features?"

Theoretical framework

Financial predictive fraud detection based on unsupervised learning possible through:

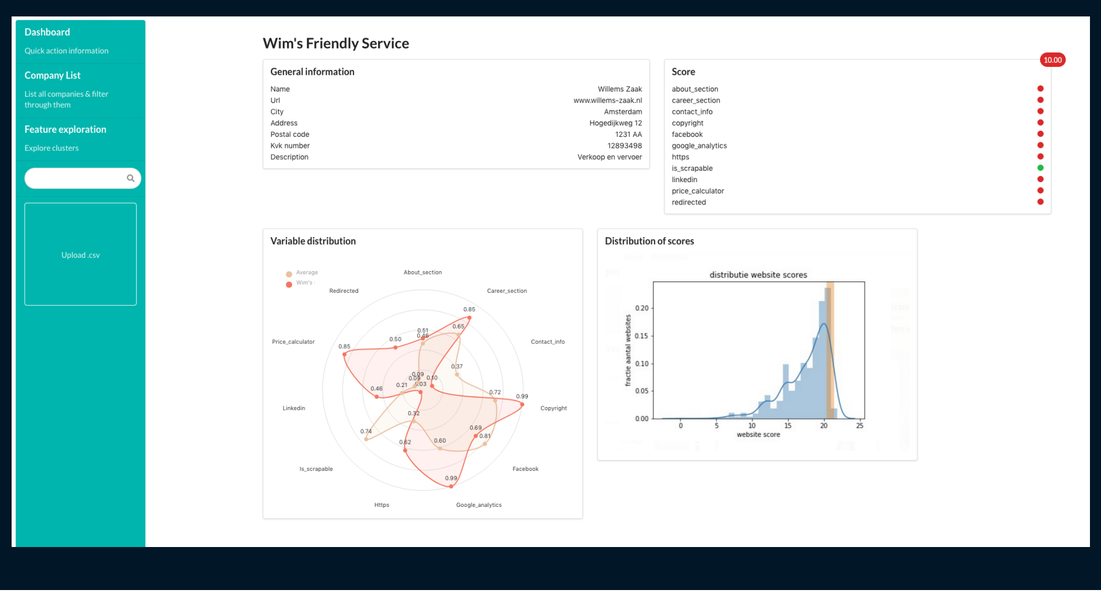
- Outlier detection
- Clustering
- Suspicion score

Machine learning for risk assessment for policing:

- Featureset based on literature, Kamer van Koophandel data and expert interviews.
- Feature identification through positive and negative features
- Correct classification only possible with enough data

User Interface

A dynamic interface is necessary since a generic approach (possible to use the model on all binary data, not only on website data) was taken. The user interface supports both quick-take-action and in-depth analysis. The interface enables to user to explore features, compare companies and filter the companies.



Methods

Stakeholder contact

- Presentation
- Questions and Answers

Collecting data

- Scraper data via custom build website scraper
- Kamer van Koophandel data

Unsupervised machine learning

- No labeled data available
- Clustering possible, but not revealing indicators, only groups

Model

Data

The data is unlabeled and mostly binary. There are a total of 17 features. Next to these features, combinations of these features were also taken into account. Example features are if a website was created with wordpress and if the copyright is up to date.

Website score

The score is based on the sum of feature presence values, which are either 1 or 0 times the fraction. The fraction is derived from the feature presence in the total collection. The score can be seen as a penalty for not having certain feature (combinations).

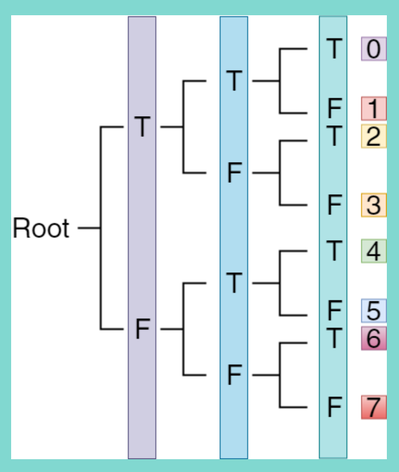
	X	Y	Z	XY	XZ	YZ	XYZ
Fraction/weight	0.6	0.1	0.05	0.07	0.04	0.03	0.002

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

Score/penalty = 0.1 + 0.07 + 0.03 + 0.002 = 0.202

Clustering with decision tree

The website score functions as an indicator score. To further analyse influencing features, decision tree clustering is used to explain and group websites with the same indicators to further support the uniqueness score.



Combining data with KvK

Selections of data can be made (e.g. the lowest scoring 100) together with the entire dataframe to be used in a statistical proportion analysis. This is then combined with the KvK data of the websites, to find significant values for the variables, for example more frequently occurring cities.

Results

Website score

The website score helps to differentiate websites with less features from websites with good features. Some papers indicate that the website with less features are more likely to be a front website, but since no labeled data is available, we can not comfirm this.

Clustering

The clusters do find groups of websites that have a unique combination of features, so our method seems to have worked. Again we can not comfirm if the decision tree does detect fraudulent companies through the lack of labeled data

Conclusion and Future work

This method provides a support tool to explore, identify, group and rank websites based on their characteristics. This can be used in the next step to further detect fraudulent websites, but to truly conclude what fraudulent website features are, labeled training data is required.

Future work

Since a generic model was created, this model can also be used in other fields or with a bigger dataset. As long as the data is binary or categorical (one hot encoding can be used to transform the data) this model works. The UI makes it easy to upload different datasets.