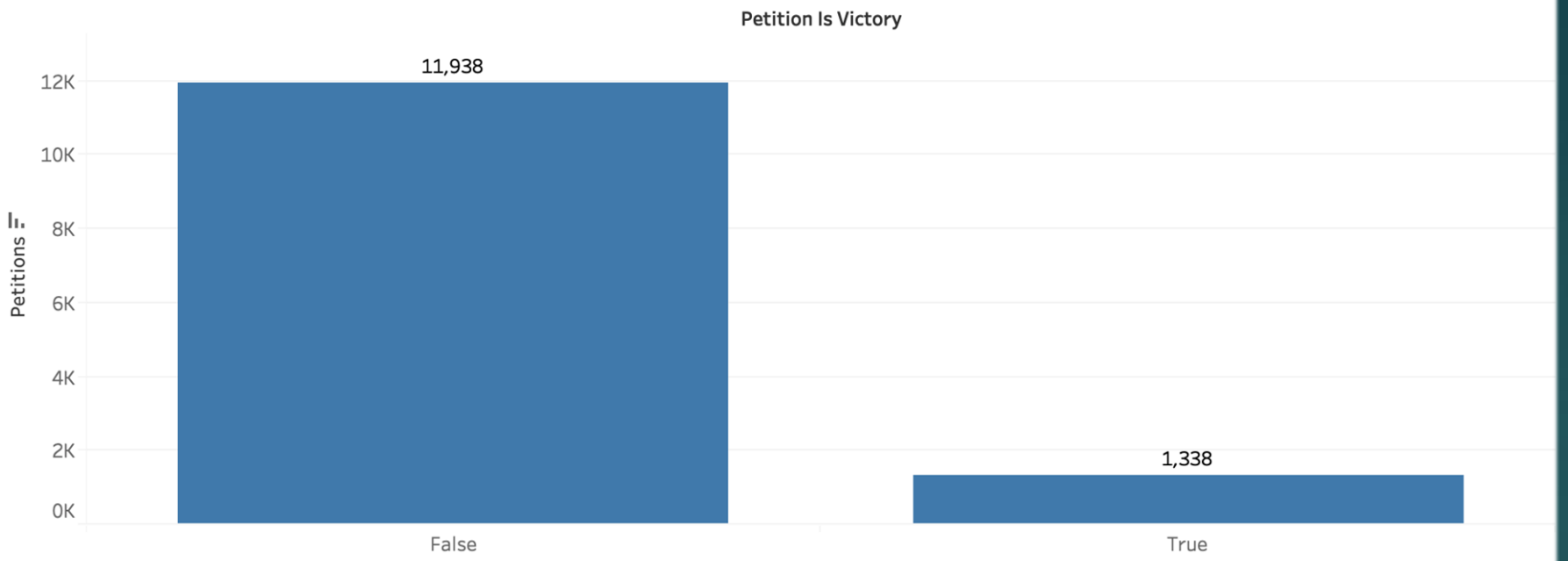
PDC Hackathon

TEAM : HIMANSHU JAGTAP

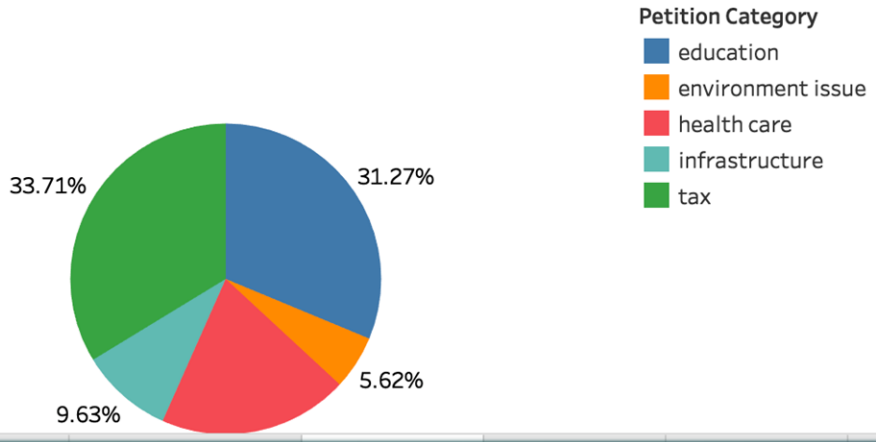
SUSHANT SONAR

