Chronic Kidney Disease Dataset Challenge

Cancer Progression Prediction

6th December 2021

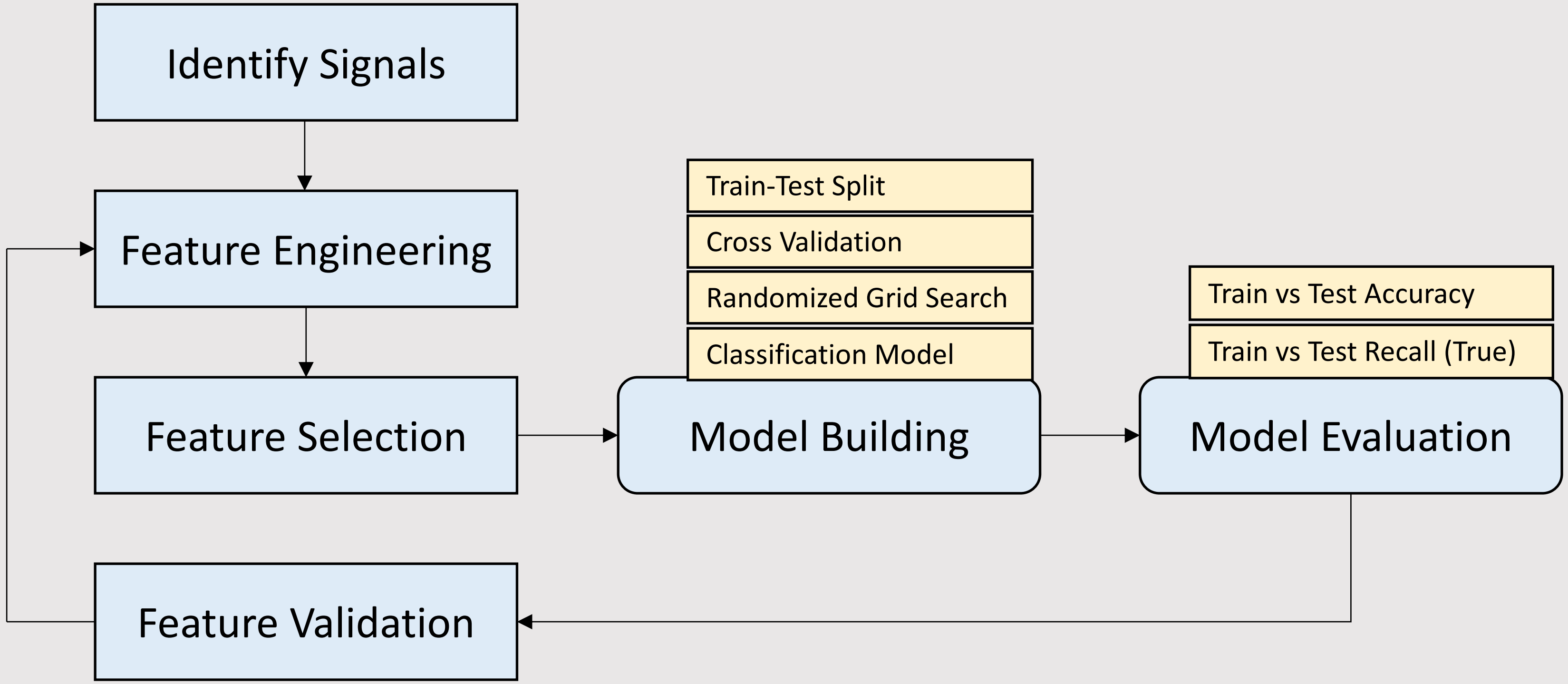
Content

1. Problem Statement

