Predicting Annual Length-Of-Stay and its Impact on Health Costs

Álvaro J. Riascos Villegas [1] Natalia Serna [2]

[1]Universidad de los Andes and Quantil

[2]University of Wisconsin and Quantil

Contents

Contenido

Introduction

- Avoidable hospitalizations are a source of increased health expenditures in many health systems.
- In the United States it is estimated that in 2006 well over 30 billion was spent on unnecessary hospital admissions (Health Heritage Prize announcement).
- To highlight the relevance of this problem consider the Health Heritage Foundation Prize Competition that offered 3 million dollars for the best predictive algorithm of future hospitalizations.

Introduction

- Though we don't have reliable estimates for the potential savings of reducing avoidable hospitalizations in Colombia, most likely they are considerable.
- We use a nice administrative dataset covering almost have of the Colombian population (POP = 48 million) of the statutary health care system (public health system) recording all services, hospitalizations, etc. during 2009 - 2011.
- Our contribution has two complementary elements:
  1. We built a machine learning algorithm that predicts fairly well next year length of stay (compared to, for example, the Health Heritage winning team)
  2. We introduce a decision model that allows to evaluate the cost effectiveness of prevention programs of patients with high risk of hospitalization.

Introduction

- Results from the predictive modelling show tree-based methods outperform linear approximations and achieve lower out-of-sample error rates compared to the winning model of the Heritage Health Prize.
- We also show that a prevention program where patient intervention is decided upon the predictions of the model can achieve significant cost savings relative to the best uniform policy (i.e, intervene all patients or no intervention).
- This holds for program efficacies greater than 40 % and intervention costs per patient ranging between $\$100,000$COP (USD$35) and $\$700,000$COP (USD$250).

Contenido

Related Literature

- Most of the literature on prediction of annual LOS has been developed from the providers' perspective rather than from the insurers' perspective.
- Authors predict LOS using a sample of patients with specific acute conditions or physiological traits that are often unobserved by the insurer:

1. Chang et. al 2016. Prediction of Length of Stay of First-Ever Ischemic Stroke.
2. Ali, et.al. Predicting Hospital Length of Stay PHLOS: A Multi-Tiered Data Mining Approach.
3. Chertow. 2005. Acute Kidney Injury, Mortality, Length of Stay, and Costos in Hospitalized Patients.
4. Clague, et.al 2002. Predictors of outcome following hip fracture. Admission time predicts length of stay and in-hospital mortality.
5. Rezaei, et.al 2013. Use of Data Mining Techniques to Determine and Predict Length of Stay of Cardiac Patients.
6. **Bayati, et.al 2014. Data-Driven decisions for reducing readmissions for heart failure: general methodology and case study.**

Related Literature

- Summing up: Our study differs from the previous ones in the sense that we predict annual LOS using information that is symmetrical between insurers, providers, and the government.
- We do not focus on users with particular health conditions but analyse a representative sample of individuals in the public health care system with heterogenous demographic and morbidity characteristics.

Contenido

Data

- To predict patient LOS and estimate its impact on health costs, we have the yearly claims of a sample of 5.7 million enrollees in the contributory system during 2009 to 2011.
- Information per individual includes: insurer to which she is enrolled, claims, provider ID, cost per service, date, diagnosis identified with the International Classification of Diseases (ICD) Codes, length-of-stay per claim, age, gender, and municipality of residence and income.
- We randomly selected 1 million patients for training and 1 million patients for testing.

Data

- We constructed several features with information from $t-2$ to $t-1$:

  Annual LOS, average LOS, maximum LOS, second maximum LOS, indicator of annual LOS greater than 30 days, standard deviation of LOS, average cost, standard deviation of cost, average income of enrollees in each insurer, standard deviation of income in each insurer, indicators of the 10 costlier diagnoses in the sample, number of hemograms, pressure tests, CTs, creatinine tests, thyroid tests, ER services, ambulatory services, hospital services, domiciliary services, drug claims, and the number of different long-term diseases affecting each patient. We also create the number of claims per month and per day of week, indicators of long-term diseases, and interactions between indicators of hospital services, ER services, domiciliary services and ambulatory services.

