Assignment 2 APA track and TESU Masters (R/Python users)

Q1. [3 points] Build a decision tree or segmentation model predicting partisanship.

a. **Describe which of the party dependent variable you will be using and why.**
b. **Describe which indicators you will suppress from the model and why.**

a. party dependent variables being used

VG_14_DV	Voted in the general election 11/2014. Y if they voted. N if they were registered but did not vote. Blank if not registered as of 11/14. Note - this dv field is in Y/N format. VG_14 in the indicators is 0/1 and should be suppressed when modeling VG_14_DV
D2	Democrat 2-way. Y if voter is a Democrat, N if Republican, blank if independent or minor party
R2	Republican 2-way. Y if voter is a Republican, N if Democrat, blank if independent or minor party
D3	Democrat 3-way. Y if voter is a Democrat, N if Republican, independent or minor party
R3	Republican 3-way. Y if voter is a Republican, N if Democrat, independent or minor party
I3	Independent 3-way. Y if voter is an independent or minor party member. N if Democrat or Republican

b.

indicators you will suppress

NH_WHITE
NH_AA
NH_NATAM
NH_ASIAN
NH_HPI
NH_OTHER
NH_MULT
HISP
COMM_LT10
COMM_609P
MED_HH_INC
COMM_CAR
COMM_CP
COMM_PT
COMM_WALK

Assignment 2 APA track and TESU Masters (R/Python users)

KIDS
KIDS_MC
M_NEV_MAR
M_MAR
M_MAR_SP
M_MAR_SNP
F_NEV_MAR
F_MAR
F_MAR_SP
F_MAR_SNP
ED_ASSOC
ED_BACH
ED_MD
ED_PROF
ED_DOC
ED_4COL
GENDER_F
GENDER_M
H_AFDLN3P
H_AFSSLN3P
H_F1
H_FFDLN2
H_FFSLN2
H_M1
H_MFDLN2
H_MFDLN3P
H_MFSLN2
H_MFSLN3P
H_MFSSLN3P
H_MMDLN2
H_MMSLN2
PARTY_D
PARTY_I
PARTY_R
HHP_D
HHP_DD
HHP_DI
HHP_DR
HHP_I
HHP_II

Assignment 2 APA track and TESU Masters (R/Python users)

HHP_R
HHP_RI
HHP_RR
VPP_12
VPP_16
VPR_12
VPR_14
VPR_16
VG_08
VG_10
VG_12
VG_14
VG_16
PP_PELIG
PR_PELIG
AP_PELIG
G_PELIG
E_PELIG
NL5G
NL3PR
NL5AP
NL2PP
REG_DAYS
UpscaleBuy
UpscaleMal
UpscaleFem
BookBuyerI
FamilyMaga
FemaleOrie
ReligiousM
GardeningM
CulinaryIn
HealthFitn
DoItYourse
FinancialM
ReligiousC
PoliticalC
MedianEduc
PRS16_PD
PRS16_PR

Assignment 2 APA track and TESU Masters (R/Python users)

Submit: Along with answers to (a) and (b), also submit

a. a model definition that can be implemented using the voterfile indicator data.

Model definition:- Decision Tree Classifier

Input data:-

OPP_SEX
AGE
HH_ND
HH_NR
HH_NI
MED_AGE
MED_AGE_M
MED_AGE_F
NH_WHITE
NH_AA
NH_NATAM
NH_ASIAN
NH_HPI
NH_OTHER
NH_MULT
HISP

Output (dependent variables):-

VG_14_DV
D2
R2
D3
R3
I3

Assignment 2 APA track and TESU Masters (R/Python users)

b. a tab delimited, excel or other common format file listing the voter_ID and model score for each voter. The model score should be in the range 0 to 100.

