Weekly Capacity Allocation for COVID Patients in a Hospital

Prepared by Umur Berkay Karakaş, Doruk Özer and Berk Yazar

1. Introduction

Mankind has struggled with deadly diseases that have been seen at regular intervals since its history. The coronavirus, which we have all followed very closely in the recent past, has taken humanity captive for the last 2 years. Large states, organizations, and groups, especially individuals, were greatly affected by this and everyone had to change their way of living drastically. For this reason, we will establish the subject of our project on a hospital in Germany during the COVID period.

In our problem, we will be working on planning the number of beds allocated for hospitalized patients. While doing this, we will have a few assumptions to rationally simplify our problem. Our demand, that is, the number of patients who need to be hospitalized, will correspond to 10% of the patients in the country and the service level will be 100% for mildly infected COVID patients. The end point we want to reach is to make a plan that will have enough beds for 95% of the patients who need to be hospitalized each week. We have taken the underage and overage costs as the cost of the beds that are insufficient and unoccupied, respectively.

Variables we consider to be predictors when cross validating will be one week lagged movement trends across different categories of places such as retail and recreation, groceries and pharmacies, parks, transit stations, workplaces, and residential, one week lagged daily cases and the logarithm of it, the stringency index, one week lagged reproduction rate (Rt value), one week lagged vaccination numbers, one week lagged booster shots and seasonal dummies for fall and winter. We thought it would be more reasonable to use one-week lags of some predictors because estimators such as daily cases, reproduction rate, booster shots, etc. are lagged in nature and they affect the number of hospitalizations per week.

2. Data Explanation

In our project, we have used the well-known COVID-19 Dataset[1] from "Our World In Data" and COVID-19 Community Mobility Dataset[2] from Google.

2.1. OWID COVID-19 Dataset

Our World in Data's dataset consists of 63 COVID-19 related metrics for each country in the world which are either fixed or updated daily/weekly. Explanation of each metric and its source can be found in OWID COVID-19 codebook. The main sources of the data are UN, COVID-19 Data Repository by CSSE at Johns Hopkins University, World Bank, OECD, European CDC and national government reports. Time dependent data is entirely dependent on the transparency of each national government and it is not altered or corrected by any means. The variables of interest in our project would be the ones that are updated daily/weekly since we aim to predict weekly COVID hospitalizations and the factors affecting COVID hospitalizations are time dependent.

2.2. Google COVID-19 Community Mobility Dataset

Google's COVID-19 Community Mobility Dataset consists of six metrics for each country in the world which are updated daily. The data reports the percentage change from the baseline in the mobility trends in six selected place categories, namely retail and recreation, supermarket and pharmacy, parks, public transport, workplaces and residential. The baseline for Germany in such categories is the mobility trends assessed between 3 January - 6 February 2020. The data is collected from Google users who have opted in to "Location History" in their Google accounts, therefore the data represents a sample of Google users and might be slightly biased.

2.3. Data Processing for Prediction

Each Coronavirus variant had different transmission and lethality characteristics. Therefore, for our predictions to be more consistent considering the recent status of Coronavirus variants, we had decided to use the data from the beginning of the emergence of the Omicron variant (approximately 1 October 2020) to 1 May 2021.

In our predictions, our variables of interest were time dependent data, therefore we had used a subset of OWID dataset and the entire Google COVID-19 Community Mobility Dataset between the dates stated above. In OWID dataset, our variables of interest were namely: weekly hospital admission per million people, logarithm of daily smoothed new cases per million, stringency index, reproduction rate, test positivity rate, percentage of fully vaccinated people and percentage of people who had boosters.

We had used one week lagged predictors in our predictions (except for stringency index) because of two following reasons: we were trying to predict hospital admissions and according to Faes et al.[3], time between symptom onset and hospitalization ranges between 3 and 10.4 days Therefore it was reasonable to use the predictor values from 7 days ago. Also, we were predicting the weekly hospitalization before the week started, therefore the recent data in our hands was the previous week's data to predict the next week's hospitalizations.

3. Prediction of Hospital Admissions

In order to get a better understanding of our data and to get better predictive results, we run regularized regression models such as Lasso along with depth controlled tree models such as regression tree, random forest and boosted tree.

