Criminal Justice Theory, Empirical Study, Machine Learning and Judge Decision-Making*

[an incomplete work]

Mingyang Chen†, Gaojie Song‡, Zhanxue Xu‡, Zhipeng Wu‡, Aokai Wang§, Zian Ren§, Boyang Xu**

Abstract

This research finds that Criminal Justice Theory can be combined with Prison Term Prediction Machine Learning Models through Empirical Study. And by employing this model, it can correct Judges' Decision-Making. We choose "treating like cases alike" (TLCA) as the Criminal Justice Theory in this study, which means in judicial documents, the similar facts description parts are between 2 documents, the similar reason writing parts are as well as the prison term parts. We first conduct an empirical study using 5081 robbery judicial documents in China. We use text mining tools and find Judges cannot always make a TLCA decision in prison term parts. Those with long prison term cases are more likely to be treated unequally. In order to solve this problem, we use empirical study to delete all unequal cases. We build 2 sets of Machine Learning Models each containing 4 models, one sets using samples before we filter, while the other one using samples after deleting unequal cases. We find after filtering by TLCA theory, Machine Learning models perform better than not doing so. Then we test the ability of our model to correct unequal decisions. The result shows our model have the ability to correct wrong prison term length.

Key Words: machine learning, empirical study, judge decision, treating like cases alike

* This paper is sponsored by Data Law Lab of China University of Political Science and Law (CDLL). Dai Li, Libo Fan, Xinyun Tu from China University of Political Science and Law (CUPL), Zhuang Liu from Hongkong University (HKU) provide me valuable suggestions on this paper. All remaining errors are our own.

† Fist author. Email: cmy0211@yeah.net

‡ Second author. Equal contribution.

§ Third author. Equal contribution.

** Forth author. Corresponding author. Email: xuboyangcupl@126.com

INTRODUCTION

Machine Learning models have already been an excellent tool in prediction. Scholars are working on combine machine learning with expert knowledge in order to develop machine learning models' ability.[1] In the field of criminal justice, Prison-Term-Prediction model (PTP model) perform really well. However, we still want to know the relationship between Criminal Justice Theory and PTP models. In order to conduct research on it, we choose "treating like cases alike" (TLCA) as the Criminal Justice Theory in this study. We collect 5081 robbery judicial documents in China. First, we complete an empirical study in order to answer whether China achieve the goal of TLCA. Then, we find the answer is no. Judges in China perform badly in prison term part. Those with long prison term cases are more likely to be treated unequally. In the second part, we try to use TLCA theory and empirical study to delete all unequal cases. We then get 2 sets of data; one is cases with no filter. The other one is using TLCA deleting unqualified cases. We build 2 sets of Machine Learning Models each containing 4 models, one sets using samples before we filter, while the other one using samples after deleting unequal cases. We find after filtering by TLCA theory, Machine Learning models perform better than not doing so. The R-square rockets and MSE drops after filter. Then we create a stacking model using all base learners and test the ability of our stacking model to correct unequal decisions. The result shows our model have the ability to correct wrong prison term length. Through this research, we find that Criminal Justice Theory can be combined with Prison Term Prediction Machine Learning Models through Empirical Study.

LITERATURE REVIEW

Theory review: Treating like cases alike

Hart believes that the justice itself could be various, but the most important rule for justice is every 2 individuals should be treated equally under the same justice. This is the origin of the principle called "treating like cases alike". Hart insists that "treating like cases alike " is an administration of judgement, not the justice itself. As long as the Judge do right verdicts, then like cases would be

[1] Devadrita Nair & Arnd Huchzermeier, Predictably Unpredictable? How Judgmental and Machine Learning Forecasts Complement Each Other, 33 Prod. & Ops. Mgmt. 1214 (2024); Jon Kleinberg, Himabindu Lakkaraju, Jon Kleinberg, Jure Leskovec, Jens Ludwig, Sendhil Mullainathan. Human Decisions and Machine Predictions, 133 Q.J. Econ. 237 (2018).

treated alike.[2] We can find in his earlier essay in 1985 that Hart has already declared this opinion. "This (treating like cases alike) is justice in the administration of the law, not justice of the law…implement just this aspect of law and which are designed to ensure that rules are applied only to what are genuinely cases of the rule or at least to minimize the risks of in-equalities in this sense."[3] Timothy J. Capurso also holds a view by analyzing judicial decision theory that if Judge X will decide a certain case one way, while Judge Y would decide that same case in another way, it may not be advantageous.[4] Later scholars developed a series of reasons to support the theory including "Unity of Legal Interpretation of Rules and Order", "Prevention of Arbitrariness", "Predictability of Punishment", "Requirement of Legal Stability", "Publicity and Authority Assurance of Judgments", "Protection of Expectant Interests" and etc.[5] However, Andrei Marmor points out treating like cases alike is not any conclusive answer. If for the same reasons above, to betray the principle of treating like cases alike is allowed.[6] But in most situation, it can raise a warning flag about possible problems when they are treated differently, or as demanding an explanation for differences in treatment, the principle is sound, and important. As a result, although treating like cases alike has some disadvantages, it is a common practice, and is called a "weak claim" according to Chen.[7] The international society has already received this theory, e.g., the International Association of Refugee Law Judge in 1990s.[8]

In conclusion, there should be same reason for same sentences.[9] According to the theory treating like cases alike, I divide treating like cases alike into a 3-part chain. The chain is case—reason—sentences. This is a rough frame as I discussed before. the following figure 1.1 illustrate this.

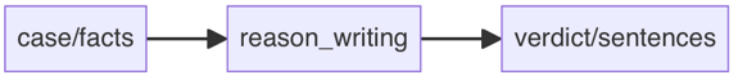


Figure 1. Theory Frame

[2] H. L. A. Hart, The Concept of Law 155-61 (2d ed. 1996).
[3] H. L. A. Hart, Positivism and the Separation of Law and Morals, 71 Harv. L. Rev. 593 (1958).
[4] Timothy J. Capurso, How Judges Judge: Theories on Judicial Decision Making, 29 U. Balt. L.F. 2 (1998).
[5] Kenneth I. Winston, On Treating Like Cases Alike, 4 The Independent Rev. 107 (1999); Andrei Marmor, Should Like Cases Be Treated Alike, 11 Legal Theory 27 (2005); Jinghui Chen, Same Case, Same Judgment: A Legal Duty or a Moral Requirement?, 3 China Legal Sci. 46 (2013).
[6] Andrei Marmor, Should Like Cases Be Treated Alike, 11 Legal Theory 27 (2005).
[7] Jinghui Chen, Same Case, Same Judgment: A Legal Duty or a Moral Requirement?, 3 China Legal Sci. 46 (2013).
[8] Hugo Storey, Consistency in Refugee Decision-Making: A Judicial Perspective, 32 Refugee Survey Q. 112 (2013).
[9] Andrei Marmor, Should Like Cases Be Treated Alike, 11 Legal Theory 27 (2005).

According to Bai's study, there is still a considerable scope where the application of the law has not achieved the same judgment for the same case and the same penalty for the same amount, but rather different judgments for the same case and different penalties for the same amount.[10] Bai and Wu even developed a model for verdict based on so-called "Judges' experiences". Through Wu's models he tries to find which judge's verdict is off-path, meaning it may be an incorrect one.[11] We can find scholars focus more on the verdict conclusion, or the sentences in the field of criminal law only. The reason writing part is usually ignored. So, this research is going to test this chain. Based on prior discussion, this research put forward 2 hypotheses.

H1.1: in judicial documents, the similar facts description parts are between 2 documents, the similar reason writing parts are.

H1.2: in judicial documents, the similar reason writing parts are between 2 documents, the similar prison term parts.

