Samantha Traversi

MDS 576

Capstone Project

Rate of SARS-CoV-2 In California Wastewater As an Indicator of Public Health

Executive summary

As the severe symptoms and hospitalization due to COVID-19 decline, the rate of reporting cases of COVID-19 for government record is also in decline. The average California resident is unlikely to disclose a positive result of a self-administered antigen test to the appropriate government agency, so it appears as if the rate of infection is in steeper decline than reality. The California Department of Public Health, and California Health and Human Services need a different standard metric to accurately measure COVID-19 infection rates within communities. Using the frequency of the SARS-CoV-2 virus within wastewater provides an accurate measurement of communal infection rates under this decline in reporting individual cases.

Table of contents

Introduction

Ever since March of 2020, the state of California has seen widespread effects of the COVID-19 pandemic, much like the rest of the world. Even though there were public health agencies that emphasized the vital importance of abiding by public health measures such as reporting, public distancing, and vaccinating once the vaccinations became available to the general public. Unfortunately, some public health measures were heavily politicized, leading to mass denial and lack of following the recommended measures. Furthermore, as the virus has evolved, the severity of symptoms have decreased for the general population as well. Due to both of these circumstances, the level of reporting of COVID-19 cases has decreased significantly even though California public health services would still like to maintain accurate metrics of viral load within communities. As a result, we look to measure the quantity of the virus within wastewater systems to maintain a measurement of COVID-19 within the community.

Business understanding

The California Department of Public Health considers a probable case of COVID-19 to be "individuals with a positive antigen test that detects the presence of viral antigens" (2022), while confirmed cases require the presence of viral genetic material, such as that which is detected with a polymerase chain reaction test. With the increase in availability of at-home testing kits, and an overall decrease in severity of symptoms of COVID-19, many probable cases of COVID-19 go unreported to Public Health departments.

Another factor that decreases the severity of the symptoms of COVID-19 is the rate of vaccinations that has been increasing across the state of California. Individuals may shed the virus, potentially continuing its circulation within local communities, but show no symptoms of infection, or such low-severity symptoms that they do not realize their illness is caused by the SARS-CoV-2 virus, and not a common rhinovirus.

SARS-CoV-2 is the virus that causes COVID-19, and it has been found in wastewater sampled from water in close proximity to treatment facilities. Even though this virus is mainly spread through the respiratory tract, the virus is also shed via fecal matter, which ends up in wastewater systems. There is no evidence at this time that the shed virus in untreated water can infect human hosts, but instead the viral content of these samples are a measure of overall public health while reporting of probable cases of COVID-19 has dropped off.

Data understanding

The principal dataset used in this project was sourced from the California Surveillance of Wastewater Systems Network, a collaborative effort between the California Department of Public Health and the California State Water Resources Control Board. The purpose of collecting

this data was to detect and quantify “SARS-CoV-2 virus shed into wastewater via feces of infected persons” (2022). Features used from this dataset include the following:

- Sample ID - A Key value assigned to each collected sample
- Epaid - Permit number for the wastewater treatment plant from where the sample was collected (unique to each treatment facility)
- Sample Collection Date - This is the date the sample was collected for analysis
- Zipcode - Zip code of the sample collection site
- Population Served - Estimated number of persons served by this treatment facility
- Sample Matrix - Wastewater matrix from which the sample was collected
- PCR Type - The type of PCR used to quantify PCR target
- PCR Target Average Concentration - Concentration of the PCR target back-calculated to unconcentrated sample basis

Another dataset used in this project was sourced from the Public Health Alliance of Southern California initiative to track Vaccine Progress. Features used from this dataset include the following:

- As of Date - This is the date of collection of cumulative vaccine data, landing once a week on Tuesdays, from January 5, 2021 to December 13, 2022
- Zipcode - Zip code based on where the individual lives
- County - The California County in which the cases were recorded
- Vaccine Equity Metric - Health Equity Metric Score Quartile with 1 being the least healthy community conditions and 4 being the most healthy community conditions
- Percent of Population Fully Vaccinated - The total number of people fully vaccinated divided by the total population
- Percent of Population Partially Vaccinated - The total number of people partially vaccinated divided by the total population

The third dataset used in this project was sourced from the California Department of Public Health initiative to track Probable Cases. Features used from this dataset include the following:

- Date - This is the date of collection of the probable cases data, landing once a week on Tuesdays, from January 5, 2021 to December 13, 2022
- Area - The California County in which the cases were recorded
- Probable Cases - Probable cases that were reported on, or in the six days prior to the date of collection of data.

