# Predicting Evacuation Location for Post-Disaster Recovery Modeling

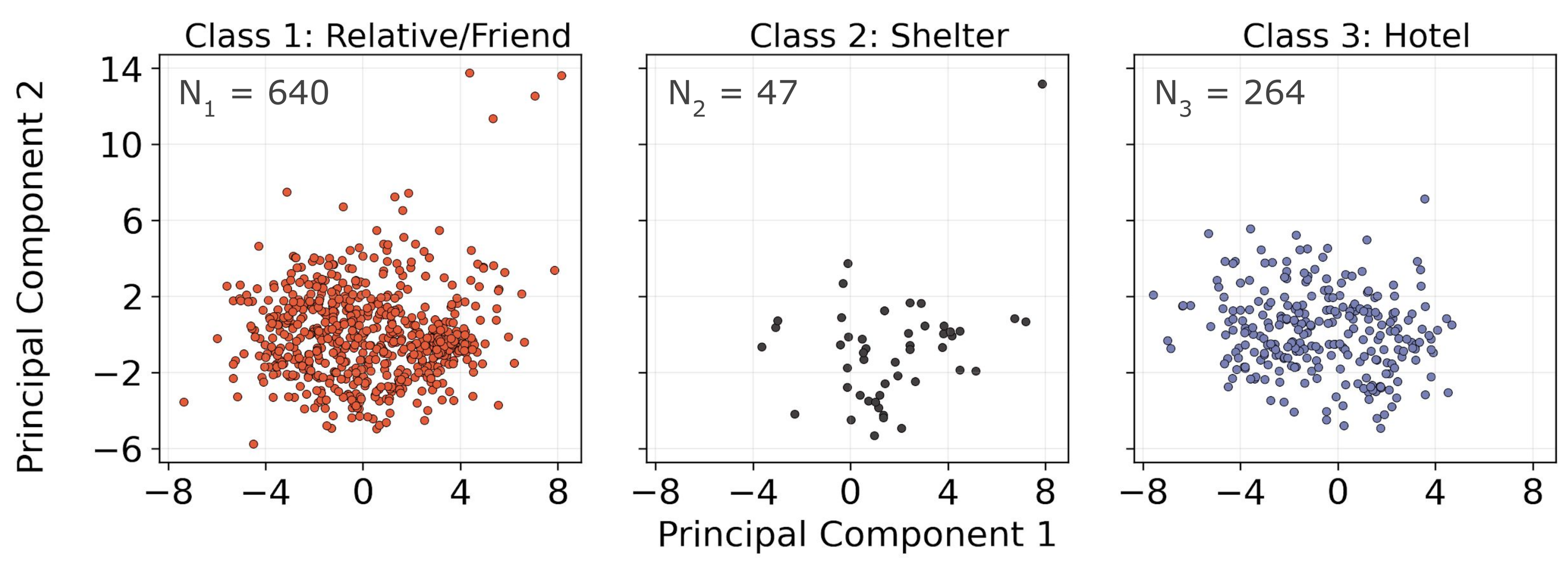
**Aaron Appelle, Ana Moura-Cook, Davyd Tamrazov**
{appellea, anamc, dt453}@stanford.edu

## Introduction / Background / Motivation

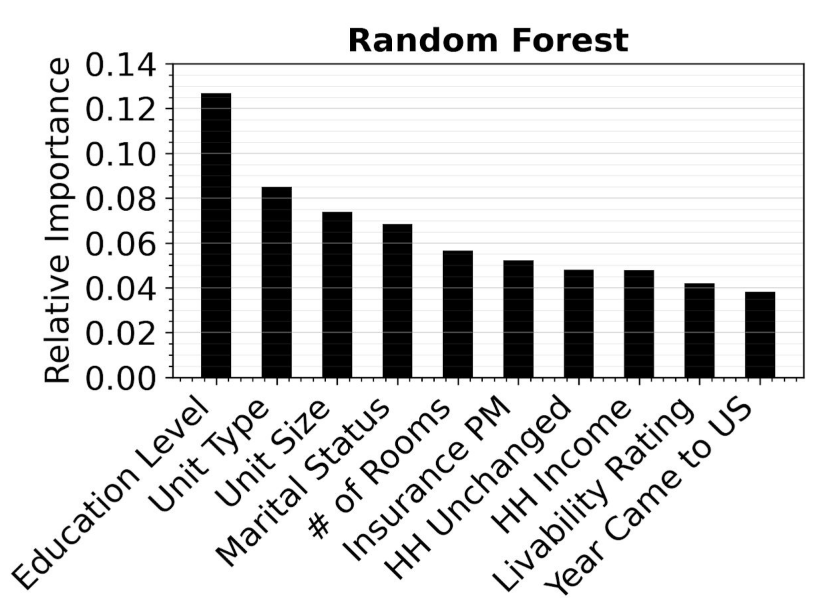
- Post-disaster recovery modeling is limited by semi-heuristic assumptions and survey-based data
- We explore machine learning techniques to predict a household's location of temporary shelter following a natural disaster
- Decision tree and neural network methods perform better in balanced accuracy than a random-guess trivial classifier

