BrainStation®

CAPSTONE PROJECT SPRINT 1

CANCER SURVIVABILITY PREDICTOR

https://github.com/r-kaba/Cancer_Survivability_Predictor

20 OCTOBER 2023

The Problem:

Use Machine Learning to predict cancer survivability using demographic and cancer specific features:

- Age
- Gender
- Race
- Tumor Mutation status
- Disease Status
- Cancer type
- Location of metastases
- Various genetic markers

Motivations/Impacts

Why do we care?

- Cancer Prognosis affects everyone involved when a patient has cancer from family and friends to the healthcare system as a whole.
- 50% life risk of cancer diagnosis and 20-25% of people in BC lifetime risk of dying
- Accurate prognosis = increased quality of life

Why do I care?

- I have been personally affected by incorrect prognosis
- I have worked in the clinical cancer industry and know its profound affect on the population

Why does industry care?

This can have affects on many different industries:

- Health care resources
- Clinical trials
- With more data we can get even more accurate predictions
- Huge increase in data in cancer with the implementation of sequencing data

The Data

MSK met2021 study

Data was pulled from the cBioPortal from the MSK met2021 study done by the Memorial Sloan Kettering Cancer Centre in New York.

25,775 Patients, 55 Features

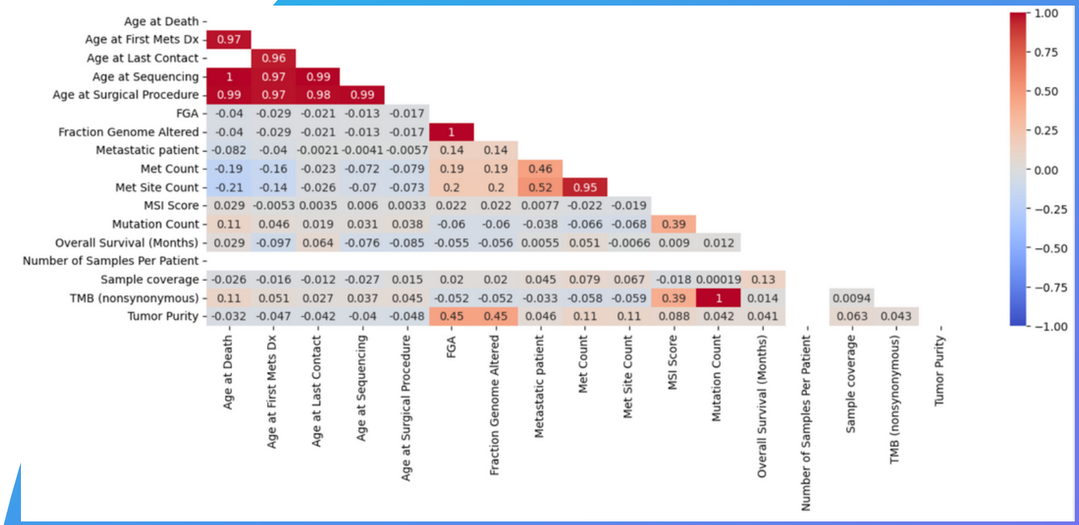
The original data set contained 25, 775 instances with 55 features. Contains numerical (17) and categorical (38) data describing the demogrpahics and cancer specific features of each patient. Cleaned to 46 features

Target variable

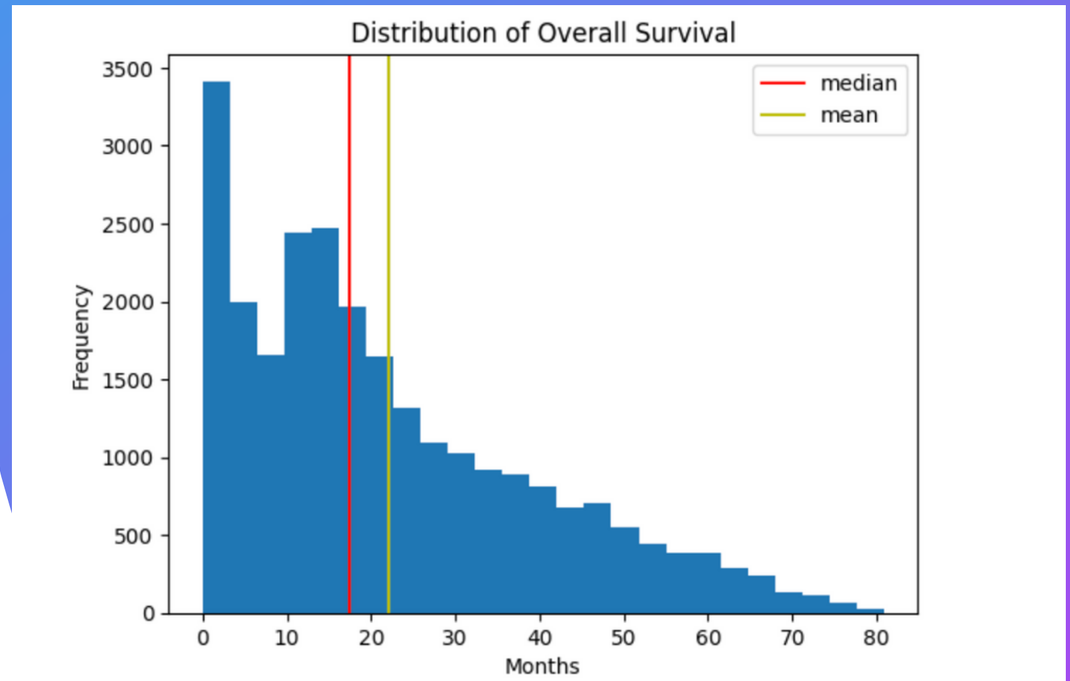
The model will aim to predict the Overall Survivability in months for each patient. This is a continuous variable.

