

**Scheduling Project**

**Team 4**

**Jaime Chaire**

**Ethan Dark**

**Jamie Liu**

**Huaiqian Yan**

Contents

[1. Summary](#_heading=h.gjdgxs) 2

[Model Metric Summary](#_heading=h.30j0zll) 2

[Economic Impact Summary - Cancellations](#_heading=h.3znysh7) 2

[Economic Impact Summary - Visit Volumes](#_heading=h.vx1227) 3

[Recommendations](#_heading=h.tyjcwt) 3

[2. Data](#_heading=h.3dy6vkm) 4

[Summary](#_heading=h.1t3h5sf) 4

[Cleaning approach](#_heading=h.4d34og8) 4

[3. Model 1](#_heading=h.2s8eyo1) 4

[Description](#_heading=h.17dp8vu) 4

[Data prep](#_heading=h.3rdcrjn) 4

[Parameters](#_heading=h.26in1rg) 4

[Evaluation](#_heading=h.lnxbz9) 4

[Learning](#_heading=h.35nkun2) 5

[4. Model 2](#_heading=h.1ksv4uv) 6

[Description](#_heading=h.44sinio) 6

[Parameters](#_heading=h.2jxsxqh) 6

[Data prep](#_heading=h.z337ya) 6

[Evaluation](#_heading=h.3j2qqm3) 6

[Learning](#_heading=h.1y810tw) 7

[5. Model 3](#_heading=h.4i7ojhp) 8

[Description](#_heading=h.2xcytpi) 8

[Parameters](#_heading=h.1ci93xb) 8

[Data prep](#_heading=h.3whwml4) 8

[Evaluation](#_heading=h.2bn6wsx) 8

[Learning](#_heading=h.qsh70q) 9

[6. Model 4](#_heading=h.3as4poj) 10

[Description](#_heading=h.1pxezwc) 10

[Data prep 1](#_heading=h.49x2ik5)0

[Parameters](#_heading=h.2p2csry) 10

[Evaluation](#_heading=h.147n2zr) 11

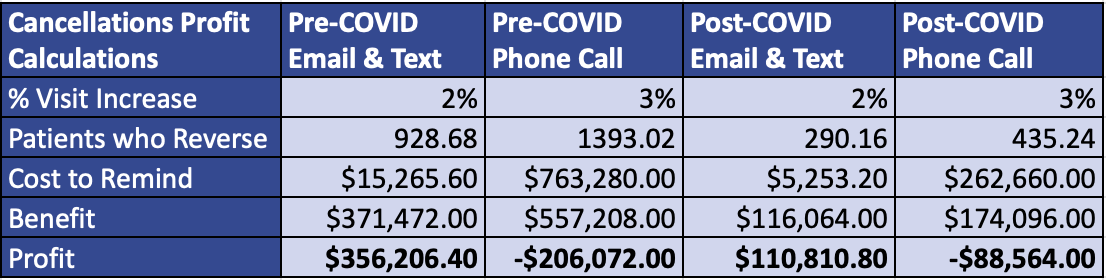
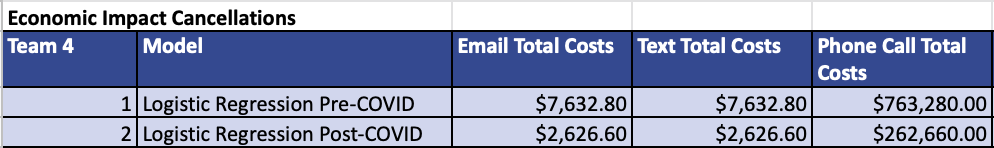
[Learning](#_heading=h.3o7alnk) 11

# Summary

This analysis focuses on increasing operational efficiency by providing insights on visit volume patterns and reducing costs associated with one of the main pain points the healthcare industry faces; cancelled appointments and no-show patients. This has great economic impact with an estimation of more than $150 billion of profit lost per year according to the US healthcare system[[1]](#footnote-0). After all, whether a patient attends their appointment or not, healthcare organizations still have to pay for the allocated resources and cover all expenses necessary. Our study using data from the specialty care units of the Children’s Hospital Colorado (CHCO) focuses on two key aspects: predicting visit volumes and appointment cancellations.

