WhiteCoat

NUS National University of Singapore | ISS INSTITUTE OF SYSTEMS SCIENCE

HEALTH WELLNESS (CHRONIC DISEASE) PREDICTION FOR Client X

CAPSTONE TEAM #7

Project Supervisor : Brian Ang (ISS)

17 Sep 2022

© 2022 National University of Singapore. All Rights Reserved

Agenda

1. Introduction
2. Project Recap
3. Data / Modeling / Deployment Pipeline
4. EDA (hold-out dataset)
5. Modeling results
   a. By Members ID
   b. By Bookings ID
   c. By Lab Report
6. Model Explainability
7. Future Work
8. Conclusion

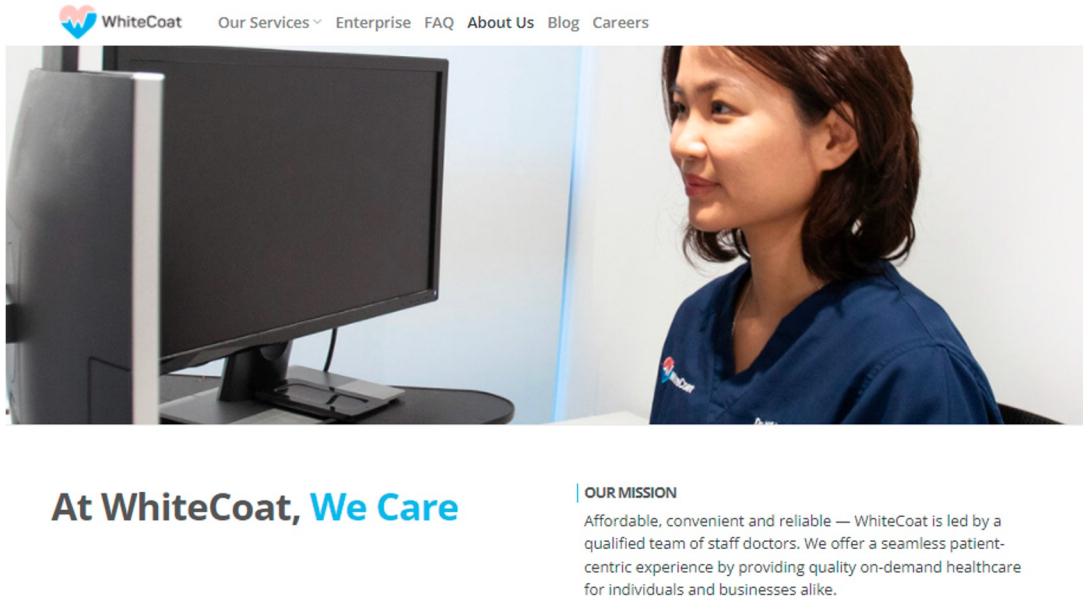
© 2022 National University of Singapore. All Rights Reserved

1.Introduction

© 2022 National University of Singapore. All Rights Reserved

Company Background

- **Company**: Client X
  - Leading telehealth service offering video consultations through their in-house app
  - Services offered include GP, mental wellness, paedetrician appointments and health screening (personal or enterprise)
  - Primarily based in Client X, with services in Client X



© 2022 National University of Singapore. All Rights Reserved

Project Objective

Aim: To build a predictive model that can predict holistic patient wellness for chronic patients

Target for prediction: Chronic disease (10 labels provided in total), eg.

Group 1: Diabetes, hypertension, hyperlipidemia
Group 2: Atopic dermatitis, asthma, vasomotor & allergic rhinitis
Group 3: Migraine, gastro-esophageal reflux disease, gout, thyroid

Product
ML model for Patient Wellness Prediction

Ideal Customer
Patients with high likelihood of suffering from Chronic Diseases

Features
Predict patients who are suffering from chronic diseases

Experience
High-risk patients could be notified to schedule appointment and go for diagnosis tests

Benefits
Early diagnosis and treatment

© 2022 National University of Singapore. All Rights Reserved

Technical Overview

Data Sources:

- Initial sampled dataset of 5000 patients provided by Client X in 2021 (training data)
- Holdout data in Jan 2022
- Client X bookings, patients, diagnoses, lab reports etc.
  - Features include age, gender, symptoms, prescriptions, 10 chronic disease labels

Model building:

- Different ways of aggregating target (chronic disease labels) for prediction
- Variety of modelling techniques explored
  - Logistic regression, random forest, linear/ non-linear SVM, XGBoost
- Model validation of holdout dataset

© 2022 National University of Singapore. All Rights Reserved

2. Project Recap

© 2022 National University of Singapore. All Rights Reserved

Deliverables

Phase I (Recap)	Phase II

• EDA of **training** dataset (sampled 2021)
 	
• Data preparation
  • Denormalization of training datasets to produce 3 output datasets (booking-level, member-level and lab report-level)
 	
  • Data dictionary of output datasets
 	
• Research on academic papers
  • Demographic features for chronic disease prediction
  • Predicting individual disease over overall patient wellness
	• EDA of **holdout** dataset (Jan 2022)
	
	• Data preparation, modelling & Deployment Pipeline
	
	• Modelling Results
	  • Performance Evaluation
	  • Explainability

© 2022 National University of Singapore. All Rights Reserved

Summary of Datasets

Datasets Overview

Filename	Description	Training data	Holdout data	Cols
tbl_Booking.csv	**Consultations by sampled patients**	**10,221**	**7,725**	**8**
tbl_Booking_Diagnosis.csv	Doctor's diagnosis of patient after consultation	65,744	9,200	4
tbl_Booking_Prescription.csv	Medication prescribed by doctor	20,242	14,926	6
tbl_Booking_Symptoms.csv	Self-reported symptoms before consultation	18,096	15,263	3
tbl_Lab_Reports.csv	Results of laboratory reports	12,956	-	6
tbl_Patient.csv	**Sampled patients**	**5,000**	**6,774**	**3**

Brief Summary

Description	Training data	Holdout data
Time Period	**Jan - Dec 2021**	**Jan 2022**
No. of Bookings	10,221	7,725
No. of Patients	5,000	6,774
No. of Diagnoses	1,249	498
No. of Symptoms	21	18
No. of Medicine	406	284
No. of Lab Reports	225	-
No. Chronic Groups	10	10

Booking Symptoms

Column	Explanation	Datatype
booking_id	Unique consult identifier	String
MDSymptom_id	Unique symptom identifier	String
MDSymptom_name	Current symptom that patient experienced	String

Booking Diagnosis

Column Name	Explanation	Datatype
booking_id	Unique consult identifier	String
MDDiagnosis_code	ICD-10 code of diagnosed illness	String
MDDiagnosis_name	Name of illness as per ICD-10	String
chronic_group	Client X's classification of chronic disease(s) to model for	String

Booking Prescription

Column Name	Explanation	Datatype
booking_id	Unique consult identifier	String
MDMedication_ID	Unique medication identifier	String
MDMedication_name	Name of medication prescribed	String
is_fulfilled_externally	Whether medication prescribed was available in Client X's inventory during that consult	Boolean
prescription_quantity	Number of medication prescribed	Integer
prescription_status	Whether patient decided to purchase the given medication or not	String

Patient

Column	Explanation	Datatype
mem_id	Unique patient identifier	String
gender	Patient gender	String
birth_year	Year of birth for each patient	Integer

Lab Reports

Column Name	Explanation	Datatype
mem_id	Unique patient identifier	String
lab_report_no	Unique lab report identifier	Integer
requested_date	Date when the lab test was ordered	Date
category	Category of lab test	String
line_item	Specific test line item within category	String
test_result	Result of the given line item	Various
unit	Unit for the given line item	String

Booking "Base"

Column	Explanation	Datatype
booking_id	Unique consult identifier	String
mem_id	Unique patient identifier	String
doctor_id	Unique doctor identifier	String
consult_waived_reason	Reason for consult fee being waived, if so. Dash = consult fee was not waived.	String
is_mc_issued	Whether MC was issued for that consult	Boolean
mc_days	Number of MC days issued, if so	Integer
is_referral_letter_issued	Whether a referral letter was issued for that consult	Boolean
booking_timestamp	When the consult occurred in Singapore Time (GMT+8)	Datetime

3. Data / Modeling / Deployment Pipeline

© 2022 National University of Singapore. All Rights Reserved

Process and Pipeline

The core of modeling project is processes and pipelines:

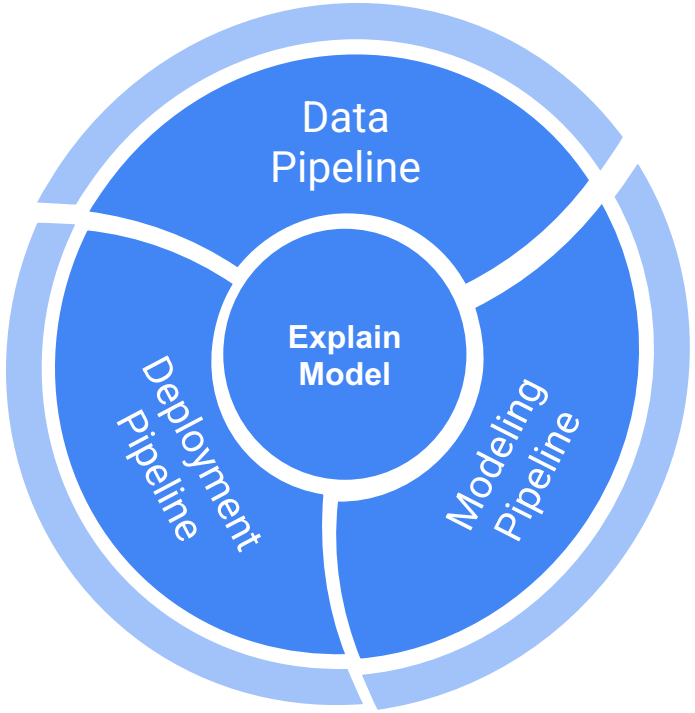
a. Data Pipeline
- i. Cleaning
- ii. Pre-processing
- **iii. Save Prepared Base Datasets**

b. Modeling Pipeline
- i. Data Preparation
- ii. Model Training / Fitting
- **iii. Save Features, Models, Parameters, Results,**

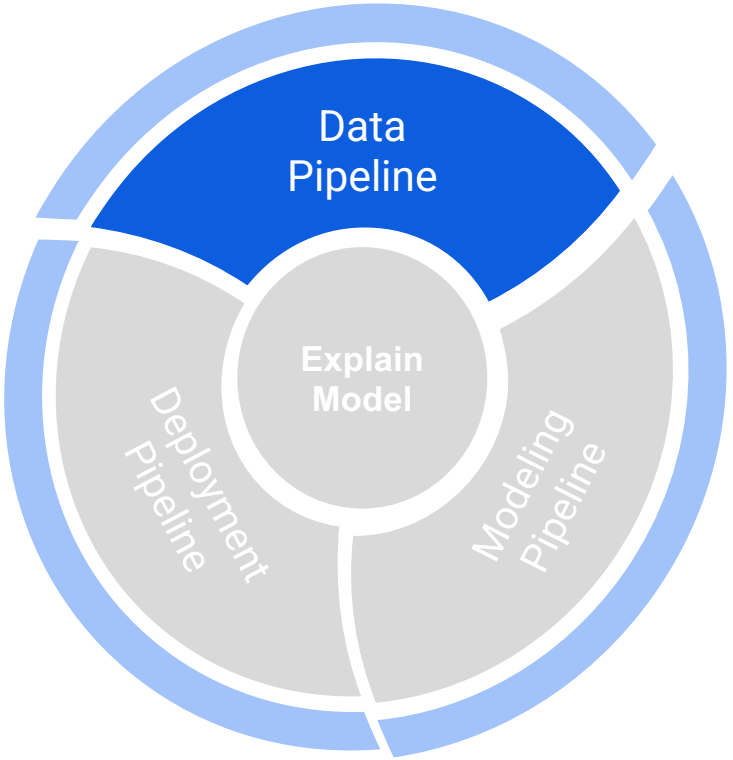
c. Deployment Pipeline
- i. Testing / Evaluation
- ii. Scoring
- **iii. Save Scored List**

d. Model Explainability Templates

A machine learning pipeline is **the end-to-end construct that orchestrates the flow of data into, and output from, a machine learning model (or set of multiple models)**.
It includes: raw data input, features, outputs, model and model parameters, prediction outputs.



a.Data Pipeline



Data Process & Pipeline
(build repeatable process)

QA & EDA after every major checkpoints

Dataset Testing
Baseline Model

Define Problem → Investigate Source Tables (Initial EDA Check & Clean) → Merge, Summarize, Derive new fields by Booking Denormalize → EDA Denormalized Booking Data → Merge, Summarize, Derive new fields Denormalize by Member → EDA of Denormalized Member Data → Run Baseline Model

Objective:
Predict Chronic Diseases (booking / member)

Target:
All Chronic, ~~individual Chronic~~, 3 High, Atopic Triad, Others

Basic Cleaning

Generate Base Dataset by Booking

Generate Base Dataset by Member

Data Dictionary

Data Pipeline Processing

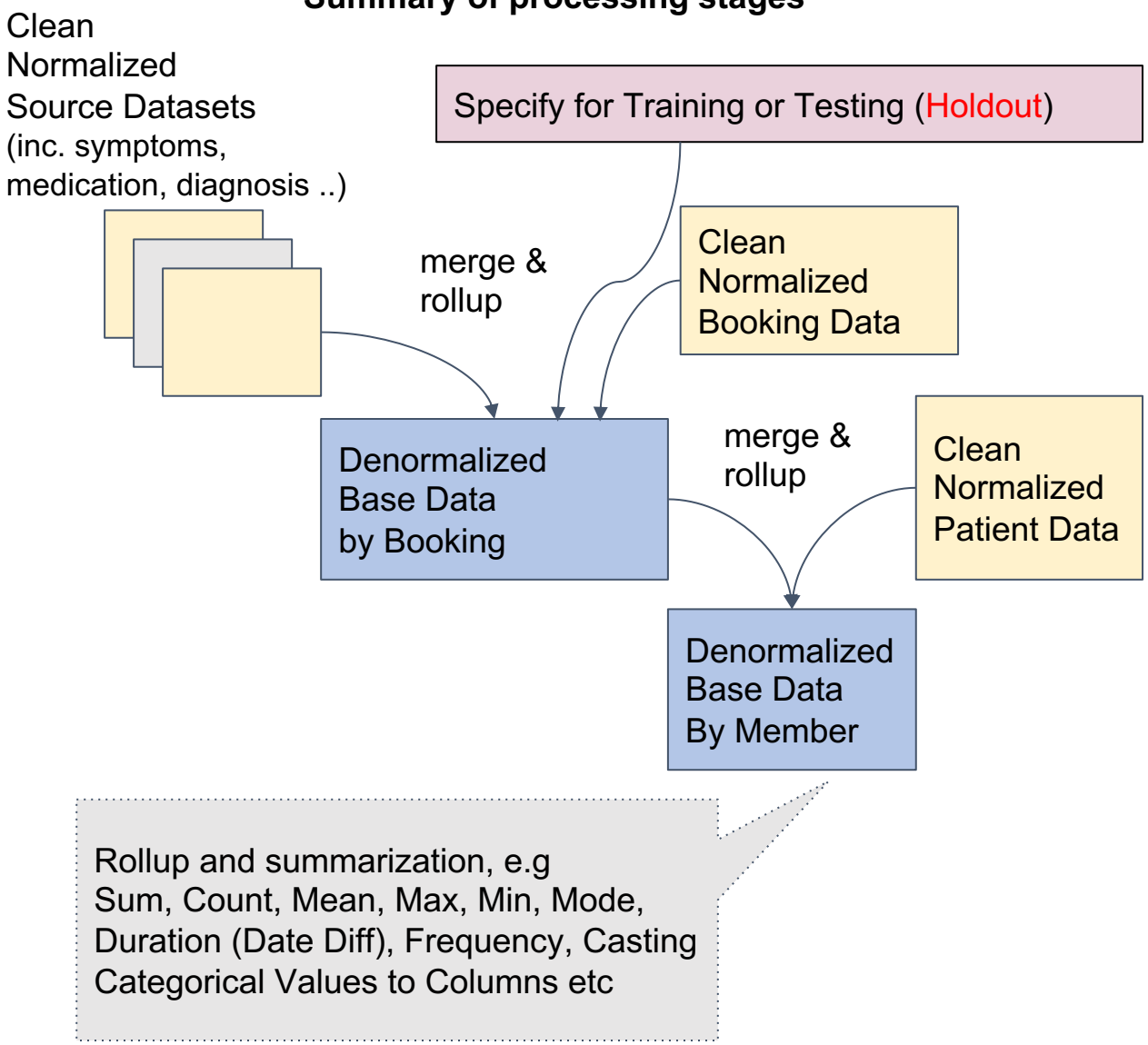
- Refactor to streamline pipeline
- Parameterized > > more generic
- Consolidate for all data preparation
- To incorporate rolling summary

