Importance Forest: A Semi-Supervised Solution to Forecasting Outages During a Hundred Year Storm

Tim Coleman, Kim Kaufeld, Mary Frances Dorn, and Lucas Mentch

Problem Overview

- Hurricane forecasts are becoming increasingly accurate, while hurricanes themselves are becoming increasingly severe
- Severe hurricanes can severely damage the power grid, leading to severe power outages across the affected area
- Fundamental issue: in era of hundred year storms happening annually, training predictive models on the historical record may lead to suboptimal predictions in a given area

Existing Outage Forecasting Techniques

- With availability of computational power, research has turned to standard statistcal learning procedures for predicting county-wide outages locally during a particular storm [2]
  - These models tend to rely on local grid information - varies across the country
- To train a global/regional model, model should only use information about the storm and commonly available ground-level information

Challenges in Forecasting

- Each county C_k affected by the hurricane has a time series (recorded every 15 minutes) of outages $O_{t,k}$ - goal is to predict only a summary of severity, defined by:

$$Y_k = \log_{10}\left(\max_t \min_k \{O_{i,k}:\ k \in [t, t+8)\}\right)$$

- Forecasting this quantity is difficult for severe storms, like Hurricane Irma, which is the strongest storm ever recorded in the Atlantic basin [1]

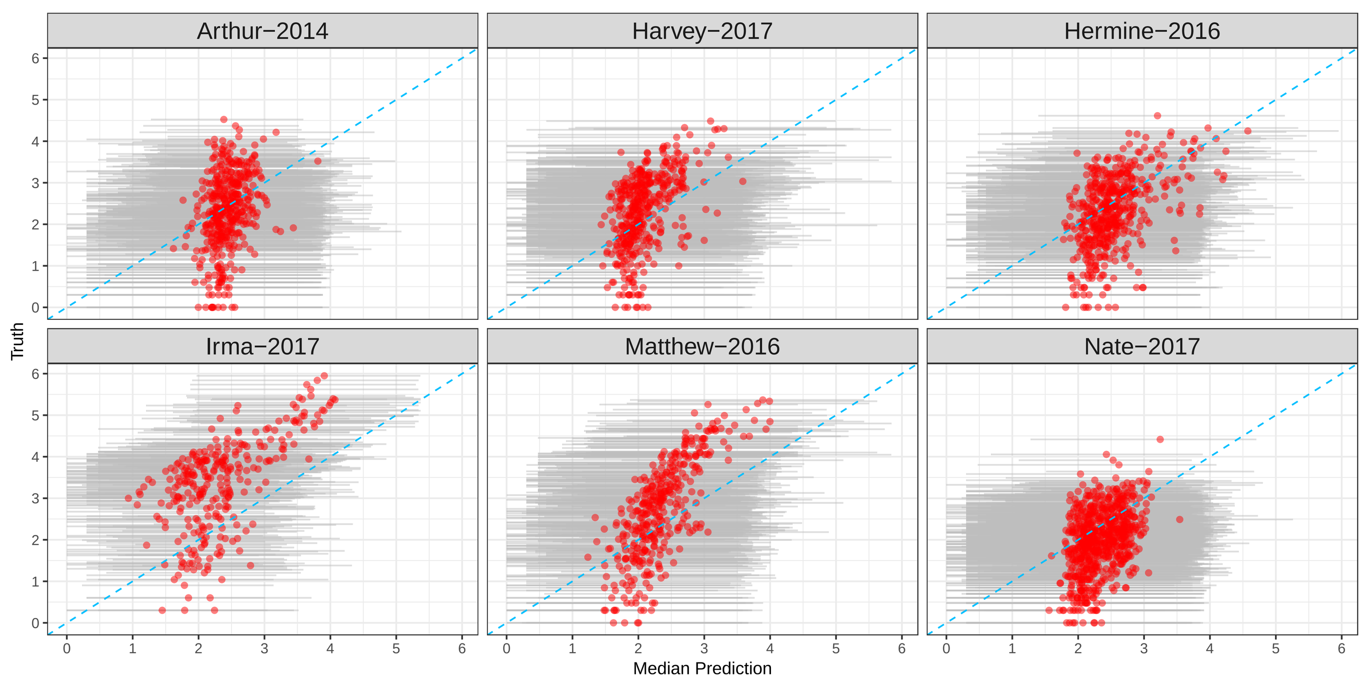

Figure: Predicted vs fitted values for a quantile regression forest, grey bars are 90% pred. intervals

References

[1] Cangialosi, J. P., Latto, A. S., and Berg, R. (2018). Hurricane irma. In *National Hurricane Center Tropical Cyclone Report.*

[2] He, J., Wanik, D. W., Hartman, B. M., Anagnostou, E. N., Astitha, M., and Frediani, M. E. (2017). Nonparametric tree-based predictive modeling of storm outages on an electric distribution network. *Risk Analysis*, 37(3):441–458.