To predict visit volumes, we created 2 times series models: one using uninterrupted data from 2018 to 2020 and the other using a split dataset of pre-COVID and post-COVID appointment data to analyze the effects COVID-19 has had on patient visit volumes. Our second analysis consisted of predicting patient cancellations with pre-COVID and post-COVID data using logistic regression by looking at data from appointments that were actually cancelled and analyzing the predictive features of those appointments that may contribute to a patient cancelling. These models will assist in determining which patients to contact through email, text, or phone call reminders and the costs associated with each reminder method. The split data time series model and the two logistic regression cancellation models analyzed and accounted for changes in behavior due to the effects of COVID-19 with a cutoff date of February 25, 2020.

**Model Metric Summary**

**Economic Impact Summary - Cancellations**

## Economic Impact Summary - Visit Volume

## 

## 

## Recommendations

Based on the time series model evaluating the volume before and during the pandemic, the volumes of appointments seem to have remained fairly level so there are no recommendations for that. However, the pandemic has caused cancelled appointments to rise at a higher rate and completed appointments to rise at a lower rate. The Children’s Hospital should look at the trend of increased rate of cancellation and allocate costs accordingly. The logistic regression analyses provide more detail on how to combat cancellations.

In terms of preventing cancellations, we recommend using email and text reminders only as an estimated cost of $10 per phone call reminder is extremely costly in comparison to the profit saved from cancellations successfully prevented. Email and text reminders are relatively inexpensive and result in significantly higher profits when they are the only methods used. The costs of phone calls outweigh the profit saved from prevented cancellations using phone calls and ultimately decrease all profits from the reminders. If CHCO wants to use phone call reminders since they have a slightly higher probability of increasing visits from preventing cancellations, they must focus on decreasing the cost per phone call.

# Data

## Summary

The dataset used for our analysis was “SpecialtyCare\_03feb2021” which consisted of 31 variables and 764,834 rows of appointment data for the hospital’s specialty care unit. Each row contained data describing various features of the appointment such as specialty type, insurance type, patient’s age, patient’s home state, appointment length, appointment time, appointment date, etc. The appointment dates ranged from January 1, 2018 to September 30, 2020.

## Cleaning approach

In order to prepare our data, we grouped ‘No Show/Cancelled’ and ‘Left without Seen’ from the combined status variable as the ‘cancelled’ variable. This was used as the dependent variable for our logistic regression models, dummy coded as a ‘1’ indicating the appointment was cancelled and a ‘0’ indicating the appointment was not cancelled. For our logistic regression models, we removed columns that were not expected to be relevant to predicting cancellations. The remaining columns containing numeric values were converted to numeric data type. After removing columns, we omitted all rows containing blank values in any of the remaining columns resulting in 707,588 remaining rows. We then removed two rows in the dataset that contained repeated column names as their values. For the models looking at the data before COVID-19 and during, a cutoff date of February 15, 2020 was used to separate the data into new datasets.

# Model 1: Visit Volumes Time Series

## Description

In our first model of predicting visit volumes, we chose a time series model using simple exponential smoothing. We assumed the future visit volumes are related to the previous data. For example, if we have 450 visits on the first day and 470 visits on the second day, it is more likely to have visit volumes within a certain range surrounding 460 on the third day. There is about three years of data in the dataset which is enough to train the model to find an appropriate pattern.

## Data prep

In this model we used the dataset “SpecialtyCare\_03feb2021.csv” in R as a data frame. Then, we only kept the “checkin\_time” column and dropped the null values in it. Because the original check in time is stored as characters, we converted it to datetime type. Next, we counted the visit volumes by date and made a new data frame to store the output. Finally, we used the daily visit time as the explained variable and dates as predictors to build the time series model.