Contenido

The Machine Learning Task

- Task: predict patient LOS in year t with claims data of $t-1$ and $t-2$. The dependent variable is $\ln(LOS+1)$ in year t.
- A similar problem was posed in the Health Heritage Price. Milestone winners used ensembles consisting of linear combinations of boosted trees models, random forests, artificial neural networks and linear regressions, restricting the sum of coefficients to 1 and truncating negative predictions.
- Models were compared and evaluated using the Root Mean Squared Error (RMSE). The winning team achieved an out-of-sample RMSE of 0.4438 which is, nonetheless, 2.5 times the average log LOS of the third year of data.

Results: prediction length of stay

Cuadro: Out-of-sample model fit

Model	MAE	RMSE	R-squared
OLS	0.4546	0.7502	0.1731
ANN	0.5032	0.7824	0.1006
RF	0.2634	0.5623	0.5354
BT	0.2721	0.5720	0.5192
ENS	0.2523	0.5609	0.5179

This table shows the out-of-sample MAE, RMSE, and R-squared of different models. OLS: Ordinary Least Squares, ANN: Artificial Neural Networks, RF: Random Forest, BT: Boosted Trees, ENS: Linear ensemble. Source: Base de Suficiencia, Ministry of Health and Social Protection. Authors' calculations.

- The RMSE of the linear ensemble is 75 % de average $ln(LOS + 1)$. While the best model in Health Heritage Prize was 249 %

Results: prediction length of stay

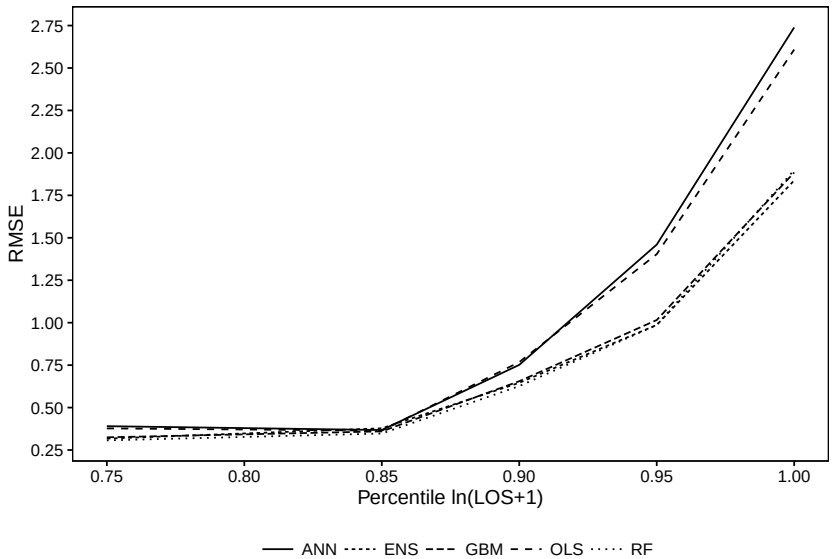

Figura: Variation in the RMSE by percentiles of the LOS distribution

Source: Base de Suficiencia, Ministry of Health and Social Protection. Authors' calculations.

