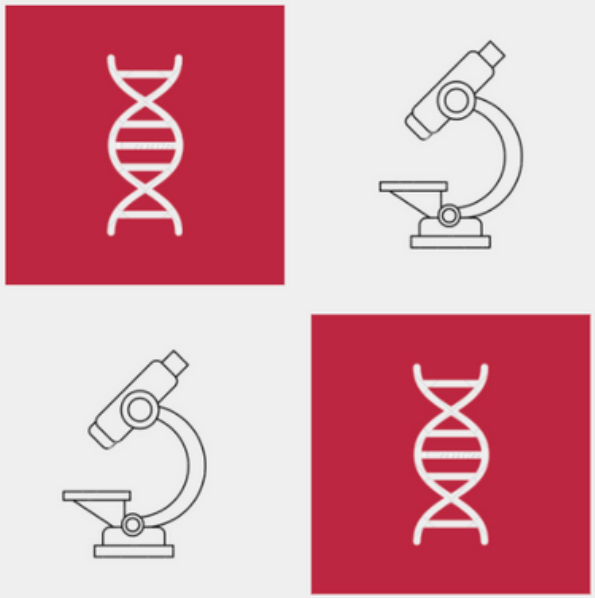
CAN WE PREDICT 7-YEAR SURVIVAL?!

