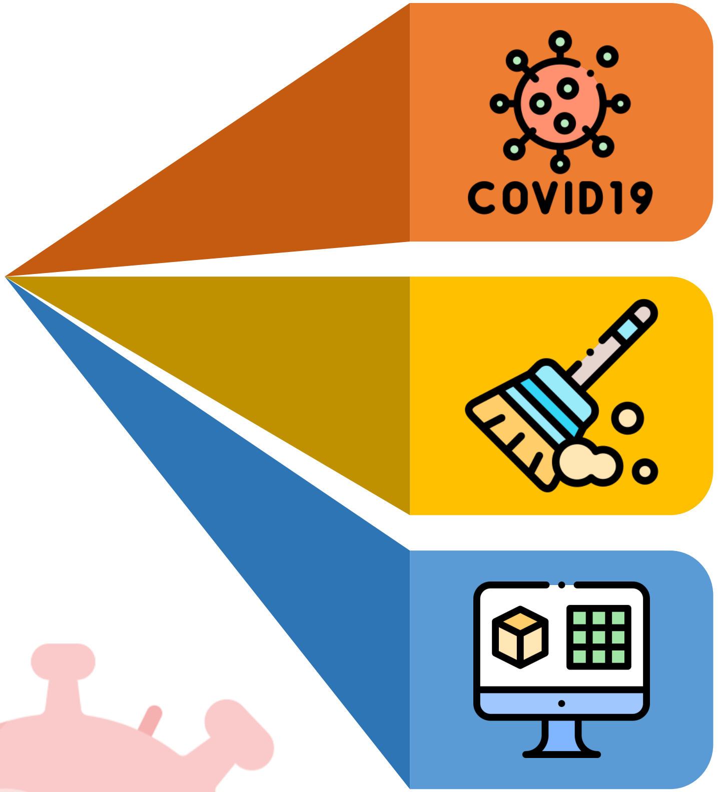
Coronavirus disease (COVID-19) pandemic analysis

Data Preprocessing and Feature Engineering with PySpark

Instructor: Charles Tsang

Presenters: Junhao Wang, Fei Chen, Lingying Sun

Roadmap



Background

Data Preprocessing & Feature Engineering

Dataset Findings & Regression Model

Background:

The Coronavirus disease has had a profound effect on the world since December 2019. The virus is highly contagious and rapidly spreads through close contact via human-to-human trans-mission. As of October 28, 2020, there have been 43.7 million people infected throughout 219countries worldwide (World Health Organization).

Problem Statement:

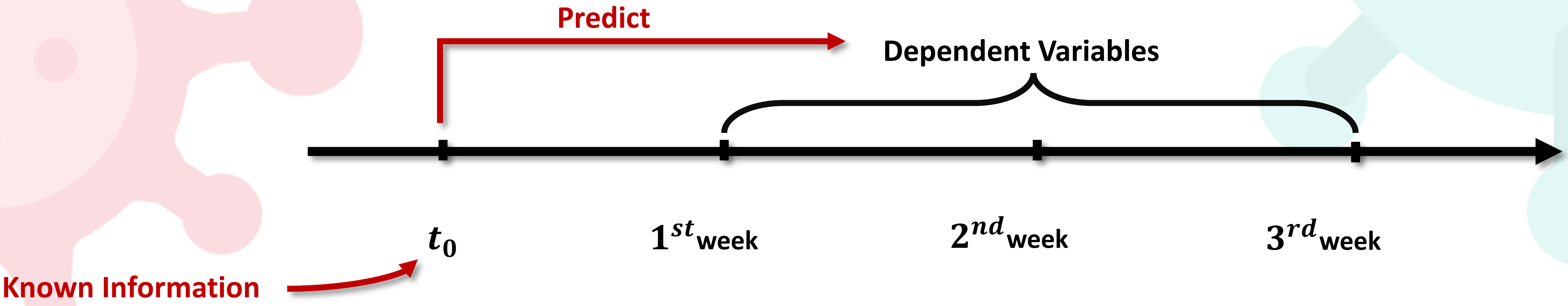
To help control the pandemic, each country imposed strict regulations such as travel bans, mandatory masks, social distancing, etc. Each country’s governmental policies have a different effect on the COVID-19 infection rate. The objective of this project is to build a model to predict future COVID-19 cases and deaths, which may help to control pandemic effectively.

