Investigating Wellness and Rx Benefit Adherence Rates as Inputs for Forecasting Future Healthcare Insurance Claims Costs

Amogh Vig, in partnership with Brown & Brown Insurance

Executive Summary

In the insurance industry, it is well-known that providing preventive healthcare benefits to employees is correlated with future healthcare claims costs to employers. Nevertheless, the precise strength of this relationship is less clear. In this project, the strength of this relationship is explored by engineering metrics to quantify employee adherence for wellness and Rx benefits to investigate if these factors serve as useful inputs for predictive models to forecast future claims costs. The results show that an Autoregressive Integrated Moving Average (ARIMA) model shows promise for forecasting future claims costs based solely on historical information about wellness benefit adherence and corresponding costs. Although the ARIMA model does suffer from underestimating the ground truth, it still achieves a low relative RMSE of 0.05. This is an impressive result considering the lack of availability of data. With more historical information and/or the inclusion of more input features, the ARIMA model would be more performant and could prove to be an industrially viable tool for utilizing employee wellness benefit adherence rates for forecasting future healthcare claims costs for employers.

Introduction

As one of the nation's leading insurance brokerage companies, Brown & Brown Insurance provides value to clients by matching employers to optimal health insurance plans for their employees. Thus, a key point of consideration for the company and its clients is the relationship between employee adherence to preventive healthcare insurance benefits (Rx and wellness visits) and the costs incurred by clients due to healthcare insurance claims. In other words, it is desirable to use metrics that measure the extent to which employees use their preventive healthcare benefits as a feature for predicting the future claims costs to their employers. In the insurance industry, it is known heuristically that this relationship exists. However, the precise strength of this relationship and its potential use as a feature for predictive modeling and grouping clients into different categories remains a largely unexplored area. Therefore, this project aims to be a first step in bridging the gap between this intuitive knowledge of experts in the healthcare insurance space and empirical results. This is accomplished by applying data science techniques and machine learning algorithms to engineer metrics that measure employee preventive healthcare insurance benefit adherence and use this as an input feature for both claustering models and predictive models that forecast future employer healthcare claims costs.

Data

All the data for this project is sourced from the Benefits Science Technologies (BST) website. The data is organized by employer group and stored in the format of Tableau dashboards. The data was pulled from BST as CSV files so that they could be read and analyzed using Python's Pandas library. The data was a combination of numerical, categorical, and date variables. For this project, 8 different employer groups were studied. Each of these groups is a present client of Brown & Brown Insurance. To protect client identity, the client names were anonymized.

Outliers and missingness were not prevalent in the raw data. However, these issues did arise when engineering new features. Outliers were handled by imposing an appropriate maximum value on the data. Additionally, to avoid losing too much information, median imputation was used to fill in missing values, as opposed to simply dropping all null values. Another issue was

that the data proved to be noisy for modeling purposes. To circumvent this, the data was smoothed by computing a rolling 6-month average of the data, and the models were trained on this smoothed version of the data (for more information about the models and their results, please see the Methodology and Results section).

Ultimately, the biggest issue with the data was the quantity of the data. For most clients, only around two years' worth of data was available. This is quite insufficient for predictive modeling. Thus, only the client for which the most data was available was chosen for the predictive modeling portion of the project. Another issue with the data is that the feature names could arbitrarily be re-named when the BST website underwent an update. For example, the 'Person_ID' column could be changed to 'personid'. To address this issue and ensure the code for this project does not break due to inconsistent column names, a function was written to standardize all column names before further processing.

The most important features used in this project, their type, and their definitions are provided in the table below.

Feature Name	Data Type	Definition
Person ID	String	Unique identifier for individual employees.
Claim Type	String	Either Rx or Medical claim.
ICD-10	String	International Classification of Diseases (ICD) – codename to uniquely identify diseases.
NDC Drug Name	String	National Drug Code (NDC) – codename to uniquely identify pharmaceutical drugs.
Days' Supply	Integer	The no. of days that are supplied by an Rx claim.
Amount Allowed	Float	The cost of a claim in USD.
Start Date	Date	The date that a claim was filed.

Methodology

For the purposes of this project, where the focus is on preventive healthcare benefits, we defined this as being either one of these two claim types:

1. Rx claims – these are prescriptive drug claims incurred by employees to manage and treat chronic conditions.
2. Medical claims with ICD-10 codes starting with the letter "Z" – these claims correspond to wellness visits that aim to prevent future health problems (essentially healthcare check-ups), as opposed to visits due to procedures for a present affliction, such as having an operation.

