Bayes : Empirical and Inference Methods in Data Science

Daniel Jiménez

Globant

27 -11 -2020

Agenda

- Introduction to the concept of Bayesian inference and Empirical Bayes.
- Bayesian Data Analysis.
- Machine Learning with Empirical Bayes.
- Conclusions.

Bibliography

- Robinson, D, “Introduction to Empirical Bayes”, (2017), Free
- McInerney, I, “An Empirical Bayes Approach to Optimizing Machine Learning Algorithms”, (2017), NY University
- Barber, D, “Bayesian Reasoning and Machine Learning”, (2019), Cambidge University

Bayesian inference introduction

This story begins in the way of **Thomas Carlyle**, and it is with the role of two men in the story.

Thomas Bayes (1761)	Alan Turing (1954)
	

And a formula

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)}$$

Bayesian inference introduction

- Bayes posits:
  - How to determine the probability of causes through observable effects.
  - Inverse probability[1]
- Turing applied:
  - Bayes through the "Bombe" Machine -> looking for configurations of Nazi machines rotors, through calculation of the possible values given the encryption pattern.

[1] It is a way of assess probabilistically all the conditions governing an assumption that has been observed on X event.

Bayesian inference introduction

What is Bayesian inference?

- Definition one: Is a method to calculate parameters or unknown quantities about known facts.
- Definition 2: Is a type of inference in witch observable evidences are used to update or infer if the hypothesis is true(certain) or false(ramdom).
- Definition 3: It´s a method that seeks to quantify the uncertainty and simulate on its truthfulness conditions, incorporating variations in the patterns.
- It´s the method that incorporates the probability of credibility on the hypotheses through a process of updating the data.

Bayesian inference introduction

As said before, let's define a problem under Bayes. what are the chances for Colombian soccer team to win a game if James plays?[2]

[2] This data collection was done through Web Scraping, the code will be available in my Github

