DS-GA 1003 Final Project Report

Case Decision Prediction from Circuit Courts

Supervisor: Dr.Daniel Chen, David Rosenberg
Team Member: Muhe Xie, Sida Ye, Xingye Zhang
Team NetID: mx419 sy1743 xz601

1. Introduction

A sample of 2526 cases have 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 (Capital Punishment, Criminal Appeals, Campaign Finance, Affirmative Action, Gay Rights, Abortion, Gender Discrimination, Sexual Harassment, Title VII, Segregation, Obscenity, Establishment Clause, Americans with Disabilities Act, Piercing the Corporate Veil, Takings, National Labor Review Board, Environmental Protection Act, National Environmental Policy Act, Federalism, First Amendment, Punitive Damages, Standing, Eleventh Amendment, Federal Communications Commission, and Contracts). 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.

2. Data Description and Processing Methodology

a. Label Creation

First, we need to build a binary target variable. Under Circuit-case folders, there are several files which belongs to different legal fields. Each file contains around 200 features. Some are related to case such as date, legal field and panel vote etc. Some are about judges' characteristics. We will need 'panel-vote', 'x_dem', 'x_repub' and a law documentation as a guideline to build the binary target variable. 'Panel_vote' has four levels, which are 0,1,2,3. It means that the number of judges in the panel vote for one decision. It might be either liberal or conservative. By the guideline from textbook "Are Judges Political?", we assume that if panel vote has positive correlation with 'X_dem', which the percentage of judges are Democratic, the panel vote represents liberal. If it has negative correlation with 'X_dem', then the vote will be coded as conservative. Under positive correlation situation, we will consider panel vote value with 2 and 3 as liberal. Then based on these data and the documentation, we will generate a binary variable as our target variable. It will be binary variable with 0 and 1. 0 represents conservative and 1 represents liberal.

b. Data merge

We want to have a dataframe which will contain caseid, n-gram data and panel vote(label) for different legal fields.

We have 100Votelevel_touse.dta, which contains case_id (start with X) and citation. We also have sunstein_data_for_updating.csv, which contains citation and issue, which represent legal fields. After merging these two files on citation, we got 2526 records with case_id, citation and issue in columns. By citation and issue, we then get dataframe with case_id, citation, issue and its corresponding panel vote from part a.

In addition to these files, we also have Vocab_map_text and Docvec_text file. Vocab_map_text contains dictionary with individual n-gram as key and unique consecutive id as value. Docvec_text contains all cases data. It has case id and its corresponding n-gram data. The next step we conduct is to find n-gram data from Docvec_text file. We get case id from merged dataset as our target match id list. By applying bash command and Python program, we find matched caseid with its corresponding n-gram in Docvec_text file. After this step, we have two datasets. One dataset has case id and panel vote. Another dataset has cased id and n-gram data. Then, we merged these two datasets into one based on caseid. Meanwhile, we drop duplicate observations in this dataset. Based on each legal field, we create 16 files with caseid, n-gram, label, legal field and issue in columns as our datasets.

In docvect file, we already have the case_ID, n_gram_token_ID and its frequency. We need to merge target variables and n_gram data together by case_ID. The challenge here is that the size of data file for n-gram is extreme large which is around 100GB. Because we need to merge n-gram files with our target variables by case ID. Though we only need to consider 2526 cases since we only have 2526 hand-labeled cases. We still need to search all files. We have to think a way to overcome this challenge.

4. Feature Generation

We try two different methodologies to do prediction. One is to combine all cases together regardless of its legal fields and another one is to build models on each individual legal field. We also apply the same process about feature selection to build the combined model as well as each individual model on different legal fields. By comparing the model performance, we think building models on each legal fields works better than the whole dataset. We will talk about one example about Campaign Finance in the following paragraphs since the process for all legal fields are similar. In the following paragraphs, we will introduce Campaign Finance as an example since we apply the same method in other legal fields.

At this stage, we have the data from unigram to 8-gram for over 2526 cases. Each row would represent a single article and each column would be a n-gram phrase in the word feature matrix. However, if we just put everything we have into one data frame and then apply the model on it, that would be inefficient and inaccuracy since the number of features is so large that is much bigger than the number of instances. So we would like to do feature selection before modeling so that to avoid overfitting problems and to reduce the computing time.

