The Pandemic's Path: Predicting Future Covid-19 Cases in Midwestern States Using Machine Learning Methods

Brenda Li (brendal)

Sophia Mlawer (smlawer)

Michelle Orden (morden)

Github: https://github.com/brendali121/CAPP-30254-Covid-Variants/tree/main/notebooks

1. Executive Summary

The coronavirus pandemic has taken the lives of over 3.5 million people worldwide, spread through virtually every country, and changed many aspects of our world[1]. A major challenge that the pandemic has posed is the lack of knowledge of what comes next. Many policy decisions such as mask mandates, stay at home orders, and virtual schooling rely on the knowledge of how many Covid-19 cases are to be expected in the coming week. The prediction of future Covid-19 cases is a complex task, as many factors contribute to the spread of the virus. In this report, we attempt to predict Covid-19 cases for a given week in the Midwest states of Illinois, Ohio, and Missouri using various machine learning models. To do so, we collected data on the number of Covid-19 cases, the number of Covid-19-related deaths, vaccine administration counts, public mask mandates, hospital utilization, mobility, and demographics information for each county in these states, dating from July 2020 to May 2021. With the successful prediction of future Covid-19 cases, policy makers and state/county leaders can better prepare their respective counties for what's to come.

2. Background and Overview of your Solution

Now over a year into this global pandemic, we have seen the destructive effects of Covid-19 on countries all across the globe. Covid-19 has taken over 3.5 million lives, overwhelmed hospital networks, disrupted schooling and education, wreaked havoc on economies, and affected our lives in all imaginable ways. At the beginning of the pandemic, part of the challenge of controlling the spread of the virus was not knowing where the next Covid-19 outbreak would occur. Unlucky states like New York[2] faced harrowing outbreaks that debilitated their hospital systems while other states worried about whether they would be next. A year into the pandemic, the U.S now has a better handle on containing outbreaks. Through the combination of restrictions, public health mandates, and a massive vaccination campaign, Covid-19 cases are currently declining in both number and severity in the U.S. However, health experts caution that the pandemic is still far from over[34], especially with the rise of dangerous variants that have devastated other countries like India[5]. Experts anticipate that the U.S. will continue to see seasonal peaks and surges throughout at least the next few years[6]. Therefore, predicting future Covid-19 cases remains an important problem and it's one we can leverage machine learning methods for.

Our machine learning model predicts Covid-19 cases a week into the future at the county level using regression-based machine learning techniques and a variety of different information for each county. This data includes information on current Covid-19 cases, hospital utilization, mobility, and general demographics. We were also able to explore using data about vaccinations and public mask mandates, though they were not incorporated into our final model due to their lack of predictive strength (see Section 5).

[1] Jordan A et. al. Coronavirus World Map: Tracking the Global Outbreak. The New York Times website. June 2, 2021. Accessed June 2, 2021.

[2] Goodman JD, Rashbaum WK. N.Y.C. Death Toll Soars Past 10,000 in Revised Virus Count - The New York Times. *The New York Times*. https://www.nytimes.com/2020/04/14/nyregion/new-york-coronavirus-deaths.html. Published April 14, 2020. Accessed June 2, 2021.

[3] Branswell H. How the Covid pandemic ends: Scientists look to the past to see the future. STAT. Published May 19, 2021. Accessed June 2, 2021. https://www.statnews.com/2021/05/19/how-the-covid-pandemic-ends-scientists-look-to-the-past-to-see-the-future/

[4] Silverstein J. When will COVID-19 end? A year into the pandemic, public health experts say: Never. CBS News. Accessed June 2, 2021. https://www.cbsnews.com/news/covid-19-endemic-disease-never-going-away/

[5] Gettleman J, Yasir S, Kumar H, Raj S, Loke A. As Covid-19 Devastates India, Deaths Go Undercounted. *The New York Times*. https://www.nytimes.com/2021/04/24/world/asia/india-coronavirus-deaths.html. Published April 24, 2021. Accessed June 2, 2021.

[6] Achenbach J. Is it now reasonable to discuss the end of the pandemic? Yes, but with caveats. *Washington Post*. https://www.washingtonpost.com/health/when-will-the-pandemic-end/2021/05/13/1fcee324-b116-11eb-ab43-bebddc5a0f65_story.html. Accessed June 2, 2021.

Using the predictions from this model, county level health officials can forecast whether Covid-19 is expected to flare up in their county and decide whether to reinstate social distancing measures and pandemic-related restrictions. Since Covid-19 restrictions are tentatively being lifted across the country right now, these predictions will be important in helping county level health officials prepare for the return of any Covid-19 surges and allow them to proactively reinstate restrictions as needed.

In order to train our models on the most complete data possible, we restricted our model to focus on the three Midwest states of Illinois, Missouri, and Ohio. Each of these states have had varying experiences with the pandemic. They've each experienced numerous surges in Covid-19 cases, deaths, and hospitalizations (see Figures 1-3 in appendix) and taken varying approaches to controlling the virus. For instance, Illinois and Ohio implemented a public mask mandate early on in the pandemic while Missouri never instituted such a mandate. Ohio captured headlines in early May when it was the first state to announce a lottery incentive for vaccinations.

3. Data

Each row in our final dataset represents a single county on a specific date. Our predictors include information about Covid-19 cases and deaths reported for that day and previously, demographics for that county, mobility, covid-related hospitalization rates in the past week, and vaccine administration rates. Our outcome variable is the number of new Covid-19 cases 7 days in the future. Our final dataset includes all counties from Illinois, Missouri, and Ohio from 7-31-2020 through 5-24-2021. We chose to limit our data to this date range because this was the date range with the most complete data across all our data sources. Figure 4 in the appendix displays summary information about each of our data sources.

Originally, we were also interested in incorporating information about covid variants trends. However, upon exploration, we found the data to be too unreliable and inconsistent across states to be used for our analyses.

Below, we describe the details of the data collected from each source:

The New York Times[7]:

We obtained data about cumulative Covid-19 cases and deaths by date and county from the NYT. Using this data, we constructed variables for the number of new Covid-19 cases and new Covid-19 deaths reported each day by each county, 7 day averages, and lagged versions of each variable. The data spans 1/21/2021 (when the first Covid-19 case was reported in the US) to 5/27/2021 (the most recent date we pulled data).

The New York Times Covid-19 data had a few data quality issues where states would amend their previously reported numbers, which would generate a huge spike in new cases for that day or return a negative new cases value. In these cases, since this was an administrative data issue, we chose to smooth out these spikes through interpolation of the previous and next day.

[7] nytimes/covid-19-data. Published June 2, 2021. Accessed June 2, 2021. https://github.com/nytimes/covid-19-data

Centers for Disease Control and Prevention[8]:

We found data about publicly mandated mask requirements from the Centers for Disease Control and Prevention. This data reported whether a state-level mask mandate was enforced. This data spans 4/10/2020 through 3/22/2021 so all data after that date was manually coded.

Department of Health and Human Services[9]:

From the Department of Health and Human Services, we collected data on Covid-19-related hospital utilization by hospital. The dataset included every CMS-approved hospital and the data was reported on the week level. The variables included the total number of patients hospitalized or admitted in the past week with laboratory-confirmed Covid-19, broken out by adults versus pediatric or by age range.

Illinois Department of Public Health[10]

Counts of vaccine administrations by county for Illinois counties were taken from the Illinois Department of Public Health. The Illinois Department of Public Health cites their data source as I-CARE (Illinois Comprehensive Automated Immunization Registry Exchange), which is a web based immunization record-sharing application developed by the Illinois Department of Public Health (IDPH).

Ohio Department of Health[11]

Counts of vaccine administrations by county for Ohio were taken from the Ohio Department of Health. We used the attribute “vaccination completed”, which indicates “individuals within that group have received all recommended Covid-19 vaccine doses and are considered fully immunized.”

Missouri Department of Health & Senior Services[12]

Counts of vaccine administrations by county for Missouri were taken from the Missouri Department of Health & Senior Services.

Census Bureau[13]

Using the 5 year 2019 American Community Survey, counts and percentages of different demographics were collected at the FIPS level. This included variables on total population, gender, age, income, race, and poverty level. These variables were then cleaned and processed.

[8] U.S. State and Territorial Public Mask Mandates From April 10, 2020 through March 22, 2021 by County by Day. Centers for Disease Control and Prevention. Accessed June 2, 2021. https://data.cdc.gov/Policy-Surveillance/U-S-State-and-Territorial-Public-Mask-Mandates-Fro/62d6-pm5i

[9] COVID-19 Reported Patient Impact and Hospital Capacity by Facility. HealthData.gov. Accessed June 2, 2021. https://healthdata.gov/Hospital/COVID-19-Reported-Patient-Impact-and-Hospital-Capa/anag-cw7u

[10] COVID-19 Vaccine Administration Data | IDPH. Illinois Department of Health. Accessed June 2, 2021. https://www.dph.illinois.gov/content/covid-19-vaccine-administration-data

[11] Vaccine Administration Metrics Dashboard. Ohio Department of Health. Accessed June 2, 2021. https://public.tableau.com/views/VaccineAdministrationMetricsDashboard/PublicCountyDash

[12] Vaccine Metrics | COVID-19 Outbreak | Health & Senior Services. Missouri Department of Health and Senior Services. Accessed June 2, 2021. https://health.mo.gov/living/healthcondiseases/communicable/novel-coronavirus/data/data-download-vaccine.php

[13] Average Household Size and Population Density - County. U.S. Census Bureau COVID-19 Site. Accessed June 2, 2021. https://covid19.census.gov/datasets/21843f238cbb46b08615fc53e19e0daf_1/data?geometry=-126.247,28.795,126.878,67.148

Google Mobility Data[14]

This dataset shows how visits and length of stay at different places change compared to a baseline. The baseline is the median value, for the corresponding day of the week, during the 5-week period 01/03/20 through 02/06/20. These values are calculated using the same kind of aggregated and anonymized data used to show popular times for places in Google Maps. The main categories are the following:

1) Retail and Recreation - Mobility trends for places like restaurants, cafes, shopping centers, theme parks, museums, libraries, and movie theaters.
2) Grocery & pharmacy - Mobility trends for places like grocery markets, food warehouses, farmers markets, specialty food shops, drug stores, and pharmacies.
3) Parks - Mobility trends for places like local parks, national parks, public beaches, marinas, dog parks, plazas, and public gardens.
4) Transit stations - Mobility trends for places like public transport hubs such as subway, bus, and train stations.
5) Residential - Mobility trends for places of residence.
6) Workplace - Mobility trends for places of work.

This data is on a county-date level from 2/15/20 to 5/25/21. There are quite a lot of values missing for the above six categories and since movement in states can vary greatly by location, we did not want to fill with the state mobility average or median. If we had more time, we would have filled the missing data with the median of surrounding counties on the same date.

See appendix figures 5 through 9 for visualizations and descriptive statistics of our data.

4. Machine Learning and Details of Solution

Given that resources are limited, state governments need to have a way to prepare for future outbreaks and make sure to prioritize the places where the outbreaks are most severe. Therefore, we decided to focus on predicting the number of Covid-19 cases next week so that governments have time to prepare or change their current mandates -whether that be through quarantine orders or mask requirements. Since we know we want to predict the number of Covid-19 cases in a future week, and have that as a variable in our training data, this is a supervised learning problem. To predict this we trained several regression models including Ridge, Lasso, Elastic Net, SVR, and Random Forest. The use of these models in epidemiologic prediction is not without precedent as seen in Wiemken et al, Weng et al, Dharani et al.[15,16,17] To accomplish our task we had to do the following tasks in the given order:

1) We cleaned the data from the sources above and combined them together on a county (using fips code)-date level. With the date range of 07/31/2020 through 05/24/2021, we had 89,993 observations and 84 variables in our final dataset.

[14] COVID-19 Community Mobility Reports. Accessed June 2, 2021. https://www.google.com/covid19/mobility/

[15] Machine Learning in Epidemiology and Health Outcomes Research. Timothy L. Wiemken and Robert R. Kelley. Annual Review of Public Health 2020 41:1, 21-36

[16] Weng SF, Vaz L, Qureshi N, Kai J. Prediction of premature all-cause mortality: A prospective general population cohort study comparing machine-learning and standard epidemiological approaches. *PLoS ONE.* 2019;**14**(3):e0214365.

