Springboard --- DSC

Capstone Project 2

Predicting Patient Outcomes with Breast Cancer Gene Expression Data

By Annie Erbsen

July, 2023

1. Introduction

Breast cancer is the most frequently occurring and the leading cause of cancer-related deaths in women. An important part of the decision making process for cancer patients is the accurate estimation of prognosis and survival duration, but the reality is that these are hard to predict because gene expression greatly impacts these metrics. Breast cancer patients with the same stage of the disease and the same clinical attributes can have different treatment responses and overall survival outcomes. This difference in outcomes may be attributed to differences in gene expressions for specific genetic mutations.

In this project I used the clinical data, z scores, and genetic mutation data for 1904 breast cancer patients to predict if a patient with primary breast cancer will die of disease or not, and determined which features most impact outcomes. The XG Boost model (tuned & undersampled) was able to correctly identify patients who will die of cancer in 81% of cases, and I identified 16 genes whose expressions have a measurable contribution towards survival.

The data is from the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) database, which is a Canada-UK project. The dataset was collected by Professors Carlos Caldas from Cambridge Research Institute in the UK and Professor Sam Aparicio from the British Columbia Cancer Centre in Canada. The maximum follow-up time for this data was 351 months.

2. Approach

2.1 Data Acquisition and Wrangling

The METABRIC dataset was sourced from Kaggle: https://www.kaggle.com/datasets/raghadalharbi/breast-cancer-gene-expression-profiles-metabric. The raw data consisted of 1904 rows and 693 columns. The columns had 3 different types of data: clinical attributes (30 columns), gene expression z scores (489 columns), and genetic mutations (174 columns).

The clinical attributes were missing quite a lot of data, but I only dropped one row, where the target feature (death from cancer) was missing. Otherwise, I dropped columns with redundant clinical features, and imputed the rest using the median value. Most of the missing data was from Cohort 4. The z scores data was clean, but the genetic mutation data was not useable in its raw format. There was a range of 0-10 mutation names that were listed for each patient and column; exploding the columns so that each unique value had its own column would have made the dataset much larger and would likely compound overfitting that comes with 'wide and short' datasets. For this project I decided to instead make each value in this section of the dataset a binary entry indicating if mutations are present or not, with the plan to investigate the individual mutations if this part of the dataset showed anything of note.

2.2 Storytelling & Inferential Statistics

The goal of this project was to predict patient outcomes; however, this dataset had 3 different survival metrics: Overall survival, death from cancer, and survival months. As I was looking for a binary outcome for cancer deaths, I kept the 'death from cancer' feature and dropped the other survival metrics.

Looking at 'death from cancer', 42.1% were living at the end of the 351 month study, 32.7% died of disease, and 25.2% died of other causes. For the purpose of this project, I grouped together the living and death from other causes patients, as I was interested in whether or not a patient dies of cancer specifically.

I computed a correlation matrix for the clinical attributes, and saw that there were no strong correlations between outcomes and features. The 'strongest' correlations with patient death by cancer were the number of positive lymph nodes (corr=0.24) and the Nottingham Prognostic Index (corr=0.27). Tumor stage was very weakly correlated with patient outcomes, but when plotting the outcome vs features, I could see that there was a clear pattern, so I included it below.

