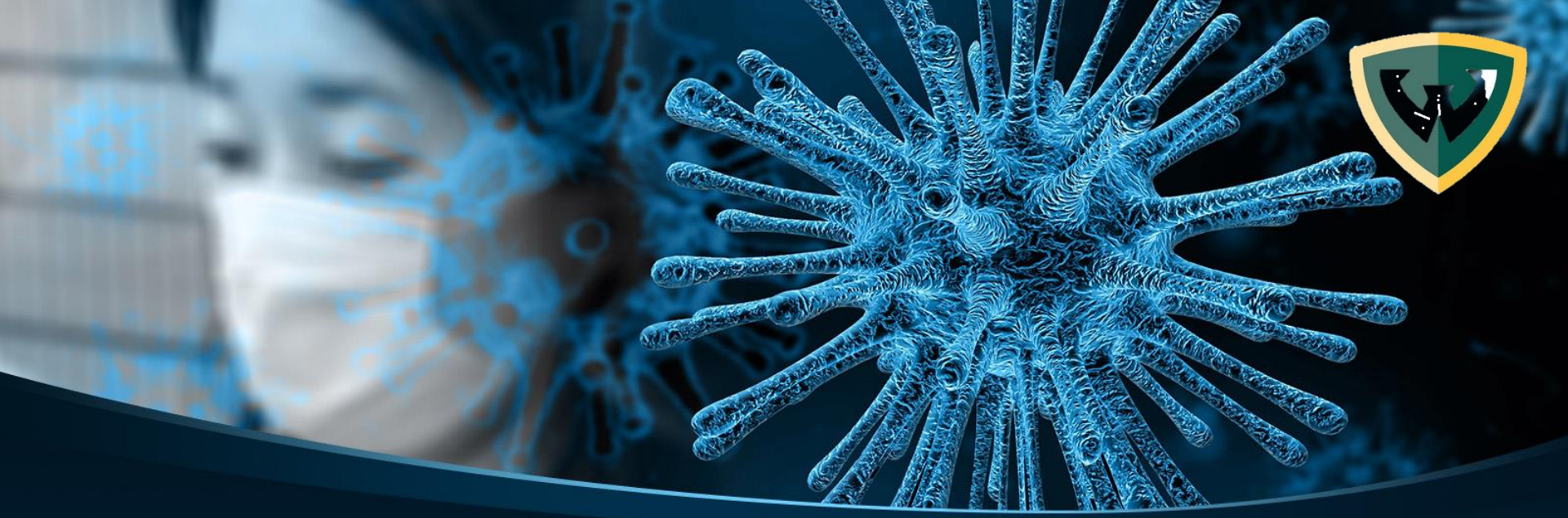


Predicting the Spread of COVID-19

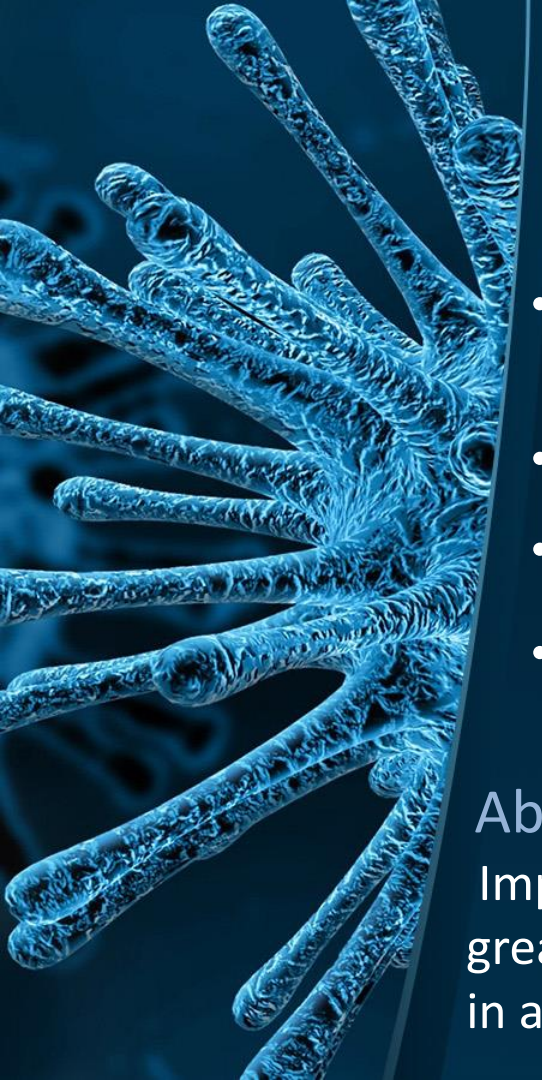
DSE 6300: Data Science Application Development

By: Kyle W. Brown

4/27/2020

Overview

- Project Proposal
- Project Schedule
- Data
- Pipeline
- Models
- Visualization
- Summary



Project Proposal

- Build a functional dashboard in Shiny or Tableau presenting COVID19 virus predictions and data.
- Stream COVID data using Kafka or Spark Streaming.
- Visualization of predictive analytics model and data.
- Predictions or time series analysis based on COVID19, environmental, and economic factors.

