Applied **Machine Learning**: Predicting **7-year Prostate Cancer** Survivability

IDS.506 - Robert Minh Duc Bui - mbui7@uic.edu

Executive Summary

- We can develop many different models model that predict 7-year survival for prostate cancer patients. An **xgBoost** model with a **manual screening step** has the **highest test accuracy**.
- Challenges include:
  - Dataset **dirtiness** - many missing values.
  - **Intentional obfuscation** - due to many recorded variables not having a clear definition.
- Final model at-a-glance:
  - **Gradient-boosted random forest** ensemble model **+ manual screening rule**
  - Text parsing + dummy variables + low variance filter + normalization + derivative features
  - Quick Performance Metrics compared to naive null model:
    - Accuracy: **acc** = **0.6875** (**+34.47%** lift)
    - F1 Measure: **F1** = **0.7187** (**+26.15%** lift)

Research **Mission**

- To develop a model that:
  - Combines factors to predict **7-year survivability** of prostate cancer patients.
  - Optimizing for accuracy - variable importance not explicitly needed.
- Pitfalls:
  - Data points are **intentionally obfuscated**, which might make human intervention to remove spurious variables more difficult.

Dataset **Introduction**

- Source: **proprietary** collected data from Enova.
- Collection Methods:
  - Unknown. Proprietary data.
- Shape:
  - Rows: **15385**
  - Columns: **33**
- Target Variable shape: binary (1,0).

- Goal: to **predict** patients' **survivability** for diagnosed **Prostate Cancer**.

Dealing with **NAs**

pos	Definition	varname	NA	rate.rest	rate.NAs	NA_method	NA_method_justification
18	Size of primary tumor 6 months after diagnosis, in mm	tumor_6_months	0.6540	0.4862	0.4798	**remove col**	too much missing.
21	Level of prostate-specific antigen in blood 6 months after diagnosis, in ng/mL	psa_6_months	0.6179	0.4859	0.4796	**remove col**	too much missing.
12	count of family members who have been diagnosed with prostate cancer	family_history	0.1034	0.4810	0.4905	force to **0**	predominant subgroup + maybe nothing to record? = 0
13	count of brothers and fathers of the patient who have been diagnosed with prostate cancer	first_degree_histor	0.1034	0.4810	0.4905	force to **0**	predominant subgroup + maybe nothing to record? = 0
14	flag indicating whether the patient has ever been diagnosed with any cancer previously	previous_cancer	0.1034	0.4810	0.4905	force to **0**	predominant subgroup + maybe nothing to record? = 0
15	flag indicating whether the patient describes himself as a smoker	smoker	0.1034	0.4810	0.4905	force to **0**	predominant subgroup + maybe nothing to record? = 0
23	How many times the patient reports drinking tea per week	tea	0.1034	0.4810	0.4905	**remove col**	not used in modeling
20	Level of prostate-specific antigen in blood at time of diagnosis, in ng/mL	psa_diagnosis	0.0909	0.4816	0.4856	**remove**	cannot make meaningful prediction without diagnosis info
10	Height of patient at time of diagnosis	height	0.0894	0.4831	0.4704	**remove**	no default value, cannot impute.
11	Weight of patient at time of diagnosis	weight	0.0856	0.4846	0.4539	**remove**	no default value, cannot impute.
22	Level of prostate-specific antigen in blood 1 year after diagnosis, in ng/mL	psa_1_year	0.0675	0.4820	0.4823	\=6/**diag**, fallback 0	impute as no change from diagnosis
8	Age of patient at time of diagnosis	age	0.0478	0.4823	0.4765	**remove**	no default value, cannot impute.
19	Size of primary tumor 1 year after diagnosis, in mm	tumor_1_year	0.0389	0.4818	0.4860	\=6/**diag**, fallback 0	impute as no change from diagnosis
24	A list of codes indicating the presence of various symptoms. Meaning has been removed.	symptoms	0.0270	0.4800	0.5550	parsed as **empty**	nothing to record = 0
3	A measurement of how abnormal the cancer cells look compared to normal cells	gleason_score	0.0207	0.4813	0.5140	**remove**	small proportion + cannot predict w/o diagnosis
17	Size of primary tumor at time of diagnosis, in mm	tumor_diagnosis	0.0197	0.4823	0.4669	**remove**	small proportion + cannot predict w/o diagnosis
9	Race of patient	race	0.0112	0.4821	0.4740	**remove col**	not used in modeling
29	brachytherapy used	brch_thrpy	0.0000	0.4820	0.0000		
27	chemotherapy used	chm_thrpy	0.0000	0.4820	0.0000		
28	crypotherapy used	cry_thrpy	0.0000	0.4820	0.0000		
2	the month and year of diagnosis	diagnosis_date	0.0000	0.4820	0.0000	**remove col**	no observable relationship + "00" dates + irregular data shape
26	hormone therapy used	h_thrpy	0.0000	0.4820	0.0000		
1	An identifier used for scoring dataset	id	0.0000	0.4820	0.0000		
6	Describes whether or not the cancer has spread to distant parts of the body	m_score	0.0000	0.4820	0.0000		
31	multiple therapies used in conjunction	multi_thrpy	0.0000	0.4820	0.0000		
5	Describes whether or not the cancer has spread to the lymph nodes	n_score	0.0000	0.4820	0.0000		
30	prostate surgically removed	rad_rem	0.0000	0.4820	0.0000		
25	external beam radiotherapy used	rd_thrpy	0.0000	0.4820	0.0000		
16	What side of the prostate the cancer has been found in	side	0.0000	0.4820	0.0000		
7	Stage of cancer	stage	0.0000	0.4820	0.0000		
32	survived 1 year from diagnosis flag	survival_1_year	0.0000	0.4820	0.0000		
4	Describes local extent of prostate tumor	t_score	0.0000	0.4820	0.0000		

After NA-filtering:
10606 rows remaining.

