2016 Trump Voter Survey

Predictive Model Selection

Andrew Graham – Fall 2022

UNIVERSITY of DENVER
DANIEL FELIX RITCHIE SCHOOL OF ENGINEERING & COMPUTER SCIENCE

Overview

Summary and Goals

Purpose

To determine the best model to predict, based on survey results from surveys, whether someone voted for Trump or not in 2016. Survey contains demographics and views on ideology, religion and racism.

Significance

To gain an understanding of how certain demographics and views align. Challenge and/or confirm assumptions about the 2016 election. Prediction models can be used in future elections to estimate the likelihood of a candidate getting votes in the future. This can also be used to predict what types of candidates voters would prefer in future elections.

Overview

Summary and Goals

How does the performance of a Naïve Bayes Classifier compare with KNN Classifier, Random Forest Classifier , and Logistic Regression in predicting whether an individual voted for Trump in 2016 given survey answers on demographics, region, ideology, and views on racism?

Overview

Data Description

- Data from survey results collected in 2016
- 64,600 datapoints
- 44,898 with answer for Trump Vote
- 1 Continuous
- 17 Nominal
  - 2 categorical
  - 4 binary
  - 11 ordinal

Output: Voted for Trump

- 1, 0 (Yes, No)
- 40% Yes, 60% No
- Balanced

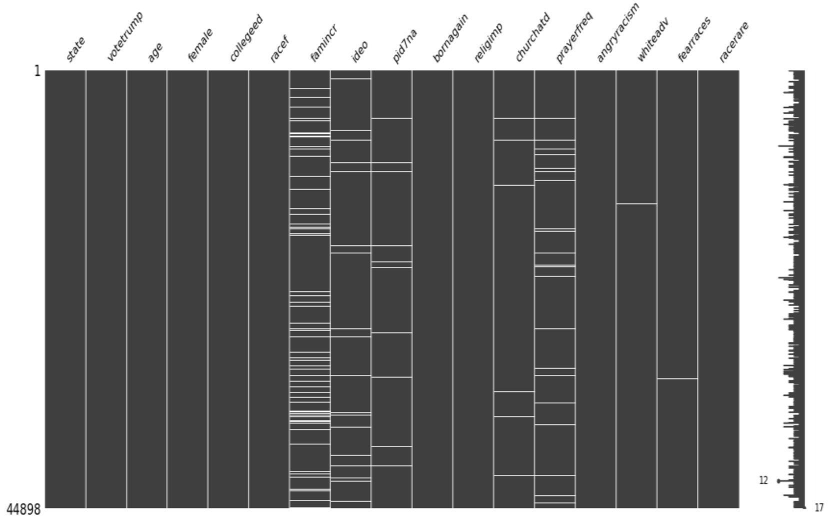
Inputs

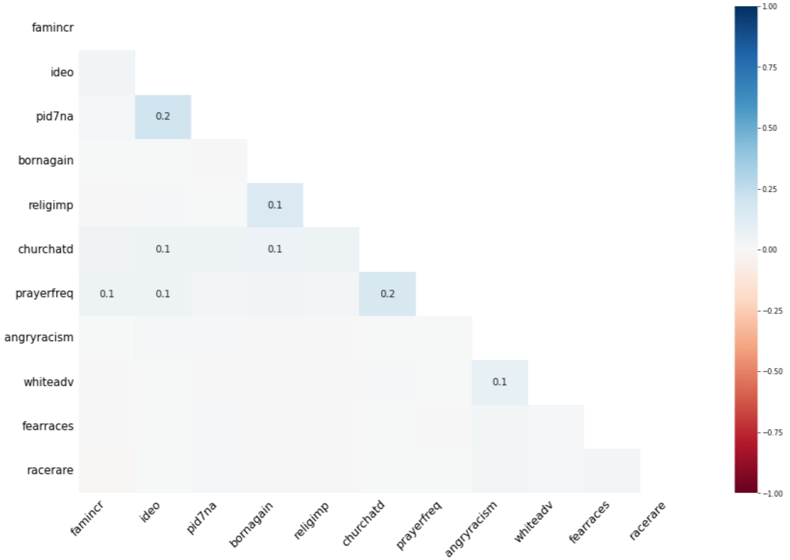
- 6 Demographic
  - Age, Sex, Education, Race, Income, State
- 2 Ideology Questions
- 4 Religion Questions
- 4 Racism Questions

