Predicting Patient Outcomes with Breast Cancer Gene Expression Data

Capstone 2 Project

Annie Erbsen
July 2023

The Problem

- Breast Cancer is the leading cause of cancer deaths among women
- Survival duration is important for treatment decisions
- Breast cancer patients with the same stage of disease and clinical attributes can have very different outcomes

Do differences in gene expressions account for the difference?

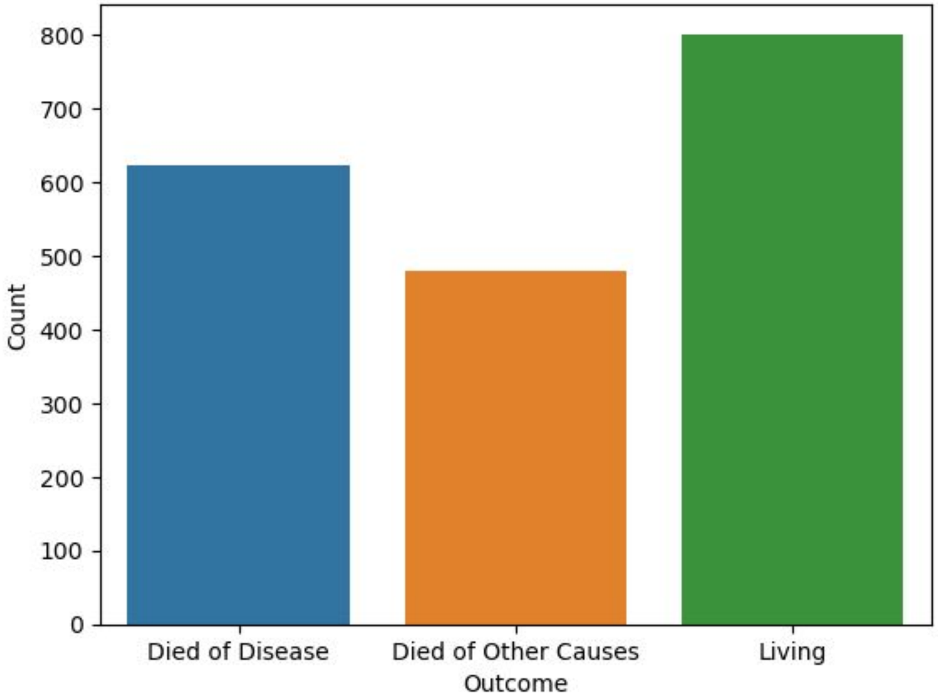
The Data

Sourced from Kaggle: Breast Cancer Gene Expression Profiles (METABRIC)

- 1904 patients
- 693 features
  - Clinical data (30), z scores(489), genetic mutations(174)

