Explainability Methods for Legal Judgment Prediction in Switzerland

University of Bern

Xinwei Li(21-117-007)

co-lab link:https://colab.research.google.com/drive/14dDmJHFzeJ1eQQDIJGEiBoP0Bk1yOruF#scrollTo=tHgYNQX7DExN

1.Introduction

Legal judgement prediction refers to the prediction of the conviction in a legal case based on the description of the facts of the case. For expectation models, we really want to find a sensible explainability model to make sense of the forecast results and make them more sound. Thus, we need to investigate 'explanatory' models and attribute predictions to them to arrive at a reasonable explanation.

Hence,my study uses this example to explain the internal logic of the results and to provide a theoretical basis for the prediction results.

2.Literature Review

My study reviews the current state of research on interpretable models, the application of interpretable models to legal judgments, and the relevant literature on the SHAP and LIME models. The literature is reviewed from both theoretical and empirical perspectives in the literature review is conducted.

2.1 The current research on interpretable models

Deep learning has been applied directly to the research process as a 'black box', when there is no reasonable explanation neither for results nor how to determine the best parameters to use. There are currently two main areas of research on interpretability:

Firstly, the deep learning model itself is studied by adjusting the internal parameters of the model and analysing the results obtained from the system to determine the impact of the parameters on the results. By adding perturbations to the input variables and probing the representation vector to assess the importance of the different variables in the system, the basis for the decisions made by the system is inferred. Secondly, it is straightforward to construct models that are inherently interpretable, only in learning more structured and interpretable models.

Ribeiro (1998) proposes a locally understandable model-independent interpretation technique (LIME), which first obtains feedback data on the response of a deep model by detecting input perturbations, he then constructs a local linear model from this data and uses this model as a simplified proxy for a deep model for a particular input value. Riberio(1998) argues that this approach can be useful in identifying the input regions that have the greatest impact on decisions for various types of models and problem domains.Ribeiro,Singh&Guestrin(2016) research problem of interpretable interpreters.The definition of interpretation given by Montavon (2017), denotes the mapping of abstract concepts into a human-understandable domain.Lipton (2018) first analyzed the interpretation of interpretability in deep learning models from four perspectives, such as trustworthiness, causal relevance, and information availability The first analysis of the meaning of interpretability in deep learning models from four perspectives, including trustworthiness, causal correlation, and information availability, pointed out that "decisions made by interpretable deep learning models tend to gain higher trust, even when the trained model diverges from the actual situation; interpretability can help humans understand the characteristics of deep learning systems and infer the relationships of variables within the system; it can help deep learning models easily cope with the inconsistency of sample distribution It can help deep learning models to cope with the inconsistency of sample distribution and migratory learning; interpretability can provide people with auxiliary information, even if it does not clarify the inner workings of the model, interpretability can also provide decision makers with a basis for judgment. Cheng et al.(2020) reviewed the history of deep learning interpretability research from three aspects: the origin of deep learning interpretability research, the research exploration period and the model building period. The research on interpretable deep learning model construction was dissected from four aspects: model agents, logical inference, network node association analysis, and improvement of traditional machine learning models.

2.2 Applications of interpretable models

In recent years, the use of natural language processing for the prediction of legal decisions has become an important direction of research method. Amir (2019) used the XGBoost method to predict accident rates for the Chicago Metropolitan Expressway and used SHAP to interpret the results and analyse the importance of individual features. Mangalath (2020) used an experimental database to predict the failure patterns of reinforced concrete columns and shear strength modes, and he used the SHAP method to rank the input variables to identify failure modes. Yin et al. (2021) argue that there is a long history of research in legal decision prediction, with early work focusing on analysing legal cases in specific scenarios using mathematical and statistical computational methods. And while machine learning methods have been increasingly applied in this area in recent years with the rapid development of artificial intelligence technology, their models are not more explicit in predicting outcome interpretations. This has prompted existing lawful judgment expectations, albeit profoundly precise, being not able to give sensible illustrative notes on the purposes behind the forecast results. They likewise need without a hypothetical help reason for direction.