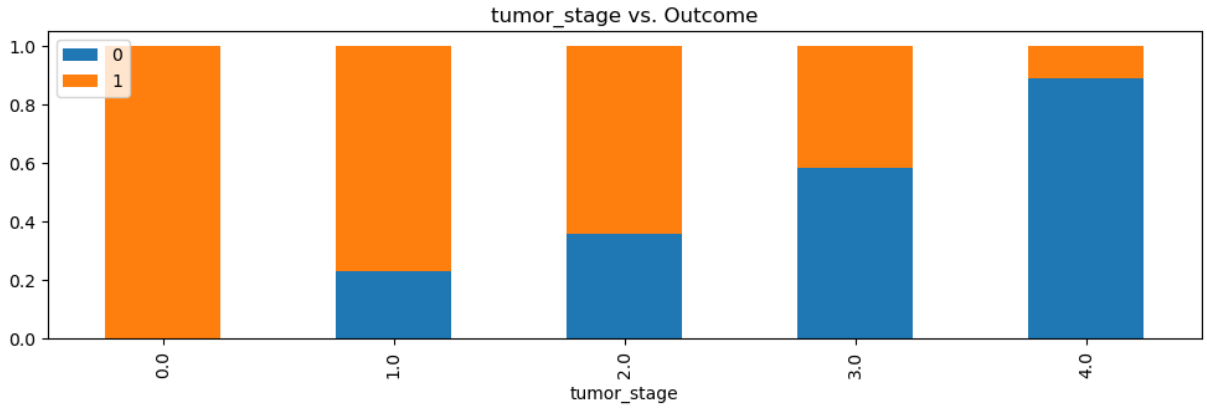

Figure 1: Tumor Stage vs. Outcome. Blue (0) indicates death from cancer.

As you can see in Figure 1, as the tumor stage increases, the proportion of patients who died of disease increases. This was unfortunately the feature missing 26% of the data, as Cohort 4 did not record this value for their patients.

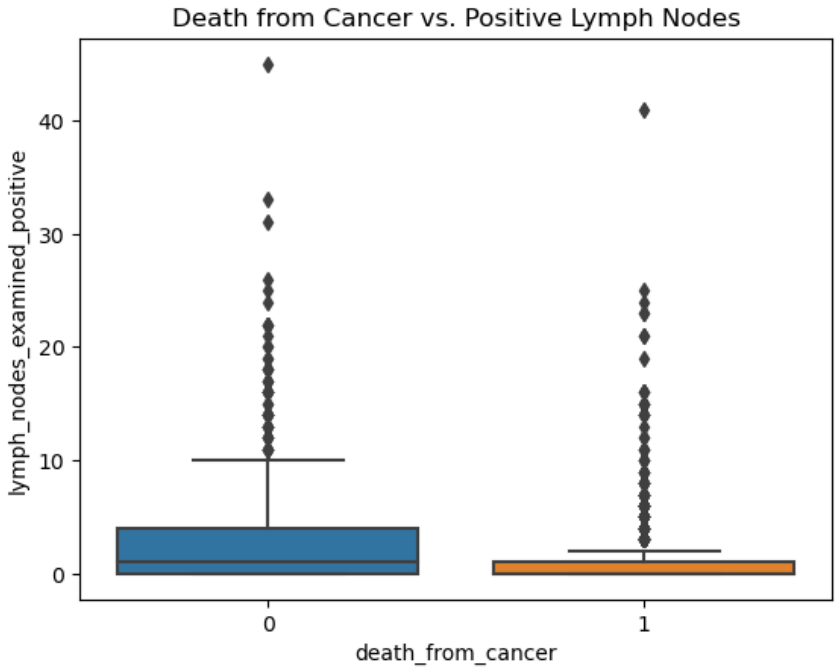

Figure 2: Death from cancer vs positive lymph nodes. Blue (0) indicates death from cancer.

The number of positive lymph nodes appears to be related to outcome. Although there are a number of outliers, we can see in Figure 2 that that the patients who did not die of disease usually had under 3 positive lymph nodes, but those who did die of disease tended to have more.