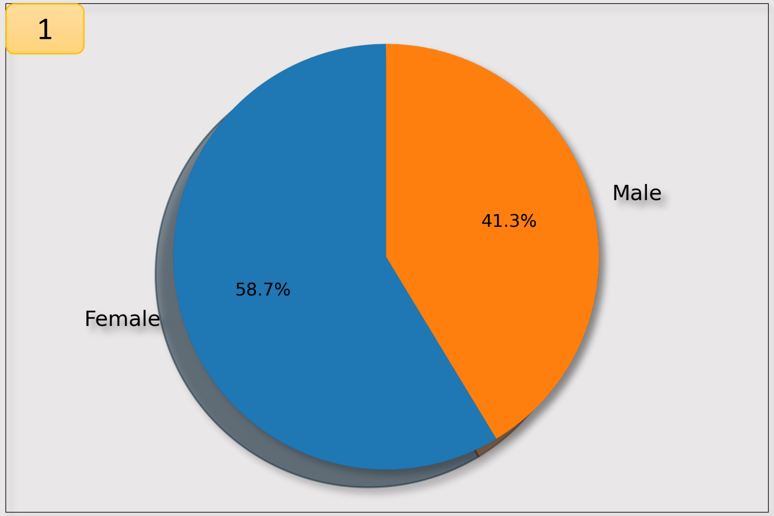
Objective

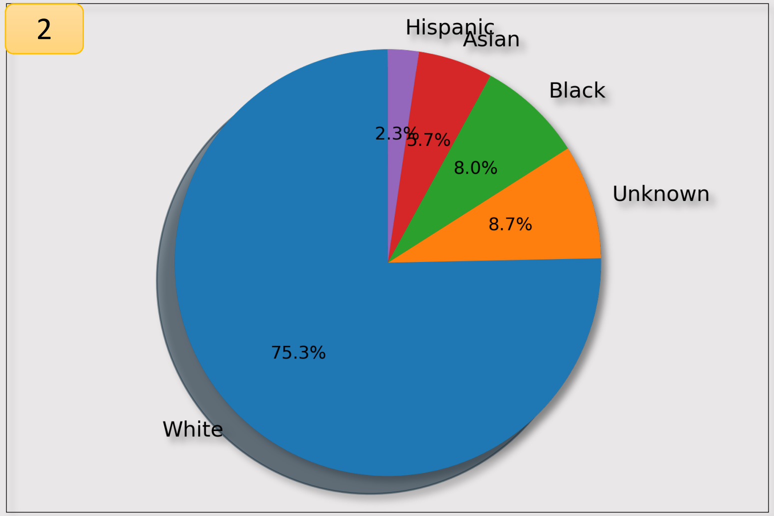
Predict Cancer prognosis/ progression given patients' lab diagnosis & medication data

