Final Project

HERITAGE HEALTH PRIZE

Phan Thi Thuy An (an.phan.sptoan@gmail.com)

INTRODUCTION

- ✓ More than 71 million individuals in the US are admitted to hospitals each year.
- ✓ Studies have concluded that in 2006 well over $30 billion was spent on unnecessary hospital admissions.

kaggle

Can we identify earlier those most at risk and ensure they get the treatment they need?

The objective is predicting days a patient will spend in the hospital in the next year base on claims data of the year before.

TABLE OF CONTENTS

01

THE DATASETS

02

DATA PROCESSING

03

PREDICTIVE MODELS

04

RESULT

TABLE OF CONTENTS

01

THE DATASETS

02

DATA PROCESSING

03

PREDICTIVE MODELS

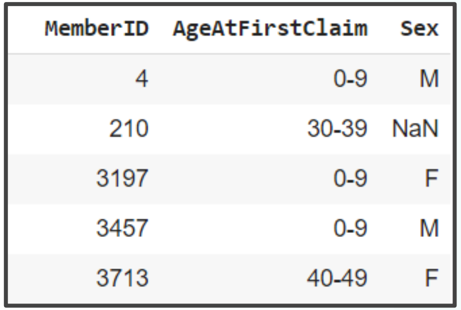
04

RESULT

THE DATASETS - HHP dataset release 3

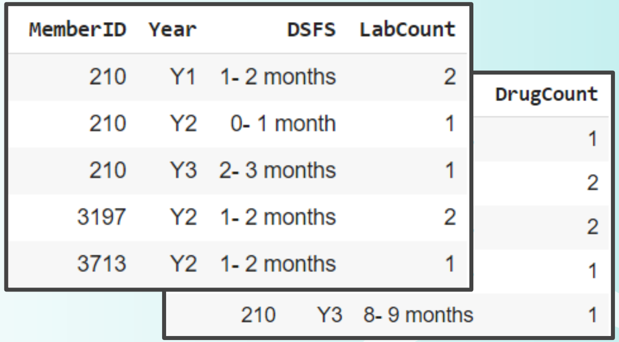
https://www.kaggle.com/c/hhp

Members Table (113000 x 3)



MemberID	AgeAtFirstClaim	Sex
4	0-9	M
210	30-39	NaN
3197	0-9	F
3457	0-9	M
3713	40-49	F

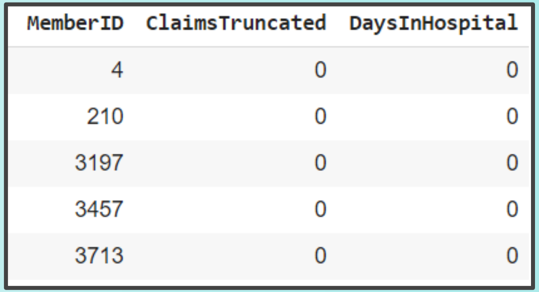
Labs Table (361484 x 4)



MemberID	Year	DSFS	LabCount
210	Y1	1- 2 months	2
210	Y2	0- 1 month	1
210	Y3	2- 3 months	1
3197	Y2	1- 2 months	2
3713	Y2	1- 2 months	1

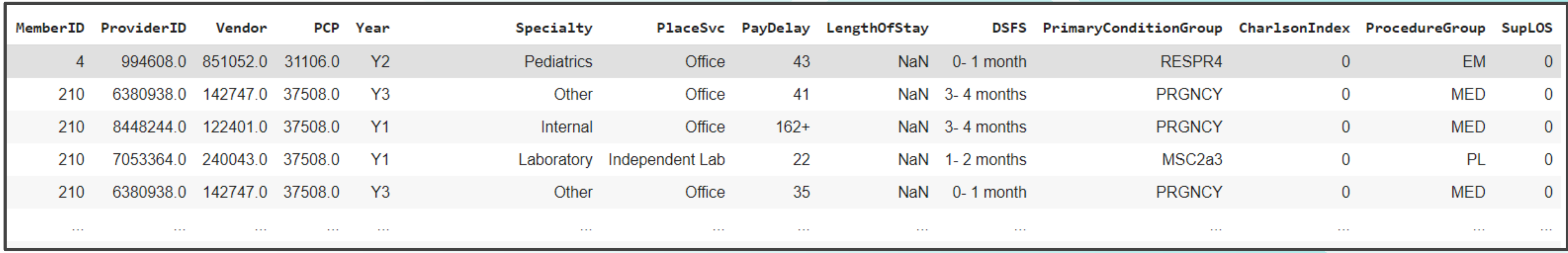
			DrugCount
			1
			2
			2
			1
210	Y3	8- 9 months	1

DaysInHospital Tables (Y2 / Y3)



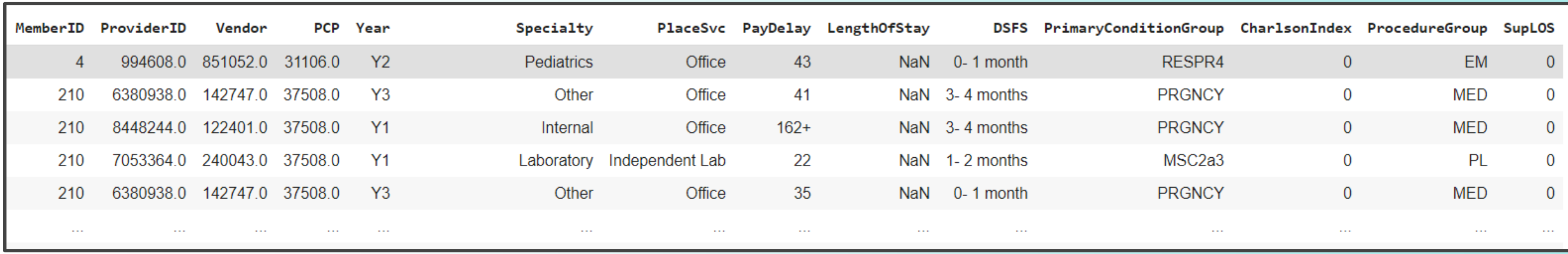
MemberID	ClaimsTruncated	DaysInHospital
4	0	0
210	0	0
3197	0	0
3457	0	0
3713	0	0

Claims Table (2668990 x 14)

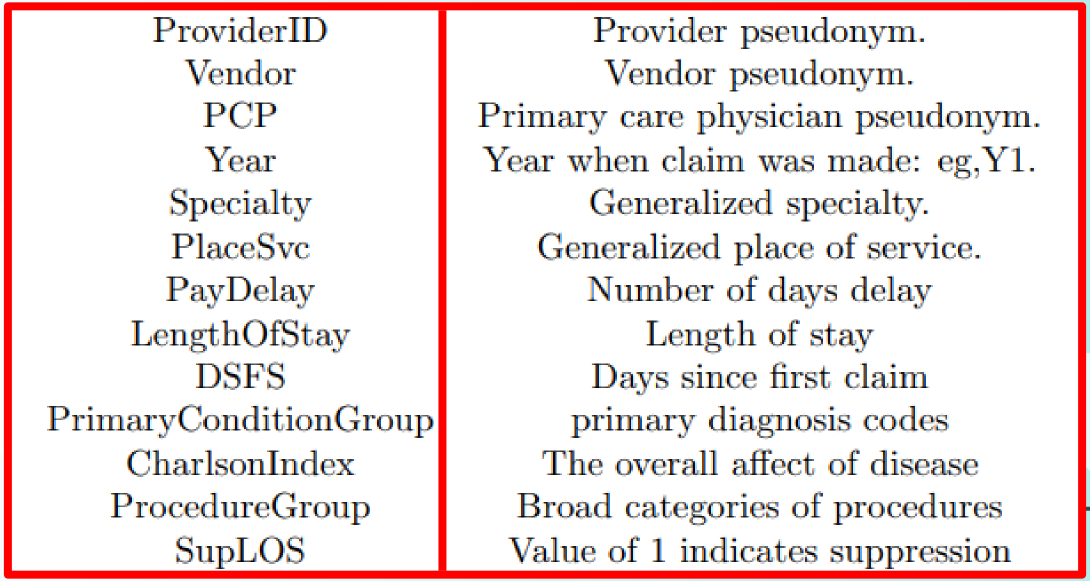


MemberID	ProviderID	Vendor	PCP	Year	Specialty	PlaceSvc	PayDelay	LengthOfStay	DSFS	PrimaryConditionGroup	CharlsonIndex	ProcedureGroup	SupLOS
4	994608.0	851052.0	31106.0	Y2	Pediatrics	Office	43	NaN	0- 1 month	RESPR4	0	EM	0
210	6380938.0	142747.0	37508.0	Y3	Other	Office	41	NaN	3- 4 months	PRGNCY	0	MED	0
210	8448244.0	122401.0	37508.0	Y1	Internal	Office	162+	NaN	3- 4 months	PRGNCY	0	MED	0
210	7053364.0	240043.0	37508.0	Y1	Laboratory	Independent Lab	22	NaN	1- 2 months	MSC2a3	0	PL	0
210	6380938.0	142747.0	37508.0	Y3	Other	Office	35	NaN	0- 1 month	PRGNCY	0	MED	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...

