Survival Analysis for Prostate Cancer Patients

IDS-506

Health Information and Management Analytics

Vivek Kumar

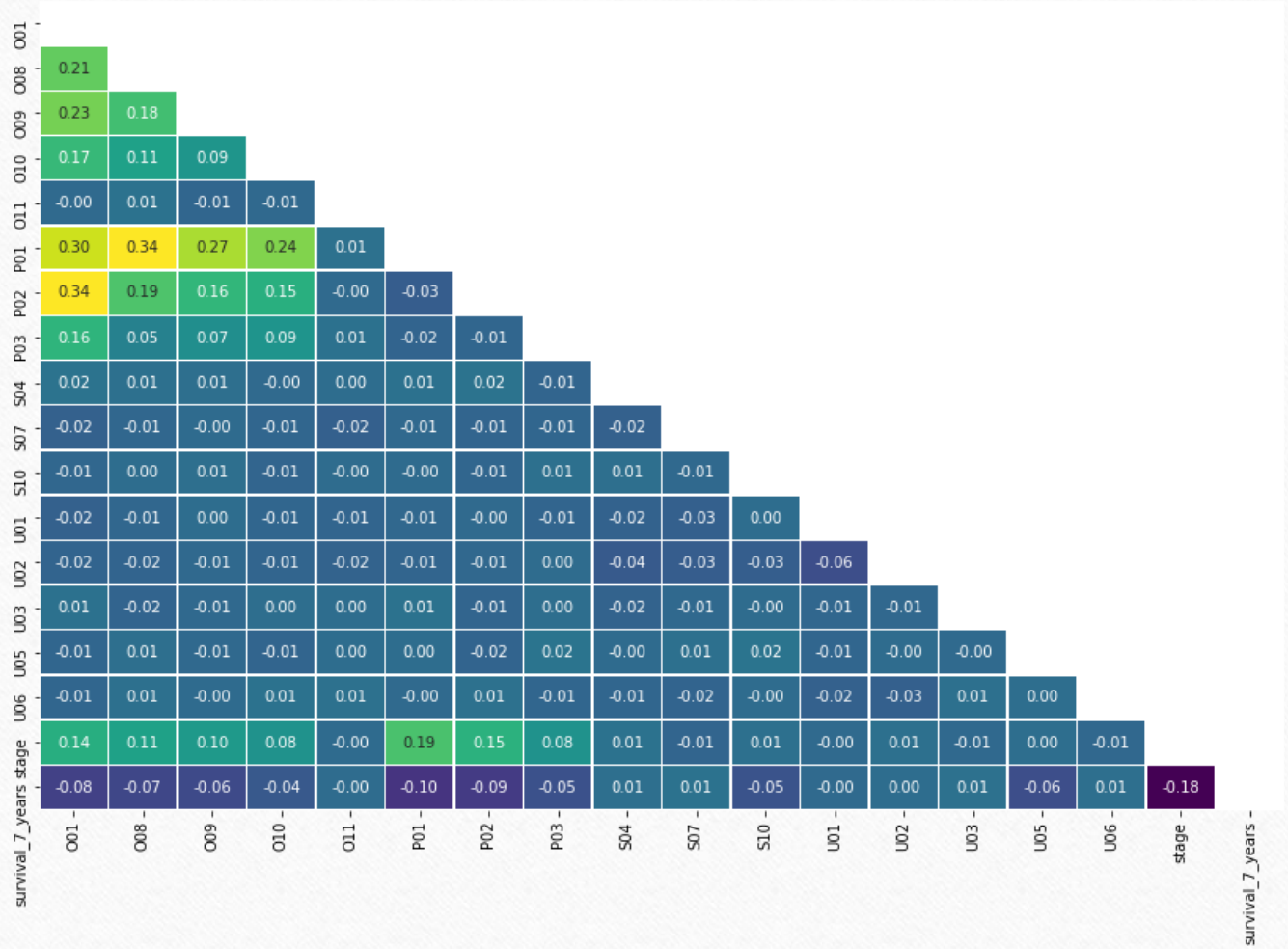
The Case

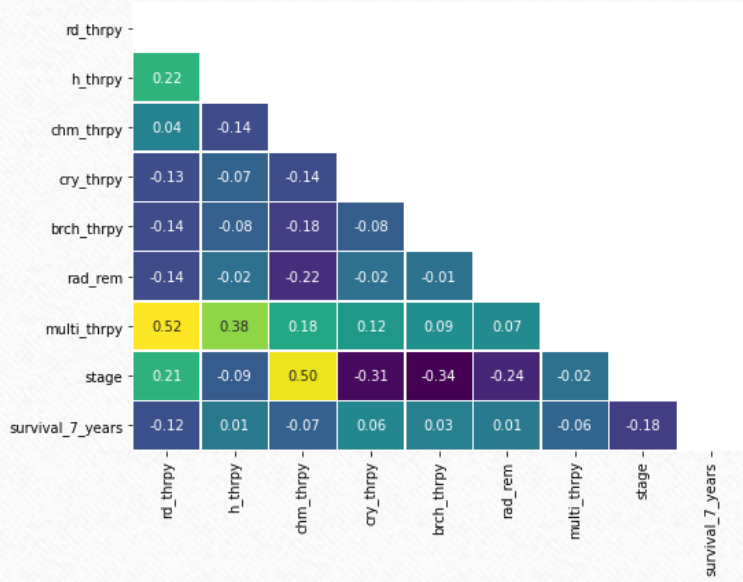
- To determine the 7-year survival of Prostate Cancer Patients based on their medical records, state of cancer, progression of their disease.
- Training Data (15,385 records, 33 attributes) with values given for target variable to train the model
- Test Data (11,531 records, 33 attributes) with all NAs for target variable

