Breast Cancer Data Analysis

Team 89

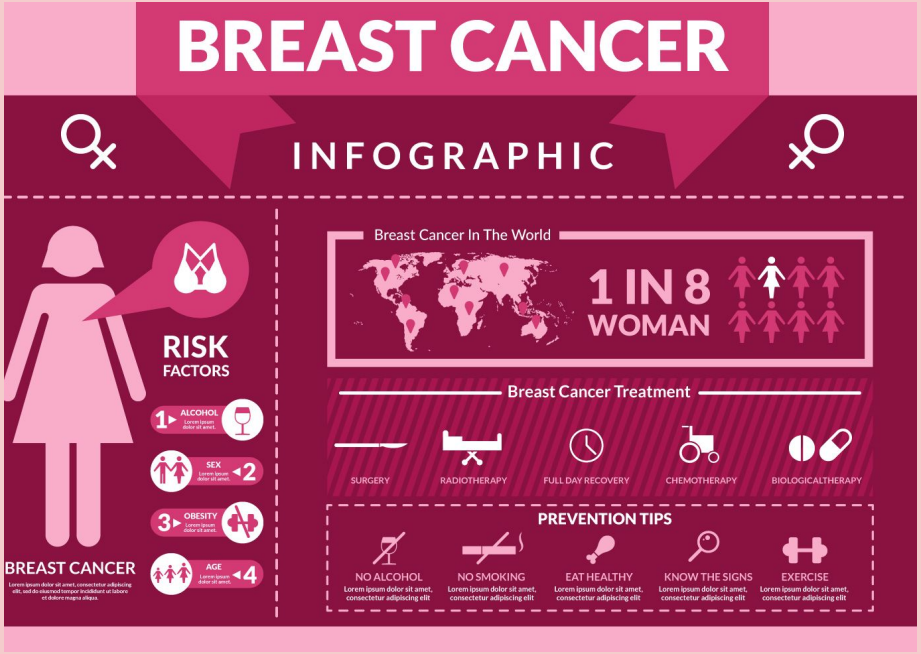
DS2500 Final Project

12/2/2022

Michelle Wang, Huiyun Wang,
Emma Shek, Sabrina Zhou

Motivation

- **Problem:** 13% chance of any woman in the US developing breast cancer
- **Solution:** identify factors that best help to predict survival in breast cancer patients
- **Impact/goal:** develop treatment to target proteins that significantly impact survival outcomes



https://static.vecteezy.com/system/resources/previews/000/257/477/original/breast-cancer-infographics-vector.jpg

Data

- Taken from data.world based on data from the Netherlands Cancer Institute(NKI)
- Includes various categorical/numerical features from real cancer patients
  - Focus on relation of various features to survival

	Patient	ID	age	eventdeath	survival	timerecurrence	chemo	hormonal	amputation	histtype	...	Contig36312_RC	Contig38980_RC	NM_000853	NM_000
0	s122	18	43	0	14.817248	14.817248	0	0	1	1	...	0.591103	-0.355018	0.373644	-0.760
1	s123	19	48	0	14.261465	14.261465	0	0	0	1	...	-0.199829	-0.001635	-0.062922	-0.682
2	s124	20	38	0	6.644764	6.644764	0	0	0	1	...	0.328736	-0.047571	0.084228	-0.695
3	s125	21	50	0	7.748118	7.748118	0	1	0	1	...	0.648861	-0.039088	0.182182	-0.524
4	s126	22	38	0	6.436687	6.318960	0	0	1	1	...	-0.287538	-0.286893	0.057082	-0.565
5	s127	23	42	0	5.037645	2.743326	1	0	1	1	...	-0.417534	-0.141338	-0.492190	0.090
6	s128	24	50	0	8.739220	8.739220	1	1	0	1	...	0.086751	-0.144424	-0.778273	0.024
7	s129	25	43	0	7.567420	7.567420	1	0	0	1	...	-0.003150	0.043824	0.442394	-0.498
8	s130	26	47	0	7.296372	7.296372	1	0	0	1	...	-0.362921	-0.038672	-0.647650	-0.760
9	s131	27	39	1	4.662560	1.114305	0	0	0	1	...	-0.845758	0.635155	-0.235659	-0.396

