Deep Learning:
Brazilian Hospitalization Data

IVETTE M TAPIA

Motivation & Description

- **Business Case:** Ability to know the procedures that are likely to be performed on a patient at the moment of hospitalization approval brings administrative efficiencies on processing the request and help managers manage capacity. Healthcare data is full complexities.
  - *Stakeholders:* administrators, payers and insurers.
- **Dependent Variable:** Procedure Performed (Group)
- **Independent Variables:** Sex, Death, Age, ICU Days, VeneraL Exam, Intermediate ICU Days, Length of Stay, Companion Nights, Principal Diagnosis, Reason Stay/Exit, Character of Hospitalization, Complexity, Type of ICU.

Data Acquisition

- The health ministry of Brazil publishes datasets on hospitalization requests and approvals of their public healthcare system.
- The hospitalization data was extracted from the DataSUS servers through the web. The website is as follows: http://www2.datasus.gov.br/DATASUS/index.php?area=0901&item=1&acao=25.
  - Data from years 2015 – 2018 was downloaded (3.5 years of data).
- Data files were as follows:
  - 2015: 325 .dbc files (1.26 GB)
  - 2016: 325 .dbc files (1.71 GB)
  - 2017: 318 .dbc files (1.71 GB)
  - 2018: 189 .dbc files (504 MB)

Git Hub Repository

— **Data Wrangling & Exploratory Analysis Code**

- Step 1: Conversion from .dbc to CSV
- Step 2: Proportional Random Sample
- Step 3: Demographic Cleaning & EDA
- Step 4: Diagnosis Cleaning & EDA
- Step 5: Hospitalization Features Cleaning and Exploratory Analysis
- Step 6: Concat all features and make train, test and validation arrays

— **Deep Learning Model**

- Deep Learning Models

— **Final Report**

- Final Report

Data Management

Step 1: Convert .dbc files to R dataframes

Step 2: Concatenate R dataframes by year

Step 3: Output concatenated files as CSV files

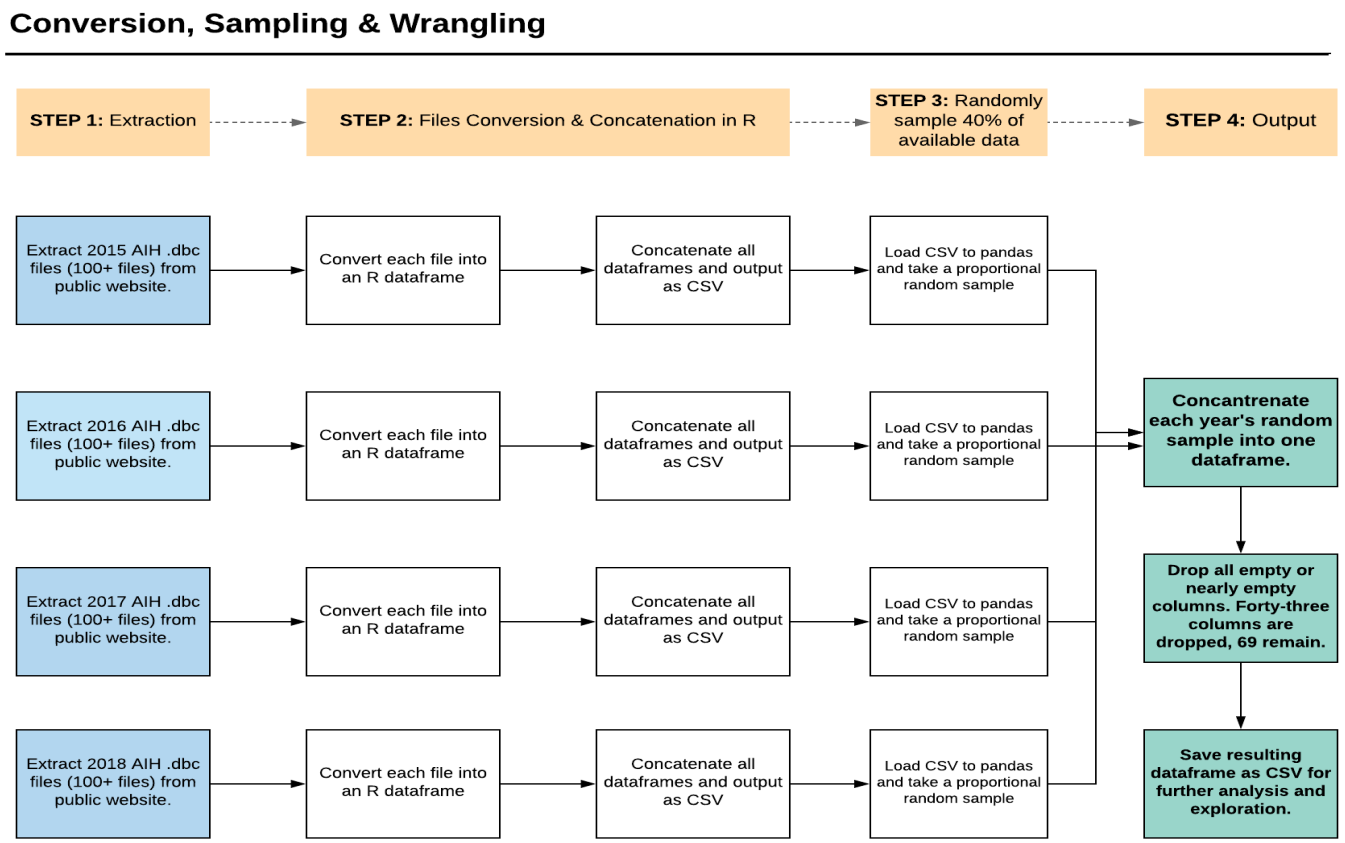
Step 4: Load CSV files to Jupyter notebook and take proportional random sample.