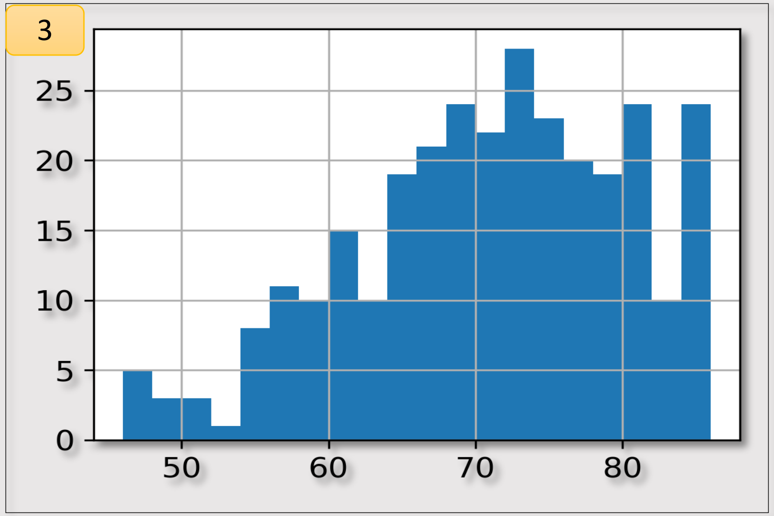
Algorithm

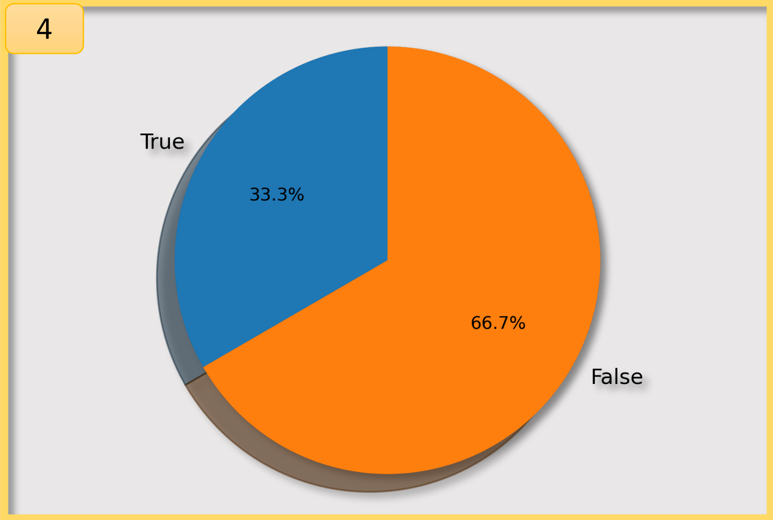


2. Data Insights- Patient Demographics







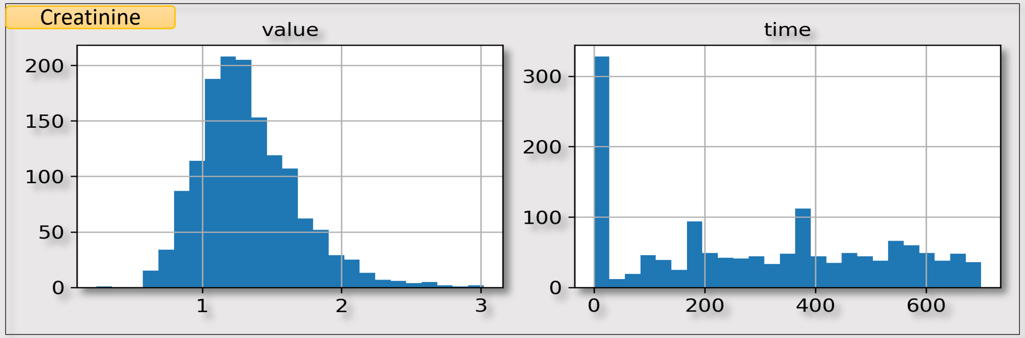


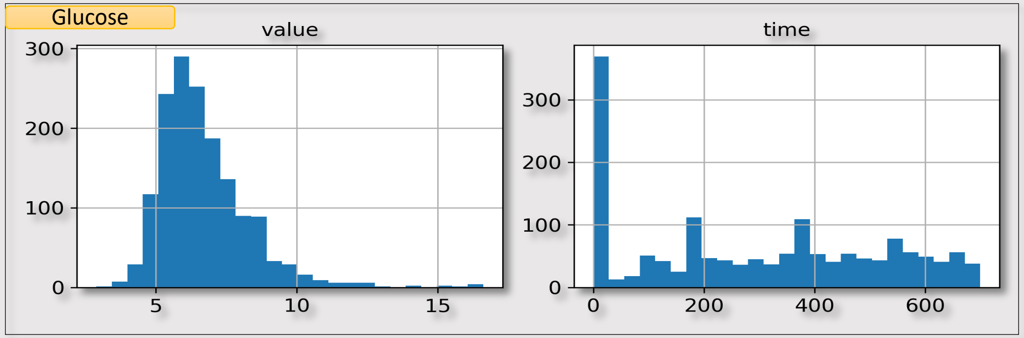
- 300 Patient IDs: id range- 0-299
- Demographic features:
  (1) Gender
  (2) Race
  (3) Age
- Class label
  (4) -> *True*: Cancer progresses
  -> *False*: Cancer do not progress
- No Missing Values
- No Class imbalance problem