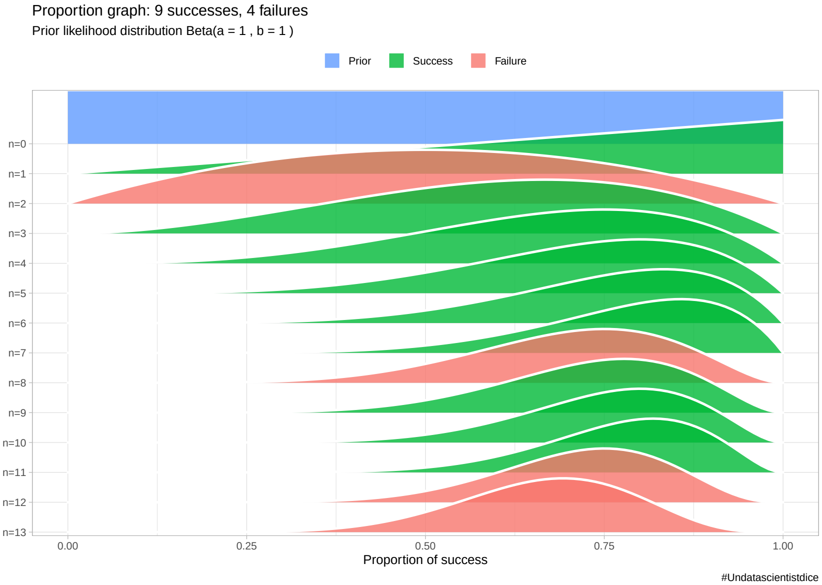
Bayesian inference introduction

Table 2: Juegos de la selección Colombia con James

Fecha	Rival	Marcador_Colombia	Marcador_Rival	Resultado	Resultado_dummy
22 de marzo de 2019	Japón	1	0	Gano	1
26 de marzo de 2019	Corea del Sur	1	2	Perdió	0
3 de junio de 2019	Panamá	3	0	Gano	1
9 de junio de 2019	Perú	3	0	Gano	1
15 de junio de 2019	Argentina	2	0	Gano	1
19 de junio de 2019	Catar	1	0	Gano	1
23 de junio de 2019	Paraguay	1	0	Gano	1
15 de octubre de 2019	Argelia	0	3	Perdió	0
15 de noviembre de 2019	Perú	1	0	Gano	1
19 de noviembre de 2019	Ecuador	1	0	Gano	1
9 de octubre de 2020	Venezuela	3	0	Gano	1
13 de noviembre de 2020	Uruguay	0	3	Perdió	0
17 de noviembre de 2020	Ecuador	1	6	Perdió	0

Bayesian inference introduction

Bayes applied to this problem, shows the following result.



Bayesian inference introduction

Using decision theory the best option is the following:

Bayesian inference introduction

In the previous example it could be seen that:

1) As evidence accumulates it allows to infer new results;
2) Hypothesis conflicts are discriminated

It was also evidenced the existence of: Transitivity, Comparison, Substitution and References in the actions.

Bayesian inference introduction

Deconstructing the theorem

Probabilidad Condicional

$$P(H_0 | E) = \frac{P(E | H_0)P(E)}{P(E)}$$

Probabilidad Previa

Probabilidad Posterior

Probabilidad de observar la nueva evidencia

Bayesian inference introduction

Deconstructing the theorem

$$\frac{P(E \mid H_0)}{P(E)} \longrightarrow$$

Mide el impacto que tiene la creencia
En la creencia de la hipótesis

Bayesian inference introduction

Deconstructing the theorem

$P(E \mid H_0) =$ Proporción de verosimilitud

$$\Lambda = \frac{P(H_0 \mid E)}{P(E \mid \text{no } H_0)}$$

Donde P(E) representa la suma de los productos de las probabilidades Excluyentes, independientes y condicionadas

$$P(H_0 \mid E_1, E_2) = \frac{\Lambda_1 \Lambda_1 P(H_0)}{[\Lambda_1 P(H_0) + P(\text{ no } H_0] + [\Lambda_2 P(H_0) + P(\text{ no } H_0)]} \longrightarrow$$

Bayes adquiere esta forma

Bayesian inference introduction

Definitions

1) **Priori (previous)** : $P(H_0)$ given the conditional weights,estimate the probability value before viewing the data.
2) H_0 : Null hypothesis is the statement that a piece of information comes from a statement.
3) **Condicional** : $P(E|H_0)$ It is a conditional probability, and describes the probability of happening **E** if the hypothesis H_0 is true.

- This is the verisimilitude function[3]
- Makes the observed value more credible
- It´s the inverse of the conditional, trying to define the unobserved variable.

[3] In a inferential manner described, if a value corresponds to a set of data.

Bayesian inference introduction

The empirical bayes, is one of the alternative Bayesian methods, where the previous (priori) is determined by the present data, therefore its approximations are more exact especially when we face a Big Data problem.

In the empirical bayes MCMC is implemented to give a better performance to the explored data, therefore, it´s more efficient computationally.

Bayesian inference introduction

Therefore **Empirical Bayes** is a hierarchical method, which allows machine learning models to better estimate the hyper-parameters since the previous data is born from the observable data and thus reduce the uncertainty, while in the field of analytics, it is efficient to justify or adjust the data of an event, thus correcting the false discoveries in the data.