Data preparation

Data Cleaning

The datasets used in this project had few null values, but there were instances where the value of the variable was equal to zero. There were also very few major outliers within the data, so overall the original data was in pretty good shape, requiring minimal cleaning. Since there were so few null values, where null values existed for measurements of rates of COVID-19 or SARS-CoV-2, those instances were dropped entirely.

One variable that was a significant signal for cleaning was the PCR Type. There was one PCR Type within this attribute that only about ten percent of the samples were categorized, “QPCR”. The resulting PCR Target Average Concentration of that sample category was on a different scale than the remaining 90% of the samples, which were all measured using the “DDPCR” type. The QPCR types were removed from the dataset for continuity of target variable measurements.

Another variable that was binned were the sample collection dates. Samples were collected on a daily basis, but it was determined that to get a general read on the health of the surrounding community, a weekly metric would provide sufficiently accurate results while also making the modeling process smoother. Samples collected from Wednesday to Tuesday were binned into the Tuesday of the final day of collection.

Data Exploration

At a glance, there are no major surprises in the data.

```
In [7]: dfWW.describe()
```

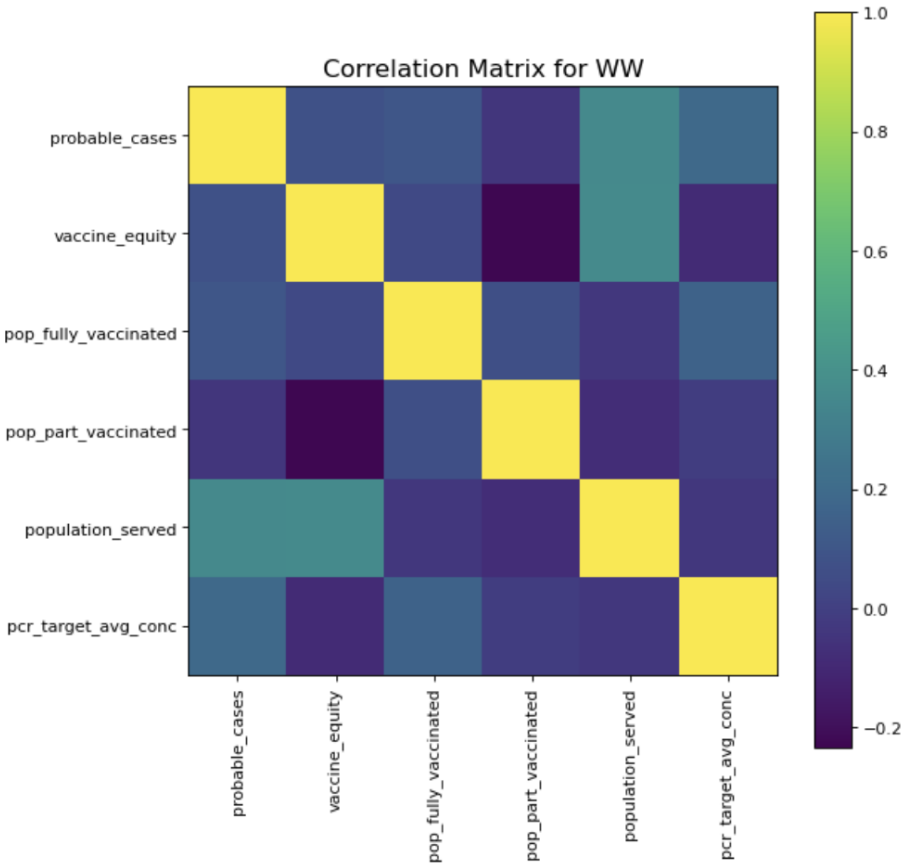
Out[7]:

	probable_cases	vaccine_equity	pop_fully_vaccinated	pop_part_vaccinated	population_served	pcr_target_avg_conc
count	7340.000000	7340.000000	7340.000000	7340.000000	7.340000e+03	7.340000e+03
mean	314.341281	2.977929	0.616174	0.081561	8.079963e+05	1.736787e+05
std	877.878736	1.028973	0.249441	0.106717	1.039373e+06	3.648874e+05
min	0.000000	1.000000	0.000000	0.000000	3.272000e+03	0.000000e+00
25%	22.000000	3.000000	0.518193	0.047398	1.530000e+05	1.493550e+04
50%	57.000000	3.000000	0.684597	0.064177	2.360000e+05	4.875400e+04
75%	226.000000	4.000000	0.791278	0.082096	1.480000e+06	1.952188e+05
max	15885.000000	4.000000	1.000000	0.978859	4.000000e+06	1.320746e+07

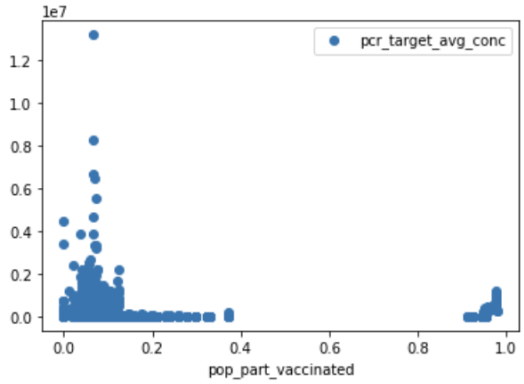
It also appears as though there are no major correlations between the data.

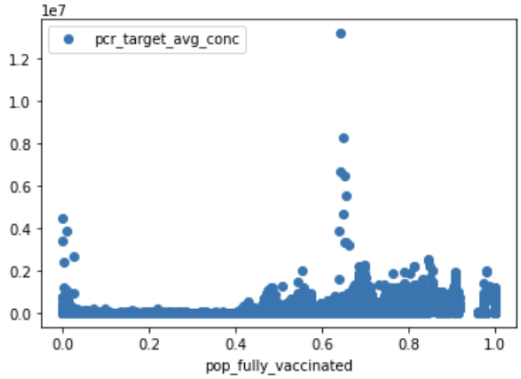
In [9]:
```
#Find variables with highest correlations
corr_matrix = dfWW.corr()
print(corr_matrix["pcr_target_avg_conc"].sort_values(ascending=False))
```

```
pcr_target_avg_conc     1.000000
probable_cases          0.191125
pop_fully_vaccinated    0.156368
pop_part_vaccinated    -0.008052
population_served      -0.035352
vaccine_equity         -0.085665
Name: pcr_target_avg_conc, dtype: float64
```

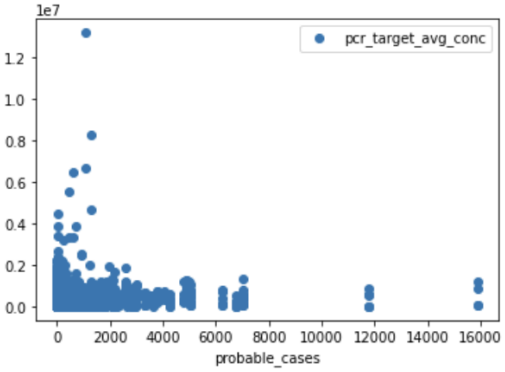


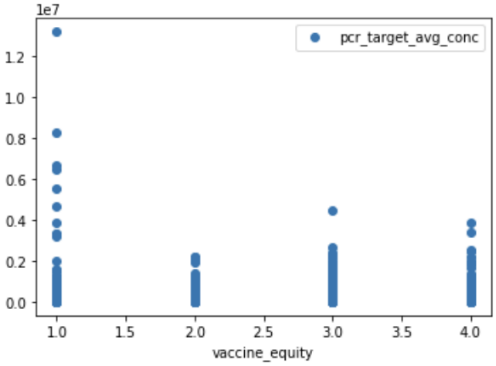
If we plot individual features against the target variable, there seems to be some correlation, but no one variable is a strong determinant for the quantity of viral particles found in wastewater systems. Below are percentages of the population that are partially and fully vaccinated plotted against the target variable.





