ALESSIA
DI GIOVANNI

MAY 12TH, 2023

FEDERICO
VINCI

A VIDEOGAME BAYESIAN ANALYSIS

COOKIE CATS IS A "CONNECTED-THREE" MOBILE GAME WHERE PLAYERS ENCOUNTER GATES AS THEY ADVANCE THROUGH DIFFERENT LEVELS OF THE GAME.

WHAT IS A GATE?

BUT WHERE SHOULD THE FIRST GATE BE PLACED? 30TH OR 40TH LEVEL?

LET'S FIND IT OUT WITH ZIGGY!

I'LL PERFORM A BAYESIAN A/B TEST!

WHAT'S THE RELATIONSHIP BETWEEN THE NUMBER OF ROUNDS PLAYED AND PLAYER RETENTION AFTER 1 WEEK?

LET'S FIND IT OUT WITH BERRY!

I'LL GO FOR A BAYESIAN LOGISTIC REGRESSION!

DATASET

90.189 PLAYERS WHO INSTALLED THE GAME RIGHT AFTER THE A/B TEST WAS INITIALISED.

- VERSION - WHETHER THE PLAYER WAS PUT IN THE CONTROL GROUP (GATE_30 - FIRST GATE AT LEVEL 30) OR IN THE SHIFTED-GATE GROUP (GATE_40 - FIRST GATE AT LEVEL 40);

- SUM GAMEROUNDS - THE NUMBER OF ROUNDS PLAYED BY EACH USER WITHIN THE FIRST 14 DAYS AFTER DOWNLOADING THE GAME;

- RETENTION 1 - WHETHER THE PLAYER CAME BACK AND PLAYED 1 DAY AFTER DOWNLOADING THE GAME;

- RETENTION 7 - WHETHER THE PLAYER CAME BACK AND PLAYED 7 DAYS AFTER DOWNLOADING THE GAME.

