Breast Cancer: How can we best extend life?

Vinu Baburaj, Matthew Greene, Hemashree Kilari,
Atharva Abhijit Kulkarni, Shashank Bettada Sathya Thirtha

Agenda

- Intro and Motivation
- Background
- Data Overview
- EDA
- Methods and Models
- Results
- Conclusions
- Further Work
- Resources

Introduction and Motivation

- Second most common cancer in women
  - Most prevalent among middle age to elderly
  - 280,000 new diagnosis/year
  - 42,000 deaths/year
- More Treatment options and better success with early detection

- **Goals:**
  - Discover which model best predicts the chance of survival
  - Analyze which treatment options may work best
  - What factors effect their chance of survival

Background

- Breast Cancer screening isn't medically recommended until the age of 40 and stops after 80
- Patient advocacy will skew the data(always fight for our loved ones; family history changes the argument)
- Dataset doesn't capture how many times a person is diagnosed
  - Remission vs Re-appear
- Doesn't capture other medical conditions that could affect treatment, survivability, or cause of death outside of the disease

Data Overview

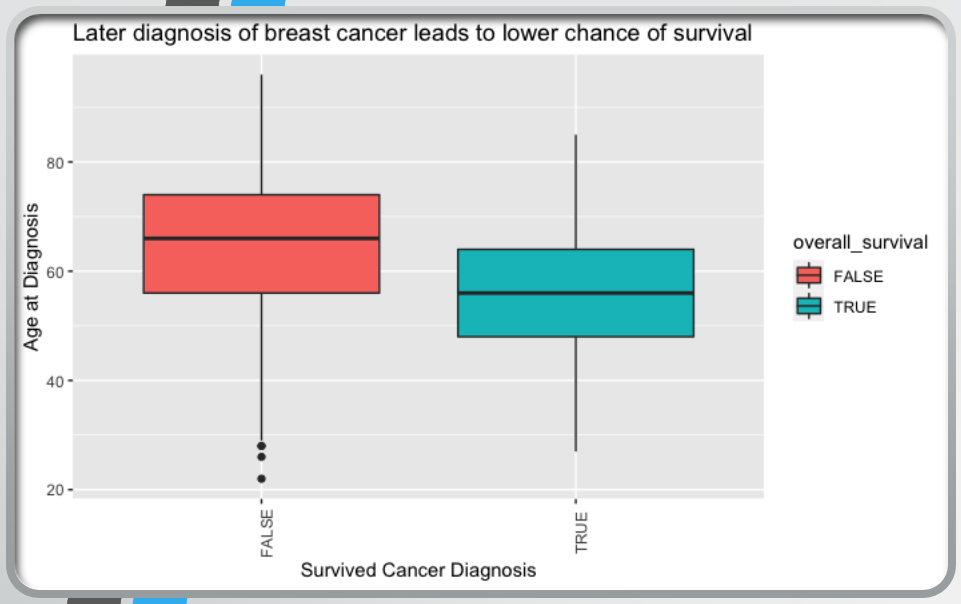
- Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) database
  - Canada-UK Project that targeted sequencing data of 1,980 primary breast cancer samples
  - dataset collected by Professor Carlos Caldas and Professor Sam Aparicio
- Dataset contains historical breast cancer data on tumors and the patients they came from
- Single File:
  - METABRIC_RNA_Mutation.csv
    - 1,904 rows and 693 columns before tidying and variable elimination
    - 662 Columns contain genetic and mutation information

