BACHELOR'S ASSIGNMENT,
BSC IN CIVIL ENGINEERING

PREDICTION OF DAMAGE TYPES DURING EXCAVATIONS

KORNÉL KISS, S2112752, K.KISS@STUDENT.UTWENTE.NL

SUPERVISORS:

LÉON OLDE SCHOLTENHUIS, JIARONG LI, UNIVERSITY OF TWENTE

ERWIN FOLMER, BOB SCHEER, KADASTER

APRIL-JUNE, 2023,

Kadaster

UNIVERSITY OF TWENTE.

PREFACE

This report has been written to summarize the findings of a ten weeks long assignment, part of my bachelor's in civil engineering program at the University of Twente. The project was carried out at Kadaster, the national land registry and mapping agency of the Netherlands. The main goal of the project was to investigate the possibility of predicting different damage types during upcoming excavations based on data gathered in the last decade.

I wish to thank my internal supervisors at the university, Leon olde Scholtenhuis and Jiarong Li, who went out of their way to provide me with as much help as possible during the entirety of the project.

I would also like to thank my external supervisors at Kadaster, Erwin Folmer for giving the opportunity to do this project, as well as giving the valuable feedback on my progress, and Bob Scheer for helping me understand some aspects of the modelling process.

Kornél Kiss,

Enschede, Netherlands

June 18., 2023

SUMMARY

The problem addressed in this document is the prediction of individual damage types, for example the breakage of water pipes or the cut of electricity or internet cables, during excavation projects using machine learning techniques. Previous studies have focused on exploring the reasons behind damages or predicting the overall probability of any damage occurring. However, there is limited research conducted on predicting individual damage types.

This research is one of the first to utilize machine learning for predicting individual damage types. The unique combined utility registry of excavations and damages of the Netherlands enables this study This prediction. , can be a powerful tool to shorten or even avoid outages. This can be done through either by warning utility providers to the risk or changing the way the excavation is done.

The project consists of several main parts. Firstly, the available data is investigated and cleaned to ensure its suitability for modeling purposes. The XGBoost machine learning method was selected, due to its successful track record in similar problem domains. Two approaches are considered for predicting damage types: binary classification and multiclassification. The binary classification approach predicts each damage type individually, treating the selected damage type as one class and all excavations without the given damage type as another class. The multiclassification approach aims to predict the most likely outcome which is either a given damage type or no damage at all.

After tuning the models, the following results were obtained for the binary classification approach: usable predictions with adjustable recall and precision were achieved for damage types of internet cable, low voltage electricity, low-pressure gas, and water strikes. To achieve a recall score of around 0.8, which represents the ratio of predicted real damages to all real damages, the precision, the ratio of predicted real damages to all predictions, varies between 0.05 and 0.01. All other damage types have a precision of less than 1% and thus not considered usable.

The multi-class model, predicting the most likely outcome, was less successful. This is mainly due to severe class imbalances. In most cases it predominantly assumes no damage will occur, and only when the excavation site is unusually large, it predicts internet cable manage. The most common damage type.

The most important features of both binary and multi-class models for damage prediction were the dimensions of the excavation site, related features such as the number of trees around, and the client and excavating companies. The damage types that could be to some extent reliably predicted where the most common one, present almost everywhere. Rarer, most dangerous type such as high-pressure gas or high-voltage electricity, cannot be predicted even with 1 percent precision. This is mainly due to the fact they do occur orders of magnitude less frequent than the common ones like water pipes or internet cables, in all cases with less than 1000 registered cases during the 2019 to 2021 period.

To summarize with the available data, it is possible to predict certain damage types in a useful manner, and possible use this prediction to minimize the time and cost of these damages. However, less common damage types cannot be predicted even by extending the data collecting timeframe.

TABLE OF CONTENTS

1. INTRODUCTION

1.1 BACKGROUND

The number of different underground utilities has been continuously increasing ever since humanity settled down. Underground water and sewage systems have been around for multi-millennia, gas pipelines for over two centuries, and in the last 50 years data, electricity cables and heat pipes have been put underground too (Orton, 2013). Consequently, it is not surprising that utility strikes, damages to cables and pipelines caused during excavations, are more common than ever. These accidents can cause disruption for both consumers and businesses, environmental damage and extra monetary cost and time delay for the constructor (Metje, Ahmad, & Crossland, 2015).

In the Netherlands there are approximately 40000 excavation damages to underground cables and pipelines every year, leading to over 25 million euros spent annually fixing these. On top of the cost there is also an uncounted number of wasted human hours due to outages. There is an ongoing effort to decrease and possibly even prevent these utility strikes. Two government agencies, Kadaster and Agentschap Telecom have been collecting and analyzing data on these accident for over a decade. The data collection and exchange among service providers and excavator companies have been also standardized to make it easier to communicate and prevent accidents in the future. To do this, the platform KLIC (Kadaster Kabels en Leidingen Informatie Centrum) was created by Kadaster maintaining one single up-to-date database listing all the underground utilities, excavation and damages. Since 2013 network providers and the excavation companies are legally obligated to register their activities and communicate through this platform.

1.2 STUDIES CONDUCTED IN THE PAST

There has been many research done focusing on preventing utility strikes through the use of sensing devices. Slightly fewer investigating their causes, among these is the often-cited paper on the ongoing research in the UK (Metje, Ahmad, & Crossland, 2015). This outlines many causes for utility strikes such as hurried construction, inaccurate maps, incorrect dept given for the different pipelines, contractors willingly risking damaging data cables in the hope of saving cost, etc. . These causes have been identified in the UK, but whether they all apply to the Netherlands, or are there any other that has not been considered, have not been investigated yet.

For predicting future strikes there have been multiple papers testing different models, one of the most detailed (Xiang & Zhou, 2021) used Bayesian Network models to predict and help to avoid utility strikes on energy pipelines in Canada. While their model has been demonstrated to be effective, creating it required significant technical expertise, and number of assumptions made.

The ongoing research (Li, 2023) focuses on creating a prediction model using different machine learning methods, and compare their effectiveness to predict the probability of utility strikes. These machine learning methods have the advantage of requiring little to no assumptions made before modelling, and easy adoptability to a wide range of problems. This research so far has proved that with the available data in the Netherlands, utility strikes can be reliably predicted, but it does not consider individual damage types, such as the probability of utility damages related to water, sewage, internet or electricity.

1.3 RESEARCH OBJECTIVE

So far there has been no research done in the Netherlands on using machine learning methods to predict different damage types during excavations individually. This would be desired by stakeholders such as Kadaster, utility providers, consumers as it could be used to avoid, make less expensive or shorten outage. For example, if we know that there is a specific data cable that is very expensive to repair or a dangerous high pressure gas pipeline in a certain excavation area, that are likely to be damaged due to certain characteristics of location, then it might be possible to avoid it by taking precautionary measures like recommending the usage certain more costly but safer excavation methods, such as hydrovac. This is when the excavation is done with a high-pressure water jet and a vacuum pump collecting the 'excavated' soil - water mixture in a tank.

Thus, the research objective is to use the available processed data and information gathered for damage prediction to create a new model that predicts the probability of different damage types to occur in a certain location for a given excavation.

To accomplish this, a statistical model must be selected. Given the limited amount of time, one machine learning model, XGBoost classification has been chosen for its many advantages. It can be trained to create a vector output containing the probabilities of the different damage types for the parameters of an excavation as inputs. XGBoost is a proven, reliable gradient boosting method, relying on the optimization of decision trees for classification. Its main advantage over other methods for the project is the fact that feature selection is not necessary when starting the training process, and thus one can achieve reasonably good results even when the user has little initial understanding on what are the important features or when causal relationships are too complex to use traditional statistical methods. It is also more efficient than most other models, which makes it faster to train and iterate.

There are numerous other sub-objectives:

- **Identifying the main causes of different damage types:** while this is not necessary for creating the model or making the predictions, this can explain why certain damage type can or cannot be predicted.
- **Exploring if different damage types often occur simultaneously, and if they can be grouped together:** if two damage types can be grouped together and thus grouped together the modelling process can be sped up by creating one model instead of many
- **Exploring if different feature often occurs together , and can be grouped together:** by grouping features together the model tunning times can be shortened.
- **Evaluating the different XGBoost modelling options:** because more than one damage can simultaneously occur, standard single outcome multiclass that would normally be used cannot be applied without consideration.
- **Presenting the final model and other findings in a manner that they can be used by others**

The following methodology (Figure 1) has been designed to accomplish the research objectives.

