ISyE 6740 - Computational Data Analytics - Summer 2020
Final Report

Team ID: 47
Team Member Names: Xuanxuan Xue (903093953), Shujie Xu (903452034), Liuyi Ye (903484423)
Project Title: Predicting Covid-19 Risk by County-Level Characteristics

1. Intro

Covid19 cases and fatality forecasting has been a popular topic ever since the beginning of 2020. There are quite a number of questions unanswered. On top of all the concerns, the society is challenged with the unpredictability of the unprecedented viruses. For instance, will the current mitigation strategies be sufficient to catch up with the speed of increasing cases? Will the current suppression method effectively reduce the reproduction number? To answer these questions, this analysis was created to test the hypotheses of whether county level demographic characteristics features can help predict the risk of fatality and cases going forward.

The machine learning algorithms and statistical analyses were engaged to verify above the hypotheses. To facilitate this study, we used a dataset with more than 300 initial variables of geo-demographic info summarized to county level. Since scientists and healthcare facilities come up with strategies as we go along, the most recently updated data as of 07/25/2020 was collected to evaluate the selected matrices. Georgia is our piloting state. Detailed explanation of data and methodology can be found in the following sections.

2. Data

We rely on data from several different sources: New York Times Covid19 dataset, County-level Socioeconomic dataset , and 2013 NCHS Urban - Rural Classification Scheme for Counties.

The main source of Covid-19 cases and deaths are taken from New York Times dataset. The Times has been tracking cases of Covid-19 and releasing a series of datafiles. We use the most up-to-date data (25 July 2020). We define risk as the relative position of a county's cumulative Covid-19 cases/deaths to the median of cases for Georgia counties. If a county has *number of cases* lower than the median, then the risk is low. If a county has *number of cases* higher than the median, then the risk is high. We create dummy variables called *"High_risk_by_deaths"* and *"High_risk_by_cases"* (1 if high risk, 0 otherwise)

The main variables of interest are taken from County-level Socioeconomics dataset, one such repository that aggregates over 10 public available governmental and academic sources such as *Census* and *American Community Survey* and includes socioeconomic factors on the county level. The variables can be categorized into 10 groups[1] and we manually pick some variables from the original dataset. A lot of variables have been studied and are confirmed to be very important predictors of variation in disease severity. For example, Desmet K et.al (2020) found that a wide range of correlates - population density, public transportation, age structure, nursing home residents, connectedness to source countries have effects on disease severity. Interestingly, they also found that Trump' s vote share in 2016 positively predicts cases and deaths.

The 2013 NCHS Urban-Rural Classification Scheme for Counties has rich information of U.S counties. We take one variable "Urbanization-level" from this source. This variable is about the urbanization level assignments for all U.S. counties and county-equivalent entities. For example, one county can be Small metro, Medium metro, Noncore, Micropolitan or Large fringe metro.

For this study, we restrict the sample to Georgia counties only. There are 159 observations in the sample. Table 1 is the summary statistics [2] and it includes variables relevant to temperature, economy, population density, age distribution, education distribution, income, unemployment, medical cares.

Table 1. Summary Statistics

Variable	Numer of Observation	mean	std	min	25%	50%	75%	max
Jan Temp AVG / F	159.0	47.597	3.777	38.7	44.55	48.2	50.8	53.6

[1] The groups are: Climate, Demographics, Education, Employment and median household income, Ethnicity, Healthcare, Housing, Identifying Variables,Population Estimates, and Transit Scores.

[2] A more detailed explanation of variables can be found here.

