Predicting Daily COVID-19 Cases in California Counties Using Decision Tree Regression and SVR Time-series Models

Jonathan Wong, Aaron Huang, Nathan Chow

Intro/Motivation

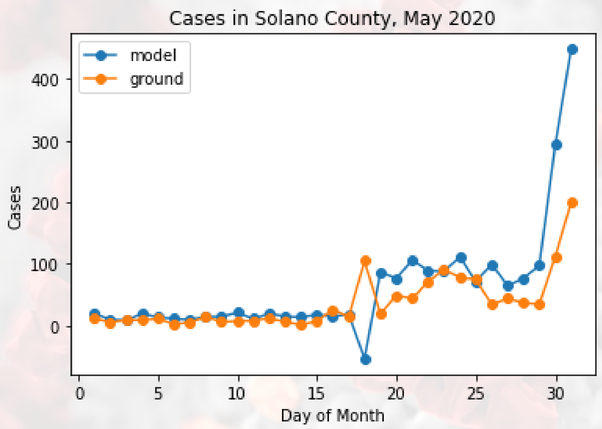
- COVID-19 is a global pandemic that has affected countless lives over the past 2 years.
- Goal: To predict the number of daily COVID-19 cases so that a specified region can be better prepared for surges in cases.
- Challenges: There is a large amount of data on COVID-19 and not all the data is relevant or consistent for predicting future COVID-cases.

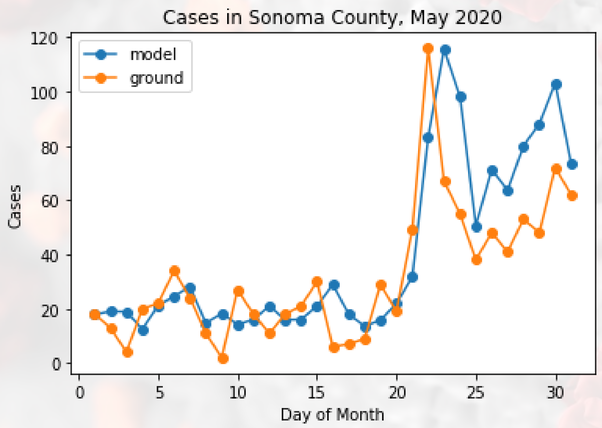
Preprocessing

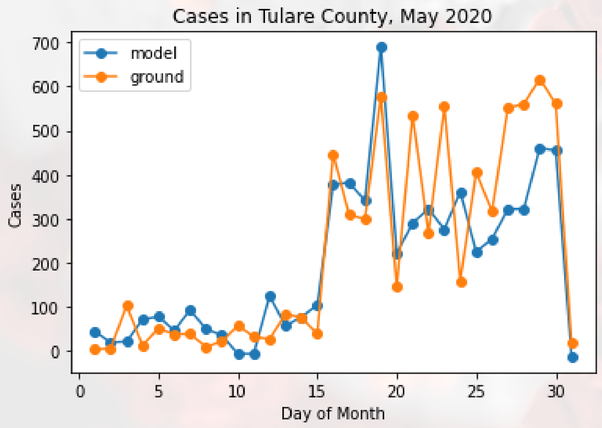
- Use the following features from COVIDcast Epidata: #confirmed cases, #cases detected, #doctor visits about symptoms, #google searches for loss of taste, #google searches for loss of smell, and #COVID-related hospital admissions.
- Data is only from California counties between 05-01-20 and 12-03-21
- Remove all data with no ground truth, 0 ground truth, and negative ground truth
- Remove all data that were missing data for 3+ features
- Impute missing values using the monthly mean of that year for that feature. Use the mean of the entire column if the value is missing for that entire month.

Results

Linear Regression Results

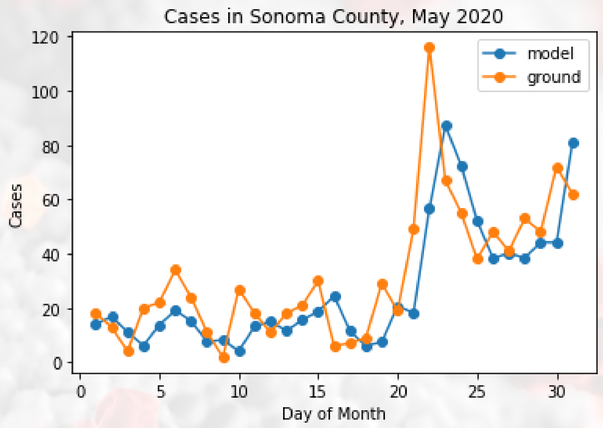


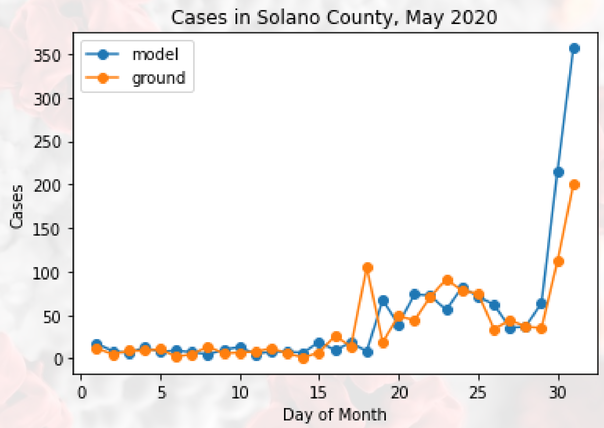


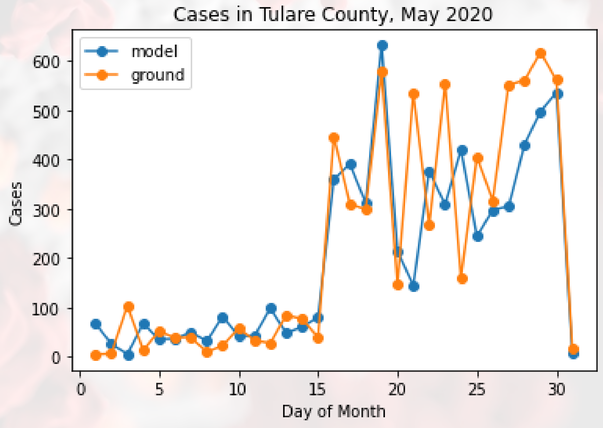


- Most important Feature: Number of cases for the previous two days
- **R^2 Score:** 0.67

SVR Results







- **Optimal Hyperparameters:**
  Kernel function: linear
- **R^2 Score:** 0.74

Modeling/Training

- Data from the previous day and current day was used to predict the next day.
- Hyperparameter tuning with 5-fold cross validation was used
- Decision tree regressor considered.
- Hyperparameter with the highest validation performance and reasonable training performance was chosen for both models.

Conclusion

- SVM performs better than the Linear Regression Model
- Decent performance indicates that Linear Regression and SVM are acceptable models to use for predicting future COVID cases.

Future Work

- Retrain models with weather data for each county.
- Limit the geographical scope of the models because the rates of change may be different per area.

Coordination

- Jonathan Wong: Preprocessing
- Nathan Chow: Model Training
- Aaron Huang: Poster and Report