Artificial-Intelligence Assisted Decision Making: A Statistical Framework

Han Hong, Xin Lin, Ke Tang and Jingyuan Wang

Standford University
Beihang University
Tsinghua University
Beihang University

December 10, 2019

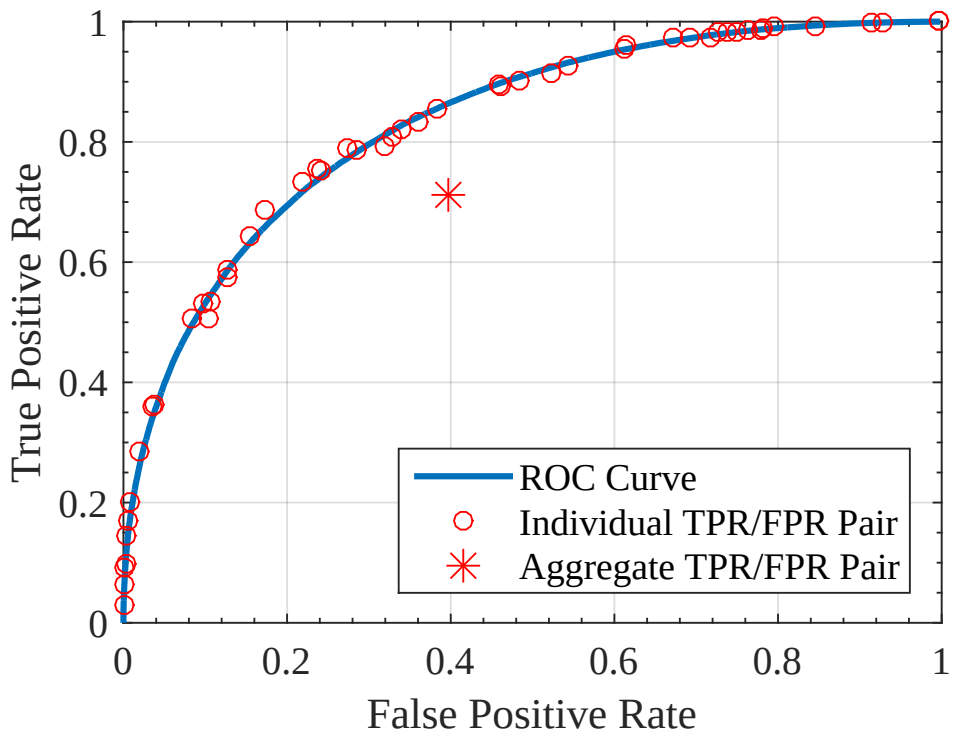
- In the era of machine learning, much attention has been drawn to the potential of machine in assisting human decision making.
- We focus on the assistance of binary classification problem, which is a foundational building block in related fields.
- The first and important thing is to compare machine and human, and filter out less capable human decision makers than machine.
- Recall the relative definitions for comparison:
  - In a sample dataset with labels $Y_i \in \{0, 1\}$, predictor $\hat{Y}_i \in \{0, 1\}$ for Y_i, and associated features $X_i, i = 1, \ldots, n$, we have TPR and FPR as

$$\text{TPR} = \sum_{i=1}^{n} Y_i \hat{Y}_i / \sum_{i=1}^{n} Y_i, \text{FPR} = \sum_{i=1}^{n} (1 - Y_i) \hat{Y}_i / \sum_{i=1}^{n} (1 - Y_i).$$

  - For machine algorithm, the decision rule is typically based on $D = \mathbb{1}(p(x) > c)$, where x is a set of features, $p(x)$ is the propensity score function learned from data, and c is a cutoff threshold. The ROC curve is generated across various cutoff threshold c.

- Much of the literature compares the performance of machines with the representative (average) person.
- The conclusions of these papers are mostly based on observing an average pair of FPR/TPR that lie strictly below the ROC curve formed by the machine algorithm, implying that machines can achieve a higher TPR for a given FPR, or a lower FPR for a given TPR.
- However, such interpretations without a deeper understanding of the human decision making process may cause inappropriate evaluation for human decision maker, as
  - The superior information quality of machine learning algorithms and the *incentive heterogeneity* of human decision makers influence the comparison.
  - The FPR/TPR pairs of human decision makers are *not precisely measured*, especially when the number of cases that the decision makers see is not large.

- For example, figure above shows a collection of human decision makers all lie approximately on the machine-learned ROC. This is the case if they employed decision rules $\hat{Y}_i = \mathbb{1}\left(\hat{p}\left(X_i\right) > c_i\right)$ with the same $\hat{p}\left(\cdot\right)$ function but with different individual cutoff points c_i.
- However, after aggregating over all decision makers, the aggregate TPR/FPR pair lies visibly below the ROC.