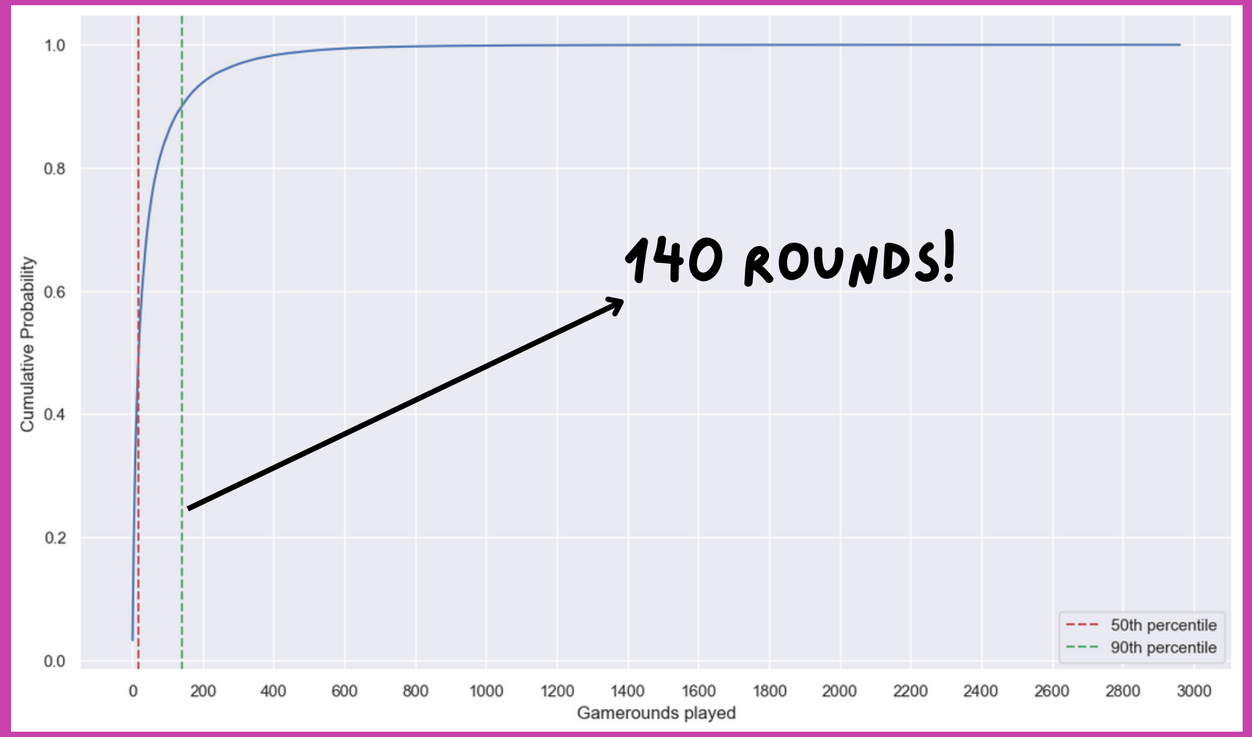
EXPLORATORY DATA ANALYSIS

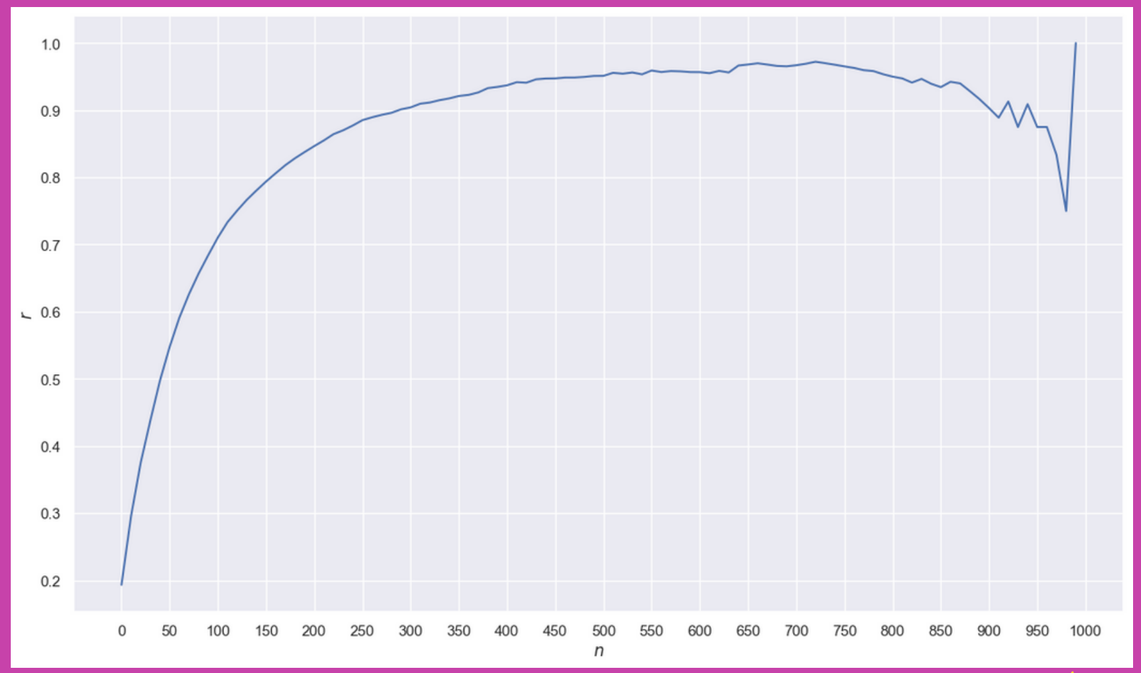
90.189 - 3.994 USERS WHO NEVER PLAYED = 86.195 OBSERVATIONS

DATASET FOR A/B TEST	GROUP A (GATE_30)	GROUP B (GATE_40)
NUMBER OF OBSERVATIONS	42.763	43.432
RETENTION AFTER ONE DAY	46,75%	46,21%
RETENTION AFTER SEVEN DAYS	19,84%	19,03%

	MEDIA	MIN	MAX
NUMBER OF ROUNDS PLAYED	54	0	2961

CUMULATIVE PROBABILITY OF USERS WHO PLAY N ROUNDS





RELATIONSHIP BETWEEN R AND N

- $R = \dfrac{\text{NUMBER OF USERS WHO PLAY N ROUNDS AND RETURN AFTER 1 WEEK}}{\text{NUMBER OF USERS WHO PLAY N ROUNDS}}$
- N LESS OR EQUAL THAN 1000

NEW DATAFRAME FOR THE BAYESIAN LOGISTIC REGRESSION

A/B TEST

FREQUENTIST APPROACH

- SETS UP NULL HYPOTHESIS & ALTERNATIVE HYPOTHESIS
- CONSIDERS THE P-VALUE
- CONSIDERS ONLY THE DATA OF THE EXPERIMENT

BAYESIAN APPROACH

- PAST & PRESENT INFORMATION ARE MESHED UP!
- POSTERIOR GIVES ROOM FOR ITS MOMENTS: EXPECTED VALUE OR VARIANCE AND MORE..

