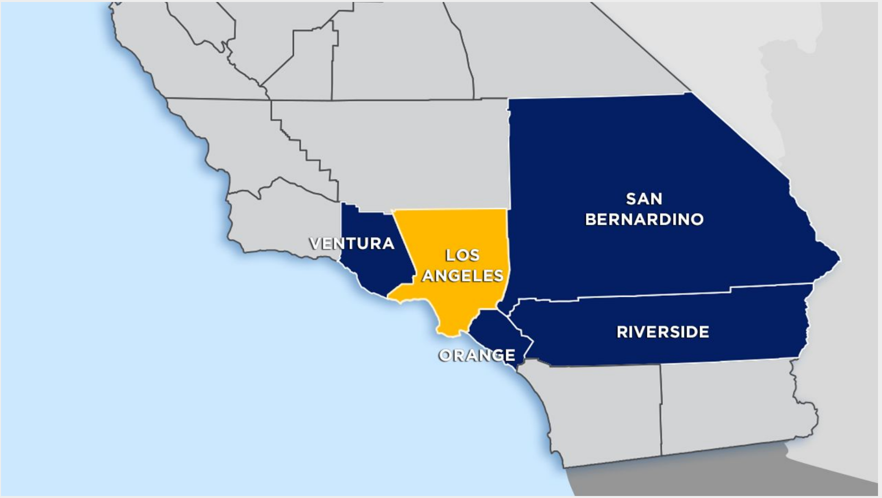
‘Yea or Nay’: LA Politicians and Machine Learning



Elijah Hernandez

Roadmap

- Research Question and Motivation
- Data Collection
- Data Cleaning
- Data Visualization
- Machine Learning
- Correctness of Results
- Conclusion

Roadmap

- Research Question and Motivation
- Data Collection
- Data Cleaning
- Data Visualization
- Machine Learning
- Correctness of Results
- Conclusion

Research Question and Motivation

What attributes are most indicative of how a Los Angeles congressional member will vote (yea or nay) and to what extent can the results be relied upon?

Motivation:

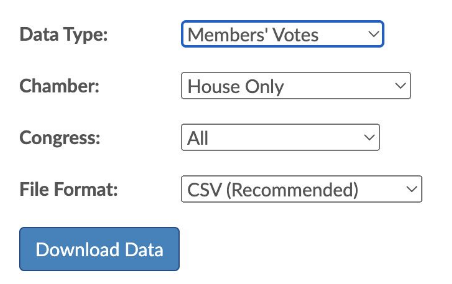
- Curiosity about the strength of influences on a person’s voting behavior
- Wanted a way to gauge LA area politicians’ general approval towards new bills, propositions, etc.

Roadmap

- Research Question and Motivation
- Data Collection
- Data Cleaning
- Data Visualization
- Machine Learning
- Correctness of Results
- Conclusion

Data Collection

voteview.com beta



	A	B	C	D	E
1	icpsr	bioname	gender	race	immigrant
2	20703	MCCARTHY, Kevin	male	white	no
3	21988	GARCIA, Mike	male	hispanic	no
4	21308	BROWNLEY, Julia	female	white	no
5	20955	CHU, Judy	female	asian	no
6	20104	SCHIFF, Adam	male	white	no
7	21309	CÁRDENAS, Tony	male	hispanic	no
8	29707	SHERMAN, Brad	male	white	no
9	29903	NAPOLITANO, Grace Flores	female	hispanic	no
10	21507	LIEU, Ted	male	asian	yes
11	21754	GOMEZ, Jimmy	male	hispanic	no
12	21508	TORRES, Norma Judith	female	hispanic	yes

Where my datasets came from:

- VoteView, a database maintained by the UCLA political science dpt
  - Members - key demographics
  - Rollcalls - what is being voted on
  - Votes - how each member voted
- Created own csv file for specific LA representative demographics

Roadmap

- Research Question and Motivation
- Data Collection
- Data Cleaning
- Data Visualization
- Machine Learning
- Correctness of Results
- Conclusion

