APPLYING DEEP LEARNING TO HOSPITALIZATION DATA

Springboard: Capstone 2

Ivette M Tapia

TABLE OF CONTENTS

No headings found.
This is an automatic table of contents. To use it, apply heading styles (on the Home tab) to the text that goes in your table of contents, and then update this table.

If you want to type your own entries, use a manual table of contents (in the same menu as the automatic one).

INTRODUCTION

[INTRO HERE]

MOTIVATION & BACKGROUND

Healthcare records have become increasingly more digitized. This is an open an opportunity to analyze and obtain patient data at an unprecedented detail and scale. While there is potential to gain greater insights, cost reduction and efficiencies in the healthcare space exist, great challenges remain. Some of these include data availability and complexity of services. Healthcare is particularly complex due to overlapping systems, diversity of providers, services and health issues.

This project would use hospitalization data from Brazil to make predictions about key features and ouctomes of an hospitalization request. Specifically, a deep learning will be used to predict: (1) procedure(s) performed, (2) hospitalization costs, and (3) hospitalization lenght given the information available in the approval provided in the hospitalization

Ability to accurately predict these three features of hospitalization can yield significant benefits. For example, knowing how many days a patient can be expected to be in the hospital will help hospital managers manage their capacity (especially in areas where beds are scarce). Length of stay and likely procedures can inform service charges and help all parties involved navigate the healthcare charge system better so likely costs are known in the front end.

Moreover, predicting healthcare expenditures can be tricky for insurers, providers and particularly consumers. One of the main factors that have been cited as a cause of rising healthcare expenditures is the inability of consumers to know in advance the cost of the healthcare services they consume.

The data that will be used is from the Authorization for Hospital Admission. This dataset is part of Brazil's SIHSUS Hospital Information System. This system manages the coordination and payment by Brazil's public healthcare system (covers around 34% of

Brazil’s population and pays for 80% of all hospitalizations). The data is publically available as .dbc files. In this application, I will be using data from 2015 – 2018. This represents 3.5 years’ of data.

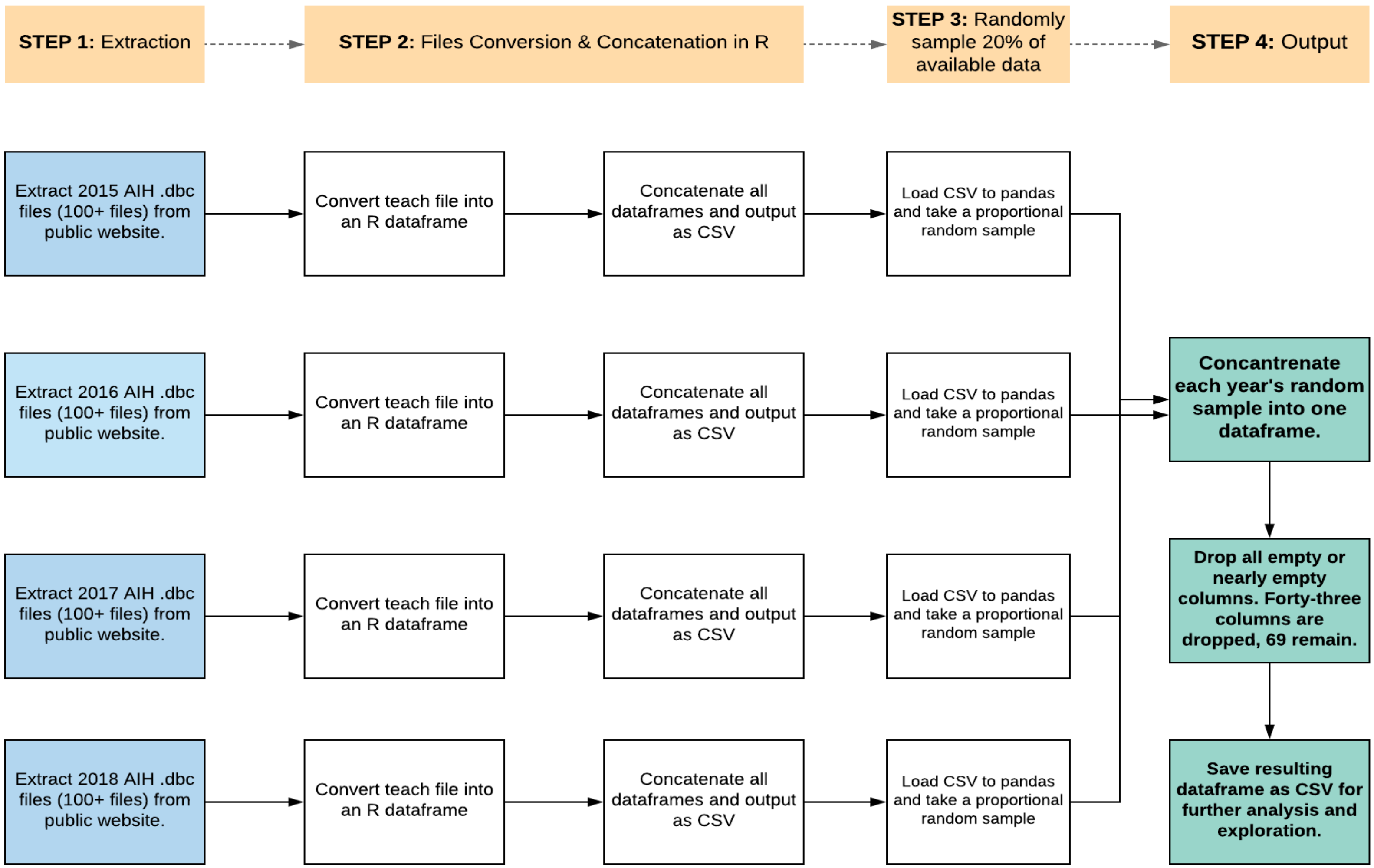
A record in the AIH database is created when a hospital or healthcare unit generates a request for hospitalization. Providers submit demographic and health information about the patient. This request is approved, reduced, rejected, or rejected due to an error. While the patient is in the hospital, the record is updated to also contain information about procedures performed and discharge. Each row of information represents an hospitalization. If a patient is hospitalized more than 30 days, a new authorization is needed and a new record (i.e. row is created).

