Final Project Guidelines

95828 Machine Learning for Problem Solving

1 Phase I: Understanding the Data and the Problem

1.1 Background and Motivation

You have graduated from Carnegie Mellon University with a Master's in Information Systems Management (MISM), and you now work at the Centers for Disease Control and Prevention (CDC), specifically focused on COVID-19 vaccination initiatives. Leveraging your data analytics and machine learning expertise, your primary goal is to support public health decisions through predictive analytics. One key area of focus is understanding how behaviors and beliefs influence health outcomes, particularly vaccine uptake and the number of COVID-19 cases.

Throughout the COVID-19 pandemic, substantial variations were observed in vaccination acceptance and test positivity rates across different communities in the United States. Public health interventions could have significantly benefited from accurate predictions regarding these outcomes. Consequently, your investigation seeks to explicitly address two key questions:

- **Can vaccine uptake (the rate of people choosing vaccination) be accurately predicted using individuals' reported behaviors and beliefs?**
- **Can COVID-related outcomes, particularly the percentage of the positive cases, be reliably predicted?**

Accurate predictions in these domains would facilitate targeted public health campaigns, better resource allocation, and enhanced preparedness for future public health crises.

1.2 Dataset Introduction

The dataset used in this study is derived from the COVID-19 Trends and Impact Survey (CTIS), conducted by the Delphi Group at Carnegie Mellon University. This dataset aggregates responses from a representative sample of Facebook users (aged 18+) at the U.S. county level. Responses were gathered during one-month time frame at the peak of the COVID pandemic (from January 07, 2021 to February 12, 2021), offering rich information to analyze temporal dynamics in health behaviors and perceptions.

The dataset contains 25627 instances where each row represents one U.S. county in a given day.

Detailed Feature Descriptions The dataset contains numerous features, each capturing various dimensions of individual behaviors, attitudes, and COVID-19 related outcomes. The full descriptions of the dataset and features can be found here, link. The following list provides detailed descriptions of key features:

- COVID Activity Indicators:
  - smoothed_wcli: Percentage reporting COVID-like illness symptoms (fever with cough or shortness of breath). *(Numerical values ranging from 0 to 100)*
  - smoothed_wtested_14d: Percentage of individuals tested for COVID-19 in the past 14 days, regardless of results. *(Numerical values ranging from 0 to 100)*
  - smoothed_wtested_positive_14d (**target**): Estimated test positivity rate (percent) among people tested for COVID-19 in the past 14 days. *(Numerical values ranging from 0 to 100)*
  - smoothed_wcovid_vaccinated (**target**): Percentage of individuals vaccinated by a COVID vaccine. *(Numerical values ranging from 0 to 100)*
- Behavioral Indicators:
  - smoothed_wwearing_mask: Percentage of respondents regularly wearing masks in public during the past 7 days. *(Numerical values ranging from 0 to 100)*
  - smoothed_wothers_masked: Percentage of respondents who say that most or all other people wear masks, when they are in public and social distancing is not possible. *(Numerical values ranging from 0 to 100)*
  - smoothed_wwork_outside_home_1d: Percentage of respondents who worked or went to school outside their home in the past 24 hours. *(Numerical values ranging from 0 to 100)*
  - smoothed_wlarge_event_1d: Percentage attending large public events in the past day. *(Numerical values ranging from 0 to 100)*
  - smoothed_wrestaurant_1d: Percentage visiting bars, restaurants, or cafes in the past day. *(Numerical values ranging from 0 to 100)*
  - smoothed_wshop_1d: Percentage visiting markets, grocery stores, or pharmacies within the past day. *(Numerical values ranging from 0 to 100)*
  - smoothed_wspend_time_1d: Percentage of respondents who "spent time with someone who isn't currently staying with you" in the past 24 hours *(Numerical values ranging from 0 to 100)*
  - smoothed_wpublic_transit_1d: Percentage of respondents who "used public transit" in the past 24 hours. *(Numerical values ranging from 0 to 100)*
- Belief Indicators:
  - smoothed_wworried_become_ill: Percentage of respondents who reported feeling very or somewhat worried that "you or someone in your immediate family might become seriously ill from COVID-19". *(Numerical values ranging from 0 to 100)*

  - smoothed_wvaccine_likely_friends: Percentage of respondents likely to vaccinate based on friends or family recommendations. *(Numerical values ranging from 0 to 100)*
  - smoothed_wvaccine_likely_who: Percentage likely to vaccinate if recommended by the WHO. *(Numerical values ranging from 0 to 100)*
  - smoothed_wvaccine_likely_govt_health: Percentage likely to vaccinate based on government health recommendations. *(Numerical values ranging from 0 to 100)*
  - smoothed_wvaccine_likely_politicians: Percentage of respondents who would be more likely to get a COVID-19 vaccine if it were recommended to them by politicians, among respondents who have not yet been vaccinated. *(Numerical values ranging from 0 to 100)*

- Other:
  - time: the end date of the week during which the data was recorded for a given county. (from 2021-01-07 to 2021-02-12)
  - geo_value: county code

1.3 Questions: Problem, Data & Their Relationship

We recommend you to keep the following questions in mind while you move forward:

(1) What specific features from the dataset are likely most relevant in predicting vaccine uptake and COVID cases?

