Predicating U.S. Presidential Election Result on County level

ECON 4130 ECONOMIC ANALYSIS FOR SOCIAL NETWORKS

YAM TSZ CHUNG

1155122291

Motivation

- The 2020 U.S. presidential election is still on-going
- A hot topic for social study
- Many online articles and studies about the elections

Motivation

- Pattern for voting on state level (My own naive observation):
  - Moderate income and multi-racial tends to vote for Biden
  - Lower income and white people tends to vote for Trump
- Can I do my own prediction on the election result using limited resource and data?
- How accurate the economic variables can predict the result
- It will be a Classification process

Target Goals

- Using multiple economic variables to predict the county vote result in 2016
  - Find the most accurate model to predict the result
- Try to predict which county will change their vote
  - Find the most accurate model to predict the vote changes

Data – the dependent variable

- 2020 election is not yet finished so in this paper we will use 2016 result to train and test our model
  - US presidential election in 2016 between Republican(Trump) and Democrat(Clinton)
  - We assign Republican win as “1” and Democrat win as “0”, this will be used in our first analysis of predicting vote result

Data – the dependent variable

- The Swing Counties
    - Compare 2012 with 2016 county vote result and assign the change:
        - "1" as those voted for Republican in 2012 but turn to Democratic in 2016
        - "2" as those voted for Democratic in 2012 but turn to Republican in 2016
        - and the remaining unchanged as "0"
    - Total of 238 counties flipped their vote from 2012 to 2016
        - 20 of them changed from Republican to Democratic → "1"
        - 218 of them changed from Democratic to Republican → "2"
    - Can those economic variables predict which counties will flip their vote?

Data – the independent variables

- We can expect there is a high autocorrelation in election, so the result in 2012 will be used as a dummy variable in our models

- Same, We assign Republican win as “1” and Democrat win as “0” in 2012

Data – the independent variables

- County level of data will be used in this paper, Kaggle provided most of the dataset for us
  - 49(independent variables) x 3112(counties)
  - Including many economic variables such as: Population, CPI, Racial, Gender, Unemployment rate etc. for each county.

Models

- We will use 3 common classification methods in this paper:
  - Logistic Regression, Decision Tree and K Nearest Neighbor
- To look for the best model and parameters to predict the election result
- Grid Search Cross Validation from scikit-learn to tune the hyperparameters and get the best model from the training set
  - A 10-Fold cross validation to chose the best parameters
- Using train_test_split method to split the dataset with testing size = 0.25 (778counties)

Models

- Hyperparameters to tune:
- Logistic Regression
  - Regularization Constant: [0.001, 0.01, 0.1, 1, 10, 100, 1000]
  - Penalty: [L2, L1]
- K Nearest Neighbor
  - Number of Neighbors: [1, 3, 7, 9, 11, 13, 15]
  - Weights: [Uniform, Distance]
  - Metric: [Euclidean, Manhattan]
- Decision Tree
  - Criterion: [Gini, Entropy]
  - Maximum Depth: [2, 4, 6, 8, 10]
  - Minimum Samples Split: [2, 4, 6, 8, 10]
  - Minimum Samples Leaf: [2, 4, 6, 8, 10]

Results

The best model to predict which candidate win in each county

- Logistic Regression ★★★
  - C = 0.1, Penalty = L2
  - With the Training set accuracy: 0.962 <<< Highest among other two methods
  - F1 score in Testing set = 0.96 <<< Highest among other two methods
- K Nearest Neighbor
  - Metric = Euclidean, No. of Neighbors = 13, Weights = Distance
  - With the Training set accuracy: 0.952
  - F1 score in Testing set = 0.94
- Decision Tree
  - Criterion = Entropy, Max_depth = 10, Min_samples_leaf = 10, Min_samples_split = 8
  - With the Training set accuracy: 0.955
  - F1 score in Testing set = 0.95

Results
The best model to predict which county will change their vote

- Logistic Regression
  - C = 0.1, Penalty = L2
  - With the Training set accuracy: 0.966
  - F1 score in Testing set predicting changes:
    - No change from 2012 to 2016   = 0.97
    - from Republican to Democratic = 0.00
    - from Democratic to Republican = 0.63
    - Overall accuracy: 0.95

- K Nearest Neighbor
  - Metric = Euclidean, No. of Neighbors = 13, Weights = Uniform
  - With the Training set accuracy: 0.963
  - F1 score in Testing set predicting change:
    - No change from 2012 to 2016   = 0.97
    - from Republican to Democratic = 0.00
    - from Democratic to Republican = 0.64
    - Overall accuracy: 0.95