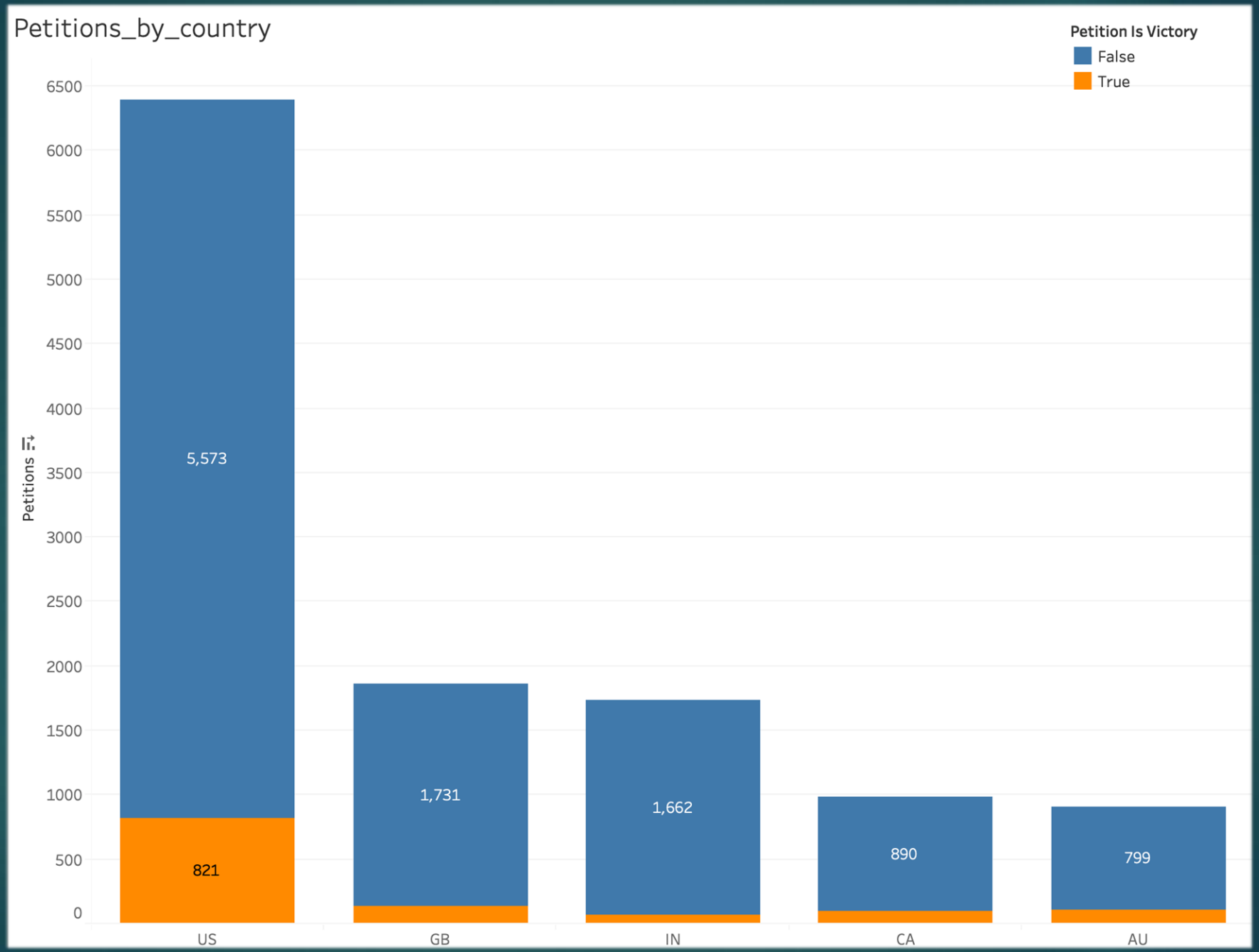
HANCEL PV

Petition_result_distribution

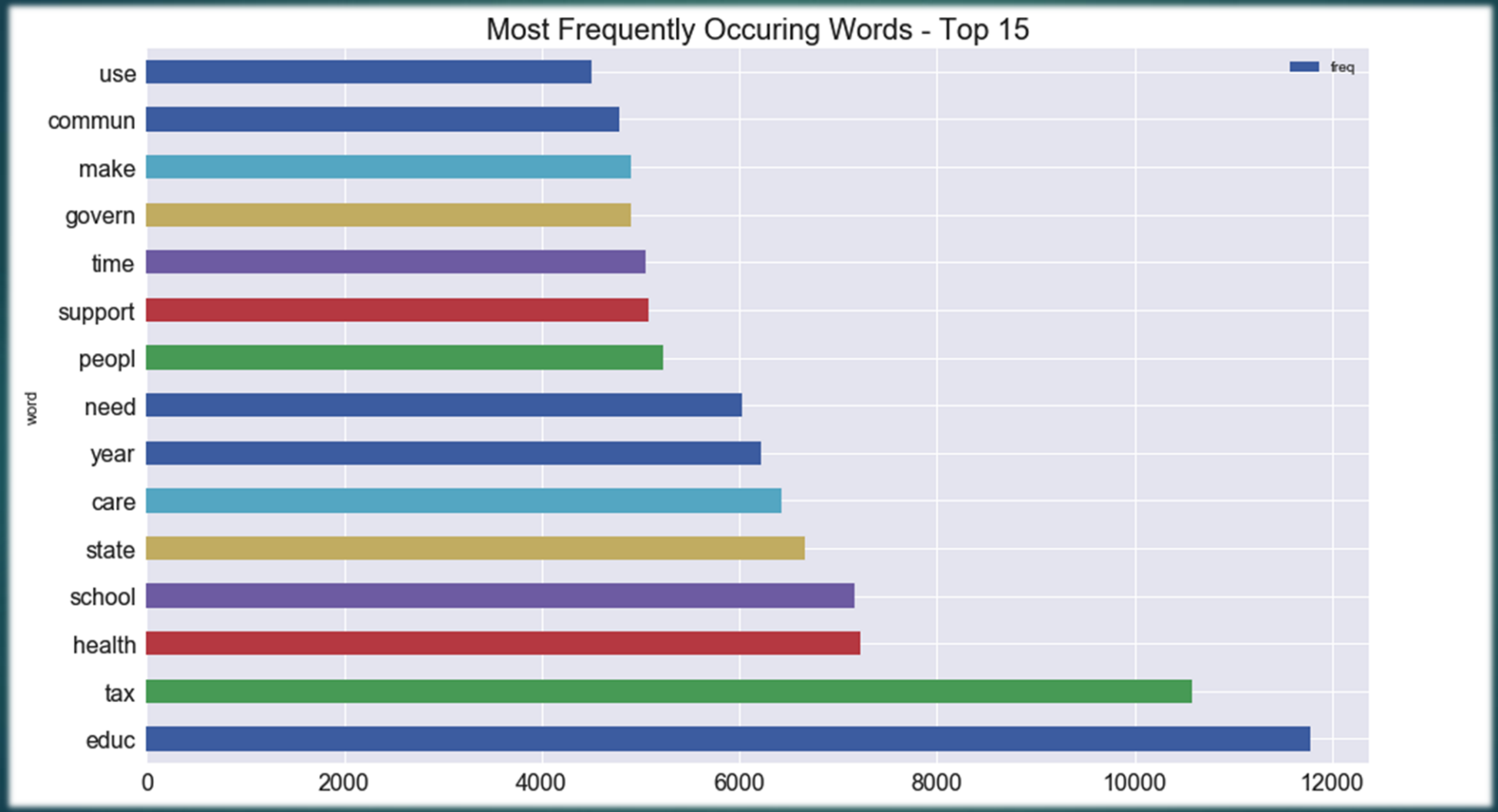


Petitions_by_category




Petitions_by_country
Petition Is Victory
False
True
Petitions
6500
6000
5500
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
5,573
821
1,731
1,662
890
799
US
GB
IN
CA
AU