COVID19

Response Variables

Dependent Variables:

- Two-week cumulative cases after next week
- Two-week cumulative deaths after next week



Predict

Dependent Variables

t_0

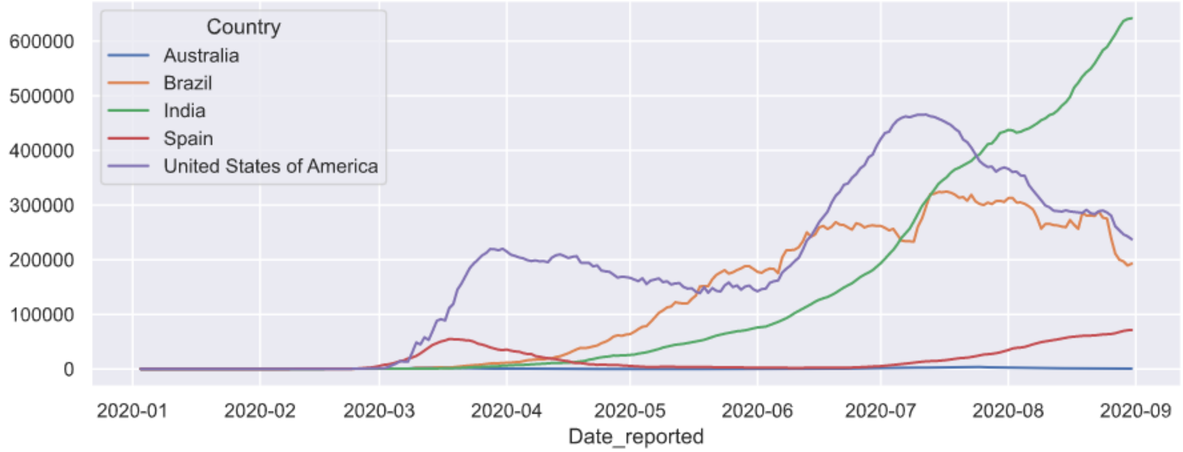
1^{st}week

2^{nd}week

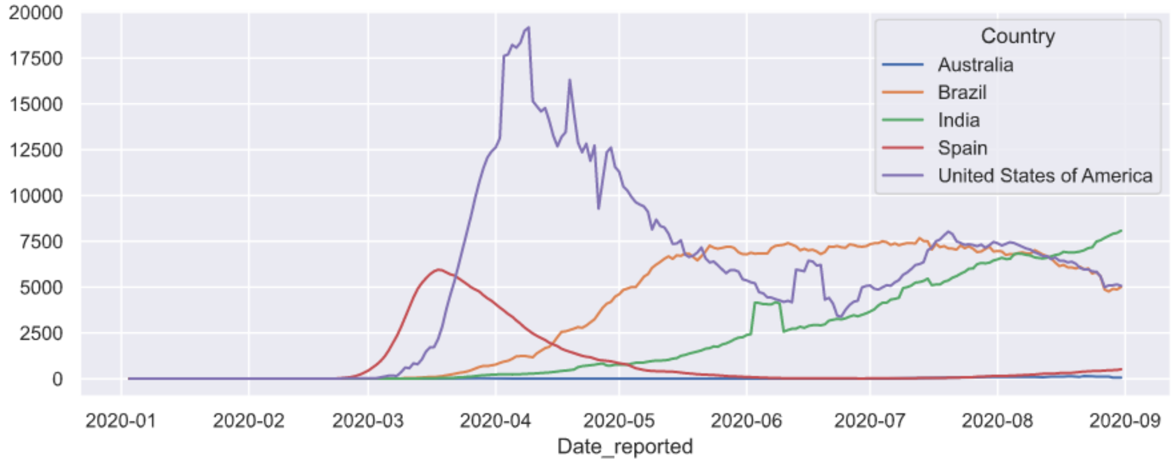
3^{rd}week