Prior ML-PTP models

In the past, there was a significant amount of work on sentencing prediction, but legal scholars and computer scientists often worked separately. Some scholars tend to use a data-driven approach, neglecting the extraction of sentencing elements. They usually rely on semantic analysis to establish a correspondence between legal provisions and case text.[12] Another method is the use of N-Gram for automatic extraction, which has the same shortcomings as the aforementioned semantic analysis.[13] These methods are prone to spurious correlations, and the algorithms have poor interpretability. Some scholars, despite being theory-driven, have used element extraction methods, but there have been serious errors in the extraction, such as treating "ringleaders" as "principal offenders." Additionally, the scope of criminal sentencing in studies using element extraction methods is too narrow.

[10] Jianjun Bai, The Constitutional Significance and Empirical Study of the Principle of Consistency in Judgment for Similar Cases, 3 China Legal Sci. 131 (2003).
[11] Jianjun Bai, Research on Sentencing Prediction Based on the Collective Experience of Judges, 38 Chinese J. L. Stud. 140 (2016); Yuhao Wu, The Boundaries of Sentencing Discretion: Collective Experience, Individual Decision-Making, and Bias Identification, 43 Chinese J. L. Stud. 109 (2021). .
[12] Haoxi Zhong, Guo Zhipeng, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu, Maosong Sun. Legal Judgment Prediction via Topological Learning; Zikun Hu et al., Few-Shot Charge Prediction with Discriminative Legal Attributes, in COLING 2018.
[13] Hongshui Shu, Research on Judicial Big Data Text Mining and Sentencing Prediction Model, 464 Legal Sci. No. 7 (2020).

Furthermore, previous research did not fully consider the issue of consistent sentencing, meaning that the output results have not been tested for consistency in similar cases. Many data-driven methods may also have algorithmic bias, so it is necessary to use algorithms with stronger interpretability in order to be applicable in practice.

In addition, a team from Zhejiang University has developed a learning model based on the theory of TLCA[14]. The working principle of this model is to find similar cases and then derive the sentence for the current case through the sentences of similar cases. However, this method cannot exclude cases where "the quality of the case itself is extremely poor." So, we decide to use empirical study to select all high-quality cases and train our model, and then compare with baseline model.

Besides, we create a stacking model using all models as base learners. We think our model, by training with high-quality data, can correct the biased result of human Judges. Based on prior discussion, we put forward 2 hypotheses:

H2.1: new model has a higher R-square than the baseline model.

H2.2: new model has a lower MSE than the baseline model.

H2.3: new stacking model can correct the result of biased cases.

METHODOLOGY

Data

1. China's Robbery Judicial Documents

This research uses 5081 judicial documents in China. I collect those documents randomly from the China Judgment Online (CJO). I choose the robbery as my sample. If there is a joint crime, I will extract the defendant only. If it is a multiple defendants' trial, I will not take it into my sample. That would make cases easier to be analyzed. The reasons for choosing robbery are: 1. Robbery is a high-incidence crime and seldom use summary procedure making the contain of documents is rich. Compared with other crime, e.g., larceny and reckless driving, the texts of robbery are longer and more suitable to be analyzed. 2. Robbery according to China's criminal law, have many aggravating circumstances e.g., robbery resulting in death or armed robbery. What is more in China's law, robbery

[14] Siying Zhou et al., Similar Case Based Prison Term Prediction, in Proc. 7th Int'l Conf. on Artificial Intelligence (CICAI 2022).

also has a special form known as "transformative robbery," which is stipulated in Article 269 of the criminal law. It states that “a person who commits theft, fraud, or snatch theft, and who uses violence or threatens violence on the spot to conceal stolen goods, resist arrest, or destroy evidence” will be legally transformed into the crime of robbery by a technical construction of the law. In the case of snatch theft, there is also a special provision known as "snatch theft by vehicle," which is considered a form of robbery. Additionally, robbery is a compound act that includes two elements: "the use of violence or threat of violence" and "the intention to acquire property," which leads to many issues regarding accomplices and attempts. For example, if a person attacks another violently and then decides to take their belongings, whether it is considered robbery or intentional injury coupled with theft. Another example is when a person violently robs and knocks out the victim, and a third party joins in to take the property; this raises questions about the determination of complicity. 3. Robbery is a type of property crime that can easily be confused with other property crimes, such as extortion and blackmail. For instance, in the case document we collected, "Chen Xiaoyong on robbery and extortion", the judge reasoned by distinguishing between "using violent and coercive means" and "using threatening means" to determine which of the defendant's multiple actions constituted robbery and which constituted extortion. In summary, robbery provides rich material for the Judge' reasoning in legal documents. Therefore, selecting robbery as the material for this study can provide a stronger diversity of reasoning-related data, which can better reflect the issues.

Besides, all documents are cases of first instance. China has a two-instance trial system and the second instance trial can correct the first instance trial. Most first instance trials are done by basic court in China. So, if we mix all cases, the outcome is not reliable. But there are problems only using cases of first instance, because some of them maybe incorrect. However, from the data of Law Yearbook of China we can confidently say it is not a matter. The rate of case reversal in China has been around 10% to 15% for a long time. In the latest yearbook data, the overall rate for criminal cases (calculated by subtracting the number of concluded first-instance cases from the number of concluded second-instance cases) is only 11.97%. These also include situations where the reasoning does not change significantly.[15] Therefore, even if there are some errors in the reasoning of the first

[15] Law Yearbook of China 1997-2007, 2021.

instance, it will not severely affect the overall structure of the sample. So only using cases of first instance is reliable.

2. Feature Extraction

Every document from our sample could be divided into 4 parts: "Facts Accused by the Authority", "Facts Found by the Court", "The Court Believes" and "The Judgment is as Follow". The second part is the fact found by court based on evidence collected. We use openpyxl in python to cut judicial documents into 3 parts.

What I need to add about is the sentence part. Sentences in China is combined of 2 parts, main sentences and additional sentences. Main sentences are usually a fixed-term imprisonment, and additional sentences is usually deprivation of political rights or fine or others. Some scholars give a formula to calculate the severity of punishment in China, but they are not reliable however because most of them do not open how they get this formula.[16] So I decide to use the main sentences only. I will prove the rationality by the PCA. The severity of sentences is in fact a multi-dimension data so we can use PCA to reduce dimensionality. In all formulas scholars give, and just as scholars all believe, that the additional sentences in China only takes a very small portion. So, after PCA, we can get a principal component that is highly similar to the main sentences. As a result, we can straightly use the main sentences because it makes sense. This method is also used in many researches in China.[17] Then I use how long months the sentences are to estimate this variable.

In order to estimate the relationship between reason writing and prison term, which is proved to be a problem so we have to focus on it, we select factors listed by China's criminal law. The end of the part is the list containing variables we use. We give all variables a dummy value. After dealing with data by ways above, we get a set of data containing facts description text, reason writing text, prison term and all factors extracted from text.

---

[16] Jianjun Bai, Quantitative Analysis of the Severity of Punishment, 6 Chinese Social Sci. 114 (2001); Jianjun Bai, Quantitative Analysis of the Severity of Crime, 6 Chinese Social Sci. 123 (2003); Jianjun Bai, Empirical Research on Sentencing Standards, 1 Chinese J. L. Stud. 97 (2008); Changming Hu, An Empirical Analysis of the Impact of Social Structural Factors on Sentencing: A Case Study of Sociological Research on Theft Crimes, 3 Legal App. 54 (2011).

[17] Yuhao Wu, The Boundaries of Sentencing Discretion: Collective Experience, Individual Decision-Making, and Bias Identification, 43 Chinese J. L. Stud. 109 (2021); Boyang Xu, You Zhou & Chunli Zhang, An Empirical Test of Social Bond Theory and Self-Control Theory on Sexual Offending: Based on an Analysis of a Sample of 260 Sexual Offenders in China, 4 Crime Research 50 (2021); You Zhou, Boyang Xu, Ivan Y. Sun, Yan Zhang & Lennon Y. C. Chang, Examining Sexual Crime Severity in China: A General-Specific Model on Sex Offending Against Adults, 34 Sexual Abuse 830 (2022).

