Data Driven Prediction of COVID-19 Cases And Greedy Algorithm for Optimal Vaccine Allocation Among Los Angeles Neighborhoods

Rojin Bakthi
University of Southern California
Viterbi School of Engineering
bakhti@usc.edu

Jackie Dong
University of Southern California
Viterbi School of Engineering
jiaqid@usc.edu

Katherine Sing
University of Southern California
Viterbi School of Engineering
ksing@usc.edu

Kaushik Tandon
University of Southern California
Viterbi School of Engineering
kaushikt@usc.edu

Abstract

As the COVID-19 pandemic rages on, the public is eagerly awaiting a vaccine for the virus. While vaccine distribution is still a few months away, it is necessary to start preparing how to best allocate the vaccine. Production and distribution limitations make it imperative that greatest-risk neighborhoods are identified and accounted for when allocating vaccines so that it reaches those who would benefit the most. This paper proposes three different models for predicting the spread of COVID-19 cases among Los Angeles neighborhoods and a greedy algorithm to create an impactful vaccine allocation. We find an increase in model accuracy when using static and dynamic features rather than solely dynamic features.

1 Introduction

The novel coronavirus COVID-19 is a world threat. The World Health Organization (WHO) declared COVID-19 a global pandemic on March 11, 2020 [1], and as of October 22, 2020, there have been 41,538,416 cases and 1,135,018 deaths worldwide [2]. Much of the coronavirus is still to be understood, including infection-related questions such as how to best predict which areas will have the most cases at any point in time given the number of existing cases and region characteristics.

Further, with expectations of a vaccine in the future, it is important to know how many cases each area is projected to have so that a vaccination allocation to prevent the most future COVID-19 cases can be produced. If a limited number of vaccines is best allocated among high priority neighborhoods, the community at large will see the greatest possible decrease in future cases.

Vaccine creation and distribution is likely to face many limitations. Trials are being conducted for various versions of the vaccine, but vaccines are not expected to exit the testing phase until the end of 2020 or the beginning of 2021 [3]. Production constraints [4] and manufacturing capacity [5] restrict the amount of vaccine doses that will be available. There are concerns about distribution as local health departments have not been fully involved in the vaccine distribution planning process and are instead waiting for guidance from state officials [6].

This paper develops three different models to predict the number of positive COVID-19 cases one month in the future from the date of input data as well as a greedy algorithm to determine the best allocation of a limited number of vaccines, using data from Los Angeles neighborhoods as our case study. We show that Support Vector Regression with data such as monthly infections and static neighborhood features is a model well suited to pandemic spread, and that it outperforms other models. While our study is limited to the Los Angeles area and a hypothetical number of vaccines, the study is extremely relevant and can be expanded to other areas.

2 Related Work

There have been other research papers working on developing machine learning methods to predict the number of infected people for infectious diseases, particularly COVID-19. However, it appears that no previous research has taken our approach in predicting the effect of vaccine allocation on the number of new cases.

Our current model builds on the previous research done by Krishnamachari et al. (2020) [7]. This model is a Susceptible, Infected, Recovered (SIR) approach that computes the daily risk score for a single community given the number of daily infected cases, population of the community, inverse of average recovery dates, and number of times actual infected cases is higher than reported ones. We used this model in our project as a baseline to compare the predicted number of infected people and assess how well our model works, using the new features such as number of essential works and nursing home densities. Their approach was to use cumulative data from LA county number of COVID cases for each neighborhood, and combine it with the population data for each community to predict the value of R_t, which is a value representing the number of individuals who are infected per infectious individual at time t. They created a visual representation of the risk score and R_t for the LA county and its communities in which the areas with high-risk, low-risk, and no very-low risk were shown.

Debanjan Parbata and Monisha Chakraborty (2020) used Support Vector Regression (SVR) to predict the total number of deaths, recovered cases, cumulative number of confirmed cases and number of daily COVID-19 cases in India [8]. Using a cumulative dataset, they calculated the difference time series to calculate the daily number of new cases. They set the number of days as the independent variable and number of daily new cases as the dependent variable. The researchers indicated that the model performed better in predicting the cumulative cases than the daily number of cases. The researchers performed 5 tasks in the paper, including: predicted the spread of coronavirus across regions; analyzing growth rates and types of mitigation across countries; predicting the outcome of the epidemic; analyzing COVID-19 transmission rate; finding the correlation between COVID-19 and weather conditions. They used the SVR model to perform the first four tasks, and used Person's method for correlating the corona virus and weather conditions. Although our research was focused on the first task, this paper helped us understand how to understand SVR parameters and tune them for our model. However, our research differs from this paper since we have used different features.

