Quantifying Uncertainty in Machine Learning

Giles Hooker
Conference on Applied Statistics in Ireland
May 15, 2023

Outline

- Machine Learning and Uncertainty
- Predictive uncertainty:
  - Classification and Calibration
  - Regression and prediction intervals
  - Conformal Prediction
- Model uncertainty
  - Definitions
  - Bootstrapping
  - Ensembles and the Infinitesimal Jackknife
- Model interpretation and comparisons between models.
- Testing Variable Importance
- ML as Plug-ins to Statistical Methods

Attempt to provide (conceptually) simple recipes for all.

Uncertainty in Machine Learning

Uncertainty quantification approached from many fields. Often divided into

Aleatoric: due to intrinsic (irreducible) randomness in nature

Epistemic: associated with lack of knowledge (e.g. about model form) or measurement ability.

Do not perfectly translate into statistical concepts used here:

Predictive uncertainty: distribution of possible outcomes given this prediction.

Inferential uncertainty: how stable is this prediction with respect to

- Examples provided in the training data
- Random numbers used in training process.

- Predictive uncertainty easiest, computationally cheap.
- Inference is harder, usually requires more computing.

A Little Notation

- Assume that we have a data set of n *observations*:

$$D = \{(X_i, Y_i)\}_{i=1}^{n}$$

  (also examples, realizations, ...)
- Data set is used to estimate (learn, train,...) a *prediction function* $f(x_1, \ldots, x_p)$
- Use: X_i = row of data set or specific new value, x = placeholder argument in $f(x)$.

Describe

- Distribution of Y using $f(x)$
- Stability of $f(x)$ with respect to D.

Beijing Housing Data

Used for illustrations, predict log(totalPrice), or totalPrice > 325 from

- Lat, Lng
- Days On Market
- online followers
- square m
- Number of
  - livingRoom
  - drawingRoom
  - kitchen
  - bathRoom
- Building Type
- Construction Date
- Renovation
- Building structure
- Ladder Ratio (resident to elevator capacity)
- Elevator
- Ownership > 5 years
- Subway access
- District

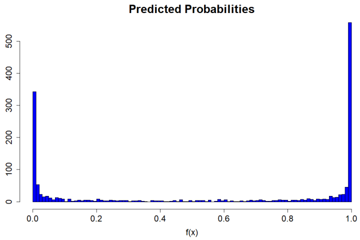
Use `ranger` and `gbm` in `R` as tools.

Prediction I: Classification and Calibration

- Given features X assign label $Y \in \{l_1, \ldots, l_k\}$, often $Y \in \{0, 1\}$.
- Usually target proportion incorrect (or average costs)
- Uncertainty *probability* $p_k(X) = P(Y = l_k | X)$
  - $p_k(X) \notin \{0, 1\}$ not always obvious.
  - Binary $\Rightarrow$ one number summary.
- Most learners threshold continuous output: $\hat{Y} = I(f(X) > c)$. (Output layer in NN, proportion votes in RF/boosting, margin in SVM).
- *Calibration*: transform $f(x) \to p(x)$ rather than threshold.
  - often useful even if $f(x) \in [0, 1]$
  - ML using misclassification $\Rightarrow$ overconfident predictions.
  - Best conducted with hold-out data

Illustration

gbm with Adaboost loss

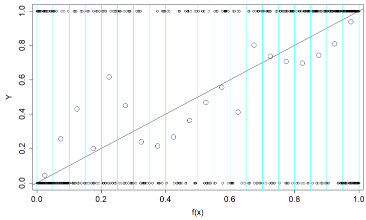


Highly overconfident!

Calibration: predicted probability corresponds to outcome proportion.

Assessing calibration:

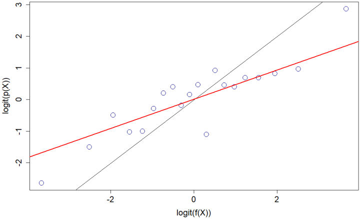
- Divide into bins based on $f(x)$
- Obtain $P(Y = 1)$ in each bin



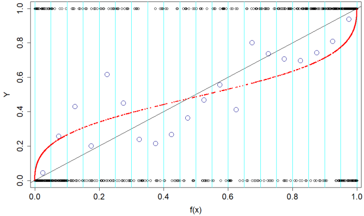
Calibration Methods

Calibrate by re-scaling $f(X)$ to give accurate probability measurements.