Known Information

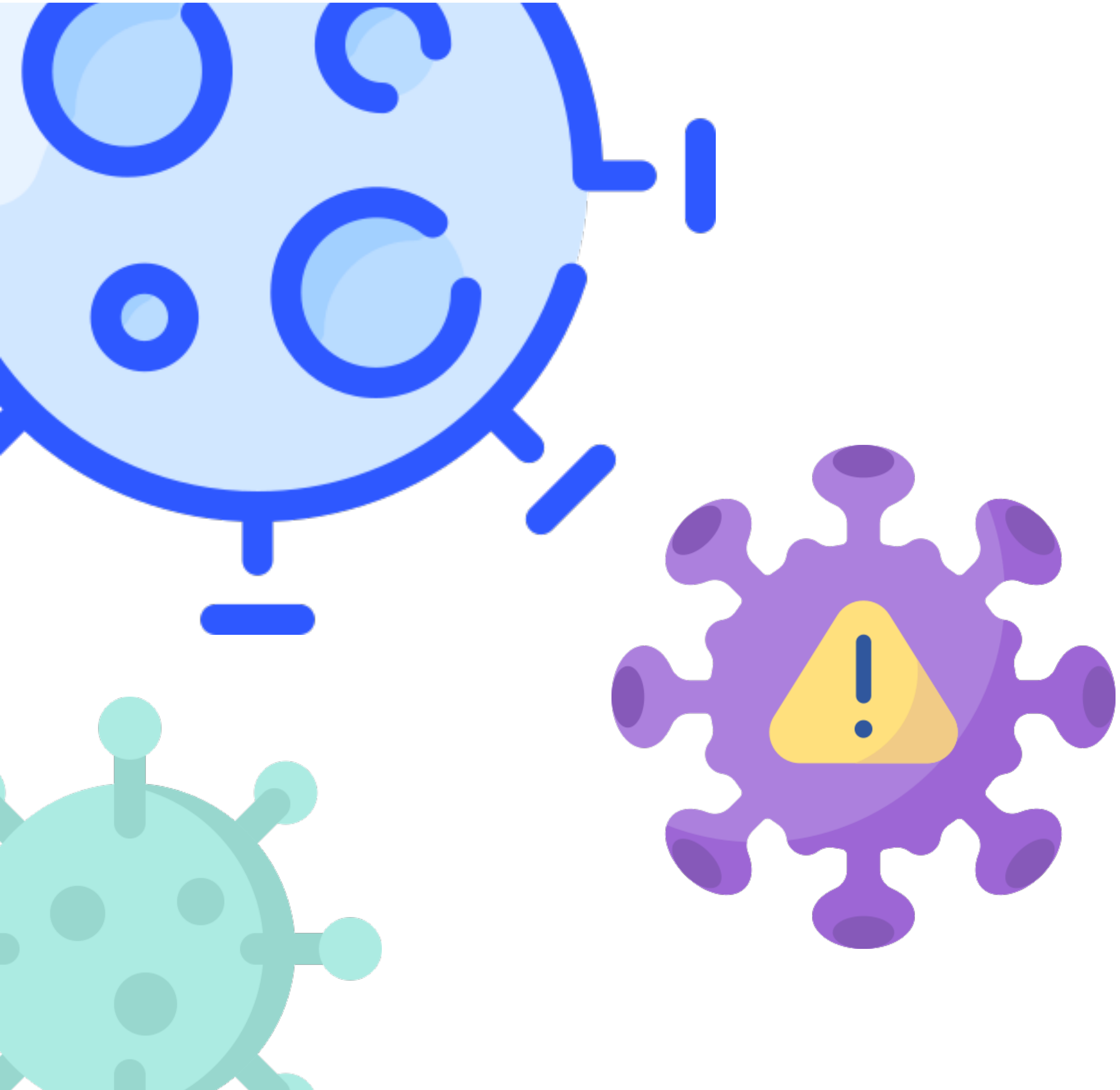
Two-week cumulative cases after next week



Two-week cumulative Deaths after next week



Explanatory Variables



Numerical Variables:

- Current Cumulative Cases
- Current Cumulative Deaths
- Daily Testing
- Total Existing Cases