BAYES THEOREM

LIKELIHOOD OF THE DATA

PRIOR DISTRIBUTION

POSTERIOR DISTRIBUTION

$$\pi(\theta|\underline{x}) = \frac{f(\underline{x}|\theta)\pi(\theta)}{f(\underline{x})} \propto f(\underline{x}|\theta)\pi(\theta)$$

MARGINAL PROBABILITY

A/B TEST

	GROUP A	GROUP B
PRIOR	$\theta^A_{1,7} \sim \text{Beta}(\alpha, \beta)$	$\theta^B_{1,7} \sim \text{Beta}(\alpha, \beta)$
LIKELIHOOD	$retention^A_{1,7} \sim \text{Binomial}(N_A, p = \theta^A_{1,7})$	$retention^B_{1,7} \sim \text{Binomial}(N_B, p = \theta^B_{1,7})$

POSTERIOR

$$Beta\left(\sum_{i=1}^{n} x_i + \alpha, n - \sum_{i=1}^{n} x_i + \beta\right)$$

THAT'S A "CLOSED FORM" EQUATION!

MCMC SAMPLING METHODS

MARKOV CHAIN MONTE CARLO METHODS COMPRISES A CLASS OF ALGORITHMS FOR SAMPLING FROM A POSTERIOR DISTRIBUTION

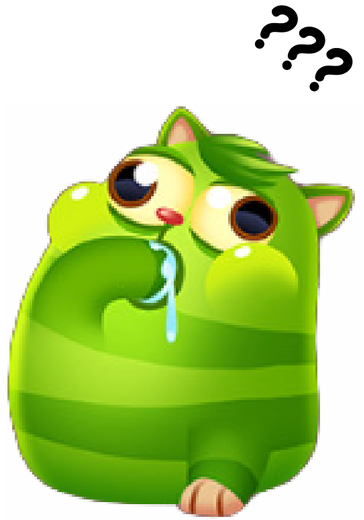
CONSTRUCTS A MARKOV CHAIN THAT HAS THE POSTERIOR DISTRIBUTION AS ITS STATIONARY DISTRIBUTION

ONE CAN OBTAIN A SAMPLE OF THE POSTERIOR DISTRIBUTION BY RECORDING STATES FROM THE CHAIN. THE MORE STEPS THAT ARE INCLUDED, THE MORE CLOSELY THE DISTRIBUTION OF THE SAMPLE MATCHES THE ACTUAL POSTERIOR DISTRIBUTION.

MARKOV CHAIN STATIONARY DISTRIBUTION

=

POSTERIOR DISTRIBUTION



GIBBS SAMPLER

OFTEN USED WHEN THE FULL CONDITIONAL DISTRIBUTION OF THE PARAMETERS IS KNOWN (SIMPLE FACTORIZATION STRUCTURE) AND CAN BE EASILY SAMPLED FROM.

1. INITIALISE THE PARAMETERS OF INTEREST BY A CHOSEN PRIOR DISTRIBUTION
2. INITIALIZE THE VALUES: START BY INITIALISING THE VALUES OF THE CONVERSION RATES FOR VERSIONS A AND B
3. UPDATE THE CONVERSION RATE OF VERSION A: SAMPLE A NEW VALUE FOR THE CONVERSION RATE OF VERSION A FROM ITS CONDITIONAL POSTERIOR DISTRIBUTION
4. UPDATE THE CONVERSION RATE OF VERSION B: SIMILARLY, SAMPLE A NEW VALUE FOR THE CONVERSION RATE OF VERSION B FROM ITS CONDITIONAL POSTERIOR DISTRIBUTION
5. REPEAT STEPS 3 AND 4 FOR A LARGE NUMBER OF ITERATIONS (E.G. 10,000).

MODELING PHASE

PYMC PYTHON'S LIBRARY TO CREATE A PYMC.MODEL CLASSES NAMED "RETENTION 1" ("RETENTION 7")

PRIOR_GATE_30, PRIOR_GATE_40: PYMC.BETA CLASSES. BETA(1,1)

ITS VARIABLES

LIFT: PYMC.DETERMINISTIC CLASS $L = \mathbb{E}[\pi(\theta^B_{1,7}|\underline{x}_B)] - \mathbb{E}[\pi(\theta^A_{1,7}|\underline{x}_A)]$