## Parameters

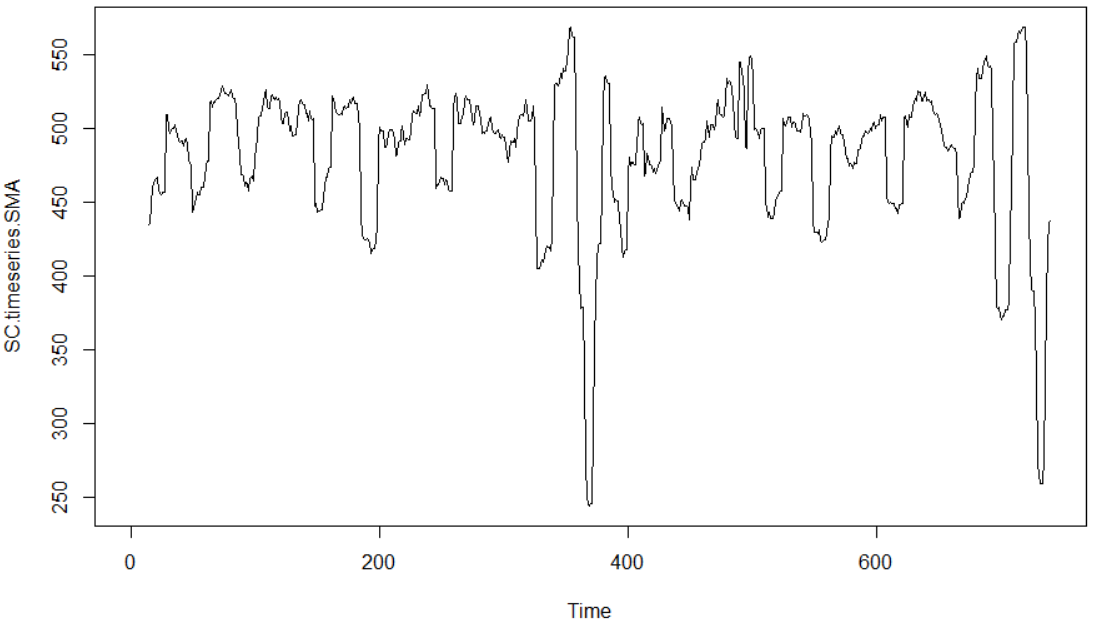
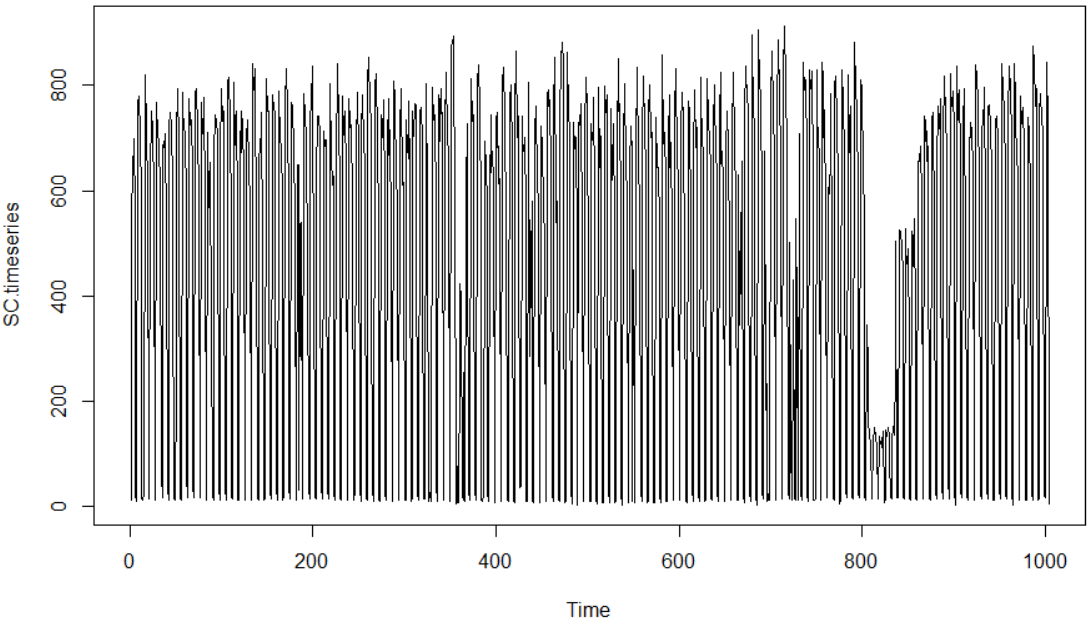
The only parameter in the model is time. We did not see a seasonality and trend in this model.

