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

Tyler Chia, Joanne Kim, Joshua Harasaki, Michael La

Introduction

- Aimed to utilize past data to predict the number of new, daily COVID cases per county using two time-series forecasting models
- The two models we evaluated were a Decision Tree Regressor and Support Vector Regressor
- This information can be used to estimate the number of resources needed to combat COVID-19 and determine the most influential features in the number of COVID cases

Data Collection

- Final Features:
  - Estimated percentage of outpatient doctor visits with confirmed COVID
  - Outpatient doctor visits about COVID-related symptoms
  - New hospital admissions with COVID-associated diagnoses
- Dates: May 1, 2020 to November 1, 2021
- Regions: 15 counties in California
- Missing values imputed using “forward fill” (filling the current value with previous available data)

Model Training/Evaluation

- Before training we created columns for the values of each feature on previous two days (t - 1) and (t - 2) to predict the number of new COVID cases at present time (t)
- Model 1 (Decision Tree Regressor):
  - Used cross-validation to tune the hyperparameters “max depth” and “splitter”
  - Max depth of 3 and splitter type of “best” produced the lowest validation RMSE
  - Evaluating tuned tree on testing data resulted in an RMSE of 908.86 and an R-squared value of 0.57
- Model 2 (Support Vector Regressor):
  - Tested three different SVR models to see if data reduction was necessary, and decided they were not
  - Used cross-validation to find best kernel
  - Evaluating final SVR with polynomial kernel produced RMSE of 1403.20 and an R-squared value of -0.025
  - **Figure 3** shows extremely small differences between training and validation RMSEs

