Predicting GDPR Fines and Penalties on Universities and Public Institutions

Alex Hartmann, Nikolaos Athanasopoulos, Athanasios Rakitzis, Vasileios Papadakis

Department of Advanced Computing Sciences, Maastricht University

June 22, 2022

Preface

This report is part of a semester project and source code and dataset can be found on GitHub.

Abstract

GDPR is a European Union law, which entered into force in 2016, with rules for how organisations and companies must handle personal data in an honest and user-friendly way. Our goal was to identify which are the main indicators regarding fine amount for Universities or Public Institutions. To accomplish that, we use data from the website GDPR Enforcement Tracker which among others, includes summaries of issued fines, country, the fined institution, date of decision and fine amount. The text summaries were used to extract features from them which are later processed and used as predictors to build our prediction model. Two regression models and a decision tree are presented, all showing good performance on this task. Among the most important predictors are violated articles as well as type of data involved. Violation of Article 32 (Implementing sufficient security measures) was identified as the most important predictor of fine amount. Also, student data is associated with higher fines while family data is associated with lower fines.

1 Introduction

The General Data Protection Regulation (GDPR) is a legal framework that sets guidelines for the collection and processing of personal information from individuals who live in the European Union (EU). The Regulation applies regardless of where the entity which collects or processes the personal data is based. That means that its guidelines must be followed by everyone that handles personal data of residents of the EU. It entered into force in 2016 and since then thousands of fines have been imposed on organisations that didn't comply. The amount ranges from some hundreds of euros to millions, so finding ways to avoid it, surely is important for them.

In this project we aim to provide insight on a way that the fine can be avoided, by identifying key factors extracted from previous fines and the summary of the fine decision.

Specifically we address the following research questions:

1. Can the data of organisations that got fined and the corresponding court decisions be used, to identify what led to the fine?
2. Can accurate fine predictions be made based on this sample regarding public organisations that have not received a fine yet?

In an effort to answer those questions we tried different approaches. First of all we do a descriptive analysis of the data to have an understanding of its structure and identify possible patterns that might be interesting. Also we use the summaries provided by the GDPR Enforcement Tracker website and apply text mining techniques to identify keywords and extract features that can later be used to train fine amount prediction models.

2 Related Work

Not much research has been done on the GDPR fine data, since the law is quite new. The only relevant paper to our work is Ruohonen and Hjerppe (2020).

It examines GDPR articles referenced in the enforcement decisions, but also predicts the amount of fines by using meta-data and text mining features extracted from the decision documents. They encountered similar problems as we did when trying to retrieve the documents and translate them but due to not having the limitation of

working only with universities and public institutions, the resulted dataset was not as small. According to their results, the most frequently referenced articles were related to the general principles lawfulness, and information security. Furthermore, they achieved good predictions using simple machine learning techinques like (log-)linear regressions and analyses of variance.

3 Data

The GDPR Enforcement Tracker lists all the fines which data protection authorities of the European Union have imposed to companies and organisations under the General Data Protection Regulation. The database is tracked by CMS, an international law firm that offers legal and tax advisory services. In total, the dataset contains 86 University and public institution fines.

Furthermore publicly available information about Universities was collected manually. When available, budget, number of students, number of employees and rankings were added to enrich the dataset. These characteristics might be related to the reasons those institutions ended up breaking the data privacy protection law.

An excerpt from the dataset can be seen in Table 5.

3.1 Data Limitations

Due to data limitations we had to abandon some of our ideas or search for alternative ways to implement them. Below we list some of the main data limitations we encountered.

- GDPR Law adopted on 2016 but became enforceable on May 2018. In only 4 years it is logical not to have many fines imposed. It is even more rare to have fines on Universities as they represent only a very small fraction of the organisations that GDPR law applies.

- The dataset is up to date as of June 22nd 2022. However, it does not contain all the fines since not all of them are made public. That limits the dataset even further.

- In an effort to extract information from text, we tried to use the legal decision documents. A link on the Enforcement Tracker website directs to those documents. The main problems with using the detailed reports are:

  Scraping the text is a challenge since they are inconsistent from country to country. Some are in web format and some others in pdf. When trying to transform it in a way that we can process it, the text is almost destroyed, due to pdf encoding or HTML elements that are blended with the text.

  Every country releases the legal document in the corresponding language. Even if we managed to retrieve the text, we would still face the challenge of translation. An automatic translation (e.g.Google translation API) would possibly lead to inaccuracies as it wouldn't be able to translate the correct legal terms in different documents.

  Some countries publish yearly reports containing all the fines of the previous year. Since we cannot detect exactly where the fine we are interested in is located in the document, scraping is difficult. Moreover, the Google translate api is often misleading because the documents are law domain specific, which makes it harder to automatically translate effectively.

3.2 Descriptive Statistics

In this section, a descriptive analysis of the dataset along with the obtained results will be discussed. The basic descriptive statistics are presented in Table 1.

Table 1: Descriptive Statistics for Fine Amount

Statistic	Value
Mean	131915€
SD	443685€
Min	290€
25th Quantile	5000€
Median	15000€
75th Quantile	50000€
Max	3700000€

It can be observed that the data range for fine amounts is very skewed, ranging from a few hundred euros all the way up to millions of Euros, with most observations clustered towards the lower end of the scale. This is why a logarithmic scale is used for most of the calculations that contain fine amounts. The next step was to calculate the fine amounts corresponding to each country in the dataset, in order to detect any patterns. The total fine amounts can be seen in Figure 1.

