Modeling Human Decision Making: A Study for Choice Prediction Competition 2018

Yuki Minai (ym1942@nyu.edu)

Yakun Wang (yw3918@nyu.edu)

Nanxi Cheng (nc2433@nyu.edu)

Abstract

Research on modeling human decision making has been getting a lot of attention in various fields such as psychology and cognitive science. While many models have been proposed to capture specific decision-making behaviors, as those models focus on a particular situation or behavior, they often result in poor generalization ability. In light of such problem, at Choice Prediction Competition (CPC18), researchers were challenged to build one model that aimed to capture 14 human anomalies. Focusing on the same problem in our project, we analyzed the performance of benchmark models, and then tried to develop our own neural network models. We used data augmentation methods and synthetic datasets to improve our model performance. As a result, the model that used part of the psychological features and augmented with synthetic data priors got a competitive result based on MSE metric. In addition, through the process of modeling, we realized the importance of both the amount of data and psychological features when fitting neural networks, and we managed to reach a balance between feature engineering efforts and data augmentation efforts.

Introduction

The modeling of human decision-making process has been a valuable research topic for years. Accurate modeling of human decision making is of great help in various areas, from economic decisions to predictions of human agent behavior in autonomous vehicle scenarios(Plonsky, Apel, Erev, Ert, & Tennenholtz, 2018). Human decision making often deviates from rational optimal choices, which has been described in many psychological theories, including Prospect theory, St. Petersburg paradox(Erev, Ert, Plonsky, Cohen, & Cohen, 2017), etc. However, the decision-making models in cognitive science and psychology have mainly been developed to capture specific interesting behavioral phenomena, with different models targeting different phenomena. As a result, many behavioral models exhibit poor prediction performance in other tasks.

As an attempt to solve this problem, two prediction competitions, CPC15 and CPC18, were introduced, in which researchers were challenged to develop a single model that can capture 14 classical choice anomalies documented in behavioral decision research(Plonsky et al., 2018). Results of the competitions suggested it is feasible to develop a model that can predict human decision making with high accuracy, capturing 14 human behavioral phenomenon.

In this project, we examine and compare the performance of several human decision making models, including BEAST, PsychForest, and our new model. We also use synthetic dataset to improve performance of the model, and analyze the results.

Dataset

In this project, we explore the modeling of human decision making by using the dataset provided by CPC18, in which a total of 210 problems are given as the training set, and another 60 are provided as the test set. The data was collected by the experiment paradigm described in (Plonsky et al., 2018). In each problem, the decision maker gets descriptions of two monetary prospects and choose between them. Such process is repeated 25 times for each problem. These 25 trials are divided into 5 blocks and 5 trials each, and in the first block the decision maker receives no feedback about the payoffs while receiving full feedback in all later blocks. The rate of choosing option B in each block is the mean of the five trials within it. A problem is defined by 12 parameters as shown in Table 1, including the descriptions of two lotteries, and game settings.

Parameter	Meaning
Ha	Expected value of payoff of Lottery A
pHa	Probability of getting payoff of Lottery A
La	Alternative outcome of Lottery A
LotNumA	Number of possible outcomes in Lottery A
LotShapeA	Shape of distribution of Lottery A
Hb	Expected value of payoff of Lottery B
pHb	Probability of getting payoff of Lottery B
Lb	Alternative outcome of Lottery B
LotNumB	Number of possible outcomes in Lottery B
LotShapeB	Shape of distribution of Lottery B
Amb	Whether prob. of outcomes in B are revealed
Corr	Correlation between two Lotteries

Table 1: Problem Parameters

Related works

In our problem setting, there are two baseline models that are of most significance, BEAST and PsychForest.