Overview

Data Cleaning

	number_missing	percent_missing
state	0	0.000000
votetrump	0	0.000000
age	0	0.000000
female	0	0.000000
collegeed	0	0.000000
racef	0	0.000000
famincr	4717	10.506036
ideo	1501	3.343133
pid7na	611	1.360862
bornagain	22	0.049000
religimp	20	0.044545
churchatd	322	0.717181
prayerfreq	904	2.013453
angryracism	47	0.104682
whiteadv	52	0.115818
fearraces	100	0.222727
racerare	86	0.191545





- No misspellings or formatting issues
- No duplicates
- Removed NAs for Target
- 7337 NA
- 16% rows with NAs
  - No dependencies
  - Randomly Distributed between outputs
- NAs Removed
- 37,561 Data Points Remaining

EDA

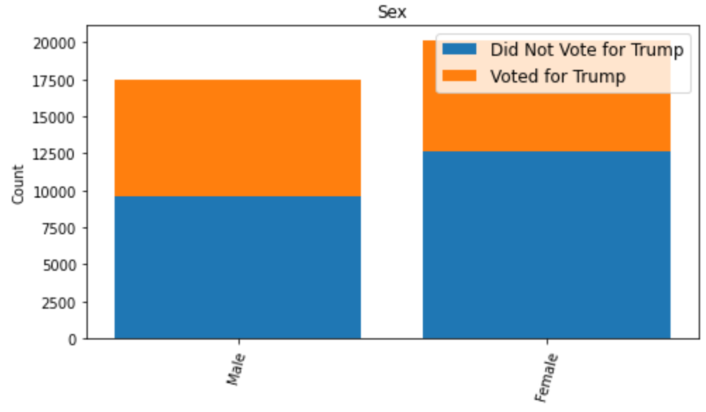
Dataset

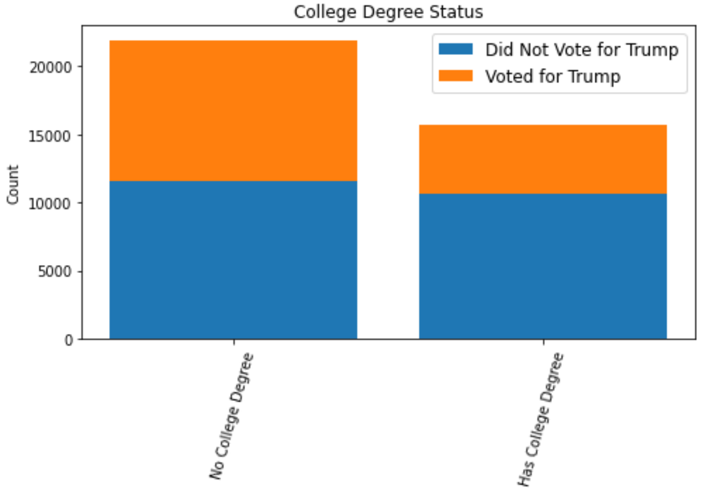
	votetrump	age	female	collegeed	famincr	ideo	pid7na	bornagain	religimp	churchatd	prayerfreq	angryracism	whiteadv	fearraces	racerare
votetrump															
age	0.18														
female	-0.08	-0.10													
collegeed	-0.15	-0.16	-0.07												
famincr	-0.00	-0.04	-0.16	0.34											
ideo	0.60	0.18	-0.08	-0.14	-0.01										
pid7na	0.73	0.12	-0.08	-0.10	0.03	0.66									
bornagain	0.23	0.03	0.04	-0.11	-0.11	0.32	0.21								
religimp	0.29	0.13	0.10	-0.10	-0.08	0.40	0.26	0.50							
churchatd	0.20	0.05	0.03	0.03	0.02	0.34	0.21	0.48	0.72						
prayerfreq	0.24	0.16	0.15	-0.09	-0.09	0.34	0.22	0.47	0.78	0.64					
angryracism	0.40	0.09	-0.15	-0.04	0.04	0.35	0.36	0.06	0.07	0.05	0.04				
whiteadv	0.62	0.14	-0.07	-0.18	-0.04	0.54	0.57	0.17	0.24	0.15	0.21	0.46			
fearraces	0.12	0.03	-0.01	-0.05	-0.03	0.14	0.09	0.07	0.10	0.07	0.06	0.10	0.04		
racerare	0.43	0.02	-0.13	-0.05	0.04	0.42	0.39	0.16	0.19	0.16	0.14	0.40	0.48	0.09	