List 1. Variables we use

Types	Variables
犯罪方法 (Crime method)	入户、冒充军警、持枪、持刀/携带凶器、转化 (Entering the house, Pretending to be a military policeman, Carrying a gun, Carrying a knife/carrying a murder weapon, and Converting robbery)
犯罪地点 (Crime place)	交通工具、金融机构 (Rob in Transportation or Financial institutions)
犯罪次数和记录 (Times and prior crime)	累犯、前科、多次犯罪 (Recidivism, Ex-offenders, Repeat offenders)
犯罪形态 (Crime patterns)	预备、未遂、中止 (Preparation, Attempt, Abort)
犯罪地位 (Crime status)	主犯、从犯、首要分子 (Principal offender, Accessory, Ringleader)
被告人因素 (Defendants' factors)	自首、立功、坦白、自愿认罪、认罪认罚、精神疾病、未成年、赔偿 (Voluntary surrender, Meritorious service, Confession, Voluntary admission of guilt, Admission of guilt and acceptance of punishment, Mental illness, Juvenile, Compensation)
被害人因素 (Victim's factors)	谅解、和解、残疾、老年、孕妇、未成年 (Understanding or Reconciliation with defendants, Disability, Old Age, Pregnant, Minor)
伤情 (Injury)	轻微伤、轻伤、重伤、死亡 (Leading to Slight injuries, Minor injuries, Serious injuries or Death)
其他 (Other factors)	黑恶势力、数额巨大、灾害、军用物资、抢险救灾救济物资、罚金 (Underworld forces, Huge amounts, During disasters, Military supplies, Rob Emergency and Disaster relief materials, Fines)

Method

For textual data, RWMD is used to calculate similarity. For prison term data, Euclidean distance is used to calculate similarity. For dummy variables, cosine similarity is used to calculate the distance. This study uses the word2vec and glove models that have been trained in the text2vec package of the R language to generate word vectors, ensuring that synonyms are considered in the calculation of text similarity, and then calculates the vector distance between texts. The RWMD method is good at classifying long texts by calculating similarity, but its disadvantage is obvious, that is, it cannot handle negations. However, this disadvantage is very suitable for the calculation of judicial document data, because judicial documents almost never contain negations (judges will not say, someone does not

have a large subjective malice, but only list the necessary conditions). Then, this study selects a case as a benchmark case, calculates the similarity of case facts, reasoning, and sentencing with this benchmark case, and then uses this data to construct a univariate linear regression model to see if there is a rule of "the more similar the case, the more similar the reasoning, the more similar the sentence." To ensure the reliability of the empirical conclusions, this study also uses cosine similarity to perform the same operation; this study also uses a more traditional method of variable extraction to verify the conclusions of this empirical analysis again; at the same time, this study uses random sampling to change the benchmark case to get more robust results.

For the prediction part, since this problem is a prediction model, this study considers the lasso model, ridge model, and linear model that perform well in prediction algorithms to predict the problem. Models are constructed through sklearn, and the data set is randomly divided into a test set and a training set according to the proportion for model training and fitting. Finally, the training results are evaluated through the mean square error function. For the data set part, this paper adopts two data sets for training, and then compares the results obtained. Below is the statics of our variables calculated in RWMD and Euclidean distance.

Table 1: statistics of variables

	Case Similarity	Reason Similarity	Prison Term Similarity
Min	0.0000	0.3921	0.00
Median	0.7172	0.6103	2.944
Mean	0.7180	0.6128	3.219
Max	1.00000	1.00000	5.094
Num.Obs.	5081	5081	5081

FINDINGS

OLS regression and Robustness

1. RWMD method and Linear Diagnostics

Table 2 shows the result of RWMD method. We randomly choose 5 cases as baseline. The C-R refers to the coefficient between case facts and reason writing. While the R-P refers to reason writing and prison term. We can find that C-R is stable and H1.1 is proved to be True. While in No. 5071, there is a strange number. H1.2 is proved to be unstable.

Table 2: OLS of RWMD

	No. 8	No. 1	No. 3923	No. 5071	No. 4806
C-R Cof.	0.15203	0.225497	0.116908	0.052300	0.16826
p-value	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
R-P Cof.	-5.0235	-3.3419	-1.0013	0.1867	-1.948
p-value	< 0.01	< 0.01	< 0.01	0.3982	< 0.01

For every linear model, we need to conduct regression diagnostics to test their stability. There may be some other model like a polynomial regression model, making our result is not reliable. Other problems will affect the model's stability as well, so regression diagnostics is a must. I use plot to conduct regression diagnostics, and the result is showed below.[18]

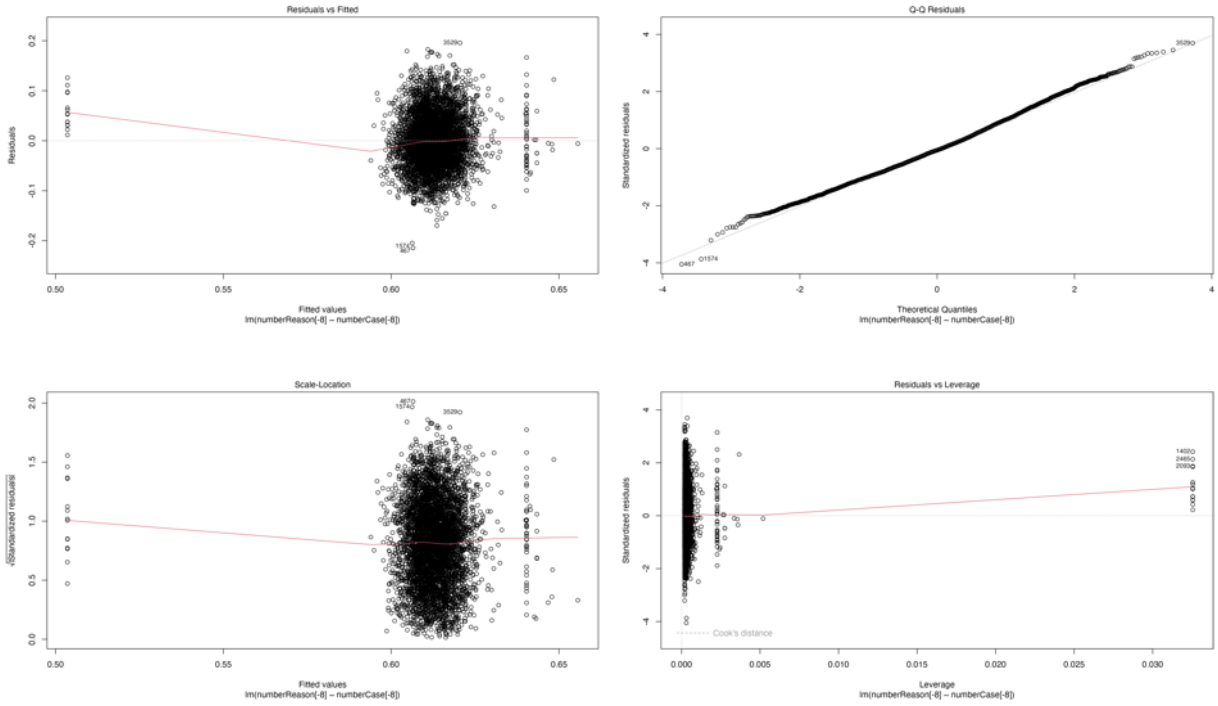


Figure 2.1 – 2.4

[18] C. Ford, Understanding Robust Standard Errors (2020), UVA Library StatLab, https://library.virginia.edu/data/articles/understanding-robust-standard-errors/ (last visited Feb. 1, 2024).