2.3 Review of the literature

According to previous studies, the construction of interpretable models has important theoretical significance and practical needs. Through interpretable analysis, researchers can clearly know the feature engineering and understand the core functions of machine learning models more.It permits users to know obviously the way of behaving and limits of the framework ,and to see plainly the intelligent thinking behind every choice. Involving legitimate decisions for instance, we concentrate on the issue of deficient thinking in expectation results and give an examination of the impacts on the aftereffects of non-judgment expectations.

Based on the research objectives, the following research question was asked.

RQ1: Can we find the models relying on spurious correlations, rather than on sensible content, to make the decisions?

RQ2: What insight on the prediction of the legal judgment outcome can we draw using the explanations?

3.Data Discription

We recently presented a dataset for legal judgment prediction including 85k Swiss Federal Supreme Court decisions. Although we achieved up to 70% Macro- F1 Score, the models still work as black boxes and are thus not interpretable.In the simulation of these two models, I use the French data as an example.

Load packges

```
gpu_info = !nvidia-smi
gpu_info = '\n'.join(gpu_info)
if gpu_info.find('failed') >= 0:
  print('Not connected to a GPU')
else:
  print(gpu_info)
```

```
 pip install shap
```

```
pip install lime
```

```
import sklearn
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.model_selection import train_test_split
import numpy as np
import shap
import pandas as pd
from nltk.corpus import stopwords
from sklearn.feature_extraction.text import TfidfTransformer, CountVectorizer
from sklearn.model_selection import GridSearchCV
from sklearn.linear_model import SGDClassifier
from sklearn import metrics
import numpy as np
import os, tqdm, json

shap.initjs()
```

Load dataset

```
label2id = {
    "approval": 0,
    "dismissal": 1
}
```

```
from google.colab import drive
drive.mount('/content/gdrive')
```

```
Mounted at /content/gdrive
```

```
pwd
```

```
'/content'
```

```
data_dict= {}
splits = ["train", "val", "test"]
for split in splits:
    data_dict[split] = pd.read_csv(f"./fr/{split}.csv")
    data_dict[split]["label"] = data_dict[split]["label"].map(label2id)
```

```
text_dict = {}
target_dict = {}
for split in splits:
    text_dict[split] = data_dict[split]["text"].values
    target_dict[split] = data_dict[split]["label"].values
```

```
target_dict["train"]
```

```
array([1, 0, 0, ..., 1, 1, 0])
```

Train model

```
import nltk
nltk.download('stopwords')
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
True
```

```
stop_words = stopwords.words('french')
```

```
#vectorizer = CountVectorizer(stop_words=stop_words, ngram_range=(1, 3))
vectorizer = TfidfVectorizer(stop_words=stop_words)
# For the real model we should use the commented lines of code, but because of the
X_train = vectorizer.fit_transform(text_dict["train"]).toarray() # sparse also work
X_val = vectorizer.transform(text_dict["val"]).toarray()
X_test = vectorizer.transform(text_dict["test"]).toarray()
```

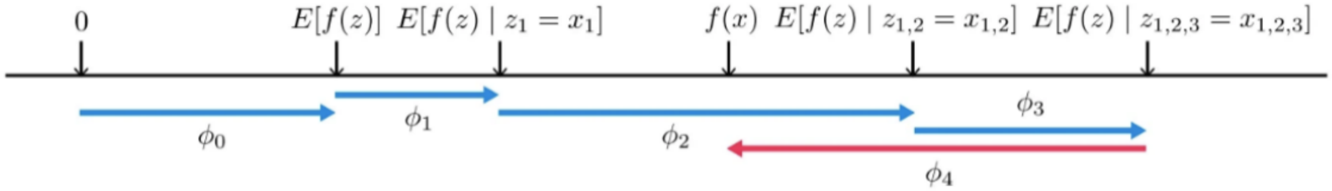
```
y_train = target_dict["train"]
y_val = target_dict["val"]
y_test = target_dict["test"]
```

```
seed = 42
model = sklearn.linear_model.LogisticRegression(penalty="l2", C=0.1)
#model = sklearn.linear_model.SGDClassifier(early_stopping=True, learning_rate='ada
#                                             eta0=1e-4, validation_fraction=0.1, m
#                                            class_weight='balanced', random_state=
# For the real model we should use the commented lines of code, but because of the

model.fit(X_train, y_train)
```

```
LogisticRegression(C=0.1)
```