## Evaluation

|  |  |  |
| --- | --- | --- |
| Mean Absolute Error: 574.534 | Mean Squared Error: 111238.1 | Root Mean Squared Error: 333.524 |

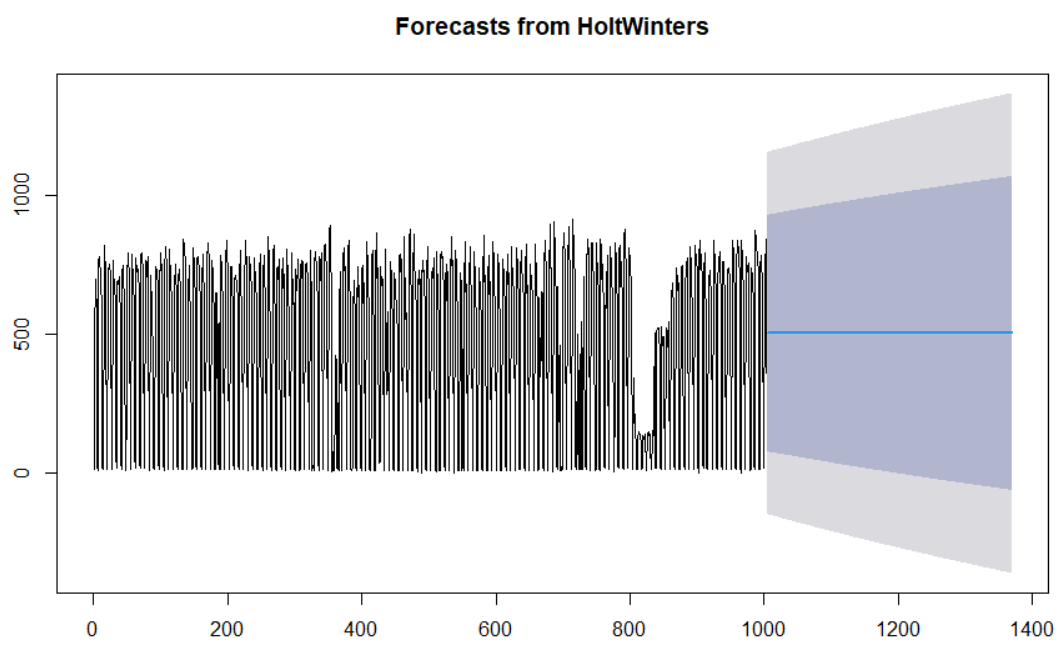
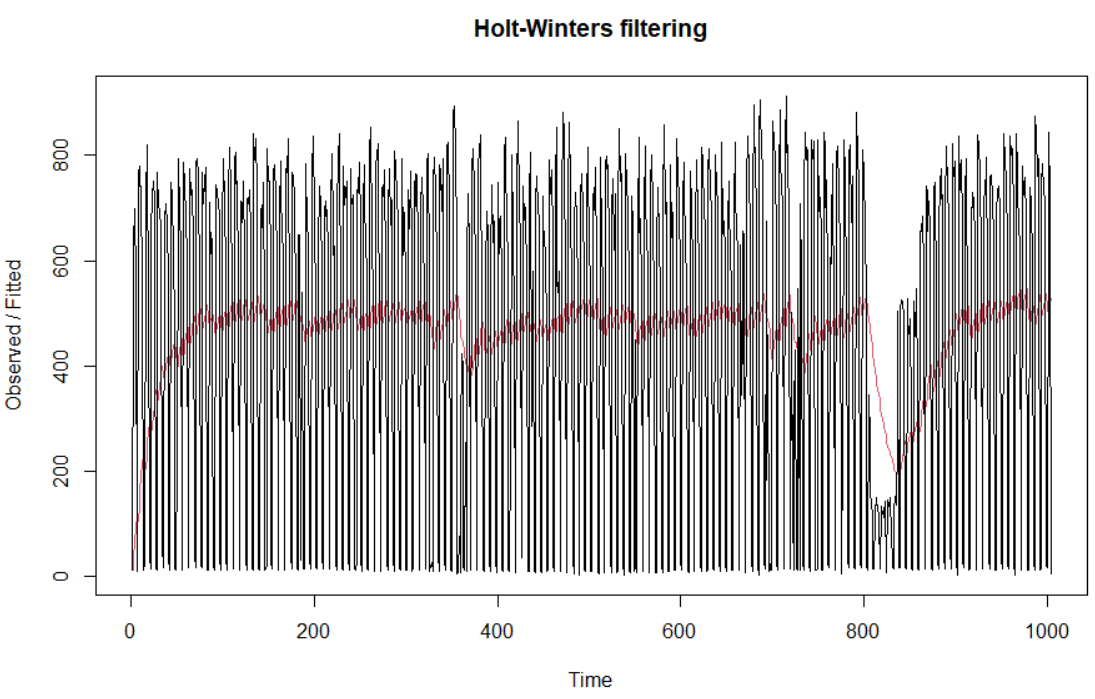
**Learning**