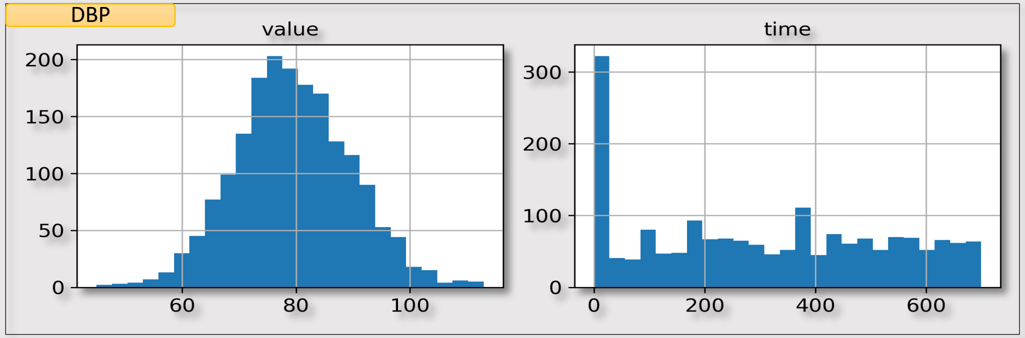
2. Data Insights: Patient Lab Measurements

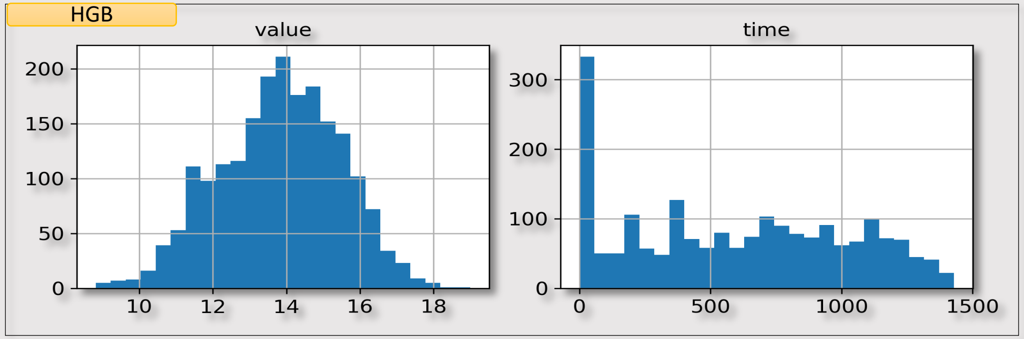
	Male		Female		High Implies
Test	Low Range	High Range	Low Range	High Range	
Creatinine	0.75	1.35	0.59	1.04	kidney problem
DBP (Diastolic Blood Pressure)	85	89	85	89	hypertension
Glucose	3.9	5.5	3.9	5.5	diabetes
HGB	13.8	17.2	12.1	15.1	cancer
ldl	100-129	160-189	100-129	160-189	heart
SBP	0	255	0	255	infection

Source	Data Availability
Medication Data	91%
Creatinine	100%
DBP	100%
Glucose	100%
HGB	100%
ldl	100%
SBP	100%