Results

Model 1: Decision Tree Regressor

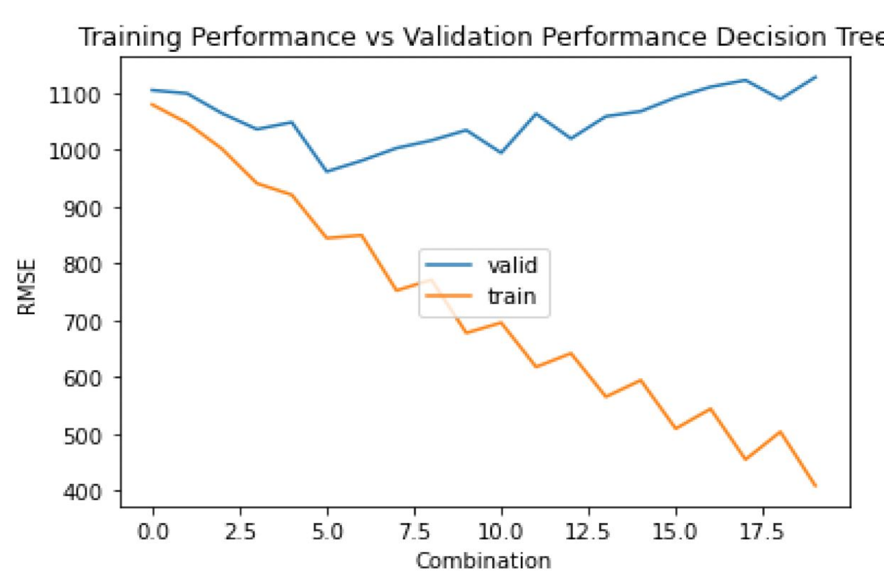

Figure 1

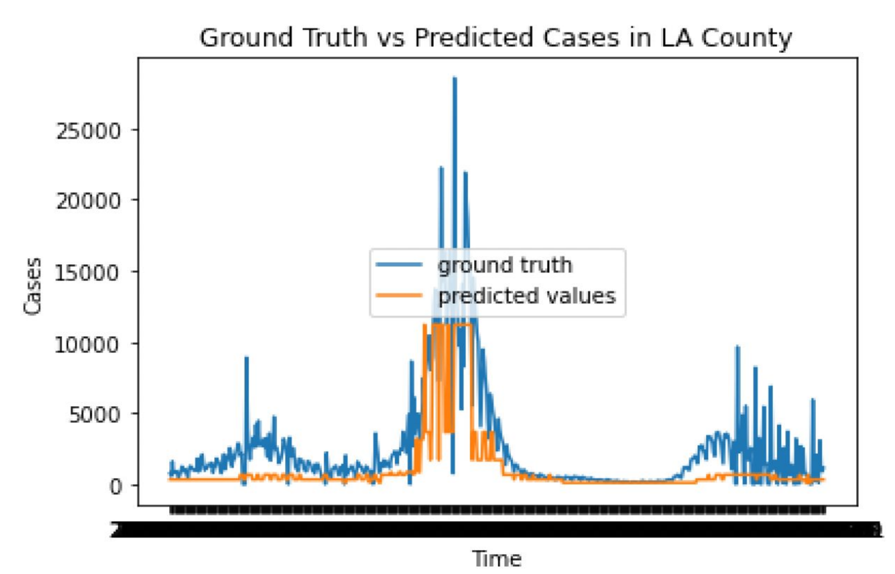

Figure 2

Model 2: Support Vector Regressor

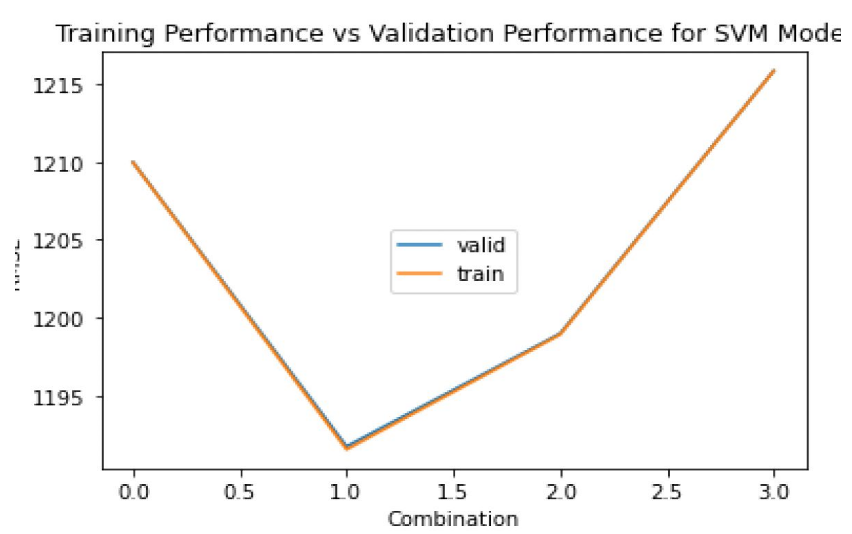

Figure 3

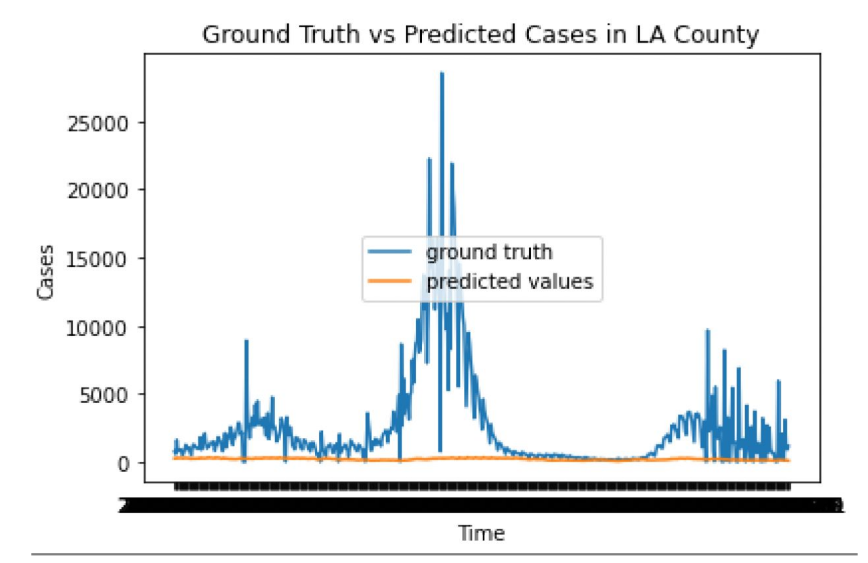

Figure 4

Decision Tree Feature Importance

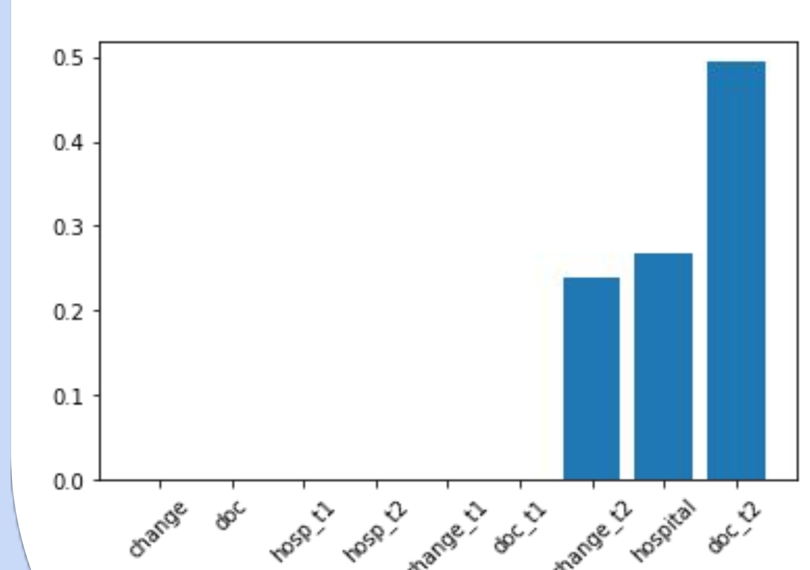

Figure 5

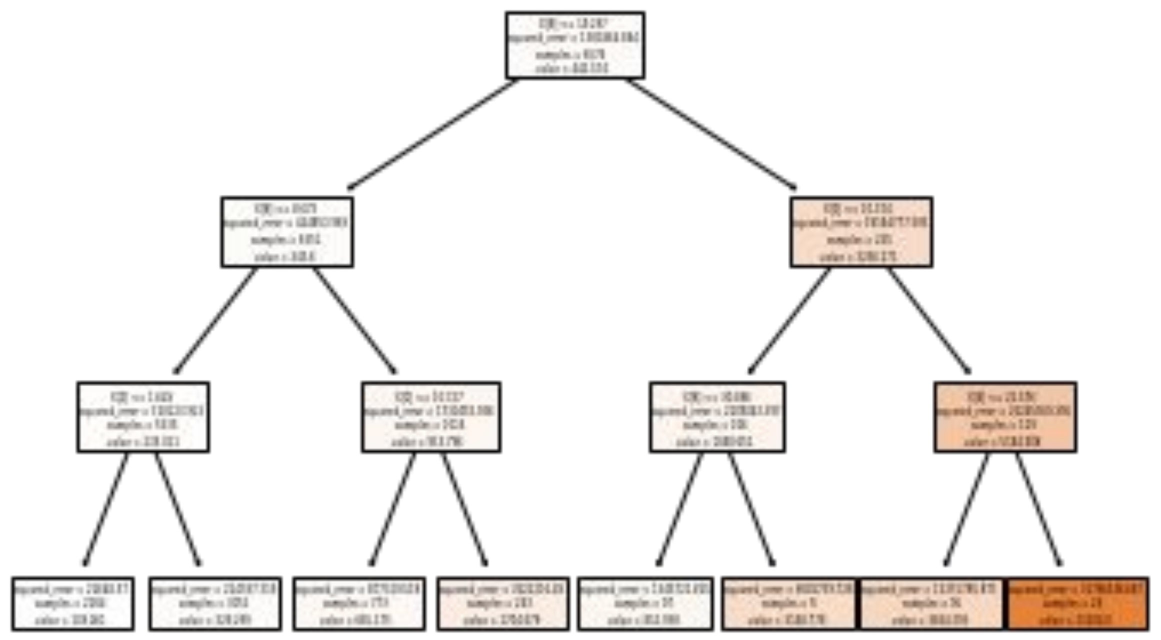

Figure 6

Conclusion

Results

- The Decision Tree Regressor performed much better than the SVM model
  - This may be due to the fact that SVRs are not optimal for large datasets
- Model 1 was able to capture the overall trend in the rise and fall of the ground truth number of COVID cases, despite suboptimal accuracy metrics (**Figure 2**)
- Model 2 performed poorly and was not able to model the general trend of COVID cases (**Figure 4**)
- The most influential feature in Model 1 was outpatient doctor visits primarily about COVID-related symptoms at time (t-2) (**Figure 5**)
- 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)

Future Works

- Explore and add more features
- Examine data across larger regions than counties, such as states
- Spend more time evaluating the dataset to choose a compatible model to prevent issues like our SVR performing poorly due to the size of our dataset
- Include more time-series columns (t-3, t-4, …) to look even further into historical data

References

- COVIDCast Epidata API

Coordination

- Coordinated well across the team. Worked on various aspects of the project during collaborative meetings

Acknowledgements This project is possible thanks to the Data Science Capstone course at UCSB, which is sponsored by the National Science Foundation.