PREDICTING WILDFIRE CAUSES

Supervised Learning Capstone

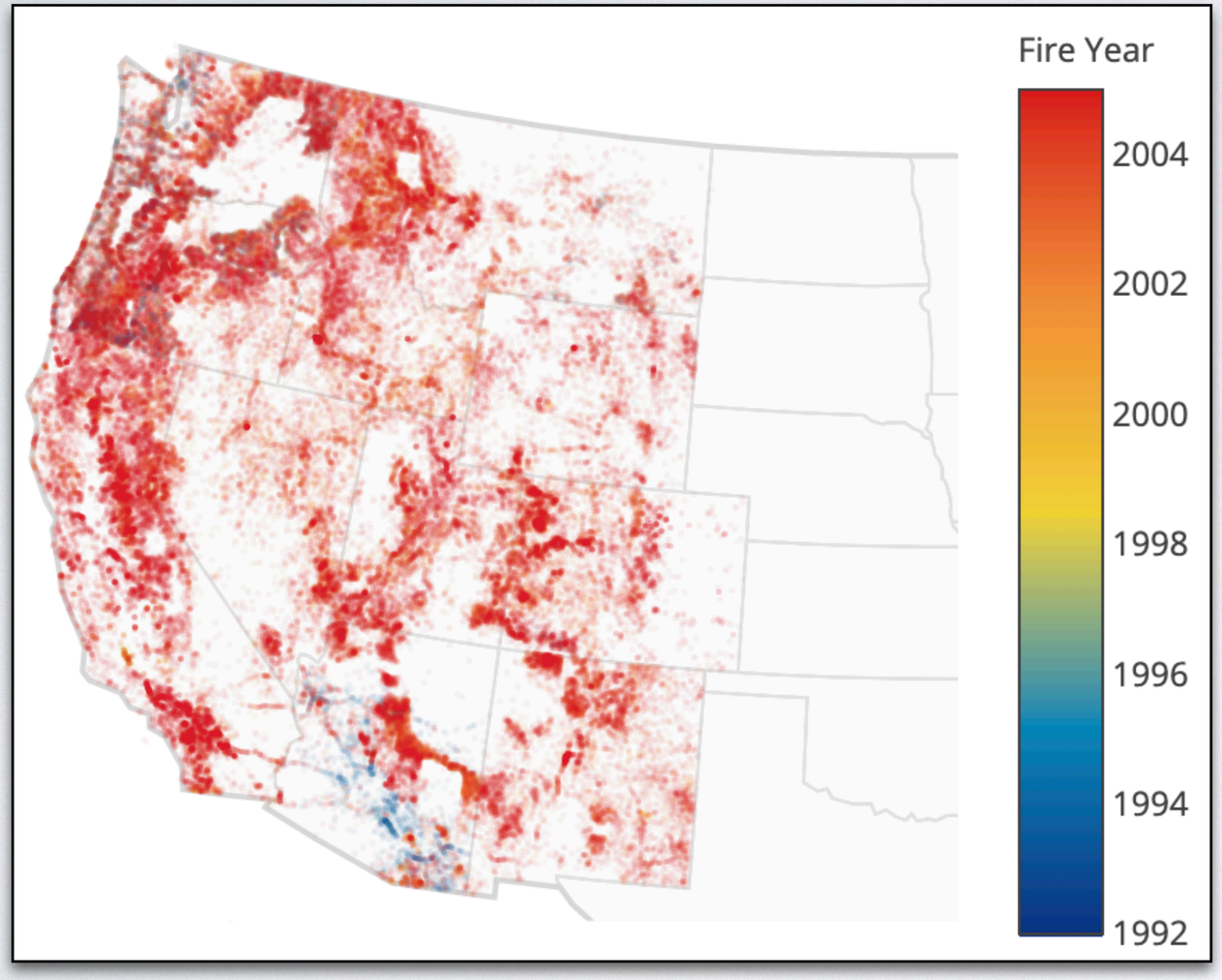
Matt Francsis

MOTIVATION

- Human-caused fires accounted for between 43 and 59% of all wildfires in the western US from 1992-2015.
- While wildfires can be beneficial to the ecosystem, they also pose serious threats to lives, property, and infrastructure.
- Cause prediction can:
  - Assist investigators bring arsonists to justice
  - Inform the public of the actual causes of wildfires
  - Act as a catalyst for abatement strategies

GOAL

Fires of the western US: 1992 - 2015



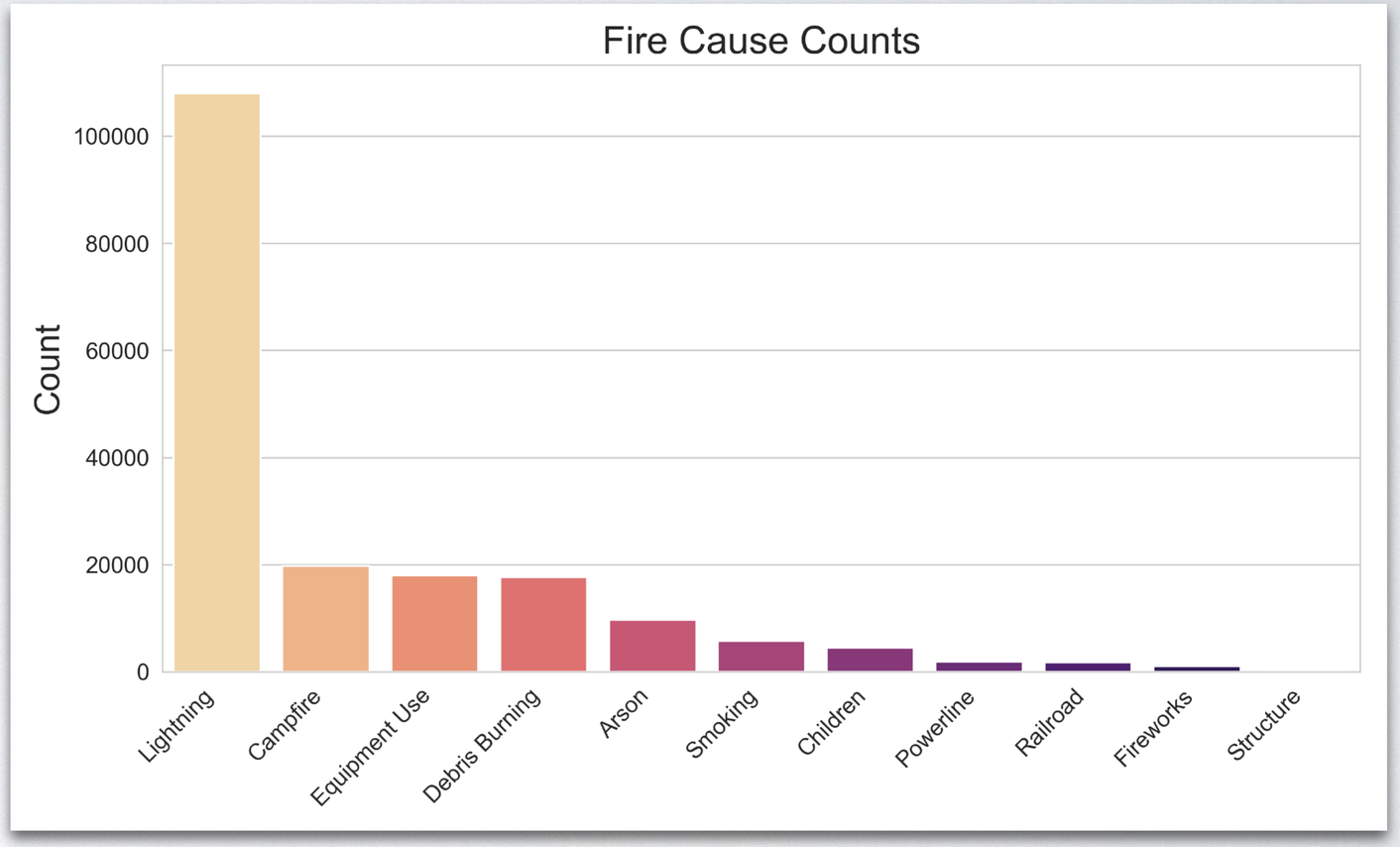
- Use supervised learning algorithms to predict the cause of forest fires in the western US