Predict Y from $\text{logit}(f(X))$



$$P(Y=1) = \frac{\exp(a + bl(X))}{1 + exp(a + bl(X))}$$

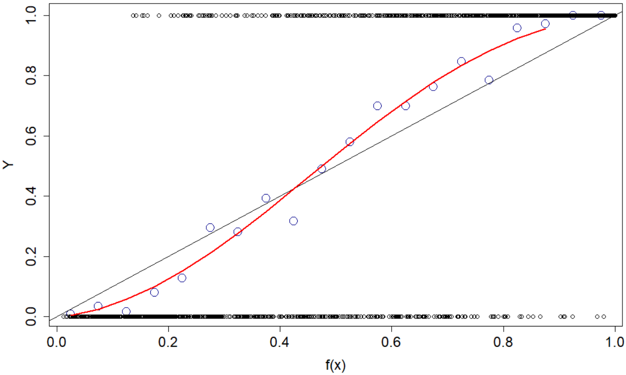


by maximizing likelihood (logistic regression)

a.k.a. Platt Scaling

Models Can Be Underconfident, Too

Random forests over-smooth $\rightarrow$ don't get as close to $\{0, 1\}$.

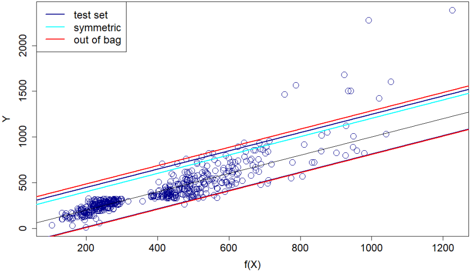


Many options

- Use bins for probabilities/regress probability on midpoints
- Pleiss et. al. 2017: simple transformations of $f(x)$ usually sufficient.

Prediction II: Regression and Prediction Intervals

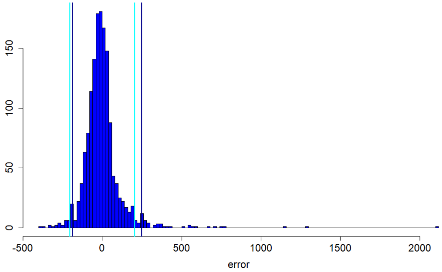
- Predicting $Y \in \mathbb{R}$ perfectly less believable than in classification.
- Usually provide a range $Y \in [a(X), b(X)]$ with target that Y in forecast range 95% of time.
- Often symmetric about prediction: $a(X) = f(X) - s(X)$, $b(X) = f(X) + s(X)$.



Ranges from Hold-Out Sets

Given data, (X_i, Y_i) *not* used to train f,

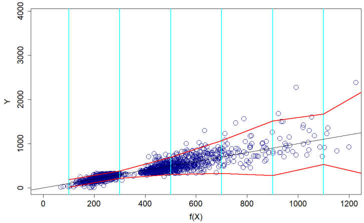
- Form residuals (errors) $e_i = Y_i - f(X_i)$
- Set $a(X), b(X)$ to be 0.025 and 0.975 quantiles (no change with X)
- Or $s(X) = 0.95$ quantile of $|e_i|$



Heteroskedasticity and Modeling Variance

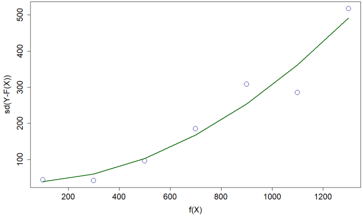
Sometimes interval width should change over X, or $f(X)$

Calculate sd or quantiles within bins defined by $f(x)$.



Intervals: $f(x) \pm 2\text{sd}(x)$

Model bin sd by midpoint

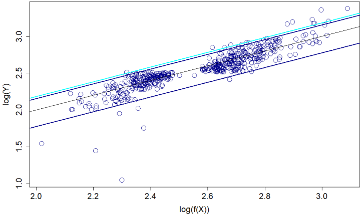


$\text{sd}(x) = g(f(x))$ results in

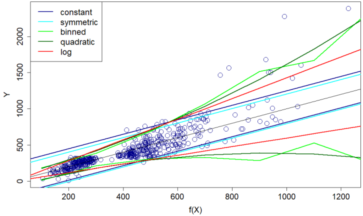
$$f(x) \pm 2g(f(x))$$

Variance Stabilizing Transformations

- $\tilde{e}_i = \log(Y) - \log(f(X))$
- Or train to predict $\log(Y)$



Transform back to
$[f(x)e^{\tilde{q}_{0.025}},\ f(x)e^{\tilde{q}_{0.975}}]$



Also $\sqrt{Y}$, $1/Y$, ...