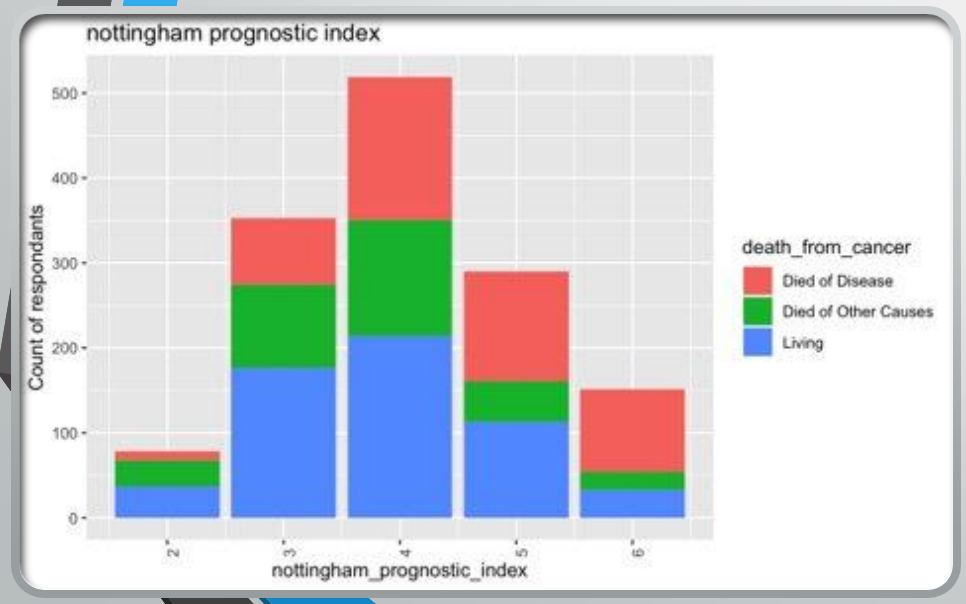
patient_id <dbl>	age_at_diagnosis <dbl>	type_of_breast_surgery <chr>	cancer_type <chr>	cancer_type_detailed <chr>	cellularity <chr>	chemotherapy <dbl>	pam50_+_claudin-low_subtype <chr>	cohort <dbl>	er_status_measured_by_ihc <chr>
2	43.19	BREAST CONSERVING	Breast Cancer	Breast Invasive Ductal Carcinoma	High	0	LumA	1	Positve
8	76.97	MASTECTOMY	Breast Cancer	Breast Mixed Ductal and Lobular Carcinoma	High	1	LumB	1	Positve
10	78.77	MASTECTOMY	Breast Cancer	Breast Invasive Ductal Carcinoma	Moderate	0	LumB	1	Positve
28	86.41	BREAST CONSERVING	Breast Cancer	Breast Invasive Ductal Carcinoma	Moderate	0	LumB	1	Positve
35	84.22	MASTECTOMY	Breast Cancer	Breast Invasive Lobular Carcinoma	High	0	Her2	1	Negative
36	85.49	MASTECTOMY	Breast Cancer	Breast Invasive Ductal Carcinoma	Moderate	0	LumA	1	Positve
60	45.43	BREAST CONSERVING	Breast Cancer	Breast Invasive Ductal Carcinoma	High	1	LumB	1	Positve
66	61.49	BREAST CONSERVING	Breast Cancer	Breast Invasive Ductal Carcinoma	High	0	LumB	1	Positve
100	68.68	MASTECTOMY	Breast Cancer	Breast Invasive Ductal Carcinoma	Low	1	Basal	1	Negative
101	46.89	MASTECTOMY	Breast Cancer	Breast Invasive Lobular Carcinoma	Moderate	0	Normal	1	Positve

Data Overview and Pre-processing

- 693 Total variables within the dataset
  - 31 Critical variables
- Tumor Stage has over 25% of its data missing
  - How can we mitigate the missing data problem?
  - Nottingham Prognostic Index
    - NPI = [0.2 * S] + N + G
    - S – Size of the lesion in centimeters
    - N - N is the node status: 0 nodes = 1, 1-3 nodes = 2, >3 nodes = 3
    - G - grade of tumor: Grade I = 1, Grade II = 2, Grade III = 3
  - Converted categorical variables with one hot encoding

Column	Percentage of Missing Data
tumor_stage	26.313
3-gene_classifier_subtype	10.714
primary_tumor_laterality	5.567
neoplasm_histologic_grade	3.782
cellularity	2.836
mutation_count	2.363
er_status_measured_by_ihc	1.576
type_of_breast_surgery	1.155
tumor_size	1.05



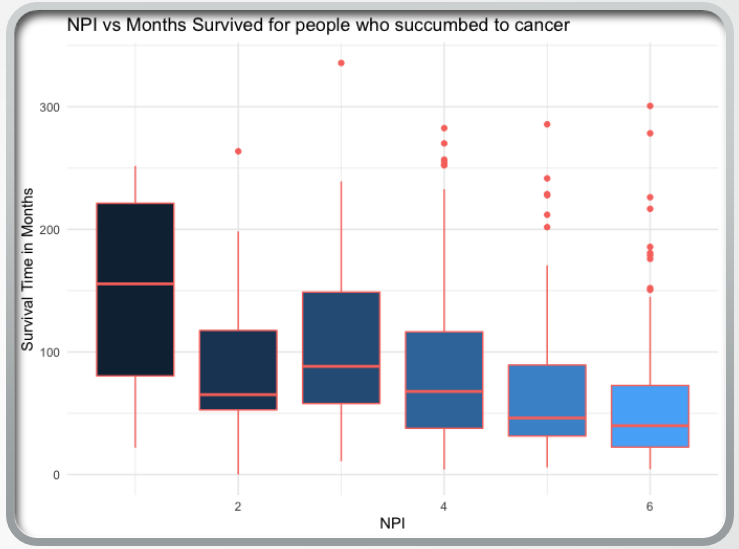


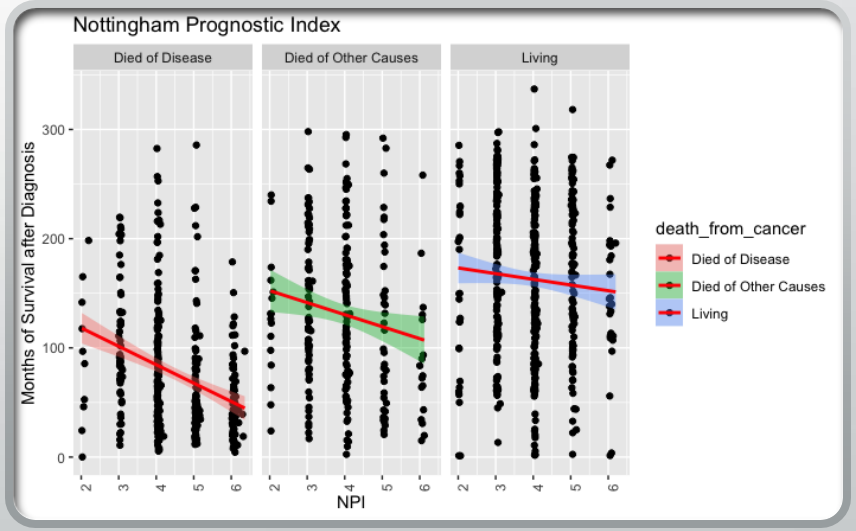
EDA: Detection and Identification is the First Step