OUTLINE

- Data set
- Model setup & metrics
- Class imbalance strategies
- Baseline models
- Models on augmented data
- Results
- Future work

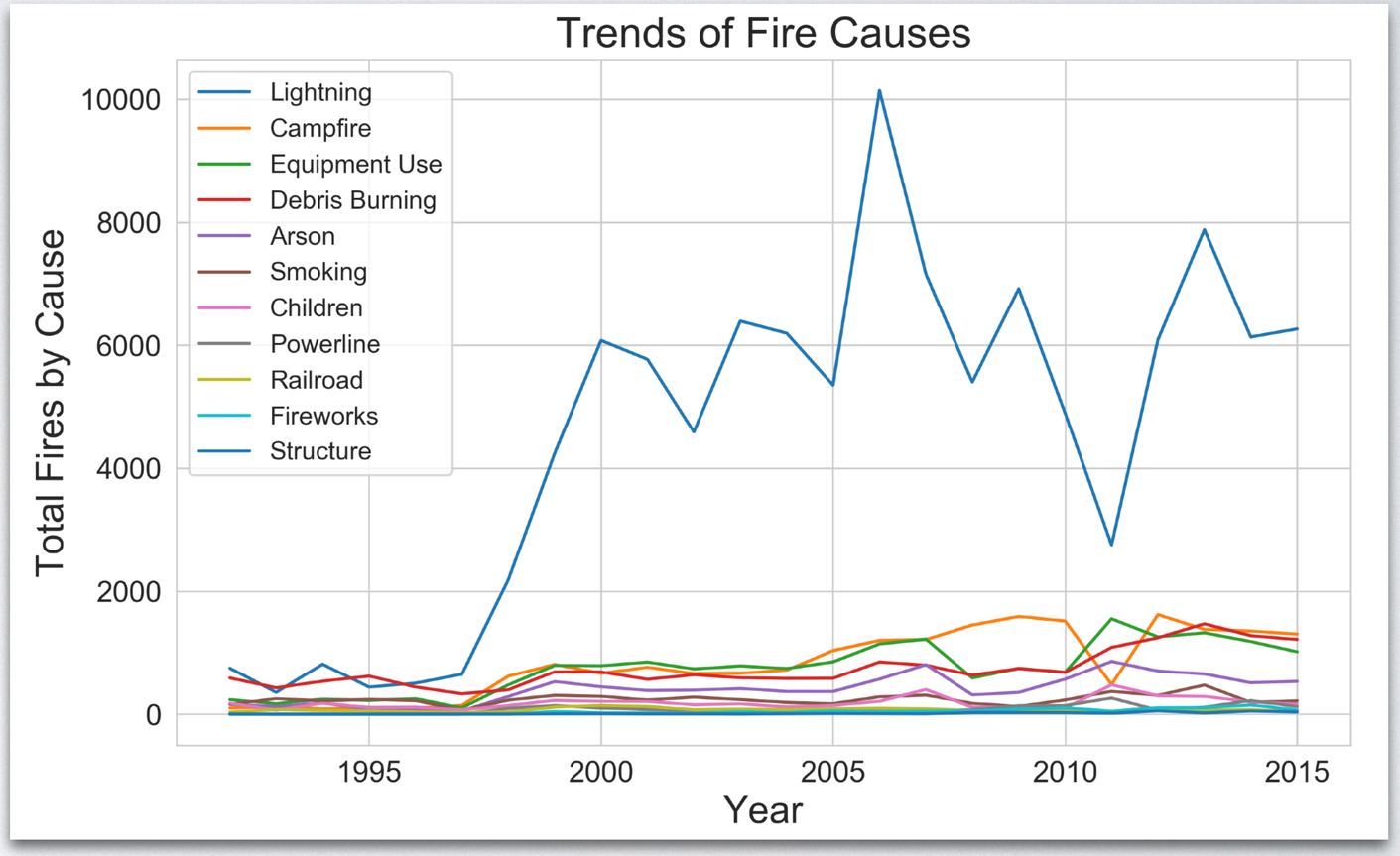
DATA SET

- Nationally compiled forest fire database from 1992 - 2015
- 1.88 million total fires recorded
- Types of data:
  - Reporting agencies
  - Report metadata
  - Fire names
  - Dates
  - Size
  - Geography



DATA SET

- Nationally compiled forest fire database from 1992 - 2015
- 1.88 million total fires recorded
- Types of data:
  - Reporting agencies
  - Report metadata
  - Fire names
  - Dates
  - Size
  - Geography