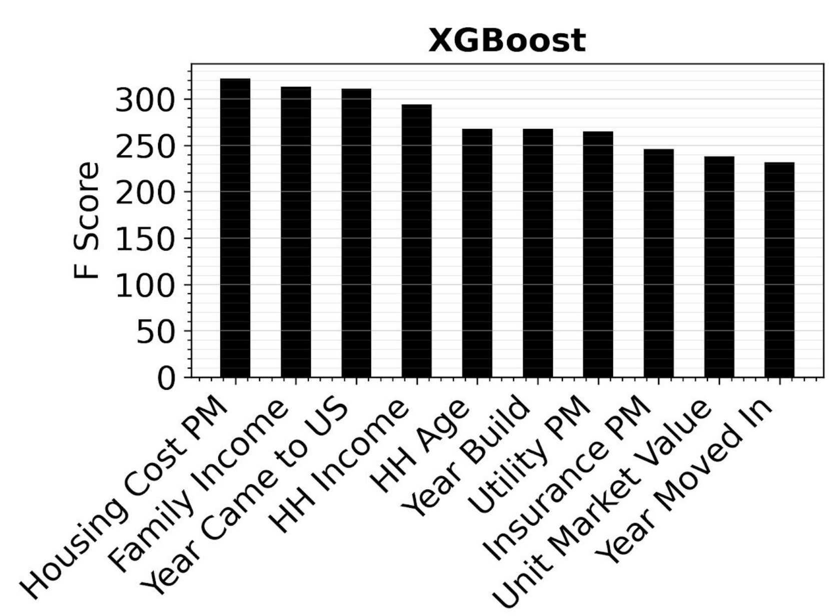
## Dataset and Feature Importance

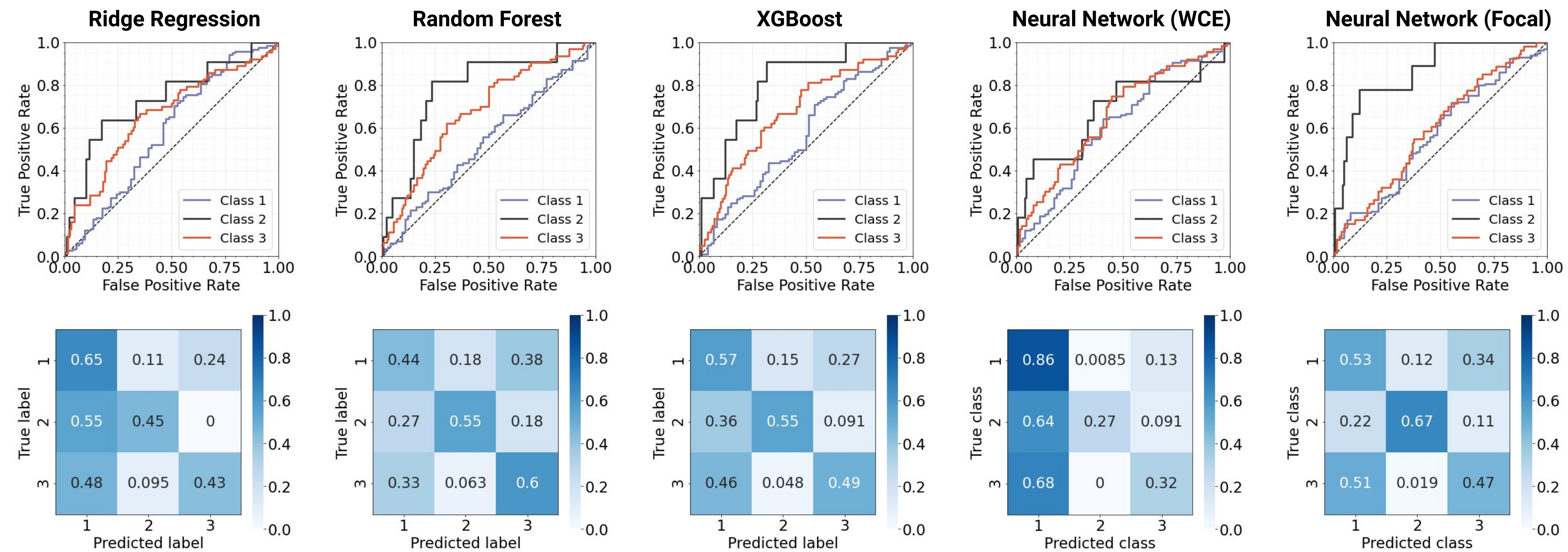
- **Data:** anonymized responses to the 2017 American Housing Survey in the greater San Francisco area (mostly binary)
- Originally 3,343 variables and 2,286 observations, after cleaning and manual feature selection: *951 observations and 61 variables*
- **PCA analysis:** PC1 and PC2 explain only 20.1% of the variance