Exploratory Data Analysis #1 - **Discretes**

variable	value	n	y		survival
			0	1	
side	both	5369	3024	2345	43.68%
	left	2045	1154	891	43.57%
	right	3192	1821	1371	42.95%
t_score	T1a	672	345	327	48.66%
	T1b	618	308	310	50.16%
	T1c	660	335	325	49.24%
	T2a	888	444	444	50.00%
	T2b	876	441	435	49.66%
	T2c	867	464	403	46.48%
	T3a	1075	646	429	39.91%
	T3b	1105	647	458	41.45%
	T3c	1013	602	411	40.57%
	T4	2832	1767	1065	37.61%
n_score	N0	6643	3240	3403	51.23%
	N1	2895	2191	704	24.32%
	NX	1068	568	500	46.82%
m_score	M0	9783	5315	4468	45.67%
	M1a	297	248	49	16.50%
	M1b	184	156	28	15.22%
	M1c	342	280	62	18.13%
stage	I	374	127	247	66.04%
	IIA	1426	597	829	58.13%
	IIB	2409	1324	1085	45.04%
	III	1790	834	956	53.41%
	IV	4607	3117	1490	32.34%
race	1	624	388	236	37.82%
	2	1560	872	688	44.10%
	3	427	231	196	45.90%
	4	7887	4449	3438	43.59%

variable	value	n	y		survival
			0	1	
family_history	0	7123	3978	3145	44.15%
	1	3020	1759	1261	41.75%
	2	396	220	176	44.44%
	3	58	36	22	37.93%
	4	7	4	3	42.86%
	5	2	2	0	0.00%
first_degree_history	0	8771	4914	3857	43.97%
	1	1672	988	684	40.91%
	2	143	84	59	41.26%
	3	16	10	6	37.50%
	4	4	3	1	25.00%
previous_cancer	0	9960	5624	4336	43.53%
	1	646	375	271	41.95%
smoker	0	10079	5717	4362	43.28%
	1	527	282	245	46.49%
tea	0	765	426	339	44.31%
	1	1837	1025	812	44.20%
	2	2475	1389	1086	43.88%
	3	2039	1157	882	43.26%
	4	1334	772	562	42.13%
	5	611	354	257	42.06%
	6	295	177	118	40.00%
	7	116	64	52	44.83%
	8	37	21	16	43.24%
	9	17	12	5	29.41%
	10	2	1	1	50.00%
	12	1	1	0	0.00%

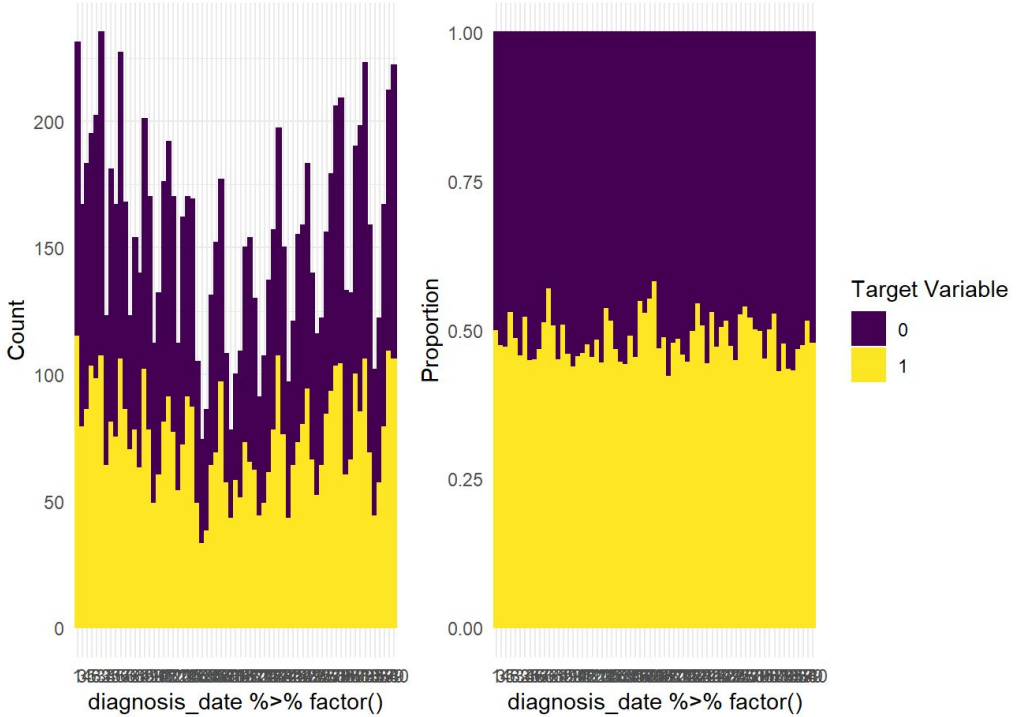
variable	value	n	y		survival
			0	1	
rd_thrpy	0	4894	2431	2463	50.33%
	1	5712	3568	2144	37.54%
h_thrpy	0	6959	3972	2987	42.92%
	1	3647	2027	1620	44.42%
chm_thrpy	0	3611	1836	1775	49.16%
	1	6995	4163	2832	40.49%
cry_thrpy	0	8054	4681	3373	41.88%
	1	2552	1318	1234	48.35%
brch_thrpy	0	8003	4631	3372	42.13%
	1	2603	1368	1235	47.45%
rad_rem	0	8744	4980	3764	43.05%
	1	1862	1019	843	45.27%
multi_thrpy	0	2322	1169	1153	49.66%
	1	8284	4830	3454	41.69%
survival_1_year	0	1107	1107	0	0.00%
	1	9499	4892	4607	48.50%