May need to make ti categorical.

- A lot of highly correlated features
- Missing values to deal with
- Blanks in my correlation heatmap

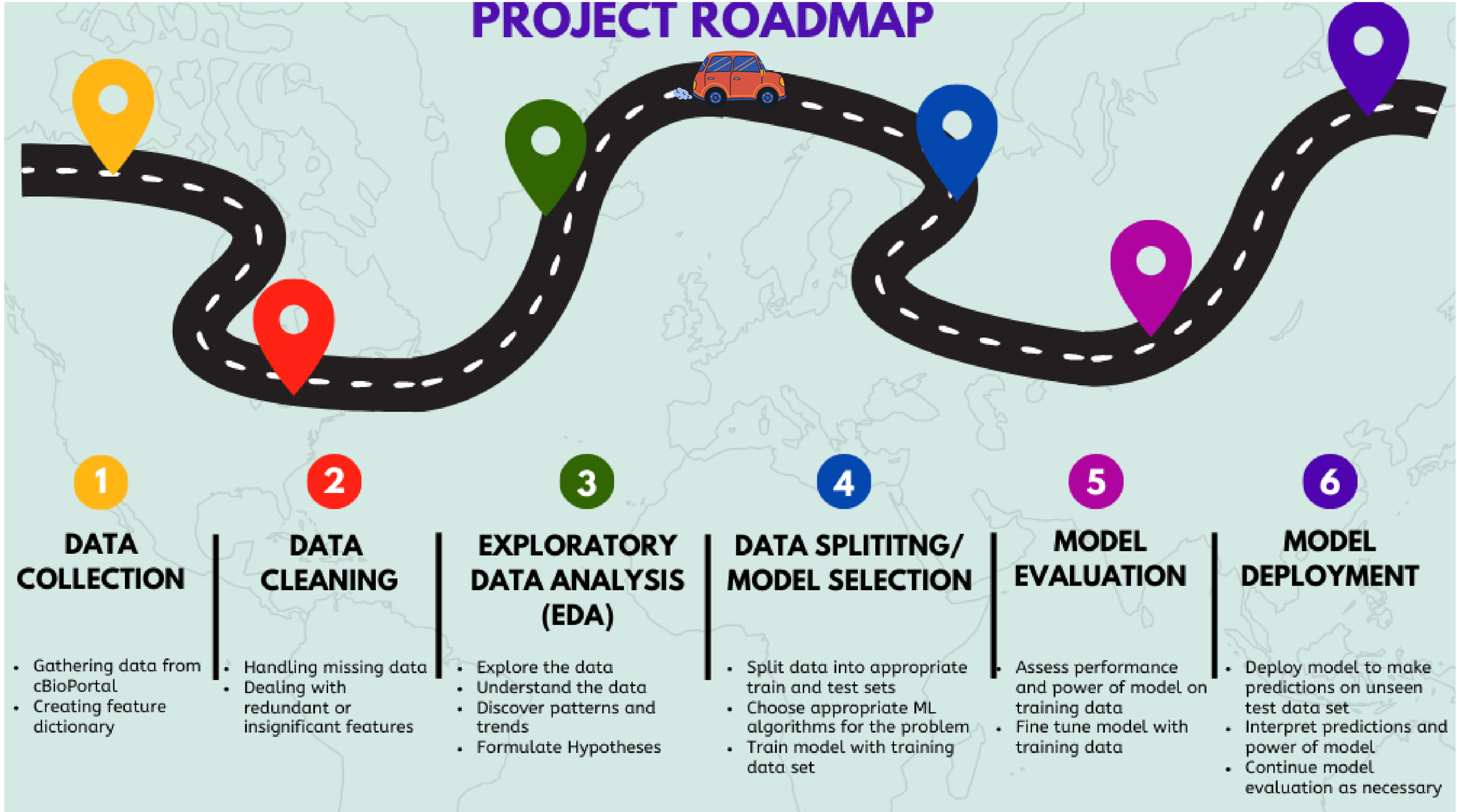


Distribution of target variable (overall Survival) is log normal distributed - high proportion survive less than 30 months



EDA

Proposed Vision using Data Science:



Next Steps

CONTINUE EDA

Ive done the initial EDA but I want to explore the data more prior to trying to fit my first iterations of the model

MODEL SELECTION/FEATURE ENGINEERING

Once I have completed the EDA, I will select a model to start. Based on my initial findings, it will likely be linear regression for overall survival month and logistic regression if I also choose overall survival status.

MODEL TRAINING AND TESTING

Once I have selected a model that seems to work, then I will train it with various hyper parameters and attempt to get the highest accuracy score possible where I can interpret and explain my model

THANK YOU