Work Outline:

part1 - clean data:

1.clean FEC data:

->step1:preprocess the data

0)remove the lines where last name == “” or first name == “” or state abbrev == “”

e.g.,

AL	1				Scattered	W				159	0.06%
AL	1									255,164	

1)add year attribute

2)delete all dirty data and reformat numeric data which was written in string format

clean the data:

->step2:merge the data

e.g.,

NY	5	H4NY07011	(I)	Gary L.	Ackerman	D	Unopposed	105,836	66.66%
NY	5	H4NY07011	(I)	Gary L.	Ackerman	IDP	Unopposed	4,084	2.57%
NY	5	H4NY07011	(I)	Gary L.	Ackerman	WF	Unopposed	2,804	1.77%

one candidate was named by three parties

part2: hypothesis analysis Preprocess:

1.reformat the data:

original data format:

	YEAR	STATE	FIRST	LAST	DISTRICT	CANDIDATE	PARTY	AMOUNT	INDUSTRYPE	CANDTOTAL	INCUMBENT	VOTES	PERCENT	PRIMARY.IN	WINNER	indrank
2	2006	CA	A	SEKHON	2	A J SEKHON	D	400	0.00264717	151105	0	68234	0.3246	Financials	0	5
3	2006	CA	A	SEKHON	2	A J SEKHON	D	1750	0.01158135	151105	0	68234	0.3246	Industrials	0	4
4	2006	CA	A	SEKHON	2	A J SEKHON	D	138055	0.91363621	151105	0	68234	0.3246	Materials	0	1
5	2006	CA	A	SEKHON	2	A J SEKHON	D	8400	0.05559048	151105	0	68234	0.3246	Not for profi	0	2

new data format:

the attributes contain:

KEY, YEAR, STATE, FIRST, LAST, DISTRICT, CANDIDATE, PARTY, INDUSTRYPERCENT, CANDTOTAL, INCUMBENT, VOTES PERCENT, WINNER,

rank 1, rank 2, rank 3, rank 4, rank 5, rank 6, rank 7, rank 8, rank 9, rank 10, rank 11, rank 12

AMOUNT of 1, AMOUNT of 2, AMOUNT of 3, AMOUNT of 4, AMOUNT of 5, AMOUNT of 6, AMOUNT of 7, AMOUNT of 8, AMOUNT of 9, AMOUNT of 10, AMOUNT of 11, AMOUNT of 12

attribute 1~19 in the table

KEY	YEAR	STATE	FIRST	LAST	DISTRICT	CANDIDATE	PARTY	INDUSTRYPE	CANDTOTAL	INCUMBENT	VOTES	PERCENT	WINNER	rank 1	rank 2	rank 3	rank 4	rank 5
MN|2|MIKE	2012	MN	MIKE	OBERMUELL	2	MIKE OBERM	D	0.03386456	389788	0	164338	0.4585	L	Not for profi	Not publicly	Consumer Di	Financials	
WI|2|JOE|K	2012	WI	JOE	KOPSICK	2	JOE KOPSICK	I	1	1000	0	6	0	L	Not for profit				
CA|50|DUN	2014	CA	DUNCAN	HUNTER	50	DUNCAN D H	R	0.06342402	843529	1	107835	0.714	W	Industrials	Not for profi	Not publicly	Consumer St	Consumer Di
MN|2|MIKE	2014	MN	MIKE	OBERMUELL	2	MIKE OBERM	D	0.06252226	379065	0	95565	0.389	L	Not for profi	Not publicly	Consumer Di	Financials	Health Care

attribute 20~44 in the table

rank 6	rank 7	rank 8	rank 9	rank 10	rank 11	rank 12	AMOUNT of	AMOUNT of	AMOUNT of	AMOUNT of	AMOUNT of	AMOUNT of	AMOUNT of	AMOUNT of	AMOUNT of	AMOUNT of	AMOUNT of	AMOUNT of 12
							314263	51325	13200	11000								
							1000											
Energy	Information Technology						425965	121450	121114	54000	53500	42700	24800					
							254915	58000	23700	22950	19500							

in this format, it is easy to do decision tree, kNN and Naive Bayes

2.give missing data a default value:

give all categorical attribute missing data a default value NULL, and all numeric attribute missing data a default value as 0.