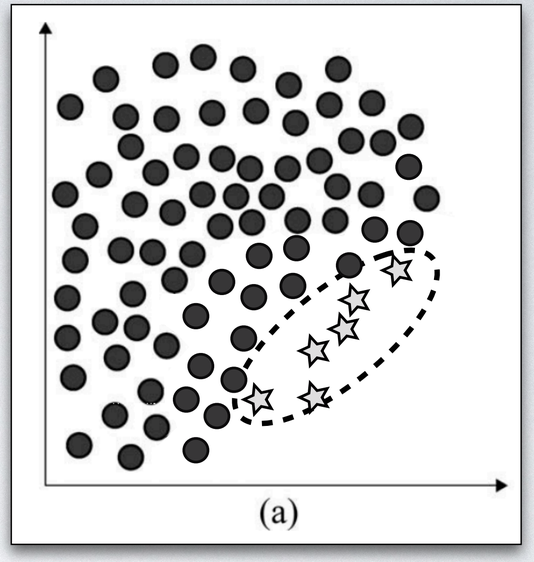
MODEL SETUP

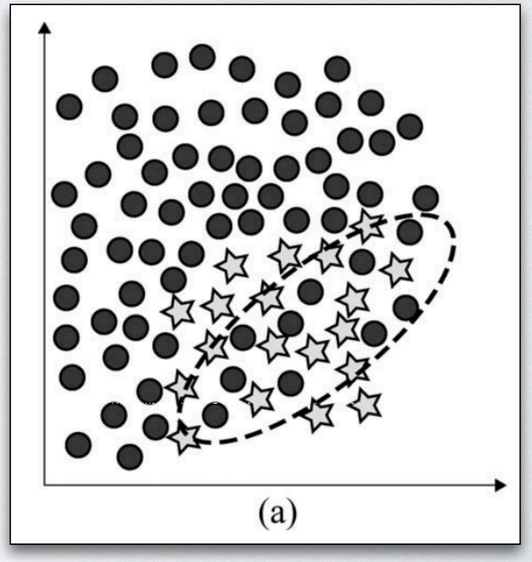
Feature	Description	Type
STAT_CAUSE_CODE	Fire cause	Categorical
DISC_MONTH	Month fire started	Categorical
DISCOVERY_TIME	Time fire started or was reported	Continuous
FIRE_LENGTH	Length (days) fire burned uncontained	Continuous
FIRE_SIZE_CLASS	Final size of fire, binned	Categorical
STATE	State fire started in	Categorical
FULL_FIPS	County fire started in	Categorical
OWNER_DESCR	Landowner type where fire started	Categorical

Data shape: (188,018, 8) **Train / Test Split: (75%, 25%)**

CLASS IMBALANCE

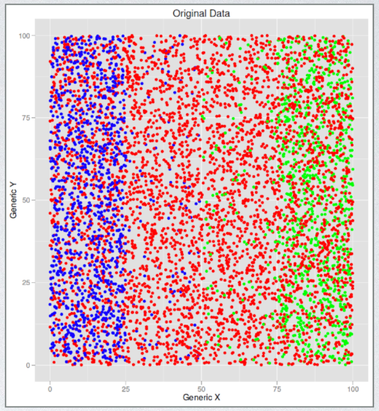
- Why class imbalance is bad:
  - Small minority classes lead to limited training losses when misclassified
  - Significant class overlap can lead to poor class boundaries
- Two main ways to deal with class imbalance:
  - Data space manipulations
  - Algorithm modifications (e.g., Fig. 3)



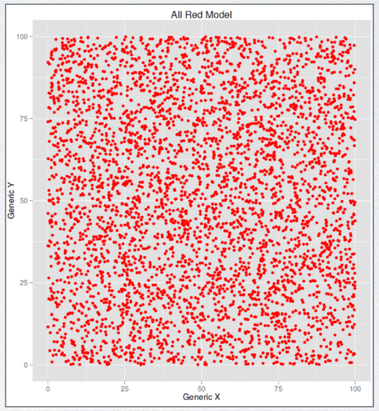


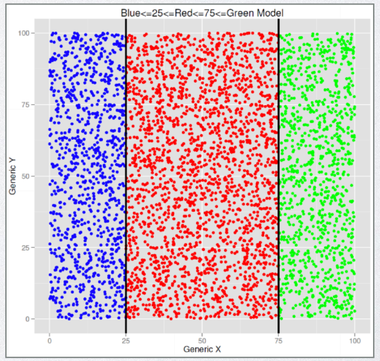
MODEL METRICS

- Accuracy is not useful
- Overview of imbalanced metrics (see §2.2)
  - Balanced accuracy
    - Arith. avg. of recall of each class
  - Geometric mean
    - Geom. avg. of recall of each class
  - Weighted F1
    - Class weighted average of precision & recall
    - Will only consider this metric
- Computation time



Given this input data





These two models have the same accuracy score

MODEL METRICS

- Accuracy is not useful
- Overview of imbalanced metrics (see §2.2)
  - Balanced accuracy
    - Arith. avg. of recall of each class
  - Geometric mean
    - Geom. avg. of recall of each class
  - Weighted F1
    - Class weighted average of precision & recall
    - Will only consider this metric
- Computation time

Confusion Matrix Example

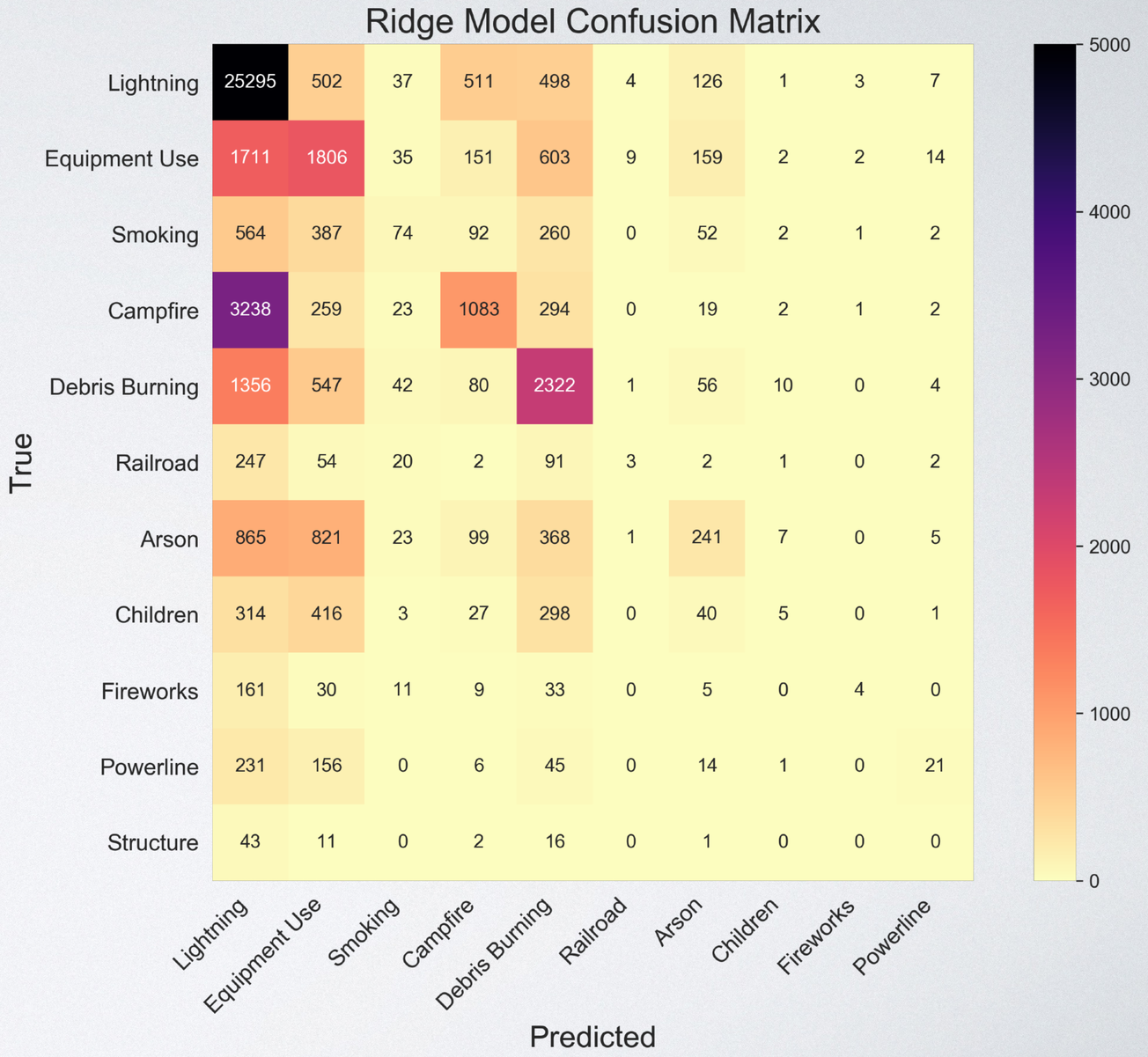
True Result \ *Predicted Result*	**A**	**B**	**C**
A	80	18	2
B	10	20	10
C	8	3	4

