Research of Criminal Cases and Identification of their Important Characteristics

Исследование уголовных дел и выявление их важных характеристик

Software Teamproject

Course project of: Morozova Yulia (DSBA), Khugaeva Dana (AMI), Sheredeko Arina (DSBA)
Supervisor: Ilya Munerman (Associate Professor)

Presentation plan

Project's GitHuB

General Idea

This project explores how NLP and ML tools can be applied to analyze large volumes of Russian court decisions under Article 228 in a structured and scalable way.

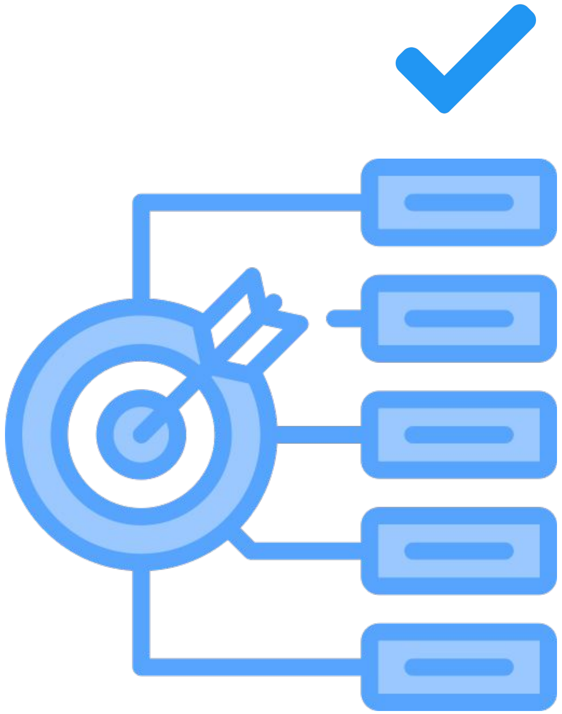
Relevance

Why it can be useful / relevant?
What tasks does it solve / contribute?

- Handle a big amount of legal information, process the data
- Help human workers with the revision of the criminal cases
- Assist students in their learning of the law (use results for AI assistant)
- Develop a model for predicting court decisions
- Contribute to the AI judge creation
- Lack of ML/NLP studies on Russian court texts

Goal

To develop **automated methods** for extracting and analyzing key features from Russian court decisions under **Article 228** using **machine learning and NLP**, in order to **identify patterns** and potential **biases** in sentencing.



Key tasks

- Data Collection & Preprocessing
- Extract features: punishment type/length, drug weight, circumstances, etc.
- Use regular expressions, manual markup, and LLMs
- Visualizations + regression analysis to find patterns

Review of Relevant Literature

Extracting Legal Facts with ML and Embeddings (Luo et al., 2021) [1]

Proposed extraction of legal facts (law articles, penalties, etc.) from English texts using TF-IDF, embeddings, Random Forest and SVMs — though models ignored domain-specific topics like drugs.

Legal Data Extraction in Russian Cases (Ivanov et al., 2022) [2]

Used Natasha and spaCy for rule-based extraction from Russian court texts. Focused on articles and penalties, but lacked modern LLMs and did not target drug-related offenses.

LLMs in Legal NLP: GPT-4 for UK Tribunals (Smith et al., 2023) [3]

Applied GPT-4 to extract structure and predict outcomes in 14,000+ UK tribunal cases. Transformer models performed on par with fine-tuned classifiers.

LLMs for Legal Design & Education (Brown et al., 2022) [4]

Explored generative AI in legal innovation — including chatbots, visual maps and tools for student engagement in legal tech courses.

Data Collection and preparation

Tool used:
A specialized parser “Если быть точным”

Filters:
Type of case: Criminal Code
- Article: 228
- Format: CSV
- Court level: district, regional
- Region: any
- Years: 2015-2024

Restriction:
Manual CAPTCHA — restriction by courts

Size:
Volume: 7.16 GB → 223 MB
Rows: 303 783
Columns: 25

Key columns:
`id`, `entry_date`, `result_date`, `court`, `codex_articles`, `penalty_type`, `result_text`, `region`, `defendants_gender`, `judge`, `case_number`, and etc.

Cleaning:
- Removing empty columns and links (dropna)
- Deleting incorrect rows
- Sampling from the received values

Extracted Characteristics

Main text source:
`result_text` - the full text of the sentence (analyzed to extract key characteristics, including the type of punishment and term).