- Computational resources available did not allow to use all the data available.
- A sample was taken based on the proportion of the year. Example: if a year's samples were 20% of the total available, from that year a random sample that equals 20% of the target sample will be taken for that year.

Step 5: Concatenate each year's samples into one dataset. Total cases extracted equals 16,614,830.

Step 6: Output resulting dataset as a CSV. This dataset is the combined random sample for all years.

Data Management Summary



Data Cleaning

Step 1: Features related to financial payments and metadata not going to be used.

Step 2: Demographic Features: Features with serious data quality issues were dropped (example. male and pregnant at risk). Variables with more than 20% missing were dropped as well. Categorical features were recoded.

- *Variables Dropped:* Instruction Level, Pregnancy at Risk, Patient's Occupation, Patient's Number of Children.
- *Variables Cleaned:* Sex, Death Indicator, Municipality of the Patient, Patient's Nationality, Patient's Race (replace missing with np.nan), Patient's Ethnicity (replace missing with np.nan), and Patient's Age.
- Output demographic features as CSV

Data Cleaning cont’

Step 3: Diagnosis: Empty columns dropped and categorical features were recoded.

- Merged with CID codebook to aid on exploratory analysis.
- *Variables Dropped:* Secondary Diagnosis
- *Variables Cleaned:* Principal diagnosis. Codebook added to aid on exploratory analysis.
- Output diagnosis features as a CSV

Step 4: Hospitalization Features: Features with more than 20% empty dropped and categorical features were recoded.

- Merged with available codebook to aid on exploratory analysis.
- *Variables Dropped:* Hospital Nights, Hospital ID
- *Variables Cleaned:* Procedure Realized, Procedure Group, ICU days, Reason Stay/ Exit, Venereal Exam, Contraception 1 & 2, Character of Hospitalization, Type of ICU, Intermediate ICU, Hospital ID
- Output hospitalization features as a CSV