Data Cleaning

```
H117_members
```

	congress	chamber	icpsr	state_icpsr	district_code	state_abbrev	party_code	occupancy	last_means	bioname	...
0	117	President	99912	99	0	USA	200	0.0	0.0	TRUMP, Donald John	...
1	117	President	99913	99	0	USA	100	0.0	0.0	BIDEN, Joseph Robinette, Jr.	...
2	117	House	20301	41	3	AL	200	NaN	NaN	ROGERS, Mike Dennis	...
3	117	House	21102	41	7	AL	100	NaN	NaN	SEWELL, Terri	...
4	117	House	21193	41	5	AL	200	NaN	NaN	BROOKS, Mo	...
...	...	...	...	...	...	...	...	...	...	...	...
452	117	House	21970	25	1	WI	200	NaN	NaN	STEIL, Bryan	...
453	117	House	21989	25	7	WI	200	NaN	NaN	TIFFANY, Thomas P.	...
454	117	House	22115	25	5	WI	200	NaN	NaN	FITZGERALD, Scott	...

	icpsr	bioname	district_code	party_code	born	economic lib-con
33	20703	MCCARTHY, Kevin	23	200	1965	0.457
64	21988	GARCIA, Mike	25	200	1976	0.370
44	21308	BROWNLEY, Julia	26	100	1952	-0.289
36	20955	CHU, Judy	27	100	1953	-0.489
26	20104	SCHIFF, Adam	28	100	1960	-0.350
45	21309	CÁRDENAS, Tony	29	100	1963	-0.387
75	29707	SHERMAN, Brad	30	100	1954	-0.343
78	29903	NAPOLITANO, Grace Flores	32	100	1936	-0.448
53	21507	LIEU, Ted	33	100	1969	-0.377
60	21754	GOMEZ, Jimmy	34	100	1974	-0.573
54	21508	TORRES, Norma Judith	35	100	1965	-0.366
38	21110	BASS, Karen	37	100	1953	-0.584
29	20310	SÁNCHEZ, Linda T.	38	100	1969	-0.508
66	22129	KIM, Young	39	200	1962	0.277

```
[ ] df_ml
```

rollnumber	icpsr	cast_code	prob	bioname	district_code	party_code	age elected	economic lib-con	vote_desc	vote_question
3	20104.0	yea	100.0	SCHIFF, Adam	28	100	60	-0.350	Adopting the Rules of the House of Representat...	On Motion to Table the Motion to Postpone to a...
4	20104.0	yea	100.0	SCHIFF, Adam	28	100	60	-0.350	Adopting the Rules of the House of Representat...	Table Motion to Refer
5	20104.0	yea	100.0	SCHIFF, Adam	28	100	60	-0.350	Adopting the Rules of the House of Representat...	On Ordering the Previous Question
6	20104.0	nay	100.0	SCHIFF, Adam	28	100	60	-0.350	Adopting the Rules of the House of Representat...	On Motion to Commit with Instructions
7	20104.0	yea	100.0	SCHIFF, Adam	28	100	60	-0.350	Adopting the Rules of the House of Representat...	On Agreeing to the Resolution

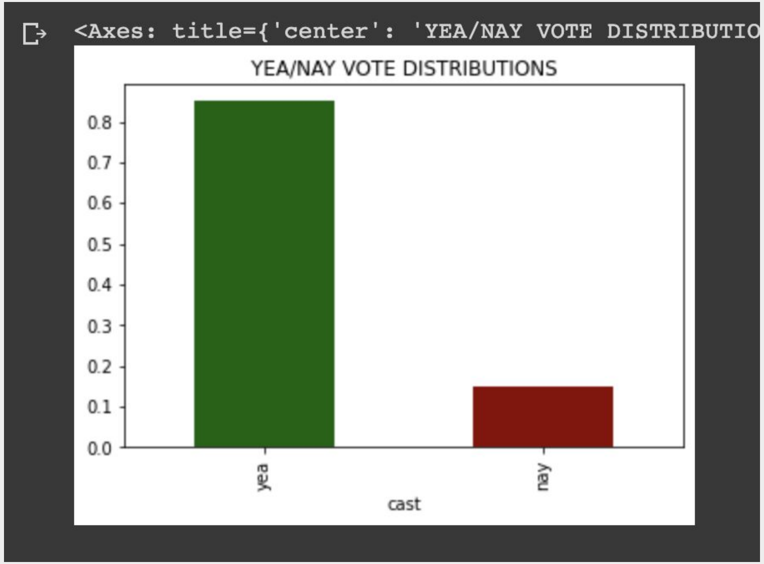
What I did:

- Filtered members dataset for only those in the Los Angeles county area
- Dropped non desired columns from data frames
- Joined all four datasets into one master one
  - ‘icpsr’ , ‘rollnumber’
- Cast binary categories into ‘1’ and ‘0’
- Converted ‘born’ year into age
- Exclude rollcalls will nan values for descriptions