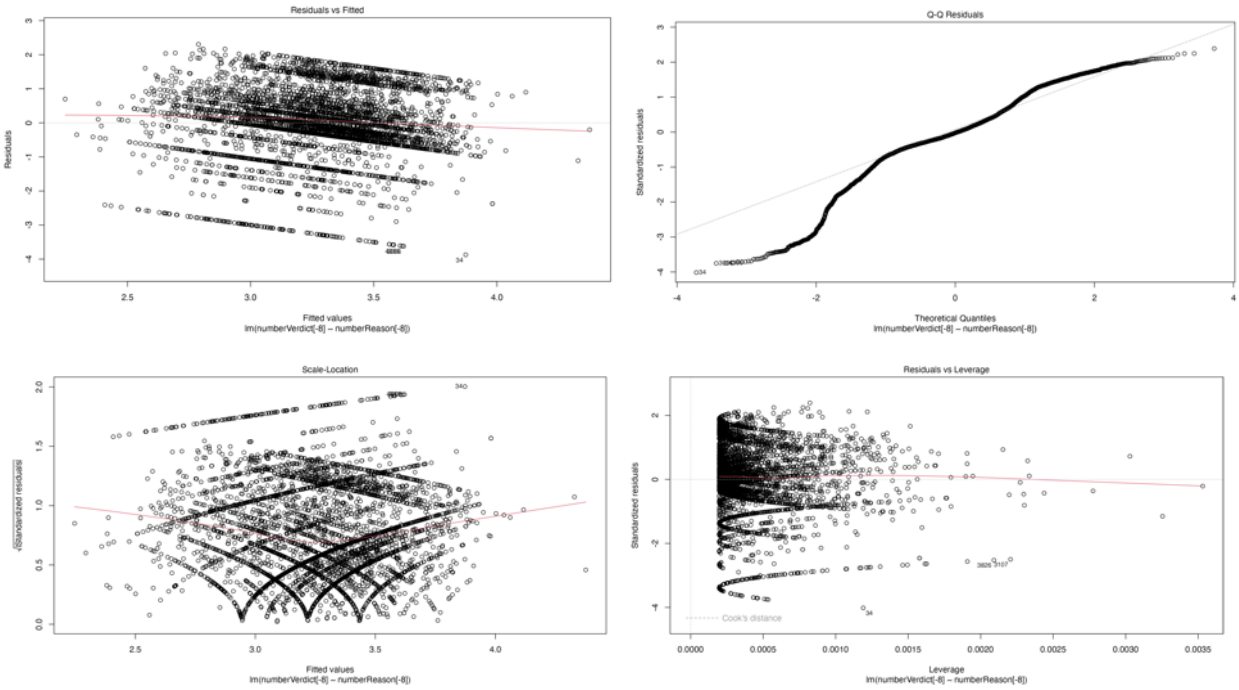


Figure 3.1 – 3.4

The diagnostic results of 2 models indicate that the variables in this model are essentially uniformly distributed, with no apparent trend of increase or decrease, suggesting that there is no issue with multicollinearity. The normality of the sample is good; there are no serious intra-relationship issues among the samples; and no points have a Cook's distance greater than 0.5, indicating the absence of significant outliers. In summary, my conclusions obtained in this study is reliable.

2. Robustness

Except for RWMD, Cosine Similarity and Correlation Similarity are all other prevalent method for text similarity. Below is the result of different algorithm.

Table 3: OLS of different Algorithm

	RWMD	Cosine	Correlation
C-R Cof.	0.225497	0.249576	0.288393
p-value	< 0.01	< 0.01	< 0.01
R-P Cof.	-3.3419	-0.98190	-0.98056
p-value	< 0.01	< 0.01	< 0.01

The benchmark case, if we change that, the result may alter, and that is a good way to test robustness of results. In table 2 I have already use RWMD to conduct this method. Below is the robustness of Cosine Similarity. We get same conclusion.

Table 4: OLS of Cosine Similarity

	No. 1017	No. 4775	No. 2177	No. 5026	No. 1533
C-R Cof.	0.081784	0.045205	0.323915	0.212049	0.226299
p-value	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
R-P Cof.	-1.0440	-0.1145	-0.6881	-0.5782	-0.6151
p-value	< 0.01	0.588	< 0.01	< 0.01	< 0.01

3. what cases are more likely to be wrong? A discussion

The variables extracted in this study have been listed in the table above. To ensure that the variable values exist in the requirements of legal texts, this study uses variable data for regression. This study conducted a full sample regression and output cases that could not show significant negative correlation. Then this study conducted an output to see what cases are different but judged the same, and analyzed their characteristics. Among all 5081 cases, this study output a total of 647 cases that do not conform to the principle of TLCA. The calculation of cosine similarity requires that it cannot be all 0, so this study added a column of data all set to 1 for calculation. This study still finds that many cases do not conform to this rule. Then this study hopes to observe the characteristics of these cases, so this study calculated the proportion of non-serious crimes in different cases judged the same and the same case judged the same according to the standard of five years as serious crime. It was found that the proportions were 18.08% and 72.899% respectively. That is to say, in the cases that are different but judged the same, a large number of cases are seriously punished. This finding also conforms to Judges are more cautious at lower sentences, and more arbitrary at higher sentences, with more discretion. So future policy on TLCA can focus on serious cases.

One thing I want to talk about more, is the unstable correlation between reason writing and prison term. There is always an unrobust significant linear model whether I use text analysis or other method. This can be explained that if we choose a high-quality benchmark, e.g., the No. 1 case in my sample, which is correctly sentenced, then we can get a good model. If the base case itself is a biased one, then the result of model would not be good. So, we can use this to elect biased cases.

Machine Learning Comparison

1. Case Filter Process

For the filter part, this study performs calculations in RStudio to determine the interrelationship between the similarity of case causes, reasoning, and sentencing. Using loop statements, this study

has screened a large number of cases that are judged the same despite being different, and then performed regression. To improve the interpretability of the algorithm, this study uses dummy variable datasets that decompose the requirements of the case itself to represent the similarity of the case, and then uses cosine distance to calculate the similarity between cases, and then performs regression. Then, each case is taken as a benchmark case and the same operation is repeated. If a case, when used as a benchmark, cannot show a significant negative correlation between case similarity and sentencing similarity, then this case should be eliminated. This forms a dataset of TLCA. This study will remove about 600 samples that do not conform to the TLCA and retrai. The logic is that if a case cannot show a significant negative correlation in regression, then this case is likely not to conform to TLCA. This study has looped 5078 times and removed the output results.

2. Comparison with baseline models

The study conducted three rounds of training and then compared the results of the three training sessions. The parameter settings for the models are as follows: For the linear regression model, this study divides the test set and validation set in a ratio of three to seven, with a random seed of 40; For ridge regression, this study still divides the test set and validation set in a ratio of three to seven, with a random seed of 48, and the coefficient alpha value is set to 1. For Lasso regression, this study still divides the test set and validation set in a ratio of three to seven, with a random seed of 4, and the coefficient alpha value is set to 0.1. For the multilayer perceptron, this study divides the test set and validation set in a ratio of two to eight, with a random seed of 4, setting one hidden layer with 100 neurons, and allowing up to 20,000 iterations. In evaluating the models, this study selects their scores and mean squared errors for evaluation. The smaller the mean squared error (MSE), the better the model performs, and the score is evaluated through the R-square. The better the fit, the closer the R-square is to 1. The result is showed below.

Table 5: Comparison with Baseline Model

Model name	Baseline R2	Baseline MSE	Aft-Filter R2	Aft-Filter MSE
Linear Reg	0.548/0.542	674.6032	0.680/0.784	245.6639
Ridge Reg	0.553/0.532	721.6177	0.701/0.687	257.6409
Lasso Reg	0.531/0.542	745.7592	0.682/0.682	264.0610
MLP Reg	0.764/0.504	747.2403	0.870/0.701	252.2937

After integrating TLCA, there is a significant enhancement in the scoring of the four models, and their MSE has also been substantially decreased. Particularly, the scoring situation of the three linear regression models is quite favorable. This to some degree indicates the efficacy of utilizing criminal law theories to assist in data filter work.

3. Empirical TLCA-test

We combine all 4 models as base learners and build a stacking model. Then we use it to re-test those cases deleted. We invite students from law school to evaluate the result by overviewing the judicial documents. Empirical study is used to measure whether the model prediction result pass TLCA-test.