$Z = \log(Y)$ reduces extreme values $\Rightarrow$ predicting Z improves test error.

More Sophisticated Alternatives

- Conditional Density Estimation
  - Making distribution parameters model-dependent: $Y \sim g(y; \theta(X))$
  - Conditional kernel density estimation using $(Y, F(X))$
  - ...
- Quantile regression
  - Train model to minimize quantile loss for the τ-quantile:

$$L_\tau(Y, a(X)) = \tau(Y - a(X))_+ + (1 - \tau)(Y - a(X))_-$$

  - based on original X
  - based on values of $f(X)$.

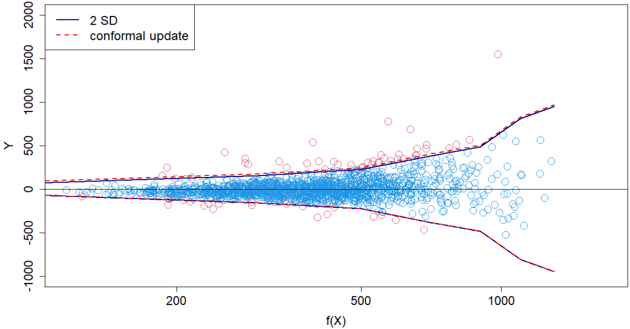
Conformal Prediction

- Provides $C(X)$ such that $P(Y \in C(X)) > 1 - \alpha$ with minimal assumptions.
- Based on *exchangeability*: on held out sample $e_i = Y_i - F(X_i)$ have same *marginal* distribution.
  - $[f(X) + q^e_{0.05}, f(X) + q^e_{0.95}]$ has 0.9 probability of covering future Y.
  - Probability: over both test set, and future X that come from test-set distribution.
- Can be used to correct UQ + obtain marginal finite sample guarantees.
- Eg. given $[a(X), b(X)]$, measure

$$r_i = \max(a(X_i) - Y_i, Y_i - b(X_i))$$

and correct to $[a(X) - q^r_{0.975}, b(X) + q^r_{0.975}]$.
- Can also apply to level of a density, risk sets, ...

Conformal Prediction



Original Conformal Prediction (Vovk and Vapnik 1998) based on:

- Retrain $f_y(X)$ with additional data (X, y).
- Find range of y so that $y - f_y(X)$ in central 95% of training residuals.

But: simple approaches (Y regressed on $f(X)$ using hold-out data) usually sufficient in practice.

Model Uncertainty

How stable is the prediction $f(X)$?

How different might $f(X)$ be

1. If we re-run the learning algorithm?
2. If we used new data?

Usually summarized by *Confidence Intervals*

- Interval includes *average* prediction 95% of time.

Reproduction intervals often more relevant

- Interval includes *re-estimated* $f(x)$ 95% of time.

Bias in $f(x) \Rightarrow$ CI need not cover "truth", but RI gives reliability.

- Normal theory:

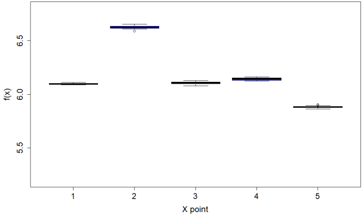
$$\text{CI} = [f(x) \pm 2sd(f(x))]$$
$$\text{RI} = [f(x) \pm 2\sqrt{2}sd(f(x))]$$

- Challenge: how to obtain $sd(f(x))$.

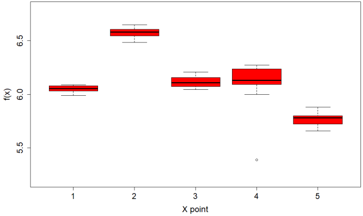
Monte Carlo versus Sample Variability

Break Housing data into 10 sets, each with 3,000 training, 1,500 test points.

Re-running random forests 20 times each for first test set.

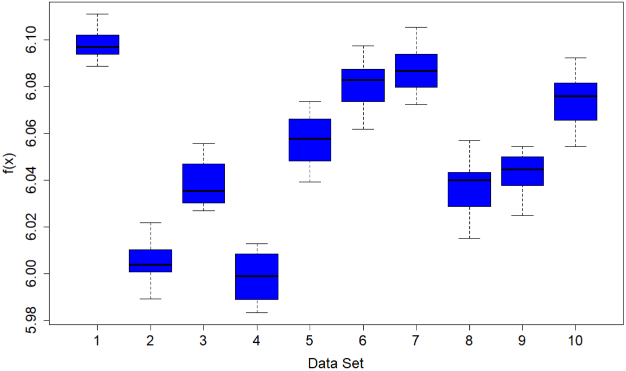


Running on each of 10 independent data sets:



Monte Carlo versus Sample Variability

Re-run 20 times for each of 10 data sets, predict at one data point.