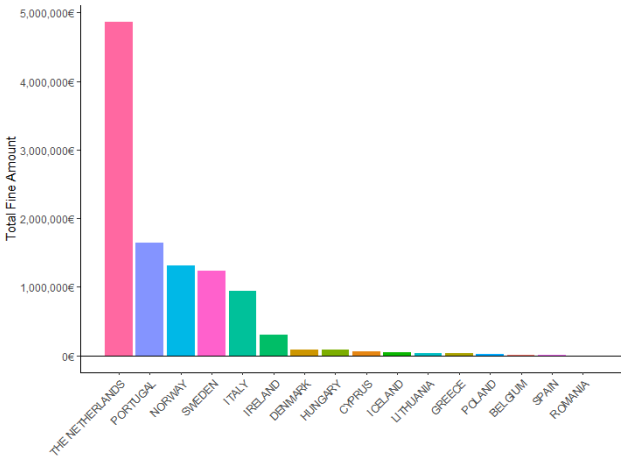

Figure 1: Fine Amounts per Country

Looking more closely at the individual fines, the Netherlands has a single fine amount of 3,700,000€, amounting to 35 percent of the total fine amount of the entire dataset. In general, the Netherlands does not have

many fines imposed, but one very significant one. The exact opposite is true for Italy and Norway, where there are many small fines imposed, but the final sum of fines is low.

Investigating developments over time, the cumulative fine amount is calculated. It can be seen in Figure 2 that the cumulative total fines can be approximated with a quadratic function of time.

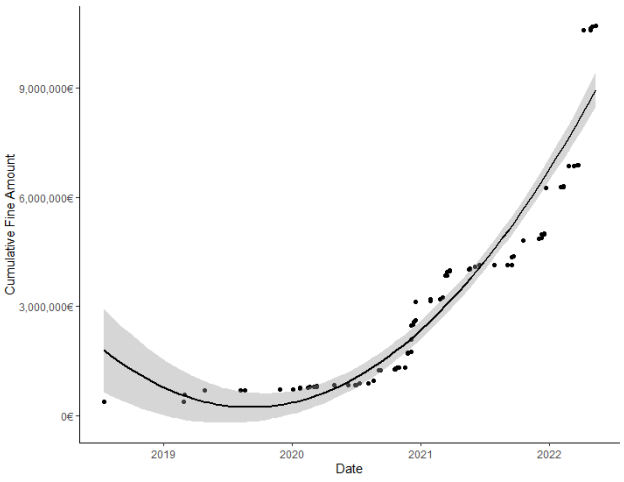

Figure 2: Cumulative Fine Amount plotted against time with 95% confidence intervals of quadratic fit as shaded regions.

The next step was to aggregate fine amounts per geographic region in the data. As a standard for categorizing countries, *EuroVoc* is used. It is a thesaurus maintained by the Publications Office of the European Union, and therefore a good source on European region classifications.

Tallying the total fines per region, the result can be seen in Figure 3. Western Europe is the region incurring the highest total fine amount, while Eastern Europe incurred the lowest.

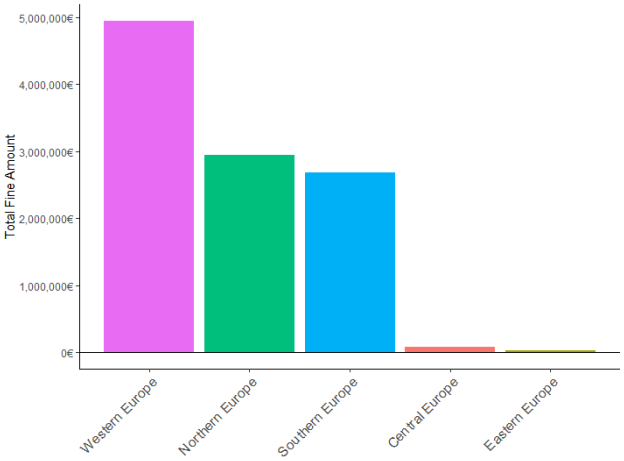

Figure 3: Sum of Fine Amounts per Region

Even though western Europe had the highest total fine amount, the number of fines imposed is the greatest in northern Europe. This can be observed in Figure 4.