Terms of punishment:

- `sentence_years`
- `sentence_months`
- `sentence_days`
- `sentence_hours`

- `is_suspended`
- `suspention_years`
- `suspention_months`

Fine:
- `fine_is_fixed`

fixed or shared
- `fine_amount`

the amount of the fine in rubles

Drugs:

- `drug_weight`
- `drug_type`

name of the substance
- `drug_purpose`

sales, storage, transportation, etc.
- `drug_amount`

significant/large/especially large volume

Factors:
- `aggravating_factors`
- `mitigating_factors`

Others:

- `location`

coordinates of the region (lat, lon)
- `result_type`

type of punishment (fine, probation, imprisonment, etc.)

Manual Annotation & "Gold" Dataset

Volume:

150 manually marked-up cases (50 per participant)

Manual Marking of Features:

All key characteristics: punishment type, drug info, circumstances, etc.

Purpose:

- To evaluate model accuracy (comparison against human labels)
- Calculate metrics: precision, recall, F1-score
- Identify which features are predicted accurately and which require improvement

Extracting characteristics from sentencing texts using LLM

Model: DeepPavlov/rubert-base-cased (Russian, cased, 12-layer, 768-hidden, 12-heads, 180M parameters)

Tasks:

- Classify `result_type` (6 punishment types)
- Classify `drug_amount` category (3 categories)
- Extract multiple classes for `mitigating_factors`

Implementation:

- Extract key sentence fragments using RE
- Fine-tune on labeled dataset of 150 cases
- Predict on 10,000 cases from the random sample

	precision	recall	f1-score
исправительные работы	1.00	0.75	0.86
лишение свободы	0.89	1.00	0.94
обязательные работы	1.00	1.00	1.00
ограничение свободы	1.00	0.50	0.67
условное осуждение	0.97	1.00	0.99
штраф	1.00	1.00	1.00
accuracy			0.98
macro avg	0.98	0.88	0.91
weighted avg	0.98	0.98	0.98

	precision	recall	f1-score
значительный	0.89	1.00	0.94
крупный	1.00	0.64	0.78
accuracy			0.91
macro avg	0.94	0.82	0.86
weighted avg	0.92	0.91	0.90

	precision	recall	f1-score
признание вины	0.93	1.00	0.97
раскаяние	0.83	1.00	0.91

for other 19 classes of mitigating_factors metric values are 0.00

	precision	recall	f1-score
micro avg	0.88	0.57	0.69
macro avg	0.08	0.10	0.09
weighted avg	0.51	0.57	0.54
samples avg	0.88	0.62	0.70

the model overfitted because of the imbalance of classes

Performance:

+ `result_type` F1 0.95
+ `drug_amount` F1 up to 0.94 (but no data for “особо крупный”)
+ `mitigating_factors` F1 0.09 (low)

Limitations:

- Small training set
- Imbalanced labels
- Complex and dispersed wording in legal texts

Find numeric characteristics using Regular Expressions

Why regex:

Regular expressions are a simple and effective tool for extracting structured info (e.g., numbers, keywords) from unstructured legal text — without model training. Was used to extract `drug_weight`.

Implementation:

- Python script using `pandas` + `re`
- Patterns searched for values near keywords like “gram”, “g.”, etc.
- Fractional and integer numbers handled
- Context-limited to prevent false matches (e.g., medical data)

Limitations:

- Didn’t convert units (mg/kg → grams)
- Skipped text fragments could miss info
- Complex wording (e.g., “one gram”) reduced accuracy
- No aggregation of multiple doses

Despite its simplicity, the RE approach worked reliably across 150 human-annotated cases.

Results summary

The extracted features **achieved high accuracy** across the board, with most exceeding 0.90.

Key legal attributes such as result_type, is_suspended, and drug_weight were identified with strong reliability, confirming the **effectiveness of our** LLM and regex-based **pipelines**.