Exploratory Data Analysis #2 - **Symptoms**

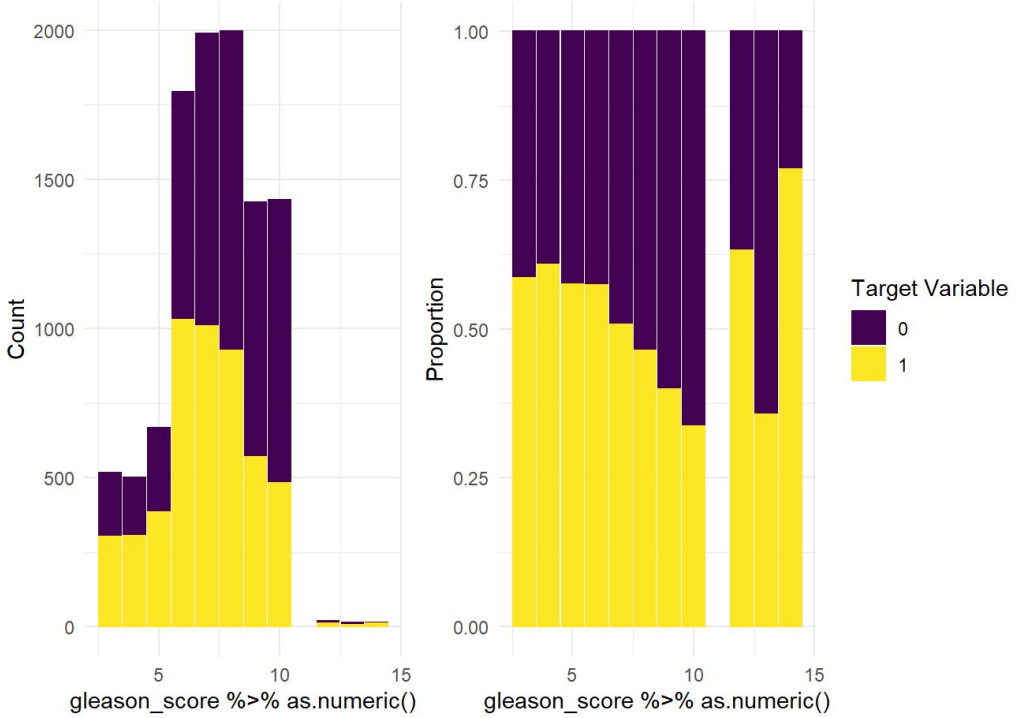
variable	value	n	y		survival
			0	1	
symptom_O01	0	10150	5163	4987	49.13%
	1	208	167	41	19.71%
symptom_O08	0	10233	5228	5005	48.91%
	1	125	102	23	18.40%
symptom_O09	0	10260	5247	5013	48.86%
	1	98	83	15	15.31%
symptom_O10	0	10282	5274	5008	48.71%
	1	76	56	20	26.32%
symptom_O11	0	8783	4517	4266	48.57%
	1	1575	813	762	48.38%
symptom_P01	0	10012	5062	4950	49.44%
	1	346	268	78	22.54%
symptom_P02	0	10133	5151	4982	49.17%
	1	225	179	46	20.44%
symptom_P03	0	10287	5270	5017	48.77%
	1	71	60	11	15.49%

variable	value	n	y		survival
			0	1	
symptom_U01	0	4067	2083	1984	48.78%
	1	6291	3247	3044	48.39%
symptom_U02	0	4694	2413	2281	48.59%
	1	5664	2917	2747	48.50%
symptom_U03	0	6777	3490	3287	48.50%
	1	3581	1840	1741	48.62%
symptom_U05	0	9339	4712	4627	49.54%
	1	1019	618	401	39.35%
symptom_U06	0	8243	4250	3993	48.44%
	1	2115	1080	1035	48.94%
symptom_S04	0	7813	4032	3781	48.39%
	1	2545	1298	1247	49.00%
symptom_S07	0	6229	3251	2978	47.81%
	1	4129	2079	2050	49.65%
symptom_S10	0	9795	4974	4821	49.22%
	1	563	356	207	36.77%