Frequently Occuring Words



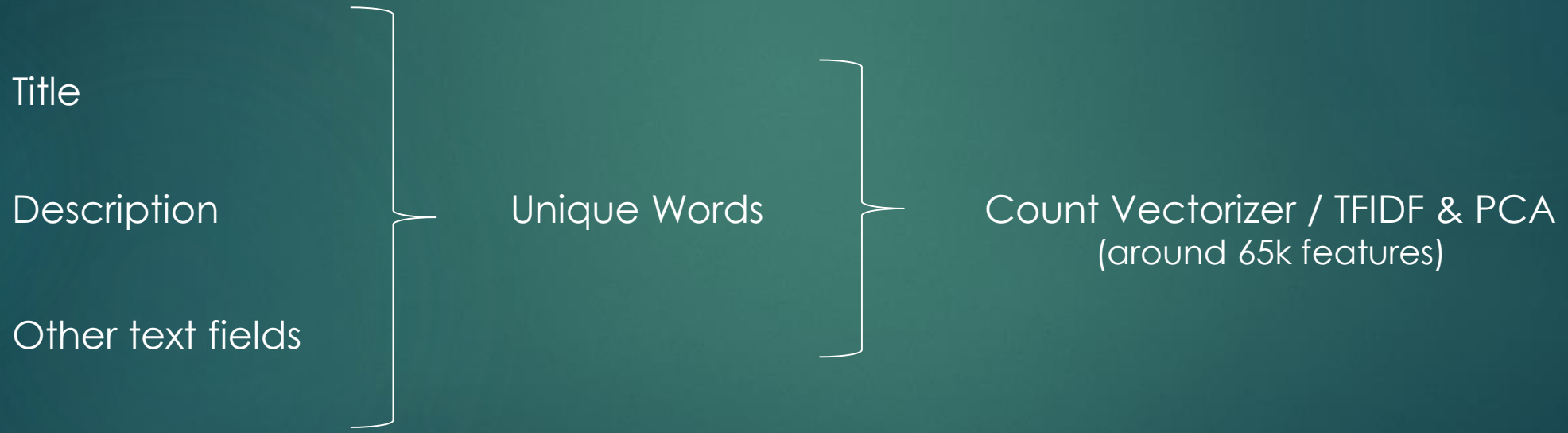
Word Cloud


WordCloud - Consolidated Vocabulary
educ
health
tax
school
state
peopl
year
need
care
support
govern
make
student
commun
time
work
mani
famili
current
allow
infrastructur
chang
right
petit
servic
children
nation
public
pleas
live
help
sign
issu
increas
use
program
includ
provid
protect
continu
citizen
human
board
like
depart
creat
stop
pay
high
countri
know
fund
develop
place
person
american
resid
class
reduc
come
day
india
action
requir
group
learn
level
propos
medic
inform
order
world
request
meet
report
district
number
everi
cost
build
univers
decis
howev
organ
social
respons
plan
water
anim
look
long
let
vote
park
receiv
million
case
abl
month
thing
share
becom
effect
individu
money
happen
end
say
process
futur
home
believ
new
ask
sever
teacher
parent
free
import
area
demand
citi
council
concern
environ
presid
want
job
addit
benefit
busi
life
elect
hous
member
gener
reason
follow
mean
law
recent
fact
act
counti
offic
unit
impact
pass
minist
alreadi
remov
result
forc
road
polici
incom
rais
differ
problem
start
given
stand
consid
access
sinc
better
way
save
local
good
child
thank
face
high
0
100
200
300
400
500
600
800
1000

Text Features

Pre Processing

- Removed urls, hastags, multiple spaces, html tags, punctuations, stopwords,
- Stemming