3 Data

Our predictive models use a combination of static and dynamic data from a variety of sources. These datasets provide census and employment data, mobility data, and COVID-19 cases at the Los Angeles neighborhood granularity. We also have access to a shapefile that defines the boundaries of each Los Angeles neighborhood. Our goal is to compile a dataset that combines static as well as dynamic features, and use data for a certain month or set of months to predict the number of cases in a future month.

3.1 Static Data

We queried Social Explorer for the American Community Survey (ACS) 5-Year Estimates for the 2014-2018 year range as our main source of census data, with categories including total population, population density, average household size, and poverty status. This data is available for all United States census tracts - we converted relevant information to 262 Los Angeles neighborhoods [9].

To quantify essential workers in each neighborhood, we used the LEHD Origin-Destination

Employment Statistics (LODES) dataset provided by the United States Census Bureau for employment data organized by field [10].

To calculate nursing home density within the neighborhoods, we obtained nursing home data from Homeland Infrastructure Foundation-Level Data (HIFLD) [11].

3.1.1 Preprocessing of Static Data

The Social Explorer census data was available at the census tract level, a relatively small section of a county. In Los Angeles County, there are 2348 census tracts. As the COVID-19 case data is available for each neighborhood in Los Angeles County, we assigned each census tract to the corresponding neighborhood and aggregated all the different columns in the dataset. For each census tract, we found the centroid latitude and longitude point and determined which neighborhood contained that point. The boundaries for each neighborhood were defined by the Los Angeles Times in a shapefile and is the same that is used in the SIR-model. The aggregation provided raw totals that were normalized to percentages. Similar logic was applied with the LODES dataset to aggregate employment statistics.

With the HIFLD nursing home data, each row provided the name, address, latitude and longitude coordinate pair, and number of beds for a nursing home. We used the coordinates to determine which Los Angeles neighborhood the nursing home is located in. For each neighborhood, we accumulated the number of nursing home beds, recording that as a percentage of the overall nursing home bed capacity.

3.2 Dynamic Data

Data for COVID-19 cases across LA County regions is provided in the github repository for the SIR model by Krishnamachari et al. [7]. The repository collects and scrapes COVID-19 data from LA Public Health Department website daily. The provided CSV file contains a cumulative count of COVID-19 cases by day for each city or region in LA county.

We also use mobility patterns from SafeGraph in our models. This dataset provides place traffic and demographic aggregations [12] that explores where people travel to and from. For each point of interest, we track which census blocks each visitor lives in. As this is provided each day, we can track movements within each community.

3.2.1 Preprocessing of Dynamic Data

3.2.1.1 Smoothing COVID Data

While the cumulative COVID-19 case data was provided in the github repository for the SIR model by Krishnamachari et al. [7], there were certain unexpected drops in the total number of cases for each neighborhood. We believe this is either incorrectly reported data or bad scraping, as the case data would return to the correct level within a few days. To account for this, if a negative change in total cases occurs, we assume that no new COVID-19 cases occurred for that neighborhood on that day. On average, this affected under 3 days of the 7 months of COVID-19 data for each neighborhood.

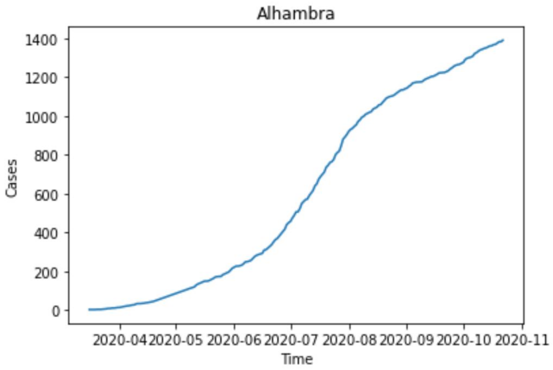


Figure 1: Example of Total Cumulative COVID-19 Cases for a LA neighborhood

We then converted the cumulative totals to a relative day by change, or the number of new cases each day. An example can be seen in Figure 2.

