Stochastic Modelling of AI Diagnostics

CBS Final Presentation

Andrés Aradillas Fernández under Professor Jing Dong

Economics & Mathematics Undergraduate Student
Cornell University

Summer 2023

Overview

1. Background & General Premise
2. The Basic Model
3. Optimization
4. Sensitivity Analysis

Background & General Premise

Inspiration for Model

- On April 11, 2018, the U.S. Food and Drug Administration approved IDx-DR, a first-of-its-kind AI-trained early diabetic retinopathy detection device
- More th an 50% of the 30 million Americans with diabetes do not see their eye doctor on a yearly basis, despite the heightened risk of eye problems
- IDx-DR acts as a tool that PCP as well as patients themselves can learn to use with little need for expertise
- But if we are to use this, how should we ensure the most severe cases are seen first, given there is some error to the AI diagnostic tool?

Research Question

How should we route patients to see the doctor based on AI triage, and who should we remove from the queue altogether?

Background on Field

Generally, most queueing problems are formulated as follows:

1. The individuals arrive according to a Poisson distribution with parameter λ.
2. They are routed in some way (which is what is of primary interest here). Attribute(s) of the individuals are conditioned on in order to route them properly.
3. However long a patient waits, they face a delay cost of c per unit time.
4. They are seen by a server (most models reduce the case to a single-server, as we will do in our model). The service time is assumed to be an Exponential distribution with parameter μ.

Sometimes, individuals arriving may belong to different classes (e.g., severity of eye problems), and they may have different service times μ_i and arrival times λ_i.

Primary Considerations

The primary considerations of a queueing problem are:

- *Maintaining a manageable customer load*: The long-run customer load is $\rho := \frac{\lambda}{\mu}$ and we want a system such that $\rho < 1$, or else more people will be arriving than can be seen per unit of time, and in the long-run there may be an infinitely long wait time.
- *Minimizing average expected delay costs*: Some patients may face more costs for waiting longer than others, and so minimizing this is another goal of an efficient queue.

However, we are also interested in not mistakenly rejecting a severe patient, and so Type I errors are also a concern.

The Model

Diagram of Model

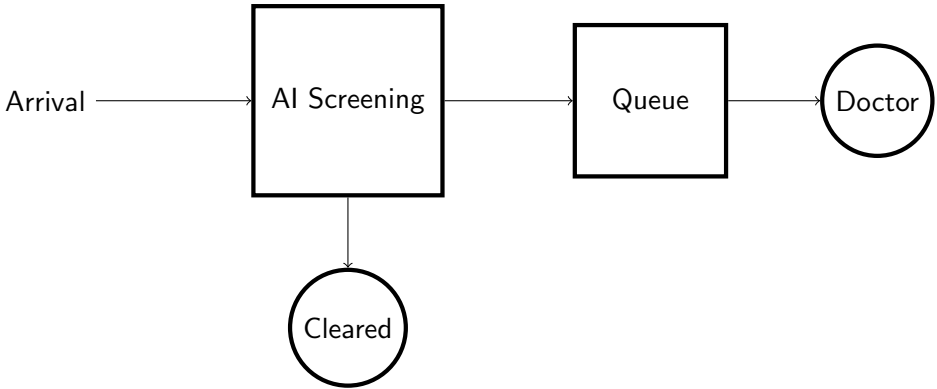


Figure: A diagram of diagnostics involving AI screening tools.

Assumptions

- Assume that we have a 'non-preemptive single-server system,' or in other words, we have one server and when a patient is being seen, the appointment cannot be interrupted to replace the patient with another.
- Assume all patients arrive according to a Poisson distribution with parameter $\lambda_0 > 0$.
- Assume service time is uniform among all patients, and is distributed according to an Exponential distribution with parameter $\mu > 0$.
- There are two classes of patients, either *diseased* (1) or *non-diseased* (0), and the proportion of them that are diseased is γ.
- Falsely rejecting a patient of class 1 results in a cost $p > 0$ per unit time
- Waiting costs are w per unit time for all.
- Cost of service is c per unit time.

Confidence Score

The exact class of a patient is not known, but a logistic regression is run based on an incoming patient's attributes, and a classifier $X \in \mathbb{R}$ is received. This can be interpreted as the probability of a patient being of class 1.

It is this 'confidence score' of X that we will condition the routing system on. This X will allow us to take expectations.