Standard Supervised Learning Approach Fails

- Cross validation (CV) error estimates are too optimistic, and lead to suboptimal model selection for the most intense storms, see Figure 2.
- CV rankings appear negatively correlated with validation rankings for Hurricane Irma

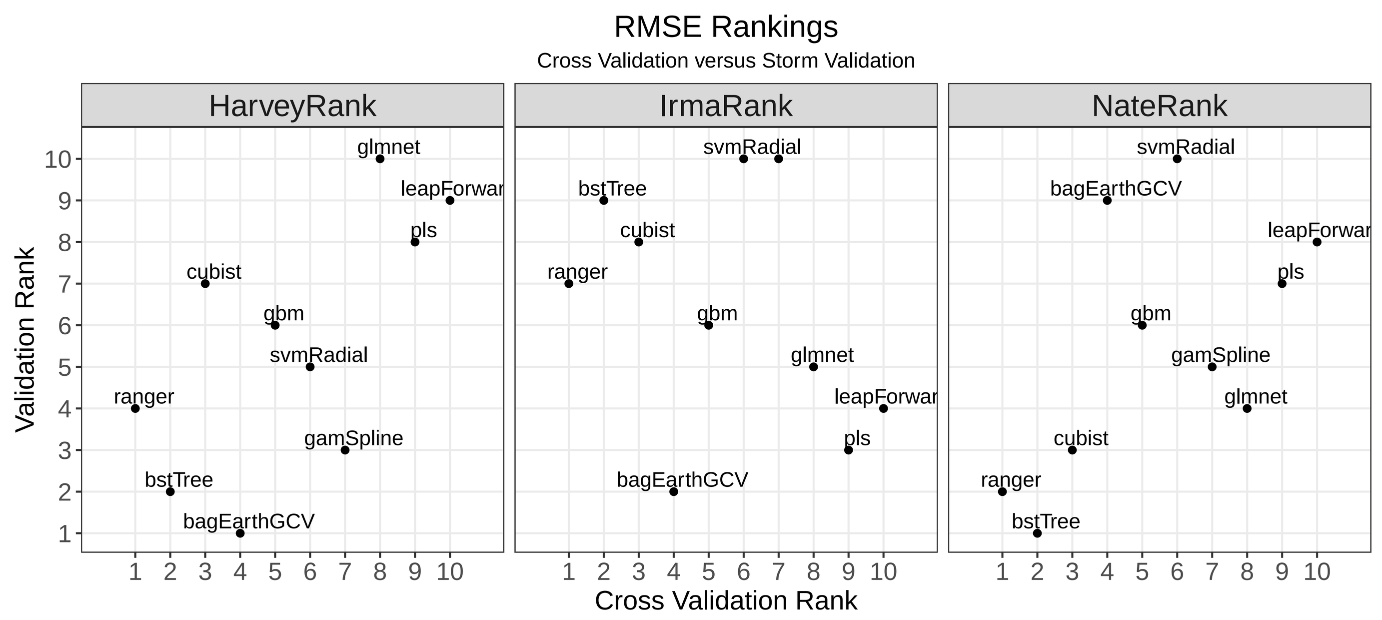

Figure: CV rank vs Validation rank for Irma, Harvey, and Matthew. Labels are `caret` model tags.

Importance Forest

- Essentially, the historical record (training data) and incoming storms (validation data) have different distributions, P_1 and P_2, so that

$$\mathbb{E}_{(X,Y)\sim P_1} L(\hat{f}(X), Y) \neq \mathbb{E}_{(X,Y)\sim P_2} L(\hat{f}(X), Y).$$

where $\hat{f}$ is an estimated regression function and $L(\cdot)$ is a loss function
- Idea: weight model training by a likelihood ratio between the test distribution and the training distribution, i.e. $w(X, Y) \propto \frac{dP_2(X,Y)}{dP_1(X,Y)}$
- Then, replace the variances/means used to recursively split the feature space with a weighted version.