Bayesian inference introduction

Given all that, the empirical bayes takes the following form

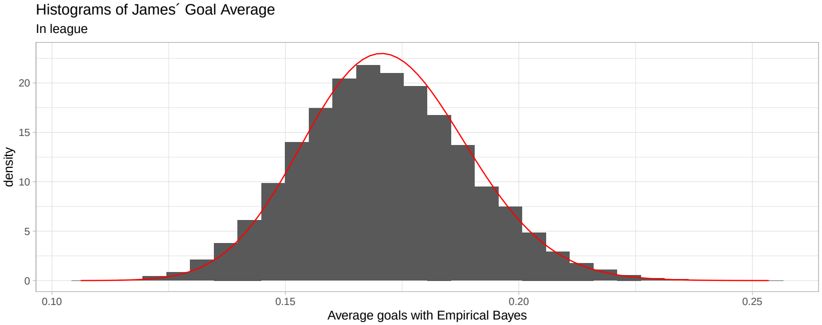
$$EB = \frac{P(E|H_0)P(E)}{P(E)} \int P(E|\eta)P(\eta)d\eta$$

Where η is a hyperparameter of the extracted sample.

Applying the theory of empirical bayes in the case of James, a simulation process was developed to obtain the answer to In how many games or chances did he score two goals?

The cornerstone of empirical Bayes

Applying the theory of empirical bayes in the case of James, a simulation process was developed to obtain the answer to In how many games or chances did he score two goals?



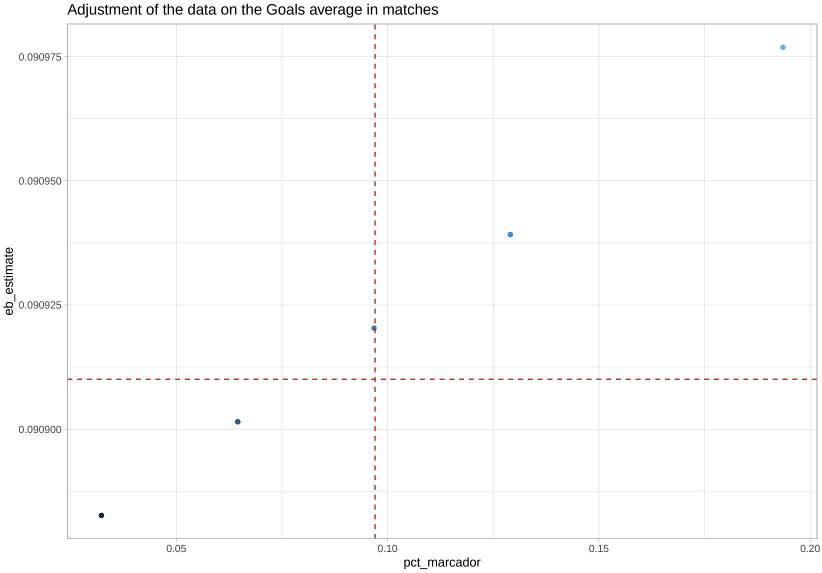
The average goal value according to FIFA (Frequentists) is 29%, while thanks to the correction of the prior with the beta, the result was 17%, here we discover the distribution of the unobserved value.

The cornerstone of empirical Bayes

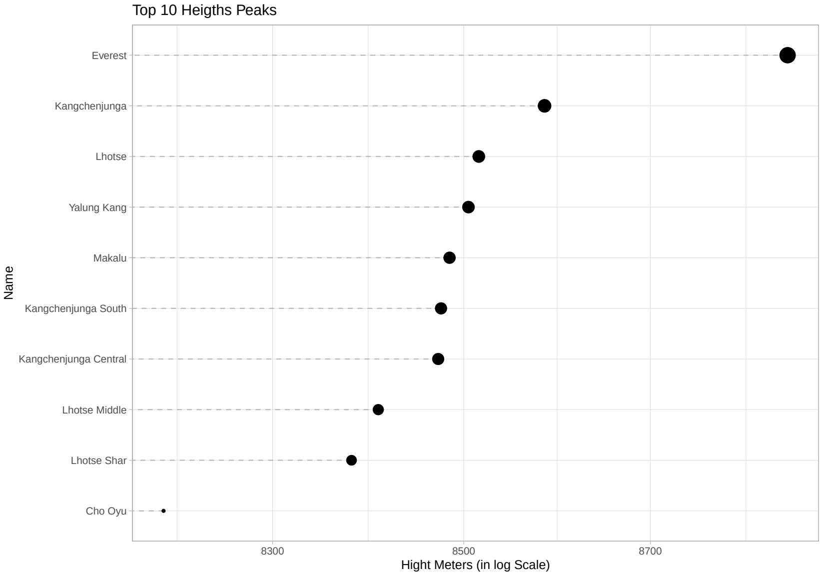
The cornerstone of empirical Bayes is the estimator of the prior beta from the data, which adapts α and β as hyper-parameters using maximum verisimilitude: that these parameters are generated by the distribution seen in the data.