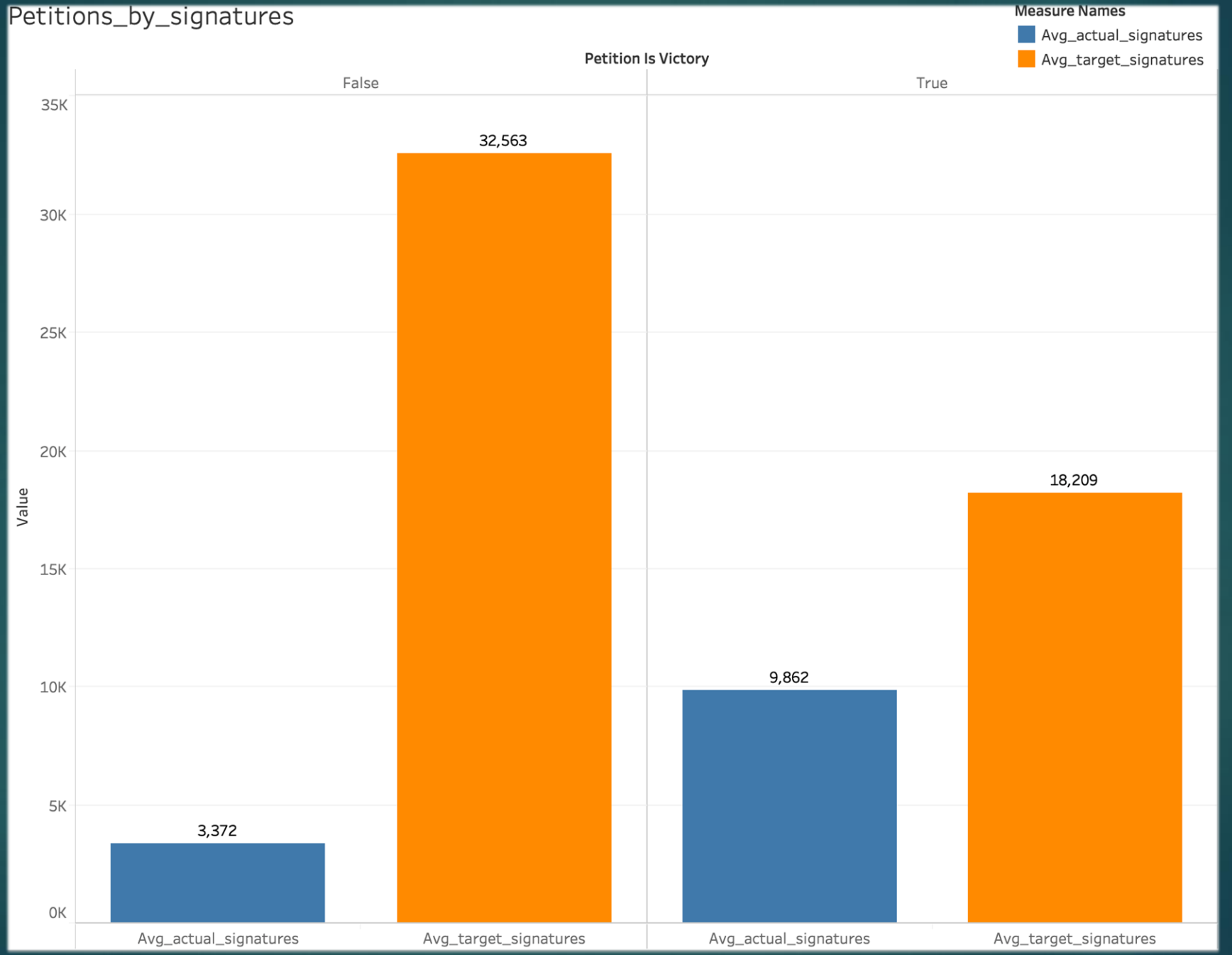
Numerical Features

- No. of words in Petition Description
- No. of words in Petition Title
- Signature Ratio