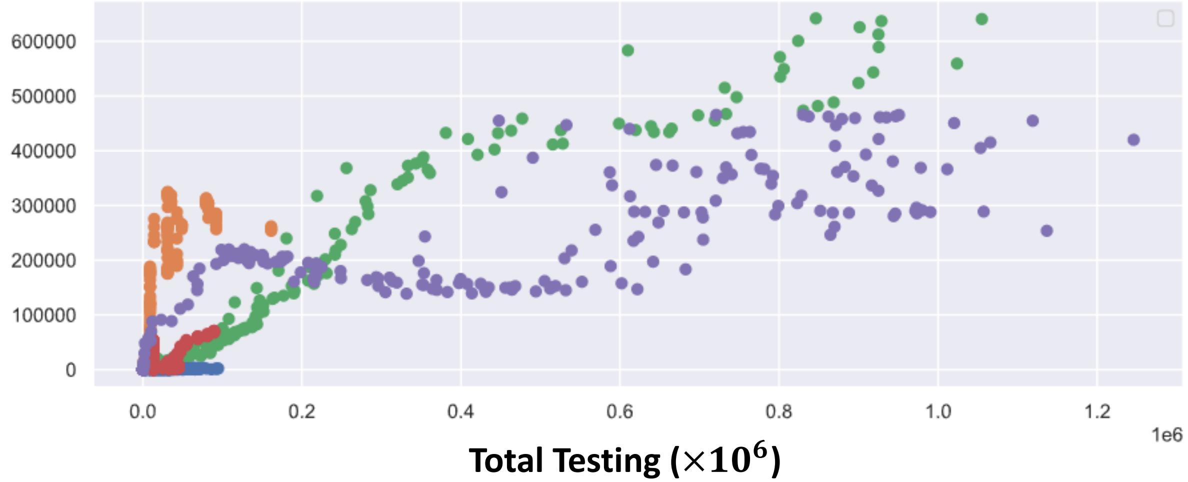
Categorical Variables:

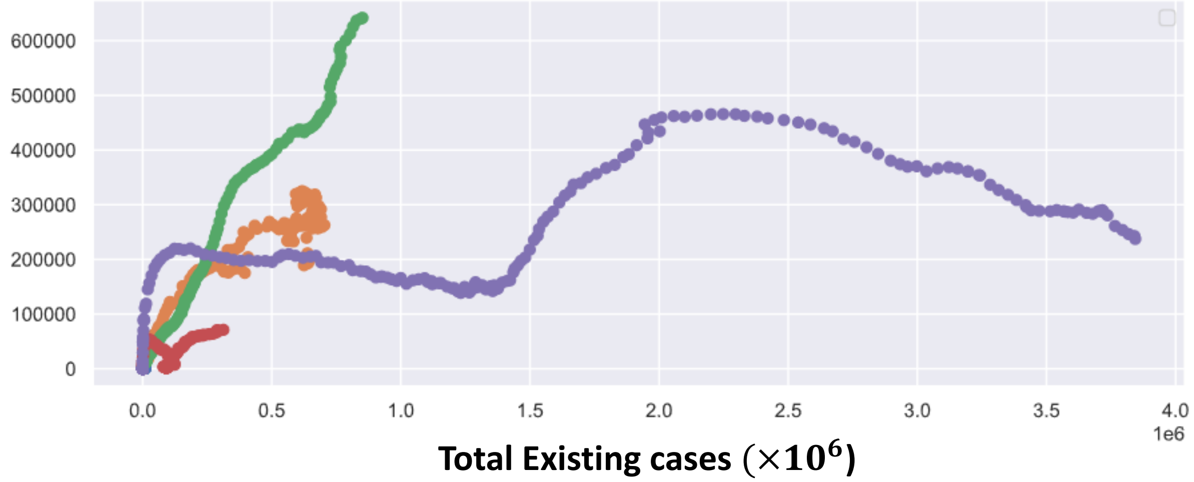
- Countries
- Travel Controls (0,1)
- Restaurant Restriction (0,1)
- School Closing (0,1,2,3)
- Workplace Closing (0,1,2,3)
- Public Events(0,1,2,3)
- Stay At Home Requirements(0,1,2,3)