4.Model 1 : SHAP

Shap was proposed by Professor Loyd Shapley to solve the problem of contribution and benefit distribution in cooperative games. The method is a post-hoc interpretation of the model, in which the core of the interpretation is to calculate the Shap Value of each of the characteristic variables. N-person cooperation in which individual members contribute should also have different benefit distributions. Often, getting the trade-off between accuracy and interpretability right can be a difficult balancing act, but SHAP value can provide both. The impact of having a particular value for a given characteristic is explained and compared with the predictions we would have made if that characteristic had some baseline value. From another perspective, the SHAP value is the value assigned to each feature through that sample. to get the optimal prediction model and the optimal set of indicators it requires to explore the impact of each feature on the outcome.To calculate the Shapley value, it is important to consider whether each feature is in the set S. As the number of features increases, the computational complexity increases , so an approximation algorithm implemented by Monte-Carlo sampling was proposed by Strumbelj et al.



Input: number of iterations M, sample points x, feature index, training data, training model;

Output: Shapley value of the nth feature;

Step1: Random sampling from the training data;

Step2: Randomly replace the features in sample x with those in z to get two new vectors;

Step 3: Calculate the marginal returns for each time and average them to get the Shapley value of the feature.

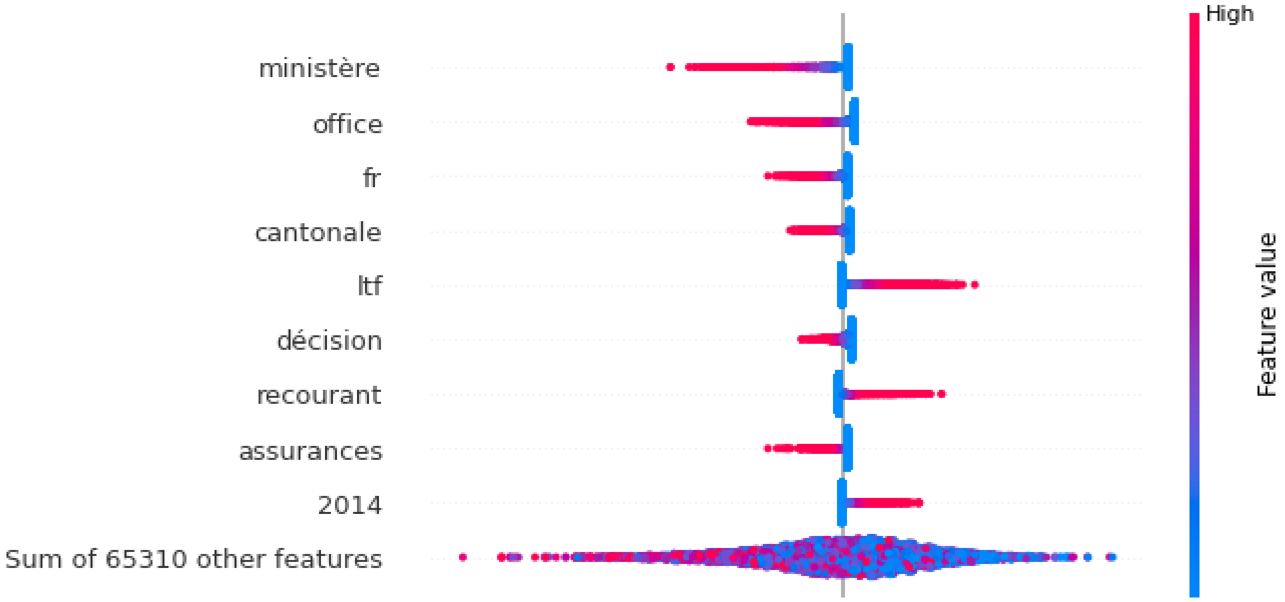
The dataset (with M characteristic variables and n samples), the original model f, and all predicted values of the original model f on the dataset are known. g is the model used in SHAP to explain f.