We need to denote the following:

- $f_0(x) := P(X = x|Y = 0)$
- $f_1(x) := P(X = x|Y = 1)$

Optimization

Since delay costs, arrival rates, and service times are all equal across both class types, the most optimal strategy in terms of minimizing average expected delay cost is to operate under a First-Come-First-Serve (FCFS) strategy for those in line, meaning the patients are placed in a line according to the order they came in.

The following optimization problem summarizes what we seek to accomplish by selecting a threshold τ of when to reject based on the classifier score X:

$$\min_{\tau} c\lambda(\tau) + w \cdot \frac{\rho(\tau)}{1-\rho(\tau)} + p\lambda_0\gamma \int_{-\infty}^{\tau} f_1(x)dx$$

where $\rho(\tau) := \frac{\lambda(\tau)}{\mu}, \lambda(\tau) := \lambda_0\gamma \int_{\tau}^{\infty} f_1(x)dx + \lambda_0(1-\gamma) \int_{\tau}^{\infty} f_0(x)dx.$

Estimation & Sensitivity Analysis

Using the optimization above, we then implemented a gradient descent algorithm to estimate the optimal threshold locally.

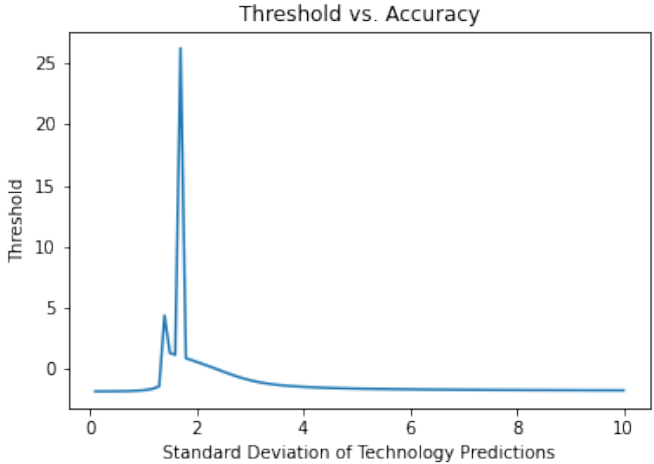
We used the following formula to iteratively find τ:

$$\tau_{k+1} = \tau_k - h \cdot \nabla g(\tau_k)$$

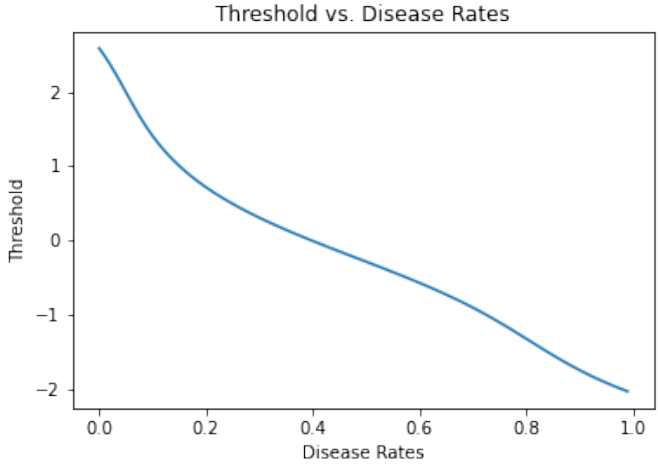
where h is the learning rate, a parameter we fine-tuned as needed, and k goes to 500.

With this solution, and a reasonable selection for initial parameters, we then conducted sensitivity analysis to see how our choice of threshold may change as one parameter changes, ceteris paribus.