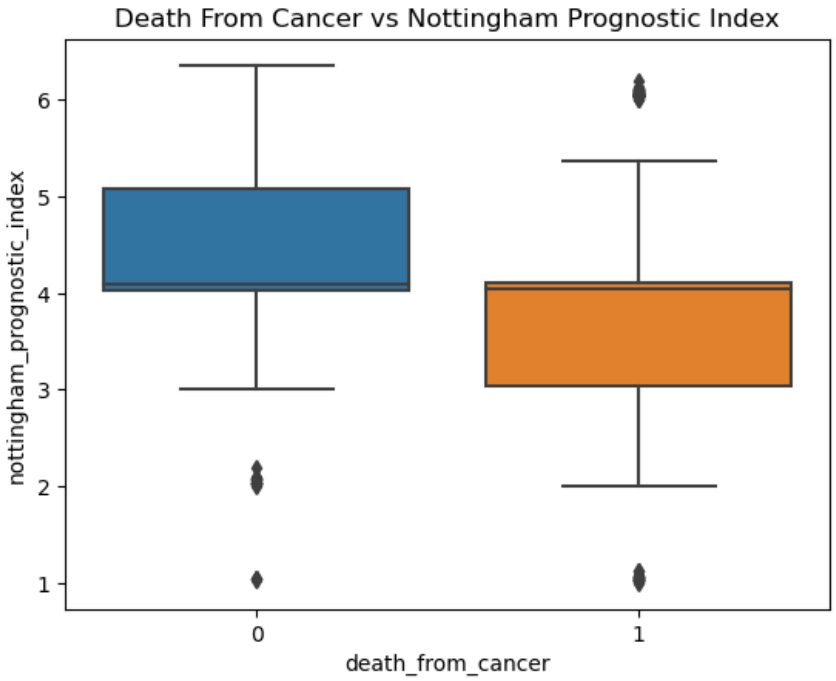

Figure 3: Death from cancer vs Nottingham Prognostic Index. Blue(0) indicates death from cancer.

The Nottingham Prognostic Index is an interesting metric; note in Figure 3 that the median Index value is similar for all outcomes, but that the data are skewed in opposite directions. From this plot, we can see that patients with a Nottingham Prognostic Index over 4 are more likely to die of disease, and those with an Index under 4 are more likely to survive.

Regarding gene z scores, only 34% of the z score columns have a corresponding column for mutations, but all but 5 of the mutation columns have a matching z score. The z scores ranged from 20.4 to -7.2, with positive z scores meaning that the gene is up-regulated, and negative z scores meaning that the gene is down-regulated.

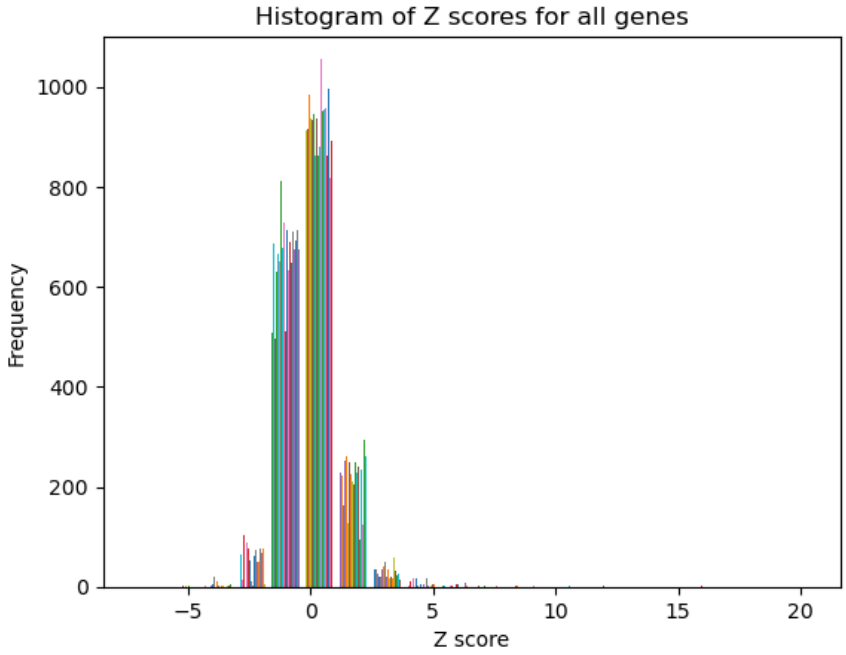

Figure 4: Z score distribution of gene expressions

As we can see above, almost all of the z scores fall between -3 and 4, and mostly between -2 and 2. I suspected that the outliers could impact outcomes, and computed the p values for z scores over 4 and under -4 with the Fisher's Exact Test. Our null hypothesis is that extreme z scores have no impact on patient outcomes and found that the p value for z scores under -4 was 0.238, and z scores over 4 had a p value of 0.22. These are not low enough to reject the null hypothesis. However, upon inspecting the correlation of z scores under -4 and outcome, it does appear that certain genes are weakly correlated to outcomes (Figure 5). Z scores over 4 didn't show any notable correlation with outcome.