We propose a statistical framework for AI assisted decision making, which is able to

- Determine whether the machine is more capable in decision making than a human decision maker is from historical data under a fixed confidence level.
- Provide decision rule based on machine's ROC with maximized level of confidence when a decision maker is replaced by machine.
- Replace less capable decision makers with machine algorithm in future decision process, but remain other ones.

Two basic assumptions are necessary for comparing the machine ROC and the individual TPR/FPR pair when it is precisely known without sampling errors.

Assumption

The machine ROC represents a propensity score model of prediction: $\mathbb{1}\left(p\left(x\right) \geqslant c_1\right)$. *The decision maker's decision is also based on a model* $\mathbb{1}\left(q\left(x\right) \geqslant c_2\right)$. $p\left(\cdot\right)$ *and* $q\left(\cdot\right)$ *can be correctly specified or misspecified. There is no incentive-features dependence across cases for each individual decision maker:* c_2 *does not depend on* x*, i.e. each decision maker uses the same weights for type I and II errors for all his/her cases.*

Assumption

The machine ROC is precisely known without any sampling error.

We focus on one individual decision maker, and we are interested in the parameter of the population pair of true positive and false positive rates for this decision maker:

$$\theta_0 = \left(\mathrm{HFPR}_0, \mathrm{HTPR}_0\right),$$

where

$$\begin{aligned}
\mathrm{HTPR}_0 =& \frac{1}{p}\mathbb{E}\, Y_i \hat{Y}_i = \frac{1}{p}\int\int y \mathbb{1}\left(q\left(x\right) \geqslant c_2\right) \mathrm{d}y\mathrm{d}x, \\
\mathrm{HFPR}_0 =& \frac{1}{1-p}\mathbb{E}\left(1 - Y_i\right) \hat{Y}_i \\
=& \frac{1}{1-p}\int\int \left(1-y\right) \mathbb{1}\left(q\left(x\right) \geqslant c_2\right) f\left(y, x\right) \mathrm{d}y\mathrm{d}x.
\end{aligned}$$

- We are interesting in whether θ_0 is above or below the ROC.
- Typically we do not observe θ_0, even though we can estimate it using a sample of observations:

$$\hat{\theta} = (\text{FPR}, \text{TPR})$$

  where the FPR and TPR can be calculated from sample.
- However, in reality only a *noisy* estimate of the TPR/FPR pair can be obtained from empirical data. How do we make probabilistic statement regarding whether θ_0 is above or below the ROC?
- We provide two approaches to address the question of sampling uncertainty. The first is based on Bayesian analysis principle, and the second is based on classical frequentist inference.

The Bayesian approach

- A Bayesian method requires specifying a prior distribution for the decision maker's θ, and the likelihood of the data given θ.
- The model is completely specified by three parameters: $p_1 = \mathbb{E} Y \hat{Y}$, $p_2 = \mathbb{E}(1 - Y)\hat{Y}$, and $p_3 = \mathbb{E} Y \left(1 - \hat{Y}\right)$.
- The data follows a multinomial distribution with four categories.
- Conceptually, $\theta = h(p)$, $h = (h_1, h_2)$. $\hat{p} = (\hat{p}_1, \hat{p}_2, \hat{p}_3)$ are sufficient statistics summarizing the data.
- The multinomial distribution is a completely specified parametric model, which allows for exact likelihood based Bayesian posterior distribution computation.

The Bayesian approach

- Given p, the data likelihood is

$$L\left(Y|p\right)=p_1^{n\hat{p}_1}p_2^{n\hat{p}_2}p_3^{n\hat{p}_3}\left(1-p_1-p_2-p_3\right)^{n\left(1-\hat{p}_1-\hat{p}_2-\hat{p}_3\right)}.$$

  Given a prior $\pi\left(p\right)$, the posterior distribution can be computed (e.g. by simulation)

$$p\left(p|Y\right)\propto\pi\left(p\right)L\left(Y|p\right).$$

- Projecting the simulated $p\left(p|Y\right)$ onto $\theta=h\left(p\right)$ gives a simulated posterior distribution for θ.
- Credible regions can be constructed by smoothing or contouring.