Feature	**Accuracy**
result_type	0.9463
sentence_year	0.8800
sentence_month	0.9267
sentence_day	0.9133
sentence_hour	0.9467
is_suspended	0.9600
suspension_years	0.9067
suspension_months	0.9800
is_fixed_fine	0.9600
fine_amount	0.9333
drug_weight	0.9400
drug_amount	0.9060

Combining All Results

Steps:

10 CSVs (0–10k cases) combined using a Python script
Duplicate headers removed
Combined Predictions:

- Regex outputs (e.g., fines, durations)
- LLM/NER predictions

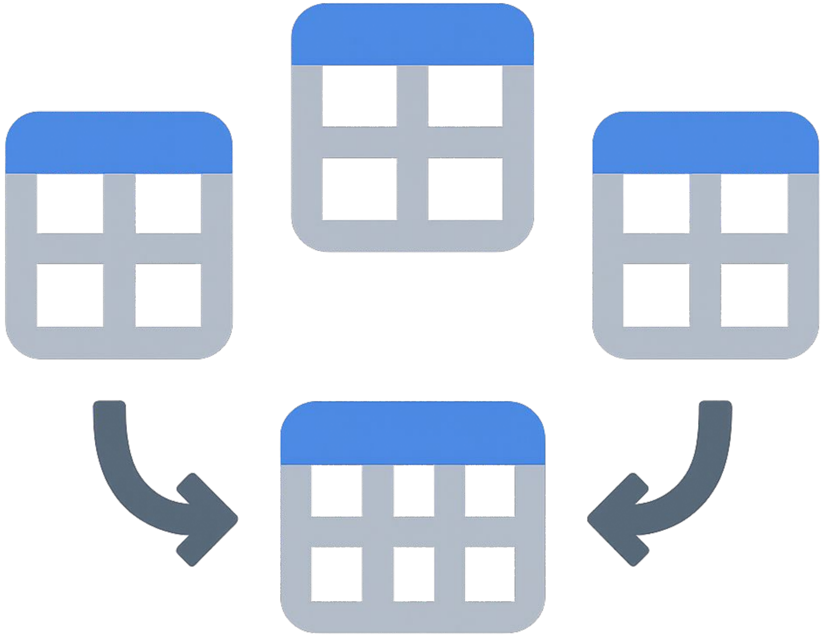
All matched by `id` into a unified DataFrame

Final Cleanup & Validation:

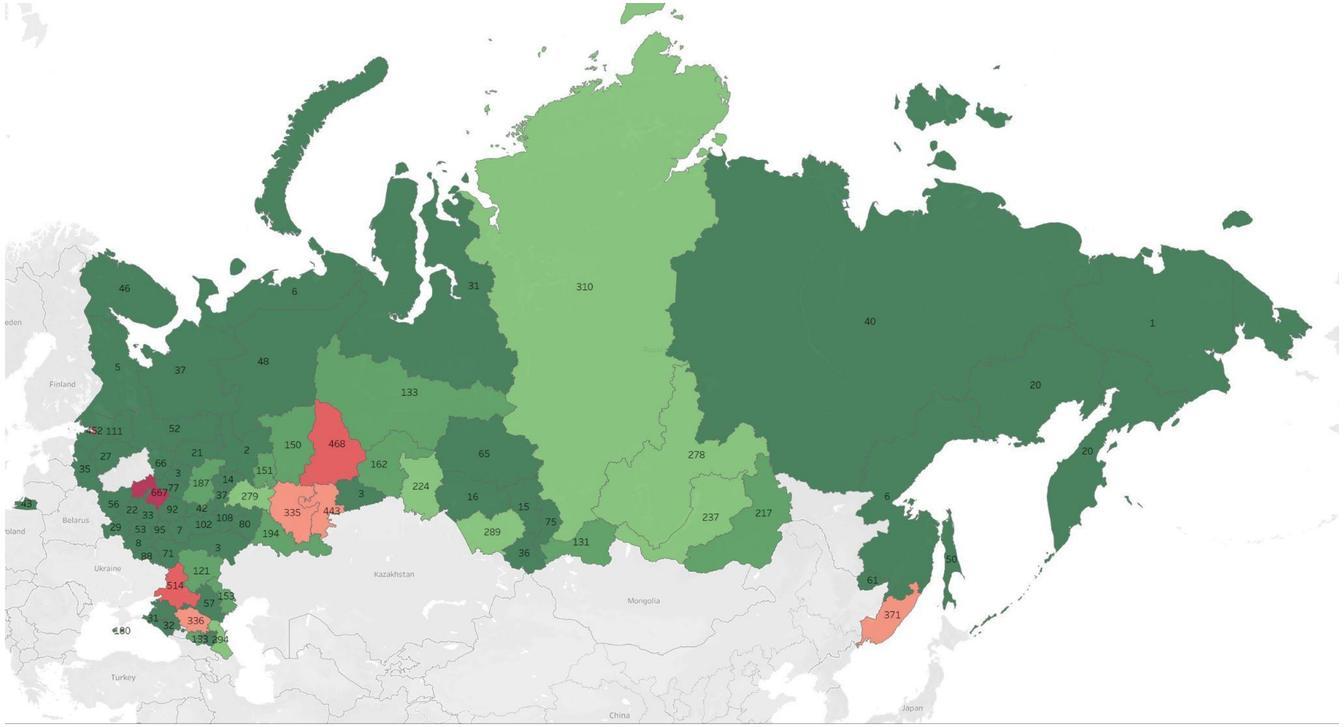
- Removed duplicates
- Verified presence of key features
- Minor columns added/removed for visualization purposes

