Bayesian non parametric modelling of Higgs pair production

Bruno Scarpa
Department of Statistical Sciences - University of Padua

September 1, 2016

joint work with Annalisa Balata (University of Padua) and Tommaso Dorigo (INFN)
Isolate the signal of the Higgs boson pairs decays in the final state characterised by 4 jets of $b$-quark: $hh \rightarrow 4b$
Goal

Isolate the signal of the Higgs boson pairs decays in the final state characterised by 4 jets of $b$-quark: $hh \rightarrow 4b$

Data

- **background**: 1 259 973 observations collected by CMS during the LHC “Run 1” in 2012
  (only if HLT-DiPFJet80-DiPFJet30-BTagCSVd07d05 trigger path is present)

- **signal**: 300 000 $hh \rightarrow b\bar{b}b\bar{b}$ events.
  Monte Carlo simulated events (Alwall et al., 2011; Gao et al. 2014)
Preselection of data: choice of the 4 jets

Events where 4 jets correspond to hadronisation of the $b$-quark

1. $b$-tagging algorithm CMVA (Das et al., 2013)

2. Selection of the first 3 jets in $b$-tag ranking, provided their CMVA is above the medium cut, 0.679

3. The fourth jet is chosen by requiring the least invariant mass difference between pairs of matched dijets
Preselection of data: choice of the 4 jets

Events where 4 jets correspond to hadronisation of the $b$-quark

1. $b$-tagging algorithm CMVA (Das et al., 2013)

2. Selection of the first 3 jets in $b$-tag ranking, provided their CMVA is above the medium cut, 0.679

3. The fourth jet is chosen by requiring the least invariant mass difference between pairs of matched dijets

Final dataset

At the end of preselection, we keep
68 454 MC signal events and
433 621 background CMS data
Available variables

For each event, the following variables are available:

- **Response variable**
  - binary variable, $y_i$ encoding signal ($y_i = 1$) or background ($y_i = 0$)

- **Kinematic explanatory variables**
  1. Transverse momentum
  2. Pseudorapidity
  3. Centrality

- **Variables related to non selected jets**
  1. Minimum, mean and maximum transverse momentum
  2. Minimum, mean and maximum pseudorapidity
  3. Minimum, mean and maximum centrality

- **Variables related to pairs of jets**

- **Other variables**

4 of 36
Available variables

For each event, the following variables are available:

**Response variable**

- binary variable, $y_i$, encoding *signal* ($y_i = 1$) or *background* ($y_i = 0$)
Available variables

For each event, the following variables are available:

**Response variable**

binary variable, $y_i$, encoding **signal** ($y_i = 1$) or **background** ($y_i = 0$)

**Kinematic explanatory variables**

- Variables related to the 4 selected jets and to the couples of dijets
  1. Transverse momentum
  2. Pseudorapidity
  3. Centrality
Available variables

For each event, the following variables are available:

**Response variable**

binary variable, $y_i$, encoding *signal* ($y_i = 1$) or *background* ($y_i = 0$)

**Kinematic explanatory variables**

- Variables related to the 4 selected jets and to the couples of dijets
  1. Transverse momentum
  2. Pseudorapidity
  3. Centrality

- Variables related to non selected jets
  1. Minimum, mean and maximum transverse momentum
  2. Minimum, mean and maximum pseudorapidity
  3. Minimum, mean and maximum centrality
Available variables

For each event, the following variables are available:

Response variable
binary variable, \( y_i \), encoding **signal** \((y_i = 1)\) or **background** \((y_i = 0)\)

Kinematic explanatory variables
- Variables related to the 4 selected jets and to the couples of dijets
  1. Transverse momentum
  2. Pseudorapidity
  3. Centrality
- Variables related to non selected jets
  1. Minimum, mean and maximum transverse momentum
  2. Minimum, mean and maximum pseudorapidity
  3. Minimum, mean and maximum centrality
- Variables related to pairs of jets
Available variables

For each event, the following variables are available:

**Response variable**

- Binary variable, $y_i$, encoding signal ($y_i = 1$) or background ($y_i = 0$)

**Kinematic explanatory variables**

- Variables related to the 4 selected jets and to the couples of dijets
  1. Transverse momentum
  2. Pseudorapidity
  3. Centrality

- Variables related to non selected jets
  1. Minimum, mean and maximum transverse momentum
  2. Minimum, mean and maximum pseudorapidity
  3. Minimum, mean and maximum centrality

- Variables related to pairs of jets

- Other variables
### Variables of the 4 selected jets and to the couples of dijets

<table>
<thead>
<tr>
<th>Name</th>
<th>Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Q_{Pt}1$</td>
<td>Transverse momentum related to the first jet</td>
</tr>
<tr>
<td>$Q_{Pt}2$</td>
<td>Transverse momentum related to the second jet</td>
</tr>
<tr>
<td>$Q_{Pt}3$</td>
<td>Transverse momentum related to the third jet</td>
</tr>
<tr>
<td>$Q_{Pt}4$</td>
<td>Transverse momentum related to the fourth jet</td>
</tr>
<tr>
<td>$Q_{Eta}1$</td>
<td>Pseudorapidity related to the first jet</td>
</tr>
<tr>
<td>$Q_{Eta}2$</td>
<td>Pseudorapidity related to the second jet</td>
</tr>
<tr>
<td>$Q_{Eta}3$</td>
<td>Pseudorapidity related to the third jet</td>
</tr>
<tr>
<td>$Q_{Eta}4$</td>
<td>Pseudorapidity related to the fourth jet</td>
</tr>
<tr>
<td>$Q_{CMVA}1$</td>
<td>CMVA related to the first jet</td>
</tr>
<tr>
<td>$Q_{CMVA}2$</td>
<td>CMVA related to the second jet</td>
</tr>
<tr>
<td>$Q_{CMVA}3$</td>
<td>CMVA related to the third jet</td>
</tr>
<tr>
<td>$Q_{CMVA}4$</td>
<td>CMVA related to the fourth jet</td>
</tr>
<tr>
<td>$Q_{Cent}$</td>
<td>Centrality of the 4 jets</td>
</tr>
</tbody>
</table>
### Available variables

#### Variables of non selected jets

<table>
<thead>
<tr>
<th>Name</th>
<th>Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>$AP_{t \text{ min}}$</td>
<td>minimum $p_t$ among non selected jets</td>
</tr>
<tr>
<td>$AP_{t \text{ mean}}$</td>
<td>mean $p_t$ among non selected jets</td>
</tr>
<tr>
<td>$AP_{t \text{ max}}$</td>
<td>maximum $p_t$ among non selected jets</td>
</tr>
<tr>
<td>$A\ell_{\text{ t min}}$</td>
<td>minimum $\eta$ among non selected jets</td>
</tr>
<tr>
<td>$A\ell_{\text{ t mean}}$</td>
<td>mean $\eta$ among non selected jets</td>
</tr>
<tr>
<td>$A\ell_{\text{ t max}}$</td>
<td>maximum $\eta$ among non selected jets</td>
</tr>
<tr>
<td>ACMV$A_{\text{ min}}$</td>
<td>minimum CMVA among non selected jets</td>
</tr>
<tr>
<td>ACMV$A_{\text{ mean}}$</td>
<td>mean CMVA among non selected jets</td>
</tr>
<tr>
<td>ACMV$A_{\text{ max}}$</td>
<td>maximum CMVA among non selected jets</td>
</tr>
<tr>
<td>$A\text{Cent}$</td>
<td>centrality of non selected jets</td>
</tr>
</tbody>
</table>
Available variables

Variables of pairs of jets, corresponding to each Higgs

<table>
<thead>
<tr>
<th>Name</th>
<th>Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>DJ1mass</td>
<td>mass of the first dijet</td>
</tr>
<tr>
<td>DJ1(p_t)</td>
<td>(p_t) of the first dijet</td>
</tr>
<tr>
<td>DJ1(\Phi)</td>
<td>(\Delta \Phi) of the first dijet</td>
</tr>
<tr>
<td>DJ1(\eta)</td>
<td>(\Delta \eta) of the first dijet</td>
</tr>
<tr>
<td>DJ1(R)</td>
<td>(\Delta R) of the first dijet</td>
</tr>
<tr>
<td>(\tau_1)</td>
<td>twist of the first dijet</td>
</tr>
<tr>
<td>DJ2mass</td>
<td>mass of the second dijet</td>
</tr>
<tr>
<td>DJ2(p_t)</td>
<td>(p_t) of the second dijet</td>
</tr>
<tr>
<td>DJ2(\Phi)</td>
<td>(\Delta \Phi) of the second dijet</td>
</tr>
<tr>
<td>DJ2(\eta)</td>
<td>(\Delta \eta) of the second dijet</td>
</tr>
<tr>
<td>DJ2(R)</td>
<td>(\Delta R) of the second dijet</td>
</tr>
<tr>
<td>(\tau_2)</td>
<td>twist of the second dijet</td>
</tr>
</tbody>
</table>
### Other variables

<table>
<thead>
<tr>
<th>Name</th>
<th>Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>$TDJP_t$</td>
<td>vectorial sum of the $P_t$ of the two dijet</td>
</tr>
<tr>
<td>$TDJ\Delta\Phi$</td>
<td>$\Delta\Phi$ between the two djets</td>
</tr>
<tr>
<td>$TDJ\Delta\eta$</td>
<td>$\Delta\eta$ between the two djets</td>
</tr>
<tr>
<td>$TDJ\Delta R$</td>
<td>$\Delta R$ between the two djets</td>
</tr>
<tr>
<td>$HHM$</td>
<td>invariant mass of the two djets</td>
</tr>
<tr>
<td>$MET$</td>
<td>Missing transverse energy</td>
</tr>
<tr>
<td>$min3cmva$</td>
<td>minimum CMVA among the first 3 jets</td>
</tr>
<tr>
<td>$avg3cmva$</td>
<td>mean CMVA among the first 3 jets</td>
</tr>
<tr>
<td>$cos\theta^*$</td>
<td>cosinus of $\theta$ on the c.o.m reference system of the two H</td>
</tr>
<tr>
<td>$cos\theta_{CS}$</td>
<td>cosinus of $\theta$ on the Collins Saper reference system</td>
</tr>
<tr>
<td>$JetsN$</td>
<td>number of jets in the event</td>
</tr>
</tbody>
</table>

### Obtained variable

<table>
<thead>
<tr>
<th>Name</th>
<th>Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>$sumQP_t$</td>
<td>sum of the transverse momenta of the selected 4 jets</td>
</tr>
</tbody>
</table>
Available variables: Correlation Plot
Approaches

Two approaches

Choice of 9 most predictive variables.

To favour interpretation and contain error propagation

Use of all 39 available variables: focus on the best classification

Estimation strategy

Training set: 50000 balanced events
Test set: 16000
Validation set: 16000
Two approaches

- Choice of 9 most predictive variables. To favour interpretation and contain error propagation
- Use of all 39 available variables: focus on the best classification
Approaches

Two approaches

- Choice of 9 most predictive variables.
  To favour *interpretation* and contain *error propagation*.
- Use of all 39 available variables: focus on the *best classification*.

**Estimation strategy**

- Training set: 50,000 balanced events
- Test set: 16,000
- Validation set: 16,000
We consider a number of typical statistical learning models to best classify signal and background. Linear and logistic regression, MARS, GAM, CART, Neural nets, Projection pursuit, etc. (e.g., Azzalini and Scarpa, 2012).
We consider a number of typical statistical learning models to best classify signal and background. Linear and logistic regression, MARS, GAM, CART, Neural nets, Projection pursuit, etc. (e.g., Azzalini and Scarpa, 2012).

Among the models with best performance on the test set:

- Random forests
- Gradient boosting
- Boosting decision tree
Random forests (Breiman, 1999)

- Refinement of bagged trees; quite popular

At each tree split, a random sample of $m$ features (variables) is drawn, and only those $m$ features are considered for splitting (Typically $m = \sqrt{p}$ or $\log_2 p$, where $p$ is the number of features). For each tree grown on a bootstrap sample, the error rate for observations left out of the bootstrap sample is monitored (out-of-bag). Random forests tries to improve on bagging by “de-correlating” the trees and reduce the variance. Each tree has the same expectation, so increasing the number of trees does not alter the bias of bagging or random forests.

Bias-variance trade off

The small $m$, the lower the variance of the random forest ensemble. Small $m$ is also associated with higher bias, because important variables can be missed by the sampling.
Random forests (Breiman, 1999)

- Refinement of bagged trees; quite popular
- At each tree split, a random sample of $m$ features (variables) is drawn, and only those $m$ features are considered for splitting (Typically $m = \sqrt{p}$ or $\log_2 p$, where $p$ is the number of features)
Random forests (Breiman, 1999)

- Refinement of bagged trees; quite popular
- At each tree split, a random sample of \( m \) features (variables) is drawn, and only those \( m \) features are considered for splitting (Typically \( m = \sqrt{p} \) or \( \log_2 p \), where \( p \) is the number of features)
- For each tree grown on a bootstrap sample, the error rate for observations left out of the bootstrap sample is monitored (out-of-bag)
Random forests (Breiman, 1999)

- Refinement of bagged trees; quite popular
- At each tree split, a random sample of $m$ features (variables) is drawn, and only those $m$ features are considered for splitting (Typically $m = \sqrt{p}$ or $\log_2 p$, where $p$ is the number of features)
- For each tree grown on a bootstrap sample, the error rate for observations left out of the bootstrap sample is monitored (out-of-bag)
- Random forests tries to improve on bagging by “de-correlating” the trees, and reduce the variance.
Random forests (Breiman, 1999)

- Refinement of bagged trees; quite popular
- At each *tree split*, a random sample of *m* features (variables) is drawn, and only those *m* features are considered for splitting. (Typically $m = \sqrt{p}$ or $\log_2 p$, where $p$ is the number of features)
- For each tree grown on a bootstrap sample, the error rate for observations left out of the bootstrap sample is monitored (*out-of-bag*)
- Random forests tries to improve on bagging by “de-correlating” the trees, and reduce the variance.
- Each tree has the same expectation, so increasing the number of trees does not alter the bias of bagging or random forests.
Random forests (Breiman, 1999)

- Refinement of bagged trees; quite popular
- At each *tree split*, a random sample of *m* features (variables) is drawn, and only those *m* features are considered for splitting (Typically $m = \sqrt{p}$ or $\log_2 p$, where $p$ is the number of features)
- For each tree grown on a bootstrap sample, the error rate for observations left out of the bootstrap sample is monitored (*out-of-bag*)
- Random forests tries to improve on bagging by “de-correlating” the trees, and reduce the variance.
- Each tree has the same expectation, so increasing the number of trees does not alter the bias of bagging or random forests.

**Bias-variance trade off**

- *The small m, the lower the variance of the random forest ensemble*
- *Small m is also associated with higher bias, because important variables can be missed by the sampling*
Interactions

Random forests overemphasizes the role of interactions
Random forests overemphasizes the role of interactions
What are the most interactive kinematic variables in predicting signal from background?
- Random forests overemphasizes the role of interactions
- What are the most interactive kinematic variables in predicting signal from background?
- Maximal subtree analysis (Ishwaran et al., 2010, 2011) for identifying interactions. Smaller minimum depth of a variable \( j \) with respect to the maximal subtree for variable \( i \) indicates interactions.
- We choose to select couples of variables with maximal subtree smaller than 0.045.
Interactions

- Random forests overemphasizes the role of interactions
- What are the most interactive kinematic variables in predicting signal from background?
- Maximal subtree analysis (Ishwaran et al., 2010, 2011) for identifying interactions. Smaller minimum depth of a variable $j$ with respect to the maximal subtree for variable $i$ indicates interactions.
- We choose to select couples of variables with maximal subtree smaller than 0.045.

<table>
<thead>
<tr>
<th>DJ1P&lt;sub&gt;t&lt;/sub&gt;</th>
<th>DJ1R</th>
<th>QCent</th>
<th>QP&lt;sub&gt;t&lt;/sub&gt;3</th>
</tr>
</thead>
<tbody>
<tr>
<td>QP&lt;sub&gt;t&lt;/sub&gt;3</td>
<td>TDJΔR</td>
<td>QP&lt;sub&gt;t&lt;/sub&gt;3</td>
<td>mhh</td>
</tr>
<tr>
<td>TDJΔR</td>
<td>mhh</td>
<td>TDJΔR</td>
<td>DJ1R</td>
</tr>
<tr>
<td>DJ1R</td>
<td>DJ2P&lt;sub&gt;t&lt;/sub&gt;</td>
<td>DJ1R</td>
<td>DJ2R</td>
</tr>
<tr>
<td>QCent</td>
<td>DJ1R</td>
<td>QP&lt;sub&gt;t&lt;/sub&gt;3</td>
<td>DJ2R</td>
</tr>
<tr>
<td>QCent</td>
<td>mhh</td>
<td>mhh</td>
<td>DJ2R</td>
</tr>
<tr>
<td>mhh</td>
<td>DJ1R</td>
<td>QP&lt;sub&gt;t&lt;/sub&gt;3</td>
<td>DJ1R</td>
</tr>
</tbody>
</table>
Idea: Put more weight on observations that are misclassified, to make the classifier work harder on those points.
Boosting (Freund & Schapire, 1997)

- **Idea**: Put more weight on observations that are misclassified, to make the classifier work harder on those points.
- Average many trees, each grown to reweighted versions of the training data.

\[ C(x) = \text{sign}\left[ M \sum_{m=1}^{\infty} \alpha_m C_m(x) \right] \]

Note: \( C_m(x) \in \{-1, +1\} \).

Boosting Decision Tree (BDT; Drucker, 1997): each error is normalized with the maximum of errors.
Boosting (Freund e Schapire, 1997)

- **Idea**: Put more weight on observations that are misclassified, to make the classifier work harder on those points.
- Average many trees, each grown to reweighted versions of the training data.
- Weighting *decorrelates* the trees, by focussing on regions missed by past trees.
Boosting (Freund e Schapire, 1997)

- **Idea**: Put more weight on observations that are misclassified, to make the classifier work harder on those points.
- Average many trees, each grown to reweighted versions of the training data.
- Weighting *decorrelates* the trees, by focussing on regions missed by past trees.
- Final classifier is weighted average of classifiers

\[ C(x) = \text{sign} \left[ \sum_{m=1}^{M} \alpha_m C_m(x) \right] \]

Note: \( C_m(x) \in \{-1, +1\} \).
Boosting (Freund e Schapire, 1997)

- **Idea**: Put more weight on observations that are misclassified, to make the classifier work harder on those points.
- Average many trees, each grown to reweighted versions of the training data.
- Weighting *decorrelates* the trees, by focussing on regions missed by past trees.
- Final classifier is weighted average of classifiers

\[
C(x) = \text{sign} \left[ \sum_{m=1}^{M} \alpha_m C_m(x) \right]
\]

Note: \( C_m(x) \in \{-1, +1\} \).

- Boosting Decision Tree (BDT; Drucker, 1997): each error is normalized with the maximum of errors
Adaboost

Details

- Start with equal observation weights $p_i = 1/n$
- At iteration $t$, draw a bootstrap sample with the current probabilities $p_1, p_2, \ldots, p_n$, compute the classifier and $e_t$, the error rate of the classifier on the original sample (for BDT normalized with the maximum error). Let $\beta_t = e_t/(1 - e_t)$
- For those points that are classified correctly, decrease their probabilities by
  $$p_i \leftarrow p_i \cdot \beta_t$$
  and renormalise them
- Do this for many (say 1000) iterations.

At the end, take a weighted vote of the classifications, with weights $\log(1/\beta_t)$ (more weight on classifiers with lower error). Boosting can improve bagging in many instances.
Gradient boosting builds additive tree models, for example, for representing logits in logistic regression.
Gradient boosting builds additive tree models, for example, for representing logits in logistic regression.

Tree size is a parameter that determines the order of interaction.
Gradient boosting (Friedman, 2002)

- Gradient boosting builds additive tree models, for example, for representing logits in logistic regression.
- *Tree size* is a parameter that determines the order of interaction.
- Gradient Boosting inherits all the good features of trees (variable selection, missing data, mixed predictors), and improves on the weak features, such as prediction performance.
Gradient boosting builds additive tree models, for example, for representing logits in logistic regression.

*Tree size* is a parameter that determines the order of interaction.

Gradient Boosting inherits all the good features of trees (variable selection, missing data, mixed predictors), and improves on the weak features, such as prediction performance.

Idea: Fit a weighted additive model (ensemble) in a forward stage-wise manner.
Gradient boosting builds additive tree models, for example, for representing logits in logistic regression.

*Tree size* is a parameter that determines the order of interaction.

Gradient Boosting inherits all the good features of trees (variable selection, missing data, mixed predictors), and improves on the weak features, such as prediction performance.

**Idea:** Fit a weighted additive model (ensemble) in a forward stage-wise manner.

In each stage, introduce a ‘weak learner’ to compensate the shortcomings of existing weak learners.
Gradient boosting (Friedman, 2002)

- Gradient boosting builds additive tree models, for example, for representing logits in logistic regression.
- *Tree size* is a parameter that determines the order of interaction.
- Gradient Boosting inherits all the good features of trees (variable selection, missing data, mixed predictors), and improves on the weak features, such as prediction performance.
- **Idea**: Fit a weighted additive model (ensemble) in a forward stage-wise manner.
- In each stage, introduce a ‘weak learner’ to compensate the shortcomings of existing weak learners.
- In Gradient Boosting, ‘shortcomings’ are identified by gradients (in Adaboost, ‘shortcomings’ are identified by high-weight data points).
Gradient boosting (Friedman, 2002)

- Gradient boosting builds additive tree models, for example, for representing logits in logistic regression.
- Tree size is a parameter that determines the order of interaction.
- Gradient Boosting inherits all the good features of trees (variable selection, missing data, mixed predictors), and improves on the weak features, such as prediction performance.
- Idea: Fit a weighted additive model (ensemble) in a forward stage-wise manner.
- In each stage, introduce a ‘weak learner’ to compensate the shortcomings of existing weak learners.
- In Gradient Boosting, ‘shortcomings’ are identified by gradients (in Adaboost, ‘shortcomings’ are identified by high-weight data points).
- Both high-weight data points and gradients tell us how to improve the model.
ROC curve of the principal statistical learning classification models
Performances of the best statistical learning classification models

<table>
<thead>
<tr>
<th>Model</th>
<th>Error test set</th>
<th>Error validation set</th>
<th>AUC test set</th>
<th>AUC validation set</th>
</tr>
</thead>
<tbody>
<tr>
<td>BDT9</td>
<td>0.2821</td>
<td>0.2695</td>
<td>0.7969</td>
<td>0.7934</td>
</tr>
<tr>
<td>RF9</td>
<td>0.2756</td>
<td>0.2685</td>
<td>0.7982</td>
<td>0.8004</td>
</tr>
<tr>
<td>GBM9</td>
<td>0.2851</td>
<td>0.2858</td>
<td>0.7888</td>
<td>0.7911</td>
</tr>
<tr>
<td>BDT38</td>
<td>0.2596</td>
<td>0.2637</td>
<td>0.8204</td>
<td>0.8227</td>
</tr>
<tr>
<td>RF38</td>
<td>0.2540</td>
<td>0.2521</td>
<td>0.8224</td>
<td>0.8263</td>
</tr>
<tr>
<td>BDT39</td>
<td>0.2340</td>
<td>0.2464</td>
<td>0.8349</td>
<td>0.8400</td>
</tr>
<tr>
<td>RF39</td>
<td>0.2320</td>
<td>0.2424</td>
<td>0.8369</td>
<td>0.8424</td>
</tr>
<tr>
<td>GBM39</td>
<td>0.2431</td>
<td>0.2477</td>
<td>0.8278</td>
<td>0.8328</td>
</tr>
</tbody>
</table>
Performances of the best statistical learning classification models

<table>
<thead>
<tr>
<th>Model</th>
<th>Error test set</th>
<th>Error validation set</th>
<th>AUC test set</th>
<th>AUC validation set</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF9</td>
<td>0.2756</td>
<td>0.2685</td>
<td>0.7982</td>
<td>0.8004</td>
</tr>
<tr>
<td>BDT9</td>
<td>0.2821</td>
<td>0.2695</td>
<td>0.7969</td>
<td>0.7934</td>
</tr>
<tr>
<td>GBM9</td>
<td>0.2851</td>
<td>0.2858</td>
<td>0.7888</td>
<td>0.7911</td>
</tr>
<tr>
<td>RF38</td>
<td>0.2540</td>
<td>0.2521</td>
<td>0.8224</td>
<td>0.8263</td>
</tr>
<tr>
<td>BDT38</td>
<td>0.2596</td>
<td>0.2637</td>
<td>0.8204</td>
<td>0.8227</td>
</tr>
<tr>
<td>RF39</td>
<td>0.2320</td>
<td>0.2424</td>
<td>0.8369</td>
<td>0.8424</td>
</tr>
<tr>
<td>BDT39</td>
<td>0.2340</td>
<td>0.2464</td>
<td>0.8349</td>
<td>0.8400</td>
</tr>
<tr>
<td>GBM39</td>
<td>0.2431</td>
<td>0.2477</td>
<td>0.8278</td>
<td>0.8328</td>
</tr>
</tbody>
</table>
Comparison RF and BDT

RF

Prediction

BDT

Prediction

Green: background
Blue: signal
A generalized mixed effects model

- Let $y_i$ be the binary variable encoding signal or background,
- the classical generalized mixed model formulation assumes

$$y_i | \pi_i \sim \text{Bern}(\pi_i)$$

$$\logit(\pi_i) = \eta_i$$

$$\eta_i = \mu_i + f(x_i)$$

where

- $x_i$ is the vector including all the explanatory variables for each event $i$
- $\beta$ is a vector of parameters
- $\mu_i$ is a random effect for each event
- $p$ is the number of available explanatory variables
A generalized mixed effects model

- Let $y_i$ be the binary variable encoding signal or background,
- the classical generalized mixed model formulation assumes

$$ y_i | \pi_i \sim \text{Bern}(\pi_i) $$

$$ \logit(\pi_i) = \eta_i $$

$$ \eta_i = \mu_i + f(x_i) $$

where

- $x_i$ is the vector including all the explanatory variables for each event $i$
- $\beta$ is a vector of parameters
- $\mu_i$ is a random effect for each event
- $p$ is the number of available explanatory variables

- Bayesian approach, assuming priors on parameters.
  (still “intrinsically” frequentist - not a subjective approach)
Dirichlet process (DP):
assume $\mu_i \sim P$ with $P \sim DP(\alpha P_0)$, $\alpha > 0$
We also assume fixed effects for explanatory variables $f(x_i) = x_i^T \beta$
Dirichlet process (DP):
assume $\mu_i \sim P$ with $P \sim DP(\alpha P_0)$, $\alpha > 0$
We also assume fixed effects for explanatory variables $f(x_i) = x_i^T \beta$

Additive model with $P$-splines

$$f(x_i) = f_1(x_{i1}) + \cdots + f_p(x_{ip})$$

$f_1(\cdot), \ldots, f_p(\cdot)$ fitted via Bayesian $P$-splines
and $\mu_i$ assumed constant and fixed
Bayesian non parametric models - BNP

- **Dirichlet process (DP):**
  assume $\mu_i \sim P$ with $P \sim DP(\alpha P_0), \alpha > 0$
  We also assume fixed effects for explanatory variables $f(x_i) = x_i^T \beta$

- **Additive model with P-splines**
  
  $f(x_i) = f_1(x_{i1}) + \cdots + f_p(x_{ip})$

  $f_1(\cdot), \ldots, f_p(\cdot)$ fitted via Bayesian $P$-splines
  and $\mu_i$ assumed constant and fixed

- **BART model (Bayesian Additive Regression Tree)**

  $f(x_i) = \sum_{j=1}^{m} g(x_i; T_j, M_j)$

  where $g(x_i; T_j, M_j)$ denotes the predicting function assigning a value
to $x_i$ given the Bayesian tree $T_j$ and parameters $M_j$
  $\mu_i$ assumed constant and fixed.
Let

\[ \eta_i = \mu_i + x_i^T \beta \]

Taking a Bayesian approach we specify prior distributions for the parameters \((\mu_i, \beta)\)
Mixed effect model with Dirichlet process

Let

\[ \eta_i = \mu_i + x_i^T \beta \]

Taking a Bayesian approach we specify prior distributions for the parameters \((\mu_i, \beta)\)

- to provide a flexible formulation for adaptively modeling the determinants of signal
Mixed effect model with Dirichlet process

Let

$$\eta_i = \mu_i + x_i^T \beta$$

Taking a Bayesian approach we specify prior distributions for the parameters $$(\mu_i, \beta)$$

- to provide a flexible formulation for adaptively modeling the determinants of signal
- to allow uncertainty in the distribution of the random intercepts, while avoiding over-shrinking and favouring clustering effects
Let

\[ \eta_i = \mu_i + x_i^T \beta \]

Taking a Bayesian approach we specify prior distributions for the parameters \((\mu_i, \beta)\)

- to provide a flexible formulation for adaptively modeling the determinants of signal
- to allow uncertainty in the distribution of the random intercepts, while avoiding over-shrinking and favouring clustering effects
- needs for an efficient algorithm for posterior computation
Mixed effect model with Dirichlet process

Let

\[ \eta_i = \mu_i + x_i^T \beta \]

Taking a Bayesian approach we specify prior distributions for the parameters \((\mu_i, \beta)\)

- to provide a flexible formulation for adaptively modeling the determinants of signal
- to allow uncertainty in the distribution of the random intercepts, while avoiding over-shrinking and favouring clustering effects
- needs for an efficient algorithm for posterior computation

Priors distributions

- \(\beta \sim \mathcal{N}_p(b, B)\)
- \(b_{p \times 1}, B_{p \times p}\) are prior mean vector and covariance matrix
- \(\mu_i \sim P\) with \(P \sim \text{DP}(\alpha P_0), \alpha > 0\), where DP indicates the Dirichlet Process.
The Dirichlet process $DP(\alpha P_0)$ represents a fully flexible prior with support on the set of distributions on the real line, allowing $P$ to be unknown with

- $P_0$ indicating the best guess for such distribution and
- $\alpha$ quantifying the confidence in such guess.

In our case, we define $P_0$ as a normal distribution $\mathcal{N}(0, \sigma^2)$ where

- $\sigma^{-2} \sim Gamma(a, b)$ (i.e. prior for $\sigma$ is inverse Gamma)
- $\alpha \sim Gamma(a_\alpha, b_\alpha)$ to favor learning of cluster effects in the data.
Dirichlet Process

We exploit the stick-breaking representation of the Dirichlet Process

### Stick-breaking representation (Sethuraman, 1994)

Let

\[ V_h \overset{iid}{\sim} \text{Beta}(1, \alpha) \quad \theta_h \overset{iid}{\sim} G_0 \]

\[ \pi_h \sim V_h \prod_{l<h} (1 - V_l) \]

\[ G = \sum_{h=1}^{\infty} \pi_h \delta_{\theta_h} \]

Therefore \( G \sim DP(\alpha, G_0) \), where \( \delta_{\theta} \) indicates a mass point concentrated in \( \theta \).
Dirichlet Process

We exploit the stick-breaking representation of the Dirichlet Process

**Stick-breaking representation (Sethuraman, 1994)**

Let

\[ V_h \overset{iid}{\sim} \text{Beta}(1, \alpha) \quad \theta_h \overset{iid}{\sim} G_0 \]

\[ \pi_h \sim V_h \prod_{l<h}(1 - V_l) \quad G = \sum_{h=1}^{\infty} \pi_h \delta_{\theta_h} \]

therefore \( G \sim DP(\alpha, G_0) \), where \( \delta_{\theta} \) indicates a mass point concentrated in \( \theta \).

- Key result: a realization of the Dirichlet process is discrete in nature
We exploit the stick-breaking representation of the Dirichlet Process

**Stick-breaking representation (Sethuraman, 1994)**

Let

\[
V_h \overset{iid}{\sim} \text{Beta}(1, \alpha) \quad \theta_h \overset{iid}{\sim} G_0
\]

\[
\pi_h \sim V_h \prod_{l<h} (1 - V_l) \quad G = \sum_{h=1}^{\infty} \pi_h \delta_{\theta_h}
\]

therefore \( G \sim DP(\alpha, G_0) \), where \( \delta_{\theta} \) indicates a mass point concentrated in \( \theta \).

- Key result: a realization of the Dirichlet process is discrete in nature
- It favours ties among random intercepts: events in the same cluster have equal random intercept values
We exploit the stick-breaking representation of the Dirichlet Process

**Stick-breaking representation (Sethuraman, 1994)**

Let

\[ V_h \sim \text{Beta}(1, \alpha) \quad \theta_h \sim G_0 \]

\[ \pi_h \sim V_h \prod_{l<h} (1 - V_l) \quad G = \sum_{h=1}^{\infty} \pi_h \delta_{\theta_h} \]

therefore \( G \sim DP(\alpha, G_0) \), where \( \delta_\theta \) indicates a mass point concentrated in \( \theta \).

- **Key result:** a realization of the Dirichlet process is **discrete** in nature.
- It favours ties among random intercepts: events in the same cluster have equal random intercept values.
- Denoting with \( S_i \) the grouping variable, the stick-breaking representation shows clustering effects among events, providing \( \mu_i = \theta_{S_i} \), with the number of clusters stochastically increasing with \( \alpha \).
Dirichlet Process

We exploit the stick-breaking representation of the Dirichlet Process

**Stick-breaking representation (Sethuraman, 1994)**

Let

\[ V_h \sim \text{Beta}(1, \alpha) \quad \theta_h \sim G_0 \]

\[ \pi_h \sim V_h \prod_{l<h} (1 - V_l) \quad G = \sum_{h=1}^{\infty} \pi_h \delta_{\theta_h} \]

therefore \( G \sim DP(\alpha, G_0) \), where \( \delta_{\theta} \) indicates a mass point concentrated in \( \theta \).

- Key result: a realization of the Dirichlet process is discrete in nature
- It favours ties among random intercepts: events in the same cluster have equal random intercept values
- Denoting with \( S_i \) the grouping variable, the stick-breaking representation shows clustering effects among events, providing \( \mu_i = \theta_{S_i} \), with the number of clusters stochastically increasing with \( \alpha \)
- This clustering property is particularly useful in our signal detection, favouring events with common kinematic features to share the same effect
We exploit the stick-breaking representation of the Dirichlet Process

**Stick-breaking representation (Sethuraman, 1994)**

Let

\[ V_h \sim \text{Beta}(1, \alpha) \quad \theta_h \sim G_0 \]

\[ \pi_h \sim V_h \prod_{l<h} (1 - V_l) \quad G = \sum_{h=1}^{\infty} \pi_h \delta_{\theta_h} \]

therefore \( G \sim DP(\alpha, G_0) \), where \( \delta_{\theta} \) indicates a mass point concentrated in \( \theta \).

- Key result: a realization of the Dirichlet process is **discrete** in nature
- It favours ties among random intercepts: events in the same cluster have equal random intercept values
- Denoting with \( S_i \) the grouping variable, the stick-breaking representation shows clustering effects among events, providing \( \mu_i = \theta_{S_i} \), with the number of clusters stochastically increasing with \( \alpha \)
- This clustering property is particularly useful in our signal detection, favouring events with common kinematic features to share the same effect
- Conditionally on the grouping indicator \( S_i \), the Gaussian base measure \( P_0 \) is conjugate, favoring the implementation of a Gibbs sampler
For posterior computation we exploit

- *blocked Gibbs sampler* algorithm by Ishwaran and James (2001)
Posterior computation

For posterior computation we exploit

- *blocked Gibbs sampler* algorithm by Ishwaran and James (2001)
- a recently proposed data-augmentation scheme based on Pólya-Gamma (PG) distribution

**Pólya-Gamma data-augmentation**

Assuming a Bayesian logistic regression setting where

\( y_i \sim \text{Bern}(1/[1 + e^{-\phi_i}]), i = 1, \ldots, n, \phi_i = x_i^T \beta \) and \( \beta \sim \mathcal{N}_p(b, B) \), the resulting Gibbs sampler alternates between two full conditional conjugate steps

- \( \omega_i \sim \text{PG}(1, x_i^T \beta) \)
- \( \beta|y, \omega, x \sim \mathcal{N}_p(\mu_\beta, \Sigma_\beta) \)

where \( \Sigma_\beta = (X^T \Omega X + B^{-1}) \) and \( \mu_\beta = \Sigma_\beta (X^T z + B^{-1}b) \),

\( z = [y_1 - 1/2, \ldots, y_n - 1/2] \) and \( \Omega = \text{diag}(\omega_1, \ldots, \omega_n) \)
## Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Test set classification error</th>
<th>False positives</th>
<th>False negatives</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic BNP</td>
<td>0.492</td>
<td>0.4121</td>
<td>0.5663</td>
<td>0.5118</td>
</tr>
</tbody>
</table>
## Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Test set classification error</th>
<th>False positives</th>
<th>False negatives</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic BNP</td>
<td>0.492</td>
<td>0.4121</td>
<td>0.5663</td>
<td>0.5118</td>
</tr>
<tr>
<td>RF39</td>
<td>0.320</td>
<td>0.2678</td>
<td>0.3678</td>
<td>0.7393</td>
</tr>
</tbody>
</table>
Drawbacks of the Dirichlet process model

- **Nonlinearities** are not allowed (variables are related to response by linear functions)
Drawbacks of the Dirichlet process model

- **Nonlinearities** are not allowed (variables are related to response by linear functions)
- **Interactions** among variables are not allowed
Drawbacks of the Dirichlet process model

- **Nonlinearities** are not allowed (variables are related to response by linear functions)
- **Interactions** among variables are not allowed

back to the generalized mixed model

\[
y_i | \pi_i \sim \text{Bern}(\pi_i) \\
\text{logit}(\pi_i) \sim \eta_i \\\n\eta_i \sim \mu_i + f(x_i)
\]
Additive model with $P$-splines

Bayesian additive model $f(x_i) = f_1(x_{i1}) + \cdots + f_p(x_{ip})$

$f_j(x_j) = M_j \sum_{r=1}^{\beta_j} \beta_{jr} B_{jr}(x_j)$

where $B_{jr}$ is the $r$-th base function and $\beta_j = (\beta_{j1}, \ldots, \beta_{jM_j})$ is a parameter vector.

Include interactions identified with Random Forests, as new variables.

<table>
<thead>
<tr>
<th></th>
<th>Test set False</th>
<th>False</th>
<th>AUC</th>
<th>classification error</th>
<th>positives</th>
<th>negatives</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P$-splines</td>
<td>0.340</td>
<td>0.3975</td>
<td>0.2874</td>
<td>0.7102</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RF39</td>
<td>0.320</td>
<td>0.2678</td>
<td>0.3678</td>
<td>0.7393</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Additive model with $P$-splines

- Bayesian additive model

$$f(x_i) = f_1(x_{i1}) + \cdots + f_p(x_{ip})$$

$f_1(\cdot), \ldots, f_p(\cdot)$ estimated via Bayesian $P$-splines, and $\mu_i = 0$. 

| $P$-splines | 0.3400 | 0.3975 | 0.2874 | 0.7102 |
| RF39       | 0.320  | 0.2678 | 0.3678 | 0.7393 |
Additive model with $P$-splines

- Bayesian additive model

$$f(x_i) = f_1(x_{i1}) + \cdots + f_p(x_{ip})$$

$f_1(\cdot), \ldots, f_p(\cdot)$ estimated via Bayesian $P$-splines, and $\mu_i = 0$.

- $P$-splines (Bayesian Penalized Splines, Lang and Brezger, 2004)

$$f_j(x_j) = \sum_{r=1}^{M_j} \beta_{jr} B_{jr}(x_j)$$

where $B_{jr}$ is the $r$-th base function and $\beta_j = (\beta_{j1}, \ldots, \beta_{jM_j})$ is a parameter vector.
Additive model with $P$-splines

- Bayesian additive model

$$f(x_i) = f_1(x_{i1}) + \cdots + f_p(x_{ip})$$

$f_1(\cdot), \ldots, f_p(\cdot)$ estimated via Bayesian $P$-splines, and $\mu_i = 0$.

- $P$-splines (Bayesian Penalized Splines, Lang and Brezger, 2004)

$$f_j(x_j) = \sum_{r=1}^{M_j} \beta_{jr} B_{jr}(x_j)$$

where $B_{jr}$ is the $r$-th base function and $\beta_j = (\beta_{j1}, \ldots, \beta_{jM_j})$ is a parameter vector.

- Include interactions identified with Random Forests, as new variables.
Additive model with \( P \)-splines

- Bayesian additive model

\[
f(x_i) = f_1(x_{i1}) + \cdots + f_p(x_{ip})
\]

\( f_1(\cdot), \ldots, f_p(\cdot) \) estimated via Bayesian \( P \)-splines, and \( \mu_i = 0 \).

- \( P \)-splines (Bayesian Penalized Splines, Lang and Brezger, 2004)

\[
f_j(x_j) = \sum_{r=1}^{M_j} \beta_{jr} B_{jr}(x_j)
\]

where \( B_{jr} \) is the \( r \)-th base function and \( \beta_j = (\beta_{j1}, \ldots, \beta_{jM_j}) \) is a parameter vector.

- Include interactions identified with Random Forests, as new variables.

<table>
<thead>
<tr>
<th></th>
<th>Test set classification error</th>
<th>False positives</th>
<th>False negatives</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P )-splines</td>
<td>0.3400</td>
<td>0.3975</td>
<td>0.2874</td>
<td>0.7102</td>
</tr>
</tbody>
</table>
Additive model with $P$-splines

- Bayesian additive model

$$f(x_i) = f_1(x_{i1}) + \cdots + f_p(x_{ip})$$

$f_1(\cdot), \ldots, f_p(\cdot)$ estimated via Bayesian $P$-splines, and $\mu_i = 0$.

- $P$-splines (Bayesian Penalized Splines, Lang and Brezger, 2004)

$$f_j(x_j) = \sum_{r=1}^{M_j} \beta_{jr} B_{jr}(x_j)$$

where $B_{jr}$ is the $r$-th base function and $\beta_j = (\beta_{j1}, \ldots, \beta_{jM_j})$ is a parameter vector.

- Include interactions identified with Random Forests, as new variables.

<table>
<thead>
<tr>
<th></th>
<th>Test set classification error</th>
<th>False positives</th>
<th>False negatives</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P$-splines</td>
<td>0.3400</td>
<td>0.3975</td>
<td>0.2874</td>
<td>0.7102</td>
</tr>
<tr>
<td>RF39</td>
<td>0.320</td>
<td>0.2678</td>
<td>0.3678</td>
<td>0.7393</td>
</tr>
</tbody>
</table>
BART (Chipman, George and McCulloch, 2010)

\[ f(x_i) = \sum_{j=1}^{m} g(x_i; T_j, M_j) \]

where \( g(x_i, T_j, M_j) \) denotes the predicting function assigning a value to \( x_i \) given the Bayesian tree \( T_j \) and parameters \( M_j \)
BART - Bayesian Additive Regression Tree

- BART (Chipman, George and McCulloch, 2010)

\[ f(x_i) = \sum_{j=1}^{m} g(x_i; T_j, M_j) \]

where \( g(x_i, T_j, M_j) \) denotes the predicting function assigning a value to \( x_i \) given the Bayesian tree \( T_j \) and parameters \( M_j \).

- Sum of trees where variable selection at each node, cut points and depth are parameters (needing for a prior distribution).
BART - Bayesian Additive Regression Tree

- BART (Chipman, George and McCulloch, 2010)

\[ f(x_i) = \sum_{j=1}^{m} g(x_i; T_j, M_j) \]

where \( g(x_i, T_j, M_j) \) denotes the predicting function assigning a value to \( x_i \) given the Bayesian tree \( T_j \) and parameters \( M_j \)

- Sum of trees where variable selection at each node, cut points and depth are parameters (needing for a prior distribution).

- To fit a BART: tailored version of Bayesian backfitting MCMC (Hastie and Tibshirani, 2000) that iteratively constructs and fits successive residuals
BART - Bayesian Additive Regression Tree

- BART (Chipman, George and McCulloch, 2010)

\[ f(x_i) = \sum_{j=1}^{m} g(x_i; T_j, M_j) \]

where \( g(x_i, T_j, M_j) \) denotes the predicting function assigning a value to \( x_i \) given the Bayesian tree \( T_j \) and parameters \( M_j \)

- Sum of trees where variable selection at each node, cut points and depth are parameters (needing for a prior distribution).

- To fit a BART: tailored version of Bayesian backfitting MCMC (Hastie and Tibshirani, 2000) that iteratively constructs and fits successive residuals

- Interactions are included in the tree model
BART - Bayesian Additive Regression Tree

- BART (Chipman, George and McCulloch, 2010)

\[ f(x_i) = \sum_{j=1}^{m} g(x_i; T_j, M_j) \]

where \( g(x_i; T_j, M_j) \) denotes the predicting function assigning a value to \( x_i \) given the Bayesian tree \( T_j \) and parameters \( M_j \)

- Sum of trees where variable selection at each node, cut points and depth are parameters (needing for a prior distribution).

- To fit a BART: tailored version of Bayesian backfitting MCMC (Hastie and Tibshirani, 2000) that iteratively constructs and fits successive residuals

- Interactions are included in the tree model

<table>
<thead>
<tr>
<th></th>
<th>Test set classification error</th>
<th>False positives</th>
<th>False negatives</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>BART</td>
<td>0.3480</td>
<td>0.3083</td>
<td>0.3846</td>
<td>0.7078</td>
</tr>
</tbody>
</table>
BART (Chipman, George and McCulloch, 2010)

\[ f(x_i) = \sum_{j=1}^{m} g(x_i; T_j, M_j) \]

where \( g(x_i, T_j, M_j) \) denotes the predicting function assigning a value to \( x_i \) given the Bayesian tree \( T_j \) and parameters \( M_j \).

- Sum of trees where variable selection at each node, cut points and depth are parameters (needing for a prior distribution).
- To fit a BART: tailored version of Bayesian backfitting MCMC (Hastie and Tibshirani, 2000) that iteratively constructs and fits successive residuals.
- Interactions are included in the tree model.

<table>
<thead>
<tr>
<th></th>
<th>Test set classification error</th>
<th>False positives</th>
<th>False negatives</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>BART</td>
<td>0.3480</td>
<td>0.3083</td>
<td>0.3846</td>
<td>0.7078</td>
</tr>
<tr>
<td>RF39</td>
<td>0.320</td>
<td>0.2678</td>
<td>0.3678</td>
<td>0.7393</td>
</tr>
</tbody>
</table>
Combinations of models

- Dirichlet process and $P$-splines

\[ f(x_i) = \mu_i + f_1(x_{i1}) + \cdots + f_p(x_{ip}) \]

where $\mu_i$ is a DP with Gaussian atoms, and the $f_j$ are Bayesian $P$-splines
Combinations of models

- **Dirichlet process and P-splines**

  \[ f(x_i) = \mu_i + f_1(x_{i1}) + \cdots + f_p(x_{ip}) \]

  where \( \mu_i \) is a DP with Gaussian atoms, and the \( f_j \) are Bayesian P-splines

- **Dirichlet process with BART atoms**

  the atoms of the Dirichlet process depends on the \( x_i \) and are described by a BART.
Combinations of models

- Dirichlet model with BART atoms and $P$-splines

\[ f(x_i) = \mu_i + f_1(x_{i1}) + \cdots + f_p(x_{ip}) \]

where $\mu_i$ is a DP with atoms described by a BART and $f_j$ are Bayesian $P$-splines
Combinations of models

- **Dirichlet model with BART atoms and \( P \)-splines**

\[
f(x_i) = \mu_i + f_1(x_{i1}) + \cdots + f_p(x_{ip})
\]

where \( \mu_i \) is a DP with atoms described by a BART and \( f_j \) are Bayesian \( P \)-splines

- **BART with \( P \)-SPLINES**

Si stima il modello

\[
f(x_i) = f_1(x_{i1}) + \cdots + f_p(x_{ip}) + \sum_{j=1}^{m} g(x; T_j, M_j)
\]

where \( f_j \) are estimated via Bayesian \( P \)-splines, and sum of \( g \) is a BART model.
## Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Classification Error</th>
<th>False Positives</th>
<th>False Negatives</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF</td>
<td>0.3200</td>
<td>0.2678</td>
<td>0.3678</td>
<td>0.7393</td>
</tr>
<tr>
<td>GB</td>
<td>0.3300</td>
<td>0.2971</td>
<td>0.3602</td>
<td>0.7234</td>
</tr>
<tr>
<td>BDT</td>
<td>0.3351</td>
<td>0.3461</td>
<td>0.3389</td>
<td>0.7198</td>
</tr>
<tr>
<td>DP</td>
<td>0.4920</td>
<td>0.4121</td>
<td>0.5663</td>
<td>0.5118</td>
</tr>
<tr>
<td>P-splines</td>
<td>0.3400</td>
<td>0.3975</td>
<td>0.2874</td>
<td>0.7102</td>
</tr>
<tr>
<td>BART</td>
<td>0.3480</td>
<td>0.3083</td>
<td>0.3846</td>
<td>0.7078</td>
</tr>
<tr>
<td>BART as atoms of DP</td>
<td>0.4620</td>
<td>0.4346</td>
<td>0.4867</td>
<td>0.5429</td>
</tr>
<tr>
<td>DP+P-splines</td>
<td>0.3300</td>
<td>0.3347</td>
<td>0.3295</td>
<td>0.7259</td>
</tr>
<tr>
<td>BART as atoms of DP + P-splines</td>
<td>0.3240</td>
<td>0.2970</td>
<td>0.3455</td>
<td>0.7321</td>
</tr>
<tr>
<td>BART+P-splines</td>
<td>0.3140</td>
<td>0.3138</td>
<td>0.3141</td>
<td>0.7417</td>
</tr>
<tr>
<td>Model</td>
<td>Classification Error</td>
<td>False Positives</td>
<td>False Negatives</td>
<td>AUC</td>
</tr>
<tr>
<td>----------------------------</td>
<td>----------------------</td>
<td>-----------------</td>
<td>-----------------</td>
<td>--------</td>
</tr>
<tr>
<td>RF</td>
<td>0.3200</td>
<td><strong>0.2678</strong></td>
<td>0.3678</td>
<td>0.7393</td>
</tr>
<tr>
<td>GB</td>
<td>0.3300</td>
<td>0.2971</td>
<td>0.3602</td>
<td>0.7234</td>
</tr>
<tr>
<td>BDT</td>
<td>0.3351</td>
<td>0.3461</td>
<td>0.3389</td>
<td>0.7198</td>
</tr>
<tr>
<td>DP</td>
<td>0.4920</td>
<td>0.4121</td>
<td>0.5663</td>
<td>0.5118</td>
</tr>
<tr>
<td>P-splines</td>
<td>0.3400</td>
<td>0.3975</td>
<td><strong>0.2874</strong></td>
<td>0.7102</td>
</tr>
<tr>
<td>BART</td>
<td>0.3480</td>
<td>0.3083</td>
<td>0.3846</td>
<td>0.7078</td>
</tr>
<tr>
<td>BART as atoms of DP</td>
<td>0.4620</td>
<td>0.4346</td>
<td>0.4867</td>
<td>0.5429</td>
</tr>
<tr>
<td>DP+P-splines</td>
<td>0.3300</td>
<td>0.3347</td>
<td>0.3295</td>
<td>0.7259</td>
</tr>
<tr>
<td>BART as atoms of DP + P-splines</td>
<td>0.3240</td>
<td>0.2970</td>
<td>0.3455</td>
<td>0.7321</td>
</tr>
<tr>
<td>BART+P-splines</td>
<td><strong>0.3140</strong></td>
<td>0.3138</td>
<td>0.3141</td>
<td><strong>0.7417</strong></td>
</tr>
</tbody>
</table>
ROC curves
Discussion

- The choice of the most appropriate model can be driven by knowledge of its strength. However, it is always better to fit different models and compare them on different data, in order to choose the best classifier.

- Combination of models may work better than single models if each model has a different strength compared to the others: committees can help when different learners have complementary strengths for a given task.

- Choice of 9 most predictive variables has been done by physicists by looking to marginal correlation and meaning of the variables. Most of models with 39 variables have some procedure of automatic selection of complexity and of variables, compromising between bias and variance. We used a test set to train and select complexity level of each model.