Table 6: TLCA-test result

	No. 1224	No. 4166	No. 3770	No. 1173	No. 2788
Judges' decision	180	37	36	171	180
Staking model	96	59	92	84	104
Pass TLCA-test? (Y/N)	Y	Y	N	Y	Y

It shows that in all 5 cases we randomly choose, 4 are corrected. Below is the analysis of law school students:

In case number 1224, the defendant merely committed robbery with a knife in a household, and it was his first offense. There were no casualties, and the amount was not huge. However, there was a statutory aggravating circumstance, but the crime did not warrant a life sentence. Nevertheless, the judge sentenced him to life imprisonment, which was obviously too harsh. After correction, the sentence was adjusted to 96 months. In case number 4166, the defendant committed assault with a knife and beat the victim. The judge sentenced him to 37 months, which was obviously too low. In other cases where there was no knife and no assault, the defendant was sentenced to 36 months of fixed-term imprisonment (for example, case number 4162, which did not belong to the same case but judged the same). Moreover, in this case, the principal offender was extracted, and the judge pointed out in the reasoning that the punishment should be increased, but the sentence was similar to that of a general robbery, which was unreasonable. After correction, it was sentenced to 59 months of imprisonment. In case number 3770, the case

involved theft in a household that transformed into robbery, and the knife was also involved. At the same time, there was a minor error in the identification of the elements by the judge. According to the judicial interpretation, this case should be a completed crime, not an attempted crime. Therefore, a higher sentence should be imposed, not just 36 months. This case may not have a significant negative correlation, which is likely due to some flaws in the reasoning itself, but it can have a negative correlation. In case number 1173, it was a robbery with a knife, and the amount was huge. In addition, the fine was paid in time. Apart from that, there were no other statutory sentencing circumstances. However, the sentence was as high as 171 months, which was obviously too high. After correction, it was 84 months. In case number 2788, the sentence was 180 months. This study found that there was a certain error in this case due to the extraction logic, and the actual extraction was 48 months. This case involved robbery in a household, which was identified as an attempted crime, but in fact, the case was already completed. This study corrected it to a completed crime, and the sentence was also 105 months. Therefore, whether it was 180 months or 48 months, it did not conform to the sentence of this case.

So, H2.1 – H2.3 are all accepted.

Conclusion

The research findings indicate the following: **First,** by integrating empirical research methods into machine learning models under the guidance of the theory of TLCA, the performance of the model can be enhanced. **Second,** under the guidance of legal theory, the model can correct the decision bias of Judges to a certain extent. This means that the relationship between legal theory, empirical research, and machine learning is such that legal theory can be integrated into data-driven machine learning through empirical research, thereby assisting Judges in correcting decision-making biases. The future integration of legal theory with computer science and machine learning technology can take this path. **Additionally,** this study reminds computer scientists that the selection and quality improvement of data are crucial for predictive results. Blindly improving the accuracy of predictions cannot well achieve the goal of fairness and justice, hence theory-driven approaches are necessary. This study also explores the future development of legal studies. It advocates that the future of legal studies should actively intersect with other disciplines.

