D22 – Algorithm for finding outliers

Project funded by the European Union's Next Generation EU through the Slovak Republic's Recovery and Resilience Plan under project no. 09I05-03-V02-00049.











INTRODUCTION

The project **"Automation of Legal Text Analysis based on Machine** Learning" (hereinafter "ALTAML") represents a key initiative aimed at integrating innovative approaches in the field of data processing and subsequent analysis, with a focus on the field of law, which also includes information and communication technology law. The aim of this project is to develop and validate effective methods for the automated analysis of legal texts using machine learning techniques. This involves, in particular, the development of tools that facilitate the processing and analysis of large volumes of data in the form of various legal documents, as well as the extraction of relevant information (attributes) from such documents, including the identification of key terms, references to legislation and other attributes.

The aim of the ALTAML project is therefore to contribute to a more efficient approach to legal information and to the acceleration of legal processes, thereby ensuring a higher degree of legal certainty and improving the accessibility of legal texts, the results of their analysis and relevant legal information for both professionals and the general public.

As part of work package KPB4**, D22 – Algorithm for finding outliers** – was developed. This document presents the proposed approach to identifying such texts using machine learning methods, specifically by utilising an Isolation Forest model based on unsupervised learning and the subsequent interpretation of results using the SHAP method.

In the context of this task, outlier legal texts are identified through the detection of so-called outliers (remote elements of the dataset). An outlier is an observation in the data that differs significantly from the others and does not correspond to the usual pattern of behaviour in the given dataset. Such cases may indicate rare, non-standard or otherwise significant legal decisions.

By using the SHAP method, it is possible to explain why specific texts were flagged as atypical/anomalous (Figure 1), whilst analysing the influence of individual attributes and their values on the model's result for a given text. At the same time, it is also possible to obtain a global view of the model's behaviour, for example through beeswarm (Figure 3) visualisations, which display the significance and distribution of the influence of individual attributes across the entire dataset.

The proposed method was experimentally validated on five datasets of court decisions that were manually annotated by lawyers. Each dataset contains court decisions from a selected area of law, such as 'contractual penalty' or 'unfair competition', which made it possible to verify the robustness of the proposed approach across different types of legal areas. The method consists of three main steps: preparation and transformation of the input data into a numerical representation, detection of atypical texts using the Isolation Forest model, and subsequent interpretation of the results using the SHAP method.

In this way, the method covers the entire process from the processing of legal attributes, through the identification of atypical texts, to the explanation of the reasons why specific texts were flagged as atypical by the model.

1. Data preparation and attribute processing

Before applying the model, the input data was transformed into a numerical representation suitable for processing by the Isolation Forest model. Legal documents were represented by structured attributes obtained from annotated datasets.

The pre-processing involved:

- processing of numerical attributes, including imputing missing values and creating missing value indicators,
- normalisation of text attributes (conversion to lower case, removal of diacritics),
- transforming binary categories (e.g. 'yes' / 'no') into numerical form,
- assigning the value 'unknown' to missing categories,
- the transformation of categorical attributes using one-hot encoding,
- removal of attributes with constant values.

The result of this process was a numerical dataset, which was subsequently used as input for the outlier detection model.

2. Outlier detection model

The Isolation Forest model, which is a machine learning method, was used to detect atypical legal decisions.

This model operates on the principle of random partitioning of the data space. Decisions that can be isolated by a smaller number of partitions based on their attributes are considered atypical. The reason is that such decisions have attribute values that differ significantly from the majority of others – for example, they contain unusual combinations of attributes or extreme values. In random data partitioning, they are therefore separated more quickly from the rest of the dataset, leading to their classification as atypical. Conversely, decisions that share similar characteristics with the majority of the data require multiple partitions to isolate them and are therefore considered typical.

This approach is suitable for situations where explicit outlier annotations are not available.