Now we proceed to calculate the posterior as follows:

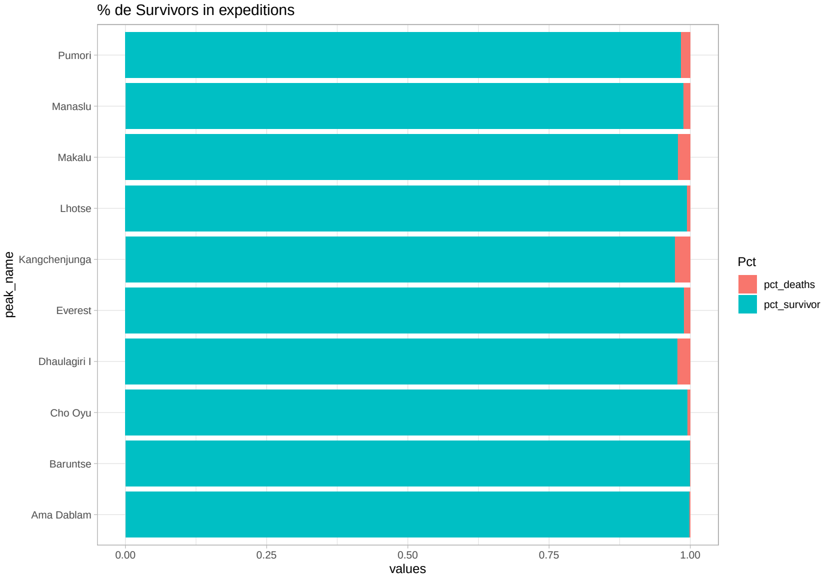
The cornerstone of empirical Bayes



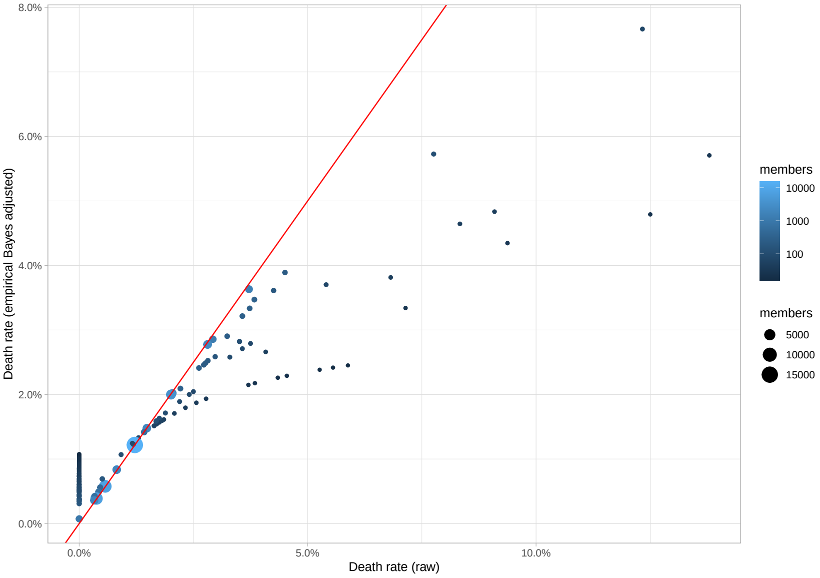
Analysis of expedition data (climbing) in the Himalayas