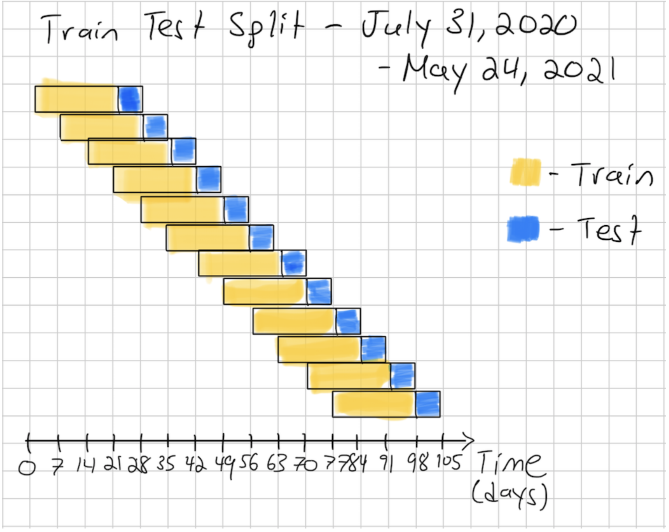
[17] Dharani NP, Bojja P, Raja Kumari P. Evaluation of Performance of an LR and SVR models to predict COVID-19 Pandemic [published online ahead of print, 2021 Feb 16]. *Mater Today Proc*. 2021;10.1016/j.matpr.2021.02.166. doi:10.1016/j.matpr.2021.02.166

2) We then split our data into training and test sets, setting aside 20% of our data to be placed in the test set. The remaining feature selection and model training steps were all conducted exclusively on the training set, while the test set was used only for final evaluation of our best-performing models.
3) Since we were hoping to explore polynomial expansions of our features as well, we decided it would be best to narrow down our features in the beginning so our models are less computationally intensive. We explored a few different feature selection methods:
    a) Correlations: We tried the manual technique of looking at correlations of every variable with every other variable in the dataset. This technique makes it easy to visualise which features are more or less correlated with other features. This is useful to identify features that are closely correlated and features that have low correlation with the target variable (both of which we would want to remove).
    b) Variance Threshold: The variance of a variable is simply the squared deviation of a variable from its mean. Variance tells us how far the data points are spread out for a given variable. We used scikit-learn's VarianceThreshold() method to remove features with zero variance by default. In our dataset, this did not remove any features since no features had zero variance.
    c) SelectKBest: The SelectKBest methodology measures the relationship between each variable and the outcome variable. Specifically, we used the mutual information score with SelectKBest to determine which variables contribute most to our target variable (Covid-19 cases next week). We purposefully kept our k large here (k=50) due to the overall number of variables, especially since we are combining this with Lasso regression to further narrow down variables for our final models. See figure 10 in the appendix for the ordering of feature importance.
    d) Lasso regularization: Lasso regression is also a useful feature selection tool, as it automatically eliminates the weights of the least important features. Therefore, we took our top 50 most important features (as determined by SelectKBest) and narrowed them down to 15 features using lasso regression with a conservative alpha of 0.2. See figure 11 in the appendix for the coefficients on these top 15 features.
4) Next, we also conducted feature engineering, which included scaling/normalizing the numeric features and one-hot-encoding the categorical features (state and county). We also explored adding polynomial features since we wouldn't expect current Covid-19 cases to have a linear relationship with future Covid-19 cases.
5) Since our data was temporal, we needed to come up with a way to create training and validation datasets that addressed the time-element of Covid-19 trends.
    a) First, we looked into using the sklearn TimeSeriesSplit function. However, this only works if the dataset is unique by date and since our data is unique by county-date, the resulting split would make no sense.
    b) Then, we had to decide if we wanted to split time using a "staircase" method or "nested" method. The "staircase" method would train a few weeks of data and then predict the next week, then train the next few weeks of data, and then predict the week after that and so on. The "nested" model would train one week and then predict the next, then train those 2 weeks and predict the third, and then train the 3 weeks and predict the fourth, etc. After reading data sources illustrating both methods[18,19], we decided that the "staircase" method made the most sense for our data. Using the

[18] Cochrane C. Time Series Nested Cross-Validation. Towards Data Science. Accessed June 2, 2021. https://towardsdatascience.com/time-series-nested-cross-validation-76adba623eb9

[19] Herman-Saffar O. Time Based Cross Validation. Medium. Published January 26, 2020. Accessed June 2, 2021. https://towardsdatascience.com/time-based-cross-validation-d259b13d42b8

process laid out in the Towards Data Science article *Time Based Cross Validation*, we created timesplit.py which does the “staircase” split for us.



6) Finally after cleaning our data, creating a train/test dataset, and selecting our features, we ran a series of regression models (described below). To evaluate which model performed best, we used Sklearn’s GridSearchCV function to test different combinations of parameters using time-based cross validation. For each model specification, we measured performance via R-squared, mean squared error, and mean absolute error.

4.1 Model Details

Ridge Regression

One of the first linear models we learned in class was the Ridge Regression model, a regularized version of linear regression. Ridge regression imposes a penalty on coefficient size, which helps prevent overfitting. The Ridge Regression cost function is as follows:

$$J(\theta) = MSE(\theta) + \alpha\frac{1}{2}\sum_{i=1}^{n}\theta_i^2$$

In this equation, the hyperparameter **α** controls the level of regularization (penalty) in the model. In our context, we used Grid Search to run Ridge Regression with the following values for **α**: 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1. See figures 12 and 13 in the appendix.

We believed that Covid-19 cases might be better represented using Polynomial features, rather than simple regressions. With this in mind, we tested the Ridge Regression approach using polynomial features to degrees 2 and 3. For this, we used alpha values of 0.5, 0.6, 0.7, 0.8, 0.9, and 1. We found that using polynomial features with Ridge regression performed worse than Ridge regression without polynomial features. See figures 14 and 15 in appendix

Lasso Regression

Following Ridge regression, we attempted to fit our data using Lasso Regression. Similar to Ridge regression, Lasso regression is a regularized version of linear regression. Unlike Ridge regression, Lasso regression reduces the weights of useless features to zero. The Lasso Regression cost function is as follows:

$$J(\theta) = MSE(\theta) + \alpha \sum_{i=1}^{n} \left|\theta_i\right|$$

We performed Lasso Regression with the following values of **α**: 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 1. See figures 16 and 17 for our results using Lasso Regression.

Additionally, we used Lasso Regression with Polynomial features to degree 2 and 3. Similar to what we saw in Ridge Regression, the Lasso Regression with polynomial features performed worse than the Lasso Regression without polynomial features (see appendix figures 18 and 19).

Elastic Net Regression

When several features are strongly correlated with one another, Lasso regression tends to behave erratically. Elastic Net Regression is a good alternative to Lasso Regression. Elastic Net mixes the regularization terms of Ridge and Lasso regression. The Elastic Net cost function is as follows:

$$J(\theta) = MSE(\theta) + r\alpha \sum_{i=1}^{n} \left|\theta_i\right| + \frac{1-r}{2}\alpha \sum_{i=1}^{n} \theta_i^2$$

For Elastic Net, we used **α** values of: 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 1. Elastic net also includes a mixing parameter, l1_ratio. L1_ratio (same as r in the above equation) controls which penalty term (Ridge vs Lasso) holds the most weight. We used the following l1_ratio values: 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 1. The results of the Elastic Net regression can be found in the appendix.

Additionally, we used Elastic Net Regression with Polynomial features to degree 2 and 3. Here, we used the same range of value for alpha and l1_ratio as before. The results for Elastic Net regression with polynomial features are found in appendix figures 20 through 22.

Random Forest Regression

We chose to explore Random Forest regressions - an ensemble technique that uses multiple decision trees and bagging. The Random Forest algorithm introduces extra randomness when growing trees; instead of searching for the very best feature when splitting a node, it searches for the best feature among a random subset of features. This is reasonable for our problem because we want to know the trends in Covid-19 cases and show us the expected dips and rises in future Covid-19 cases. We explored different combinations of parameters to see which performed the best in terms of MSE, MAE, and R-squared comparisons. We tuned different parameters like criterion (MSE or MAE), maximum features (auto, square root, and log), maximum depth (None, 5, and 10), and minimum samples split (2 or 5). Unfortunately, due to the length of time it took to run these combinations (over 15+ hours), we were not able to explore more hyperparameters. Results can be seen in figures 23 through 26 in the appendix.

Support Vector Regression (SVR).

We chose to explore support vector regressions–a regression model from the support vector machine model family–in order to leverage the ability to run regressions while defining a level of acceptable error. This is reasonable for our problem because we are more concerned with predicting surges and approximate trends in Covid-19 cases rather than the exact number of Covid-19 cases in the future. For our models, we explored a few options for the underlying kernel (linear, polynomial, and radial basis function) as well as ranges of different values for C (the regularization parameter) and epsilon (the error tolerance). After finding that the best performing models were those with a linear kernel, we also experimented with a different Sklearn implementation of linear SVR–svm.LinearSVM– which has more flexibility in the choice of penalties and loss functions and scales better to larger numbers of samples. See appendix for some MSE, MAE, and R-squared comparisons. Results can be seen in figures 27 through 30 in the appendix.

5. Evaluation and Results

To evaluate our model performance, we used three evaluation metrics: mean squared error (MSE), mean absolute error (MAE), and R-squared. Since our problem is a regression problem, we elected to look at mean squared error to evaluate how close our predicted Covid-19 cases were to the true Covid-19 cases. Since mean squared error is sensitive to outliers, and Covid-19 trends may be susceptible to sharp peaks, we decided to also add mean absolute error to our evaluation metrics in case we find models that may not be the best at minimizing MSE but are great at minimizing MAE. Finally, we also examined the R-squared of each model to evaluate how well our models are explaining the data.

In determining our set of best-performing models, we considered the models that generated the least MSE, the models that generated the least MAE, and the models that generated the highest R2. Based on the cross-validation on the training data, we found that one of the random forest models performed the best in terms of MSE and R2 and a support vector regression model performed best in terms of MAE. We also identified a few other better-performing model specifications among each of the types of models we ran so we could compare the performance of a few different types of models on the test data.

Running our set of best-performing models on the test data, we found that the random forest regression model with max_features set to “log2” and all other parameters set to their defaults performed the best in terms of MSE, MAE, and R-squared,

Table 1. Performance of best-performing models on test data

Model	**MSE**	**MAE**	**R2**
RandomForestRegressor(max_features='log2')	137.09	3.28	0.99
RandomForestRegressor(criterion='mae')	151.20	3.42	0.99
LinearSVR(C=1, epsilon=0)	593.90	6.13	0.94
SVR(C=20, epsilon=0, kernel='linear')	588.74	6.13	0.94
Lasso(alpha=1)	363.34	6.66	0.96
Ridge(alpha=1)	580.54	6.87	0.94
ElasticNet(alpha=0.1)	671.02	7.63	0.93

From this best-performing random forest model, the features that were identified to have the greatest importance were the seven day average of new Covid-19 cases, new Covid-19 cases that day, the number of adults admitted into hospitals with confirmed Covid-19, the number of adults between the ages of 50 and 59 admitted into hospitals with confirmed Covid-19, and the number of adults over age 80 that were admitted into hospitals with confirmed Covid-19. See appendix figure 31.

While this best performing random forest model had the lowest bias among all the best-performing models, it did also unfortunately have the highest variance among all models. However, based on our validation against the test set, this model does not seem to be overfit to the training data as it still generated the lowest MSE and MAE on the test set. See figures 32 through 34 in the appendix.

Finally, we took our best performing model, the random forest model, and used it to predict Covid-19 cases in Cook County over time (appendix figure 35) and for this current week (appendix figure 36).

6. Ethics

We chose to include demographic information on each county, including race, as predictive features in our above models. We did so because we know that Covid-19 cases are disproportionately distributed by race and ethnicity, and that race in the United States is linked to factors such as geographic location, socioeconomic status, health, societal expectations, and more. In their article, "Race, Ethnicity, and Age Trends in Persons Who Died from COVID-19 — United States, May–August 2020," researchers found that between May and August of 2020, "...Black persons still accounted for 18.7% of overall deaths despite representing just 12.5% of the U.S. population. Similarly, Hispanic persons were disproportionately represented among decedents: 24.2% of decedents were Hispanic compared with 18.5% of the U.S. population".[20] Although race and Covid-19 cases seem to be related, it is important to consider the potential biases and issues that may arise in including race in a Machine Learning model. As we've seen in class, if there is societal bias in a given training set, such bias will be embedded in any predictions that the model produces. We should also note that when minimizing the average mean squared error, as we did in all of our above models, we are minimizing the error in the overall/majority population. This could be potentially problematic, as the model will not be fit to represent minority populations.

In addition to the inclusion of race and demographic information to our machine learning models, it is also important to note the potential implications that our predictions may have on policy decisions. Over the past year and a half, states and cities have been faced with the decisions to impose stay-at-home orders, and restrict the openings of businesses, schools, public facilities, outdoor parks and recreation, and more. There is a delicate balance between restricting such openings to prevent people from contracting/spreading Covid-19, while also allowing for our country's economy to progress.

On one side of this argument, evidence suggests that state/city-wide lockdowns and stay at home orders do, in fact, decrease the spread of Covid-19. In the article, "The effect of state-level stay-at-home orders on COVID-19 infection rates," researchers studied the effectiveness of state-wide stay at home orders in 42 states and the

[20] Rossen LM, Branum AM, Ahmad FB, Sutton P, Anderson RN. Excess Deaths Associated with COVID-19, by Age and Race and Ethnicity - United States, January 26-October 3, 2020. MMWR Morb Mortal Wkly Rep. 2020;69(42):1522-1527. Published 2020 Oct 23. doi:10.15585/mmwr.mm6942e2

District of Columbia. They found that the issuing of stay-at-home orders play a significant role in “flattening the curve” (decreasing infection rates)[21].

On the other side, stay-at-home orders have disproportionate negative effects on people who work low-wage jobs, people of color, women, single mothers, and other underrepresented groups[22]. In their article, “Early Signs Indicate That COVID-19 Is Exacerbating Gender Inequality in the Labor Force,” the authors found that during the period of February-April 2020, labor force participation of mothers of young children decreased by 3.2 percentage points among mothers with children younger than 6, where fathers exit rates are 1-2 percentage points lower in comparison. Additionally, outbreaks have occurred more frequently in communities of color where there’s a higher percentage of people working in in-person jobs. However, shutdowns may more negatively impact these communities than communities where people are able to work in stay-at-home jobs. Therefore, it’s a difficult line since we obviously want to protect the most vulnerable communities from getting an outbreak, but doing so may make these communities even more vulnerable economically. This bias could be helped by making sure these communities have access to vaccines.

In summary, although government officials could potentially use machine learning models and predictions to further inform their policy decisions, it is critical to consider how different populations are disproportionately affected by Covid-19, stay-at-home orders, and lockdowns.