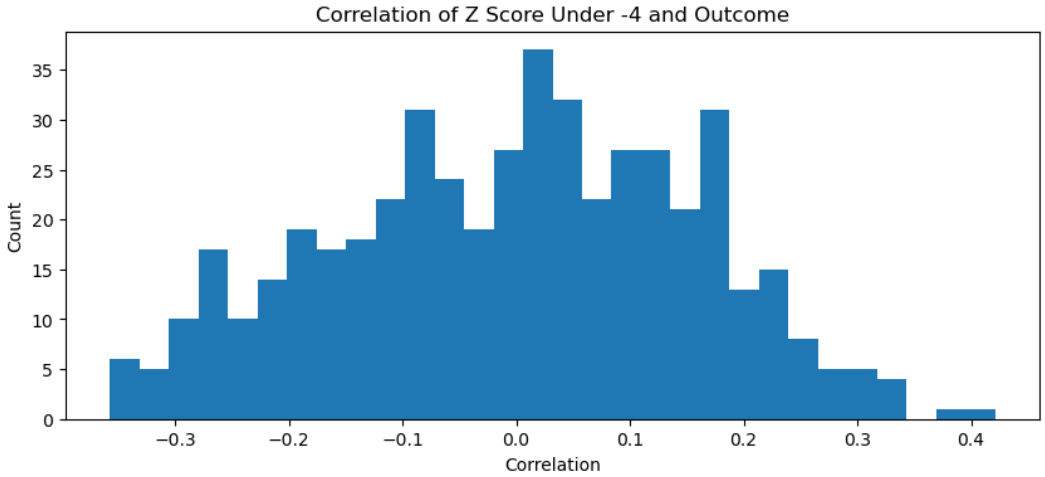

Figure 5: Correlation of Z Score under -4 and outcome.

To find which down-regulated genes have a meaningful correlation with outcomes, I did a Mann Whitney U Test, as the data were not normally distributed. I found that there are 127 genes in our dataset that have a p value under 0.05, which means that our null hypothesis can be rejected in those cases, and the genes may impact survival outcomes.

The 10 genes with the lowest p values were CYB5A, JAG2, NFKB2, SETDB1, MUC16, PRKCZ, RAN, RB1, SETD1A, and HSD17B4.

2.3 Baseline Modeling

I chose logistic regression as my baseline model. I performed this model on 4 separate train/test splits: the full dataset, PCA with 25 components, PCA with 190 components (the number to make the dataset not 'wide and short'), and PCA with 360 components (accounts for 90% of variance).

F1 scores for predicting cancer death:

Full data set: 0.46
PCA-25: 0.3
PCA-190: 0.46
PCA-360: 0.44

With the exception of the test/train split with 25 PCA components, these iterations all performed similarly with the logistic regression model. As PCA reduces interpretability, I decided to move forward with the baseline model using our full dataset with no PCA.

2.4 Extended Modeling

The performance of the baseline Logistic Regression model was likely poor due to imbalanced classes; the class we were interested in (class 1 the patients who died of disease), were the minority class. As we saw earlier, 33% of our data belongs to class 1. As a solution to this, I oversampled and undersampled the data, and ran all models on the oversampled and undersampled data.

I performed oversampling via the Synthetic Minority Over-Sampling Technique (SMOTE) which synthesizes new data from within the minority class. I undersampled with NearMiss. Both of these techniques should minimize overfitting.

In addition to the baseline Logistic Regression model, I created models with Random Forest, XGBoost, and LGBM boost. I chose these additional models because they work well with binary classification, can handle non-linear relationships between variables, and are suited for a large numbers of features. XGBoost in particular is less prone to overfitting. I ran each model with oversampling and undersampling, and then performed hyperparameter tuning and ran the oversampled and undersampled sets again with the new parameters. For each iteration of a model, I also performed 5 fold cross validation. Because the goal of this project is to identify true cancer deaths, I used the class 1 F1 score and recall values as the primary metrics of interest, and also checked the confusion matrix to make sure that the true positives were maximized, and the false negatives were minimized. I am not as concerned with false positives, as living longer than expected is a much better situation for patients than having a more aggressive cancer/dying sooner from cancer than expected.

The best model was the undersampled & tuned XG Boost model. It should be noted that this model had the same CV, F1 and recall scores for class 1 as the undersampled LGBM model, but that the undersampled & tuned XG Boost model performed slightly better for class 0; due to it being overall the best performer, it was selected as the model used moving forward.

However, the class 1 F1 score was only 0.57, which is only slightly better than flipping a coin. Upon inspecting the precision and recall scores, the recall is actually quite good at 0.81, but with precision scores of 0.44; this model does correctly identify most true positives, but there is a high false positive rate. This could be counteracted by further testing in the clinic, and closely monitoring disease progress.