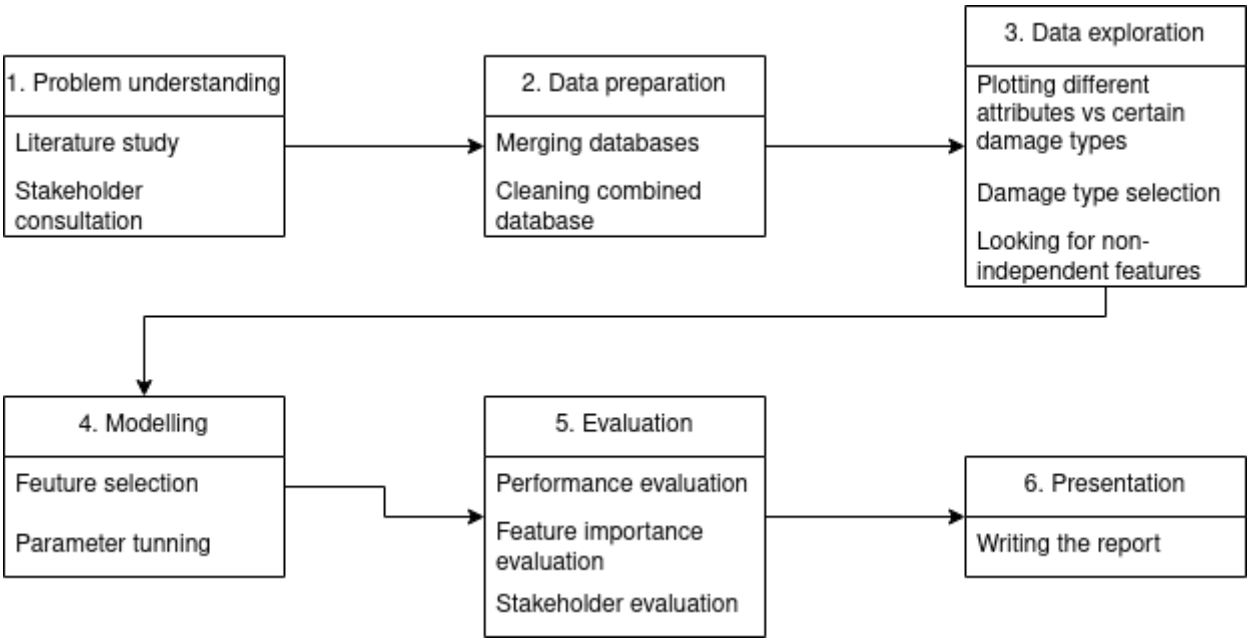


Figure 1 Methodology to answer the research questions

2. DISCUSSION

2.1 USED DATA AND ITS PREPARATION

Given that this project heavily relied on the research done by (Li, 2023), most of these data had already been combined into one database and cleaned, and only the specific damage types had to be added. Then, this combined database was cleaned from missing or incorrect values (e.g., incorrect damage type, mismatch between excavation registry and damage registry). In order to use XGBoost classification, the different damage types also had to be given individual columns with zeros and ones, showing if, in the case of an excavation, a certain type of accidents happened or not. If during a utility strike, multiple damage types occurred, they had to be registered under two or more columns.

The two main proprietary databases used, are graafmelding (excavation data), and schaderapport (damage data), listing of all recorded damage instances. Other databases, providing other, likely less relevant features such as the soil and the land use types, and tree density in the area are provided by opensource databases (BRO soil map, Bestand Bodemgebruik, Bomen in Nederlans).

The following data is available about the location of the excavations (both from KLIC and other sources):

- Information about the location of utilities nation wide
- The type of excavation that has been carried out, type of the project, its size and duration
- Soil type
- The company that carried out the excavation
- Geotechnical information
- Information about the surrounding area, land use, number of trees and their density

On total after one-hot encoding (binary information one a feature presence) there are over a hundred such features available.

As for utility strikes, the following twelve damage type category are in the database. The presented numbers are from 2021:

Table 1 Underground utility damages in 2021

Damage type listed	What is carried	Counted cases	Average cost [Euro]
Datatransport	internet	25875	3385
Laagspanning	low voltage electricity	11730	1165
Gas lage druk	natural gas (low pressure)	4468	1553
Water	Water	3305	1416
Middenspanning	medium voltage electricity	506	3698
riool vrijverval	gravity sewage	147	974
riool onder druk	pressurized sewage	101	2258
gas hoge druk	natural gas (high pressure)	77	13714
Overig	Other	25	496
Warmte	heat pipe	17	50038
Hoogspanning	high voltage electricity	9	2517
buisleiding gevaarli	natural gas, aviation fuel	5	5

As it can be seen in the table, the cases numbers and their costs vary widely. It is also important to note that in most cases during a utility strike more than one damage types are registered, whether this is because multiple damages occurred or because one damage can be put in multiple category is something to consider.

During the data preparation phase, it also became clear that many of the registered damages are badly filled in. Many of these could be used as datapoint for any damage occurring but not for predicting specific damage types. Many Klic-number has been registered as 00G000000, 20G000000, or not given at all. What is worse is that these registries also tend to have their damage types empty or undefined. Upon visual inspection it is clear that most of these are registered by internet providers and consequently likely many data cable strikes. Given the questionable quality of this data, and data cables already the most numerous category, they were removed from the database. The final count of the different damage types between 2019 and 2021 are the following in decreasing order (Table 2):

Table 2 The number of registered utility strikes in each category after matching databases between 2019-2021

Damage type	Number of registered cases
datatransport	48970
Laagspanning	25870
gas_lage_druk	8956
Water	6405
Middenspanning	1057
gas_hoge_druk	255
riool_vrijverval	196
riool_onder_druk	182
Overig	38
Warmte	22
Hoogspanning	21
buisleiding_gevaarli	7
landelijk_hoogspanni	3
Wees	1

After combining the cleaned database used to predict any damages occurring (Li, 2023) with the cleaned damage registry, it became clear that in some occasion not only multiple damage types occurred during a single excavation, but in even the same damage types multiple times(Table 3). If the same damage type occurred multiple times during one excavation, they are ctonsidered to be a single event, if more than 20 of the same damage type occurred that excavation has been discarded as upon visual inspection every excavation like that had unlikely Klic numbers(they were made up).

Table 3 Example: number of excavations with a certain number of data cable damages

Number of internet cable damage	Number of excavations with this many damages
0	28456
1	27236
2	3481
3	917
4	639
5	232
6	104
9	58
7	55
8	28
10	17
11	13
16	10
13	8
12	5
14	3
22	2
17	2
2712	1
3047	1
15	1
19	1

After removing duplicates, merging according to Klic numbers and combining identical damage type during the same excavation, a total of 64801 usable registered excavation damages occurred during 2221673 excavations between 2019 and 2021.

Table 4 Presence of excavations during all excavations between 2019 and 2921

Damage type	No damage occurring	At least one damage of this type occurring
datatransport	2191202	30471
laagspanning	2201626	20047
gas_lage_druk	2214001	7672
water	2216384	5289
middenspanning	2220760	913
gas_hoge_druk	2221446	227
riool_vrijverval	2221622	51
riool_onder_druk	2221542	131

2.2 GROUPING DAMAGE TYPES AND FEATURES

2.2.1 Grouping damage types

Seeing the relatively low numbers of certain damages, one question that came up is, if they can be grouped together. This would have been useful in accelerating the modeling process. To evaluate this a Pearson correlation matrix was created(Figure 1):

	datatransport	gas_hoge_druk	gas_lage_druk	laagspanning	middenspanning	riool_onder_druk	riool_vrijverval	water
datatransport	1.000000	0.013357	0.087463	0.132357	0.022050	0.004134	0.001050	0.075796
gas_hoge_druk	0.013357	1.000000	0.012310	0.013163	0.004189	-0.000078	-0.000048	0.011385
gas_lage_druk	0.087463	0.012310	1.000000	0.087046	0.014707	0.003545	0.001319	0.071753
laagspanning	0.132357	0.013163	0.087046	1.000000	0.027657	0.007326	0.003517	0.080509
middenspanning	0.022050	0.004189	0.014707	0.027657	1.000000	-0.000156	-0.000097	0.019972
riool_onder_druk	0.004134	-0.000078	0.003545	0.007326	-0.000156	1.000000	0.012198	0.002031
riool_vrijverval	0.001050	-0.000048	0.001319	0.003517	-0.000097	0.012198	1.000000	0.001694
water	0.075796	0.011385	0.071753	0.080509	0.019972	0.002031	0.001694	1.000000

Figure 2 Correlation among different damage types

The Pearson's correlation matrix gives a value between -1 and 1 for each feature combination. A value of ±0.3 means weak, ±0.5 medium and greater than ±0.5 a strong correlation. As it can be seen above there is little to no correlation between damage types and thus, they cannot be grouped.

2.2.2 Removing irrelevant features?

By removing some features that consistently correlate to one another, the XGBoost model could be made more efficient, and thus the model training could be sped up. The do this the Pearson's correlation matrix was used again.