Abstract:

Implement a predictive model to identify which factors have the greatest impact on the spread and strength of the COVID-19 virus in a geographic area.

Project Schedule

Shortened Agile Framework

- 2 (1) week iterations
- 2 checkpoints indicating updates
- Checkpoint 2 is midterm
- Agile Schedule used for task list tracking

Project Schedule

Project Task List

	Completed
Objectives	√
Data	√
Pipeline	√
Machine learning	√
Visualization	√
Dashboard	√
Integration	

Final Week Schedule

Thu	Fri	Sat	Sun	Mon	Tue
4/16/2020	4/17/2020	4/18/2020	4/19/2020	4/20/2020	4/21/2020
Final Modeling	Model update, submit models by 7:00 p.m.	• Group model selection • Kafka setup	• Start ppt • Kafka topics, records, cluster	• Finish ppt parts •Finish pipeline	Finalize ppt

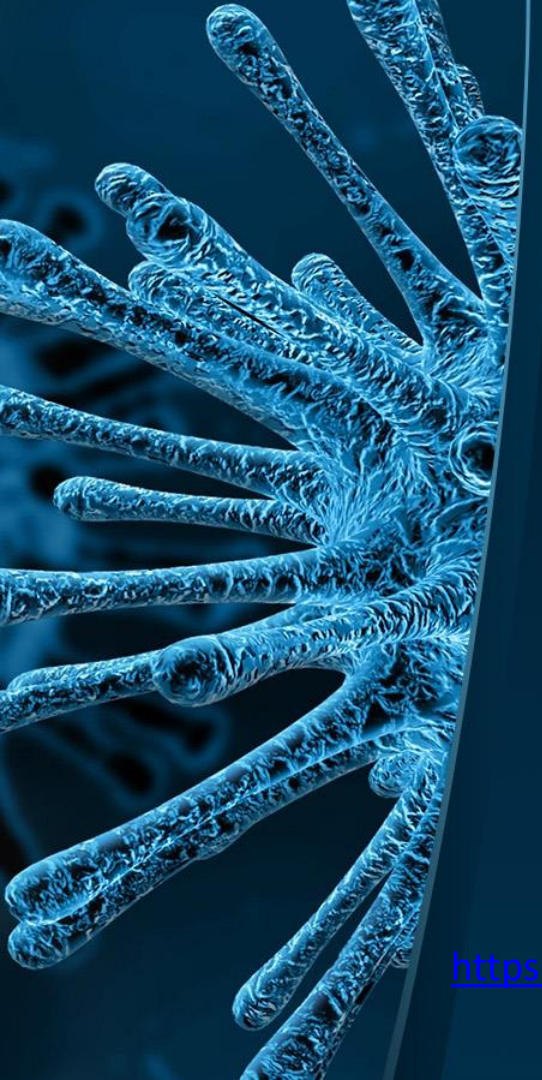
Data

John Hopkins University, GitHub repo

https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_daily_reports_us

Daily Reports for United States

Province_State	Last_Update	Lat	Long_	Confirm	Deaths	Recovered	Active	FIPS	Incident_Rate	People_Tested	People_Hospitalized	Mortality_Rate	Testing_Rate	Hospitalization_Rate
Alabama	4/15/2020 22:56	32.3182	-86.9023	4075	118	0	3957	1	86.91	34077	525	2.90	726.76	12.88
Alaska	4/15/2020 22:56	61.3707	-152.404	293	9	106	284	2	49.02	8664	34	3.07	1449.44	11.60
American Samoa	4/15/2020 22:56	-14.271	-170.132	0	0	0	0	60	0	3	0	0	5.39	0
Arizona	4/15/2020 22:56	33.7298	-111.431	3964	142	385	3822	4	54.46	45310	590	3.58	622.50	14.88
Arkansas	4/15/2020 22:56	34.9697	-92.3731	1569	33	489	1536	5	60.60	21834	130	2.10	843.33	8.29
California	4/15/2020 22:56	36.1162	-119.682	26686	861	0	25825	6	68.06	216486	5163	3.23	552.14	19.35
Colorado	4/15/2020 22:56	39.0598	-105.311	7956	328	0	7628	8	140.40	39580	1556	4.12	698.45	19.56
Connecticut	4/15/2020 22:56	41.5978	-72.7554	14755	868	0	13887	9	413.85	50143	1908	5.88	1406.42	12.93



