Final Project Report for Stats 170B, Spring 2022

Project Title: COVID Predictors of Mortality and Positivity

Student Names:
Christie Yang, 20665413, christy9@uci.edu
Micah Fadrigo, 11471092, mfadrigo@uci.edu
Roni Asatourian, 76142466, rasatour@uci.edu

Github: https://github.com/mfadrigo/oc-positivity-plus

1. Introduction and Problem Statement

We conducted a detailed spatiotemporal epidemiologic analysis that examines the social, economic, and demographic associations with COVID-19 in Orange County (OC) from January 22, 2020 - January 25, 2021. We are interested in the heterogeneity in risks for test positivity and death within Orange County. We used a combination of datasets that include: individual records of PCR test results, individual records of deaths, and ZIP code-level sociodemographic data.

COVID-19 has appeared in different ways across social, economic, and demographic groups, and this is especially true for the diverse and high-density population of OC. Certain communities are more privileged and face less challenges in education, household income, access to healthcare, and life expectancy - all of which are social, economic, and demographic factors - which can ultimately affect the manifestation of COVID-19. Identifying predictors behind test positivity and mortality disparities can give us insight into how we can better address the needs of diverse populations and mitigate risk. At the individual level, several factors (e.g. age, race, gender, and ZIP codes with high educational and health insurance attainment) strongly affected death and test positivity.

2. Related Work:

OPTIMIZING FOR ESTIMATION APPROACH

Since we are trying to examine heterogeneity in risks for COVID-19 death and test positivity, our main objective is estimation. Estimation is a matter of finding the most appropriate parameter(s) that best describe the multivariate distribution of historical data - in which we are trying to estimate the odds of COVID-19 death and test positivity as a function of geographic, demographic, and economic risk factors. This is different from prediction, which uses the given data to compute the random value of the unseen data. In epidemiologic research, logistic regression is often used to study the relationships between a disease in two modalities and risk factors (qualitative or quantitative). The study that our project is based on utilizes logistic regression, tests several specifications of the models, and chooses the best models based on Bayesian Information Criterion (BIC) Score.

Reference: Parker, D. M., Bruckner, T., Vieira, V. M., Medina, C., Minin, V. N., Felgner, P. L....Boden-Albala, B. (2021). Predictors of Test Positivity, Mortality, and Seropositivity during the Early Coronavirus Disease Epidemic, Orange County, California, USA. Emerging Infectious Diseases, 27(10), 2604-2618. https://doi.org/10.3201/eid2710.210103.

OPTIMIZING FOR PREDICTION APPROACH

Our experiments were made by differing the undersampling coefficient ranging from (0.013 - 1) which is equivalent to totally undersampling and no oversampling totally oversampling and no undersampling. Clearly after undersampling the majority class, the minority class is oversampled upto the new size of majority class therefore low values for the coefficient are equivalent to low training set size.

For each coeff value we used three classifiers: 1- XGBOOST 2- Logistic Regression 3- Random Forest. (Since the problem is actually a binary classification problem we used classifiers to solve it). For each experiment we recorded the four common performance measures along with the time elapsed for training and predicting and confusion matrix.

References:

1) Mukherjee, Mimi, and Matloob Khushi. "SMOTE-ENC: A Novel SMOTE-Based Method to Generate Synthetic Data for Nominal and Continuous Features." Applied System Innovation, vol. 4, no. 1, Mar. 2021, p. 18, https://doi.org/10.3390/asi4010018.
2) Chen, Tianqi, and Carlos Guestrin. "XGBoost." Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Aug. 2016, https://doi.org/10.1145/2939672.2939785.
3) https://medium.com/analytics-vidhya/smote-nc-in-ml-categorization-models-fo-imbalanced-datasets-8adbdcf08c25

3. Data Sets

We utilized two separate datasets reflecting the mortality and positivity of COVID-19 data in Orange County, California. Both datasets came from the Orange County Health Care Agency (OCHCA) and include test results and cases from 01/22/2020 - 01/25/2021. We selected these dates to include both the first and second wave of the pandemic in hopes of capturing the different variables that factored into the spread of COVID-19. In order to consider other non-individual level variables, zip code data (Fig. 3) that included features like: percentage of adults with a bachelor's degree or health insurance, median household income, and population density, were joined with the test positivity and mortality dataset by zip code. This data came from the 2018 American Community Survey. Additionally, public hospital data was joined with both the test positivity and mortality dataset based on the specific date of the record. As seen in figure 4, this data was sourced from the California State Dashboard[1] and contains statewide hospitalization counts on the number of available ICU beds. This data was useful in the mortality dataset as a variable to consider hospital bed shortages during certain times of the pandemic. It was used as a method to incorporate a time factor when predicting mortality.