Code template for illustration

Contents

1 Import Package
2 Set Global Variable
2.1 Set input filenames
2.2 Set for Testing or Training
2.3 Set for Member or Booking Modeling
2.4 Set previous booking related source file (for Testing)
2.5 Set minium count for feature considerations
3 Helper Function
4 Step1 - Intro Prep & Clean
5 Step1a - Prep & Clean Booking Diagnosis
6 Step1b - Prep & Clean Booking Prescription
7 Step1c - Prep & Clean Booking Symptoms
8 Step2 - Merge Data Tables
9 Step3 - Intro Cast & Transpose for Member
10 Step3a - Cast & Transpose for Chronic
11 Step3b - Cast and Transpose for Diagnosis Code
12 Step3c - Cast & Transpose for Symptoms
13 Step4 - Merge Summarized Data
14 Step5 - Finalise Booking Data
15 Step6 - Data Aggregate by Member
16 Step7 - Define Group Targets
17 Step8 - Finalise Member Data

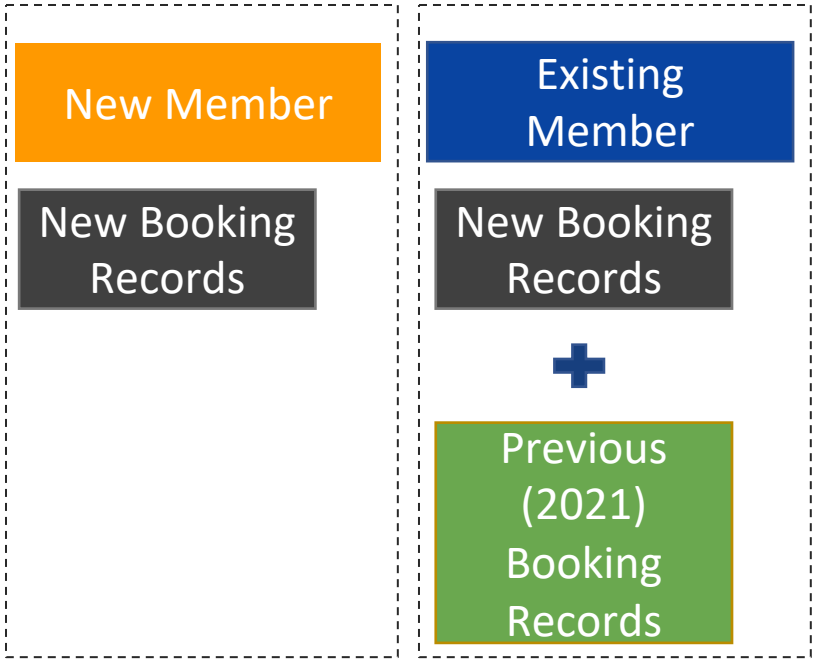
Summary of processing stages



Rolling Data & Other Consideration

- New holdout dataset include 714 existing member record
- For existing member, must include previous booking records for aggregation
- To include rolling summary booking data for existing member

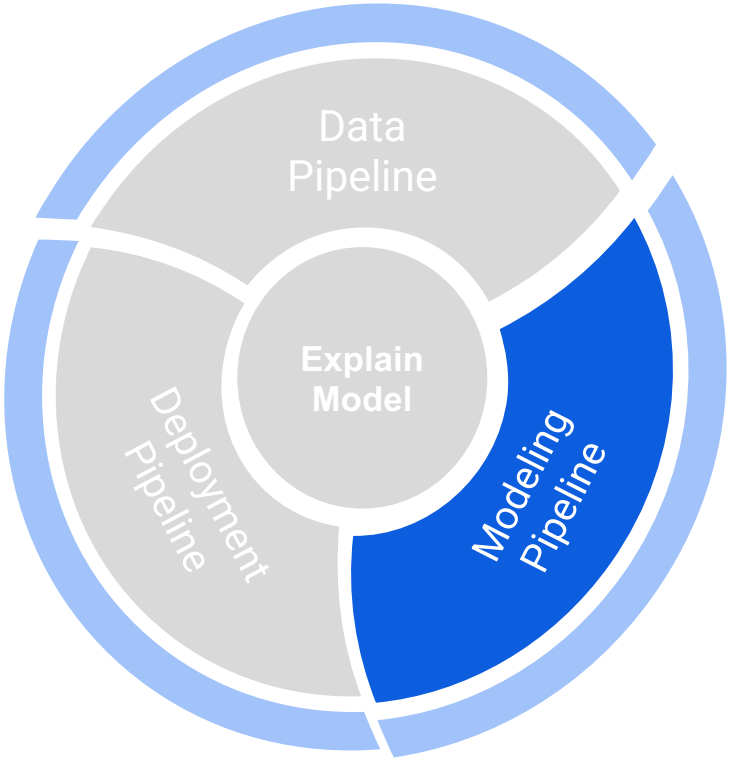
New Holdout Data (2022)



User future consideration:

- Rolling Timeline consideration
- Analytics DataMart [Out of Scope]
- Others:
  - Longer Modeling Data Timeframe,
  - Monitoring Metrics for model drift:
    - Data Drift
    - Concept Drift

b. Modeling Pipeline



Modeling Pipeline

(build repeatable process)

Prepared **Denormalized** data with feature engineering

Test and Select Feature

Further pre-process:
1. Handle rare / extreme observation
2. Data Standardisation
3. One-hot encoding

Explain model e.g. using **SHAP, LIME**

Create a Baseline to validate base data and modeling feasibilities

Define Problem → Prepare Modeling Data → Perform PCA, MCA, re-group → New Features → Train-Test Split → Pre-process & Transform → Train Model → Tune & Evaluate Model → Explain Model → Create Baseline for Checking

Predict Chronic, (booking / member)

Target:
- **all Chronic**
- ~~**individual Chronic**~~
- **3 Highs**
- **Atopic Triad**
- **Others**

Perform PC:
1. Dimension reduction
2. Handle multicollinearity

1. Train Test Split e.g. **70:30%**
2. **Random Stratified Sampling**
3. Test different approach on Train dataset to handle **extreme imbalanced** data (oversample, undersample, SMOTE etc)

1. Test model algorithms
2. Tune model parameters
3. Select Features
4. Evaluation Metrics:

Modeling Pipeline Output for Deployment

Prepare Modeling Data → Perform PCA, MCA, re-group → New Features → Train- Test Split → Pre-process & Transform → Train Model → Tune & Evaluate Model → Explain Model → **Model pickles for Deploy:**
- Selected Features
- Standard Scaler
- Models & Parameters
- Train/Test Result

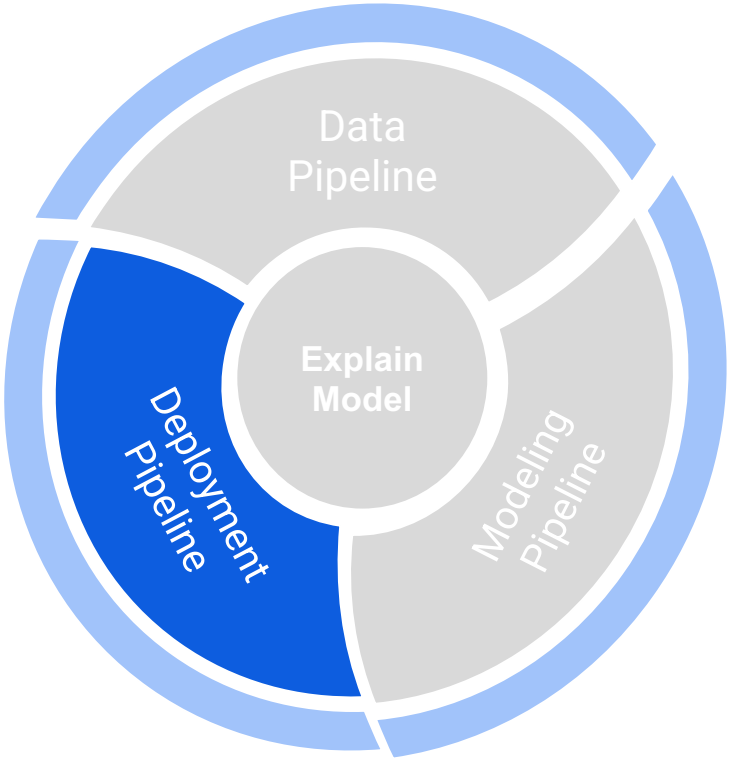
Code template for illustration

Contents

1 Import Library
2 Global Setting
2.1 Import File
2.2 Include Outlier Removal
2.3 Include SMOTE
2.4 Adjust Train Test Ratio
2.5 Choose SelectFromModel Algo
2.6 Choose to run Shap or not
3 Helper Function
4 Import Data
5 Basic Training Data Investigation¶
6 One Hot encoding
7 Define input and output
8 Train Test Split
9 Oversample for Training
10 Remove Outliers
11 Standardized Data
12 Model Selection
12.1 Logistic Model
12.2 Support Vector Machine (Linear)
12.3 Support Vector Machine (Non-Linear)
12.4 Random Forest
12.5 XGBoost
13 Feature Selected
14 ROC
15 Model Explanability

- Refactor, streamlined and made generic for multiple models
- Parameterize:
  - for model pre-processing options,
  - for model parameters,
- Dump pickles for deployment, Testing / Scoring:
  - Standardization
  - Selected input features
  - Saved Models & parameters
  - Train / Test Evaluation Result

c. Deployment Pipeline



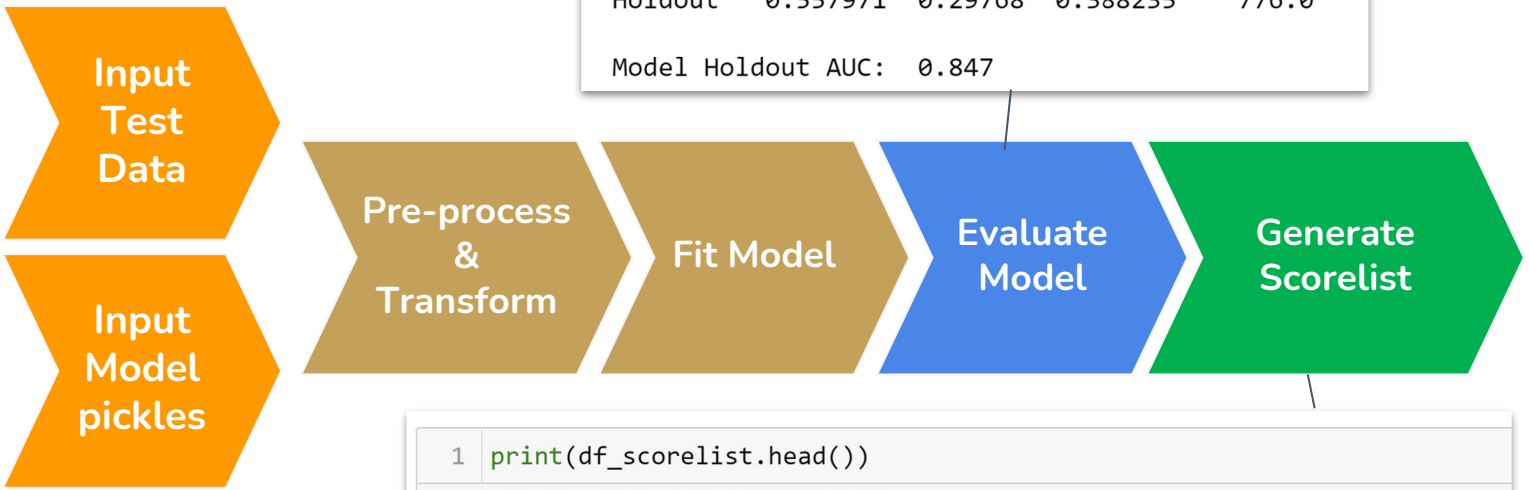
Deployment Pipeline: Testing & Scoring

- Streamlined code for Scoring, and Testing (if Target label available)

Code template for illustration

Contents

1 Import Library
2 Global Setting
3 Import Test Data
4 Import Model Training Pickles
5 Prepare Scoring Data
6 Model Result
7 Compare Result
8 Generate Score List



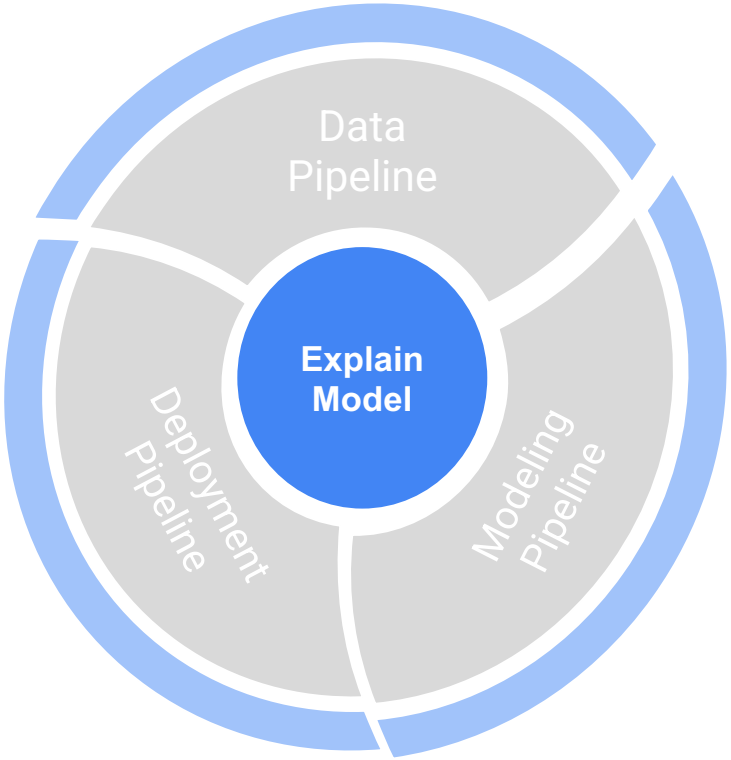

```
Model:  LogisticRegression

          precision   recall  f1-score  support
Train      0.738872  0.42132  0.536638    591.0
Test       0.733766  0.44664  0.555283    253.0
Holdout    0.557971  0.29768  0.388235    776.0

Model Holdout AUC:  0.847
```

```
1  print(df_scorelist.head())
```

executed in 14ms, finished 16:10:18 2022-08-30

	member_id	true_target	prediction
0	00059B0F-87F5-4D64-A5C8-441457913282	0	0
1	000789AE-E161-4FDE-B9B0-9377AB383A1B	1	0
2	0008AEAF-0767-46A2-830F-C98A9B88FF34	0	0
3	000C3E29-2C23-4EB0-A1FB-837EC7EA9E59	0	0
4	001C4B02-1827-4F58-85E8-AB985B27496A	0	0

d. Model Explainability



Model Explainability Templates

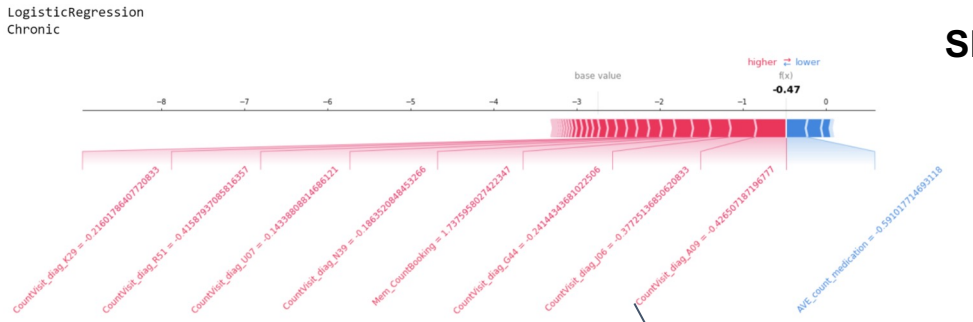
- Include Model Explainability Templates for Shap, Global Surrogate Model & LIME