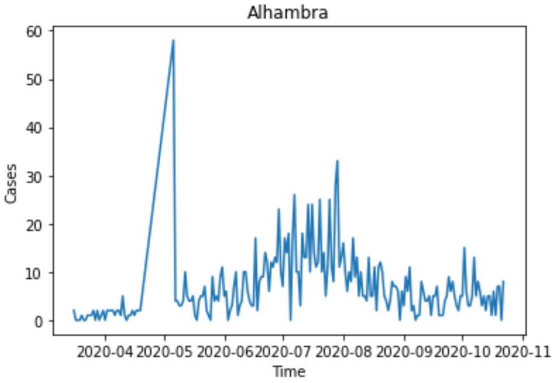


Figure 2: Example of Daily COVID-19 Cases for a LA neighborhood

3.2.1.2 Inconsistencies in region names

The LA county shapefile and COVID-19 dataset have inconsistent names for some of the regions. To address these inconsistencies, we first obtained a list of names which appear in COVID-19 dataset but not in LA county shapefile. Then, given the latitude and longitude of these missing regions in COVID-19 dataset, we cross-checked where these regions lie within the LA county shapefile. Having acquired a mapping between the missing region name from COVID-19 dataset and its corresponding region in LA county dataset, we manually located both regions on Google Map to see if the mapping is accurate. We use the accurate mappings to align the region names in COVID-19 dataset to LA county shapefile.

3.2.1.3 Obtaining list of regions whose neighbors have COVID-19 data

As we use the total number of cases within a region's neighbors as a feature, we ensured that we only consider regions whose neighbors do not have missing COVID-19 data. We first identified regions which do not have COVID-19 data, and removed these regions as well as their neighbors.

Next, we eliminated all the regions which were on the border of Los Angeles county, as these regions will have some neighbors located outside of the county. We built a polygon bigger than Los Angeles county that covered the entire region. Then, we subtracted the county from the covering polygon to obtain the outline of the county. Finally, we located and removed regions which were touching the covering polygon as bordering regions.

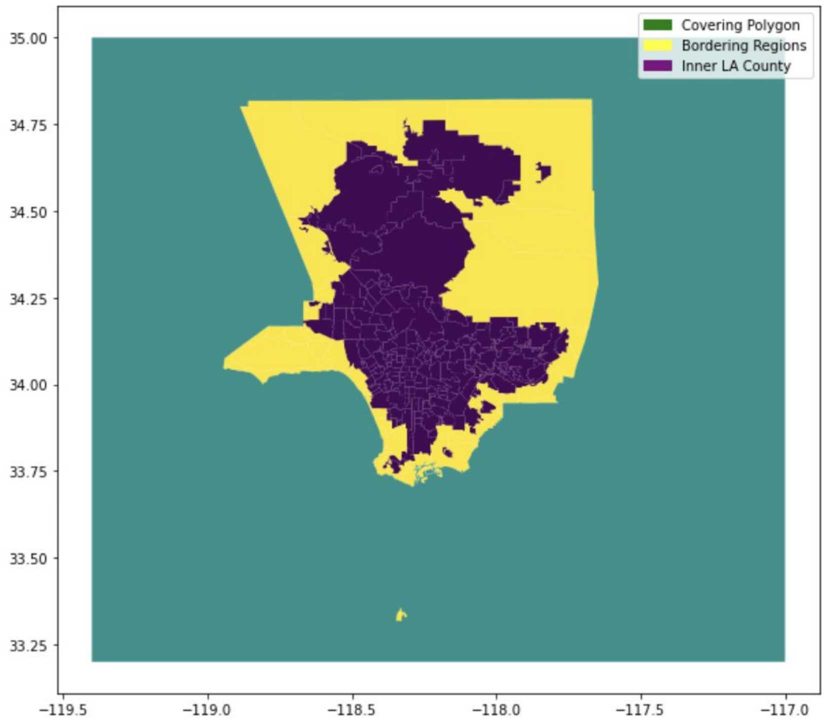


Figure 3: Bordering regions of Los Angeles County

These two steps bring down the number of regions from 272 to 52. After we acquired a list of valid regions, we computed the total number of cases in the neighbors areas of each region.

3.2.2 SafeGraph Mobility Data

The processing and specific use of SafeGraph mobility data in our models is still a work in progress and will be explained in more detail in a future version of this report. This data is provided at census block granularity and will be aggregated into Los Angeles neighborhoods. We will track inter-community and intra-community movements over different time periods to evaluate if this provides an impact on the accuracy of COVID-19 case predictions.

3.2.3 Combining Dynamic and Static Features

We combined dynamic and static features together to build our final dataset. For dynamic features, we computed monthly cumulative COVID-19 cases for each region, number of cases in the neighboring areas of each region, and total number of cases in the entire LA county. For static features, we have all the census data converted to percentages as described above. The target value is the number of cases in the next month.