Figure 1 illustrates the dataset on test positivity which contains all of the PCR test data for patients in Orange County. This dataset included patient information like age, sex, race, ethnicity, and zip code. Additionally each patient had a different description of the PCR test described by: **test_result** and **covid_positive**, for the results of the test; **posted_date,** reflecting the date patients took the test; and **time_days**, is the number of days since the start date in January. Originally, the dataset on test positivity had 2,605,761 cases with 12 different features. After cleaning and filtering the data to fit our time frame and include additional zip code and hospital data, the test positivity had 1,865,151 entries but with 33 total features. It is also important to note that for those patients that had multiple positive PCR tests, only the first positive test was recorded in the data.

The next dataset focuses on mortality cases among those who have already tested positive for COVID-19, as shown in figure 2. This dataset contains 211,251 death cases in the twelve month period. However, out of the 211,251 cases only 2,671 cases were attributed to COVID-19. The test mortality dataset had 33 total features including reported city, death date, gender, sex, and zip code information. Compared to our positivity data, this dataset contains a column named, **death_due_to_covid**, which is the binary response on whether or not the patient died due to COVID-19.

[1] California Hospital Dataset URL. https://data.ca.gov/dataset/covid-19-hospital-data1

	id	posted_date	time_days	adj_time_days	test_result	age	sex	race	ethnicity	zip	...	population	name
0	1	2020-07-02	162	-1.471365	positive	29	female	NaN	Unknown	90620	...	45.113	Buena Park
1	100	2020-08-09	200	-0.967804	negative	22	female	NaN	Unknown	92692	...	47.222	Mission Viejo
2	1000	2020-06-30	160	-1.497868	positive	50	male	NaN	Unknown	92805	...	70.401	Anaheim
3	100000	2020-05-01	100	-2.292965	negative	53	male	NaN	Unknown	92677	...	63.297	Laguna Niguel
4	100001	2020-05-01	100	-2.292965	negative	23	male	NaN	Unknown	92708	...	56.004	Fountain Valley

Figure 1. A sample of the cleaned Positivity data

	id	zip	posted_date	age	gender	race	ethnicity	death_date	death_due_to_covid	age_group	...	population	name
0	1	90620	2020-07-02	29	female	unknown	Unknown	2020-01-22	0	25-29	...	45.113	Buena Park
1	100	92692	2020-09-15	22	female	unknown	Unknown	2020-01-22	0	20-24	...	47.222	Mission Viejo
2	1000	92805	2020-06-30	50	male	unknown	Unknown	2020-01-22	0	50-59	...	70.401	Anaheim
3	1000079	92802	2020-12-23	12	female	hispanic	Hispanic or Latino	2020-01-22	0	10-14	...	42.709	Anaheim
4	1000087	92802	2020-12-23	13	female	hispanic	Hispanic or Latino	2020-01-22	0	10-14	...	42.709	Anaheim

Figure 2. A sample of the cleaned Mortality data

	zip	percent_insured	percent_bachelors	med_income	population	name	area_km	pop_density	adj_pop_density	adj_med_income	adj_perc_bach
1	90620	83.4	30.1	8.8082	45.113	Buena Park	17.051480	2.645694	0.255812	-0.196087	-0.814268
2	90621	74.2	29.8	6.0799	35.153	Buena Park	10.370346	3.389762	0.812408	-1.089492	-0.831189
3	90623	80.5	47.6	9.9576	15.554	La Palma	4.794937	3.243838	0.703251	0.180294	0.172801
4	90630	87.4	42.1	9.2073	47.993	Cypress	16.014339	2.996877	0.518512	-0.065398	-0.137420
5	90631	81.0	30.1	7.9500	67.619	La Habra	35.130463	1.924797	-0.283452	-0.477112	-0.814268

Figure 3. A sample of the 2018 American Community Survey zip code data

avail_icu_beds <dbl>	perc_avail_beds <dbl>	covid_icu_beds <dbl>	adj_perc_avail_beds <dbl>	adj_avail_icu_beds <dbl>	adj_covid_icu_beds <dbl>
326	89.30385	208	-0.3941672	1.3844667	0.8102243
255	90.74765	162	-0.2112683	0.5534664	0.4113276
386	90.51032	212	-0.2413326	2.0867204	0.8449109
374	93.24868	85	0.1055608	1.9462697	-0.2563908
374	93.24868	85	0.1055608	1.9462697	-0.2563908
70	66.86192	523	-3.2370983	-1.6118161	3.5417995