The figure above explains the reasons for the differences between the sample predictions and the sample means. Where the blue arrow indicates a positive SHAP value, meaning that the characteristic variable has a positive effect on the predicted value of the original model. The red arrow indicates a negative SHAP value, meaning that the eigenvariable has a negative effect on the predicted value of the original model. The presence of Z4 has a negative effect on f(x). The combined effect of the four characteristic variables of this sample X brought the original model prediction of this sample x from the mean to f(x).

```
explainer = shap.Explainer(model, X_train, feature_names=vectorizer.get_feature_nam
shap_values = explainer(X_test)
```

```
Function get_feature_names is deprecated; get_feature_names is deprecated in 1
```

```
shap.plots.beeswarm(shap_values)
```



Understanding of the outcomes:

As displayed in the figure over, the x-axis addresses Shapley Value, where every information point addresses a bunch of information, and various varieties address the size of the component.

For example, for the feature of 'cantonale', a larger value of this feature will have a positive impact on the model's interpretation as positive (guilty).

```
ind = 0 # index of the test sample
shap.plots.force(shap_values[ind])
```

```
ind = 1 # index of the test sample
shap.plots.force(shap_values[ind])
```

```
shap.summary_plot(shap_values)
```

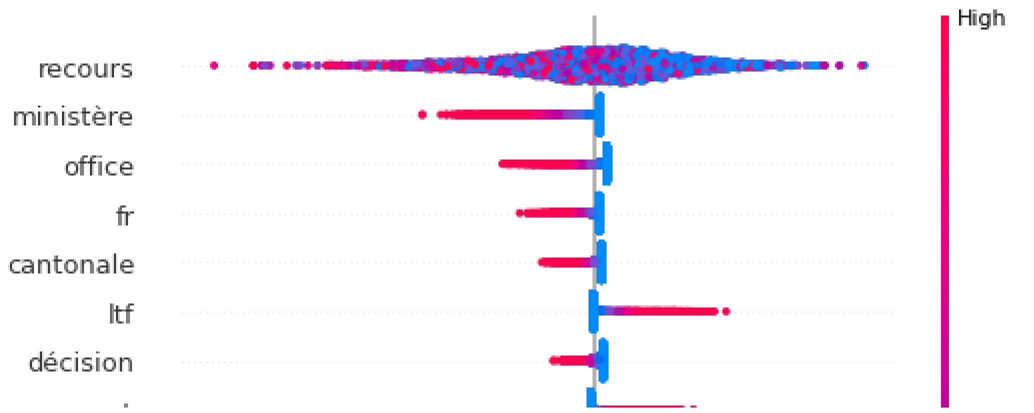


Figure above is the complete result.

5.Model 2: LIME

Local Interpretable Model-Agnostic Explanation (LIME) is an interpretable model based on a model agent, i.e.: a simple model is used to interpret a complex model. The predictions of the model are interpreted by perturbing the input samples. The analytical idea of the model is illustrated in the following diagram.

The specific analysis steps of the LIME model are as follows.