3.1. Lasso Regression

In lasso, we initially run our model with full predictors. We manually tried different alpha values and evaluated the results by observing significant predictors, RMSE of the training set and RMSE of the test set. We figured out that the best alpha value is 0.2. In choosing the alpha value, we wanted a lower $RMSE_{training\_set}$ / $RMSE_{test\_set}$ ratio to avoid overfitting and around 4 or 5 significant predictors left after regularization. After regularization, the predictors that are left are log of one week lagged cases, stringency index, one week lagged RT, one week lagged booster shots and one week lagged park mobility. In Figure 1, it can be observed that R^2 is 0.978 and adjusted R^2 is 0.971 both of which look quite good. RMSE values for training and test sets for OLS model with reduced predictors are 4.63 and 17.49. In Figure 2, the difference between

predicted and real hospital admissions can be seen. Even though the test set has almost three times bigger RMSE than the training set, predictions can be considered acceptable.

3.2. Random Forest

In the random forest model, we tried to find the maximum number of features to be used in order to minimize the test RMSE. As a result, the maximum number of features that minimizes test RMSE was 12 for which training RMSE was 7.42 and test RMSE was 18.29. In Figure 3, the prediction of the random forest model and the real values can be seen. From Figure 4, it can be inferred that the top four most important features in random forest model were log of one week lagged cases, one week lagged grocery pharmacy mobility, one week lagged positivity rate and one week lagged RT.

3.3. Regression Tree

In the regression tree model, we tried to find the maximum depth to be used in order to minimize the test RMSE. As a result, the maximum depth that minimizes test RMSE was 4 for which training RMSE was 2.36 and test RMSE was 19.19. In Figure 5, the prediction of the regression tree model and the real values can be seen. From Figure 6, it can be inferred that the top four most important features in the regression tree model were one week lagged total vaccinations, log of one week lagged cases, one week lagged retail and recreation mobility and one week lagged RT.

3.4. Boosted Tree

In the boosted tree model, we tried to find the maximum depth and the learning rate to be used in order to minimize the test RMSE. As a result, the maximum depth and the learning rate pair that minimizes test RMSE was 3 and 0.3 for which training RMSE was 2.44E-07 and test RMSE was 15.46. In Figure 7, the prediction of the boosted tree model and the real values can be seen. From Figure 8, it can be inferred that the top four most important features in the boosted tree model were one week lagged total vaccinations, log of one week lagged cases, one week lagged grocery and pharmacy mobility and one week lagged RT.

3.5. Results of Predictions

Model specification, training and test RMSE for each model can be seen from Table 1. Random forest uses 12 predictors and has higher RMSE for training and test sets than any other model. Regression tree seems to have an overfitting problem as the test RMSE is almost seven times higher than the training RMSE. Boosted tree has the same problem with its corresponding RMSE values. Therefore, we have decided to use predictions of lasso regression in our prescriptive analysis.

4. Prescriptive Analysis

Preparing a weekly COVID bed reservation calendar for a hospital in Germany is a type of newsvendor problem. The problem mainly concerns the mismatch between supply and demand where the demand is the number of patients who got COVID and applied to the hospital each week and the supply is the number of capacity allocated daily for people who got COVID for an entire week. In addition, the capacity that is reserved for COVID patients is initially reserved for regular patients. In other words we are trading off from the capacity which are reserved for regular patients and that causes the mismatch of our problem. The mismatch can be explained as the excessive and insufficient reservation of COVID beds. The objective of our problem is to reserve the beds for 95% of the COVID patients i.e. 95% service level for COVID patients. Underage cost would be the cost of allocating less number of beds than the number of COVID patients and overage cost would be the cost of allocating more number of beds than the number of COVID patients, so the beds could have been used for regular patients but they were empty.

4.1. Implementation

From the predictive part, the weekly hospital admission for a hospital in Germany is predicted by implementing regression. The test set is used for validation purposes of our regression model. In the prescriptive part we used training and test data combined for estimation of weekly hospital admissions. In other words, the model which uses all of the dataset that estimates weekly COVID hospital admission is the input as the demand for our newsvendor

problem. After estimating the weekly COVID patient admission, our main task was determining the cost function which we want to minimize by determining the optimal number of capacity to be allocated for the COVID patients. Since the objective of our problem would be to reserve the capacity for 95% for the COVID patients who made an admission to our hospital weekly, the ratio of (underage cost /(underage cost + overage cost)) is determined as 95% and the coefficients for underage cost and overage cost is determined according to this ratio.