4 COVID-19 Cases Predictive Model

We used linear regression, support vector regression (SVR), and gradient boosting regression (GBR). The models displayed varied performance, but a 20% increase in r^2 can be seen in all models when static features are considered in combination with dynamic features, as opposed to considering only dynamic features.

4.1 Time Based Cross Validation

Time based cross validation is used to split the data into train and test sets, where we train the model on a particular time interval and test it on the following time interval. Two different approaches are employed to split the data.

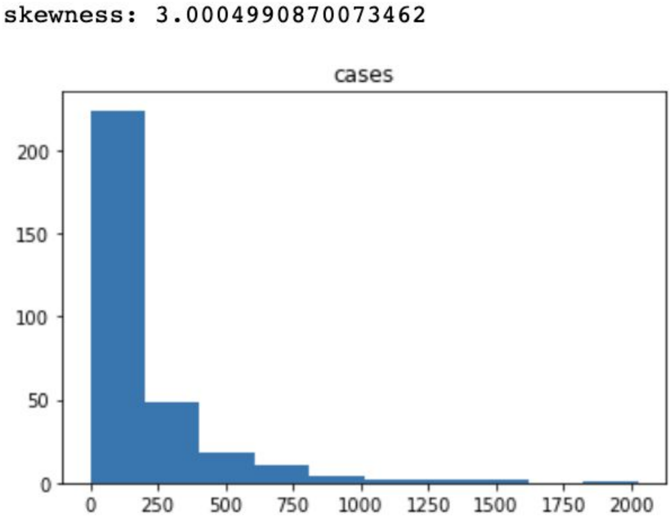
Given COVID-19 time series case data from April to August, the one split approach uses data from April to July as train set, and data in August as test set. The multiple splits approach divides the data into subintervals of duration of three consecutive months. We train on data from April and May and test on data from June, train from May and June and test in July, and finally train from June and July and test in August.

4.2 Including Static Features

We employed two different methods to train the model. We train one set of models using only dynamic features, and train another set of models using dynamic features as well as static features. The purpose is to analyze whether there is an improvement in model performance after static features are taken into consideration.

4.3 Data Normalization

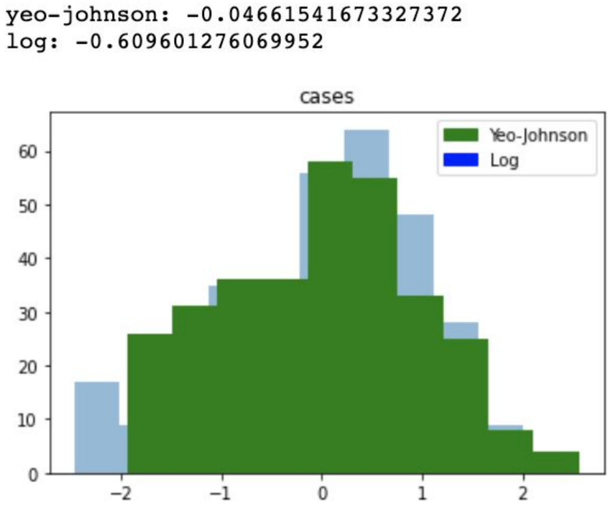
We normalized the training data in order to improve performance of the machine learning models. In particular, machine learning models such as SVR require zero mean and unit variance. Furthermore, some features in our model are either left skewed or right skewed, as shown in the following figure.



```
skewness: 3.0004990870073462
```

Figure 4: Right Skewed Feature

The ‘cases’ feature is right skewed. The skewness is computed using the Fisher-Pearson coefficient. Skewness of 0 means the data is not skewed. Skewness smaller than 0 means left skewed data, and skewness larger than 0 means right skewed data. We applied log transformation and Yeo-Johnson power transformation and compared the results.



```
yeo-johnson: -0.04661541673327372
log: -0.609601276069952
```

Figure 5: Log Transformation and Yeo-Johnson Transformation

Yeo-Johnson transformation performs better and reduces skewness to -0.047. In general, Yeo-Johnson transformation is better at reducing skewness. Therefore, after splitting the data into train test sets, we applied Yeo-Johnson transformation to all skewed features in train sets and standardized the features so that they have 0 mean and unit variance. Predictions from our models greatly improved after data is normalized.

4.4 Models

For each machine learning model, we trained and tested on four different sets of data:

1. data without static features, with one split validation
2. data with static features, with one split validation
3. data without static features, with multiple splits validation
4. data with static features, with multiple splits validation

We used local search to locate the optimal hyperparameters for all models.

4.5 Results

Out of the three machine learning models which we used, linear regression and SVR are our best performing models. We show the results from SVR models in greater detail. Results for all models can be found in section 4.4.2.