APPENDIX A

下面是实证研究的流程图：

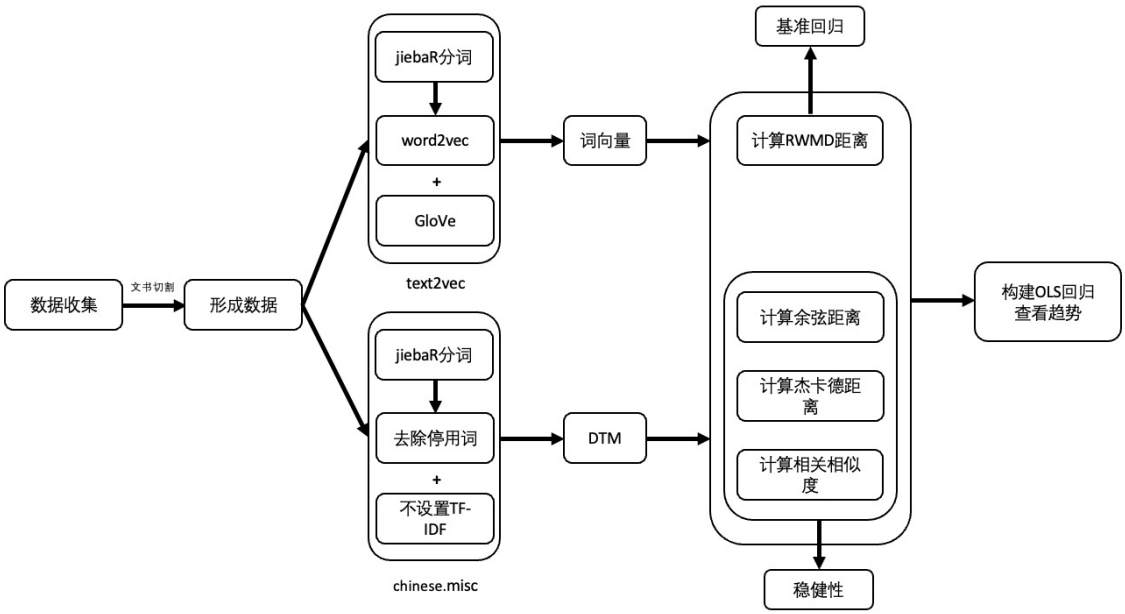


Figure A1. 实证研究流程图

下面是建模流程图：

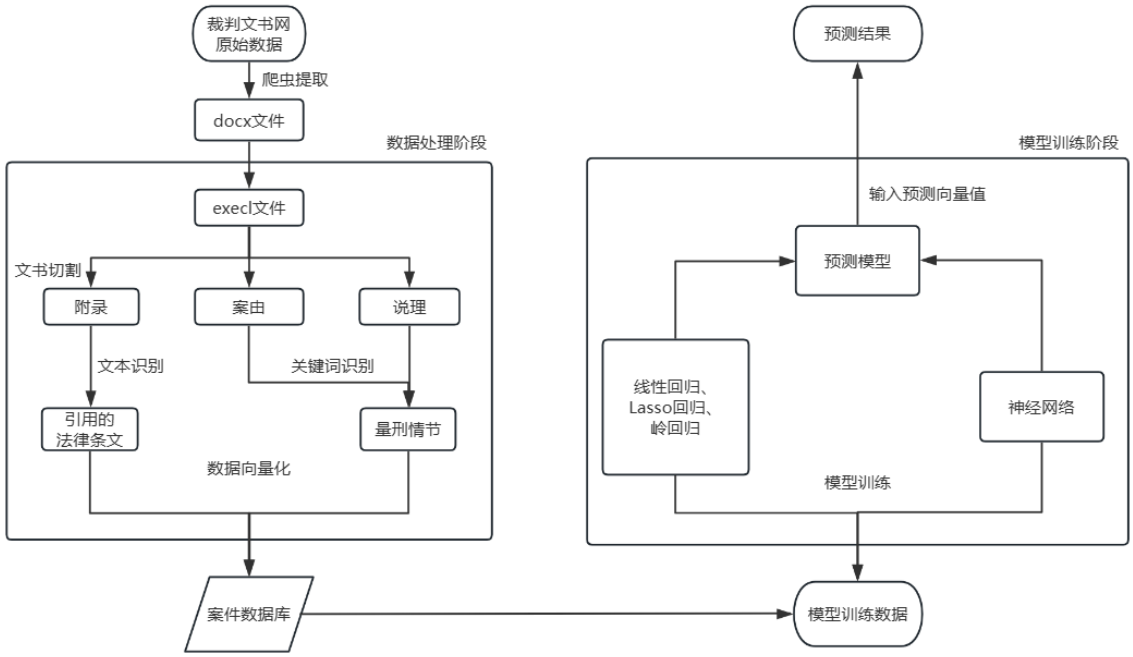


Figure A2. 预测模型流程图

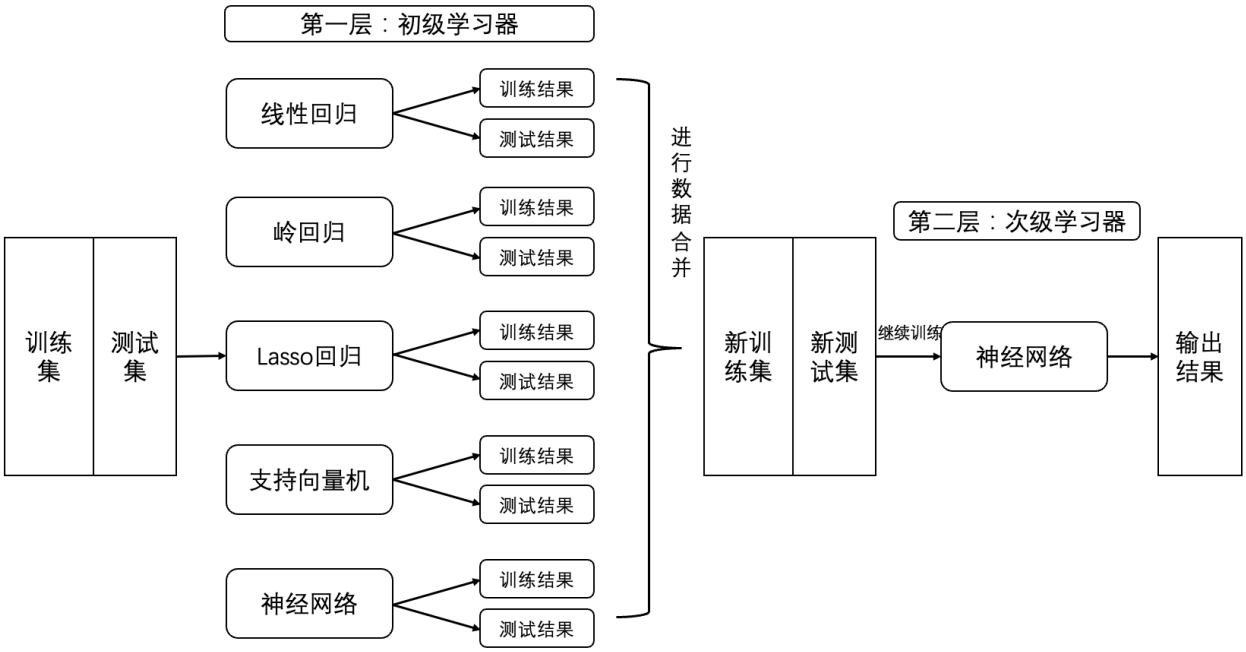


Figure A3. Stacking 集成学习流程图

APPENDIX B

下面是各种数据的直方图：

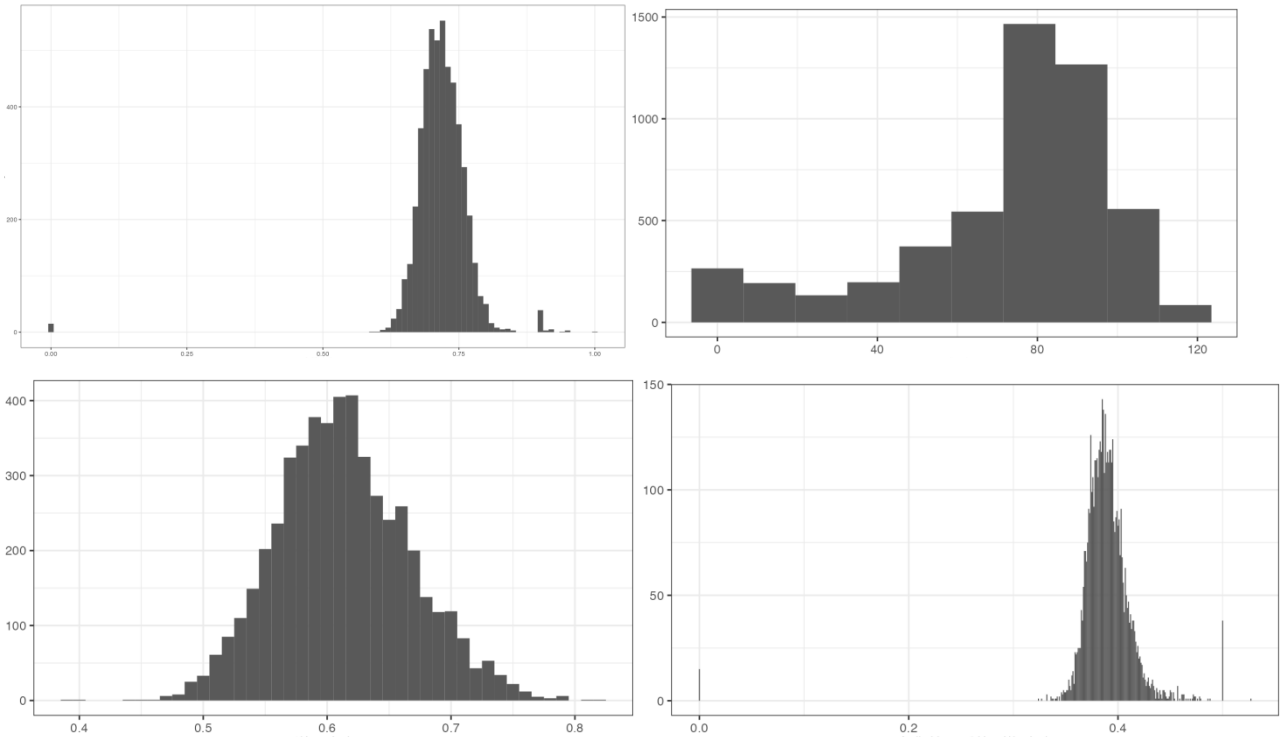


Figure B1.1 – 1.4. 案情、刑期、说理、log1p 处理后的刑期（从左往右）

下面是词云图，用来发现切割是否合理：

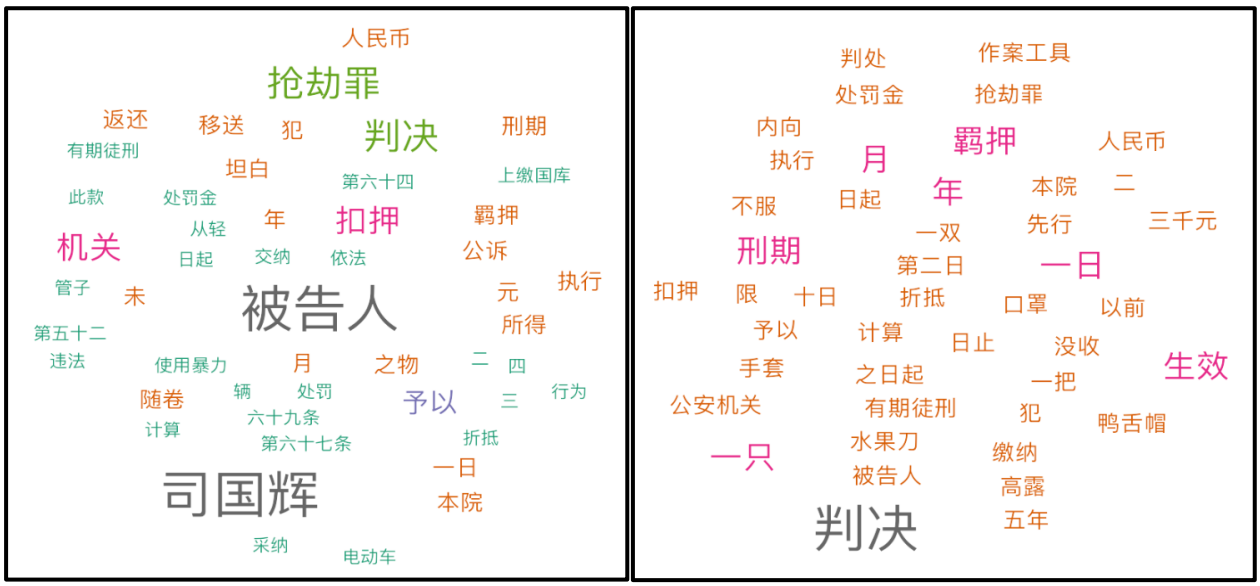


Figure B2.1 – 2.2. 词云图

下面是机器学习模型的拟合效果图：

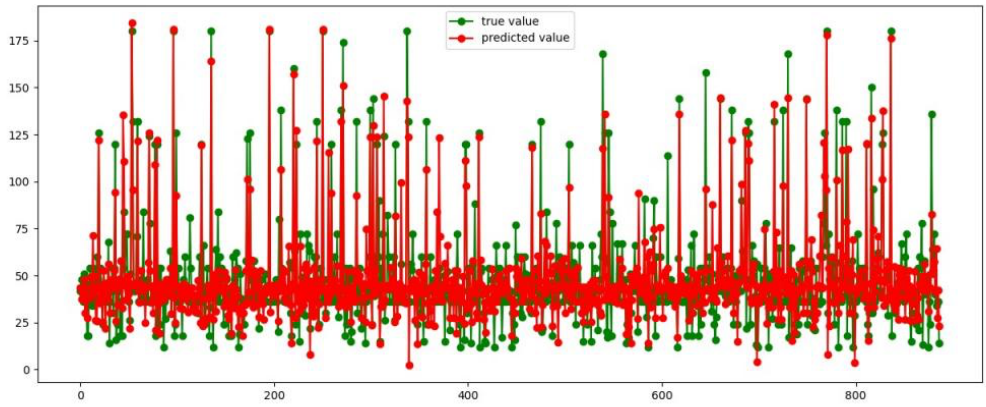


Figure B3.1. 线性回归拟合效果

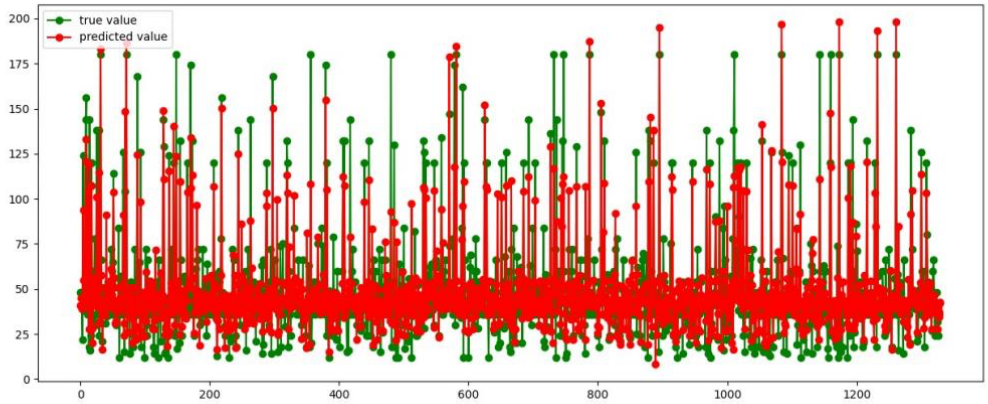


Figure B3.2. 多层感知机拟合效果

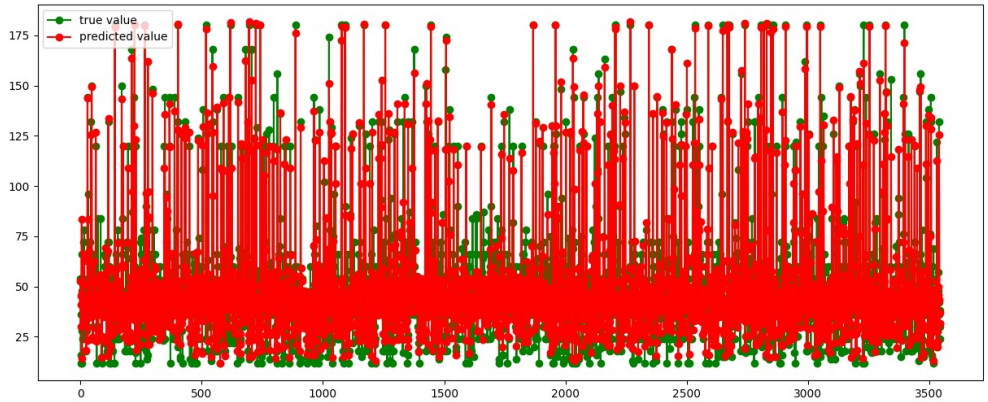


Figure B3.3. 集成学习拟合效果