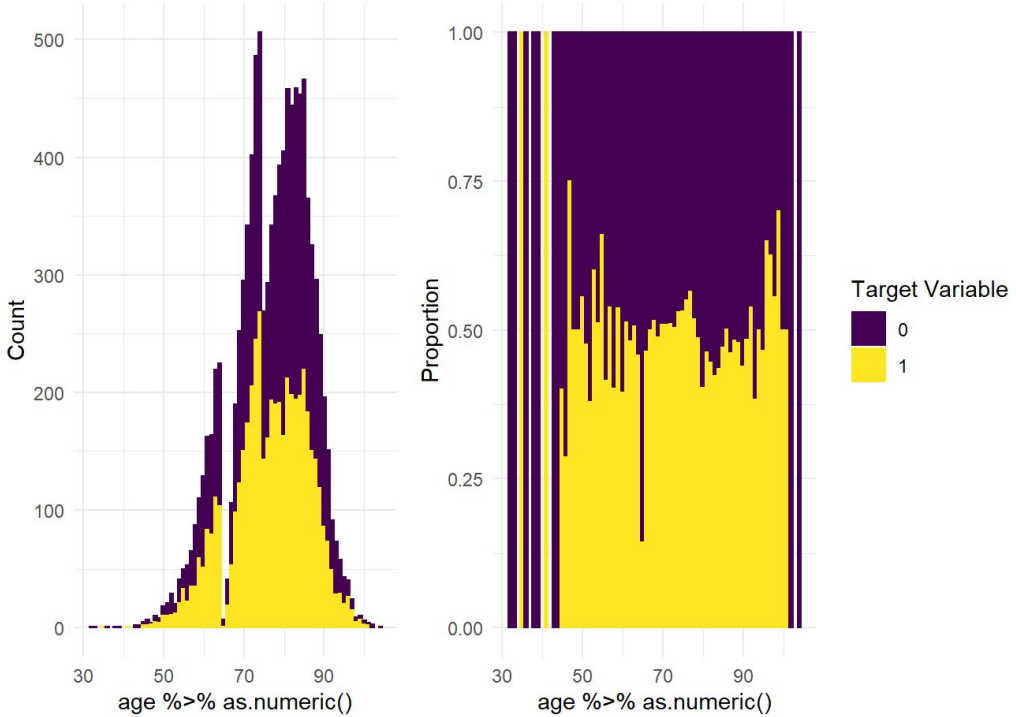
Exploratory Data Analysis #3 - **Numerics**



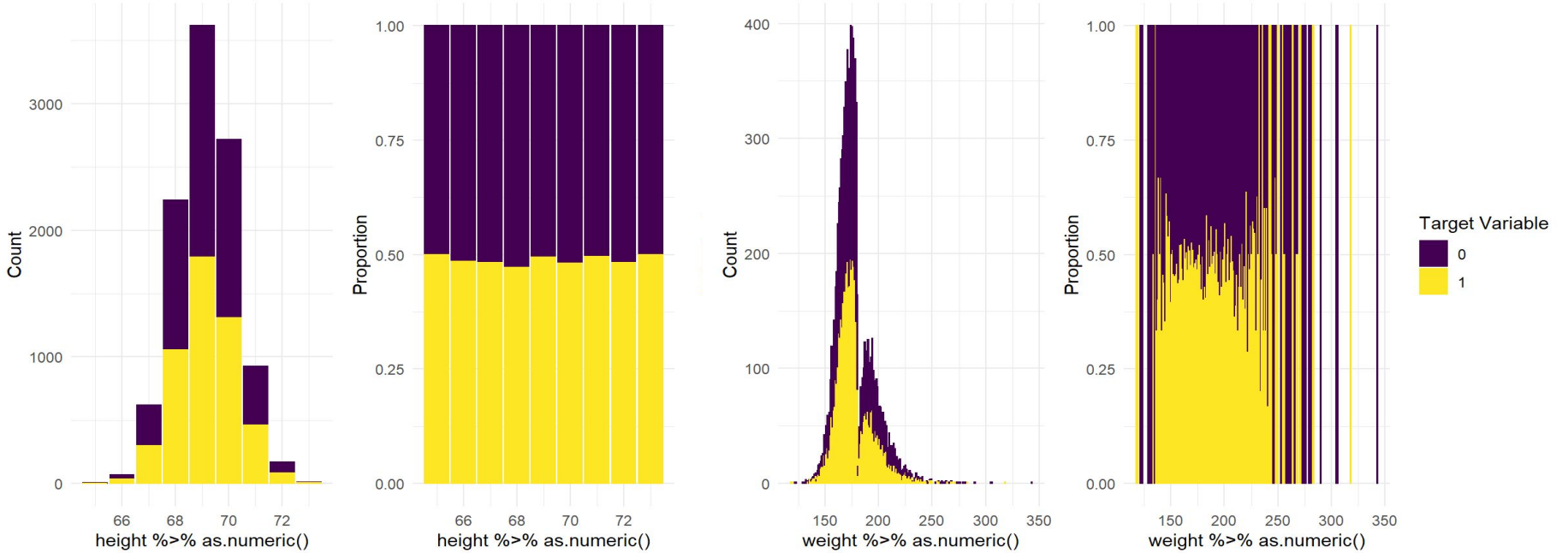
Exploratory Data Analysis #4 - **Numerics**