Results: Classification problem

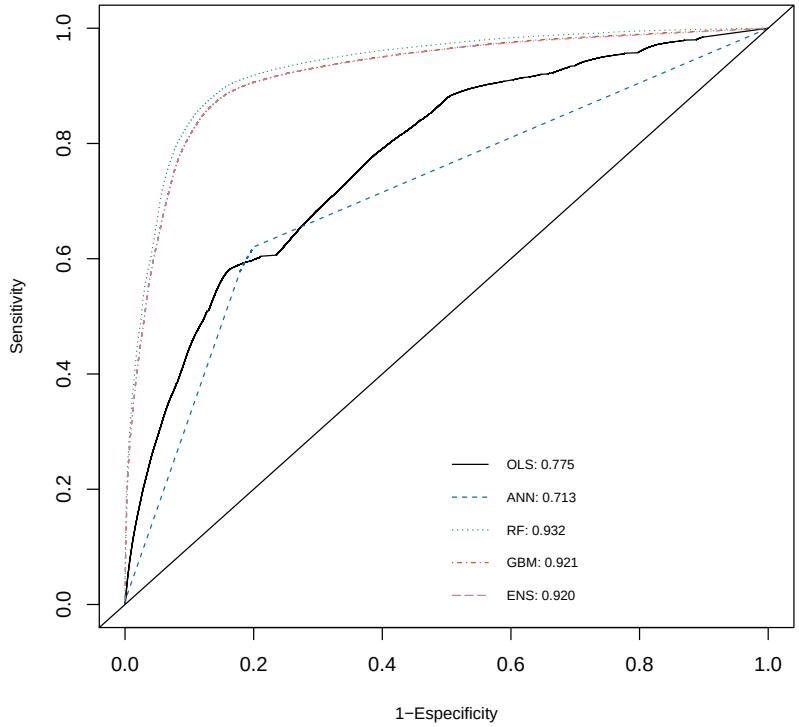

Figura: Prediction accuracy

Source: Base de Suficiencia, Ministry of Health and Social Protection. Authors' calculations.

Results: Variables relative importance

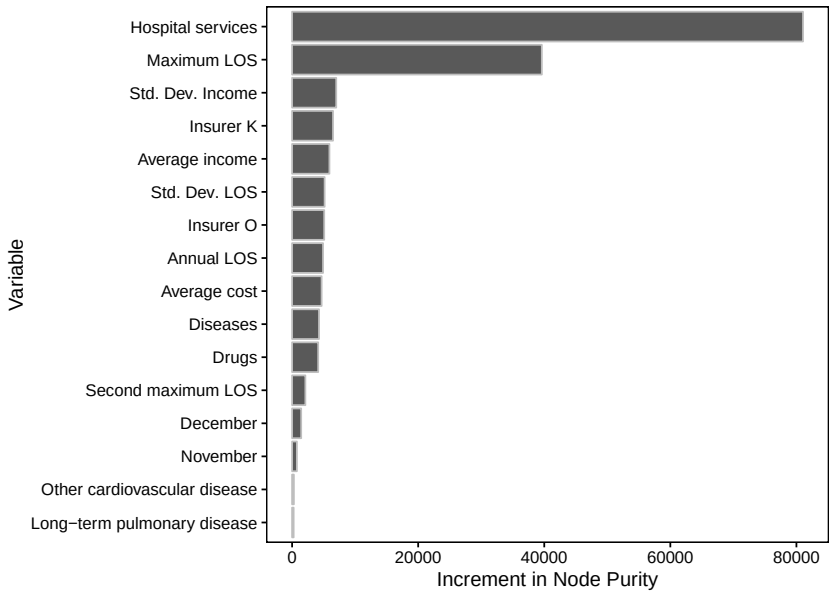

Figura: Risk factors in the random forest model

Source: Base de Suficiencia, Ministry of Health and Social Protection. Authors' calculations.

Contenido

Modelling Framework: Decision model

- Following Bayati, et.al 2014 the expected cost of hospitalization for patient i is the product between the probability of being hospitalized and the cost of hospitalization in the risk pool she belongs to:

$$C_0(\hat{p}_i) = \hat{p}_i c_g \tag{1}$$

- If insurers undergo the prevention program for this patient, the probability of being hospitalized decreases with the program's efficacy, but hospitalization costs increase linearly with the cost of the intervention per patient. If intervened, expected health costs are:

$$C_1(\hat{p}_i) = (1-\alpha)\hat{p}_i c_g + f \tag{2}$$

- Thus, a patient must be intervened if

$$\pi(\hat{p}_i|\alpha, f) = C_0(\hat{p}_i) - C_1(\hat{p}_i) \geq 0 \tag{3}$$

Modelling Framework: Decision model

- To measure the incremental cost-effectiveness of the prevention program, we compare the costs generated by a program where intervention is decided upon the inequality in expression with two base scenarios (given α and f):
  1. The *no-intervention policy*.
  2. The *best uniform policy*.