Figure 4. A sample of the California COVID-19 Hospital data

3.1 Exploratory Data Analysis Visualizations

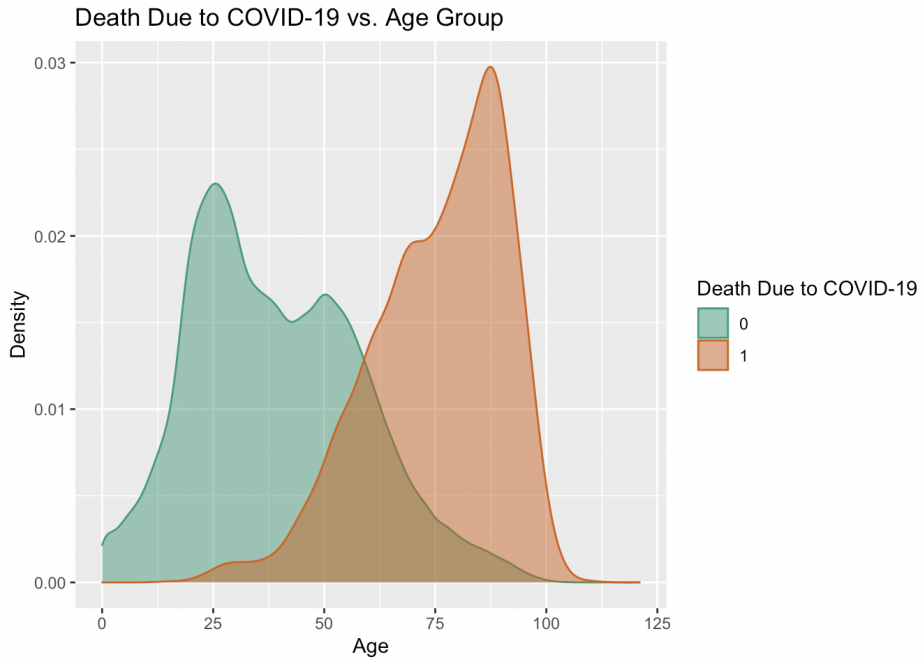


Figure 5. Age groups that range from zero to thirty-nine years old have little to no deaths due to COVID-19, while ages ranging from forty years old to eighty years old and above have a higher risk of death due to COVID-19. Since the distributions associated with death due to COVID-19 are oppositely skewed, this tells us that there is heterogeneity in risk of death within the age variable.

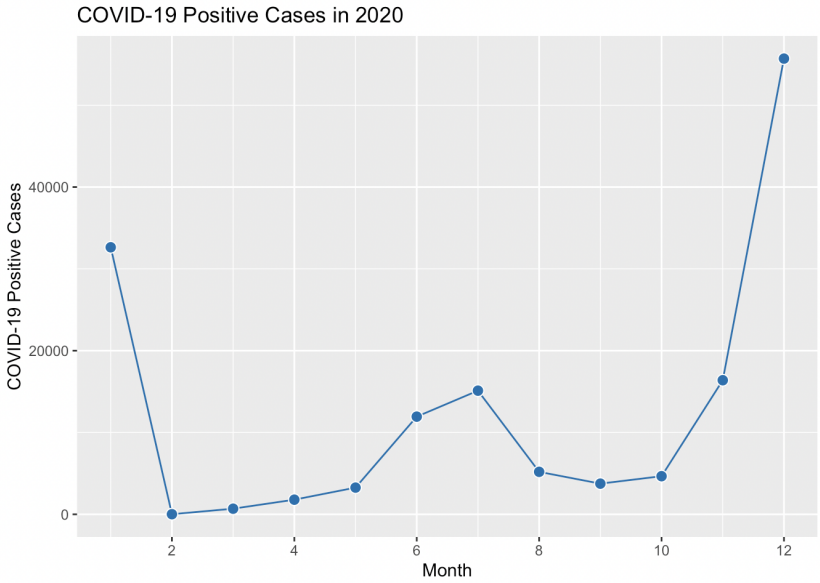


Figure 6.
In order to visualize the trends of positivity throughout the year, we created a scatter plot illustrating the number of positive tests for each month of the year. In general, the dataset captures the first and second wave of COVID-19. The first wave occurs from January to March and the second is captured at the end of the year from August to December.