Feb Temp AVG / F	159.0	55.486	4.208	46.3	51.8	55.6	59.45	61.9
Mar Temp AVG / F	159.0	55.738	3.548	47.1	53.2	56.0	58.7	61.3
Apr Temp AVG / F	159.0	64.267	2.625	57.6	62.2	64.6	66.5	68.5
May Temp AVG / F	159.0	75.32	2.567	67.2	73.85	75.7	77.4	78.8
Jun Temp AVG / F	159.0	77.758	2.896	68.9	75.8	78.4	80.2	81.9
Jul Temp AVG / F	159.0	80.784	1.883	73.1	80.15	81.1	82.05	83.4
Aug Temp AVG / F	159.0	80.484	2.071	72.6	79.6	81.2	81.9	82.8
Sep Temp AVG / F	159.0	78.972	1.803	71.2	78.4	79.4	80.0	82.2
Oct Temp AVG / F	159.0	69.213	3.268	60.6	66.75	70.0	71.7	75.2
Nov Temp AVG / F	159.0	52.449	2.913	44.9	50.4	52.8	54.75	58.3
Dec Temp AVG / F	159.0	51.749	2.983	44.5	49.4	52.0	54.25	58.1
Density per square mile of land area - Population	159.0	193.579	378.84	8.5	35.25	66.3	155.0	2585.7
Total_age0to17	159.0	15759.44	34565.961	272.0	2485.0	5141.0	13567.5	249129.0
Total_age85plusr	159.0	922.22	1744.735	59.0	239.5	428.0	828.5	14296.0
Total_age65plus	159.0	9184.962	16657.402	423.0	2073.0	4217.0	8184.5	122730.0
Total_age18to64	159.0	41215.818	92627.372	913.0	6826.0	13385.0	34174.0	697977.0
Percent of adults with a bachelor’s degree or higher 2014-18	159.0	18.196	9.153	7.0	12.0	15.2	21.4	51.7
Less than a high school diploma 2014-18	159.0	5684.994	9446.918	350.0	1588.0	2805.0	5732.0	70656.0
Unemployment_rate_2018	159.0	4.438	0.919	3.0	3.7	4.2	5.0	7.7

Median_Household_Income_2018	159.0	47507.082	13626.51	28298.0	38399.0	43439.0	51704.0	105921.0
Active Physicians per 100000 Population 2018 (AAMC)	159.0	228.7	0.0	228.7	228.7	228.7	228.7	228.7
MD and DO Student Enrollment per 100000 Population AY 2018-2019 (AAMC)	159.0	28.6	0.0	28.6	28.6	28.6	28.6	28.6
Total nurse practitioners (2019)	159.0	30.283	65.681	0.736	5.272	10.35	25.486	480.661
ICU Beds	159.0	15.774	52.475	0.0	0.0	0.0	10.0	538.0
2013 NCHS scheme	159.0	3.27	1.381	0.0	2.0	4.0	4.0	5.0
High_risk_by_deaths	159.0	0.478	0.501	0.0	0.0	0.0	1.0	1.0
High_risk_by_cases	159.0	0.497	0.502	0.0	0.0	0.0	1.0	1.0

3. Method

In order to figure out if the risk of Covid-19 is predictable by local demographic data, and to compare the performances of each machine learning models, the following modeling have been fitted to this project:

(a) Logistic regression: simple binary logistic regression
(b) Ridge: *L1*-norm regularized binary logistic regression
(c) Lasso: *L2*-norm regularized binary logistic regression
(d) Decision Tree: classification Trees
(e) Random Forest: random decision trees
(f) Elastic Net: regularized binary logistic regression regression that linearly combines *L1*-norm and *L2*-norm

(g) GMM:
Since there were only two clusters: 0 for low risk, 1 for high risk, therefore the Gaussian mixture number of components was set to 2.

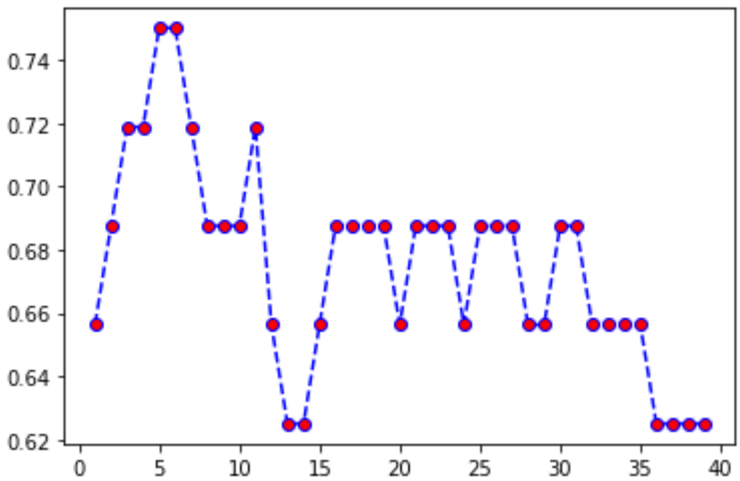
(h) KNN:
To tune the best fitted model, the number of nearest neighbors has iterated from 1 to 40. As Plot 1 shown below, k=6 has the highest accuracy rate of 75%.