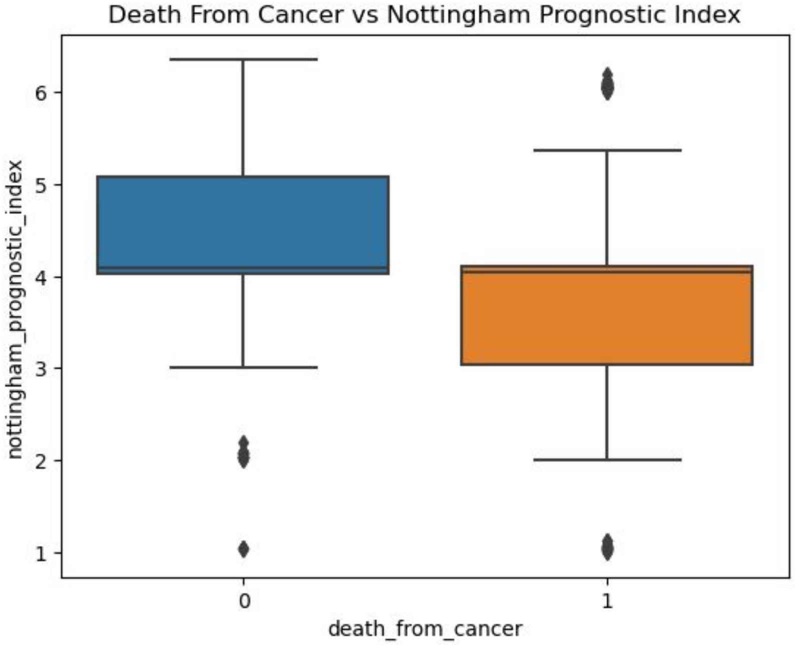
➡ **Target feature: death_from_cancer**

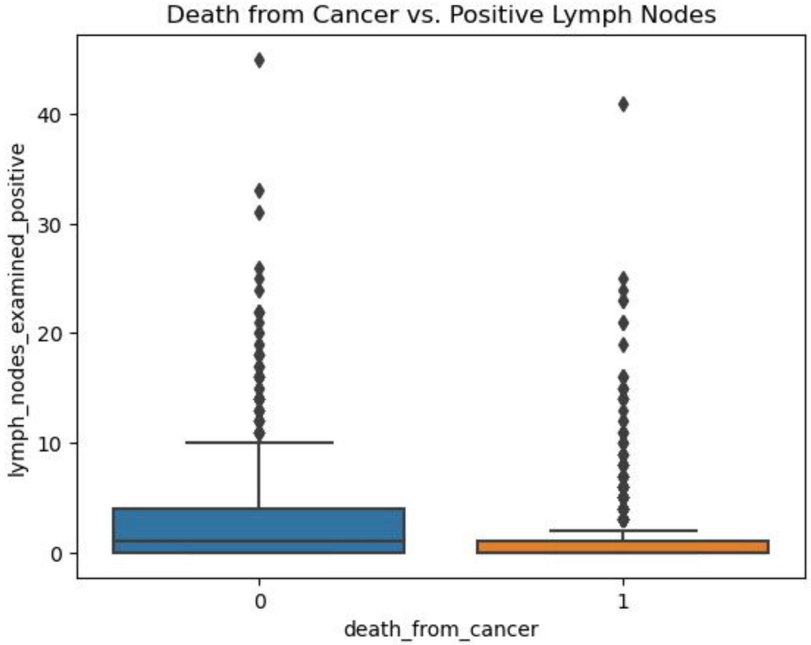
Survival Outcomes (351 months)



