Predictive Modeling of Property Hazards Prior to Inspection

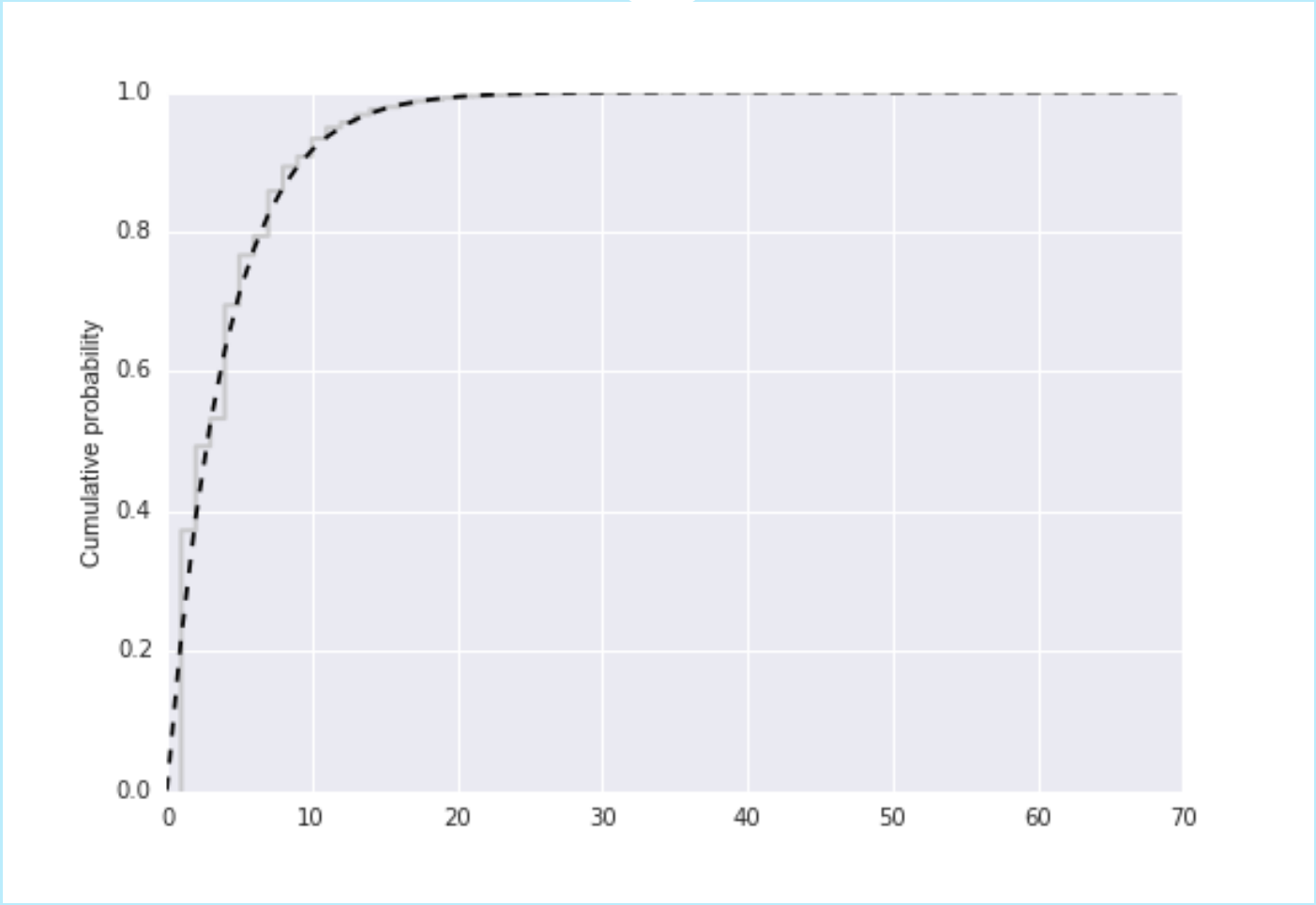
CAPSTONE PROJECT FOR SLIDERULE'S DATA SCIENCE INTENSIVE COURSE

BY:
JUSTIN GEIMAN

Property Hazard Data Set

