Explainability and understanding for deep learning models

Sanji Fernando

July 12, 2019

OPTUM Labs®

Clinical reviews for reimbursement

AI case study

- For many inpatient stays, health care payers require 3rd party, independent physician review of the medical record to support the level of care reimbursement.
- Many admissions do not support a higher level of reimbursement — yet clinicians still end up reviewing these cases.
  - Hundreds of clinicians at Optum review the medical records and notes every day and provide their independent assessment of the level of care that should be supported for reimbursement.
- These skilled clinicians combine clinical guidelines and experience to review medical notes to determine if an inpatient stay is supported.
- The record of their work includes the notes the clinicians reviewed and their expert determination.

We trained an AI solution to automate the review of notes and determine which cases require a clinician review.

OPTUM Labs®

Trained deep learning model

Pre-screen medical notes for clinicians

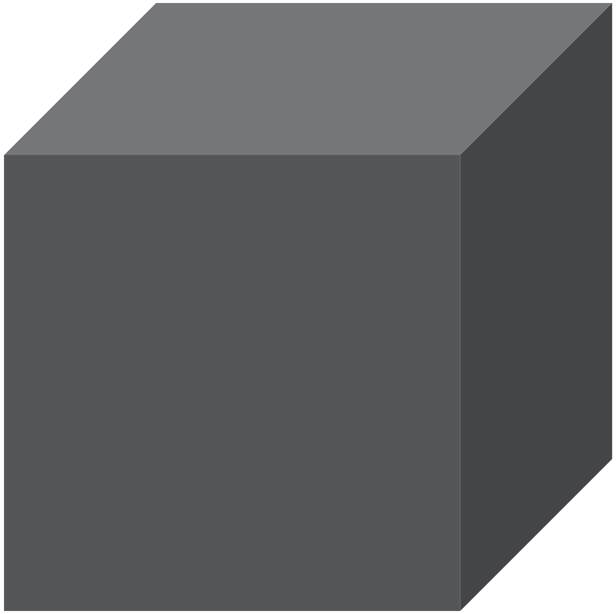
- We trained a deep learning model on over 200K medical notes — and the decisions from clinicians after reviewing these notes.
- The trained model automates the review of all medical notes, and determines which notes should be reviewed by physicians.
- The deep learning model “reads” the medical notes and “learns” from the previous determinations made by physicians.
- We used a variation of a Recurrent Neural Network (RNN), a model architecture that is often used to determine sentiment in unstructured text.
- A score is generated by the deep learning model — and model designers select a threshold over which a chart should be reviewed by a physician.

OPTUM Labs®

The black box problem

How can we understand if the model is working correctly?

- Deep learning models turn words/text into numeric vectors.
- These vectors are passed to a series of connected equations.
- They resolve to a numeric score without detail on what the basis was for the score.
- But a model does not specify for clinicians what aspects of the medical notes would support the model’s assessment.
- This leads to key questions as to what is driving the determination made by the model.



OPTUM Labs®

Techniques for better understanding

Leveraging an attention function in our model

- We implemented model architecture that helps us better understand what terms are important to the model.
- It creates a set of intermediate scores of the words in the notes, helping us understand what may be drawing the “attention” of the model.
- We created a tool for users to see what terms in the medical notes had higher attention scores.
- Using this tool, clinicians are able to see which terms have generated more “attention” — and confirm these are the aspects of the medical notes they would have keyed on as well.