  Defn : Signatures Collected/Target

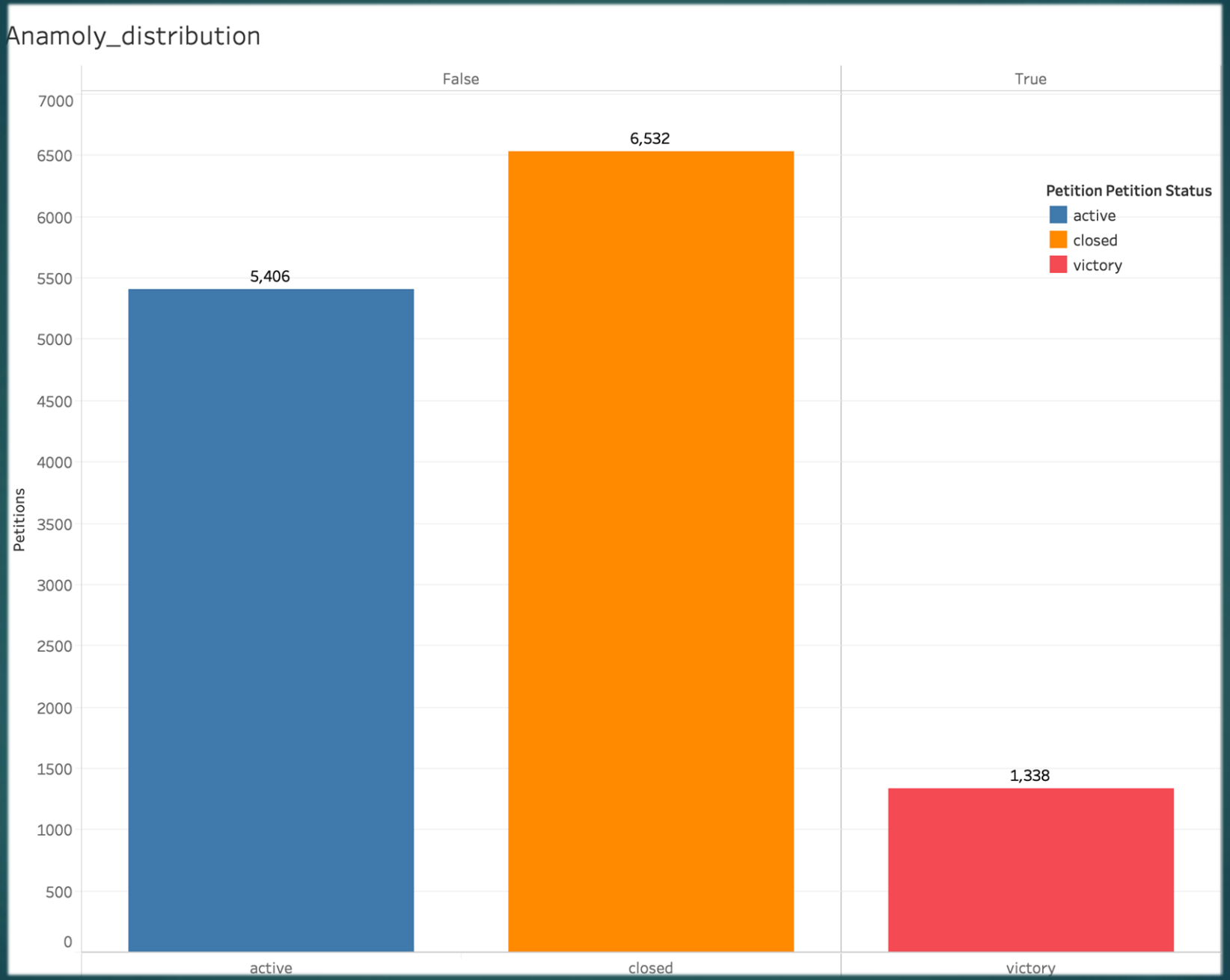
- Ratio of Victory with respect to Sponsorship

  Defn : Victory Percent w.r.t Sponsored and Non Sponsored Petitions


Petitions_by_signatures
Measure Names
Avg_actual_signatures
Avg_target_signatures
Petition Is Victory
False
True
35K
30K
25K
20K
15K
10K
5K
0K
Value
32,563
3,372
9,862
18,209
Avg_actual_signatures
Avg_target_signatures
Avg_actual_signatures
Avg_target_signatures

Observation

- We observed that there a was a feature called ‘petition_status’ which was leaking information about the target variable.
- All the ’active’ petitions are tagged as False for Petition_is_victory, in the train set.
- Using the petition_status feature, one can achieve 100% accuracy on test set, as this becomes the sole driver for predictions.
- In our models, we have filtered out the ‘active’ category records before training.
- Please refer to next slide for the details and graph.


Anamoly_distribution
False
True
7000
6500
6000
5500
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
Petitions
6,532
5,406
1,338
Petition Petition Status
active
closed
victory
active
closed
victory

Models

- Petition Category is predicted using the available features
- Predicted value of Petition Category is used as a feature to predict petition victory

