BREAST CANCER SURVIVAL MODELING

Utilizing data for breast cancer tumors to try and predict survival

DATA OVERVIEW

- Source: CBioPortal.org – Repository of large scale cancer genomics datasets
- Study: Nature 2012 & Nat Commun 2016
- 2509 Unique Cases of Breast Cancer Tumors
- 35 Columns

EDA METHODOLOGY

- Out of 2,509 total cases, if all NaNs were dropped 1,092 cases would remain
- Created a column ‘died_survived’ that classifies patients who survived breast cancer or died of other causes as 0 and classifies patients who died of breast cancer as 1
- Compared the correlation of all initial columns to ‘died_survived’ and dropped those with correlation <=(+-0.10) from the data frame
- By dropping the low correlation columns (19) an additional 261 cases (1,353 total) were able to be included in the modeling

DROPPED COLUMNS

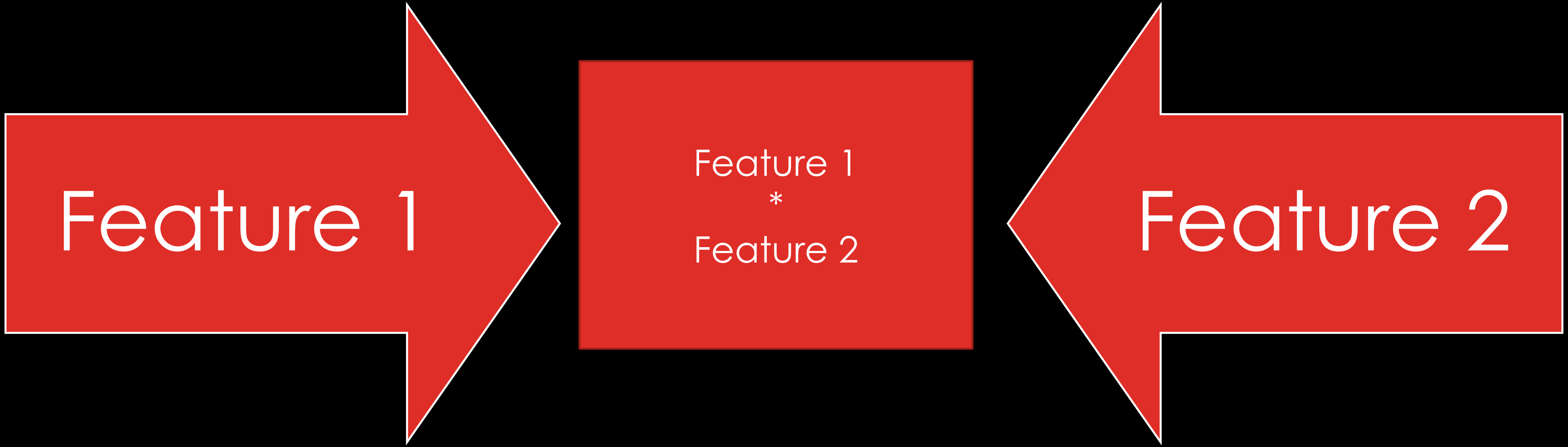
Column	Description
Integrative Cluster	4 Sub-types of breast cancer molecular classifiers
Cellularity	The proportion of cancer within the residual tumor bed
Mutation Count	Number of cancer mutations detected
Cancer Type Detailed	Specific type of breast cancer
Primary Tumor Laterality	Whether the tumor was aligned to the right or left
ER Status	Either positive or negative, whether or not the cancer cell grows in response to estrogen
ER status measured by IHC	Either positive or negative, whether or not the cancer cell grows in response to estrogen
Cohort	Group of patients (1 - 9)
HER2 status measured by SNP6	Human Epidermal Growth Factor Receptor 2. Approximately one in five breast cancers are driven by amplification and overexpression of HER2

Column	Description
Age at Diagnosis	Age at which breast cancer was diagnosed
Tumor Other Histologic Subtype	Subtype of tumor (8 types)
Hormone Therapy	Whether or not patient was treated with hormone therapy
Inferred Menopausal State	Whether patient was pre or post menopausal
Oncotree Code	Code for the type of breast cancer based on the OncoTree cancer classifier tree (6 Types)
Radio Therapy	Whether or not Radio Therapy was performed
Number of Samples Per Patient	How many samples were taken from the patient
Sample Type	All sample types were ‘Primary’
Overall Survival Status*	Living or deceased. Dropped because of ‘dead_survived’ column used for y variable
Overall Survival (Months)*	Length of time patient survived with breast cancer. Dropped due to co-linearity concerns