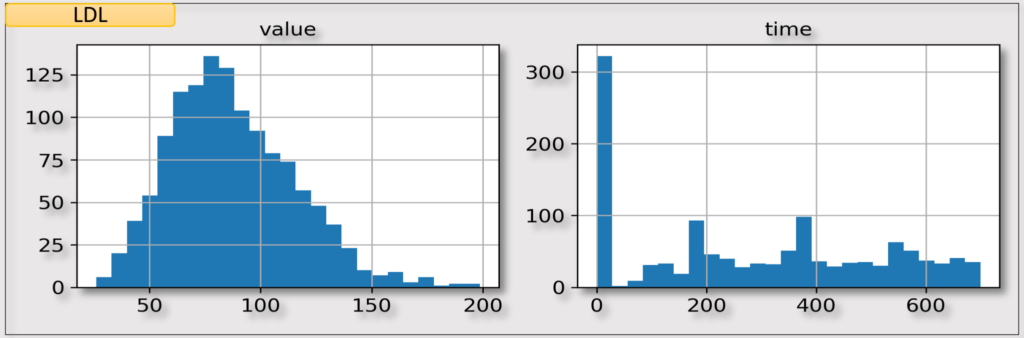


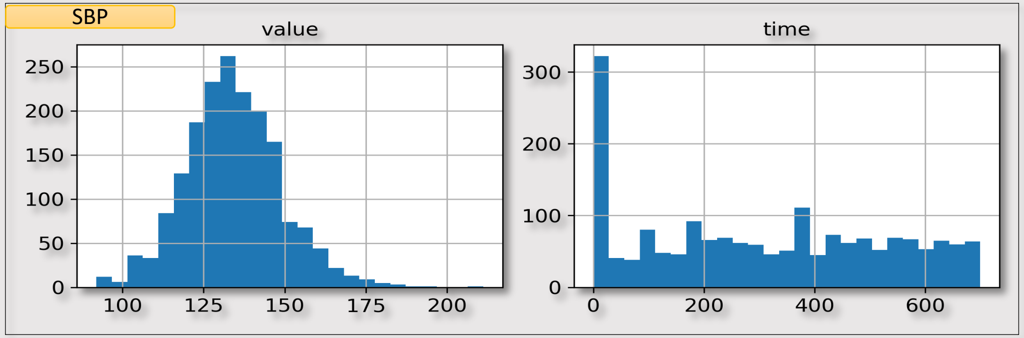






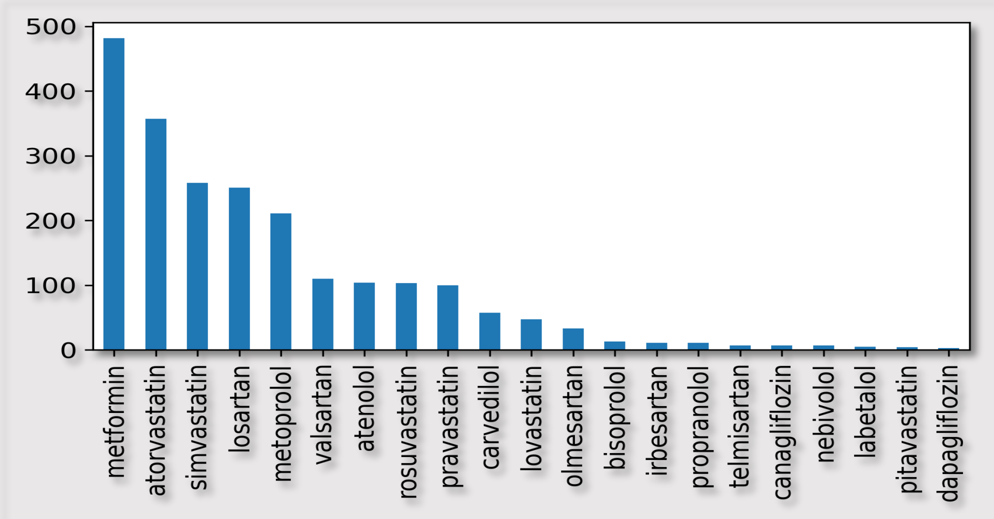
2. Data Insights: Patient Lab Measurements