Here is the predicted score for first 100 voters,

Predict for multiple observations,

Output variable, y = FX_indicators_2020_df_new['D2_Y']

prob = clf.predict_proba(X_test[0:100])

```
array([[0., 1.],
       [1., 0.],
       [1., 0.],
       [0., 1.],
       [0., 1.],
       [0., 1.],
       [0., 1.],
       [0., 1.],
       [0., 1.],
       [0., 1.],
       [1., 0.],
       [0., 1.],
       [0., 1.],
       [1., 0.],
       [0., 1.],
       [1., 0.],
       [0., 1.],
       [1., 0.],
       [1., 0.],
       [0., 1.],
       [0., 1.],
       [0., 1.],
       [1., 0.],
       [0., 1.],
       [1., 0.],
       [0., 1.],
       [0., 1.],
       [1., 0.],
       [0., 1.],
       [0., 1.],
       [0., 1.],
       [0., 1.],
       [1., 0.],
       [0., 1.],
       [0., 1.],
       [0., 1.],
       [0., 1.],
```

Assignment 2 APA track and TESU Masters (R/Python users)

```
[0., 1.],
[0., 1.],
[0., 1.],
[0., 1.],
[0., 1.],
[1., 0.],
[0., 1.],
[0., 1.],
[1., 0.],
[0., 1.],
[0., 1.],
[1., 0.],
[0., 1.],
[1., 0.],
[0., 1.],
[1., 0.],
[0., 1.],
[0., 1.],
[1., 0.],
[0., 1.],
[0., 1.],
[0., 1.],
[0., 1.],
[1., 0.],
[0., 1.],
[0., 1.],
[1., 0.],
[1., 0.],
[1., 0.],
[0., 1.],
[1., 0.],
[0., 1.],
[0., 1.],
[1., 0.],
[0., 1.],
[1., 0.],
[1., 0.],
[0., 1.],
[1., 0.],
[0., 1.],
[0., 1.],
[0., 1.],
[1., 0.],
[0., 1.],
[1., 0.],
[0., 1.],
[0., 1.],
[1., 0.],
[1., 0.],
[1., 0.],
[0., 1.],
```

Assignment 2 APA track and TESU Masters (R/Python users)

```
[0., 1.],
[0., 1.],
[1., 0.],
[0., 1.],
[0., 1.],
[1., 0.],
[0., 1.],
[0., 1.],
[0., 1.],
[1., 0.],
[0., 1.],
[0., 1.]])
```

c. R or Python code you used. (See attached file)

Q1 [3 points] Build a decision tree predicting partisanship.

Assignment 2 APA track and TESU Masters (R/Python users)

Q2. [3 points] Build a logistic regression predicting partisanship.

a. **Describe which party dependent variable you will be using and why .**

   Same as Q1

b. **Describe which indicators you will suppress from the model and why.**

   Same as Q1

Submit: Along with answers to (a) and (b), also submit

a model definition that can be implemented using the voterfile indicator data.

Model definition:- Logistic regression

Input data:-

OPP_SEX
AGE
HH_ND
HH_NR
HH_NI
MED_AGE
MED_AGE_M
MED_AGE_F
NH_WHITE
NH_AA
NH_NATAM
NH_ASIAN
NH_HPI
NH_OTHER
NH_MULT
HISP

Assignment 2 APA track and TESU Masters (R/Python users)

Output (dependent variables) :-

VG_14_DV
D2
R2
D3
R3
I3

b. a tab delimited, excel or other common format file listing the voter_ID and model score for each voter. The model score should be in the range 0 to 100.

Here is the predicted score for first 100 voters,
Predict for multiple observations,
Output variable, y = FX_indicators_2020_df_new['D2_Y']

```
prob = logreg.predict_proba(X_test[0:100])
```

```
array([[0.36649819, 0.63350181],
       [0.35557967, 0.64442033],
       [0.38286689, 0.61713311],
       [0.38448646, 0.61551354],
       [0.35201947, 0.64798053],
       [0.38005647, 0.61994353],
       [0.34797492, 0.65202508],
       [0.36088116, 0.63911884],
       [0.36173476, 0.63826524],
       [0.37803327, 0.62196673],
       [0.37626537, 0.62373463],
       [0.34665739, 0.65334261],
       [0.37895216, 0.62104784],
       [0.35365187, 0.64634813],
       [0.36658645, 0.63341355],
       [0.36881575, 0.63118425],
       [0.3998148 , 0.6001852 ],
       [0.40381716, 0.59618284],
       [0.37148861, 0.62851139],
       [0.38129579, 0.61870421],
       [0.37637112, 0.62362888],
       [0.38620881, 0.61379119],
```