Was done for:

- Analysis & statistics
- Dashboards (e.g., Tableau)
- Model performance evaluation



Visualizations



To enhance the spatial analysis, a new feature — **region coordinates** — was added to the dataset.

Steps Taken:

1. Selected region names from 10,000 cases.
2. Manually created a region-to-coordinates dictionary (latitude & longitude).
3. Added latitude, longitude columns to the dataset.
4. Grouped by region → counted number of cases per region.

Output:
An interactive choropleth map showing the distribution of Article 228 cases across Russian regions.

Color gradient:
Light green → few cases
Bright red → high case concentration

Insights:
Easily identifies **regions with anomalously high incidence**, enabling deeper regional analysis.

Visualizations

Gender	Amount
Women	777
Men	8480
Mixed	50
Unknown	476

Preprocessing:

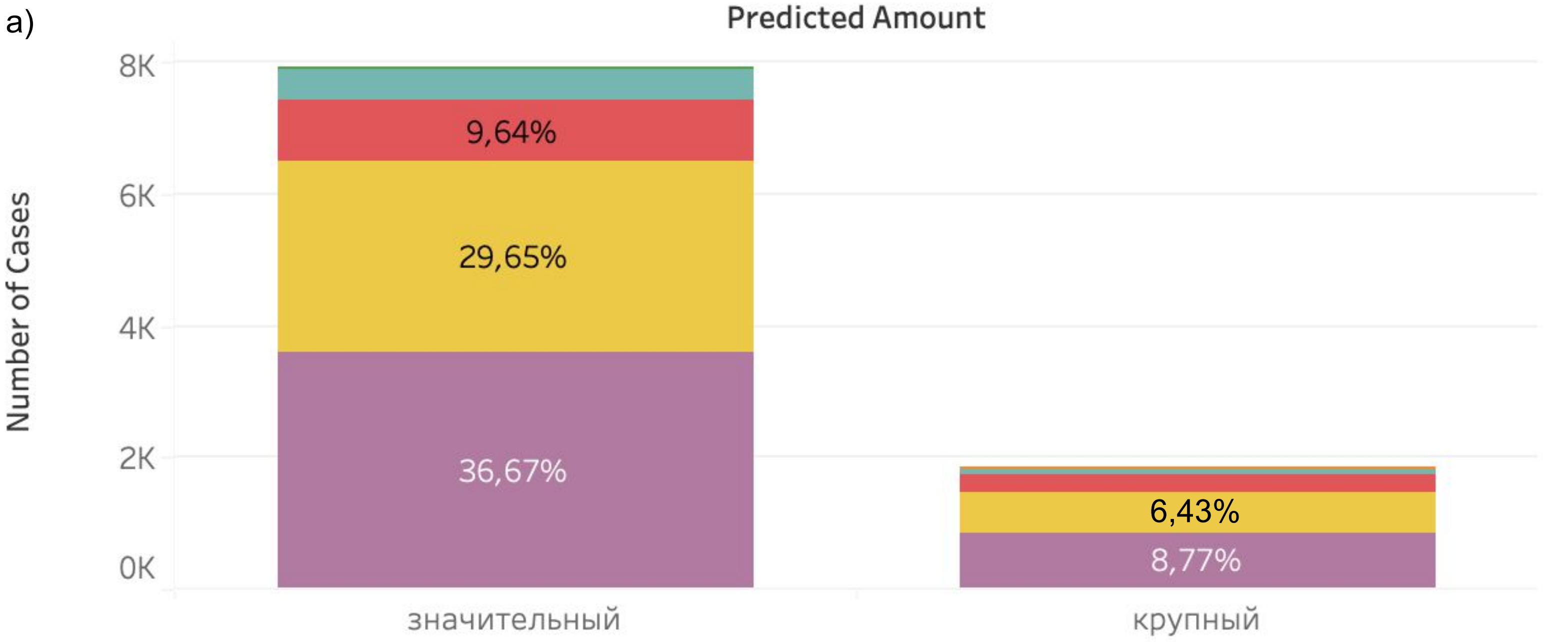
- Normalized the `defendants_gender` column.
- Cleaned ambiguous entries (e.g., `"M, F"`, `"F."`, `"-"`)