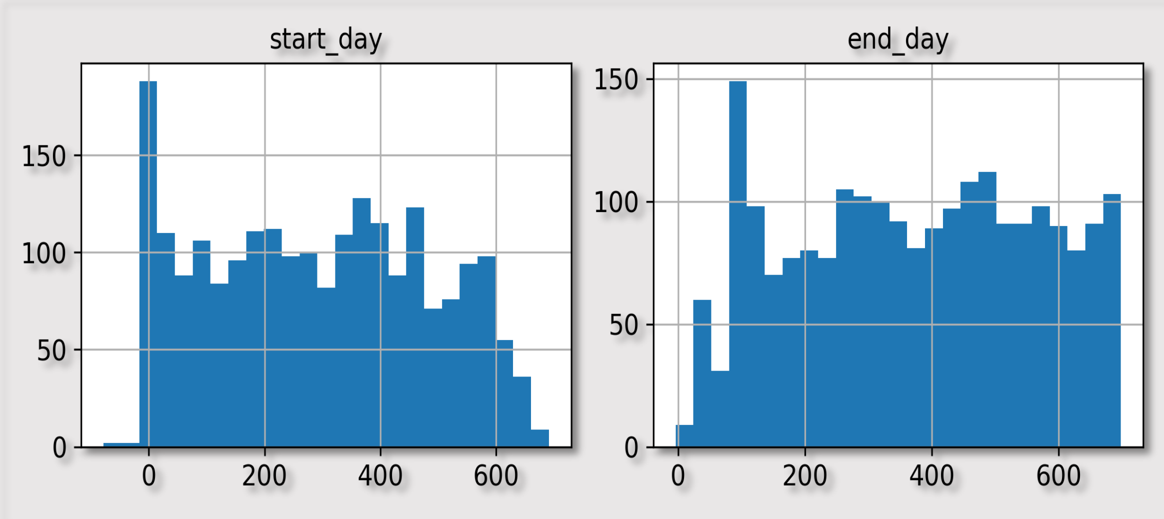


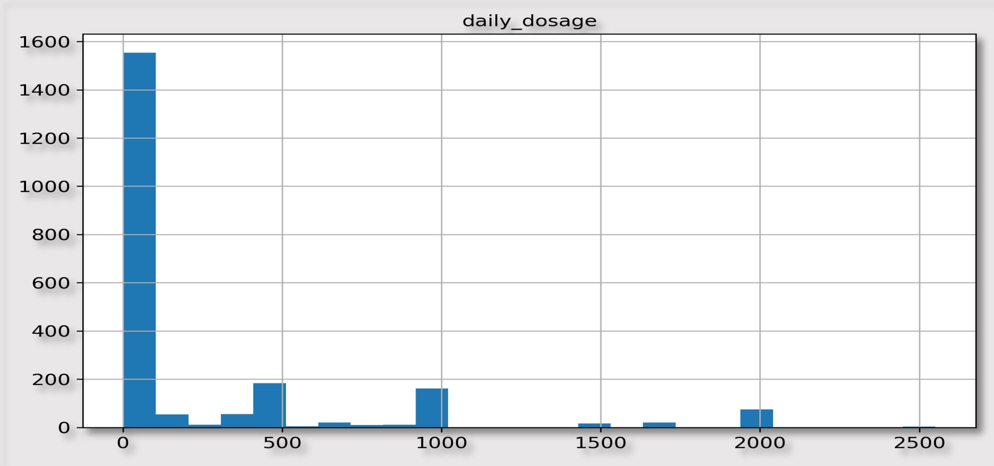


- All lab measurements span over 0-699 days, except for HGB, for which data is present across 0-1499 days (! could double check with the client if this is expected)
- For 100% of patients, there is at least 1 lab measurement available for all Test types
- These attributes have varying scales (there is a need for feature scaling)
- Few histograms here are *tail-heavy:* they extend too much farther to the right of the median than to the left (need for feature transformation to make them bell-shaped)