At first, we try to delete word with count less than 2 and greater than 50 since there are so many words with count 1. However, in the later process, we find that the outcome is not what we except. So, we decide to remain all n-gram instead of deleting n-gram with frequency equals to 1.

The second steps we conducted was that to delete words which appear in no more than a threshold based on the number of observations we have in each individual legal field.

Third, we want to use random forest to do the feature selection. We applied this method at third step because in previous steps, the number of features is too large. If we use random forest in previous steps, the computational cost would be extreme large.

The last step is looking into real n-gram word is another step to reduce the number of features. It also can remove some redundant features with a positive correlation. Since our data include from unigram to 8-gram for each article, some of the phrases must contain the same words. i.e. 'violate the' (bigram) and 'violate the law' (trigram) Those kinds of phrases are very similar and we believe that removal of multicollinearity could improve the performance of the our models. We find consecutive n-gram ids and then find its key, which is the real word in Vocab_map_text.

Finally, the resulting number of word features in campaign finance was approximately 40 and our word feature matrix was 62x45, with rows representing an individual review and the columns composing the word dictionary.

Once we had created our word feature matrices, we randomly segmented the data of the two word feature matrices which are count and TF-IDF, into train and test datasets, with 80% of data allotted to the training dataset and 20% to our test dataset. Once we segmented our training and test data, we began testing various models on each word feature matrix.

5. Model Fitting

The first model we built is Logistic Regression. It is flexible, interpretable, and is less prone to overfitting. We then want to find top word feature with positive and negative coefficients. To guard against overfitting and to ensure that our model would generalize well, we used L1 regularization.

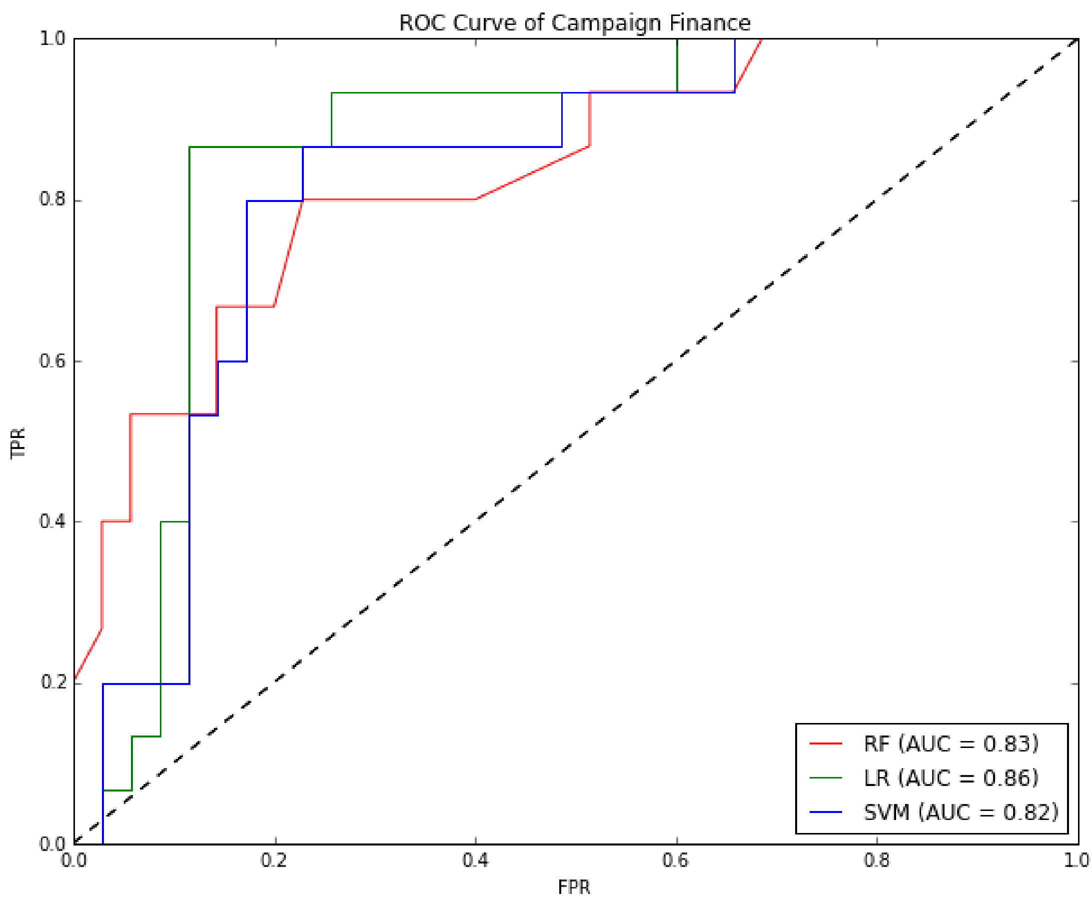
Then we will try SVM with linear kernel. Also it generates a sparse set of support vectors which are good for testing. And SVM always works very good in practice.