We can also see how there is no strong correlation between probable cases and the quantity of viral particles found in the wastewater system. The 1-4 measurement of community health does show some correlation here (communities that were given a low rating for overall health tend to have the highest quantities of SARS-CoV-2 in wastewater).





Feature Engineering and Data Merging

Several variables were merged into this dataset from other repositories, also collected by California government institutions. The vaccine progress among communities in California was added as an attribute, matched to sample zip codes and dates. Since the COVID-19 vaccination takes at least ten days to take effect, these metrics were matched to the samples from the next week (i.e. the vaccinations recorded from January 6 to January 12, 2021 were matched to samples collected from January 13 to January 19, 2021 to match community vaccination level that was currently in effect.

Another variable that was merged into this dataset is probable cases of COVID-19 divided by county. These cases were collected on a daily basis; however, since the majority of the remaining data was chunked into weeks instead of days, the cases were binned into weeks. They were also matched to the sample counties and dates to be merged into this dataset.

The target variable, the average concentration of the target pcr in wastewater samples, was converted to an integer. Many of the numbers were listed as float point numbers with a wide variety in digits to the right of the decimal. Since there was such a large range of values, they were binned into quartiles and given values similar to riptide warning flags (green, yellow, orange, red). Here we can see the mean of each numerical attribute within these bins.

Out[25]:

class	probable_cases	vaccine_equity	pop_fully_vaccinated	pop_part_vaccinated	population_served	pcr_target_avg_conc
green	143.280807	2.912759	0.564643	0.106731	902511.375136	6423.150491
orange	269.420708	3.032153	0.617124	0.062672	764077.235422	100256.041417
red	701.475490	2.807734	0.719839	0.083323	765834.186275	558828.490741
yellow	142.883924	3.159128	0.563004	0.073531	799636.900272	28906.099728

Modeling

After some initial auto-modeling tests for classification models, the two models that performed with the highest accuracy were XGBoost and Gradient Boosted Machine. Screenshots of hand-tuned models are included below:

XGBoost

In [24]:
```
#XGBoost split data into input and output
X = dfWW[['probable_cases', 'vaccine_equity', 'pop_fully_vaccinated', 'pop_part_vaccinated', 'population_served']]
Y = dfWW['class']
```

In [25]:
```
#split data into train and test sets
seed = 7
test_size = 0.33
X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=test_size, random_state=seed)
```

In [26]:
```
#Fit model to training data
model = XGBClassifier()
model.fit(X_train, y_train)
```

Out[26]:
```
XGBClassifier(base_score=0.5, booster='gbtree', callbacks=None,
              colsample_bylevel=1, colsample_bynode=1, colsample_bytree=1,
              early_stopping_rounds=None, enable_categorical=False,
              eval_metric=None, feature_types=None, gamma=0, gpu_id=-1,
              grow_policy='depthwise', importance_type=None,
              interaction_constraints='', learning_rate=0.300000012,
              max_bin=256, max_cat_threshold=64, max_cat_to_onehot=4,
              max_delta_step=0, max_depth=6, max_leaves=0, min_child_weight=1,
              missing=nan, monotone_constraints='()', n_estimators=100,
              n_jobs=0, num_parallel_tree=1, objective='multi:softprob',
              predictor='auto', ...)
```

In [28]:
```
#make predictions for test data
y_pred = model.predict(X_test)
predictions = [round(value) for value in y_pred]
```

In [29]:
```
#Evaluate predictions
accuracy = accuracy_score(y_test, predictions)
print("Accuracy: %.2f%%" % (accuracy * 100))
```

```
Accuracy: 64.88%
```

Gradient Boosted Machine

```
In [32]: X, y = make_classification(n_samples=500, n_features=20, n_informative=15, n_redundant=5, random_state=7)
         print(X.shape, y.shape)

         (500, 20) (500,)

In [33]: #Evaluate Gradient Boosting Algorithm
         from numpy import mean
         from numpy import std
         from sklearn.datasets import make_classification
         from sklearn.model_selection import cross_val_score
         from sklearn.model_selection import RepeatedStratifiedKFold
         from sklearn.ensemble import GradientBoostingClassifier

In [34]: #Define dataset
         X, y = make_classification(n_samples=500, n_features=20, n_informative=15, n_redundant=5, random_state=7)

In [35]: #define the model
         model = GradientBoostingClassifier()

In [37]: #define evaluation method
         cv = RepeatedStratifiedKFold(n_splits=10, n_repeats=3, random_state=1)

In [38]: #Evaluate the odel on the dataset
         n_scores = cross_val_score(model, X, y, scoring='accuracy', cv=cv, n_jobs=-1)

In [39]: #report performance
         print('Mean Accuracy: %.3f (%.3f)' % (mean(n_scores), std(n_scores)))

         Mean Accuracy: 0.883 (0.049)
```