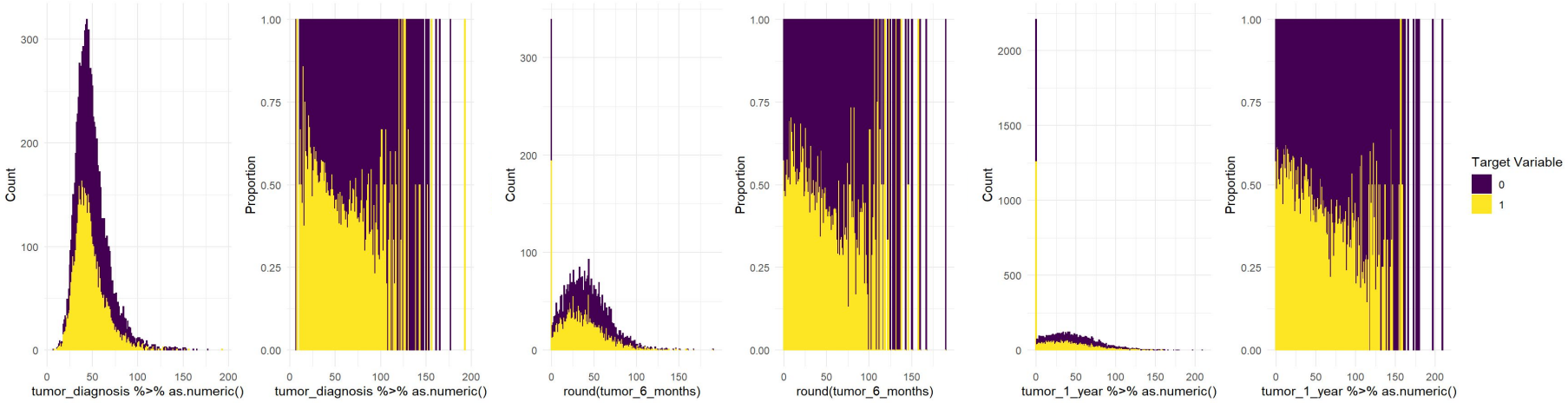
Exploratory Data Analysis #5 - **Numerics**



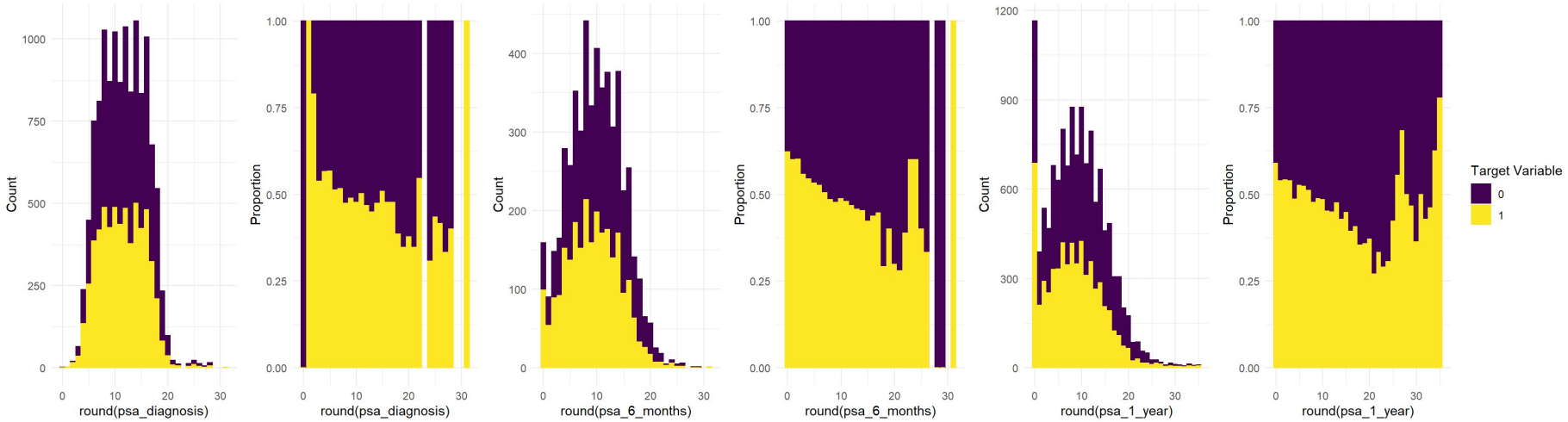
Exploratory Data Analysis #6 - **Numerics**