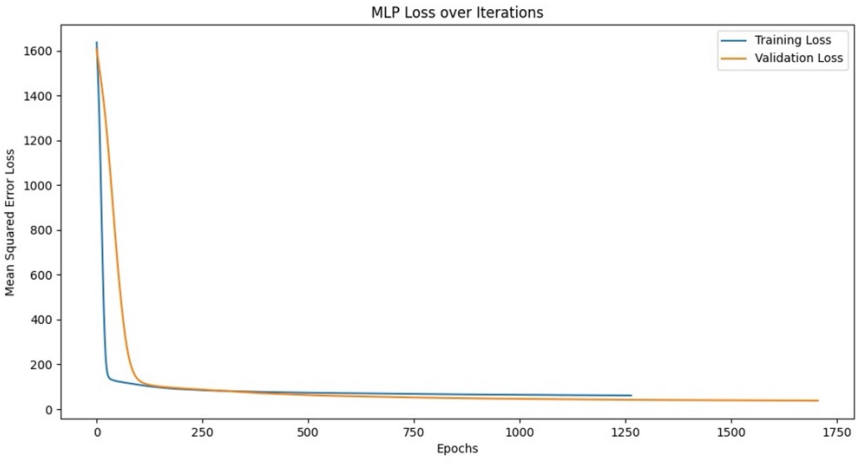


Figure B4. 多层感知机损失曲线

Tasks	Law Articles		Charges		Prison Terms
Evaluation Metrics	F_{micro}	F_{macro}	F_{micro}	F_{macro}	Score
nevermore	**0.958**	**0.781**	**0.962**	**0.836**	77.57
jiachx	0.952	0.748	0.958	0.815	69.64
xlzhang	0.952	0.760	0.958	0.811	69.64
HFL	0.953	0.769	0.958	0.811	**77.70**
大师兄	0.945	0.757	0.951	0.816	73.16
安徽高院类案指引研发团队	0.946	0.756	0.950	0.803	72.24
AI_judge	0.952	0.766	0.956	0.811	–
只看看不说话	0.948	0.738	0.954	0.801	77.54
DG	0.945	0.717	0.949	0.755	76.18
SXU_AILAW	0.940	0.728	0.950	0.791	76.49
中电28所联合部落	0.934	0.740	0.937	0.772	75.77

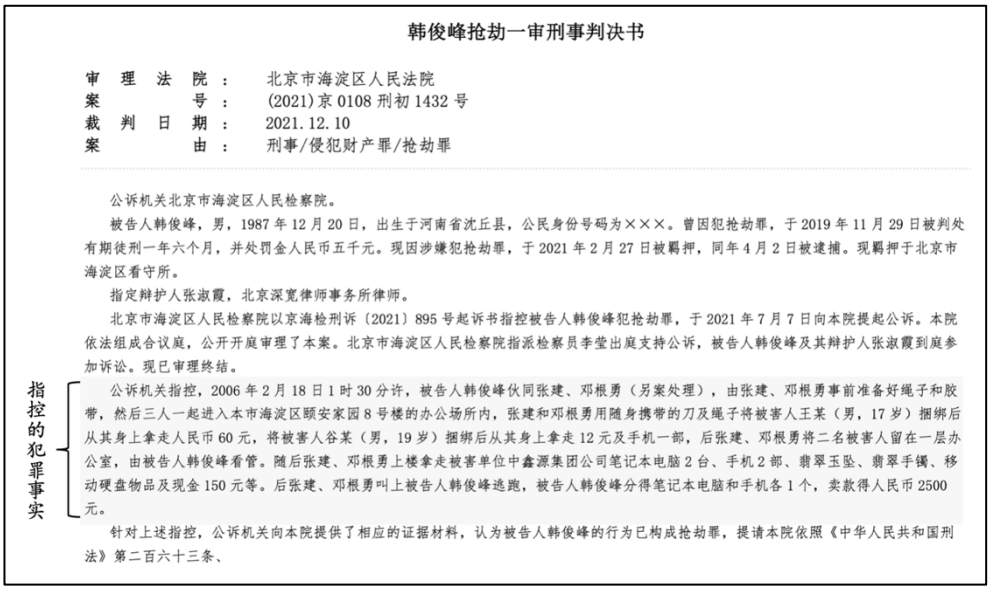
Table 1: Performance of participants on CAIL2018 .

