Blue and Gold Health

Executive Summary

Team 15: Casey Tattersall, Sarah Zhang, Yuki Kitamura, Rishika Singhal

In this project, our goal is to help Blue and Gold Health analyze the year-over-year change in allowed dollars (as well as other related statistics) for 2022. We were given the results of an XGBoost machine learning model, which had been trained on part of the 2022 data, and used the results of the historical trend data and claims data to generate two alternative predictions for the trend of 2022. We believe that the following analysis can help Blue and Gold Health decide how to incorporate machine learning models in their future predictions.

Assumptions

Throughout our analysis, we made various assumptions, but we will address two especially important ones here. The background information did not clarify when the treatments that were deferred due to COVID-19 during 2021 would be carried out; for the sake of simplicity, we assumed that all deferred treatments would be done in 2022. Also, we assumed that the output of the XGBoost data (pred_trend_all) is referring to the predicted change in allowed dollars from 2021-22, but since the model's output did not come with any commentary or explanation, we cannot be sure of this.

Trend Results Analysis and Implications - Historical Data

Our general plan with the historical data was to generate predicted trends for both unit cost and utilization in 2022 that account for both the underlying trend and the event trend. We then combined these two values to estimate the yearly trend for allowed dollars (as utilization * unit cost = allowed dollars).

We were given year-over-year trends from 2014 to 2020 for each combination of LOB and benefit. For each combination, we manually removed outliers and conducted linear regression or calculated the mean to predict the trend for 2022. To look at aggregate trends (for example, the trend for one LOB across all benefit types) we combined these results using a weighted average constructed from the total allowed dollars each group accounted for in the 2021 claims data. After calculating the underlying trends, we then incorporated the event trends to create an adjusted core trend. While more detailed results can be found in our Excel and Python workbooks, this process resulted in a final prediction of a 9.9% year-over-year increase in utilization and a 5.4% year-over-year increase in unit cost, which assumes a 15.8% increase in allowed dollars (1.099 * 1.054 = 1.158).

Trend Results Analysis and Implications - Claims Data

We were given claims data from all of 2021 and the first half of 2022. To calculate year-over-year trends, we needed to estimate the data for the entire year of 2022. We did this by simply doubling the data from the first half of the year, as our seasonality analysis from 2021 concluded that there is no noticeable difference in claim volume between the first half and second half of the year. While we would like to conduct this analysis over multiple years to be sure, we were only given one year of complete data.

Now, with complete data from 2021 and 2022, we were able to calculate various statistics for each year (Util/K, Unit Cost, PMPM, Member Months, and Allowed Dollars) and calculate the difference between the two. However, since our claims data records what actually happened, this percentage change includes both the underlying trend and the event trends. We believed it would be better to separate the two, which we did by estimating the percentage of the overall claims that were caused by the special events, and letting the underlying trend equal the difference between the actual trend and the event trend (for example, in 2021, 9.2% of IP claims were deferred, so our normal underlying trend would be 9.2% larger than the actual trend that we calculated using the claims data). These isolations can be seen for Util/K and Unit Cost in Figures 1-4 in the appendix.

Benefits and Disadvantages of the XGBoost Model

XGBoost (Extreme Gradient Boosting) is a supervised machine learning algorithm used for regression, classification, and user-defined prediction problems on large datasets where the number of transactions and composition of groups are different. The model is highly efficient and fast on classification and regression tasks, which makes it a perfect fit for trend analysis. However, the model requires careful tuning of the hyperparameters, a process which can be time consuming and computationally expensive. Specifically, the number of trees (n_estimators) is very critical in terms of overfitting. The model is also sensitive to outliers and has limitations on sparse and unstructured data, so the performance of the model tends to decrease when the dataset has more noise (ex: extreme outliers like those in our data from 2020 and 2019).

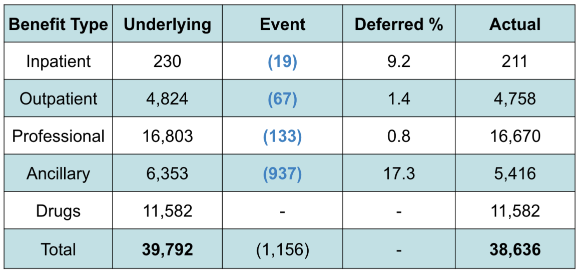
Recommendation

We believe that the XGBoost model has the potential to be incredibly useful in estimating trends in future years, but we have a few recommendations regarding how it should be implemented. First, since the model is especially sensitive to outliers, we recommend preprocessing the data and removing some of these outliers before training the model. Additionally, attempting to isolate the underlying trend and event trend for each year would be useful, as we don't want our future predictions to be affected by one-time events in the past. We also recommend increasing the size of the dataset, as the current model was run using only 3 or 6 months of data, which we believe makes it sensitive to overfitting.

