Springboard Data Science Career Track

Capstone Project 3

"Using Mobility Markers to Predict the Number of

New COVID-19 cases per Day"

By Jeremy Silva

August, 2020

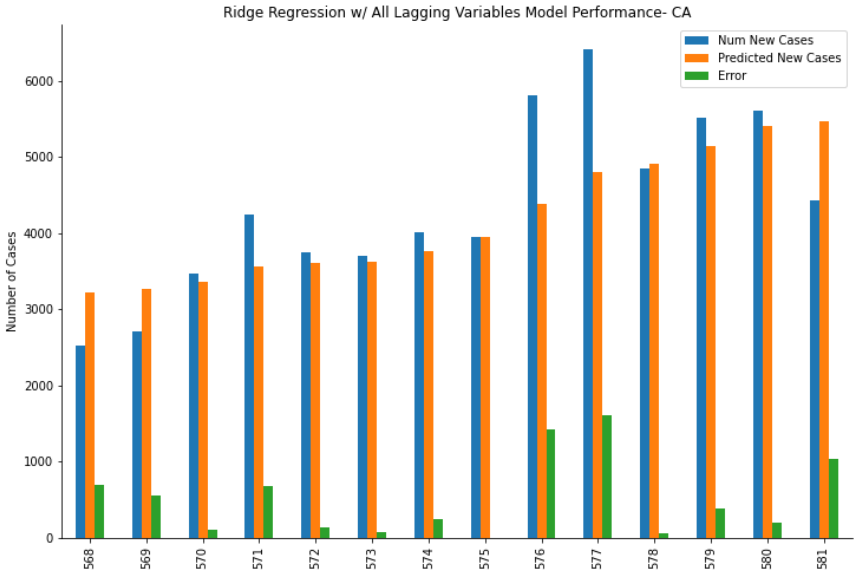
Section 1: Introduction

Problem Statement and Objective

Covid-19 is currently the most pressing problem facing the United States and the rest of the world. States and local governments across the country are issuing stay at home orders and implementing other policies to reduce the movement of people within their districts. Thus making it clear that top officials see a clear link between mobility within a state and that state's Covid-19 infection rate. The goal of this project is to use raw smartphone mobility data made public by Apple and Google to build a model which can forecast the Covid-19 infection rate for a given State. The foreseen utility of this model is twofold: 1) An accurate forecasting model for the Covid-19 infection rate would be a valuable tool for public officials 2) A model that uses mobility markers as input variables could help further our understanding of the relationship between mobility and infection rate.

Synopsis of Key Data Science Findings

The final model for this iteration of the project was able to forecast the Number of New Covid-19 Infections over a 14 day prediction window with a test set **Mean Average Error of 517 (cases)** and a test set **R2 of 0.58** for the state California.



** The remainder of this report will cover the end to end implementation details of the project but all source code can also be found in the Jupyter Notebooks housed in [this Github Repository](#).

Section 2: Approach

2.1 Data Acquisition and Wrangling

Data Sources

3 Data Sets were aggregated for the purposes of this project (All are available for download by the general public).

1) New York Time Covid-19 U.S Infection Dataset- This dataset contains the number of new COVID-19 cases and related deaths reported on a daily basis for every state in the United States.
2) Google Mobility Report Data- This dataset contains mobility on the city, state and country level worldwide. The data is gathered from google maps and other google products. The segment of the dataset used for this project contains the following mobility marker categories: Retail and Recreation, Grocery and Pharmacy, Parks, Transit Stations, Workplaces and Residential. All values are recorded as a change from a baseline which was determined from activity preceding the wide scale COVID-19 outbreak in the U.S.
3) Apple Mobility Trends Report- This dataset contains mobility data on the state and city level gathered from Apple Maps. The mobility data fields are Driving GPS Hits, Walking GPS Hits, and Transit GPS Hits. All are represented as a change from a baseline number which was calculated from activity preceding the wide scale COVID-19 outbreak in the U.S.