4.5.1 SVR

For each configuration of the dataset, we graph the predicted values against measured values. The dashed line in each graph represents perfect predictions. We also present a table which contains mean squared error (MSE), r squared value (r^2), mean absolute error (MAE), and average measured value (avg).

4.5.1.1 One Split Without Static Features

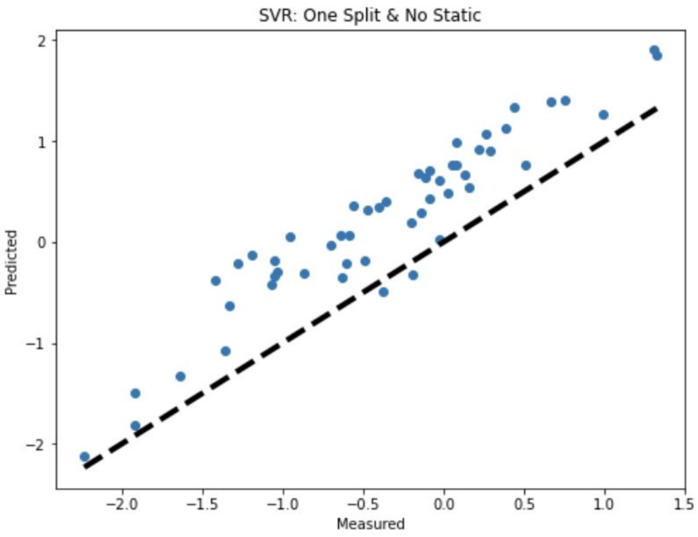


Figure 6: SVR, One Split Without Static Features

Metric	Value
MSE	0.424
r^2	0.402
MAE	0.596
avg	-0.374

Table 1: Metrics for Figure 6

4.5.1.2 One Split With Static Features

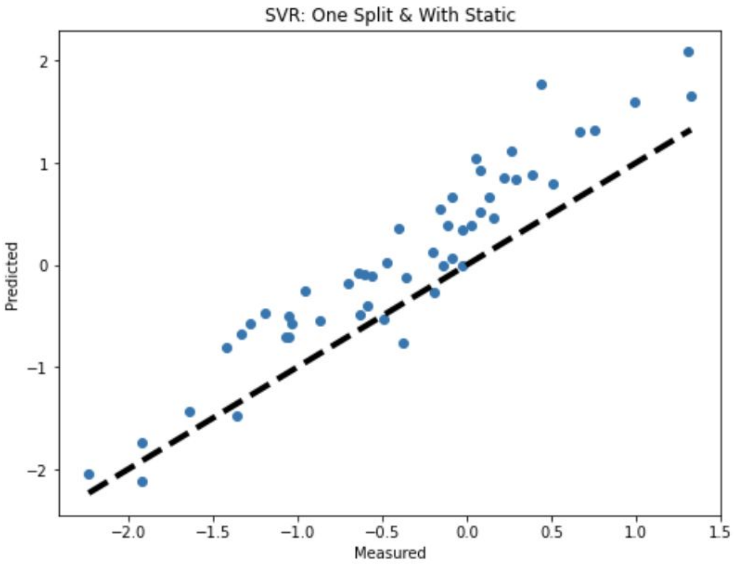


Figure 7: SVR, One Split With Static Features

Metric	Value
MSE	0.286
r^2	0.671

MAE	0.466
avg	-0.374

Table 2: Metrics for Figure 7

From table 1 and 2, we see that through incorporating static features, we improved r^2 value from 0434 to 0.645. The same increase can be seen in all other models as well. This demonstrates that static features indeed play a role in affecting COVID-19 cases.