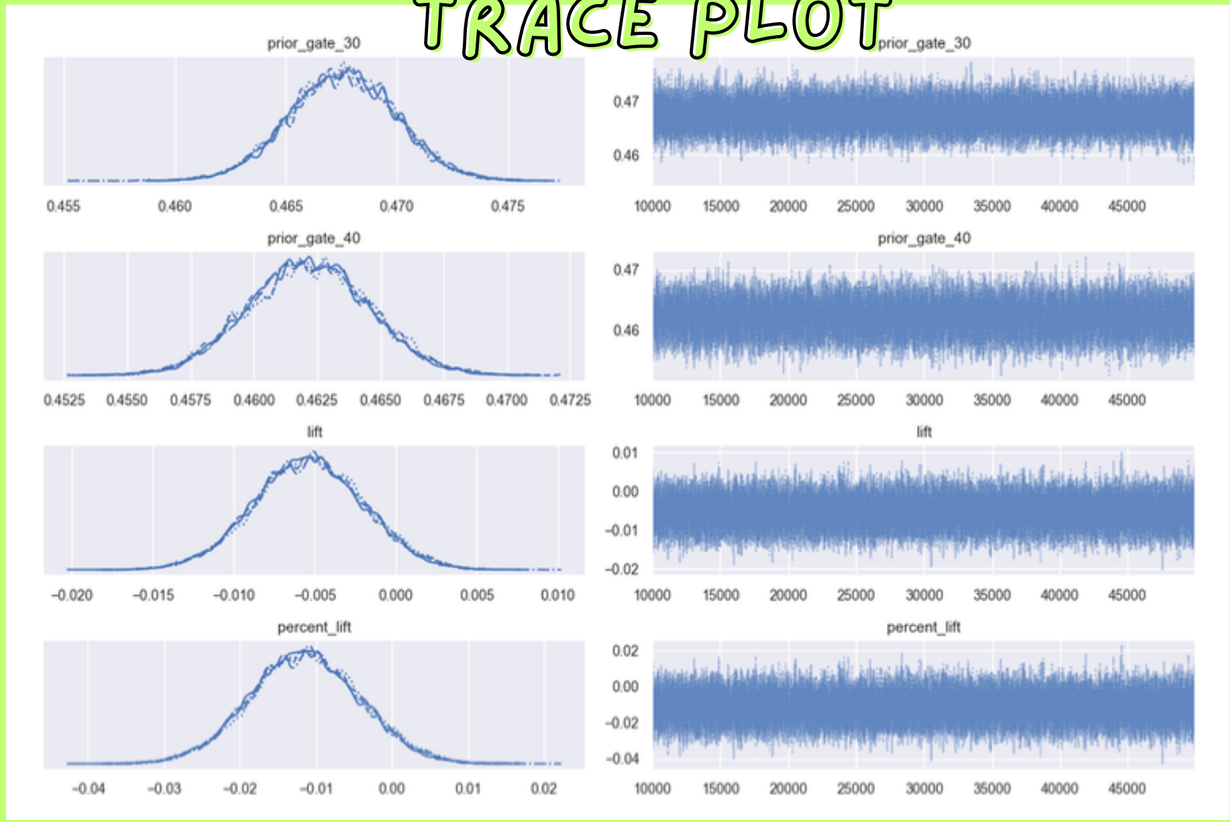
PERCENT_LIFT: PYMC.DETERMINISTIC CLASS $PL = \frac{\mathbb{E}[\pi(\theta^B_{1,7}|\underline{x}_B)]}{\mathbb{E}[\pi(\theta^A_{1,7}|\underline{x}_A)]} - 1$