BEAST BEAST(Best Estimate And Sampling Tools) is a pure behavioral simulation-based model with outstanding performance. In CPC15, all 12 top models are variants of BEAST, and in CPC18, one of the baseline models provided is a refinement of BEAST assuming subjective detection of dominance (BEAST.sd). In BEAST, the prediction of choice is calculated by considering three terms: a) estimation of expected value, b) average of samples drawn from the lottery distributions, and c) estimation noise. For the second term,

the samples are drawn by using one of four simulation tools that assume pessimism, equal weighting, sign and unbiased respectively. The estimation noise fluctuates according to the complexity of problems, being lower for easy problems and higher otherwise.

PsychForest PsychForest(Psychological Forest) is another baseline model provided in CPC18. It is based on off-the-shelf random forest model that uses engineered psychological features and foresight from BEAST model. The features used in PsychForest and their corresponding brief descriptions are illustrated in Table 2. These features are designed according to research in social science and psychology, capturing phenomena including sensitivity to difference between EV, pessimism under ambiguity, reflection effect, etc.

Methods

In this section we introduce the methods we used in this project.

Synthetic datasets

The original data provided in the competition is collected from real human responses. While being relatively large in its kind, its scale is still insufficient for many models, especially for neural networks which normally require large amount of data to perform well. To better utilize such significant branch of cognitive models, it would ideal to augment the original data. In *Cognitive Model Priors for Predicting Human Decisions*(Bourgin, Peterson, Reichman, Griffiths, & Russell, 2019), Bourgin et al. presented a way to solve this problem of data scarcity. They suggested that by using a certain psychological theory-based models a synthetic dataset can be generated to pretrain a neural network model, after which real human data are used to fine-tune the network. Intuitively, in this way, we would be first training a neural network to approximate the psychological theory-based model, and then doing slight adjustment to adapt to the real human data. The transfer of foresight knowledge between the theoretical model and neural network is through the generated synthetic dataset. In our exploration of the CPC18 challenge, we adopted this approach to generating synthetic datasets for training neural networks. Details are in following sections.

Generate New Problems As described in the Dataset section, a problem is uniquely defined by a 12 dimension vector. To generate new problems, we referred to the problem selection algorithm introduced by (Plonsky et al., 2018) in Appendix D. Using this procedure, 1k new problems are generated to produce the synthetic dataset.

Construct Cognitive Priors Theory-based models might provide more generalization than conventional machine learning methods (Bourgin et al., 2019), which is desired for a good model. To leverage this feature, Bourgin et al. proposed to extract and transfer foresight of theoretical models by using them to perform predictions on the new problems sampled in the last step. Specifically, BEAST model is used. The synthetic dataset is then complete and used for pretraining the neural networks.

Neural Networks

In this part, we would show the neural network models we tried. The features in Table 1 were treated as *raw features*, as they were parameters that directly defined the problem space. We further divided the features in Table 2 into two clusters: *deterministically* computed ones and *simulation-based* ones (all features of $pBetter$ type). We argue that, since one goal of neural networks is to minimize feature engineering efforts, we should avoid computing the expensive simulation-based features and excluded them in the neural network models. Therefore, in all below models, only raw features and easily computed deterministic features were used.

Data augmentation by replacing two option One important aspect of implementing deep learning is to avoid overfitting and one method often used to do it is data augmentation. With image dataset, for example, rotating, shifting, and scaling are used to increase the number of a dataset. In (Hartford, Wright, & Leyton-Brown, 2016), the authors augmented their dataset by replacing the conditions for a column player and for a row player in a simple two-player game. They assumed that subjects are indifferent to the order in which options are presented, implying invariance to permutations of two choices. In order to augment our dataset, by referring their method, two choice conditions for each game were replaced (Hartford et al., 2016). This makes the amount of data double to prevent overfitting.

Regular neural network In order to evaluate the performance of regular neural network architecture, a three-layer network of 50 hidden units for each was used as the preliminary neural network model architecture. ReLu activation was implemented as the activation function for the first two layers. We later took (Bourgin et al., 2019) as reference and adjusted the network structure to a network with larger hidden space: three hidden layers, each consisting of 200, 275, 100 hidden units respectively. The dropout rate of each layer is set to 0.15, and the optimizer is SGD.