	OPPERVLAKTE	AANTAL_COORDINATEN	LENGTE	polygon_complex
OPPERVLAKTE	1.000000	0.107427	0.808838	-0.271885
AANTAL_COORDINATEN	0.107427	1.000000	0.431024	-0.099331
LENGTE	0.808838	0.431024	1.000000	-0.384929
polygon_complex	-0.271885	-0.099331	-0.384929	1.000000
Bebouwd_exclusief_bedrijfsterrein	0.401618	0.095655	0.340701	-0.149561
Bedrijfsterrein	0.282147	0.027743	0.236164	-0.094543
	0.300687	0.019288	0.231036	-0.065973

Figure 3 Some of the features that show strong correlations to each other

Features such as the client company- excavation company and the area of the excavation site- polygon area – the number of coordinates of the polygon - length of the polygon's edge correlate to each other. This is not surprising giving that regular client- contractor relationships are common, and that certain dimensions of an excavation site are related. Ultimately, they were left in the database separately as their number was small and XGBoost can do the feature selection by itself.

2.3 XGBOOST MODEL

2.3.1 Classification models that have been considered but rejected

One of the most challenging aspects of this project was choosing the right classification model. Most problem can either be solved by binary classification (whether something is true of false) or

multiclassification where the model has to decide which group a give case fits in the best. In the case of modelling damages, there is a possibility, even likely that one case can be correctly classified multiple ways. Unfortunately, the later solution is more complicated and have relatively little available documentation. Usually, it simplified or divided into smaller parts so they could be solved by the either of the first two. To do this, multiple approaches have been considered. Some that have been rejected:

One utility strike per row

These means that if during an excavation for example three different type of damages occurred then they would be treated as three separate events. This is statistically wrong because theoretically:

If features *a, d* and *e* cause event *x* 100% of the time

and the following event has been registered:

Features *a,b,c,d,e* caused event *x,y,z*

And it is treated as if three individual cases were

a,b,c,d,e caused event *x*

a,b,c,d,e caused event *y,*

a,b,c,d,e caused event *z*

Then the statistical model will think that the combination of *a,d* and *e* only result in 33% of the time in *x*.

XGBoost ignores both empty cells and false values when runs, only carrying about positive values, thus one can delete cells with features values and only cause slightly worse predictions. But if target values are removed then the database no longer represents reality.

Modelling only the damages

By removing all excavations that have no registered damages the modelling process and iteration time could be greatly accelerated, unfortunately this approach decreases the efficacy of the model on new data, but at least statistically correct. For example, if the following four excavation has been registered:

a, b, e caused *x*

a, b,f caused nothing

b,e,g caused *x*

a, b, g caused nothing

then it is a logical assumption that feature e causes event *x*. But if I remove the cases where no *x* occurred:

a, b, e caused *x*

b,e,g caused *x*

the maximum one can say that feature g is likely not necessary condition for event x to occur. If this data is used to train the model than it will be significantly worse when tested on a new dataset.

Removing the less important damage type when there are multiple damage types

If x and y that I want to predict, and there are way more damages with x then it seems logical to remove x when both x and y occurs during an excavation and then applying a traditional multiclassification.

2.3.2 Modelling, evaluation and tuning

Modelling

The two types of XGBoost models were created in the end. First binary classification models for each model individually, and one multi-class classification model to determine the most likely outcome for an excavation. The models were written in Python, using the dmlc XGBoost package, and were trained with all the available features of the dataset.

Evaluation

The models were evaluated by ROC curve (Receiver operating characteristic curve), precision-recall curve and analyzed by the feature importance vector.

ROC is the curve plotted of percentages of the positively classified positives divided by all actual positive cases, the true positive rate, on one axis, and the negatively classified negatives divided by all actual negative cases, the false positive rate, on the other. The area under the ROC curve is called AUC (Area under curve), which could be used to compare the performance of different or used during training.

The recall is the predicted real damages divided by all real damages. The precision is the predicted real damages divided by all predictions. The precision-recall curve is created is these two values plotted on x and y.

Both the ROC curve and the precision-recall curve are function of the threshold above which something is positively classified. The threshold is on an interval between 0 and 1 , by default it is 0.5 or with other words above 50% probability something is classified positive.

Another important performance factor was the efficiency of the model. This showed how much time it took to train the model, which was important given the likely need for many iterations to achieve higher accuracy. I did not expect the training time to be longer than ten minutes before results stopped improving without manual iteration. If this were not the case, the model should be changed, simplified, or subdivided in order to be more manageable.

Tuning

When tuning the models, it became clear that tuning the different parameters for so many damage types to achieve the best possible results would take very long. The simplest method to tune the parameters is one by one and see how the evaluation metric, in this case the harmonic average of precision and recall, changes. Then doing this till a point where no more improvement can be observed. The most common method to make this faster is to define realistic intervals for the parameters one wants to tune, then dived all intervals by a pre-selected number of times and find the combination with the best evaluation score. This method is called GridSearchCV, and it also takes quite long because the script has to evaluate many combinations that based on previous results could otherwise be removed. It also does not find the best combination only the one closest to it.

To avoid manually having to spend to tuning the model the BayesSearchCV package can be used. This creates a statistical model on top of the XGBoost model for parameter tuning and thus only evaluate increasingly promising combinations on smaller and smaller intervals. The entire process is automated and does not need manual intervention.

One iteration with a manually selected parameter combinations took about 6 minutes for one damage type, while finding the best possible combination to get the highest precision-recall score, using the BayesSearchCV package, took about 40 minutes for each binary models. For the multi-class model, it took about 4 hours.

```
search_space = {
    'clf__max_depth': Integer(2,8),
    'clf__learning_rate': Real(0.001, 1.0, prior='log-uniform'),
    'clf__subsample': Real(0.5, 1.0),
    'clf__colsample_bytree': Real(0.5, 1.0),
    'clf__colsample_bylevel': Real(0.5, 1.0),
    'clf__colsample_bynode' : Real(0.5, 1.0),
    'clf__reg_alpha': Real(0.0, 10.0),
    'clf__reg_lambda': Real(0.0, 10.0),
    'clf__gamma': Real(0.0, 10.0),
    'clf__scale_pos_weight' : Real(110, 111)  #10000.0)
}
```

Figure 4 Parameters of the XGBoost model that have been tuned

3. RESULTS

3.1 MODELLING THE PROBABILITY OF EACH DAMGE TYPES INDIVIDUALLY

Given the technical simplicity of setting up binary classifications, it was relatively easy to evaluate, iterate, and later used to answer questions on for example the acceptable ratio of precision and recall based on the cost of different actions.

3.1.1 'damage_datatransport'

Internet cable strikes can be reliably matched to about 1.5% of all excavations, or more precisely 30471 out of 2221673. Unfortunately, about one third of the internet cable damages were unmatchable. The overall performance of the model is fairly well matching the model used to predict any damage types using the same database (Li, 2023). This is not surprising given that internet cable strikes are the most numerous among all registered damage types, making up more than half of all cases. The AUC and recall scores of the two models are almost identical, 0.821-0.829 and 0.64-0.66. The biggest difference is the precision, the datatransport model classifying about twice as many false positives 0.05-0.09 when the cutoff is 50% (at what probability a case is classified positive). This difference in performance can be mainly accounted by the number of cases when one can match the damage to the excavation, but the damage type is missing.

```
Logistic ROC AUC 0.821
Logistic PR AUC 0.109
[[461182  86527]
 [  2790   4920]]
Balanced accuracy 0.740
              precision    recall  f1-score   support

         0.0       0.99      0.84      0.91    547709
         1.0       0.05      0.64      0.10      7710

    accuracy                           0.84    555419
   macro avg       0.52      0.74      0.51    555419
weighted avg       0.98      0.84      0.90    555419
```

Model	XGBoost
Features used	63
Training time (s)	28
Predicting time	2
AUC score	0.829
PR score	0.186
Balanced accuracy	0.749
Precision	0.090
Recall	0.660
F1-score	0.160

Figure 5 Different evaluation metrics of the datatransport damage model : Metrics of any damage prediction model (Li, 2023)

The recall and precision scores have been calculated for 50% cutoff, but they do not have to be. Based on other factors such as cost and danger other combinations can be read of the curve bellow. For example, if correctly classifying only 20% of the damages that happened than the incorrectly classified ones are around 80%.

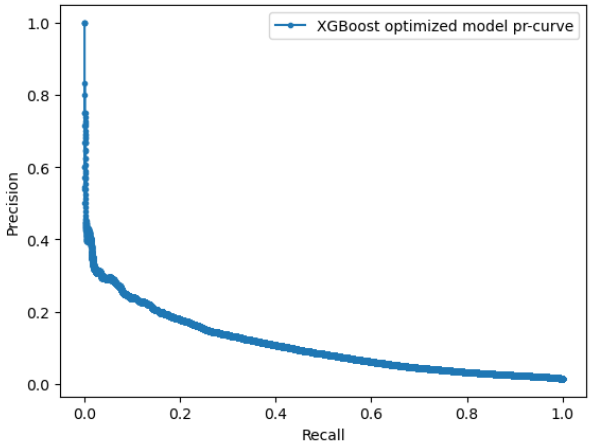


Figure 6 Precision-Recall curve of the datatransport model

All of the top ten most important features for prediction are either related to the dimensions of the excavation site or to the companies (client, contractor) involved. This means that per m2, the probability of damage is mostly the same, no matter the location and other features like the soil type, only the companies involved.