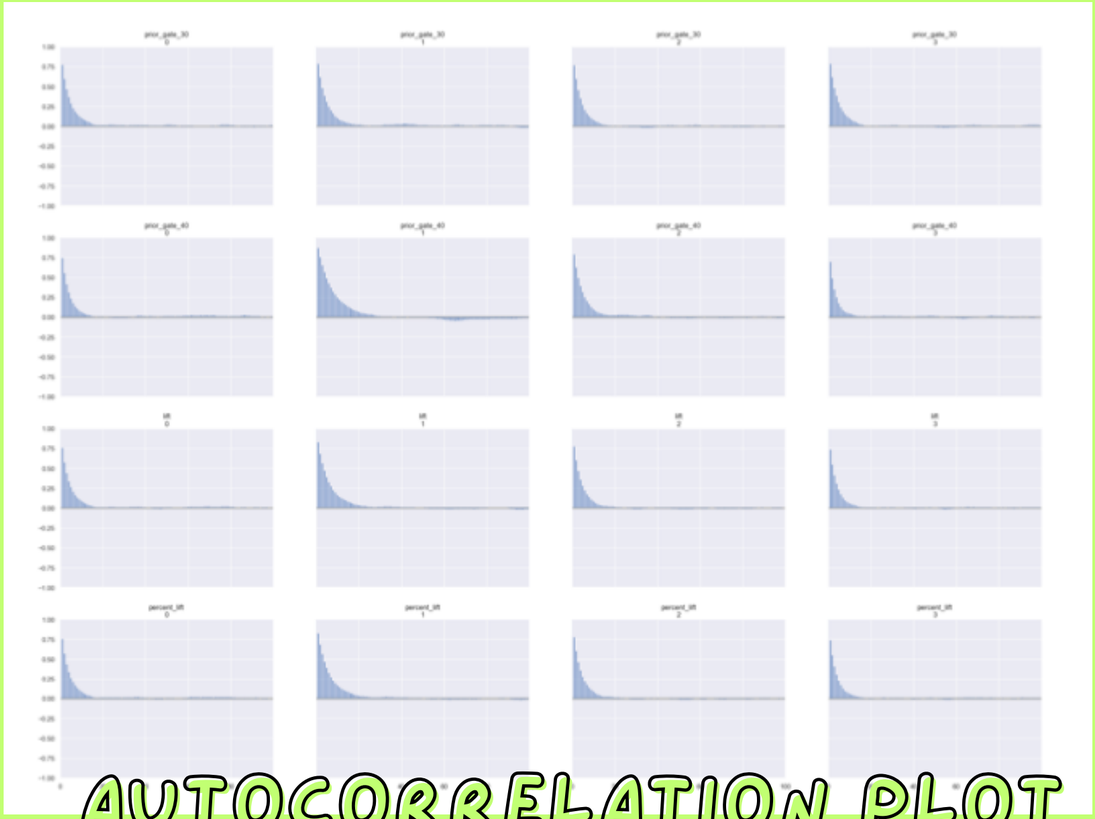
LIK_GATE_30, LIK_GATE_40: PYMC.BINOMIAL CLASSES. SHAPES THE LIKELIHOOD OF THE DATA. P DRAWS VALUES FROM THE PRIOR DISTRIBUTION

THE "RETENTION_1" ("RETENTION_7") PYMC.MODEL WILL SAMPLE THE POSTERIOR DISTRIBUTION THROUGH A GIBBS SAMPLER FOR 50000 OBSERVATIONS IN 4 CHAINS.