Prostate Cancer Case

Anisha Joshi | Healthcare Analytics

Our Agenda for Today

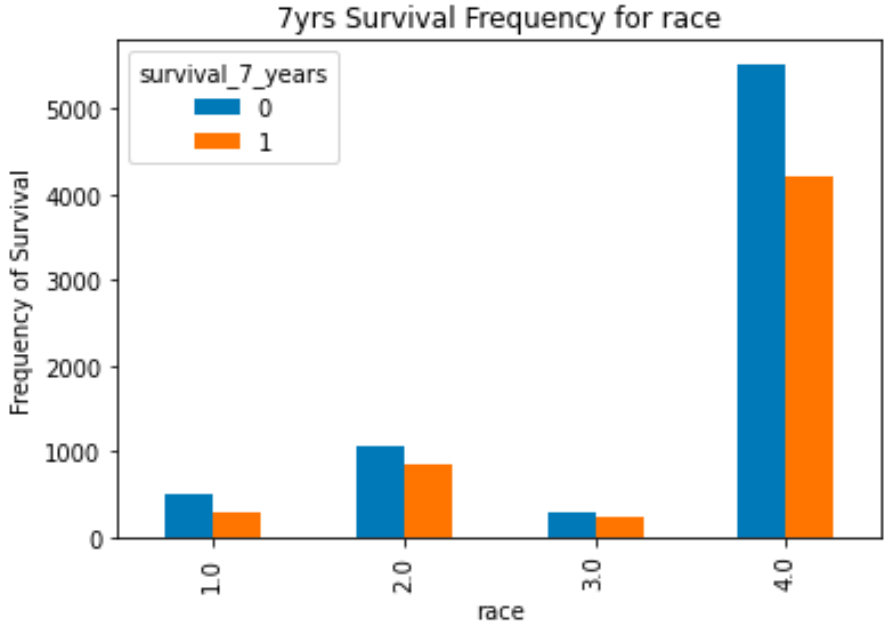


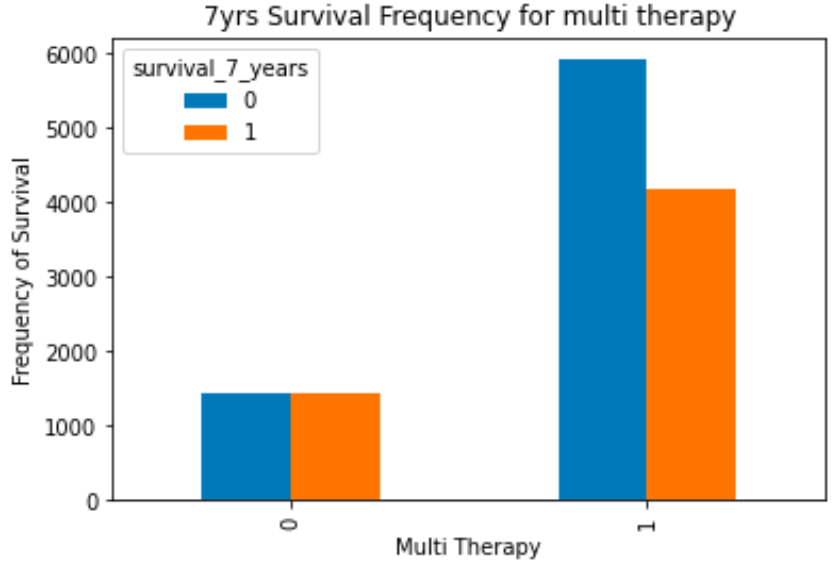
LIST OF KEY CONCEPTS

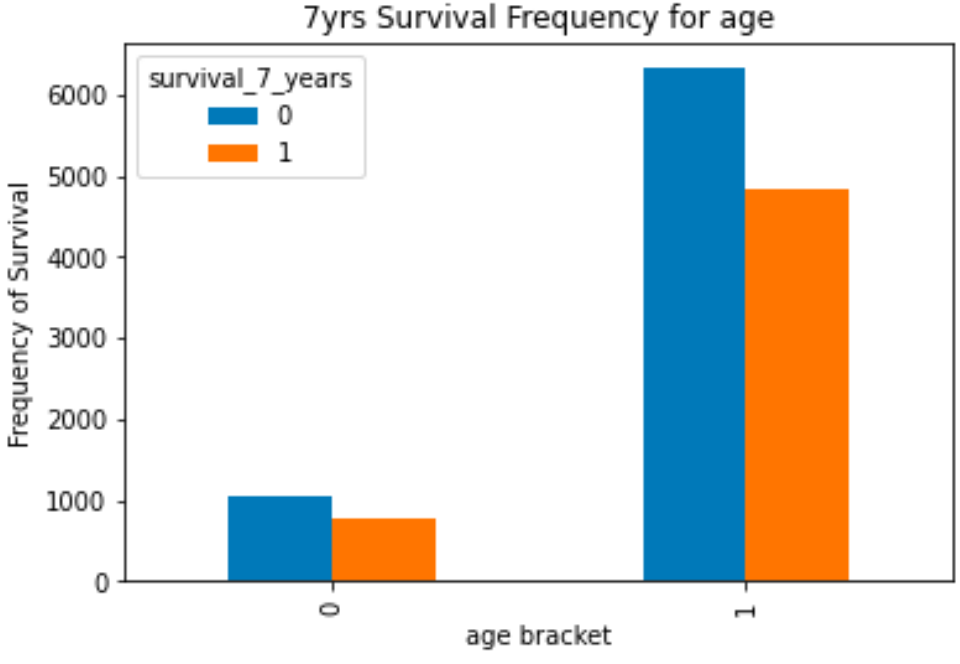
-Introduction to data
-Exploratory Data Analysis
-Feature Engineering
-Feature Selection
-Model
-Evaluation
-Predict on Test Data

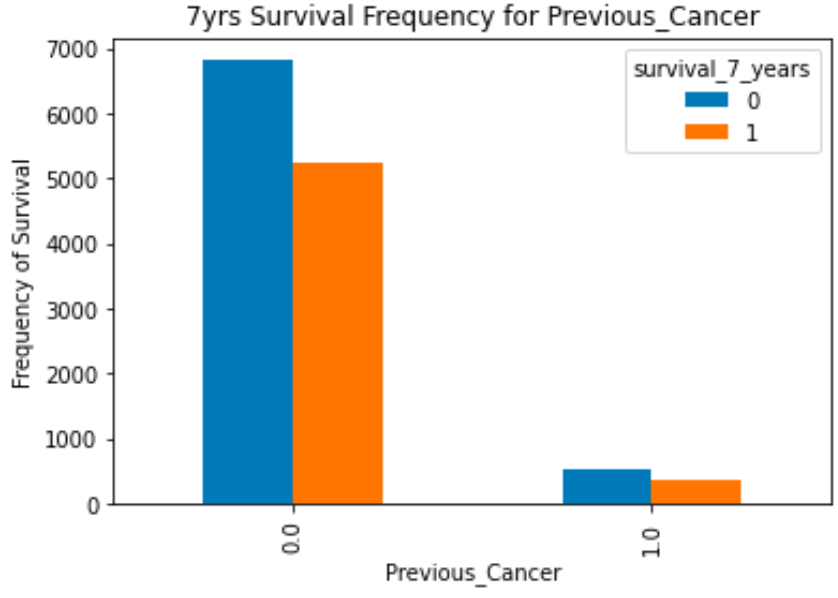
Assuming the provided data is of 2021 & 2022

