Final exam

Oskar Escobar Sossa

1. **Overview: Summarize the dataset and its origins**

The data comes from Choice Prediction competition 2018 (CPC18), Part of these data were previously used in CPC15. We will work with two datasets containing the same variables, one is the calibration data that includes 510,750 consequential choices of human decision makers choosing between two risky and/or uncertain prospects with up to 10 possible outcomes each, this will be our training data. The other is the competition data that contains the same information but for 3750 choices, this will be our test data.

Among the variables we can include Location, Gender, Age, Game identification variables, Expected value of (High) lottery in Option A or B, Probability to get payoff drawn from lottery in Option A or B, Low payoff in Option A or B, Shape of lottery, Number of lottery outcomes in Option A or B. Payoffs, both received and missed.

2. **Methodology**

Basically, we will run three neural net models for prediction. In the first model we will use the variables included in the original dataset with a bit of cleaning to support our prediction. In the second model we will include some risk-related variables, and finally in model 3 we will include some attention-related variables.

In the end, we include an extra model, for this case we will exclude the variables with the lowest correlation and change the layers of the model, Trying to maximize the accuracy of our model by only keeping the "best" predictors. At the end we also present a table with the correlation between our decision variable and all the other variables, including our created features.

2.1 First model

For the first model, we will work with the basic variables provided in the original dataset with a bit of cleaning, for example, some variables that are descriptive strings will be turn into numeric in order to our Neural Network be able to find patterns, like male or female which are replaced with 1 and 2. Some other changes are:

- **Location:** Rehovot and Technion are labeled as 1 and 2 respectively
- **Gender:** Female are labeled as 1 and male as 2
- **Condition:** ByFB is labeled as 1 and ByProb as 2.
- **Lottery shapes:** These are labeled as "-" for lotteries with a fixed payoff, Left-skewed and Right skewed lotteries and Symmetric lotteries. We will label fixed lotteries as 1, Symmetric as 2, Left-skewed as 3 and Right-skewed as 4

Then we drop variables that identify the player but we keep variables that identify the game. CP18 has 210 different games.

We will take our ABKK_experiment_pilot.jl as a base for the code, we will use the functions: get_processed_data, train, test. But adapting these to fit our narrative, instead of 6 outputs, in this case we only have 2 possible outcomes.

2.2 Second model

Now we will run our second model by adding risk variables, this variables will be:

- **Risk_LotA (or B):** Based on the shape of the lottery we classify the games by their risk levels. Something interesting is that we were lucky enough to use this classification when we transformed the lottery shape into variables. A lottery with a certain value has a level of risk equal to 1, Left-skewed equal to 2, symmetric equal to 3 and right-skewed equal to 4. We will explain this more in the Features section.
- **Expected value for each lottery**: Although we do not count with all the lottery information, we can have an estimate of the expected value for the lottery using the highest and lowest pay with their respective probabilities.
- **Expected value of B is higher.** Even though we know that in reality, people tend to use expected utility rather than expected value, we will also include a dummy if the expected value of lottery B is higher.
- **Expected value of highest payoff and comparison:** As a personal opinion, I consider that in hypothetical games like these, people tend to focus more on the expected gains rather than the losses given to the nature of the experiment, this is why I will compute the expected highest outcome of each lottery (highest payoff times the probability of occurrence) and compare if it is higher for lottery B.

2.3 Third model

Now, we are interested in seeing the variables related to attention, we are already including some of this variables from the first model, this are:

- **Reaction Time (RT)**: The time taken by individuals to make a choice between option A and option B. A shorter reaction time may indicate a higher level of attention or a more confident decision, while a longer reaction time may suggest either indecision or distraction. It is hard to determine which one
- **Button:** The on-screen side of the chosen button (e.g., "L" or "R"). Analyzing the distribution of choices between the left and right buttons can provide insights into attentional biases or preferences.
- **Block**: The number of time-blocks within the current game. Analyzing changes in choice behavior across different blocks can provide insights into attentional fatigue or learning effects over time.
- **Trial**: The trial number within a game. Analyzing changes in choice behavior across different trials within a game can provide insights into attentional changes or strategies over time.
- **Feedback**: A binary variable indicating whether full feedback was provided to the participant regarding payoffs in the current trial. Participants who receive feedback may pay more attention to their choices and adjust their decisions accordingly.
- **Mean reaction time and the comparison with subject:** we estimated the mean response time by game and by block and also create a dummy equal to 1 if the response time is higher than the mean response time of the block for each observation
- **Probability of an option and consistency** we will estimate the probability of choosing an option by game and block.

- We will also estimate the probability of choosing an option by game, block and subject .
- Then, we create a dummy if the subject probability to choose an option is higher than the group probability.
- We will also compare if the subject probability to choose an option is consistent with the probability of the group, it is consistent if the difference between both probabilities is less than 25% in absolute value.

2.4 Extra model

At the end we will run an extra model by excluding the variables with low correlation with our choice variable. For this we will present a correlation table to analyze the performance of all our existing and created variables.

3. Feature analysis

We will explain a bit more about our new variables.

Risk: For the risk variables, we have as the most important the way we measured the level of risk of the game. For example, a certain lottery is not risky at all so we put a value of 1, a Left-skewed lottery has lower payoffs in general with a small chance of get a high payoff, so the risk is a bit higher, however, since I consider this a "pessimistic" game then is likely that the users don't see too much risk since they know the majority of outputs are low (value of 2). Symmetric lotteries are more risky than the previous, this time there is a good and a bad outcome with similar probabilities, (value of 3). And finally, Right-skewed lotteries are the more risky because even if the payments are relatively high, the possibility of failure implies a bigger loss for the player (value of 4).

