Project 1: Prediction of hospitalization days based on Electronic Health Records

Pablo Freyria Dueñas

February 22, 2021

Abstract

In 2006, it was found that unnecessary hospitalizations acounted for $30Bn USD in the U.S. alone. This motivated a competition in which total days in hospital are to be predicted for each patient based on their insurance claims from previous years. This study aims to provide a general overview of a leaderboard method, to build our own models for this primary objective and, as secondary objective, to determine whether sex plays a significant role on hospitalization rates.

1 Background

More than 71 million individuals are admitted annually to hospitals in the United States, and in 2006 it was found that over $30Bn USD were spent on unnecessary hospital admissions. In 2012, and sponsored by Heritage Health Price, Kaggle launched a competition in which an algorithm was to be developed to predict, based on available patient data, how many days will each patient spend on a hospital on the next year. This would help care providers to develop new care plans and strategies to reach patients before the emergency occurs, avoiding unnecessary hospitalizations. This report provides a brief explanation of a leaderboard, and develops an original method. As a secondary objective, this report studies whether the probability of being hospitalized is different for men and women.

2 Competition methods

The team "Market Makers" proposed a solution that involved, cleaning the data, building predictive models, and an ensemble learner. They modeled DIH both as a function of previous year data and of previous two years. A square error loss function was used where the outcomes were added 1 and transformed to the log scale. Four classes of algorithms were used in building the predictive models: gradient boosting machines, neural networks, bagged trees and linear models. Gradient boosting machines were found to be the

most accurate with a RMSE[1] of 0.461 and linear models the least with a RMSE of 0.466.

The individual predictive models were cross validated using n-folds, so each model had a cross-validated prediction for all data in the training data. This was used to develop a linear model that aimed to minimize the loss functions by a weighted average of individual models predictions. Overfitting was prevented by averaging the coefficients estimated with a randomly assigned sample of 50% of candidate models. The weights found in this process were applied to the log scale of the loss function. Ensembling improved the reported RMSE to 0.459.

Finally, these models only work when we expect the future to behave like the past, an assumption that will eventually stop to hold. They suggest using the historical data of important variables to identify mayor changes in their distribution and to build a model that excludes said variables to avoid overfitting to the changing past.

3 Own methods

3.1 Data description and processing

The original data covered three consecutive years and consisted on 2.7M claims of 113K members, who filled 820K drug prescriptions and took 360K laboratory tests. The target data is to predict hospitalization days by member ID and with an indicator of whether they had truncated claims[2] for year 4. Results from year 2 and 3 were used in training the model.

Data was distributed among 4 tables, all linked by a member ID variable with no missing entries. The 4 tables contained individual (that is 1 row per event) information about members, claims, drug prescriptions and laboratory tests. Additionally, 2 tables with the days in hospital (DIH) per member ID (one for each year) and the truncated indicator were used as the target data to train the algorithms used. A table with the same format was also provided to predict the DIH for Y4.

Out of the 113K members with information at least one year, 20.3K were missing data from either their age at first claim (5.7K missing) or their sex (17.6K). The sex variable was simulated based on its distribution given age (or overall if age also missing) and then age was imputed as the average given sex. Numerical variables originally coded as interval categories were transformed to the lowest value of the interval, and values over the 95% percentile[3] were truncated to the value plus one.

[1] In the log scale of outcomes

[2] Means that the dollar amount of claim was below the deductible, so no insurance was used

[3] Excluding zero

A final table for modeling was built by aggregating the individual information by member ID and year. Categorical data, such as claim condition group, was processed by adding a column for each category and putting the row count for each member ID and year combination. To avoid colinearity, the column with the lowest variance by variable was dropped. Numerical variables were averaged over the same member ID and year. The categorical variables related to the place of care (provider ID, vendor and PCP) were not incorporated in the final table as they would add 22K columns with this methodology.

Around 30% of members in which DIH on Y4 are to be predicted have not information for 2 consecutive years, so only data from the previous year was used to model DIH. This resulted in a final data table with 218K rows and 99 columns. 147K rows had observed outcomes that were used to predict DIH for 71K members on Y4. Model performance was measured using RMSE, without the log scale used in the competition methods.

3.2 Methods development and results

First, the final data table is split into the modeling (data with observed outcome) and predicting data. The modeling table is split into a training and testing data set (70% and 30% respectively). 85% of the modeling data corresponds to 0 DIH, so the split into training and testing was done to maintain this proportion. Then, the training data is used to screen for relevant variables[4] by building a linear model with main terms and then only with square terms. The 10% variables with lowest p-values were selected in each model to build a mixed model and ultimately select again the 10% variables with lowest p-values[5] and following the heredity principle[6] for the final model. This was based on 10 variables, and using only the main terms, achieved an RMSE of 1.51 in the testing data and predicted a total of 33,595 DIH for Y4.

Then, a tree regression model was built on the selected variables. The first node was split on 24 number of claims and then, on both branches, with the indicator of claims truncated. This model achieved a RMSE of 1.52 on the testing set and predicted a total of 32,601 DIH for Y4.

To answer the secondary objective a Chi square test was performed, with a resulting test statistic of 16.24 and a p value of 5*1e-5, so we reject the null hypothesis of equal hospitalization probability in favor of women being more likely to be hospitalized (14.5% vs 15.6%).

[4]Based on : "Bayesian analysis of definitive screening designs when the response is nonnormal". Victor Aguirre. DOI: 10.1002/asmb.2160 and a False Discovery Rate selection

[5]Sample size is too big and using an absolute significance threshold isn't meaningful

[6]All linear terms are included for square terms