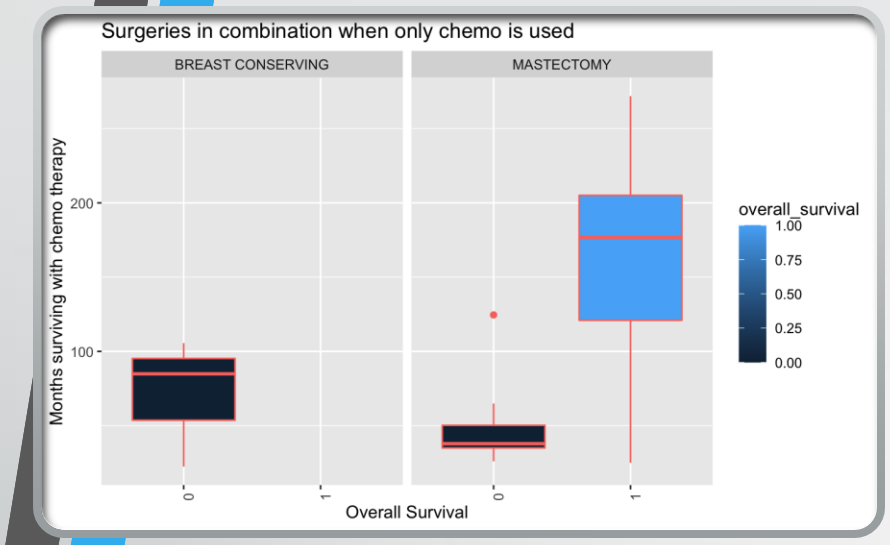
- Age appears to be most significant factor in determining if you survive the diagnosis
  - Data confirms our assumption about diagnosis age based on medical advice
- NPI also plays a significant factor, normally distributed
  - Initial trend as NPI increases, more patients die of Breast Cancer

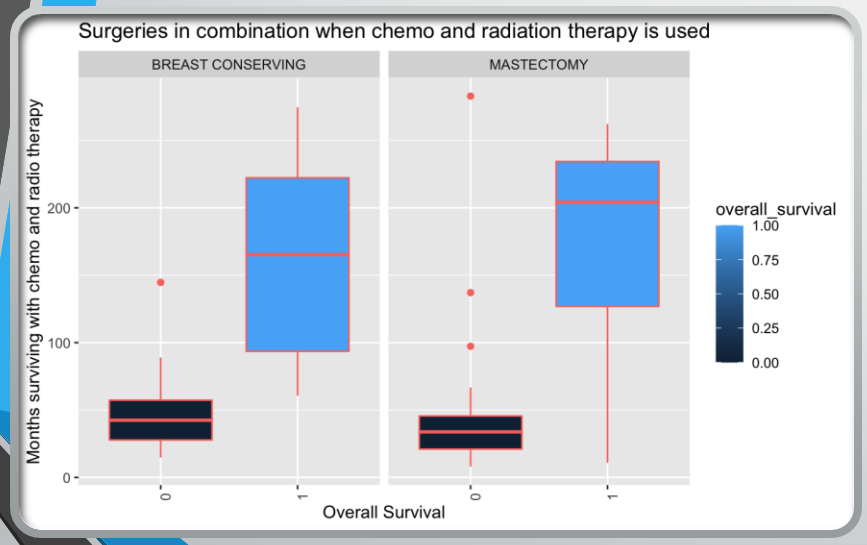
EDA: Detection and Identification is the First Step

- Steep trend slope for Died of Disease as NPI increases
  - Suggests that as NPI increases the life expectancy is low
- Died of Other Causes slope suggests that Breast Cancer may have contributed to the patient's death
- Living: still have declining slope
- For those still living, some missing data that would be helpful is what their current health status is
  - May be on death bed