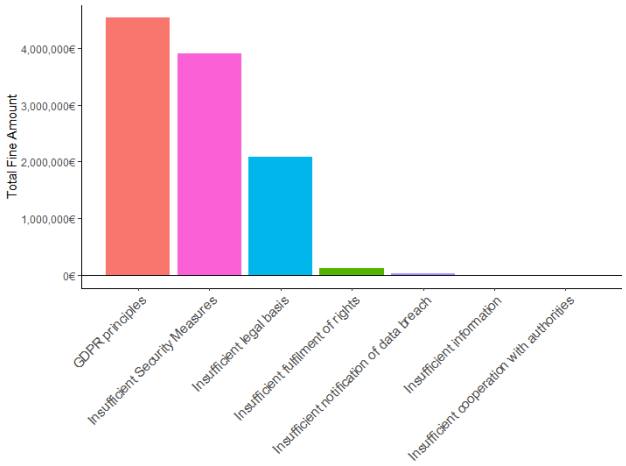

Figure 5: Total Fine Amount by type of violation

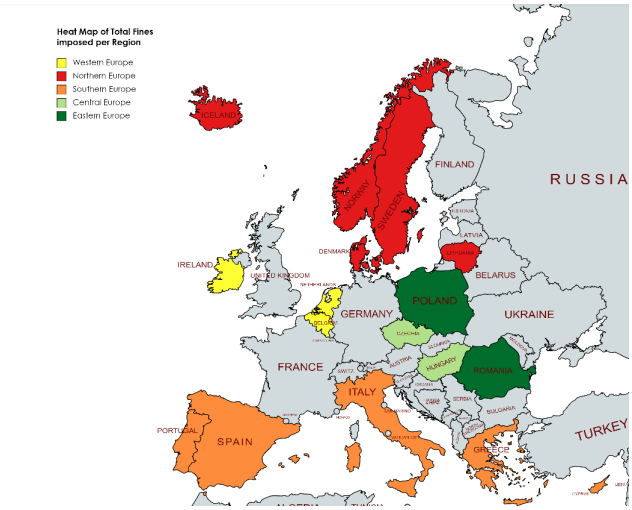

Figure 4: Heat Map of fine counts by region. The color gradient shows the amount of fines imposed on any country in the region.

Since the dataset also provides fine categories, it is interesting to examine the relationship of fine amount with those. In Figure 5, one can see that the largest amount is fined on the basis of violation of the four basic principles of GDPR. The second most fined type of violation is on the basis of insufficiency of security measures.

In terms of articles, the count of fines referring to each article can be seen in Figure 6. Article 5 is referenced most often, followed by Article 32.

4 Text Mining

Our main goal with text mining is to extract, modify or create valuable variables/features that could be used as potential predictors for our model in order to predict fine

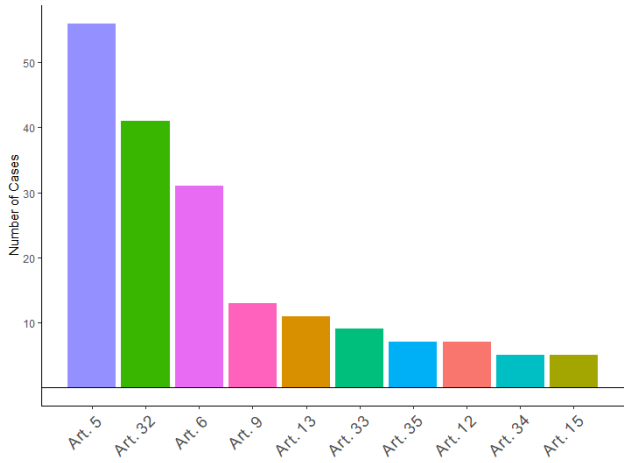

Figure 6: Number of Cases referencing each article

amounts. The methods that are used to achieve this purpose are described below.

4.1 Preprocessing

The first step of any text mining related work and hugely important is the preprocessing our text in order to make it ready for analyzing. First, reports are tokenized using SpaCy (Honnibal and Montani, 2017) and punctuation characters replaced with white space.
Then, we lowercase all tokens so we do not count each word more than once. Also we remove stopwords which do not add semantic value, like "and" or " however" etc. For stopword removal an extended list of stopwords is used along with many non-useful words for our purpose that we manually added by looking at later results, like "datatilsynet" which is for example the Data Protection Agency(DPA) of Norway. Finally, we lemmatize our tokens, aiming to remove inflectional endings only and to return the lemma of each word.
After we finished with preprocessing our data, we run a few tests to deduct token importance (e.g. with tf-idf, which shall be explained later down the report) and we noticed that one specific token, "data" was converted to "datum" and thus we decided to not lemmatize it. Lastly, we also removed sequences of numbers because sometimes they were shown to be important tokens, but they would not provide us with any relevant information, since they were just the corresponding fine amounts of each report.

4.2 Feature Extraction

4.2.1 Article Extraction

First of all we were interested in extracting all article violations corresponding to each fine report. As one can assume, we deemed them to be extremely important for predicting the fine amount, and we wanted to understand which of the articles influenced fine amounts the most. By extracting each article seperately from the dataset, we were able to describe them as binary variables ("1" being article violated in this case, "0" otherwise) in order to use them as predictors for our model, which will be explained later on.

4.2.2 Vectorization

Text Vectorization is the process of converting text into numerical representations, in order for text to be processable by the computer. There are multiple ways to implement it, with some being more advanced that others and having clear advantages.
Our goal with using text vectorization is to create word embeddings which indicate the most important words in each summary in order to use them as features, and also to be able to "feed" these embeddings as input in our different models and algorithms.

The first method we used is TF-IDF vectorization. TF-IDF, formally known as Term Frequency - Inverse Document Frequency, is one of the simplest and efficient methods there are to embed our text. TF-IDF works by counting word frequency in each document, and multiplying by the fraction of the number of documents divided by the word occurance in the specific document. With this procedure, TF-IDF gives low scores on words with high frequency and high document appearance (like conjunction words) while giving high scores to words that appear frequently in one document and fewer times on the rest of the documents.
In our case, we manually picked a few tokens based on their IDF values, which seemed to be influential for our purpose, (e.g. "child", "university" etc.) which is to use them as input for our Factor Analysis and then reduce the dimensionality.
Lastly , we attempted to use bigrams instead of words them as input for our factor analysis. However, since our data are structured summaries in the same exact way and extremely homogenous, factor analysis grouped the bigrams sequentially in those sentences, and thus did not provide any meaningful results.