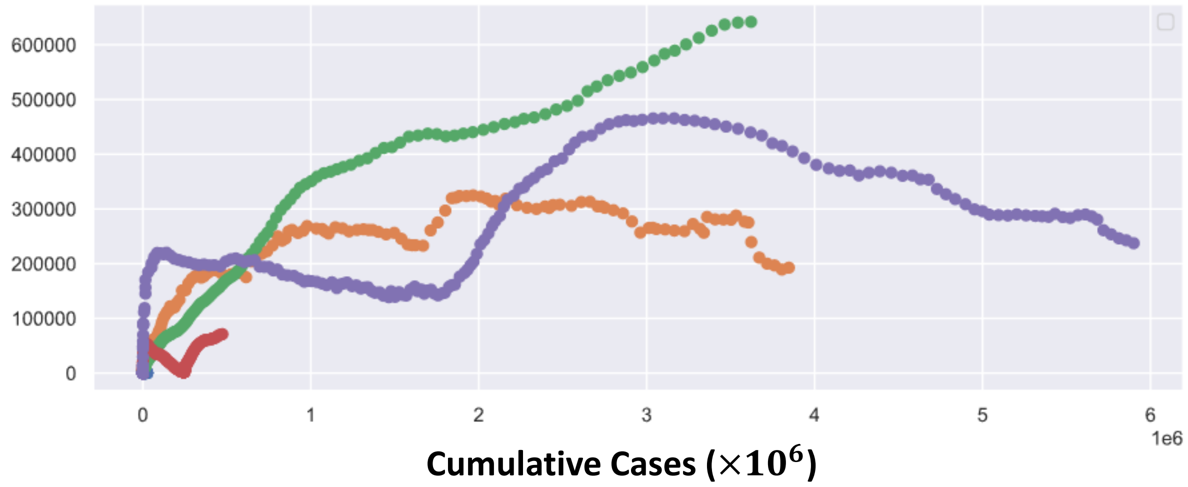
Dataset Findings – Numerical Variables

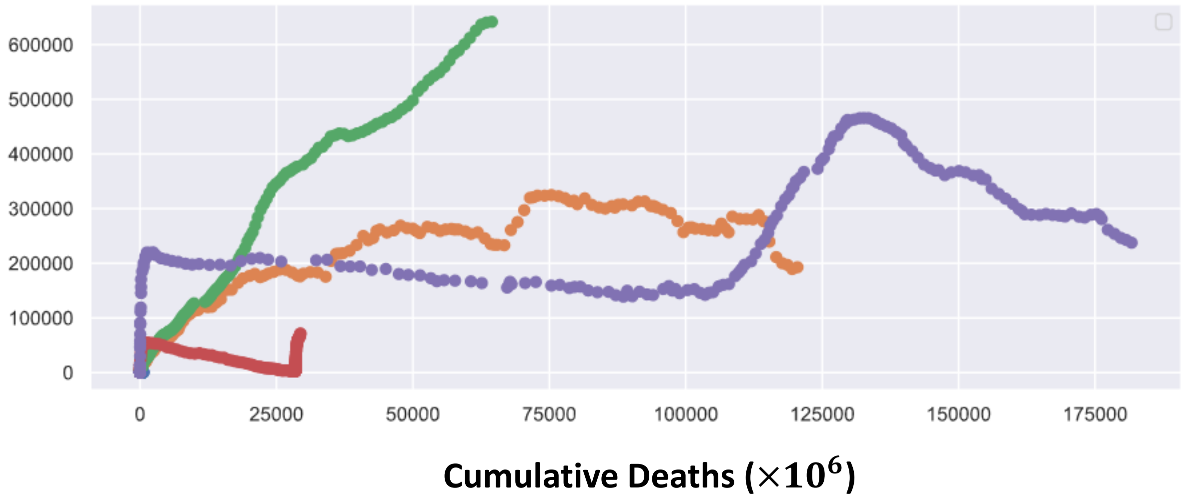
● Australia
● Brazil
● India
● Spain
● United States of America