Bias-Prediction Model To eliminate the Neural Networks' tendency of underestimating the irrationality from human participants, we decided to build a model that encode the cognitive bias explicitly. It uses the same basic assumption as used in the BEAST simulation, where Option A would be strictly preferred over Option B, after trial r, if and only if:

$$[BEV_A(r) - BEV_B(r)] + BIAS_{AB}(r) + e(r) > 0$$

where BEV_j is the best estimate of the expected value of option j, e is an error term and $BIAS_{AB}$ is the bias towards option A. In BEAST, $BIAS_{AB}(r) = ST_A(r) - ST_B(r)$, where ST_j is the value of option j based on the use of sampling tools, while in our model, this bias is directly encoded. The model structure is shown in Figure 1. Each neural network module consists of two feed-forward layers with batch normalization and ReLU

Feature Set	Features	Description
Domain-relevant	dEV	Difference between 2 lotteries' expected value
	dSD	Difference between 2 lotteries' standard deviations
	$dMins$	Difference between 2 lotteries' minimal outcomes
	$dMaxs$	Difference between 2 lotteries' maximal outcomes
Psychological	dEV_0, dEV_{FB}	(Est.) difference between expected values(EV), before/after feedback
	$pBetter_0, pBetter_{FB}$	(Est.) prob. that one lottery produces better outcome, before/after feedback
	$dUniEV, pBetter_U$	Uniform heuristic that treats both lotteries' distribution as uniform
	$dSignEV, pBetter_{S0}, pBetter_{SFB}$	Sign heuristic that neglects the magnitude of outcomes
	$dMins$	Minimax heuristic that favors choice with higher minimal outcome
	$SignMax$	Whether gains are possible
	$RationMin$	Ratio between minimal outcomes
	Dom	Whether one choice dominates the other

Table 2: PsychForest Features

activation. The hidden features are 15-dimensional vectors.

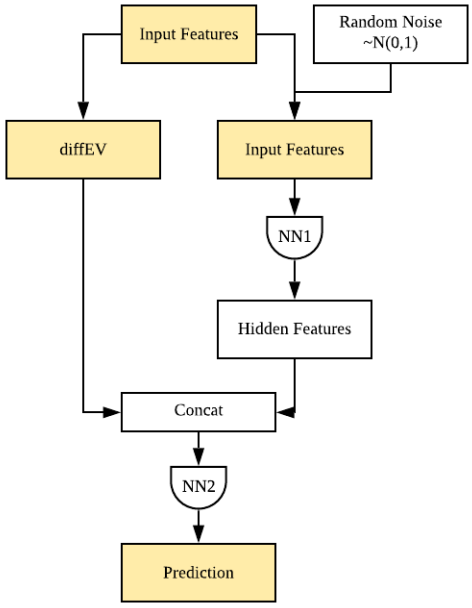

Figure 1: Network Structure of the Bias-Prediction Model

Results and Discussion

BEAST Simulation and Traditional Machine Learning Models

In this part, we first replicated the two benchmark models mentioned in the contest description: the BEAST model and the PsychForest model. Besides, as it could be noticed that PsychForest was just a random forest model using features extracted from BEAST simulations, we also implemented a XGBoost model using the exact same features. XGBoost is known as one of the best traditional machine learning models. By implementing such a model, we aimed to quickly come up with a model comparable to the benchmarks, so that we could further perform our own analysis based on it.

These three models are trained only on provided datasets, without any data augmentation nor synthetic datasets. The first three lines of Table 3 showed their performance on the evaluation set. BEAST model is very competitive and the newly added XGBoost model outperformed the PsychForest one.