RETENTION 1

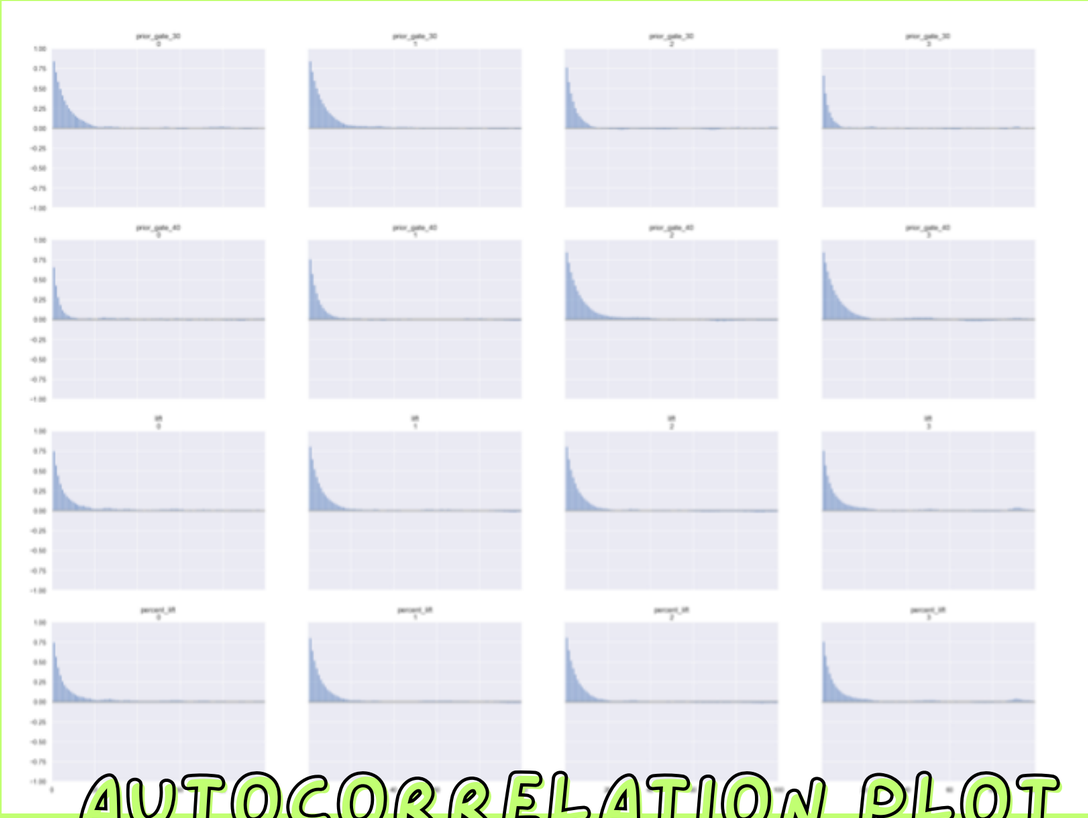
TRACE PLOT





AUTOCORRELATION PLOT

RETENTION 7



AUTOCORRELATION PLOT

TRACE PLOT

RESULTS, POSTERIOR 1

	MEAN	SD	HDI_3%	HDI_97%	R_HAT
POSTERIOR_GATE_30	0.468	0.002	0.463	0.472	1.002
POSTERIOR_GATE_40	0.462	0.002	0.458	0.467	1.002
LIFT	-0.005	0.003	-0.012	0.001	1.003
PERCENT LIFT	-0.011	0.007	-0.025	0.002	1.002

RESULTS, POSTERIOR 7

	MEAN	SD	HDI_3%	HDI_97%	R_HAT
POSTERIOR_GATE_30	0.198	0.002	0.195	0.202	1.002
POSTERIOR_GATE_40	0.190	0.002	0.187	0.194	1.003
LIFT	-0.008	0.003	-0.003	-0.003	1.001
PERCENT LIFT	-0.041	0.013	-0.016	-0.016	1.003

LOGISTIC REGRESSION

THE RETENTION AFTER ONE WEEK IS A BINARY CATEGORICAL RESPONSE VARIABLE CONVERTED TO 0-1.

$$y_i = \begin{cases} 1 & \text{IF THERE IS RETENTION} \\ 0 & \text{OTHERWISE} \end{cases}$$

$$y_i|\pi_i \sim Bern(\pi_i)$$

EXPECTED VALUE

$$E(yi|\pi_i) = \pi_i$$

GOAL: EXPECTED VALUE AS A LINEAR FUNCTION OF THE PREDICTOR

SUM OF GAME ROUNDS

BAYESIAN LOGISTIC MODEL

$$logit(\pi_i) = log\left(\frac{\pi_i}{(1-\pi_i))}\right) = \beta_0 + \beta_1 X_i.$$

LOGISTIC RESPONSE FUNCTION

$$\pi_i = \frac{e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}} = \frac{1}{1+e^{-(\beta_0+\beta_1 X_i)}}$$

BAYESIAN APPROACH

THE POSTERIOR DISTRIBUTION IS GIVEN BY

$$\pi(\underline{\beta}|\underline{y}) \propto f(\underline{y}|\underline{\beta})\pi(\underline{\beta})$$

LIKELIHOOD FUNCTION

$$f(\underline{y}|\underline{\beta}) = \prod_{i=1}^{n} f_{y_i}(y_i|\underline{\beta}) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{y_i}$$

PRIOR DISTRIBUTION

β_0 AND β_1 CAN ASSUME ANY REAL VALUE, SO NORMAL DISTRIBUTIONS ARE APPROPRIATE FOR THEM

$$\pi(\underline{\beta}) = \prod_{j=0}^{1} \pi(\beta_j) = \prod_{j=0}^{1} \frac{1}{\sqrt{2\pi\sigma_j^2}} exp\left(\frac{-1(\beta_j-\mu_j)^2}{2\sigma_j^2}\right)$$

POSTERIOR DISTRIBUTION NOT KNOWN: METROPOLIS-HASTINGS

METROPOLIS-HASTINGS

METROPOLIS-HASTINGS IS A MCMC ALGORITHM WHICH IS OFTEN USED WHEN IT IS NOT POSSIBLE (OR CONVENIENT) TO DIRECTLY SAMPLE FROM THE CONDITIONAL DISTRIBUTIONS.

GIBBS SAMPLER

↓

A VALUE FOR THE PARAMETER IS SAMPLED FROM ITS CONDITIONAL DISTRIBUTION, GIVEN THE CURRENT VALUES OF THE OTHER PARAMETERS

METROPOLIS-HASTINGS

↓

A VALUE FOR A PARAMETER IS PROPOSED USING A PROPOSAL DISTRIBUTION AND IT CAN BE ACCEPTED OR REJECTED BASED ON THE ACCEPTANCE PROBABILITY, DESIGNED TO ENSURE THAT THE TARGET DISTRIBUTION IS STATIONARY UNDER THE MARKOV CHAIN DEFINED BY THE SAMPLER.