The last model we want to try is Random Forest, which can significantly decrease overfitting. Random Forest also allows for more complexity in the fitting process, we would have a better chance of accurately fitting the model to the data while minimizing overfitting due to the bagging feature to reduce the variance.

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.

6. Model Evaluation

We would like to use AUC. Meanwhile, F1-score might be another performance measure since these we do not have preference on either positive label or negative label.



Campaign Finance	AUC
Logistic regression	0.86
Random Forest	0.83
Linear SVM	0.82

Based on the AUC performance for logistic regression, random forest and svm, we finally choose logistic regression with tf-idf as our final model. In order to find some meaningful words, we choose the word features with larger positive coefficients as well as negative coefficients. After this process, we get a list of important words. Then we will match the word in Vocab_map_text file, it will return the real n-gram to us. This process takes around 5 to 6 hours. We put these important words in a table under Summary and Conclusion section.

7. Error Analysis

For example, in Campaign Finance legal field, we list a wrong predicted case. The true outcome is 1, but the predicted label is 0. The reason it occurs is that this observation has lots of non-zero vectors in features with negative coefficients . For example, ‘compel’, ‘congress’ and ‘fund run’ occurs in this observation. These word feature has negative coefficient.

In our most beginning approach, we delete all n-gram with frequency with 1. The AUC is not very good. After the error analysis, we decide to keep all n-gram and then apply random forest to do feature selection. After this improvement, the AUC is increased about 0.04 from 0.83 to 0.87.

In addition, some words occur in both positive and negative word list in a legal field. This is due to the size of the word features. For some legal fields, we only pick less than 60 word features. While selecting important we consider top 30 positive words and top 30 negative words and we rank the words from highest to lowest, then we pick up top 30 from head and top 30 from tail. In this case, if the total number of words is less than 60, there is going to be some overlaps between positive word list and negative word list. We modify our code by adding a if condition to avoid this situation happen.

8. Summary and Conclusion

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.

AUC for each legal field

AUC	Logistic Regression with tf-idf
11th Abrogation	0.845
Abortion	0.642
ADA	0.751
Affirmative Action	0.653
Campaign Finance	0.876
Capital Punishment	0.650
EPA	0.72
FCC	0.96
First Amend	0.695
Homosexual Rights	0.873
NEPA	0.783
NLRB	0.715
Obscenity	0.855
Piercing Corp Veil	0.719
Sex Discrimination	0.752
Title 7	0.78

The following table shows important word features in each legal field.

Legal Field	Positive	Negative
11th Abrogation	accru plaintiff knew known, amend constitut unit state, argu eleventh amend, compel undertak approach, congress unequivoc express, congruent proport	cite discrimin public, compet privat enterpris doe mean, congress articl, enact statut, plaintiff appel argument exact relief seek
Abortion	appeal concern, Life health, court held proper function legislatur, health servic, pregnant minor	mother result, clinic district court abus discret, plaintiff deni,
ADA	Administr, Distress, punit damag, rehabilit act claim	medic condit, Dispar, extend
Affirmative Action	black candid, board compli, racial balanc, impermiss	Arbitrari, argu constitut, Conscienc, constitut violat
Campaign Finance	advertis influenc outcom vote, argument appel consid definit, challeng present, case controversi district, disclosur sourc	Expens, inform elector mean provis, compel court went histori, buckley court limit
Capital Punishment	duti make reason, Involuntari, materi reason probabl, mental health	consid mitig, Attack, Inelig, counti jail
EPA	act impos, board character, Chevron, Elimin, interst transport hazard wast	factor demonstr, id statut silent ambigu respect, requir provis
FCC	arrier transport,	recov intrast contribut state