7. Limitations/Further steps

There are a few additional avenues we would have liked to explore further that were unfortunately out of the scope of this project. These include:

- **Incorporating all Midwestern states into our model**. We chose to limit our model to the three states with the most reliable county-level vaccine data. However, we would have liked to have been able to secure vaccine data from all Midwestern states, perhaps through contacting state departments of health directly.
- **Incorporating Covid-19 variant data into our model.** While we did explore several sources of Covid-19 variant data, we found the data to be too inconsistent and sparse at the time given that variants were just beginning to be tracked in the US. However, we feel that Covid-19 variants are an important factor in the current state of the pandemic, and therefore would have liked to have been able to incorporate it into our model.
- **Explore other time-based cross validation approaches.** We would have liked to also have tried the nested time-based cross validation approach to see if it would perform better.
- **Investigate overfitting.** Our best fitting model had low bias but high variance. Though it performed well on the test data, we would have still liked to investigate our model further and confirm that we don’t have any concerns about overfitting.

[21] Castillo RC, Staguhn ED, Weston-Farber E. The effect of state-level stay-at-home orders on COVID-19 infection rates. Am J Infect Control. 2020;48(8):958-960.
doi:10.1016/j.ajic.2020.05.017

[22] Landivar LC, Ruppanner L, Scarborough WJ, Collins C. Early Signs Indicate That COVID-19 Is Exacerbating Gender Inequality in the Labor Force. Socius. January 2020. doi:10.1177/2378023120947997

APPENDIX

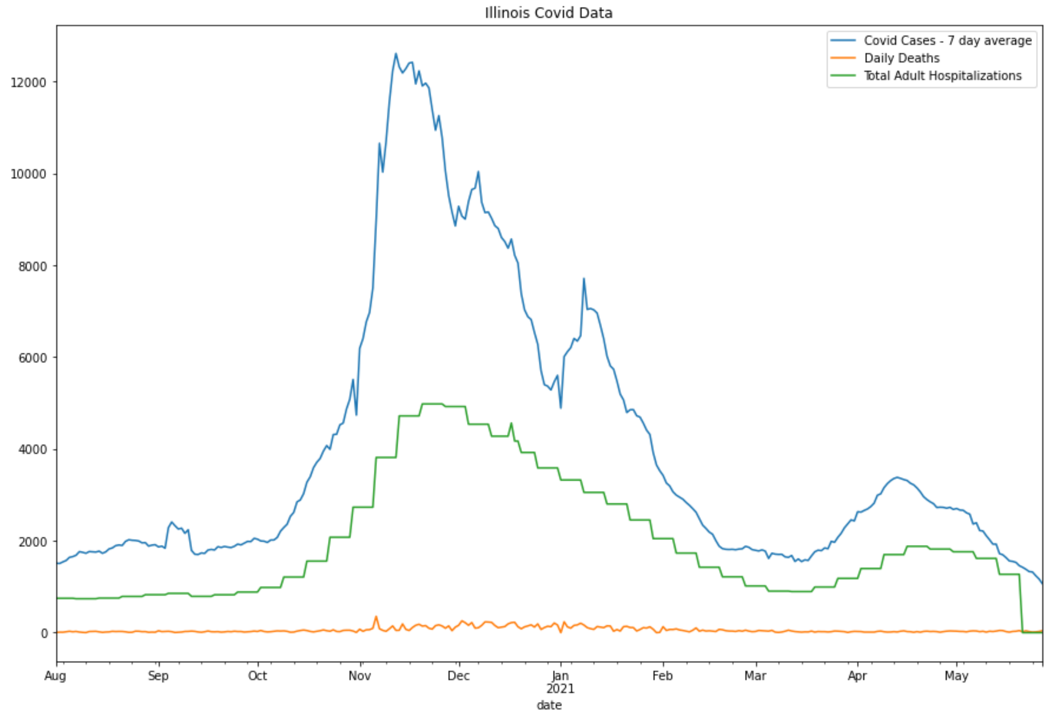


Fig. 1: Illinois Covid-19 Cases, Deaths, and Hospitalizations over Time

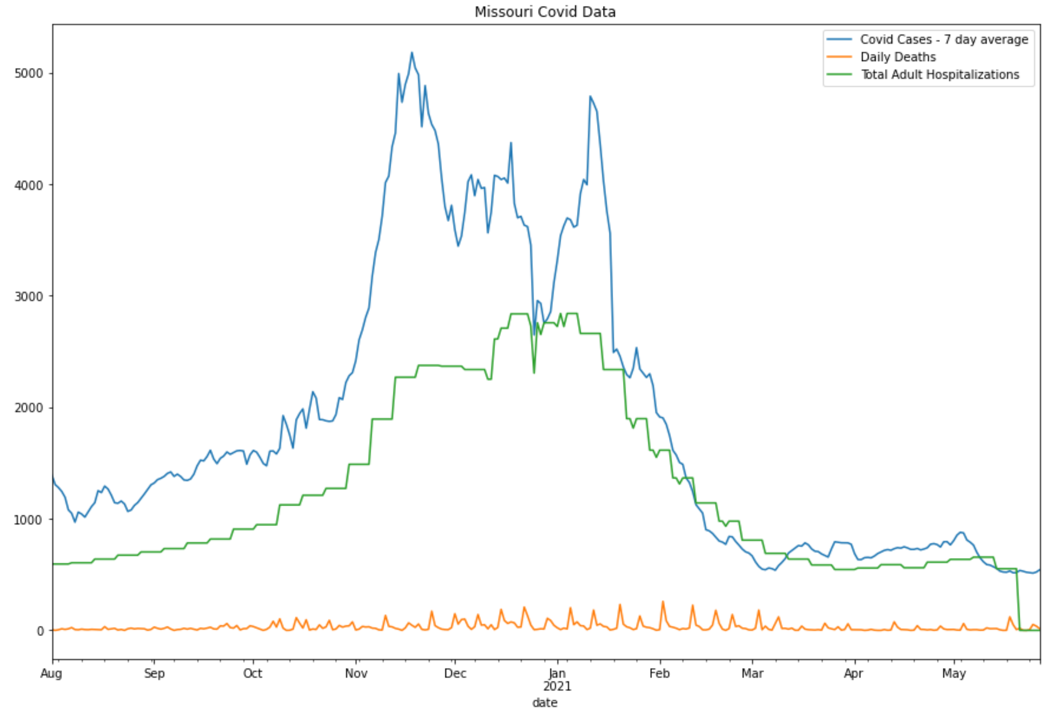


Fig. 2: Missouri Covid-19 Cases, Deaths, and Hospitalizations over Time

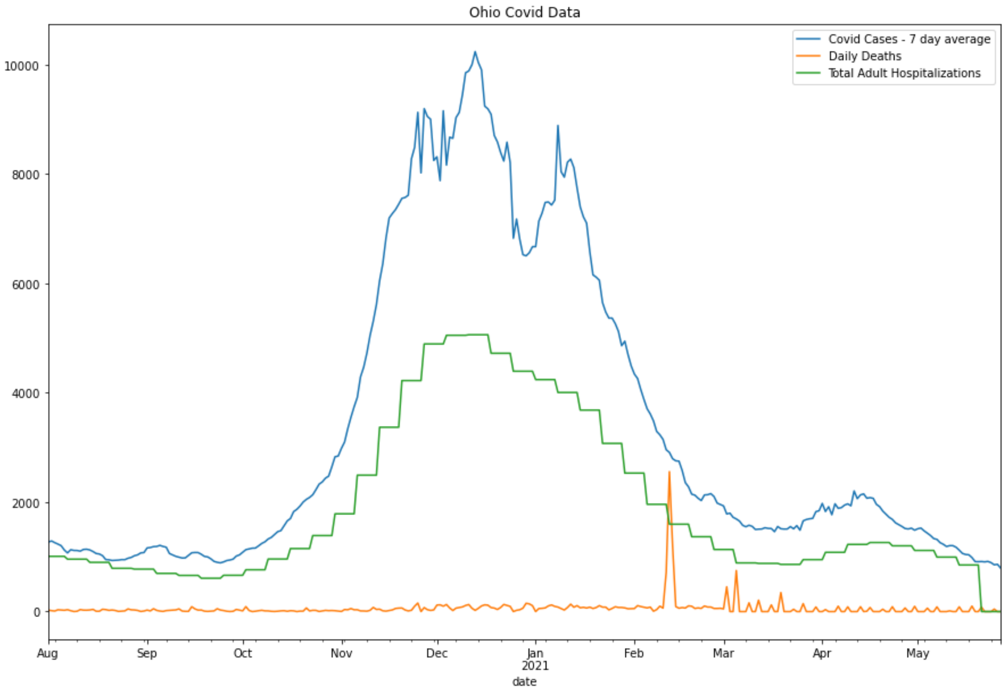


Fig. 3: Ohio Covid-19 Cases, Deaths, and Hospitalizations over Time

Data	Source	Aggregation Level	Time Range
Number of Covid-19 Cases, Deaths	The New York Times	County-Date	1/21/2020 - 5/27/2021
Vaccination Administration	State-level Departments of Health	State-Date	12/13/2020 - 5/27/2021
Public Mask Mandates	Centers for Disease Control and Prevention	County-Date	4/10/2020 - 3/22/2021
Hospital Utilization	Department of Health and Human Services	Hospital-Week	7/31/2020 - 5/24/2021
Travel/Mobility	Google	County-Date	2/15/20 to 5/25/21
Population Density, Sex, Age, Race, Poverty	Census Bureau	County	2019

Fig. 4 Data Sources Summary

Variable	Mean	Std	Min	25%	50%	75%	Max
New cases (7 day average)	31	125	0	2	7	21	4654
New cases (7 day average, 2 weeks ago)	31	125	0	2	7	21	4654
New cases	31	132	0	1	6	20	6697
Total aged 45 - 54	13113	42957	241	1877	4177	8645	658679
White (total)	78252	202651	1982	13694	30578	60881	2946314
Total aged 35 - 44	12521	44667	197	1645	3699	8126	701726
Cumulative Cases	5935	23958	6	463	1463	3880	549205
Asian (total)	3549	23439	0	44	171	576	379444
Other race (total)	3130	29161	0	28	134	629	500069
Adult Covid-19-related hospital admissions (7 day sum)	16	62	0	0	3	10	1641
White (percent)	91	8	46.5	89.6	94.3	96.5	99.3
Retail and recreation mobility	-3	16	-97	-11	-3	6	213
Adult Covid-19-related hospital admissions for ages 80+ (7 day sum)	4	13	0	0	0	3	300
Percent 65 and older	19	3	8.3	16.8	18.5	20.5	33.5
Adult Covid-19-related hospital admissions for ages 50 to 59 (7 day sum)	3	11	0	0	0	3	278

Fig. 5 Summary statistics of the final features

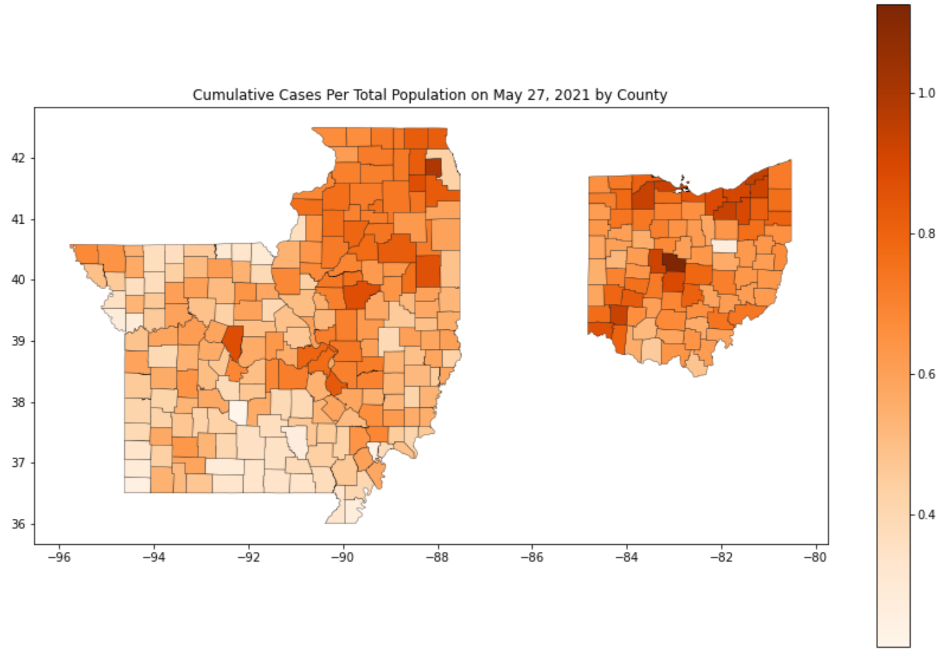

Cumulative Cases Per Total Population on May 27, 2021 by County
42
41
40
39
38
37
36
−96
−94
−92
−90
−88
−86
−84
−82
−80
1.0
0.8
0.6
0.4