Bert Embeddings was the second method we used to vectorize text. Bert embeddings are vector embeddings like Word2Vec but they are more advanced in the fact that they take semantics and other intricacies into account, for example polysemy. They are context-informed embeddings that capture other forms of information that result in more accurate feature representations, which in turn results in better model performance.
For our task we decided to use two pretrained BERT models and compare their results, *base Bert* (Devlin et al., 2018a) and a Bert model called *Legal-Bert* (Chalkidis et al., 2020).

4.2.3 Dimensionality Reduction

Dimensionality Reduction was used in order to reduce the number of variables we selected using the IDF scores. Our proposed method of dimensionality reduction is factor analysis, which is a statistical technique that reduces a set of variables by extracting all their commonalities

into a smaller number of factors. When observing large numbers of variables, some common patterns emerge, which are known as factors. These serve as an index of all the variables involved and can be utilized for later analysis, in our case as predictors for our models.
By tinkering with different number of selected features and number of factors we reached satisfying results. We found that using 3 factors was best to describe our features, and we named the factors *"Family","University/Student","Sensitive/Personal/Employee"*. For example for *Family* there were strong factor loadings for words "parent" and "child".

In order to test if our input feature data was suitable for Factor Analysis, we used use two methods: Kaiser-Meyer-Olkin criterion (KMO) (Kaiser, 1970) and Bartlett's test. KMO is a test that measures sampling adequacy for each variable in the model and for the complete model. Bartlett's Test of sphericity tests the null hypothesis that the true correlation matrix is an identity matrix. For KMO we got an acceptable but mediocre result of 0.62 and Bartlett's test was significant ($\chi^2(45) = 117.93$, $p = .00$).

4.2.4 Keyword Extraction

Keyword extraction is an automated method of extracting the most relevant words and phrases from unstructured text input. It is a text analysis method frequently used to automatically extract keywords from emails or long texts in order to see the most important words in that specific document.
There are multiple keyword extraction algorithms, like KeyBert (Grootendorst, 2020) which uses Bert embeddings and cosine similarity of words vectors to a document vector to detect the most important keywords, and statistical methods like YAKE (Campos et al., 2020) which follows an unsupervised approach that builds upon local text statistical features extracted from single documents.

We used the following approach: After extracting all the keywords for all summaries , we kept only the 20 most frequent keywords for the whole corpus ,used them as features and then as input for our factor analysis. However promising ,this method did not yield better results than our original TF-IDF vectorization selection , since by manually looking through the factors we noticed them being worse in terms of interpretability. On a side remark, YAKE worked better than KeyBert, as indicated in Piskorski et al. (2021) even though it only relies on statistical methods and does not take semantics into account.

4.2.5 Topic Modeling

Topic modeling is an unsupervised machine learning technique that's capable of scanning a set of documents, detecting word and phrase patterns within them, and automatically clustering word groups and similar expressions that best characterize a set of documents. In our project , we wish to use it in order to create topics that can be used as predictors for our models ,after extracting the relevant topics from our summary corpus. We tried 4 different methods of topic modeling to see which one yielded the best results. Since evaluating topic modeling results is not easy and is more of an intuitive task (Egger and Yu, 2022), we decided to evaluate them on the following principles:

- Are the identified topics understandable and distinquishable?
- Are the topics coherent?
- Is the topic model helpful for it's purpose i.e. being used as a predictor for fine amounts?

Statistical Topic Modeling

Latent Dirichlet Allocation LDA is one of the most popular statistical topic modeling methods, and thus we decided to implement it first. Since it requires to specify the number of topics to cluster at, we tinkered with the desired number of topic and evaluated their outputs. We found that 3 topics worked the best for this method. However even after filtering extremes, the results were poor in interpretability, since the word distributions of each topic had similar words and were not distinquishable enough to be given a topic name easily.
We create 2 Gensim LDA models (Rehurek and Sojka, 2010) , one which was fitted using the Bag-of-Words corpus and and other with TF-IDF vectors. We found the former proved to work better with our data, however in the scheme of things results were not satisfying.

Non-Negative Matrix Factorization NMF is another popular method for topic modeling that we used. It became popular because of its ability to automatically extract sparse and easily interpretable factors. While LDA is a probabilistic model, NMF is a matrix factorization and multivariate analysis technique. We fit the NMF sklearn model using the TF-IDF matrix and specified the number of topics to be 4. The reason we used 4 and not 3 compared to LDA was because from the output we noticed an extra topic which seemed interesting which we named as *Publishing* as it contained words related to publishing personal data.
All in all, NMF performed better for us as it returned more intuitive and distinquishable topics than LDA.

State-of-the-Art Topic Modeling With the latest technological advances in machine learning , we can leverage transformers and pre-trained word embeddings for topic modeling applications. The 2 methods we will test are **BerTopic** (Grootendorst, 2022) and **Top2Vec** (Angelov, 2020). Theoretically, those methods should yield better results that the traditional ones since they take semantics of words into account. Before we analyze them it should be noted that preprocessing is not required for these methods, as it is built in.

