Final Project- Data Demo

Rohan Dalal-PSTAT 131

Contents

DATA MEMO

Introduction to Data source

I will be choosing COVID data from Covidcast package for my Final Project. The covidcast R package, which provides access to the COVIDcast Epidata API published by the Delphi group at Carnegie Mellon University. According to the covidcast R package website,This API provides daily access to a range of COVID-related signals Delphi that builds and maintains, from sources like symptom surveys and medical claims data, and also standard signals that we simply mirror, like confirmed cases and deaths. (see website here) Here is a list of the signals, we can see all the documentation for each one. This includes information about when the first data points were collected, if the data is available on a daily, or weekly basis, what regions we can call the signal for, and so on.

Data overview: Fetching/Merging/Prepping data

I plan to choose five signals to predict cases across California counties. Predictor : "visits","admits", "chngVisits" ,"covidChngVisits","gsymptoms" and Outcome : "Cases"

- "Cases": Get the number of daily new Covid cases for all the counties in California,for a given date range (example :from May 2020 to July 2020) by fetching the "US Facts Cases and Deaths" data source (https://cmu-delphi.github.io/delphi-epidata/api/covidcast-signals/usa-facts.html). This will be the Ground Truth(label)
- "visits": Get the daily percentages of doctor visits that are related to Covid in California for a given date range (example :from May 2020 to July 2020) by fetching the "Doctor Visits" data source (https://cmu-delphi.github.io/delphi-epidata/api/covidcast-signals/doctor-visits.html).

- “admits” :Get the daily hospital admissions for covid diagnosed that are related to Covid in California for a given date range (example :from May 2020 to July 2020) by fetching the “Doctor Visits” data source (https://cmu-delphi.github.io/delphi-epidata/api/covidcast-signals/hospital-admissions.html).
- “chngVisits”: Get the Estimated percentage of outpatient doctor visits primarily about COVID-related symptoms in California for a given date range (example :from May 2020 to July 2020) by fetching the “Doctor Visits” data source (https://cmu-delphi.github.io/delphi-epidata/api/covidcast-signals/chng.html).
- “covidChngVisits”: Get the Estimated percentage of outpatient doctor visits with confirmed COVID-19, based on Change Healthcare claims data in California for a given date range (example :from May 2020 to July 2020) by fetching the “Doctor Visits” data source (https://cmu-delphi.github.io/delphi-epidata/api/covidcast-signals/chng.html).
- “gsymptoms”:Get Sum of Google search volume for anosmia and ageusia related searches in California for a given date range (example :from May 2020 to July 2020) by fetching the “Doctor Visits” data source (https://cmu-delphi.github.io/delphi-epidata/api/covidcast-signals/google-symptoms.html).

Data PreProcessing

- Get the required signals and merge data to create a csv file, clean up and tidy data.
- Observation Count: After collecting needed predictors from the data sources contains about 5428 observations, not all the predictors have missing value but some do.
- Analyze the data fetched for datatype and null/missing values – Dealing with missing/NA data : Method:1-dropping rows with missing values-disadvantage is smaller data set for modeling Method:2-imputation method-disadvantage it might limit the effectiveness of the model

I am planning to do exploratory analysis to see accuracy and effectivenes compariosn by both methods.

Motivation/Goal

Goal is to: -to build predictive models that forecast the future of the pandemic so that we can see one step ahead and prepare accordingly using the past data. -to build a predictive model that uses historical COVID cases and related data to forecast the short-term future number of COVID cases in a particular region.

Project Timeline

- April 8 -April 14: Load and tidy data
- April 14 -April 24: Exploratory analysis and Model selection
- April 24- May 10: Test and Run models
- May 10 -May 24 ; work on draft paper
- May-25 - June 2 : Any edits and finalize paper