Assignment 2 APA track and TESU Masters (R/Python users)

```
[0.39100798, 0.60899202],
[0.376874  , 0.623126  ],
[0.38785209, 0.61214791],
[0.34246204, 0.65753796],
[0.36667952, 0.63332048],
[0.37492729, 0.62507271],
[0.38327612, 0.61672388],
[0.37804038, 0.62195962],
[0.36663227, 0.63336773],
[0.3794972 , 0.6205028 ],
[0.36928639, 0.63071361],
[0.39505837, 0.60494163],
[0.36221214, 0.63778786],
[0.40968086, 0.59031914],
[0.36338309, 0.63661691],
[0.37766958, 0.62233042],
[0.41286603, 0.58713397],
[0.35998714, 0.64001286],
[0.38254325, 0.61745675],
[0.38488398, 0.61511602],
[0.35075816, 0.64924184],
[0.3508171 , 0.6491829 ],
[0.41069812, 0.58930188],
[0.40444855, 0.59555145],
[0.35359856, 0.64640144],
[0.38750034, 0.61249966],
[0.35857808, 0.64142192],
[0.40801714, 0.59198286],
[0.37687341, 0.62312659],
[0.36173019, 0.63826981],
[0.36035987, 0.63964013],
[0.35633788, 0.64366212],
[0.37941422, 0.62058578],
[0.37990665, 0.62009335],
[0.40567422, 0.59432578],
[0.3688912 , 0.6311088 ],
[0.35412268, 0.64587732],
[0.35321153, 0.64678847],
[0.39058938, 0.60941062],
[0.39647648, 0.60352352],
[0.34685538, 0.65314462],
[0.36025635, 0.63974365],
[0.36505831, 0.63494169],
[0.37777163, 0.62222837],
[0.37075151, 0.62924849],
[0.35481505, 0.64518495],
[0.37626983, 0.62373017],
[0.41413124, 0.58586876],
[0.38724111, 0.61275889],
[0.36837809, 0.63162191],
[0.38124817, 0.61875183],
```

Assignment 2 APA track and TESU Masters (R/Python users)

```
[0.37496948, 0.62503052],
[0.37548551, 0.62451449],
[0.40629296, 0.59370704],
[0.39770152, 0.60229848],
[0.34762547, 0.65237453],
[0.39671686, 0.60328314],
[0.39272698, 0.60727302],
[0.3937235 , 0.6062765 ],
[0.3963592 , 0.6036408 ],
[0.38861034, 0.61138966],
[0.37030513, 0.62969487],
[0.36091096, 0.63908904],
[0.37994463, 0.62005537],
[0.35186324, 0.64813676],
[0.40244798, 0.59755202],
[0.36805501, 0.63194499],
[0.35172708, 0.64827292],
[0.38779475, 0.61220525],
[0.37224029, 0.62775971],
[0.39365134, 0.60634866],
[0.38236211, 0.61763789],
[0.37938696, 0.62061304],
[0.37151693, 0.62848307],
[0.36795888, 0.63204112],
[0.39566584, 0.60433416],
[0.38328899, 0.61671101],
[0.38793247, 0.61206753]])
```

c. R or Python code you used. (See attached file)

Q2 [3 points] Build a logistic regression predicting partisanship.

Assignment 2 APA track and TESU Masters (R/Python users)

Q3. [3 points] Build a model predicting turnout.

a. **In the simulation, the upcoming election is 11/14. The models will be built to predict turnout in the 11/10 election for use in predicting turnout in the upcoming 11/14 election. Explain why we would want to predict turnout in 11/10 rather than the more recent 11/12 election.**

   We would want to predict turnout in 11/10 rather than the more recent 11/12 election because it is general presidential election upcoming in 11/14 and need turnout data for 11/10 presidential election to compare with.

b. **Describe which indicators you will suppress from the model and why?**

   indicators you will suppress

NH_WHITE
NH_AA
NH_NATAM
NH_ASIAN
NH_HPI
NH_OTHER
NH_MULT
HISP
COMM_LT10
COMM_609P
MED_HH_INC
COMM_CAR
COMM_CP
COMM_PT
COMM_WALK
KIDS
KIDS_MC
M_NEV_MAR
M_MAR
M_MAR_SP
M_MAR_SNP
F_NEV_MAR
F_MAR
F_MAR_SP