KEY	YEAR	STATE	FIRST	LAST	DISTRICT	CANDIDATE	PARTY	INDUSTRYPE	CANDTOTAL	INCUMBENT	VOTES	PERCENT	WINNER	rank 1	rank 2	rank 3	rank 4	rank 5	r
MN|2|MIKE	2012	MN	MIKE	OBERMUELL	2	MIKE OBERN	D	0.03386456	389788	0	164338	0.4585	L	Not for profi	Not publicly	Consumer Di	Financials	NULL	N
WI|2|JOE|K	2012	WI	JOE	KOPSICK	2	JOE KOPSICK	I	1	1000	0	6	0	L	Not for profi	NULL	NULL	NULL	NULL	N
CA|50|DUN	2014	CA	DUNCAN	HUNTER	50	DUNCAN D	R	0.06342402	843529	1	107835	0.714	W	Industrials	Not for profi	Not publicly	Consumer St	Consumer Di	E
MN|2|MIKE	2014	MN	MIKE	OBERMUELL	2	MIKE OBERN	D	0.06252226	379065	0	95565	0.389	L	Not for profi	Not publicly	Consumer Di	Financials	Health Care	N

rank 6	rank 7	rank 8	rank 9	rank 10	rank 11	rank 12	AMOUNT of	AMOUNT of	AMOUNT of	AMOUNT of	AMOUNT of	AMOUNT of	AMOUNT of	AMOUNT of	AMOUNT of	AMOUNT of	AMOUNT of	AMOUNT of 12
NULL	NULL	NULL	NULL	NULL	NULL	NULL	314263	51325	13200	11000	0	0	0	0	0	0	0	0
NULL	NULL	NULL	NULL	NULL	NULL	NULL	1000	0	0	0	0	0	0	0	0	0	0	0
Energy	Information	NULL	NULL	NULL	NULL	NULL	425965	121450	121114	54000	53500	42700	24800	0	0	0	0	0
NULL	NULL	NULL	NULL	NULL	NULL	NULL	254915	58000	23700	22950	19500	0	0	0	0	0	0	0

3.filter some too sparse attribute or unnecessary attributes in hypothesis analysis

PARTY	INDUSTRYPE	CANDTOTAL	INCUMBENT	WINNER	rank 1	rank 2	rank 3	rank 4	rank 5	AMOUNT of	AMOUNT of	AMOUNT of	AMOUNT of	AMOUNT of
1	0.03386456	389788	0	L	1	8	5	2	0	314263	51325	13200	11000	0
2	1	1000	0	L	1	0	0	0	0	1000	0	0	0	0
3	0.06342402	843529	1	W	4	1	8	3	5	425965	121450	121114	54000	53500
1	0.06252226	379065	0	L	1	8	5	2	9	254915	58000	23700	22950	19500

e.g., last name, first name, etc. are not important in the model to predict whether the man will win the campaign.

4.bin the data(amount of no.1 donation industry to amount of no.12 donation of data and candTotal) This is for NaiveBayes analysis, which only accepts categorical attributes.

PARTY	INDUSTRYPE	CANDTOTAL	INCUMBENT	VOTES	PERCENT	WINNER	rank 1	rank 2	rank 3	rank 4	rank 5
1	0.03386456	389788	0	164338	0.4585	L	1	8	5	2	0
2	1	1000	0	6	0	L	1	0	0	0	0
3	0.06342402	843529	1	107835	0.714	W	4	1	8	3	5
1	0.06252226	379065	0	95565	0.389	L	1	8	5	2	9
1	0.04799427	69800	0	83176	0.3347	L	1	5	4	3	13

Part3: Hypothesis Analysis

Decision Tree Model:

Summary:

I want to show you the result first and then describe the details. So let's take a look at the result first:

Compare the three models

	DT	kNN	Naïve Bayes
Prediction accuracy	0.881	0.839	0.879
Prediction variance	0.105	0.135	0.106
Area under the ROC curve	0.842	0.893	0.871
95% CI	0.812-0.872(DeLong)	0.867-0.919(DeLong)	0.842-0.899(DeLong)
hypothesis 1	Wrong. Feature "rank1 amount" and feature "rank5 amount" are more important than others	Wrong. These features are the same important.	Wrong. These features are the same important.
hypothesis 2	Right.	Wrong.	Right.
hypothesis 3	Feature Candtotal is important, but YrPercentChange is	Both of them are not important.	Feature Candtotal is important, but YrPercentChange is

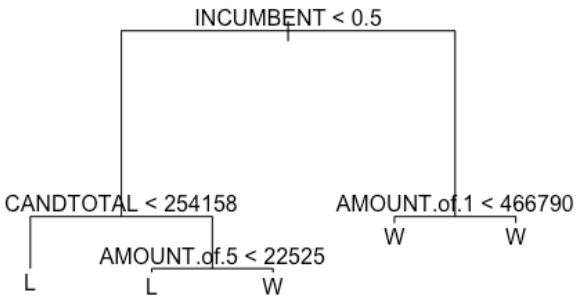
	not.		not.

[note]
[hypothesis 1]: in prediction, the feature of rank1 amount of money is more important than rank2 amount of money, and rank2 amount of money is more important than rank3 amount of money, ...
[hypothesis 2]:is incumbent feature important?
[hypothesis 3]: Candtotal and YrPercentChange important?
[Prediction] In this experiment, I am trying to predict whether the candidate can win the campaign.

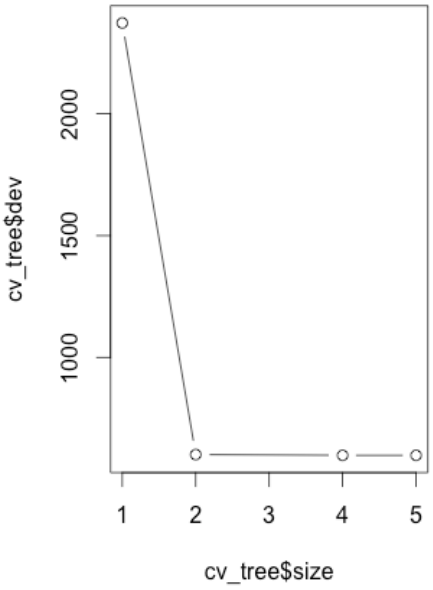
[Conclusion]
Compare these three models, all of them work pretty well, which really surprises me. I never thought the model can predict the winner of campaign with accuracy above 0.8.
And in these three models, decision tree seems work best.

Here are the details of the three models' results:

1.Tree structure after pruning:



I use cv_tree$dev~cv_tree$size to decide the tree's size and do pruning: in this case, I choose size=3 according to the following graph:



2.tree hypothesis results:

[hypothesis 1]: in prediction, the feature of rank1 amount of money is more important than rank2 amount of money, and rank2 amount of money is more important than rank3 amount of money, ...

[result]Wrong.
rank1 amount of money and rank5 amount of money is really important in tree model. rank2 amount of money~rank4 amount of money are not important.

[hypothesis 2]:is incumbent feature important?
[result]It is the most important feature. It is the root of the tree, which means its information gain is the biggest in all attributes.

[hypothesis 3]: Candtotal and YrPercentChange important?
[result]Candtotal is important.
But YrPercentChange is not important. It is not a node in the decision tree.

3.[prediction analysis]

What am I doing here:

I try to use the data we collect to predict whether the politician will win the campaign or not.