Created unified categories:
`Men`, `Women`, `Mixed`, `Unknown`

Key Insight:

- Clear **gender imbalance**: majority of cases involve **male defendants**.
- Rare instances of mixed or unspecified gender cases.

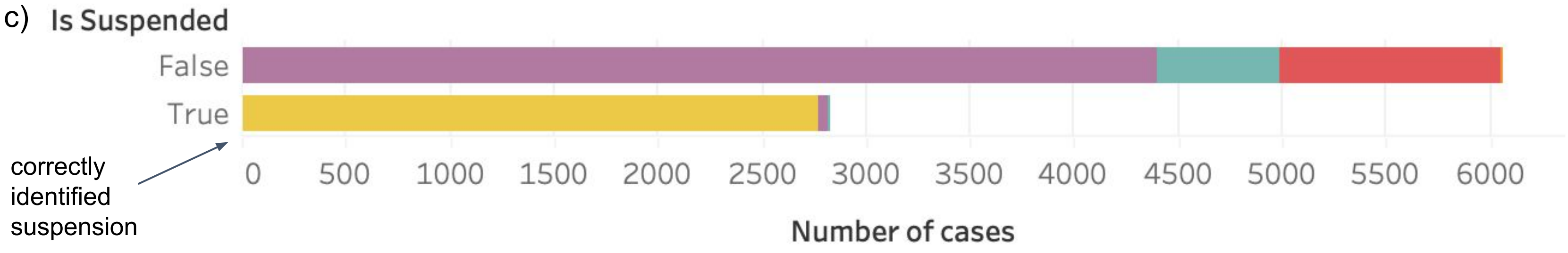
a)



b)

Predicted Result Type	
штраф	4 445
условное осуждение	3 530
лишение свободы	1 203
обязательные работы	599
исправительные работы	4
ограничение свободы	2

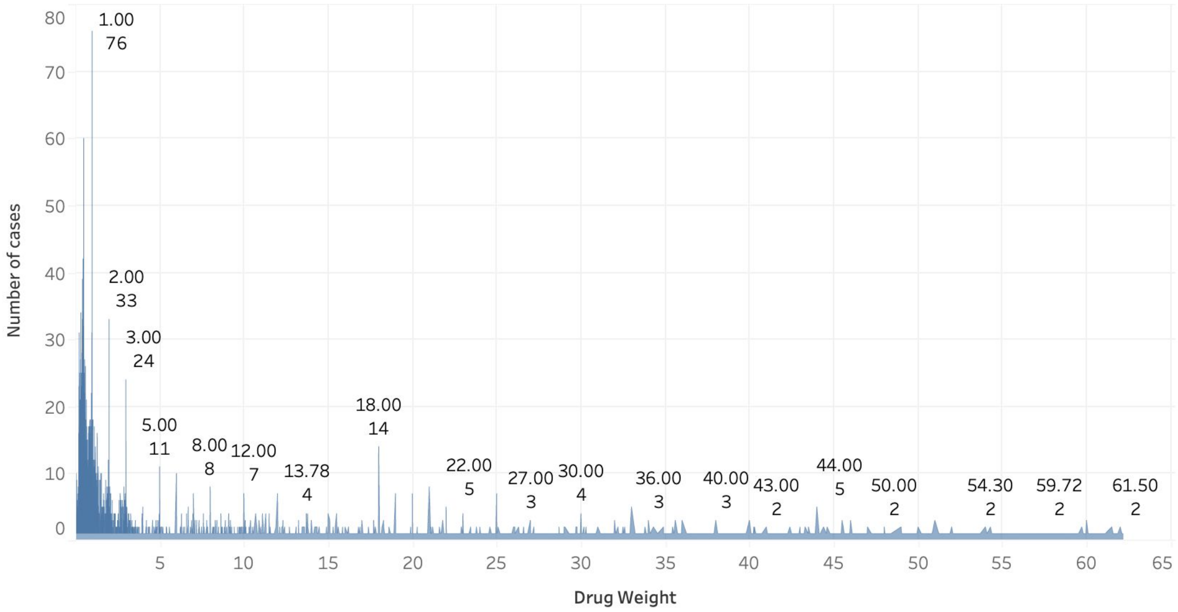
c)