4.5.1.3 Multiple Splits Without Static Features

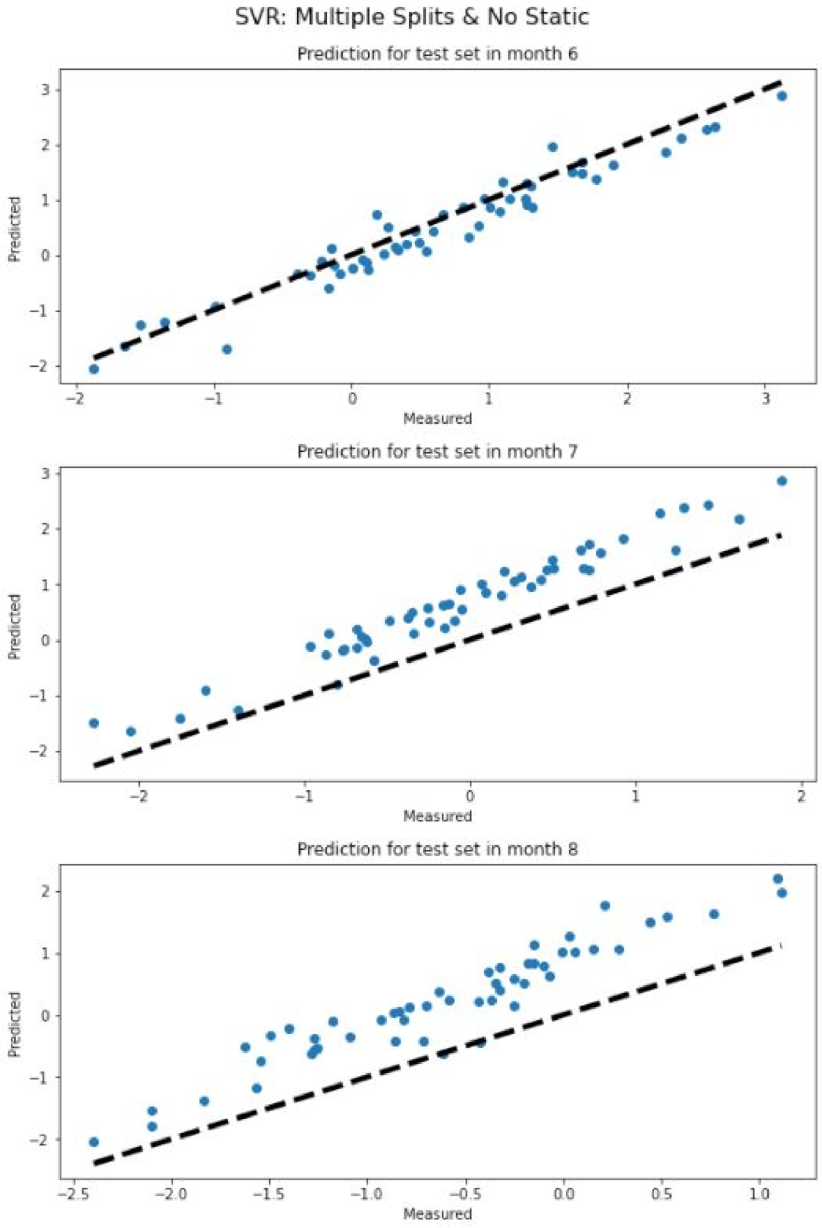


Figure 8: SVR, Multiple Splits Without Static Features

Metric	**Month 5**	**Month 6**	**Month 7**
MSE	0.083	0.551	0.744

r^2	0.926	0.470	0.141
MAE	0.238	0.702	0.808
avg	0.623	-0.080	-0.597

Table 3: Metrics for Figure 8

Note that “prediction for test set in month 6” means that we trained on data for month 4 and 5, then tested on data whose features are from month 6, and target values are from month 7.