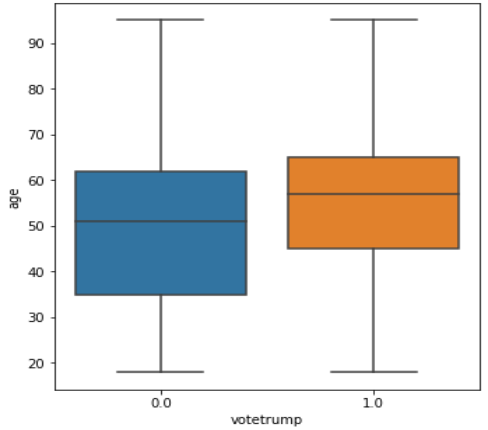
- Ideology Features >60% correlation with a vote for Trump
- Views on whether whites have an advantage correlates 62%
- Views on racism correlates >40%
- Religion views correlate ~ 20-30%
- Demographics , no significant correlation

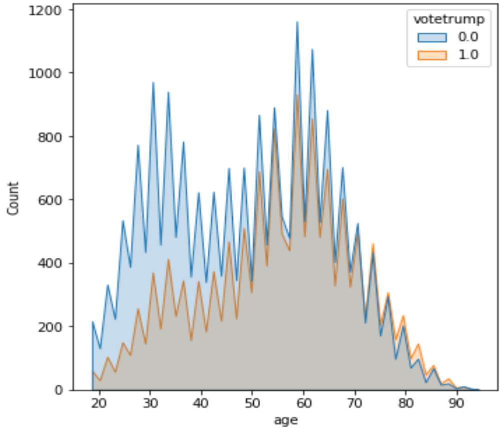