To have a big picture of the features for time series analysis on this dataset, we plotted a graph taking time as X variable and the number of visits per day as Y variable.



According to the left plot, we can tell the average daily visits is basically constant each year and no seasonality exists. These allow us to apply an additive model. To see if there is a trend component, we calculated the simple moving average of the time series with a span of fourteen and got a plot with smoother fluctuation.

From the right plot, we can tell that there is no clear trend of the fluctuations. If a time series can be described using an additive model with constant level and no seasonality, simple exponential smoothing is an appropriate method to make short-term forecasts. The smoothing is controlled by the parameter alpha. The value of alpha lies between 0 and 1. The closer it is to 0, the more weight is placed on recent observations for the prediction. Our alpha value is about 0.0456 which is very close to 0.



The red line on the left plot is the prediction of the model for the time window of the data set.

The plot on the right is the prediction of the future visit volumes within 365 days. The blue line represents the average daily visit volumes the model uses. The purple area represents what the volume may be under a 80% confidence interval, and the gray area is under a 95% confidence interval.

# Model 2: Pre-Covid and Post-Covid Time Series

## Description

## The Pre-COVID and Post-COVID time series models show visit volume trends within each time period. This way, we can evaluate how the volume should have been without COVID-19 and compare that with the actual volume. We are also able to evaluate how the volume should look based on habits during the COVID period to avoid previous outliers. The time series will display forecasts for future volumes overall and seasonally.

## Data prep

## After the initial data cleansing for all of the analyses, I added a column named “one” which would give sum-total volumes when aggregated by date. For the first part of the model, I created two new data frames for Pre- and Post-COVID values. I removed all columns except for “contact\_date” and “one” so I could group by the date and sum up the volumes, then set the date as the index. With limited data values, I resampled Pre-COVID values by week and Post-COVID values by day to create seasonality. I ran a for loop for each to find the best ARIMA variation by picking the variation with the lowest AIC. Now with the seasonality parameter figure, I was able to run this part of the model.

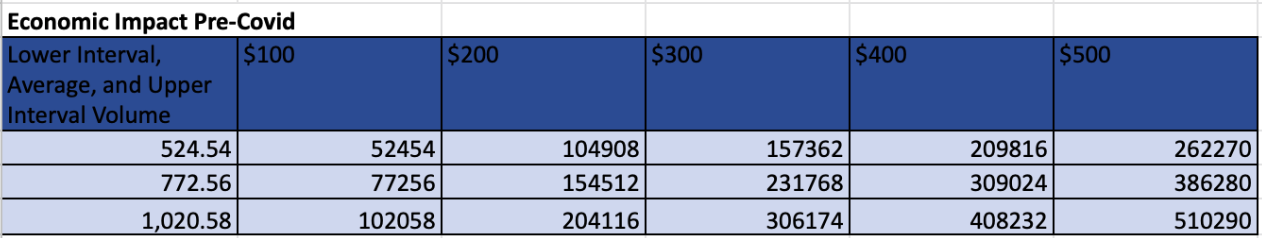
## For the second part of the model, I created six new data frames: two data frames each for Pre- and Post-COVID looking at cancellations and completions and two data frames merging the previous by time period. I removed all the same columns as before. The merged data frames only served to create the plots with volumes separated by status. With the other status-specific data frames, I used Prophet to create volume forecasts for each status in each time period. I then merged the status forecasts for each period to evaluate the forecasts side-by-side. With the forecasts set up, I was able to complete the model.