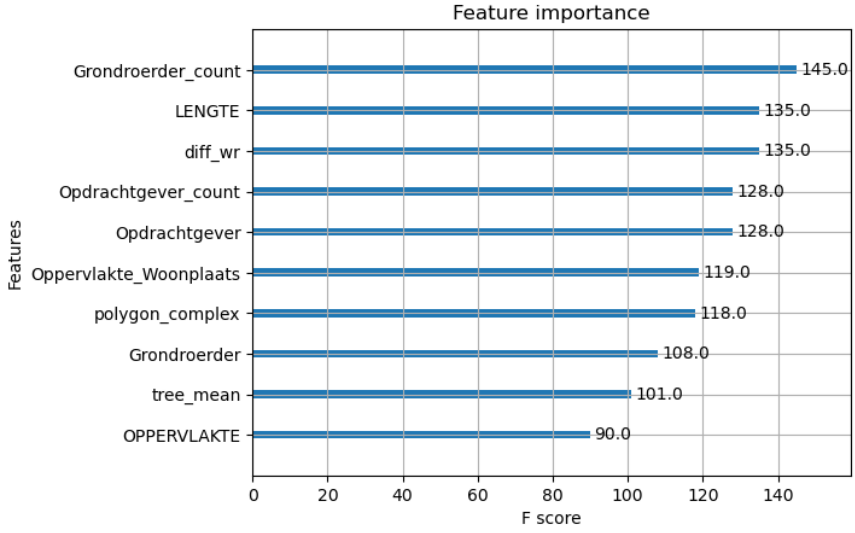


Figure 7 Feature importance of the datatransport model

3.1.2 'damage_laagspanning'

There are 20047 matched low voltage electric cable strikes or approximately 1% of all excavations. While the precision has decreased the recall increased at 50% cutoff, so ultimately the performance of this model is very similar to the internet cable one.

```
Logistic ROC AUC 0.836
Logistic PR AUC 0.082
[[448380 102055]
 [  1475   3509]]
Balanced accuracy 0.759
              precision    recall  f1-score   support

         0.0       1.00      0.81      0.90    550435
         1.0       0.03      0.70      0.06      4984

    accuracy                           0.81    555419
   macro avg       0.51      0.76      0.48    555419
weighted avg       0.99      0.81      0.89    555419
```

Figure 8 Different evaluation metrics of the laagspanning damage model

3.1.3 'damage_gas_lage_druk'

For low pressure gas pipe damage, there are 7672 matched events. Compared to the previous cases this number is much lower, and it is reflected in the precision required to achieve the same 60%-70% recall rate.

```
Logistic ROC AUC 0.828
Logistic PR AUC 0.034
[[439454 114064]
 [   614   1287]]
Balanced accuracy 0.735
              precision    recall  f1-score   support

         0.0       1.00      0.79      0.88    553518
         1.0       0.01      0.68      0.02      1901

    accuracy                           0.79    555419
   macro avg       0.50      0.74      0.45    555419
weighted avg       1.00      0.79      0.88    555419
```

Figure 9Different evaluation metrics of the gas lage druk damage model

3.1.4 'damage_water' - 5289

Similarly, to low pressure electricity, the matched water pipe strikes numbers are much lower than internet or low voltage electricity, and thus the precision of the prediction is accordingly lower too. Whether these are still usable or not depends on the use case.

```
Logistic ROC AUC 0.808
Logistic PR AUC 0.032
[[430078 123984]
 [   443    914]]
Balanced accuracy 0.725
              precision    recall  f1-score   support

         0.0       1.00      0.78      0.87    554062
         1.0       0.01      0.67      0.01      1357

    accuracy                           0.78    555419
   macro avg       0.50      0.72      0.44    555419
weighted avg       1.00      0.78      0.87    555419
```

Figure 10 Different evaluation metrics of the water damage model

3.1.5 'damage_middenspanning' and others damage types with less than 1000 matched cases.

The quality of the predictions for damage types with less than 1000 matched cases are very poor. In order to classify over 70% of real damages, more than 100% of the predicted cases are wrong. These damage types include high voltage, high pressure, sewer, heat pipes etc. .

```
Logistic ROC AUC 0.828
Logistic PR AUC 0.003
[[387204 168006]
 [    45    164]]
Balanced accuracy 0.741
              precision    recall  f1-score   support

         0.0       1.00      0.70      0.82    555210
         1.0       0.00      0.78      0.00       209

    accuracy                           0.70    555419
   macro avg       0.50      0.74      0.41    555419
weighted avg       1.00      0.70      0.82    555419
```

Figure 11 Different evaluation metrics of the middenspanning model

Part of the reason for this low number is likely due to the fact that most of these utilities are also less frequent. To prove this point, the most important feature to predict them, is their presence at the location, while none of the previous damage type had its matching utility's existence in their top ten features.

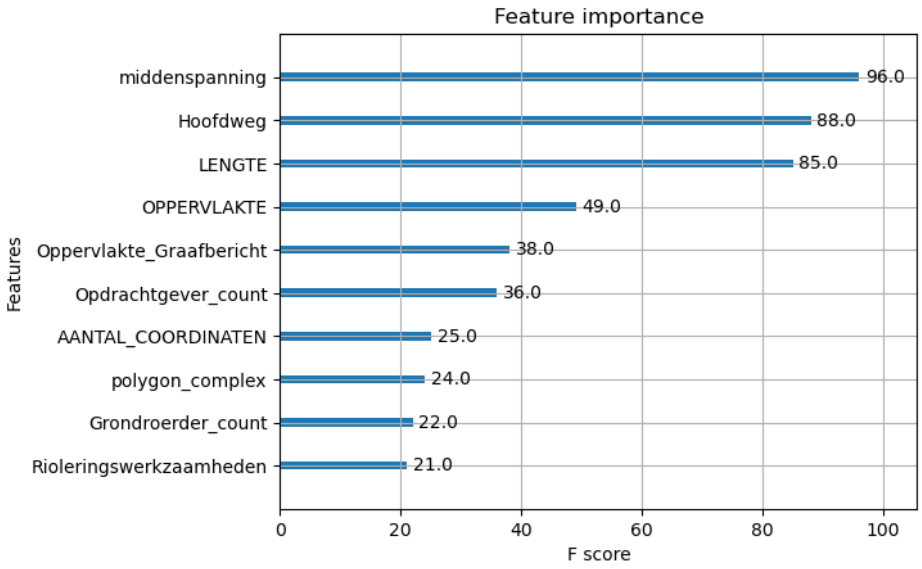


Figure 12 Feature importance of the model for medium voltage electricity cable damages prediction

3.2 MODELLING THE PROBABILITY OF EACH DAMAGE TYPES SIMULTANIOUSLY

The other possible useful model type is the simple multiclass classification, predicting the mostly likely outcome . This means that for each excavation only one prediction is made as if more than one outcome would be impossible. To prepare the data for this, excavations with more than one damage types registered has to be removed. Luckily damage types that could not be predicted with binary classification can be ignored, decreasing the number of excavations that has to be removed before the model training. In this case the possible classes are no damage, data cable, low voltage, low pressure gas and water. Unfortunately, even after preparing the model for the severe imbalance between no damage and damage and even between damage types by scaling the types and evaluate them weighted by the imbalance, the results were poor or more precisely not too useful. The model correctly assumes that no damage is the most likely outcome in most cases. Only for very large excavation sites classified the model the damages to internet cables more likely.

Ultimately the multi-class model does not have any benefit over a series of binary classification, that can be better tuned, other than the elegance of execution.

```
------------------ Confusion Matrix -----------------

[[541401     46      2     34      0]
 [  6248     21      0      8      1]
 [  1381      1      0      1      0]
 [  3824      3      0      9      0]
 [   915      0      0      1      0]]

Accuracy: 0.98
Balanced Accuracy: 0.20

Micro Precision: 0.98
Micro Recall: 0.98
Micro F1-score: 0.98

Macro Precision: 0.29
Macro Recall: 0.20
Macro F1-score: 0.20

Weighted Precision: 0.96
Weighted Recall: 0.98
Weighted F1-score: 0.97

--------------- Classification Report ---------------

              precision    recall  f1-score   support

           0       0.98      1.00      0.99    541483
           1       0.30      0.00      0.01      6278
           2       0.00      0.00      0.00      1383
           3       0.17      0.00      0.00      3836
           4       0.00      0.00      0.00       916

    accuracy                           0.98    553896
   macro avg       0.29      0.20      0.20    553896
weighted avg       0.96      0.98      0.97    553896
```

Figure 13 Evaluation of the multi-class classification

4. CONCLUSION

The main objective of the project has been partially accomplished. Damages related to common utilities, water, internet, low-voltage electricity and low-pressure gas can be predicted, with 80% recall rate for 5-1% precision. Whether this can be used in practice it depends on the given utility and factors such as the cost of different interventions and the amount of direct or indirect danger to human life.