EXPLORATORY ANALYSIS

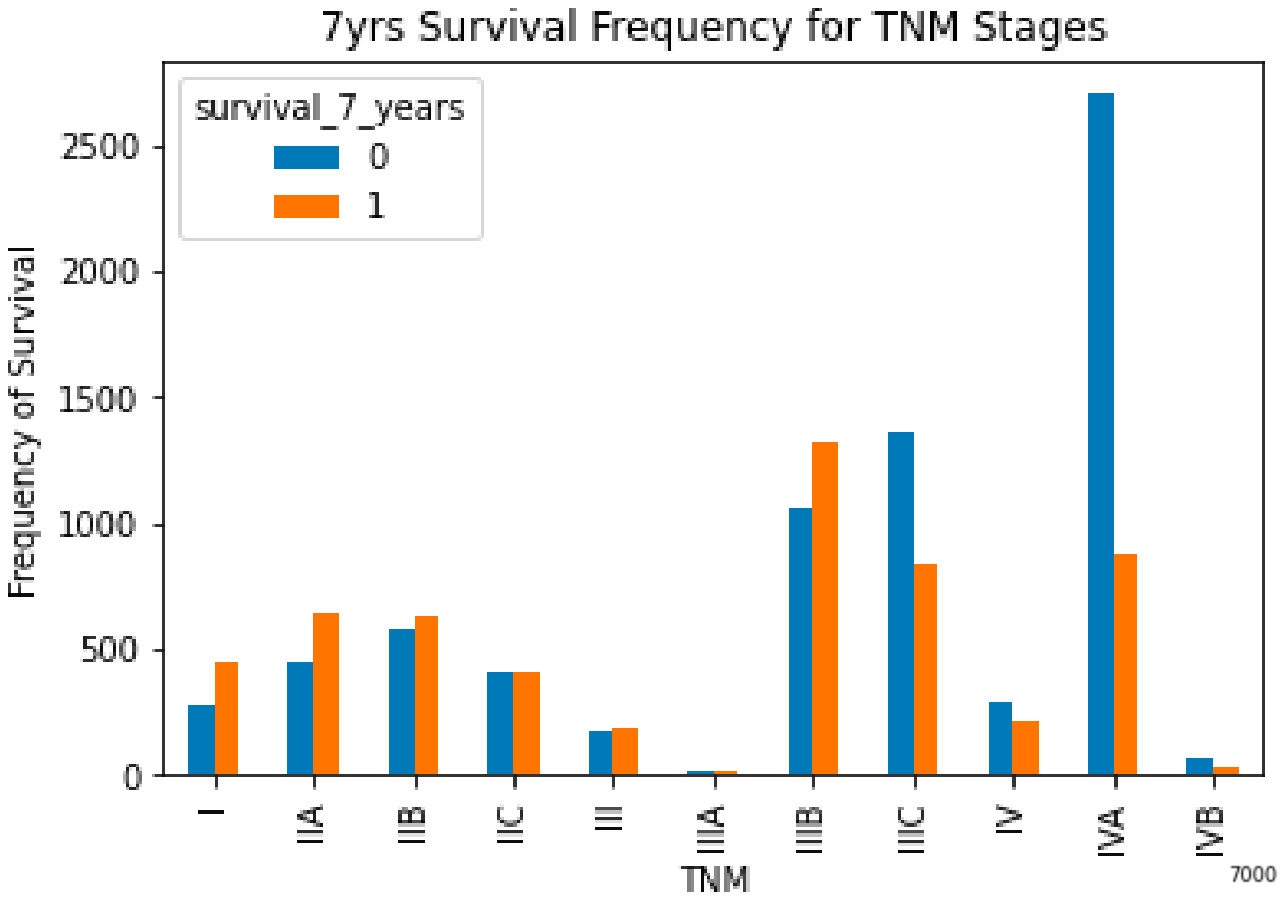


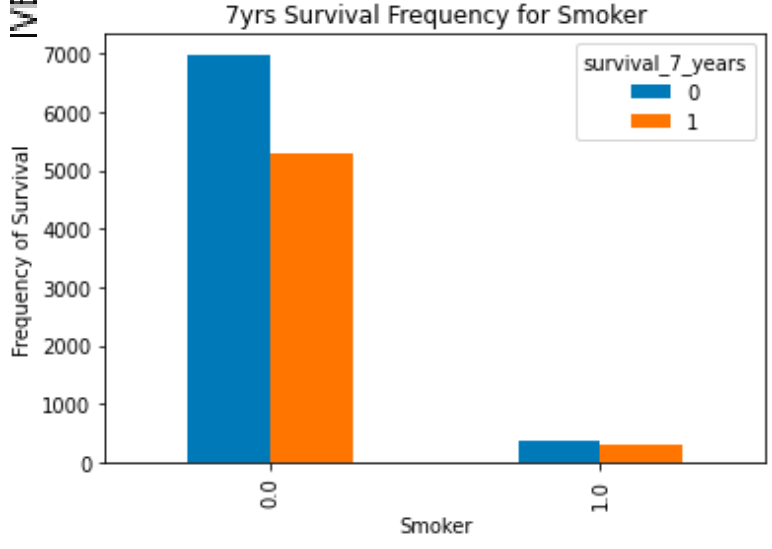






EXPLORATORY ANALYSIS





Subject Matter

The TNM system for prostate cancer is based on 5 key pieces of information:

- The extent of the main (primary) **tumor** (**T** category)
- Whether cancer has spread to nearby lymph **nodes** (**N** category)
- Whether cancer has spread (**metastasized**) to other parts of the body (**M** category)
- The **PSA level** at the time of diagnosis
- The **Grade Group** (based on the **Gleason score**), is a measure of how likely the cancer is to grow and spread quickly. This is determined by the results of the prostate biopsy (or surgery).

Instead of these 5 categories have TNM score

Ref - https://www.cancer.org/cancer/prostate-cancer/detection-diagnosis-staging/staging.html

Feature Engineering

Delete Date, ID and combined symptoms

SYMPTOMS

Dummy One Hot Key Encoding, Add later, Sym > 3 = 1 as Sym

AGE

In different Brackets Initially, later 65+

DIFF. IN TUMOUR SIZE

From date of diagnosis to 6 months/year

HEIGHT & WEIGHT

Replaced with BMI

RFE - Feature Selection

With Estimater as Logistic Regression

```
print(rfe.ranking_)
```

```