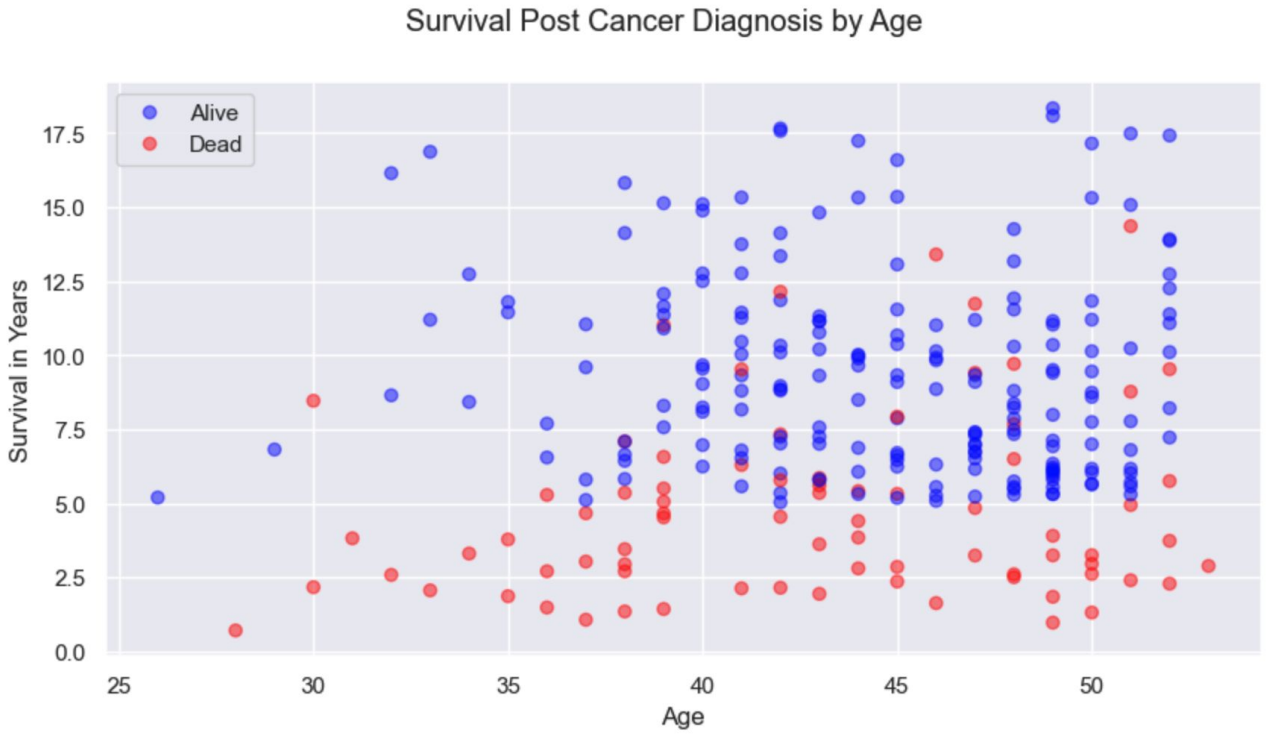
10 rows × 1570 columns

Data Description

- Approximately 1500 columns
  - Lack clear description/ data definitions
- Only 273 rows
- Necessary to scale normalize data
- Y-variable: **survival** (number of years survived by patient after diagnosis)

Visualizing the Data

- Age, survival, death



Methods

Random Forest Regressor

Which **proteins** are the most helpful in **predicting survival outcomes?**

Multiple Regression

Can we **predict how many years a patient will survive** based off expression of their proteins?

Methods

Random Forest Regressor

Which **proteins** are the most helpful in **predicting survival outcomes?**

Multiple Regression

Can we **predict how many years a patient will survive** based off expression of their proteins?