We decided to use Gurobi optimizer to minimize the cost function and find the optimal quantity for the capacity allocation for COVID people. The constraints for the cost function is determined based on the 95% service level for COVID patients. The optimization is constrained with underage cost, when the total number of weekly admission of COVID patients exceeds the capacity we allocated for the COVID patients and the optimization is constrained with overage cost when the capacity we allocated for the COVID people exceeds the total number of weekly admission of COVID patients. So we added the constraints to the cost function and set this cost function to Gurobi and Gurobi had minimized the cost function based on these constraints. The minimization is based on finding the optimal number of capacity allocation for COVID people and Gurobi finds the capacity to be allocated for each day of the next week as 125.

4.2. N-Step Ahead Capacity Estimation

It is important to emphasize we are using available predictor's values to estimate next week's hospital admissions for COVID people and the predictors are also handled weekly. That is to say, if we want to estimate the two week ahead capacity allocation or if we want to extend the weekly COVID bed allocation to 4 weeks in one month we have two options. First option is that we can fit estimators for the predictors and estimate the values for the predictors for next weeks that we want to estimate and use those values as predictors for those weeks' capacity estimation. However, using estimated values for predictors would be costly and might be more biased. It is possible that using predictors for further steps ahead would increase the variance. That is to say, we should use available real data and wait for next week for the next week's predictor values and make our estimations based on those values for both the predictive part and prescriptive part. In other words, we should estimate the weekly COVID patient admission using the past week's data and use that weekly COVID hospital admission prediction as demand in our

newsvendor problem. So, we should repeat the predictive and prescriptive part for each new week to find optimal COVID capacity allocation for that new week. In addition, if we change the predictive part to monthly estimation of COVID people's hospital admission then our prescriptive part would estimate the optimal number of COVID bed allocation for next month. However, we decided to find weekly optimal capacity allocation because COVID has a lot of weekly variations based on the government's attitudes, precautions, and restrictions. All of the weekly variations are able to be estimated by the predictors we used. Because of the weekly variations, it is better to find weekly optimal capacity allocation instead of monthly optimal capacity allocation.

5. Summary

We analyzed the problem under two main subjects, predictive and prescriptive. In the predictive part, we used a couple of machine learning algorithms and evaluated their error performances. We ran into two major problems here, the high RMSE value and the overfitting issue. Due to these, we decided to use the model we would reduce with Lasso regression in our prescriptive analysis. With our findings with fewer predictors after running the Lasso regression, it was concluded that we should allocate a capacity of 125 people every day of the week. It is important to underline that this problem is dynamic rather than static. Since this analysis is done with the weekly estimations, different estimations will be used to estimate different timings and these will give us various results. Still, the basis of our project was on a weekly calculation, therefore we followed a path accordingly and came up with this result.

6. References

1. Mathieu, E., Ritchie, H., Ortiz-Ospina, E. et al. A global database of COVID-19 vaccinations. Nat Hum Behav (2021). https://doi.org/10.1038/s41562-021-01122-8

2. Google LLC. Google COVID-19 Community Mobility Reports. https://www.google.com/covid19/mobility

3. Faes, C., Abrams, S., Van Beckhoven, D., Meyfroidt, G., Vlieghe, E., Hens, N., & Belgian Collaborative Group on COVID-19 Hospital Surveillance (2020). Time between Symptom Onset, Hospitalisation and Recovery or Death: Statistical Analysis of Belgian COVID-19 Patients. International journal of environmental research and public health, 17(20), 7560. https://doi.org/10.3390/ijerph17207560

Appendix

Model	*Training RMSE*	*Test RMSE*	*Spec.*
Lasso & OLS	*4.63*	*17.49*	*alpha = 0.2*
Random Forest	*7.42*	*18.29*	*max_features = 12*
Regression Tree	*2.36*	*19.19*	*max_depth = 4*
Boosted Tree	*2.44E-07*	*15.46*	*max_depth, lr = 3, 0.3*

Table 1. Model comparison

OLS Regression Results

Dep. Variable:	hosp_admission	**R-squared:**	0.978
Model:	OLS	**Adj. R-squared:**	0.971
Method:	Least Squares	**F-statistic:**	127.4
Date:	Wed, 08 Jun 2022	**Prob (F-statistic):**	3.68e-11
Time:	13:27:24	**Log-Likelihood:**	-59.024
No. Observations:	20	**AIC:**	130.0
Df Residuals:	14	**BIC:**	136.0
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-201.0377	15.790	-12.732	0.000	-234.904	-167.171
log_one_week_cases	47.4188	3.029	15.657	0.000	40.923	53.915
str_index	0.3399	0.162	2.099	0.054	-0.007	0.687
one_week_rt	20.0709	7.791	2.576	0.022	3.360	36.782
one_week_booster	-1.4642	0.111	-13.248	0.000	-1.701	-1.227
one_week_parks	-0.1087	0.056	-1.939	0.073	-0.229	0.012