[25 24 15 10  6 23  1 17 16 26  5  1 12  1  7 13  1  1 11 18  2 22 19 21
  8  4 14  9  1  3  1  1 20  1  1]
```

Index numbers before age in bins - 6,13,15,16,27,29,30,32,33,43

Index number after Age 65+ >> 6,11,13,16,17,28,30,31,33,34

'RD_THRPY','RAD_REM','SURVIVAL_1_YEAR', 'STAGE_I','STAGE_IIA','TNM_SC_IIIB', 'TNM_SC_IIIC','TNM_SC_IVA','TNM_SC_IVB', 'MORE_THAN65'

Final data

15,385

Rows

33

Features

12,972

Final Rows

10

Final Features

After Dummy Variables - 12972, 36
Removed all the Null Value Rows

Evaluation Metric & Model Performance

FN - PLAY A CRUCIAL ROLE

Metric - Recall is important
Based on the Given Dataset

Recall 62% ***Accuracy 65%***

```
print(classification_report(y_test,y_pred))
```

	precision	recall	f1-score	support
0	0.70	0.67	0.68	2185
1	0.60	0.62	0.61	1707
accuracy			0.65	3892
macro avg	0.65	0.65	0.65	3892
weighted avg	0.65	0.65	0.65	3892

Predicted	0	1	All
Actual			
0	1464	721	2185
1	642	1065	1707
All	2106	1786	3892

Logistic Regression Model

Train-Test

Stratified Sampling of 70:30

LR with Recursive Feature Selection

Found 10 features that were able to define the model accuracy of 65%

Predict

RFE and model are fit on all available data, then the predict() function is used for *test predictions*