PCA

Can we **simplify** the complexity of our data while still **retaining patterns** to visualize the association between protein expression and survival?

Interpreting Our Results

To validate our models, we compute the cross-validated r^2 value among the given cancer patient data

- If the value is close to 1, then we can effectively **predict** breast cancer survival outcomes
- If the value is close to 0, then we are effectively **guessing** breast cancer survival outcomes
- If the value is negative, we are **ineffectively** predicting breast cancer survival outcomes which is worse than guessing blindly

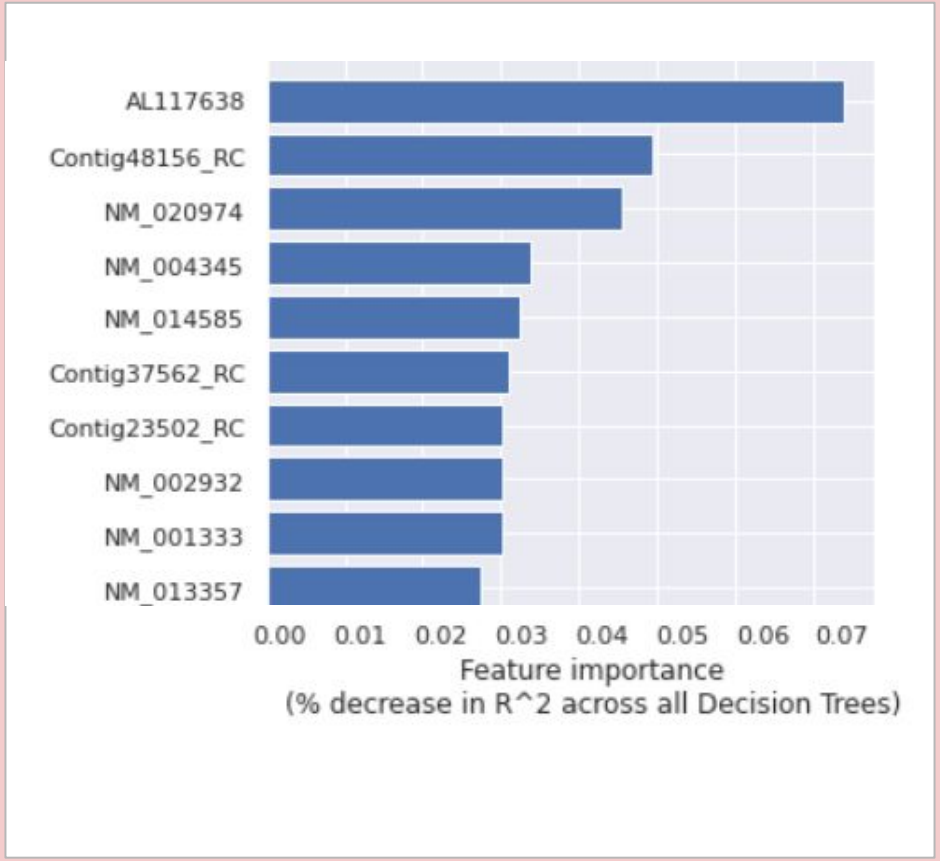
Random Forest Regressor

Goal: identify proteins most important in predicting survival outcomes

CV r2 using **top 43 proteins: 0.25**

Most important proteins include:

- AL117638
- Contig48156_RC
- NM_020974



Multiple Regression

Survival can be predicted via the equation:

survival = 9.15 + 0.17 NM_004405 - 0.22 NM_006157 + 0.70 Contig45397_RC - 1.09 Contig29982_RC + 1.62 V00522 - 1.03 Contig49589_RC + 0.40 Contig39556_RC - 0.68 ...