Learning $\ell(X, Y)$

- Assume that P_1 and P_2 satisfy $P_1(X, Y) = P(Y|X)P_1(X)$, $P_2(X, Y) = P(Y|X)P_2(X)$ so that the conditional distribution of outages is the same for all storms.
- Let $Z \sim \text{Bernoulli}(\alpha)$, and then assume that

$$X|Z \sim ZP_1(X) + (1 - Z)P_2(X)$$

- Then, make the following calculations

$$\frac{P(Z=1|X)}{P(Z=0|X)} = \frac{\frac{dP_2(X)P(Z=1)}{P(X)}}{\frac{dP_1(X)P(Z=0)}{P(X)}} = \ell(X)\frac{P(Z=1)}{P(Z=0)}. = \ell(X)\frac{\alpha}{1-\alpha}$$

- Use a classifier to learn $\pi_i = (Z_i = 1|X_i)$, giving unnormalized weight

$$w_i = \frac{\pi_i + 1/n}{1 - \pi_i + 1/n}$$

which must be renormalized within each node, and the $1/n$ terms prevent 0 and infinite weights

Tuning the Model - Weighted OOB measure

- Model selection only works if generalization error estimates work
- Well known that random forest out of bag (OOB) measures are asymptotically equal to leave-one-out cross validation
- Let $B_i = \sum_{k=1}^{B} I(X_i \notin \mathcal{D}_k^*)$, and $T^w(x;\xi)$ be a a weighted tree prediction at x using randomization ξ, then

$$\text{OOB}^w_{m,B} = \sum_{i=1}^{n} \frac{w_i}{\sum_{j=1}^{n} w_j}\left(\frac{1}{B_i}\sum_{k=1}^{B} T_w(X_i;\xi_k) I(X_i \notin \mathcal{D}_k^*) - Y_i\right)^2.$$

- **Result**: if $w(x)$ is consistent for $\ell(x)$, then $\text{OOB}^w_{m,B}$ is consistent for the out of bag error for a random forest trained on P_2

Simulations

- We simulate covariates X over $\mathcal{X} = [0,1]^{30}$, according to the two distributions

$$[X^{(1)}, ..., X^{(5)}] \sim \text{Dirichlet}(\alpha) \ \blacktriangleright\ \alpha \text{ changes between } P_1, P_2$$

$$X^{(6)}, ..., X^{(30)} \overset{iid}{\sim} \text{Uniform}(0,1). \ \blacktriangleright\ \text{Rest of features are the same}$$

- We let $\alpha_1 = \lambda^{\{1:5\}}$, $\alpha_2 = \lambda^{\{5:1\}}$, for some $\lambda > 0$ - higher λ means higher divergence between P_1, P_2
- Generate response from 5 different distributions

Model #	Data Generating Model
1	$Y = 5X^{(1)} + \epsilon$
2	$Y = 5\sin(4\pi X^{(1)}) + \epsilon$
3	$Y = 10\sin(\pi X^{(1)}X^{(2)}) + 20(X^{(3)} - 0.5)^2 + 10X^{(4)} + 5X^{(5)} + \epsilon$
4	$Y = 5e^{2X^{(1)}X^{(2)}+X^{(3)}} \times \text{XOR}(X^{(5)} > X^{(6)}, 1, -.5) + \epsilon$
5	$Y = 5\sum_{j=1}^{5}\left(X^{(j)}\right)^2 + \epsilon$

Simulation Results

- Table below shows results aggregated across $\lambda \in \{1, 5, 10\}$

Model Type	Model #	RMSE	MAE	PCT	Width
Unweighted	1	3.9616	**3.4663**	0.5058	6.5175
Unweighted	2	3.8796	3.3256	0.5584	7.0050
Unweighted	3	**2.9041**	**2.3882**	0.7327	6.8184
Unweighted	4	**5.7940**	**5.4039**	0.4347	8.2436
Unweighted	5	2.2801	1.8113	0.8371	6.4980
Weighted	1	**3.9572**	3.4704	0.4499	5.8064
Weighted	2	**3.8507**	**3.3211**	0.5089	6.3950
Weighted	3	2.9122	2.3908	0.6809	6.1502
Weighted	4	6.0542	5.5165	0.3740	7.3749
Weighted	5	**2.2581**	**1.7883**	0.8016	5.7683

- Weighting is most effective on simpler models (1, 2, 5) in terms of RMSE

Acknowledgments

Lucas Mentch was supported in part by NSF DMS-1712041. This research was supported in part by the University of Pittsburgh Center for Research Computing through the resources provided.

PittCRC Center for Research Computing

Department of Statistics, University of Pittsburgh, Pittsburgh, PA 15260

Presenting Author: Tim Coleman   email: `tsc35@pitt.edu`