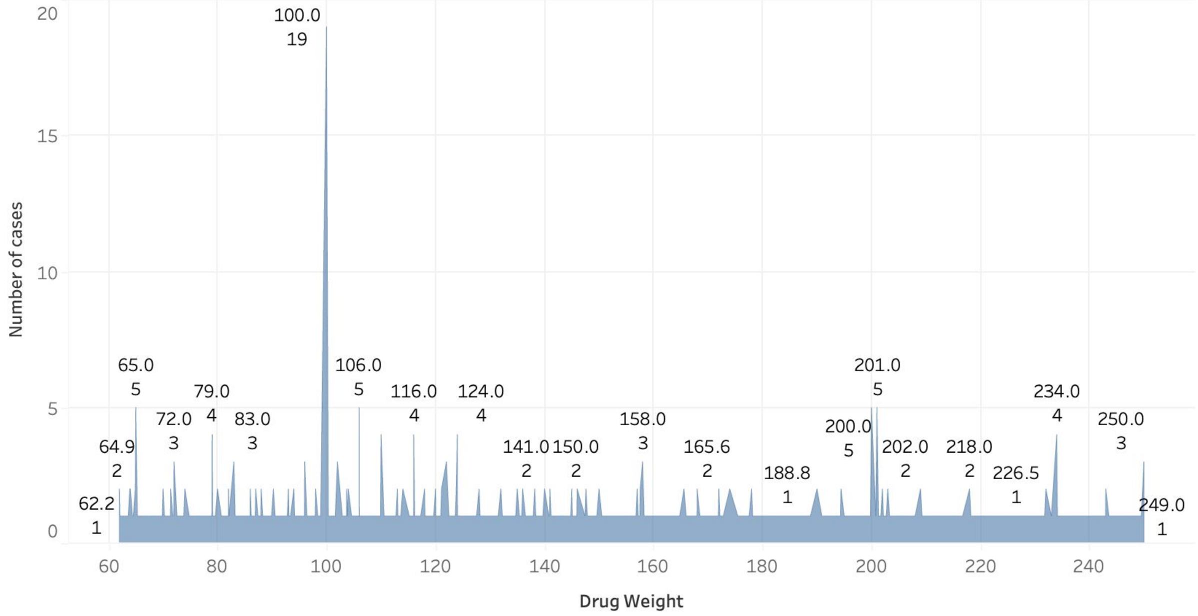
The dashboard: a) Distribution of result_type by drug_amount category, b) Overall distribution of predicted result_type, c) Suspended vs Non-Suspended sentences by result_type

Number of cases distribution for drug weight for different scale: a) from 0 to 62 grams, b) from 62 to 250 grams

a)



b)



The distribution of drug weights shows **sharp peaks at legally significant thresholds** like 1g, 100g, and 200g, suggesting influence from law enforcement practices. These patterns indicate that weight values *may reflect* not only actual possession but also **legal boundaries** and **potential procedural manipulation**.

Hypothesis 1

H: Larger drug amounts reduce the probability of receiving lenient punishments (fines or suspended sentences) compared to imprisonment.