The model was applied to individual datasets separately, and for each decision the following were calculated:

- **an anomaly score**, which expresses the degree to which a decision differs from others based on its attributes. This score is derived from how quickly a given decision is isolated within the model – decisions that are isolated more quickly achieve lower scores and are considered more atypical,
- an **indicator showing whether a decision is marked as an outlier**, generated by comparing the score with the model's internal threshold (a given contamination parameter, in our case 10%),
- **a ranking of decisions by degree of deviation**, which allows decisions to be ordered from the most atypical to the most typical.

3. Interpretation of results using SHAP

The SHAP (SHapley Additive exPlanations) method was used to interpret the model results, which allows the model's decisions to be explained at the level of individual attributes.

Using SHAP, it is possible to:

- explain why **a specific court decision** was flagged as anomalous,
- determine which attributes and their specific values contributed most to this decision,
- analyse the direction of influence of individual attributes (whether they contribute to the atypicality or, conversely, to the 'typicality' of the case).

For each document, a detailed explanation can be displayed in the form of so-called waterfall plots, which illustrate the contribution of individual attributes to the final score.

Figure Figure 1 shows a so-called waterfall chart for a specific decision. This chart illustrates how individual attributes and their values progressively influence the model's final score.

The graph is based on the model's baseline value, to which the contributions of individual attributes are successively added or subtracted. Each row represents a single attribute along with its specific value for the given court decision.

In the Isolation Forest model, it is important to note that a lower final score indicates a higher degree of anomaly. Therefore, attributes with a negative contribution lower the model's resulting score for a given court decision, thereby shifting it towards atypical cases, whilst attributes with a positive contribution increase this score and contribute to its classification among more common decisions.

The resulting value of the graph f(x) represents the final model score for a given decision. From the graph, it is thus possible to identify which attributes contributed most to the decision being assessed as atypical, and which attributes, conversely, mitigated this effect.

It is also possible to interpret specific attributes in detail from the graph. For example:

- The attribute "**Výmera (v mesiacoch) = 204"** (eng. Duration (in months)) has a significant negative contribution (−1.04), which means that its value significantly reduces the model's final score for the given court decision, thereby shifting it towards atypical cases. Unlike binary attributes, this is a numerical attribute, and the specific value (204 months) represents an extreme value within the dataset, which leads to its significant influence on the model's result. It is also the attribute with the greatest influence on the model's result for this decision.

These values show that the combination of specific attributes and their values can significantly influence the model's output and lead to a decision being flagged as anomalous.

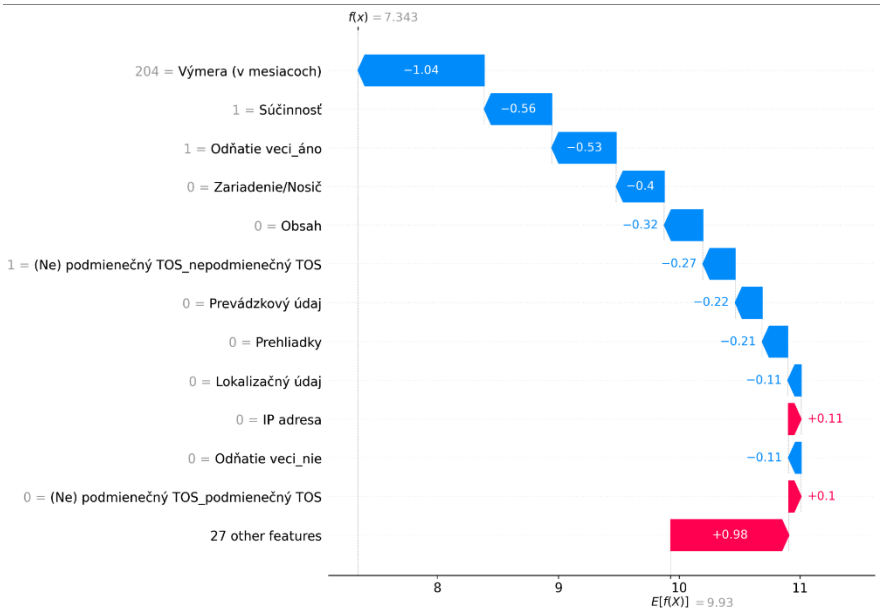


Figure 1 - Explanation of the anomalous nature of a specific court decision using a waterfall chart

At the global level, the following were used:

- an attribute importance graph (Figure 2),
- a beeswarm visualisation showing the distribution of attribute influence across the entire dataset (Figure 3),
- analysis of interactions between attributes using SHAP interaction values (Figure 4).