- Decision Tree ★★
  - Criterion = Gini, Max_depth = 4, Min_samples_leaf = 10, Min_samples_split = 10
  - With the Training set accuracy: 0.964
  - F1 score in Testing set predicting changes:
    - No change from 2012 to 2016 = 0.98
    - from Republican to Democratic = 0.00
    - from Democratic to Republican = 0.70 <<< Highest among other two methods
    - Overall accuracy: 0.95

Conclusions & Suggestions

- The variables are accurate when predicting which candidate will win the county vote

- But not that accurate when predicting the swing counties
  - The best model provide us with a 70% accuracy only on swing from Democratic to Republican
  - It fail to predict those swing from Republican to Democratic

- We can rely on the model to predict who win in each county but not as an election strategy because it fail to predict the changes

Conclusions & Suggestions

- Control for more variables which seems to stable across time
  - E.g.: Racial percentage, Gender
- More variable to account for:
  - Election campaign held by each candidate in each county
  - Percentage of people using social media regularly(?)
- Will the policies implement by the last president have huge effect on how the people will vote?
  - Performance of last president
  - Economic Condition

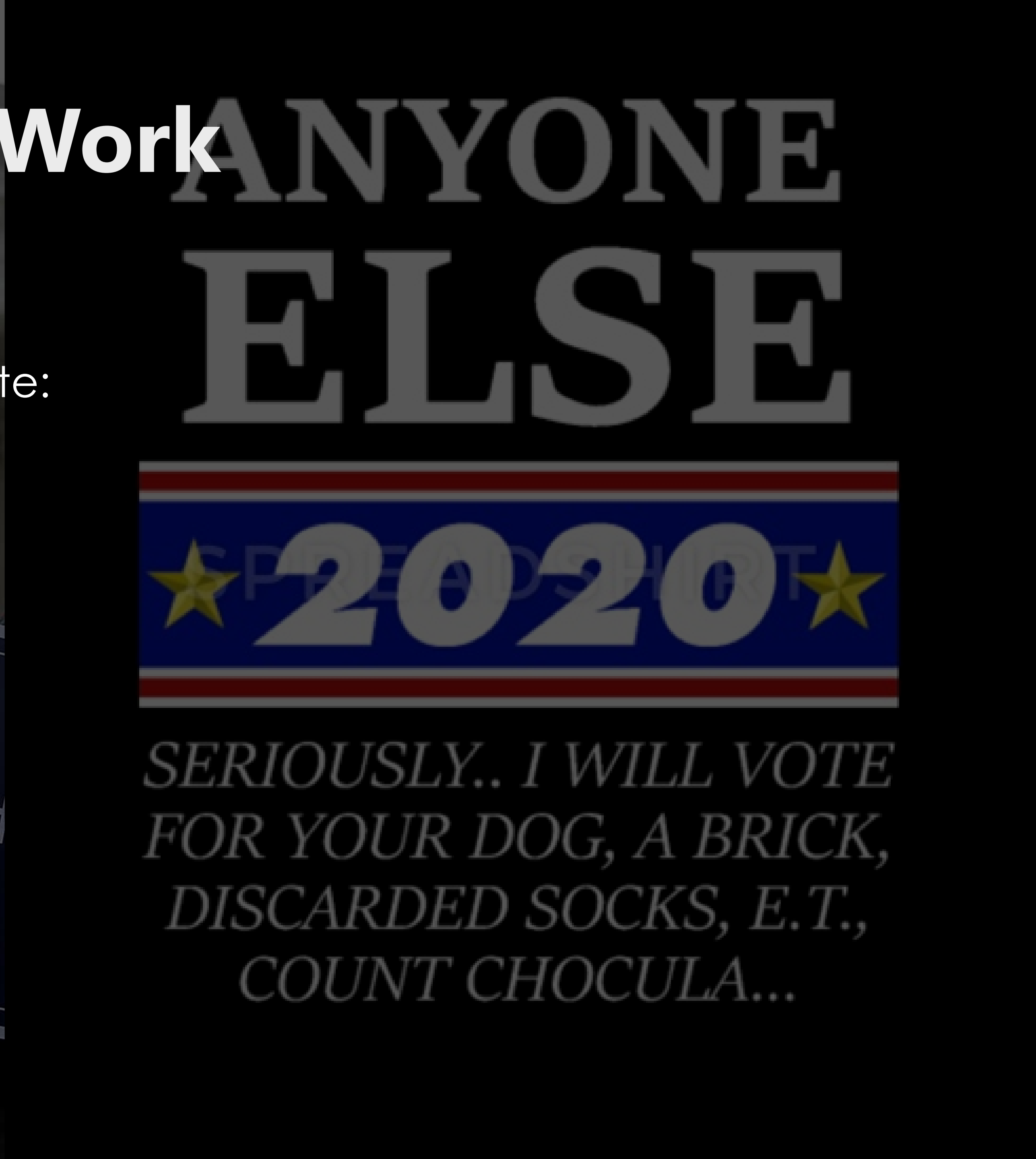
Conclusion & Future Work

- After applying the models to the 2020 result, will it be overfitting or accuracy as it was
  - If high autocorrelation is true, then the result should be close as the prediction in this paper
  - Number of votes increased significantly in 2020 (152,507,250) compare with 2016 (123,716,997)
  - Social media and its influence growth so much in these 4 years, how this will affect our model and prediction from 2016 to 2020