THE DATASETS - HHP dataset release 3



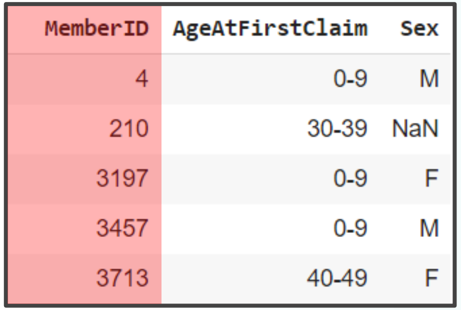
MemberID	ProviderID	Vendor	PCP	Year	Specialty	PlaceSvc	PayDelay	LengthOfStay	DSFS	PrimaryConditionGroup	CharlsonIndex	ProcedureGroup	SupLOS
4	994608.0	851052.0	31106.0	Y2	Pediatrics	Office	43	NaN	0- 1 month	RESPR4	0	EM	0
210	6380938.0	142747.0	37508.0	Y3	Other	Office	41	NaN	3- 4 months	PRGNCY	0	MED	0
210	8448244.0	122401.0	37508.0	Y1	Internal	Office	162+	NaN	3- 4 months	PRGNCY	0	MED	0
210	7053364.0	240043.0	37508.0	Y1	Laboratory	Independent Lab	22	NaN	1- 2 months	MSC2a3	0	PL	0
210	6380938.0	142747.0	37508.0	Y3	Other	Office	35	NaN	0- 1 month	PRGNCY	0	MED	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...



ProviderID	Provider pseudonym.
Vendor	Vendor pseudonym.
PCP	Primary care physician pseudonym.
Year	Year when claim was made: eg,Y1.
Specialty	Generalized specialty.
PlaceSvc	Generalized place of service.
PayDelay	Number of days delay
LengthOfStay	Length of stay
DSFS	Days since first claim
PrimaryConditionGroup	primary diagnosis codes
CharlsonIndex	The overall affect of disease
ProcedureGroup	Broad categories of procedures
SupLOS	Value of 1 indicates suppression

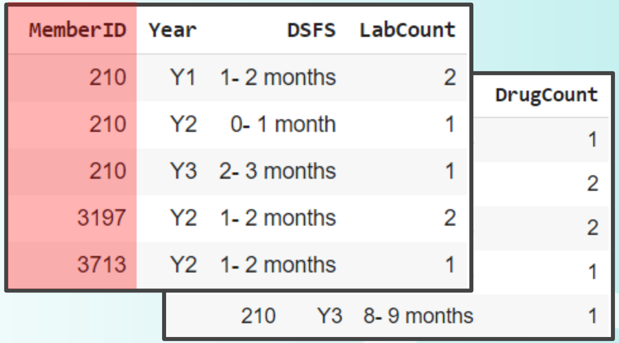
THE DATASETS - HHP dataset release 3

Members Table (113000 x 3)



MemberID	AgeAtFirstClaim	Sex
4	0-9	M
210	30-39	NaN
3197	0-9	F
3457	0-9	M
3713	40-49	F

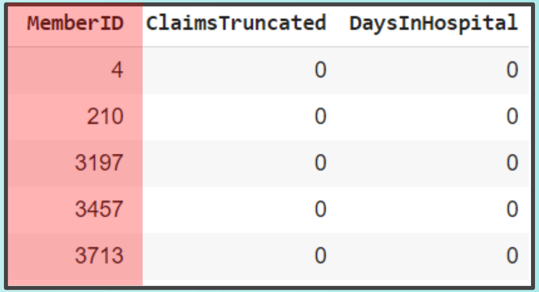
Labs Table (361484 x 4)



MemberID	Year	DSFS	LabCount
210	Y1	1- 2 months	2
210	Y2	0- 1 month	1
210	Y3	2- 3 months	1
3197	Y2	1- 2 months	2
3713	Y2	1- 2 months	1

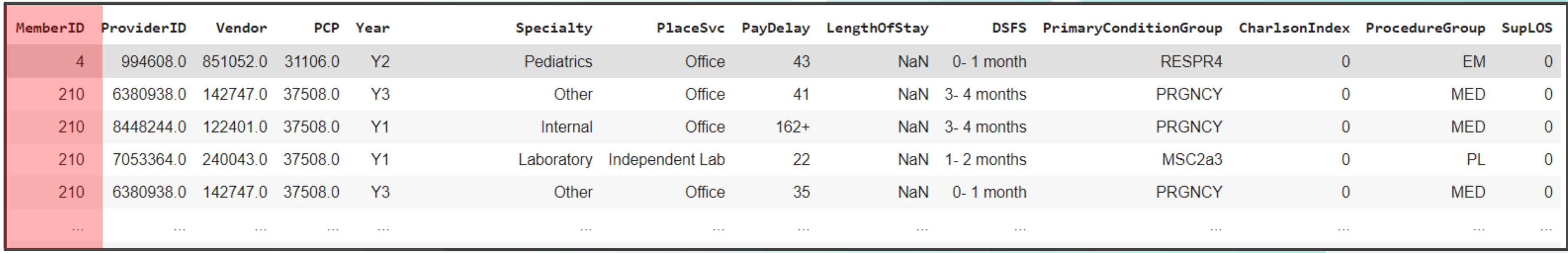
			DrugCount
			1
			2
			2
			1
210	Y3	8- 9 months	1

DaysInHospital Tables (Y2 / Y3)



MemberID	ClaimsTruncated	DaysInHospital
4	0	0
210	0	0
3197	0	0
3457	0	0
3713	0	0

Claims Table (2668990 x 14)



MemberID	ProviderID	Vendor	PCP	Year	Specialty	PlaceSvc	PayDelay	LengthOfStay	DSFS	PrimaryConditionGroup	CharlsonIndex	ProcedureGroup	SupLOS
4	994608.0	851052.0	31106.0	Y2	Pediatrics	Office	43	NaN	0- 1 month	RESPR4	0	EM	0
210	6380938.0	142747.0	37508.0	Y3	Other	Office	41	NaN	3- 4 months	PRGNCY	0	MED	0
210	8448244.0	122401.0	37508.0	Y1	Internal	Office	162+	NaN	3- 4 months	PRGNCY	0	MED	0
210	7053364.0	240043.0	37508.0	Y1	Laboratory	Independent Lab	22	NaN	1- 2 months	MSC2a3	0	PL	0
210	6380938.0	142747.0	37508.0	Y3	Other	Office	35	NaN	0- 1 month	PRGNCY	0	MED	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...

To facilitate the prediction models, we combine 5 tables into one consistent dataset.

TRAINING AND TESTING DATA

Split claims by year

- Y1: 865689 Claims, 76038 Patients.
- Y2: 898872 Claims, 71435 Patients.
- Y3: 904429 Claims (missing outcome).

TRAINING AND TESTING DATA

Split claims by year

- Y1: 865689 Claims, 76038 Patients.
- Y2: 898872 Claims, 71435 Patients.

TRAINING AND TESTING DATA

SOLUTION 1

- Y1: 865689 Claims, 76038 Patients.

  DaysInHospital_Y2

- Y2: 898872 Claims, 71435 Patients.

  DaysInHospital_Y3

TRAINING AND TESTING DATA

SOLUTION 1

- Y1: 865689 Claims, 76038 Patients.

  DaysInHospital_Y2

- Y2: 898872 Claims, 71435 Patients.

  DaysInHospital_Y3

TRAINING AND TESTING DATA

SOLUTION 1

TRAINING

- Y1: 865689 Claims, 76038 Patients.

  DaysInHospital_Y2

- Y2: 898872 Claims, 71435 Patients.

  DaysInHospital_Y3

TESTING

TRAINING AND TESTING DATA

SOLUTION 2

- Y1: 865689 Claims, 76038 Patients.

  DaysInHospital_Y2

- Y2: 898872 Claims, 71435 Patients.

  DaysInHospital_Y3

TRAINING AND TESTING DATA

SOLUTION 2

- Y1: 865689 Claims, 76038 Patients.

  DaysInHospital_Y2

- Y2: 898872 Claims, 71435 Patients.

  DaysInHospital_Y3

TRAINING AND TESTING DATA

SOLUTION 2

- Y1: 865689 Claims, 76038 Patients.