Code template for illustration

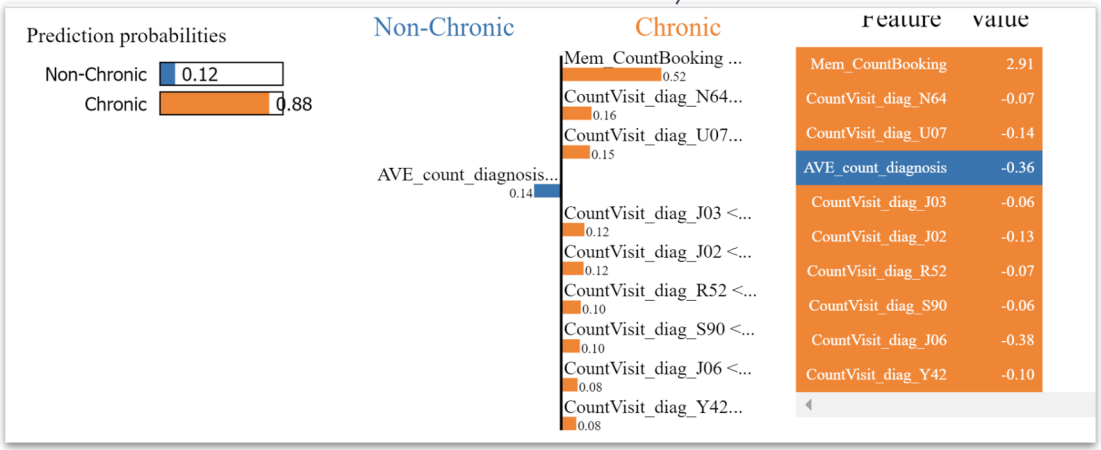
Contents

1 Import Library
2 Global Setting
3 Import Data
4 Import Model Training Pickles
5 Prepare Scoring Data
6 Model Result
7 Explain with Shap for Random Forest
8 Explain with Shap for XGB
9 Explain with Shap Logistic Explainer
10 Global Surrogate Model
11 LIME (Local Explainer)
12 Check index for Predictions that are positive

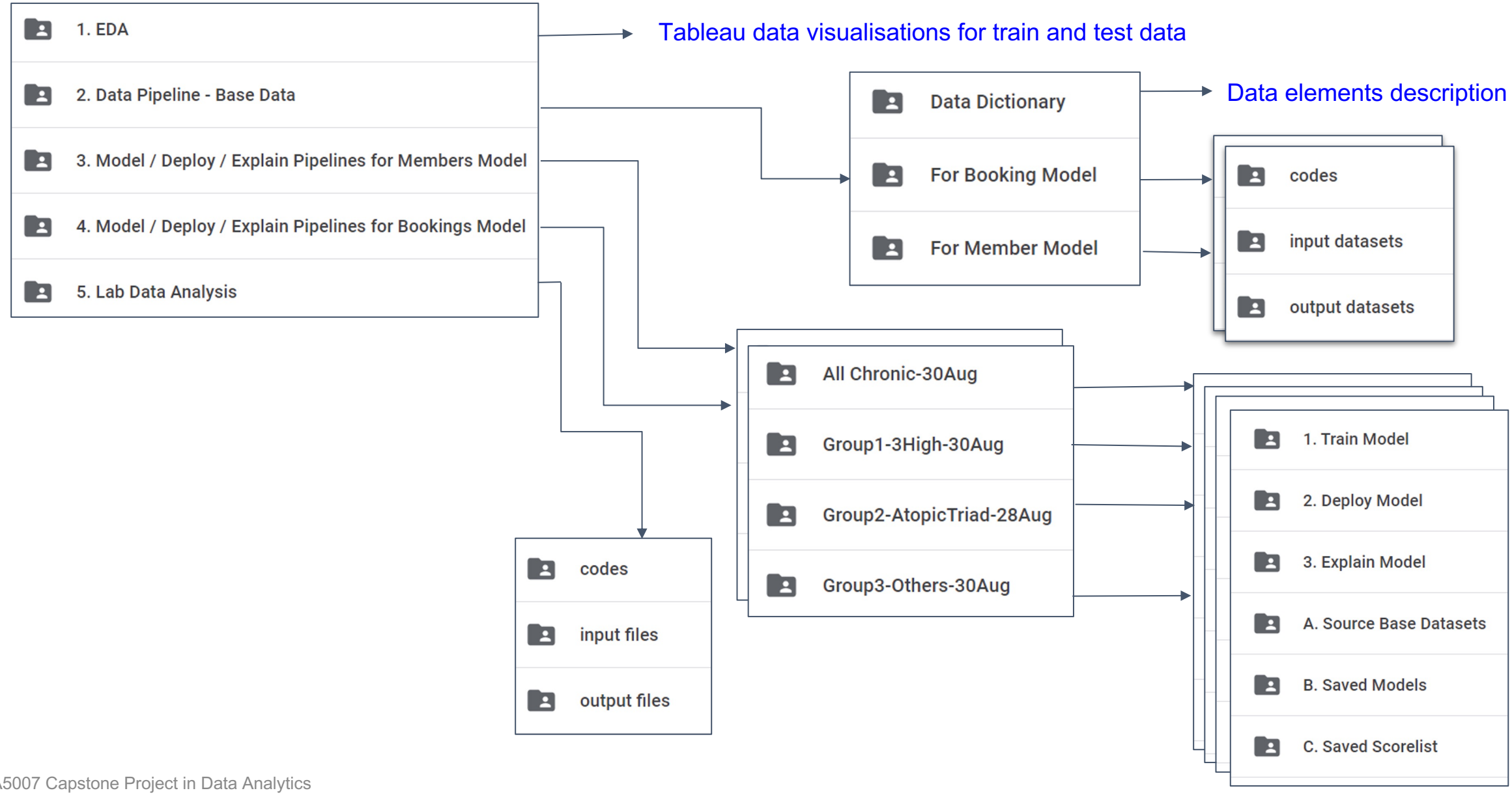
Shap



LIME



Codebase Handover to Client X



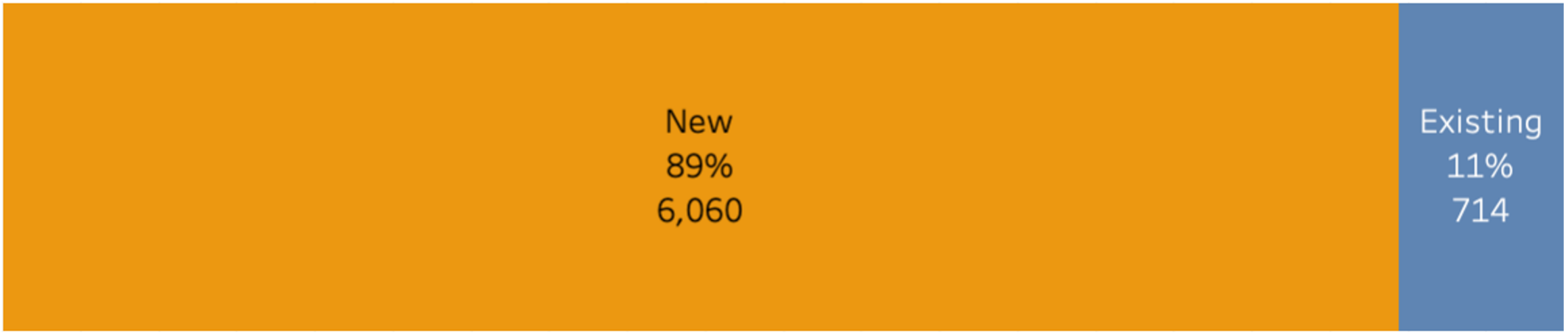
4. EDA (hold-out dataset)

© 2022 National University of Singapore. All Rights Reserved

New vs Existing Patient in 2021

Train/Test split by time - examine difference for potential data drift.

About 11% of the patients in Jan 2022 (Holdout datasets) are existing patients from 2021 (Training datasets)



Doctor Consultations

- Consultation were mostly done by 6 doctors in 2021

- Consultations in Jan 2022 are mostly done by 10 doctors including the 6 main ones from 2021
  - 1 doctor with only 139 bookings became the top doctor in number of consultations in Jan 2022 (likely newly joint in late 2021)
  - 2 new doctors in 2022 but yet to pick up in number of bookings

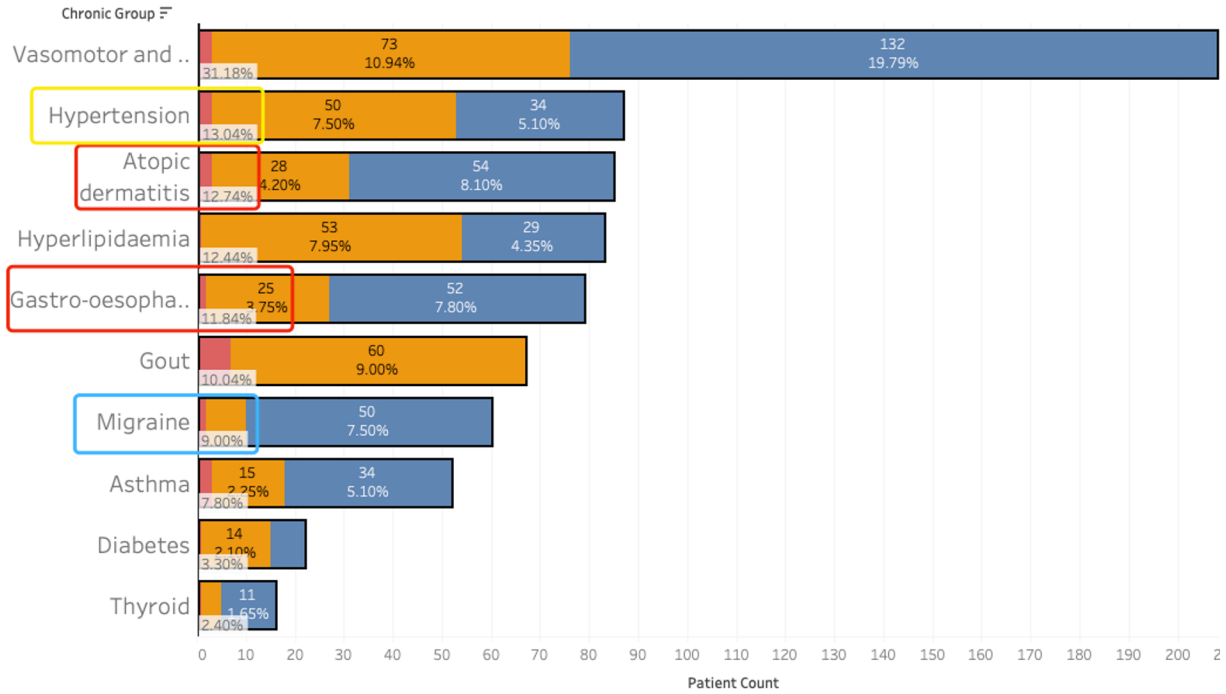
Doctor ID	Jan 2022 Bookings Count	2021 Bookings Count
DE28DA35-C28E-49FF-B77C-28E6268F0EF9	892	139
7C2FDBF3-2942-47AD-AE19-ECC6F4E115E9	849	1938
77EFF8BB-D181-4BFB-90A4-75FA494F4BBB	800	1543
251C07D2-FC5F-4229-9F7E-C662B754AFC6	783	1746
B045D798-DC64-4CA9-9FC4-28691E82EB72	778	1242
CCFD1554-88E9-4141-9D28-32A8194B8AD8	715	1101
EB022DC9-CBDB-4E11-950B-ECBB2287E2C3	666	1393
A66FDEC0-AA68-4EA4-9B49-CE24C8DD7955	508	116
84599BF3-5C45-4786-9D7C-16CDFA73F1CF	471	111
438F511A-DE86-496F-8AC6-BBE153A376D5	391	115
B428EC44-9615-4894-BA85-B6626419EC13	229	86
B2643AC3-F5D0-4BEA-AC16-B6B373E86AB8	196	36
C23693AF-97A7-4140-8E05-8D355540EFEE	147	77
44F5B766-8B80-456E-8797-5CF9893D6AE4	128	220
FB818D86-8723-4599-9152-C5A54701D544	48	24
AFD29C73-97BB-414A-AA15-43226CB4D966	46	new
602732C7-7471-4B0E-BBE1-1BAD2C0FDC6C	30	13
471F4903-195A-41AA-847A-41B16EFF3272	24	35
8CF93EBC-5210-4756-91F0-0F744D42FDB9	15	new
E0DA11A6-7ACF-48A7-B83C-EF2C97EE856B	7	128
B756C654-073C-459B-B509-F8AD89BA5028	2	3
Grand Total	**7725**	**10221**

Chronic Group Distribution

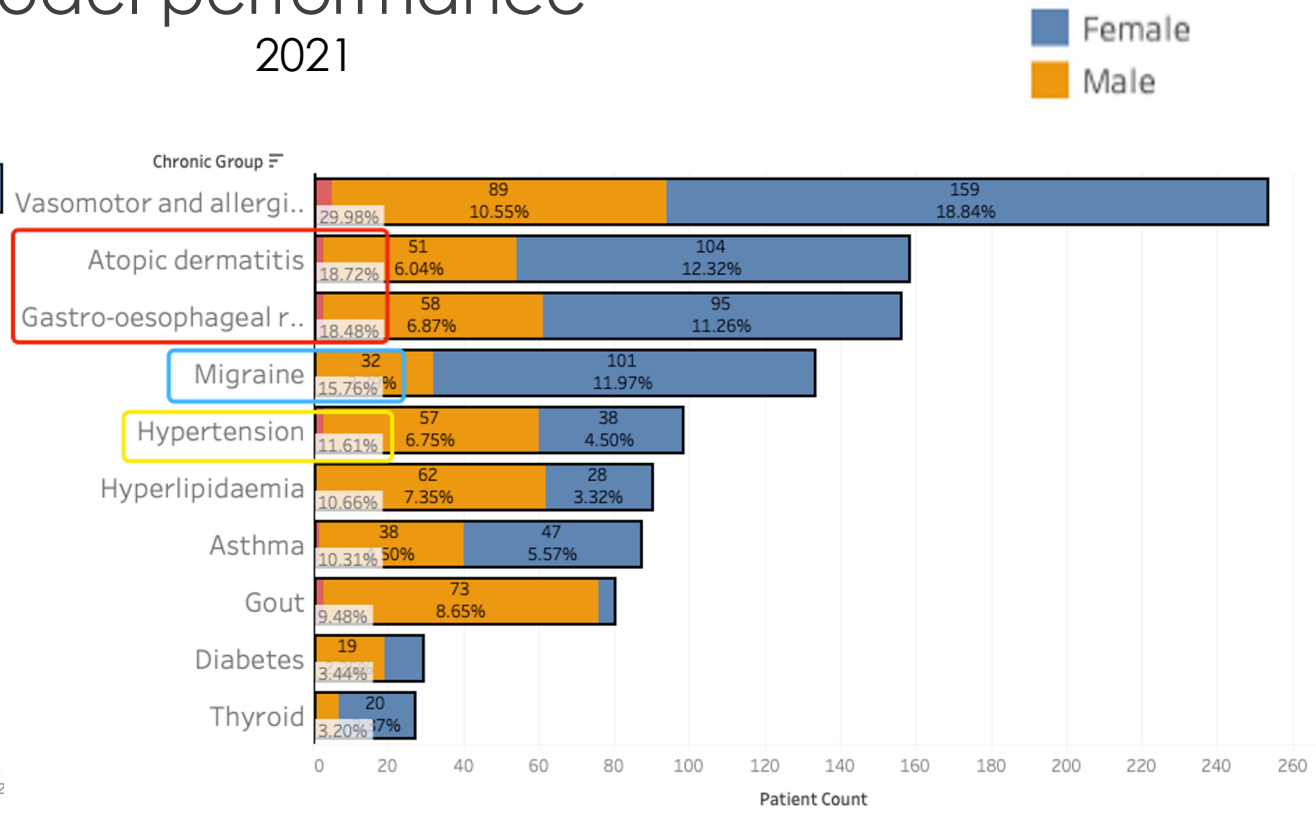
More patients with hypertension and less patients with atopic dermatitis, gastro-oesophageal reflux disease & migraine compared to 2021 data

