Kaggle Project: Campaign Contributions in the United States

Please read all of the guidelines carefully before submitting the lab. ☺ There are ***100 points*** in total. **You can work alone or in a group of two in this project.**

Due date: Sunday, April 23, 11:59 PM. For this assignment, no late submission is possible.

Deliverables:

1) The code of the project in **.ipynb** format (one file)
2) Your submission on Kaggle platform
3) No lab report is required.

Guidelines – Before You Start

1) **Please do not post any of your code or solutions online. This is a Kaggle competition.**
2) Three high-ranked teams will receive a bonus of 0.5% if they choose to present their methods and classification strategy in class.
3) You will be using the **`Python`** programming language for this project. You need to write your codes in an empty **.ipynb** file.
4) Make sure that you provide many comments to describe your code and the variables that you created.
5) Please use the ***IEEE*** journal template on ***overleaf.com***. Here is the link: https://www.overleaf.com/latex/templates/preparation-of-papers-for-ieee-sponsored-conferences-and-symposia/zfnqfzzzxghk
   To be able to work on ***overleaf.com***, you will need to register first (you can also compile your ***LaTeX*** file locally.)
6) For some of the code, you may need to do a little bit of "*Googling*" or review the documentation.

What is the Campaign Contribution Data?

The datasets provided with this assignment contain information on the contributions to the campaigns of the US politicians at the state and the federal level. The contribution data has been collected from various sources and covers the 1989-2017 period. The two main datasets (**`training`** and **`test`** data) include aggregated information on the campaign-related behavior of contributors and campaign outcomes. Each data point represents the aggregated campaign behavior of a single contributor over the course of the data collection period. The remainder of the datasets include information on election outcomes, the networks of money between contributors and politicians at the state and federal level. For the main dataset, there are 288,080 observations (=contributors) in total. The **`training`** data consists of 60% of all contributors (=172,848 observations), and the **`test`** data consists of 40% of all contributors (=115,232 observations). The points in both **`training`** and **`test`** datasets are randomly ordered.

Kaggle Project – Data Dictionary

The following datasets have been provided in the assignment folder:

Aggregated Campaign Contributor Data:
- training_data.csv
- test_data.csv

Bipartite Networks Between Contributors and Candidates:
- all_candidates_state_bipartite_weighted_network.csv
- federal_contributor_top100_contributors_network.csv
- state_contributor_top100_contributors_network.csv
- winning_candidates_state_bipartite_weighted_network.csv

Sample Solution File:
- sample_solutions.csv

Below you can find a description of the columns in individual files.

training_data.csv and ***test_data.csv***:

index : The index value associated with contributors (found only in the **test** dataset)

winner_ratio (target variable): The percentage of the candidates who received a contribution and won the elections they were running for (found only in the **training** dataset)

general_sector: Job sector of the contributor

city: City of the contributor

zip_code: ZIP code of the contributor

specific_sector: Specific job sector of the contributor

state: US state of the contributor

contributor_type: Type of the contributor (individual or non-individual)

candidacy_count: Number of the candidacies invested by the contributor

candidacy_democratic_count: Number of Democratic candidacies invested by the contributor

candidacy_republican_count: Number of Republican candidacies invested by the contributor

contribution_count: Number of contributions made by the contributor

contribution_democratic_count: Number of contributions made to Democratic candidate***s***

contribution_republican_count: Number of contributions made to Republican candidates

politician_challenger_count: Number of challengers invested by the contributor

politician_count: Number of all politicians invested by the contributor

politician_democratic_count: Number of Democratic politicians invested by the contributor

politician_incumbency_count: Number of incumbent politicians invested by the contributor

politician_open_pos_count: Number of open positions invested by the contributor

politician_republican_count: Number of Republican politicians invested by the contributor

contribution_democratic_sum_2010_usd: Total amount of all contributions (2010 USD) made to Democratic candidates

contribution_republican_sum_2010_usd: Total amount of all contributions (2010 USD) made to Republican candidates

contribution_sum_2010_usd: Total amount of all contributions made by the contributor (2010 USD)

governor_contributions_sum_2010_usd: Number of contributions (2010 USD) made to candidates running for governorship

house_and_assembly_contributions_sum_2010_usd: Number of contributions (2010 USD) made to candidates running for state-level houses and assemblies

senate_contributions_sum_2010_usd: Number of contributions (2010 USD) made to candidates running for state-level senator positions

us_house_contributions_sum_2010_usd: Number of contributions (2010 USD) made to candidates running for US House positions

us_senate_contributions_sum_2010_usd: Number of contributions (2010 USD) made to candidates running for US Senate positions