Fig. 6: Cumulative Cases Per Total Population on May 27, 2021

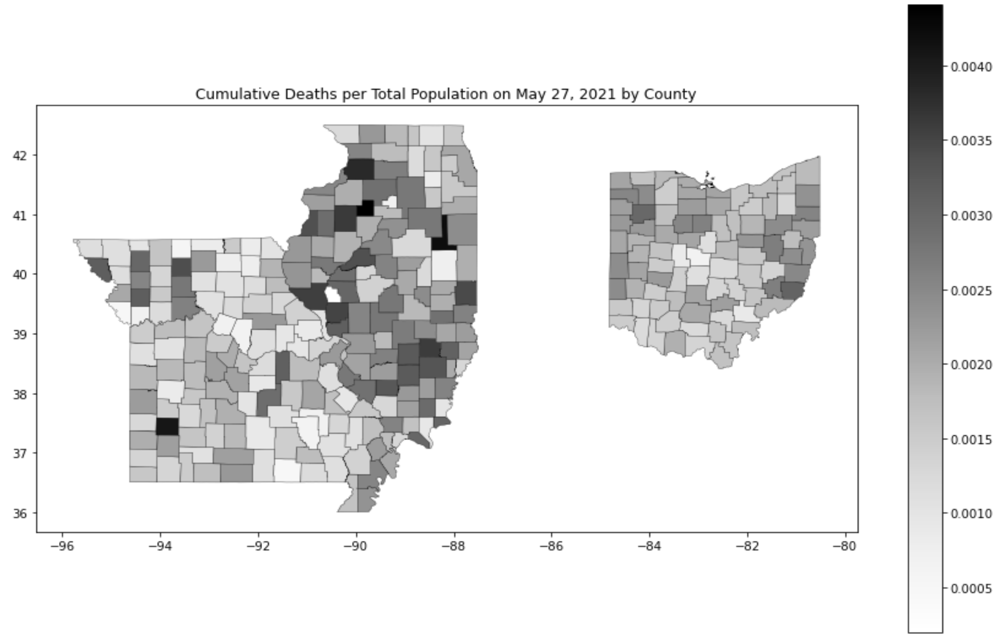


Fig. 7: Cumulative Deaths Per Total Population on May 27, 2021

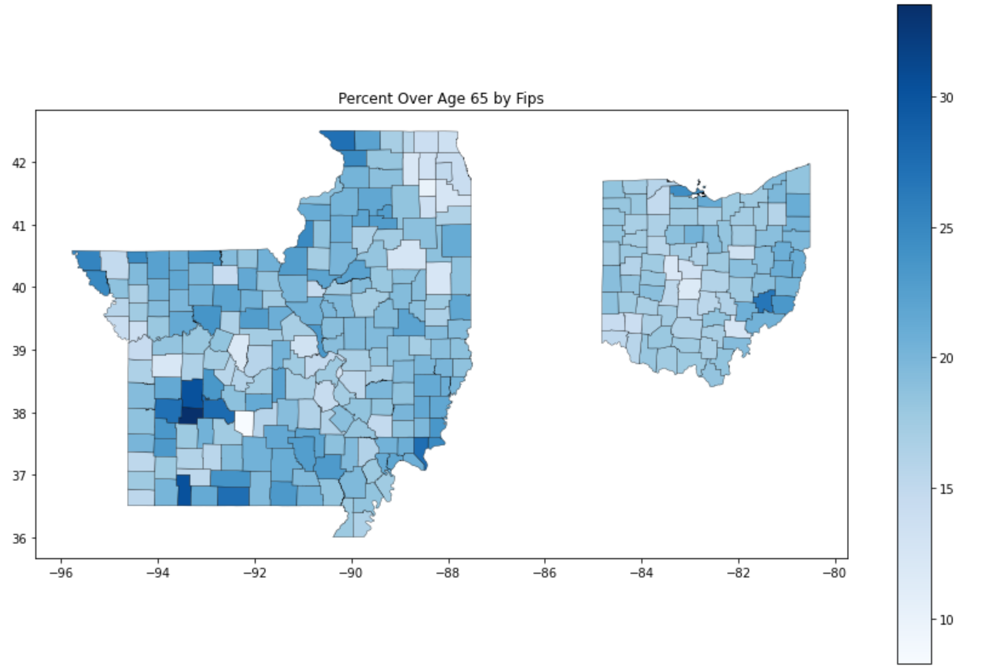


Fig. 8: Percent of Population over Age 65

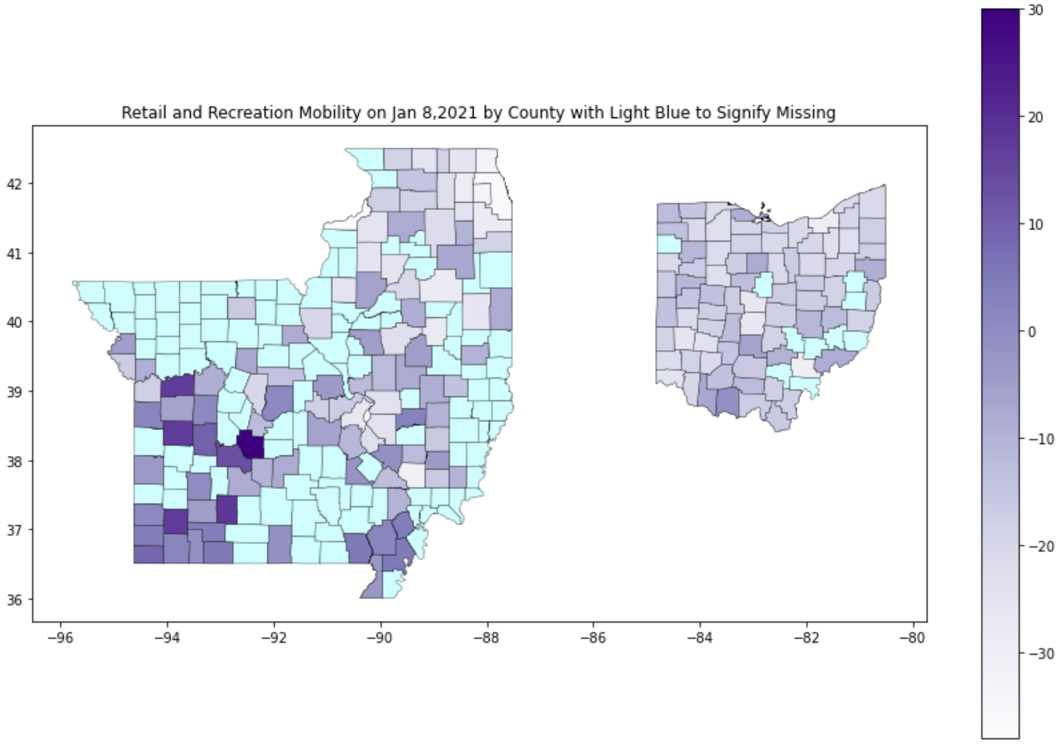
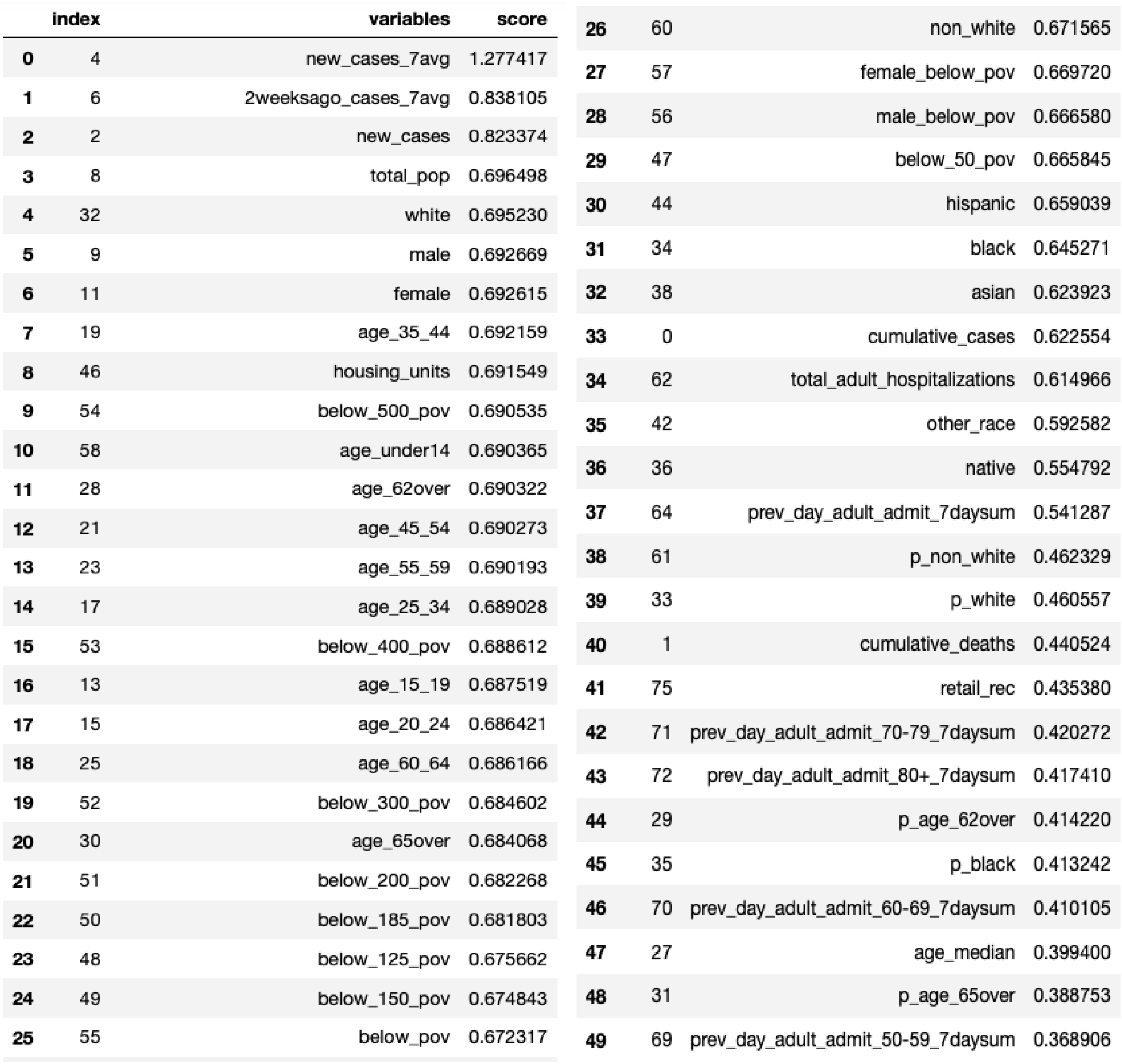
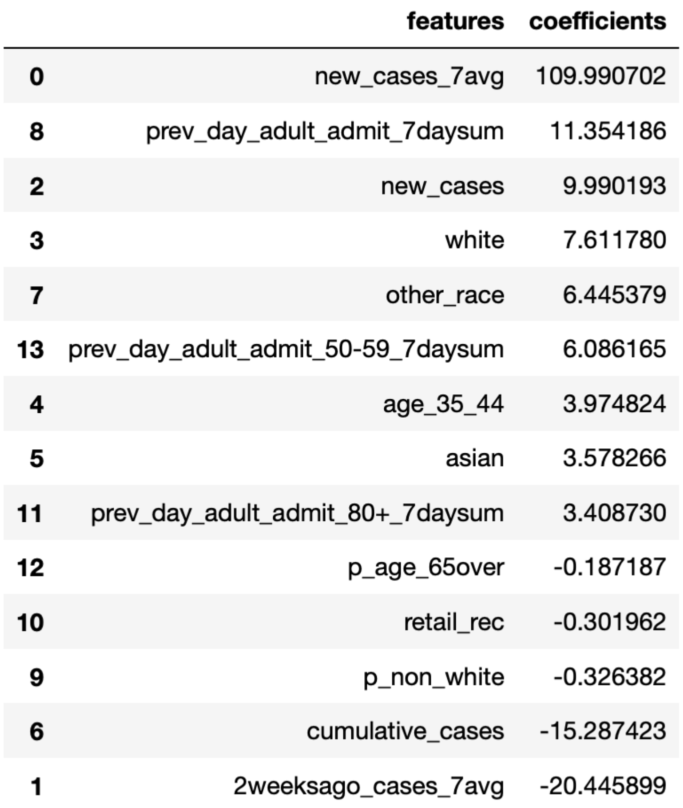

Retail and Recreation Mobility on Jan 8,2021 by County with Light Blue to Signify Missing
42
41
40
39
38
37
36
−96
−94
−92
−90
−88
−86
−84
−82
−80
30
20
10
0
−10
−20
−30

Fig. 9: Retail and Recreation Mobility on May 27, 2021 (Relative to Mobility in January 2020)



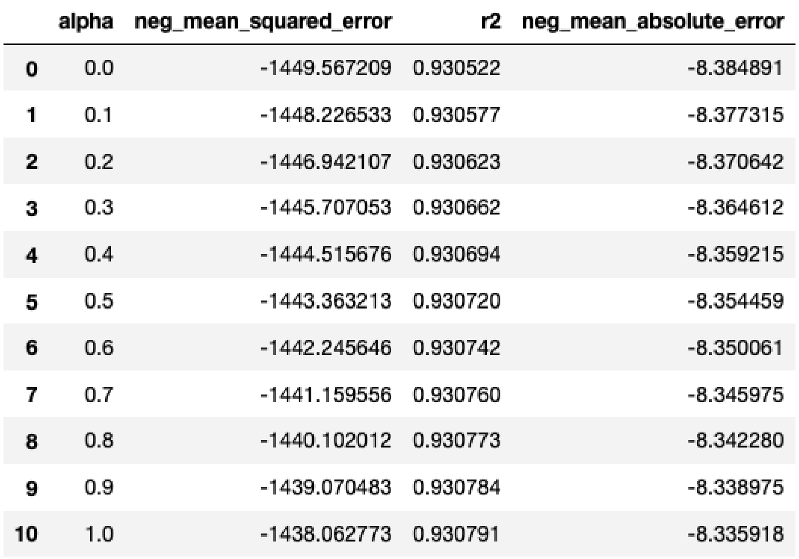
	index	variables	score
0	4	new_cases_7avg	1.277417
1	6	2weeksago_cases_7avg	0.838105
2	2	new_cases	0.823374
3	8	total_pop	0.696498
4	32	white	0.695230
5	9	male	0.692669
6	11	female	0.692615
7	19	age_35_44	0.692159
8	46	housing_units	0.691549
9	54	below_500_pov	0.690535
10	58	age_under14	0.690365
11	28	age_62over	0.690322
12	21	age_45_54	0.690273
13	23	age_55_59	0.690193
14	17	age_25_34	0.689028
15	53	below_400_pov	0.688612
16	13	age_15_19	0.687519
17	15	age_20_24	0.686421
18	25	age_60_64	0.686166
19	52	below_300_pov	0.684602
20	30	age_65over	0.684068
21	51	below_200_pov	0.682268
22	50	below_185_pov	0.681803
23	48	below_125_pov	0.675662
24	49	below_150_pov	0.674843
25	55	below_pov	0.672317
26	60	non_white	0.671565
27	57	female_below_pov	0.669720
28	56	male_below_pov	0.666580
29	47	below_50_pov	0.665845
30	44	hispanic	0.659039
31	34	black	0.645271
32	38	asian	0.623923
33	0	cumulative_cases	0.622554
34	62	total_adult_hospitalizations	0.614966
35	42	other_race	0.592582
36	36	native	0.554792
37	64	prev_day_adult_admit_7daysum	0.541287
38	61	p_non_white	0.462329
39	33	p_white	0.460557
40	1	cumulative_deaths	0.440524
41	75	retail_rec	0.435380
42	71	prev_day_adult_admit_70-79_7daysum	0.420272
43	72	prev_day_adult_admit_80+_7daysum	0.417410
44	29	p_age_62over	0.414220
45	35	p_black	0.413242
46	70	prev_day_adult_admit_60-69_7daysum	0.410105
47	27	age_median	0.399400
48	31	p_age_65over	0.388753
49	69	prev_day_adult_admit_50-59_7daysum	0.368906