Damage types that could be predicted such as high-pressure gas, high-voltage electricity, etc. have all in common that they are not present in most places, and thus the number of registered damages is order or magnitudes lower than the ones' that could be predicted. Given the large size difference even including extra years in the process would not help.

Answers found for the subobjective: Only features related to the size of the excavation site correlate and no damage type does to another. Out of all modelling options by far the simplest binary classification proved to be the most useful. It is both the easiest to train while provides the most of information, while more complex models provided no extra benefit.

5. RECOMMENDATION

In order to evaluate the usefulness of the damage prediction of the common utility strikes, further research is needed on the possible interventions such as the cost of fixing certain damages, standing standby, and what kind of dangers different damages present.

Many of the registered damages could not be used because the Klic number was wrong, and there is no matching excavation in the database. To avoid this Klic numbers could be generated according to an 'secret' algorithm and then when a damage registered its number could be check if it is at least a valid number or not. While this would not allow the prediction of less common damages, but it would certainly improve the precision of the common ones.

The most important features where the ones related to size of the excavation or the companies involved. As the size does not affect the probability of damage per unit area, the only thing that can be influenced is the companies that were involved. Thus, the companies with the highest and lowest damage rates should be investigated to see what do they do different from other that influences their performance.

6. REFERENCES

Li, J. (2023). Exploration of Machine Learning Approaches in predicting Excavation Damages. *Yet to be published*.

Lim, S., & Chi, S. (2019). Xgboost application on bridge management systems for proactive damage estimation. *Advanced Engineering Informatics, 41*, 100922. doi:https://doi.org/10.1016/j.aei.2019.100922

Metje, N., Ahmad, B., & Crossland, S. M. (2015, September). Causes, impacts and costs of strikes on buried utility assets. *ice | proceedings, 168*(3), 165-174. doi:https://doi.org/10.1680/jmuen.14.00035

Orton, H. (2013, July-August). History of underground power cables. *IEEE Electrical Insulation Magazine, 29*(4), 52-57. doi:10.1109/MEI.2013.6545260.

Xiang, W., & Zhou, W. (2021). Bayesian network model for predicting probability of third-party damage to underground pipelines and learning model parameters from incomplete datasets. *Reliability Engineering and System Safety*, 205. doi:https://doi.org/10.1016/j.ress.2020.107262

7. APPENDIX

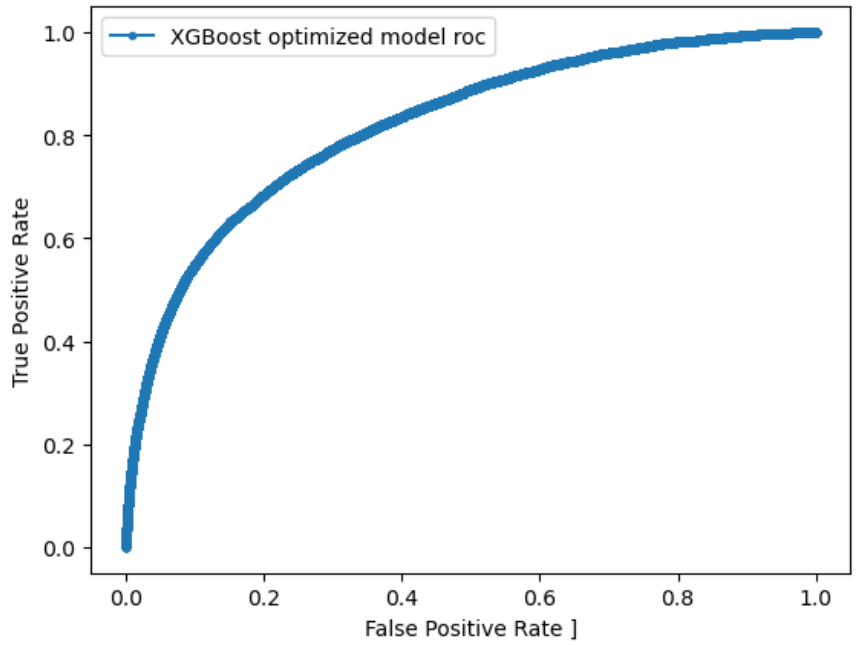
7.1 EVALUATION METRICS OF THE INDIVIDUAL XGBOOST MODELS

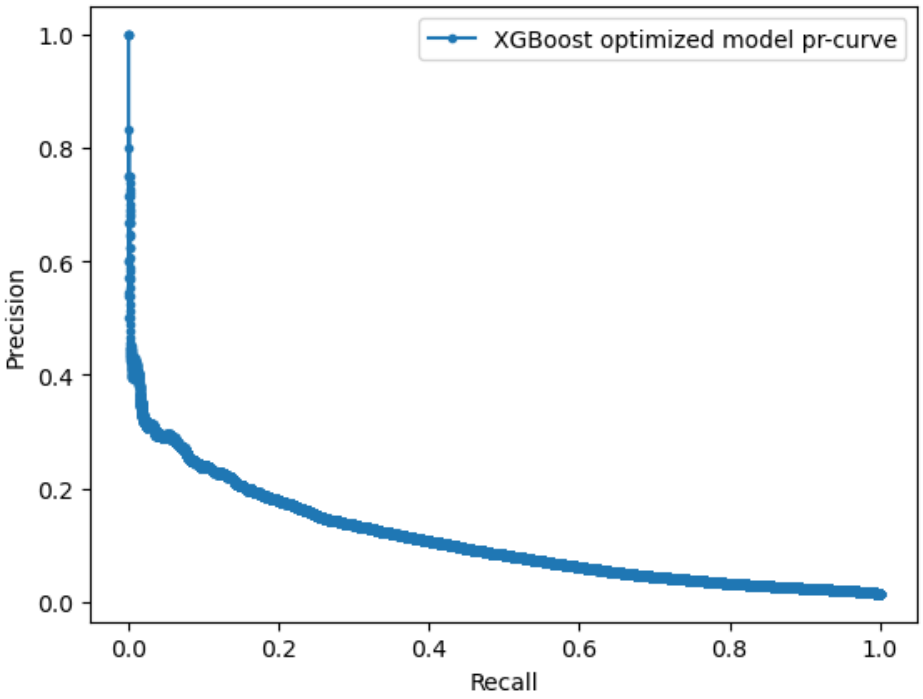
7.1.1 'damage_datatransport'

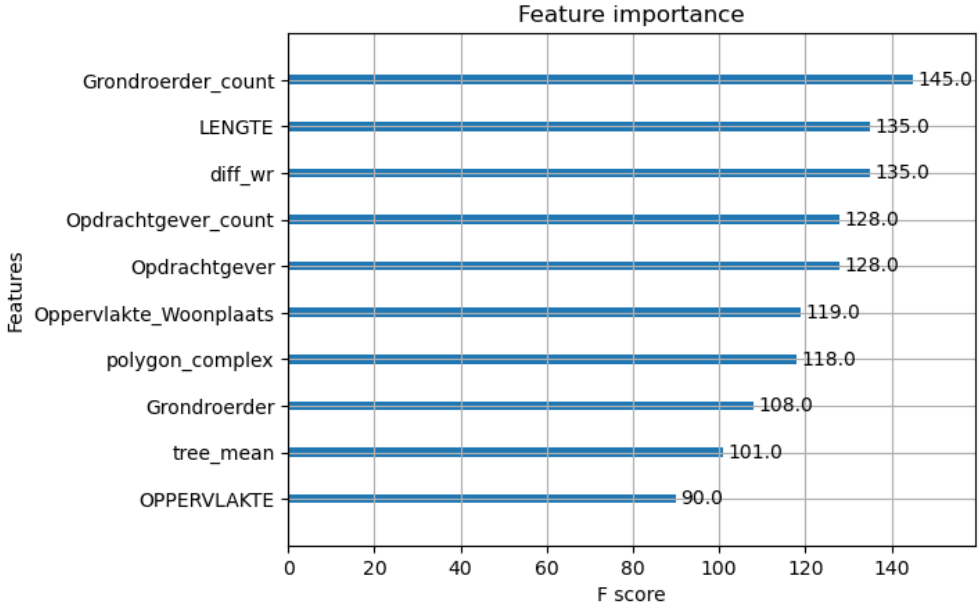
```
Logistic ROC AUC 0.821
Logistic PR AUC 0.109
[[461182  86527]
 [  2790   4920]]
Balanced accuracy 0.740
              precision    recall  f1-score   support

         0.0       0.99      0.84      0.91    547709
         1.0       0.05      0.64      0.10      7710

    accuracy                           0.84    555419
   macro avg       0.52      0.74      0.51    555419
weighted avg       0.98      0.84      0.90    555419
```