DaysInHospital_Y2

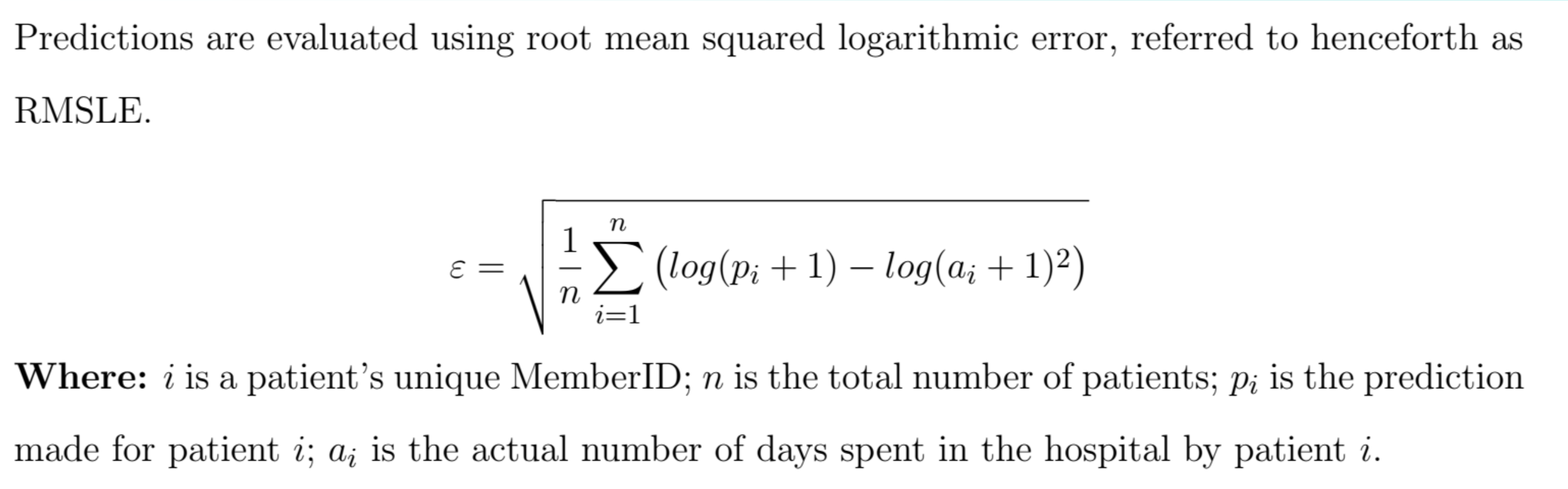
- Y2: 898872 Claims, 71435 Patients.

DaysInHospital_Y3

TRAINING | **TESTING**

EVALUATION

RMSLE



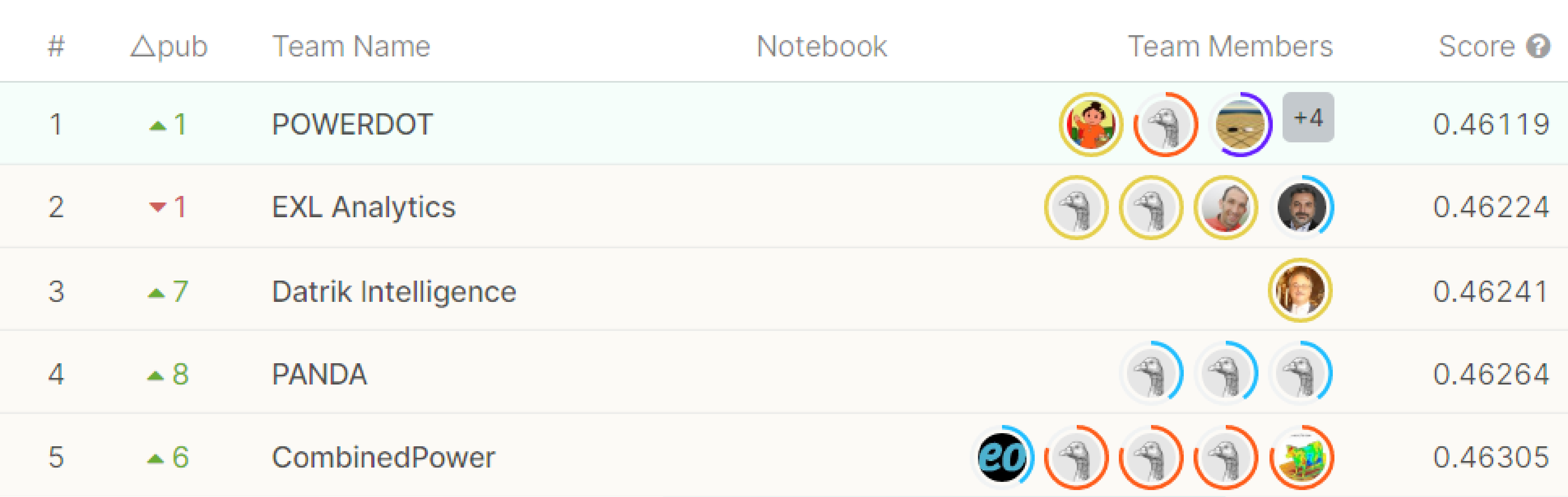
Predictions are evaluated using root mean squared logarithmic error, referred to henceforth as RMSLE.

$$\varepsilon = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(log(p_i+1) - log(a_i+1)^2\right)}$$

Where: i is a patient's unique MemberID; n is the total number of patients; p_i is the prediction made for patient i; a_i is the actual number of days spent in the hospital by patient i.

EVALUATION

RMSLE



#	Δpub	Team Name	Notebook	Team Members	Score
1	▲1	POWERDOT		+4	0.46119
2	▼1	EXL Analytics			0.46224
3	▲7	Datrik Intelligence			0.46241
4	▲8	PANDA			0.46264
5	▲6	CombinedPower			0.46305

TABLE OF CONTENTS

01

THE DATASETS

02

DATA PROCESSING

03

PREDICTIVE MODELS

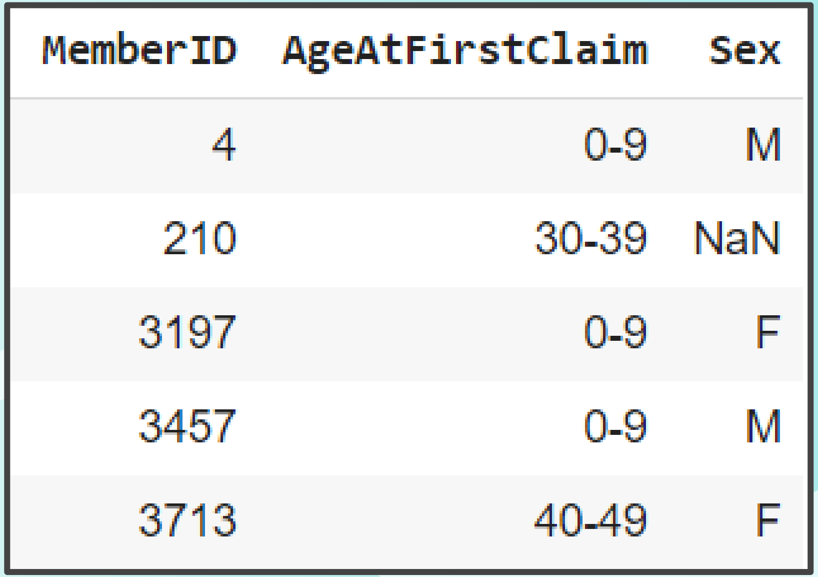
04

RESULT

DATA PROCESSING

FEATURE ENGINEERING

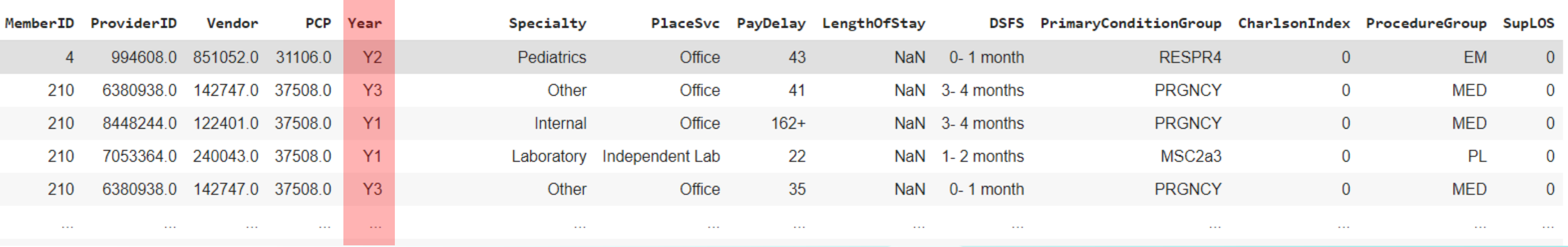
AgeAtFirstClaim	- Replace by Mean of each interval - Fill NanN with 45
Sex	Onehot-Encoding with 3 columns Female, Male, Unknown



MemberID	AgeAtFirstClaim	Sex
4	0-9	M
210	30-39	NaN
3197	0-9	F
3457	0-9	M
3713	40-49	F