Here is the result:

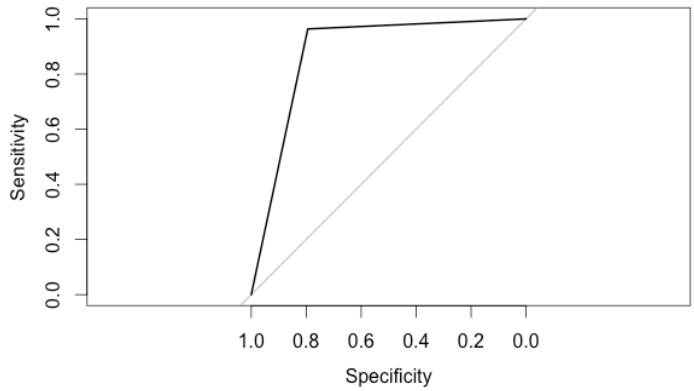
tree model	confusion matrix	accuracy	variance	TRUE	false positive	true positive

run1	tree_predict_after_prune L W L 222 15 W 45 258	0.889	0.099	300	0.150	0.850
run2	tree_predict_after_prune L W L 204 14 W 44 278	0.893	0.096	322	0.137	0.863

[Note] L represents Loser and W represents Winner

Similarly, I run it 10 times, the mean(accuracy) is 0.881 and variance is 0.105, which is amazing for me.

4.[ROC analysis]:



(Note: I cannot set smooth=TRUE because there is not enough nodes. I try "trainset:testset = 1:1, but it is still not enough")
result:
Area under the curve: 0.8929
95% CI: 0.8672-0.9187 (DeLong)
the area under the curve is pretty high, which means the model works pretty well.

kNN analysis(Lazy Model)
[hypothesis 1]: in prediction, the feature of rank1 amount of money is more important than rank2 amount of money, and rank2 amount of money is more important than rank3 amount of money, ...
[Answer] Wrong. seems that they are the same important.
the below is the result if we discard the feature rank1 amount/rank2 amount/...

[situation 1]without attribute: rank1 amount of money

```
          result
testTarget   L   W
        L 176  85
        W   6 273
```

Accuracy: 0.8314815

Variance: 0.14038

[situation 2]without attribute: rank2 amount of money

	result	
testTarget	L	W
	L 194	70
	W	14 262

Accuracy: 0.8444444
Variance: 0.1316017

[situation 3]without attribute: rank3 amount of money

	result	
testTarget	L	W
	L 179	72
	W	18 271

Accuracy: 0.8333333
Variance: 0.1391466

[situation 4]without attribute: rank4 amount of money

	result	
testTarget	L	W
	L 193	63
	W	16 268

Accuracy: 0.8537037
Variance: 0.1251254

[situation 5]without attribute: rank5 amount of money

	result	
testTarget	L	W
	L 200	57
	W	17 266

Accuracy: 0.862963
Variance: 0.1184773

[hypothesis 2]: is incumbent feature important?
[Answer]: No it is not very important. It is different from DT and Naive Bayes.

	result	
testTarget	L	W
	L 211	64
	W	8 257

Accuracy: 0.8666667
Variance: 0.1157699

[hypothesis 3]: Candtotal and YrPercentChange important?
[Answer]: No. Candtotal is not important.

	result	
testTarget	L	W
	L 184	75

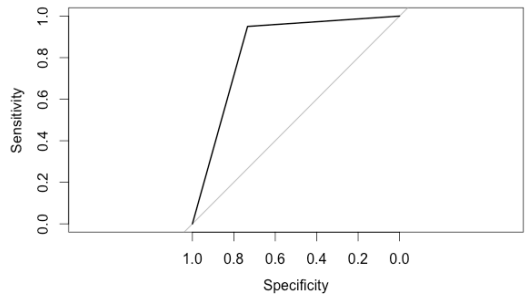
W 12 269
Accuracy: 0.8388889
Variance: 0.1354051

without attribute: percentage
result
testTarget L W
L 195 70
W 9 266
Accuracy: 0.8537037
Variance: 0.1251254
There are not big change of accuracy after I discard the attribute Candtotal or YrPercentChange

Result Analysis

I try to use the data to predict whether the guy can win the campaign or not. Here is the result:
Mean Accuracy = 0.83925926
Variance = 0.13503195
The result is amazing. We can predict the result with accuracy of 0.84.