Data Wrangling

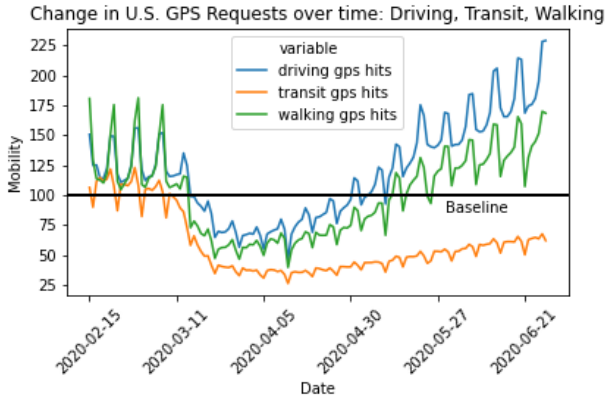
For the sake of brevity I will not go into extensive detail regarding the data wrangling and cleaning process but rather I will cover the key alterations made that are relevant to the subsequent modeling. Here are the major steps:

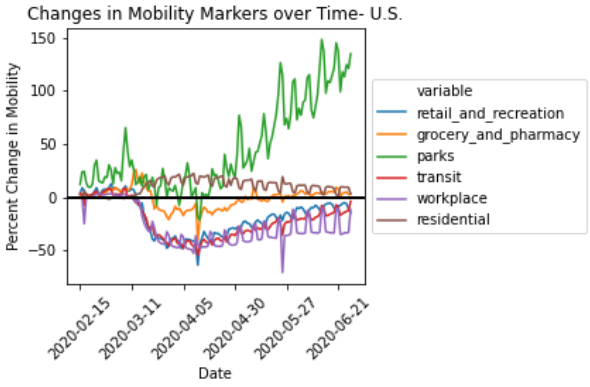
1) The three data sets were joined by matching on the state level. The datasets were re-oriented for the sake of concatenation but no changes were made to the underlying data.
2) A number of new cases column was added. As downloaded, the data had the number of cases recorded as a running total. For modeling purpose I added a new column, 'Num New Cases', that holds the number of new cases recorded for that day. The number was

calculated by taking the number of cases and subtracted out the number of cases the day prior.

3) Adding lagging variables. This was done during the modeling phase and can be found in the modeling notebook rather than the data wrangling notebook. The reason for these lagging variables will be discuss in detail later. Two lagging variables were computed for each field: an average of the previous 7 days and an average of the previous 14 days.

Visualizing the Data





Just by visualizing the mobility markers for the U.S. as a whole we can see that there are some notable trends in how the mobility markers change over time. This tells us that these markers could potentially be used as input variables to a model. If there was no trend over time then they would not be very informative as input variables.