The other variables are more related with expected values, in this case we don't have information about possible utilities so we could estimate Von Newman - Morgensten utilities, and although the expected values do not pass the Bernoulli Paradox, I consider that in hypothetical decision making it may be a good measure. And, as mentioned above, given the nature of a game in which agents do not really lose anything, I consider they pay more attention to the possible rewards, therefore giving an extra value to the expected high outcome.

Attention: About the attention variables, many are already included in the original dataset (a possible reason for the good behavior of our first model) but then, we introduce more variables. I consider the most important the average response time and the consistency with the group. Sadly, we do not count with enough information for the response time of all participants. As a measure of the time required to analyze each game, we take the mean by game and block in order to control for possible learning across the repeated games. If a person takes a longer time than the group, it may be a signal that it is paying more attention to analyzing the problem with more depth. About the probability of choosing B and consistency with the group. Even though each person can have different decisions, as a group the tendency should prevail, that is why we compare the probability of each agent to choose a certain option related to the group. If a person is not following that trend it may be a signal that the individual is ont paying attention.

4. Results and Discussions

Now, we will present the results of each model, specifically the accuracy and loss for each model. We will also include the correlation between the decision variable and the new variables we have created to analyze how good (or bad) they performed. For all models we will use only one layer that goes from all our variables to the prediction for B (2 values)

The results from the first model are:

Accuracy: 0.7629333333333334

Confusion Matrix:

1485 413
476 1376

Loss test data 0.57048315

For the second model, where we incorporate the risk variables we have:

Accuracy: 0.7205333333333334

Confusion Matrix:

1365 533
515 1337

Loss test data 0.5916656

For the third model, where we incorporate the attention variables we have created, the results are:

Accuracy: 0.764

Confusion Matrix:

1541 357
528 1324

Loss test data 0.5486565

Correlation analysis

As an addition, we will run an extra model by looking at our correlation with all the variables and the choice variable (see appendix). We will drop all the variables with a correlation (in absolute value) less than 0.01. This variables include (meanRT Moretime Forgone Condition block Trial Payoff Gender GameID Age RT Location B_more Order meanB meanB_subject Set Feedback consistent Ha) Something important to notice is that we have many missing values with the RT (response time) variable so that may have influenced in getting such a low correlation. Something important to notice is that our **Risk** variables have proven to have a very high correlation with the choice, specifically **ExB_best, ExHB_best** are the two variables with higher correlation among all the variables and **ExA, ExB** are among the top 10. After this analysis, we ran our extra model and got the following results.

This last model is the one with the highest accuracy across all our models, we also modified the layers to fit better. We can confirm that by cleaning our low correlation variables and fixing the layers of the model we can get a better accuracy rate. Something important to notice here is that models with one or two layers presented higher accuracy, however I decided to keep models with more layers to exercise all this content and manipulation of models.

Accuracy: 0.77333333333333

Confusion Matrix:

1519 379

471 1381

Loss test data 0.54134154

5. **Additional Discussion**

BLP models and neural network models offer different approaches to analysis of consumer choice behavior. A first difference is that BLP models are based on a theoretical model based on market, product characteristics, utility maximizing agents. With neural networks we don't need to assume any theoretical model, just take as many variables as possible related with decision making and let the algorithm predict decisions.

Some advantages about BLP models is that it allows us to see clearly the effect of changes in product and market attributes, and see these results in a simple economic way like elasticities or prices. For neural networks an advantage is that we don't need to build a complex economic model, that the algorithm may discover new non-expected effects of certain variables.

Therefore we can conclude that the BLP model is easier to interpret both in inputs and results in economic language, something that is not as simple in neural networks models. However, the latter is better when we talk about flexibility to include new variables, since the BLP model is based on a theoretical demand and market structure then it is hard to include new conditions or variables. Neural networks provide a good prediction ability but BLP provide a more structured way of seeing results and analyzing the economic components.

Appendix: Correlation with choice variable

Location	-0.0031215029	LotShapeA	-0.0291133055	Trial	-0.0013509159	ExB_best	0.3665390331
Gender	-0.0023311015	LotNumA	-0.0201304794	Button	-0.0138438024	ExHA	-0.0574926155
Age	0.0027809694	Hb	-0.0433459661	Payoff	-0.0022323170	ExHB	-0.0480289044
Set	-0.0081366165	pHb	-0.0266028520	Forgone	-0.0007218868	ExHB_best	0.1086998228
Condition	0.0010198978	Lb	0.1059391556	Apay	-0.0517599865	RT	-0.0027972912
GameID	-0.0026525129	LotShapeB	0.0862931886	Bpay	0.0392521819	meanRT	0.0001257702
Ha	-0.0087181406	LotNumB	0.0691412394	Feedback	0.0086162178	Moretime	-0.0003608920
pHa	-0.0155434483	Amb	-0.0117367967	block	-0.0013138985	meanB	0.0048162794
La	-0.0876612230	Corr	-0.0355075002	ExA	-0.0872563365	meanB_subject	0.0051953947
consistent	-0.0086682205	Order	-0.0045390032	ExB	0.0762102132	B_more	0.0041057624