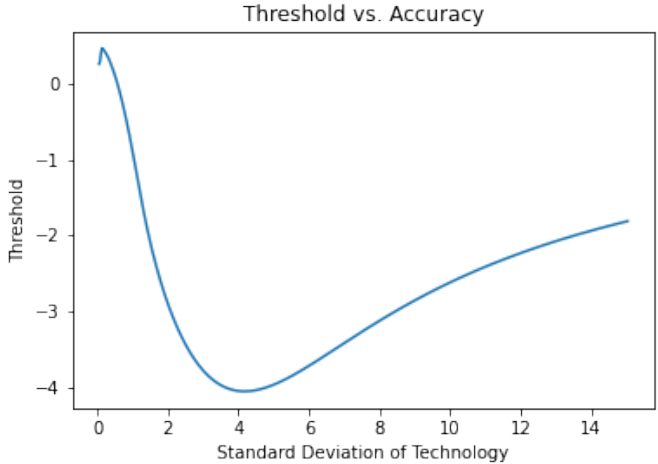
Sensitivity Analysis Results



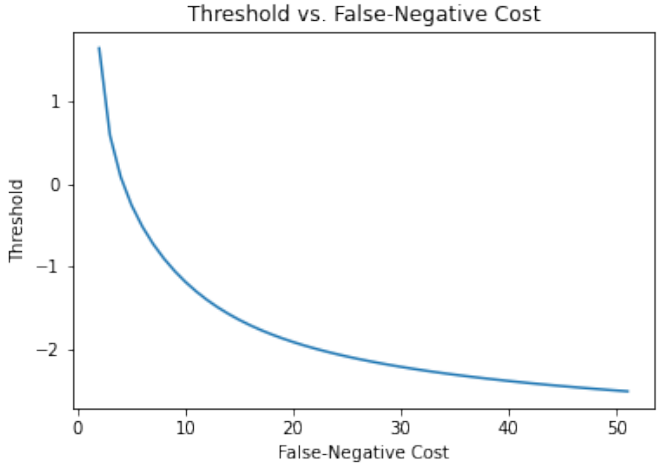
Sensitivity Analysis Results



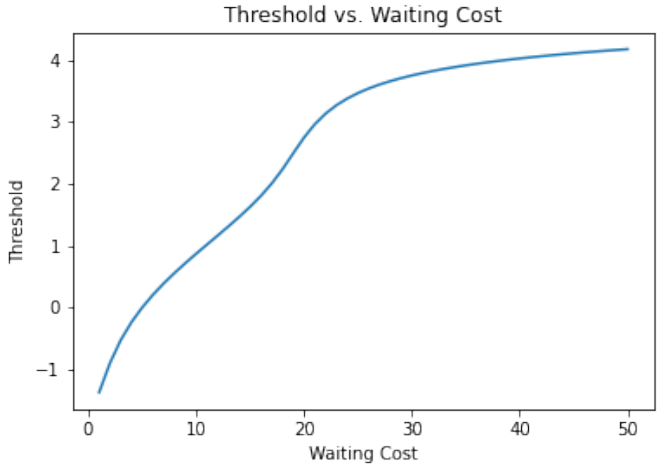
Sensitivity Analysis Results



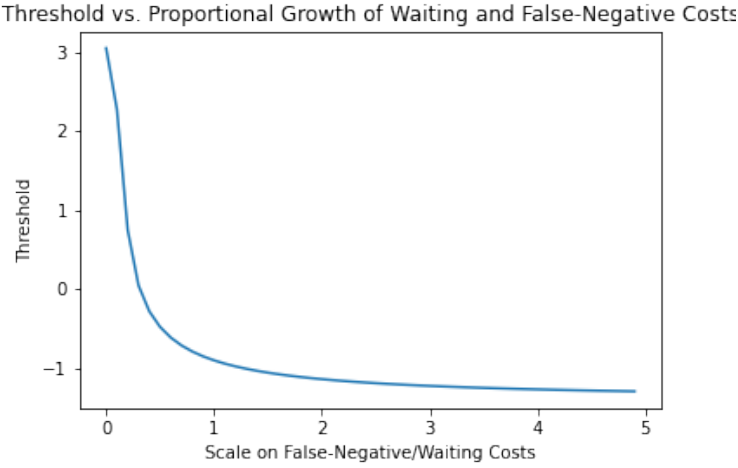
Sensitivity Analysis Results



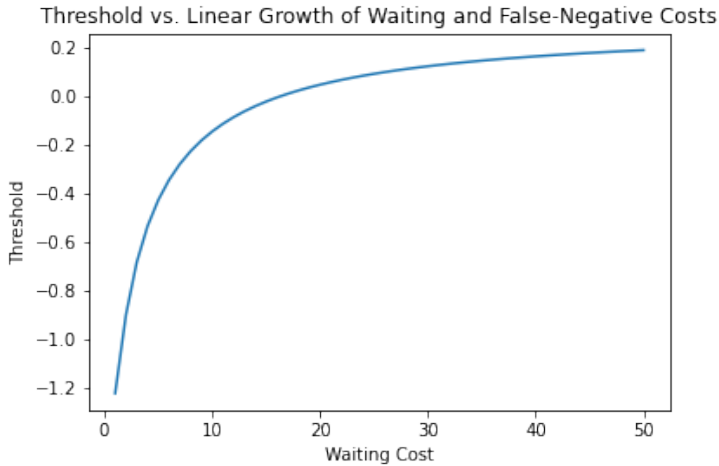
Sensitivity Analysis Results



Sensitivity Analysis Results



Sensitivity Analysis Results



Questions?