Assignment 2 APA track and TESU Masters (R/Python users)

F_MAR_SNP
ED_ASSOC
ED_BACH
ED_MD
ED_PROF
ED_DOC
ED_4COL
GENDER_F
GENDER_M
H_AFDLN3P
H_AFSSLN3P
H_F1
H_FFDLN2
H_FFSLN2
H_M1
H_MFDLN2
H_MFDLN3P
H_MFSLN2
H_MFSLN3P
H_MFSSLN3P
H_MMDLN2
H_MMSLN2
PARTY_D
PARTY_I
PARTY_R
HHP_D
HHP_DD
HHP_DI
HHP_DR
HHP_I
HHP_II
HHP_R
HHP_RI
HHP_RR
VG_10
PP_PELIG
PR_PELIG
AP_PELIG
G_PELIG
E_PELIG

Assignment 2 APA track and TESU Masters (R/Python users)

NL5G
NL3PR
NL5AP
NL2PP
REG_DAYS
UpscaleBuy
UpscaleMal
UpscaleFem
BookBuyerI
FamilyMaga
FemaleOrie
ReligiousM
GardeningM
CulinaryIn
HealthFitn
DoItYourse
FinancialM
ReligiousC
PoliticalC
MedianEduc
PRS16_PD
PRS16_PR

Submit: Along with answers to (a) and (b), also submit

1. **a model definition that can be implemented using the voterfile indicator data.**

Model definition:- Logistic Regression

Assignment 2 APA track and TESU Masters (R/Python users)

Input data:-

H_F1
H_FFDLN2
H_FFSLN2
H_M1
H_MFDLN2
H_MFDLN3P
H_MFSLN2
H_MFSLN3P
H_MFSSLN3P
H_MMDLN2
H_MMSLN2
PARTY_D
PARTY_I
PARTY_R
HHP_D
HHP_DD

VG_10

Output (dependent variables):-

VG_14_DV

2. **a tab delimited, excel or other common format file listing the voter_ID and model score for each voter. The model score should be in the range 0 to 100.**

Here is the predicted score for first 100 voters,
Predict for multiple observations,

y = FX_indicators_2020_df_new['VG_14_DV_Y']

prob = logreg.predict_proba(X_test[0:100])

```
array([[0.46026005, 0.53973995],
       [0.4499961 , 0.5500039 ],
       [0.40116768, 0.59883232],
```

Assignment 2 APA track and TESU Masters (R/Python users)

```
[0.42193183, 0.57806817],
[0.36156994, 0.63843006],
[0.41022164, 0.58977836],
[0.46166972, 0.53833028],
[0.42799743, 0.57200257],
[0.47989737, 0.52010263],
[0.49978097, 0.50021903],
[0.42009635, 0.57990365],
[0.42952228, 0.57047772],
[0.43665795, 0.56334205],
[0.45723728, 0.54276272],
[0.49895898, 0.50104102],
[0.49797114, 0.50202886],
[0.41053118, 0.58946882],
[0.38936271, 0.61063729],
[0.48470134, 0.51529866],
[0.40513993, 0.59486007],
[0.4346949 , 0.5653051 ],
[0.41561879, 0.58438121],
[0.41026864, 0.58973136],
[0.36220669, 0.63779331],
[0.42084104, 0.57915896],
[0.48994987, 0.51005013],
[0.4633573 , 0.5366427 ],
[0.44633286, 0.55366714],
[0.43667058, 0.56332942],
[0.48130692, 0.51869308],
[0.45876911, 0.54123089],
[0.4987457 , 0.5012543 ],
[0.44557911, 0.55442089],
[0.41499552, 0.58500448],
[0.35832738, 0.64167262],
[0.45938694, 0.54061306],
[0.4597174 , 0.5402826 ],
[0.44648278, 0.55351722],
[0.44323869, 0.55676131],
[0.46080008, 0.53919992],
[0.37664521, 0.62335479],
[0.36951713, 0.63048287],
[0.41917619, 0.58082381],
[0.49203875, 0.50796125],
[0.40626808, 0.59373192],
[0.36309564, 0.63690436],
[0.42913387, 0.57086613],
[0.45340178, 0.54659822],
[0.48613928, 0.51386072],
[0.43341922, 0.56658078],
[0.37057838, 0.62942162],
[0.39167138, 0.60832862],
[0.4882744 , 0.5117256 ],
[0.37846818, 0.62153182],
```