Step 1:Training on the entire data;

Step 2:Select a sample with interpretable N perturbations to generate a local sample;

Step 3:Define a function for the similarity measure and select the k features that best describe the output of the complex model;

Step 4: Use the complex model to predict them;

Step 5: Fit a simple model to the sampled samples to obtain the weights of the model.

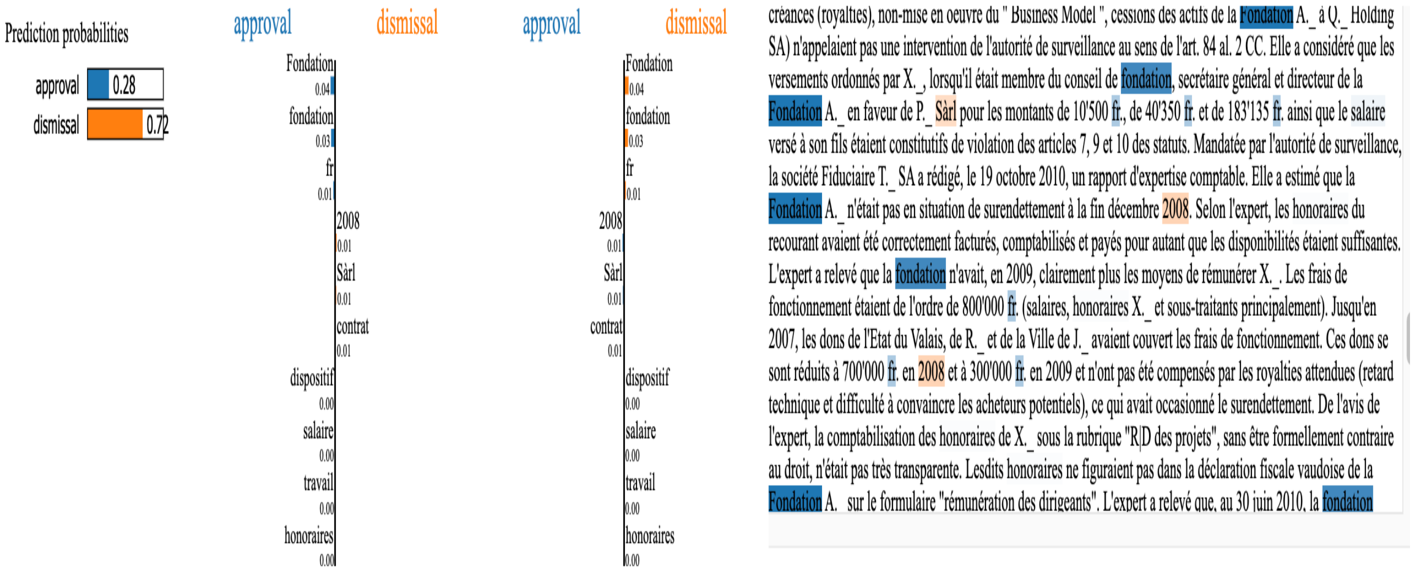
```
from lime import lime_text
explainer = lime_text.LimeTextExplainer(class_names=["approval", "dismissal"])

# print("Question: \n", val_df["question_text"][idx])
# print("Probability (Insincere) =", c.predict_proba([val_df["question_text"][idx]]
# print("Probability (Sincere) =", c.predict_proba([val_df["question_text"][idx]])[
# print("True Class is:", class_names[val_df["target"][idx]])


#weights = {0: weights}
from sklearn.pipeline import Pipeline, make_pipeline
pipeline = make_pipeline(vectorizer, model)


idx= 0
exp = explainer.explain_instance(text_dict["test"][0], pipeline.predict_proba, num_


exp.as_list(label=1)
```

```
[('Fondation', -0.03763943680848953),
 ('fondation', -0.03309152272301504),
 ('fr', -0.014799521297092886),
 ('2008', 0.013097728338323273),
 ('Sàrl', 0.010312740337403757),
 ('contrat', 0.006716946258752594 6),
 ('dispositif', -0.0031840084532425944),
 ('salaire', -0.0026986884250304634),
 ('travail', -0.0016927636350414976),
 ('honoraires', -0.0007668849590715892)]
```

```
exp.as_list(label=0)
```

```
[('Fondation', 0.03763943680848954),
 ('fondation', 0.03309152272301504),
 ('fr', 0.014799521297092885),
 ('2008', -0.013097728338323273),
 ('Sàrl', -0.010312740337403764),
 ('contrat', -0.006716946258752596),
 ('dispositif', 0.003184008453242598),
 ('salaire', 0.0026986884250304664),
 ('travail', 0.0016927636350414995),
 ('honoraires', 0.0007668849590715892)]
```

```
exp.show_in_notebook()
```