Demographics: *Exploratory Data Analysis*

- Gender: More hospitalizations are of females.
- Race: Majority of hospitalizations are of individuals of white and brown races. Racial breakdown seem to closely follows the overall demographics of Brazil.
- Ethnicity: This field is only used if individual's race is indigenous. There seems to be a overwhelming majority of one indigenous ethnicity.
- Nationality: Overwhelming Brazilian. I have dropped this variable since there does not seem to be great variation.
- Municipality: There are ~5,000 municipalities. While big cities comprise the top number of hospitalizations, their totals are still not a significant proportion of the overall dataset.
- Age: Age is not normally distributed. It has three peaks: in young ages, middle age and late middle age.
- Visuals Used: Histograms, Frequencies, Correlation, ECDF, Bootstrap replicates

Diagnosis: *Exploratory Data Analysis*

- Top diagnosis are related to birth, pneumonia, bacterial diseases, heart disease and vesicular biliary. **There are 8,721 unique diagnoses.**
- The distribution of principal diagnosis is highly unbalanced. While there are diagnosis that are more common than others, the top diagnosis are still a fraction of the total. As such, there is a lot of heterogeneity in this variable.
- The point above is evidenced by the large confidence intervals for the mean diagnosis counts.
- There are 8,721 unique diagnosis.
- Visuals Used: Histograms, Frequencies, Correlation, ECDF, Bootsrap replicates

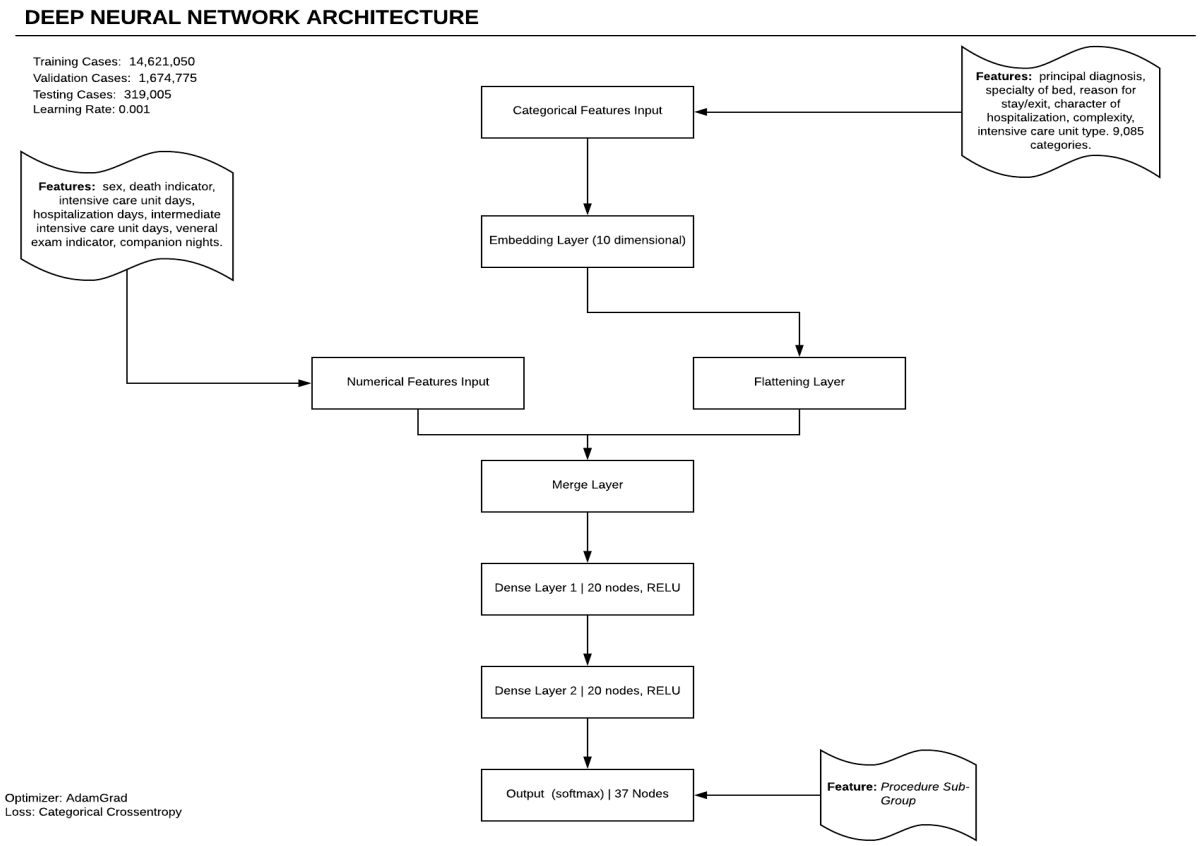
Hospitalization: *Exploratory Data Analysis*