	conduct fund line texa counsel contend,	author, order feder communic commiss
First Amend	materi fact, purpos regul, materi fact, purpos regul	concur result reach judg hall opinion analysi amend, direct narrowli, essenti curti content statut inclus, provid ineffect remot support govern
Homosexual Rights	claim revers, equal protect claus fourteenth amend, homosexu engag consensu sodomi court state great, prohibit homosexu conduct, sexual prefer	know corps, militari matter difficult think clearer,
NEPA	accord proper forest servic fish, area plan, increas risk, litig cost save, nativ	caprici violat. result destruct, project narrowli
NLRB	animus design rid compani financi, care includ critic, complaint alleg violat section mention, Limit, involv face simpl jurisdict question doe	alleg februari precis interrog employe threaten employe, allow mason van atter time acclim posit, expens attribut unit employ communic worker finerti
Obscenity	appel court, constitut protect, regard invalid major opinion miller recogn	core pornographi, materi public, sexual relat
Piercing Corp Veil	oblig make, issu liabil, refund	agreement district, alleg complaint, court review
Sex Discrimination	district court grant summari judgment favor, dismiss complaint, complain, opinion	deni, discrimin, evid

Title 7	Faith, Reason trier fact, conclud	prima faci case racial discrimin, discrimin retali

9. Reference

[1] Cass R. Sunstein, Are Judges Political? An Empirical Analysis of the Federal Judiciary.
[2] Tf-Idf Document: http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html
[3] Random Forest Feature Selection: http://blog.datadive.net/selecting-good-features-part-iii-random-forests/
[4] Random Forest Documentation: http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
[5] Logistic Regression Document: http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

10. 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.