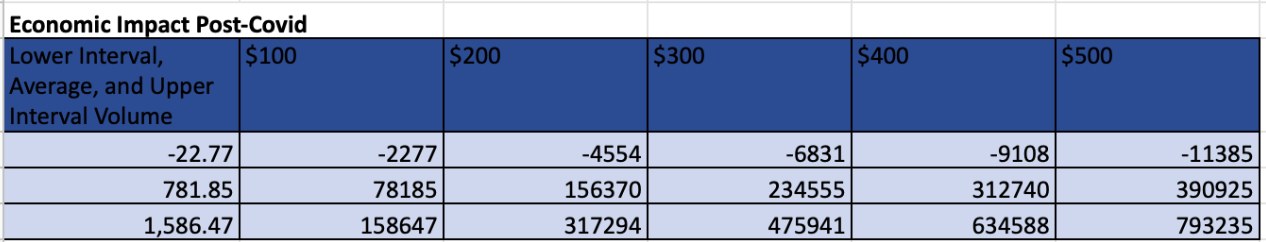
## Parameters

## The time periods for Pre- and Post-COVID were split on the date February 15th, 2020. The seasonality samples for Pre-COVID examples were by week and for Post-COVID examples were by day. The best ARIMA values for the Pre-Covid examples were (1, 1, 1)x(0, 1, 1, 12) and for Post-COVID examples were (1, 0, 1)x(1, 1, 1, 12). The volume predictions for the first part of the model looked ahead 60 steps.

## The second model future data frame looked at 50 periods for the Pre-COVID and 150 periods for the Post-COVID. The trend graph used the trend forecasts, and the estimates graph used the y-hat forecasts.

## Evaluation





|  |  |  |
| --- | --- | --- |
| Pre-COVID Volume MAE: 121.3 | Pre-COVID Volume MSE: 29938.29 | Pre-COVID Volume RMSE: 173.03 |
| Post-COVID Volume MAE: 318.77 | Post-COVID Volume MSE: 171587.12 | Post-COVID Volume RMSE: 414.23 |

**Learning**

Looking at the first part of the model, our Pre-COVID seasonality shows consistent volumes around 750 until certain times where the volume seems to drop. Our Post-COVID seasonality has a bit more of a variation since the sample was daily instead of weekly, however the volumes seem to be consistently higher than Pre-COVID, hovering above 1000. For Pre-COVID, it seems the only residuals are around the Christmas/New Years holidays. If we look at the one-step ahead forecast, we can see the trend disregard the outlier and continue to hover around the average volume. That results in the future forecast float between 650-900, averaging around 750-800. For Post-COVID, it seems that even though the values float from high to zero, the Normal Q-Q and the Histogram for seasonality seem to show consistency without too many outliers. This results in the one-step ahead forecast looking very similar to the observed volumes. The future forecast looks at the average of the graph, projecting that the volume would continue to hover around 800. After looking at both forecasts, it seems that COVID does not have too much of an effect on the average volume.

The second part of the model gave us insight about these time periods for cancellations and completed visits. For Pre-COVID, the seasonality Prophet charts show us a consistent trend of cancellations around 500. However, the Pre-COVID completed volume trends upwards, with each year showing higher averages. The year 2018 averaged around 260 completed visits, year 2019 averaged around 290, resulting in a forecast of the year 2020 averaging around 320. These observations are verified by the side-by-side forecast analysis. The trend line for cancelled visits only raises by 25 units, where the line for completed visits raises by 75 units. The volume estimates for each show how rapidly the completed appointments approach cancellations. The yearly seasonal pattern charts show consistency as well, with the only slight downward trend being from June to September for completed visits. For Post-COVID, the seasonality Prophet charts show us a large trend upwards in cancellations, going from an average of around 500 to an average of around 1000. The chart for completed visits also seems to trend upwards at a slower rate, going from an average of around 200 to around 500. The merging of the two forecasts again verifies what we have observed. The trend line for cancellations shoots up 100 units, where the line for completed raises 200. The estimates side-by-side show the averages being very similar until the future forecasts. Both charts have very similar seasonality patterns. They both seem to trend down from March until October.

# Model 3: Cancellations Pre-COVID

## Description

This model consists in predicting cancellations, focusing on appointments made before covid times. We began by creating a logistic regression using all variables, avoiding assumptions made about the relevance of any specific variable. Next, we randomly split the dataset into two subsets: pre\_test, and post\_tests. Pretest data was used to test our model and posttest to evaluate the significant coefficients. These partitions were randomly allocated using a ratio of 80:20.