(i) Naive Bayes: Naive Bayes Model

(j) K-means:
Again, because there are only two clusters so that the k has been set to 2.

(k) Neural Network: For this method, Adam was specified as the optimizer, and the Rectified Linear Unit was used as the activation function. In general, ReLU Performs better as compared to traditionally used activation functions such as Sigmoid and Hyperbolic-Tangent functions. 20 hidden layers were utilized in this model.

Figure 1. Accuracy Rate over N keighbors



(l) Support Vector Machine:
This method constructs a hyperplane or a set of hyperplanes on the high dimensional space, which was later used in classification.

4. Result

We have chosen 2 sets of dependent variables: *high_risk_by_case* and *high_risk_by_death*. One county will be classified to 1 (high risk) if the confirmed case or death is higher than the median of Georgia, otherwise it will be classified to 0. Data set was split to 80% training data and 20% te sting data. In order to keep it a fair game, random seed has been set to 12345678 for every model.

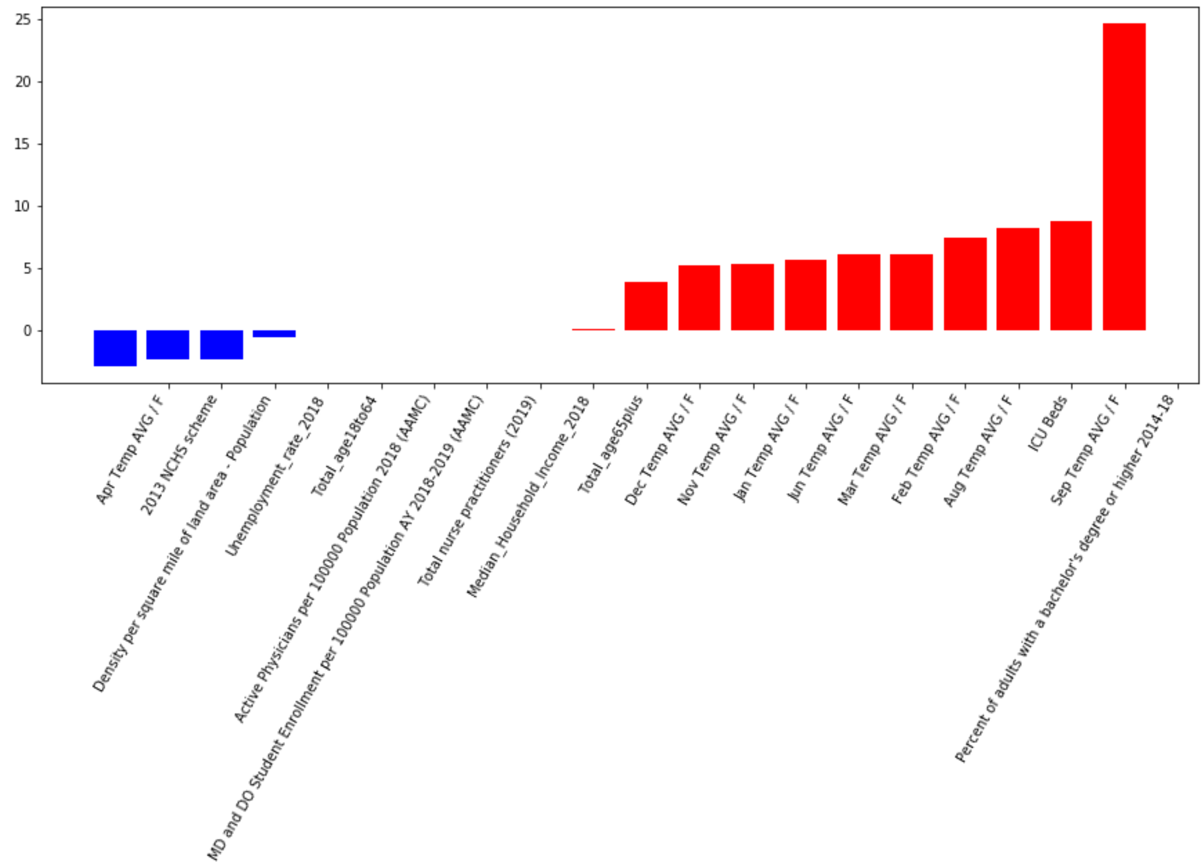
Firstly, we ran 12 models to predict risk level by confirmed cases. As Table 2 shown below, most of the models have accuracy rates from 75% to 81%, except GMM and K-means. Decision Tree and Neural Network have the best performance with accuracy rate of 81.25% among these models.