Data

NY Times COVID-19, GitHub

date	county	state	fips	cases	deaths
1/21/2020	Snohomish	Washington	53061	1	0
1/22/2020	Snohomish	Washington	53061	1	0
1/23/2020	Snohomish	Washington	53061	1	0
1/24/2020	Cook	Illinois	17031	1	0
1/24/2020	Snohomish	Washington	53061	1	0
1/25/2020	Orange	California	6059	1	0
1/25/2020	Cook	Illinois	17031	1	0
1/25/2020	Snohomish	Washington	53061	1	0

https://raw.githubusercontent.com/nytimes/covid-19-data/master/us-counties.csv

Data

STATWORX COVID19 API

date	day	month	year	cases	deaths	country	code	population	cases_cum	deaths_cum
4/16/2020	16	4	2020	30148	4928	United_States_of_America	US	327167434	639664	30985
4/15/2020	15	4	2020	26922	2408	United_States_of_America	US	327167434	609516	26057
4/14/2020	14	4	2020	25023	1541	United_States_of_America	US	327167434	582594	23649
4/13/2020	13	4	2020	27620	1500	United_States_of_America	US	327167434	557571	22108
4/12/2020	12	4	2020	28391	1831	United_States_of_America	US	327167434	529951	20608
4/11/2020	11	4	2020	35527	2087	United_States_of_America	US	327167434	501560	18777
4/10/2020	10	4	2020	33901	1873	United_States_of_America	US	327167434	466033	16690
4/9/2020	9	4	2020	33323	1922	United_States_of_America	US	327167434	432132	14817
4/8/2020	8	4	2020	30613	1906	United_States_of_America	US	327167434	398809	12895

Kafka Pipeline

Kafka directory to Wayne State's Grid

```
[gd3384@cehg ~]$ cd $KAFKA_HOME
```

Creating a COVID19 Test Topic

```
[gd3384@cehg kafka-broker]$ bin/kafka-topics.sh --zookeeper cehpn:2181 --create
--topic COVID19-kafka-test --replication-factor 1 --partitions 1
Created topic "COVID19-kafka-test".
```

Quick Test and Consumer contents for Topic in command

```
[gd3384@cehg kafka-broker]$ bin/kafka-console-producer.sh --broker-list cehpn:90
92 --topic COVID19-kafka-test
>bin/kafka-console-consumer.sh --zookeeper cehpn:2181 --topic COVID19-kafka-test
 --from-beginning
>
```

Connecting SparkR and Kafka to WSU Grid

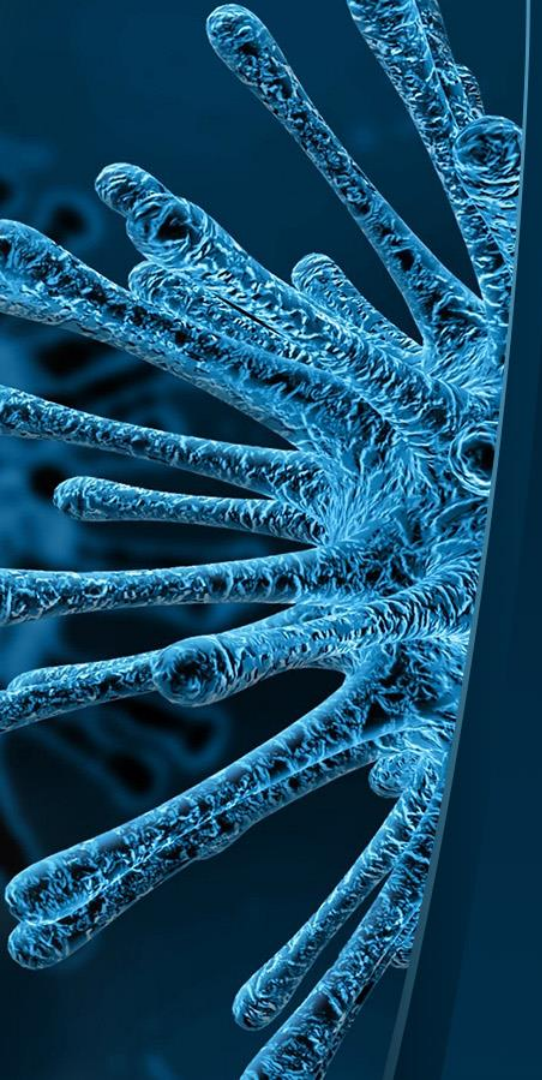
Connected SparkR and Kafka to Grid

```
sc <- spark_connect(master = "local", config = list(
  sparklyr.shell.packages = "org.apache.spark:spark-sql-kafka-0-10_2.11:2.
))
```

```
> sc <- spark_connect(master = "local", config = list(
+   sparklyr.shell.packages = "org.apache.spark:spark-sql-kafka-0-10_2.11:2.4.0"
+ ))
* Using Spark: 2.4.3
```