The following figureFigure 2 illustrates the global importance of individual attributes using mean absolute SHAP values.

This type of visualisation provides an overview of which attributes have the greatest influence on the model's decision-making across the entire dataset.

Each row represents a single attribute, and the value on the X-axis (SHAP value) expresses the average absolute influence of that attribute on the model's final score. The higher the value, the greater the influence of that attribute on the model's decision-making, regardless of whether this influence increases or decreases the anomaly score.

It is important to emphasise that this graph does not show the direction of the influence (i.e. whether the attribute increases or decreases the anomaly), but only its magnitude.

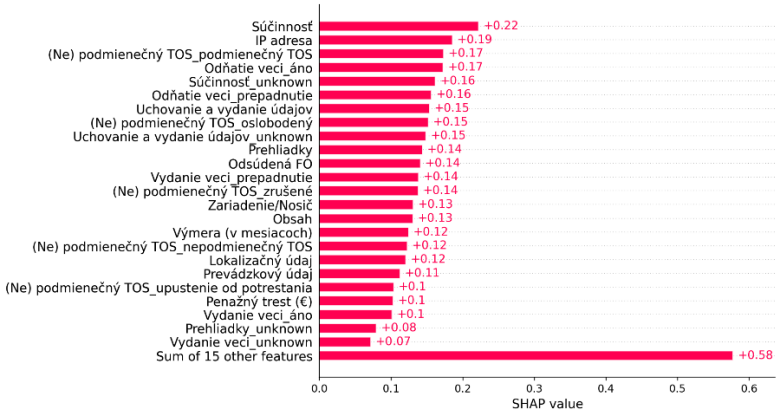


Figure 2 - Global importance of attributes expressed using the average absolute SHAP value, showing the attributes with the greatest influence on the model's decision-making.

The following figureFigure 3 shows a so-called beeswarm visualisation of SHAP values, which provides a more detailed view of the influence of individual attributes across the entire dataset.

Each row represents a single attribute, and each point corresponds to a specific court ruling. Each row contains as many points as there are rulings in the dataset. The position of a point on the X-axis represents the SHAP value, i.e. the influence of that attribute on the model's final score for a specific ruling. In the case of the Isolation Forest model, values shifted further to the left reduce the final score (i.e. contribute to outlier behaviour), whilst values closer to zero or further to the right contribute to more common decisions.

The colour of a point represents the value of an attribute – blue corresponds to lower attribute values, whilst red represents higher values. By combining the position and colour of the points, it is possible to identify how specific attribute values influence the model's output.

The graph also highlights differences between types of attributes. Some attributes have only two values (e.g. binary yes/no attributes), which is reflected in the fact that the points are shown in only two colours – blue and red. Conversely, attributes with continuous or numerical values (e.g. 'Duration (in months)') create a colour spectrum ranging from blue to red, reflecting the wider range of possible values and their gradual influence on the model's final score.

Specific attributes can also be interpreted from the graph. For example, for the **'Súčinnosť'** (eng. cooperation with the court) attribute, it can be seen that decisions with a value of 1 (yes) are systematically shifted to the left , thus contributing to the anomaly. Conversely, for the **'Prehliadky'** (eng. inspections) attribute, the opposite effect applies – decisions with a value of 0 tend to contribute to the anomaly, suggesting that the presence of this attribute is rather typical in the dataset.