Roadmap

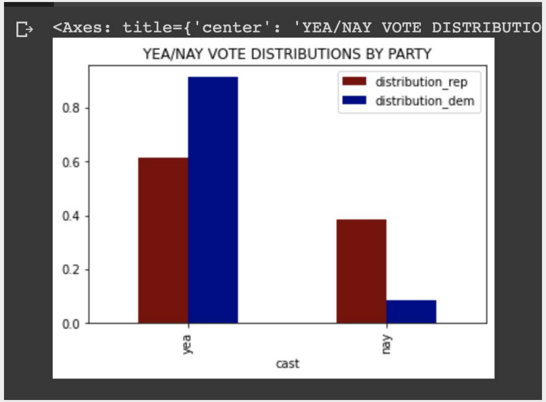
- Research Question and Motivation
- Data Collection
- Data Cleaning
- Data Visualization
- Machine Learning
- Correctness of Results
- Conclusion

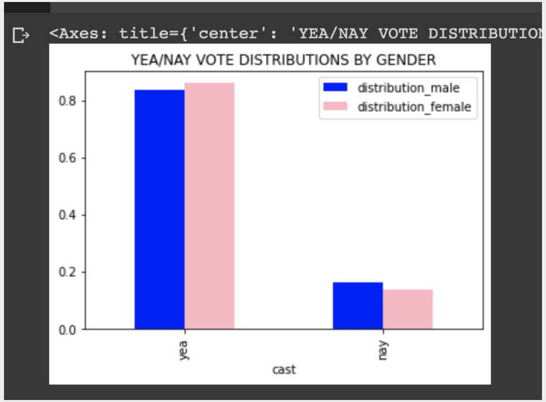
Data Visualization



- Comparing differences between ‘YEA/NAY’ vote distributions for each category
- Will use this as a metric to determine the order of inputs for machine learning model

Data Visualization

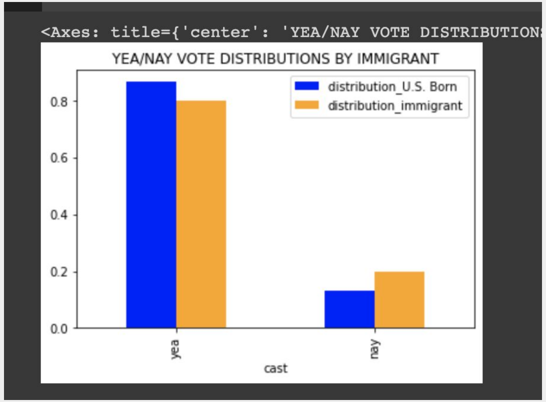


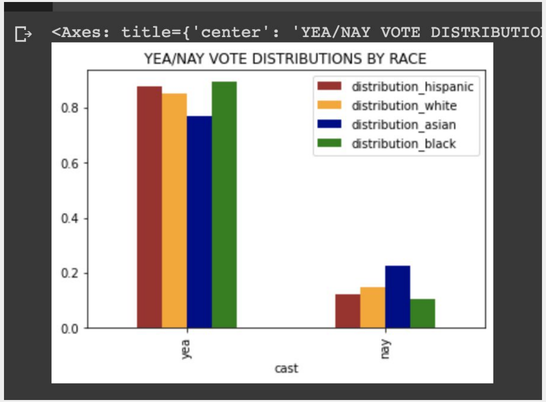


- Democrats split heavier towards 'YEA' votes than Republicans
- Republicans split heavier towards 'NAY' votes than Democrats