Sample variance 20 times Monte Carlo

Bootstrapping

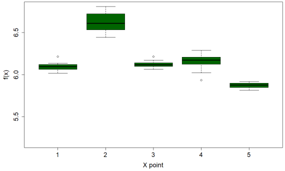
"Simulate" new data. For $b \in 1, \ldots, B$:

1. Resample training data with replacement
2. Fit $f^b(x)$ on resampled data

use collection of *bootstrapped* $f^b(x)$ for inference:

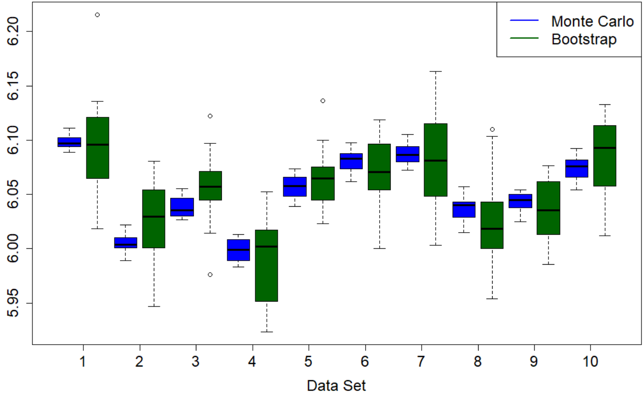
- $sd(f(x))$ from standard deviation among bootstraps ($B \approx 50$)
- CI based on quantiles of $f^b(X) - f(x)$ ($B \approx 500$)
  - *subtract and reverse* quantiles: $[f(x) - q_{0.975}(x), f(x) - q_{0.025}(x)]$

Theoretical guarantees for bootstrapping may not hold in ML.



Bootstrap and Monte Carlo Variance

20 Bootstraps per independent data set, 20 re-runs with same data.

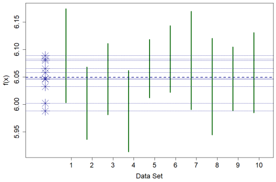


Bootstrap Confidence Intervals

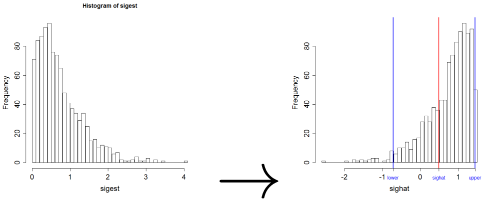
Simple intervals are
$f(x) \pm 2s(x)$

- $s(x)$ = between bootstrap standard deviation

Most commonly used in ML.



Asymmetric intervals *reversed* in order to cover mean.



Random Forests and the Infinitesimal Jacknife

Bagging-type ensemble methods have bootstrapping built in.

Random Forests:

1. Train trees $T_b(x)$ on bootstrap data sets
2. Report average prediction: $f(x) = \frac{1}{B} \sum T_b(x)$

Do we want to bootstrap a bootstrap? And wouldn't we then average more? Bootstrap is about $T(x)$, *not* $\frac{1}{b} \sum f^b(x)$.

Infinitesimal Jackknife (IJ) gives variance calculation for average *without* bootstrapping:

$$\mathrm{var}(f(x)) \approx \sum_{i=1}^{n} \mathrm{cov}(\mathrm{T}(x), N_i)^2$$

- $\mathrm{T}(x)$ = vector of values from trees.
- N_i = number of times obs i replicated in each bootstrap.

Infinitesimal Jacknife

IJ defined for a large class of models:

$$U_i = \lim_{\epsilon \to 0} \frac{\hat{\theta}((1-\epsilon)\hat{P} + \epsilon \delta_{X_i}) - \hat{\theta}(\hat{P})}{\epsilon}, \quad \hat{V}_{ij} = \sum U_i^2$$

allows a covariance estimate between $\hat{\theta}^1$ and $\hat{\theta}^2$

$$\widehat{\text{cov}}(\hat{\theta}^1, \hat{\theta}^2) = \sum U_i^1 U_i^2$$

- Results in familiar estimates for M-estimators:

$$\hat{\theta} = \text{argmax} \sum M(Z_i, \theta) \Rightarrow U_i = \left[\sum d^2 M_i / d\theta^2\right]^{-1} dM_i / d\theta$$