XGBoost optimized model pr-curve
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

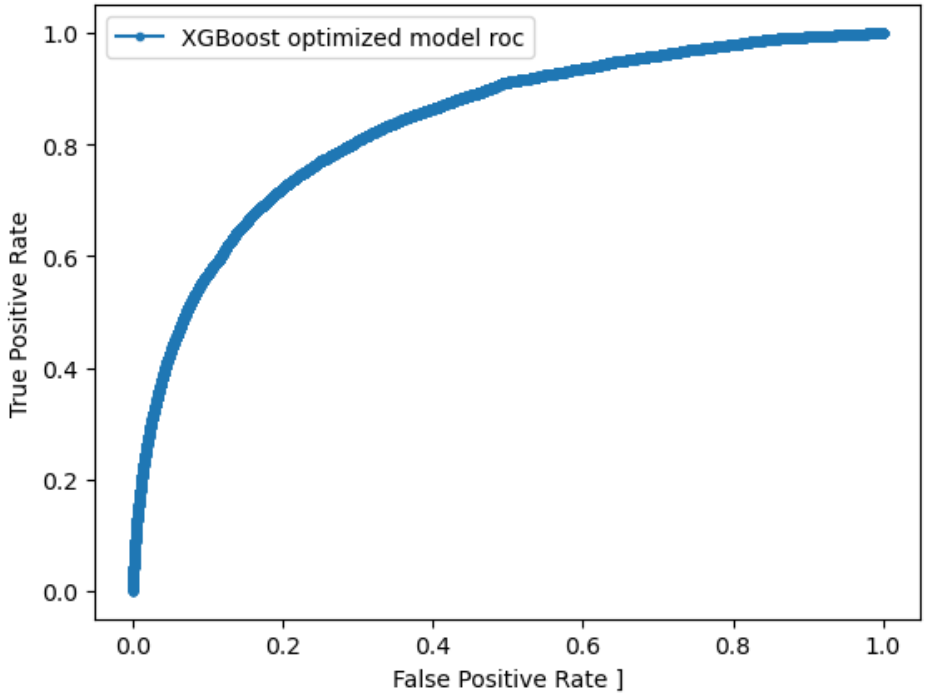

Feature importance
Features
Grondroerder_count
LENGTE
diff_wr
Opdrachtgever_count
Opdrachtgever
Oppervlakte_Woonplaats
polygon_complex
Grondroerder
tree_mean
OPPERVLAKTE
145.0
135.0
135.0
128.0
128.0
119.0
118.0
108.0
101.0
90.0
0
20
40
60
80
100
120
140
F score

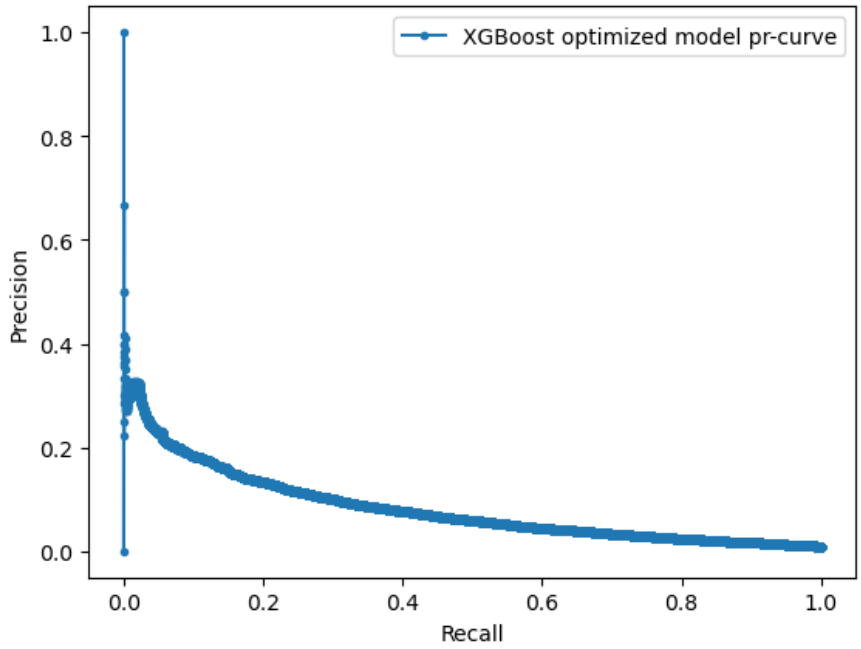
7.1.2 'damage_laagspanning' - 20047

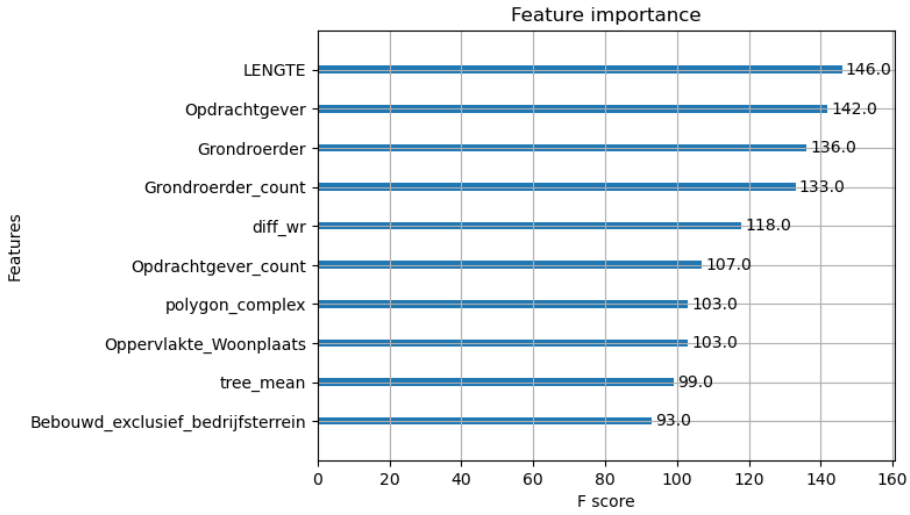
```
Logistic ROC AUC 0.836
Logistic PR AUC 0.082
[[448380 102055]
 [  1475   3509]]
Balanced accuracy 0.759
              precision    recall  f1-score   support

         0.0       1.00      0.81      0.90    550435
         1.0       0.03      0.70      0.06      4984

    accuracy                           0.81    555419
   macro avg       0.51      0.76      0.48    555419
weighted avg       0.99      0.81      0.89    555419
```