(2) How accurately can vaccine uptake and COVID cases be predicted using historical survey data?

(3) What actionable insights or policy implications might result from successful predictions of vaccine uptake and COVID test positivity rates? Specifically:

- How could accurate predictions of vaccine uptake inform targeted outreach and vaccination programs?
- How could predictions of COVID cases improve health policy decisions and resource allocation, potentially reducing transmission and disease impact?

2 Phase II: Data Exploration and Analysis

In this phase, we will perform detailed data cleaning, feature preparation, and exploratory data analysis to set the groundwork for subsequent modeling phases. The objectives are to ensure data quality, understand key characteristics of the variables, and identify important relationships between the predictors and our two outcome variables: vaccine uptake and number of COVID cases.
We begin by carefully reading, inspecting, and preparing the COVID-19 Trends and Impact Survey (CTIS) dataset:

1. Data Cleaning and Preparation:

    - Identify and handle potential outliers or anomalies through visual exploration (e.g., using histograms, boxplots, or scatterplots).
    - Clearly document and justify all decisions made regarding the removal of outliers.
    - Are there any missing values? Handle the missing values properly (either drop them or consider mean imputation or other advanced imputation methods.) Justify how you handle the missing values.

2. Feature Selection and Transformation:

    - Based on the initial data examination, select relevant features for the prediction tasks, clearly distinguishing between predictors and target variables. Justify your feature selection method in your report.
    - Ensure the chosen target variables are clearly identified:
        - Vaccine uptake: `covid_vaccinated`
        - Positive COVID cases: `smoothed_wtested_positive_14d`
    - Decide whether feature transformations (logarithmic, scaling, or standardization/normalization) are necessary based on feature distributions and statistical properties.

3. Correlation Analysis:

    - Explore correlations among features using correlation matrices, scatterplot matrices, or heatmaps.
    - Identify and document any strong correlations (positive or negative) among features.
    - Specifically examine and document correlations between potential predictors and the two target variables.
    - Identify pairs of features with strong correlations. Discuss potential issues with multicollinearity (if any) and decide how to address them in subsequent modeling phases.

At the conclusion of this phase, you should have a cleaned and well-understood dataset to begin the next step.

3 Phase III: Predictive Modeling and Policy Decision-Making

In Phase III, we will build predictive models using the cleaned dataset from Phase II. We aim to predict two main outcomes: vaccine uptake and COVID-19 test positivity rates. Subsequently, we will use these predictive models to inform policy-making decisions relevant to public health strategy.

3.1 Predictive Modeling

Your predictive modeling should address two clearly defined questions:

1. **Can vaccine uptake be accurately predicted using behavioral and belief indicators?**
2. **Can the positive COVID cases be reliably predicted from these indicators?**

Follow these steps for modeling:

1. Clearly define your prediction setup:
   - Split the dataset into training, validation, and testing sets.
   - Justify your data splitting strategy clearly. Explain potential limitations and benefits. You can consider splitting the train/test data using temporal information: e.g, use the week on 1/30 for training and use the week on 3/2 for testing. This is only optional, but you need to justify your splitting strategies for whatever method you use.
2. Consider using multiple predictive models to benchmark performance (**you should compare at least three models in your report. For example, you can have one baseline models and two other models you learned**), you may use:
   - Linear Regression (with Ridge and Lasso Regularization)
   - Neural Networks
   - Any other models that you learned in or outside the course

   Clearly explain your rationale for choosing each model.
3. Evaluate your models rigorously:
   - Demonstrate which metrics you are using and why they are appropriate in the setting (RMSE, MAE, etc).
   - Conduct k-fold cross-validation (with clearly explained choice of k).
   - Report average performance and standard deviation (format: mean $\pm$ std).
4. Choose and clearly justify the best-performing models for each prediction task. Document and visualize key predictive relationships between predictors and your outcomes.

Your report must clearly document the predictive performance results, model comparison/validation, and your rationale for choosing the best-performing models.

3.2 Decision-making and Policy Recommendations

Using your observation from data analysis and the results from the prediction tasks, provide some policy recommendations relevant to public health interventions by the CDC. The questions can be open-ended, but you need to justify your answer using our empirical results and your data analysis and demonstrate how your policy recommendations are aligned with them. Here are some suggested questions you can think about, and you can propose some questions by yourself for your investigation as well:

(1) Vaccine effectiveness on COVID outcomes:

- Using your model predictions, assess the relationship between vaccine uptake and the number of COVID cases.
- Specifically, investigate: *Does higher vaccine uptake correlate with lower number of COVID cases?* Provide clear visualizations (scatter plots, regression lines, etc.) and statistical evidence (correlations or regression results).

(2) Vaccine policy implications:

- What vaccine policy recommendations would you suggest for the CDC to improve vaccination uptake and lower the new confirmed COVID cases?
- Clearly justify your recommendations based on model insights and quantified outcomes.

Provide clear explanations of your insights, policy recommendations, and any assumptions involved.