ML Models Comparison

github.com/david26694/model-comparison-training

David Masip

Pagantis

2019/09/16

Contents

- Problem description and approach
- Frequentist and bayesian methods
- tidyposterior examples

Problem formulation

Given two models trained on the same dataset, we want to assess which one is better according to a given metric.

Requirements:

- We should not use test set to assess.
- We should embrace uncertainty.

Solution:

- Metric in cross-validation folds (trust your local cv).
- **Always use the same folds.**

Example

Throughout the session we'll be using OK Cupid data.

```
head(recipes::okc)
```

```
## # A tibble: 6 x 6
##     age diet              height location            date       Class
##   <int> <chr>              <int> <chr>               <date>     <fct>
## 1    22 strictly anything     75 south san francisco 2012-06-28 other
## 2    35 mostly other          70 oakland             2012-06-29 other
## 3    38 anything              68 san francisco       2012-06-27 other
## 4    23 vegetarian            71 berkeley            2012-06-28 other
## 5    29 <NA>                  66 san francisco       2012-06-27 other
## 6    29 mostly anything       67 san francisco       2012-06-29 stem
```

Example (continuation)

Basic feature engineering (one hot encoding, date parsing). Lasso, Xgboost and random forest trained.

```r
load(file = "../data/aucs.RData")

aucs
```

```
## # A tibble: 10 x 4
##    id     roc_auc_xgb roc_auc_lasso roc_auc_rf
##    <chr>        <dbl>         <dbl>      <dbl>
##  1 Fold01       0.659         0.648      0.652
##  2 Fold02       0.666         0.632      0.660
##  3 Fold03       0.662         0.629      0.652
##  4 Fold04       0.648         0.636      0.647
##  5 Fold05       0.657         0.630      0.653
##  6 Fold06       0.661         0.628      0.654
##  7 Fold07       0.654         0.646      0.648
##  8 Fold08       0.642         0.625      0.634
##  9 Fold09       0.643         0.617      0.629
## 10 Fold10       0.660         0.631      0.653
```

Frequentist methods 1: paired t-test

- Doesn't account for correlation among folds (test considers independent samples).

```
t.test(aucs$roc_auc_xgb, aucs$roc_auc_rf, paired = T)
```

```
##
##     Paired t-test
##
## data:  aucs$roc_auc_xgb and aucs$roc_auc_rf
## t = 6.3989, df = 9, p-value = 0.0001254
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.004615056 0.009662469
## sample estimates:
## mean of the differences
##             0.007138762
```

Frequentist methods 1: paired t-test (continuation)

```
t.test(aucs$roc_auc_xgb, aucs$roc_auc_lasso, paired = T)
```

```
##
##     Paired t-test
##
## data:  aucs$roc_auc_xgb and aucs$roc_auc_lasso
## t = 7.2409, df = 9, p-value = 4.864e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.01579726 0.03015267
## sample estimates:
## mean of the differences
##              0.02297496
```

Frequentist methods 2: Correlated t-test

Paired t-test that accounts for the correlation between samples (Nadeau and Bengio, 2003).

- There is no unbiased estimator for the correlation.
- Correlation parameter is estimated through an heuristic.

Frequentist methods 3: ANOVA

$$auc = b_0 + b_1 m_1 + b_2 m_2$$

- Can compare multiple models.
- Uses models to compare models.
- Doesn't account for correlation.
- Doesn't answer *Which models are different?*

Frequentist methods 3: ANOVA (continuation)

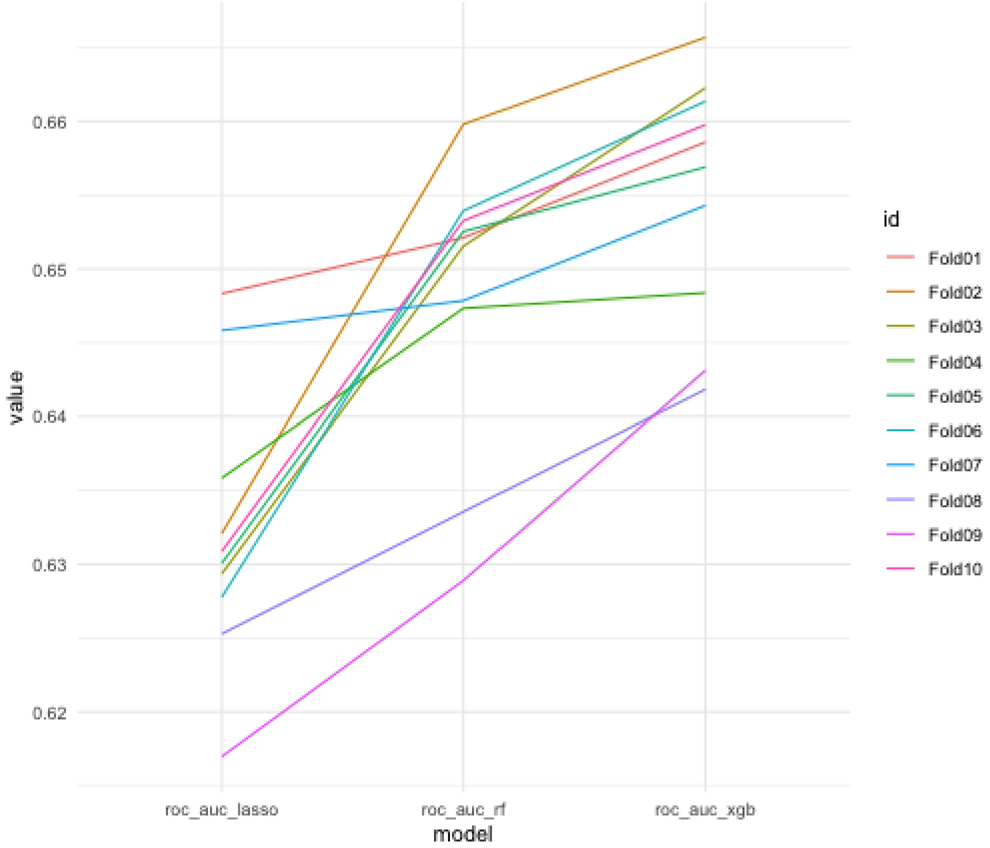
```r
# Convert dataframe from wide to long
anova_df <- aucs %>% gather(model, value, -id)


# Anova finds differences
anova(lm(value ~ model, anova_df))
```

```
## Analysis of Variance Table
##
## Response: value
##           Df    Sum Sq    Mean Sq F value    Pr(>F)
## model      2 0.0027653 0.00138266  16.942 1.709e-05 ***
## Residuals 27 0.0022035 0.00008161
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Correlated structure!

- Folds 1 and 2 are easy.
- Folds 8 and 9 are hard.



Solution: random intercepts

Linear model

$$y_i = b_0 + b_1 x_i + e_i$$

Random intercepts model

$$y_{ij} = b_0 + b_1 x_{ij} + u_j + e_i$$

Not same as dummy variables, u is a random variable.

Bayesian methods

Methods:

- Correlated t-test (Benavoli et al., 2017): *the probability of the bayesian t-test and p-value of the frequentist t-test are numerically equivalent*.

- ANOVA with random intercepts: tidyposterior and Max Kuhn talk.

Disclaimer: I'm not a bayesian activist.

Bayesian/frequentist differences

- Frequentist methods: We assume both methods are equal have the same AUC and compute

$$P(x|AUC_1 = AUC_2)$$

- Bayesian methods: the parameter has a distribution of possible values. We have prior knowledge, and update the distribution according to the data. In our case,

$$P(AUC_1 - AUC_2|x)$$

  This is what we want to estimate!

Frequentist pitfalls (Benavoli et al, 2017)

p-value depends on sample size

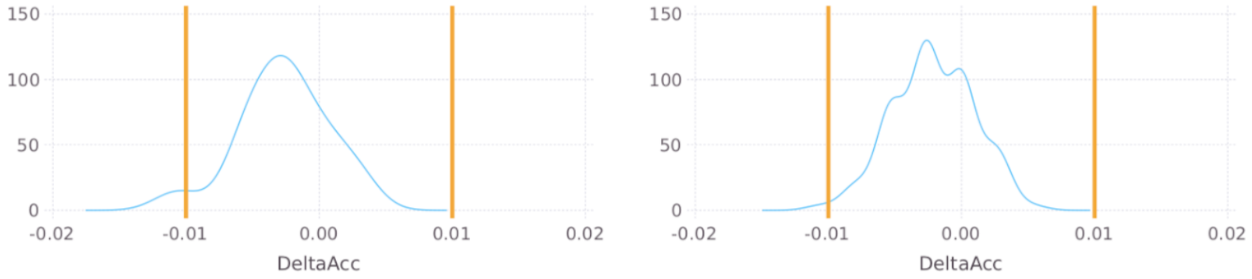


Figure 1: Density plot for the differences of accuracy between *nbc* and *aode* for the dataset *hepatitis* considering only 15 of the 100 data (left) or all the data (right). Left: the null hypothesis cannot be rejected ($p = 0.077 > 0.05$) using half the data. Right: the null hypothesis is rejected when all the data are considered ($p = 0.048 < 0.05$), despite the very small effect size.

Frequentist pitfalls (Benavoli et al, 2017)

p-value ignores magnitude

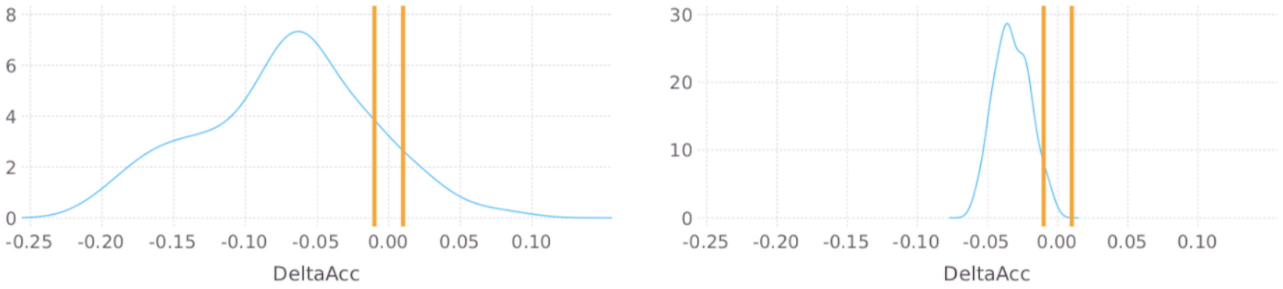
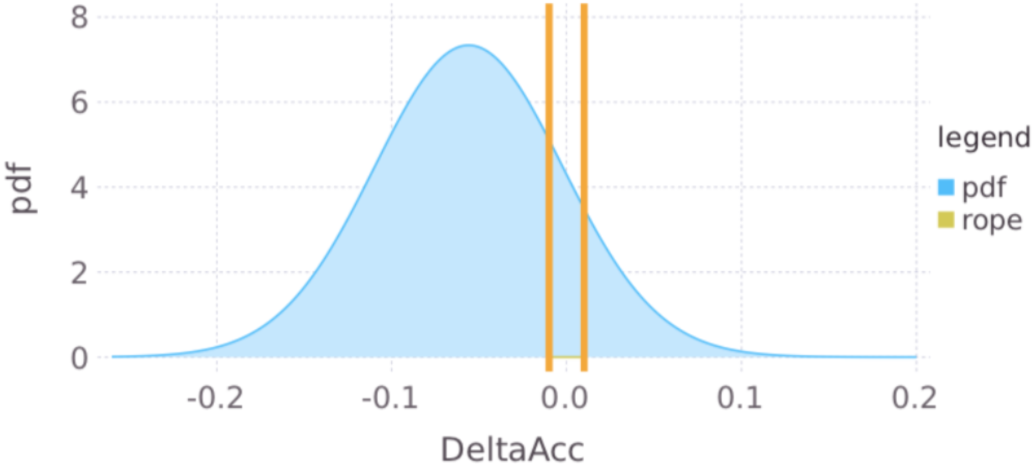


Figure 2: Density plot for the differences of accuracy (DeltaAcc) between *nbc* and *aode* for the datasets *ecoli* (left) and *iris* (right). The null hypothesis is rejected ($p < 0.05$) with similar p-values, even though the two cases have very different uncertainty. For *ecoli*, the uncertainty is very large and includes zero.

Region of practical equivalence

ROPE: range of values of the metrics' difference where we think the models as equivalent (definition left to the modeller). Three probabilities:

- $P(AUC_1 \gg AUC_2)$
- $P(AUC_1 \approx AUC_2)$
- $P(AUC_1 \ll AUC_2)$



Bayesian comparison

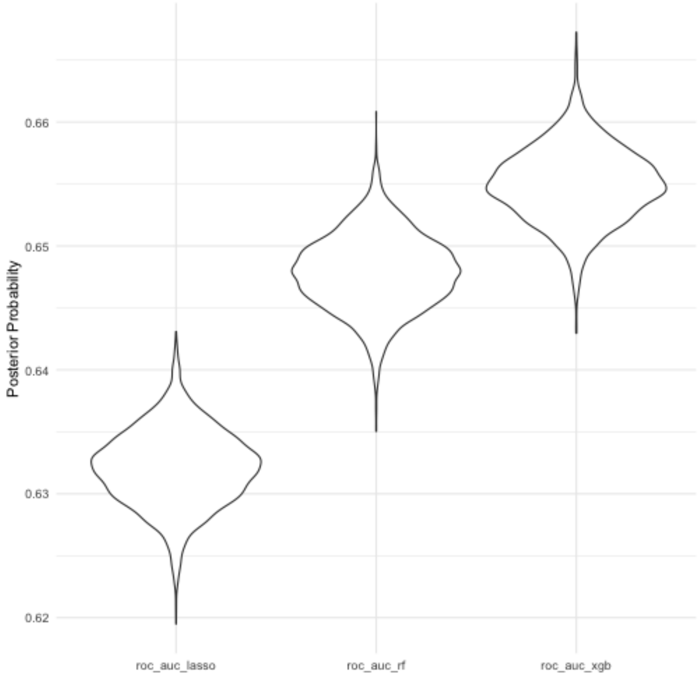
```r
library(tidyposterior)

# Model and generate posteriors
bayesian_lm <- perf_mod(aucs, transform = logit_trans)

# Compare posteriors
bayesian_comparison <- contrast_models(bayesian_lm)
```

Plot differences

```
ggplot(tidy(bayesian_lm))
```



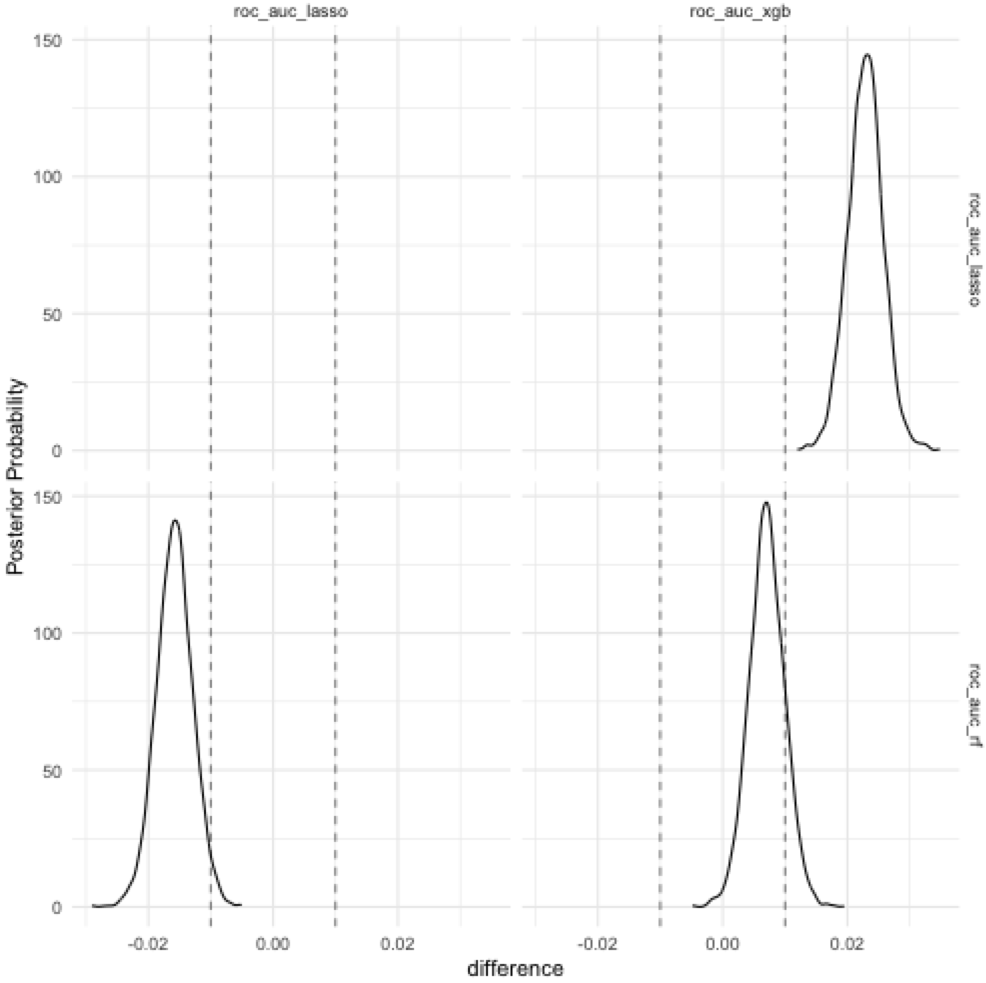
```r
# ROPE = 1%
summary(bayesian_comparison, size = 0.01) %>%
  select(contrast, pract_neg, pract_equiv, pract_pos)
```

```
## # A tibble: 3 x 4
##   contrast                     pract_neg pract_equiv pract_pos
##   <chr>                            <dbl>       <dbl>     <dbl>
## 1 roc_auc_lasso vs roc_auc_rf      0.976      0.0245     0
## 2 roc_auc_xgb vs roc_auc_lasso     0          0          1
## 3 roc_auc_xgb vs roc_auc_rf        0          0.849      0.151
```

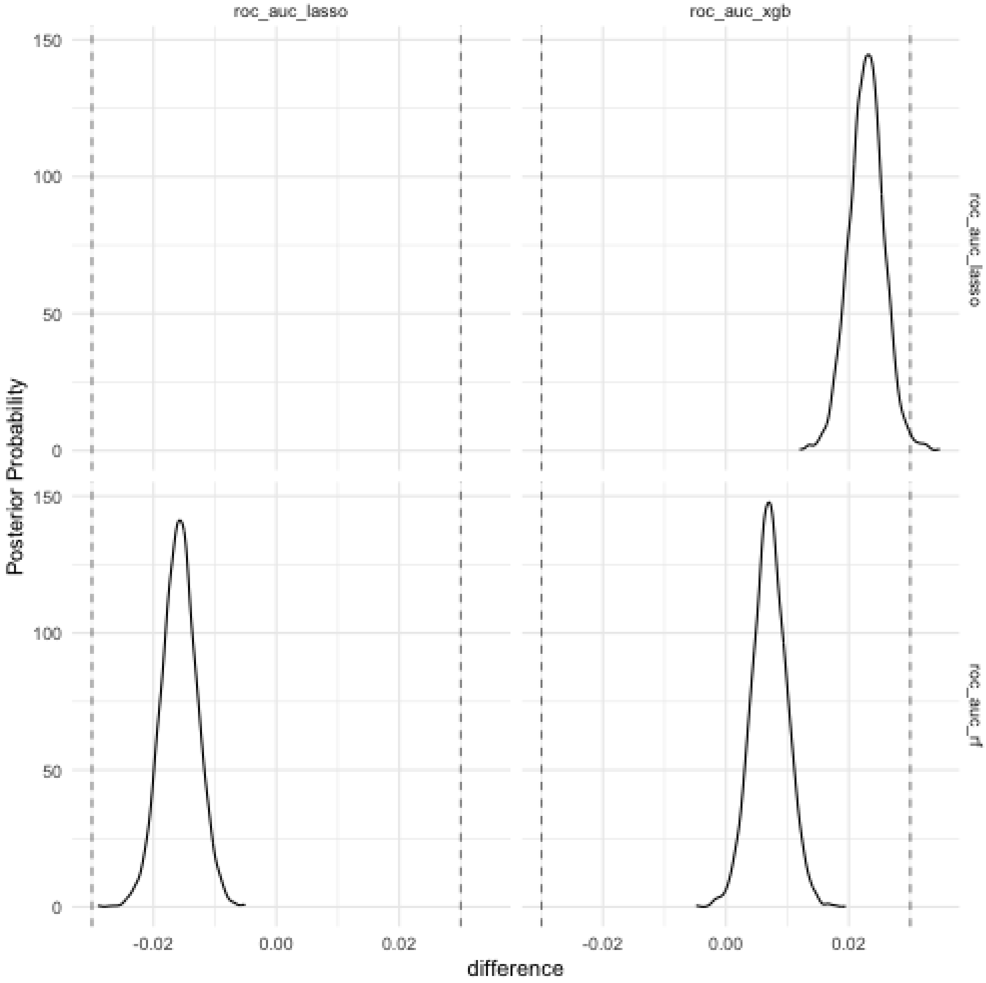
```r
# ROPE = 3%
summary(bayesian_comparison, size = 0.03) %>%
  select(contrast, pract_neg, pract_equiv, pract_pos)
```

```
## # A tibble: 3 x 4
##   contrast                     pract_neg pract_equiv pract_pos
##   <chr>                            <dbl>       <dbl>     <dbl>
## 1 roc_auc_lasso vs roc_auc_rf          0       1        0
## 2 roc_auc_xgb vs roc_auc_lasso         0       0.990    0.0095
## 3 roc_auc_xgb vs roc_auc_rf            0       1        0
```

```r
# ROPE = 1%
ggplot(bayesian_comparison, size = 0.01)
```



```r
# ROPE = 3%
ggplot(bayesian_comparison, size = 0.03)
```



Bayesian pitfalls

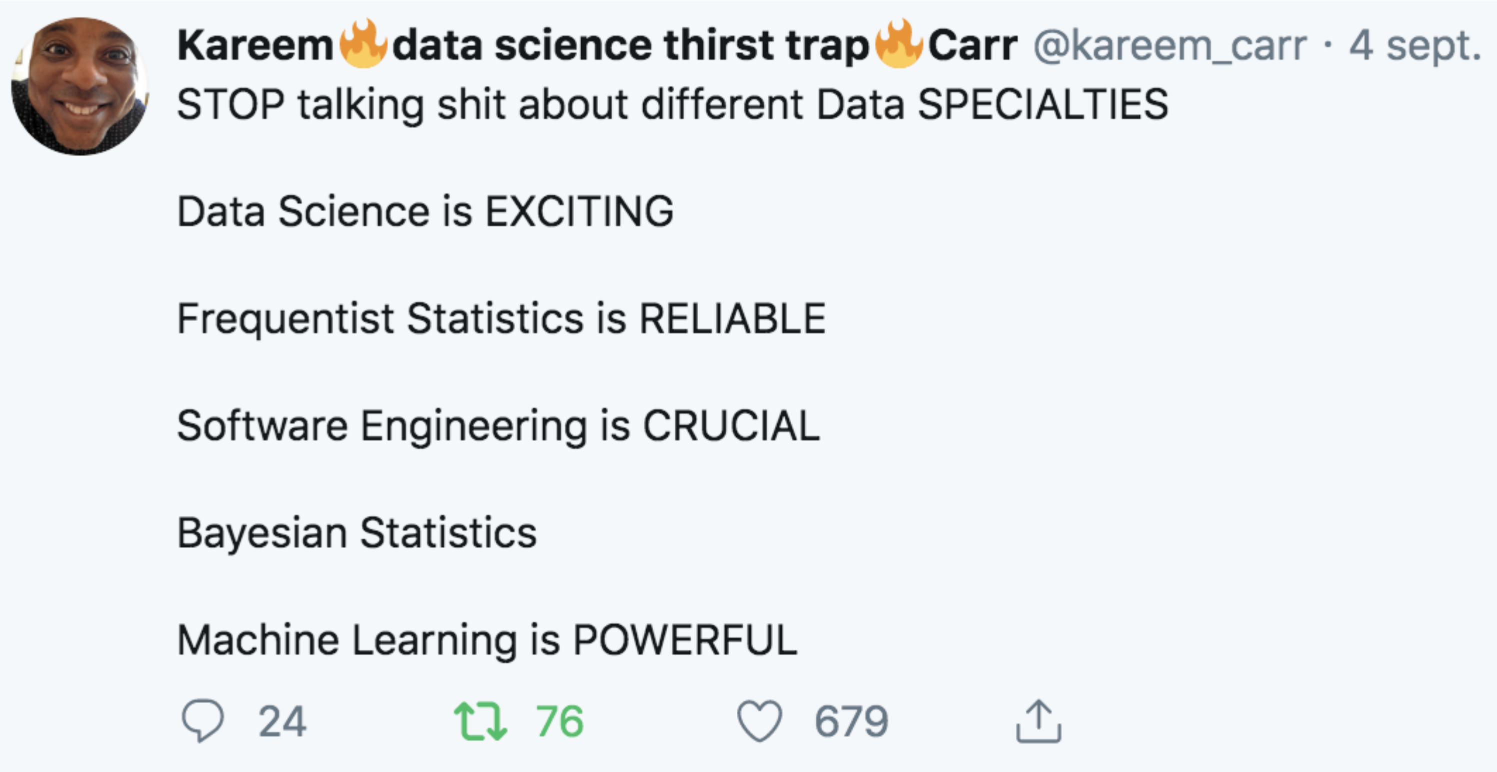
- Choosing priors.
- Defining ROPE.
- End up thinking black and white-ish.

Conclusions

- Use cross-validation to compare models.

- Think about practical differences.

Thanks and questions



Kareem🔥data science thirst trap🔥Carr @kareem_carr · 4 sept.

STOP talking shit about different Data SPECIALTIES

Data Science is EXCITING

Frequentist Statistics is RELIABLE

Software Engineering is CRUCIAL

Bayesian Statistics

Machine Learning is POWERFUL

💬 24 🔁 76 ♡ 679

References

- Inference for the Generalization Error (Nadeau and Bengio, 2003).
- Time for a Change: a Tutorial for Comparing Multiple Classifiers Through Bayesian Analysis (Benavoli et al, 2017).
- Comparing posteriors: Estimating Practical Differences Between Models (Max Kuhn, 2018 New York R Conference).
- tidyposterior package.