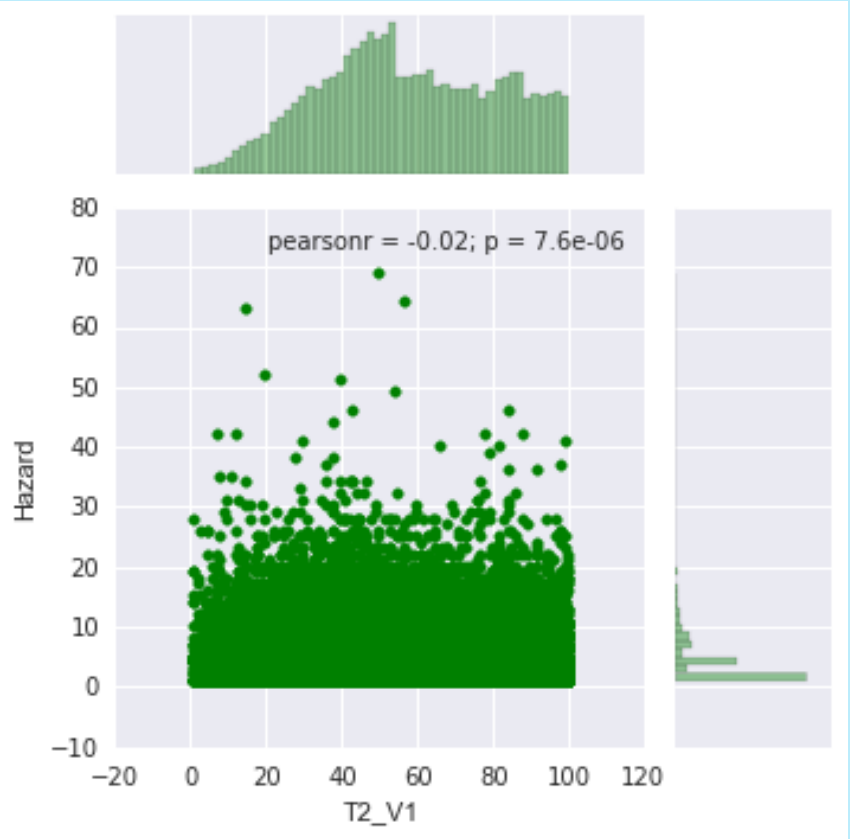
- Kaggle competition
- Training & Test Data: 51,000 properties each
- **Goal:** Predict property hazard scores from variables that are available before the inspection.

