PH290: Kaggle Heritage Health Prize Competition Project

Jin Rou New
SID: 25944841

1 Summary

The most important predictors of the number of days a patient will spend in hospital in the following year are: the patient's sex and age at first claim, total number of claims in the current year, number of pregnancy-related claims, hospital inpatient claims and emergency-related claims in the current year, pay delay of claim in the current year, number of unique health providers among the patient's claims in the current year. Both Random Forest (RF) and a Gradient Boosting Machine (GBM) were used to fit the model and the latter gives marginally better performance. My analysis also showed that the most important predictors of pay delay on claims are: Procedure Group (Pathology and Lab, Evaluation and Management, Surgery (genital system and integumentary system)), Specialty (Surgery, Diagnostic Imaging, Radiology), Primary Condition (Other renal diseases, Cancer B) and Place of Service (Office).

2 Introduction

The Heritage Health Prize Competition is a Kaggle competition in which the goal is to create an algorithm that predicts how many days a patient will spend in the hospital in Year 4, given previous years' claim data from members of the Heritage Provider Network. The accuracy of the algorithm is evaluated based on the root mean log squared error (RMLSE). In this paper, instead of predicting the number of days in hospital in Year 4 given data from Years 1-3, I will predict the number of days in hospital in Year 3 given data from Years 1-2. This is because the competition is closed to submissions and Year 4 data is not released so it is no longer possible to calculate the RMLSE for Year 4.

3 Data processing and feature engineering

The data is organized into a relational database with separate tables on member data, claims data, lab test data, prescription data and days in hospital data. The first step is to join all the tables into one and aggregating the data such that each row corresponds to a member instead of a claim. Categorical variables with ordinal values were converted to numeric ones (e.g. age 0-9 to 5). Missing values for quantitative variables were inputed by the mean of that variable for that year.

This data processing step occurs concomitantly with the feature engineering step, in which the available member-level information is summarized in a variety of ways, e.g. information from the drugs table by year were summarized into the following variables by year: total number of drug claims, total/mean/minimum/maximum/range of/standard deviation of/last drug count, claim-level vendor info was aggregated into number of unique vendors per member per year, claim-level Primary Condition Group (PCG) info for each was condensed into number of claims for each PCG per member per year. This resulted in a total of 143 member-level features for 76038 members in Year 1 in the training set and 71435 members in Year 2 in the test set.

4 Proposed method

I applied both a Random Forest (RF) and a Gradient Boosting Machine (GBM) to the processed data set. Relative importance of the top 20 most important predictors of both models are given in Figure 1. The following predictors belong in the top 20 most important predictors for both models: Sex, PrimaryConditionGroupPRGNCYCount, PayDelayMean, AgeAtFirstClaimNum, PayDelayTotal, PlaceSvcInpatientHospitalCount, ClaimCount, ProviderIDCount, SpecialtyEmergencyCount. Most of these make sense, although it is odd that PayDelay would be an important predictor; though this could be related to some underlying confounder.