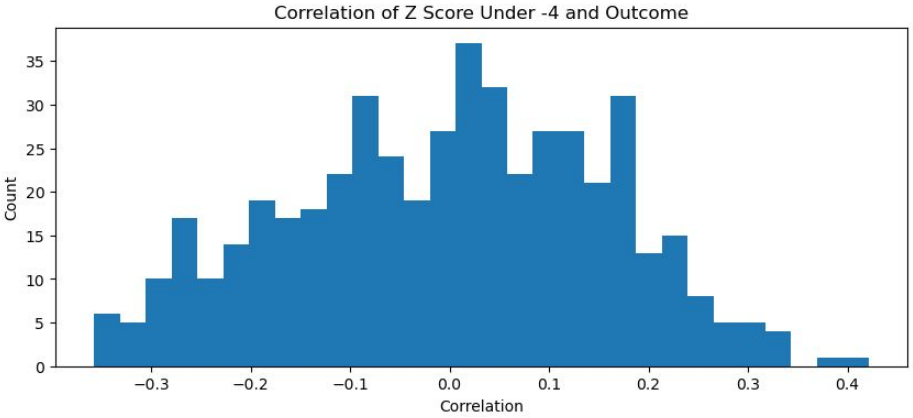
33% 25% 42%

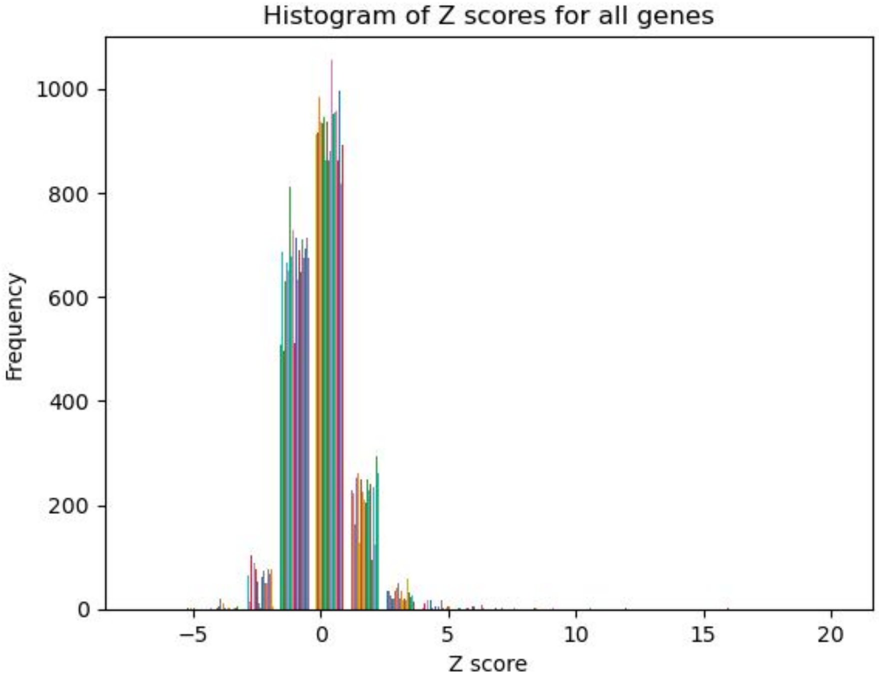
Clinical Attributes Correlated with Outcome