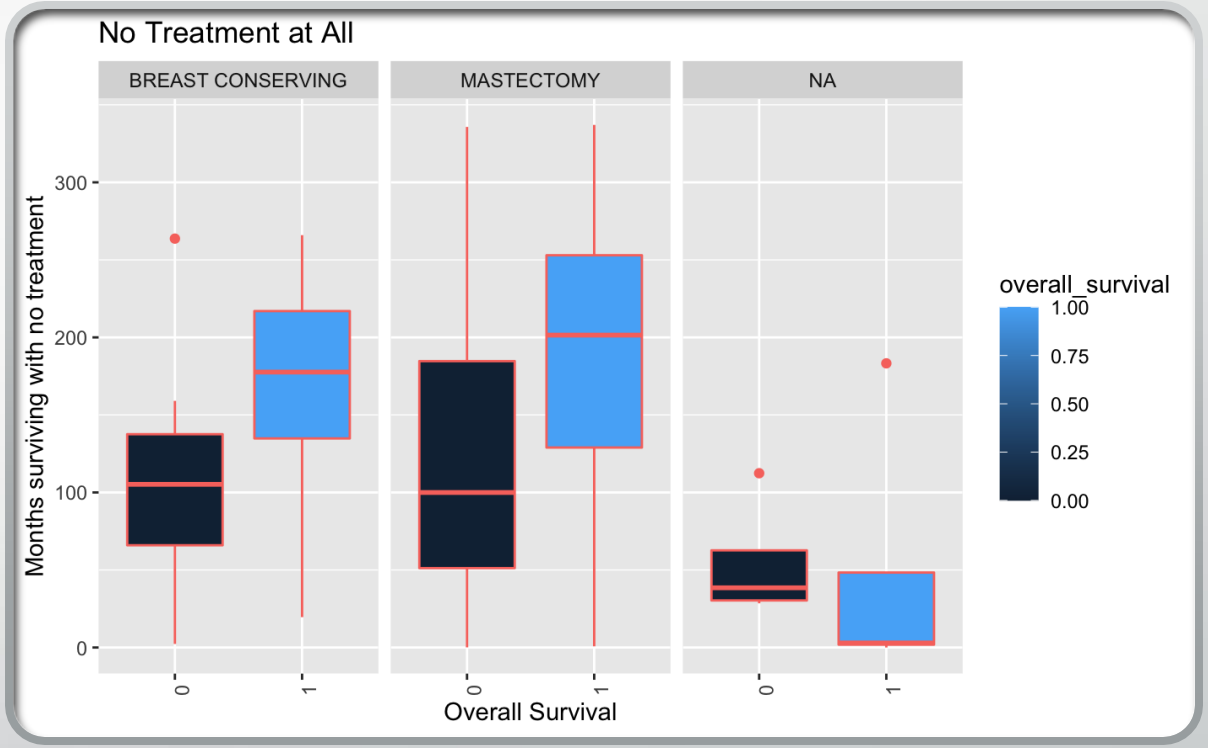


EDA: Treatment and Surgery

- Combinations of treatments and surgeries prevalent throughout and seem massively more effective.
- Treatments vs number of times diagnosed would be helpful here
  - Understanding what combo of procedures based on times and interval of diagnoses
  - Suggests multiple diagnosis and/or series of treatments
  - Not discernable from the data