2.2 Baseline Modeling

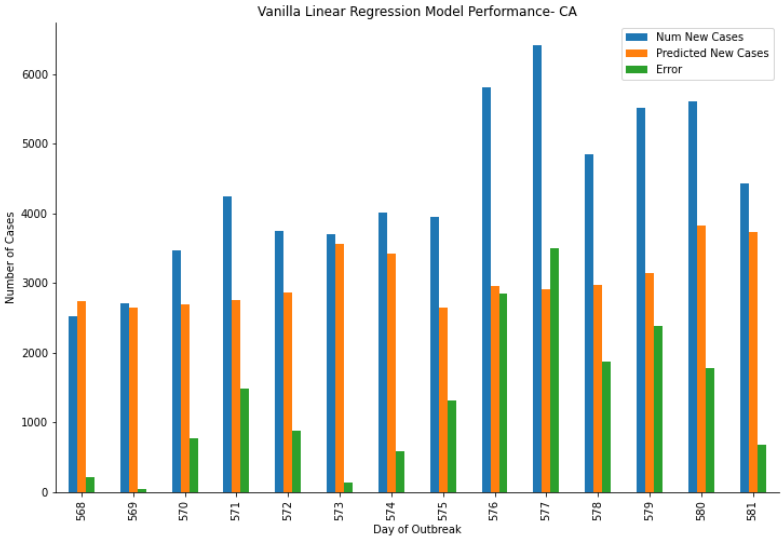
Basic Approach to Modeling

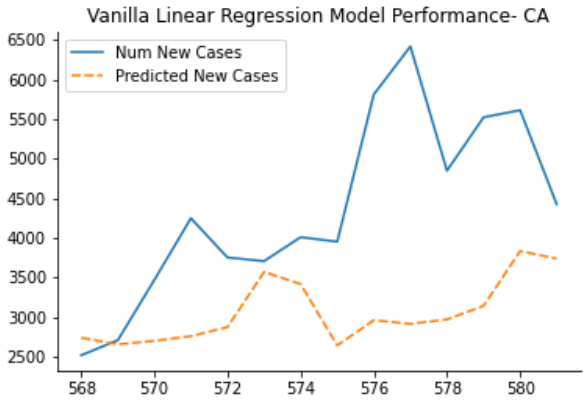
In the context of this problem, models need to be built on a per state basis. Thus for the initial modeling I chose to use the data for California. Once I settled upon the best modeling pipeline I then used that pipeline to build models for 10 other states to see if the pipeline could generate useful models for other states as well.

Baseline Model

The baseline model build was an out of the box linear regression model using all 9 of the mobility markers as X variables and the number of new cases as the target variable. Performance for the baseline model was extremely poor as demonstrated below.

Model	R2 Train/ R2 Test	MAE Train/ MAE Test
Vanilla Linear Regression	0.77/ -1.22	362/ 1323





Takeaway: From these figures we can see that with the baseline model we have extremely high error bars and we can see from the line graph that our predictions barely track the line of the actual number of new cases.

2.3 Extended Modeling and Findings

The Process

After building the baseline model I continued to iterate by tweaking input variables and hyperparameters. Here is how the basic progress of model development went:

First: I built the Baseline Vanilla Regression Model.

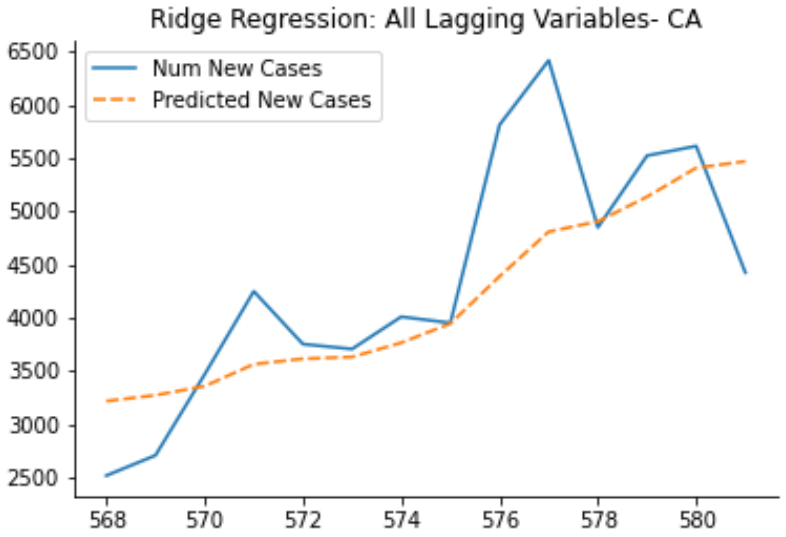
Second: I attempted to tackle the overfitting by running GridSearchCV on a set of hyperparameters for Lasso, Ridge and ElasticNet regressions.

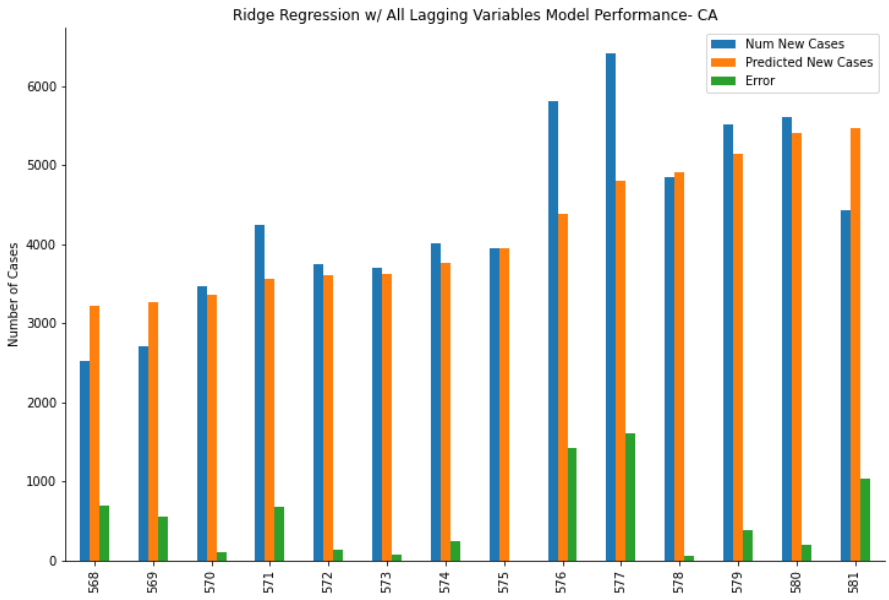
Third: I used two lagging variables for all mobility markers as input variables. Then I utilized GridSearchCV to find the best hyperparameters.