Data Drift over time -> impact model performance

Jan 2022



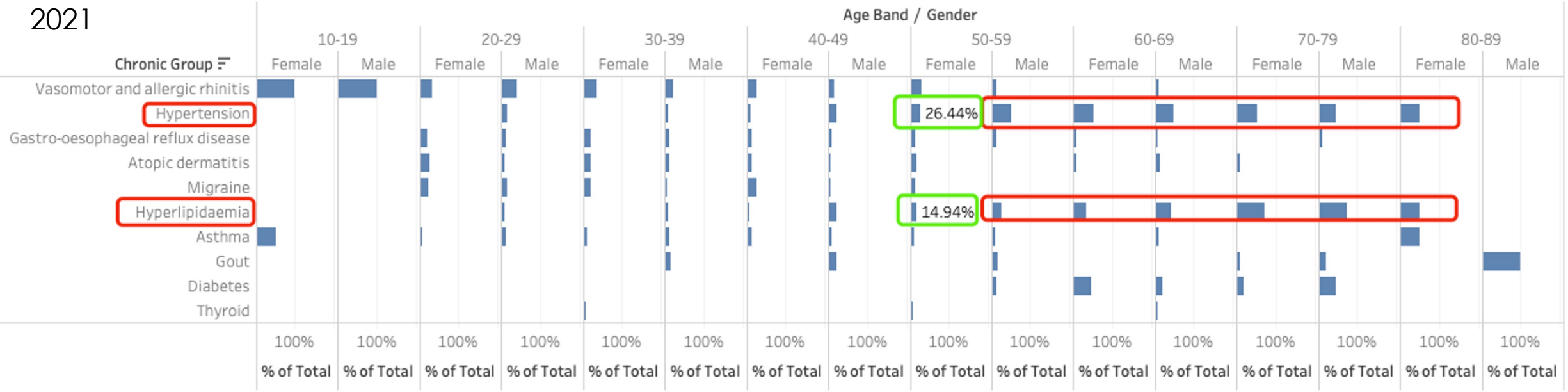
2021



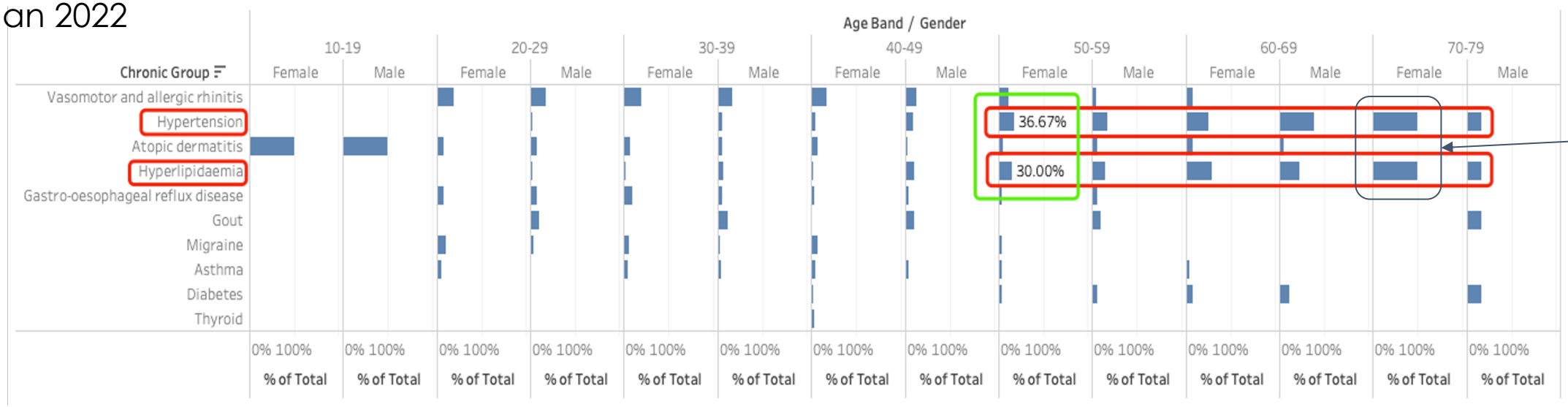
Chronic Group Distribution

Data drift over time

2021



Jan 2022



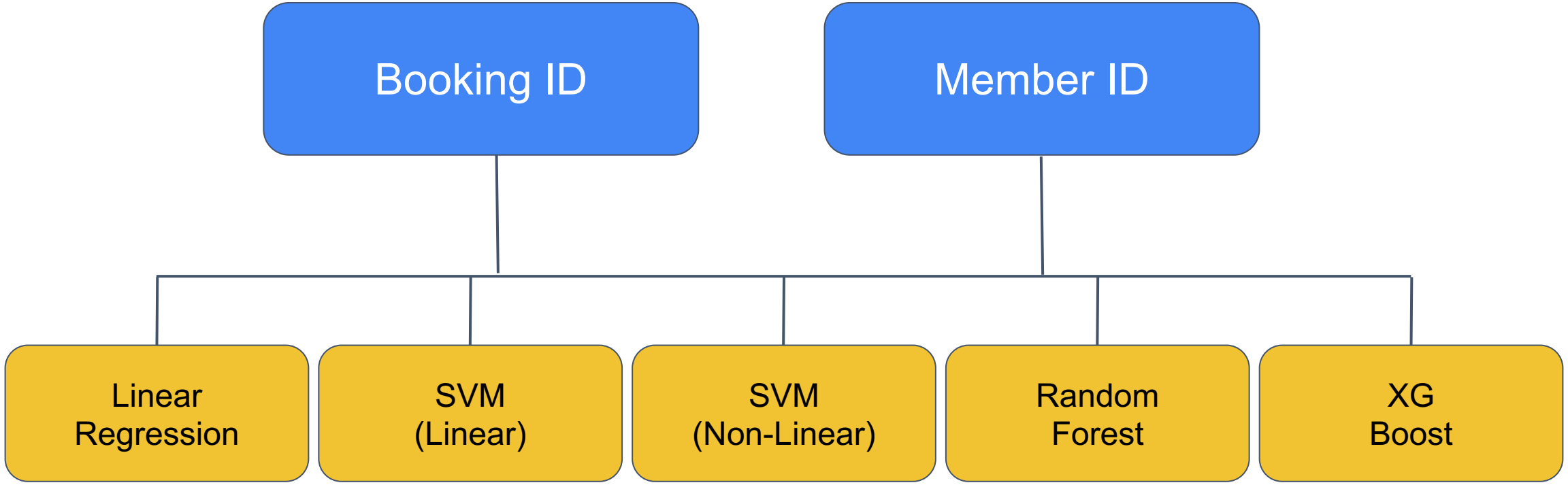
Due to 1 female patient having both hypertension & hyperlipidaemia

5. Modeling Results

a. By Booking ID
b. By Members ID
c. By Lab Report

© 2022 National University of Singapore. All Rights Reserved

Multiple Base Models



Initial Target Labels

S/No	Chronic Disease	Layman Term / Also Known As	All Chronics
1	**Hypertension**	High Blood Pressure	✓
2	**Hyperlipidaemia**	High Cholesterol	✓
3	**Diabetes**	High Blood Sugar	✓
4	**Atopic dermatitis**	Eczema	✓
5	**Asthma**	Lung Disorder	✓
6	**Vasomotor and allergic rhinitis**	Allergy	✓
7	**Migraine**	Severe Headache	✓
8	**Gastro-esophageal reflux disease**	Acid Reflux / Heartburn	✓
9	**Gout**	Arthritis	✓
10	**Thyroid**	Thyroid	✓

Base Model Target Labels Attempted:

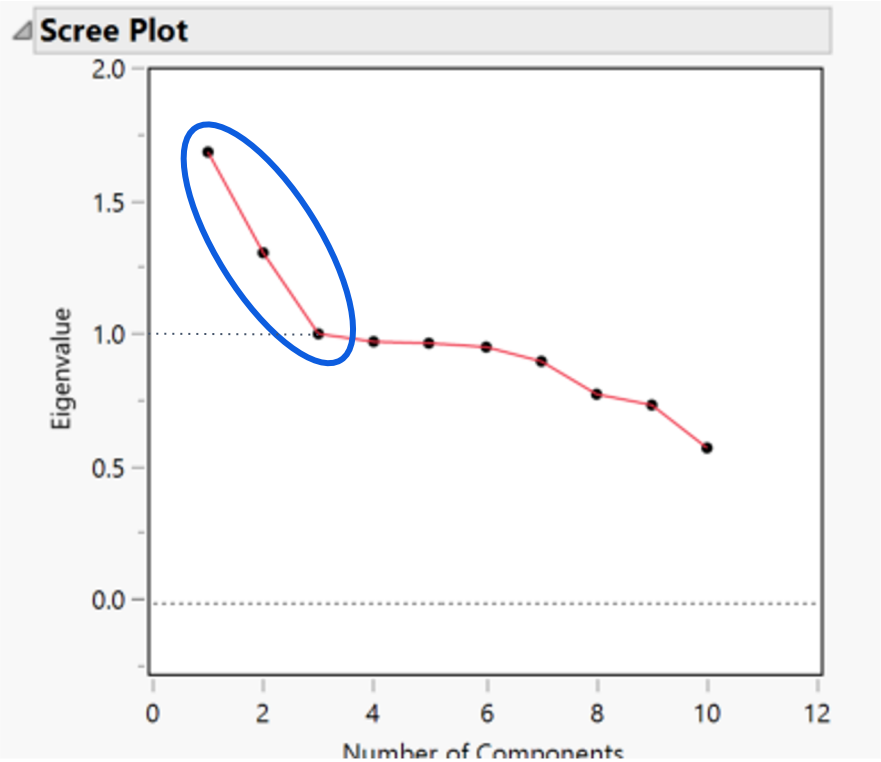
- All Chronics
- ~~Hypertension~~
- ~~Hyperlipidaemia~~
- ~~Diabetes~~
- ~~Atopic dermatitis~~
- ~~Asthma~~
- ~~Allergy~~
- ~~Migraine~~
- ~~Gastro~~
- ~~Gout~~
- ~~Thyroid~~

Only the All Chronic showed promising result, while individual chronic disease as target label failed to do so.

Researched further to see if the chronic diseases could be grouped together.

Grouping Targets on PCA

- From the Scree Plot, we can see that 3 PC are above eigenvalues 1.
- The rotated loading of the 3 PC suggests that PC1(Factor 1) consist of Hypertension, Hyperlidaemia and Diabetes.
- PC2(Factor 2) consist of Dermatitis, Allergic and Asthma.
- The rest can be grouped as the PC3(Factor 3).



Rotated Factor Loading

	Factor 1	Factor 2	Factor 3
Target_Dermatitis	-0.030308	0.519810	-0.115518
Target_Migraine	0.002105	0.128865	0.457773
Target_Allergic	-0.018755	0.594531	0.339253
Target_Gastro	0.100874	0.046046	0.717289
Target_Hypertension	0.748029	0.078915	-0.035530
Target_Asthma	-0.010802	0.701398	-0.075644
Target_Hyperlipidaemia	0.777413	0.075914	-0.008150
Target_Thyroid	0.145787	0.302099	0.141424
Target_Diabetes	0.671094	-0.081840	0.011843
Target_Gout	0.149066	0.203463	-0.447603

© 2022 National University of Singapore. All Rights Reserved

Final Target Groupings

S/No	Chronic Disease	Layman Term / Also Known As	All Chronics	3 High (Group 1)	Atopic Triad (Group 2)	Others (Group 3)
1	**Hypertension**	High Blood Pressure	✓	✓		
2	**Hyperlipidaemia**	High Cholesterol	✓	✓		
3	**Diabetes**	High Blood Sugar	✓	✓		
4	**Atopic dermatitis**	Eczema	✓		✓	
5	**Asthma**	Lung Disorder	✓		✓	
6	**Vasomotor and allergic rhinitis**	Allergy	✓		✓	
7	**Migraine**	Severe Headache	✓			✓
8	**Gastro-esophageal reflux disease**	Acid Reflux / Heartburn	✓			✓
9	**Gout**	Arthritis	✓			✓
10	**Thyroid**	Thyroid	✓			✓

Result Emphasis

Client X's focus are on All Chronics and 3 High. The other 2 groups are good to have but not their main interest.

All Chronics
3 High (Group 1)
Atopic Triad (Group 2)
Others (Group 3)

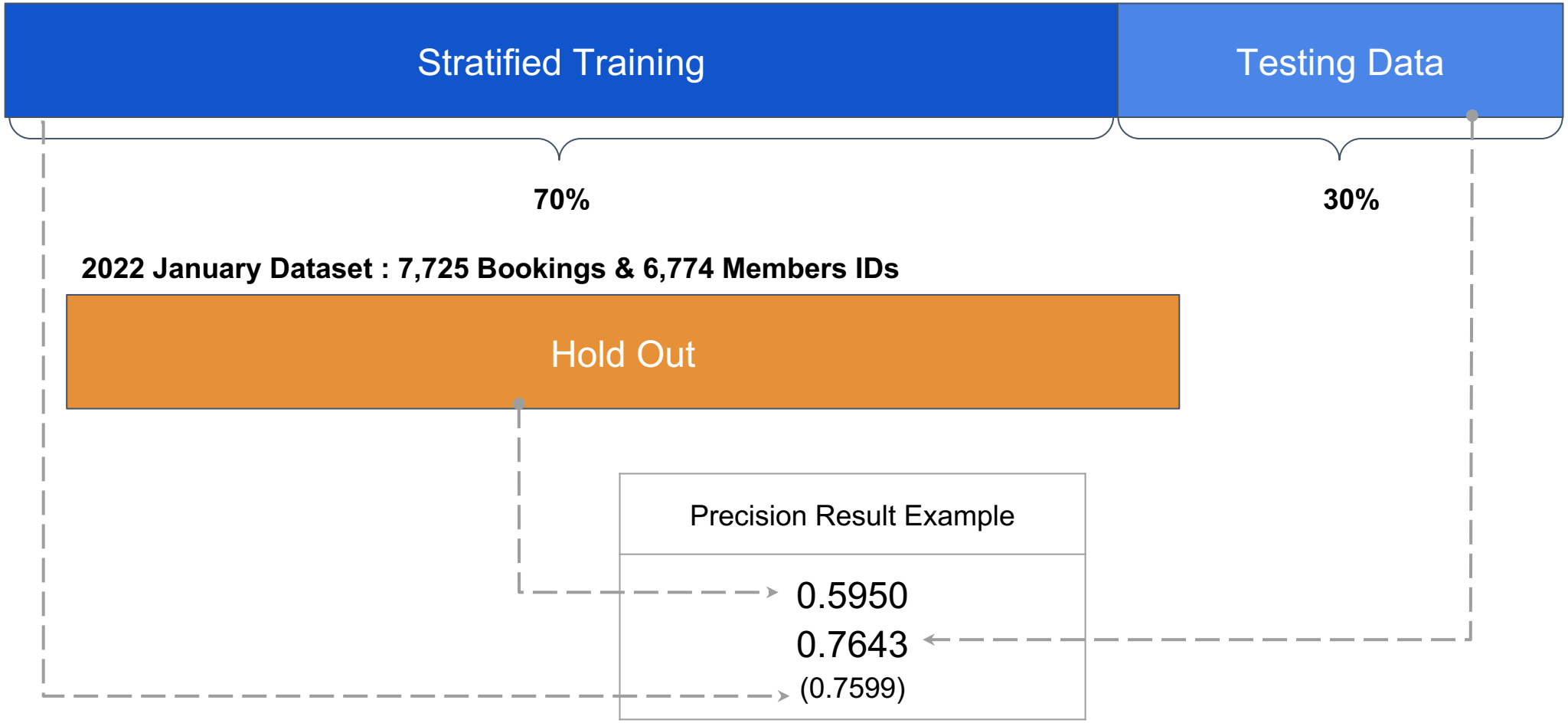
In detecting anomalies, Recall is often used to measure the model's ability to capture the true positive rate. However, Client X stated Precision as their preferred rate to look at.