4. Overall Technical Approach

4.1 Data Preprocessing

Our data preprocessing code is from https://github.com/CatalinaMedina/oc-positivity-plus which stems from a research study which is called “Predictors of Test Positivity, Mortality, and Seropositivity during the Early Coronavirus Disease Epidemic, Orange County, California, USA”. Our study is similar, except that it only considers predictors of test positivity and mortality and covers a larger time period of the pandemic. We did not have access to the data associated with the house-crowding ZIP code-level

predictor. In order to reflect these changes, we removed the corresponding parts of the code and changed the start and end date.
The data preprocessing code can be broken down into 4 main components:

1. Combine ZIP Code-level Datasets
    a. Join **zip-population** and **zip-area** (by ZIP code) and calculate population density for every ZIP code given population size and areaKM.
    b. Join the **dataset produced in part 1A** and **income-by-zip** (by ZIP code) to add the covariate that captures **median household income**. Join this newly created dataset with **education-by-zip** to add the covariate that captures the percentage of the population that has a Bachelor's degree. Repeat with **insurance-by-zip** to add the covariate that captures the percentage of the population that has health insurance.
2. Clean and Merge PCR Test Data
    a. Collapsed levels of **sex** and **race** and added new covariates such as **time_days** and its scaled version.
    b. Removed observations that had inconsistencies for **age**, **sex**, and **race** and observations that tested in the same day. Kept all observations up to and including their first positive.
    c. Merged with ZIP code-level data produced in part 1.
3. Clean and Merge Case Data (similar to step 2)
4. Save datasets produced in part 2 and 3 into **usable_tests** and **usable_cases**, respectively.

4.2 Logistic Regression & Data Analysis

Diagnostics

The first assumption is satisfied since we have a binary outcome type (death due to COVID-19). The second assumption was satisfied in which the smoothed scatter plots showed that most variables were quite linearly associated with the mortality outcome in logit scale. The third assumption was assessed by checking the absolute standardized residuals and Cook's distance. This assumption was not satisfied, however, we decided to keep the influential values in our model. Multicollinearity was assessed by checking GVIF values and no multicollinearity was detected since all values were below 5. Given that space is a confounder, independence was not satisfied. We had a sufficiently large sample size of 211, 251 total observations.

Factors Associated with Dying From COVID-19

We first did a fixed effects model which does not include a random intercept for ZIP code and a predictor for the number of ICU beds occupied by COVID-19 patients. It contains: age, gender, race, education, health insurance, population density, time, and median household income. All predictors except population density were statistically significant at the 5% significance level. Although time was significant, we anticipated that the model would improve significantly once we incorporated the hospital bed predictor. Since the hospital bed data is joined with the mortality dataset based on time (not ZIP code), time is necessary in the model when the hospital bed predictor is in the model. Otherwise, the time of an individual's first positive test is not relevant by itself when considering risk factors of dying from COVID-19. We tried 3 variations of models, where each model differed by the variable that was used for the hospital bed predictor. Model 2 used **`adj_covid_icu_beds`** (scaled # occupied icu beds by covid patients), model 3A used **`adj_perc_avail_beds`** (percentage of available icu beds), and model 3B used **`adj_avail_icu_beds`** (number of available icu beds). Since the total number of hospital beds in Orange County was most likely not constant over time, we anticipated that **`adj_covid_icu_beds`** in model 2 would not be as precise. We anticipated model 3A and model 3B to be better fits than model 2 because **`adj_perc_avail_beds`** and **`adj_avail_icu_beds`** both account for the total number of beds changing. However, model 2 had the lowest BIC score and was the best model.

For model 4, we wanted to incorporate a predictor that would signify whether or not it would be more dangerous (higher risk of death) to be sick at different points in time. We decided to add an interaction term between time and the hospital bed predictor because the effect of the number of occupied

ICU beds in OC on the odds of dying from COVID-19 could potentially depend on time. For example, an increase by 100 in the number of occupied ICU beds in OC may result in a higher or lower risk of dying from COVID-19 for an individual depending on when they first tested positive. From the model comparison, the interaction term improved the model and was also significant.

Space is a confounder in which people within close spatial proximity will have similar risk of dying from COVID-19. Each individual belongs to a certain zip code, and individuals within each zip code are correlated. When we have clustered data with correlations within the clusters, we can control for this by adding a random intercept for cluster ID, or zip code in this case. From the model comparison, adding a random intercept for zip code improved the model.