Example setup for reading Kafka stream with SparkR in Grid

```
stream_read_kafka(
  sc,
  options = list(
    kafka.bootstrap.server = "host1:9092, host2:9092",
    subscribe = "<topic-name>"
  )
)
```



Models

- Regressions
  - Linear
  - Mulivariate
  - Non-Linear
- Principal Component Analysis
- Time Series
- AutoML Deep Learning

Multivariable Regression

```
# Linear Regression of JHU Time Series

# Mortal Rate as the Predictor
mortalUSlm <- read.csv("https://raw.githubusercontent.com/WorldCapital/COVID19-Project/master/COVID-19/JHU_county
mortalUSlm <- lm(Mortality_Rate ~ Confirmed + Deaths + Recovered + FIPS + Incident_Rate  + People_Tested  + Peop
confint(mortalUSlm)
summary(mortalUSlm)
#Multiple R-squared:  0.6843, Adjusted R-squared:  0.5264
#F-statistic: 4.334 on 9 and 18 DF,  p-value: 0.003951


mse = mean(mortalUSlm$residuals^2)
print(paste0("MSE= ", mse))
#[1] "MSE= 0.717658874682637"

print(paste0("RMSE= ", RMSE(mortalUSlm$residuals)))

#[1] "RMSE= 0.847147492873961"
```

Non-Linear Regressions

```
#Non-Linear Model NY Times using Deaths
nyTimesNLM <- read.csv("https://raw.githubusercontent.com/WorldCapital/COVID19-Project/master/COVID-19/nyt-us-co
nyTimesNLM <- lm(deaths ~ fips + cases, I(cases^2), data = ny_Times_Counties)
confint(nyTimesNLM)
summary(nyTimesNLM)
#Residual standard error: 0.6038 on 55656 degrees of freedom
#(3543 observations deleted due to missingness)
#Multiple R-squared:  0.9265, Adjusted R-squared:  0.9265
#F-statistic: 3.51e+05 on 2 and 55656 DF,  p-value: < 2.2e-16

mse = mean(nyTimesNLM$residuals^2)
print(paste0("MSE= ", mse))

#[1] "MSE= 0.364578917674488"

print(paste0("RMSE= ", RMSE(nyTimesNLM$residuals)))

#[1] "RMSE= 0.603803707900579"
```

AutoML Models

```
#Split data into Train/Validation/Test Sets
split_h2o <- h2o.splitFrame(conv_data.hex, c(0.6, 0.2), seed = 1234 )
train_conv_h2o <- h2o.assign(split_h2o[[1]], "train" ) # 60%
valid_conv_h2o <- h2o.assign(split_h2o[[2]], "valid" ) # 20%
test_conv_h2o  <- h2o.assign(split_h2o[[3]], "test" )  # 20%


#Model
# Set names for h2o
target <- "cases_cum"
predictors <- setdiff(names(train_conv_h2o), target)
# Run the automated machine learning
automl_h2o_models <- h2o.automl(
  x = predictors,
  y = target,
  training_frame    = train_conv_h2o,
  leaderboard_frame = valid_conv_h2o
)
```

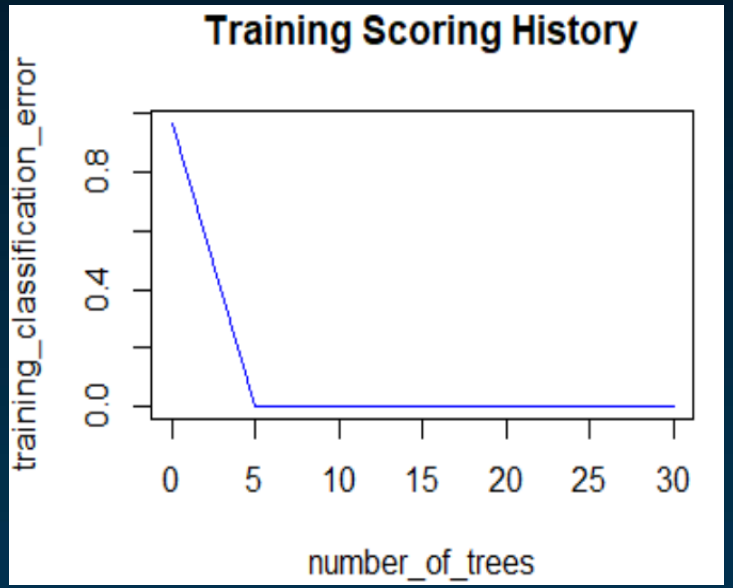
```
#Cross-Validation Metrics Summary:
#                              mean           sd cv_1_valid cv_2_valid cv_3_valid cv_4_valid cv_5_valid
#accuracy                 0.30769232  0.094211146 0.46153846 0.23076923 0.23076923 0.30769232 0.30769232
#err                       0.6923077  0.094211146 0.53846157  0.7692308  0.7692308  0.6923077  0.6923077
#err_count                       9.0    1.2247449        7.0       10.0       10.0        9.0        9.0
#logloss                  3.4635184    0.4602315  2.7115996  3.6106849  3.9662728   3.548047  3.4809875
#max_per_class_error              1.0          0.0        1.0        1.0        1.0        1.0        1.0
#mean_per_class_accuracy 0.86153847   0.01884223  0.8923077 0.84615386 0.84615386 0.86153847 0.86153847
#mean_per_class_error    0.13846155   0.01884223 0.10769231 0.15384616 0.15384616 0.13846155 0.13846155
#mse                      0.6727001   0.08535113 0.53059614 0.71209335  0.7577791  0.6815926 0.68143946
#r2                       0.9982088 6.2457175E-4  0.9988421 0.99843067 0.99719256 0.99811006  0.9984687
#rmse                    0.81877226  0.053760055 0.72842026  0.8438563  0.8705051  0.8255862 0.82549345
```