OPTUM Labs®

Visual tools for better understanding

We have translated specific attention weights into a visual interface

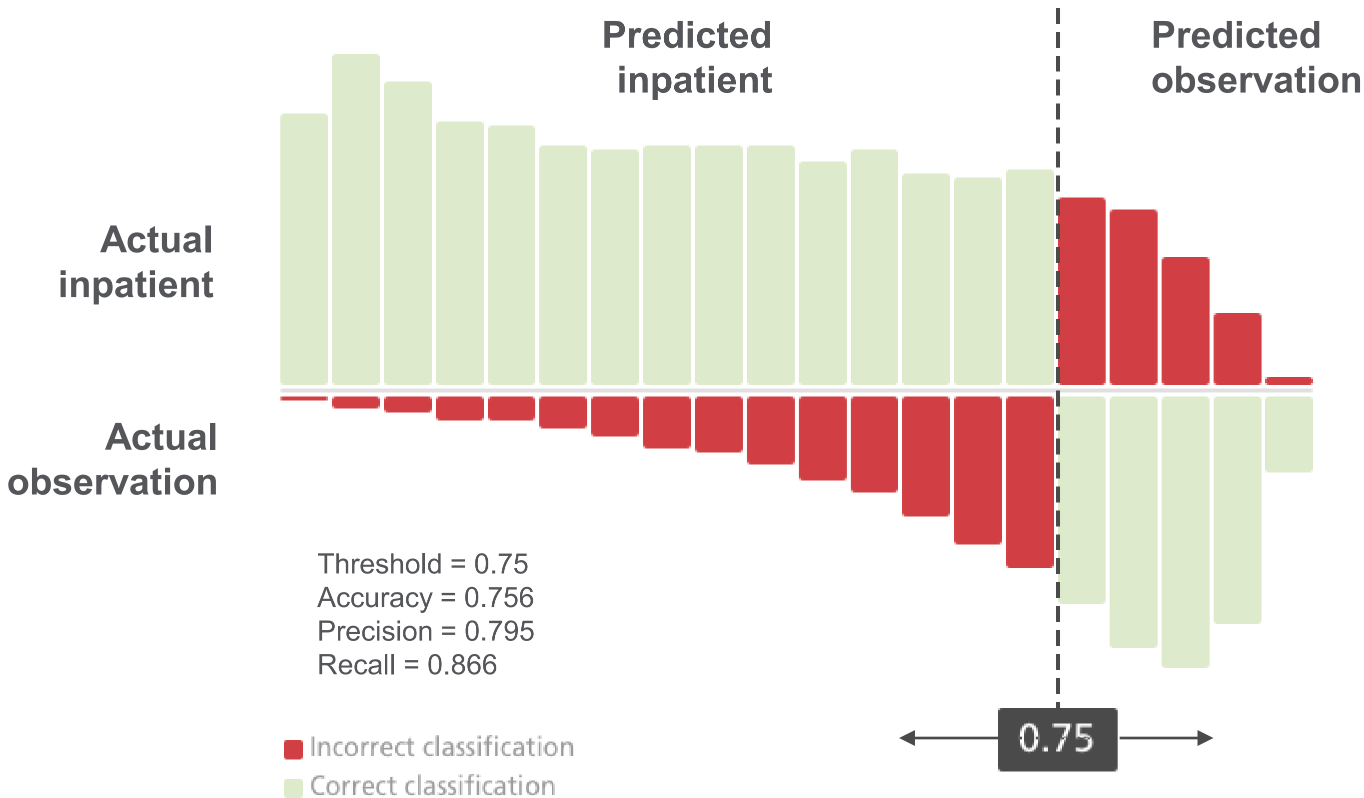
chest view pm history atypical chest pain comparison technique xr chest view findings heart is borderline in thoracic aorta is and lungs are with an increase in there is of interstitial lung disease is again there are no focal areas of lung consolidation or suspicious lung no pleural sternotomy hardware leads and impression chronic interstitial lung no change since report generated on workstation all poa and collapse occurred while working with pt pain rbbb compared to ekg last admission cabg recent acute diastolic chf exacerbation adm from hypoxia due to pulmonary edema from acute diastolic chf exacerbation and idiopathic pulmonary fibrosis pt was discharged on home oxygen 2l fibrosis copd clinically exacerbation to severe tricuspid regurgitation pulmonary 60 ckd htn however sbp has been on lower side and on last discharge all bp meds were discontinued lisinopril losartan metoprolol type to observation medical telemetry troponins ekg to hold as pt remains hypotensive obtain orthostatic vital signs precautions ssi hold metformin all home medications as medically indicated orders prophylaxis with heparin plan discussed with pt and family all of their questions were problem medical history chronic dm diabetes mellitus htn hypertension hyperlipidemia historical no qualifying data acute hypoxia secondary to pulmonary edema from acute diastolic chf exacerbation idiopathic pulmonary fibrosis with ambulation home with oxygen supplement acute diastolic chf exacerbation pulmonary fibrosis troponin elevation secondary to chf exacerbation moderate to severe tricuspid regurgitation possible pulmonary 60 hyp
dementia hld history open heart surgery tonsillectomy home medicati
mg ml inh every hours prn aspirin 81 mg enteric coated tablet 81 mg
mg po qhs furosemide 20 mg po qday glipizide mg tablet extended re
nitrostat mg sublingual tablet mg sl q5min prn proair hfa 90 inhalation

- **The darker** the highlighting, the **higher its importance** or influence on status score
- **Darker highlights** denote **how frequently**, and **in what context**, it was used in prior IP recommendations
- AI reviews words that precede and succeed the target word to **understand its context** and **to assign its weight**

OPTUM Labs®

Understanding tradeoffs

We have also developed visual tools for customers to understand the tradeoffs when selecting the threshold used to classify medical notes.



Model score distribution and classification customization

OPTUM Labs®

Key takeaways

How to approach explainability and understanding

When possible, express and share how the model operates to key stakeholders.

- Expose training data and the labels (answers) used to train the model.
- Apply techniques like Attention to increase visibility on inputs to the model, and create a dialogue with clinicians on how the model works.
- Understand that the output of a model is a score, and understand the tradeoffs for false positives and false negatives for any prediction/inference.

But we still have more work…

- We need AI model approaches that can answer "why" models scored an input a certain way.
- New breakthroughs in causal models may lead to more interpretable models.
  - Invariances is a very new technique described by Facebook on how to train multiple deep learning networks on different slices of training data and isolate specific features that are central to each model's inference.
  - Probabilistic programming employs Bayesian methods to combine learning for data with an assertion of priors reflecting expertise (like known clinical knowledge).

OPTUM Labs®