Model	MSE$\times$100
BEAST	0.69
PsychoForest	0.83
Tuned XGBoost	0.76
PsychoForest without BEAST prediction	1.19
Tuned XGBoost without BEAST prediction	1.17

Table 3: Performance (MSE) for CPC18 benchmarks and XGBoost

Analysis of Feature Importance Before developing our neural network models, we would like to secure some guidance by first analyzing which of the features are most important in the model when it came to the prediction. The package lime (short for local interpretable model-agnostic explanations) was applied to the XGBoost model to provide case-based quantitive interpretations (Ribeiro, Singh, & Guestrin, 2016). For each of the block 1, 3 and 5, explanation of a randomly sampled problem is shown in Figure 2. In every case, the features were ordered by the absolute values of their local weights, which could be approximately interpreted as their importance in it. It is obvious that the XGBoost model depended heavily on the BEAST predictions. Indeed, out of 300 cases (60 problems by 5 trials each), the BEAST prediction got the first rank in all of them. Hence, we came to the conclusion that the BEAST simulation results actually dominated the XGBoost prediction. The findings still holds for the PsychForest model. As we argued before, it could be unfair to integrate the predictions from BEAST into a machine learning model if we would like to compare the performance

of those two kinds of models. In order to provide a more reasonable benchmark model for our next steps, we excluded the BEAST simulation results from the features and reran the models. The last two lines of Table 3 shows the adjusted MSE scores.

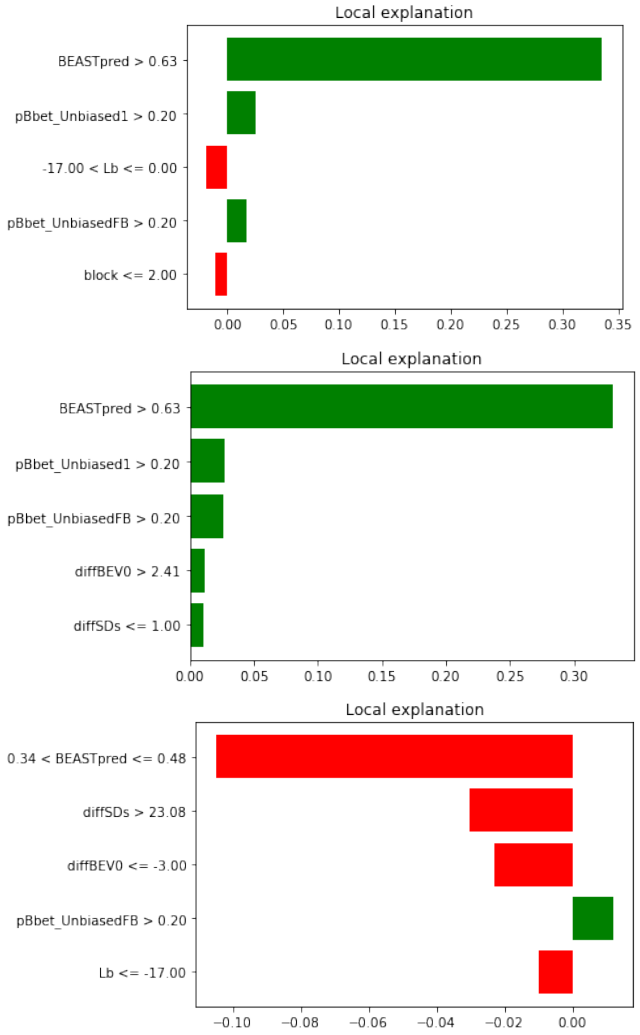

Figure 2: Local Explanations. Included Cases from Top to Down: Problem 25, Block 1; Problem 18, Block 3; Problem 45, Block 5

Note that we still keep the other simulation-based features in these models. However, we excluded them in our following neural network attempts.

Regular neural network performance

In a previous study, the authors showed that a deep neural network architecture can outperform the best available behavioral models in a different task of prediction human choices (Hartford et al., 2016). Although their model architecture depends on their task setting and it is difficult to apply the model in general problem settings including our problem, their research encourages us to evaluate a performance of deep neural network architecture. As the first step, we compared performances of a regular neural network and baseline models (i.e., BEAST model and PsychForest model).