Analytic posterior calculation

- Consider a Dirichlet prior on the $K(=4)$ simplex $p_1, \ldots, p_K$, such that $p_1 + \ldots + p_K = 1$, with parameters $\alpha_1, \ldots, \alpha_K$.
- $\alpha_1 = \ldots = \alpha_K = \alpha$ is called a symmetric Dirichlet distribution.
- Given symmetry, $\alpha = 1$ corresponds to a uniform prior on $p_1, \ldots, p_K$.
- The posterior distribution is also Dirichlet with parameters

$$\hat{\alpha}_k = \alpha_k + n\hat{p}_k, \quad k = 1, \ldots, K.$$

with uniform prior, the posterior Dirichlet distribution has parameters

$$\hat{\alpha}_k = 1 + n\hat{p}_k, \quad k = 1, \ldots, K.$$

- Simulate the posterior distribution of $\theta = (\theta_1, \theta_2)$, for $s = 1, \ldots, S$, by making draws from the Dirichlet distribution with parameters $\hat{\alpha}_k = 1 + n\hat{p}_k, \quad k = 1, \ldots, K.$

- The posterior probability that the θ lies below the ROC is

$$\int_{\theta \text{ below ROC}} p\left(\theta \middle| Y\right) d\theta.$$

- The *maximum* posterior probability of a set of θ that are dominated simultaneously by a single point on the ROC (chosen as *best* point):

$$\text{maxposteriorprob} \equiv \sup_{a \in (0,1)} \int_{\theta_1 \geqslant a, \theta_2 \leqslant g(a)} p\left(\theta \middle| Y\right) d\theta.$$

  where the machine ROC has been represented by $\theta_2 = g\left(\theta_1\right)$.
- Both of these can be computed by simulation: for $\theta_r, r = 1, \ldots, R$ simulated draws from the posterior distribution, they are estimated by

$$\frac{1}{R} \sum_{r=1}^{R} \mathbb{1}\left(\theta_r \text{ below ROC}\right)$$

  and

$$\sup_{a \in (0,1)} \frac{1}{R} \sum_{r=1}^{R} \mathbb{1}\left(\theta_{1r} \geqslant a,\ \theta_{2r} \leqslant g\left(a\right)\right).$$

The frequentist approach

Consider the following heuristic procedure.

- Form a confidence set $\hat{S}$ for θ_0, for example based on bootstrapping or the asymptotic normal distribution of $\hat{\theta}$ around θ_0.
- For each $s \in \hat{S}$, define $A_s = \{\theta : \theta_1 \leqslant s_1, \theta_2 \geqslant s_2\}$. In other words, A_s is the set of points that are *better* than s (namely, lower FPR and higher TPR).
- Next define

$$A = \bigcap_{s \in \hat{S}} A_s$$

  A is the set of points that are *better* than all the points in $\hat{S}$ in the above sense.
- Next define

$$\bar{A} = A \cap \text{ROC}$$

  Namely, $\bar{A}$ is the set of points on the ROC that simultaneously dominates all the points in $\hat{S}$.

- How do we provide a statistical probabilistic statement justifying this procedure?
- In a classical approach, θ_0 is a fixed number and not a random variable. It is either above or below the ROC, without an associated probability. The confidence set itself is a random set, such that

$$P\left(\theta_0 \in \hat{S}\right) \approx 1 - \alpha.$$

Namely, if the world is to repeat 100 times, about 95 times $\theta_0 \in \hat{S}$.

- Consequently, $\bar{A}$ is a random set, that is possibly empty, such that

$$P\left(\text{all points on } \bar{A} \text{ dominates } \theta_0\right) \approx 1 - \alpha.$$

- What about the probability of the random event of $A \cap \text{ROC} \neq \varnothing$?
- Obviously $P_{\theta_0}\left(A \cap \text{ROC} \neq \varnothing\right)$ depends on θ_0 and is unknown.
- It can be estimated by replacing θ_0 with $\hat{\theta}$ and iterates on the bootstrap procedure. Depending on where the true θ_0 lies, $P_{\theta_0}\left(A \cap \text{ROC} \neq \varnothing\right)$ can be anywhere between 0 and 1.

- We turn to a Hypothesis testing setup, to formulate tests of

$$H_0 : \theta_0 \text{ lies below ROC} \quad \text{against} \quad H_1 : \theta_0 \text{ lies above ROC.}$$

The roles of the null and alternative hypotheses can be reversed, depending on which one we are more willing to theorize on, *a priori*.

- Suppose we represent the ROC curve by $\theta_2 = g(\theta_1)$, where $g(\cdot)$ is known, increasing, and concave. Then we can rewrite

$$H_0 : \theta_{20} \leqslant g(\theta_{10}) \quad \text{against} \quad H_1 : \theta_{20} \geqslant g(\theta_{10}).$$

- Let $t = \theta_2 - g(\theta_1)$, then the hypotheses can be restated as

$$H_0 : t_0 \leqslant 0 \quad \text{against} \quad H_1 : t_0 \geqslant 0.$$

An asymptotic test can then be based on combining the Delta method with a one-sided t-test.