INITIAL MODELS

LOGISTIC REGRESSION

- Ridge Regression
  - Solver = lbfgs
  - C = 0.9
- Did not predict any **structure** fires
- Low precision for **lightning** class
- Low accuracy for all classes

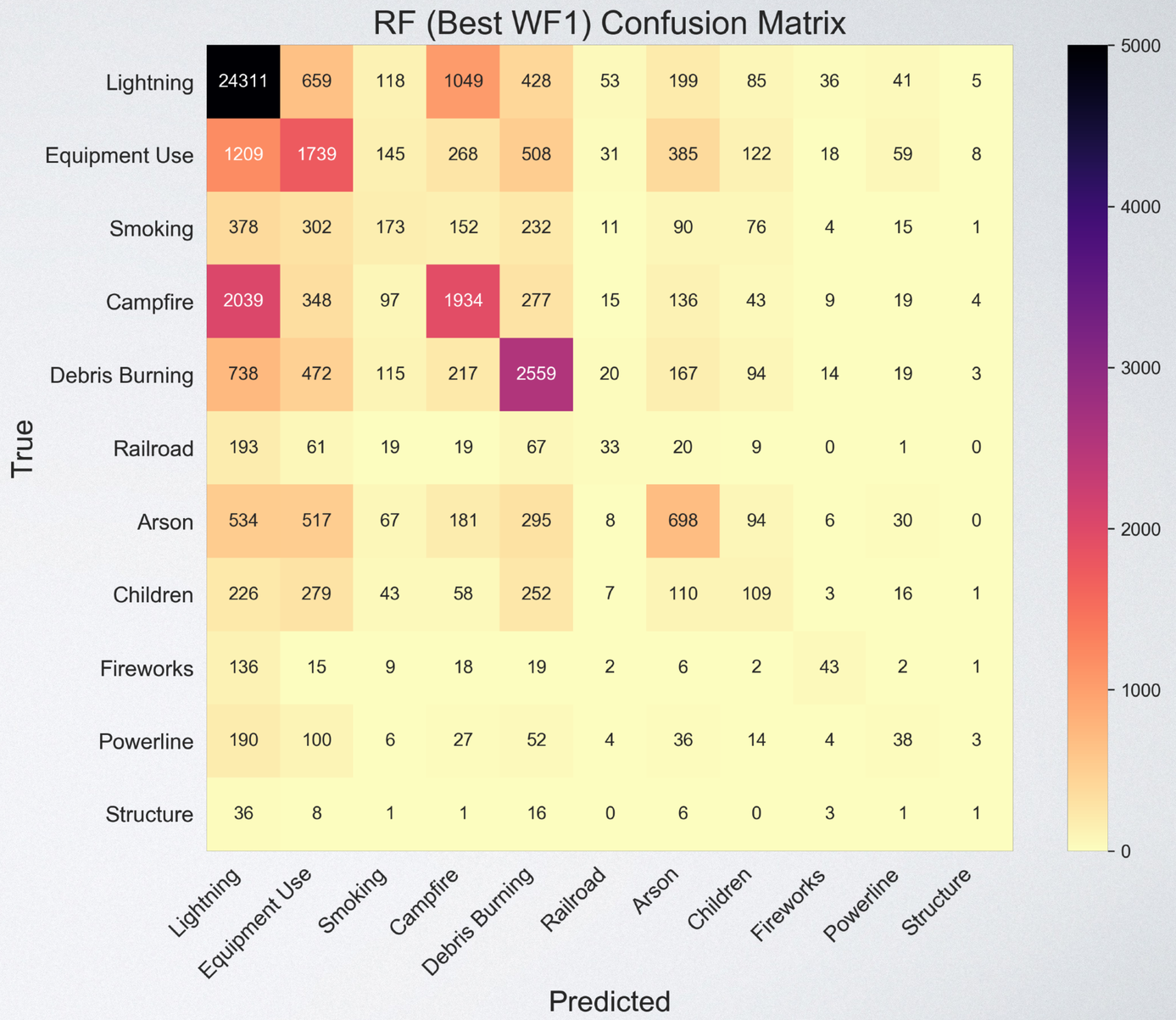
Weighted F1	**0.604**



RANDOM FOREST

- Hyperparameters:
  - N_estimators = 500
  - Max_depth = 20
- Predicted all classes
- Better accuracy for all classes except **lightning** compared to Ridge