Our model recommendations weigh on Precision but counter check Recall to ensure model viability.

Precision
Recall
F1-Score

Training, Testing & Hold Out

2021 Sample Dataset : 10,221 Bookings & 5,000 Member IDs



a. Booking ID Model Results

© 2022 National University of Singapore. All Rights Reserved

Booking Baseline - All Chronic :

Legend	Support	Feature
Holdout	697	42
Test	446	
(Train)	1041	

Target Class	Precision	Recall	F1-score	AUC
Logistic	0.5950 0.7643 (0.7599)	0.4448 0.4798 (0.4986)	0.5090 0.5895 (0.6021)	0.93
SVM (Linear)	0.5406 0.7550 (0.7379)	0.4491 0.5139 (0.5045)	0.4906 0.6059 (0.6048)	0.922
SVM (Non-Linear)	0.5511 0.7231 (0.7107)	0.4720 0.5269 (0.5687)	0.5085 0.6096 (0.6318)	0.92
Random Forest	0.6780 0.7845 (0.8505)	0.3716 0.4081 (0.4371)	0.4801 0.5369 (0.5774)	0.92
XGBoost	0.6873 0.7833 (0.8115)	0.3816 0.4215 (0.4457)	0.4908 0.5481 (0.5764)	0.93

XGBoost is recommended here for the high Precision, and Recall is fairly close to other models.

Booking Baseline - 3 High :

High Blood Pressure, High Cholesterol, Diabetes

Legend	Support	Feature
Holdout	151	45
Test	108	
(Train)	252	

Target Class	Precision	Recall	F1-score	AUC
Logistic	0.7347 0.7294 (0.8021)	0.4768 0.5741 (0.5952)	0.5783 0.6425 (0.6834)	0.97
SVM (Linear)	0.6930 0.6939 (0.8000)	0.5232 0.6296 (0.6190)	0.5962 0.6602 (0.6980)	0.97
SVM (Non-Linear)	0.8136 0.7778 (0.8671)	0.3179 0.4537 (0.4921)	0.4571 0.5731 (0.6278)	0.97
Random Forest	0.8140 0.8000 (0.9483)	0.2318 0.3704 (0.4365)	0.3608 0.5063 (0.5978)	0.96
XGBoost	0.7308 0.7500 (0.9222)	0.3775 0.4722 (0.6111)	0.4978 0.5795 (0.7351)	0.96

SVM (Linear) is recommended here as the Recall is the highest amongst all the model. While Random Forest has the highest Precision, the Recall is low at 23%. Logistic can also be considered as an alternative model with slightly better precision and Recall just a notch below SVM (Linear).

Booking Baseline - Atopic Triad :

Eczema, Asthma, Allergy

Legend	Support	Feature
Holdout	337	46
Test	198	
(Train)	462	

Target Class	Precision	Recall	F1-score	AUC
Logistic	0.5108 0.5918 (0.6620)	0.2819 0.2929 (0.3052)	0.3633 0.3919 (0.4178)	0.92
SVM (Linear)	0.4656 0.6082 (0.6587)	0.2611 0.2980 (0.2965)	0.3346 0.4000 (0.4090)	0.92
SVM (Non-Linear)	0.5564 0.6143 (0.7188)	0.2196 0.2172 (0.2489)	0.3149 0.3209 (0.3698)	0.92
Random Forest	0.6923 0.8667 (0.9600)	0.0534 0.0657 (0.1558)	0.0992 0.1221 (0.2682)	0.91
XGBoost	0.5794 0.6721 (0.8074)	0.1840 0.2071 (0.2359)	0.2793 0.3166 (0.3652)	0.93

Logistic is recommended with the highest Recall, while the Precision is still reasonable. Random Forest here has very high Precision but the Recall is unsuitable for consideration.

Booking Baseline - Others :

Migraine, Gastro, Gout, Thyroid

Legend	Support	Feature
Holdout	229	46
Test	166	
(Train)	387	

Target Class	Precision	Recall	F1-score	AUC
Logistic	0.2727 0.4375 (0.5065)	0.1179 0.1687 (0.2016)	0.1646 0.2435 (0.2884)	0.93
SVM (Linear)	0.2281 0.4333 (0.5422)	0.0568 0.0783 (0.1163)	0.0909 0.1327 (0.1915)	0.92
SVM (Non-Linear)	0.1446 0.2051 (0.2903)	0.5284 0.5361 (0.7080)	0.2270 0.2967 (0.4117)	0.92
Random Forest	0.0000 0.0000 (0.0000)	0.0000 0.0000 (0.0000)	0.0000 0.0000 (0.0000)	0.92
XGBoost	0.0000 0.5000 (0.7500)	0.0000 0.0060 (0.0233)	0.0000 0.0119 (0.0451)	0.94

Overall performance for all models were not suitable for recommendation. The results were inconsistent and unstable.

b. Member ID Model Results

Member Baseline - All Chronic :

Legend	Support	Feature
Holdout	776	34
Test	253	
(Train)	591	

Target Class	Precision	Recall	F1-score	AUC
Logistic	0.5580 0.7338 (0.7389)	0.2977 0.4466 (0.4213)	0.3882 0.5553 (0.5366)	0.85
SVM (Linear)	0.6122 0.7442 (0.7557)	0.2461 0.3794 (0.3350)	0.3511 0.5026 (0.4642)	0.82
SVM (Non-Linear)	0.5708 0.7692 (0.8940)	0.1714 0.2372 (0.4281)	0.2636 0.3625 (0.5789)	0.82
Random Forest	0.7192 0.8056 (0.9492)	0.1353 0.2292 (0.2842)	0.2278 0.3569 (0.4375)	0.81
XGBoost	0.5764 0.7483 (0.8476)	0.2577 0.4466 (0.4704)	0.3562 0.5594 (0.6050)	0.86

Random Forest Model has better overall results.
Best for hold out dataset precision **(72%)** and one of the best for modeling test precision **(81%)**.
Recall is low overall for all.

Member Baseline - 3 High :

High Blood Pressure, High Cholesterol, Diabetes

Legend	Support	Feature
Holdout	156	25
Test	46	
(Train)	109	

Target Class	Precision	Recall	F1-score	AUC
Logistic	0.4565 0.6000 (0.7119)	0.2692 0.3261 (0.3853)	0.3387 0.4225 (0.5000)	0.91
SVM (Linear)	0.5373 0.5238 (0.8182)	0.2308 0.2391 (0.3303)	0.3229 0.3284 (0.4706)	0.89
SVM (Non-Linear)	0.2529 0.6316 (0.9710)	0.1410 0.2609 (0.6147)	0.1811 0.3692 (0.7528)	0.86
Random Forest	0.7857 0.8750 (0.9310)	0.0705 0.1522 (0.2477)	0.1294 0.2593 (0.3913)	0.92
XGBoost	0.7692 0.6471 (0.8363)	0.1923 0.2391 (0.4220)	0.3077 0.3492 (0.5610)	0.93

Extreme Gradient Boost Model has better overall results. Good result for hold out dataset precision (77%) and modeling test precision (64%). Recall is low overall. Random Forest has better precision, but XGB has better balance between Precision and Recall.

Member Baseline - Atopic Triad :

Eczema, Asthma, Allergy

Legend	Support	Feature
Holdout	391	49
Test	130	
(Train)	304	

Target Class	Precision	Recall	F1-score	AUC
Logistic	0.4568 0.5286 (0.6212)	0.1893 0.2846 (0.2697)	0.2676 0.3700 (0.3761)	0.85
SVM (Linear)	0.5833 0.6111 (0.8929)	0.0358 0.0846 (0.0822)	0.0675 0.1486 (0.1506)	0.83
SVM (Non-Linear)	0.5952 0.7333 (0.9803)	0.064 0.085 (0.3289)	0.1155 0.1517 (0.4926)	0.80
Random Forest	0.6207 0.7333 (0.9574)	0.0460 0.0846 (0.1480)	0.0857 0.1517 (0.2564)	0.85
XGBoost	0.4872 0.6000 (0.8471)	0.1458 0.2077 (0.2368)	0.2244 0.3086 (0.3702)	0.87

Extreme Gradient Boost Model has relatively slightly better overall results. (Followed by Logistic Regression)
Relatively acceptable result for hold out dataset precision (49%) and modeling test precision (60%). The other ensemble model Random Forest has better holdout precision **(62%)**, but holdout recall is too low for Random Forest.
Given the low precision and recall, recommendation is to model the Chronic in this group individually when given more data.

Member Baseline - Others :

Migraine, Gastro, Gout, Thyroid

Legend	Support	Feature
Holdout	285	40
Test	113	
(Train)	263	

Target Class	Precision	Recall	F1-score	AUC
Logistic	0.0754 0.5581 (0.6352)	0.5333 0.2124 (0.2053)	0.1322 0.3077 (0.3103)	0.69
SVM (Linear)	0.0000 0.0000 (0.0000)	0.0000 0.0000 (0.0000)	0.0000 0.0000 (0.0000)	0.68
SVM (Non-Linear)	0.0329 0.3265 (0.9937)	0.1474 0.1416 (0.6008)	0.0538 0.1975 (0.7488)	0.51
Random Forest	0.2727 0.6000 (1.000)	0.1580 0.0265 (0.0989)	0.2000 0.0508 (0.1799)	0.68
XGBoost	0.1112 0.6667 (0.7619)	0.3930 0.0354 (0.0608)	0.1734 0.0672 (0.1127)	0.73

The results under this modeling sub-group are overall **not consistent and not acceptable**.
Given the bad precision and recall result, the results under this group are all rejected.
Recommendation is to model the Chronic in this group individually when given more data.

c. Lab Report Results

Dataset by Lab Report Number

Booking Symptoms

Column	Explanation	Datatype
booking_id	Unique consult identifier	String
MDSymptom_id	Unique symptom identifier	String
MDSymptom_name	Current symptom that patient experienced	String

Booking Diagnosis

Column Name	Explanation	Datatype
booking_id	Unique consult identifier	String
MDDiagnosis_code	ICD-10 code of diagnosed illness	String
MDDiagnosis_name	Name of illness as per ICD-10	String
chronic_group	WhiteCoat's classification of chronic disease(s) to model for	String

Booking Prescription

Column Name	Explanation	Datatype
booking_id	Unique consult identifier	String
MDMedication_ID	Unique medication identifier	String
MDMedication_name	Name of medication prescribed	String
is_fulfilled_externally	Whether medication prescribed was available in WhiteCoat's inventory during that consult	Boolean
prescription_quantity	Number of medication prescribed	Integer
prescription_status	Whether patient decided to purchase the given medication or not	String

Patient

Column	Explanation	Datatype
mem_id	Unique patient identifier	String
gender	Patient gender	String
birth_year	Year of birth for each patient	Integer

Lab Reports

Column Name	Explanation	Datatype
mem_id	Unique patient identifier	String
lab_report_no	Unique lab report identifier	Integer
requested_date	Date when the lab test was ordered	Date
category	Category of lab test	String
line_item	Specific test line item within category	String
test_result	Result of the given line item	Various
unit	Unit for the given line item	String

Booking "Base"

Column	Explanation	Datatype
booking_id	Unique consult identifier	String
mem_id	Unique patient identifier	String
doctor_id	Unique doctor identifier	String
consult_waived_reason	Reason for consult fee being waived, if so. Dash = consult fee was not waived.	String
is_mc_issued	Whether MC was issued for that consult	Boolean
mc_days	Number of MC days issued, if so	Integer
is_referral_letter_issued	Whether a referral letter was issued for that consult	Boolean
booking_timestamp	When the consult occurred in Singapore Time (GMT+8)	Datetime

Phase 1: Completed Final Lab Dataset

Note: only 6 CAT columns are illustrated here*

CAT|LIPID PROFILE|HDL Cholesterol|mg/dL	CAT|LIPID PROFILE|LDL Cholesterol|mg/dL	CAT|LIPID PROFILE|Total Cholesterol|mg/dL	CAT|LIPID PROFILE|Triglycerides|mg/dL	CAT|HAEMATOLOGY|Basophils|%	CAT|HAEMATOLOGY|Basopl
desirable	desirable	desirable	desirable	not tested	not tested
high	desirable	desirable	desirable	low	normal
borderline high	desirable	desirable	desirable	normal	normal
desirable	desirable	borderline high	desirable	low	normal
desirable	desirable	desirable	desirable	normal	normal
desirable	high	high	desirable	low	normal
desirable	desirable	desirable	desirable	not tested	not tested
desirable	high	high	desirable	not tested	not tested
desirable	desirable	desirable	desirable	not tested	not tested
desirable	desirable	desirable	desirable	normal	normal
desirable	desirable	desirable	desirable	normal	normal
desirable	desirable	borderline high	desirable	high	normal
desirable	desirable	desirable	desirable	low	normal
borderline high	desirable	borderline high	borderline high	normal	normal
desirable	desirable	desirable	desirable	low	normal
borderline high	borderline high	borderline high	desirable	low	normal

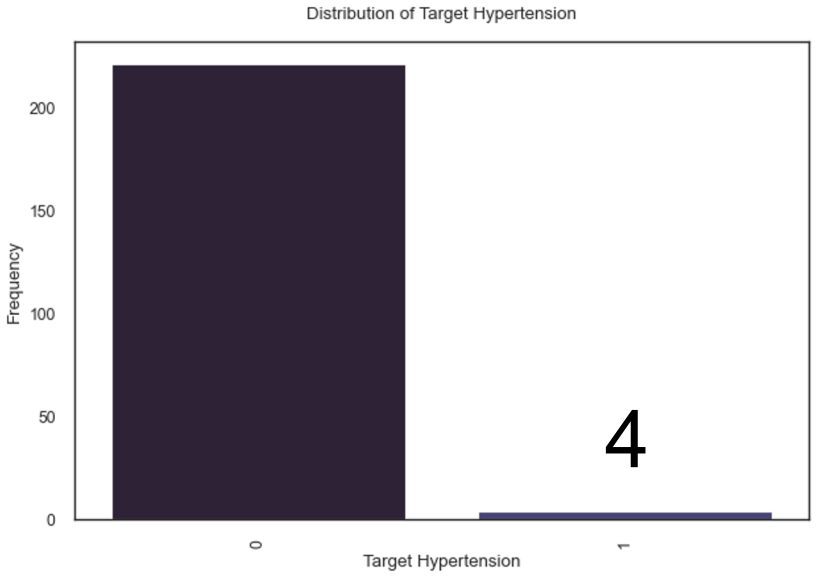
Notes:

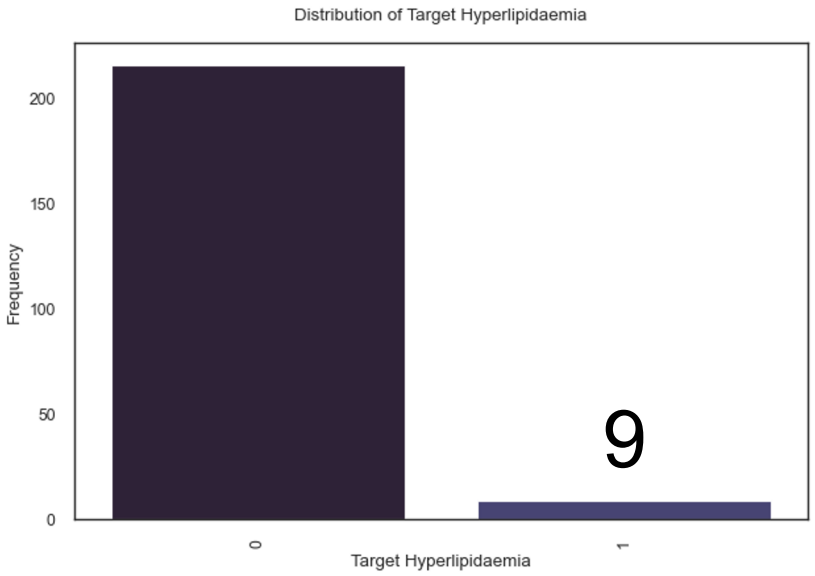
1. Total number of CATegorical columns created in dataset: 43

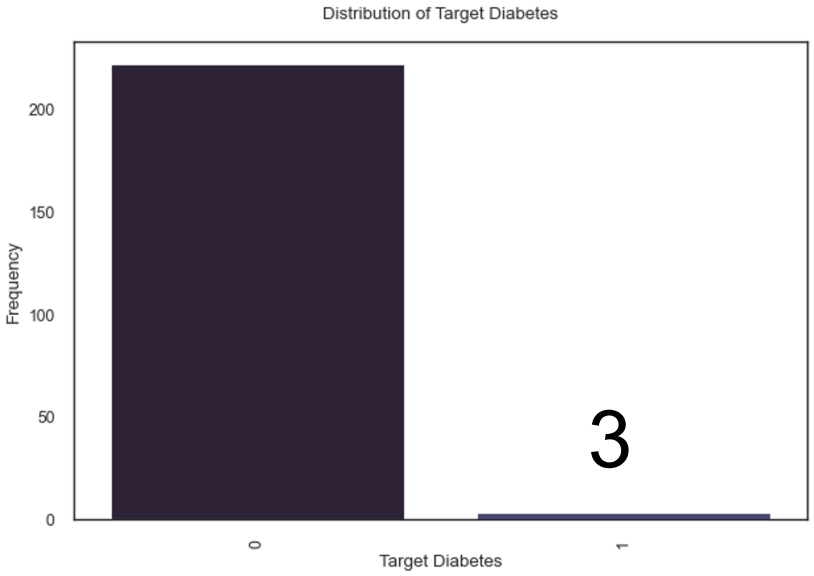
© 2022 National University of Singapore. All Rights Reserved

Dataset with lab_report_no (Chronic cases)

Total number of reports : **225**







Hypertension cases : 1.7%

Hyperlipidaemia cases : 4%

Diabetes cases : 1.3%

Observations: Imbalance data for all top 3 chronic cases in the Lab Reports

© 2022 National University of Singapore. All Rights Reserved

Target_Hypertension (4 cases)

	52	55	82	223
gender	Female	Male	Male	Male
CAT|LIPID PROFILE|HDL Cholesterol|mg/dL	desirable	borderline high	high	high
CAT|LIPID PROFILE|LDL Cholesterol|mg/dL	desirable	desirable	desirable	desirable
CAT|LIPID PROFILE|Total Cholesterol|mg/dL	desirable	desirable	desirable	desirable
CAT|LIPID PROFILE|Triglycerides|mg/dL	desirable	desirable	desirable	borderline high
CAT|HAEMATOLOGY|Basophils|%	low	not tested	not tested	not tested
CAT|HAEMATOLOGY|Basophils|x109/L	normal	not tested	not tested	not tested
CAT|HAEMATOLOGY|Eosinophils|%	normal	not tested	not tested	not tested
CAT|HAEMATOLOGY|Eosinophils|x109/L	normal	not tested	not tested	not tested
CAT|HAEMATOLOGY|Haematocrit (PCV)|nan	normal	not tested	not tested	not tested
CAT|HAEMATOLOGY|Haemoglobin|g/dL	normal	not tested	not tested	not tested
CAT|HAEMATOLOGY|Lymphocytes|%	high	not tested	not tested	not tested
CAT|HAEMATOLOGY|Lymphocytes|x109/L	normal	not tested	not tested	not tested
CAT|HAEMATOLOGY|MCH|pg	normal	not tested	not tested	not tested
CAT|HAEMATOLOGY|MCHC|g/dL	normal	not tested	not tested	not tested
CAT|HAEMATOLOGY|MCV|fL	normal	not tested	not tested	not tested
CAT|HAEMATOLOGY|Monocytes|%	normal	not tested	not tested	not tested
CAT|HAEMATOLOGY|Monocytes|x109/L	normal	not tested	not tested	not tested
CAT|HAEMATOLOGY|Neutrophils|%	low	not tested	not tested	not tested
CAT|HAEMATOLOGY|Neutrophils|x109/L	low	not tested	not tested	not tested
CAT|HAEMATOLOGY|Peripheral Blood Film|nan	Red cells appear normochromic and normocytic.	not tested	not tested	not tested
CAT|HAEMATOLOGY|Platelets|x109/L	normal	not tested	not tested	not tested
CAT|HAEMATOLOGY|RDW|%	low	not tested	not tested	not tested
CAT|HAEMATOLOGY|Red cell count|x1012/L	normal	not tested	not tested	not tested
CAT|HAEMATOLOGY|Total White cell count|x109/L	low	not tested	not tested	not tested
CAT|URINALYSIS|Urine Clarity|nan	not tested	not tested	Clear	Clear
CAT|URINALYSIS|Urine Colour|nan	not tested	not tested	Pale Yellow	Pale Yellow
CAT|THYROID FUNCTION|TSH|mIU/L	normal	not tested	not tested	not tested

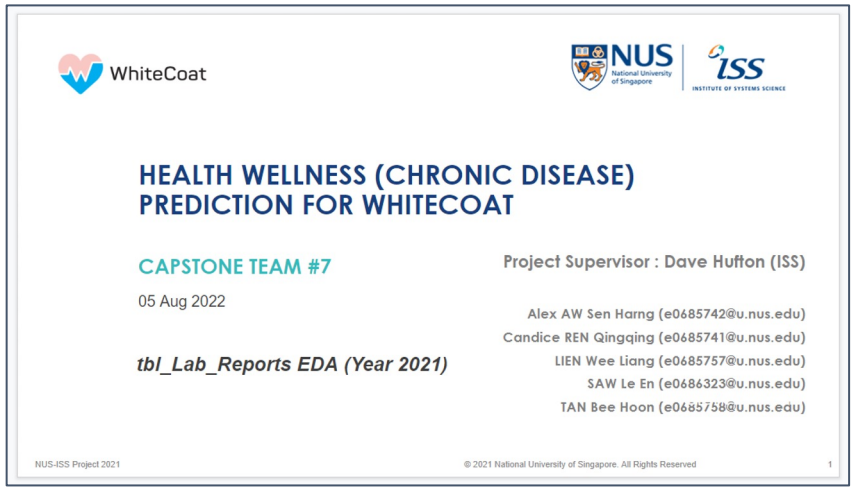
Observations:

1. A number of members did not take the required tests to sieve out their respective chronic cases.

1. Some tests shows normal healthy test results, sometimes due to tests made in earlier visits.

1. Similar observations was made for *Target_Hyperlipidaemia* and *Target_Diabetes* cases.

Lab Test Results (EDA) for Doctors

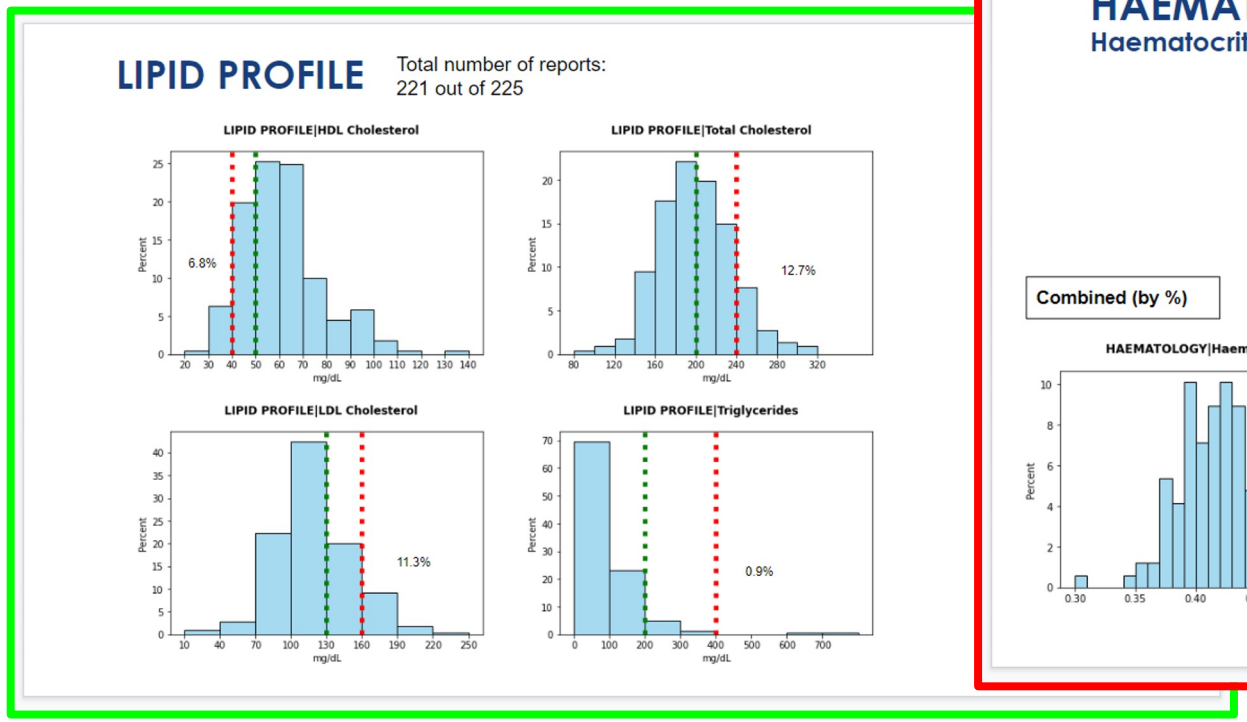


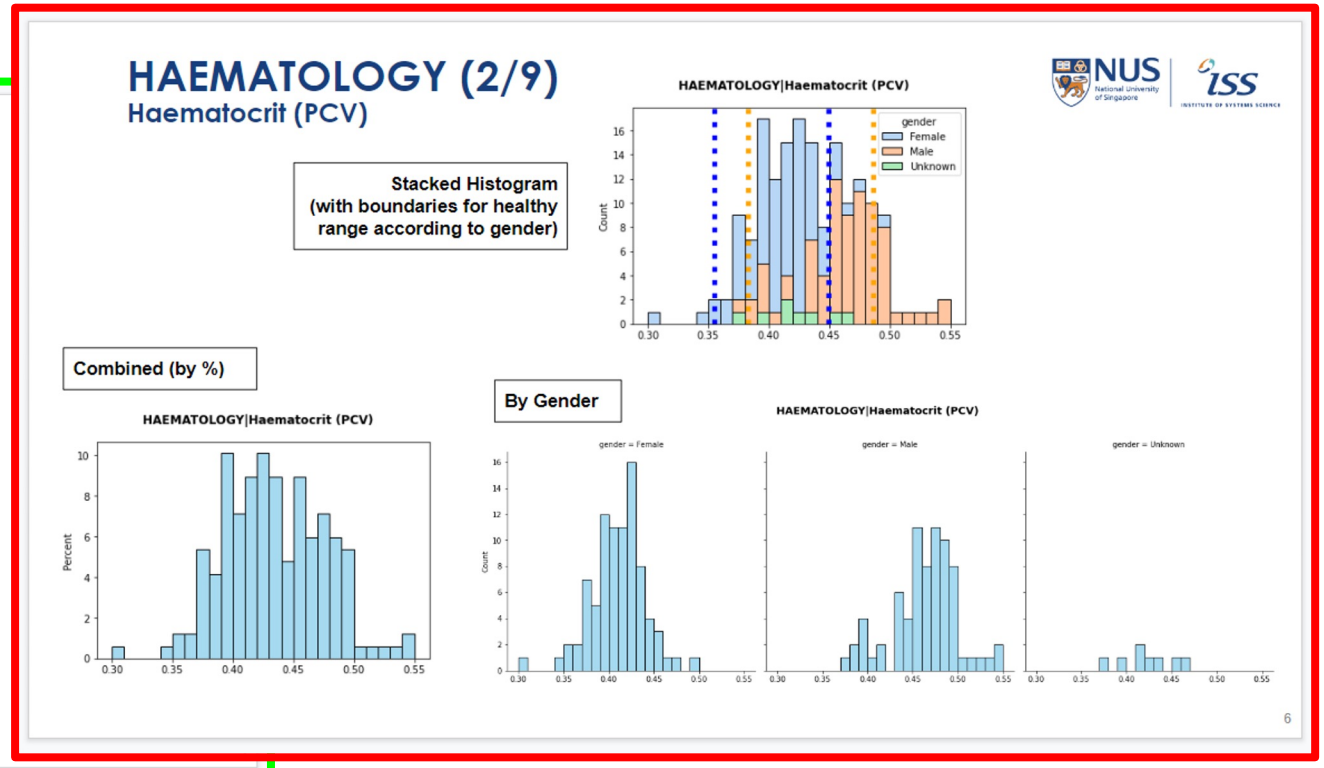
EDA LAB TEST RESULTS

1. Lipid Profile
2. Haematology
3. Diabetes Mellitus Profile
4. Urinalysis
5. Urine Chemistry
6. Thyroid Function
7. Kidney Profile
8. Serum Creatinine