Assignment 2 APA track and TESU Masters (R/Python users)

```
[0.4768054 , 0.5231946 ],
[0.41363442, 0.58636558],
[0.41047508, 0.58952492],
[0.38736752, 0.61263248],
[0.44965637, 0.55034363],
[0.39554292, 0.60445708],
[0.43223459, 0.56776541],
[0.40488355, 0.59511645],
[0.46376725, 0.53623275],
[0.4502964 , 0.5497036 ],
[0.4281639 , 0.5718361 ],
[0.4594279 , 0.5405721 ],
[0.39775966, 0.60224034],
[0.46941786, 0.53058214],
[0.3614075 , 0.6385925 ],
[0.39093387, 0.60906613],
[0.36010785, 0.63989215],
[0.47422446, 0.52577554],
[0.45574036, 0.54425964],
[0.46123681, 0.53876319],
[0.45883623, 0.54116377],
[0.38681161, 0.61318839],
[0.36214022, 0.63785978],
[0.48458782, 0.51541218],
[0.39665493, 0.60334507],
[0.45594857, 0.54405143],
[0.36208222, 0.63791778],
[0.41931186, 0.58068814],
[0.39931089, 0.60068911],
[0.43145396, 0.56854604],
[0.36873204, 0.63126796],
[0.46943019, 0.53056981],
[0.43997932, 0.56002068],
[0.40006593, 0.59993407],
[0.40953939, 0.59046061],
[0.38133715, 0.61866285],
[0.45687139, 0.54312861],
[0.42099516, 0.57900484],
[0.40389077, 0.59610923],
[0.4581496 , 0.5418504 ],
[0.48432622, 0.51567378],
[0.40334594, 0.59665406],
[0.40003139, 0.59996861],
[0.41020834, 0.58979166],
[0.41008973, 0.58991027],
[0.45994911, 0.54005089]])
```

Assignment 2 APA track and TESU Masters (R/Python users)

3. **R or Python code you used.**

Q3. [3 points] Build a model predicting turnout (LogisticRegression)¶

Q4. [3 points] Presidential vs off-year turnout.

Even though this is a presidential election year, we are building models predicting turnout in an off-year.

How does off-year turnout differ from presidential years, and how would models predicting off-year turnout likely be different from a presidential year model?

1. Here is the predicted score for first 100 voters (predicting turnout in an off-year 2010) for 2014 presidential year

y = FX_indicators_2020_df_new['VG_14_DV_Y']

logreg.predict(X_test[0:100])

```
array([1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1],
```

2. a presidential year model (2014) as per data file

y[0:100]

```
3     1
4     1
8     0
9     1
14    1
..
```

Assignment 2 APA track and TESU Masters (R/Python users)

```
199    1
200    1
204    1
205    1
206    1
```

Q5. [3 points] Build a different model predicting turnout.

a. **Describe how the model is different from the model built for question 3. Did you use a different algorithm? Are the indicators used different?**

   The model/ algorithm is different from the model built for question 3 (logistic regression) as we use now Decision Tree Classifier model.

   Indicators are same.

b. **Describe which indicators you will suppress from the model and why.**

   Same as Q3

Submit: Along with answers to (a) and (b), also submit

1. **a model definition that can be implemented using the voterfile indicator data.**

   Model definition: - Decision Tree Classifier model.

   Input data:-

Assignment 2 APA track and TESU Masters (R/Python users)

H_F1
H_FFDLN2
H_FFSLN2
H_M1
H_MFDLN2
H_MFDLN3P
H_MFSLN2
H_MFSLN3P
H_MFSSLN3P
H_MMDLN2
H_MMSLN2
PARTY_D
PARTY_I
PARTY_R
HHP_D
HHP_DD

VG_10

Output (dependent variables):-

VG_14_DV

2. a tab delimited, excel or other common format file listing the voter_ID and model score for each voter. The model score should be in the range 0 to 100.