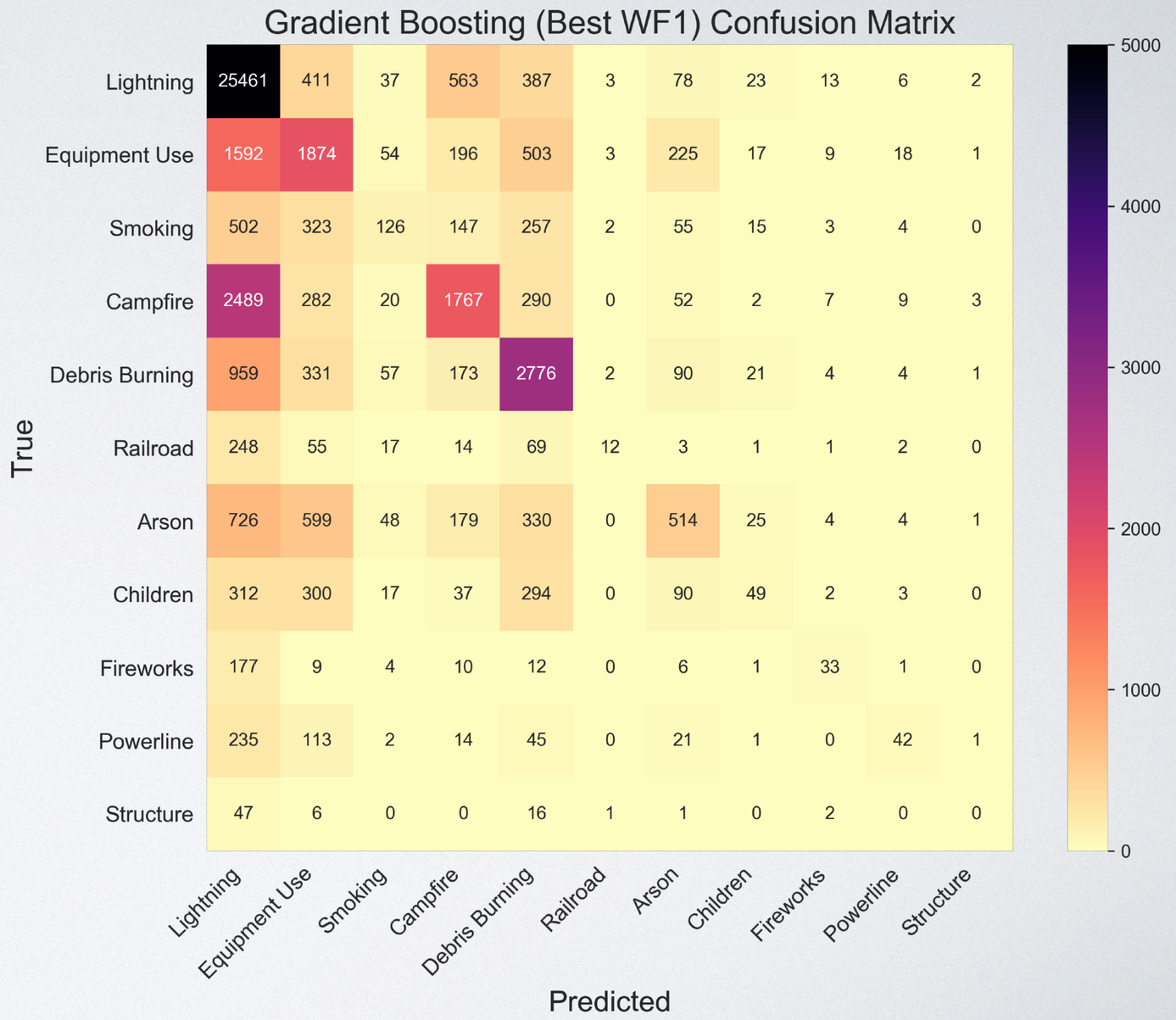
Weighted F1	**0.660**



GRADIENT BOOSTING

- Hyperparameters:
  - N_estimators = 250
  - Max_depth = 5
  - Learning_rate = 0.1
- Lower recall for small minority classes than RF
- Higher **lightning** accuracy than RF

Weighted F1	**0.656**



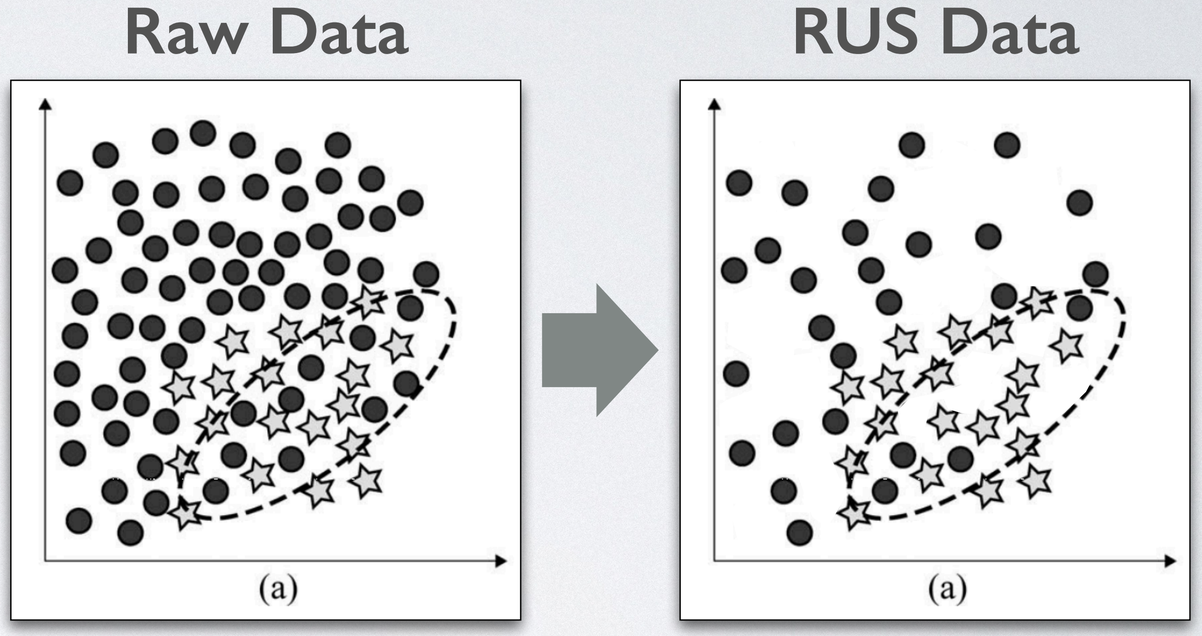
INITIAL MODEL RESULTS

- Logistic regression
  - Comparable score to tree-based models, but failed to predict a class
- Random forest
  - Best models for minority class prediction
- Gradient boosting
  - Near identical scores to RF, but poorer minority class prediction
  - Computationally inefficient for large, multi-class problems