Table 4 compares the performance of baseline models with that of a regular neural networks, with and without data augmentation, also with and without using synthetic datasets.

Model	MSE×100
NN	7.60
NN with data augmentation	3.77
NN with synthetic data	1.88
NN using both techniques	1.63
BEAST	0.69
PsychForest	0.83

Table 4: Performance (MSE×100) for NN

Approach to generating synthetic dataset is described in previous section. Regarding the data augmentation method, two choice conditions for each game were replaced by referring to the approach used in the previous study (Hartford et al., 2016). Although both methods to increase the amount of our dataset contributed to improve the model's performance, the performance comparison showed that even the neural network model with both synthetic datasets and the data augmentation method couldn't outperform baseline models.

We suspect that this may be because the deep neural network model sometimes fails to capture the irrational choice behaviors of subjects. Figure 3 shows that the observed probability from subjects (a) and the predicted probability by each model (b, c, and d) to choose option B as the function of the gap of the expected value between each option (expected value for the option B minus expected value for the option A). In general, the probability to choose option B becomes linearly higher as the expected value gap becomes bigger in all figures, which means that the subjects and models make rational decisions based on the expected value for each option. However, observed probabilities shown in Figure 3(a) sometimes deviate from this linear trend and the distribution has relatively high variability. While predictions by two baseline models shown in Figure 3(b) and Figure 3(c) also have a relatively high variability sometimes deviating from the linear trend, the neural network model tends to make more stable predictions with a lower variability as shown in Figure 3(d). According to the previous study, two baseline models can capture human's irrationalities by assuming high sensitivity to 4 humans' tendencies: pessimism, bias toward equal weighting, sensitivity to payoff sign, and an effort to minimize the probability of immediate regret (Erev et al., 2017). However, the neural network model didn't assume any specific humans' tendencies, which, we suspect, creates the performance gap between the neural network models and two baseline models.

The above analysis revealed that the regular neural network model lack the capacity of capturing the cognitive bias. An intuitive way to solve it is to enlarge the hidden space. We