MODELING A BAYESIAN LOGISTIC REGRESSION

DATAFRAME WITH NUMBER OF ROUNDS PLAYED LESS OR EQUAL THAN 1000

PYMC PYTHON'S LIBRARY TO CREATE A PYMC.MODEL CLASS NAMED "LOGISTIC"

$$Pr(\texttt{retention\_7} = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \texttt{sum\_gamerounds})}}$$

PRIOR_BETA_0, PRIOR_BETA_1: PYMC.NORMAL CLASSES. N(0,10)

ITS VARIABLES

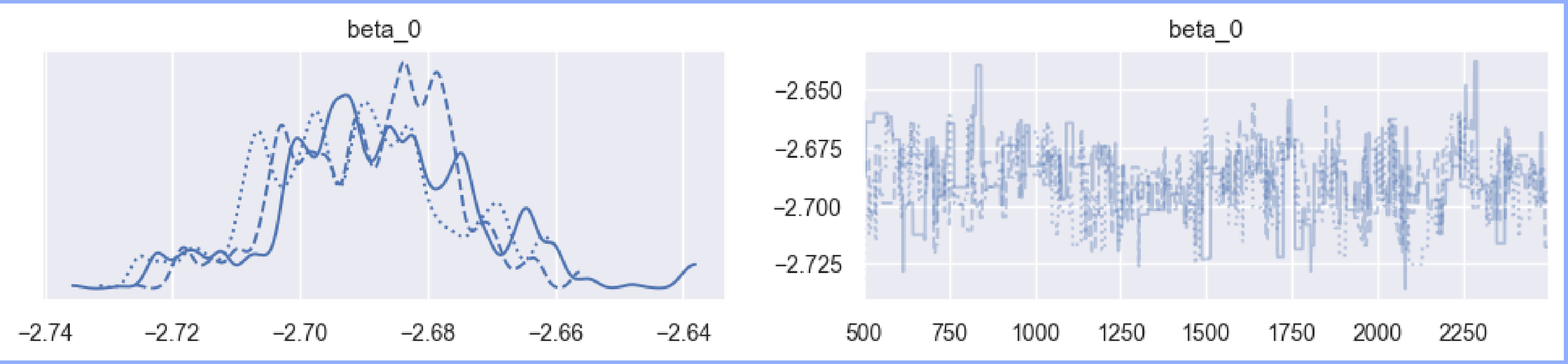
P: PYMC.DETERMINISTIC CLASS: REPRESENTS THE PROBABILITY P OF SUCCESS IN A LOGISTIC REGRESSION MODEL

```
p = pm.Deterministic('p', pm.math.sigmoid(prior_beta_0 + prior_beta_1 * log_df["sum_gamerounds"]))
```

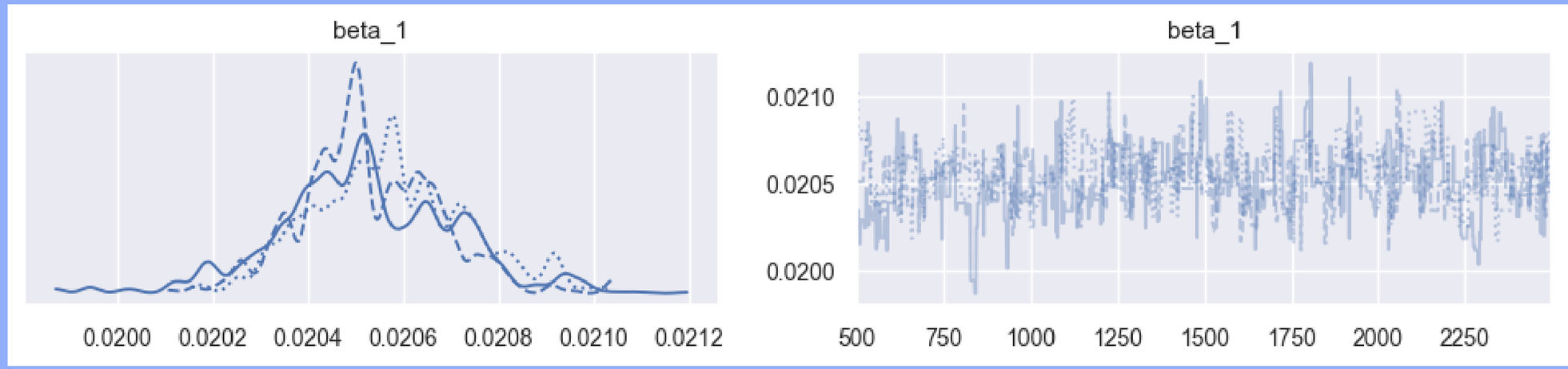
POSTERIOR: PYMC.BINOMIAL CLASS. MODELS THE OBSERVED DATA RETENTION_7 AS A BINARY OUTCOME WITH PROBABILITY P