For data preprocessing, functions were written to standardize column names, validate data type, and check for missingness. Missingness was addressed through median imputation. Once the

data was processed, a join was used to consolidate the data into data frames. These steps ensured that the data was ready for further analysis and modeling.

A key part of the analysis was feature generation. For measuring employee adherence for Rx claims, we computed the Medical Possession Ratio (MPR) using the Days Supply and Date features for each unique combination of Person ID and NDC Drug Name in the data. The formula for MPR is as follows:

$$MPR = \frac{Total\ Days'\ Supply\ of\ Drugs}{Total\ No.of\ Days}$$

Where the *Total Days' Supply* was computed by taking the sum of all the *Days' Supply* values in the given time window, and the *Total No. of Days* was similarly computed as the difference in days between the first and last dates in the given time window. An MPR of 1.0 indicates perfect adherence. Thus, outliers were addressed by capping the MPR at a value of 1.0.

The second engineered feature was Preventive Claim Ratio (PCR), serving as a metric to measure the extent to which employees were utilizing their provided preventive wellness visits. The formula for PCR is as follows:

$$PCR = \frac{No.\ of\ Preventive\ Medical\ Claims}{Total\ Medical\ Claims}$$

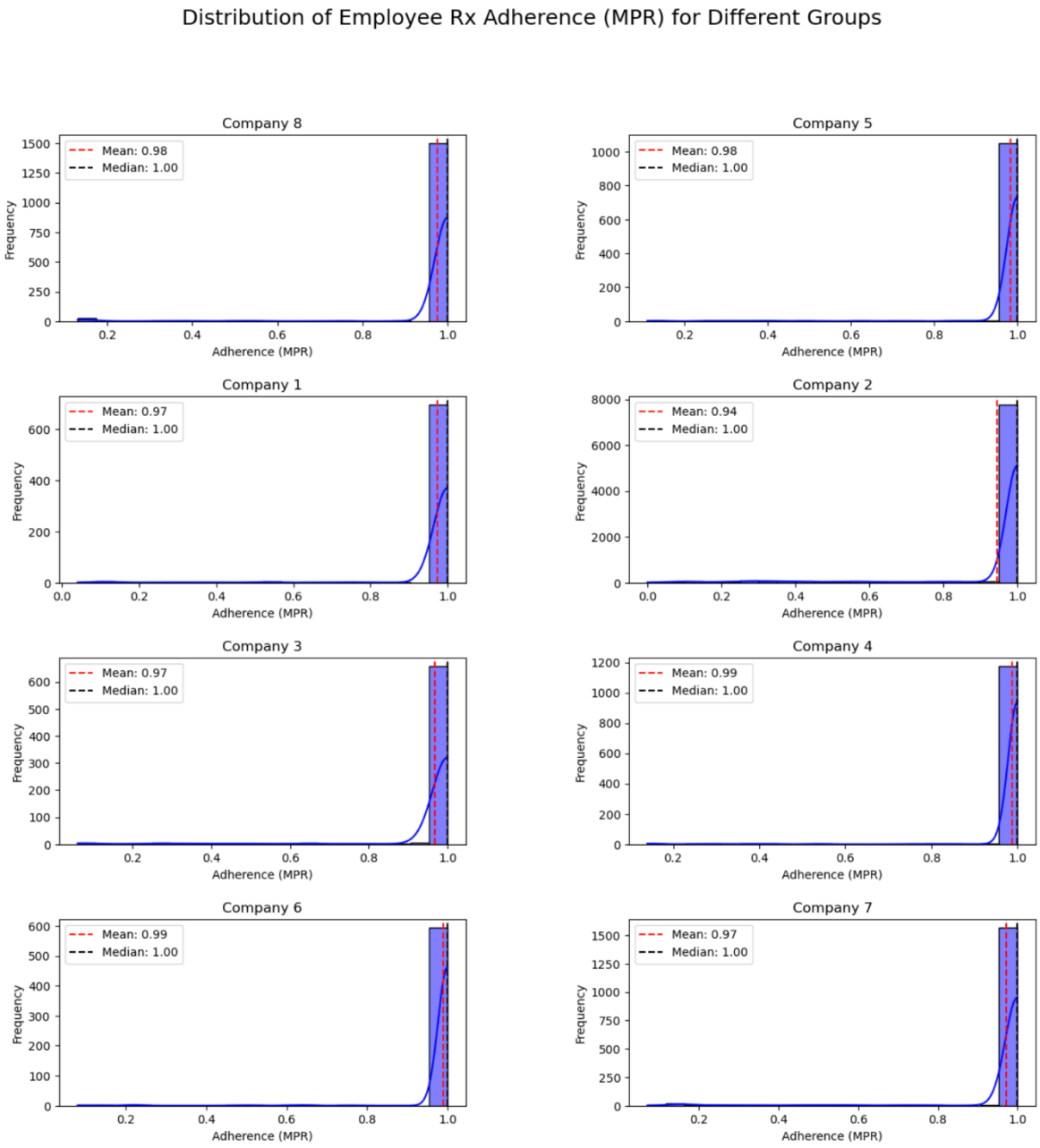
Where the *No. of Preventive Medical Claims* is calculated by counting the number of medical claims with an ICD-10 code starting with the letter "Z" in each month, and the *Total Medical Claims* is similarly calculated by counting the total number of recorded medical claims for that month. The engineered PCR data proved to be quite noisy for modeling purposes. This was addressed by taking a 6-month rolling average over the PCR values, as well as for the corresponding costs. The models were trained on the smoothed version of the data.

