Legal Case Outcome Classification

Sida Ye, Xingye Zhang, Muhe Xie

NYU Center for Data Science

Introduction

A sample of 2526 cases has been hand-coded for meaning, like pro-plaintiff or pro-defendant, pro-business or pro-environment, pro-criminal defendant rights or pro-prosecutor, etc., in 16 politically salient legal areas (11th Abrogation, Abortion, ADA, Affirmative Action, Campaign Finance, Capital Punishment, EPA, FCC, First Amend, Homosexual Rights, NEPA, NLRB, Obscenity, Piercing Corp Veil Sex Discrimination, Title 7). By using only the text and other variables, we want to predict the binary outcome decision of liberal and conservative for cases in different legal areas. Finding important word features which have impact on liberal or conservative decision in each individual legal fields is another task.

Data Description

The major data we worked on is the circuit case data and the ngram data. Circuit case data of different legal fields contains the panel vote which is directly related with our target label(liberal or conservative). Ngram data composed of 1-8 gram data generated for case text.

After processing procedure, the final dataset we works on contains label data (0 or 1 stands for liberal or conservative), case field (Gay Rights, Abortion, Gender Discrimination, etc), and ngram data (a dictionary contains the ids which represent specific ngram items).

	caseid	citation	panelvote	issue	field	n_gram
0	X42KB3	475 F2d 65	1.0	16	Obscenity	{'1207777881': 1, '2045835947': 1, '1090701101...
1	XE6R3S	342 F3d 1233	0.0	5	Capital Punishment (vote against)	{'1713155254': 1, '618852088': 1, '1713155252':...
2	X6B818	348 F3d 537	1.0	13	Title 7	{'2094016335': 1, '2294056281': 1, '187420851?...
3	XN5AGRQNB5G0	136 F3d 276	0.0	11	Sex Discrimination	{'719834498': 1, '719834499': 1, '719834496': ...
4	X36AM3	181 F3d 1342	0.0	13	Title 7	{'1359643443': 1, '1359643442': 1, '1359643441...
5	XST0R8003	306 F3d 203	0.0	22	11th Abrogation	{'462249592': 1, '807063769': 2, '20750631': 1...
6	X3P948	655 F2d 848	0.0	2	Abortion (vote pro-choice)	{'3159879990': 1, '143889': 1, '186185077?: 1,...
7	X3519O	120 F3d 476	1.0	11	Sex Discrimination	{'2018877588': 1, '29105507': 1, '2018877589':...
8	X50LNA	246 F3d 1083	0.0	3	ADA (vote for Plaintiff)	{'2187097428': 1, '2187097429': 1, '2179348040...
9	X36C4Q	188 F3d 932	0.0	3	ADA (vote for Plaintiff)	{'995860351': 1, '1445801826': 1, '667238026':...
10	X40T2C	163 F3d 1012	0.0	23	NLRB - Chevron/Liberal-conservative	{'2144747202': 1, '250918325': 1, '1501777563':...
11	X40TAO	163 F3d 137	1.0	21	FCC - Chevron/Liberal-conservative	{'365652655': 2, '280770585': 1, '280770586': ...
12	XARF3I	368 F3d 123	0.0	11	Sex Discrimination	{'2307961897': 1, '199807126': 5, '2218784089'...

Figure 1: Basic dataset screen shot(panelvote is the label, 0 represents conservative, 1 represents liberal)

1	1871036902	1819593224	1018927461	815663622	460321111	365349837	62632255
2	0	3	0	0	3	10	9
3	2	2	0	12	36	2	0
4	2	0	0	6	14	0	0
5	0	0	0	4	6	20	6
6	0	5	0	24	2	0	0
7	2	0	0	6	0	6	4
8	0	4	0	9	4	0	3
9	0	0	0	2	14	0	0
10	0	4	6	0	12	0	8
11	3	2	0	0	0	0	0
12	0	0	3	0	0	0	2
13	0	2	0	7	10	7	0
14	0	0	0	2	4	5	0
15	0	0	0	0	0	3	15
16	0	0	0	0	0	0	0
17	4	2	6	0	0	2	2
18	4	2	6	0	0	2	2
19	0	0	0	4	8	0	5

Figure 2: training feature data screen shot

Feature Selection

We select the features from each legal fields by choosing the random forest feature importance which are grater than 0. We will talk about one example about Campaign Finance below since the process for all legal fields are similar.