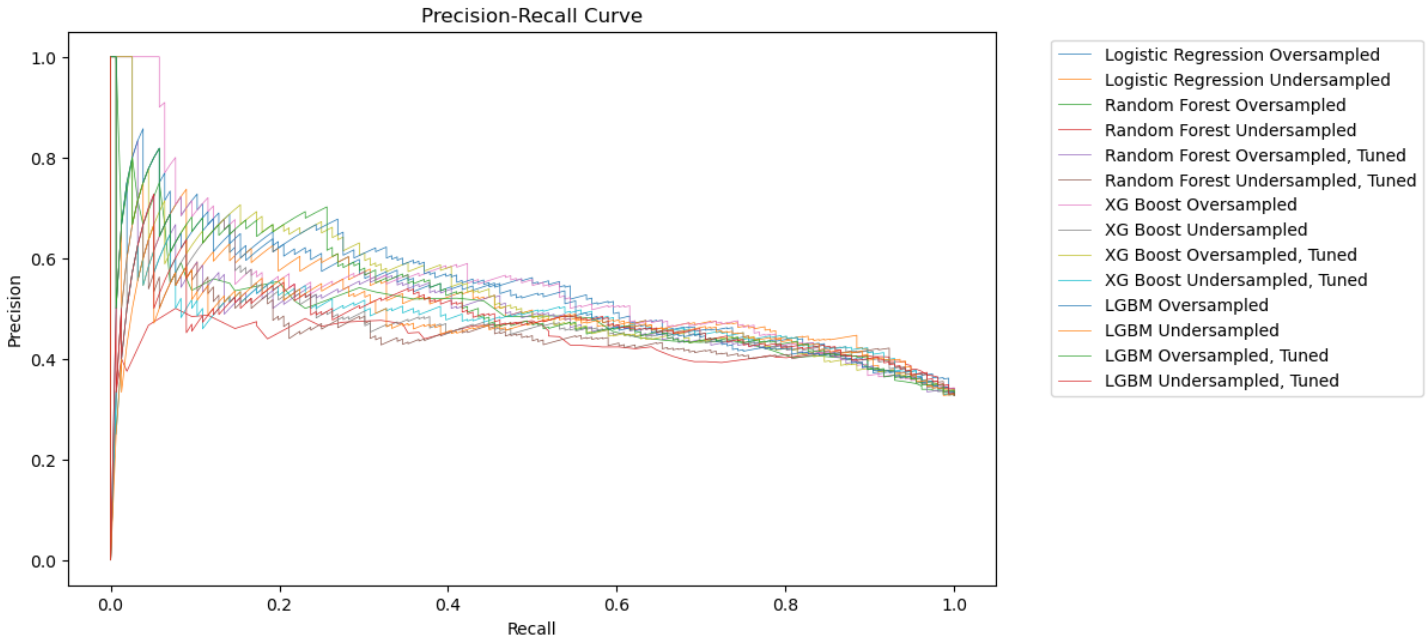

Figure 6: Precision Recall Curves for all models.

3. Findings

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

Figure 7: Performance metrics for all models, ranked from worst to best

After selecting the undersampled & tuned XG Boost model, I performed SHAP (Shapley Additive Explanations), which in our case was able to quantify how much each feature contributes towards patient death. We started off with 693 features in the original dataset, and by looking at the summary plot below (Figure 8) it is easy to see that there are a relatively small number of features that potentially have a large contribution towards patient outcome. Interestingly, none of the genetic mutation data appears here, indicating that the number of mutations on a gene may not impact the outcome very much, but most of the significant features were gene z scores. There were also a few clinical attributes that look to be important, and they are what I would expect,

as these are part of how doctors already anticipate patient outcomes: tumor size, Nottingham Prognostic Index, and the number of positive lymph nodes. These metrics were also identified in section 2.2 as the features having the highest correlation with outcomes.