- Urgent type of hospitalization is by far the most common character of hospitalization.
- Improved is the most common reason reason for stay/exit. *
- Bootstrap mean replicates shows a 95% confidence interval for days of stay is between 5.388 and 5.396. This is a very tight interval. This range contains our sample mean of 5.39.
- Top 5 procedure performed sub-groups are: Clinical Treatment(other specialties), Birth, Obstetric Surgery, Surgery of the Osteomuscular System, Surgery of the Digestive System.
- The top 5 procedures performed are: Normal birth, cesarean birth, influenza treatment, bacterial treatment, psiquiatric treatment.
- Categorical variables are highly unbalanced, notably procedure performed and procedure sub-group.
- Correlations show that: (1) specialty of bed is moderately correlated with procedure realized and (2) venereal exam is moderately correlated with for reasons for stay/exit.
  - Visuals used: Histograms, Frequencies, Correlation, ECDF, Bootstrap replicates

Feature Selection & Engineering

- Numerical Features were normalized (0 to 1)
- Output variable was one-hot encoded
- Three arrays were created: numerical, categorical and output variable.
- Cleaned and not used: Contraception, Hospital ID, municipality, race, ethnicity, nationality. (missing values + ethics)
  - This was a decision based on judgment, resources available and ethics.
- Data divided on train, validation and test. (88%, 10%, 2%)
  - Training Arrays: 14,621,050 cases
  - Validation Arrays: 1,674,775 cases
  - Testing Arrays: 319,005 cases

Modeling Part I: Architecture Design



Baseline Model

Modeling Challenges

- Several categorical features with large number of categories.
  - Embedding layer was used to avoid converting to one-hot encoding and using large amounts of memory.
  - Alternative approach was to use the ‘hashing trick’.
- Large number of output categories.
  - Softmax layer with 37 outputs
- Unbalanced Classes: Notably diagnoses & procedures.
  - Random oversampling was used on categories that had less than 100k cases. SMOTE-NC was tested, but performed similarly to random oversampling.

Evaluation Strategy

- **Primary metric is accuracy.**
  - *While accuracy is not the only evaluation metric available and not appropriate for all problems (precision, recall, ROC, & AUC were other options) in this case the main interest is the extent the neural network can predict the correct procedure for a patient.*
- Secondary metric is the false positive rate.
- Tertiary metric is average accuracy (i.e. accuracy by class).
- Beyond these three-metrics, the extent of model overfitting, complexity, and training time were also taken into account when evaluating models.

Modeling Part I: *Model Architecture Tests*

Model Number	Embedding Layer	Layers	Nodes (at each dense layer)	Regularization	Learning Rate	Oversampling
Model 1	Yes	2	20	No	0.001	No
Model 2	Yes	4	20	No	0.001	No
Model 3	Yes	2	20	L2	0.001	No
Model 4	Yes	6	20	No	0.001	No
Model 5	Yes	2	20	L1	0.001	No
Model 6	Yes	2	20	Dropout	0.001	No
Model 7	Yes	2	20	No	0.001	Random
Model 8	Yes	2	20	No	0.001	SMOTE
Final Model	Yes	2	20	Dropout	0.001	Random