EDA

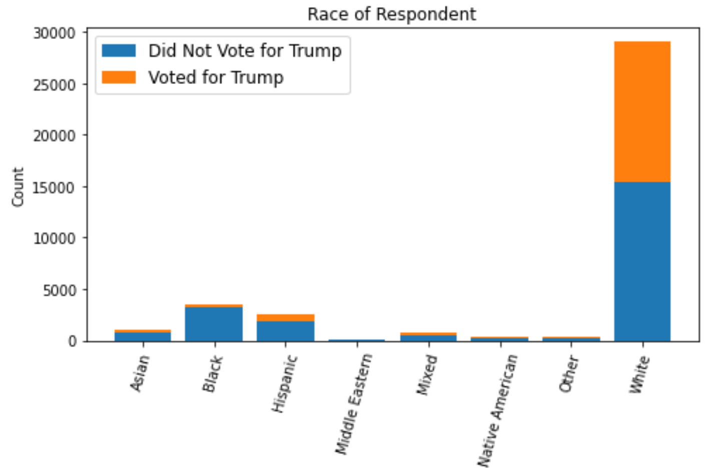
Demographics

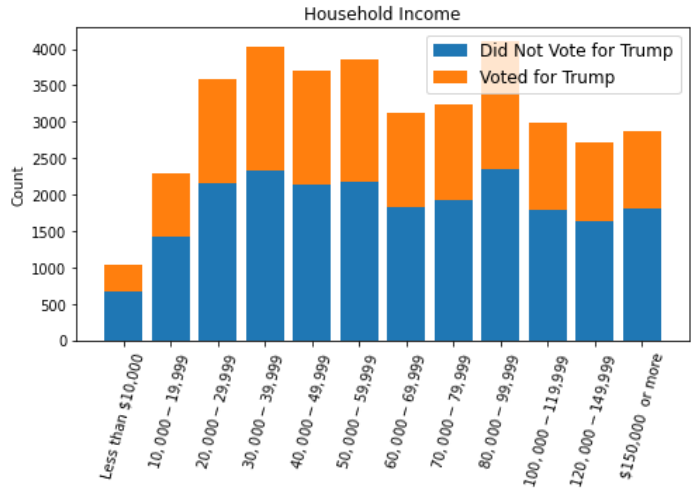










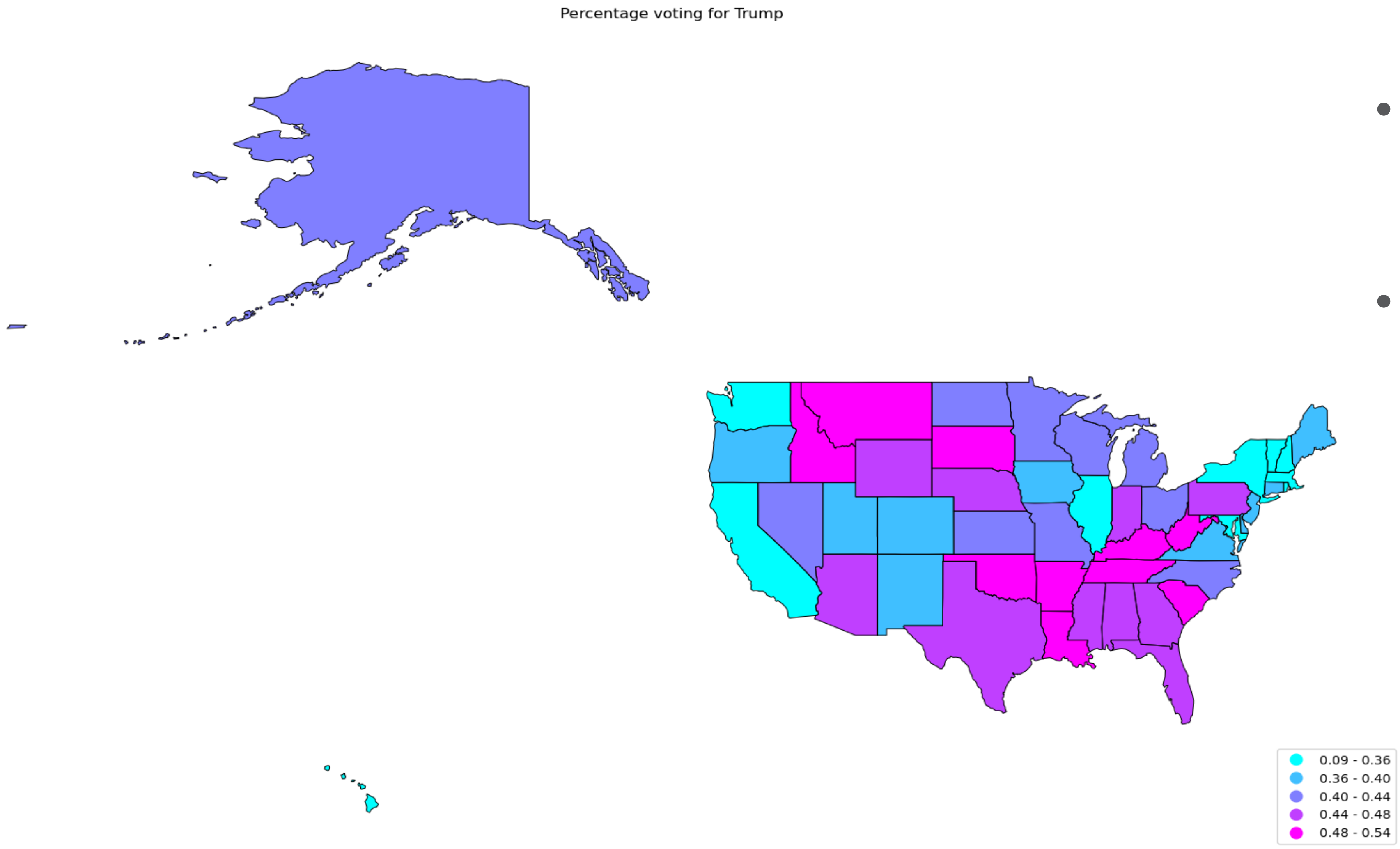


- Younger voters less like to vote for Trump
- Non-white very unlikely to vote for Trump
- Other demographics have minor differences