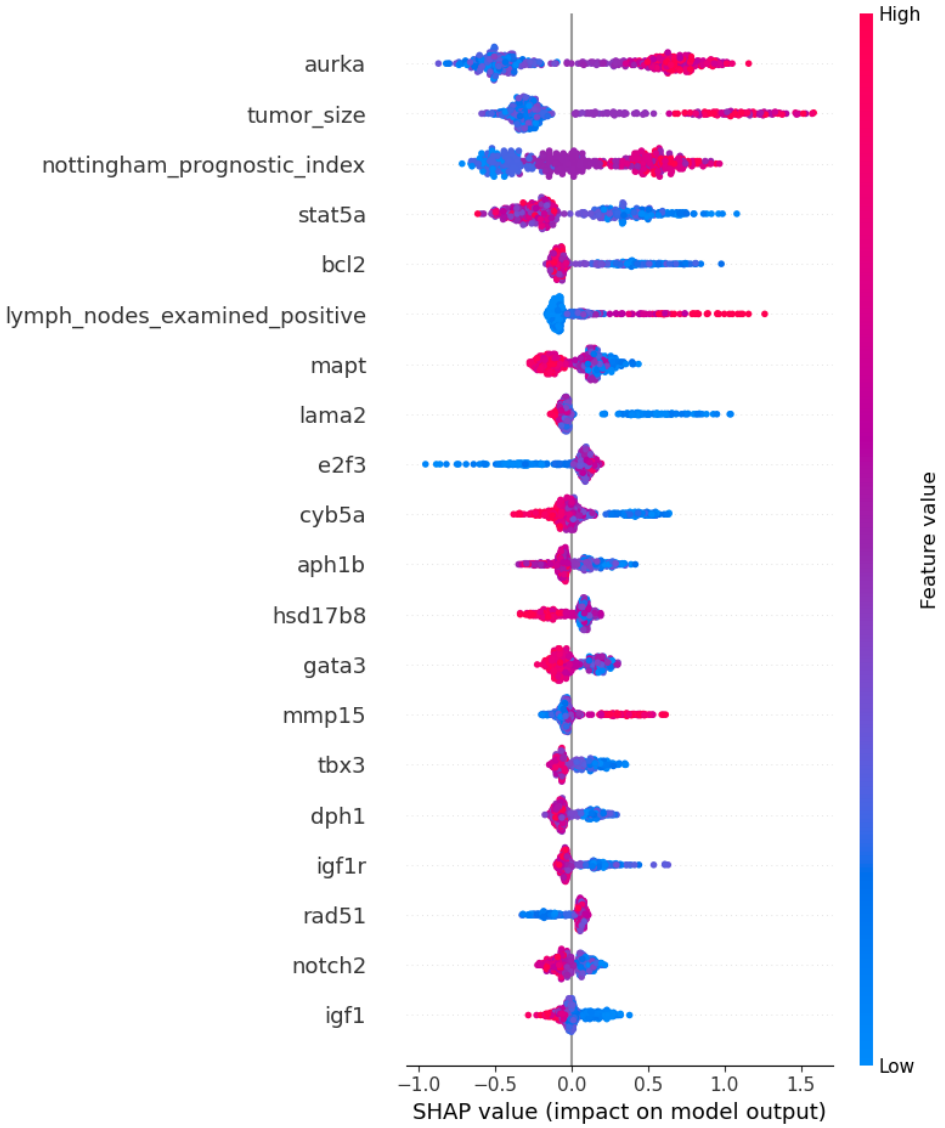

Figure 8: Features with highest SHAP values, corresponding with contribution towards death from cancer

In Figure 8, note that the positive SHAP values contribute towards death from cancer, and the negative SHAP values contribute towards *no* death from cancer (in this case, survival or death from other causes). The feature values show a clear break right around a SHAP value of 0 for all features in the above figure, indicating that they could be important prognostic indicators.

The z scores of certain genes measurably contribute towards patient outcome, which can be seen on the plots of the z scores vs SHAP values of individual genes in Figure 9. For most of the genes inspected in the figure above, the data falls into 2 distinct clusters: those with SHAP values over 0 (contributing towards cancer death), and those with SHAP values under 0 (contributing towards no cancer death, which in this case was either survival or death from other causes). In these cases, we can clearly see that that the range of the z score is very important. It should also be noted that some genes interact and need to be looked at together, as the presence of a certain range of z scores for one gene can correlate with higher or lower SHAP values for

another gene. For example, for both the APH1B and CYB5A genes, most of the data points with low SHAP values also had AURKA gene z scores under 0.