Model Results

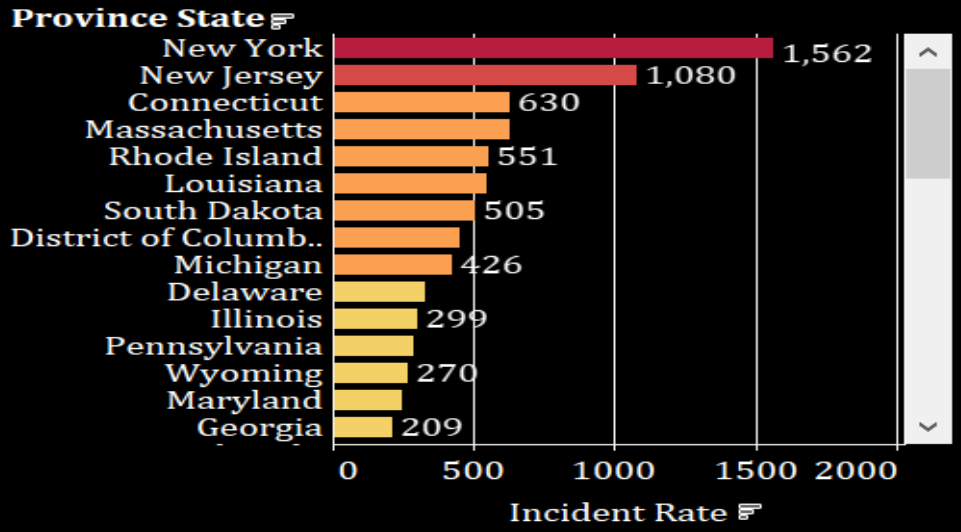
Model	r^2	rmse	mse
DeepLearning_grid__2_AutoML_20200419_000459_model_6	99.80%	87.24%	76.10%
GBM_grid__1_AutoML_20200419_000459_model_3	99.80%	84.66%	71.67%
GBM_grid__1_AutoML_20200419_000459_model_19	99.80%	80.01%	64.02%
GBM_grid__1_AutoML_20200419_000459_model_11	99.80%	75.51%	57.02%
GBM_grid__1_AutoML_20200419_000459_model_14	99.80%	75.48%	56.97%
GBM_grid__1_AutoML_20200419_000459_model_17	99.80%	74.26%	55.15%
Mortality_Rate_JHU_lm	52.64%	71.77%	84.72%
cases_cum_gbm_AutoML	99.81%	67.69%	82.16%
deaths_cum_gbm_AutoML	99.76%	49.60%	70.25%
Deaths_JHU_nlm	92.65%	36.46%	60.38%