Appendix: Data Files Screenshot

1. Docvec_text

part-00000


```
('X3FM41||28sep2000||Civil Procedure||2000||4||MICHAEL, M. BLANE||BUTZNER, JOHN||TRAXLER, WILLIAM BYRD, JR.||MICHAEL, M. BLANE||MICHAEL||1||1||1',
'374710:33||380681:1||380754:1||380755:1||380756:1||380757:1||380758:1||380759:1||381223:3||381224:1||381225:1||381226:1||381227:1||381228:1||
381229:1||381230:2||381231:1||381232:1||381233:1||381234:1||381235:1||381236:1||381237:1||381238:1||381239:1||381240:1||382248:1||382267:1||
382268:1||382269:1||382270:1||382271:1||382272:1||386146:3||386201:3||386207:3||386208:1||386209:1||386210:1||386211:1||386215:1||386216:1||
386217:1||386218:1||386219:1||386220:1||386221:1||386222:1||434848:1||443674:1||443685:1||443686:1||443687:1||443688:1||443689:1||506253:1||
506473:1||506488:1||506489:1||506490:1||506491:1||506492:1||512551:1||512817:1||512818:1||512819:1||512820:1||512821:1||512822:1||537671:1||
537684:1||537685:1||537686:1||537687:1||537688:1||537689:1||544808:2||544936:1||544937:1||544938:1||544939:1||544940:1||544941:1||544988:1||
544989:1||544990:1||544991:1||544992:1||544993:1||580565:3||580693:1||580694:1||580695:1||580696:1||580697:1||580698:1||580872:1||580873:1||
580874:1||580875:1||580876:1||580877:1||580920:1||580921:1||580922:1||580923:1||580924:1||580925:1||583682:1||583882:1||583883:1||583884:1||
583885:1||583886:1||583887:1||596296:1||596538:1||596576:1||596577:1||596578:1||596579:1||596580:1||642558:1||642665:1||642666:1||642667:1||
642668:1||642669:1||642670:1||645522:1||645535:1||645536:1||645537:1||645538:1||645539:1||645540:1||666849:1||666862:1||666863:1||666864:1||
666865:1||666866:1||666867:1||678430:3||678636:2||678637:2||678638:1||678639:1||678640:1||678641:1||678642:1||678643:1||678644:1||678645:1||
680687:1||680712:1||680713:1||680714:1||680715:1||680716:1||691847:1||691854:1||691855:1||691856:1||691857:1||691858:1||691859:1||698075:1||
698082:1||698088:1||698093:1||698094:1||698095:1||698096:1||727299:1||727731:1||727732:1||727733:1||727734:1||727735:1||727736:1||738255:1||
738384:1||738385:1||738386:1||738387:1||738388:1||738389:1||739850:4||740262:1||740263:1||740264:1||740277:1||740293:1||740294:1||740503:1||
740514:1||740515:1||740516:1||740517:1||740518:1||740791:2||740817:1||740818:1||740819:1||740820:1||740821:1||740838:1||740839:1||740840:1||
740841:1||740842:1||7082759:2||8242391:1||8244075:1||8244076:1||8244077:1||8244078:1||8244079:1||8244080:1||8582750:1||8583885:1||8583896:1||
8583897:1||8583904:1||8583905:1||8583906:1||11280604:4||11328459:1||11328460:1||11328461:1||11328462:1||11328463:1||11328464:1||11328465:1||
11438042:1||11438566:1||11438567:1||11438568:1||11438569:1||11438570:1||11438571:1||12177788:1||12178302:1||12227280:1||12227281:1||12227282:1||
12227283:1||12227284:1||14257649:1||14264360:1||14265337:1||14265342:1||14265343:1||14265344:1||14265345:1||16922792:3||17045917:1||17045966:1||
17045967:1||17045968:1||17045969:1||17045970:1||17045972:1||17051569:1||17051642:1||17051653:1||17051654:1||17051655:1||17051656:1||17051657:1||
17142685:1||17144740:1||17146467:1||17146468:1||17146469:1||17146470:1||17146471:1||17282942:1||17501604:1||17501809:1||17501810:1||17501811:1||
17501812:1||17501813:1||17501814:1||19138098:1||19575263:1||19575264:1||19575265:1||19575266:1||19575267:1||19575268:1||19575269:1||19723319:2||
19747616:1||19747646:1||19747647:1||19747648:1||19747649:1||19747650:1||19747651:1||19912182:1||19912459:1||19912460:1||19912461:1||19912462:1||
19912463:1||19912464:1||20562748:11||20753339:1||20758292:1||20758293:1||20758294:1||20758295:1||20758296:1||20758297:1||21764921:1||21779779:1||
21779844:1||21779845:1||21779846:1||21779847:1||21779848:1||24458953:1||24458954:1||24458955:1||24458956:1||24458957:1||24458958:1||24458959:1||
24568212:1||24568255:1||24568256:1||24568257:1||24568258:1||24568259:1||24568260:1||24801874:1||24819852:1||24819858:1||24819859:1||24819860:1||
24819861:1||24819862:1||25692083:1||25692376:1||25692387:1||25692388:1||25692389:1||25692390:1||25692391:1||26968717:1||26981123:1||26981139:1||
26981140:1||26981141:1||26981142:1||26981143:1||27165906:1||27245890:1||27245921:1||27245922:1||27245923:1||27245924:1||27245925:1||27752663:1||
```


2 Vocab_map_text (dictionary)


```
(u'access nuclear power station requir govern regul', 9187494)
(u'access nuclear power station requir govern regul photo', 9187495)
(u'access nuclear wast', 9187496)
(u'access nuclear wast fund', 9187497)
(u'access nuclear wast fund herrington', 9187498)
(u'access nuclear wast fund herrington nevada', 9187499)
(u'access nuclear wast fund herrington nevada affect', 9187500)
(u'access nuclear wast fund herrington nevada affect unit', 9187501)
(u'access nuclear wast fund id', 9187502)
(u'access nuclear wast fund id fund', 9187503)
(u'access nuclear wast fund id fund activ', 9187504)
(u'access nuclear wast fund id fund activ activ', 9187505)
(u'access nuclear wast fund pre', 9187506)
(u'access nuclear wast fund pre site', 9187507)
(u'access nuclear wast fund pre site character', 9187508)
(u'access nuclear wast fund pre site character activ', 9187509)
(u'access nuclear weapon', 9187510)
(u'access nuclear weapon mind', 9187511)
```


3 Merged raw dataframe

	caseid	citation	panelvote	issue	field	n_gram
0	X42KB3	475 F2d 65	1.0	16	Obscenity	{'1207777881': 1, '2045835947': 1, '1090701101...
1	XE6R35	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, '1874208517...
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)	{'315987990': 1, '143889': 1, '1861859077': 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'...

4. Training data format

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