According to SHAP on the tuned and oversampled XG BOOST model, the following genes contribute towards cancer death, with SHAP values over 0.1:

High z scores associated with high SHAP values:

- AURKA
- E2F3

Low z scores associated with high SHAP values:

- STAT5A
- BCL2
- MAPT
- LAMA2
- CYB5A
- APH1B
- HS17b8
- GATA3

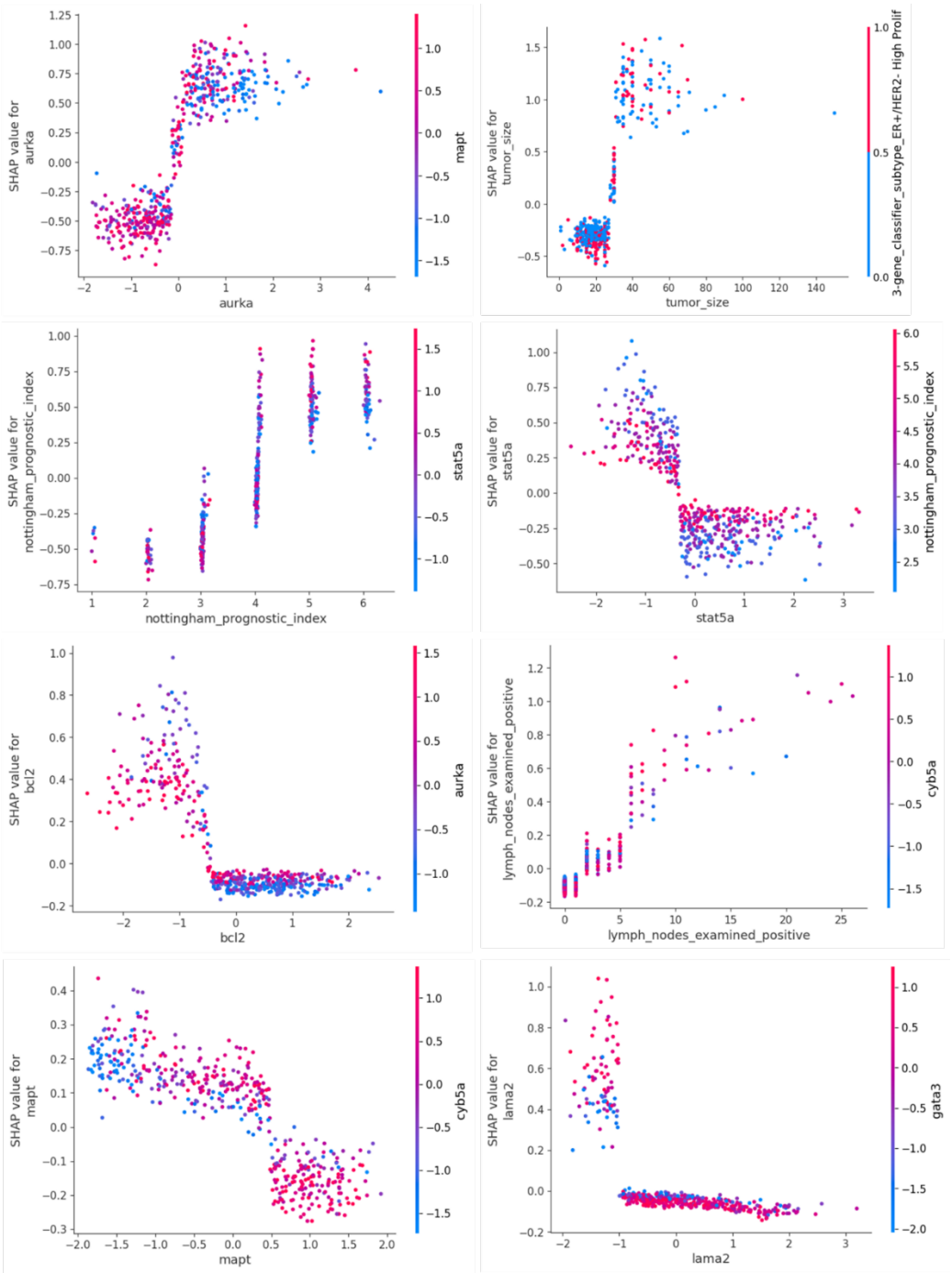

Figure 9: Plots of z scores and gene interactions for the highest 8 SHAP values. The x axis on all plots is the z score of that gene.