Project_Capstone_tbl_Lab_reports_EDA_2021.pptx

Lab Test results are converted into an EDA document for doctors' reference





Merging Lab Results to Members' dataset

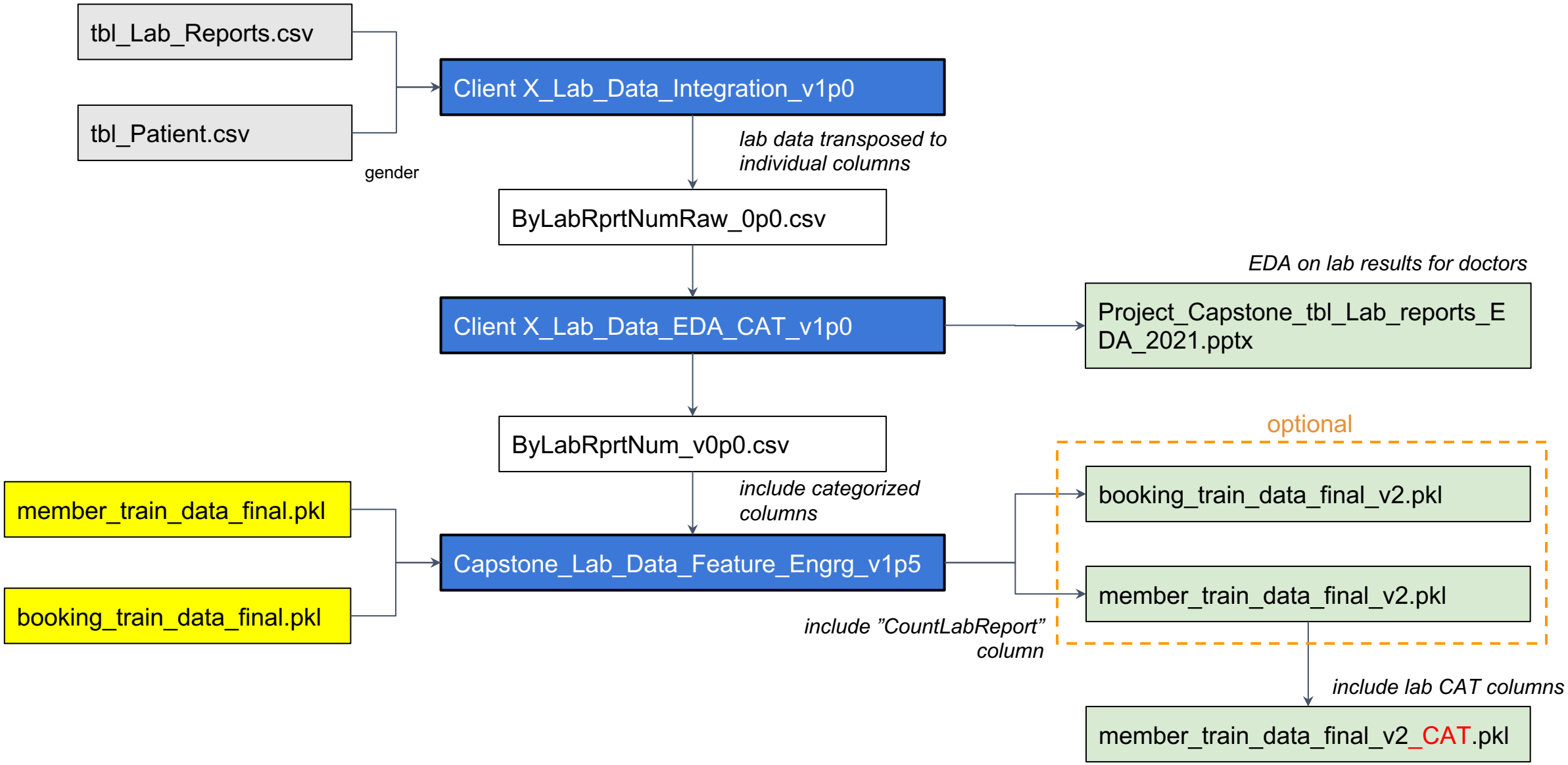
member_train_data_final_v2_CAT.pkl

Lab Test results combined with members' database for doctors' reference

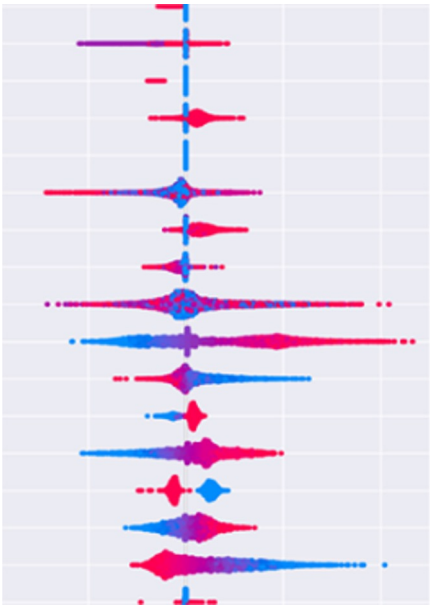
mem_id	gender	CountLabReport	CAT|LIPID PROFI	CAT|LIPID PROF	CAT|LIPID PROF	CAT|LIPID PRO	CAT|HAEMATOLOG	CAT|HAEMA	CAT|HAEM
03A57929-EED6-4439-	Female	2	desirable	borderline high	high	desirable	low	normal	normal
45BC9539-D233-48C3-	Female	2	desirable	desirable	desirable	desirable	normal	normal	normal
7D83353F-5138-4CC4-	Female	2	desirable	desirable	borderline high	desirable	normal	normal	normal
DDE51D88-C8C1-45B0	Female	2	desirable	desirable	desirable	desirable	normal	normal	normal
0004A803-9EAB-4E3E-	Female	1	desirable	desirable	desirable	desirable	low	normal	normal
0199B08C-5213-40F2-	Female	1	desirable	borderline high	borderline high	desirable	normal	normal	normal
052E1B67-A42F-4AB2-	Unknown	1	desirable	high	high	desirable	normal	normal	normal
056EEBE5-076B-40DD-	Female	1	high	high	high	borderline high			
058C24BA-1F7D-46DC-	Female	1	desirable	desirable	borderline high	desirable	normal	normal	normal
06E20473-6833-41B9-	Male	1	desirable	borderline high	borderline high	desirable	normal	normal	normal
09A6F07C-CE7F-4516-	Female	1	desirable	desirable	desirable	desirable	normal	normal	normal
0A38767E-F9D4-4CAB-	Female	1	desirable	high	high	desirable	high	normal	normal
0C231304-1F89-40E6-	Unknown	1	desirable	borderline high	borderline high	desirable	low	normal	normal
0F4CADAD-DFC0-4273	Female	1	borderline high	desirable	desirable	desirable			
10E502D6-C8C3-4D18-	Female	1	desirable	desirable	borderline high	desirable			
11089833-9780-4D36-	Female	1	desirable	borderline high	borderline high	desirable			
11A28AAA-F7E4-4A03-	Male	1	high	desirable	desirable	desirable	normal	normal	normal
16543599-0181-4BA6-	Female	1	desirable	high	borderline high	desirable			
17899DFC-16A7-4836-	Unknown	1	borderline high	desirable	desirable	desirable	normal	normal	normal
18ADCB76-7F6A-456F-	Male	1	desirable	borderline high	borderline high	desirable	high	normal	high

LabData data flow

Legend: input file | codes | pipeline output | output file



6. Modeling Explainability



© 2022 National University of Singapore. All Rights Reserved

Explainability Models

01	**Global Explainability**	• How each feature contributes to the 'overall/global' prediction • Identify important features for the whole model • Methods used: **SHAP, Global Surrogate Method**
02	**Local Explainability**	• How each feature contributes to an individual prediction • Identify important features for this instance only • Methods used: **SHAP, LIME**

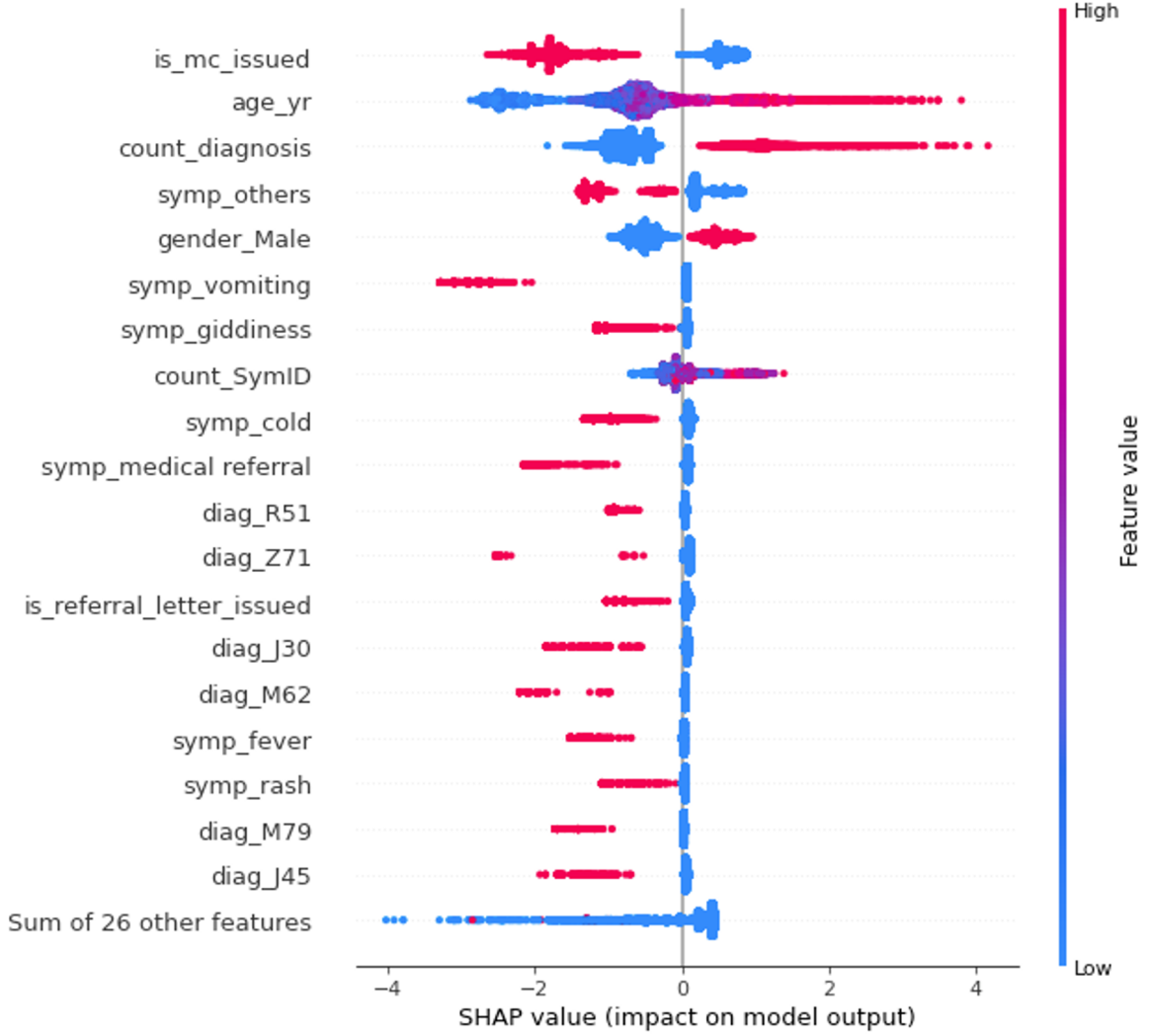
Model by Booking for All Chronic (XGB) Feature Importance - SHAP

For XGBoost model, we selected 45 features out of 154 total features.

The SHAP values indicate the importance of these features.

Being older (Age), have higher count_diagnosis and fewer MC days have positive impact on the prediction.

Being younger (age), have fewer count_diagnosis and longer MC days have negative impact on the prediction.



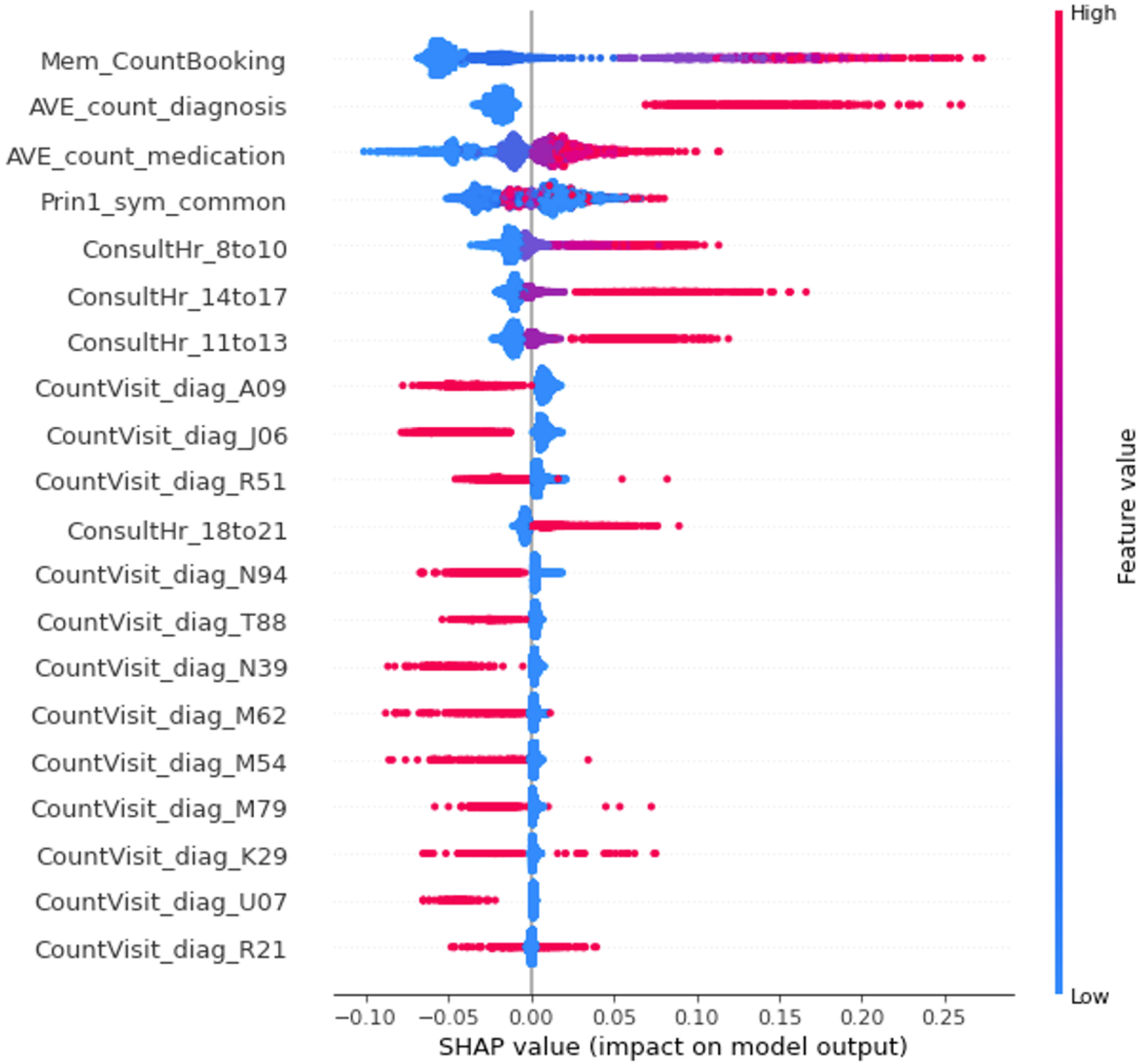
Model by Member for All Chronic (RF) Feature Importance - SHAP

For Random Forest model, we selected 34 features out of 121 total features.

The SHAP values indicate the importance of these features.

Higher Mem_CountBooking & AVE_count_diagnosis have positive impact on the prediction.

Lower AVE_count_medication have negative impact on the prediction.



Model by Member for All Chronic (XGB) Feature Importance - SHAP

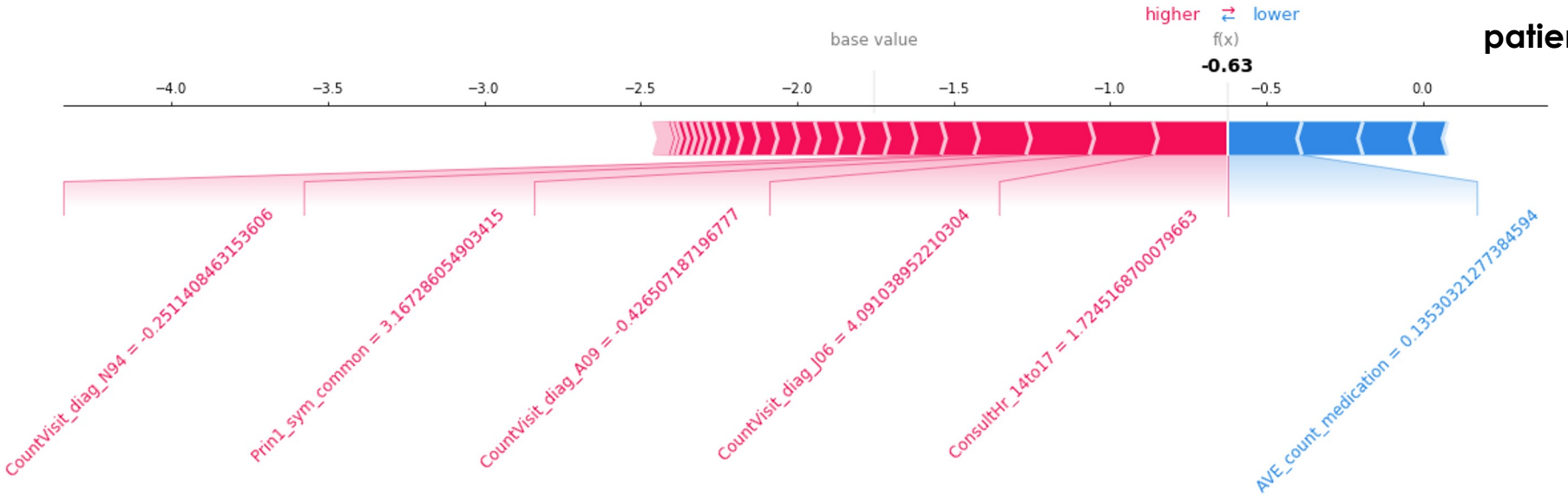
The force plot shows how features contributed to the model's prediction for a specific observation.