FEATURE ENGINEERING

- Utilized 2nd degree polynomial feature engineering to create new variables for potential inclusion in the model

- Ran all features through a Ridge Regression to see which has the most impact and included the top 4



Feature 1

Feature 1
*
Feature 2

Feature 2

FEATURES INCLUDED IN MODEL

Feature Name	Description
Chemotherapy	Whether or not Chemotherapy was performed
Neoplasm Histologic Grade	1 to 3 score of how fast growing and normal the cells are
HER2 Status	Human Epidermal Growth Factor Receptor 2. Approximately one in five breast cancers are driven by amplification and overexpression of HER2
Lymph nodes examined positive	The number of lymph nodes in which cancer was detected
Nottingham prognostic index	Determines prognosis following surgery. Calculated score using three pathological criteria. Lower is more survivable
PR Status	Whether or not the breast cancer cells have progesterone receptors
Tumor Size	Size of the tumor
Tumor Stage	Stage of the tumor (0 to 4)

Feature Name	Description
3-Gene classifier subtype	Used to identify the four primary molecular subtypes of breast cancer
Pam50 + Claudin-low subtype	Genetic classifier of breast cancer subtypes (7 subtypes)
Engineered Feature Names	**Description**
Nottingham * Claudin-low subtype_LumB	Combination of Nottingham prognostic index and Pam50 + Claudin-low subtype_LumB
PR Status * Claudin-low subtype_LumB	Combination of PR Status and Pam50 + Claudin-low subtype_LumB
Neoplasm * Pam50 + Claudin-low subtype_Her2	Combination of Neoplasm Histologic Grade and Pam50 + Claudin-low subtype_Her2
Chemo * Nottingham	Combination of Chemotherapy and Nottingham prognostic index

MODELING APPROACH

- Given that the model is attempting to predict whether or not a patient will survive breast cancer, we want to minimize both false positives and false negatives.

- False Positive: Model predicts the patient will die of breast cancer when they will actually survive. Measured by Specificity.

- False Negative: Model predicts the patient will survive breast cancer when they will actually die. Measured by Sensitivity.

- F1 Score: Measure of both Sensitivity and Specificity. Primary evaluation metric for this analysis.

MODELING RESULTS

Model	Test F1 Score	Train F1 Score	Specificity	Sensitivity	Accuracy	Precision	ROC AUC Score	Cross Val Score
Logistic Regression	44.9%	44.4%	89.7%	34.8%	71.1%	63.5%	64.5%	70.8%
Random Forrest	48.0%	87.7%	75.0%	46.7%	65.5%	49.1%	60.0%	65.1%
Bagging Classifier	43.4%	53.3%	64.3%	47.0%	58.4%	40.0%	61.7%	67.8%
Support Vector Classifier	40.9%	45.6%	90.6%	30.4%	70.2%	62.5%	62.5%	70.0%
AdaBoost	39.1%	89.5%	74.1%	36.5%	61.4%	42.0%	58.5%	NA
KNN	35.5%	47.3%	89.3%	26.1%	67.9%	55.6%	61.9%	70.5%
Decision Tree	26.0%	32.0%	94.6%	16.5%	68.1%	61.3%	58.8%	67.6%

Baseline Model = 33.85% Patients Die of Breast Cancer

CONCLUSIONS

- Logistic Regression is the best performing model overall with the highest accuracy and second highest F1 score

- The Bagging Classifier model has the best balance between Sensitivity and Specificity, but is not very accurate

- Random Forrest model has the best F1 Score but has a lower accuracy than the Logistic Regression Model

- Overall the models performed far better in Specificity than Sensitivity measures. There are few overall false positive results

ABOUT ME

- Recent Data Science Immersive Graduate from General Assembly Jun 2020
- 5 years experience in financial product development and management at Sate Street
- Built and managed the iNAV platform. Backup low touch Net Asset Value Calculation Tool for Mutual Funds
- 2 years experience as an Emerging Markets Stock Broker