WHAT IS DEEP LEARNING?

DATASET EXTRACTION, CONVERSION & SAMPLING

1. Extraction from Webpage
2. Conversion from dbc to CSV using R
3. Conversion to pandas df
4. Proportional Random Sample

Conversion and Data Wrangling Process



DATA WRANGLING

EXPLORATORY DATA ANALYSIS

1. Patient Demographics
2. Diagnosis
3. Hospitalization Services
4. Financial Information

FEATURE ENGENIEERING

SOURCES

1. Choi et al. Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. 2016. https://arxiv.org/abs/1511.05942
2. de Araujo Lima et al. Successful Brazilian Experiences in the Field of Health Information – Final Report. 2016. https://www.measureevaluation.org/our-work/health-information-systems/health-information-system-strengthening-in-lac-region-2005-2010/his-brazil-english-august2007.pdf
3. Malhotra et al. Long Short Term Memory Networks for Anomaly Detection in Time Series. 2015. Long Short Term Memory Networks for Anomaly Detection in Time Series. https://www.elen.ucl.ac.be/Proceedings/esann/esannpdf/es2015-56.pdf
4. Ralf Carvalho & Alex Paino. Deep Learning for Fraud Detection. https://engineering.siftscience.com/deeplearning-fraud-detection/
5. Sapronova et al. Deep learning for wind power production forecast. 2017. http://ceur-ws.org/Vol-1818/paper3.pdf
6. Shipmon et al. Time Series Anomaly Detection. https://ai.google/research/pubs/pub46283
7. Tang and Mendis. Unsupervised fraud detection in Medicare Australia. 2011. https://pdfs.semanticscholar.org/e7b6/9c1ba648de68d30aacb84b47ab43fcdf75e9.pdf
8. Teodoro et al. ORBDA: An openEHR benchmark dataset for performance assessment of electronic health record servers. 2017. https://www.ncbi.nlm.nih.gov/pubmed/29293556
9. Tianwei Yue & Haohan Wang. Deep Learning for Genomics: A Concise Overview. 2018. https://arxiv.org/abs/1802.00810
10. Viscondi et al. Cost Description of Prevention and Treatment of Cervical Cancer. https://www.researchgate.net/profile/Juliana_Viscondi/publication/299483604_COST_DESCRIPTION_OF_PREVENTION_AND_TREATMENT_OF_CERVICAL_

CANCER_IN_BRAZIL_IN_2012/links/59b6786ea6fdcc7415bd3940/COST-DESCRIPTION-OF-PREVENTION-AND-TREATMENT-OF-CERVICAL-CANCER-IN-BRAZIL-IN-2012.pdf