DATA PROCESSING

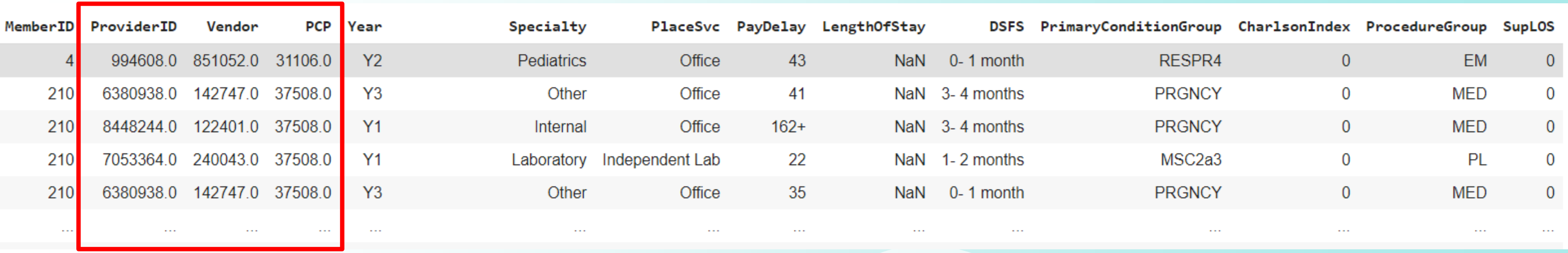
FEATURE ENGINEERING



MemberID	ProviderID	Vendor	PCP	Year	Specialty	PlaceSvc	PayDelay	LengthOfStay	DSFS	PrimaryConditionGroup	CharlsonIndex	ProcedureGroup	SupLOS
4	994608.0	851052.0	31106.0	Y2	Pediatrics	Office	43	NaN	0- 1 month	RESPR4	0	EM	0
210	6380938.0	142747.0	37508.0	Y3	Other	Office	41	NaN	3- 4 months	PRGNCY	0	MED	0
210	8448244.0	122401.0	37508.0	Y1	Internal	Office	162+	NaN	3- 4 months	PRGNCY	0	MED	0
210	7053364.0	240043.0	37508.0	Y1	Laboratory	Independent Lab	22	NaN	1- 2 months	MSC2a3	0	PL	0
210	6380938.0	142747.0	37508.0	Y3	Other	Office	35	NaN	0- 1 month	PRGNCY	0	MED	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...

DATA PROCESSING

FEATURE ENGINEERING

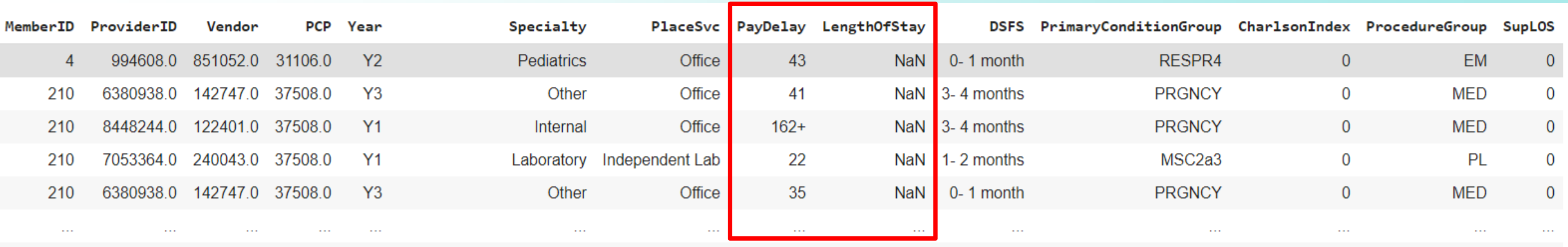


MemberID	ProviderID	Vendor	PCP	Year	Specialty	PlaceSvc	PayDelay	LengthOfStay	DSFS	PrimaryConditionGroup	CharlsonIndex	ProcedureGroup	SupLOS
4	994608.0	851052.0	31106.0	Y2	Pediatrics	Office	43	NaN	0- 1 month	RESPR4	0	EM	0
210	6380938.0	142747.0	37508.0	Y3	Other	Office	41	NaN	3- 4 months	PRGNCY	0	MED	0
210	8448244.0	122401.0	37508.0	Y1	Internal	Office	162+	NaN	3- 4 months	PRGNCY	0	MED	0
210	7053364.0	240043.0	37508.0	Y1	Laboratory	Independent Lab	22	NaN	1- 2 months	MSC2a3	0	PL	0
210	6380938.0	142747.0	37508.0	Y3	Other	Office	35	NaN	0- 1 month	PRGNCY	0	MED	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...

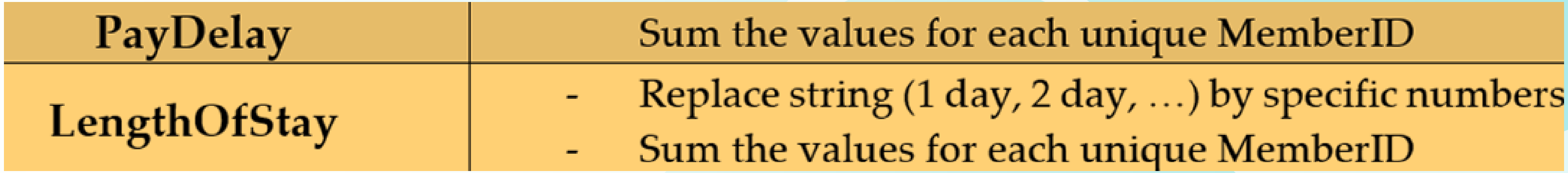
- Count values of Provider to find number of Claims
- Count distinct value for unique MemberID

DATA PROCESSING

FEATURE ENGINEERING



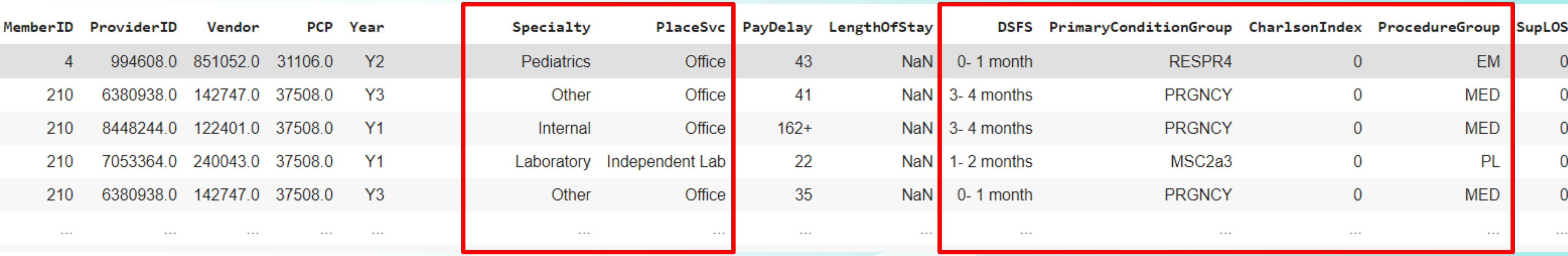
MemberID	ProviderID	Vendor	PCP	Year	Specialty	PlaceSvc	PayDelay	LengthOfStay	DSFS	PrimaryConditionGroup	CharlsonIndex	ProcedureGroup	SupLOS
4	994608.0	851052.0	31106.0	Y2	Pediatrics	Office	43	NaN	0- 1 month	RESPR4	0	EM	0
210	6380938.0	142747.0	37508.0	Y3	Other	Office	41	NaN	3- 4 months	PRGNCY	0	MED	0
210	8448244.0	122401.0	37508.0	Y1	Internal	Office	162+	NaN	3- 4 months	PRGNCY	0	MED	0
210	7053364.0	240043.0	37508.0	Y1	Laboratory	Independent Lab	22	NaN	1- 2 months	MSC2a3	0	PL	0
210	6380938.0	142747.0	37508.0	Y3	Other	Office	35	NaN	0- 1 month	PRGNCY	0	MED	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...



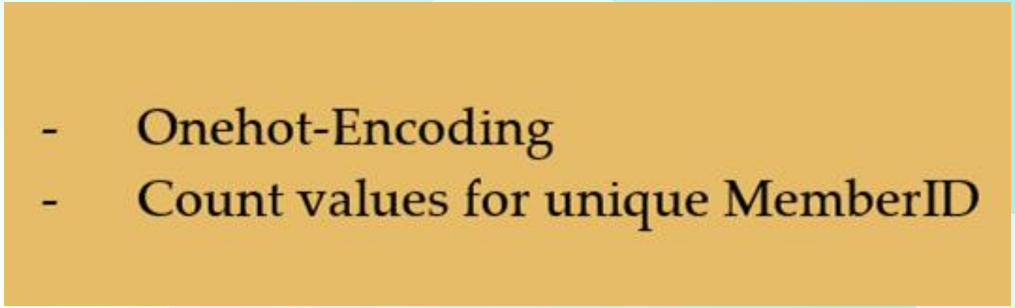
PayDelay	Sum the values for each unique MemberID
LengthOfStay	- Replace string (1 day, 2 day, ...) by specific numbers - Sum the values for each unique MemberID