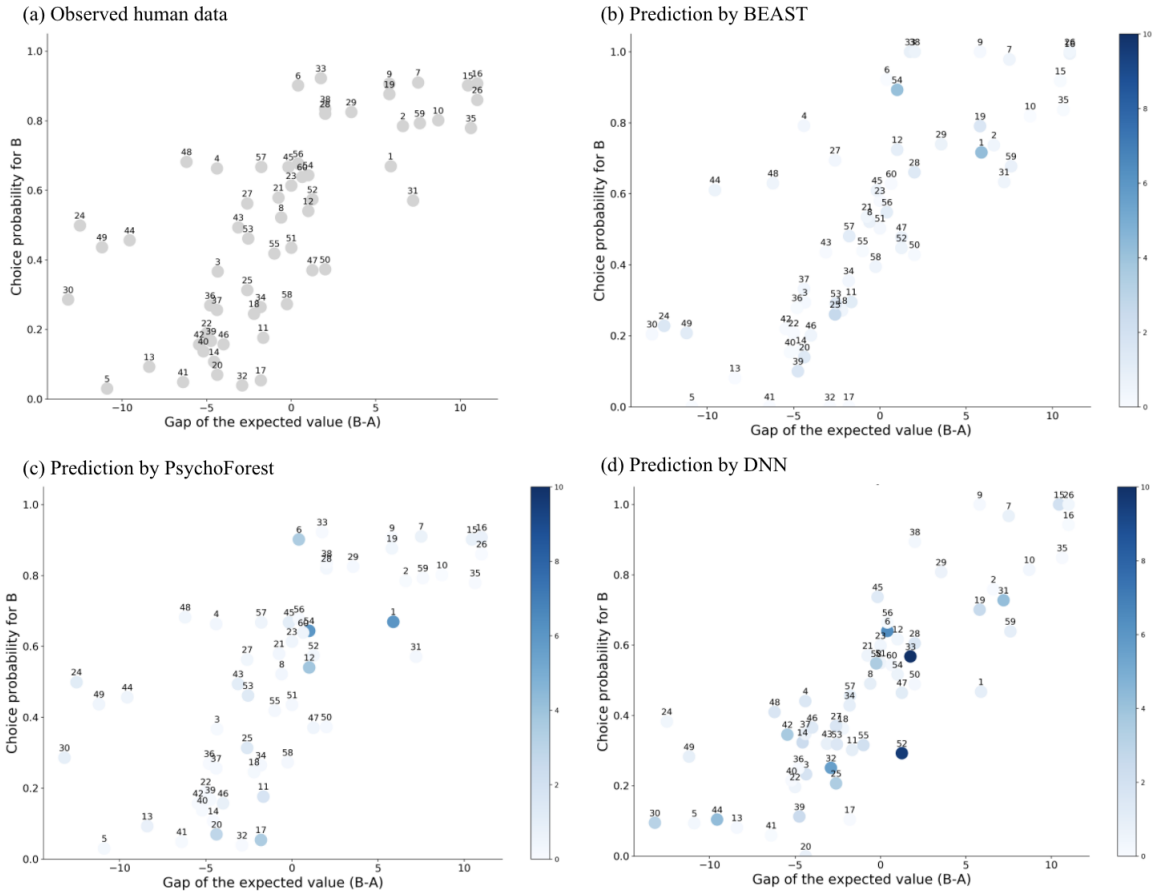

Figure 3: The choice probability as the function of the gap of the expected value
The annotated labels denote the gameIDs and the color gradations denote the values of MSE. (a) choice probabilities by human subjects, (b) choice probabilities by BEAST model, (c) choice probabilities by PsychForest model, (d) choice probabilities by DNN

were not able to do so under data scarcity, but as we have proved the effectiveness of synthetic dataset and data augmentation approach, we could now try to build a much larger neural network. Dropouts were also deployed here to force structure regularization. The result of the refined neural network is shown in Table 5. This model worked surprisingly well.

	With SynthData1k	No SynthData
PF Features[1]	0.96	1.49
Only Raw Features	5.91	6.34

Table 5: Results of Refined Neural Network

The Bias-Prediction Model

See Table 6 for the results comparison. As this model has more parameters than the previous neural network model, it overfitted extremely easily when neither data augmentation nor synthetic dataset were used. With both techniques applied, the model could reach a descent result, outperforming the PsychForest model without BEAST predictions. However, this model did not defeat the adjusted regular network model proposed before. This might due to the small hidden space allocated in the model. We kept a small hidden space here with the aim of avoiding overfitting, but neural network models still got no advance over traditional models when the data is scarce, and when we adopted the large synthetic datasets, the model here did not have enough capacity to curve the complex underlying patterns.

Model	MSE×100
Bias-Prediction Model with data augmentation	1.62
Bias-Prediction Model with synthetic data	1.29
Bias-Prediction Model using both techniques	1.05

Table 6: Performance (MSE) for Bias-Prediction Model

[1] Only deterministic psychological features

Rationality in Human Decisions

It is natural to assume that humans would always want to maximize their gain in a gamble. By this standard, we may find a "rational" choice for each problem: choose the one choice with higher expected payoff. In reality, however, human behaviors quite often deviate from such rationality, and capturing such phenomena is also part of the goal of this competition. To explore this, we analyzed the correlation between prediction results from multiple models and the absolute "rational" choices. By calculating the correlation factor between the "rational" choice and multiple models, we obtain the following tables.