BerTopic For BerTopic , we utilized various pretrained embeddings. However, the results were disappointing, as the topics were filled with stopwords like "of" ,"the" etc. and in general did not make sense. We believe that some technical problem was behind that, but despite our tinkering with different embeddings or code, we couldn't find a way to fix it.

Top2Vec was the last method we used and the most successful one. It creates jointly embedded documents and word vectors utilizing Sentence Transformers and visualizes the vectors in one vector space. From that we can deduct the similarity of the word vectors to the document vectors. By reducing the dimension using **UMAP** (McInnes et al., 2018), focusing on dense areas using **HDBSCAN** (Malzer and Baum, 2020) and then calculating the centroid of topic vectors for each dense area, we can output the words that are most similar to each topic, while at the same time excluding outliers . Other than that, Top2Vec has other core differences to LDA and NMF; For instance, the number of topics is not predifined, Top2Vec automatically outputs as many topics as it deems necessary. Afterwards we can hierarchically reduce the number of topics to what we want. With regard to our results and dataset, 7 topics proved to be the most distinquishable and interpretable . Those topics we named *University, Without Legal Basis, Family Data, Publishing, Employee Rights Violation, Health Data, Governmental Fines.*
All in all, Top2Vec yielded the most explainable results and distinquishable enough to be used as predictors for our model.11

5 Method

In order to predict fines, two generalized linear modeling approaches are used. First, an OLS with log-link, second a logistic regression with logit-link. For OLS, the fines were log-transformed due to the skewed nature of their distribution. Note that dispersion tests were run to check the validity of log-linking the OLS analysis. It was concluded that no overdispersion is present, thus confirming the validity of a simple log-link. For the binary logistic regression fine variable was categorized into "high" and "low" using a median cut ($\mu_{1/2}$ = 15000€) for lack of validated cutoff for this specific use case. A third quantitative approach is a decision tree. The aim here is to develop a simple set of rules to predict fine classes with good accuracy. The tree is trained on the same predictors as the other two models. Overfitting is controlled by setting a minimum of ten samples per leaf. Thus, outlier influence is greatly reduced.

For feature selection, the same approach is applied to all three. First, the model is fitted using all available predictors. The predictor with the largest p-value is removed from the model and it is refitted. For each step, both the Akaike information criterion (AIC) as well as the Bayesian information criterion (BIC) are evaluated. The model that minimizes these criteria is termed the best. In total, 26 predictors are taken into consideration after removing very rare categories (and articles). These consist of Articles, extracted factor scores, topics and regions. After concluding quantitative analyses, a qualitative analysis of highest fines is presented.

6 Results

6.1 OLS

Then, the aforementioned optimization process is started. In Figure 7, one can see the development of information criteria over optimization steps. Both criteria are minimized after 24 steps. The model retained eight predictors. After inspection of the leverage values

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T, \tag{1}$$

where $\mathbf{X}$ is the model matrix, one observation was removed from the analysis that had a fine of 3700000€.

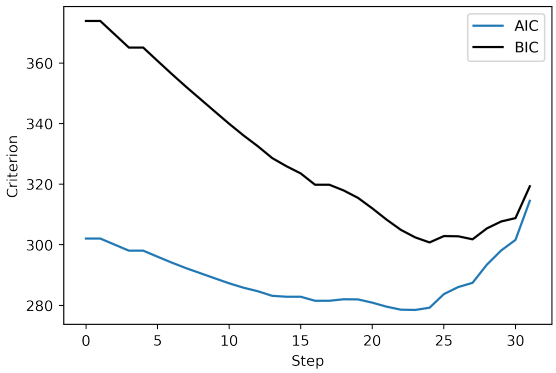

Figure 7: Information criteria by predictor removal steps

The model explains a significant proportion of variance in the fine variable with $F(8, 71) = 13.81$ and adjusted $R^2 = .57$. Further, the model achieves a MAE of .89 logarithmic units. Exponentiating the dependent variable and predictions, the mean real-scale error is 63515.37€. Note though that this is highly inflated by the non-linear nature of the model specification.

All retained variables yield significant coefficients at $\alpha = .05$. Regarding the violated articles, articles 9,13 and 32 are significant predictors of fine amount. All three lead to higher fine amounts if violated (compared to any other article). Between the three articles, there is no statistically significant difference in coefficients, although the largest coefficient in this sample belongs to article 13. Violating it instead of any other article (except 9 and 32), while holding all other variables constant, leads to a 6.77-fold increase in fine amount. Regarding geographical predictors, three were retained in the final model. Their effects are compared to organizations that are located either in central or eastern Europe. Northern Europe yields the largest coefficient. Organizations located in northern Europe receive fines roughly 8 times as large as organizations in central or eastern Europe. Northern and southern Europe are also associated with larger fines, albeit in

Table 2: OLS coefficient table for logarithmic fine amount prediction

Group	Variable	e^B	B	SE	t	p
	Intercept	859.97	6.76	.44	15.54	.00
	Article 9	4.28	1.45	.43	3.42	.00
Articles	Article 13	6.77	1.91	.47	4.12	.00
	Article 32	4.48	1.49	.34	4.38	.00
	Northern Europe	8.18	2.10	.46	4.58	.00
Regions	Southern Europe	3.63	1.29	.46	2.79	.01
	Western Europe	7.58	2.03	.62	3.28	.00
Topics	Government	5.13	1.63	.53	3.07	.00
	University/Student	3.33	1.20	.36	3.32	.00

Note. adj. $R^2 = .57$, $F(8, 71) = 13.81$, $p = .00$

less extreme fashion. Finally, regarding the extracted topics, two were retained after model optimization. Government related cases are fines the highest, while any case that includes student data is also fined more than others (any topic that is neither government-related nor student-related).

