COVID-19 Impact On Counties

Katherine Lin
Byram Hills High School

Supervised by Linda Zhao
My COVIDpath to Data Scientist

First COVIDcase

Jan

Mar

Stay at home

Soaring cases

Despair

Graph showing the increase in cases over time.
My COVIDpath to Data Scientist

Prof Zhao’s Modern Data Mining Course

What can I do to help?

First COVIDcase

Jan

Mar

Stay at home
Soaring cases
Despair

Jun

Mo
DATA MINING
My COVIDpath to Data Scientist

Data Scientist
Aim to be a data scientist
Use data for social good

Prof Zhao’s Modern Data Mining Course

What can I do to help?

COVID+ County study

Stay at home
Despair

First COVID case

Jan

Mar

May

Jun

Jul

2021
COVID-Death Rate vs County Socioeconomic Status

Race group
Is one racial group affected more?

Education level
More educated
More informed and careful?

Age group
Elderly is affected more. Is this true?

Income group
High income less affected?
1. Important factors related to COVID death rate

1. Predict COVID death rate
Roadmap

1. Data
   - Data source
     - CDC + USDA

2. Model
   - Modelling
     - Exploratory Data Analysis
     - LASSO, Random Forest, Boosting, Neural Network

3. Findings
Data source
CDC + USDA

Model
Explanatory Data Analysis
LASSO, Random Forest,
Boosting,
Neural Network

Findings
Roadmap

1
2
3
Data
Data

1. Infection
   New cases by county

2. Fatality
   Fatality by county
Data

1. Income
   Poverty level and average household income

2. Jobs
   Employment type, rate, and change

3. Demographics
   Population size, density, education level, race, age, household size, and migration rates

4. County Classifications
   Type of county
Merging Data
Roadmap

1. Data source
   - CDC + USDA

2. Modelling
   - Exploratory Data Analysis
   - LASSO, Random Forest, Boosting, Neural Network

3. Findings
EDA

Accumulative Deaths and Death Rate
(8/19/2020)
Death rate by county
(8/19/2020)
Histogram of Death Rate

Very Skewed!
Transformation

\[
\log \left( \frac{\text{COVID-19 deaths} + 1}{\text{County population}} \right)
\]
Histogram of log(Death Rate)

Very Skewed!

Bell shape!
LASSO Regression

Input: 40+

LASSO

Output: 20
Roadmap

1. Data source
   - CDC + USDA

2. Modelling
   - Exploratory Data Analysis
   - LASSO, Random Forest, Boosting, Neural Network

3. Findings

Model

Data

Findings
Worst States

NJ
LA
CT
MA
DE
Best States

OR  CA  ME  WI  NM  WA
Finding: Covid Factors affecting death rate

- **Age**: % of 65+
- **Population**: % of 65+
- **Race group**: % of White
- **Foreign born**: 
- **High creative class**: 
- **Job type**: Varied by job types

Image: A circle with a virus icon, connected by lines to the factors listed above.
Finding

Predictive Models

Prediction Error

Random Forest
0.47

Boosting
0.50

Deep Learning
0.54

Relaxed LASSO
0.56
My COVIDpath to Data Scientist

Thank you!

Acknowledgement:
Professor Linda Zhao and Jeffrey Cai