Table 2: Accuracy of models of predicting risk level by cases

	Logistic	Ridge	Lasso	Decission Tree	Random Forest	Elastic Net	GMM	KNN	Navie Bayes	K Means	Neural Network	SVM
Accuracy	78.13%	78.42%	78.7%	**81.25%**	78.13%	79.29%	46.88%	75%	75%	50%	**81.25%**	78.13%

Figure 2 shows the top 10 features which were used to form a hyperplane with the negative and positive weights in Support Vector Machine:

Figure 2. Top 10 features from SVM, Covid-19 cases



The average temperature is negatively correlated with the severity of cases in different counties. So it seems that higher the temperature, lower the severity of COVID19 infection in this area. This has been a verified fact in many published articles[3].

2013 NCHS indicates the urbanization of the region. And that' s an indicator of population density & transportation frequency as well. The higher the urbanization, the higher the chance of virus reproduction in this region.

ICU beds could be an indicator of the hospital size and an outcome of the aggregated infected cases.

We can also see that people older than 65 are strongly correlated with the infected probability. This has also been verified in the published articles[4].

[3] Richard Gray, 23rd March 2020, Will warm weather really kill off Covid-19?, BBC. Retrieved from https://www.bbc.com/future/article/20200323-coronavirus-will-hot-weather-kill-covid-19

[4] Centers for Diseases and Control Prevention https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/older-adults.html

Secondly, we ran 12 models to predict risk level by death cases. As Table 3 shown below, the models accuracy rates are scattered from 46.88% to 80.01%. Ridge regression has the best performance with an accuracy rate of 80.01%. However, the overall performance of 12 models is not as good as when high_risk_by_case was set to be the dependent variable.