2. Data Insights: Patient Medication Data







- Medication logs missing for 10% of patients (! could double check with the client, if this is expected)
- Need for missing value imputation (check with the client if 0 works or drop those patient ids)
- Drugs dosage values have different scales (need for feature scaling)
- On high level, only half of total available drugs are prescribed (data sparsity problem)

3. Feature Engineering

Feature	Feature Category	Type	Pre-processing Steps
Lab Test Longitudinal - First Value	For every Lab Test Type	Numerical	1. Transformation 2. Standardization
Lab Test Longitudinal - Last Value		Numerical	
Lab Test Longitudinal - First Time		Numerical	
Lab Test Longitudinal - Last Time		Numerical	
Lab Test Longitudinal - Average Value		Numerical	
Lab Test Longitudinal - Median Value		Numerical	
Lab Test Longitudinal - Maximum Value		Numerical	
Lab Test Longitudinal - Minimum Value		Numerical	
Lab Test Longitudinal - Last – Minimum Value		Numerical	
Lab Test Longitudinal - Weighted Moving Average Value		Numerical	
Drug - Medication Indicator (1/0)	For every Drug Type	Numerical	
Drug - Maximum Dosage Duration		Numerical	
Drug - Average Daily Dosage		Numerical	
Drug - Last Dosage		Numerical	
Gender	Demographic	Categorical	Label Encoding
Race	Demographic	Categorical	One Hot Encoding
Age	Demographic	Numerical	As is

4. Feature Selection: Medication Data

Drug	False Case	False Count	True Case	True Count	Total Count	% of Total
atenolol	66.12%	36	33.88%	32	68	4%
atorvastatin	76.34%	210	23.66%	71	281	16%
bisoprolol	84.18%	5	15.82%	5	10	1%
canagliflozin	84.83%	6	15.17%	1	7	0%
carvedilol	67.70%	24	32.30%	25	49	3%
dapagliflozin	100.00%	3	0.00%		3	0%
irbesartan	0.00%		100.00%	8	8	0%
labetalol	100.00%	5	0.00%		5	0%
losartan	71.90%	120	28.10%	63	183	11%
lovastatin	66.70%	29	33.30%	15	44	3%
metformin	61.48%	221	38.52%	157	378	22%
metoprolol	51.01%	88	48.99%	85	173	10%
nebivolol	0.00%		100.00%	7	7	0%
olmesartan	25.58%	12	74.42%	15	27	2%
pitavastatin	100.00%	3	0.00%		3	0%
pravastatin	53.99%	44	46.01%	37	81	5%
propranolol	100.00%	11	0.00%		11	1%
rosuvastatin	69.58%	63	30.42%	18	81	5%
simvastatin	75.36%	119	24.64%	67	186	11%
telmisartan	6.95%	1	93.05%	6	7	0%
valsartan	45.72%	48	54.28%	45	93	5%

- Medication Data only created pertaining to highlighted drugs
- Drug selection explanation: those which create maximum split (at least 60:40) AND are given to >= 3% of patients
- This analysis is done only based on Train Set

5. Model Building & Model Evaluation

Feature Used	Modelling Technique	Model Description	Train - Accuracy	Test - Accuracy	Train - Recall	Test - Recall	Train - Precision	Test - Precision
Only Lab Longitudinal	Logistic Regression	{'class_weight' = {0:0.5, 1:1.75}}	77%	70%	77%	89%	60%	52%
Lab + Demog	Logistic Regression		76%	73%	91%	89%	60%	55%
Lab + Demog + Med	Logistic Regression		78%	75%	89%	79%	63%	58%
Only Lab Longitudinal	Random Forest	{'bootstrap': True, 'class_weight': {0: 0.2, 1: 0.8}, 'max_depth': 3, 'min_samples_leaf': 2, 'min_samples_split': 10, 'n_estimators': 50}	85%	73%	96%	84%	70%	55%
Lab + Demog	Random Forest		78%	63%	95%	89%	62%	46%
Lab + Demog + Med	**Random Forest**		**90%**	**77%**	**98%**	**84%**	**78%**	**59%**
Only Lab Longitudinal	SVM	{'C'=100, 'class_weight'={0: 0.5, 1: 1.75}, 'gamma'=0.01)	85%	77%	95%	89%	70%	59%
Lab + Demog	SVM		87%	72%	94%	79%	75%	54%
Lab + Demog + Med	SVM		94%	72%	100%	58%	85%	55%