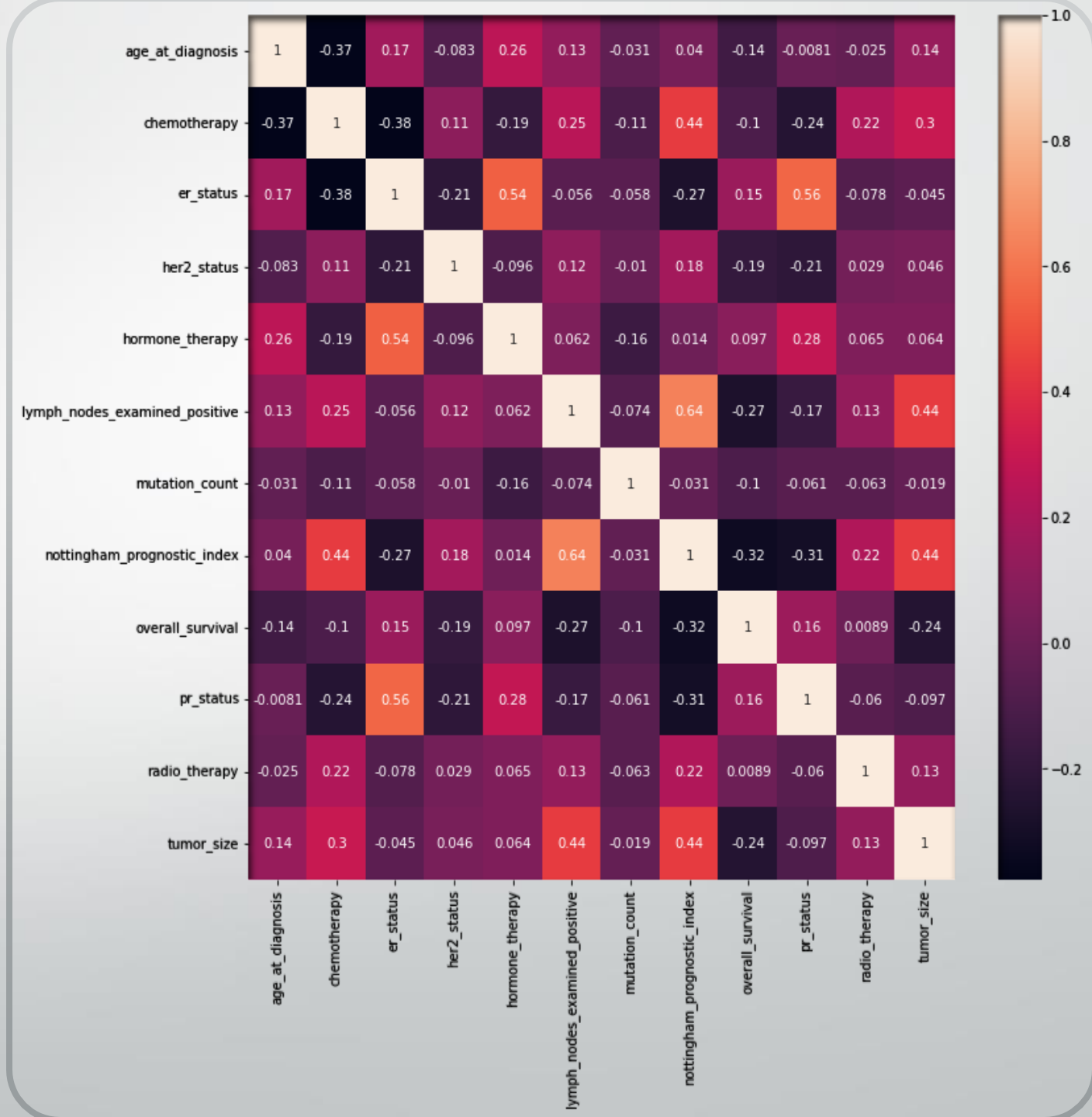
EDA: Surgery, No Treatment

- Only cases where there was either just surgery or no treatment or surgery
- Surgical options appear to play a large factor on longevity
- No surgery suggests 2 specific cases
  - Early diagnosis that doesn't have a treatment plan yet
  - Late diagnosis with little time to live



EDA: Feature Selection

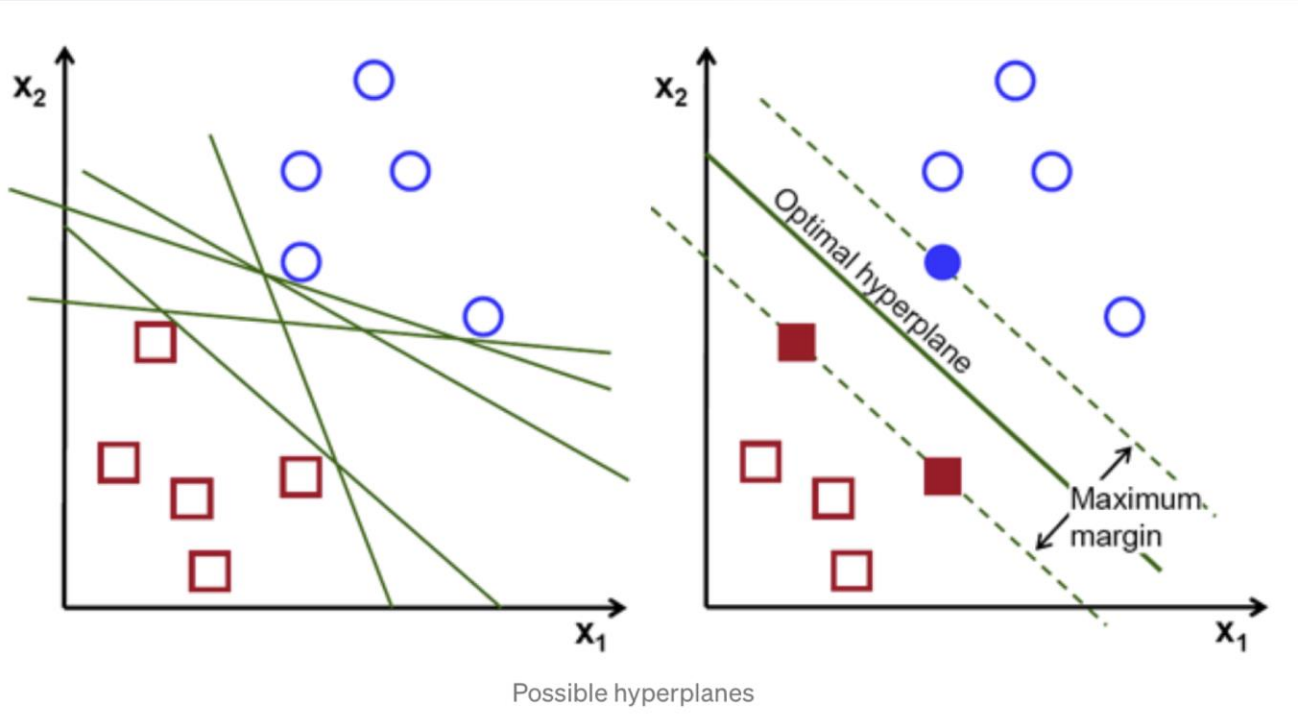
- Age at Diagnosis
- Chemotherapy, Hormone and Radio Therapy
- NPI
  - Product of tumor size, grade/stage, pos nodes examined