Z Scores & Outcome Correlation

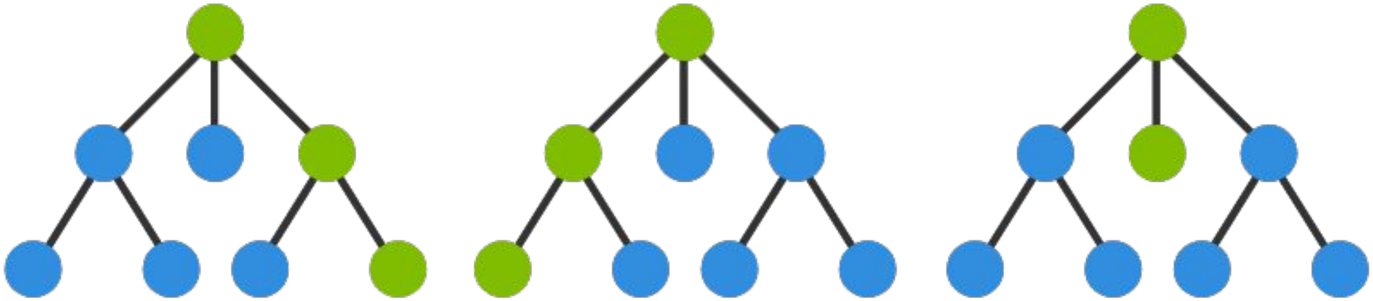




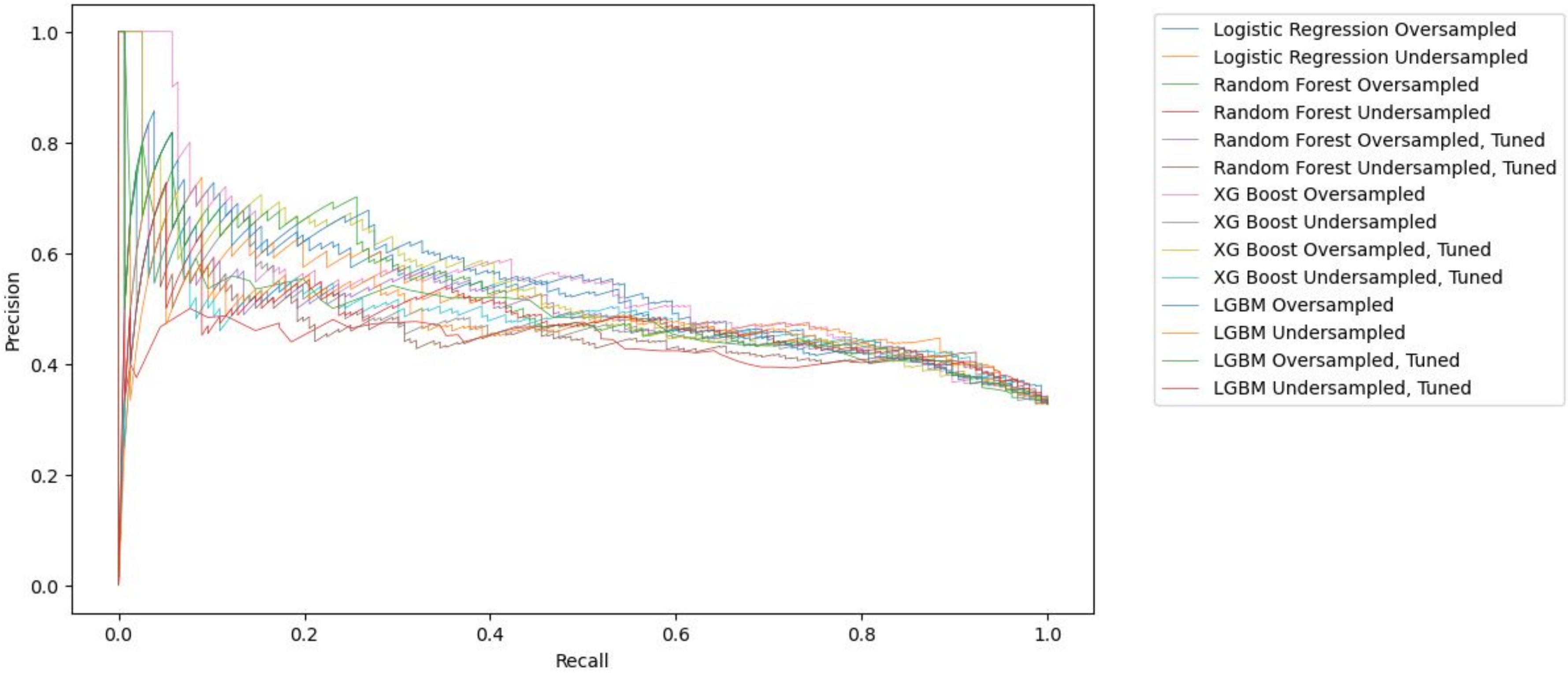
The Models

- Base model: Linear Regression
  - Poor performance