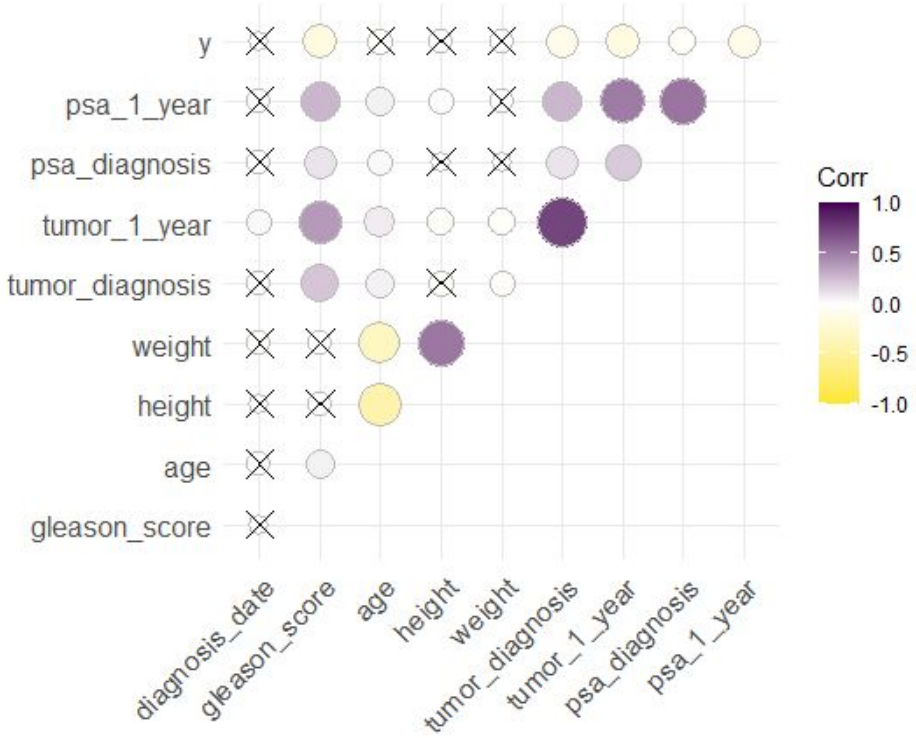
Exploratory Data Analysis #7 - **Numerics**



Exploratory Data Analysis #8 - **Numerics**



Exploratory Data Analysis #9 - **Numerics CorrMap**



Train-Test **Splitting**

- Split proportion:
  - 80% training
  - 20% testing

Phase 1 Modeling: **Basic** Models

- Processing steps:
  - **Tokenizing** symptom string
  - Deriving **Term-Frequency** from tokens
  - Set **Reference levels** for Dummy variables
  - Create **Dummy** variables
  - Filtering for **no-variance** predictors

Basic Models #1: Naive **Sampling** (no features)

- `y_hat` is predicted completely blindly based on distribution of training `y=(0,1)`.

```
blind <- function(dframe){
  dframe %>%
    mutate(
      y_hat = sample(c("0","1"), replace = T, prob = probs.train, size = nrow(dframe))
    )
}
```

model	accuracy	f_meas	lift
Naive Sampling	0.5113101	0.5695309	0.0000000

Basic Models #2: **Manual** Model

```r
manual1 <- function(dframe) {
  dframe %>%
    mutate(
      y_hat = case_when(
        # 1-yr Dead Rule
        survival_1_year == 0 ~ 0,
        # Stage Rule
        stage %in% c("IIB", "III", "IV") ~ 0,
        # Metastatisation Rule
        m_score != "M0" ~ 0,
        n_score != "N0" ~ 0,

        # Symptoms Rule
        `_symptoms_o01` == 1 ~ 0,
        `_symptoms_o08` == 1 ~ 0,
        `_symptoms_o09` == 1 ~ 0,
        `_symptoms_o10` == 1 ~ 0,
        `_symptoms_p01` == 1 ~ 0,
        `_symptoms_p02` == 1 ~ 0,
        `_symptoms_p03` == 1 ~ 0,
        `_symptoms_s10` == 1 ~ 0,
        TRUE ~ 1
      )
    )
}
```

- Manual model with few rules, and no decision tree. Simple flag-setting.

model	accuracy	f_meas	lift
Manual Rules	0.6140434	0.7310345	0.2009217

Basic Models #3: **Naive Bayes**

- Default, untuned Naive Bayes model.

model	accuracy	f_meas	lift
Naive Sampling	0.5113101	0.5695309	0.0000000
Naive Bayes	0.5730443	0.7224265	0.1207373
Manual Rules	0.6140434	0.7310345	0.2009217

Basic Models #4: base **GLM** LogReg

- Default, untuned GLM Logistic Regression Classifier.

term	odds_ratio	p.value	sig
n_score_N1	0.44	0.0000000	***
gleason_score	0.89	0.0000000	***
tumor_1_year	0.99	0.0000000	***
tf_symptoms_u05	0.65	0.0000002	***
rd_thrpy	0.75	0.0000134	***
tf_symptoms_s10	0.64	0.0001020	***
rad_rem	0.79	0.0012385	**
weight	0.99	0.0012616	**
brch_thrpy	0.81	0.0020365	**

model	accuracy	f_meas	lift
Naive Sampling	0.5113101	0.5695309	0.0000000
Naive Bayes	0.5730443	0.7224265	0.1207373
Manual Rules	0.6140434	0.7310345	0.2009217
base GLM	0.6743638	0.7075751	0.3188940

Phase 2 Modeling: **Intermediate** Models

- Additional Feature Engineering:
  - Calculating BMI & binning into **weight classes** (under/normal /overweight/obese)
  - Calculating **PSA** & **tumor** size change: **1-year delta**
  - **Centering & scaling** all numerics *(for regularised GLM family only - xgBoost & tree family does not benefit from scaling)*

model	accuracy	f_meas	lift
Naive Sampling	0.5113101	0.5695309	0.0000000
Naive Bayes	0.5730443	0.7224265	0.1207373
Manual Rules	0.6140434	0.7310345	0.2009217
base GLM	0.6743638	0.7075751	0.3188940

Phase 2 Modeling: **Intermediate** Models

- Cross-validation split: **4-fold CV**, 75/25. CV on `d.train` only.
- Parameter Tuning:
  - Parameters:
    - **Elastic-net GLM**: 6x6 standard sampling grid
      - `penalty` : how much regularisation?
      - `mixture` : what type of regularisation?
    - **xgBoost**: length 30 Latin hypercube sampling grid
      - `learn_rate` : learning rate.
      - `loss_reduction` : required loss reduction for further node-split
      - `min_n` : minimum data points at a node for further node-split
      - `mtry` : proportion of predictors randomly sampled at split
      - `sample_size` : proportion of data used in fitting
      - `tree_depth` : max tree depth.