Y-AXIS: Two-week cumulative cases after next week



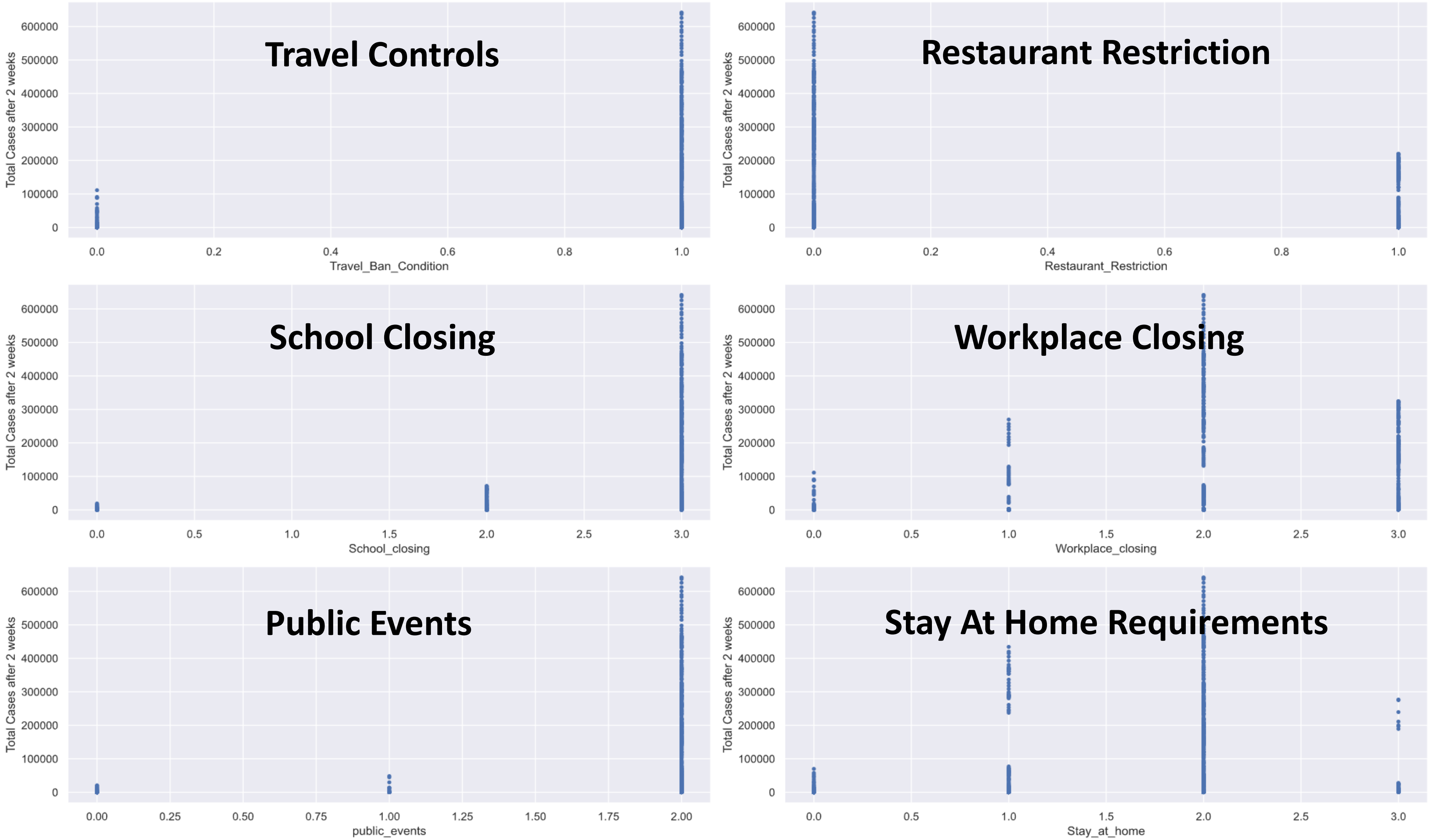






Dataset Findings – Categorical Variables

Y-AXIS: Two-week cumulative cases after next week



Travel Controls

Restaurant Restriction

School Closing

Workplace Closing

Public Events

Stay At Home Requirements

Regression Model

Lasso Regression Model

$$\mathcal{L}_{reg} = RSS + \lambda \sum_{i=0}^{k} |\beta_i| = \sum_{i=1}^{k} [Y_i - \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}\right)]^2 + \lambda \sum_{i=0}^{k} |\beta_i| \ , \lambda = 5$$