UNIVERSITY of DENVER
DANIEL FELIX RITCHIE SCHOOL OF ENGINEERING & COMPUTER SCIENCE

EDA

Location

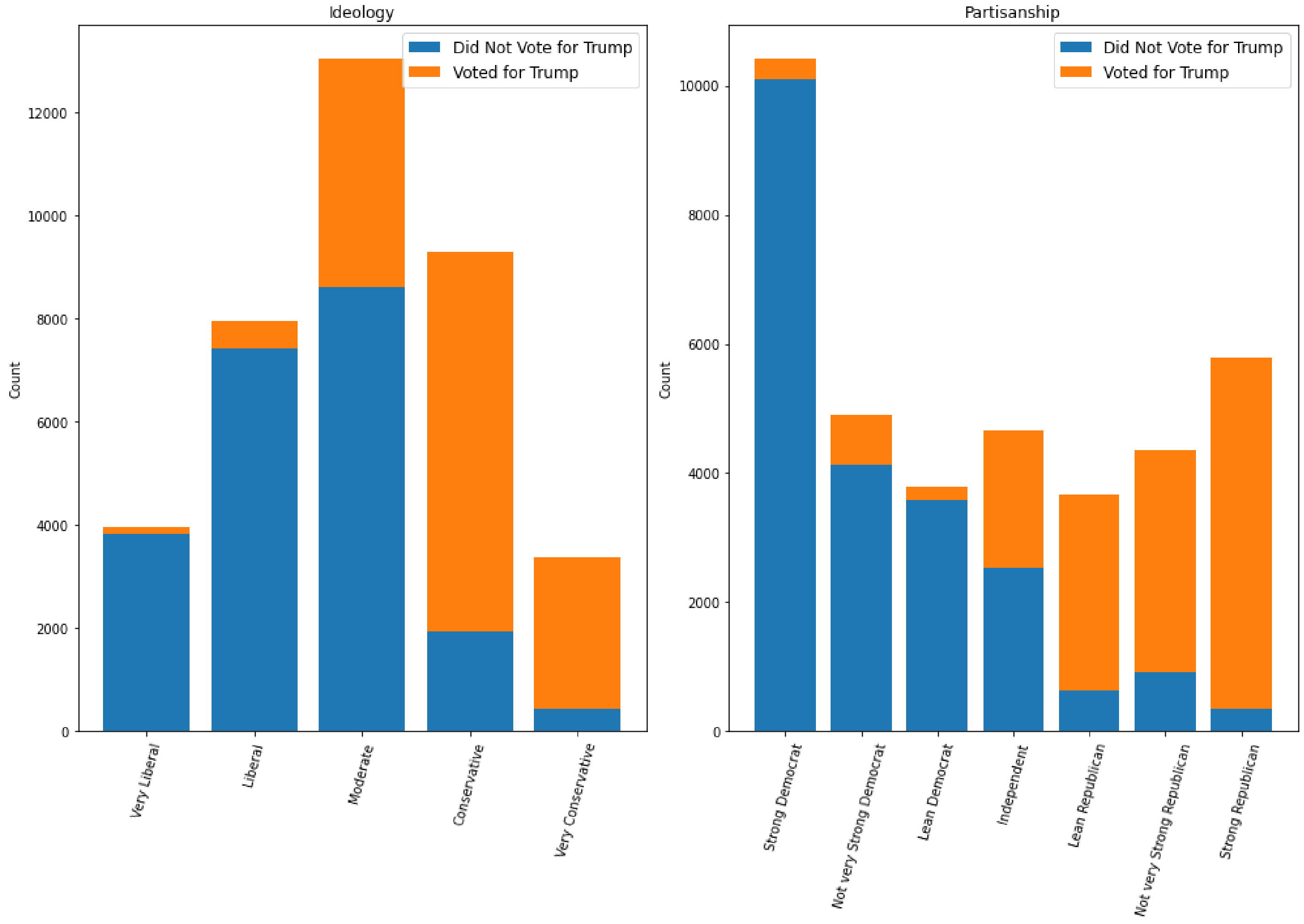


- South and Midwest most likely to vote for Trump
- Most states not very predictive
  - Exception Northeast
  - California
  - Oregon
  - Hawaii