- Marginal difference between gender split on votes

Data Visualization

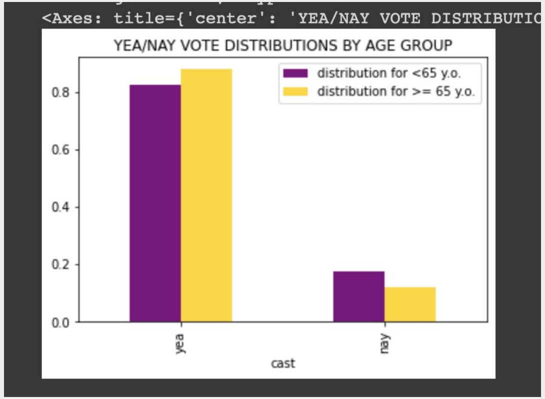


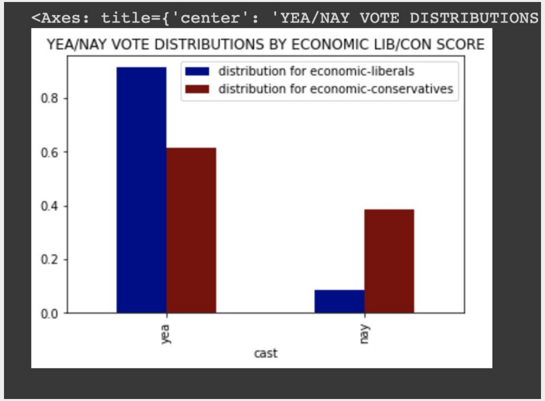


- Marginal difference between status as immigrant/naturalized citizen split on votes

- Black representatives most likely to vote ‘YEA’
- Asian representatives most likely to vote ‘NAY’

Data Visualization





- Marginal difference between age group split on votes

- Economic liberals largely split towards voting ‘YEA’
- Economic conservatives largely split towards voting ‘NAY’
- *this split largely mirrors that by political party*

Roadmap

- Research Question and Motivation
- Data Collection
- Data Cleaning
- Data Visualization
- Machine Learning
- Correctness of Results
- Conclusion

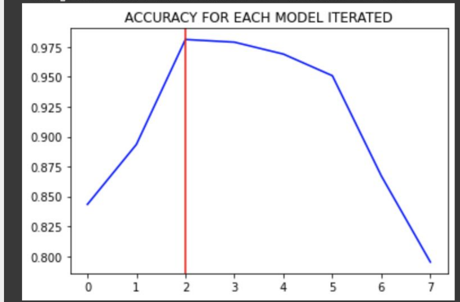
Machine Learning

```
import numpy as np

model_indices = np.argmax(np.asarray((scores)), axis=0)+1
features[:model_indices]
```

```
['vote_desc', 'vote_question', 'party_code']
```

```
<matplotlib.lines.Line2D at 0x7ff32169f820>
```




```
grid_cv.fit(X_train, y_train)
grid_cv.best_params_
```

```
{'kneighborsclassifier__metric': 'manhattan',
 'kneighborsclassifier__n_neighbors': 3}
```

- Classification model
- K nearest neighbors
- Model selection
  - Testing x argument inputs according to the order of greatest comparative difference to least

- Hyperparameter tuning
- GridSearch
  - Number of neighbors
  - Distance metric

Roadmap

- Research Question and Motivation
- Data Collection
- Data Cleaning
- Data Visualization
- Machine Learning
- Correctness of Results
- Conclusion

Correctness of Results

```
cross_val_score(final_model,
                X=df_ml[['vote_desc', 'vote_question', 'party_code']],
                y=df_ml['cast_code'],
                scoring="accuracy", cv=10).mean()
```

```
0.983588799878196
```

```
confusion_matrix(y_train, y_train_)
```

```
array([[1000,  102],
       [  26, 6032]])
```

PRECISION: TP / (TP + FP)

YES: 6032 / (6032 + 102) = 6032 / 6134 = **0.98337137**

NO: 1000 / (1000 + 26) = 1000 / 1026 = **0.97465887**

RECALL: TP / (TP + FN)

YES: 6032 / (6032 + 26) = 6032 / 6058 = **0.99570815**

NO: 1000 / (1000 + 102) = 1000 / 1102 = **0.90744102**

F1 score: 1 / (1/2)((1/precision) + (1 /recall))

YES: 1 / (1/2)((1/0.98337137) + (1 /0.99570815)) = **0.98951118**

NO: 1 / (1/2)((1/0.97465887) + (1 /0.90744102)) = **0.93984962**

How I Measured the ‘Correctness’:

- Initially set ‘scoring’ equal to accuracy as a way to select a model and preferred parameters
- Estimated test error through implementing cross validation
- Printed a confusion matrix in order to compute a F1 score for each ‘yea’ or ‘nay’ prediction

Roadmap

- Research Question and Motivation
- Data Collection
- Data Cleaning
- Data Visualization
- Machine Learning
- Correctness of Results
- Conclusion

Conclusion

Key Takeaways

- Political party alone was determined to be the most indicative of a ‘yea’ or ‘nay’ vote
- Model possesses relatively high rate of success in its predictions

Applications

- Potential basis for campaign funding considerations

Ethical Concerns

- Though high predictive value, still NOT actual vote

Questions?