DATA AUGMENTATION

UNDER SAMPLING

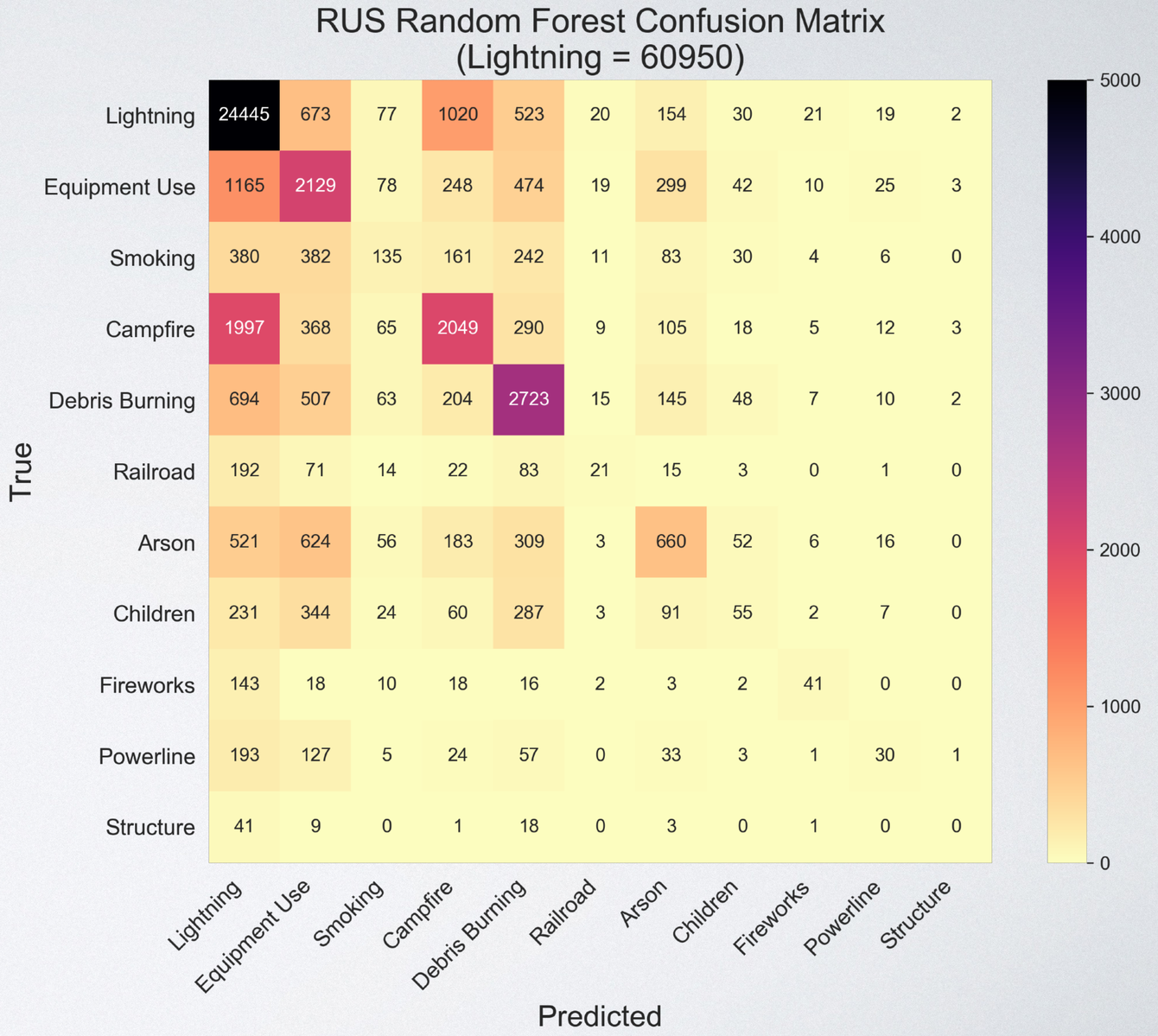
- **Random Under Sampling**
  - Under sample the majority class(es) by randomly picking samples to drop



RANDOM UNDER SAMPLED RANDOM FOREST

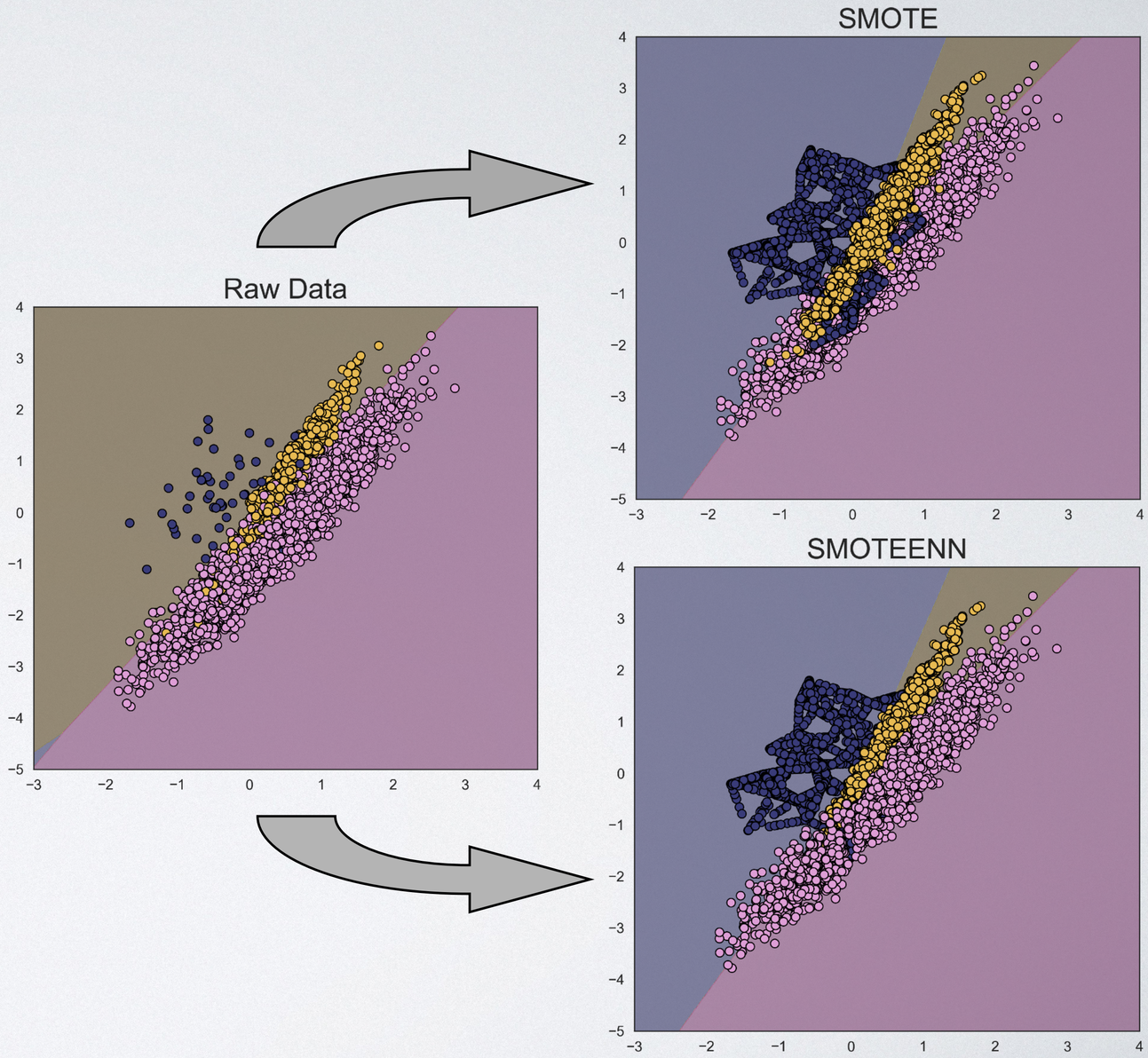
- Lightning RUS to 60,950
- Hyperparameters:
  - N_estimators = 250
  - Max_depth = 20
- Very similar performance to baseline RF
- Slightly better accuracy for the top half classes and slightly worse for the bottom half classes