Figure B5.1. CAIL2018 模型的准确度得分情况

Figure B5.2. 本研究模型的得分情况（多层感知机和集成学习数据）

APPENDIX C

下面是文书切割方法的展示：



韩俊峰抢劫一审刑事判决书

审 理 法 院 ： 北京市海淀区人民法院
案 号 ： (2021)京 0108 刑初 1432 号
裁 判 日 期 ： 2021.12.10
案 由 ： 刑事/侵犯财产罪/抢劫罪

公诉机关北京市海淀区人民检察院。

被告人韩俊峰，男，1987 年 12 月 20 日，出生于河南省沈丘县，公民身份号码为×××。曾因犯抢劫罪，于 2019 年 11 月 29 日被判处有期徒刑一年六个月，并处罚金人民币五千元。现因涉嫌犯抢劫罪，于 2021 年 2 月 27 日被羁押，同年 4 月 2 日被逮捕。现羁押于北京市海淀区看守所。

指定辩护人张淑霞，北京深宽律师事务所律师。

北京市海淀区人民检察院以京海检刑诉〔2021〕895 号起诉书指控被告人韩俊峰犯抢劫罪，于 2021 年 7 月 7 日向本院提起公诉。本院依法组成合议庭，公开开庭审理了本案。北京市海淀区人民检察院指派检察员李莹出庭支持公诉，被告人韩俊峰及其辩护人张淑霞到庭参加诉讼。现已审理终结。

指控的犯罪事实

公诉机关指控，2006 年 2 月 18 日 1 时 30 分许，被告人韩俊峰伙同张建、邓根勇（另案处理），由张建、邓根勇事前准备好绳子和胶带，然后三人一起进入本市海淀区颐安家园 8 号楼的办公场所内，张建和邓根勇用随身携带的刀及绳子将被害人王某（男，17 岁）捆绑后从其身上拿走人民币 60 元，将被害人谷某（男，19 岁）捆绑后从其身上拿走 12 元及手机一部，后张建、邓根勇将二名被害人留在一层办公室，由被告人韩俊峰看管。随后张建、邓根勇上楼拿走被害单位中鑫源集团公司笔记本电脑 2 台、手机 2 部、翡翠玉坠、翡翠手镯、移动硬盘物品及现金 150 元等。后张建、邓根勇叫上被告人韩俊峰逃跑，被告人韩俊峰分得笔记本电脑和手机各 1 个，卖款得人民币 2500 元。

针对上述指控，公诉机关向本院提供了相应的证据材料，认为被告人韩俊峰的行为已构成抢劫罪，提请本院依照《中华人民共和国刑法》第二百六十三条、

第二十五条第一款之规定，对被告人韩俊峰定罪处罚。

被告人韩俊峰对起诉书的指控事实和指控罪名没有提出异议。其辩护人发表的辩护意见为，被告人韩俊峰到案后如实供述自己的罪行，有立功情节，且本罪系漏罪，提请法庭对其从轻处罚。

查明的事实

经审理查明，2006 年 2 月 18 日 1 时 30 分许，被告人韩俊峰伙同他人一起进入本市海淀区颐安家园 8 号楼的办公场所内，将被害人王某（男，17 岁）捆绑后从其身上拿走人民币 60 元，将被害人谷某（男，19 岁）捆绑后从其身上拿走 12 元及手机，后由韩俊峰看管二被害人，另外二人上楼拿走被害单位中鑫源集团公司笔记本电脑 2 台、手机 2 部、翡翠玉坠、翡翠手镯、移动硬盘物品及现金 150 元等。

2006 年 2 月 18 日，被害人及被害单位报案，同日立案。2021 年 2 月 27 日，被告人韩俊峰在河南监狱门口被抓获归案，到案后如实供述了犯罪事实，并自愿认罪认罚。

另查明，被告人韩俊峰因犯抢劫罪于 2019 年 8 月 30 日被羁押。2019 年 11 月 29 日，河南省沈丘县人民法院以（2019）豫 1624 刑初 679 号刑事判决书判决被告人韩俊峰犯抢劫罪，判处有期徒刑一年六个月，罚金人民币五千元，2021 年 2 月 27 日释放，当日被北京市公安局海淀分局刑事拘留。张建因涉嫌犯抢劫罪，2021 年 4 月 14 日被刑事拘留，同年 5 月 21 日被逮捕。邓根勇因涉嫌犯抢劫罪，于 2021 年 4 月 14 日被刑事拘留，同年 5 月 26 日被逮捕。2021 年 11 月 16 日，北京市海淀区人民检察院以犯罪事实不清，证据不足为由对张建、邓根勇二人作出不起诉的决定。

上述事实，被告人韩俊峰及其辩护人在开庭审理过程中亦无异议，并有被告人韩俊峰的供述，被害人王某、谷某的陈述，证人赵某、焦某的证言，被抢物品清单，营业执照复印件，现场勘验笔录，现场平面示意图，现场照片，现场提取痕迹、物证登记表，鉴定书，受案登记表，立案决定书，接受刑事案件登记表、回执，不起诉决定书，到案经过，传唤证，工作说明，刑事判决书，刑事裁定书，释放证明，缴费通知书，身份证明材料等证据证实，足以认定。

法院说理部分

本院认为，被告人韩俊峰伙同他人抢劫他人和单位财物，其行为已构成抢劫罪，应予惩处。北京市海淀区人民检察院指控被告人韩俊峰犯抢劫罪的事实清楚，证据确凿，指控罪名成立。被告人曾因犯抢劫罪被判处刑罚，在判决宣告以后刑罚执行完毕以前发现在判决宣告以前还有本罪没有判决，应依法与前罪并罚。针对公诉人及辩护人关于被告人韩俊峰具有立功情节的意见，因未能查证属实，本院不予认定。鉴于被告人韩俊峰到案后如实供述自己的罪行，本院依法对其从轻处罚。辩护人的相关辩护意见，本院酌予采纳。依照《中华人民共和国刑法》第二百六十三条、第六十九条、第七十条、第六十七条第三款、第五十三条第一款、第六十四条及《中华人民共和国刑事诉讼法》第十五条之规定，

刑期提取的位置

判决如下：

一、被告人韩俊峰犯抢劫罪，判处有期徒刑四年六个月，罚金人民币二万元；与（2019）豫 1624 刑初 679 号刑事判决书所判处的有期徒刑一年六个月，罚金人民币五千元并罚，决定执行五年六个月，罚金人民币二万五千元。

（刑期从判决执行之日起计算；判决执行以前先行羁押的，羁押一日折抵刑期一日，即自 2019 年 8 月 30 日起至 2025 年 1 月 27 日止。罚金限自本判决生效后十日内缴纳。）

二、责令被告人韩俊峰向被害人王某退赔人民币六十元，向被害人谷某退赔人民币十二元，向被害人单位北京中鑫源房地产开发集团有限公司退赔人民币一百五十元。

如不服本判决，可在接到本判决书的第二日起十日内，通过本院或者直接向北京市第一中级人民法院提出上诉。书面上诉的，应提交上诉状正本一份，副本二份。

审判长：段倩倩
人 民 陪 审 员：姜毅
人 民 陪 审 员：闵增云
二〇二一年十二月十日
书记员：耿雁