- Supposing that $\hat{\theta}$ is below the ROC.
- We look for pairs (a, b), such that (a, b) lies on the ROC, that the following level of confidence in the superiority of the machine algorithm:

$$P\left(\hat{\theta}_1 \geqslant a, \hat{\theta}_2 \leqslant b\right)$$

- This coverage probability can be estimated by bootstrapping

$$P^*\left(\hat{\theta}_1^* \geqslant a, \hat{\theta}_2^* \leqslant b\right).$$

- An implementable program is therefore

$$\max_{b=g(a)} P^*\left(\hat{\theta}_1^* \geqslant a, \hat{\theta}_2^* \leqslant b\right).$$

then the pair (a, b) with maximized P^* is chosen as the *best* point on ROC.

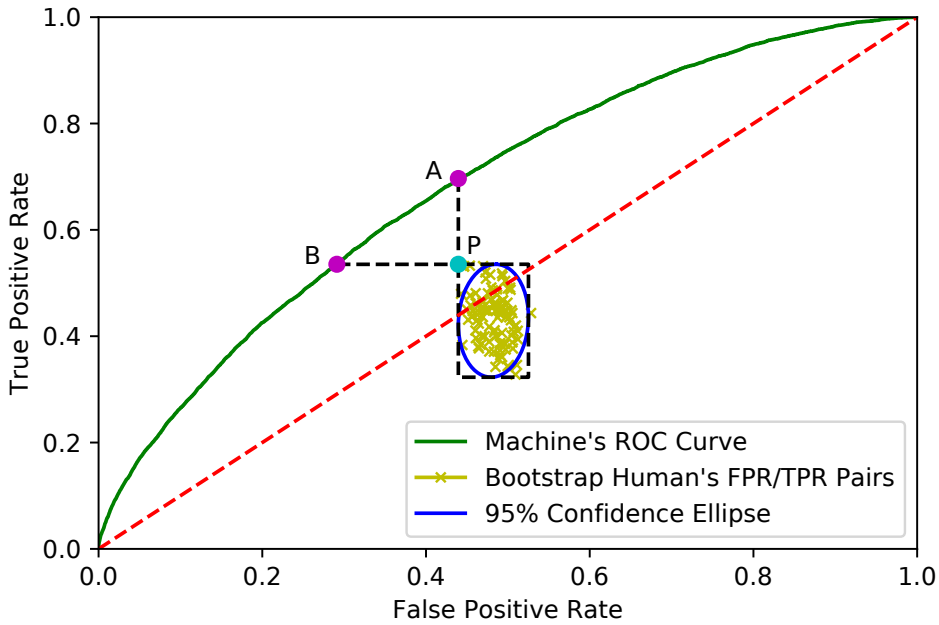
Replacing A Subset of Human Decision Makers

- Conventional confidence and credibility levels suggest that we are conservative when making decisions to replace humans with machines.
- Specifically, only when there exists a point on the machine's ROC curve, which is better than any point in a confidence set of the human's FPR/TPR of level α, the machine's decision is considered to be better than the human's decision in an α confidence sense.
- In other words, only when a sufficiently large portion of the FPR/TPR joint distribution corresponding to α confidence level is below the machine's ROC curve, we will replace the human's decision by the machine's.