- **Top features ranked from PCA:** Number of rooms, Owner/renter status, Monthly utility cost, Unit size (sq ft), Number people in household, Type of housing unit, Market value of unit, Type of household, Marital Status, US Citizenship
- When compared to feature ranking from decision tree methods (below), common significant variable themes include *wealth*, *stage of life*, and *time spent living in the region*




$$\hat{\theta} = \left(X^\top X + \lambda I\right)^{-1} X^\top y$$

A multi-output linear regression using L2 regularization

$$I_G(n) = 1 - \sum_{i=1}^{J} (p_i)^2$$

Sequentially splits the data at a number of features to minimize Gini impurity

$$\ell(\theta)^{(t)} = \sum_{i=1}^{n} \ell_{\text{CE}}(y^{(i)}, \hat{y}^{(t)}) + \gamma T + \frac{1}{2}\lambda \sum_{i=1}^{T} w_j^2$$

Uses gradient descent in an iterative fashion to "boost" weak models based on previous trees

$$\ell(x^{(i)}) = w_{y^{(i)}} \left(-S_{y^{(i)}}^{(i)} + \log \sum_{j=1}^{k} \exp(S_j^{(i)})\right)$$

Re-weighted cross entropy loss (WCE) to bias towards minority classes

$$\ell(x^{(i)}) = -\alpha(1 - p_{y^{(i)}}^{(i)})^\gamma (\log(p_{y^{(i)}}^{(i)}))$$

Focal loss makes easily-classified examples contribute less to the loss

## Results and Discussion

- Neural network with Focal Loss[1] has the highest balanced accuracy: best for predicting the number of households which will evacuate to shelters
- Neural network with weighted cross-entropy loss has a higher AUROC for Class 3 while maintaining high total accuracy: best for predicting evacuation to hotels
- XGBoost[2] is the most interpretable and performs relatively well across error metrics and classes, but it is not best-in-class for any one metric
- Categorical and imbalanced data is a challenge for every model considered

| *Metric* | Baseline | Ridge Regression | Decision Tree | | Neural Network | |
|---|---|---|---|---|---|---|
| | *Naive* | | *Random Forest* | *XGBoost* | *Weighted CE* | *Focal Loss* |
| Total Accuracy | 0.6126 | 0.5654 | 0.5026 | 0.5445 | **0.6492** | 0.5864 |
| Balanced Accuracy | 0.3333 | 0.5109 | 0.5310 | 0.5367 | 0.4845 | **0.5739** |
| Weighted F1 Score | 0.4654 | 0.5740 | 0.5171 | 0.5575 | **0.6180** | 0.5976 |
| AUROC (Class 1) | 0.5000 | 0.5835 | 0.5337 | 0.5698 | **0.6185** | 0.5678 |
| AUROC (Class 2) | 0.5000 | 0.7354 | 0.7864 | 0.8131 | 0.6843 | **0.8651** |
| AUROC (Class 3) | 0.5000 | 0.6612 | **0.6773** | 0.5820 | 0.6598 | 0.5956 |

## Future Work

- Apply LDAM and deferred re-weighting[3] until after the initial stage
- Explore data undersampling and better class weighting schemes
- Establish more data-driven relationships to improve the underlying semi-heuristic assumptions in existing recovery models

## References

[1] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," 2018.

[2] T. Chen and C. Guestrin, "Xgboost," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* Aug 2016. [Online]. Available: http://dx.doi.org/10.1145/2939672.2939785

[3] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017.