Results: Potential cost savings of a prevention program

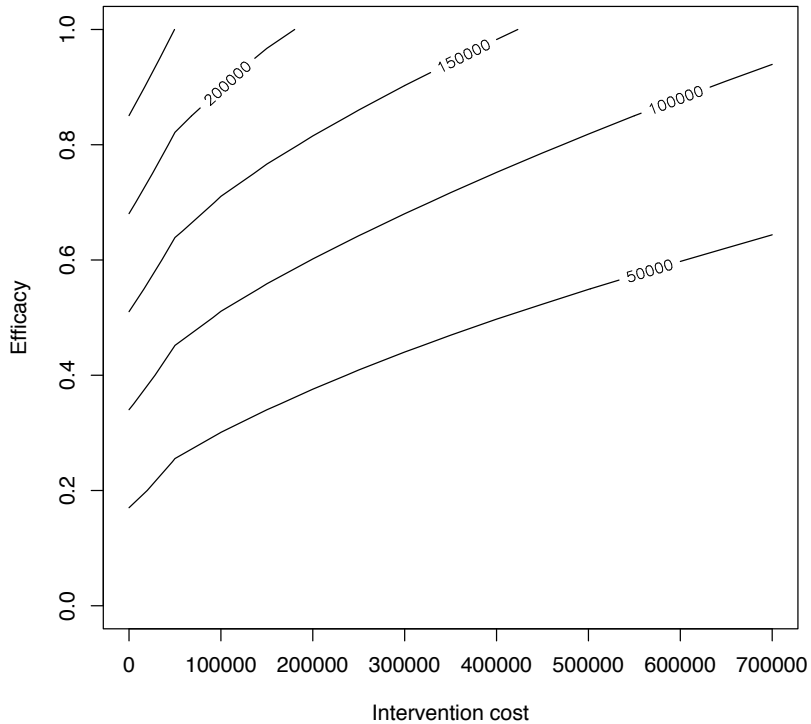

Figura: Cost savings over no-intervention policy

Source: Base de Suficiencia, Ministry of Health and Social Protection. Authors' calculations.

Results: Potential cost savings of a prevention program

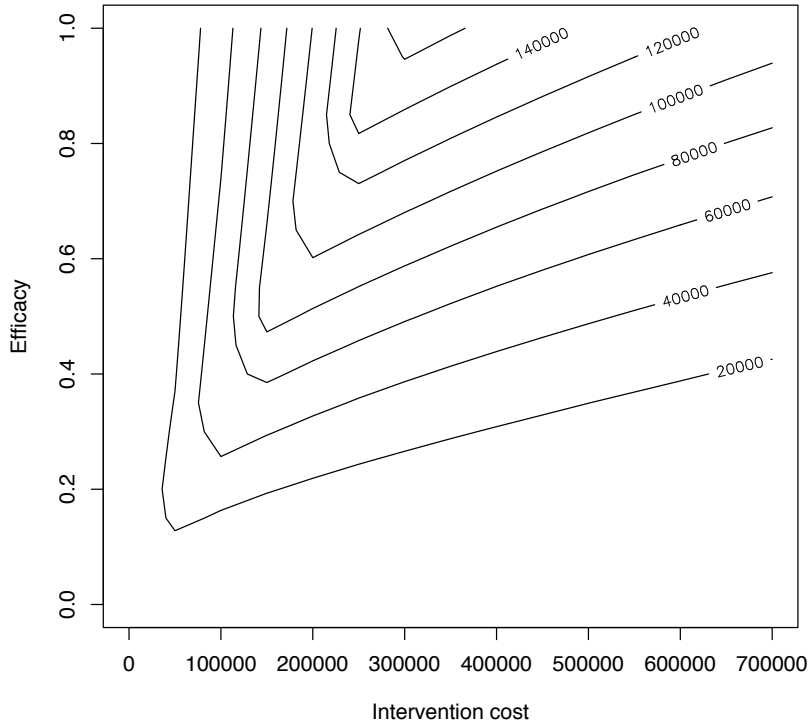

Figura: Cost savings over best uniform policy

Source: Base de Suficiencia, Ministry of Health and Social Protection. Authors' calculations.

Contenido

Conclusions

- It is possible to use ML algorithms to predict length of stay or futuree hospitalizations using currently available data.
- Using this model it is possible to inform decision makers and allow them to make better decisions regarding health prevention programs.
- This methodology could greatly reduce public health costs in Colombia.