Our logistic regression model was then used to create a confusion matrix helping us predict the appointment volume predicted to cancel, allowing us to estimate the cost benefit of targeting with email, text or a reminder call to people predicted to cancel.

## Data prep

To start with the data prep, the team focused on creating a dataframe including only appointments made before the COVID-19 virus had any impact in the health industry, making a cutoff date of February 15, 2020 resulting in 541,606 observations. We transformed categorical variables to their own binary column column with 0, 1 values, resulting in 39 variables. For the geographical column we decide to convert the column into a variable defining instate as 1 and out of state patients as 0. Also, we created 3 different columns bucketing appointments made in the morning, afternoon or after hours, with the idea that there could be a correlation between cancellations and the time of the appointment. After running the first model including all variables we opted to remove 22 variables that were not significant in the model including columns such as check out time or cancellation reason.

As we cleaned the data, we opted to delete any row that had NA’s after giving 0 values when applicable. NA’s in the dataset were a very small proportion of the data which had no significant impact when removing the NAs.

For the dependent variable we created a Cancellation column combining “No Show/Cancelled” with “Left without Seen” giving a more accurate representation of cancelled appointments and completed appointments.

## Parameters

The parameters that the team used to run our regression model were based on out of the most significant coefficients when running our first model with all the variables in our first test model. After running our first model we selected 24 significant independent variables with p-value of <.05 which are the base for our final regression model. Our significant variables were the following, weekend appointment, patient’s age, visit type, insurance types, including Medicaid, CHP, Medicare, selfpay, Gov insurance, and private insurance. Also, patient from Colorado, and the infusion center, rehabilitation, urology, neurological surgery, surgery, infectious disease, endocrinology, rheumatology, ophthalmology, nutrition, developmental peds, allergy and immunology, gastroenterology, nephrology, orthopedics, cardiology, neurology, pulmonology, otolaryngology, dermatology specialties.

To create our confusion matrix we randomly split the dataset into two subsets: pretrain, and pretest. Pretest data was used to test our model and posttrain to evaluate the significant coefficients. These partitions were randomly allocated using a ratio of 80:20. Our linear regression model was then used to create a confusion matrix helping us predict the appointment volume predicted to cancel, allowing us to estimate the cost benefit of targeting with email, text or a reminder call to people predicted to cancel.

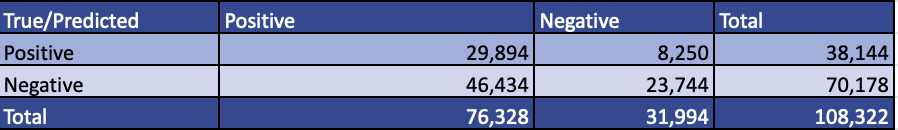
As mentioned before, for our dependent variable is a combination from the data column combine\_status merging “No Show/Cancelled” with “Left without Seen” to attain a more concise dataset.

When evaluating the final model, a confusion matrix was created using a probability threshold of 0.3 to maximize the sensitivity rate, since we want to include all potential cancellations in our selection to get those patients prioritized to send a reminder email, text, or phone call.

## Learning

It is evident that cancelled appointments in CHCO Denver can create a significant loss over a fiscal year. After running our model we are able to determine what variables significantly impact the cancellation of an appointment and what patients we should prioritize with an appointment reminder. The logistic regression model gave us a total profit savings of $356,206.40 when sending a text and email reminder, and a total profit lost of -$206,072 when doing phone call reminders. It is important to point out that the return on investment for targeting patients with a phone call is not favorable, since the cost per call is 100 times more than an email or text and only 2% more for the conversion rate. The total cost for reminder calls from our model was $763,280 outweighing revenues of $557,208 giving us a net loss of -$206,072. On the other hand, patients targeted with email or text only cost $15,265.60 resulting in a saved profit of $356,206.40. To prevent losses and maximize profits it's important to only use email and text as appointment reminders to effectively increase successful appointments in the CHCO Denver.