Replacing A Subset of Human Decision Makers

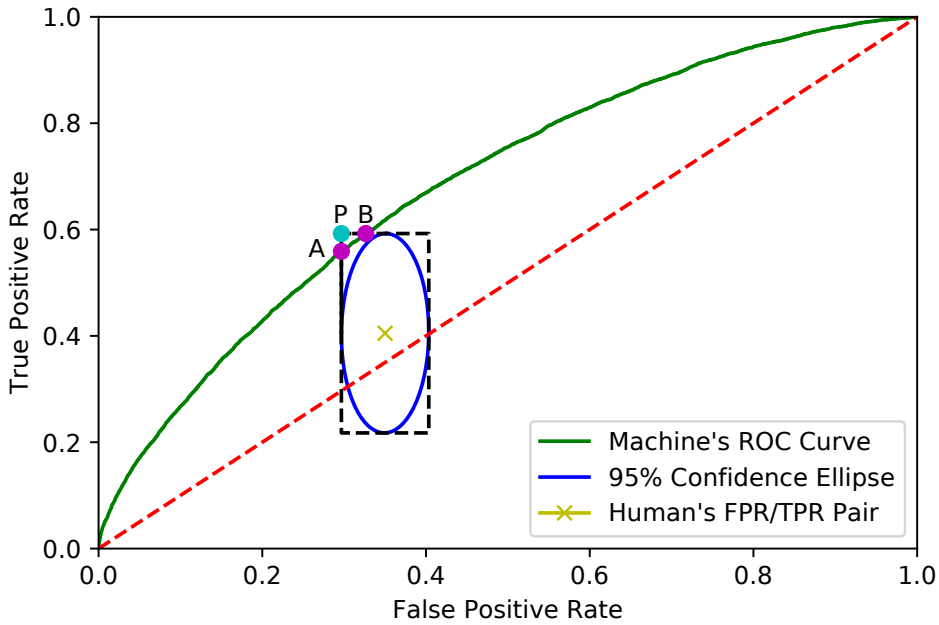
- Next, consider an illustration using a conventional oval-shaped confidence region based on the asymptotic normal distribution of the estimated human's FPR/TPR pairs.
- Denote by $(HFPR_1, HTPR_H)$ the point that corresponds to the highest TPR on the oval, and $(HFPR_S, HTPR_2)$ the point that corresponds to the smallest FPR on the oval.
- Regarding the position of the human's oval shape and machine's ROC curve, we have three cases.

- Case 1: The human's oval area and the point $P(HFPR_S, HTPR_H)$ are all below the ROC curve.



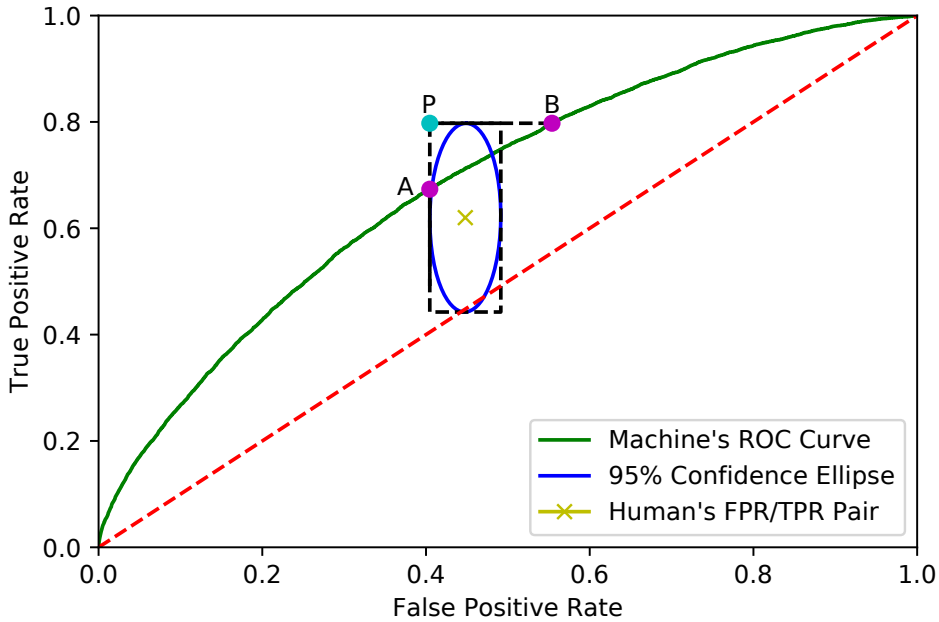
- In this case, the human decision maker is "worse" than the machine, and hence can be replaced by the machine. Any point from A to B corresponds to a better machine's decision.

- Case 2: The entire human's oval confidence area is below the ROC curve, but the point $P(HFPR_S, HTPR_H)$ is above the ROC curve.



- There's no a nonempty fraction on ROC that is better than *any* point in human's confidence set so as to account for the randomness of the estimated human TPR/FPR pair. The decision maker is not replaced.

- Case 3: The human's oval area has a certain area above the machine's ROC curve, therefore the human decision maker is not replaced by machine.



