Tyler Chia, Joanne Kim, Michael La, Joshua Harasaki

Forecasting Short-Term Future COVID-19 Cases Based on Historical Data

The onslaught of COVID-19 has proved to be a detrimental event in society and has restricted much of our everyday lives. As data scientists, we aimed to utilize past data to predict the number of COVID cases using machine learning. This is extremely valuable information to not only estimate the number of resources (i.e. hospital beds, COVID test kits, other supplies) that we need to combat the virus, but also to find the most influential factors that determine the number of COVID cases. With this goal in mind, we created two time-series forecasting models that predicted the number of new daily COVID cases in each county. These models were trained on the estimated percentage of outpatient doctor visits with confirmed COVID, outpatient doctor visits primarily about COVID-related symptoms, and new hospital admissions with COVID-associated diagnoses: all for 15 counties in California from May 1st, 2020 to November 1st, 2021, making up a total of 8,220 data points.

For the data collection process, we pulled COVID-related data from the COVIDCast Epidata API. Originally, we decided to use 5 features including the three listed above, but we decided not to use the sum of Google search volume for anosmia and ageusia related searches and the sum of Google search volume for solely anosmia related searches. The reason we decided not to use these two in our model was due to the fact that they were not comparable across geographic regions. This would not work for this project as we focused only on the counties of California. In addition, we decided to reduce the data to dates ranging from May 1, 2020 to November 1, 2021 as there were many values of zero for new daily COVID cases towards the beginning of 2020 which would have skewed our models to predict values closer to zero. We also had issues with merging the features into one large dataframe as different features contained different counties with certain dates missing from the data. In order to combat this issue of missing values, we imputed the missing data using a “forward fill” which takes the previous day’s value, specific to each county. Another issue that we ran into was the fact that our labels column, the number of daily COVID cases, had negative values in it, which was very confusing since it is illogical to have a negative value for the number of daily COVID cases. We solved this by taking the absolute value as we believed that the negative values were a result of a mistake in inputting the data. Finally, in order to create a time series-like model, we created two new columns for each feature. These columns were the number of cases at time (t - 1) and (t - 2) for each value. In other words, we created new columns that contained the value for each feature at the previous day as well as the value two days prior. This would ultimately allow us to predict the number of new COVID cases at the present day by looking at the past.

Once again, the main goal of this project was to be able to forecast short-term future COVID-19 cases based on historical data. Our results reflected the messiness of dealing with real

world data, as not only did we take many steps to prepare the data for our models (i.e., missing values, reformatting, etc.), but we also did not see great predictive results. We created both an interpretable model in a decision tree regression model as well as a more complex model in a support vector regression model. After tuning the hyperparameters of the decision tree in order to minimize the mean squared error, we found that the best tree had a max depth of 3 and a splitter type equal to "best." This means that the decision tree with the best accuracy made three splits for a variable, using a splitting criteria that selects the best feature to split after shuffling them. The decision tree's testing RMSE was 908.86 and its R-squared value was approximately 0.57. On the other hand, the tuned SVR had a kernel set equal to "poly," meaning that the model used a polynomial function to transform the dataset and reduce the dimensionality. We found that it produced a testing RMSE of 1403.20 and an R-squared value of approximately -0.025. It is evident that even after cross validation, we still saw high RMSE values and suboptimal R-squared values. Simply put, neither model accurately predicted the number of new daily COVID cases per county using historical data. We can attribute the high mean squared error to the extremely wide range of values of the original dataset in addition to the features simply not being good predictors of new COVID cases. Nevertheless, when comparing the accuracy metrics, the decision tree performed much better than the SVR model. For instance, it was able to capture the overall trend in the rise and fall of new daily COVID cases from May 2020 to November 2021, which is something our complex model was not able to do. One of the reasons may be that SVR is computationally costly and is not optimal for large datasets. In real life, the data obtained about COVID-19 cases would likely be even larger than our dataset – in which case, in alignment with what our results showed, it may be better to utilize a decision tree rather than an SVR model in a real scenario. Furthermore, our model focused only on a few feature variables, but there are many more features that can potentially be examined in the real world. In this case, where the dataset may be small but has more dimensionality, SVR appears to be an appropriate model to evaluate. Overall, these aspects of a real dataset should be taken into consideration when determining which model to utilize.

With respect to the feature importances in our decision tree regression model, we were able to determine that the most influential feature in terms of deciding how the tree was split was outpatient doctor visits primarily about COVID-related symptoms at time (t - 2). The next two most important features were new hospital admissions with COVID-associated diagnoses at time (t) and estimated percentage of outpatient doctor visits with confirmed COVID at time (t - 1). Something that we found strange was that the most important feature in our model was one that contained a value from two days prior to the observation. However, because this is a time series related model, it makes sense that the values from previous days are extremely useful in predicting the number of COVID cases for the present day.