R-Square Measures

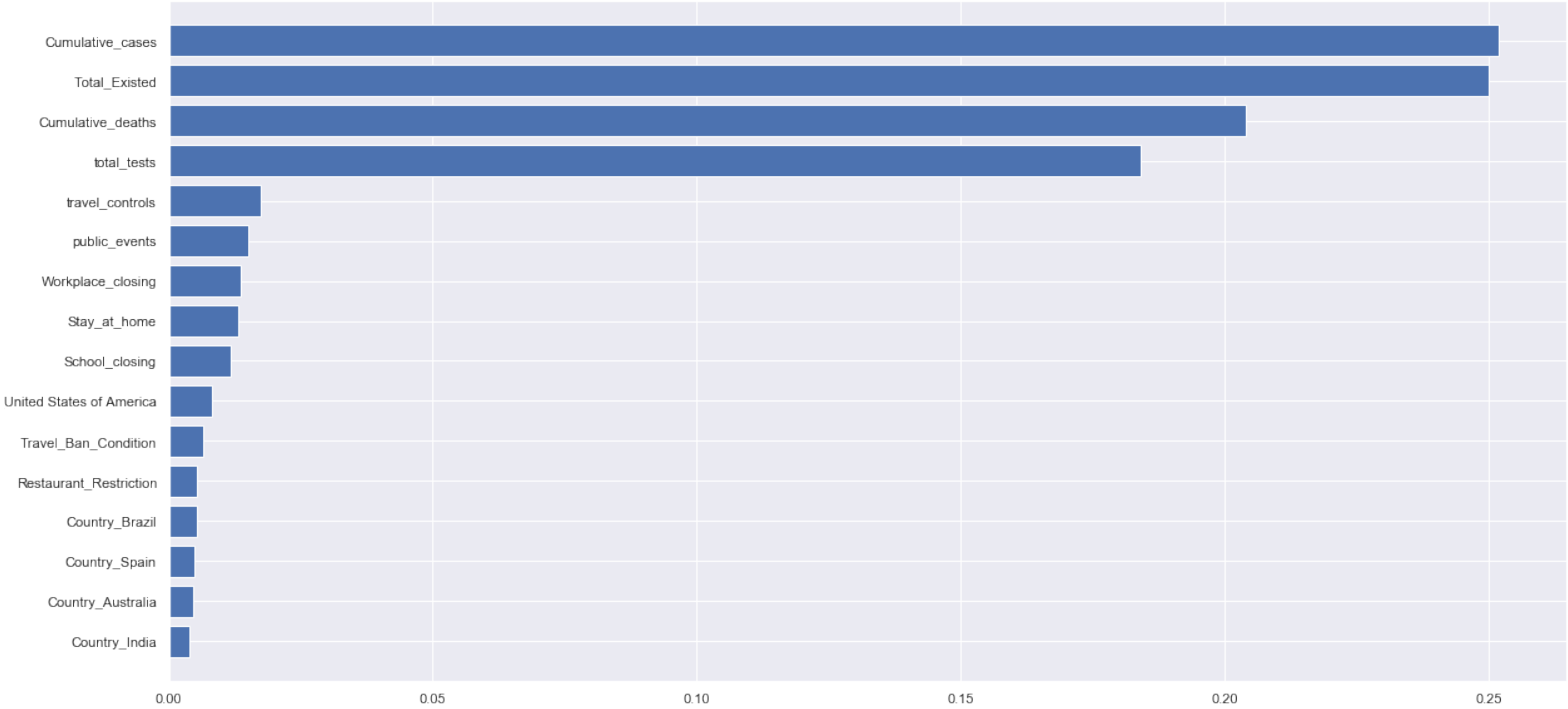
Cases prediction: 85.39

Deaths prediction: 60.99

Model Predictions (Example)

United States on September 23rd	Predicted Cases	True Cases	Predicted Deaths	True Deaths
	454682	677593	9772	9464

Feature Importance Exploring Using Extra Trees Classifier



Thank You!