Intermediate Models #1: base GLM w/ FE

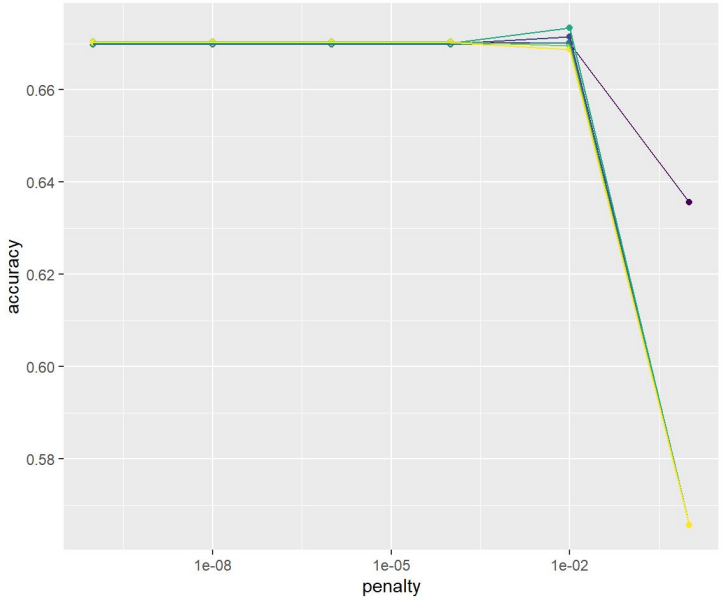
- **Unpenalised GLM** logistic classifier model. Uses feature-engineered dataset.

term	estimate	p.value	sig
n_score_N1	-0.3688596	0.0000000	***
gleason_score	-0.2207874	0.0000000	***
tumor_1_year	-0.3135726	0.0000000	***
tf_symptoms_u05	-0.1323507	0.0000002	***
rd_thrpy	-0.1435097	0.0000122	***
tf_symptoms_s10	-0.1026216	0.0000985	***
wgt_class_obese	-0.0924944	0.0004049	***
rad_rem	-0.0921186	0.0012108	**
brch_thrpy	-0.0868727	0.0024341	**

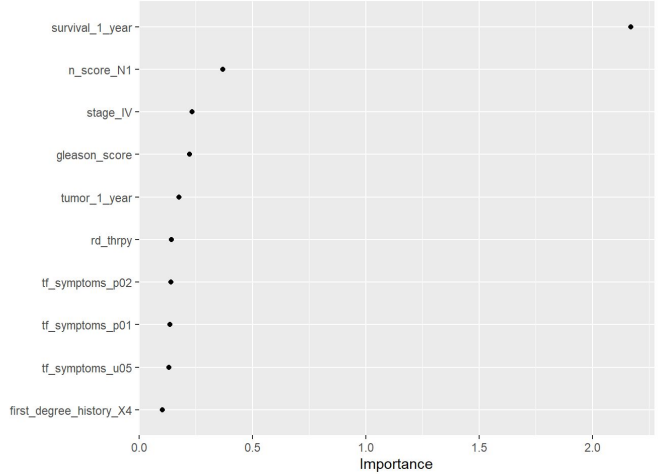
model	accuracy	f_meas	lift
Naive Sampling	0.5113101	0.5695309	0.0000000
Naive Bayes	0.5730443	0.7224265	0.1207373
Manual Rules	0.6140434	0.7310345	0.2009217
base GLM	0.6743638	0.7075751	0.3188940
engineered GLM	0.6757776	0.7084746	0.3216590

Intermediate Models #2: elastic-net **GLM**

- Tuned, **penalised GLM** logistic classifier model. Uses feature-engineered dataset.