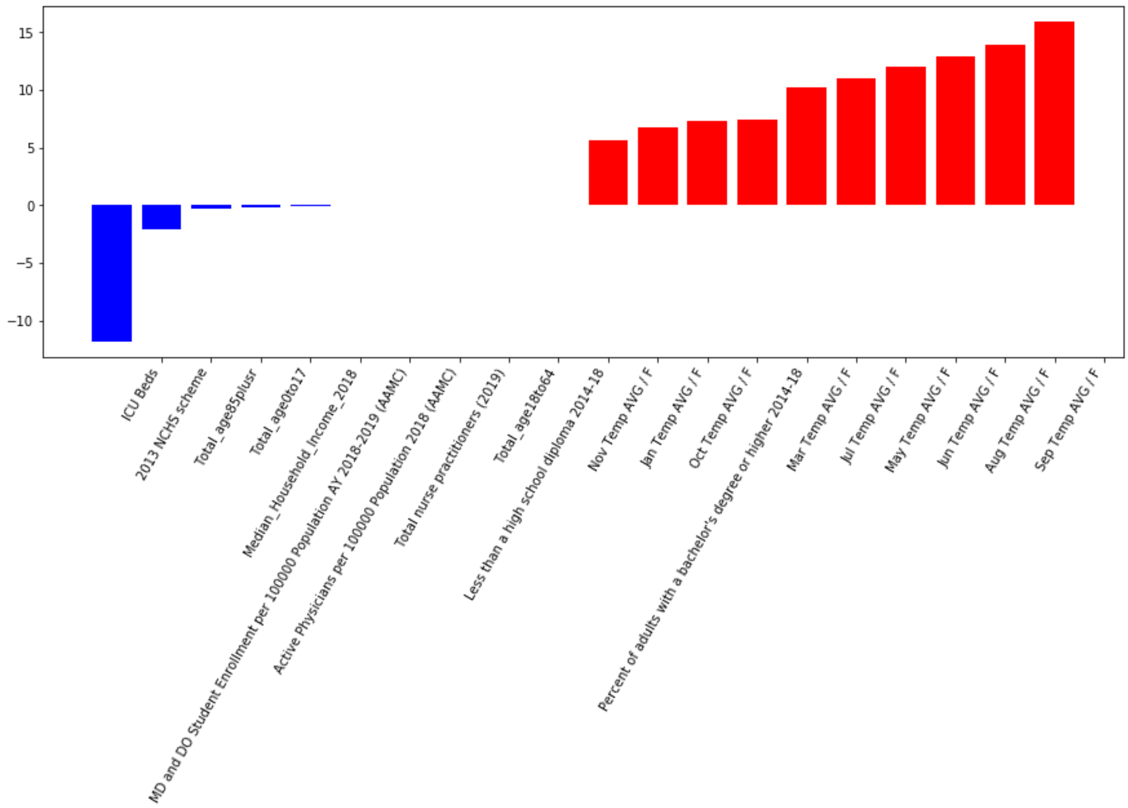
Table 3: Accuracy of models of predicting risk level by death cases

	Logistic	Ridge	Lasso	Decission Tree	Random Forest	Elastic Net	GMM	KNN	Navie Bayes	K Means	Neural Network	SVM
Accuracy	68.75%	**80.01%**	77.49%	62.5%	65.63%	77.94%	46.88%	68.75%	68.75%	50%	68.75%	65.62%

Figure 3 shows the top 10 features which were used to form a hyperplane with the negative and positive weights in Support Vector Machine:

Compared with Cases prediction, for death prediction ICU Beds was assigned with a much heavier weight. It' s the same with education level (Less than a high school diploma, Percent of adults with a bachelor' s degree or higher).

Figure 3. Top 10 features from SVM, Covid-19 deaths



5. Conclusion

In this project, we collected and manipulated county-level demographic characteristic data and focused on Georgia state’ s 159 counties, fitted above 12 machine learning models. The team had predicted risk levels by confirmed cases and death cases separately. Decision tree and Neural Network have the best predicting power with accuracy rate of 81.25% on predicting risk level by cases, and Ridge Regression has the best predicting power with accuracy rate of 80.01% on predicting risk level by death cases. The overall performance of 12 models to predict risk level by cases have better predicting power than predict risk level by death cases. This result echoed with our hypothesis that we can use county-level characteristics applied with machine learning models to predict Covid-19 risks.

6. Collaboration

Xuanxuan Xue	Explored and collectd data, fitted GMM, KNN, Naive Bayes, K-Means models, wrote Method, Result and Conclusion parts of this report
Shujie Xu	Manipulated data, fitted neural network, CART, Support Vector Machine, wrote Intro, Method, and Result parts of this rep

	ort
Liuyi Ye	Collected and cleaned data, fitted linear models (a) - (f), the Data part of the report

7. Reference:

Desmet, K. and Wacziarg, R., 2020. *Understanding Spatial Variation in COVID-19 across the United States* (No. w27329). National Bureau of Economic Research.

Bordalo, P., Coffman, K.B., Gennaioli, N. and Shleifer, A., 2020. *Older People are Less Pessimistic about the Health Risks of Covid-19* (No. w27494). National Bureau of Economic Research.

Papageorge, N.W., Zahn, M.V., Belot, M., van den Broek-Altenburg, E., Choi, S., Jamison, J.C. and Tripodi, E., 2020. *Socio-Demographic Factors Associated with Self-Protecting Behavior during the COVID-19 Pandemic* (No. 13333). Institute of Labor Economics (IZA).