Data Transformation

- Transformed Stage Variable : {I : 1, IIA : 2, IIB : 3, III : 4, IV : 5}
- Encoded Symptoms to 16 columns, i.e. individual columns representing presence or absence of a symptom
- Dropped 't_score', 'n_score' and 'm_score' as these combined together forms stage
- One hot encoding for Race attribute
- For surivial_1_year = 0, updated psa_1_year, psa_6_months, tumor_1_year and tumor_6_months with preceding existing values
- For records having equal values in psa_1_year and psa_diagnosis, missing values in psa_6_months was replaced by either of the two
- For records having equal values in tumor_1_year and tumor_6_months, missing values in tumor_6_months was replaced by either of the two

Correlation of Symptoms & Therapy attributes with Target Variable





Logistic Regression Models

t=0

	precision	recall	f1-score	support
0	0.62	0.76	0.69	1035
1	0.58	0.41	0.48	822
accuracy			0.61	1857
macro avg	0.60	0.59	0.58	1857
weighted avg	0.60	0.61	0.60	1857

t=6 months

	precision	recall	f1-score	support
0	0.67	0.85	0.75	517
1	0.55	0.31	0.40	312
accuracy			0.64	829
macro avg	0.61	0.58	0.57	829
weighted avg	0.62	0.64	0.62	829

t=1 year

	precision	recall	f1-score	support
0	0.72	0.74	0.73	1095
1	0.62	0.60	0.61	785
accuracy			0.68	1880
macro avg	0.67	0.67	0.67	1880
weighted avg	0.68	0.68	0.68	1880

Odds Ratio

t=0

gleason_score	0.878858
stage	0.785405
age	1.000098
height	1.043981
weight	0.995780
family_history	0.985934
first_degree_history	0.947477
previous_cancer	0.961802
smoker	1.230887
tumor_diagnosis	0.993838
psa_diagnosis	0.990064
tea	0.969972
rd_thrpy	0.693475
h_thrpy	1.258384
chm_thrpy	1.015377
cry_thrpy	0.935814
brch_thrpy	0.861442
rad_rem	0.902120
multi_thrpy	0.876293
O01	0.761985
O08	0.449631
O09	0.379361
O10	0.877619
O11	1.132676
P01	0.467985
P02	0.373967
P03	0.490846
S04	1.074480
S07	1.058264
S10	0.524444
U01	1.007259
U02	1.056496
U03	1.002971
U05	0.664183
U06	1.064224
race_1.0	0.778450
race_2.0	1.066810
race_3.0	1.021797
race_4.0	1.026756

t=6months

0	gleason_score	0.948034
1	stage	0.779225
2	age	0.998959
3	height	1.043201
4	weight	0.994791
5	family_history	0.985569
6	first_degree_history	0.942193
7	previous_cancer	0.834588
8	smoker	1.337849
9	tumor_6_months	0.990043
10	psa_6_months	0.963581
11	tea	0.983514
12	rd_thrpy	0.791503
13	h_thrpy	1.275550
14	chm_thrpy	0.995765
15	cry_thrpy	0.830029
16	brch_thrpy	0.967656
17	rad_rem	0.907272
18	multi_thrpy	0.718686
19	O01	1.034065
20	O08	0.399972
21	O09	0.474068
22	O10	0.937208
23	O11	1.062522
24	P01	0.298008
25	P02	0.466383
26	P03	0.716510
27	S04	1.218289
28	S07	1.313301
29	S10	0.585628
30	U01	1.133938
31	U02	1.090613
32	U03	0.932133
33	U05	0.646738
34	U06	0.931103
35	race_1.0	0.772572
36	race_2.0	0.923036
37	race_3.0	0.946628
38	race_4.0	0.846105

t=1year

0	gleason_score	0.909606
1	stage	0.802313
2	age	1.001154
3	height	1.036747
4	weight	0.996591
5	family_history	1.058453
6	first_degree_history	0.935572
7	previous_cancer	0.987842
8	smoker	1.123575
9	tumor_1_year	0.992161
10	psa_1_year	0.992991
11	tea	0.975422
12	rd_thrpy	0.698252
13	h_thrpy	1.172697
14	chm_thrpy	0.986993
15	cry_thrpy	0.938269
16	brch_thrpy	0.874851
17	rad_rem	0.858601
18	multi_thrpy	0.894466
19	O01	0.730338
20	O08	0.615076
21	O09	0.444223
22	O10	1.433118
23	O11	1.028107
24	P01	0.428078
25	P02	0.401007
26	P03	0.401923
27	S04	1.089684
28	S07	1.078927
29	S10	0.570666
30	U01	0.998479
31	U02	1.020869
32	U03	1.049348
33	U05	0.682950
34	U06	1.053596
35	race_1.0	0.744192
36	race_2.0	1.016619
37	race_3.0	0.964678
38	race_4.0	0.946701

Survival Analysis

- The Survival Function is given by:

$$S(t) = P(T>t)$$

where,

T is the lifetime of a member of the population

t denotes time

S(t) is the survival curve at time t

Kaplan-Meier Curve