Model Evaluation

Even though the XGBoost model only had an accuracy rating of approximately 65%, and the Gradient Boosted Machine had a mean accuracy of approximately 88%, I am going to recommend deploying XGBoost as the preferred algorithm in this case, with the use of Gradient Boosted Machine as a secondary evaluation of this model.

After further exploration of the data, and further business understanding, it is clear that as the pandemic wanes, metrics such as probable cases of COVID-19 do not serve as strong predictors of overall community health. Also, even with a fully vaccinated individual, they will still shed the virus once exposed to SARS-CoV-2, even though they will not have any symptoms of infection. As a result of these two considerations, the Gradient Boosted Machine is most likely overfit at this time.

Model deployment

The first step to begin deploying this model is to build a data pipeline that can pull the incoming data from the variety of sources that were utilized in this project, namely the California Surveillance of Wastewater Systems Network, the Public Health Alliance of Southern California, and the California Department of Public Health. Due to the fact that the vaccination database is only updated on a weekly basis, there is no need to pull the data more frequently than that.

There was some data manipulation that was performed manually for this project, but should be able to be automated within the pipeline, such as accumulating the daily probable cases into weekly bins.

Finally, this model should be carefully considered for data drift. As several of the data sources are non-linear, it is possible that this particular model is already on the border of what it is capable of. The next sprint should prioritize employing a time series analysis technique for this data, and reviewing mean values of the probable cases parameter week on week. If the pattern of reporting probable cases continues to decline, the model should be adjusted to remove that feature entirely.

Conclusion and Future Steps

In conclusion, there were some predictive results for the rate of SARS-CoV-2 in wastewater samples, but more research needs to be done in order to get the right data to make a stronger prediction. Part of the problem is that over time, the rate of virus within wastewater samples tended to increase continuously, while reporting of probable cases had more of a parabolic arch shape, and the vaccine equity measurement of community health generally had a parabolic shape as well. These features may have been strong predictors before reaching the turn of their curve, but over both 2021 and 2022, they weren't strong predictors for the viral load of wastewater samples.

Potential future steps that can be taken for the next sprint include modeling the quantity of SARS-CoV-2 within wastewater as a time series analysis instead of a classification system, especially within specific counties of California.

Another potential future step for a later sprint would be further segmenting out this data into accurate boundaries to show the communities that these wastewater treatment plants serve. Once the data is geographically segmented, measuring viral load in samples taken from these locations should provide a more correct reading on the overall health of these communities.

References

California Surveillance of Wastewater Systems (2022). *COVID-19 Wastewater Surveillance Data. California*. California Open Data Portal. https://data.ca.gov/dataset/covid-19-wastewater-surveillance-data-california/resource/16bb2698-c243-4b66-a6e8-4861ee66f8bf

Public Health Alliance of Southern California (2022). *Statewide COVID-19 Vaccines Administered By County*. California Open Data Portal. https://data.ca.gov/dataset/covid-19-vaccine-progress-dashboard-data/resource/eef88868-0cfc-4655-8a5a-3d1af1d23498

Public Health Alliance of Southern California (2022). *Statewide COVID-19 Vaccines Administered By Zip Code*. California Open Data Portal. https://data.ca.gov/dataset/covid-19-vaccine-progress-dashboard-data-by-zip-code

California Department of Public Health (2022). *COVID-19 Probable Cases*. California Open Data Portal. https://data.ca.gov/dataset/covid-19-probable-cases/resource/a74f0542-b337-4bf0-9239-a7876f6e1c11

California Department of Technology (2022). *County and Zip Code References*. California Open Data Portal. https://data.ca.gov/dataset/county-and-zip-code-references