Inference was checked for validity using conventional measures. The residuals were visually inspected for heteroskedasticity and tested for normality using the Shapiro-Wilk test ($W = 0.99$, $pvalue = 0.50$). Both assumptions hold. Therefore, nonrobust standard errors lead to correct inference and the model does not show signs of misspecification.

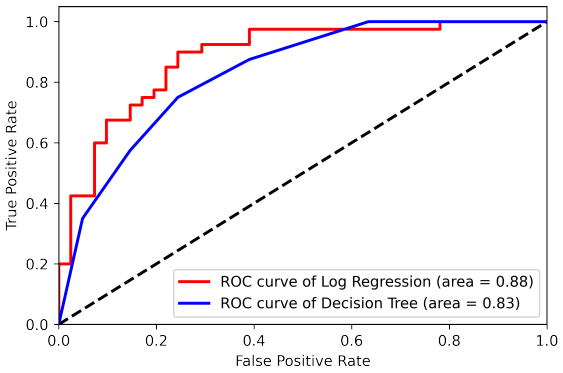

Figure 8: Receiver-Operator characteristics curves for both classifiers

6.2 Logistic Regression

Logistic regression is presented as an alternative approach to log-linked OLS. The coefficients for the BIC-optimized model can be seen in Table 3.

Overall, the model has good explanatory power, as indicated by a Cox-Snell pseudo R^2 of .49.

All six retained predictors have significant associations with fine category. The largest effect belongs to article 13, which increases the odds ratio $\frac{P(\text{"high"})}{P(\text{"low"})}$ by a factor of 27, should it be violated. Violating article 9 is also associated with higher probability of yielding a "high" fine, as is article 32. The remaining four predictors all have negative coefficients. The more a fine has to do with employee data in general, the lower the probability of yielding a large fine.

Also, fines that regard "Family data" are less likely to be classified as high than those that fit any other topic the most. The classification model reaches a train set accuracy of 82.72%. Also, with an ROC area under curve of .88 (see Figure 8), the logistic regression performs quite well. In Table 4, common performance metrics for binary classifiers are summarized. Here, it is apparent that the logistic regression has very good recall with $R = .88$, thus detecting most "high" fines as such.

6.3 Decision Tree

In addition to the two generalized linear models, a decision tree model is presented. It operates on the same 32 predictors that the other models "chose" from. In order to prevent the common issues with overfitting when training decision tree models, only leaves with at least 10 samples were retained. In Figure 9, one can see the derived decision process. As also seen in the regression models, violating article 32 is the most influential predictor in categorizing fines. If an organization violates article 32, the only way for it not to be fined high is to not be University or student-related and to not violate article 5. On the other hand, if article 32 was not violated, the only way to predict a high fine is if article 9 was violated.

Overall, the decision tree reaches good performance with training accuracy of 75.31% (note that this tree is pruned substantially already). Using the proportion of "high" fines on each leaf, probability predictions can be calculated from the decision tree. ROC analysis yields an area under the curve of .83 (see Figure 8), which is less than for the logistic regression but still relatively good. Remarkably, only looking at the first decision step (Article 32 violation), one can categorize a case as high or low with over 70% accuracy ($\frac{28+29}{12+12} = .70$) which further underlines the predictive power of knowing whether article 32 was violated.

Table 4: Performance metrics for both classifiers

Model	Accuracy	Precision	Recall	AUC
Log. Reg.	.83	.80	.88	.88
Dec. Tree	.75	.75	.75	.83

Table 3: Logistic regression coefficient table for binary fine classes

Group	Variable	e^B	B	SE	z	p
	Intercept	.06	-2.76	.74	-3.72	.00
	Article 9	20.87	3.04	1.07	2.85	.00
Articles	Article 13	27.13	3.30	1.28	2.58	.01
	Article 32	16.83	2.82	.82	3.46	.00
Factors	Employee data	.46	-.78	.35	-2.23	.03
Region	Northern Europe	12.41	2.52	.85	2.97	.00
Topics	Family data	.07	-2.65	.95	-2.787	.01