- In summary, decision makers can be divided into two groups by the relative position of point $P(HFPR_S, HTPR_H)$ and machine's ROC:

Less capable decision makers

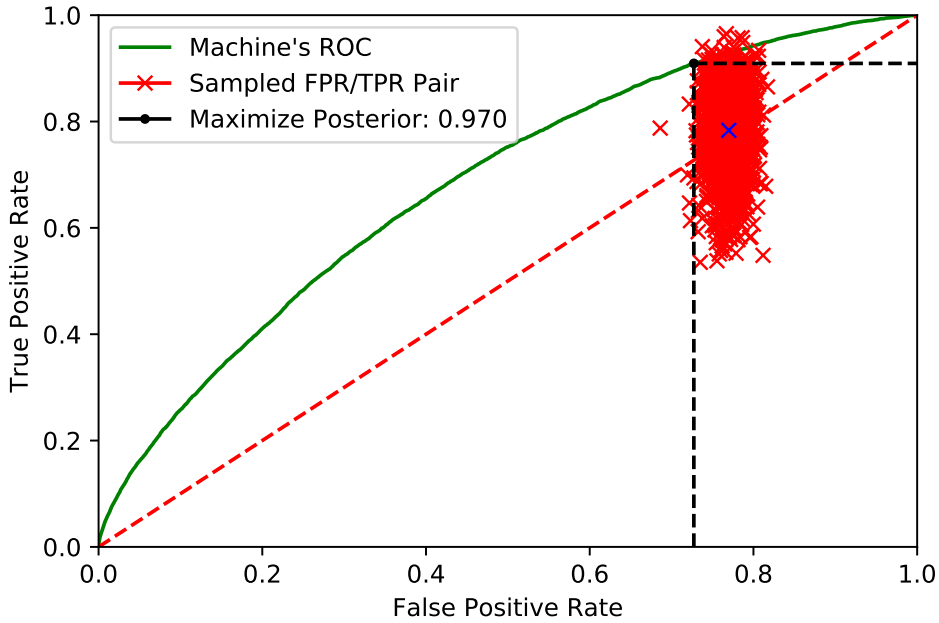
The decision maker's representative point $P(HFPR_S, HTPR_H)$ is below machine's ROC curve, which corresponds to case 1. These doctors are replaced in future decision making.

Other decision makers

These decision makers are either more capable than machines or their capability can not be precisely measured likely due to small amount of data. They hence are not replaced by the machine algorithm in future decision making.

- Given a decision to replace a human diagnosis by the machine learning algorithm, the remaining question of which point on the machine ROC curve to use can be addressed by the approaches of bayesian or frequentist.

- Note however that the previous discussion applies to any valid confidence region of any shape.
- For example, for the Bayesian approach, whether the human decision maker is replaced by machine is determined by whether the maximized posterior exceeds α, where α is a confidence level, as the figure shows.



- For the frequentist approach, it is determined by whether maximized P^* exceeds α.

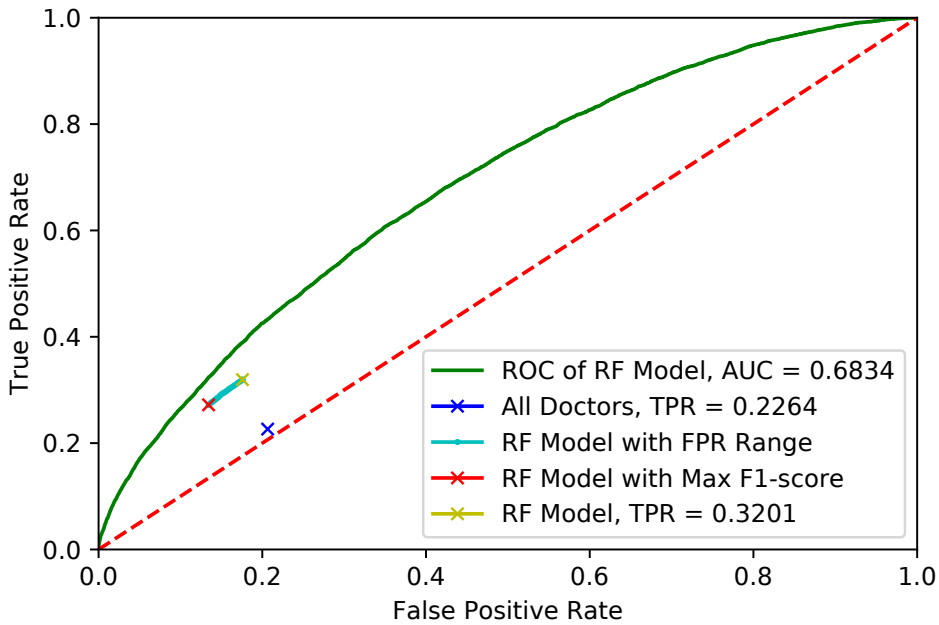
Application

- National Free Pre-Pregnancy Checkups (NFPC), a national-wise free health checkup for couples planning a pregnancy in China.
- More than 300 features are available for each observation, including age, demographic characteristics, results from medical examination and clinical test, disease and medication history, pregnancy history, as well as lifestyle and environmental information of both wife and husband.
- Contains binary pregnancy outcome, which is denoted as normal (0) or defect (1).
- Doctor's diagnosis are re-grouped into 2 levels, i.e. normal pregnancy diagnosis (0) and risky pregnancy diagnosis(1).