Conclusion

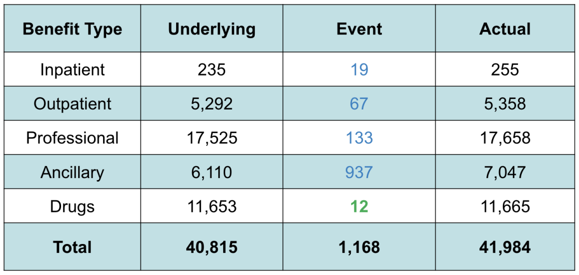
Our final comparison in our year-over-year trend for allowed dollars can be seen below in figure 5. It is important to note that the historical data and claims data model account for a decrease in utilization in 2021 due to deferred treatments, but the XGBoost model does not (as it was only using data from 2022). We accounted for this difference by increasing the predicted trend for each LOB by 3.8%, as we calculated that each LOB would experience a 3.8% decrease in utilization in 2021. This is the "adjusted XGBoost prediction" in the figure. While there is some variance in the results, this is to be expected as the three trends are calculated using different data and different methods. Since all three are relatively similar, we will conclude that the actual trend for the year 2022 is within the same range.

Appendix



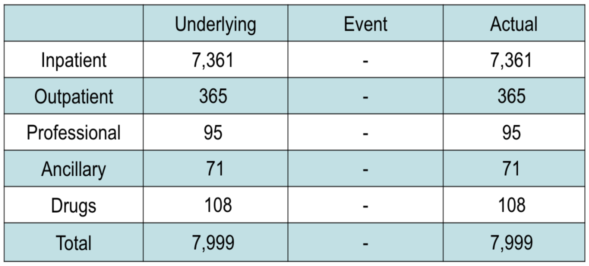
Benefit Type	Underlying	Event	Deferred %	Actual
Inpatient	230	(19)	9.2	211
Outpatient	4,824	(67)	1.4	4,758
Professional	16,803	(133)	0.8	16,670
Ancillary	6,353	(937)	17.3	5,416
Drugs	11,582	-	-	11,582
Total	**39,792**	(1,156)	-	**38,636**

Figure 1: 2021 Utilization/K



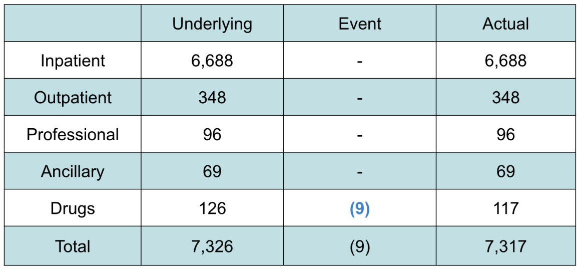
Benefit Type	Underlying	Event	Actual
Inpatient	235	19	255
Outpatient	5,292	67	5,358
Professional	17,525	133	17,658
Ancillary	6,110	937	7,047
Drugs	11,653	12	11,665
Total	**40,815**	**1,168**	**41,984**

Figure 2: 2022 Utilization/K



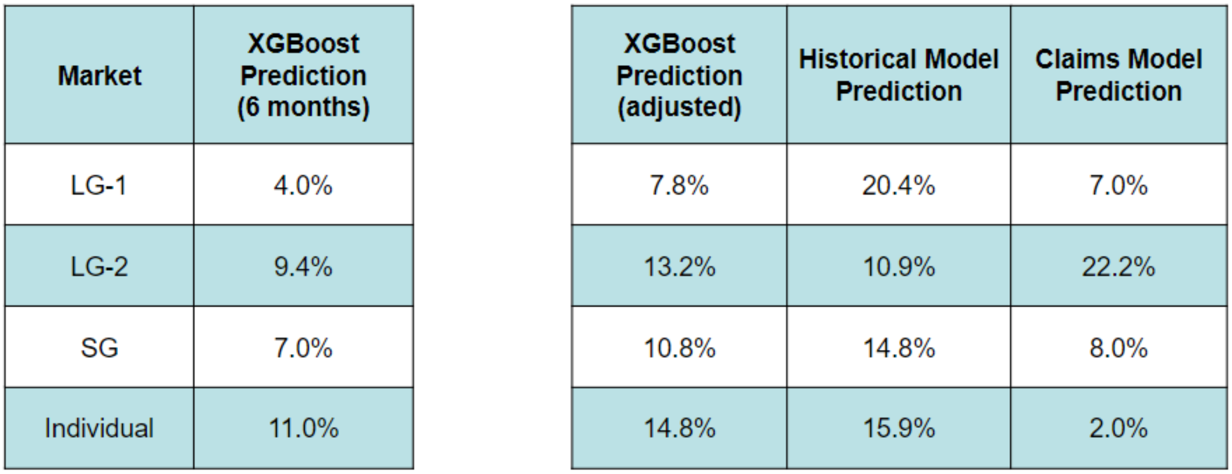
	Underlying	Event	Actual
Inpatient	7,361	-	7,361
Outpatient	365	-	365
Professional	95	-	95
Ancillary	71	-	71
Drugs	108	-	108
Total	7,999	-	7,999

Figure 3: Unit Cost in 2021



	Underlying	Event	Actual
Inpatient	6,688	-	6,688
Outpatient	348	-	348
Professional	96	-	96
Ancillary	69	-	69
Drugs	126	(9)	117
Total	7,326	(9)	7,317

Figure 4: Unit Cost in 2022



Market	XGBoost Prediction (6 months)
LG-1	4.0%
LG-2	9.4%
SG	7.0%
Individual	11.0%

XGBoost Prediction (adjusted)	Historical Model Prediction	Claims Model Prediction
7.8%	20.4%	7.0%
13.2%	10.9%	22.2%
10.8%	14.8%	8.0%
14.8%	15.9%	2.0%

Figure 5: Final Comparison (trend for allowed dollars).