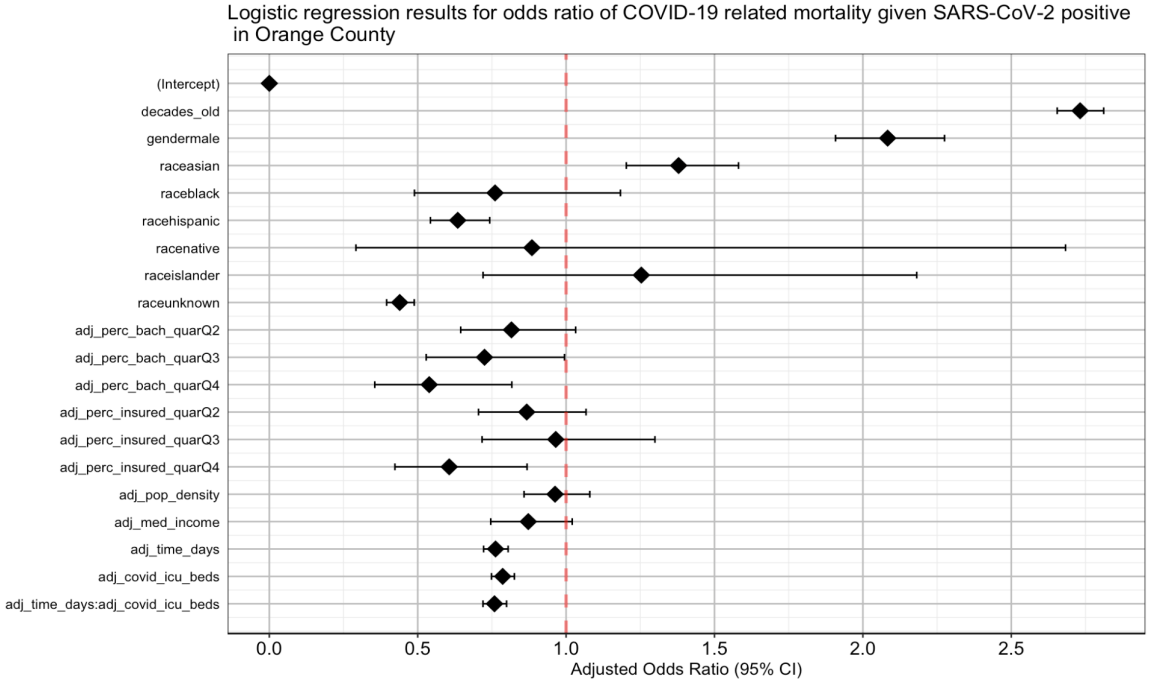


Figure 7. Forest plot demonstrating the odds ratio and confidence intervals for model 5

Age and gender showed nearly identical results compared to the published study. Similar to the published study, Asians had a 38% higher risk of dying from COVID-19 compared to non-hispanic whites (not as drastic as the published study which reported 54% higher risk). Unlike the published study, we found that hispanics and those with an unknown race had a 37% and 56% (respectively) lower odds of dying from COVID-19 compared to non-hispanic whites. Similar to the published study, only the highest levels of education and health insurance were predictive of mortality outcomes. The odds of dying from COVID-19 for an individual in a ZIP code in the third or fourth quartile of college degree was 27% and 46% lower (respectively) than that of a similar individual in a ZIP code in the first quartile of college degree. The odds of dying from COVID-19 for an individual in a ZIP code in the fourth quartile was 39% lower than that of a similar individual in a ZIP code in the first quartile of health insurance. Similar to the published study, the number of COVID-19 patients in hospital beds was a significant predictor of death. Unlike the published study, the interaction term between time and the hospital bed predictor as well as the random intercept for ZIP code was significant.

Factors Associated with Testing Positive of COVID-19

Unlike mortality, the positivity model requires a GAM model rather than the GLM model that was used. This is due to a couple of things. First, GAM's are semi-parametric extensions of the GLM model with the assumption that any additional functions added are smoothed. This allows for highly

non-linear relationships that we expected due to the fact that different time periods of the pandemic would affect the positivity rates greatly. By fitting the data independently using thin plate regression splines, it improved our model greatly. This was also based on the assumptions of independent observations and no extreme outliers. After comparing five different models with different variations of interaction terms and smoothing functions, we found that model 5 had the lowest BIC score and best results for explaining test positivity for people in Orange County. The final factors associated with model 5 can be found in **Appendix A**.With the addition of all the regular features like age, race, and gender, there also was an interaction term between time and median income. As well as the smoothing spline function on time and also a random intercept for the zip code feature.

Based on model 5 and as seen in figure 8, we found a couple interesting results for testing positive. First, individuals in ages 10-14 and 15-19 had almost 1.8 times the odds of testing positive compared to individuals in ages 0-4. Second, males had 1.13 times the odds of females. Third, Hispanic individuals had 2.578 times the odds of testing positive compared to individuals identifying as White. This is interesting because while our results show they are 2.578 times more likely to test positive, they actually have a 37% lower chance of actually dying from COVID-19 compared to White individuals.