## Evaluation



|  |  |
| --- | --- |
| Sensitivity: 0.784 | Specificity: 0.338 |
| Precision: 0.392 | Accuracy: 0.495 |

# Model 4: Cancellations Post-COVID

## Description

A logistic regression analysis was conducted to predict the probability of an appointment being cancelled accounting for the effects of COVID-19. The binary column ‘cancelled’ was used as the predictor variable and consisted of either a value of ‘1’ indicating the appointment was cancelled or a ‘0’ indicating that the appointment was not cancelled. 39 independent variables were used to explain the potential relationship between cancellations and the appointments consisting of the independent variables.

The goal of the analysis was to produce a model that can classify an appointment as likely to be cancelled to assist in determining which patients to contact prior to the predicted cancellation. This model predicted saving $110,810.80 in profit from cancellations that were reversed due to the reminder emails and texts, as phone call reminders were too costly in comparison.

## Data prep

The binary ‘cancelled’ column was used as the dependent variable using the values ‘No Show/Cancelled’ and ‘Left without Seen’ from the combined\_status column. Using appointment times, we created dummy coded columns based on whether the appointment was a morning, afternoon, or after hours appointment. The following variables were also dummy coded to a ‘0’ or ‘1’ depending on their unique value: tele\_clinic\_visit, insurance\_type, and location. The state column was dummy coded to be a ‘1’ if the patient was from Colorado, and ‘0’ if the patient was from out of state.

Before dummy coding the specialty column, specialties that had less than 10,000 appointments were filtered out of the data to ensure the specialties had a higher chance of being significant. Columns that seemed insignificant in terms of cancellations or repetitive to the existing dummy columns were removed. The resulting dataset consisted of 39 columns and over 707,000 rows. The cutoff date of February 15, 2020 was used to subset the data for post-COVID cancellation analysis. The final dataset consisted of 165,000 rows of appointments.

## Parameters

The post-COVID data set was split into a training set containing 80% of the data to be used to build the models and a test set to validate the model containing the rest of the 20% of data. The first logistic regression model was run using all columns and resulted in 13 insignificant variables with p-values above 0.05. After dropping the insignificant variables, a second and final model ran consisting of the rest of the 26 variables which were all significant to the model with p-values below 0.05.

To evaluate the final model, a confusion matrix was created using a probability threshold of 0.3 to maximize the sensitivity rate, as contacting all possible cancellations was prioritized over ensuring only accurately predicted cancellations were contacted through a reminder email, text, or phone call.

## Evaluation

|  |  |
| --- | --- |
| Sensitivity: 0.886 | Specificity: 0.272 |
| Precision: 0.448 | Accuracy: 0.518 |

## Learning

## The following variables were significant in increasing the probability of an appointment being cancelled: weekend appointment, patient’s age, self-pay insurance, patient from Colorado, and the infusion center, rehabilitation, urology, neurological surgery, surgery, infectious disease, endocrinology, rheumatology, ophthalmology, nutrition, developmental peds, allergy and immunology, gastroenterology, nephrology, orthopedics, cardiology, neurology, pulmonology, otolaryngology, dermatology specialties. Due to the limited amount of data on appointments during COVID in the dataset, this model had significantly less data to train and test on. Increasing the amount of recent data to use with the model would allow for the model to optimize to the true effects of COVID and increase accuracy and precision.

## Based on our economic impact analysis, this logistic regression model successfully predicted and prevented cancellations using email and text reminders resulting in saving $110,810.90 in total profit. With a goal of prioritizing sensitivity in order to capture all possible cancellations, an estimated cost of $10 per phone call and $400 profit per appointment would result in the hospital spending significantly more on phone call expenses in comparison to profit saved from the actual prevented cancellations. Phone calls proved to be more costly than profitable in terms of preventing cancellations with a total cost of $262,660 to contact each predicted cancellation and only $174,096.80 in saved profits from cancellations successfully prevented by phone calls. On the other hand, email and text reminders combined only cost $5,253.20 and resulted in saved profits of $110,810.80 from successfully prevented cancellations. To prevent losing profit, only emails and texts should be used for reminding patients as they are cost effective to use in comparison to the actual number of prevented appointment cancellations.

1. (Gier, 2017 www.scisolutions.com/uploads/news/Missed-Appts-Cost-HMT-Article-042617.pdf) [↑](#footnote-ref-0)