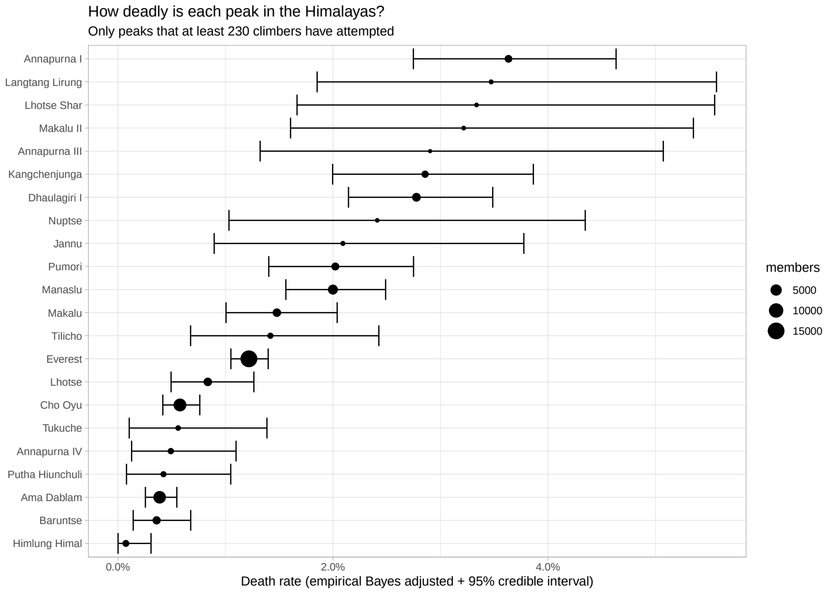
Bayesian data analysis



Bayesian data analysis



Bayesian data analysis



Machine Learning with Bayes

we will work the Churn database of Kaggle

```
##   customerID           gender          SeniorCitizen      Partner
##  Length:7043        Length:7043        Min.   :0.0000   Length:7043
##  Class :character   Class :character   1st Qu.:0.0000   Class :character
##  Mode  :character   Mode  :character   Median :0.0000   Mode  :character
##                                        Mean   :0.1621
##                                        3rd Qu.:0.0000
##                                        Max.   :1.0000
##
##   Dependents            tenure      PhoneService       MultipleLines
##  Length:7043        Min.   : 0.00   Length:7043        Length:7043
##  Class :character   1st Qu.: 9.00   Class :character   Class :character
##  Mode  :character   Median :29.00   Mode  :character   Mode  :character
##                     Mean   :32.37
##                     3rd Qu.:55.00
##                     Max.   :72.00
##
##  InternetService    OnlineSecurity     OnlineBackup       DeviceProtection
##  Length:7043        Length:7043        Length:7043        Length:7043
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  TechSupport        StreamingTV        StreamingMovies      Contract
##  Length:7043        Length:7043        Length:7043        Length:7043
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  PaperlessBilling   PaymentMethod      MonthlyCharges    TotalCharges
```