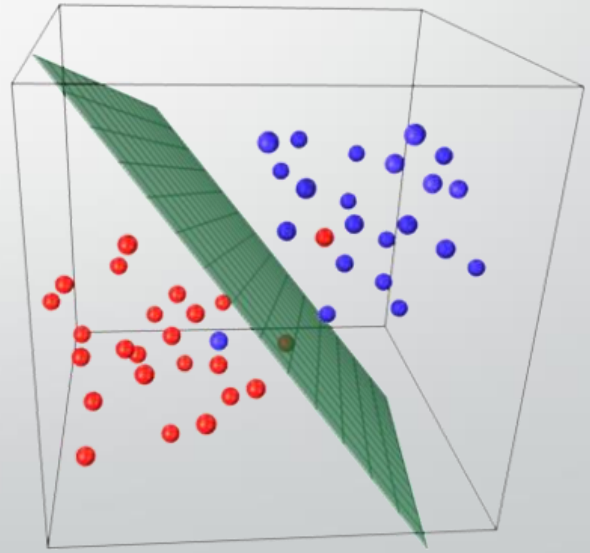


Methods and Models:

- **Support Vector Machine(SVM):**

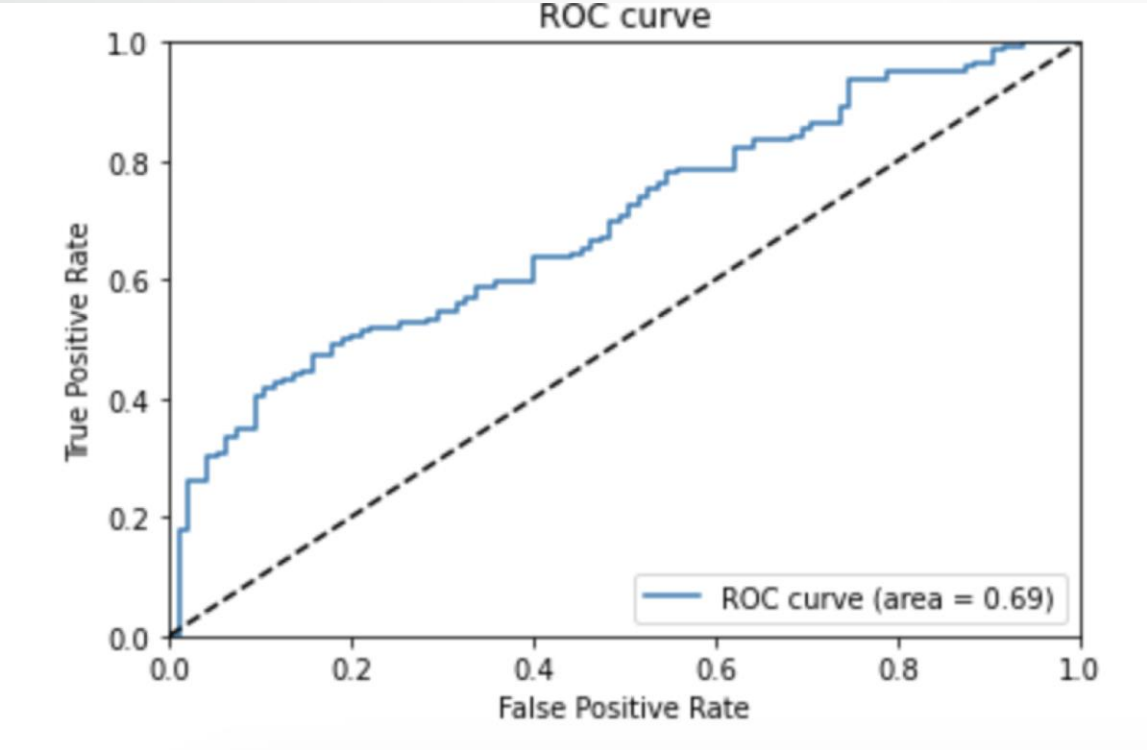
SVM constructs a **hyperplane** in multidimensional space to separate different classes. SVM generates optimal hyperplane in an iterative manner, which is used to minimize an error. The core idea of SVM is to find a maximum **marginal** hyperplane(MMH) that best divides the dataset into classes.





The accuracy of the SVM model : **69%**

ROC curve:

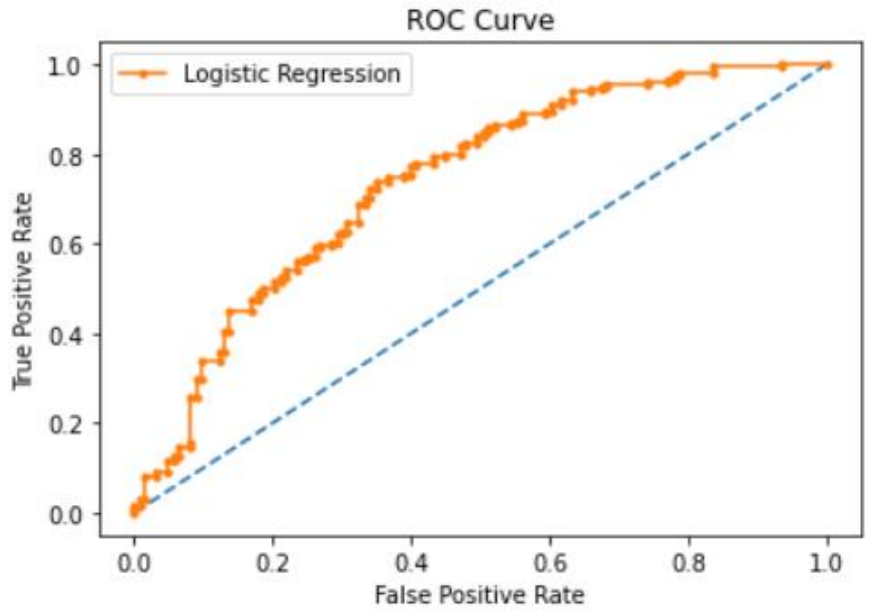


Logistic Regression:

- Logistic Regression is a supervised Machine Learning algorithm which determines the outcome in a binary format.
- The assumptions made which building a Logistic Regression model are:

1. The target variable must be binary.
2. The observations are independent.
3. There is no multicollinearity among the variables.