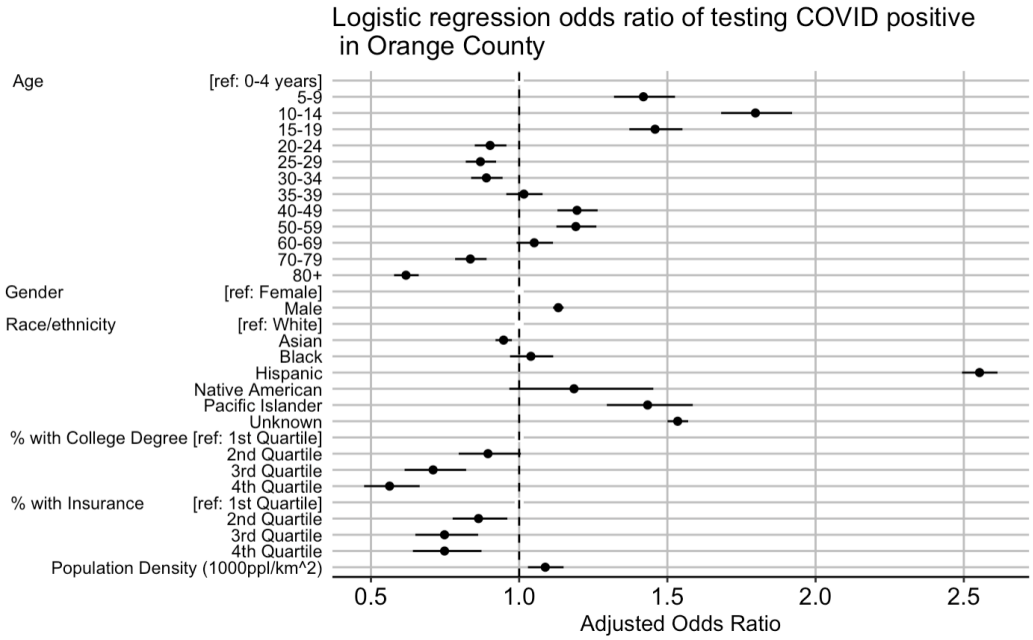


Figure 8. Forest plot demonstrating the odds ratio and confidence intervals for model 5.

5. Software

Below is the self-written code written for this project:

Name	Description
EDA.Rmd	An exploratory data analysis markdown for visualizing the data and analyzing assumptions
mortality_model.Rmd	A data analysis notebook for model comparison and fitting of the mortality dataset. Also includes code for visualizations of results

positivity_model.Rmd	A data analysis notebook for model comparison and fitting of the positivity dataset. Also includes code for visualizations of results
MLprediction.jpynb	A jupyter notebook for ML techniques for modeling mortality and dealing with oversampling/undersampling

This is the publicly-available code and software used for this project:

Name	Description
clean-covid-data.R	The data preprocessing code provided through a third party that cleaned and merged the datasets provided. Modified by us to accommodate this projects objectives
scikit-learn	A Python library used for classification and regression for the models built
XGBoost	An open source software library optimized for distributed gradient boosting
randomForest	An ensemble learning library for classification and regression using random inputs
RStudio	An open-source software environment for R that we used for version control with Git and utilized for the majority of our modeling.
Git/GitHub	A source code management system used to coordinate work among our team.

6. Experiments and Evaluation

OPTIMIZING FOR ESTIMATION APPROACH

BIC Score is a method for assessing model fit penalized for the number of estimated parameters. This method is suitable for understanding the main drivers behind our outcome variable.

Positivity Model Comparison using BIC

	df	BIC
Model 1	30.0000	424240.7
Model 2	31.0000	421216.9
Model 3	33.0000	420797.9
Model 4	103.8347	415296.1
Model 5	115.1509	414772.7

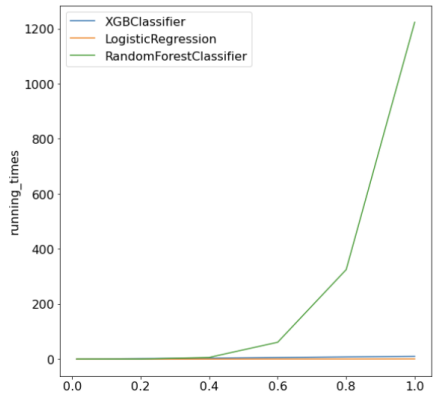
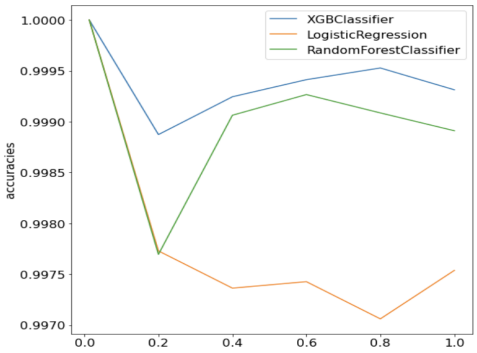
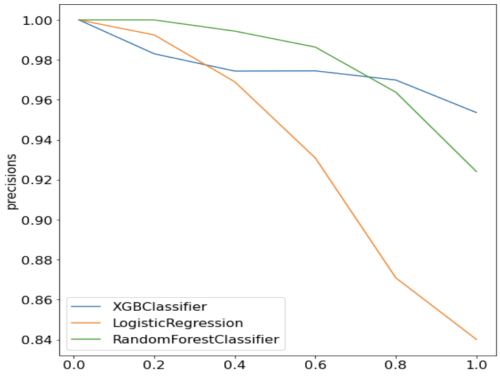
Mortality Model Comparison using BIC