Weighted F1	0.663



OVERSAMPLING *(Then Undersampling)*

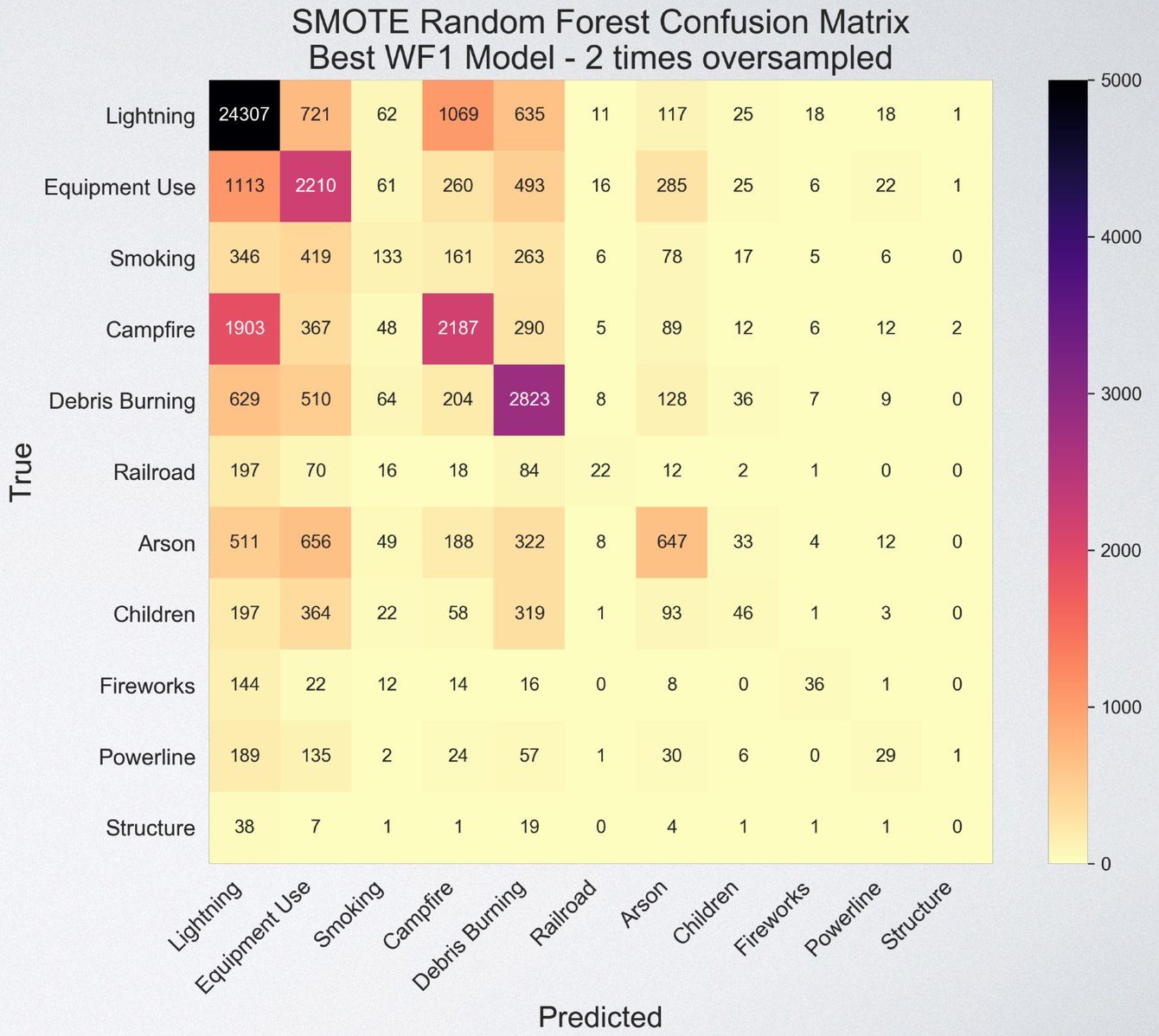
- **SMOTE**
  - Synthetic samples on a line connecting a sample with it's KNN of the same class
- **SMOTEENN**
  - SMOTE but removes all points where a majority of it's KNNs are from a different class



SMOTE RANDOM FOREST

- Minority classes over-sampled by a factor of two
- Hyperparameters:
  - N_estimators = 500
  - Max_depth = 20
- Best weighted F1 score
- Very similar to baseline RF
- Slightly better with middle class prediction