UNIVERSITY of DENVER
DANIEL FELIX RITCHIE SCHOOL OF ENGINEERING & COMPUTER SCIENCE

EDA

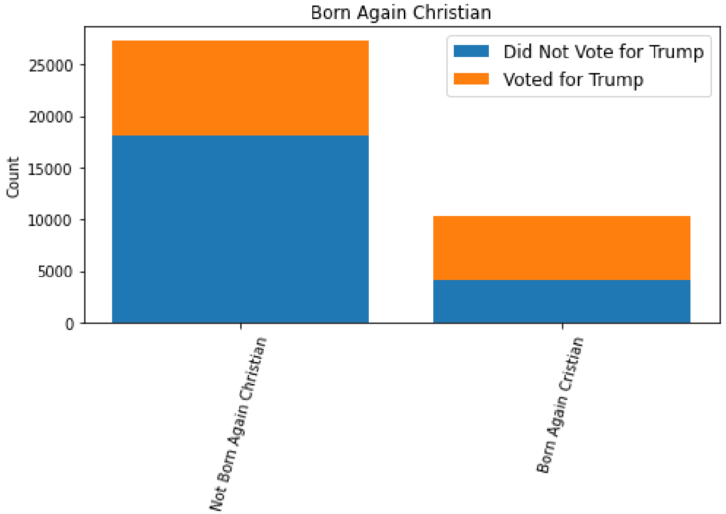
Ideology

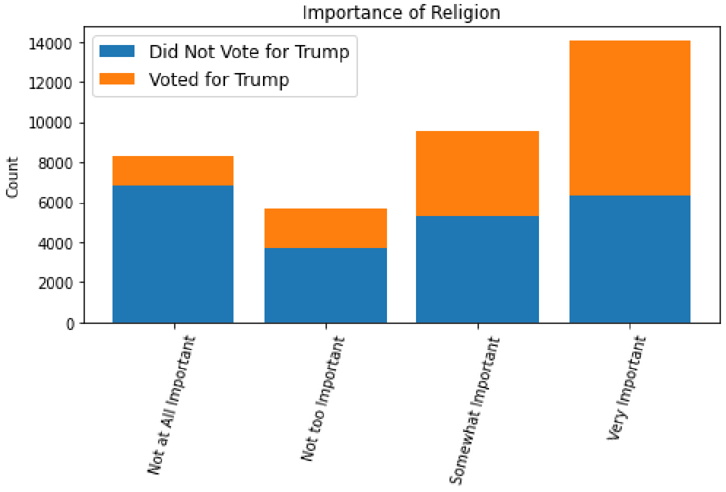


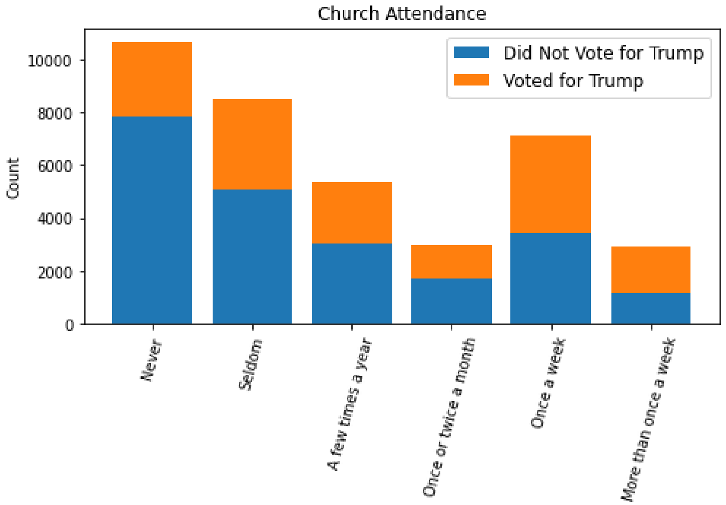
- As expected, the left ideology unlikely to vote for Trump, right likely
- Independents split
- Moderates split by lean to not vote for Trump