In addition to the main goal of using preventive benefit adherence for predictive modeling, we had a secondary goal to investigate whether this feature could also be used for training unsupervised models that can classify clients into different categories. A K-Means algorithm was implemented for this purpose. Key steps taken to prepare the data for this included normalizing the data to avoid bias, median imputation, and generating an Elbow Plot to determine the optimal number of clusters.

For the main project goal of forecasting future healthcare claims costs, two different time series models were implemented and compared. Namely, these models were an autoregressive integrated moving average (ARIMA) model and a Seasonal Autoregressive Integrated Moving Average Exogenous (SARIMAX) model. For both models, hyperparameter optimization is performed using a grid search approach to minimize the Akaike Information Criterion (AIC).

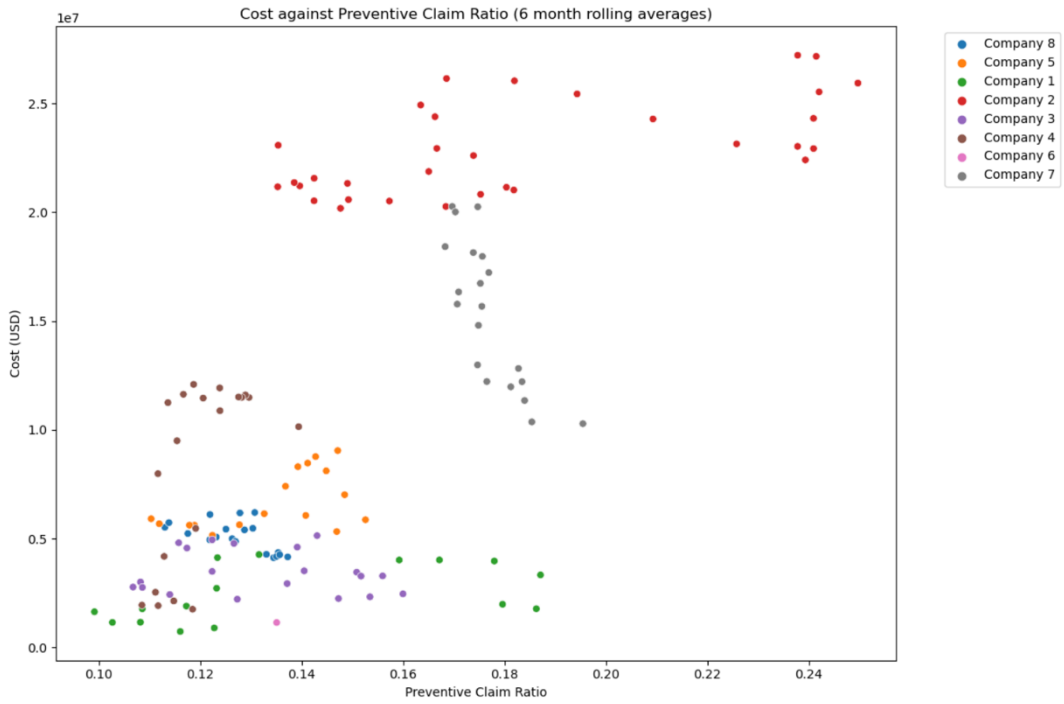
Results

The histograms below show the distribution of Rx adherence rates for the 8 clients.



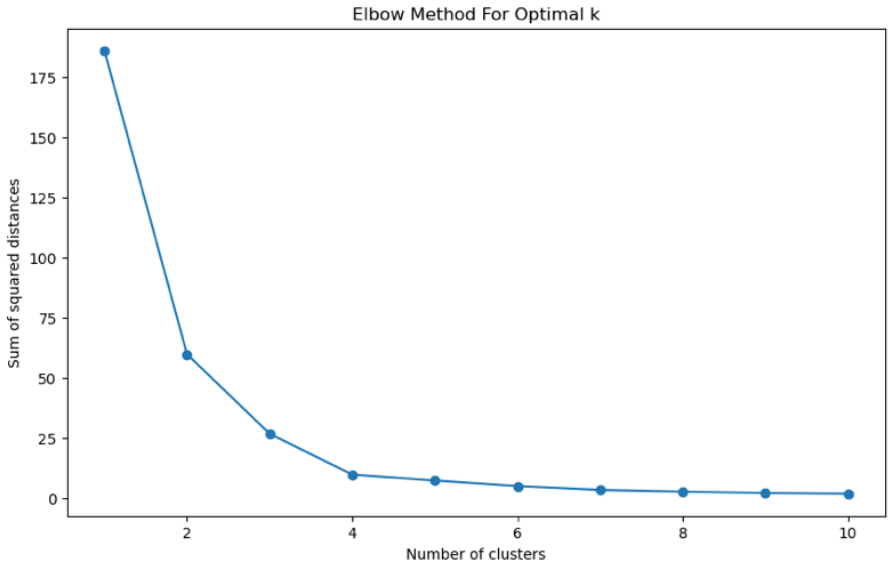
From these results, the distribution of employee Rx adherence is similar across the 8 clients. Thus, this feature is unlikely to be a useful input feature for our clustering model.

PCR proved to be the more useful feature for modeling. The following scatterplot shows PCR plotted against cost for the 8 clients.



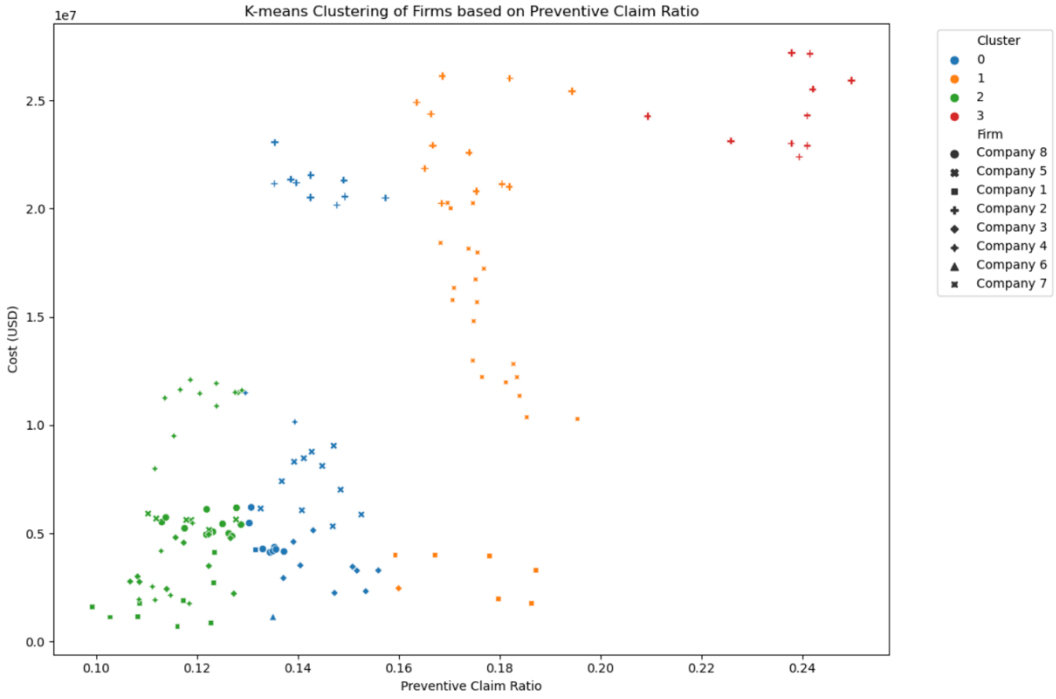
We can observe that different clients cluster in different locations, indicating that their differences are statistically significant. Although there is some overlap between the clients, the visualization provided evidence that they could be sorted into distinct groups based on their specific relationship between PCR and cost. Therefore, PCR was ultimately chosen as our sole metric for employee preventive healthcare benefit adherence, and this feature was used as the key input for both the clustering and forecasting models.

The results from the K-Means algorithm are presented below, starting with the elbow plot.



From the elbow plot above, we can discern that a choice of 4 clusters appears to be optimal for our data. The K-Means algorithm result using 4 clusters is presented below.

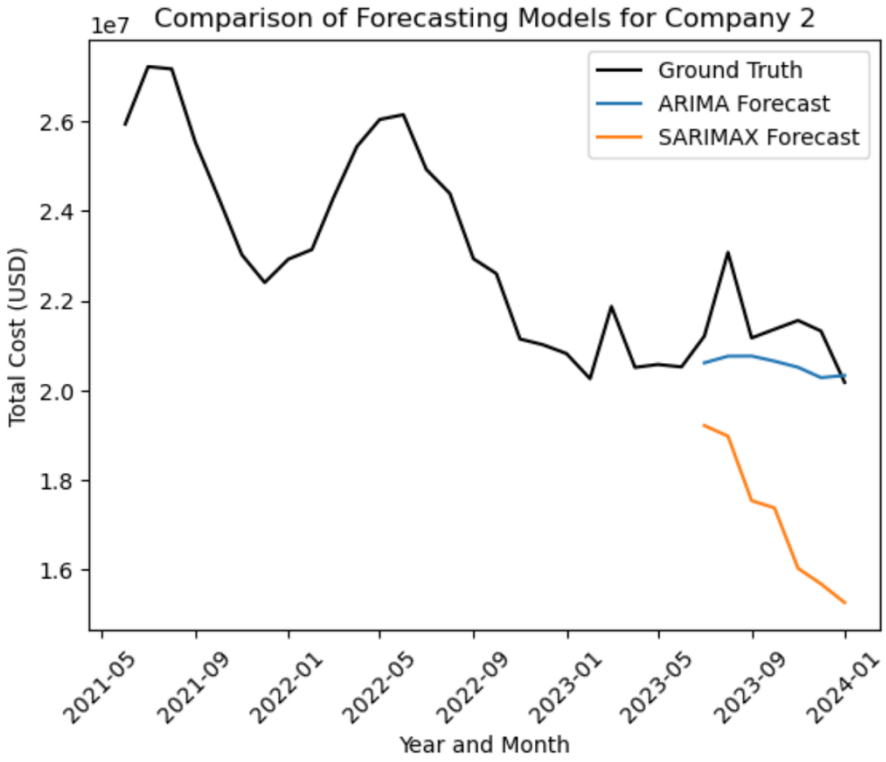
```
Silhouette Score for k=4: 0.69
Cluster     0   1   2   3
Firm
Company 1   6   6   9   0
Company 2  15  12   0  10
Company 3  14   1  10   0
Company 4   7   0  18   0
Company 5  16   0   6   0
Company 6   6   0   0   0
Company 7   5  20   0   0
Company 8  13   0  12   0
```



We can observe that the K-Means algorithm achieves a Silhouette Score of 0.69. This is a relatively high score since the scale ranges from -1 to +1. Therefore, this suggests that the clustering model has performed well for the given feature, and we can be confident in the distinction between the clusters found. Except for Company 2, the clients seem to be neatly sorted into the different clusters.

The results from the ARIMA and SARIMAX forecasting models are presented below. Due to data limitations, this phase of the project was conducted only for the client for which the most data was available (Company 2).

```
ARIMA RMSE: 1105941.98
SARIMAX RMSE: 4407349.95
ARIMA Relative RMSE: 0.05
SARIMAX Relative RMSE: 0.21
```



From the resulting visualization, both models underestimate the ground truth. SARIMAX appears to capture the overall trend better but suffers from more extreme underestimation of the ground truth, resulting in a high relative RMSE of 0.21. In contrast, the ARIMA model achieves a low relative RMSE of 0.05, despite underestimating the ground truth. Therefore, the ARIMA model has a comparatively high accuracy for forecasting future healthcare claims costs and is therefore the best model for this purpose. However, it is still desirable to have a lower relative RMSE for the purpose of financial forecasting in industry.

Conclusion and Next Steps

In this project, we successfully demonstrated that Preventive Claim Ratio could serve as a key input for training models to cluster clients into different categories and forecast future healthcare insurance claims costs. These models would solve Brown & Brown Insurance's business need of developing analytical tools that help brokers better serve clients based on their specific characteristics and future financial outlook.

For next steps, it would be highly beneficial to repeat this study but incorporate more information about the specific health insurance plans that employees are enrolled in. In this present work, employees were only grouped according to their employer. However, many employers offer more than one health insurance plan for their employees. Thus, conducting a more granular study that maps employees to their specific insurance plan and includes specific plan features, such as the plan type, deductible, copay, coinsurance, etc. as additional inputs to the clustering and forecasting models may yield superior results to what was achieved here. Additionally, future work could experiment with other types of clustering and forecasting models to investigate if there are alternatives to K-Means and ARIMA that perform better.