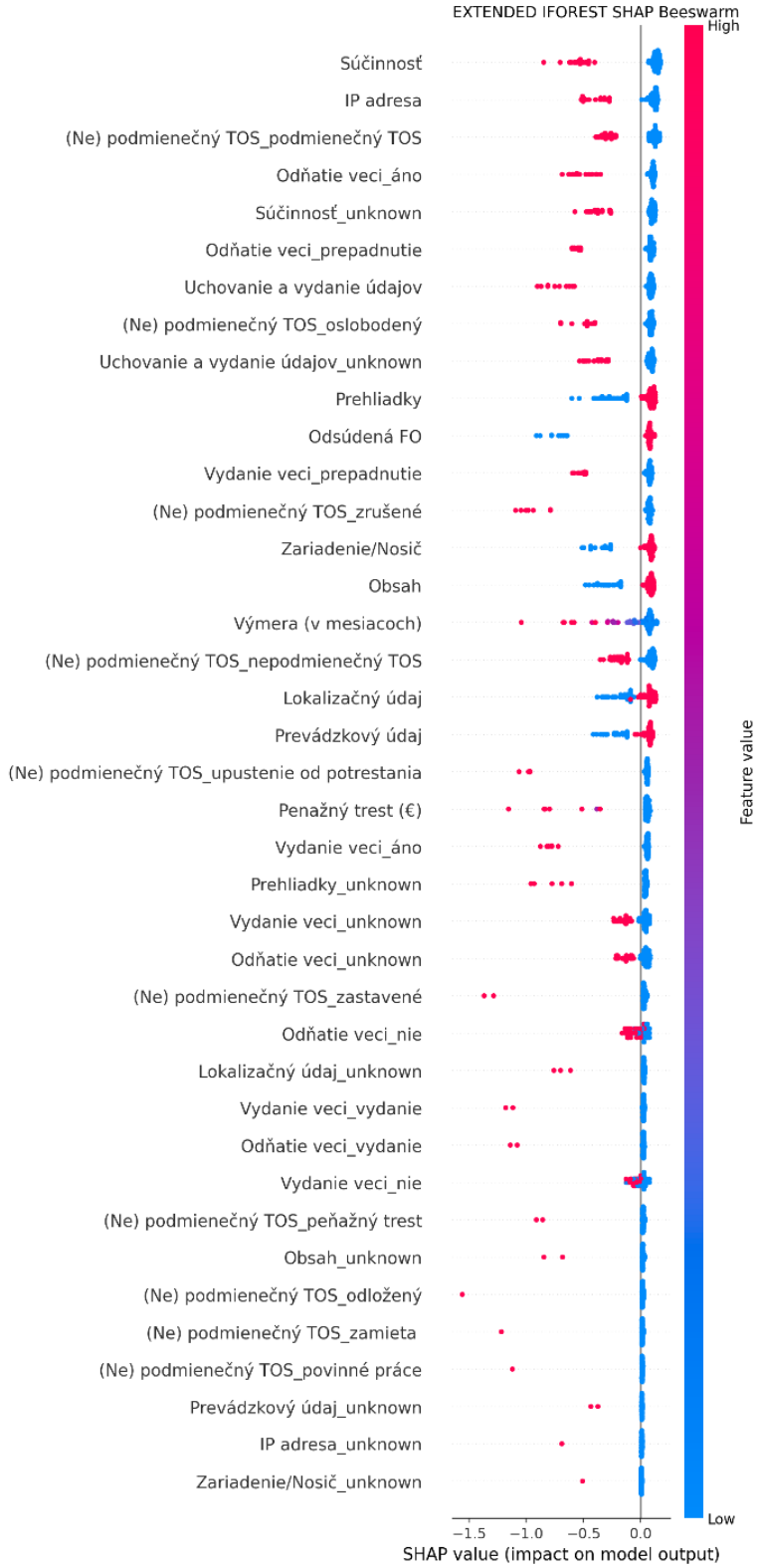


Figure 3 - Beeswarm visualisation of SHAP values showing the influence of attributes on the model's final score across all decisions.