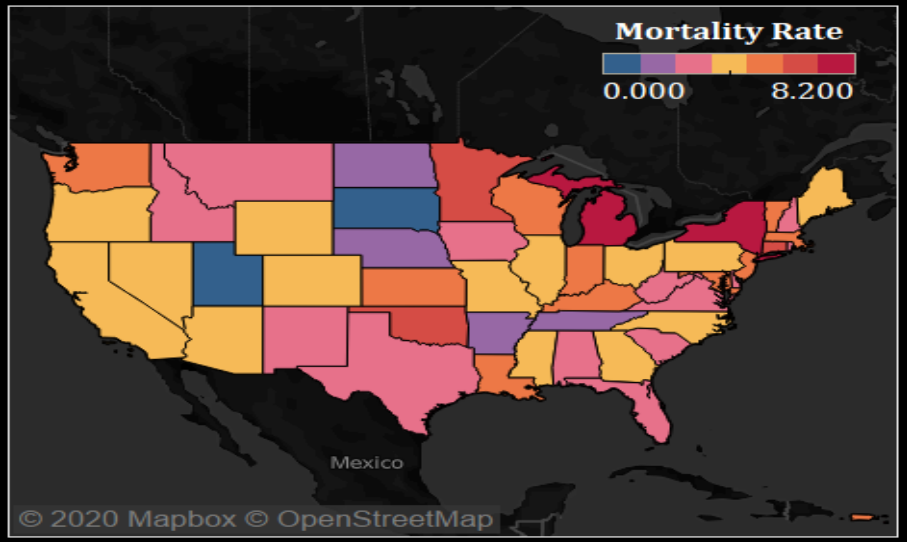
GBM Multinomial AutoML Leader

cases_cum_gbm_model

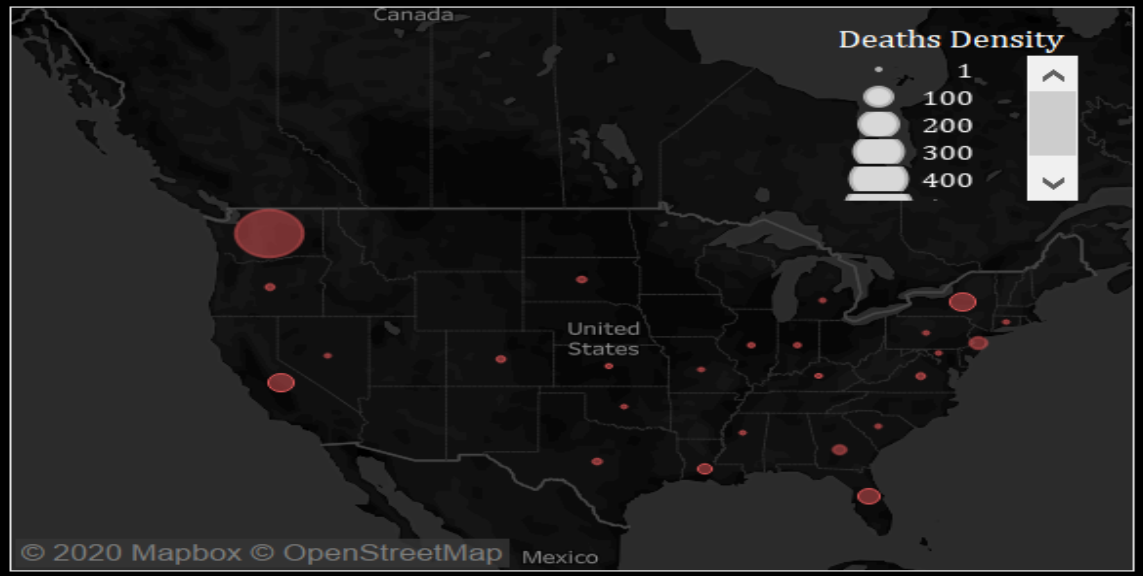


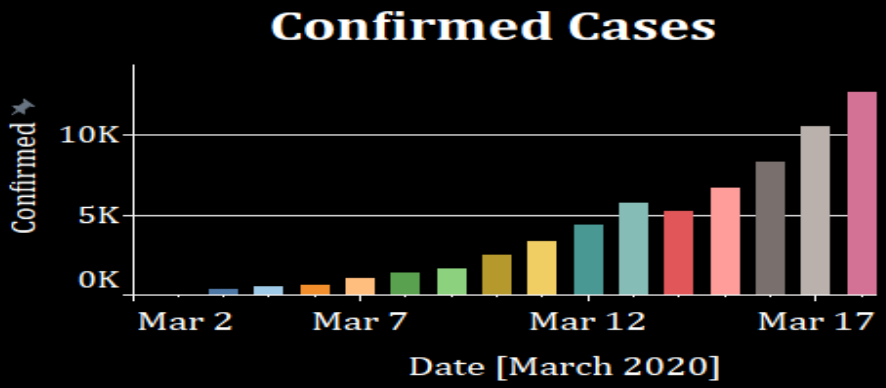
COVID-19 U.S.A. Dashboard

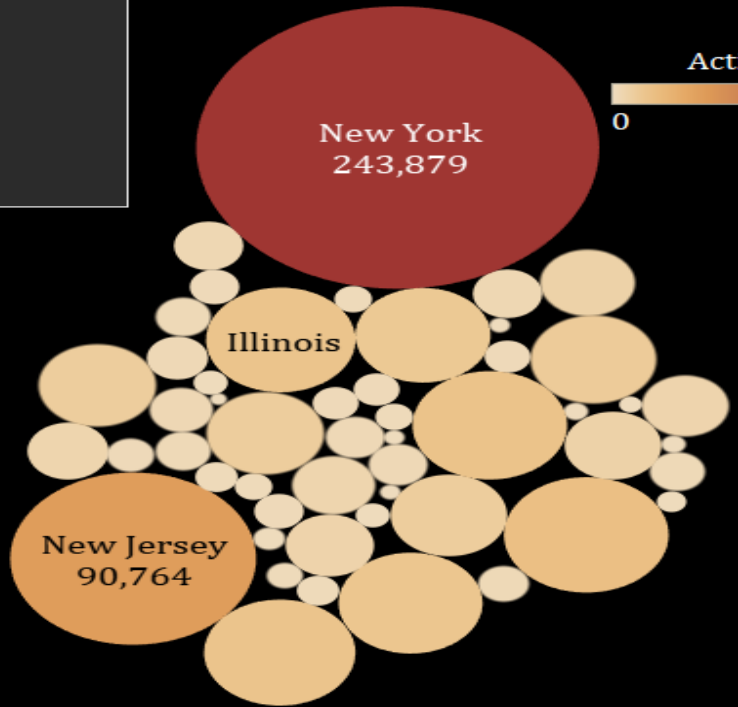


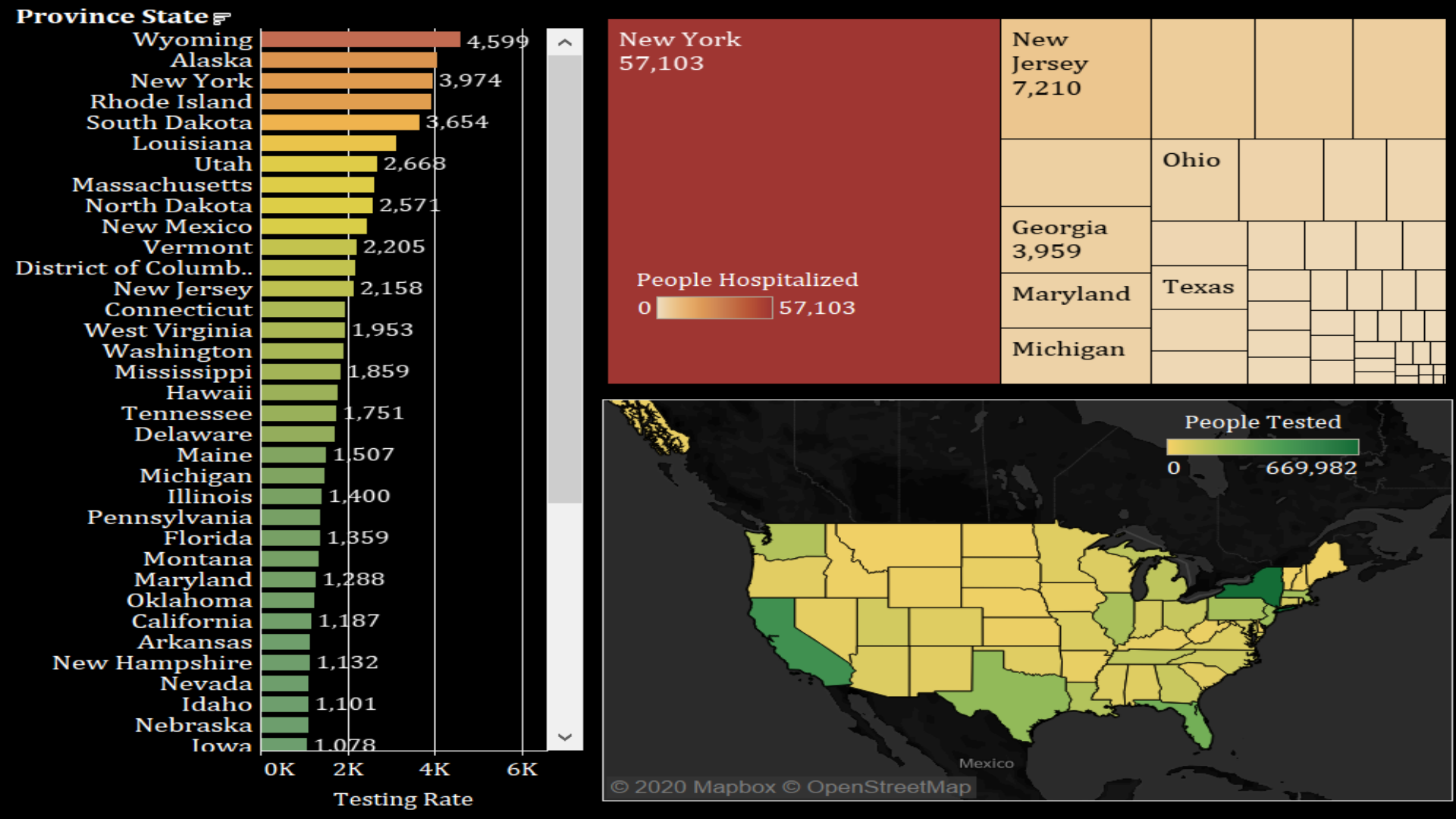


Province State	Confirmed..	Deaths..	Recovered..	Active	Fips	Incident Rate	People Tested
Alabama	5593	196	0	5397	1	119.28	48760
Alaska	335	9	196	326	2	56.04	12159
Arizona	5473	231	1265	5242	4	75.19	56601
Arkansas	2276	42	863	2234	5	87.91	29713
California	37344	1421	0	35923	6	95.24	465327
Colorado	10891	506	0	10385	8	192.19	48704
Connecticut	22469	1544	0	20925	9	630.22	69918
Delaware	3200	89	599	3111	10	328.62	16553
District of Columbia	3206	127	645	3079	11	454.27	15502