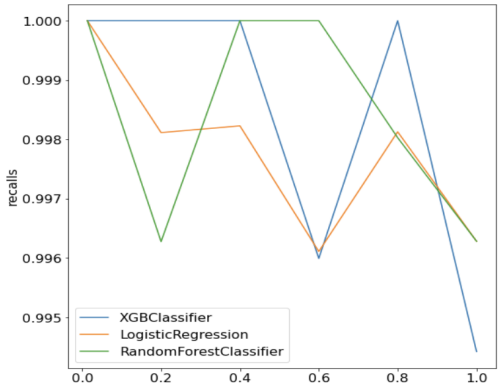
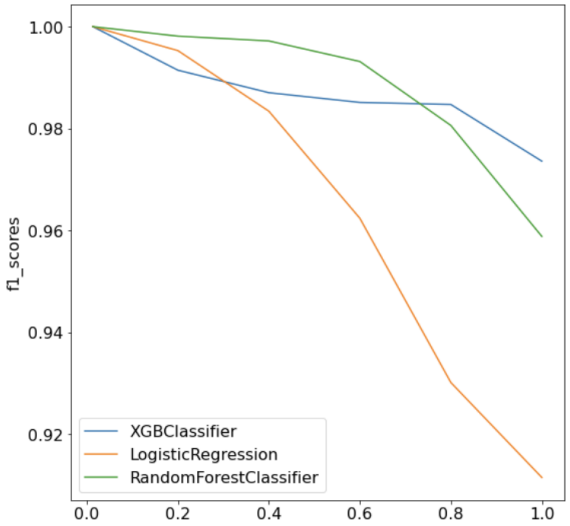
	df	BIC
model_no_hospital	18	18237.10
model_covid_beds	19	17791.32
model_avail_beds	19	17812.38
model_avail_beds_not_percentage	19	18027.42
model_inter_time_bed	20	17693.74
model_random_zip	21	17684.43

OPTIMIZING FOR PREDICTION APPROACH

Accuracy: The percent of data which are correctly classified. (poor measure for imbalanced data)
Precision: The percent of correct predictions within positives.
Recall: The percent of True class which are correctly predicted.
F1_score: The harmonic average of precision and recall.
Exploring performance measures: Common evaluation metrics for classification problems are accuracy, precision, recall, and F1 score. Precision and recall are not good independent performance measures

because we can design dummy classifiers to simply boost all of them (Not both at the same time). Therefore, we usually use the F1 score which is a harmonic average of them and cannot be hacked! Among accuracy and F1 score, accuracy is very biased in unbalanced scenarios and can be hacked especially in binary classification problems. So the very robust performance measure under unbalanced scenarios is the F1 score as stated in the literature.

Timing: (Changes on how we are accounting for running time) The diagram shows that random forest takes an extremely long time (exponentially with the size of the training set) in contrast to the two other classifiers (which appeared linear behavior with increasing training size). The random forest is practically not usable for large training set sizes.	
Accuracy: As can be seen in the diagram the highest accuracy is for the case of no over sampling but the differences are so small (the minimum accuracy is 99.7%) so no significant difference can be seen within the classifiers as well as related to training set size.	
Precision: Precision always decreases with the increasing training size, the lowest decrement was for XGBOOST and the highest for logistic regression which is a rather considerable decrement compared to the other two classifiers which is as expected for baseline classifier compared to advanced classifiers.	

Recall: The values of recall are very high and the variations in them are too small compared to the values (the minimum recall is about 99.5%). So no major conclusion can be made based on the recall values except that the recall remains constant among the classifiers and train set size.	
F1_score: F1_score always decreases with the increasing training size, the lowest decrement was for XGBOOST and the highest for logistic regression. This is a rather considerable decrease compared to the other two classifiers, which is as expected. As F1_score is the best performance measure (among the above measurements) for the imbalanced data, we can conclude that the best performance was earned for **XGBOOST,** slightly better than random forest, with excellent performance equal to 97.5% for totally oversampling (worst) case. Linear regression earned poor performance for rather high training sets due to its model simplicity. For small training set size (majorly undersampling) linear regression earned competitive performance compared to XGBOOST.	