4.5.1.4 Multiple Splits With Static Features

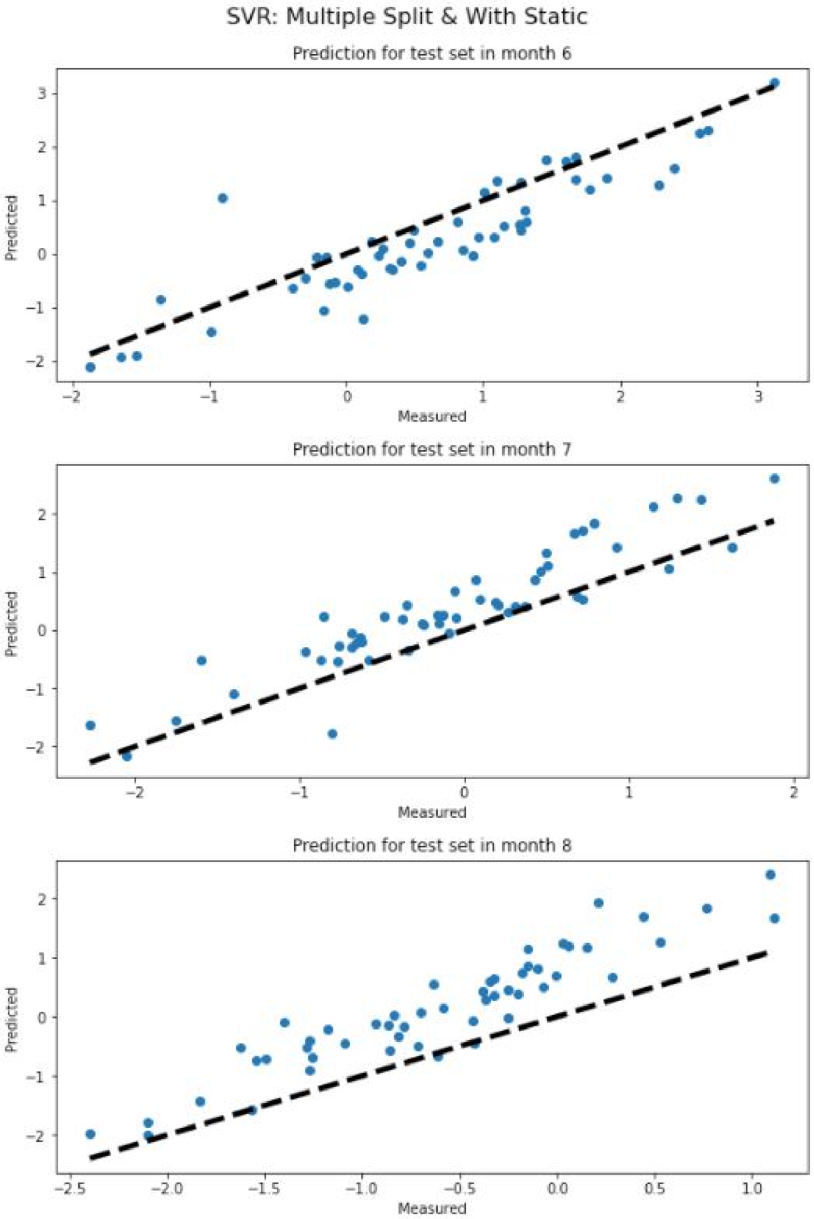


Figure 9: SVR, Multiple Splits With Static Features

Metric	**Month 5**	**Month 6**	**Month 7**
MSE	0.348	0.341	0.673

r^2	0.713	0.669	0.303
MAE	0.477	0.493	0.732
avg	0.623	-0.080	-0.597

Table 4: Metrics for Figure 9

From table 3 and 4, we see that our model performs extremely well on prediction for earlier months but decreases for later months. This is likely because the peak of COVID-19 in LA county hit during mid July to mid August, but cases gradually reduced after August. Using data at the peak of the virus to predict cases for months after the peak could potentially lead to the model consistently predicting higher cases. To increase model performance for later months, we need to take into account more dynamic features, such as number of elderly population affected or migration data, which also varied considerably during the peak and after the peak to analyze how those features impact the change in cases.

4.5.2 Metrics for All Models

Model	**Linear Regression**	**SVR**	**GBR**
One split, no static feat	0.409	0.402	0.382
One split, with static feat	0.645	0.671	0.581
Multiple splits, no static feat	0.928, 0.505, 0.101	0.926, 0.470, 0.141	0.876, 0.516, -0.0338
Multiple splits, with static feat	0.778, 0.606, 0.363	0.713, 0.669, 0.303	0.831, 0.643, 0.054

Table 5: r^2 for all models

Model	**Linear Regression**	**SVR**	**GBR**
One split, no static feat	0.433	0.424	0.447
One split, with static feat	0.337	0.286	0.391
Multiple splits, no static feat	0.089, 0.528, 0.699	0.083, 0.551, 0.744	0.140, 0.491, 0.651
Multiple splits, with static feat	0.269, 0.438, 0.621	0.348, 0.341, 0.673	0.174, 0.405, 0.627

Table 6: Mean Squared Error for all models

Model	**Linear Regression**	**SVR**	**GBR**
One split, no static feat	0.610	0.596	0.592
One split, with static feat	0.500	0.466	0.540
Multiple splits, no static feat	0.229, 0.684, 0.788	0.238, 0.702, 0.808	0.291, 0.646, 0.764
Multiple splits, with static feat	0.433, 0.583, 0.705	0.477, 0.493, 0.732	0.330, 0.576, 0.744

Table 7: Mean Absolute Error for all models

4.5.3 Feature Importance

The GBR model computes feature importances based on impurity scores. We showed the top 10 most important features in the table below, computed from "one split with static data".