Evaluation Results: *Model Architectures*

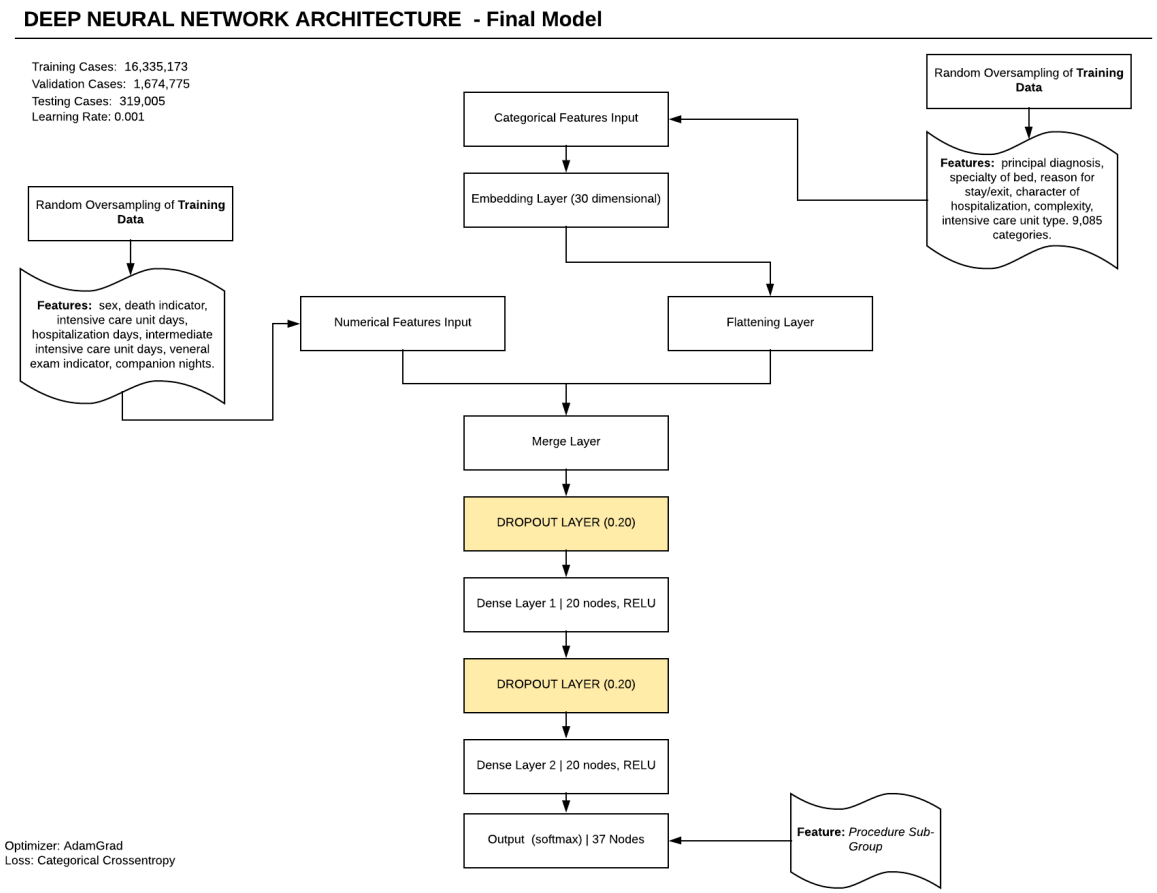
Model Number	Training & Testing Data				Validation Data		Total Training Time
	Training Accuracy	Testing Accuracy	Average Accuracy*	False Positive Rate	CV (k = 5)	Diff Cross Validation & Training	
Model 1	89.95%	89.90%	66.1%	0.003	85.85% (+/-0.74%)	4.10	3.5 hours
Model 2	89.56%	89.53%	60.5%	0.004	85.98% (+/- 1.08%)	3.58	4.25 hours
Model 3	84.21%	84.18%	36.6%	0.005	77.31% (+/- 1.11%)	6.90	3.5 hours
Model 4	89.97%	89.92%	64.1%	0.003	86.25% (+/- 0.69%)	3.72	5.9 hours
Model 5	85.43%	85.38%	39.6%	0.005	70.03% (+/- 1.35%)	15.13	3.5 hours
Model 6	85.96%	88.89%	55.3%	0.004	82.83% (+/- 0.67%)	3.13	4.25 hours
Model 7	90.30%	90.64%	82.6%	0.003	85.14% (+/- 0.37%)	5.16	4 hours
Model 8	90.68%	90.66%	81.2%	0.003	85.80%(+/- 0.23%)	4.88	5 hours
Final Model	87.96%	90.65%	80.4%	0.003	86.45%(+/- 0.53%)	1.51	19 hours

