MISSION PREDICTABLE: USING MACHINE LEARNING TO PREDICT COVID-19'S IMPACT ON TEXAS

Team Data Liberation
Gia Gillis & Amanda Wright
August 14, 2020

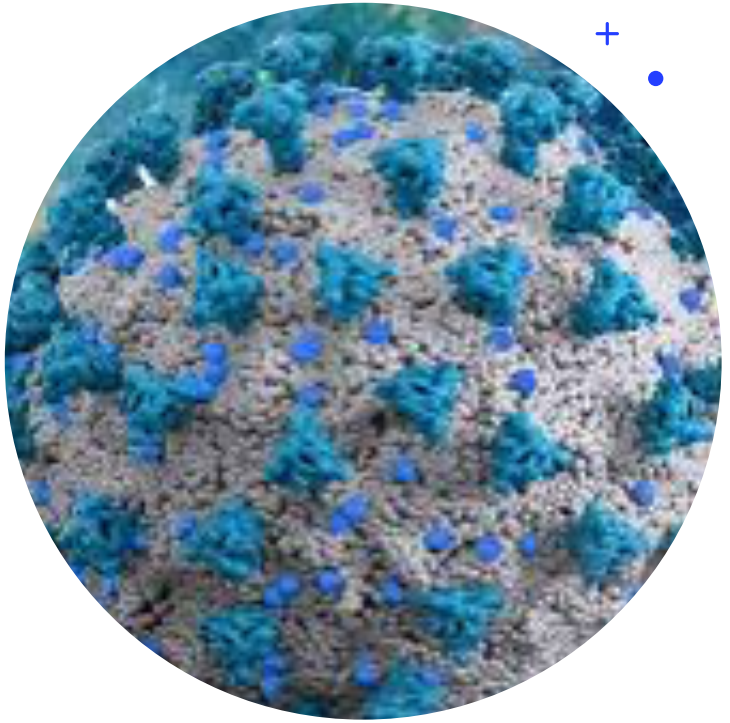
TEXAS

- Second largest state in the US (261,231.71 sq. miles)
  - Divided into 254 separate counties
- Second most populous state in the US (~ 29 million)
- Second largest economy in the US (10th largest in the world)
- Home of 3 of the US's 10 largest cities:
  - Houston (4th), San Antonio (7th), Dallas (9th)

https://www.mapsofworld.com/usa/states/texas/texas-county-map.html

COVID-19

- Disease caused by the SARS-CoV-2 coronavirus
- Common symptoms include fever, dry cough, and tiredness
- 20% of people with COVID develop severe breathing problems
- COVID-19 is spread by person to person transmission through small droplets expelled from the nose or mouth
- The current US morality rate (case/fatality ratio) is 3.2%



https://www.technologynetworks.com/diagnostics/product-news/indica-labs-and-octo-launch-covid-19-digital-pathology-repository-335166

QUESTION: CAN CHARACTERISTICS OF A TEXAS COUNTY EXPLAIN COVID-19'S IMPACT ON THAT COUNTY?

Gather data about Texas counties

- We found 23 different data sets containing a total of 123 features
- Features were diverse and included:
  - Population demographics
  - Education and employment
  - Health data
  - Info on airports, correctional facilities, military institutions, nursing facilities, and hospitals
  - Voting records
  - Home values
  - COVID policies
  - Land use
  - Transportation

Obtain targets that reflect COVID's impact on a county

- Total COVID cases and deaths were downloaded from the Texas's Department of Public Health website on 8/4/20
- Targets and calculated targets:
  - Total cases
  - Total deaths
  - Deaths per case
  - Cases per 1000 population
  - Deaths per 1000 population

Coding and computing tools used

- All code was written in Python using the pandas and sci-kit learn libraries
- Code was written in Jupyter Notebooks and executed using AWS Sagemaker
- Data was stored in AWS S3

Clean data and assemble the final dataset

- Data cleaning
  - Dropped irrelevant features
  - Formatted data frames
  - Changed strings to integers
  - Replaced categorical data with dummy variables
- Standardized county names
- Merged datasets

Feature reduction analysis

- Feature reduction measures
  - Removed 19 highly correlated features
  - Removed features with low variance
  - Performed PCA analysis
- Only removing the correlated features improved model performance

Run and optimize models

- Split data into a 30% test set and 70% train set
- Scaled data using normalization
- Trained regression models:
  - Linear Regression – Lasso
  - Linear Regression – Ridge
  - Linear Regression – Elastic Net
  - Support Vector Regressor
  - Decision Tree Regressor
  - Random Forest Regressor

Model Selection Results

Average R^2 score from a 3-fold cross validation

	Deaths per 1000	Cases per 1000	Deaths per Case	Total Deaths	Total Cases
Linear Regression - Ridge	-0.649	-123.5	-1.391	0.678	0.810
Linear Regression - Ridge (parameters optimized)	0.014	-1.517	0.001	0.651	0.785
Linear Regression - Lasso	-0.014	-1.521	-0.011	-0.008	0.883
Linear Regression - Lasso (parameters optimized)	.011	-1.52	-0.011	0.531	0.877
Linear Regression – Elastic Net	-0.014	-17.370	-0.011	0.678	0.751
Decision Tree Regressor	-0.582	-13.713	-1.256	0.261	0.289
Random Forest Regressor	-0.112	-19.592	-0.906	0.626	0.770
Support Vector Regression	-0.014	-0.056	-1.923	-0.057	-0.070

Models with positive R^2 values are shaded in pink

Conclusion 1:
The more populated a county, the more COVID cases and deaths

- Many models produced a high R^2 value when the target was total cases or total deaths.
- Important features in these models were population related suggesting that the more highly populated a county, the more likely the county is to have COVID cases or deaths.
  - Population and COVID cases have a 0.966 correlation coefficient
  - Population and COVID deaths have a 0.948 correlation coefficient

Conclusion 2:
This feature set is not predictive for the calculate targets

- None of the models satisfactorily explained the variation in deaths per 1000, cases per 1000, or deaths per case using the current feature set.

Next steps

- Additional feature engineering
- Obtain new features, especially those related to county and population actions during the pandemic, then repeat the modelling