Note. Cox-Snell pseudo-R^2: $R^2_{CS} = .49$

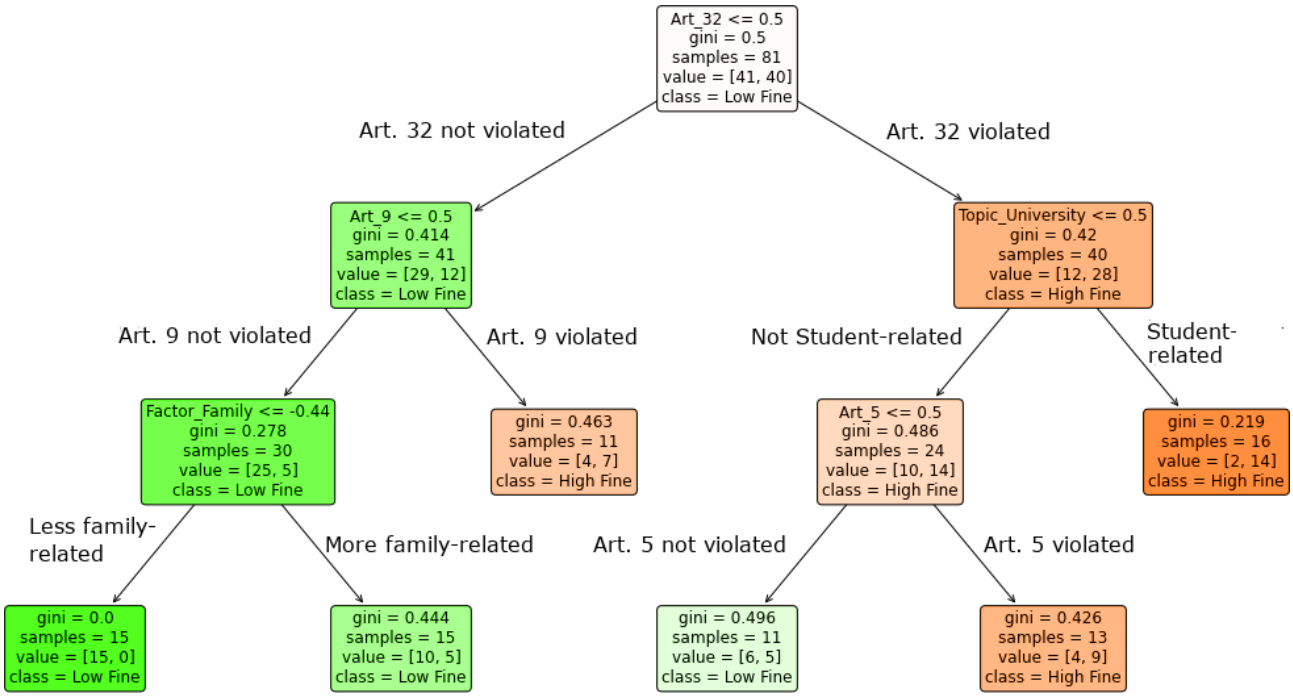

Figure 9: Trained decision tree classifier with decision labels.
Note. Colors represent a gradient based on the proportions of high/low fines on each leaf. The "value" field represents the distribution of low and high fines, so [15, 0] means that there are 15 low fines on that leaf and 0 high ones.

6.4 Qualitative Keyphrase Analysis

In addition to the quantitative analyses, a qualitative approach based on keywords is examined. Using results from YAKE, the fines above the 75th percentile are analyzed. Running the YAKE keyword/keyphrase extractor on single summaries, most keywords are locations and DPA names. This is likely due to the short nature of the summaries, which does not allow for more detailed keyword extraction. It is concluded that this approach does not yield meaningful results for the purpose of this project.

7 Discussion

7.1 Interpretation

Three quantitative approaches were presented, their results must now be compared and interpreted. Articles 9, 13 and 32 were present in both regression models, yielding positive coefficients in both cases (an overview of Article coefficients can be seen in Figure 10). The decision tree did not include article 13, but yielded article 32 as most important predictor and also included Article 9. Article 9 prescribes the special treatment of particularly sensitive data, such as political beliefs. Violating this leads to higher fines. Article 13 prescribes the proper information of data subjects about treatment and collection of their data. Thus, organizations that collect data without properly informing people about their actions leads to higher fines. Article 32 prescribes security measures that must be taken in order to preserve and protect collected data. Thus, in addition to not informing people about data collection, not implementing sufficient measures to protect the collected data leads to higher fines. This is also possibly the most important factor. Violation of Article 32 lead to higher fines in all analyses, underlining the importance of implementing adequate security measures.

In the OLS approach, regionality had significant influences on fine amount, while it did not influence the probability of yielding a high fine, neither according to logistic

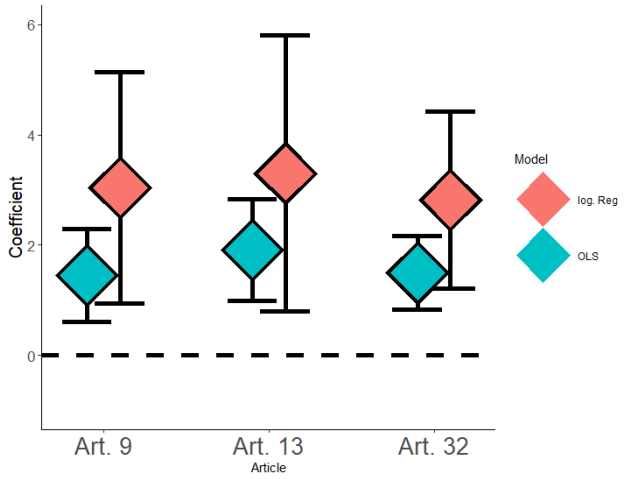

Figure 10: Regression coefficients for both models by article. Error bars indicate 95% CIs.

regression, nor the decision tree. Therefore, it appears as if the country that issues the fine is correlated with the fine amount in general, while the amount of extreme fines does not relate to the region the fine was issued in.

Further, OLS regression yielded positive significant effects for fines relating to government bodies as well as fines relating to student data. Thus, violating GDPR with regard to government-related data as well as student data leads to higher fines. Especially student data is interesting for the purpose of this study, indicating that student data ought to be treated with extra care in order to avoid large fines. That said, neither of these predictors are part of the final binary logistic regression model. Still, logistic regression yields a negative effect for the "University" factor. The decision tree included the "University/Student data"-topic as a fine-increasing factor. These two seem to contradict each other, unless one takes into account that the topic has a larger focus on student data as opposed to general University involvement. Thus, it seems that Universities can generally expect lower fines than other institutions, but increase the risk of receiving a high fine if the violation has to do with student data.

Employee data, family data and the act of publishing the data are related to the high/low classification of fines though. Any GDPR violation regarding any of these three topics is associated with lower probability of yielding a high fine. Especially employee rights do not seem to be as crucial as others. Also, interesting when taking the article inference from earlier into account, violations regarding the publishing of data yield less high fines than other violations. Thus, it seems like informing about collecting data, as well as protecting it once collected, is more important than making sure one has the right to publish it. This is because publishing without proper justification is related to lower fines.

7.2 Limitations

In terms of limitations for this study, one apparent one is the sample size, which, after removing outliers turned out to 80. Post-hoc power calculations reveal that, in order to achive 95% power with the given OLS results, a sample size of 27 would have been needed. Thus, sample size itself is not an issue for the OLS model, and very good statistical power was achieved. For the logistic regression, 35 observation would have been required to achieve the power goal.

Still, the sample size restricted feature extraction and selection in that some attributes were simply very sparse in the dataset. For example, some articles could not be used for analyses since they were referenced in only a few summaries.

Another limitation is that the sample is (by design) not representative of the population of organizations that are subject to GDPR surveillance. Therefore, inference made during this study really only applies to public institutions and Universities.

8 Future Work

There are some ways that future research on this topic can have more promising results and be more informative. The main points are:

- In it's current state, the GDPR fines on public institutions and universities are not that many. In the future more fines will be issued , and thus data supplementation is recommended since more feature predictors might return better results.
- If data are enough, research that includes only universities might be possible. That will lead to a more homogeneous dataset, and will allow enrichment and expansion of it with relevant data that can only be found in universities such that student count, ranking etc. A possible connection between for example, between the reputation or size of university would be very interesting.
- Despite the efficiency of summaries, information extraction would be greatly improved if the detailed court reports of each fine are used. However, as mentioned before, scraping difficulties along with the 24 official languages that the reports consist of can prove to be a problem. Potential improvements could be done using Google Translate API to translate the documents(with severe limitations, since documents are domain specific Law documents) or state of the art tools like Multilingual Bert Embeddings (Devlin et al., 2018b).

9 Conclusion

In conclusion, both research questions could be answered. Regarding the first, some interesting features were extracted using common information extraction practices. These are important factors in predicting fine amounts. For example, overarching topics were extracted that predict fines with good precision and accuracy. The involvement of student data is generally associated with

higher fines while family and employee-related data violations incur lower fines. The most important predictor was the absence of adequate security measures to secure collected data. Other important ones are the processing of particularly sensitive data such as political beliefs as well as providing fully informed consent to data subjects before collecting data.

To answer the second research question, three quantitative modeling approaches were presented and discussed. Overall, both the prediction of fine amount and the classification into low and high fines worked well, yielding good performance metrics. These multiple approaches also allow for flexible modeling choices depending on the user's goal. The regression approaches might yield better performance, while the decision tree approach is more easily applied and can work as a simple rule set for decision makers to use as guideline.

References

Angelov, D. (2020). Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*.

Campos, R., Mangaravite, V., Pasquali, A., Jorge, A., Nunes, C., and Jatowt, A. (2020). Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289.

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., and Androutsopoulos, I. (2020). LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018a). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018b). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Egger, R. and Yu, J. (2022). A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. *Frontiers in Sociology*, 7.

Grootendorst, M. (2020). Keybert: Minimal keyword extraction with bert.

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Honnibal, M. and Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Kaiser, H. F. (1970). A second generation little jiffy.

Malzer, C. and Baum, M. (2020). A hybrid approach to hierarchical density-based cluster selection. In *2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, pages 223–228. IEEE.

McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Piskorski, J., Stefanovitch, N., Jacquet, G., and Podavini, A. (2021). Exploring linguistically-lightweight keyword extraction techniques for indexing news articles in a multilingual set-up. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, pages 35–44.

Rehurek, R. and Sojka, P. (2010). Software framework for topic modelling with large corpora. In *In Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks*. Citeseer.

Ruohonen, J. and Hjerppe, K. (2020). Predicting the amount of gdpr fines. *arXiv preprint arXiv:2003.05151*.

Appendix

Name	Country	Date	Fine	GDPR Article	...	Summary
Università Telematica Internazionale Uninettuno	Italy	16/12/2021	1000	Art. 5 (1) c)	...	A professor had filed a complaint with....
Sahlgrenska University Hospital	Sweden	3/12/2020	341300	Art. 5 (1) f), Art. 5 (2), Art. 32 (1), Art. 32 (2)	...	The Swedish DPA fined ... for failing to...

Table 5: Small part of the dataset

Figure 11: WordCloud Visualization for Topic *"Health Data"*