Fig. 10: SelectKBest Results



	features	coefficients
0	new_cases_7avg	109.990702
8	prev_day_adult_admit_7daysum	11.354186
2	new_cases	9.990193
3	white	7.611780
7	other_race	6.445379
13	prev_day_adult_admit_50-59_7daysum	6.086165
4	age_35_44	3.974824
5	asian	3.578266
11	prev_day_adult_admit_80+_7daysum	3.408730
12	p_age_65over	-0.187187
10	retail_rec	-0.301962
9	p_non_white	-0.326382
6	cumulative_cases	-15.287423
1	2weeksago_cases_7avg	-20.445899

Fig. 11: Lasso Regularization Results



	alpha	neg_mean_squared_error	r2	neg_mean_absolute_error
0	0.0	-1449.567209	0.930522	-8.384891
1	0.1	-1448.226533	0.930577	-8.377315
2	0.2	-1446.942107	0.930623	-8.370642
3	0.3	-1445.707053	0.930662	-8.364612
4	0.4	-1444.515676	0.930694	-8.359215
5	0.5	-1443.363213	0.930720	-8.354459
6	0.6	-1442.245646	0.930742	-8.350061
7	0.7	-1441.159556	0.930760	-8.345975
8	0.8	-1440.102012	0.930773	-8.342280
9	0.9	-1439.070483	0.930784	-8.338975
10	1.0	-1438.062773	0.930791	-8.335918

Fig. 12: Ridge Regression Results for Different Alphas

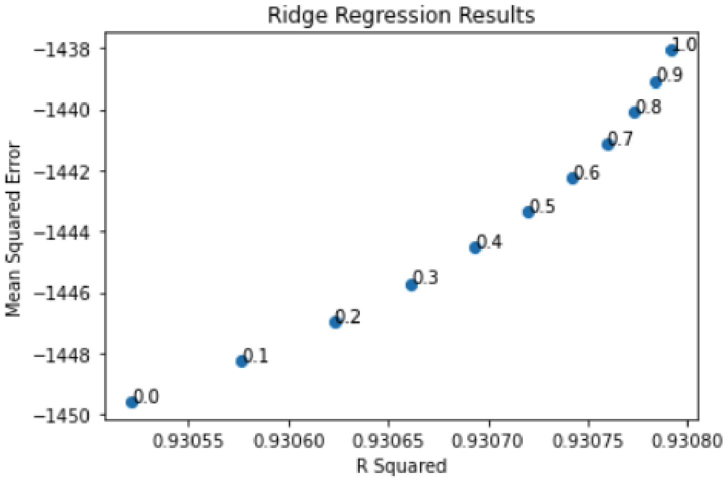
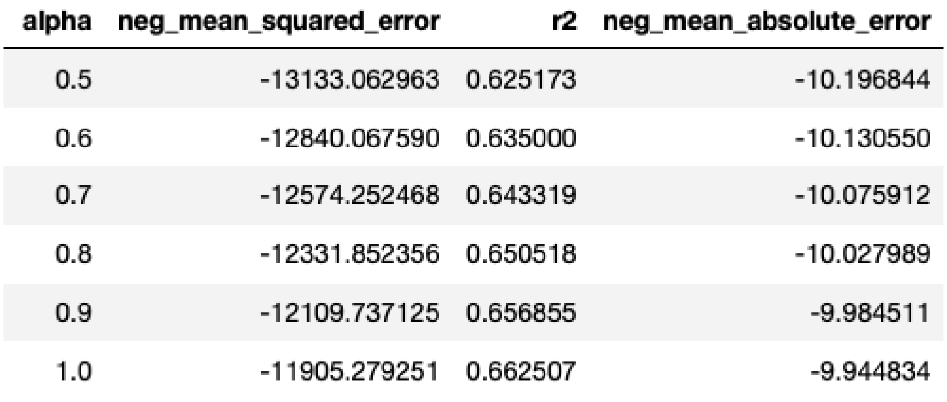
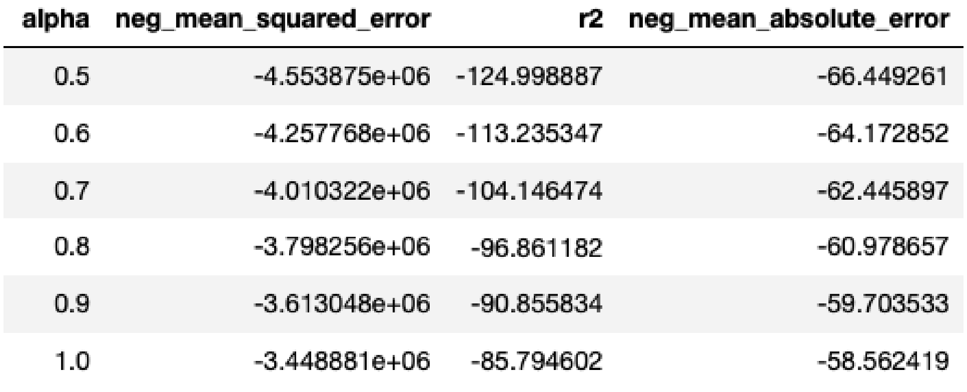


Fig. 13: Ridge Regression Results for Different Alphas



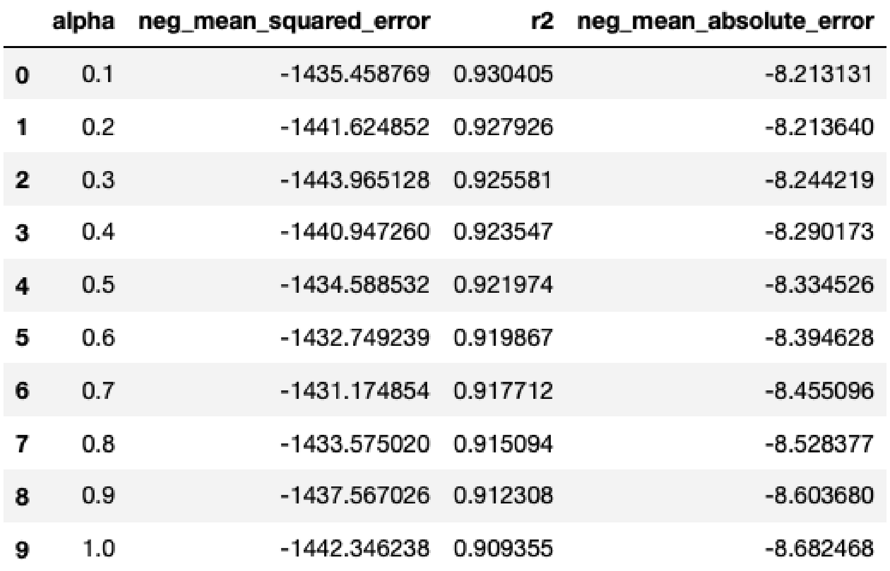
alpha	neg_mean_squared_error	r2	neg_mean_absolute_error
0.5	-13133.062963	0.625173	-10.196844
0.6	-12840.067590	0.635000	-10.130550
0.7	-12574.252468	0.643319	-10.075912
0.8	-12331.852356	0.650518	-10.027989
0.9	-12109.737125	0.656855	-9.984511
1.0	-11905.279251	0.662507	-9.944834

Fig. 14: Ridge Regression Results: Polynomial Features to Degree 2



alpha	neg_mean_squared_error	r2	neg_mean_absolute_error
0.5	-4.553875e+06	-124.998887	-66.449261
0.6	-4.257768e+06	-113.235347	-64.172852
0.7	-4.010322e+06	-104.146474	-62.445897
0.8	-3.798256e+06	-96.861182	-60.978657
0.9	-3.613048e+06	-90.855834	-59.703533
1.0	-3.448881e+06	-85.794602	-58.562419

Fig. 15: Ridge Regression Results: Polynomial Features to Degree 3



	alpha	neg_mean_squared_error	r2	neg_mean_absolute_error
0	0.1	-1435.458769	0.930405	-8.213131
1	0.2	-1441.624852	0.927926	-8.213640
2	0.3	-1443.965128	0.925581	-8.244219
3	0.4	-1440.947260	0.923547	-8.290173
4	0.5	-1434.588532	0.921974	-8.334526
5	0.6	-1432.749239	0.919867	-8.394628
6	0.7	-1431.174854	0.917712	-8.455096
7	0.8	-1433.575020	0.915094	-8.528377
8	0.9	-1437.567026	0.912308	-8.603680
9	1.0	-1442.346238	0.909355	-8.682468

Fig. 16: Lasso Regression Results

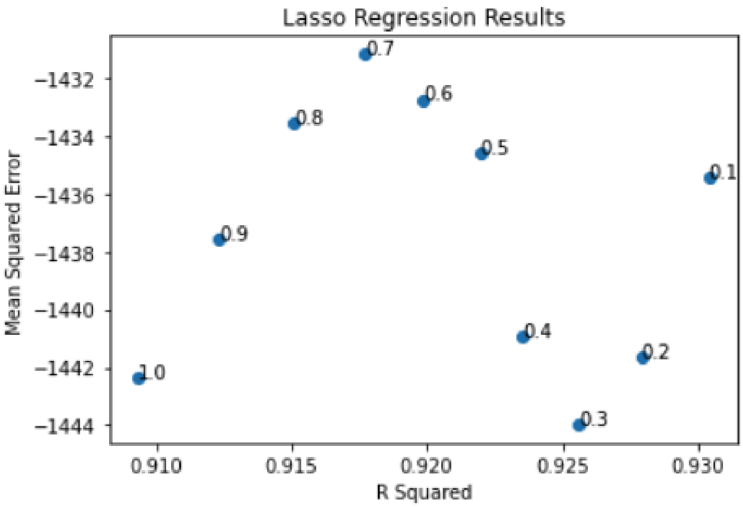
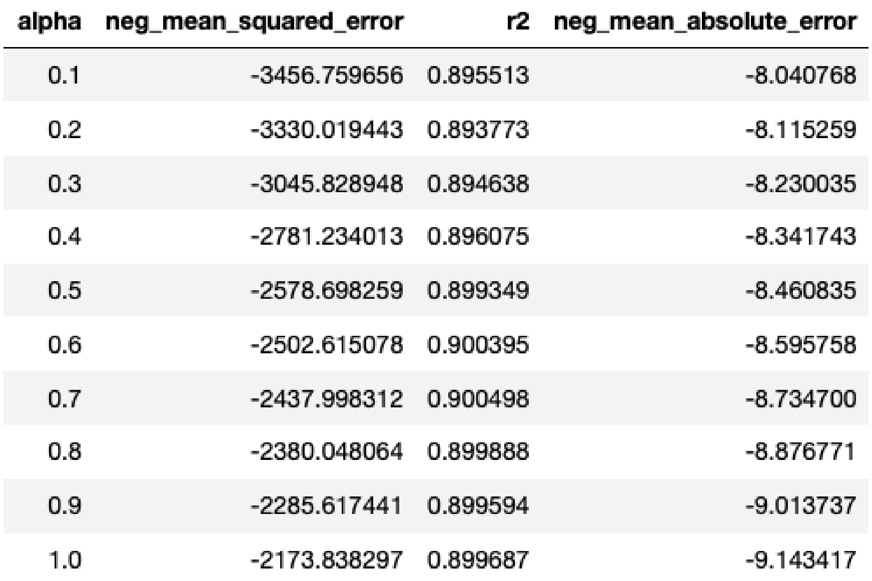
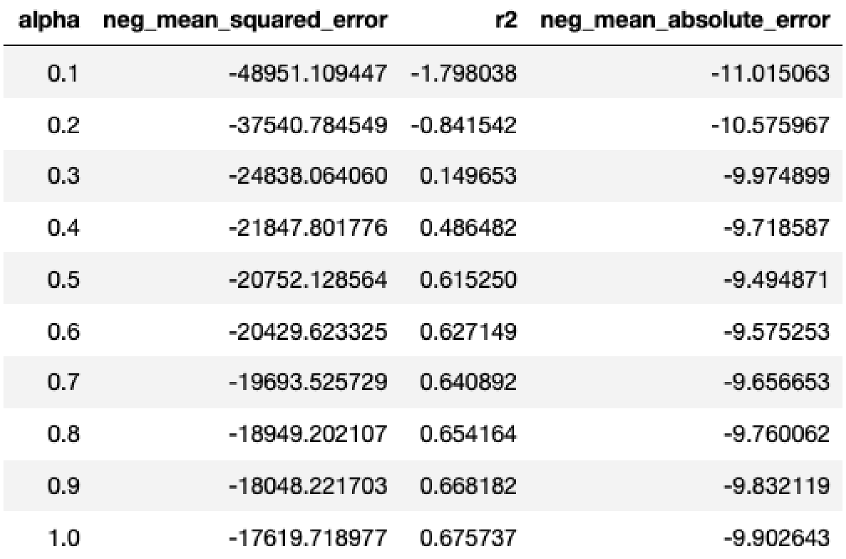