The following figureFigure 4 shows a heatmap of interactions between attributes based on SHAP values. This visualisation captures the extent to which pairs of attributes jointly influence the model's final score.

Each row and column represents a single attribute, and the value in a cell expresses the average absolute strength of the interaction between that pair of attributes across the entire dataset. Using the absolute value ensures that the overall strength of the interaction is taken into account, regardless of its direction (i.e. whether the interaction increases or decreases the model's final score). Without using absolute values, positive and negative interactions could cancel each other out, thereby obscuring truly significant relationships between attributes. The higher the value (lighter colour), the stronger the combined influence of these attributes on the model's decision-making.

The values on the diagonal represent the individual (main) influence of each attribute, whilst the values off the diagonal capture the actual interactions between different attributes.

The graph shows that most interactions between attributes are relatively weak, with individual attributes playing a dominant role. Stronger interactions (lighter areas off the diagonal) indicate pairs of attributes whose combination has a more significant influence on the model's outcome than either attribute alone.

This analysis allows us to identify not only important attributes but also their interrelationships, which may be key to understanding the model's behaviour.

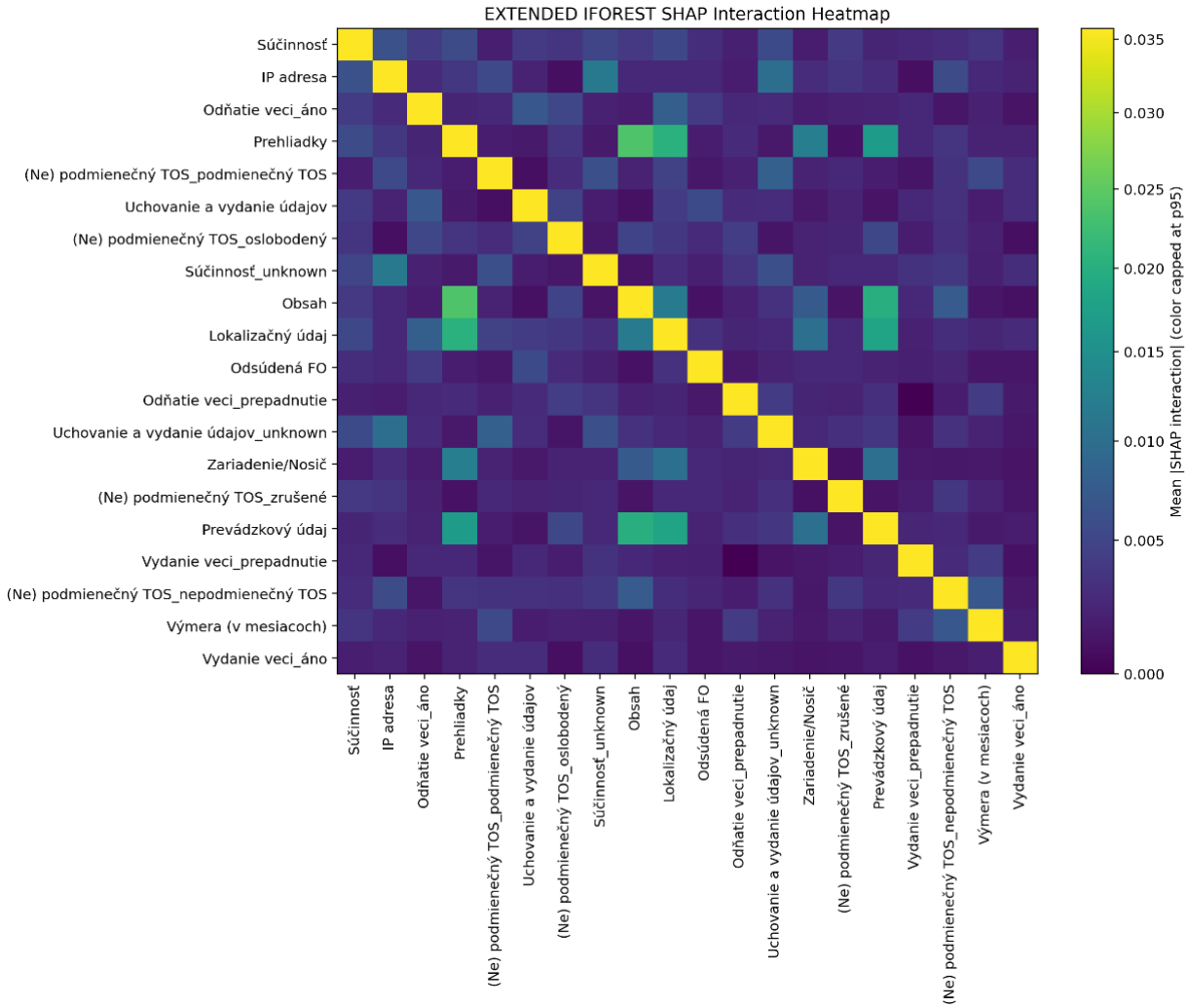


Figure 4 - SHAP interaction heatmap showing the strength of the mutual influence between pairs of attributes on the model's final score.

These approaches enable a better understanding of the model's behaviour and the identification of the main factors leading to documents being labelled as anomalous, whilst also providing a transparent and interpretable explanation of the model's decisions at both the individual decision level and across the entire dataset.

4. The proposed method and its outputs

The result of the solution is a method for identifying and interpreting atypical legal decisions, which has been applied to five datasets.

For each dataset, a separate output folder is created (e.g. *ALTAML_{dataset}_anomaly_detection/iforest*), where *{dataset}* represents the name of the specific dataset, containing the complete detection results and their interpretations.

