Eric Wang

Ali Vakilimafakheri

Mac Kul

COVID-19 Project Summary Report

**Abstract**

COVID-19 is a serious disease that has affected the lives of billions around the world. Due to the severity of the disease, our team was tasked to build and train a time series model in order to forecast the future of the pandemic. We utilized both linear regression and support vector regression to train models for each county separately because we do not believe that there exists a one-size-fits-all model, even if all the counties in our dataset are part of California. Our team discovered that preference of a model is highly dependent on the data. For certain counties, a linear regression model performed quite well on testing data, while a support vector regression model failed to produce similar results一 and vice versa. In our report, we delve deeper and go over our data preparation, methodology, and conclusions.

**The Procedure**

**Data and Data Preprocessing**

Our team decided to pick five features we believed were relevant for predicting covid cases to incorporate into our dataset: percentage of hospital admissions, percentage of doctor visits, percentage of outpatients, number of restaurant visits, and percentage of people with COVID-like illness. Our intuition suggested that these features were highly correlated with the number of covid cases and would serve as great predictors. However, there was one challenge that we faced一 making sure each feature had enough counties, dates, and data in the Covidcast API. We sorted through multiple features and ran diagnostic checks by plotting the coverage of each feature for each county in California over the span of 7 months. All but one of our features, percentage of hospital admissions, had an abundance of data for each county. We used said feature as a bottleneck because it contained the least amount of data and chose only the counties where their coverage was at least 85%. As for our ground truth values, we opted to use the proportion of covid cases rather than the number of covid cases because we did not include the past number of covid cases as a feature. For each county, the missing values for that specific county were imputed using the average of the values for that specific county for each of the signals used in our data. With the subsetted counties and finalized features, we moved on to preprocessing our data. We wanted our model to use the features from one and two days prior to predict the current number of cases, so we shifted each of the features to achieve this一 in effect, we would be predicting the future. After aggregating the t-1 and t-2 features with the current ground truths, we proceeded with our training phase.

**Methodology**

For our interpretable model, our team decided to use linear regression rather than a decision tree regressor after running multiple training tests. Even after tuning the hyperparameters on our decision tree, the predictions were too simple, which we believe was the case because the tree picked up on a trend for a particular feature. For our linear regression model, we started off by standardizing and normalizing the data so that large feature values do not disproportionately affect the regression line. With the 18 counties in our dataset, we trained a linear regression model for each county. Our decision to do this came from the fact that we did not believe one model would generalize all the counties at once very well. An argument can be made that because these counties are from the same state, their trends would more or less be similar; however, due to the gravity of the situation that is COVID-19, we did not want to take a risk on the aforementioned assumption. We used a normal train test split rather than a time series split because the data already uses the features from the previous two days to predict the current ground truth value. We then used linear regression to fit the data from each county, printed out their MSEs (mean squared errors), and plotted their predictions vs. ground truth. For support vector regression, we utilized the same train test split with standardized data. After some experimentation, we decided to use a radial basis kernel (rbf) and sigmoid kernel with a regularization parameter in the range (1, 5, 0.1) (*with the third value representing the increment*), and a polynomial kernel with degrees in the range (2, 10, 1) , independent coefficient in the range (0, 1, 0.1), and a regularization parameter in the range (1, 5, 0.1) for our hyperparameters. We used a 5-fold cross validation in order to find the optimal hyperparameters for each model, printed the MSEs, and plotted their predictions vs. the ground truth values.

**Results and Conclusions**

Our team found that some counties were better explained by linear regression compared to support vector regression一 and vice versa. For example, county 6001 had an MSE value of 26.73 with the best linear regression model and an MSE value of 41.12 with its best support vector regression model (rbf kernel with regularization parameter = 4.9), and county 6019 had an MSE value of 384 with its best linear regression model and an MSE value of 300 with its best support vector regression model (polynomial kernel with degree = 3, independent coefficient = 0.9, and regularization parameter = 4.9).We assume that this phenomenon occurred due to the data used for each county and due to more features being required for the counties. In the case of county 6001, the ground truth values do not spike very radically or often and the training and testing data followed similar trends and values. Due to this, and the simplistic nature of linear regression, we are inclined to believe the data is a major reason for why the linear regression model performed quite well. Meanwhile, the support vector regression model may have picked up on a more complex trend and incorrectly predicted a result, or another feature(s) is needed to account for the spikes in ground truth values for that particular county. Similar logic may be extended to county 6019, where the trends and values in the training and testing data were **not** similar and more complex, which resulted in a **better** performance for the support vector regression model.

With these results and assumptions in mind, our team concluded that it is **possible** to build predictive models to forecast the pandemic, but only for **similar** regions (because their important features may be similar as well). We strongly believe that non-similar regions require different/more features to accurately predict the proportion of covid cases due to cultural/social differences, so a one-size-fits-all model is a bit naive. Although our results have shown that the choice of region may affect whether a linear regression model or support vector regression model is better, we think that, generally, **a support vector regression model should be used to forecast the pandemic**. This is because the viral nature of COVID-19 probably means that most data will have more spikes, and so a complex model (support vector regression) may be better suited. Additionally, our intuition suggests that support vector regression is a better version of linear regression due to how well they can handle multi-dimensional data and there is always room for improvement with different choices of hyperparameters or hyperparameter ranges.

We illustrate the importance of different features in different regions with our interpretable model, linear regression. For county 6001, the weights of percentage of hospital admissions (t-2) & (t-1), percentage of doctor visits (t-2) & (t-1), percentage outpatients (t-2) & (t-1), number of restaurant visits (t-2) & (t-1), and percentage of people with COVID-19 like symptoms (t-2) & (t-1) are: -0.57, 2.46, -5.34, 3.26, 1.37, 3.44, -0.98, -0.44, -0.72, and 2.12, respectively. Because the proportion of covid cases is generally between 0 to 100, the weights are actually quite good for each of the features. As for county 6019, the weights of the same features are: 4.63, -1.86, -3.14, 2.32, 29.99, -23.01, 0.85, -1.16, 5.82, and -0.48, respectively. The weights for the features here are quite good as well, with the percentage of outpatients seemingly having a much larger impact on the proportion of covid cases. As demonstrated above, the significance of each feature seems to rely heavily on the region at hand. Features from one region may be great predictors of COVID-19 but not so great in others, which tends to drastically affect model performance. Although we have seen differing significances for each feature in different counties, they tended to remain quite important for forecasting the covid cases. Thus, we believe that policies should take into account our aforementioned features in order to be successful. For example, if policymakers notice an increase in some features, strict guidelines should be enforced as soon as possible in order to halt the spread of COVID-19, or else hundreds of thousands of lives are at risk.

*It may seem counterintuitive that a feature, hospital admission for example, has both a negative and positive weight depending on the feature day (t-2 or t-1), but that may be the result of multicollinearity一 a problem in linear regression where the presence of one variable may lead to less statistically significant independent variables, often due to collinearity between multiple variables.*