   Here is the predicted score for first 100 voters,
   # Predict for multiple observations,

   y = FX_indicators_2020_df_new['VG_14_DV_Y']

   prob = clf.predict_proba(X_test[0:100])

```
[0., 1.],
[1., 0.],
[0., 1.],
[0., 1.],
[0., 1.],
[0., 1.],
[0., 1.],
```

Assignment 2 APA track and TESU Masters (R/Python users)

```
[1., 0.],
[1., 0.],
[1., 0.],
[1., 0.],
[0., 1.],
[1., 0.],
[1., 0.],
[0., 1.],
[1., 0.],
[1., 0.],
[1., 0.],
[0., 1.],
[0., 1.],
[0., 1.],
[1., 0.],
[0., 1.],
[1., 0.],
[1., 0.],
[0., 1.],
[1., 0.],
[0., 1.],
[0., 1.],
[1., 0.],
[0., 1.],
[0., 1.],
[1., 0.],
[1., 0.],
[1., 0.],
[0., 1.],
[0., 1.],
[0., 1.],
[0., 1.],
[1., 0.],
[0., 1.],
[0., 1.],
[0., 1.],
[0., 1.],
[1., 0.],
[0., 1.],
[0., 1.],
[0., 1.],
[1., 0.],
[1., 0.],
[1., 0.],
[1., 0.],
[1., 0.],
[0., 1.],
[1., 0.],
[1., 0.],
[1., 0.],
[0., 1.],
```

Assignment 2 APA track and TESU Masters (R/Python users)

```
[0., 1.],
[0., 1.],
[0., 1.],
[1., 0.],
[0., 1.],
[0., 1.],
[0., 1.],
[1., 0.],
[0., 1.],
[0., 1.],
[0., 1.],
[0., 1.],
[1., 0.],
[1., 0.],
[0., 1.],
[1., 0.],
[1., 0.],
[1., 0.],
[1., 0.],
[0., 1.],
[1., 0.],
[1., 0.],
[1., 0.],
[1., 0.],
[0., 1.],
[1., 0.],
[0., 1.],
[0., 1.],
[1., 0.],
[1., 0.],
[0., 1.],
[0., 1.],
[1., 0.],
[1., 0.],
[1., 0.],
[1., 0.],
[0., 1.],
[1., 0.],
[1., 0.],
[0., 1.],
[1., 0.],
[1., 0.]])
```

3. **R or Python code you used.**

Q4. [3 points] Build a model predicting turnout (Decision Tree Classifier model)

Assignment 2 APA track and TESU Masters (R/Python users)

Q6. [5 points]

a. **Build versions of the models from questions 1 and 2 using the small dataset and the full dataset.**

   Version 1 the small dataset: FX_indicators_2020_rand_10k (decision tree)

   Version 2 the small dataset: FX_indicators_2020_rand_10k (regression model)

   Version 3 the full dataset: FX_indicators_2020 (decision tree)

   Version 4 the full dataset: FX_indicators_2020 (regression model)

   for predicting partisanship

b. **Describe the differences that you see in both the range of scores and granularity (how many distinct values there are for each score).**

   Please see attached Python codes for build versions of the models.

c. **Which method, segmentation or logistic regression seemed to benefit the most from using the larger dataset?**

   logistic regression seemed to benefit the most from using the larger dataset.

   Submit: Along with answers to (a) and (b), also submit

1. **a tab delimited, excel or other common format file listing the voter_ID and model score for each voter. The model score should be in the range 0 to 100.**

   Please see attached Python codes for build versions of the models.

   Here is the predicted score for first 100 voters,
   Predict for multiple observations,
   Output variable, y = FX_indicators_2020_df_new['D2_Y']

   prob = logreg.predict_proba(X_test[0:100])

Assignment 2 APA track and TESU Masters (R/Python users)

prob = clf.predict_proba(X_test[0:100])

2. **R or Python code you used.**

   See attached files for python codes.