- For ensembles, derivative is given by *inclusion probability* weights

$$F(x) = \sum_{S \subset 1,\ldots,n} \left(\prod_{i \in S} w_i\right) f(x; Z_S) \to U_i = \text{cov}(f, N_i)$$

Bias Corrections for IJ Variance Estimates

In IJ formula

$$\mathrm{var}(f(x)) \approx \sum_{i=1}^{n} \mathrm{cov}(\mathrm{T}(x), N_i)^2$$

$\mathrm{cov}(\mathrm{T}(x), N_i)$ taken over infinite trees.

Over-estimate variance for common B.

Bias in IJ estimated from

$$\frac{k}{B}\mathrm{var}(\mathrm{T}(x))$$

(k = subsample size; more accurate formulae are trivially different)

$$\mathrm{var}(f(x)) \approx \sum_{i=1}^{n} \mathrm{cov}(\mathrm{T}(x), N_i)^2 - \frac{k}{B}\mathrm{var}(\mathrm{T}(x))$$

For small B this can be negative!

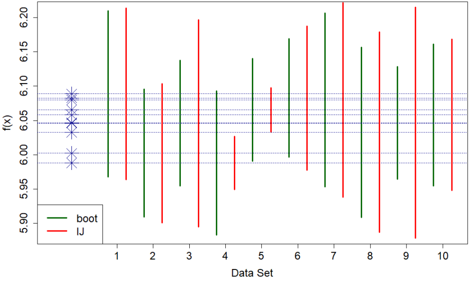
Theory applies to *subsamples with replacement* sized $k < n$, rather than bootstrap samples. We use $k = n/2$.

Infinitesimal Jacknife Reproduction Intervals

10 data sets lets us get an idea of coverage

For each data set

1. Fit random forest $f(x)$ + predict at new point (blue *)
2. Produce IJ variance estimate + calculate RI (red)
3. 20 bootstraps to estimate variance + calculate RI (green)

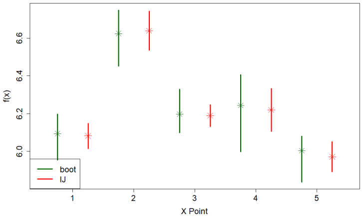


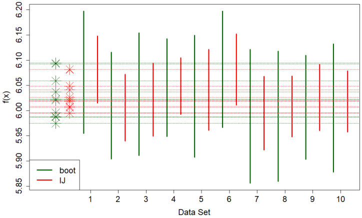
Note: 5,000 trees used in this forest.

If You Were Going to Bootstrap Anyway....

- IJ variance calculations only use ensemble structure.
- Bootstrap for UQ
- **Or** use $f(x) = \frac{1}{B}\sum f^b(x)$ and IJ
- Applies to *any* learner (if $B \approx 1000$)

IJ and ensemble of 500 boosted tree models:





IJ gives more variable CI's, but corresponds to more stable ensemble estimates.

What About Bayes?

Alternative notion of inference:

- Model data: $P(Y|f(X))$ and prior belief $P(f(x))$
- Use conditional distribution $P(f(x)|D)$ to describe uncertainty.
- Credible intervals from posterior distributions.

Does *not* ask *What if our data were (believably) different?*

Few Bayesian ML methods:

- Gaussian Process models
- **B**ayesian **A**dditive **R**egression **Trees**
- Some neural network methods (usually approx Bayesian)

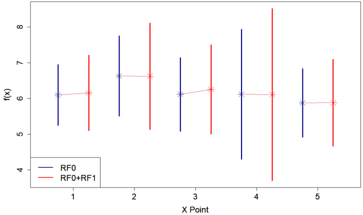
Tend to be computationally expensive, but in-built UQ.