Methodology: Multinomial logistic regression model using

- predicted_amount_bin = 1 if drug amount is large
- predicted_amount_bin = 0 if drug amount is significant

Results:

- Statistically significant effect for suspended sentence:
  - Coefficient: –0.2403
  - p-value: 0.004
- Near-significant effect for fine:
  - Coefficient: –0.1421
  - p-value: 0.075

```
                          MNLogit Regression Results
==============================================================================
Dep. Variable:        result_type_cat   No. Observations:                 9783
Model:                        MNLogit   Df Residuals:                     9773
Method:                           MLE   Df Model:                            5
Date:                Mon, 26 May 2025   Pseudo R-squ.:               0.0004303
Time:                        19:53:48   Log-Likelihood:                -11342.
converged:                      False   LL-Null:                       -11347.
Covariance Type:            nonrobust   LLR p-value:                   0.08213
==============================================================================
result_type_cat=условное осуждение      coef    std err          z      P>|z|
-----------------------------------------------------------------------------
const                                 1.1237      0.037     29.978      0.000
predicted_amount_bin                 -0.2403      0.083     -2.905      0.004
-----------------------------------------------------------------------------
          result_type_cat=штраф         coef    std err          z      P>|z|
------------------------------------------------------------------------------
const                                  1.3360      0.037     36.507      0.000
predicted_amount_bin                  -0.1421      0.080     -1.783      0.075
------------------------------------------------------------------------------
```

The results suggest that **higher drug amounts significantly reduce the likelihood of receiving suspended sentences** and may also reduce chances of receiving a fine, compared to imprisonment. This **supports** the hypothesis.

Hypothesis 2

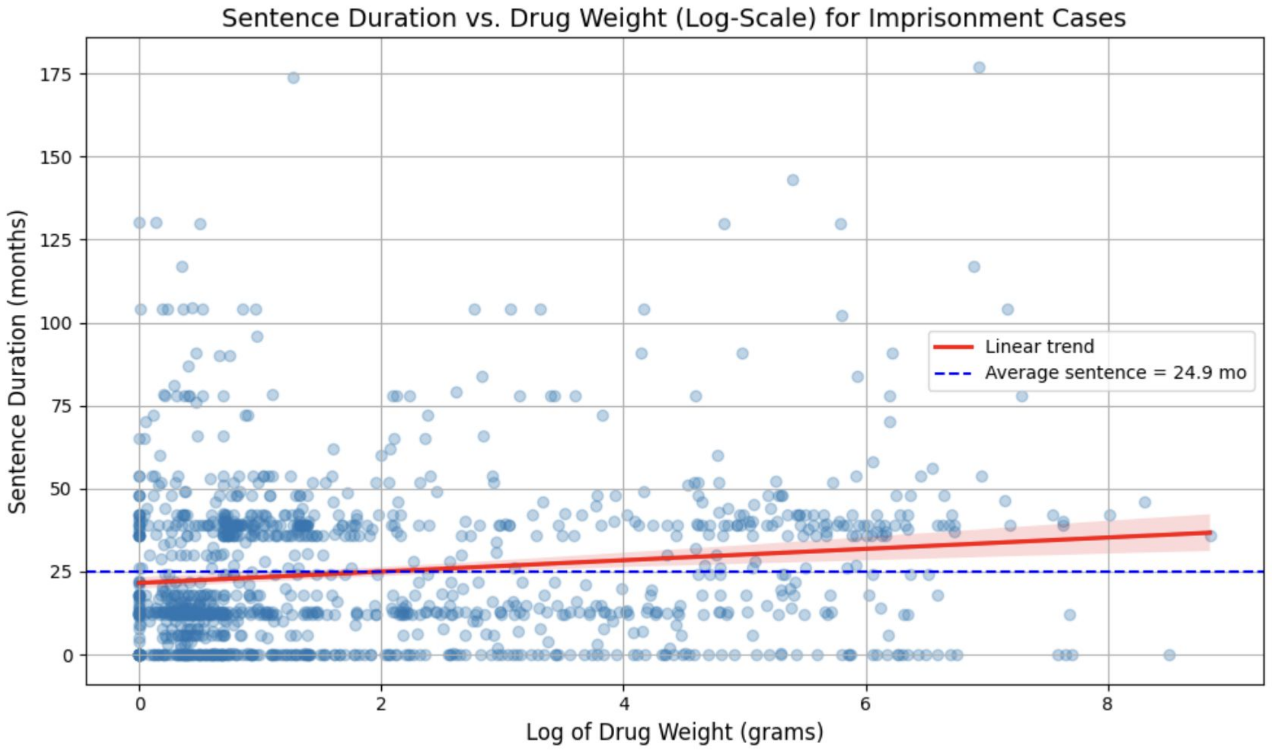
H: Among convicted individuals, larger drug weights are associated with longer prison sentences.