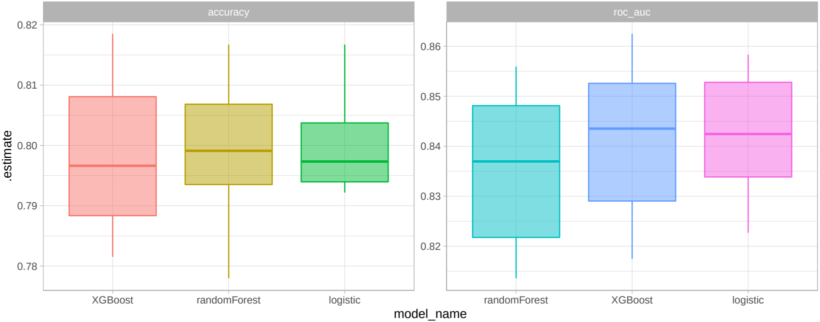
Machine Learning with Bayes

Some feature engineering is done

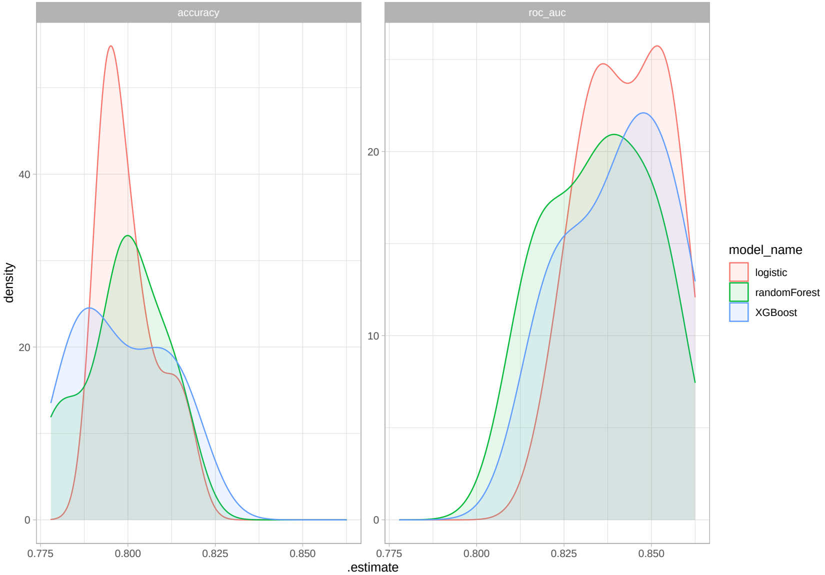
```
## # A tibble: 6 x 21
##   customerID gender SeniorCitizen Partner Dependents tenure PhoneService
##   <fct>      <fct>  <fct>         <fct>   <fct>       <dbl> <fct>
## 1 7590-VHVEG Female 0             Yes     No              1 No
## 2 5575-GNVDE Male   0             No      No             34 Yes
## 3 3668-QPYBK Male   0             No      No              2 Yes
## 4 7795-CFOCW Male   0             No      No             45 No
## 5 9237-HQITU Female 0             No      No              2 Yes
## 6 9305-CDSKC Female 0             No      No              8 Yes
## # ... with 14 more variables: MultipleLines <fct>, InternetService <fct>,
## #   OnlineSecurity <fct>, OnlineBackup <fct>, DeviceProtection <fct>,
## #   TechSupport <fct>, StreamingTV <fct>, StreamingMovies <fct>,
## #   Contract <fct>, PaperlessBilling <fct>, PaymentMethod <fct>,
## #   MonthlyCharges <dbl>, TotalCharges <dbl>, Churn <fct>
```

Machine Learning with Bayes

Three classification models are developed: Logistic Regression, Random Forest, Xgboost and the objective is to correctly choose a model.



Machine Learning with Bayes



Note that the logistics model seems to have better performance

Machine Learning with Bayes

Table 3: Model metrics

model_name	.metric	mean
logistic	accuracy	0.8002528
logistic	roc_auc	0.8424128
randomForest	accuracy	0.7988325
randomForest	roc_auc	0.8351292
XGBoost	accuracy	0.7984751
XGBoost	roc_auc	0.8408523

Machine Learning with Bayes

And here generated the conditions for posterior model.