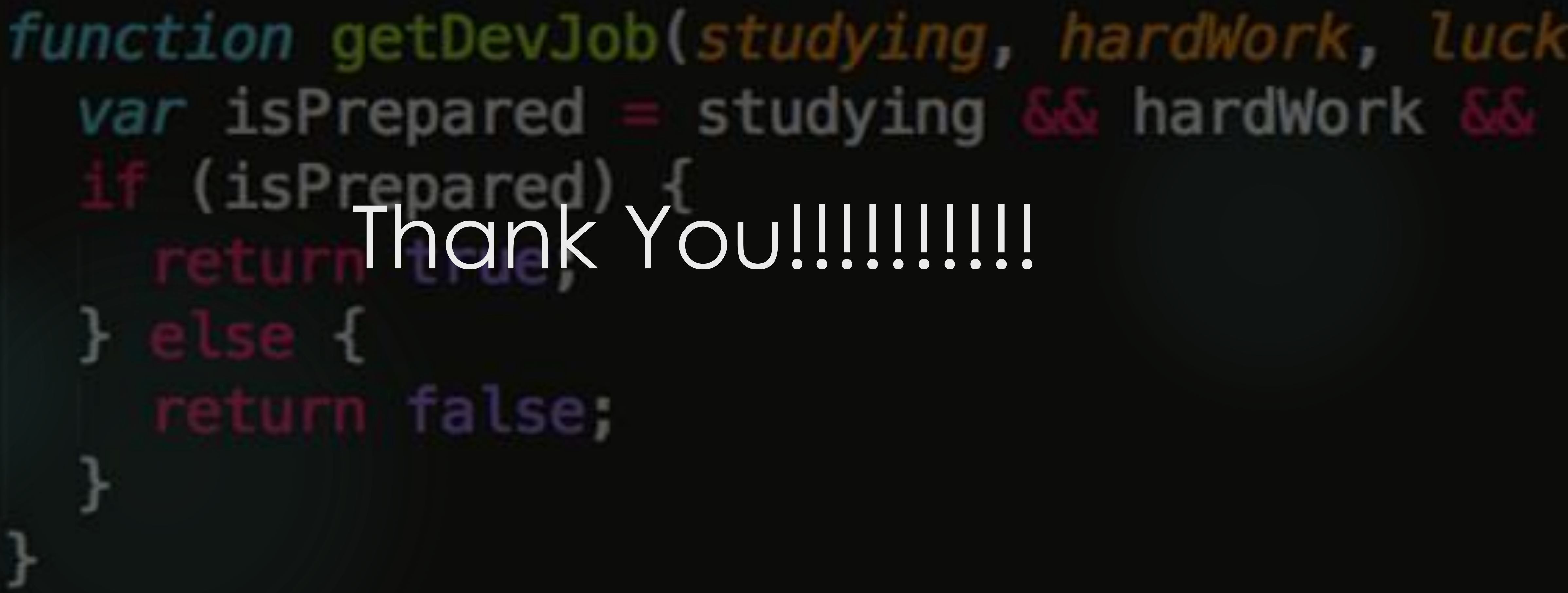
- Winning more votes don't guarantee the candidate will become the next president of USA
  - In 2016, Clinton have 1,325,481 more votes than Trump, but Trump win by 74 electoral votes
  - The battle strategy is not winning most of the votes but win in the state that contains more electoral votes
  - Improve the accuracy by using Neural Network (even find the best strategy?)

Conclusion & Future Work

- Some variable that difficult to estimate:
  - “Anyone But Trump” in 2020 election
  - Any Voter Fraud?
  - Media influences


ANYONE ELSE
2020
SPREADSHIRT
SERIOUSLY.. I WILL VOTE FOR YOUR DOG, A BRICK, DISCARDED SOCKS, E.T., COUNT CHOCULA...

Thank You!!!!!!!!!!



```
function getDevJob(studying, hardWork, luck
  var isPrepared = studying && hardWork &&
  if (isPrepared) {
    return true;
  } else {
    return false;
  }
}
```

HOPE TO SEE YOU AGAIN IN 4140