Data collection efforts using covidcast package.

```
#install.packages(’covidcast’)
library(covidcast)
# Cumulative COVID cases per 100k people on 2020-12-31
df <- covidcast_signal(data_source = "usa-facts",
                   signal = "confirmed_cumulative_prop",
                   start_day = "2020-12-31", end_day = "2020-12-31")
summary(df)
```

```
## A ‘covidcast_signal‘ dataframe with 3142 rows and 12 columns.
##
## data_source : usa-facts
## signal      : confirmed_cumulative_prop
## geo_type    : county
##
## first date                          : 2020-12-31
## last date                           : 2020-12-31
## median number of geo_values per day : 3142
```

```
# This looks at the people who reported COVID-like symptoms from their fb-survey
# from dates 5-1-2020 to 5-7-2020 in all counties
data <- covidcast_signal("fb-survey", "smoothed_cli", start_day = "2020-05-01",
                         end_day = "2020-05-07")
head(data)
```

```
##   data_source       signal geo_value time_value      issue lag missing_value
## 1   fb-survey smoothed_cli     01000 2020-05-01 2020-09-03 125             0
## 2   fb-survey smoothed_cli     01001 2020-05-01 2020-09-03 125             0
## 3   fb-survey smoothed_cli     01003 2020-05-01 2020-09-03 125             0
## 4   fb-survey smoothed_cli     01015 2020-05-01 2020-09-03 125             0
## 5   fb-survey smoothed_cli     01031 2020-05-01 2020-09-03 125             0
## 6   fb-survey smoothed_cli     01045 2020-05-01 2020-09-03 125             0
##   missing_stderr missing_sample_size     value    stderr sample_size
## 1              0                   0 0.8254101 0.1360033   1722.4551
## 2              0                   0 1.2994255 0.9671356    115.8025
## 3              0                   0 0.6965968 0.3247531    584.3194
## 4              0                   0 0.4282713 0.5485655    122.5577
## 5              0                   0 0.0255788 0.3608268    114.8318
## 6              0                   0 1.0495589 0.7086324    110.6544
```

```
cases <- covidcast_signal(data_source ="usa-facts", "confirmed_incidence_num",
                         start_day = "2020-05-01",
                         end_day = "2020-05-07",
                         geo_type="state", geo_values="ca")
head(cases)
```

```
##   data_source                  signal geo_value time_value      issue lag
## 1   usa-facts confirmed_incidence_num        ca 2020-05-01 2021-09-16 503
## 2   usa-facts confirmed_incidence_num        ca 2020-05-02 2021-09-16 502
## 3   usa-facts confirmed_incidence_num        ca 2020-05-03 2021-09-16 501
## 4   usa-facts confirmed_incidence_num        ca 2020-05-04 2021-09-16 500
## 5   usa-facts confirmed_incidence_num        ca 2020-05-05 2021-09-16 499
## 6   usa-facts confirmed_incidence_num        ca 2020-05-06 2021-09-16 498
##   missing_value missing_stderr missing_sample_size value stderr sample_size
## 1             0              5                   5  1913     NA          NA
## 2             0              5                   5  2213     NA          NA
## 3             0              5                   5  1379     NA          NA
## 4             0              5                   5  1142     NA          NA
## 5              0              5                   5  2406     NA          NA
## 6              0              5                   5  2592     NA          NA
```

```
visits <- covidcast_signal(data_source ="doctor-visits", "smoothed_cli",
                        start_day = "2020-05-01",
                         end_day = "2020-05-07",

                          geo_type="state", geo_values="ca")
head(visits)
```

```
##     data_source       signal geo_value time_value      issue lag missing_value
## 1 doctor-visits smoothed_cli        ca 2020-05-01 2020-07-04  64             0
## 2 doctor-visits smoothed_cli        ca 2020-05-02 2020-07-05  64             0
## 3 doctor-visits smoothed_cli        ca 2020-05-03 2020-07-06  64             0
## 4 doctor-visits smoothed_cli        ca 2020-05-04 2020-07-07  64             0
## 5 doctor-visits smoothed_cli        ca 2020-05-05 2020-07-08  64             0
## 6 doctor-visits smoothed_cli        ca 2020-05-06 2020-07-09  64             0
##   missing_stderr missing_sample_size    value stderr sample_size
## 1              5                   5 3.943336     NA          NA
## 2              5                   5 3.793779     NA          NA
## 3              5                   5 4.363606     NA          NA
## 4              5                   5 4.876648     NA          NA
## 5              5                   5 4.456923     NA          NA
## 6              5                   5 4.050397     NA          NA
```