Fig. 17 Lasso Regression Results - R-squared versus MSE



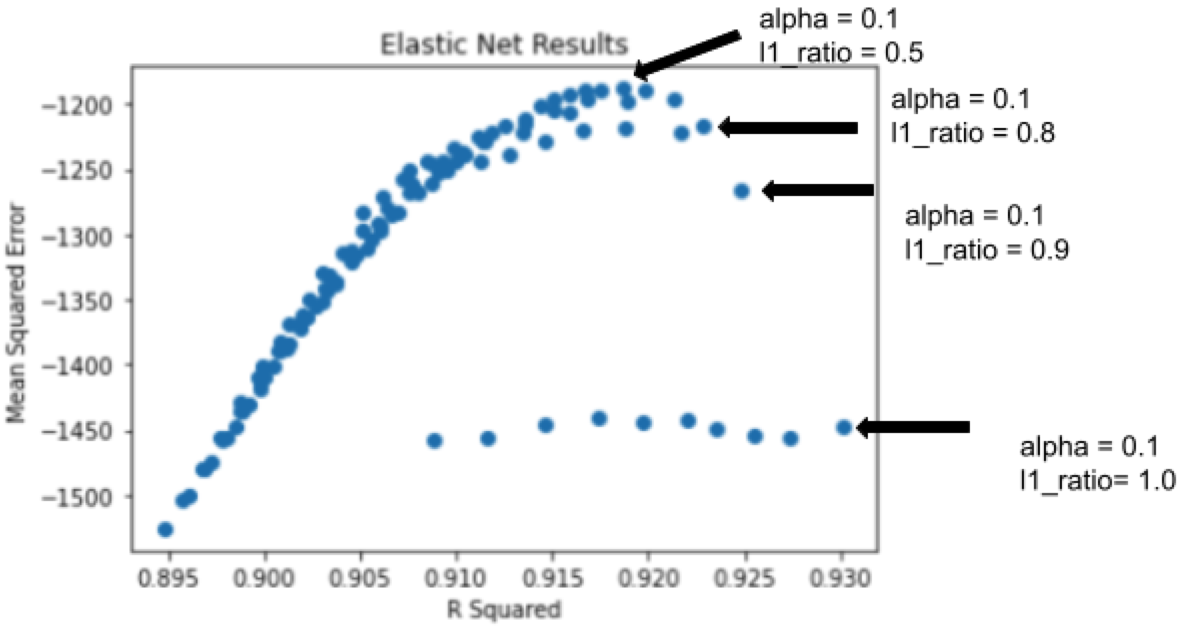
alpha	neg_mean_squared_error	r2	neg_mean_absolute_error
0.1	-3456.759656	0.895513	-8.040768
0.2	-3330.019443	0.893773	-8.115259
0.3	-3045.828948	0.894638	-8.230035
0.4	-2781.234013	0.896075	-8.341743
0.5	-2578.698259	0.899349	-8.460835
0.6	-2502.615078	0.900395	-8.595758
0.7	-2437.998312	0.900498	-8.734700
0.8	-2380.048064	0.899888	-8.876771
0.9	-2285.617441	0.899594	-9.013737
1.0	-2173.838297	0.899687	-9.143417

Fig 18: Lasso Regression Results: Polynomial Feature Degree 2



alpha	neg_mean_squared_error	r2	neg_mean_absolute_error
0.1	-48951.109447	-1.798038	-11.015063
0.2	-37540.784549	-0.841542	-10.575967
0.3	-24838.064060	0.149653	-9.974899
0.4	-21847.801776	0.486482	-9.718587
0.5	-20752.128564	0.615250	-9.494871
0.6	-20429.623325	0.627149	-9.575253
0.7	-19693.525729	0.640892	-9.656653
0.8	-18949.202107	0.654164	-9.760062
0.9	-18048.221703	0.668182	-9.832119
1.0	-17619.718977	0.675737	-9.902643

Fig 19: Lasso Regression Results: Polynomial Feature Degree 3



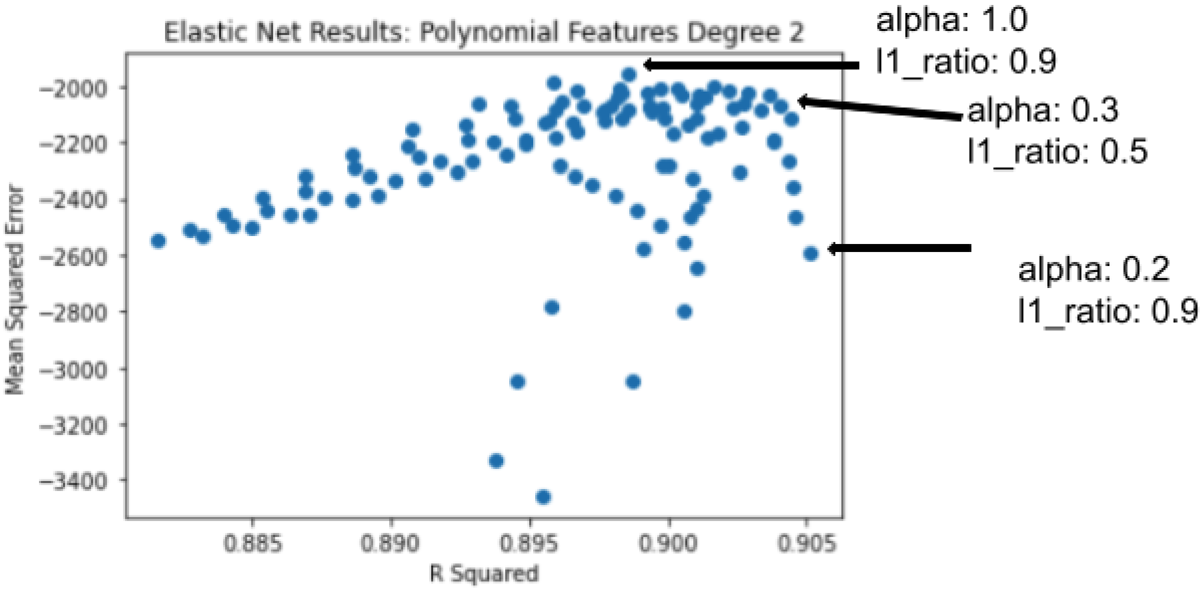
alpha = 0.1
l1_ratio = 0.5

alpha = 0.1
l1_ratio = 0.8

alpha = 0.1
l1_ratio = 0.9

alpha = 0.1
l1_ratio= 1.0

Fig. 20: Elastic Net Results

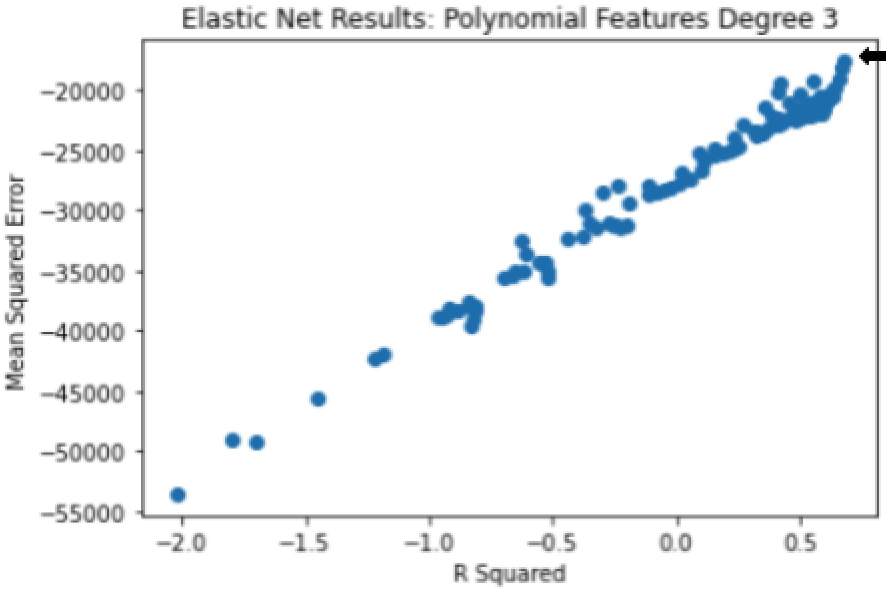


alpha: 1.0
l1_ratio: 0.9

alpha: 0.3
l1_ratio: 0.5

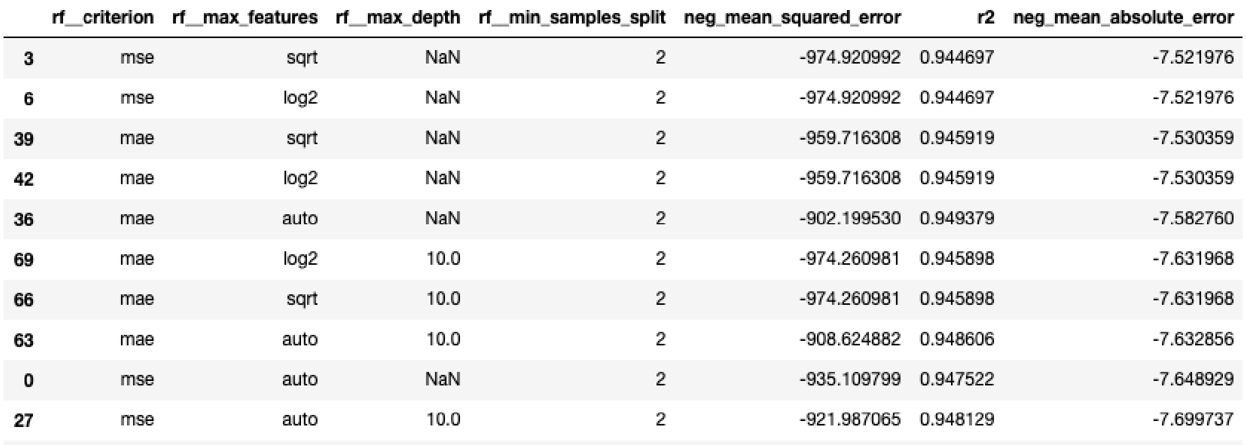
alpha: 0.2
l1_ratio: 0.9

Fig. 21: Elastic Net Results (Polynomial Degree 2)



alpha: 1.0
l1_raio: 1.0

Fig. 22: Elastic Net Results (Polynomial Degree 3)



	rf__criterion	rf__max_features	rf__max_depth	rf__min_samples_split	neg_mean_squared_error	r2	neg_mean_absolute_error
3	mse	sqrt	NaN	2	-974.920992	0.944697	-7.521976
6	mse	log2	NaN	2	-974.920992	0.944697	-7.521976
39	mae	sqrt	NaN	2	-959.716308	0.945919	-7.530359
42	mae	log2	NaN	2	-959.716308	0.945919	-7.530359
36	mae	auto	NaN	2	-902.199530	0.949379	-7.582760
69	mae	log2	10.0	2	-974.260981	0.945898	-7.631968
66	mae	sqrt	10.0	2	-974.260981	0.945898	-7.631968
63	mae	auto	10.0	2	-908.624882	0.948606	-7.632856
0	mse	auto	NaN	2	-935.109799	0.947522	-7.648929
27	mse	auto	10.0	2	-921.987065	0.948129	-7.699737

Fig. 23: Random Forest Regression: Top 10 Model Specifications (by Negative Mean Absolute Error)

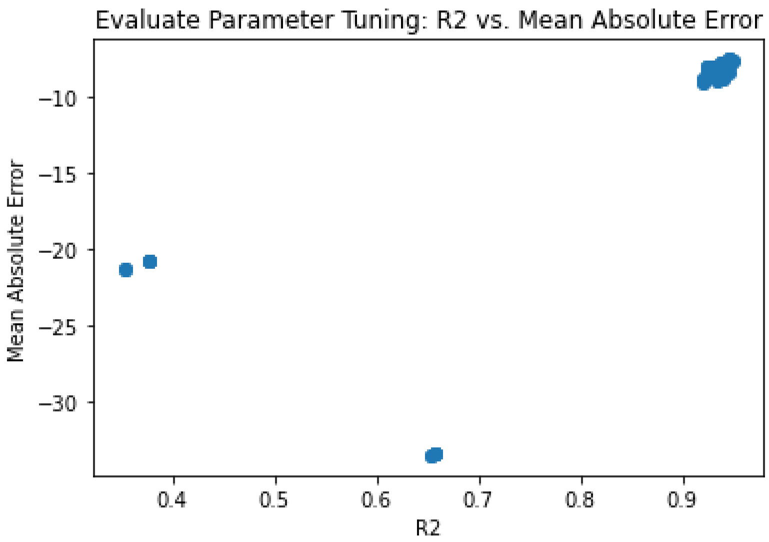


Fig. 24: Random Forest Regression Results (R-Squared versus Negative MAE)