Interestingly, when inspecting the correlation of z scores and outcome in the exploratory data analysis section, we saw that the negative z scores were more highly correlated with patient death from cancer. Upon cross-checking the genes with the highest 10 SHAP values (listed above) with significant correlation p values, the genes CYB5A, APH1B, AURKA, and LAMA2 appear in both lists. However, when evaluating how certain gene expressions contribute towards outcomes, we must consider that gene interactions also play an important role.

The gene CYB5A had the lowest p-value, indicating it is statistically likely to contribute towards patient death. There have been studies on CYB5A which show a connection with lowered patient life expectancy, mostly in the context of renal, pancreatic, and liver cancer[5,6]. However, there is not much published research about this

gene and breast cancer. Negative z scores for both CYB5A and MAPT genes correspond with higher SHAP values, as you can see in Figure 9.

The AURKA gene overall had the highest SHAP value, as well as a significant p value, and research has shown that it is linked with higher rates of breast cancer recurrence. It should be noted that AURKA was one of the only genes where a *positive* z score is linked with increased patient death rates. However, as you can see in Figure 8, AURKA interacts strongly with MAPT, and the negative patient outcomes usually have a positive z scores for AURKA, and negative z scores for MAPT.

Mutations on the LAMA2 gene are known to be associated with poorer prognosis[8], and as we can see above in Figure 8, z scores over -1 have almost no contribution towards outcome, but z scores between -2 and -1 have SHAP values between 0.2 and 1.2. Almost all of the SHAP values over 0.5 were also associated with positive GATA3 z scores.

4. Conclusions and Future Work

The tuned & undersampled XG Boost model correctly identifies breast cancer patients who will die of disease (true positives) 81% of the time, although there is a high false positive rate. This can be counteracted by further testing and tracking of disease progress. This could be a useful tool for clinicians or researchers to better tailor treatments for patients, potentially resulting in better patient quality of life.

Additionally, the down-regulation of a number of genes (and the up-regulation of 2 genes) corresponds with more contribution towards patient death of disease. A more in-depth literature review of gene regulation in breast cancer should be done to see if any new conclusions made in this project.

In the process of building a model to predict cancer death and discover contributing features, questions come to mind that could potentially improve the model. In future work, it would be interesting to investigate the following questions:

1. What does the chosen model look like when we remove the genetic mutation section of the data? Does this change the output of the model at all? Simplifying the data needed for the model would make it easier and less expensive to implement for future patients.
2. Modeling based on the survival time could give different results than binary outcome.
3. I would like to look more deeply at the genetic mutation section of the data. As noted earlier, I simplified the original dataset with the full genetic mutations to be binary (mutations are present or not). Looking back at the exploratory data analysis this project, I see that some of the genes with high SHAP values are also among the most commonly occurring genes with mutations. Would expanding the original genetic mutation data with the 'get dummies' method make our model better or worse? Would we be able to pull more insights from SHAP with this information included? Are certain mutations associated with higher or lower SHAP values?

5. Client Recommendations

I recommend that the selected model be used in a clinical setting where patients have the clinical and genetic data available to make this survival prediction. It should not be used as a stand-alone tool, but could be useful to identify patients at higher risk of dying of disease.

It is also recommended that further genetic research be done on the interactions between the identified gene expressions which measurably contribute towards patient outcomes. In particular, the genes to further inspect are AURKA, STAT5A, BCL2, MAPT, LAMA2, E2F3, CYB5A, APH1B, HSD17B8, GATA3, MMP15, TBX3, DPH1, IGFLR, RAD51, NOTCH2. Understanding these relationships would not only improve anticipating patient outcomes, but could also lead to improved treatments.

6. Consulted Resources

1. Original dataset & discussion
2. Machine Learning–Based Interpretation and Visualization of Nonlinear Interactions in Prostate Cancer Survival
3. Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival

   Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions
4. Polymorphisms in the carcinogen detoxification genes *CYB5A* and *CYB5R3* and breast cancer risk in African American women
5. Human Protein Atlas
6. Role of CYB5A in Pancreatic Cancer Prognosis and Autophagy Modulation
7. GATA-3 expression in breast cancer has a strong association with estrogen receptor but lacks independent prognostic value
8. Molecular features of untreated breast cancer and initial metastatic event inform clinical decision-making and predict outcome: long-term results of ESOPE, a single-arm prospective multicenter study