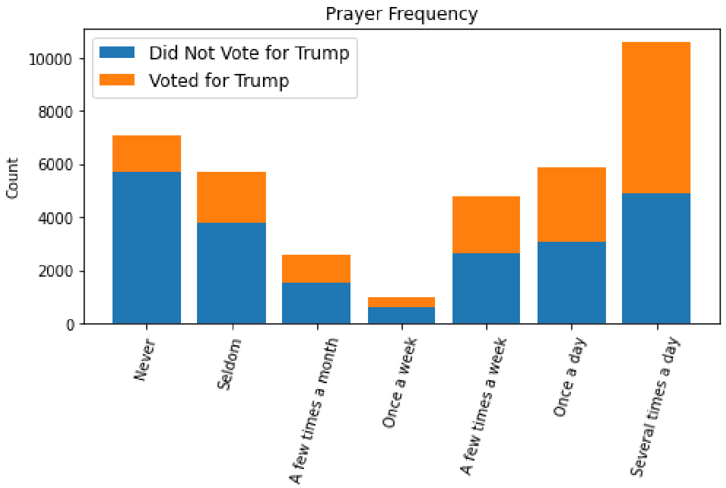
EDA

Views on Religion







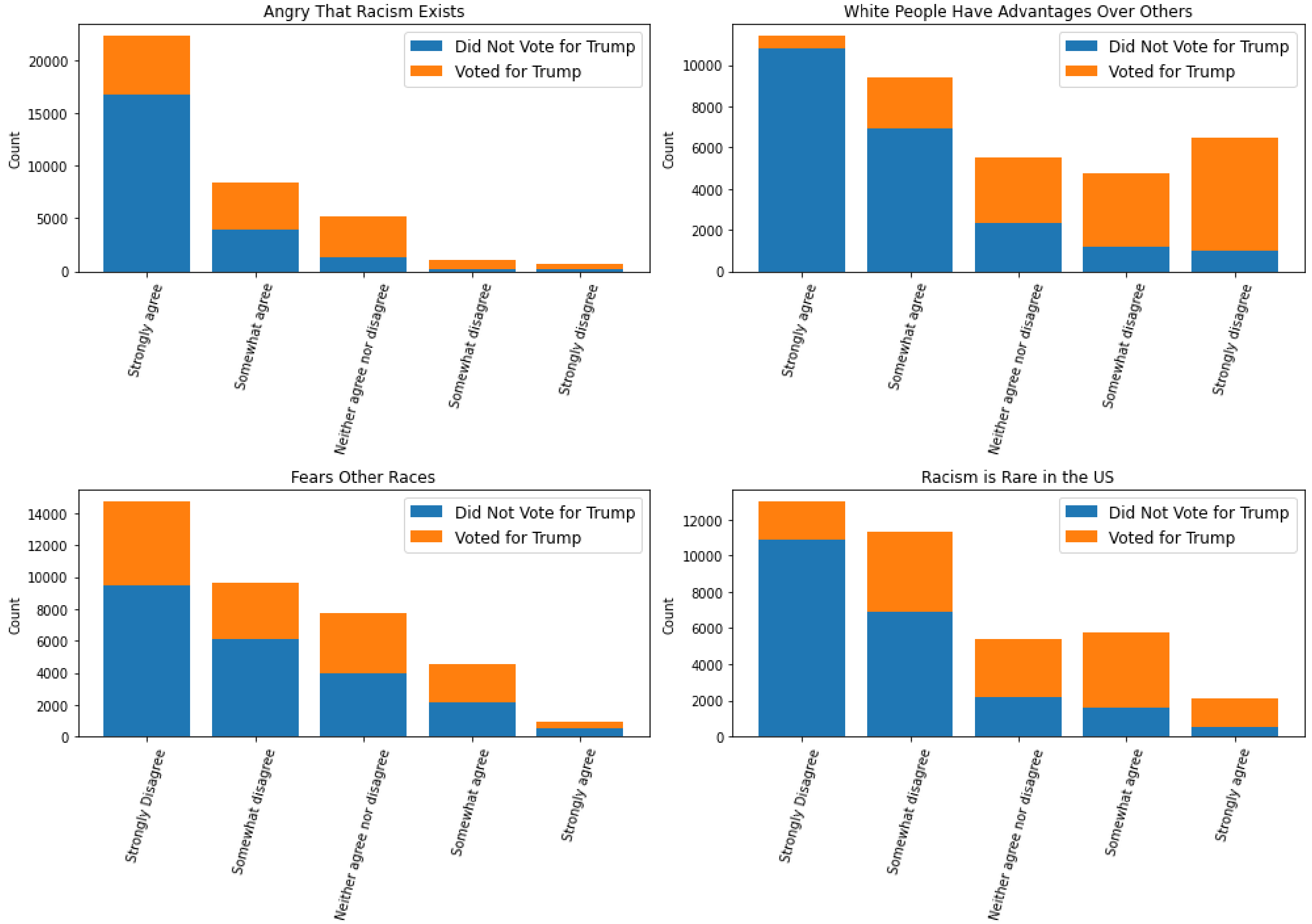


- The more religious, the more likely to vote for Trump
- Church Attendance the affect is a bit less significant
- Prayer frequency much more significant

UNIVERSITY of DENVER
DANIEL FELIX RITCHIE SCHOOL OF ENGINEERING & COMPUTER SCIENCE

EDA

Views on Racism

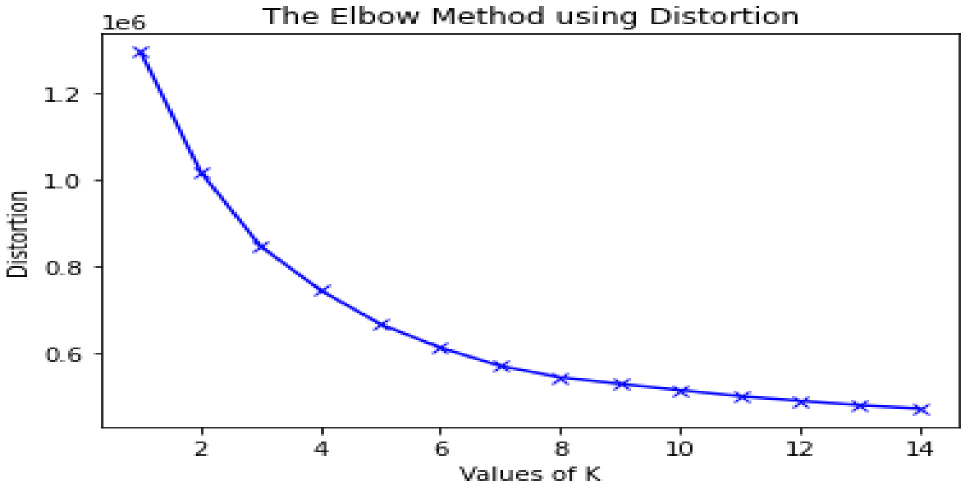


- Most angry that racism exists, slight correlation.
  - May indicate different definitions
- The rejection of white privilege highly correlated with a vote for Trump
- Fearing other races not correlated (Likely people did not admit to this)
- Rejecting the commonality of racism correlates with a vote for Trump