THE "LOGISTIC" PYMC.MODEL WILL SAMPLE THE POSTERIOR DISTRIBUTION THROUGH A MCMC METROPOLIS-HASTINGS ALGORITHM FOR 2500 OBSERVATIONS IN 4 CHAINS.

DIAGNOSTICS



TRACEPLOTS AND POSTERIOR DISTRIBUTIONS FROM DIFFERENT CHAINS

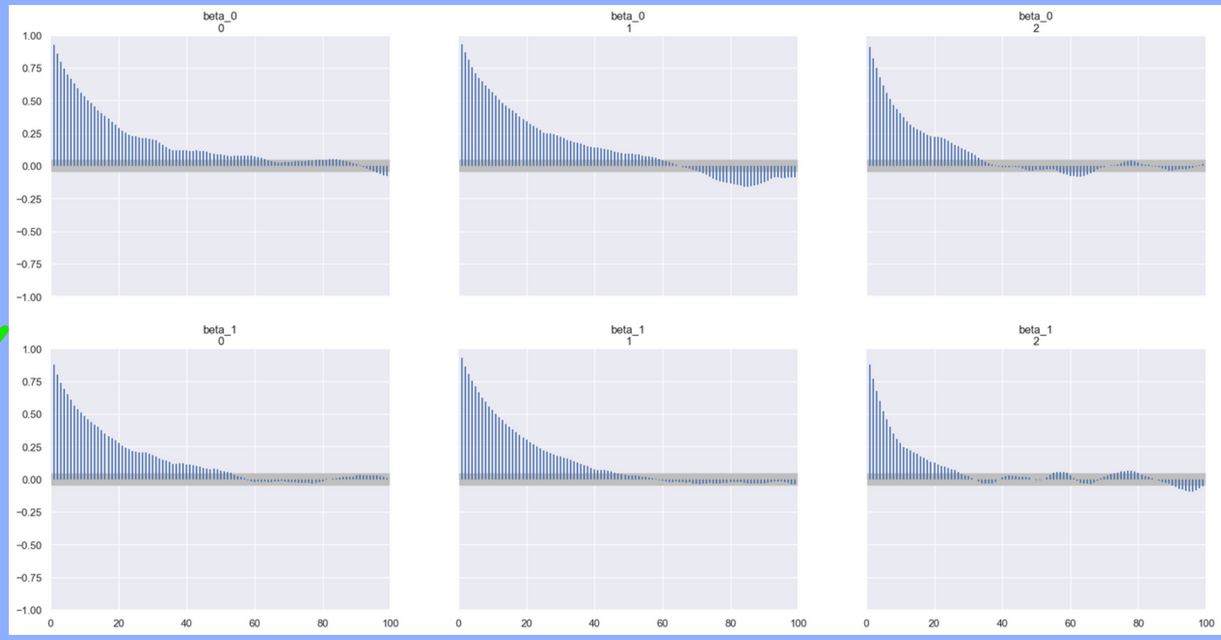


RESULTS

THERE IS SAMPLE AUTOCORRELATION ✗

BUT

THE POTENTIAL SCALE REDUCTION IS CLOSE TO 1! ✓



THE PROBABILITY THAT THE EVENT OCCURS, REGARDLESS SUM GAMEROUNDS BEING CONSIDERED, IS APPROXIMATELY 6, 36%.

	MEAN	SD	HDI_3%	HDI_97%	R_HAT
BETA_0	-2,690	0,014	-2,719	-2,664	1,02
BETA_1	0,021	0,000	0,020	0,021	1,02

WHENEVER THE PREDICTOR $X_j = X$, WHEN $X + 1$ WITHIN THE PREDICTOR'S SCALE, THEN LOGIT(π) GOES UP BY AN AMOUNT $\beta 1 = 2,1\%$

=> ODDS GOES UP BY E $\beta 1$

THANKS FOR YOUR ATTENTION!