- The original data set includes 1,137,010 couples, who are diagnosed by 28,716 doctors.
- To improve the precision of statistical inference, we focus on doctors who diagnosed more than 300 patients, and we have a total of 584 such doctors corresponding to 584,181 cases.
- Samples are divided into two sub-samples stratified for each doctor:
  - The first subsample is for training a good machine algorithms and find less capable doctors relative to machine.
  - The second subsample is for testing the performance of machine assisting human decision. Patients from less capable doctors are diagnosed by machine, while other patients are diagnosed by their original doctors.
- Random Forest (RF) is trained as machine algorithm, which is a well-known and commonly used ensemble learning algorithm proposed by Breiman.
- The confidence level α is set to 95% for all experiments.

Result of Asymptotic Confidence Set with Random ROC Sampling

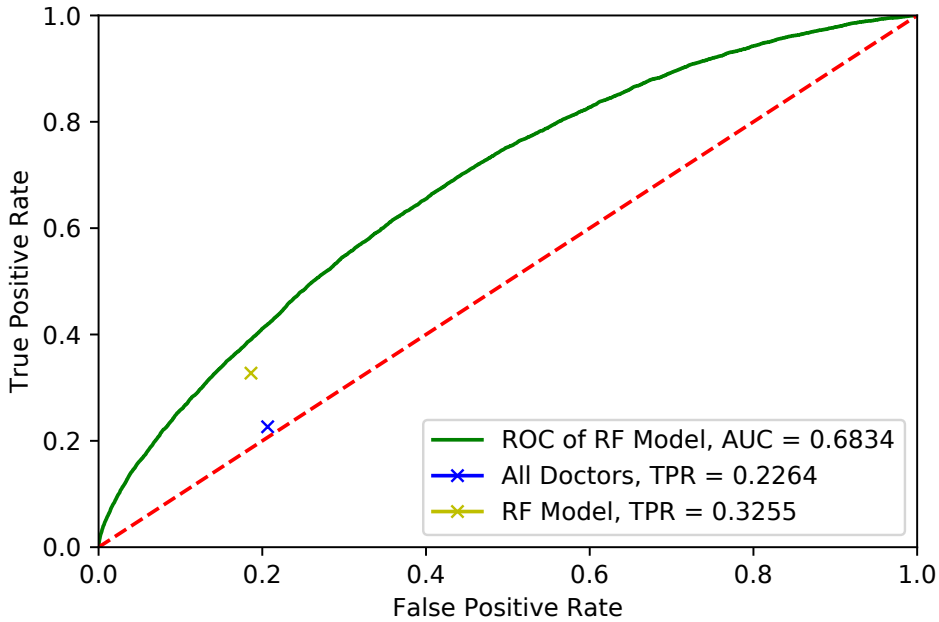
- Aggregated FPR of all doctors is at 0.2065, TPR is at 0.2264.
- After replacing less capable doctors (64% of all doctors) with machine, the overall FPR is at 0.1770, TPR is at 0.3201.



- 41.4% improvement of TPR and 14.3% reduction of FPR.

Result of Bayesian Approach

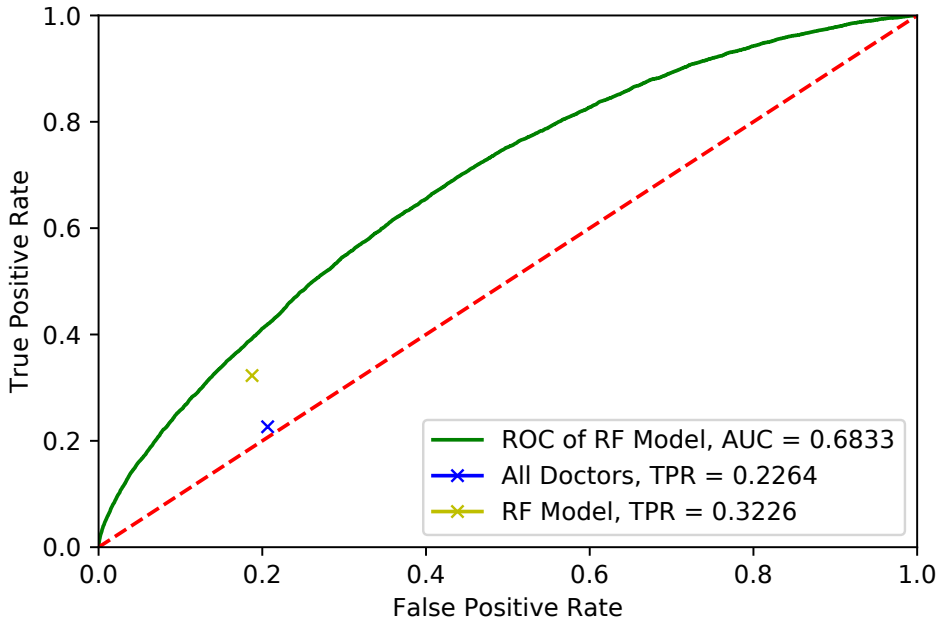
- Aggregated FPR of all doctors is at 0.2065, TPR is at 0.2264.
- After replacing less capable doctors (44% of all doctors) with machine, the overall FPR is at 0.1871, TPR is at 0.3255.