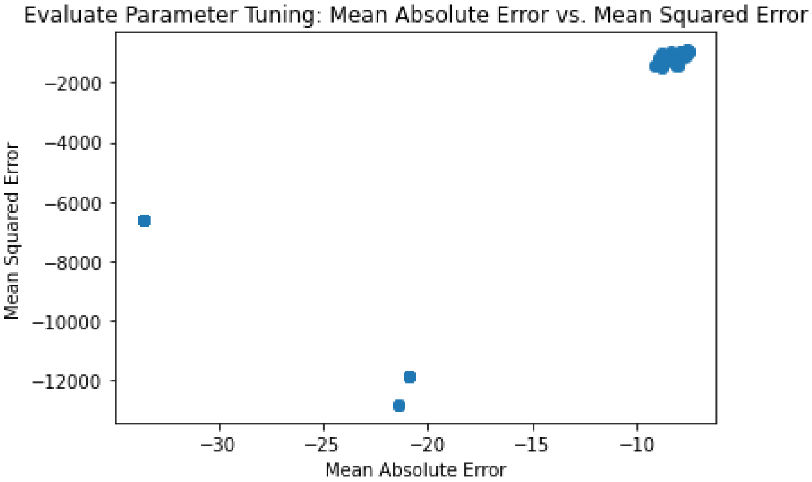


Fig. 25: Random Forest Regression Results (Negative MAE versus Negative MSE)

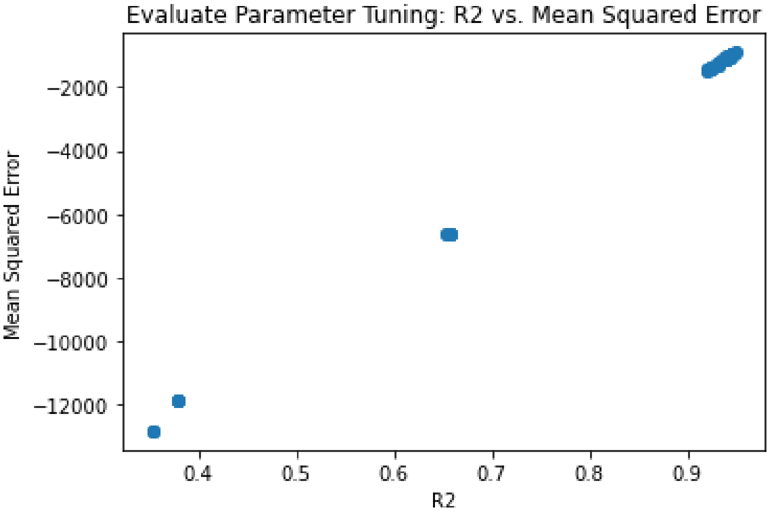
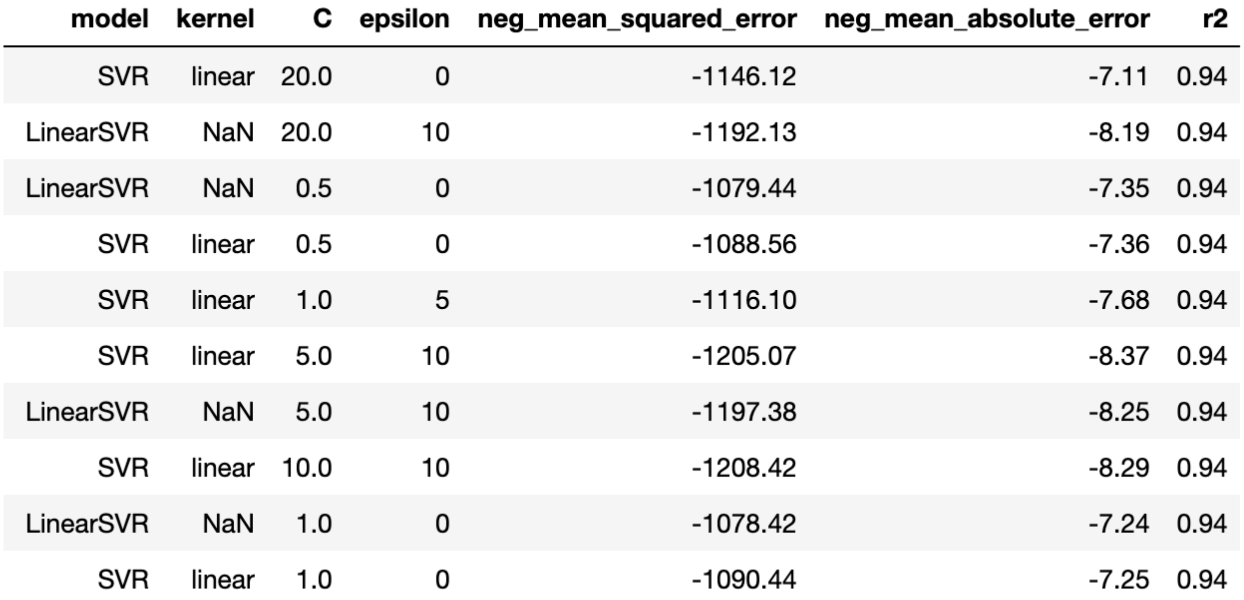


Fig. 26: Random Forest Regression Results (R2 versus Negative MSE)



model	kernel	C	epsilon	neg_mean_squared_error	neg_mean_absolute_error	r2
SVR	linear	20.0	0	-1146.12	-7.11	0.94
LinearSVR	NaN	20.0	10	-1192.13	-8.19	0.94
LinearSVR	NaN	0.5	0	-1079.44	-7.35	0.94
SVR	linear	0.5	0	-1088.56	-7.36	0.94
SVR	linear	1.0	5	-1116.10	-7.68	0.94
SVR	linear	5.0	10	-1205.07	-8.37	0.94
LinearSVR	NaN	5.0	10	-1197.38	-8.25	0.94
SVR	linear	10.0	10	-1208.42	-8.29	0.94
LinearSVR	NaN	1.0	0	-1078.42	-7.24	0.94
SVR	linear	1.0	0	-1090.44	-7.25	0.94

Fig. 27: SVR Results: Top 10 Model Selections

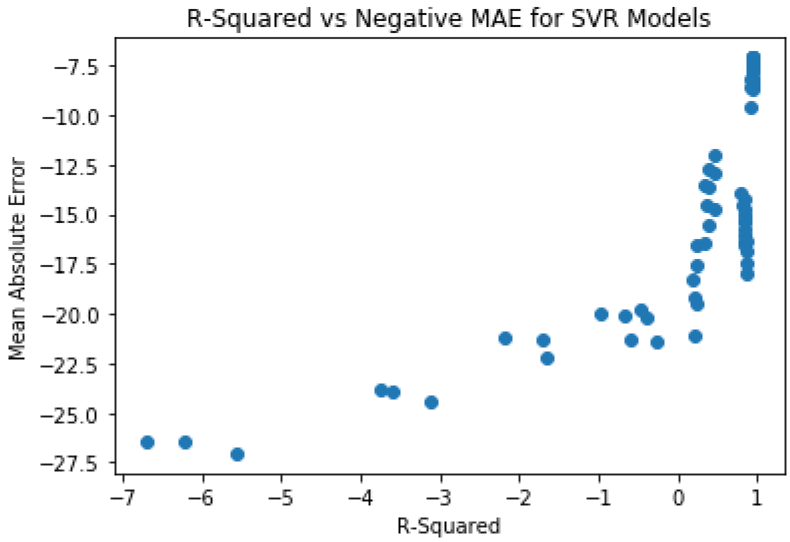


Fig. 28: SVR Results (R-Squared versus Negative MAE)

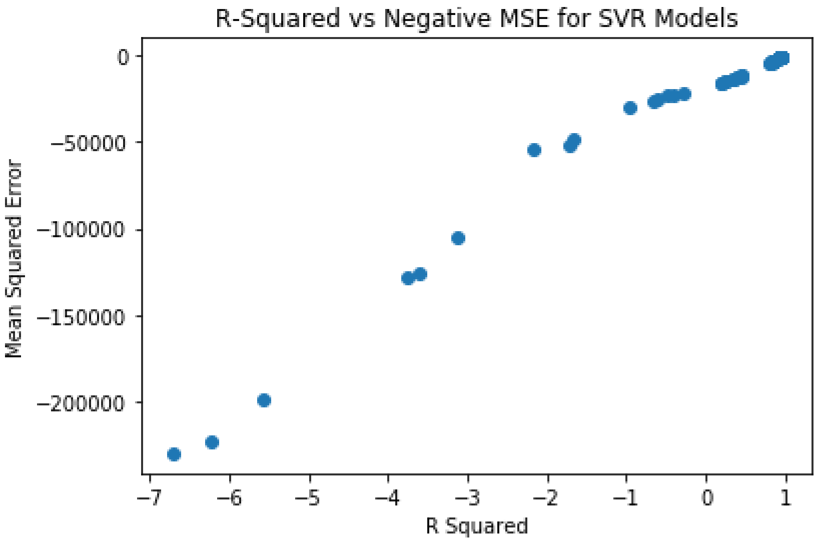


Fig. 29: SVR Results (R-Squared versus Negative MSE)

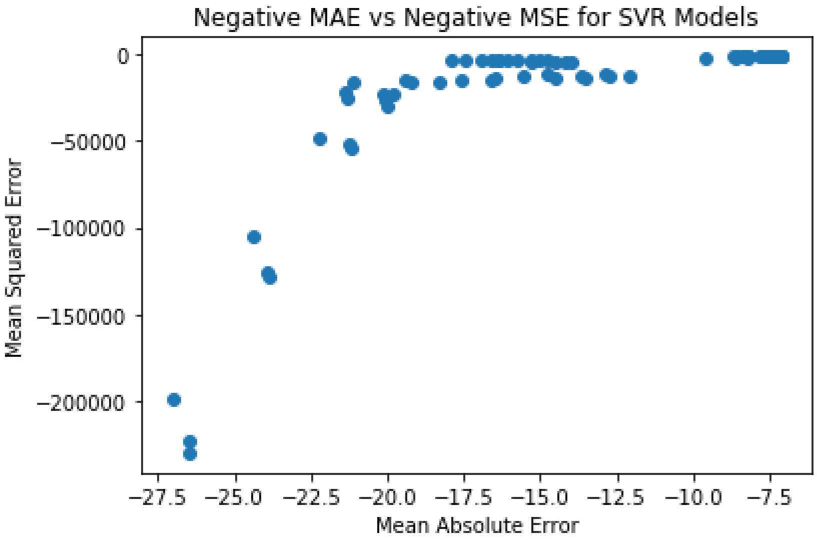
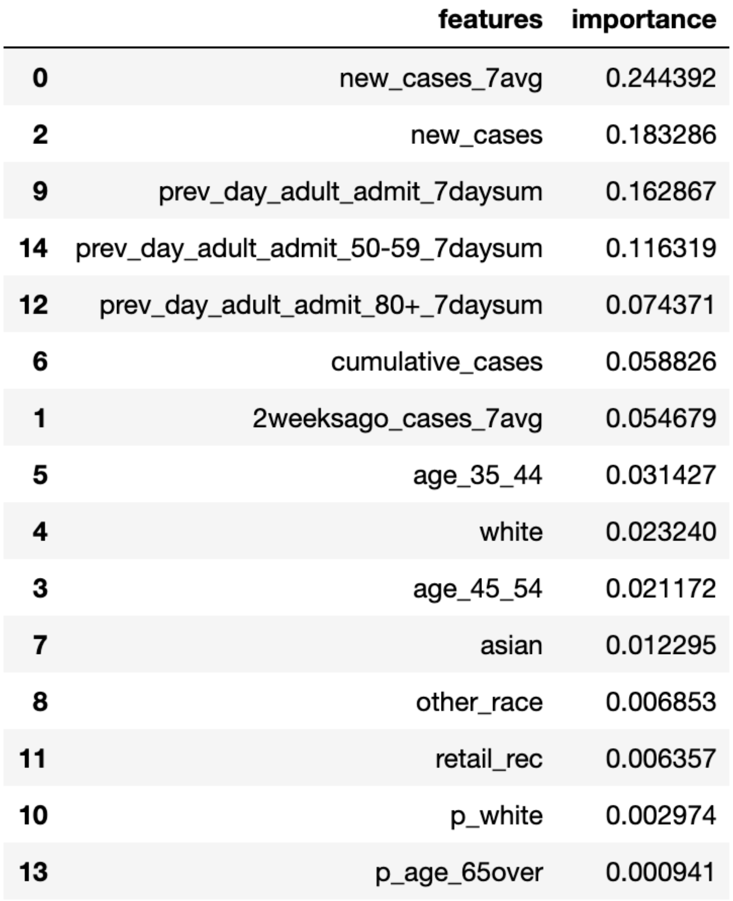


Fig. 30: SVR Results (Negative MAE versus Negative MSE)



	features	importance
0	new_cases_7avg	0.244392
2	new_cases	0.183286
9	prev_day_adult_admit_7daysum	0.162867
14	prev_day_adult_admit_50-59_7daysum	0.116319
12	prev_day_adult_admit_80+_7daysum	0.074371
6	cumulative_cases	0.058826
1	2weeksago_cases_7avg	0.054679
5	age_35_44	0.031427
4	white	0.023240
3	age_45_54	0.021172
7	asian	0.012295
8	other_race	0.006853
11	retail_rec	0.006357
10	p_white	0.002974
13	p_age_65over	0.000941