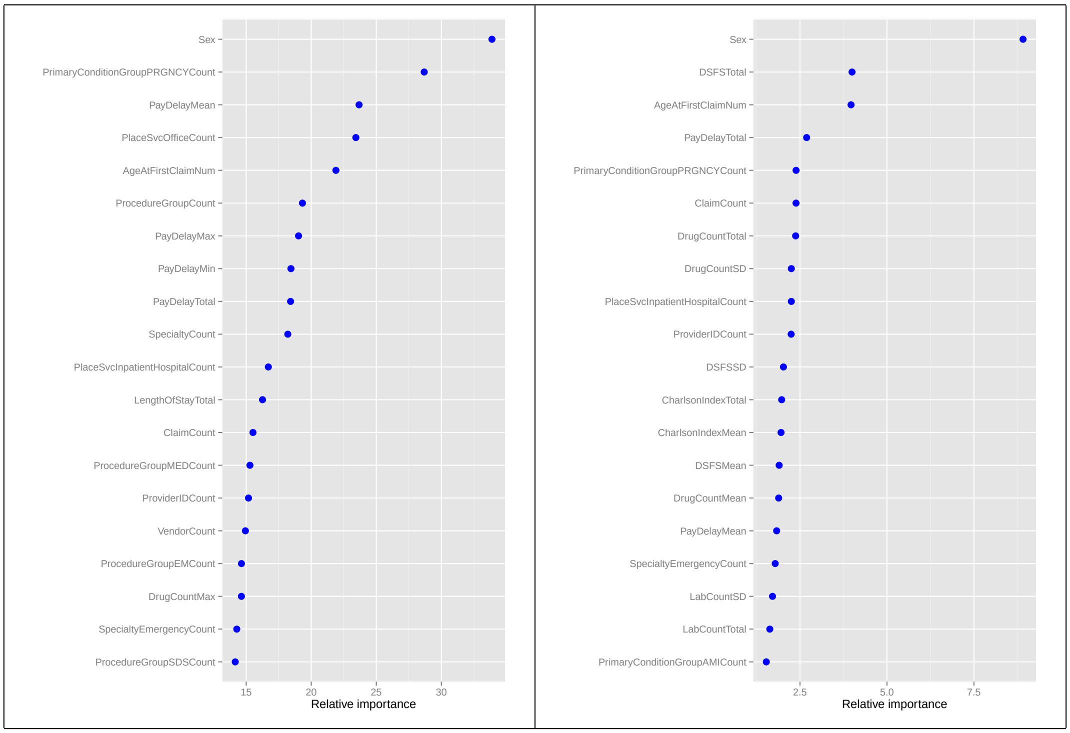

Figure 1: **Variables ranked by variable relative importance for the models Random Forest (left) and Gradient Boosting Machine.**

5 Method benchmarking

All winning methods employed a combination of many different models. Models are combined by blending and the final predictions are linear combinations of the predictions of all models in the blend, with weights given based on a ridge regression performed on the leaderboard scores (e.g. [1-3]). The base models that make up the final combination include: Stochastic Gradient Descent Models [1-3], Gradient Boosting Machines [2-5], Neural Networks [4-5], Linear Models [4-5], Tree Ensemble [2-3], Random Forests [4-6], Additive Groves [5], Multivariate Adaptive Regression Splines [5], Regularized Greedy Forests [6], Gradient Boosting Decision Trees [6] and optimized constant value models [2-5].

A fair comparison between my method and the winning methods was not possible given that the winning methods used all available data to predict the days in hospital for Year 4 while I used data up to Year 2 to predict the days in hospital for Year 3. Moreover, I was unable to reproduce exactly the winning methods to check their performance with Year 2 data on Year 3 predictions.

Instead, we can compare my methods against the base models of 1) setting the prediction to 0 days for all patients 2) setting the prediction to the mean of days in hospital in Year 2 for all patients. The first model gives an RMLSE of 0.4262385 while the second gives a RMLSE of 0.3871215. In comparison, the RF results in a RMLSE of 0.3771039 while GBM gives 0.3765514.

6 Additional analyses

I was interested in the pay delay on claims. The mean pay delay is about 43 days (assuming no pay delay more than 162 days), which should be possible to optimize for greater operational efficiency and member satisfaction. Applying a GBM on claim-level data and looking at variable relative importance, I found that the 10 most important predictors of pay delay are (as shown in Figure 2): ProcedureGroupPL, ProcedureGroupEM, SpecialtySurgery, PrimaryConditionGroupRENAL3, SpecialtyDiagnosticImaging, ProcedureGroupRAD, PrimaryConditionGroupCANCRB, PlaceSvcOffice, ProcedureGroupSGS and ProcedureGroupSIS.

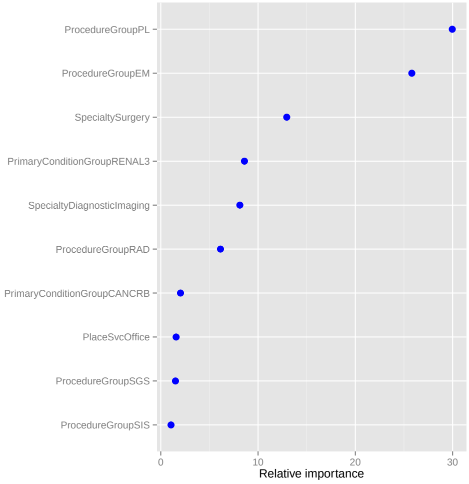

Figure 2: **Variables ranked by variable relative importance for the Gradient Boosting Machine on claim-level pay delay data.**

7 Discussion

While the RF and GBM perform better than the base models, there is likely still room for improvement, such as parameter tuning using cross-validation. More models such as Generalized Linear Models, Neural Nets etc. can be fitted and a mix of the models can be blended to yield a lower RMLSE.

8 References

1. Mestrom, W. (2011). My milestone 1 solution to the Heritage Health Prize.

2. de Grijs, E. and Mestrom, W. (2012a). Our milestone 2 solution of the Heritage Health Prize.

3. de Grijs, E. and Mestrom, W. (2012b). Our milestone 3 solution of the Heritage Health Prize.

4. Brierley, P., Vogel, D. and Axelrod, R. (2011). Heritage Provider Network Health Prize Round 1 Milestone Prize: How We Did It – Team 'Market Makers'

5. Brierley, P., Vogel, D. and Axelrod, R. (2012). Heritage Provider Network Health Prize Round 2 Milestone Prize: How We Did It – Team 'Market Makers'

6. Johnson, R. and Zhang, T. (2012). Heritage Provider Network Health Prize Round 3 Milestone: Team crescendo's Solution.