Distribution of Hazard Scores Fits Exponential Distribution

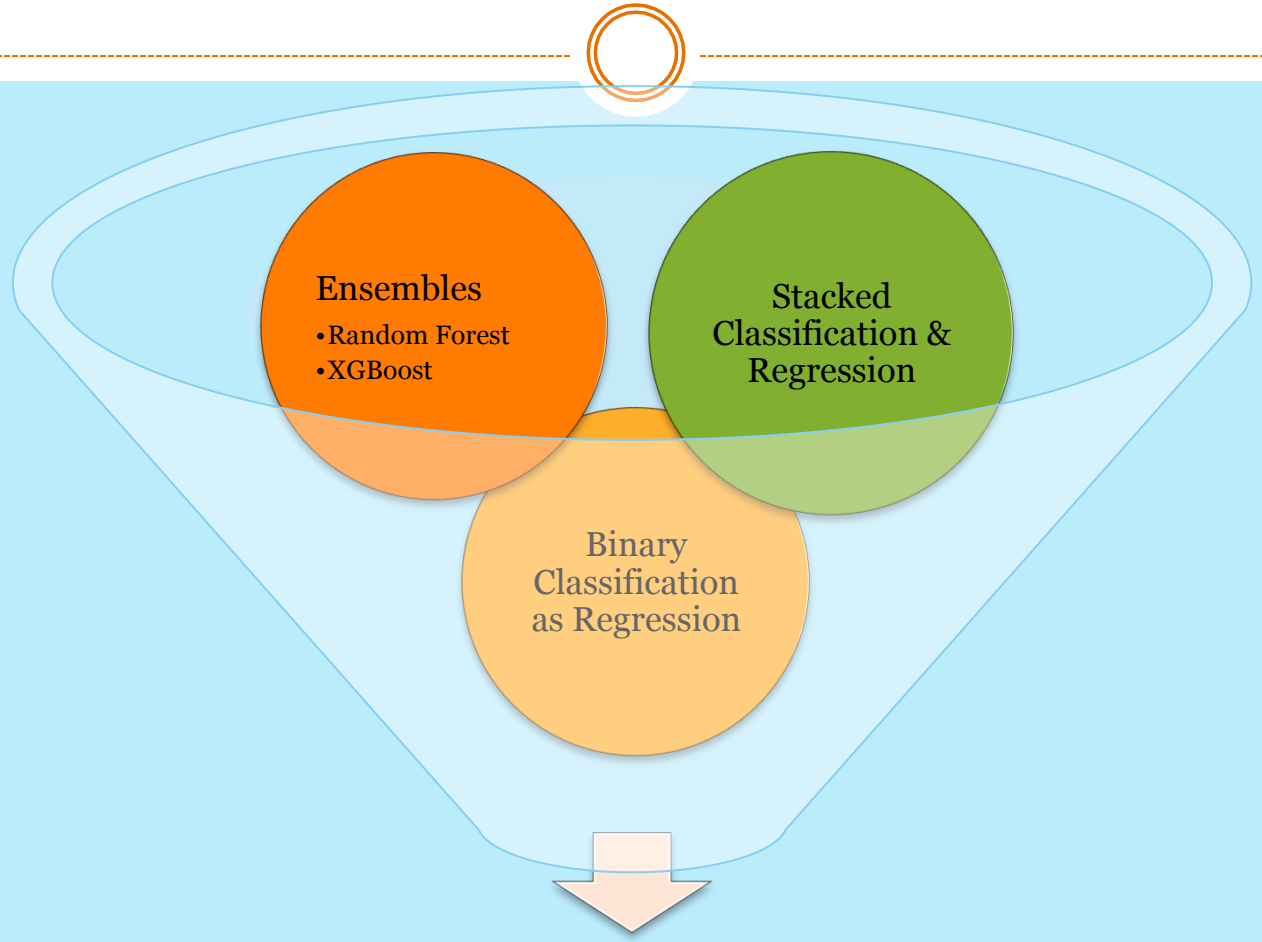


Property Features

- 32 **anonymous** feature variables
- 16 numerical features
- 16 categorical (Yes/No, A/B/C/D) features
- No features had strong relationships with hazard score



Modeling Approach

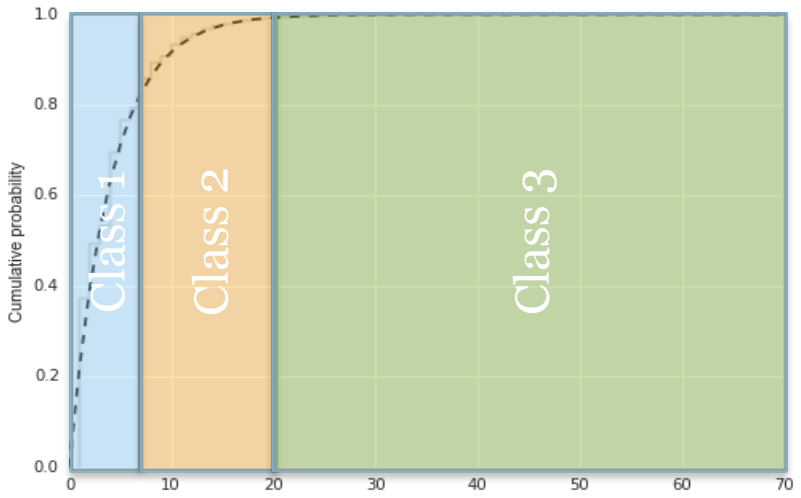


Ensemble Methods

- Combine multiple decision trees to create model that generalizes well to new data
- Examples:
  - Random Forest
  - Extra Trees
  - Gradient Boosted Trees
  - XGBoost