Real human data	BEAST	PsychForest	NN
0.7631	0.7911	0.8276	0.7909

Table 7: Correlation With "Rational" Choices

	Block 1	Block 2	Block 3	Block 4	Block 5
Human data	0.7209	0.7787	0.7707	0.7718	0.7760
BEAST	0.7421	0.7988	0.8073	0.8097	0.8072
PsychForest	0.7867	0.8241	0.8385	0.8427	0.8456
NN	0.7716	0.7808	0.7852	0.8131	0.8056

Table 8: Correlation With "Rational" Choices Per Block

As can be seen from Table 7, all three models seem more correlated with rationality than the actual human data, which might suggest that there is still room for improvement in terms of capturing human decision anomalies. Besides, from the first line in Table 8 it is clear that the degree of rationality of human choices exhibits a evolving progress, with a relative large gap between the first and second block, which may be the influence of feedback. Such phenomenon is captured by all three models, although in NN model it is not as evident as in the other two models.

Comparison with results from the Original Cognitive Priors Study

In (Bourgin et al., 2019), by generating 13 thousand problems, their model using only *raw features* reached MSE($\times$100) score as low as 0.48 on the evaluation set, which is more competitive than our results. However, we found that generating cognitive priors is extremely expensive. It took around 2 days to generate 10 thousand problems. Referring to Table 5, if we deploy the same model structure as that of the original study and only use 1 thousand synthetic problems for pre-training, the performance, although slightly improved compared with no synthetic dataset, would be really disappointing, with MSE($\times$100) being 5.91. By taking advantage of deterministic psychological features, our model managed to reach a balance between feature engineering effort and prior data generation effort.

Conclusion

In this paper, we evaluated methods to augment the dataset and different machine learning models to build a model for capturing human decision-making behaviors. In the evaluation of features using XGBoost, it was revealed that the prediction result of the BEAST model was the most important feature. Since the BEAST model makes predictions based on simulations, the effectiveness of the simulation-based approach to predict the humans' anomaly behaviors were shown. However, in order to evaluate the performances of other new models such as a neural network, the features based on the simulation were excluded from the fitting processes. Through the evaluation of a neural network model, it was shown that the synthetic dataset and the data augmentation method contributed to the performance improvement of the model. Our best performed model is an adjustment of the regular neural network model, where we enlarged the hidden space and adopted dropout to force sparse connections. It got a 0.96 MSE$\times$100 score. Although it still could not outperform the pure cognitive BEAST model, it beat the PsychForest model, which got a score of 1.19 without direct information from BEAST predictions. Also, it's easier to train and deploy, compared to the time-consuming BEAST simulations and 13-thousand synthetic dataset generation process.

Unlike areas in which deep learning models such as CNN and RNN have achieved great success, there are no general models that can produce high accuracy from limited data in human decision modeling. We believe that further research is necessary to develop a model that can identify important features from data and capture human irrational behavior in general settings.

References

Bourgin, D. D., Peterson, J. C., Reichman, D., Griffiths, T. L., & Russell, S. J. (2019). Cognitive model priors for predicting human decisions. *arXiv preprint arXiv:1905.09397*.

Cpc18: Choice prediction competition 2018. (2018). `https://cpc-18.com/`. (Accessed March 30, 2020)

Erev, I., Ert, E., Plonsky, O., Cohen, D., & Cohen, O. (2017). From anomalies to forecasts: Toward a descriptive model of decisions under risk, under ambiguity, and from experience. *Psychological review*, *124*(4), 369.

Hartford, J. S., Wright, J. R., & Leyton-Brown, K. (2016). Deep learning for predicting human strategic behavior. In *Advances in neural information processing systems* (pp. 2424–2432).

Plonsky, O., Apel, R., Erev, I., Ert, E., & Tennenholtz, M. (2018). When and how can social scientists add value to data scientists? a choice prediction competition for human decision making. *Unpublished Manuscript*.

Plonsky, O., Erev, I., Hazan, T., & Tennenholtz, M. (2017). Psychological forest: Predicting human behavior. In *Thirty-first aaai conference on artificial intelligence.*

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, san francisco, ca, usa, august 13-17, 2016* (pp. 1135–1144).