Fig. 31. Feature importance according to best performing model

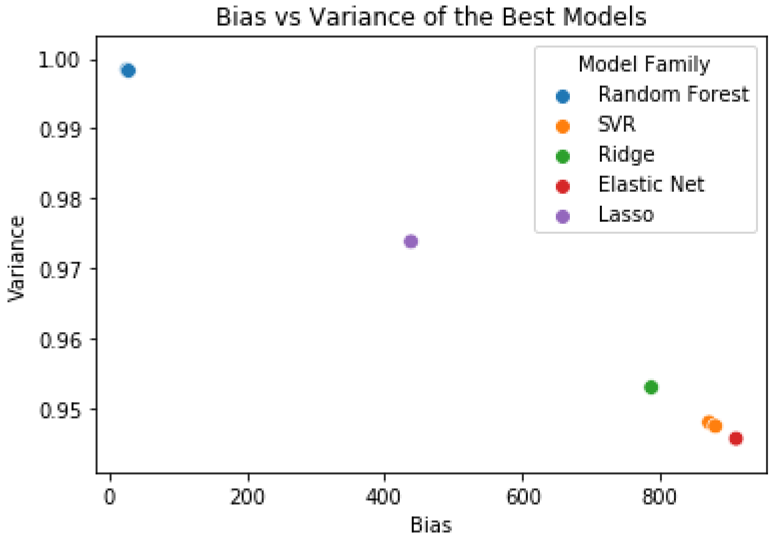


Fig. 32. Bias vs Variance of the Best Models

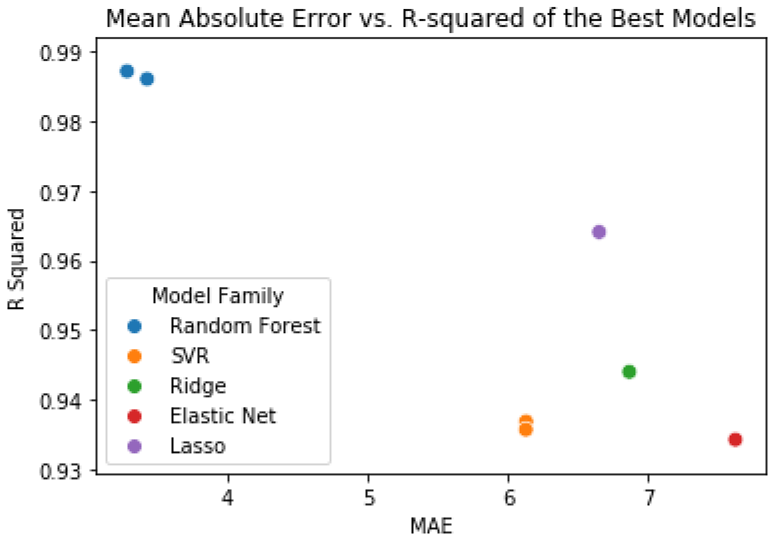


Fig. 33. MAE vs R-squared of the Best Models

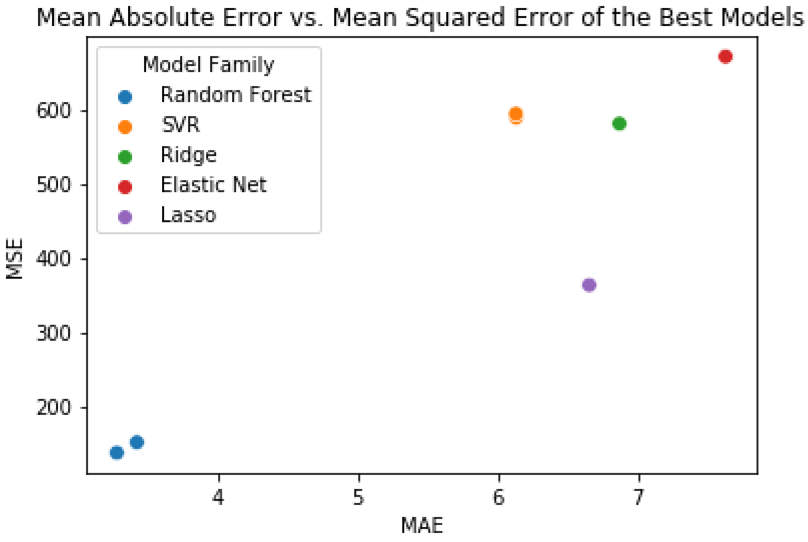


Fig. 34. MAE vs MSE of the Best Models

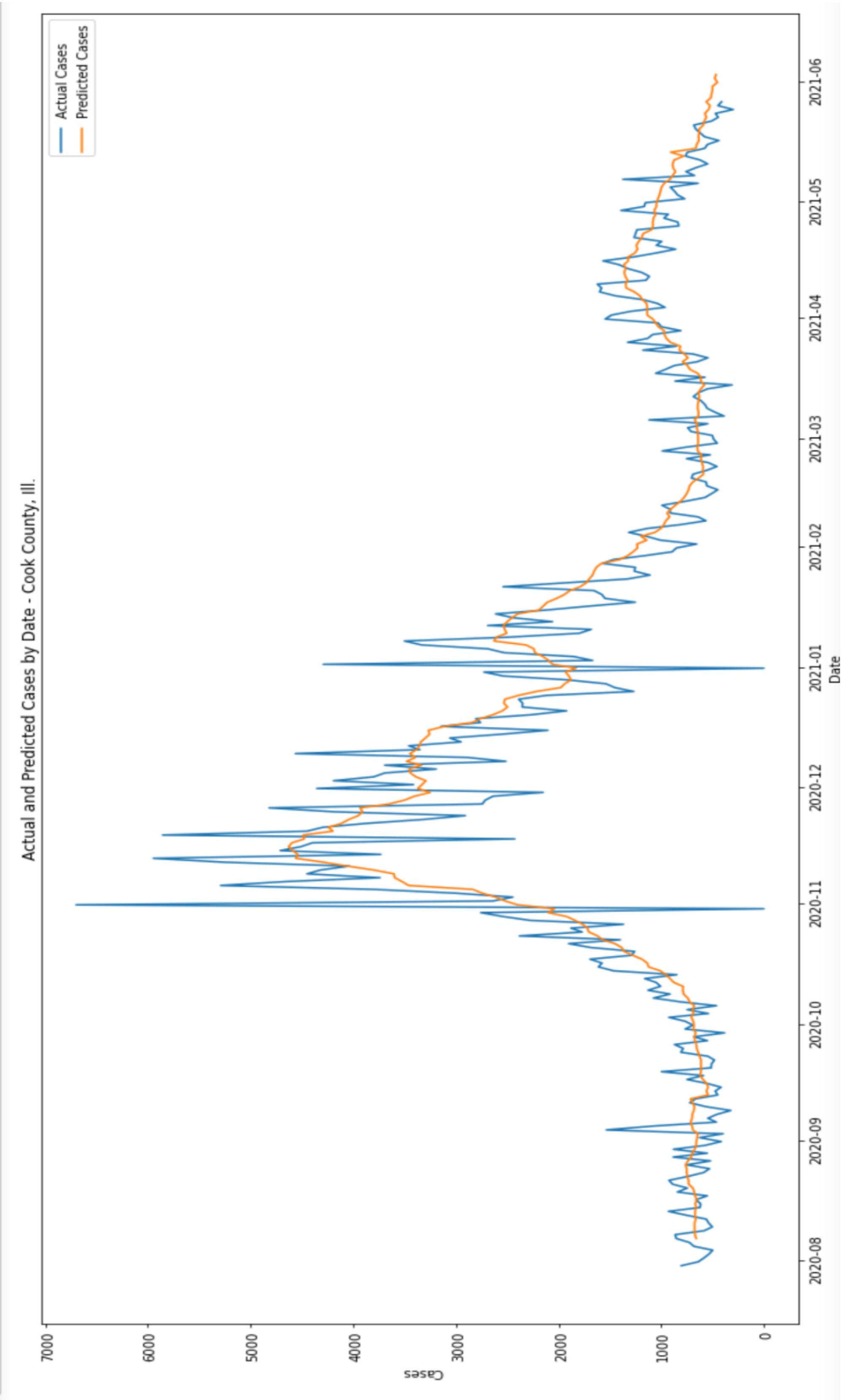


Fig 35: Actual vs Predicted Cases by Date - Cook County, IL

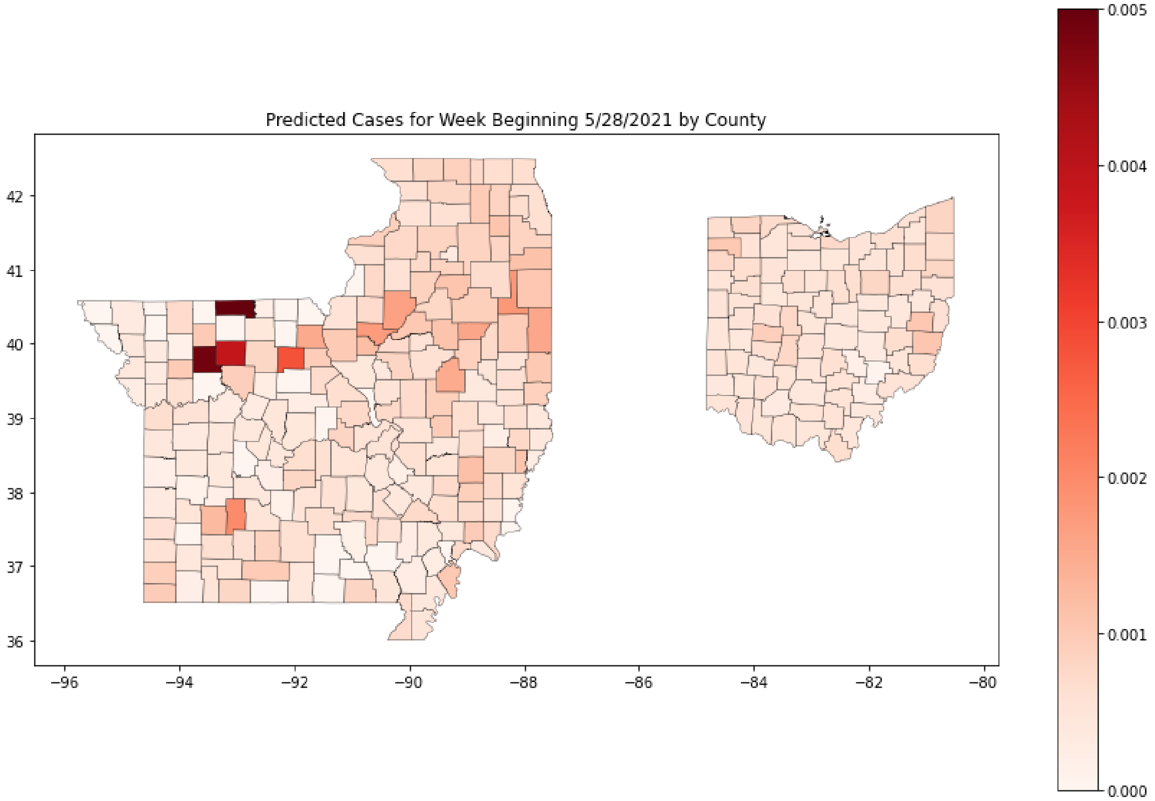
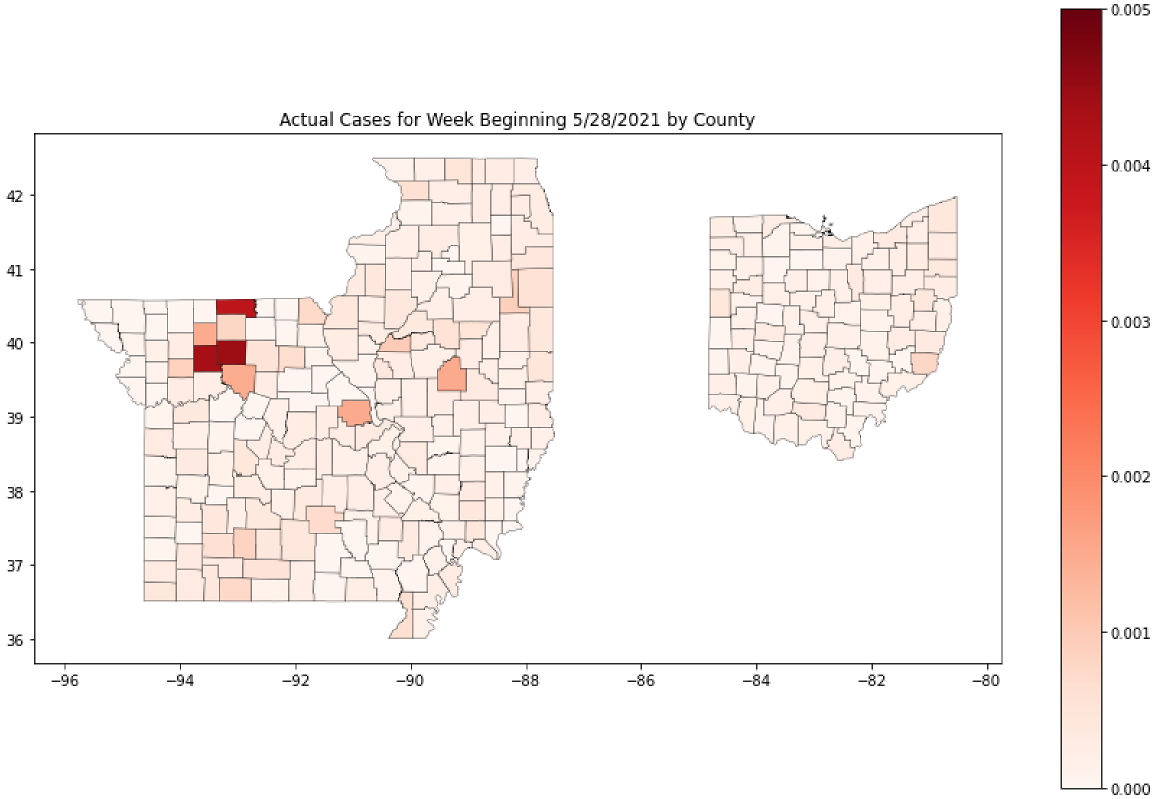


Fig 36: Actual vs Predicted Cases by Date - Cook County, IL


Actual Cases for Week Beginning 5/28/2021 by County
42
41
40
39
38
37
36
−96
−94
−92
−90
−88
−86
−84
−82
−80
0.005
0.004
0.003
0.002
0.001
0.000