The proposed approach is not merely a direct application of the Isolation Forest model, but also involves data processing, the transformation of legal attributes into a numerical representation, and the integration of explainability methods (SHAP), which enable the results to be interpreted in the context of legal decisions.

The specific outputs are:

- **{dataset}_rows_with_iforest_scores.csv**
  contains the original decisions supplemented with anomaly scores, the identification of atypical decisions (outliers) and a ranking by degree of deviation.
- **{dataset}_iforest_shap_values.csv**
  contains SHAP values for individual attributes and decisions, i.e. local numerical explanations of the model.
- **{dataset}_iforest_global_shap_importance.csv**
  contains the global importance of attributes calculated as the average absolute SHAP value.
- **{dataset}_iforest_shap_bar_mean_abs.png**
  visualises the most important attributes according to the average absolute SHAP value (Figure 2).
- **{dataset}_iforest_shap_beeswarm.png**
  shows the global behaviour of the model and the direction of the attributes' influence on decisions (Figure 3).
- **waterfalls/**
  contains a SHAP waterfall plot for each decision, i.e. a detailed local explanation of a specific case (Figure 1).
- **{dataset}_iforest_interaction_pairs.csv**
  contains pairs of attributes sorted by the strength of their SHAP interaction.
- **{dataset}_iforest_interaction_heatmap.png**
  visualises the strongest interactions between attributes (Figure 4).
- **{dataset}_iforest_top_pair_per_row.csv**
  contains the most significant interaction pair for the most anomalous decisions.

- **detector_summary.csv**
  summarises the results for a specific dataset (e.g. number of decisions, number of attributes, number of marked atypical decisions).

In addition, a common file, **all_datasets_detector_summary.csv**, is generated, which summarises the results across all processed datasets.

The proposed method is also flexible and allows for the analysis of various aspects of legal decisions, including their outcomes (verdicts), which are represented as attributes in the data. This makes it possible to identify atypical decisions not only based on individual attributes, but also based on combinations of them.

5. Source files

The project's source code is managed in the UPJŠ internal GitLab repository. It is available at the following address https://gitlab.science.upjs.sk/altaml/method-for-outlier-detection and also includes an executable implementation of the proposed method across the five datasets used.