Fourth and Final: I used two lagging variables for all the mobility markers and two lagging variables for the target variable as input variables. I then used GridSearchCV to find the best hyperparameters and ultimately built a final Ridge Regression Model.

Final Model for California

Ridge Regression w/ Lagged Mobility Markers and Lagged Target Variable	0.86/ 0.58	240/ 517





Takeaway: As we can see with this final model we have significantly reduced the error bars and are now tracking the true new cases line with our prediction line much better.

Below is a summary of how the performance metrics evolved over modeling iterations.

Model	R2 Train/ R2 Test	MAE Train/ MAE Test
Vanilla Linear Regression	0.77/ -1.22	362/ 1323
ElasticNet Regression	0.77/ -1.61	361/ 1489
Vanilla Linear Regression w/ Lagging Mobility Markers	0.83/ 0.29	253/ 685
Vanilla Linear Regression w/ Lagged Mobility Markers and Lagged Target Variable	0.87/ 0.46	243/ 654
Ridge Regression w/ Lagged Mobility Markers and Lagged Target Variable	0.86/ 0.58	240/ 517

Additional 10 State Modeling

After settling on a model for California I tried using the same general pipeline to build models for 10 additional states. For each state I computed all appropriate lagging variables, ran a GridSearchCV to find the best regression parameters and then subsequently build a model with those parameters. Below is a summary of the results.

	State	R2_Test	MAE_Test
0	Alabama	0.070175	186.975599
1	Washington	0.177980	86.895380
2	Oregon	0.138774	57.128520
3	Ohio	-0.533123	197.980633
4	Florida	0.741066	771.028057
5	Colorado	0.181774	44.681297
6	Minnesota	-0.066840	82.210855
7	Texas	0.508102	853.697622
8	Virginia	-0.318555	81.986198
9	New York	-49.194226	536.949783

Takeaway: As we can see the pipeline used to build a model for California generalized to other states with varying levels of success. Visualizations for each state, similar to those generated for California, can be viewed in this notebook.

Section 3 Conclusions and Future Work

Key Takeaways

1) As seen in the case of California, using lagged mobility markers to predict COVID-19 Infection rate over a 14 day window can be moderately accurate.
2) The fact that lagging the variables drastically improves model performance tells us that there is likely a time dependent relationship between mobility and COVID-19 infection rate.
3) A general modeling pipeline that works for one state will not necessarily work for another state.

Future Work

1) Being that lagging the variables had such a significant impact on performance, intuition tells us that we could further improve our model for California but bringing in some form of time series analysis.
2) More work needs to be done as far as figuring out how to build a modeling pipeline that can be generalized and used to build accurate models for many states.
3) As more data comes in the models need to be reevaluated. An increase in training data may be able to improve performance on its own.
4) It may be fruitful to further explore how individual mobility markers interact with the infection rate.

Section 4 Stakeholder Recommendations

While the model for California reached a decent level of accuracy the model would need to be further tested and iterated upon before being ready for deployment in making decisions for things like reopening procedures and hospital staffing.

As far as other states, it is clear that more individualistic modeling efforts need to be done on a per state basis in order to consistently produce accurate models. However, the results from California show us that there is potential in using mobility markers in the modeling process when trying to predict Covid-19 infections rate.