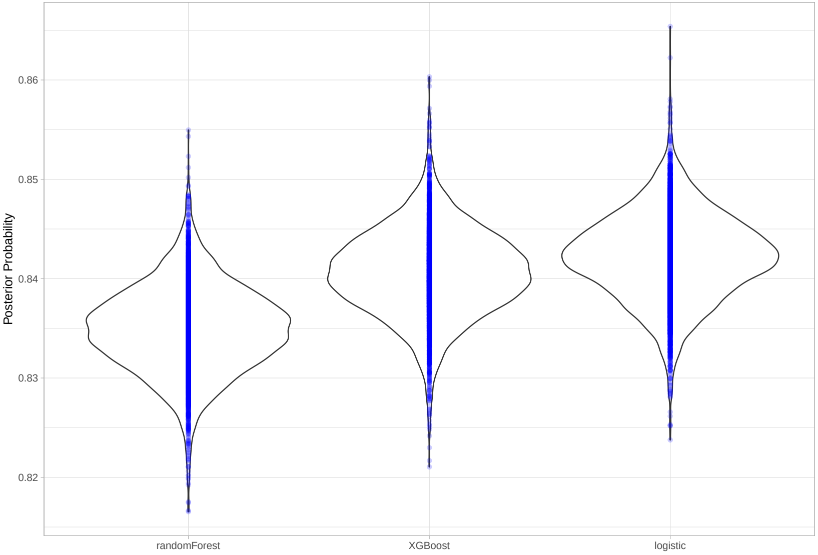
```
## # A tibble: 10 x 4
##    id     logistic randomForest XGBoost
##    <chr>     <dbl>        <dbl>   <dbl>
##  1 Fold01    0.855        0.851   0.856
##  2 Fold02    0.823        0.819   0.822
##  3 Fold03    0.841        0.837   0.844
##  4 Fold04    0.834        0.814   0.817
##  5 Fold05    0.844        0.837   0.847
##  6 Fold06    0.831        0.829   0.835
##  7 Fold07    0.852        0.851   0.854
##  8 Fold08    0.858        0.856   0.862
##  9 Fold09    0.834        0.818   0.827
## 10 Fold10    0.853        0.840   0.843
```

Machine Learning with Bayes

Let's explore a bit the variance that the models produce

```
##
## SAMPLING FOR MODEL 'continuous' NOW (CHAIN 1).
## Chain 1:
## Chain 1: Gradient evaluation took 0.000113 seconds
## Chain 1: 1000 transitions using 10 leapfrog steps per tr
## Chain 1: Adjust your expectations accordingly!
## Chain 1:
## Chain 1:
## Chain 1: Iteration:    1 / 2000 [  0%]  (Warmup)
## Chain 1: Iteration:  200 / 2000 [ 10%]  (Warmup)
## Chain 1: Iteration:  400 / 2000 [ 20%]  (Warmup)
## Chain 1: Iteration:  600 / 2000 [ 30%]  (Warmup)
## Chain 1: Iteration:  800 / 2000 [ 40%]  (Warmup)
## Chain 1: Iteration: 1000 / 2000 [ 50%]  (Warmup)
## Chain 1: Iteration: 1001 / 2000 [ 50%]  (Sampling)
## Chain 1: Iteration: 1200 / 2000 [ 60%]  (Sampling)
## Chain 1: Iteration: 1400 / 2000 [ 70%]  (Sampling)
```