DATA PROCESSING

FEATURE ENGINEERING



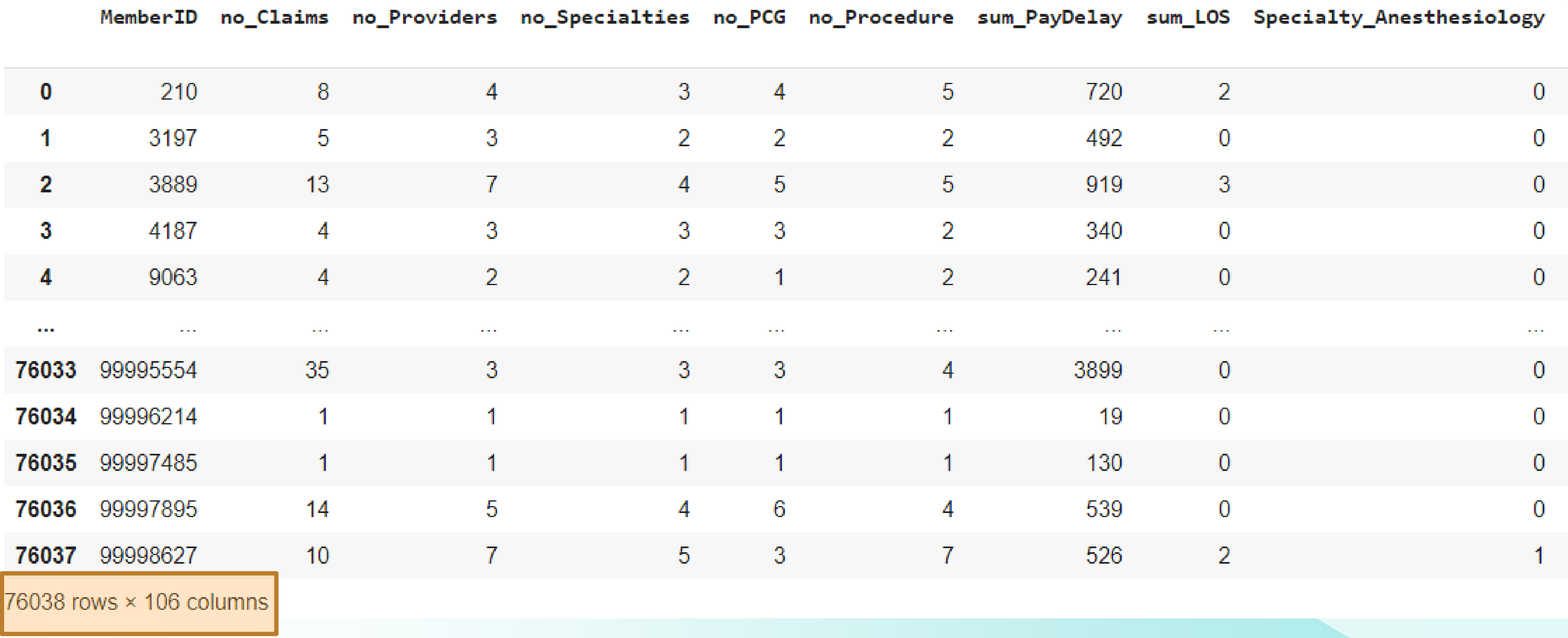
MemberID	ProviderID	Vendor	PCP	Year	Specialty	PlaceSvc	PayDelay	LengthOfStay	DSFS	PrimaryConditionGroup	CharlsonIndex	ProcedureGroup	SupLOS
4	994608.0	851052.0	31106.0	Y2	Pediatrics	Office	43	NaN	0- 1 month	RESPR4	0	EM	0
210	6380938.0	142747.0	37508.0	Y3	Other	Office	41	NaN	3- 4 months	PRGNCY	0	MED	0
210	8448244.0	122401.0	37508.0	Y1	Internal	Office	162+	NaN	3- 4 months	PRGNCY	0	MED	0
210	7053364.0	240043.0	37508.0	Y1	Laboratory	Independent Lab	22	NaN	1- 2 months	MSC2a3	0	PL	0
210	6380938.0	142747.0	37508.0	Y3	Other	Office	35	NaN	0- 1 month	PRGNCY	0	MED	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...



- Onehot-Encoding
- Count values for unique MemberID

DATA PROCESSING

FEATURE ENGINEERING

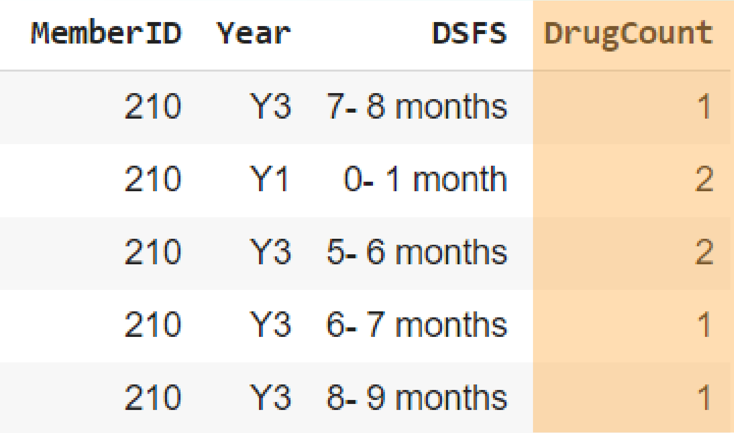


	MemberID	no_Claims	no_Providers	no_Specialties	no_PCG	no_Procedure	sum_PayDelay	sum_LOS	Specialty_Anesthesiology
0	210	8	4	3	4	5	720	2	0
1	3197	5	3	2	2	2	492	0	0
2	3889	13	7	4	5	5	919	3	0
3	4187	4	3	3	3	2	340	0	0
4	9063	4	2	2	1	2	241	0	0
...	...	...	...	...	...	...	...	...	...
76033	99995554	35	3	3	3	4	3899	0	0
76034	99996214	1	1	1	1	1	19	0	0
76035	99997485	1	1	1	1	1	130	0	0
76036	99997895	14	5	4	6	4	539	0	0
76037	99998627	10	7	5	3	7	526	2	1

76038 rows × 106 columns

DATA PROCESSING

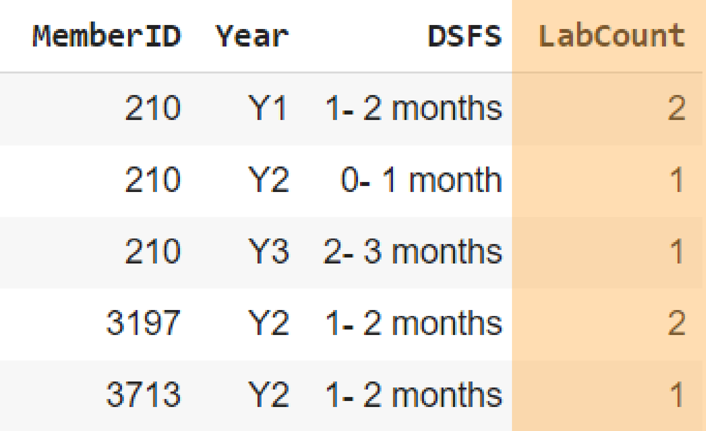
FEATURE ENGINEERING



MemberID	Year	DSFS	DrugCount
210	Y3	7- 8 months	1
210	Y1	0- 1 month	2
210	Y3	5- 6 months	2
210	Y3	6- 7 months	1
210	Y3	8- 9 months	1

SUM

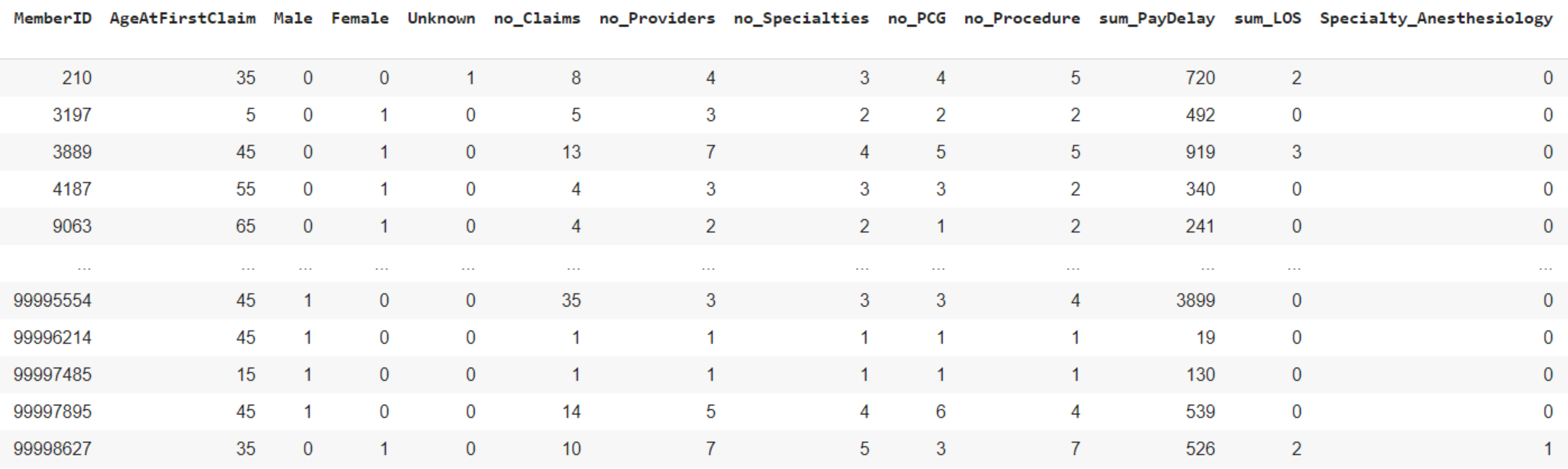
SUM



MemberID	Year	DSFS	LabCount
210	Y1	1- 2 months	2
210	Y2	0- 1 month	1
210	Y3	2- 3 months	1
3197	Y2	1- 2 months	2
3713	Y2	1- 2 months	1