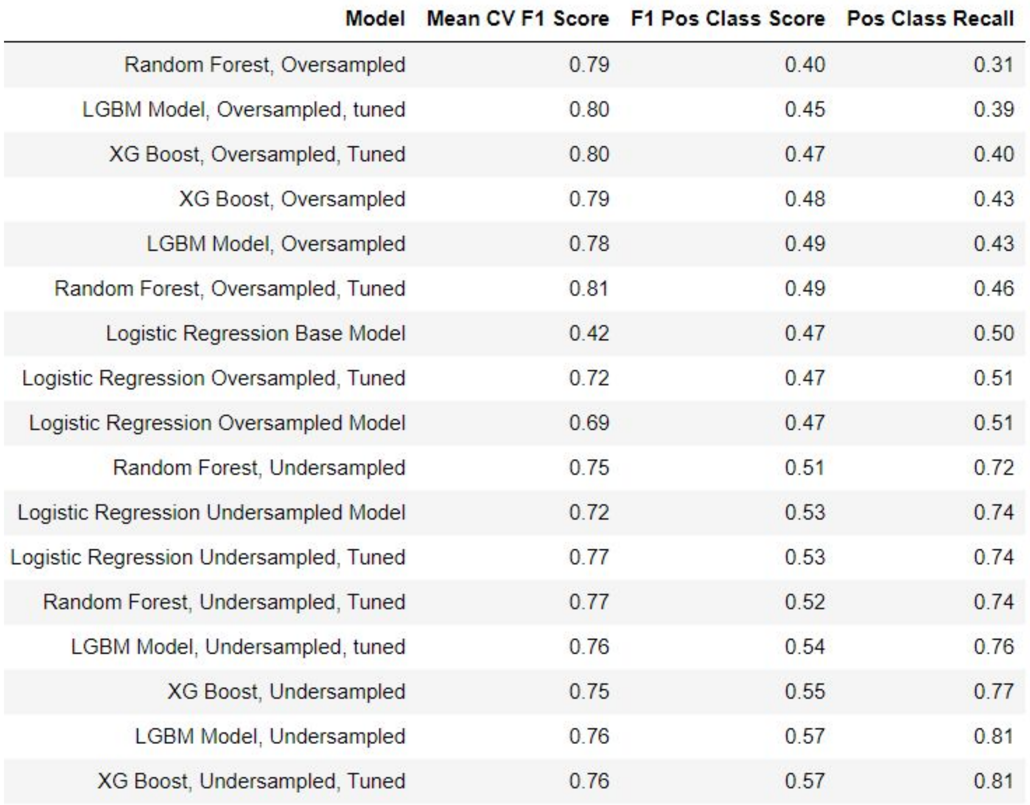
- Additional models: Random Forest, XG Boost, LGBM
  - Undersampled & Oversampled
  - Baseline model & hyperparameter tuning for all model types