Florida	28309	893	0	27416	12	133.33	288627


Canada
Deaths Density
1
100
200
300
400
United States
© 2020 Mapbox © OpenStreetMap
Mexico


Confirmed Cases
Confirmed
10K
5K
0K
Mar 2
Mar 7
Mar 12
Mar 17
Date [March 2020]

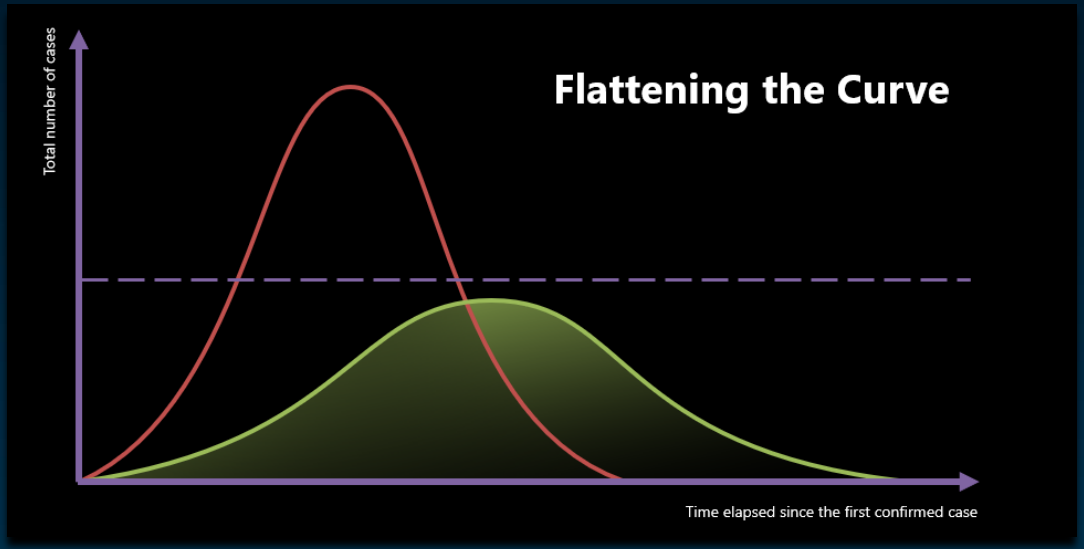

Active
0
244K
New York
243,879
Illinois
New Jersey
90,764


Province State
Wyoming 4,599
Alaska
New York 3,974
Rhode Island
South Dakota 3,654
Louisiana
Utah 2,668
Massachusetts
North Dakota 2,571
New Mexico
Vermont 2,205
District of Columb..
New Jersey 2,158
Connecticut
West Virginia 1,953
Washington
Mississippi 1,859
Hawaii
Tennessee 1,751
Delaware
Maine 1,507
Michigan
Illinois 1,400
Pennsylvania
Florida 1,359
Montana
Maryland 1,288
Oklahoma
California 1,187
Arkansas
New Hampshire 1,132
Nevada
Idaho 1,101
Nebraska
Iowa 1,078
0K 2K 4K 6K
Testing Rate
New York
57,103
People Hospitalized
0 57,103
New Jersey
7,210
Ohio
Georgia
3,959
Texas
Maryland
Michigan
People Tested
0 669,982
Mexico
© 2020 Mapbox © OpenStreetMap

Summary

- Provided analysis of COVID19 pandemic to determine any relationships or trends.
- Followed shortened Agile framework for project schedule using checkpoints.
- Created COVID19 test topic with Kafka pipeline and setup SparkR to read the Kafka stream into WSU's Grid.
- Selected GBM Multinomial model based AutoML model based on 0% training error, 99% R-Squared, and 82.16% RMSE.
- Designed Tableau dashboard to display COVID-19 pandemic data.

Any Questions or Comments?



Thank you