- The accuracy of the Logistic Regression model is 70%.

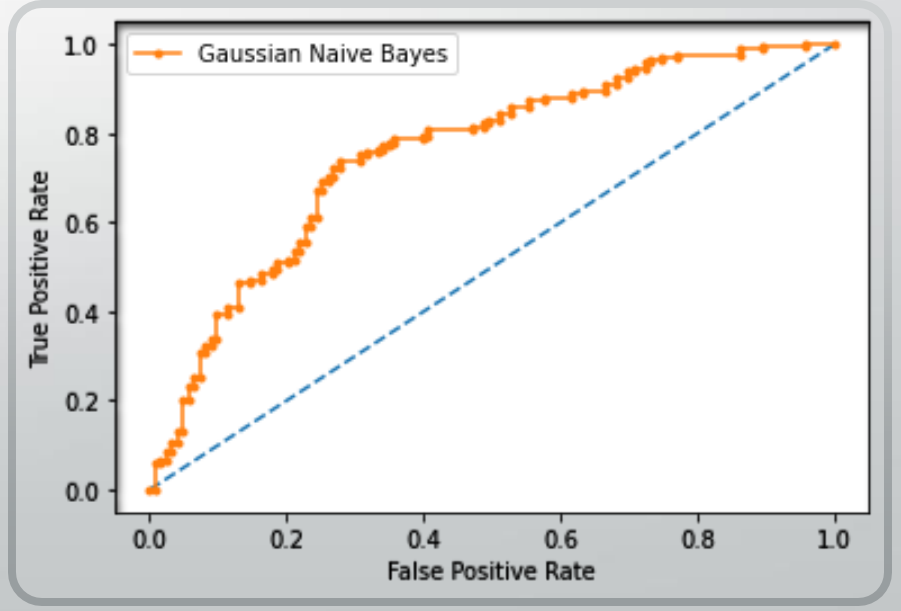


Naïve Bayes:

- Naïve Bayes classifier is a probabilistic machine learning model used for supervised machine learning.
- We perform stepwise feature selection to find the most relevant variables.
- Works on the following two assumptions:

1. Variables are independent of each other.
2. Variables follow a normal distribution.

- We achieved an accuracy of **70%** with our Naïve Bayes model.



Random Forests

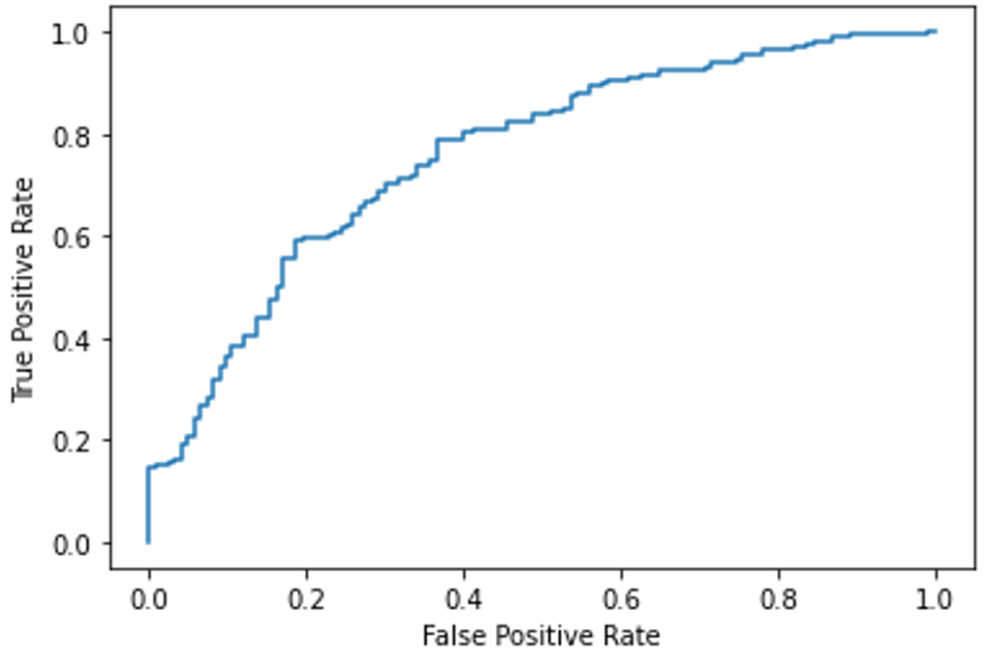
Random Forest is a **supervised machine learning algorithm** that grows and combines multiple decision trees to create a decision forest.

Random Forest grows multiple decision trees which are merged together for a more accurate prediction.