Machine Learning with Bayes

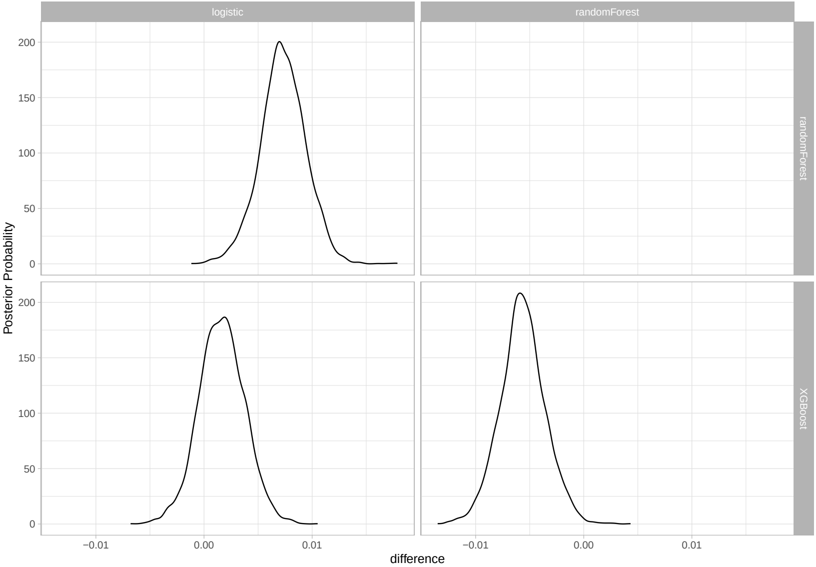


Machine Learning with Bayes

A model contrast (comparative) is generated to evaluate the strength of randomness between models.

```
## # Posterior samples of performance differences
## # A tibble: 6 x 4
##   difference model_1  model_2      contrast
##        <dbl> <chr>    <chr>        <chr>
## 1    0.00744 logistic randomForest logistic vs. randomForest
## 2    0.00608 logistic randomForest logistic vs. randomForest
## 3    0.00552 logistic randomForest logistic vs. randomForest
## 4    0.00701 logistic randomForest logistic vs. randomForest
## 5    0.00753 logistic randomForest logistic vs. randomForest
## 6    0.00626 logistic randomForest logistic vs. randomForest
```

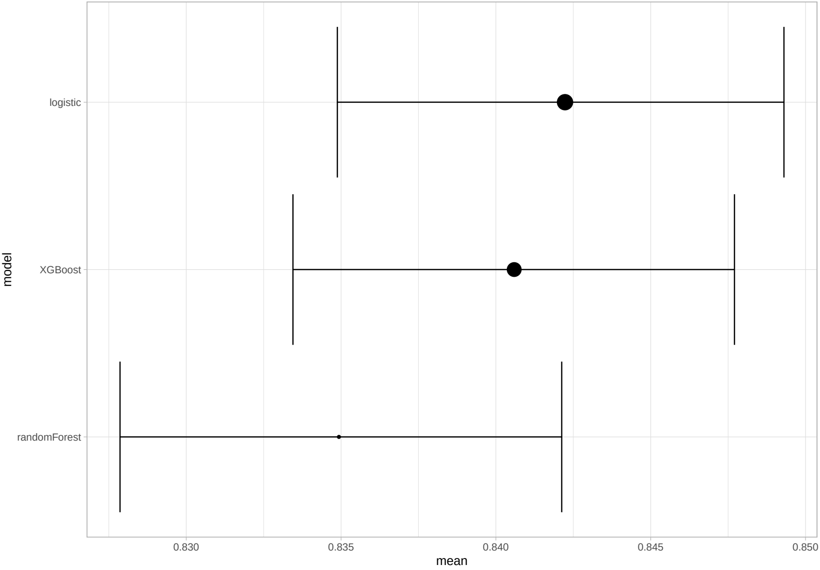
Machine Learning with Bayes



The model to be selected is the one whose differences are greater than zero.

Machine Learning with Bayes

The credibility intervals are created and we conclude on the model



Machine Learning with Bayes

```
## # A tibble: 3 x 5
##   model         mean lower upper   diff
##   <chr>        <dbl> <dbl> <dbl>  <dbl>
## 1 logistic     0.842 0.835 0.849 0.0144
## 2 randomForest 0.835 0.828 0.842 0.0143
## 3 XGBoost      0.841 0.833 0.848 0.0143
```

Machine Learning with Bayes

According to what was previously analyzed, the selection of the logistic model for this problem is confirmed.

Conclusions

1) Classical Bayes allows us to determine and measure uncertainty, and Empirical to adjust our assumptions.
2) Empirical Bayes is computationally efficient for adjusting the beliefs of frequentist presented data.
3) Bayes in Machine Learning allows us to adjust the verisimilitude of the data, and thus adjust the hyperparameters of the models in order to select one that has better performance

Repo

Github Repo

https://github.com/carlosjimenez88M/Bayes_presentation