Omnibus:	0.122	**Durbin-Watson:**	1.933
Prob(Omnibus):	0.941	**Jarque-Bera (JB):**	0.085
Skew:	0.095	**Prob(JB):**	0.959
Kurtosis:	2.745	**Cond. No.**	1.33e+03

Figure 1. OLS results with predictors left from Lasso

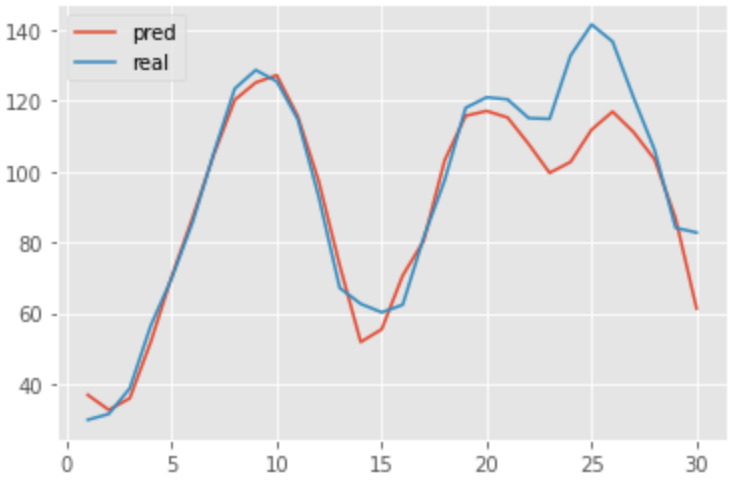

Figure 2. Predicted and real weekly hospital admissions for OLS with reduced model