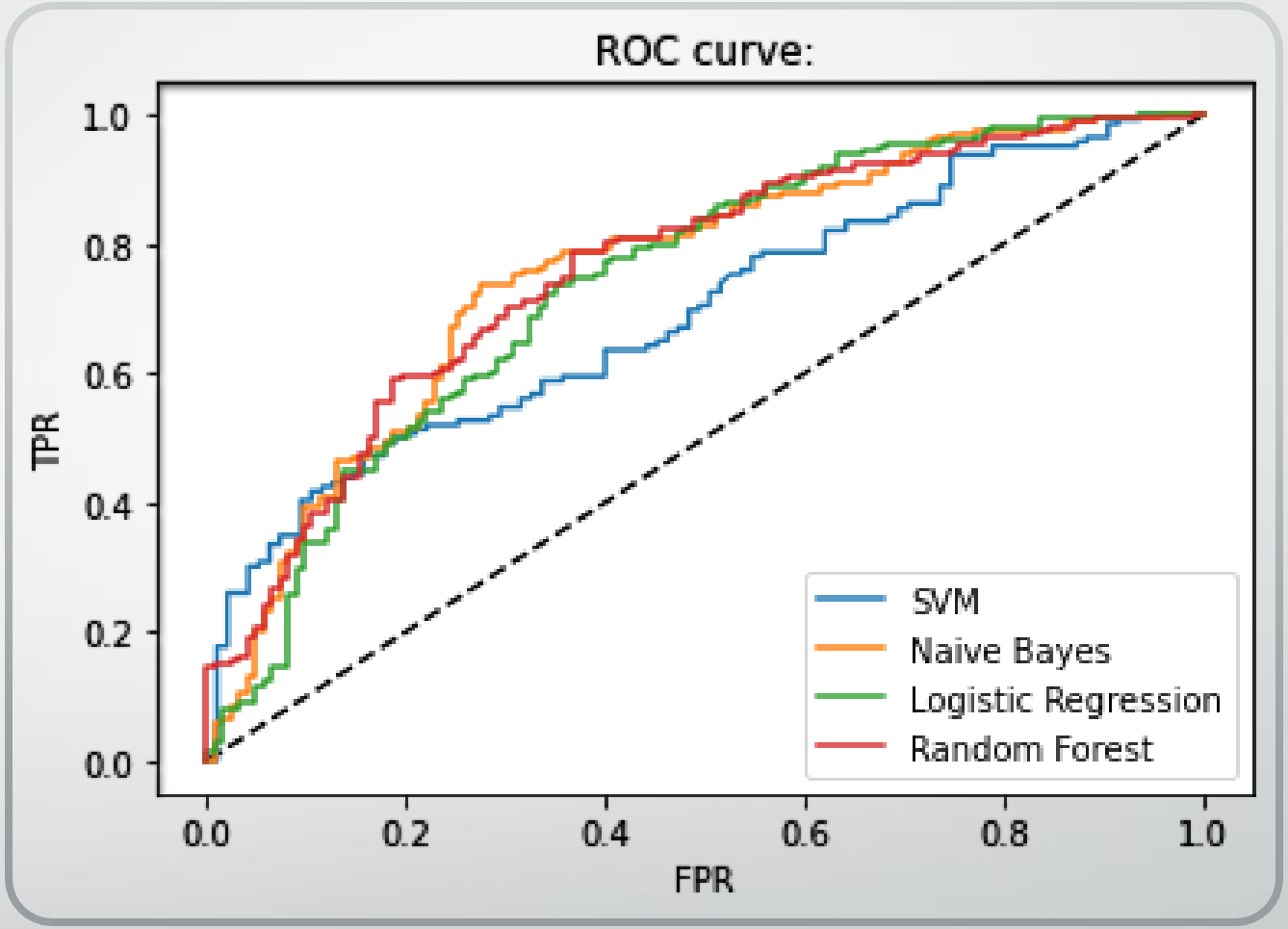
Stepwise Feature Selection performed using the Random Forest estimator.

The features extracted after the SFS were fitted into random forest model after tuning its hyperparameters.

The accuracy of the Random Forests model : **72%**



All Model ROCs:



Results and Conclusions:

- The best model for predicting survivability that we tested was Random Forests at 72%.
- Age at diagnosis and NPI are extremely significant
- Things that may be affecting accuracy of the models
  - Generality of variables(chemo, radiation, etc)
    - Doesn't indicate num of rounds
    - No current health state, just alive or dead
    - Number of remissions and returns per patient not present
  - Missing data that is omitted during tidying
  - Lack of specific professional knowledge about variables

Further Work

- Expanded analysis on the genetic mutation markers included in the data set.
- More sampling and recent data(data from 2016)
- Need more explanatory data
  - Number of times cancer has been diagnosed (initial vs remission vs reappears)
  - Number of rounds of each treatment(radio, chemo, hormone)
  - Family History – are patients genetically disposed to Breast Cancer
  - More samples would be helpful
  - Current health status at time of sampling

References:

- Dr. George Alvarez, MD
- https://www.kaggle.com/datasets/raghadalharbi/breast-cancer-gene-expression-profiles-metabric
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3974073/
- https://www.cancer.gov/types/breast/patient/breast-treatment-pdq