UNIVERSITY of DENVER
DANIEL FELIX RITCHIE SCHOOL OF ENGINEERING & COMPUTER SCIENCE

Model Evaluation

Preprocessing



- Clustering using KNN used to see if there is an effect on performance
- K for KNN chosen to be 6
- Train Test split 80:20
- Clustering and non-clustered datasets

```
from sklearn.model_selection import train_test_split

# Non Cluster
X = tv.drop(columns=['votetrump'])
y = tv['votetrump']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.20, random_state = 32)


# Cluster
X_clust = tv_clust.drop(columns=['votetrump'])
y_clust = tv_clust['votetrump']

X_train_clust, X_test_clust, y_train, y_test = train_test_split(X_clust, y_clust, test_size = 0.20, random_state = 32)
```

Model Evaluation

Multi-Model Comparison

```
dfs = []
models = [
        ('LogReg', LogisticRegression(max_iter= 10000)),
        ('RF', RandomForestClassifier()),
        ('KNN', KNeighborsClassifier()),
        ('GNB', GaussianNB())
    ]
names = []
target_names = ['Did Not Vote for Trump', 'Voted for Trump']
scores = []
for name, model in models:
    kfold = model_selection.KFold(n_splits=5, shuffle=True, random_state=42)
    cv_results = model_selection.cross_validate(model, X_train, y_train, cv=kfold, scoring=['accuracy','f1'])
    clf = model.fit(X_train, y_train)
    y_pred_test = clf.predict(X_test)

    scores.append([name,cv_results['test_accuracy'].mean(),accuracy_score(y_test, y_pred_test),
                    cv_results['test_f1'].mean(),f1_score(y_test, y_pred_test)])

results = pd.DataFrame(data= scores,columns=['Model','Train_Accuracy','Test_Accuracy', 'Train_F1', 'Test_F1'])
results.sort_values(by=['Test_Accuracy'])
```

- 4 Models Trained
- Random Forest and Logistic Regression Best Performers
- Clustering did not improve results
- Optimize models without Clustering
- No Overfitting

Without Clustering

	Model	Train_Accuracy	Test_Accuracy	Train_F1	Test_F1
1	RF	0.892572	0.892054	0.868902	0.867635
0	LogReg	0.890042	0.890057	0.866023	0.865341
2	KNN	0.878794	0.877679	0.852389	0.851126
3	GNB	0.810536	0.815786	0.782754	0.788767

With Clustering

	Model	Train_Accuracy	Test_Accuracy	Train_F1	Test_F1
1	RF	0.891440	0.893917	0.867699	0.870343
0	LogReg	0.890675	0.889924	0.866744	0.865331
2	KNN	0.879093	0.878477	0.852876	0.851906
3	GNB	0.812234	0.818315	0.784607	0.791889

Model Evaluation

Model Optimization/ Selection

```
logR = LogisticRegression(max_iter= 10000)

param_grid_logR = {'C' : [0.001,0.01,0.1,1,10,100]}


grid_logR = GridSearchCV(estimator=logR, param_grid=param_grid_logR,scoring='None',cv=8)
grid_logR = grid_logR.fit(X_train,y_train)
```

```
rf = RandomForestClassifier(n_jobs=14)

param_grid_rf = {'max_depth': [1,2,4,8,16],
            'n_estimators':[100,150,200],
            'max_features':['sqrt',1,5,10,20],
            'min_samples_split': [1,2,4,6]}


grid_rf = GridSearchCV(estimator=rf, param_grid=param_grid_rf,scoring='None',cv=8)
grid_rf = grid_rf.fit(X_train,y_train)
```

	Model	Accuracy	F1
0	Random Forest	0.895381	0.872403
1	Log Reg	0.895381	0.872403

- Both Models perform equally
- Select simplest model
- Select Logistic Regression
    - Parameter C: 0.1

Conclusion

Final thoughts Lessons Learned

- Ideology, views on white advantages, and the race of the respondent were the most significant predictors
- Income not as significant as initially thought
- There seems to be an accuracy ceiling of about 90%
  - The 10% may not be captured in the questions analyzed
  - Fuller datasets with full survey answers would be useful
- Logistic Model was the simplest and performed the best