Error analysis using confusion matrix: Since misclassifications are clearly viewed in the confusion matrix, we plotted confusion matrices against resampling coefficient for 3 classifiers. The plots can be found in **Appendix A.** The number of errors clearly increases with increasing training size, the lowest increment is for XGBOOST (29 false predictions), the second place is for random forest (46 wrong predictions) and the last place for logistic regression (102 wrong predictions). We can conclude that XGBOOST has the lowest error rate.

7. Notebook Description

The logistic regression models for positivity and mortality were both done in their own RMarkdown notebooks (`analysis/mortality_model.Rmd`, `analysis/positivity_model.Rmd`). Besides the models themselves, each of the notebooks contain information about model assumptions, model comparison, and model interpretations. The basic workflow was as follows: download the individual-level, zip-code level, and hospital datasets and save them locally, run the `analysis/clean-covid-data.R` file to obtain and save the cleaned datasets, and run either the `mortality_model.Rmd` or `positivity_model.Rmd` to fit the models and visualize the results (adjusted odds ratios). The machine learning models for mortality were done in a jupyter notebook and can be run as is.

8. Members Participation

In general, throughout the project, Micah Fadrigo and Christie Yang did mostly paired programming and split up the work between positivity and mortality modeling that reflected the original research study. Roni Asatourian strictly focused on a modeling approach that optimized for prediction rather than estimation.

Team Member	Workload Ratio	Task and Description
Christie Yang Micah Fadrigo	50% 50%	Building the logistic regression models for mortality and positivity
Micah Fadrigo		Cleaning code modifications done to merge our datasets together
Christie Yang		Positivity EDA and model result visualizations
Micah Fadrigo		Mortality EDA and model results visualizations
Roni Asatourian		ML models for mortality
Christie Yang Micah Fadrigo Roni Asatourian	45% 45% 10%	Writing the final report
Christie Yang Micah Fadrigo	50% 50%	Managing Github repository

9. Discussion and Conclusion

1) The main method that we learned a lot about during this project was the general additive model (GAM). This was a fairly new method to work with and was tricky because of the large number of options that were available for the smoothing spline functions. While the generalized linear model assumes a linear parametric relationship, GAM has a lot of strengths that fit really well with the goal of our positivity model. To account for the effect of the time of the pandemic, GAM fits the data in each knot independently and applies a smoothing function. This places different weights on different times of the pandemic, which is a big strength. One limitation of GAM is dealing with the over-fitting aspect. To try to make up for overfitting, we used simpler models and used BIC for model selection.
2) Something that was more difficult than expected was dealing with the unbalanced mortality data, since we did not expect the deaths due to COVID-19 to be that low. The different sampling methods were more difficult to implement than expected.
3) We were all familiar with logistic regression and the tools we used were not anything out of the ordinary. We anticipated ways to improve our model (for example, by changing the hospital bed predictor to available beds in OC), however these ways ended up not improving our model which did not intuitively make sense. Perhaps this is a problem that concerns how the data was collected and/or processed.
4) Even political context is relevant to this problem as many of the mask mandate decisions are influenced by political decisions such as garnering votes, behavior of the public in OC, availability and access to vaccines, hospitals, and other resources. This privilege in terms of availability of vaccines can contribute to lower hospitalizations / death. Incorporating political predictors in the future (once the appropriate data is found) could potentially improve the models for both test positivity and mortality. Accounting for vaccine accessibility would also mean expanding the time range that the data captures, to include the time period where the vaccines were distributed.

10. Appendix A

A.1

Model 5 Positivity model log odds equation:

Model 5:
Generalized additive model

$$\begin{aligned}\log(O_i) = \beta_0 &+ \overrightarrow{\beta}_{\text{Age Group}}\overrightarrow{\text{Age Group}_i} + \beta_{\text{Gender}}\text{Gender}_i + \overrightarrow{\beta}_{\text{Race}}\overrightarrow{\text{Race}_i} \\ &+ \overrightarrow{\beta}_{\text{College}}\overrightarrow{\%\text{ with College Degree Quartile}_i} \\ &+ \overrightarrow{\beta}_{\text{Insurance}}\overrightarrow{\%\text{ with Medical Insurance Quartile}_i} \\ &+ \beta_{\text{Population Density}}\text{Population Density}_i + \beta_{\text{Median Income}}\text{Median Income}_i \\ &+ \beta_{\text{Time}}f(\text{Time}_i) + \beta_{\text{Interaction}}\text{Median Income}_i \times f(\text{Time}_i),\end{aligned}$$

with a random intercept for zip code.

A.2

Confusion Matrix on Mortality modeling: