Pitfalls of Variable Importance Measures in Machine Learning
Berlin School of Economics and Law

M Loecher
Motivation

Linear Models

Machine Learning

Trees

The textbook story

Overfitting

Outlook
Motivation
Prediction versus Understanding

- Variables are seldom equally relevant
- Find ranking in “impact”
- Relative importance of regressor variables is an old topic
Data

- Sinking of the Titanic

- Kaggle’s Two Sigma Connect:

  [Two Sigma Connect renthop]

  NY Rental Listing Inquiries competition

- Swiss Fertility
Linear Models
### Titanic Survival I

**Important Predictor Variables**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sexmale</td>
<td>-0.489***</td>
<td>(0.031)</td>
</tr>
<tr>
<td>Pclass</td>
<td>-0.193***</td>
<td>(0.023)</td>
</tr>
<tr>
<td>SibSp</td>
<td>-0.052***</td>
<td>(0.017)</td>
</tr>
<tr>
<td>Parch</td>
<td>-0.013</td>
<td>(0.019)</td>
</tr>
<tr>
<td>Age</td>
<td>-0.007***</td>
<td>(0.001)</td>
</tr>
<tr>
<td>Fare</td>
<td>0.0003</td>
<td>(0.0003)</td>
</tr>
<tr>
<td>PassengerId</td>
<td>0.0001</td>
<td>(0.0001)</td>
</tr>
<tr>
<td>Constant</td>
<td>1.341***</td>
<td>(0.082)</td>
</tr>
</tbody>
</table>

- **Observations**: 714
- **$R^2$**: 0.401
- **Adjusted $R^2$**: 0.395
- **Residual Std. Error**: 0.382 (df = 706)
- **F Statistic**: 67.625*** (df = 7; 706)

**Note:** *p<0.1; **p<0.05; ***p<0.01
**Titanic Survival II**

**Correlated Features**

<table>
<thead>
<tr>
<th>Survived</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Sexmale</td>
<td>-0.511*** (0.032)</td>
</tr>
<tr>
<td>Fare</td>
<td>0.002*** (0.0003)</td>
</tr>
<tr>
<td>Age</td>
<td>-0.002 (0.001)</td>
</tr>
<tr>
<td>Constant</td>
<td>0.720*** (0.040)</td>
</tr>
</tbody>
</table>

| Observations   | 714      |
| R²             | 0.322    |
| Adjusted R²    | 0.319    |
| Residual Std. Error | 0.406 (df = 710) |
| F Statistic    | 112.383*** (df = 3; 710) |

**Note:** *p < 0.1; **p < 0.05; ***p < 0.01*
## Titanic Survival III

### Interactions

<table>
<thead>
<tr>
<th></th>
<th>Survived</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sexmale</td>
<td>$-0.826^{***}$ (0.077)</td>
</tr>
<tr>
<td>Pclass</td>
<td>$-0.244^{***}$ (0.025)</td>
</tr>
<tr>
<td>Pclass:Sexmale</td>
<td>$0.138^{***}$ (0.032)</td>
</tr>
<tr>
<td>Constant</td>
<td>$1.269^{***}$ (0.059)</td>
</tr>
</tbody>
</table>

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Observations</td>
<td>891</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.381</td>
</tr>
<tr>
<td>Adjusted $R^2$</td>
<td>0.378</td>
</tr>
<tr>
<td>Residual Std. Error</td>
<td>0.384 (df = 887)</td>
</tr>
<tr>
<td>F Statistic</td>
<td>$181.661^{***}$ (df = 3; 887)</td>
</tr>
</tbody>
</table>

*Note:* $^*p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$
Interest in new rental on RentHop

- bathrooms: number of bathrooms
- bedrooms: number of bathrooms
- latitude
- longitude
- **price**: in USD
- **interest_level**: ’high’, ’medium’, ’low’
- street_address
- photos: a list of photo links.
- building_id

- created
- description
- display_address
- features: a list of features about this apartment
NY rent data set 1

“Location, Location, Location, ..?”

<table>
<thead>
<tr>
<th>Dependent variable:</th>
<th>price</th>
</tr>
</thead>
<tbody>
<tr>
<td>latitude</td>
<td>$-6,442.635$ (6,600.829)</td>
</tr>
<tr>
<td>longitude</td>
<td>$-3,559.672$ (3,638.745)</td>
</tr>
<tr>
<td>bathrooms</td>
<td>$1,994.054^*$ (1,060.910)</td>
</tr>
<tr>
<td>bedrooms</td>
<td>$677.732$ (481.168)</td>
</tr>
<tr>
<td>Constant</td>
<td>$-128.134$ (20,096.990)</td>
</tr>
</tbody>
</table>

Observations: 10,000
R$^2$: 0.001
Adjusted R$^2$: 0.001
Residual Std. Error: 44,902.160 (df = 9995)
F Statistic: $3.173^{**}$ (df = 4; 9995)

Note: *p<0.1; **p<0.05; ***p<0.01
NY rent data set II

<table>
<thead>
<tr>
<th>Dependent variable:</th>
<th>interest_level</th>
</tr>
</thead>
<tbody>
<tr>
<td>bathrooms</td>
<td>$-0.178^{***}$ (0.015)</td>
</tr>
<tr>
<td>latitude</td>
<td>$-0.083$ (0.091)</td>
</tr>
<tr>
<td>longitude</td>
<td>$-0.051$ (0.050)</td>
</tr>
<tr>
<td>bedrooms</td>
<td>$0.048^{***}$ (0.007)</td>
</tr>
<tr>
<td>price</td>
<td>$-0.00000$ (0.00000)</td>
</tr>
<tr>
<td>Constant</td>
<td>$1.146^{***}$ (0.278)</td>
</tr>
</tbody>
</table>

Observations 10,000  
$R^2$ 0.015  
Adjusted $R^2$ 0.015  
Residual Std. Error 0.620 (df = 9994)  
F Statistic $30.774^{***}$ (df = 5; 9994)

Note: *p<0.1; **p<0.05; ***p<0.01
"Variable importance is not very well defined as a concept. There is no theoretically defined variable importance metric..."

- Change in $R^2$ when the variable is added to the model last
- Average order-dependent $R^2$ allocations over all $p!$ orderings (LMG)

+ Direction/sign of contribution
+ Uncertainty "for free"
+ Easy to understand !?
  - Marginal versus conditional
  - Confounding effects
  - Slave to linearity
  - Interactions must be coded apriori
Machine Learning
Which machine learning algorithms?

Data-driven advice for applying machine learning to bioinformatics problems

Randal S. Olson†*, William La Cava†*, Zairah Mustahsan, Akshay Varik, and Jason H. Moore†

Institute for Biomedical Informatics, University of Pennsylvania
Philadelphia, PA 19104, USA
†E-mails: rso@ralson.com, lacava@upenn.edu and jhmoore@upenn.edu

As the bioinformatics field grows, it must keep pace not only with new data but with new algorithms. Here we contribute a thorough analysis of 13 state-of-the-art, commonly used machine learning algorithms on a set of 165 publicly available classification problems in order to provide data-driven algorithm recommendations to current researchers. We present a number of statistical and visual comparisons of algorithm performance and quantify the effect of model selection and algorithm tuning for each algorithm and dataset. The analysis culminates in the recommendation of five algorithms with hyperparameters that maximize classifier performance across the tested problems, as well as general guidelines for applying machine learning to supervised classification problems.
Boosting and Random Forests

![Bar chart showing mean ranking for various models](image-url)
<table>
<thead>
<tr>
<th>Wins</th>
<th>GTB</th>
<th>RF</th>
<th>SVM</th>
<th>ERF</th>
<th>SGD</th>
<th>KNN</th>
<th>DT</th>
<th>AB</th>
<th>LR</th>
<th>PA</th>
<th>BNB</th>
<th>GNB</th>
<th>MNB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gradient Tree Boosting</td>
<td>32%</td>
<td>45%</td>
<td>38%</td>
<td>67%</td>
<td>72%</td>
<td>78%</td>
<td>76%</td>
<td>82%</td>
<td>90%</td>
<td>95%</td>
<td>95%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Random Forest</td>
<td>9%</td>
<td>33%</td>
<td>23%</td>
<td>62%</td>
<td>65%</td>
<td>71%</td>
<td>69%</td>
<td>71%</td>
<td>76%</td>
<td>85%</td>
<td>95%</td>
<td>90%</td>
<td></td>
</tr>
<tr>
<td>Support Vector Machine</td>
<td>12%</td>
<td>21%</td>
<td>25%</td>
<td>55%</td>
<td>65%</td>
<td>56%</td>
<td>62%</td>
<td>67%</td>
<td>74%</td>
<td>79%</td>
<td>95%</td>
<td>93%</td>
<td></td>
</tr>
<tr>
<td>Extra Random Forest</td>
<td>8%</td>
<td>14%</td>
<td>30%</td>
<td>58%</td>
<td>63%</td>
<td>61%</td>
<td>64%</td>
<td>67%</td>
<td>70%</td>
<td>81%</td>
<td>93%</td>
<td>91%</td>
<td></td>
</tr>
<tr>
<td>Linear Model trained via Stochastic Gradient Descent</td>
<td>8%</td>
<td>16%</td>
<td>9%</td>
<td>15%</td>
<td>38%</td>
<td>41%</td>
<td>44%</td>
<td>41%</td>
<td>61%</td>
<td>66%</td>
<td>89%</td>
<td>87%</td>
<td></td>
</tr>
<tr>
<td>K-Nearest Neighbors</td>
<td>4%</td>
<td>8%</td>
<td>7%</td>
<td>8%</td>
<td>35%</td>
<td>42%</td>
<td>45%</td>
<td>52%</td>
<td>53%</td>
<td>70%</td>
<td>88%</td>
<td>85%</td>
<td></td>
</tr>
<tr>
<td>Decision Tree</td>
<td>2%</td>
<td>2%</td>
<td>20%</td>
<td>8%</td>
<td>42%</td>
<td>38%</td>
<td>43%</td>
<td>48%</td>
<td>57%</td>
<td>69%</td>
<td>80%</td>
<td>82%</td>
<td></td>
</tr>
<tr>
<td>AdaBoost</td>
<td>1%</td>
<td>7%</td>
<td>10%</td>
<td>15%</td>
<td>30%</td>
<td>35%</td>
<td>32%</td>
<td>39%</td>
<td>47%</td>
<td>59%</td>
<td>76%</td>
<td>77%</td>
<td></td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>5%</td>
<td>10%</td>
<td>3%</td>
<td>8%</td>
<td>11%</td>
<td>31%</td>
<td>33%</td>
<td>35%</td>
<td>37%</td>
<td>54%</td>
<td>79%</td>
<td>81%</td>
<td></td>
</tr>
<tr>
<td>Passive Aggressive</td>
<td>2%</td>
<td>6%</td>
<td>1%</td>
<td>5%</td>
<td>0%</td>
<td>18%</td>
<td>28%</td>
<td>28%</td>
<td>13%</td>
<td>50%</td>
<td>81%</td>
<td>79%</td>
<td></td>
</tr>
<tr>
<td>Bernoulli Naive Bayes</td>
<td>0%</td>
<td>2%</td>
<td>2%</td>
<td>4%</td>
<td>10%</td>
<td>13%</td>
<td>18%</td>
<td>15%</td>
<td>22%</td>
<td>25%</td>
<td>62%</td>
<td>68%</td>
<td></td>
</tr>
<tr>
<td>Gaussian Naive Bayes</td>
<td>0%</td>
<td>1%</td>
<td>3%</td>
<td>2%</td>
<td>6%</td>
<td>6%</td>
<td>11%</td>
<td>12%</td>
<td>9%</td>
<td>10%</td>
<td>22%</td>
<td>45%</td>
<td></td>
</tr>
<tr>
<td>Multinomial Naive Bayes</td>
<td>1%</td>
<td>1%</td>
<td>2%</td>
<td>2%</td>
<td>2%</td>
<td>5%</td>
<td>10%</td>
<td>14%</td>
<td>4%</td>
<td>5%</td>
<td>13%</td>
<td>39%</td>
<td></td>
</tr>
</tbody>
</table>
Trees
Trees details

- **Greedy**: At each split we minimize squared error or node impurity.
- **All Interactions**: data “thin out” exponentially fast.
- **Piecewise Constant**: no smoothness, inferior for regression.
- **Model complexity**: depth of tree, typically single pruned trees.
- **Boosting**: many shallow trees sequentially minimize loss.
- **Random Forests**: many deep trees grown in parallel on bootstrapped samples. **Column sampling** leads to additional parameter.
Column Subsampling

![Box plots showing test misclassification error for random forest and gradient boosting models with varying numbers of relevant and noise variables.](image)
Swiss Fertility
LR versus Forests

Figure 3. Main effects plot for the linear model (top, with 95% bands) and RF-CART ($mtry = 1$ (black) and $mtry = 2$ (gray)). Rugs at the bottom represent individual data values for the 182 Swiss provinces.
Variable Importance I

- **gini importance**: the mean decrease in impurity of a feature is computed by measuring how effective the feature is at reducing uncertainty (classifiers) or variance (regressors).

- For a single decision tree $T$:

  $$I^2_i(T) = \sum_{t=1}^{J-1} \hat{i}_t^2(v(t) = l)$$

  as a measure of relevance for each predictor variable $X_i$. The sum is over the $J - 1$ internal nodes of the tree.

- For ensembles it is simply averaged over the trees

  $$GiniImp^2 = \frac{1}{M} \sum_{m=1}^{M} I^2_i(T_m)$$
The textbook story
**FIGURE 10.14.** Relative importance of the predictors for the California housing data.
FIGURE 10.17. Partial dependence of median house value on location in California. One unit is $100,000, at 1990 prices, and the values plotted are relative to the overall median of $180,000.
Overfitting
random forests are averages of large numbers of individually
grown regression/classification trees.

both “row and column subsampling”: each tree is based on a
random subset of the observations, and each split is based on a
random subset of $mtry$ candidate variables.

The tuning parameter $mtry$ can have profound effects on
prediction quality as well as the variable importance measures
outlined below.
Due to the CART bootstrap row sampling, 36.8% of the observations are (on average) not used for an individual tree; those “out of bag” (OOB) samples can serve as a validation set to estimate the test error, e.g.:

\[
E \left( Y - \hat{Y} \right)^2 \approx OOB_{MSE} = \frac{1}{n} \sum_{i=1}^{n} \left( y_i - \bar{\hat{y}}_{i,OOB} \right)^2
\]  

(1)

where \( \bar{\hat{y}}_{i,OOB} \) is the average prediction for the \( i \)th observation from those trees for which this observation was OOB.
The default method to compute variable importance is the *mean decrease in impurity* (or *gini importance*) mechanism: At each split in each tree, the improvement in the split-criterion is the importance measure attributed to the splitting variable, and is accumulated over all the trees in the forest separately for each variable. Note that this measure is quite like the $R^2$ in regression on the training set.

The widely used alternative *reduction in MSE when permuting a variable* as a measure of variable importance or short *permutation importance* is defined as follows:

$$VI = OOB_{MSE,perm} - OOB_{MSE}$$ (2)
Gini importance can be highly misleading
Let us go one step further and add a Gaussian noise feature, which we call **PassengerWeight**:
Categorical Features

Coding passenger ID as factor makes matters worse:

Variable Importance: DRF

- PassengerId
- Sex
- Pclass
- Fare
- Age
- PassengerWeight
- SibSp
- Parch

h2o: PassengerId factor coding
NY Rent, Gini Importance

Random Forest **regressor** predicting apartment rental price from 4 features + a column of random numbers. Random column is last, as we would expect but the importance of the number of bathrooms for predicting price is highly suspicious.

Random Forest **classifier** predicting apartment interest level (low, medium, high) using 5 features + a column of random numbers. Highly suspicious that random column is much more important than the number of bedrooms.
permuting each column and computing change in out-of-bag $R^2$. 

permuting each column and computing change in out-of-bag accuracy.
Collinear features

drop column importance

permutation importance

drop column importance with noise
Why Random Forests

- No pruning
- Column Subsampling
- Bootstrap

Distribution of the p values of $\chi^2$ tests of each categorical variable $X_2, \ldots, X_5$ and the binary response for the null case simulation study, where none of the predictor variables is informative.
Outlook
Summary/Recommendations

▶ RF default importance not reliable
▶ use permutation importance for all models
▶ Boosting or extremely randomized trees for VI
▶ Careful about conditional versus marginal importance
▶ Social Sciences, Bioinformatics and Economics