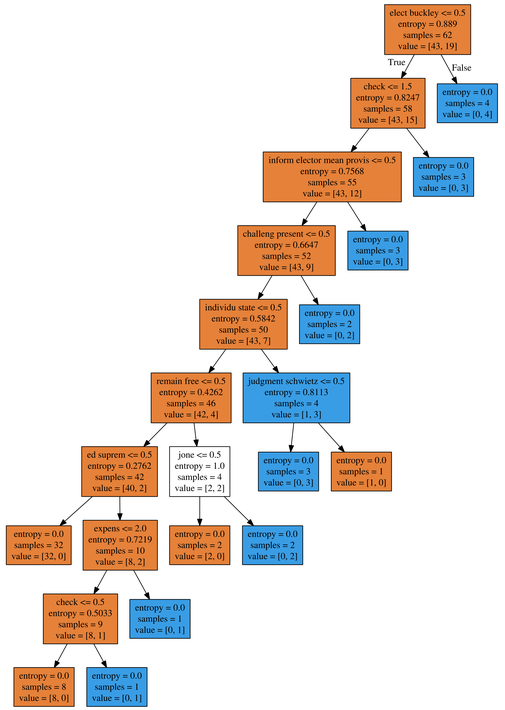

Figure 3: Tree plot for Campaign Finance field

Modeling and Results

For modeling part, we tried Logistic Regression, Linear SVM and RandomForest. For each model, we will use cross validation to choose the best parameters which decide the regularization strength. First we choose the parameter in a range of different magnitudes. When we get an optimal range, we then zoom in to get the best parameters. After comparing the AUC score, Logistic Regression always has the highest AUC score amoung all models. Also, It is flexible, interpretable, and is less prone to overfitting. To guard against overfitting and to ensure that our model would generalize well, we used L1 regularization. Thus, we chose to use Logistic Regression model. The following AUC plots shows results about Campaign Finance legal field.

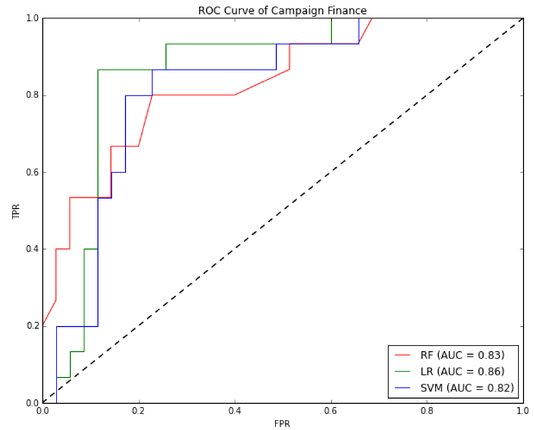

Figure 4: AUC plot for three different models

Summary

By applying Logistic Regression algorithm with tf-idf features, we find better performance on predicting legal case outcomes with AUC value equals 0.86.

We also found some important meaningful word features, which contribute significantly to prediction result. For example, in Campaign Finance legal, we have following words which are important.

The term 'Advertisement influence outcome vote' has power on predicting liberal outcome. Since election needs advertisement and candidates will raise funds to buy advertisement. So, in this case, extend the limitation of fund raising is a liberal outcome.

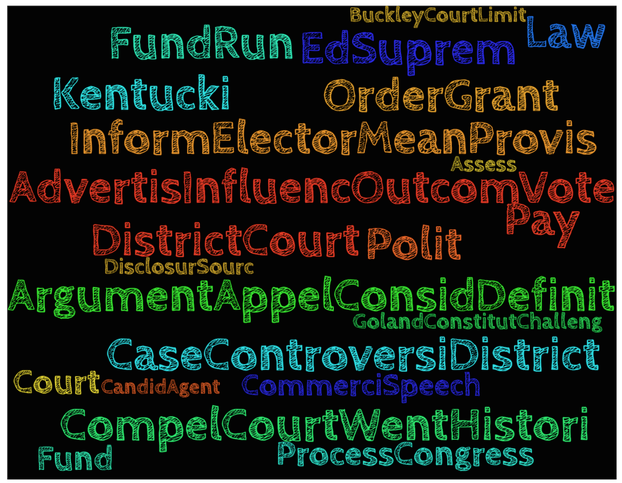

Figure 5: Some meaningful tokens in Campaign Finance field

Partial Reference

- Cass R. Sunstein, Are Judges Political? An Empirical Analysis of the Federal Judiciary.
- TF-IDF document `http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html`.

Acknowledgments

First and foremost, we want to thank our advisor Daniel Chen for providing extraordinary expertise, patience and having faith in us. Without him we would not have completed our project. Also, we would like to express our appreciation to David Rosenberg for his understanding, invaluable help and inspiration in our study of machine learning.

Advisors: Dr. Chen , Prof. Rosenberg