Models

Cross Validation

- Stratified K-Fold Cross Validation with 3 folds

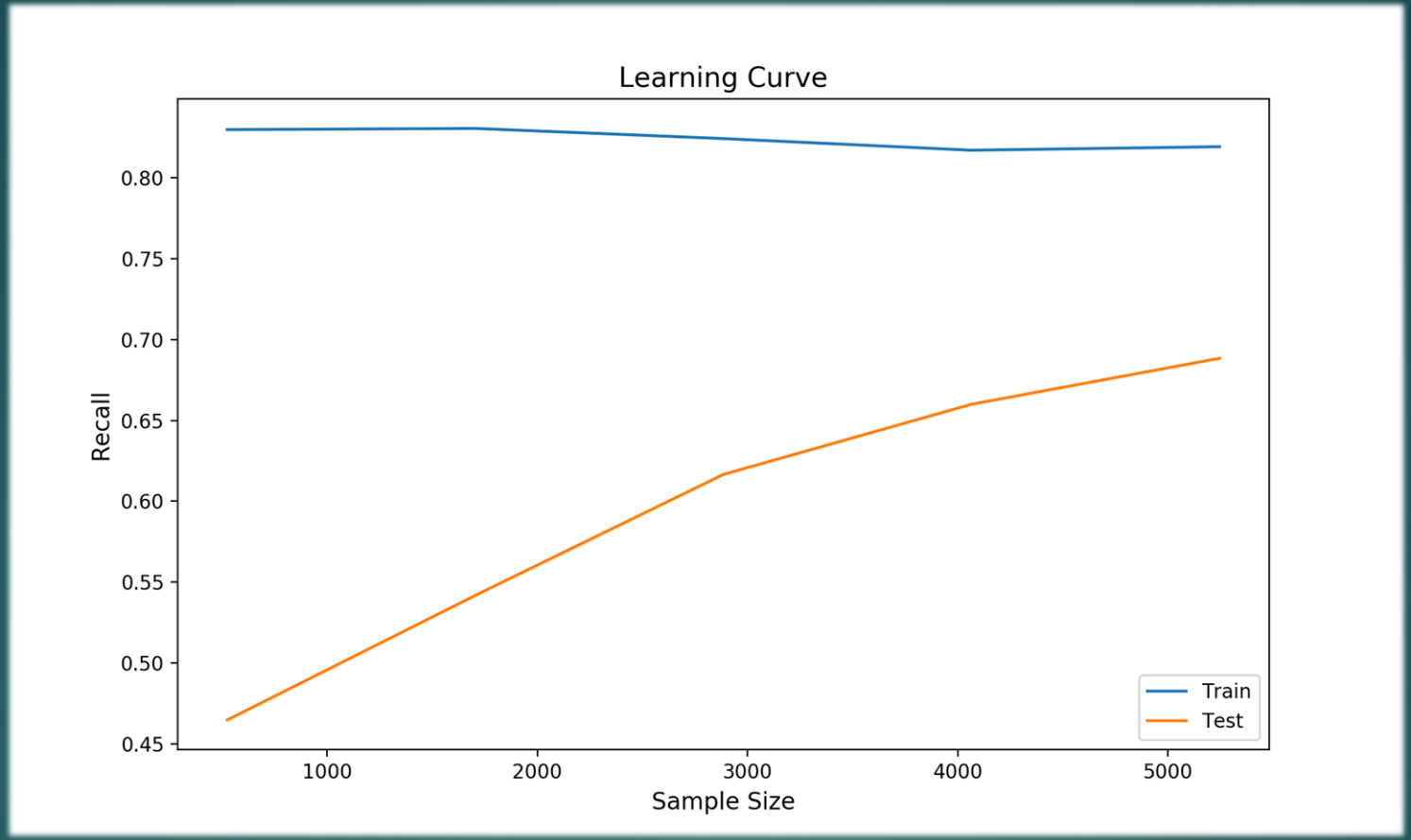
https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html

- Parameter optimized : F1 Score

https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html

- Parameters are optimized using GridSearch

Validation Curve



- We can infer from the plot that, as the sample size increases, the recall for the test set is also increasing.
- This is a strong indicator, that the results of the model will improve on providing more data.

Results

- Model 1 – petition_category

Model	F1 Score	Accuracy
Logistic Regression	99.01	98.47

- Model 2 – petition_is_victory

Model	F1 Score	Recall for positive class
Logistic Regression	0.66	0.75
XGBoost Classifier	0.69	0.66
Voting Classifier (Soft)	0.74	0.67

* Voting Classifier is a combination of Log Regression & XGB Classifier.

Future Scope

- From learning curve, it is clear that model is overfitting. To control for the same, we can either increase no. of datapoints or reduce the number of features further
- Use better vector representation of textual data like n-grams, word-embeddings
- Train model using other machine learning models like LSTM with attention mechanism