6. Feature Validation

1. Demographic + Lab Longitudinal Features gives good performance
2. Random Forest outperforms Logistic and SVM
3. Adding medication data drops model performance – need domain specific knowledge
4. Adding demographic data (i.e., Age, Gender, Race) gives Accuracy & Recall (True) lift.
5. Running model with multiple features gives good fit on training but fails against test data – Overfitting – need for more data
6. Logistics generalizes better than Random Forest (whereas RF fits 100% on training data) – need for more data

6. Feature Validation: Feature Importance

Feature	Feature Description	Feature Importance Score
last_minus_1st_ldl	Last Value Minus First Value for LDL Lab Test	0.194799
weighted_average_ldl	Average Test Value Change Per Day for LDL Lab Test	0.124618
last_minus_1st_SBP	Last Value Minus First Value for SBP Lab Test	0.100481
last_minus_1st_glucose	Last Value Minus First Value for Glucose Lab Test	0.080617
last_minus_1st_DBP	Last Value Minus First Value for DBP Lab Test	0.060523
weighted_average_SBP	Average Test Value Change Per Day for SBP Lab Test	0.056314
weighted_average_glucose	Average Test Value Change Per Day for Glucose Lab Test	0.047286
weighted_average_HGB	Average Test Value Change Per Day for HGB Lab Test	0.046495
last_minus_1st_HGB	Last Value Minus First Value for HGB Lab TesT	0.04511
weighted_average_DBP	Average Test Value Change Per Day for DBP Lab Test	0.044946

7. Misclassification Analysis

Analyze False Negative Cases (Type 2 Error)

There are total 7 such cases (out of 60) where Model predicts: FALSE but actual Stage_Progress is TRUE

Row Labels	FALSE	TRUE	Grand Total
creatinine	0.052389937	0.017037037	0.040458333
DBP	-2.670628931	3.220123457	-0.6825
glucose	-0.436477987	0.516790123	-0.11475
HGB	-0.302389937	-0.211111111	-0.271583333
ldl	-11.87106918	9.363580247	-4.704375
SBP	-5.314465409	5.976666667	-1.503708333
Grand Total	-3.423773585	3.14718107	-1.206076389

Case: Patient ID 2 and 45

- Patients average test value change per day is close to 0 whereas in Training data: average test values ranges from -2.76 to + 3.22 (DBP as an example)

pid	test_name	weighted_average	Stage_Progress	Pred_Stage_Progress
2	creatinine	-0.003697701	TRUE	FALSE
2	DBP	0.056232555	TRUE	FALSE
2	glucose	-0.007201138	TRUE	FALSE
2	HGB	-0.021693369	TRUE	FALSE
2	ldl	0.116184713	TRUE	FALSE
2	SBP	0.01525459	TRUE	FALSE

pid	test_name	weighted_average	Stage_Progress	Pred_Stage_Progress
45	creatinine	0.000329547	TRUE	FALSE
45	DBP	-0.913354472	TRUE	FALSE
45	glucose	0.004339264	TRUE	FALSE
45	HGB	-0.032712704	TRUE	FALSE
45	ldl	0.047667587	TRUE	FALSE
45	SBP	-0.918631285	TRUE	FALSE

8. Model Limitations & Next Steps

Next Steps:

1. Collect more data
2. Feature using domain expert consultation to design features
3. Ensemble of multiple model output scores

Limitations:

1. Overfitting in few model technique scenarios