Precision Recall Curves



Model Comparisons

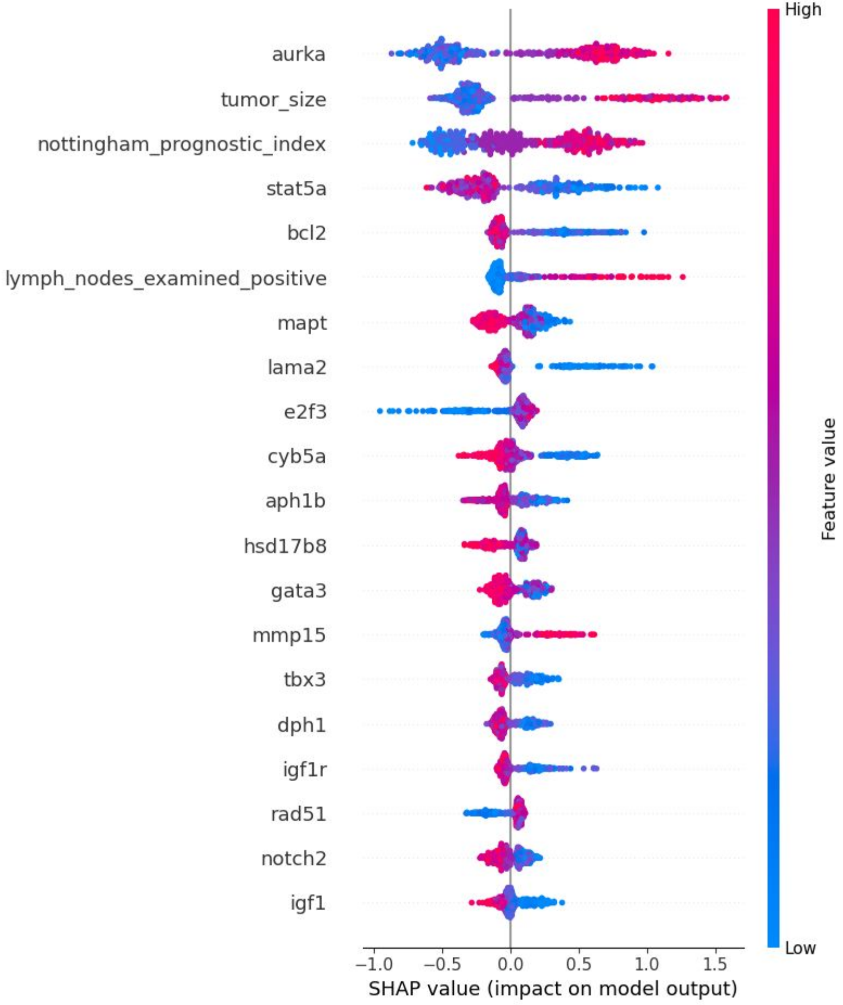


Model	Mean CV F1 Score	F1 Pos Class Score	Pos Class Recall
Random Forest, Oversampled	0.79	0.40	0.31
LGBM Model, Oversampled, tuned	0.80	0.45	0.39
XG Boost, Oversampled, Tuned	0.80	0.47	0.40
XG Boost, Oversampled	0.79	0.48	0.43
LGBM Model, Oversampled	0.78	0.49	0.43
Random Forest, Oversampled, Tuned	0.81	0.49	0.46
Logistic Regression Base Model	0.42	0.47	0.50
Logistic Regression Oversampled, Tuned	0.72	0.47	0.51
Logistic Regression Oversampled Model	0.69	0.47	0.51
Random Forest, Undersampled	0.75	0.51	0.72
Logistic Regression Undersampled Model	0.72	0.53	0.74
Logistic Regression Undersampled, Tuned	0.77	0.53	0.74
Random Forest, Undersampled, Tuned	0.77	0.52	0.74
LGBM Model, Undersampled, tuned	0.76	0.54	0.76
XG Boost, Undersampled	0.75	0.55	0.77
LGBM Model, Undersampled	0.76	0.57	0.81
XG Boost, Undersampled, Tuned	0.76	0.57	0.81

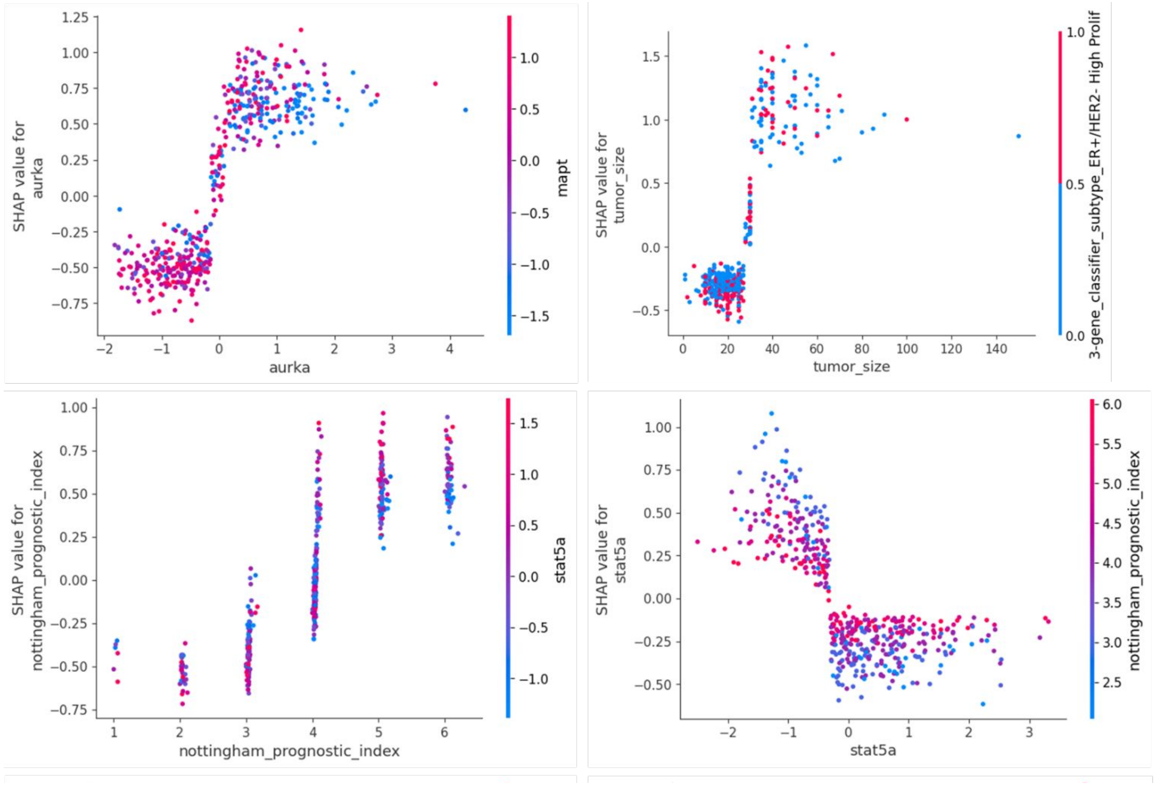
Best Model!