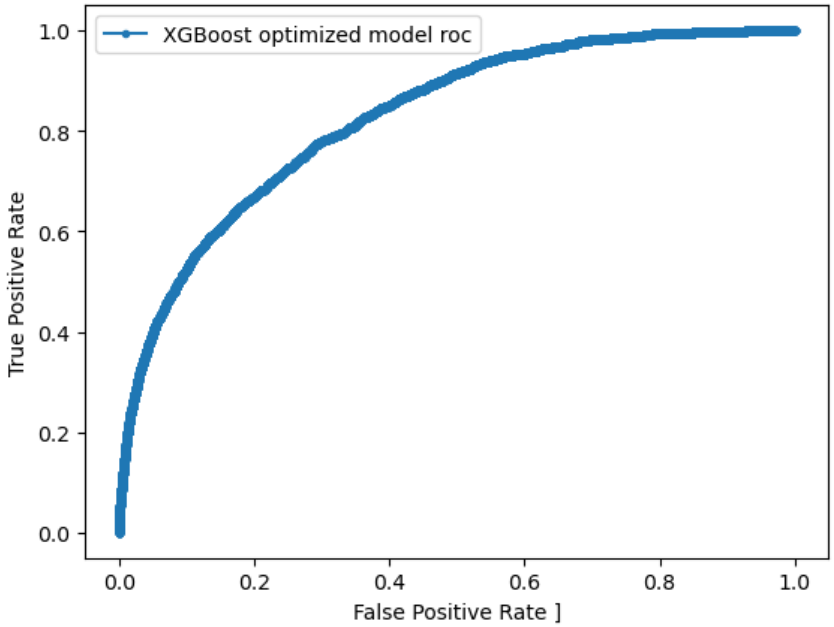


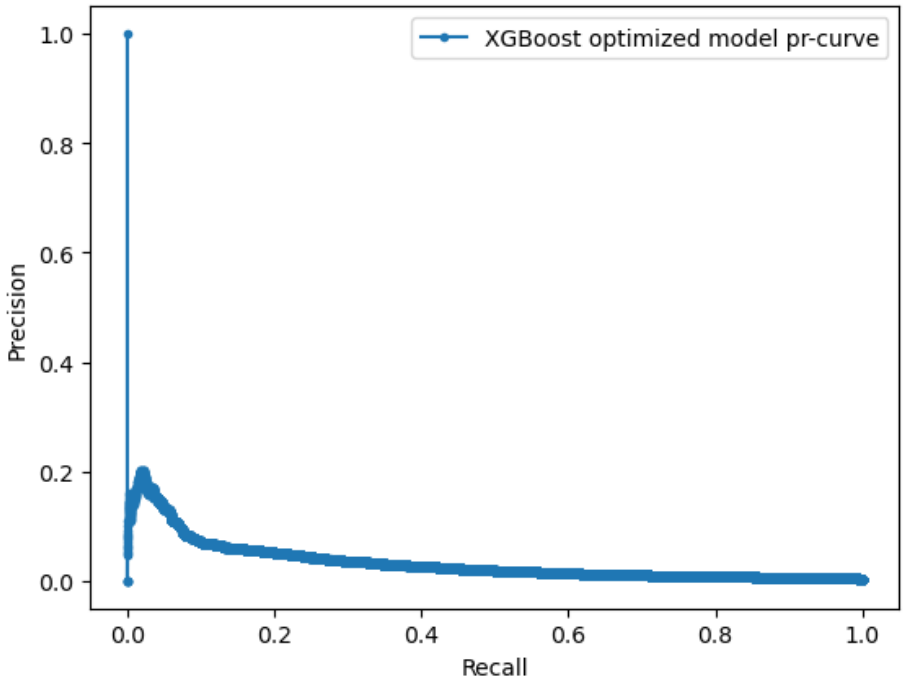
7.1.3 'damage_gas_lage_druk' - 7672

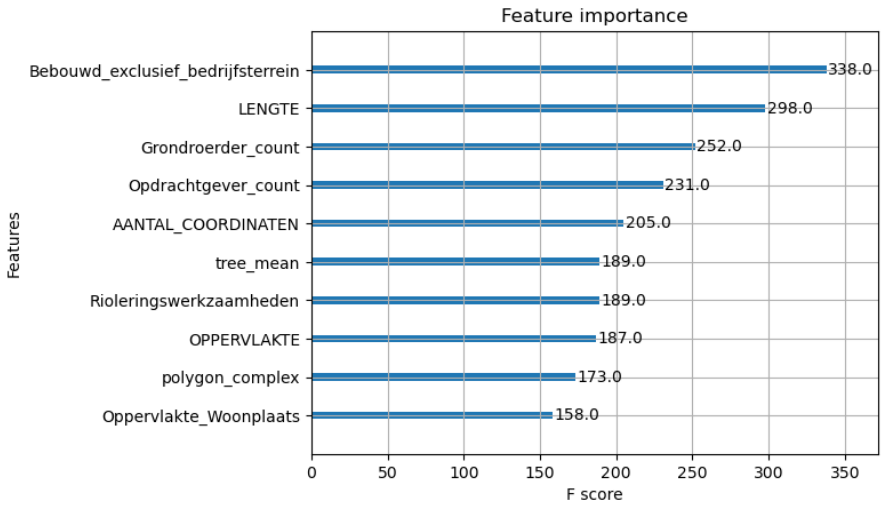
```
Logistic ROC AUC 0.828
Logistic PR AUC 0.034
[[439454 114064]
 [   614   1287]]
Balanced accuracy 0.735
              precision    recall  f1-score   support

         0.0       1.00      0.79      0.88    553518
         1.0       0.01      0.68      0.02      1901

    accuracy                           0.79    555419
   macro avg       0.50      0.74      0.45    555419
weighted avg       1.00      0.79      0.88    555419
```


XGBoost optimized model roc
True Positive Rate
False Positive Rate]
0.0
0.2
0.4
0.6
0.8
1.0


XGBoost optimized model pr-curve
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0