Feature	Importance
COVID-19 Cases	0.851
County total cases	0.069
% Households: Male Householder, No Wife Present	0.021
% Population for Whom Poverty Status Is Determined: 1.50 to 1.99	0.012
% Total Population: 18 to 24 Years	0.008
% Households: 5-Person Household	0.006
% Households: 3-Person Household	0.005
% Workers 16 Years and Over: Did Not Work At Home	0.004
% Workers 16 Years and Over: Worked At Home	0.003
% Households: No Wage or Salary Income	0.003

Table 8: Top 10 Most Important Features

We can see from this table that important features include the number of cases, household status and household population, employment status, and poverty level. These features are among those which are argued to be the most important factors that affect spread of COVID-19. It shows that our model and methodologies are promising. With inclusion and processing of more features, we believe that our model can further improve and show more insights on factors affecting COVID-19 cases.

5 Planned Tasks

5.1 Extensions to Predictive Model

We plan to extend our predictive models and improve the overall results in a few ways. First, we would like to transform the SIR model constructed by Krishnamachari et al. [7]. into a prediction for the number of cases and use it as a baseline for our predictions. We also plan to include several new dynamic features, such as SafeGraph's mobility data, to take into consideration other possible factors which could impact the spread of COVID-19. This feature would allow us to better understand daily spread within and between various neighborhoods. Finally, we would like to increase the number of regions that we are considering through filling missing data and try training the models at different time frames, such as in weeks or days.

5.2 Greedy Algorithm for Vaccine Allocation

We aim to create a greedy algorithm to predict the best vaccine allocation distribution. We assume a budget of 10,000 vaccine units and provide them in blocks of 1000 to each region in LA county. Using the predicted number of cases for each neighborhood generated by the regression model, the algorithm will allocate 1000 vaccines to each neighborhood and calculate the change in the number of infected people before and after vaccine allocation. The algorithm will optimize the vaccine allocation in a way that minimizes COVID-19 spread among LA-county neighborhoods. We will create spatial distribution maps to visualize the number of predicted COVID-19 cases before and after vaccine allocation.

This section will be expanded upon in a future version of this report.

Acknowledgments

We would like to thank Professor Bistra Dilkina, Prathik Rao, and the city of Los Angeles Data Science Team for providing helpful resources as well as feedback and advice on what to

potentially explore in our project.

References

[1] J. Ducharme, “World Health Organization Declares COVID-19 a 'Pandemic.' Here's What That Means,” Time, 11-Mar-2020. [Online]. Available: https://time.com/5791661/who-coronavirus-pandemic-declaration/. [Accessed: 18-Sep-2020].

[2] John Hopkins University, “COVID-19 Map,” Johns Hopkins Coronavirus Resource Center. [Online]. Available: https://coronavirus.jhu.edu/map.html. [Accessed: 22-Oct-2020].

[3] M. Peiris and G. M. Leung, “What can we expect from first-generation COVID-19 vaccines?,” The Lancet, Sep. 2020.

[4] R. Khamsi, “If a coronavirus vaccine arrives, can the world make enough?,” Nature News, 09-Apr-2020. [Online]. Available: https://www.nature.com/articles/d41586-020-01063-8. [Accessed: 22-Oct-2020].

[5] S. Koch, “Adding up manufacturing capacity for COVID-19 vaccines,” BioCentury, 03-Jun-2020. [Online]. Available: https://www.biocentury.com/article/305365/adding-up-manufacturing-capacity-for-covid-19-vaccines. [Accessed: 18-Sep-2020].

[6] G. Galvin, “As CDC’s Deadline for Submitting COVID-19 Vaccine Distribution Plans Nears, States Say They’re Still Short on Guidance,” Morning Consult, 09-Oct-2020. [Online] Available: https://morningconsult.com/2020/10/09/covid-19-vaccines-state-distribution-plans/. [Accessed: 19-Oct-2020]

[7] B. Krishnamachari, 2020 COVID-19 Risk Estimation for L.A. County using a Bayesian Time-varying SIR-model. https://github.com/ANRGUSC/covid19_risk_estimation

[8] D. Parbat and M. Chakraborty, A python based support vector regression model for prediction of COVID19 cases in India. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7261465/

[9] Social Explorer. Census Tract Data [Data file]. Retrieved from: https://www.socialexplorer.com/tables/ACS2017_5yr/R12411264

[10] LEHD Origin-Destination Employment Statistics 2017 Employment Data for Los Angeles Country [Data file]. Retrieved from: https://lehd.ces.census.gov/data/lodes/LODES7/ca/rac/

[11] Nursing Home Densities from Homeland Infrastructure Foundation-Level Data. Retrieved from: https://hifld-geoplatform.opendata.arcgis.com/datasets/nursing-homes/data

[12] SafeGraph Data for Neighborhood Patterns. Retrieved from: https://www.safegraph.com/neighborhood-patterns