Extensions: Combinations of Ensembles

Random forests can be boosted:

1. Fit $f_0(x)$ to (X_i, Y_i)
2. Fit $f_1(x)$ to $(X_i, Y_i - f_0(X_i))$
3. Predict $f_0(x) + f_1(x)$

20% test error reduction



$$\mathrm{var}(f_0(x) + f_1(x)) = \mathrm{var}(f_0(x)) + \mathrm{var}(f_1(x)) + 2\mathrm{cov}(f_0(x), f_1(x))$$

and

$$\mathrm{cov}(f_0(x), f_1(x)) \approx \sum_{i=1}^{n} \mathrm{cov}(\mathrm{T}^0(x), N_i^0)\mathrm{cov}(\mathrm{T}^1(x), N_i^1)$$

(or ensure *same* subsamples for f_0 and f_1)

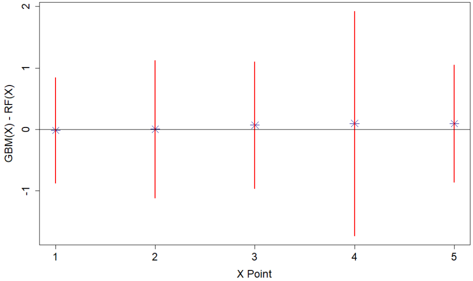
Comparisons of Ensembles

Do different learners give us different answers?

$$\text{var}(f_0(x) - f_1(x)) = \text{var}(f_0(x)) + \text{var}(f_1(x)) - 2\text{cov}(f_0(x), f_1(x))$$

for example

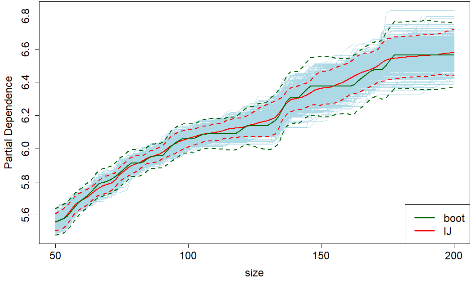
- Models learned with/without features
- Different ensemble methods



Interpretation and Inherited Variability

Bootstrap/IJ recipe

- Explanation/summary on each member of bootstrapped ensemble.
- Bootstrap intervals for original summary, or IJ for average of bootstraps.



- Alternative summary $= g(f(X_1), \ldots, f(X_k))$ + inherit variability in f
- More complex, $k \times k$ *covariance* requires much larger B.

Testing Variable Importance

Hypothesis $H_0 = E(Y|X) = f(X_{-1})$, or $Y \perp X_1 | X_{-1}$

Conditional Randomization Test:

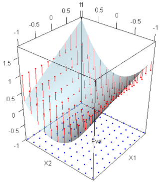
- Simulate $\tilde{X}_1 \sim X_1 | X_{-1}$
- Re-fit F on $(\tilde{X}_1, X_{-1}, Y)$
- Measure change in F/accuracy/other statistic
- Compare to original data.

Knockoffs augment data to $(\tilde{X}, X, Y)$ to run all tests at once (under more conditions).

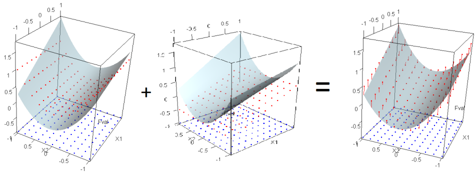
NOTE: simply training on (X_{-1}, Y) changes statistical properties of F.

Testing Variable Importance II

- Examine $D = F(\tilde{X}_1, X_{-1}) - F(X_1, X_{-1})$ over feature space.
- Often in the form of a grid:
- Form χ^2 statistic: $\hat{\boldsymbol{D}}^T \hat{\Sigma}_D^{-1} \hat{\boldsymbol{D}} \sim \chi^2_N$

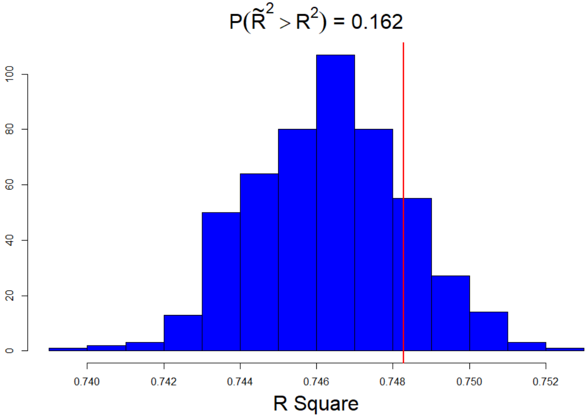


Extensible to further tests for structure
$F(X_1, X_2, X_3) = G_{-2}(X_1, X_3) + G_{-1}(X_2, X_3)$.



Testing the Effect of Renovation

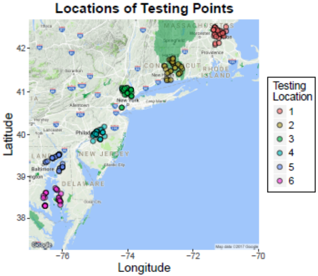
Renovation condition has four levels.