ROC analysis



Area under the curve: 0.8419
95% CI: 0.8121-0.8717 (DeLong)
the area under the curve is pretty high, which means the model works pretty well.

Naïve Bayes Model

1.hypothesis analysis

[hypothesis 1]: in prediction, the feature of rank1 amount of money is more important than rank2 amount of money, and rank2 amount of money is more important than rank3 amount of money, ...
Wrong. seems that they are the same imporant.

the below is the result if we discard the feature rank1 amount or rank2 amount/...
from which, we can see that the accuracy, variance and confusion matrix don't change a lot.

[situation 1]without attribute: rank1 amount of money
Confusion Matrix:
pred L W
L 252 38

W 27 223
Accuracy: 0.8796296
Variance: 0.1060778

[situation 2]without attribute: rank2 amount of money
Confusion Matrix:

pred L W
L 228 32
W 36 244
Accuracy: 0.8740741
Variance: 0.1102728

[situation 3]without attribute: rank3 amount of money
Confusion Matrix:
pred L W
L 249 37
W 20 234
Accuracy: 0.8944444
Variance: 0.09458874

[situation 4]without attribute: rank4 amount of money
Confusion Matrix:
pred L W
L 247 36
W 27 230
Accuracy: 0.8833333
Variance: 0.1032468

[situation 5]without attribute: rank5 amount of money
Confusion Matrix:
pred L W
L 235 40
W 31 234
Accuracy: 0.8685185
Variance: 0.114406

[hypothesis 2]:is incumbent feature important?
Yes. whether the man/woman is the incumbent is the most important feature. If we discard this feature, the model's performance is as following:
pred L W
L 134 17
W 135 254
Accuracy: 0.7185185
Variance: 0.2026249
The accuracy drops a lot.

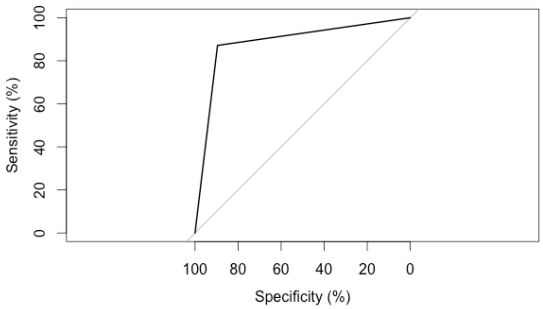
[hypothesis 3] Candtotal and YrPercentChange important?
No. Candtotal and YrPercentChange are not important. The top 5 money donation attributes

work well.
The accuracy doesn't change much after I delete the attribute YrPercentChange or CandTotal.
without attribute: YrPercentChange
pred L W
L 242 34
W 25 239
Accuracy: 0.8907407
Variance: 0.09750223

without attribute: total amount of money(Candtotal)
pred L W
L 240 36
W 37 227
Accuracy: 0.8648148
Variance: 0.1171271

[Prediction Result Analysis]
Mean Accuracy = 0.8788889
Variance = 0.10642135
The result is amazing. We can predict the result with accuracy of 0.84.

[ROC analysis]



Area under the curve: 87.05%
95% CI: 84.22%-89.87% (DeLong)
the area under the curve is pretty high, which means the model works pretty well.

Code citation:
Naïve bayes:
#code citation: https://cran.r-project.org/web/packages/e1071/e1071.pdf
#code citation: https://cran.r-project.org/web/packages/pROC/pROC.pdf

kNN:
#code citation: https://cran.r-project.org/web/packages/pROC/pROC.pdf
#code citation: https://www.youtube.com/watch?v=GtgJEVxI7DY
#code citation: https://www.youtube.com/watch?v=DkLNb0CXw84
#code citation: https://cran.r-project.org/web/packages/class/class.pdf

Decision Tree:
#code citation: https://www.youtube.com/watch?v=GOJN9SKl_OE
#code citation: https://cran.r-project.org/web/packages/ISLR/ISLR.pdf
#code citation: https://cran.r-project.org/web/packages/pROC/pROC.pdf