Risk Identification and Prediction for Long COVID

Mid-Project Assessment

Emma Jin, Elliot Kim, Victoria Song
CS 66 (Fall 21)

The Long COVID Problem

- **Long COVID:** persistence of symptoms post-COVID-19
  - **Symptoms:** Headaches, Fatigue, Cough/Shortness of breath, Anxiety, Heart problems, Muscle aches, Dizziness
- **Affected population:** subpopulation of COVID-19 patients who retain COVID symptoms
- **Goals:** identify risk factors of Long COVID and predict likelihoods that a patient would suffer from Long COVID
- **Dataset:** COVID-19 Fall 2020 & Winter 2021 Community Supplement from MCBS (Medicare Current Beneficiary Survey)

Overall Plan

- Find usable dataset
- Pre-process data
- Identify models to test
- Train, test, tune, validate models
- Select highest-performing model and identify risk factors
- Construct ranking system of Long COVID risk factors
- Repeat with Long COVID patient subgroups by lasting symptom

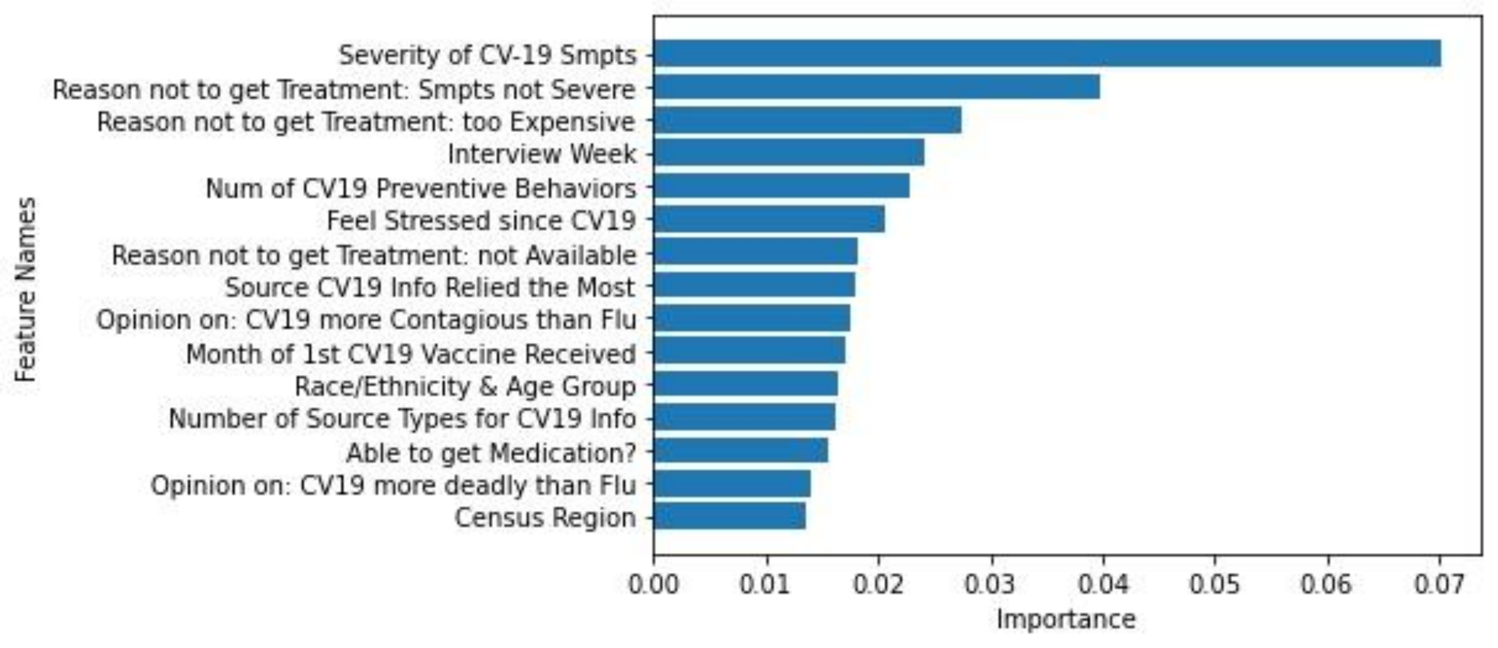
Progress

- ☑ Data pre-processing: Removed nuisance features (interview IDs), mapped feature code labels to human-readable language
- ☑ Training & testing: Compared Logistic Regression, Decision Tree, Random Forest, SVM (RBF), Naive Bayes (Complement)
- ☑ Tuning & validation: Performed grid search, k-fold validation
- ☑ Selected best model: RandomForest with max_features = 0.5, min_samples_leaf = 1 (accuracy of 0.723)
- ☑ Examine the feature importance and risk factors of the model chosen

Table: Average Test Scores of Different Models

	Original Model	Tuned Model
Logistic Regression	0.65651	0.68258
Decision Tree	0.66205	0.60044
Random Forest	0.68698	**0.71697**
SVM	0.71191	0.67591
Naive Bayes (ComplementNB)	0.57618	0.59934

Risk Factors Importance Using Random Forest



Next Steps

- ❑ Analyze and discuss prominent risk factors
- ❑ Construct ranking system of Long COVID risk factors
- ❑ Repeat with Long COVID patient subgroups by lasting symptom

Thank you!