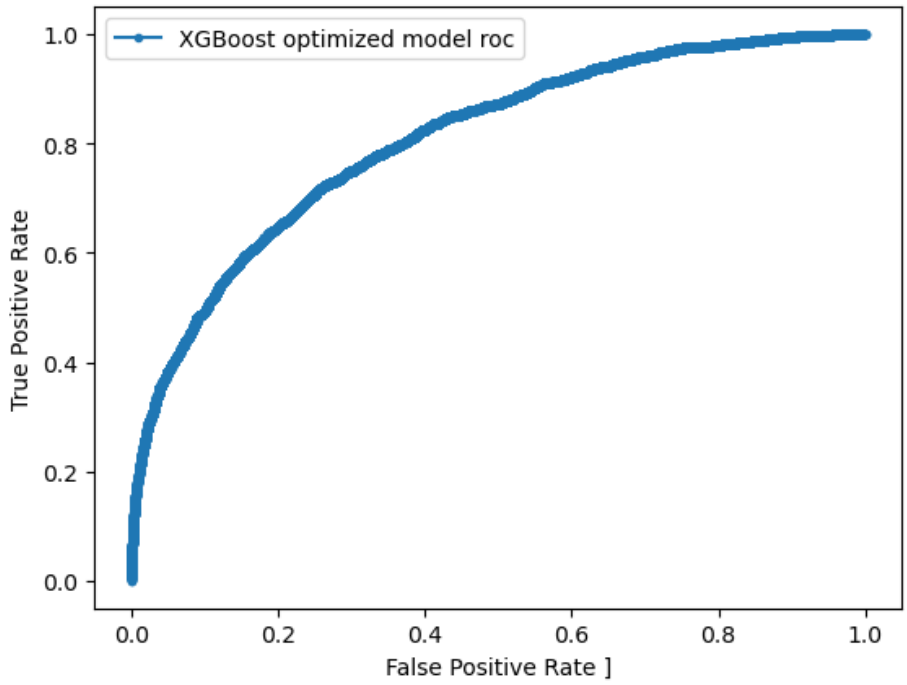


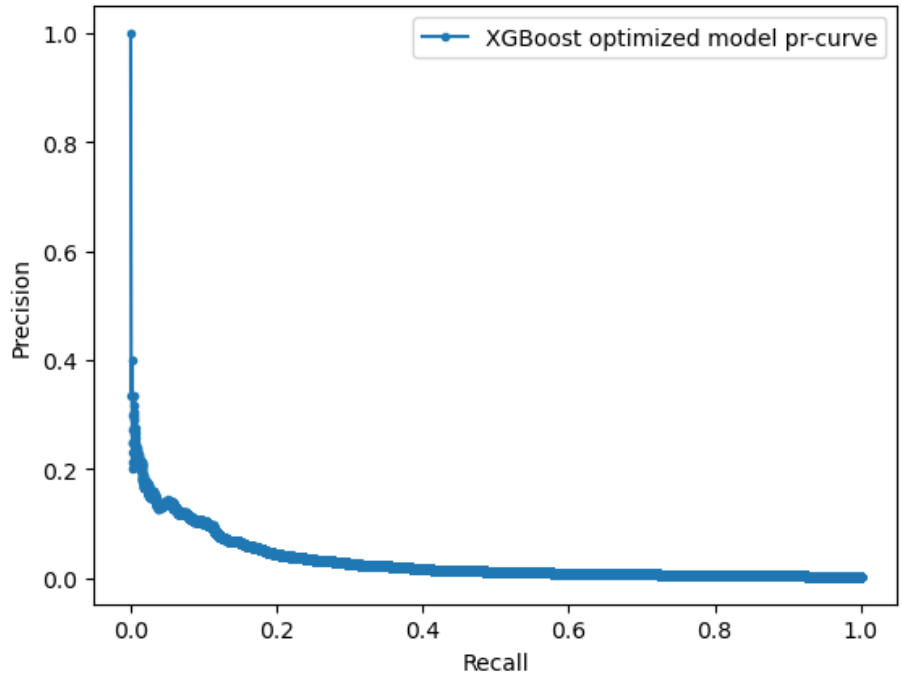
7.1.4 'damage_water' - 5289

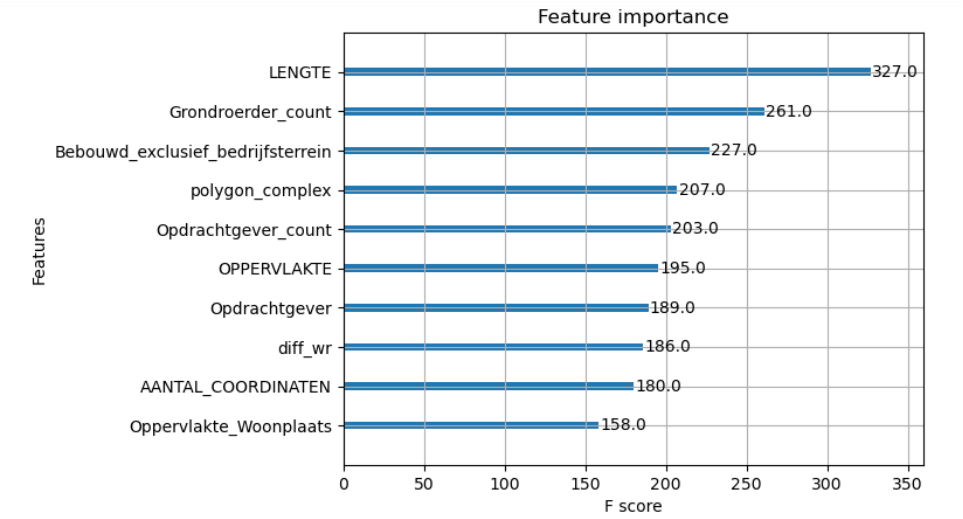
```
Logistic ROC AUC 0.808
Logistic PR AUC 0.032
[[430078 123984]
 [   443    914]]
Balanced accuracy 0.725
              precision    recall  f1-score   support

         0.0       1.00      0.78      0.87    554062
         1.0       0.01      0.67      0.01      1357

    accuracy                           0.78    555419
   macro avg       0.50      0.72      0.44    555419
weighted avg       1.00      0.78      0.87    555419
```


XGBoost optimized model roc
True Positive Rate
False Positive Rate]
0.0
0.2
0.4
0.6
0.8
1.0


XGBoost optimized model pr-curve
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0

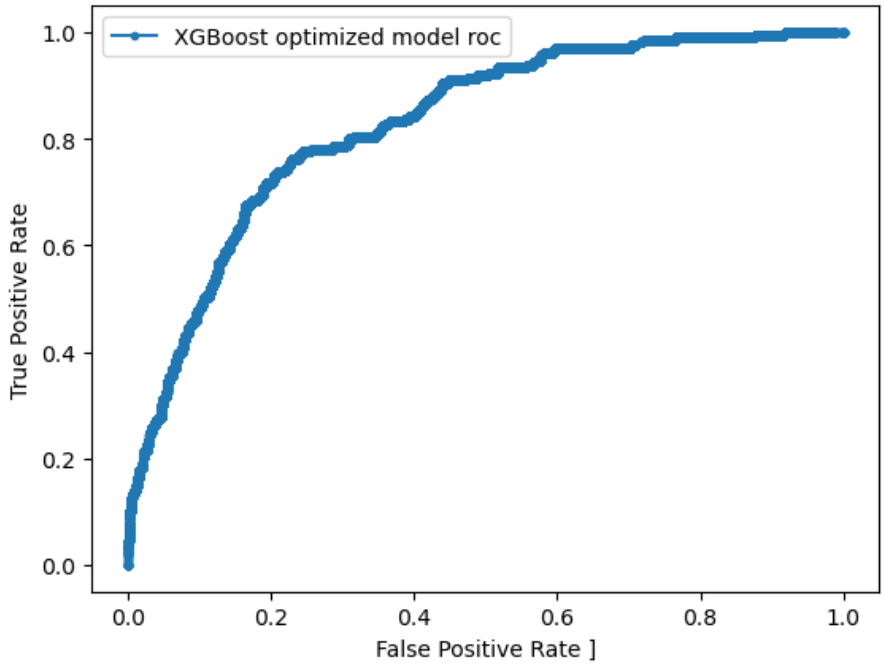


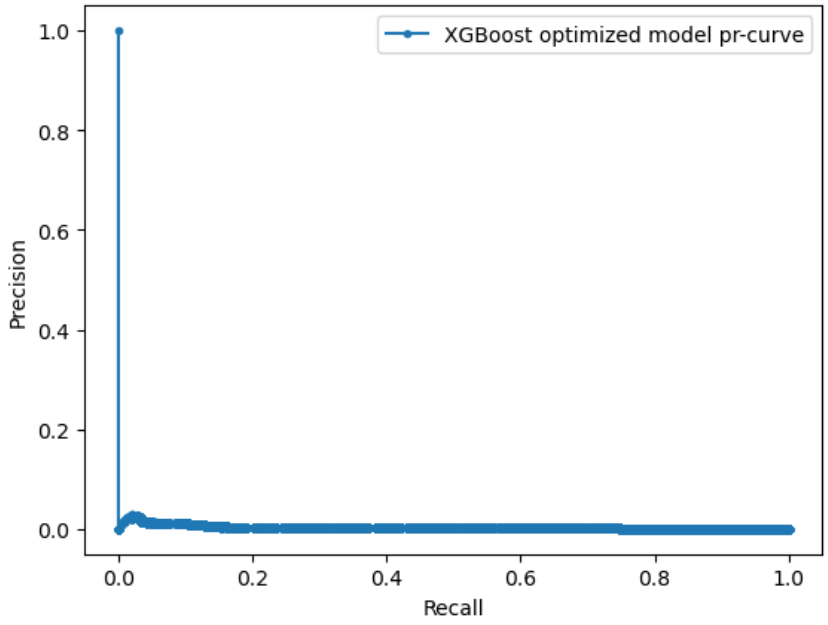
7.1.5 'damage_middenspanning' - 913 and others that do not have enough data

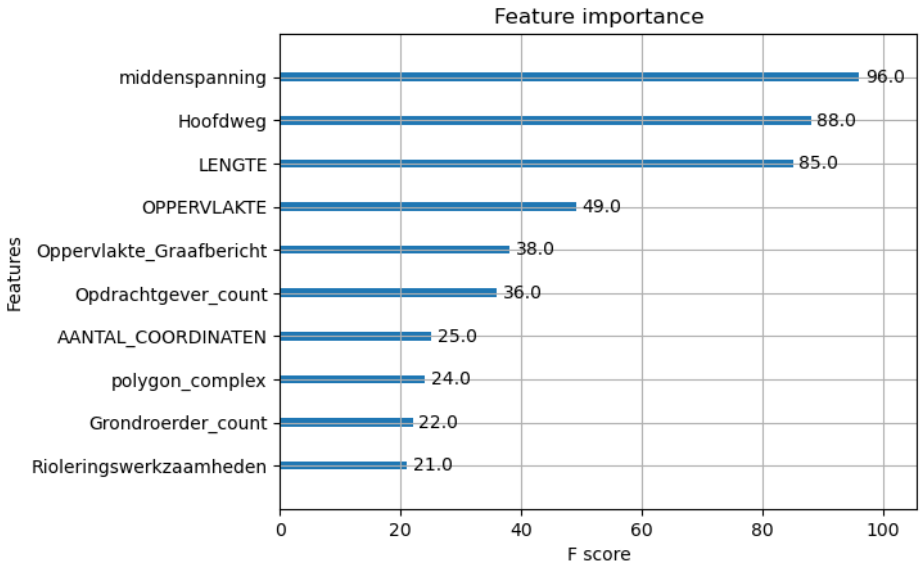
```
Logistic ROC AUC 0.828
Logistic PR AUC 0.003
[[387204 168006]
 [    45    164]]
Balanced accuracy 0.741
              precision    recall  f1-score   support

         0.0       1.00      0.70      0.82    555210
         1.0       0.00      0.78      0.00       209

    accuracy                           0.70    555419
   macro avg       0.50      0.74      0.41    555419
weighted avg       1.00      0.70      0.82    555419
```


XGBoost optimized model roc
True Positive Rate
False Positive Rate]
0.0
0.2
0.4
0.6
0.8
1.0


XGBoost optimized model pr-curve
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0



7.2 EVALUATION METRICS OF THE MULTI-CLASS CLASSIFICATION

```
------------------ Confusion Matrix -----------------

[[541401     46      2     34      0]
 [  6248     21      0      8      1]
 [  1381      1      0      1      0]
 [  3824      3      0      9      0]
 [   915      0      0      1      0]]

Accuracy: 0.98
Balanced Accuracy: 0.20

Micro Precision: 0.98
Micro Recall: 0.98
Micro F1-score: 0.98

Macro Precision: 0.29
Macro Recall: 0.20
Macro F1-score: 0.20

Weighted Precision: 0.96
Weighted Recall: 0.98
Weighted F1-score: 0.97

--------------- Classification Report ---------------

              precision    recall  f1-score   support

           0       0.98      1.00      0.99    541483
           1       0.30      0.00      0.01      6278
           2       0.00      0.00      0.00      1383
           3       0.17      0.00      0.00      3836
           4       0.00      0.00      0.00       916

    accuracy                           0.98    553896
   macro avg       0.29      0.20      0.20    553896
weighted avg       0.96      0.98      0.97    553896
```