Modeling Part II. Parameter Testing

- Smaller batch size of 32 and larger embedding dimensions yield better performance in terms of accuracy.
- A smaller learning rate needs a larger number of epochs to learn more, as such the larger learning rate of 0.001 tested yielded higher accuracy in these experiments.

Batch Size	Learning Rate	Embedding Dimension	5 fold CV (epochs = 25)	
			Training Accuracy	5 Fold CV Evaluation
128	0.0001	10	44.53%	44.54% (+/- 4.41%)
128	0.0001	30	55.55%	55.80% (+/- 2.77%)
32	0.0001	10	59.05%	59.23% (+/- 5.23%)
32	0.001	10	85.80%	85.74% (+/- 0.55%)
32	0.0001	30	67.64%	67.67% (+/- 0.66%)
32	0.001	30	88.39%	88.32% (+/- 0.22%)
128	0.0001	10	44.53%	44.54% (+/- 4.41%)

Modeling Part III. Final Model



Final Model Performance

Model Number	Training & Testing Data				Validation Data		Total Training Time
	Training Accuracy	Testing Accuracy	Average Accuracy*	False Positive Rate	CV (k = 5)	Diff CV & Training	
Final Model	87.96%	90.65%	80.4%	0.003	86.45%(+/-0.53)	1.51	19 hours

LIME Explainer

The LIME explainer suggests that key features driving the predictions are:

1. Principal Diagnosis
2. Reasons for stay/exit
3. Complexity
4. Specialty of bed

Main Findings

- A well performing model in terms of accuracy as constructed.
- Opportunities remain to optimize the model, however costs and trade-offs must be taken into account.
- The network still had instances of having difficulty predicting less common procedures. An oversampling strategy improved performance across classes, it did not completely solved it.
- The addition of layers from baseline did not improve model performance.
- The LIME explainer suggests that principal diagnosis, reasons for stay/exit, complexity and specialty of bed are key features.
- Regularization was needed to ameliorate overfitting. The final model ovefitted by 2.69%.

Recommendation & Next Steps

- Hyper Parameter Tuning and Optimization
  - Dropout Rate
  - Learning Rate
  - Batch Size
  - Embedding Dimensions
  - Nodes
  - Regularization Term
- Further data engineering: from the graph, it is clear that there is a cluster of categories which the network is having difficulty with. It is possible that these maybe profiles of patients or procedures that are related.
- Pruning the Network: removing nodes that do not contribute much to the predictions.

Recommendation & Next Steps Cont’

- Develop business processes to review and use predictions, specially on procedures that are less common or ‘rare’.

- Before embarking in further optimization and testing is important to consider the trade-offs and the cost-benefit ratio to the business. Building, training and tuning a neural network is time consuming and computationally expensive.
  - ✓ *A key question is how much the business is willing to invest to increase network performance by a few percentage points more.*

Summary

- A deep neural network was built to predict procedures performed given a patient's hospitalization request information.
- The dependent variable was the procedures performed group.
- Independent variables included
- Final model key metrics were as follows:
  - Testing accuracy of 90.3%.
  - False positive rate as 0.003.
  - Overfitting: 2.69%
  - Training Time: 19 hours

Summary Cont’

- Opportunities remain to optimize and fine tune parameters, however costs and benefits must be taken into account.
- The network still has some difficulty predicting less common procedures. An oversampling strategy improved performance across classes, it did not completely solved it.
- Business procedures regarding review is strongly encouraged, specially for less common procedures.
- Using the Lime explainer we found that: principal diagnosis, reason for stay/exit, specialty of bed, and complexity were key features.
- It is possible to build a model with more specific procedures, however this model would be more complex and computationally very expensive (1,600+ outputs).