SHAP Summaries

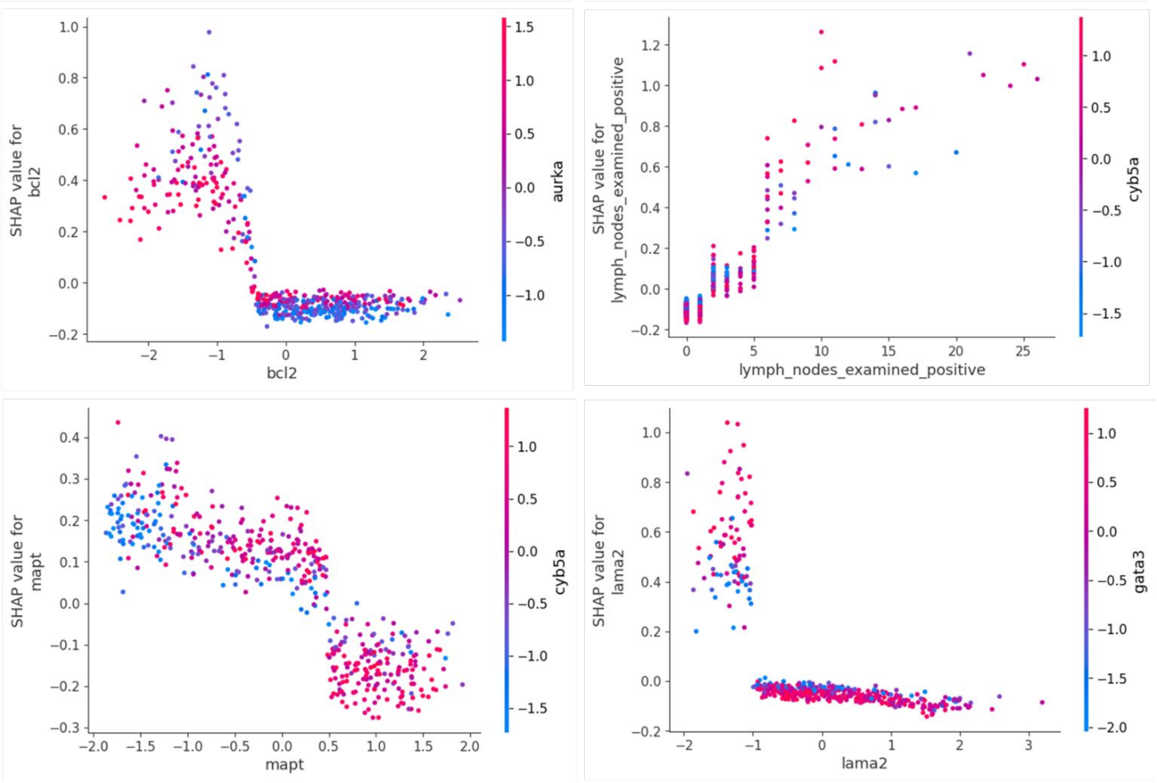
- Higher SHAP = more contribution towards death from cancer.
- Shape & color change at SHAP = 0
- Most positive SHAP values associated with down-regulated genes
- Genes with largest impact:
  - AURKA
  - STAT5A
  - BCL2
  - LAMA2



SHAP Breakdown by Feature



SHAP Breakdown by Feature



Conclusions

- Best Model: Undersampled & Tuned XG Boost: 0.81 recall score
  - Useful model for correctly identifying cancer death
  - High false positive rate

- Gene expressions can contribute towards patient outcomes
  - Down-regulated genes more associated with cancer death
  - The interactions of gene expressions are important

Future Work

- Remove genetic mutation data to make more efficient model

- Expand genetic mutation data to see if adding in the specific mutations makes a better model

- Model based on survival time instead of binary outcome

Recommendations

- Use model to identify patients more likely to die of breast cancer
- Conduct literature review and further genetic research on the identified genes of interest