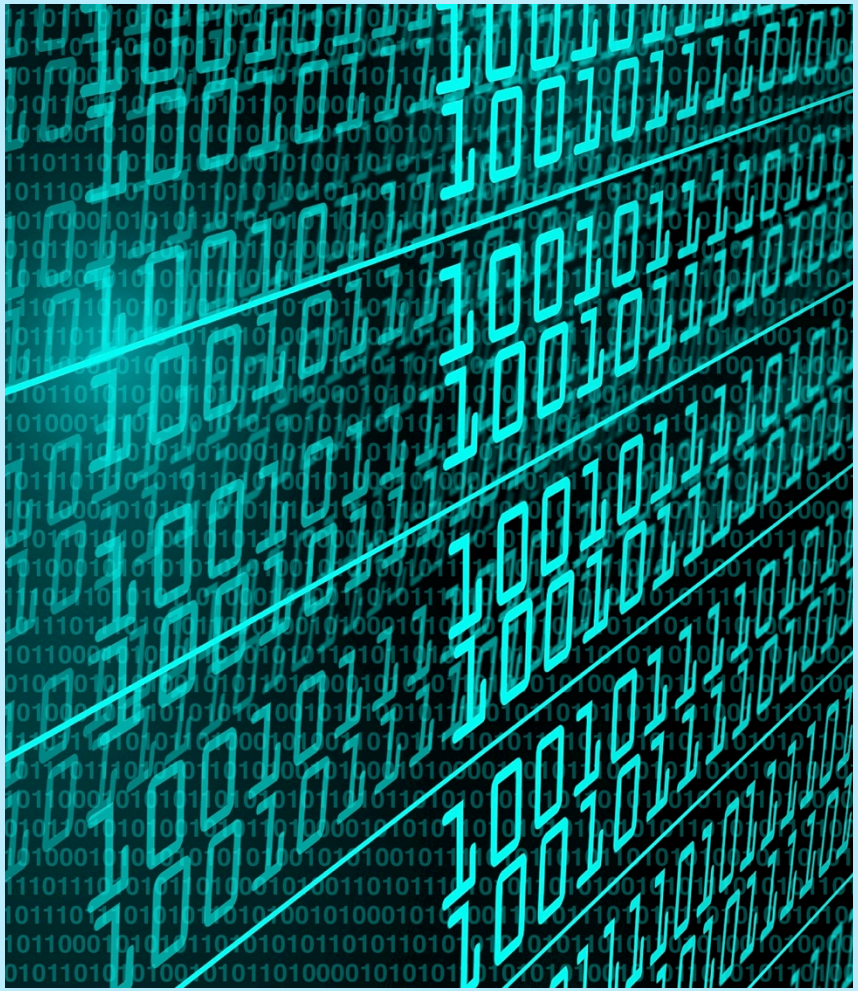
Stacked Classification & Regression

- Separate hazard scores into bins (classes)
- Fit regression models for each class



Binary Classification

- Create series of binary classification problems (1 = True, 0 = False)
- Predict True (1) if hazard predicted to be greater than some threshold
- Sum binary classification predictions over wide range of threshold values
- Approach used by Kaggle competition winner



Best Model

- Normalized Gini Coefficient: 0.375
- Model – Ensemble of 4 models:
  1. Random Forest Regression
  2. XGBoost Regression
  3. Stacked Classification & Regression with XGBoost
  4. Binary Classification using XGBoost
- Outputs of each model averaged to achieve final results.

Conclusions & Recommendations

- The distribution of hazard scores appeared to follow an exponential probability distribution.
- Tree-based ensemble methods such as Random Forests and XGBoost worked well for the property hazard data set.
- Overall prediction accuracies were low. A business decision is required to determine whether the predictions are accurate enough to be useful.

Conclusions & Recommendations

- Feature engineering, with domain-specific knowledge and known variables, may increase prediction accuracies. Feature engineering was difficult for this project given the anonymous features in the data set.
- Ensembling models from diverse approaches proved to provide the best overall performance.