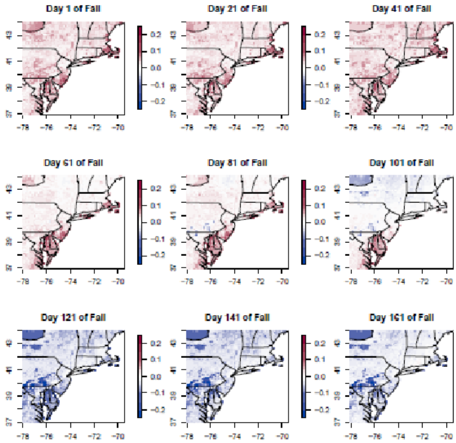
Conditional Randomization Test: predict and re-simulate as multinomial; test statistic is out-of-bag R^2 from `ranger`.

Effect of Changes in Maximum Temperature

Coleman et. al. 2018



Testing Location	Test Statistic	P-Value
1	84.51	4.346E-07
2	98.57	3.119E-09
3	102.93	6.373E-10
4	160.10	0.000E+00
5	180.11	0.000E+00
6	89.59	7.585E-08



ML Plug-Ins for Causal Inference

Double Machine Learning; "treatment" D in model

$$Y = \theta D + G(X) + \epsilon$$
$$D = M(X) + \eta$$

so that

$$EY|X = \theta M(X) + G(X)$$

and estimating θ from Y, $EY|X$ and D is biased.

However

$$(Y - E(Y|X)) = (D - E(D|X))\theta + \epsilon = \eta\theta + \epsilon$$

suggests

- Regress Y on X to get $\hat{F}(X)$ and residual $\hat{\epsilon}$
- Regress D on X to get $\hat{M}(X)$ and residual $\hat{\eta}$
- Regress $\hat{\epsilon}$ on $\hat{\eta}$ to estimate θ.

Sample Splitting and Double Robustness

Why do things this way?

$$\hat{\theta} = \frac{\sum \hat{\epsilon}_i \hat{\eta}_i}{\hat{\eta}_i^2} = \frac{(\eta_i \theta + \epsilon_i + \mathrm{err}(\hat{F}(X_i)))(\eta_i + \mathrm{err}(\hat{M}(X_i)))}{\hat{\eta}_i^2}$$

$$= \frac{1}{\sum \hat{\eta}_i^2} \Big[\sum \eta_i^2 \theta + \epsilon_i \eta_i + \epsilon_i \mathrm{err}(\hat{M}(X_i)) + \eta_i \mathrm{err}(\hat{F}(X)_i)) + \mathrm{err}(\hat{F}(X_i))\mathrm{err}(\hat{M}(X_i)) \Big]$$

If θ estimated independently of $\hat{F}$, $\hat{M}$ then

- $\epsilon_i \mathrm{err}(\hat{F}(X_i))$, $\eta_i \mathrm{err}(\hat{M}(X_i))$ are mean zero.
- $\mathrm{err}(\hat{F}(X_i))\mathrm{err}(\hat{M}(X_i)) \to 0$ if at least one does.

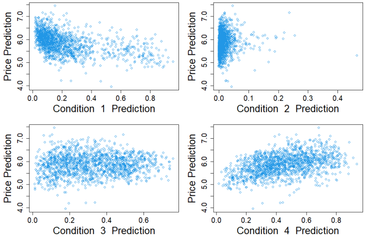
and Uncertainty Quantification comes from regression for $\hat{\theta}$.

Effect of Renovation

Weak relationship between predicted price and predicted renovation condition.

Weak relationship but may produce some correlation.

Multivariate regression of errors on errors:




```
           Estimate Std. Error Pr(>|t|)
Condition1 0.001623   0.009203 0.860
Condition2 0.022005   0.055402 0.691
Condition3 0.056309   0.023567 0.017
Condition4 0.095079   0.022607 2.76e-05

M R-sq:  0.01243,   Adj R-sq:  0.00979
F: 4.708 on 4 and 1496 DF,  p: 0.00089
```

A Little More Generally

Use ML to account for nuisance regression parameters (call this $F(X)$).

But must be able to account for smoothing biases in estimating $\hat{F}$.

1. Double ML: use an appropriate score equation

$$\psi(\theta, F) = 0$$

   such that $\nabla_F E\psi(\theta, F)(\hat{F} - F) = 0$, eg via sample splitting.