DATA PROCESSING

FEATURE ENGINEERING

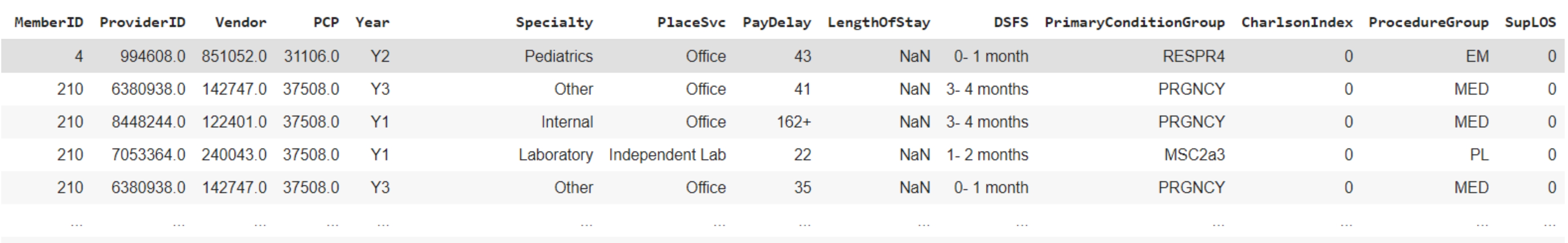


MemberID	AgeAtFirstClaim	Male	Female	Unknown	no_Claims	no_Providers	no_Specialties	no_PCG	no_Procedure	sum_PayDelay	sum_LOS	Specialty_Anesthesiology
210	35	0	0	1	8	4	3	4	5	720	2	0
3197	5	0	1	0	5	3	2	2	2	492	0	0
3889	45	0	1	0	13	7	4	5	5	919	3	0
4187	55	0	1	0	4	3	3	3	2	340	0	0
9063	65	0	1	0	4	2	2	1	2	241	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...
99995554	45	1	0	0	35	3	3	3	4	3899	0	0
99996214	45	1	0	0	1	1	1	1	1	19	0	0
99997485	15	1	0	0	1	1	1	1	1	130	0	0
99997895	45	1	0	0	14	5	4	6	4	539	0	0
99998627	35	0	1	0	10	7	5	3	7	526	2	1

After merging 5 tables, we get **114 unique features** in total.

DATA PROCESSING

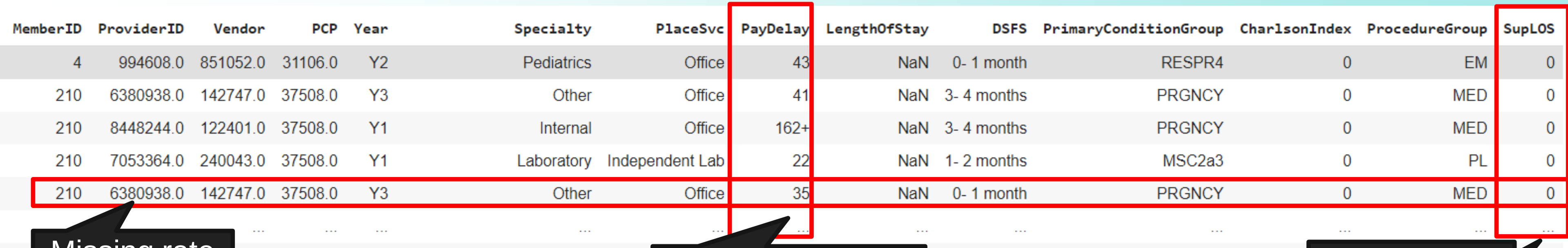
DROPPED DATA



MemberID	ProviderID	Vendor	PCP	Year	Specialty	PlaceSvc	PayDelay	LengthOfStay	DSFS	PrimaryConditionGroup	CharlsonIndex	ProcedureGroup	SupLOS
4	994608.0	851052.0	31106.0	Y2	Pediatrics	Office	43	NaN	0- 1 month	RESPR4	0	EM	0
210	6380938.0	142747.0	37508.0	Y3	Other	Office	41	NaN	3- 4 months	PRGNCY	0	MED	0
210	8448244.0	122401.0	37508.0	Y1	Internal	Office	162+	NaN	3- 4 months	PRGNCY	0	MED	0
210	7053364.0	240043.0	37508.0	Y1	Laboratory	Independent Lab	22	NaN	1- 2 months	MSC2a3	0	PL	0
210	6380938.0	142747.0	37508.0	Y3	Other	Office	35	NaN	0- 1 month	PRGNCY	0	MED	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...

DATA PROCESSING

DROPPED DATA

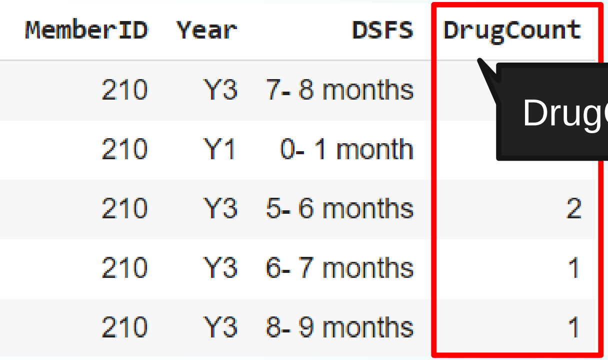


MemberID	ProviderID	Vendor	PCP	Year	Specialty	PlaceSvc	PayDelay	LengthOfStay	DSFS	PrimaryConditionGroup	CharlsonIndex	ProcedureGroup	SupLOS
4	994608.0	851052.0	31106.0	Y2	Pediatrics	Office	43	NaN	0- 1 month	RESPR4	0	EM	0
210	6380938.0	142747.0	37508.0	Y3	Other	Office	41	NaN	3- 4 months	PRGNCY	0	MED	0
210	8448244.0	122401.0	37508.0	Y1	Internal	Office	162+	NaN	3- 4 months	PRGNCY	0	MED	0
210	7053364.0	240043.0	37508.0	Y1	Laboratory	Independent Lab	22	NaN	1- 2 months	MSC2a3	0	PL	0
210	6380938.0	142747.0	37508.0	Y3	Other	Office	35	NaN	0- 1 month	PRGNCY	0	MED	0
		...	...	...	...	...	...	...	...	...	...	...	...

Missing rate > 50%

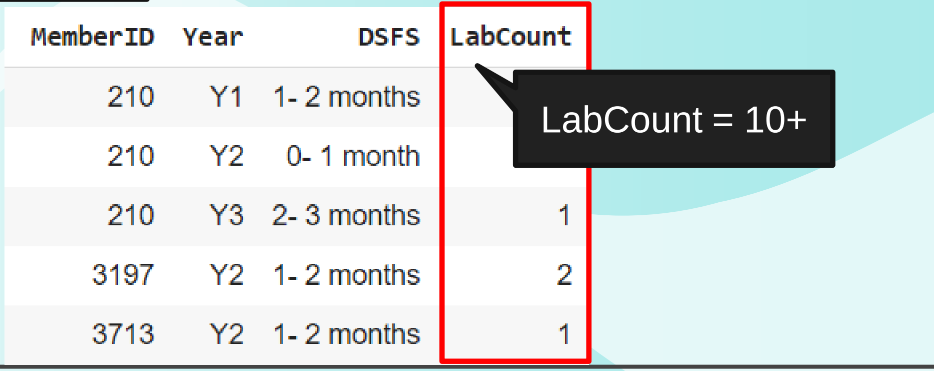
PayDelay = 162+

SupLOS = 1



MemberID	Year	DSFS	DrugCount
210	Y3	7- 8 months	
210	Y1	0- 1 month	
210	Y3	5- 6 months	2
210	Y3	6- 7 months	1
210	Y3	8- 9 months	1

DrugCount = 7+



MemberID	Year	DSFS	LabCount
210	Y1	1- 2 months	
210	Y2	0- 1 month	
210	Y3	2- 3 months	1
3197	Y2	1- 2 months	2
3713	Y2	1- 2 months	1