- 43.8% improvement of TPR and 9.39% reduction of FPR.

Result of Frequentist Approach

- Aggregated FPR of all doctors is at 0.2065, TPR is at 0.2264.
- After replacing less capable doctors (45% of all doctors) with machine, the overall FPR is at 0.1885, TPR is at 0.3226.



- 42.5% improvement of TPR and 8.73% reduction of FPR.

Geographic Characteristics of Replaced Doctors

- These approaches provide methods for filtering incapable doctors that have poorer diagnostic capability than machines.
- Do the replaced doctors have similar geographic characteristics?
- Regression analysis on whether a doctor is replaced by machine is needed.
- For each doctor, a dummy variable "incapable" is set representing whether she/he is replaced by machine (1) or not (0).
- Two geographic variables for each doctor are generated:
  - Provincial GDP of doctor's location from China City Statistical Yearbook of 2014, which coincides with the start year of data.
  - A dummy variable indicating whether the doctor comes from township (lower-level) clinic (1) or not (0).
- Conduct regression analysis of "incapable" on these two variables.

Geographic Characteristics of Replaced Doctors

- Doctors from township clinic have significantly higher probability to be replaced by machines relative to those from urban clinics.
  - Bayesian approach: 50.5% of township clinic doctors are replaced, while the rate of replacement in urban clinic is 40.2%.
  - Frequentist approach: 50.2% of township clinic doctors are replaced, while the rate of replacement in urban clinic is 39.3%.
- The provincial GDP has a negative coefficient in the regression, although with less strength of statistical significance than the township dummy.

Summary

Artificial-intelligence technology tends to benefit rural area more than developed cities in decision making.

Table: Replaced Doctors and Geographic Characteristics

Panel A: Bayesian Approach

	(1) OLS	(2) OLS	(3) OLS	(4) Logit	(5) Logit	(6) Logit
gdp_all_province	-0.0971		-0.0773	-0.393		-0.317
	(-0.74)		(-0.59)	(-0.74)		(-0.59)
township=1		0.103	0.101		0.417	0.411
		(2.44)	(2.40)		(2.43)	(2.39)
Constant	0.490	0.402	0.427	-0.0373	-0.399	-0.296
	(10.59)	(12.63)	(8.02)	(-0.20)	(-3.05)	(-1.37)
Observations	563	563	563	563	563	563
R^2	0.001	0.011	0.011			

t statistics in parentheses

Table: Replaced Doctors and Geographic Characteristics

Panel B: Frequentist Approach

	(1) OLS	(2) OLS	(3) OLS	(4) Logit	(5) Logit	(6) Logit
gdp_all_province	-0.184		-0.164	-0.752		-0.678
	(-1.40)		(-1.25)	(-1.40)		(-1.25)
township=1		0.108	0.105		0.439	0.427
		(2.56)	(2.48)		(2.55)	(2.47)
Constant	0.513	0.393	0.447	0.0533	-0.433	-0.214
	(11.09)	(12.39)	(8.41)	(0.28)	(-3.30)	(-0.98)
Observations	563	563	563	563	563	563
R^2	0.004	0.012	0.014			

t statistics in parentheses

Conclusion

- By assuming that preference of individual decision maker is constant, we can make comparison on the ability of decision making between human and machine, and identify the less capable decision makers.
- Given that a machine is chosen to replace a human decision maker, we propose decision rules based on the machine's ROC, including randomization and maximizing the level of statistical confidence.
- Replacing these less capable decision makers with machines, we would obtain improved performance in the quality of future decisions.

The End