2. Targeted ML: $\theta = G(P)$ for $P =$ data distribution
   - Define Efficient Influence Function:

$$g(z) = \frac{d}{d\epsilon} G((1-\epsilon)P + \epsilon\delta_z)$$

   - Maximize the likelihood of $(1-\eta)\hat{P} + \eta g$
   - Modify $\hat{P}$, iterate until $\hat{\eta} = 0$.

Both require tailoring to specific situations, but provide a natural means to plug ML into stats.

Conclusions and Best Practices

Predictive uncertainty

- treat $f(x)$ as covariate
- apply appropriate statistical model
- simple models of Y on $f(X)$ usually works fine (on test set)

Model uncertainty:

- Bootstrap a generic method, lot of effort just for SD.
- Ensembles $\Rightarrow$ infinitesimal jackknife avoids bootstrapping a bootstrap
- Still requires more work than optimal for prediction
- Does allow comparison/combinations of models, UQ for explanations.

Formalized inference:

- Tests of structure of $f(X)$ dependent on retraining or high dimensional covariances.
- Sample splitting does allow for pre-estimated ML models as plug-in to inference. But conditions may apply.

Brief Bibliography and Resources

On Platt Scaling:

Guo, Chuan, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. "On calibration of modern neural networks." International Conference on Machine Learning, pp. 1321-1330. PMLR, 2017.

https://www.analyticsvidhya.com/blog/2016/07/platt-scaling-isotonic-regression-minimize-logloss-error/

Conformal Inference:

Shafer, Glenn, and Vladimir Vovk. "A tutorial on conformal prediction." Journal of Machine Learning Research 9, no. 3 (2008).

Lei, Jing, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. “Distribution-free predictive inference for regression.” Journal of the American Statistical Association 113, no. 523 (2018): 1094-1111.

Angelopoulos, A. N., & Bates, S. (2021). “A gentle introduction to conformal prediction and distribution-free uncertainty quantification.” arXiv preprint arXiv:2107.07511.

Bootstrap:

Efron, Bradley, and Robert J. Tibshirani. An introduction to the bootstrap. CRC press, 1994.

Kohavi, Ron. "A study of cross-validation and bootstrap for accuracy estimation and model selection." Ijcai. Vol. 14. No. 2. 1995.

Random Forest/Infinitesimal Jackknife Papers

Variance Estimation Methods:

Sexton, Joseph, and Petter Laake. "Standard errors for bagged and random forest estimators." Computational Statistics & Data Analysis 53.3 (2009): 801-811.

Stefan Wager, 2014, "Confidence Intervals for Random Forests; The Jackknife and the Infinitesimal Jackknife", *Journal of Machine Learning Research*.

Zhou, Zhengze, Lucas Mentch and Giles Hooker, 2021, "V-Statistics and Variance Estimation" *Journal of Machine Learning Research*, 22(287):1-48.

Ghosal, Indrayudh and Giles Hooker, 2021, "Boosting Random Forests to Reduce Bias; One-Step Boosted Forest and its Variance Estimate", JCGS, 30(2):493-502

Additional References:

Mentch, Lucas and Giles Hooker, 2016, "Quantifying Uncertainty in Random Forests via Confidence Intervals and Hypothesis Tests", *Journal of Machine Learning Research*, 17(3):1-41

Stefan Wager and Susan Athey, 2018, "Estimation and Inference of Hererogeneous Treatment Effects using Random Forests", *Journal of the American Statistical Association*.

Coleman, T., L. Mentch, D. Fink, F. La Sort, G. Hooker, W. Hochachka and D. Winkler, 2020, "Statistical Inference on Tree Swallow Migrations", JRSSC 69(4):973-989.

Causal Inference, Double Machine Learning Papers

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters: Double/debiased machine learning. The Econometrics Journal, 21(1).

Gruber, Susan and van der Laan, Mark J., "Targeted Maximum Likelihood Estimation: A Gentle Introduction" (August 2009). U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 252. https://biostats.bepress.com/ucbbiostat/paper252

Levy, J. (2019). Tutorial: Deriving the efficient influence curve for large models. arXiv preprint arXiv:1903.01706.

Van der Laan, Mark J., and Sherri Rose. Targeted learning: causal inference for observational and experimental data. Vol. 4. New York: Springer, 2011.

Fink, D., Johnston, A., Strimas-Mackey, M., Auer, T., Hochachka, W. M., Ligocki, S., & Rodewald, A. D. (2022). A Double Machine Learning Trend Model for Citizen Science Data. arXiv preprint arXiv:2210.15524.