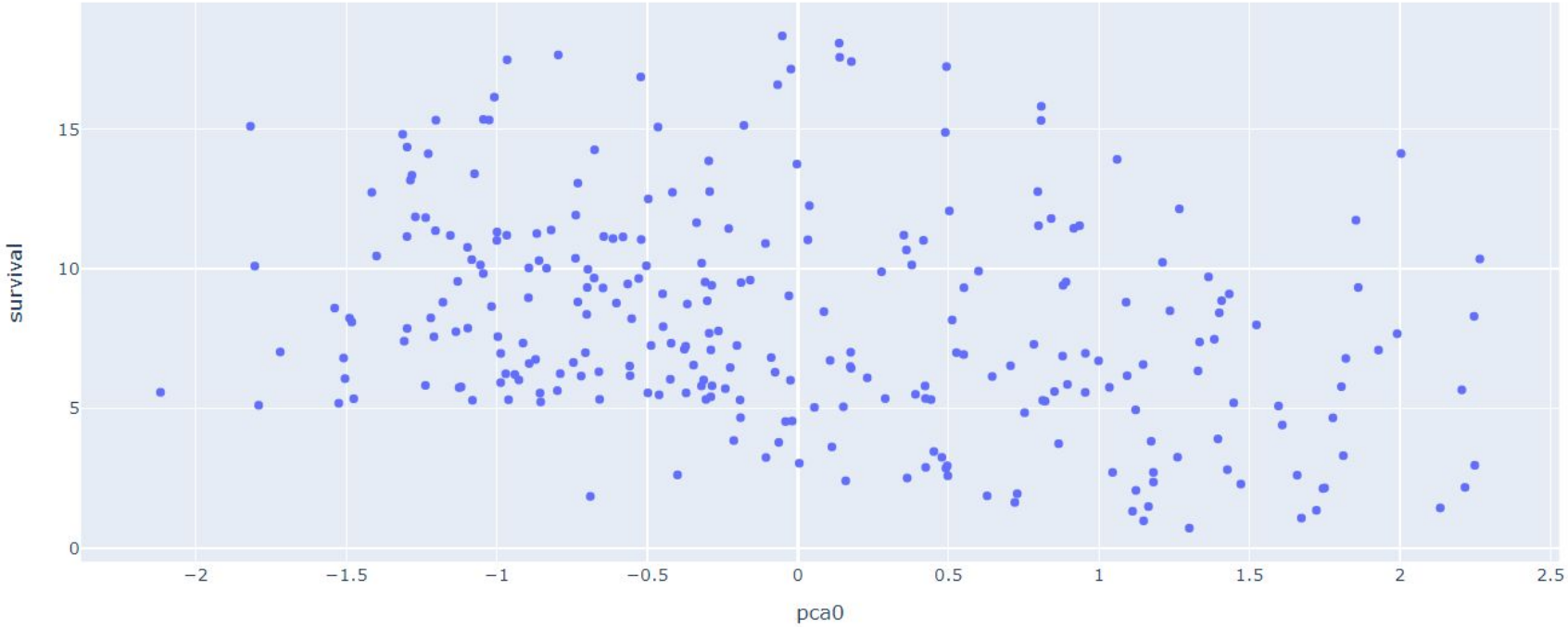
- **r2 = 0.38**
  - 38% of the variability in the survival can be explained by the proteins
  - Higher than r2 obtained in Random Forest Regressor

Cont.- Correlation Between x-Variables

survival = 9.15 + 0.17 NM_004405 - 0.22 NM_006157 + 0.70 Contig45397_RC - 1.09 Contig29982_RC + 1.62 V00522 - 1.03 Contig49589_RC + 0.40 Contig39556_RC - 0.68 ...

- **Pearson's correlation coefficient** calculated between every x variables
- Highest 3 correlations:
  - NM_001168, NM_003258: 0.85
  - NM_018410, NM_005733: 0.85
  - NM_016359, NM_003981: 0.83
- High correlation between x variables
  - **Difficult to interpret the coefficients of the equation**

PCA



Takeaways

We were **unable to create a prediction model** that seemed to predict breast cancer survival outcomes.

Future work

- Change dataset
  - Defined columns
  - Proteins known to significantly impact survival outcomes
  - More patients
- Predict treatment type that would be associated with best survival outcome (classifier)

Project **should not be used to predict survival** of breast cancer patients

- Complexity of the disease
- Flaws in the dataset

Thank you! Any Questions?