Important features for making this prediction are shown below with red features pushing the score higher and blue features pushing the score lower.

The features with higher impact are closer to the red/blue dividing boundary and have bigger bar size.

Model Score = -0.63 (higher than base value) for this patient and likely to lead to Chronic prediction.

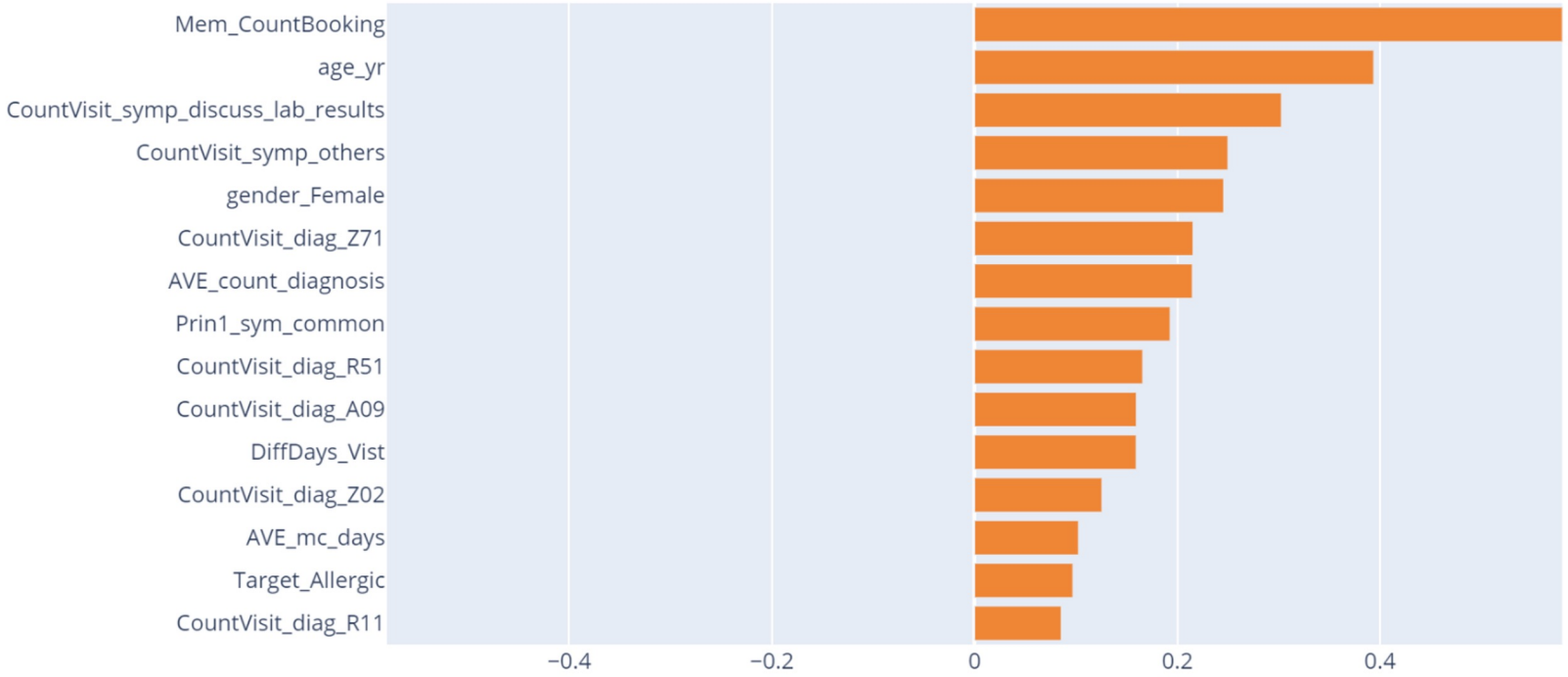
patient_id = 500



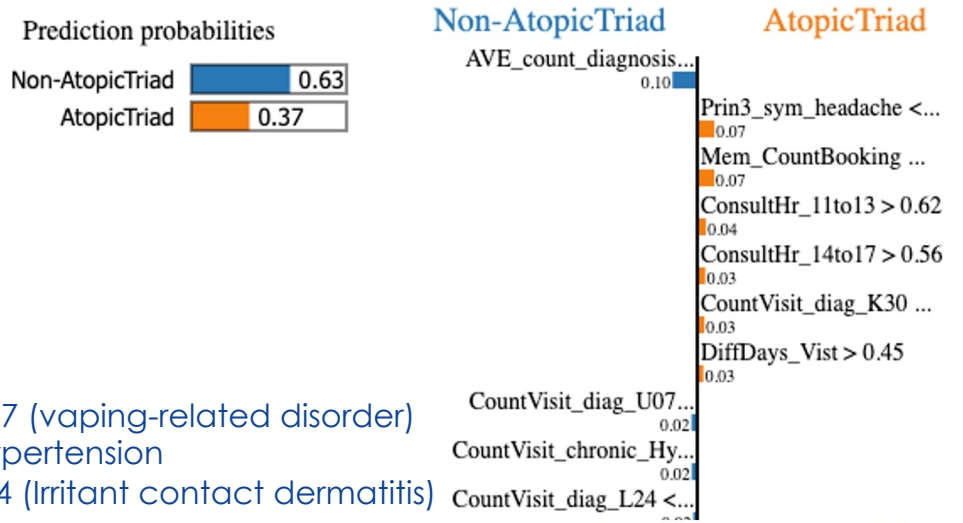
Model by Member for Target 3H (Logistic) Global Surrogate Method (Global Explainability)

Mimic model explainer created for selected logistic model with 25 selected features for global explainability.

Mem_CountBooking was the most important feature followed by age_yr, CountVisit_symp_discuss_lab_results, CountVisit_symp_others, gender_female, CountVist_diag_Z71 (Other counselling & medical advice) and AVE_count_diagnosis.



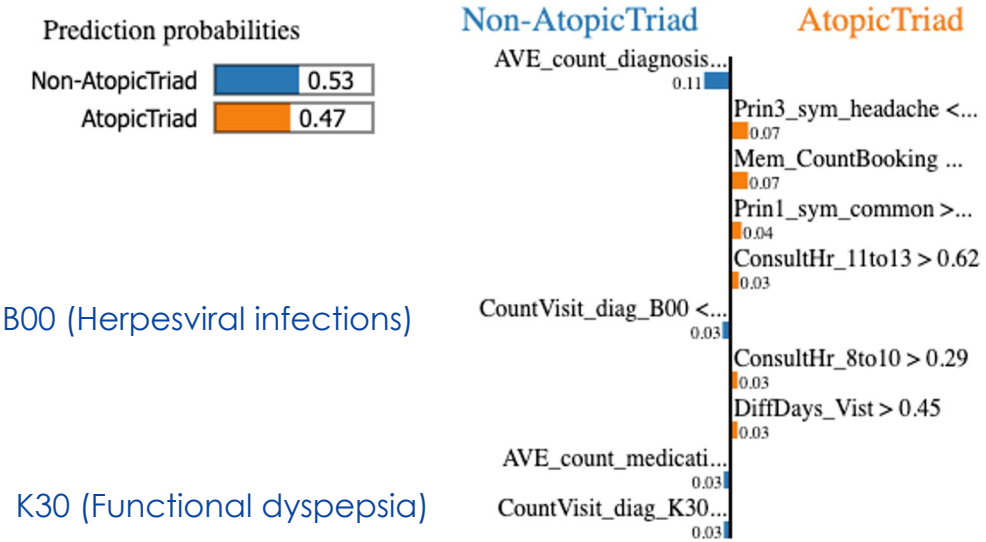
Model by Member for Atopic Triad (RF) LIME (Local Explainability)



U07 (vaping-related disorder)
Hypertension
L24 (Irritant contact dermatitis)

member_id = 34

Feature	Value
AVE_count_diagnosis	-0.34
Prin3_sym_headache	-0.47
Mem_CountBooking	1.13
ConsultHr_11to13	1.85
ConsultHr_14to17	1.73
CountVisit_diag_K30	4.31
DiffDays_Vist	0.95
CountVisit_diag_U07	-0.14
CountVisit_chronic_Hypertension	-0.12
CountVisit_diag_L24	-0.08



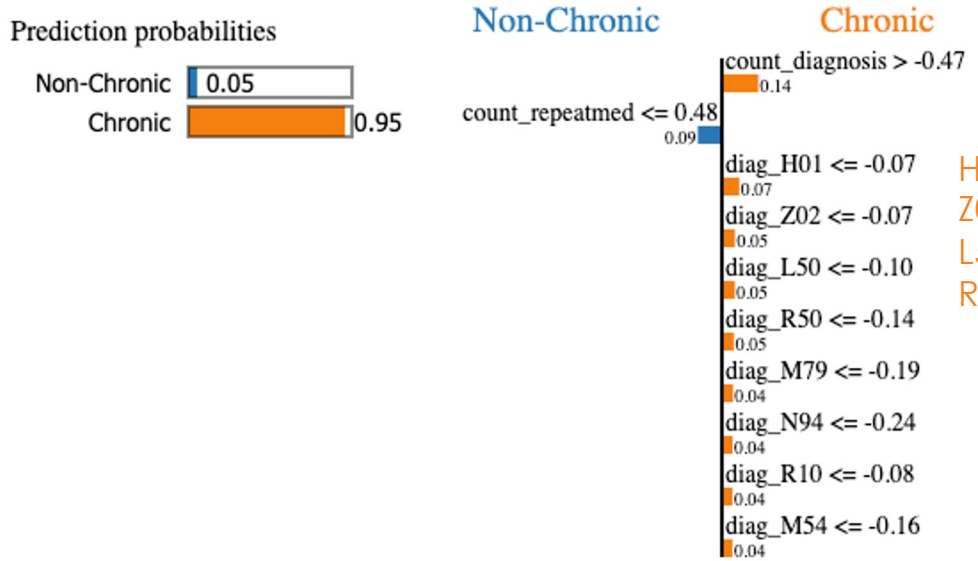
B00 (Herpesviral infections)

K30 (Functional dyspepsia)

member_id = 16

Feature	Value
AVE_count_diagnosis	-0.34
Prin3_sym_headache	-1.33
Mem_CountBooking	2.28
Prin1_sym_common	0.29
ConsultHr_11to13	3.09
CountVisit_diag_B00	-0.07
ConsultHr_8to10	1.24
DiffDays_Vist	2.29
AVE_count_medication	-0.59
CountVisit_diag_K30	-0.16

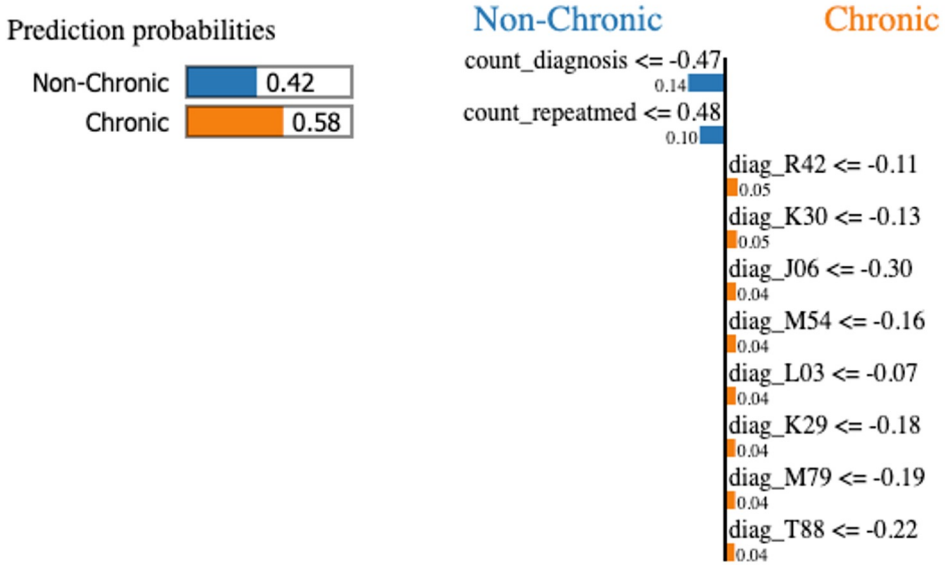
Model by Booking for All Chronic (Logistic) LIME (Local Explainability)



booking_id = 39

H01 (Other inflammation of eyelid)
Z02 (Encounter for administrative exam)
L50 (diseases of skin & subcutaneous tissue)
R50 (Fever of other and unknown origin)

Feature	Value
count_diagnosis	5.15
count_repeatmed	0.48
diag_H01	-0.07
diag_Z02	-0.07
diag_L50	-0.10
diag_R50	-0.14
diag_M79	-0.19
diag_N94	-0.24
diag_R10	-0.08
diag_M54	-0.16



booking_id = 9

R42 (Dizziness and Giddiness)
K30 (Functional dyspepsia)
J06 (Acute upper respiratory infections)
M54 (Low Back Pain)

Feature	Value
count_diagnosis	-0.47
count_repeatmed	0.48
diag_R42	-0.11
diag_K30	-0.13
diag_J06	-0.30
diag_M54	-0.16
diag_L03	-0.07
diag_K29	-0.18
diag_M79	-0.19
diag_T88	-0.22

7. Future Work

© 2022 National University of Singapore. All Rights Reserved

Future Work

1. Enhanced Client X datasets
   a. Additional independent factors
      i. Patients' BMI / weight
      ii. Patients' family chronic history
      iii. Patients' lifestyle, etc.
      iv. Doctors' notes
   b. Further increase dataset size
      i. Use the full booking set (e.g. the remaining 90% in 2021)

2. Coding & Model fine-tune
   a. Try more options to overcome imbalanced data
   b. Include Lab Test categorical results in members modeling
   c. Further modularize code (e.g. configuration set-up for different chronic group) for easier software management.

3. Solutions implementation for production deployment (e.g. Streamlit)

8. Conclusion

© 2022 National University of Singapore. All Rights Reserved

Conclusion

1. Timely completion of EDA, dataset cleaning, chronic detection models, explainability models and roll-out database
   a. 7 project meetings (Kick-off to closure) with Client X
   b. >20 internal meetings for alignment and discussions

1. Value creation for Stakeholders (Client X)
   a. Models to sieve out chronic cases through two methods
      i. booking dataset (individual patient visits)
      ii. member dataset (growing patient database)
   b. Explainability models
      i. to decipher most crucial factors for detection of chronic cases
   c. Lab data results
      i. lab measurement statistics for doctors' interpretation
   d. Research papers' findings
   e. Grouping of related chronics (e.g. Atopic Triad) through MCA/PCA analysis

© 2022 National University of Singapore. All Rights Reserved

References

[1] http://Client X.com.sg

[2] Deepika, K., Shedole, S. “Predictive analytics to prevent and control chronic diseases”, 2016

[3] Agarwal, A., Beha, S., Jain, R., Pulumati, P., Rao, V. “Predictive Modeling for Wellness and Chronic Conditions,”, 2014

[4] Wang, A., Chen, G., Li, L., An, N., Xia, Y. “A Logistic Regression and Artificial Neural Network-based Approach for Chronic Disease Prediction: a Case Study of Hypertension,” 2014

[5] Alanazi, R. “Identification and Prediction of Chronic Diseases Using Machine Learning Approach. Journal of Healthcare Engineering,” 2022

© 2022 National University of Singapore. All Rights Reserved

Thank You

Questions & Answers

© 2021 National University of Singapore. All Rights Reserved