candidacy_democratic_ratio: Percentage of Democratic candidacies invested by the contributor

candidacy_republican_ratio: Percentage of Republican candidacies invested by the contributor

contribution_democratic_count_ratio: Percentage of Democratic candidates invested by the contributor

contribution_republican_count_ratio: Percentage of Republican candidates invested by the contributor

governor_contribution_ratio: Percentage of all contributions made to people running for governor positions

house_and_assembly_contribution_ratio: Percentage of all contributions made to people running for state-level house and assembly positions

politician_challenger_ratio: Percentage of all contributions made to challengers

politician_democratic_ratio: Percentage of all contributions made to Democratic politicians

politician_incumbency_ratio: Percentage of all contributions made to incumbents

politician_open_pos_ratio: Percentage of all contributions made to open positions

politician_republican_ratio: Percentage of all contributions made to Republican politicians

senate_contribution_ratio: Percentage of all contributions made to candidates running for state-level senates

us_house_contribution_ratio: Percentage of all contributions made to candidates running for US House

us_senate_contribution_ratio: Percentage for all contributions made to candidates running for US Senate

All network datasets:

- Columns represent the contributors (for instance, all contributors from a certain state) and indices represent the candidates
- The cross-section of columns and indices represent the amount of money invested in the candidate
- Networks include data on the connections between top-100 contributors who invested in federal-level candidates, top-100 contributors who invested in state-level candidates, and contributors who invested in all and winning candidates. The contributors become 'connected' if they invest in the same candidate.
- The network datasets can be used to extract network-related information about candidates.

Model Creation and Prediction (100 points)

For this Kaggle challenge, you are only required to develop a prediction model [and submit the related code].

Please do not post any of your code or solutions online.

This part of the analysis needs to be submitted by the deadline (no late submission will be accepted).

In this part of the analysis, graduate and undergraduate students will be graded/ranked separately.

Please use the dataset provided to you. You can use any external dataset that could be helpful.

For this part of the analysis, you will need to train a model that predicts the number of cases based on all of the remaining variables in the dataset, and report the **RMSE-Value** of your best model. **You will mainly be graded on the RMSE-Value of your model** (more information provided below). Some guidelines (please also review the information shared through lectures):

- The maximum RMSE-value should be around ***0.42***. This corresponds to a sample solution where winner ratio for every lobbyist is '0.5' (assuming that there will be on average one winner and one loser associated with a contributor). RMSE-values around ***0.42*** will be accepted as benchmark for the analysis, and therefore be considered as the maximum score you should obtain.
- You can use **any** classification/prediction model (including, but not limited to, logistic regression, decision trees, neural networks etc.)
- Your model should run on a laptop [equivalent to a modern MacBook Pro] in a reasonable amount of time (in a few hours at a maximum) for grading purposes.
- You can use **any** feature engineering method to transform your dataset, such as:
    - Dimensionality reduction methods such as PCA, t-SNE, spectral embedding
    - Logarithmic, polynomial, and other transformations
    - Different word vectorization techniques
    - Different weighting strategies
- You are free to create a new column (or a stream of data) based on the existing columns and use your new column as an independent variable.
- You are welcome to use **any** external dataset to enrich your training and test datasets.
- You are welcome to create **any** logical condition (if, else etc.) to label the target variable (if you do so, please describe why you made these choices).

Please use your real name when you sign up for the Kaggle project. To participate in the competition, please click: https://www.kaggle.com/t/6071a015d828439ab8bd2b1be9834450

Model Evaluation

Your submission will be evaluated using the ***RMSE*** cost function. If you are unsure, please review what ***RMSE*** means before starting on the project. **Please also report all of your code in the .ipynb file and your final *RMSE* value both in the .ipynb file and in the report.**

Please use your actual name on Kaggle! [Or, please indicate your nickname in the report!]

Make sure that all of your code is running!

Save the code file you have created as "**kaggle_campaign_contribution_lab.ipynb**" in the folder you have created at the beginning.

General Rules and Grading

You will be graded based on the following criteria:

- Code: Cleanliness/understandability (i), executability (ii), format (iii) [We need to be able to run all parts of the code using the datasets provided.]
- Ranking: Ranking in the **Kaggle** competition

* Three high-ranked teams will receive a bonus of 0.5% if they choose to present their methods and classification strategy in class.

An interesting scene that explains Frank Underwood's views (not necessarily the right views coming from a fictional US president) about lobbyists in the brilliant TV show, *House of Cards*:
https://www.youtube.com/watch?v=h52A7_vBxWw