Methodology: Linear regression model (OLS) on cases with result_type = ‘лишение свободы’

- sentence_total_months - dependent variable
- log(drug_weight) - independent variable

Results:

- Coefficient: +2.1545
- p-value: 0.023
- R^2: 0.004



The coefficient indicates that an increase in drug weight **is associated** with a longer prison sentence. However, the **low R^2** value shows that **drug weight alone explains very little variance in sentence length** — other factors also play an important role.

Hypothesis 3

H: Women receive more lenient punishments (fines or suspended sentences) than men, without considering other factors.

Methodology:
A simplified logistic regression model:

- `female = 1` if the defendant is a woman
- `female = 0` other

Results:

- Sample size: 9,257 cases
- Coefficient for the variable `female`: 0.0987
- p-value: 0.316

The positive coefficient suggests a slightly higher probability of receiving a lenient sentence for women. However, the effect is **not statistically significant**. Therefore, the data does not provide sufficient evidence to support the hypothesis.

Logit Regression Results

Dep. Variable:	soft_punishment	No. Observations:	9257
Model:	Logit	Df Residuals:	9255
Method:	MLE	Df Model:	1
Date:	Tue, 27 May 2025	Pseudo R-squ.:	0.0001143
Time:	11:32:20	Log-Likelihood:	-4480.3
converged:	True	LL-Null:	-4480.8
Covariance Type:	nonrobust	LLR p-value:	0.3116

	coef	std err	z	P>|z|	[0.025	0.975]
const	1.4517	0.028	52.416	0.000	1.397	1.506
female	0.0987	0.098	1.003	0.316	-0.094	0.292

Vector of future development

- **Expanding the training sample**

- **Improving extraction of complex features**

  `aggravating_factors, mitigating_factors, drug_type`

- **Expanding analytical capabilities**
  - Perform **factor analysis** to identify key influences on court decisions
  - Use **clustering** to group cases by substance type, region, or sentence st
  - Test **fairness hypotheses**, such as regional or gender biases in sentenc

- **Multifactorial forecasting of court decisions**

  Develop predictive models that estimate:
  - Type and severity of the sentence
  - Influence of drug amount, article part, region, and circumstances

Conclusion

Our project team:

1. Developed a pipeline to extract several features from court decisions under Article 228 using LLMs and regular expressions.

2. Achieved high accuracy for key characteristics: result_type, drug_weight, sentence length (accuracy > 0.90 in most cases).

3. Identified statistical patterns in sentencing – drug amount impacts the severity of punishment.

4. Found no significant effect of gender on sentence leniency in a simplified model.

5. Conducted visual and regression analysis, that revealed hidden trends and raised questions about fairness and consistency.

This project demonstrates that modern NLP tools and Regular Models **can effectively extract** and **analyze** complex legal data and paving the way for **scalable, data-driven legal analysis** in Russian judicial practice.

Sources

1. **Extracting Legal Facts with ML and Embeddings (Luo et al., 2021)**
2. **Legal Data Extraction in Russian Cases (Ivanov et al., 2022)**
3. **LLMs in Legal NLP: GPT-4 for UK Tribunals (Smith et al., 2023)**
4. **LLMs for Legal Design & Education (Brown et al., 2022)**
5. **How can we… use AI to predict the outcome of court cases?** ***(Professor Felix Steffek, Faculty of Law and Dr Helena Xie, Centre for Business Research)***
6. **Extracting Proceedings Data from Court Cases with Machine Learning** ***(Bruno Mathis)***
7. **Predicting Indian Supreme Court Judgments, Decisions, Or Appeals** ***(Sugam Sharma, Ritu Shandilya and Swadesh Sharma)***
8. **USING AI SYSTEMS IN JUDICIAL WORK (INVESTIGATION, EVIDENCE AND LEGAL RESEARCH)** ***(Herzi Said, Mourad Khelifa)***
9. **Rethinking the field of automatic prediction of court decisions** ***(Masha Medvedeva)***
10. **Icons for the presentation**
11. **S1E5 | Using machine learning to predict court decisions of the ECtHR | TLOT Podcast (a podcast episode with one of the article authors - Masha Medvedeva)**

HSE UNIVERSITY • EST 1992 •