LabCount = 10+

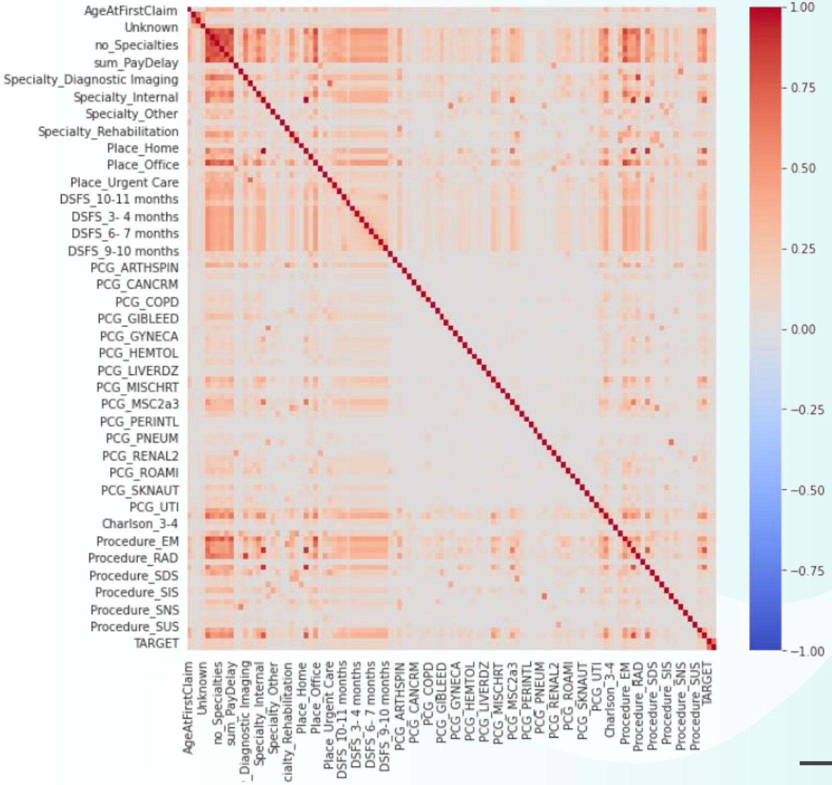
DATA PROCESSING

3 versions of the dataset:

- **Full version** of each year (supporting Solution 1)
- **Full version** of 2 years (supporting Solution 2)
- **Dropped version** of each year (supporting Solution 1)

DATA PROCESSING

FEATURES SELECTION



After calculating **correlation matrix,**

we keep **27 features**

correlation with target > 0.1 and with others < 0.9

TABLE OF CONTENTS

01

THE DATASETS

02

DATA PROCESSING

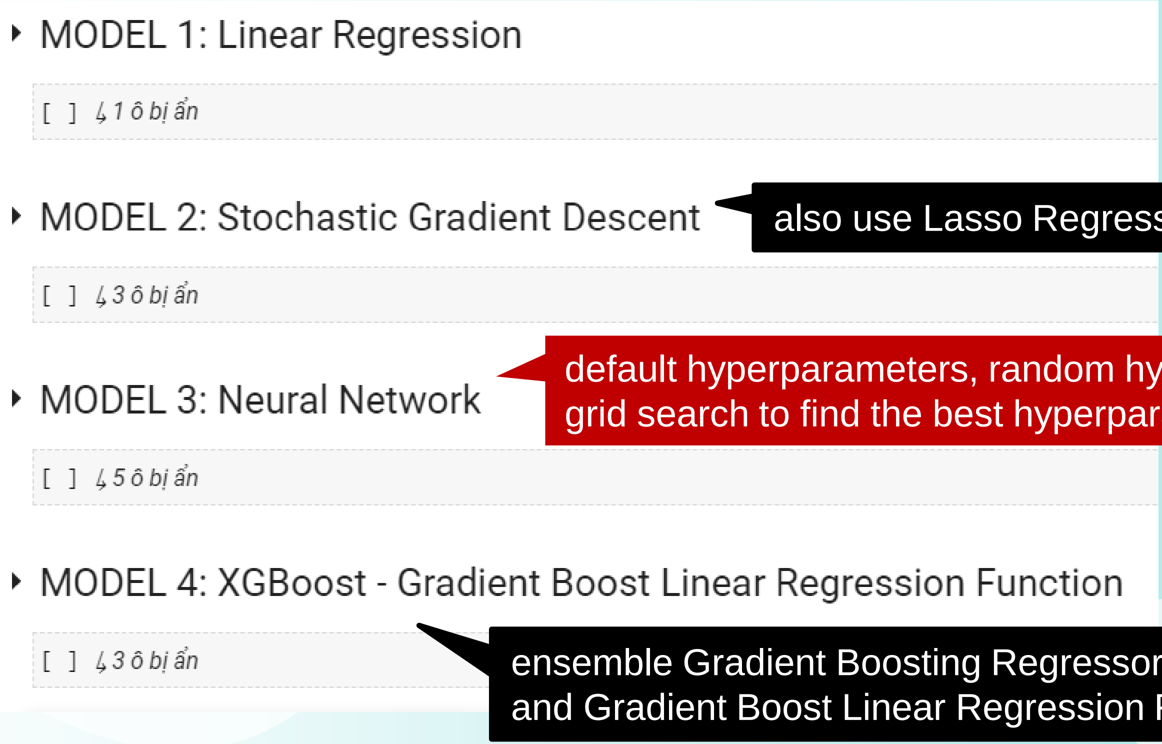
03

PREDICTIVE MODELS

04

RESULT

PREDICTIVE MODELS



MODEL 1: Linear Regression

[ ] ↳ 1 ô bị ẩn

MODEL 2: Stochastic Gradient Descent

[ ] ↳ 3 ô bị ẩn

MODEL 3: Neural Network

[ ] ↳ 5 ô bị ẩn

MODEL 4: XGBoost - Gradient Boost Linear Regression Function

[ ] ↳ 3 ô bị ẩn

also use Lasso Regression

default hyperparameters, random hyperparameters, grid search to find the best hyperparameters

ensemble Gradient Boosting Regressor of Scikit-learn and Gradient Boost Linear Regression Function

PREDICTIVE MODELS

The sets of 4 Models are applied for

❑ full data of each year.

❑ full data of each year with Features Selection.

❑ dropped data of each year.

❑ dropped data of each year with Features Selection.

❑ full data of 2 years.

❑ full data of 2 years with Features Selection.