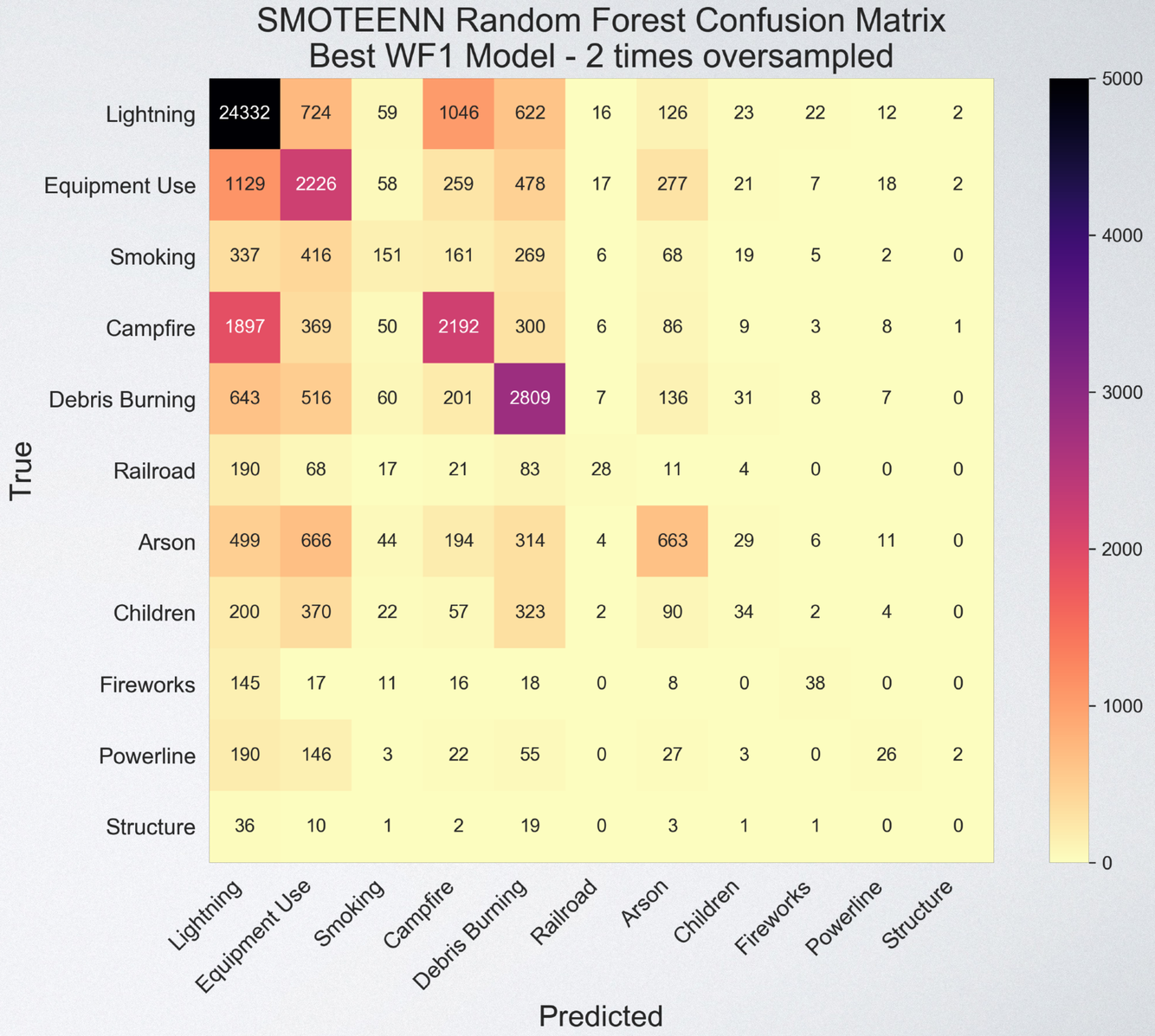
Weighted F1	0.666



SMOTEENN RANDOM FOREST

- Minority classes over-sampled by a factor of two
- Hyperparameters:
  - N_estimators = 250
  - Max_depth = 20
- Essentially identical to SMOTE model

Weighted F1	0.655

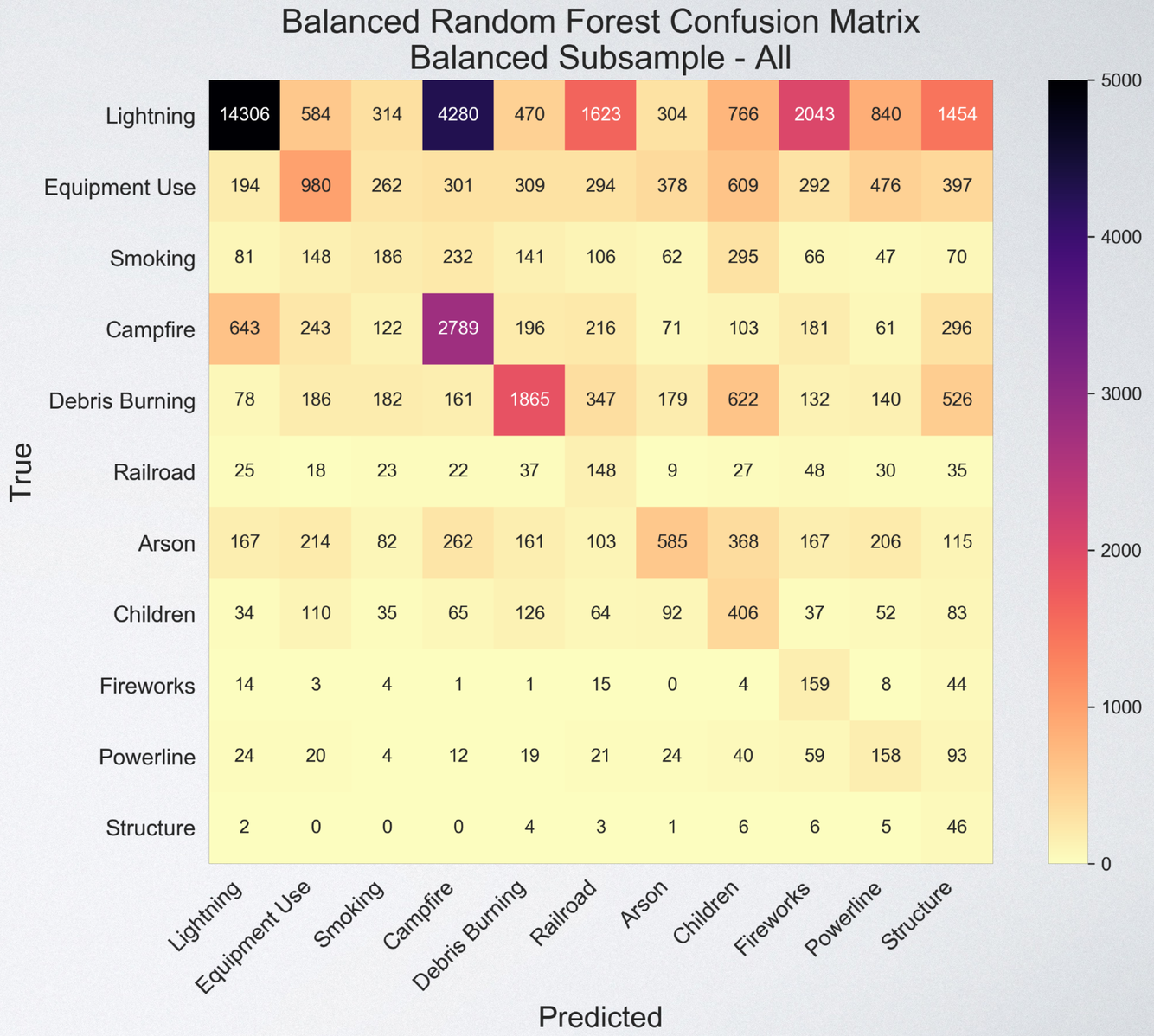


ALGORITHM MODIFICATION

BALANCED RANDOM FOREST

- RUS each tree to force a balanced sample
- Hyperparameters:
  - N_estimators = 250
  - Max_depth = 20
  - Class_weight = 'balanced_subsample'
  - Sampling_strategy = 'all'
- Much lower **lightning** recall
- Much lower precision for low minority classes

Weighted F1	**0.529**



RESULTS

Model	Weighted F1	Balanced Acc
Ridge	0.604	0.210
Random Forest	0.660	0.272
Gradient Boosting	0.656	0.267
RUS RF (L=60950)	0.663	0.282
SMOTE RF (OvSmp * 2)	**0.666**	0.284
SMOTEENN RF (OvSmp * 2)	0.655	0.270
Balanced Random Forest	0.529	**0.402**

FUTURE WORK

- More feature engineering
  - Elevation
  - Weather
  - Land cover type
  - Land use type
  - County-level demographic data
- Utilize different data sampling techniques
- Impute missing county FIPS codes
- GIS mapping work
- Unsupervised clustering analyses
- Time-series analyses