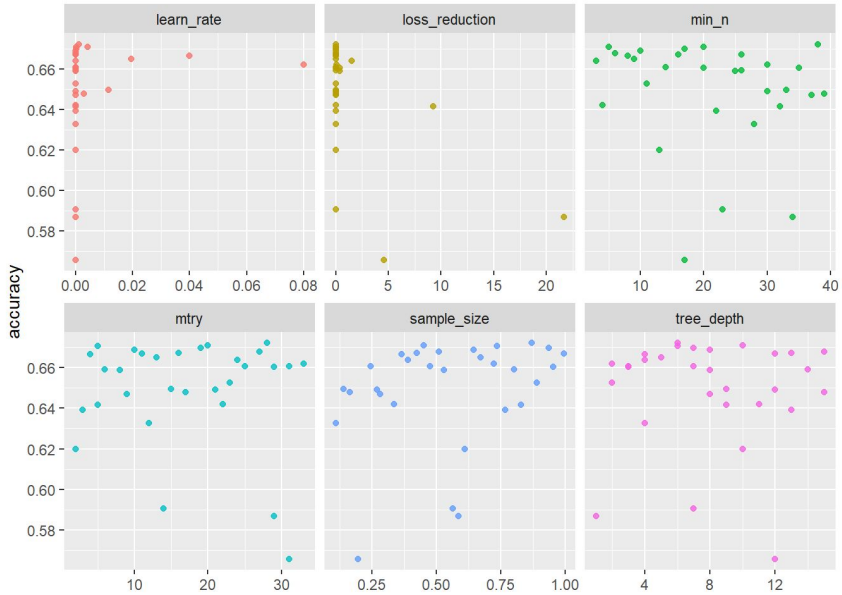
term	estimate
survival_1_year	0.9837585
n_score_N1	-0.3640233
gleason_score	-0.2035380
tumor_1_year	-0.1341893
rd_thrpy	-0.1143969
tf_symptoms_p01	-0.1019169
tf_symptoms_u05	-0.0986897
tf_symptoms_p02	-0.0930415
tumor_delta1	-0.0741324
tf_symptoms_s10	-0.0699019
stage_III	0.0565106
tf_symptoms_p03	-0.0533405
stage_IV	-0.0532372
rad_rem	-0.0516380
wgt_class_obese	-0.0467116
tf_symptoms_o08	-0.0464231
tf_symptoms_o09	-0.0384841
brch_thrpy	-0.0321105
multi_thrpy	-0.0207395
m_score_M1b	-0.0100925
m_score_M1a	-0.0097641
family_history_X1	-0.0096943
stage_IIB	-0.0091137
t_score_T1c	-0.0070674
m_score_M1c	-0.0051965
t_score_T3c	-0.0025840
t_score_T1b	0.0018512
wgt_class_underweight	0.0005630
t_score_T2a	0.0003439
age	0.0000000

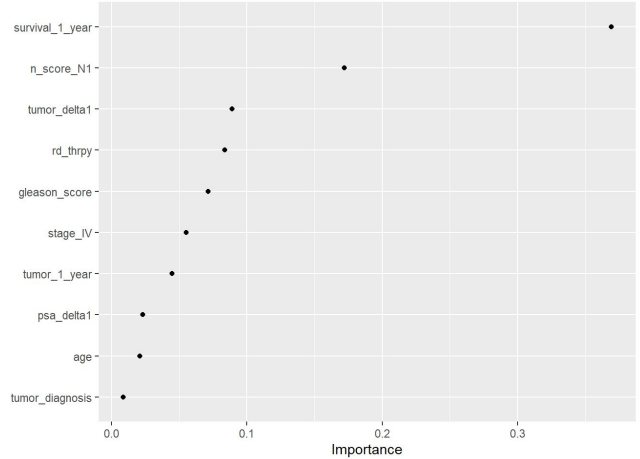


model	accuracy	f_meas	lift
engineered GLM	0.6757776	0.7084746	0.3216590
Elastic-net GLM	0.6795476	0.7145256	0.3290323

Intermediate Models #3: **xgBoost**

- Tuned, **xgBoost** classifier model. Uses feature-engineered dataset.





model	accuracy	f_meas	lift
engineered GLM	0.6757776	0.7084746	0.3216590
Elastic-net GLM	0.6795476	0.7145256	0.3290323
xgBoost	0.6833176	0.7152542	0.3364055

Intermediate Models: Model Performances

model	accuracy	f_meas	lift
Naive Sampling	0.5113101	0.5695309	0.0000000
Naive Bayes	0.5730443	0.7224265	0.1207373
Manual Rules	0.6140434	0.7310345	0.2009217
base GLM	0.6743638	0.7075751	0.3188940
engineered GLM	0.6757776	0.7084746	0.3216590
Elastic-net GLM	0.6795476	0.7145256	0.3290323
xgBoost	0.6833176	0.7152542	0.3364055

Phase 3 Modeling: the Human Touch

- Recall that one of the variables is `survival_1_year`. This is a perfect predictor for y=0 (not y=1), but some models might not be able to converge on this rule, which could be a way to squeeze out some more performance from the model.
- Additionally, we can add more screening rules from the manual model earlier.
- Thus, we stack this screening step on a model. Experimental code:

```
# Stacking a check for survival_1_year on top of any fitted model
manual_stack <- function(dframe,modelfit){

  df_man <- dframe[dframe$survival_1_year == min(dframe$survival_1_year),]%>%
    mutate(.pred_class = 0)

  df_alg <- dframe[dframe$survival_1_year != min(dframe$survival_1_year),]
  df_alg <- df_alg %>% cbind(predict(modelfit, new_data = df_alg))

  result <- rbind(df_man,df_alg)
  return(result)
}
```

Phase 3 Modeling: the Human Touch

- Final Metrics:

model	accuracy	f_meas	lift
Naive Sampling	0.5113101	0.5695309	0.0000000
Naive Bayes	0.5730443	0.7224265	0.1207373
Manual Rules	0.6140434	0.7310345	0.2009217
base GLM	0.6743638	0.7075751	0.3188940
engineered GLM	0.6757776	0.7084746	0.3216590
Elastic-net GLM	0.6795476	0.7145256	0.3290323
Manual + Elastic-net	0.6795476	0.7145256	0.3290323
xgBoost	0.6833176	0.7152542	0.3364055
Manual + xgBoost	0.6875589	0.7187102	0.3447005