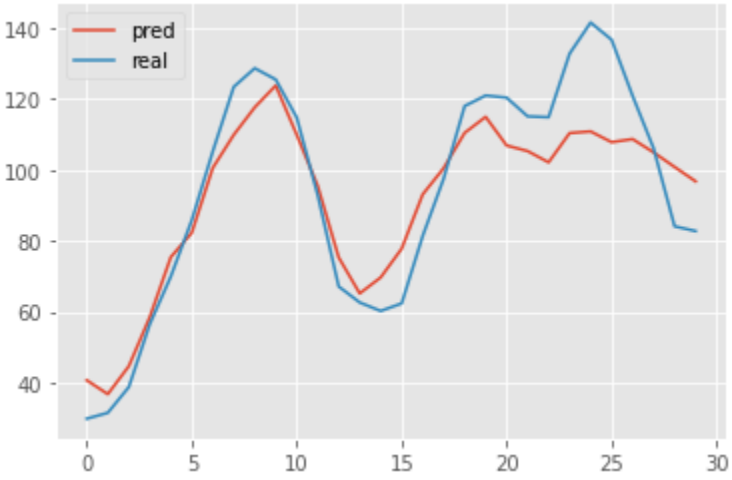

Figure 3. Predicted and real weekly hospital admissions for random forest model

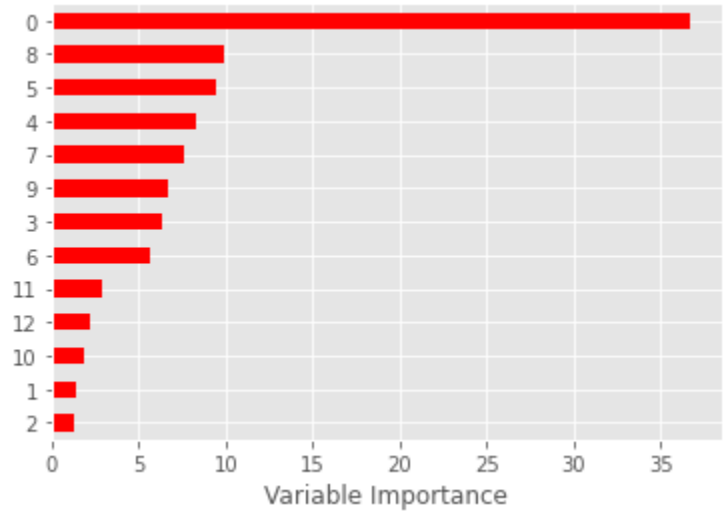

Figure 4. Variable importance graph for random forest

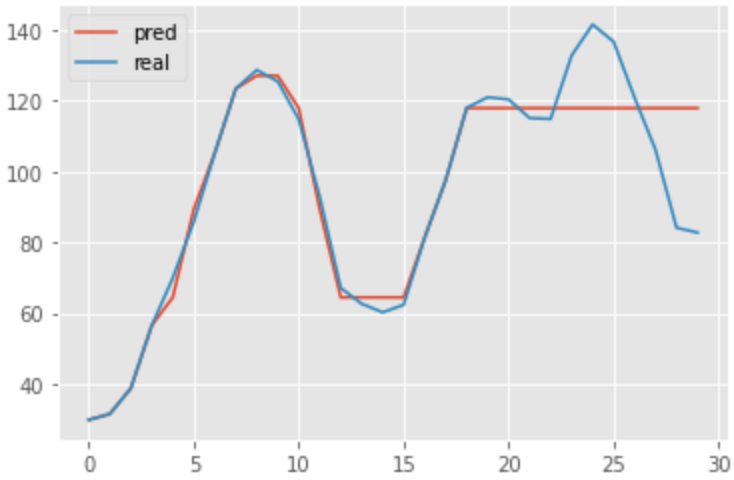

Figure 5. Predicted and real weekly hospital admissions for regression tree model

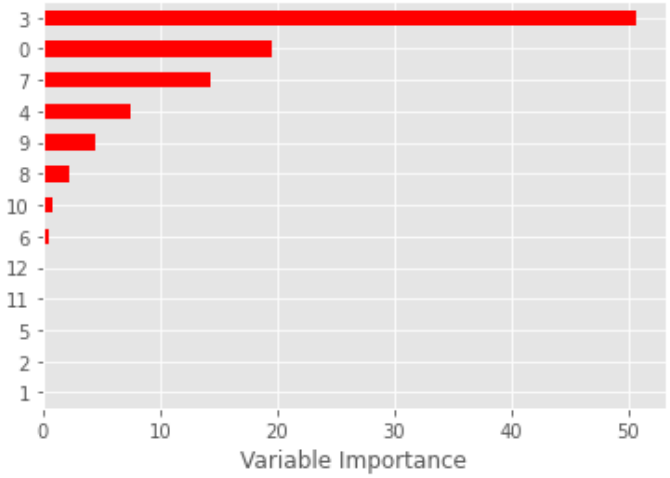

Figure 6. Variable importance graph for regression tree model

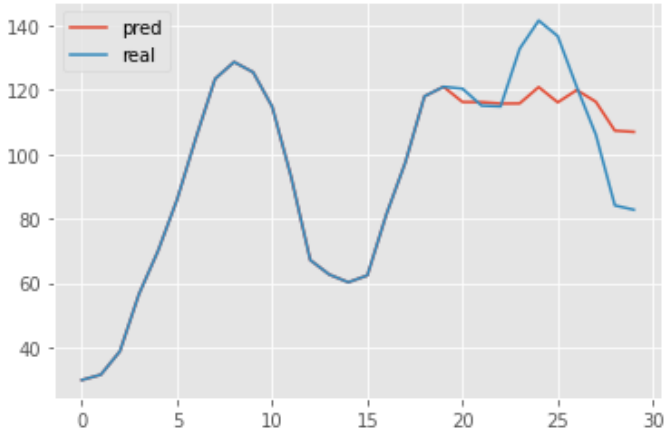

Figure 7. Predicted and real weekly hospital admissions for boosted tree model

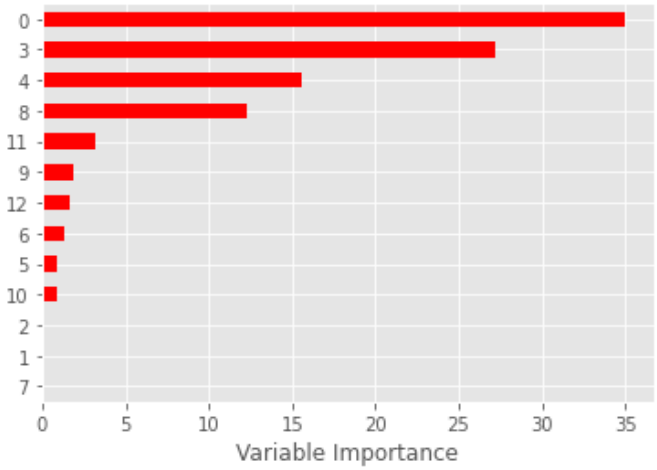

Figure 8. Variable importance graph for boosted tree model