créances (royalties), non-mise en oeuvre du " Business Model ", cessions des actifs de la Fondation A._ à Q._ Holding SA) n'appelaient pas une intervention de l'autorité de surveillance au sens de l'art. 84 al. 2 CC. Elle a considéré que les versements ordonnés par X._, lorsqu'il était membre du conseil de fondation, secrétaire général et directeur de la Fondation A._ en faveur de P._ Sàrl pour les montants de 10'500 fr., de 40'350 fr. et de 183'135 fr. ainsi que le salaire versé à son fils étaient constitutifs de violation des articles 7, 9 et 10 des statuts. Mandatée par l'autorité de surveillance, la société Fiduciaire T._ SA a rédigé, le 19 octobre 2010, un rapport d'expertise comptable. Elle a estimé que la Fondation A._ n'était pas en situation de surendettement à la fin décembre 2008. Selon l'expert, les honoraires du recourant avaient été correctement facturés, comptabilisés et payés pour autant que les disponibilités étaient suffisantes. L'expert a relevé que la fondation n'avait, en 2009, clairement plus les moyens de rémunérer X._. Les frais de fonctionnement étaient de l'ordre de 800'000 fr. (salaires, honoraires X._ et sous-traitants principalement). Jusqu'en 2007, les dons de l'Etat du Valais, de R._ et de la Ville de J._ avaient couvert les frais de fonctionnement. Ces dons se sont réduits à 700'000 fr. en 2008 et à 300'000 fr. en 2009 et n'ont pas été compensés par les royalties attendues (retard technique et difficulté à convaincre les acheteurs potentiels), ce qui avait occasionné le surendettement. De l'avis de l'expert, la comptabilisation des honoraires de X._ sous la rubrique "R|D des projets", sans être formellement contraire au droit, n'était pas très transparente. Lesdits honoraires ne figuraient pas dans la déclaration fiscale vaudoise de la Fondation A._ sur le formulaire "rémunération des dirigeants". L'expert a relevé que, au 30 juin 2010, la fondation

Understanding of the outcomes:

From the results in the above figure, we can see the correlation between the frequency of the corresponding words and the judgment results. The bar graph represents the importance given to the most relevant word, which is also highlighted in the text. Colors indicate to which category the word contributes.Examples of specific text are highlighted in the image to the right.

6.Conclusions and limitations

The following conclusions can be drawn from the theoretical mechanism study and the experimental analysis.

For the SHAP model, through experiments, we can find the following advantages and limitations：

1. the calculation of SHAP value is assigned to different features ;
2. comparative analysis can be carried out, where the SHAP value of the same feature is compared between different samples. On the other hand, the SHAP model also has the following disadvantages:
3. the computation is very time-consuming, with exponential levels of complexity;
4. the SHAP may be misleading, since what is calculated is not the result obtained by training the model after removing this feature.

For the LIME model, the LIME algorithm works relatively well. Subject classification, part-of-speech tagging, etc. At the same time, because the assumption of LIME is model-independent, it requires less data and has wide applicability. However, the speed of the LIME algorithm is relatively slow, and the requirements for the GPU are relatively high.