TABLE OF CONTENTS

01

THE DATASETS

02

DATA PROCESSING

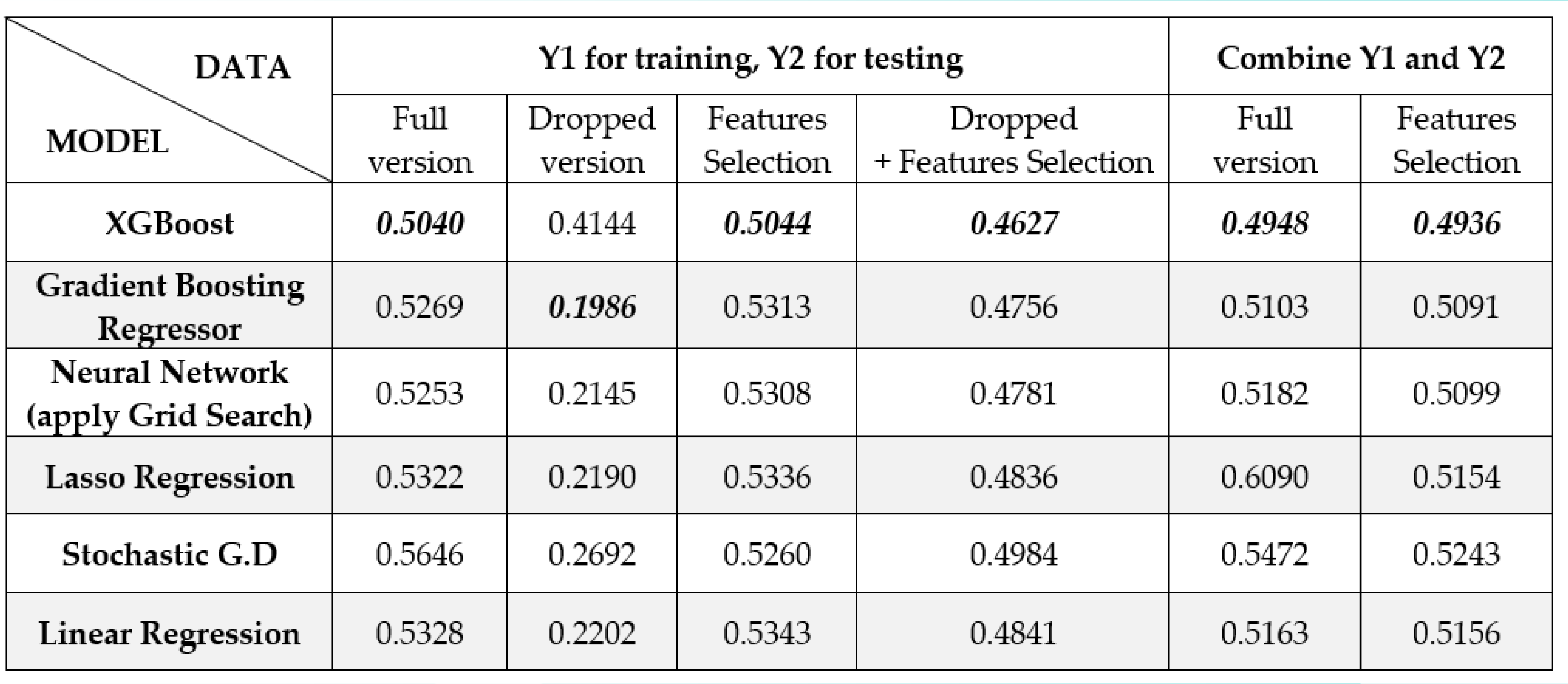
03

PREDICTIVE MODELS

04

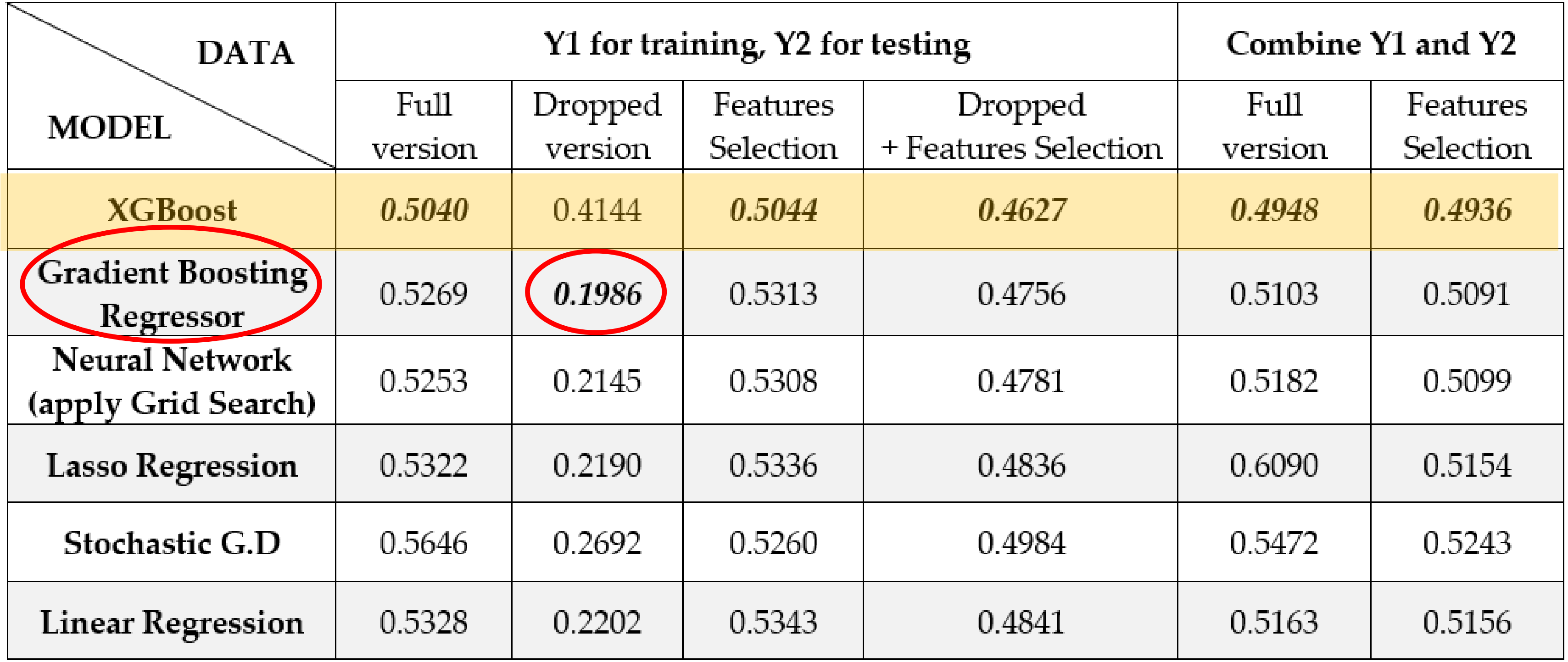
RESULT

RESULT



DATA MODEL	Y1 for training, Y2 for testing				Combine Y1 and Y2	
	Full version	Dropped version	Features Selection	Dropped + Features Selection	Full version	Features Selection
XGBoost	***0.5040***	0.4144	***0.5044***	***0.4627***	***0.4948***	***0.4936***
Gradient Boosting Regressor	0.5269	***0.1986***	0.5313	0.4756	0.5103	0.5091
Neural Network (apply Grid Search)	0.5253	0.2145	0.5308	0.4781	0.5182	0.5099
Lasso Regression	0.5322	0.2190	0.5336	0.4836	0.6090	0.5154
Stochastic G.D	0.5646	0.2692	0.5260	0.4984	0.5472	0.5243
Linear Regression	0.5328	0.2202	0.5343	0.4841	0.5163	0.5156

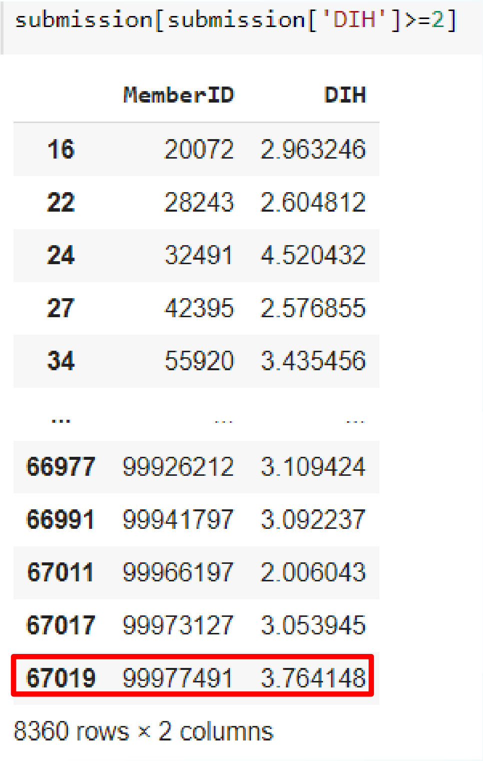
RESULT



DATA MODEL	Y1 for training, Y2 for testing				Combine Y1 and Y2	
	Full version	Dropped version	Features Selection	Dropped + Features Selection	Full version	Features Selection
XGBoost	***0.5040***	0.4144	***0.5044***	***0.4627***	***0.4948***	***0.4936***
Gradient Boosting Regressor	0.5269	***0.1986***	0.5313	0.4756	0.5103	0.5091
Neural Network (apply Grid Search)	0.5253	0.2145	0.5308	0.4781	0.5182	0.5099
Lasso Regression	0.5322	0.2190	0.5336	0.4836	0.6090	0.5154
Stochastic G.D	0.5646	0.2692	0.5260	0.4984	0.5472	0.5243
Linear Regression	0.5328	0.2202	0.5343	0.4841	0.5163	0.5156

RESULT

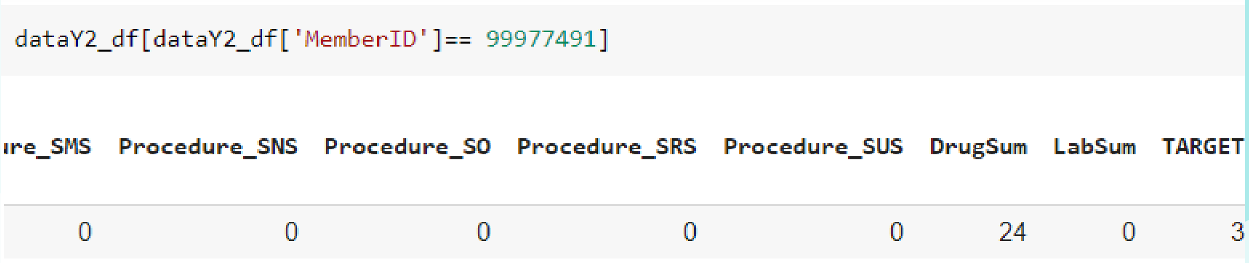
Base on this result, we apply ensemble Gradient Boosting Regressor of Scikit-learn for the dropped version data of each year



```
submission[submission['DIH']>=2]
```

	MemberID	DIH
16	20072	2.963246
22	28243	2.604812
24	32491	4.520432
27	42395	2.576855
34	55920	3.435456
...	...	...
66977	99926212	3.109424
66991	99941797	3.092237
67011	99966197	2.006043
67017	99973127	3.053945
67019	99977491	3.764148

8360 rows × 2 columns



```
dataY2_df[dataY2_df['MemberID']== 99977491]
```

re_SMS	Procedure_SNS	Procedure_SO	Procedure_SRS	Procedure_SUS	DrugSum	LabSum	TARGET
0	0	0	0	0	24	0	3

LESSON LEARNT

- ❑ 70% amount of time to prepare data.
- ❑ Features Selection may help reducing the time taken, but cleaning data tends to show better results.
- ❑ Beside ensembling method, data extraction and features selection also should be done in more different ways.

POTENTIAL IMPROVEMENT

Apply another way for processing data

remove (1) the patients whose **length of stay** (LOS) in hospital tended to be longer, (2) they tended to be **older**, (3) they tended to have **more claims**.

POTENTIAL IMPROVEMENT

Divided into **2 stages**:

The 1st stage is Classification, which define whether the patient will be in hospital in the next year or not.

Then, the classified result becomes input of Regression - the 2nd stage.

SOURCE CODE

https://github.com/anphantt2406/Heritage-Health-Prize.git

Thank You For Your Listening!