Heritage Health Prize

Loc DOAN, Ha PHAM, and Van NGUYEN

Contents

1 Member contributions **3**

2 Problem **4**

3 Preprocessing data **5**

4 Candidate models **7**

5 Result **10**

Appendices **13**

A Raw Data **13**

B Features **14**

List of Figures

1 Summary of missing value in Claims Dataset 5
2 Counting claim information per Year and MemberID 5
3 DaysInHospital Year 2 7
4 Comparison between actual and prediction DIH in Year 3 8
5 Comparison predictions between algorithms 9
6 Heat map of base predictions 10

List of Tables

1 Result 10
2 Members Data 13
3 Claims (level 2) Data 13
4 LabCount 14
5 DrugCount 14
6 Outcome Data 14
7 List of features 15

Abstract

According to the latest survey from the American Hospital Association, more than 71 million individuals in the United States are admitted to hospitals each year. However studies have concluded that in 2006 well over $30 billion was spent on unnecessary hospital admissions. The Heritage Provider Network believes that there is a better way to identify earlier those most at risk and ensure they get the treatment they need. To do this, we need to create an algorithm that predicts how many days a patient will spend in a hospital in the next year. Once known, health care providers can develop new care plans and strategies to reach patients before emergencies occur, thereby reducing the number of unnecessary hospitalizations. This will result in increasing the health of patients while decreasing the cost of care. (see more details at the webpage https://www.kaggle.com/c/hhp).

In the scope of the project "Heritage Health Prize", we would like to apply the knowledge gained from the machine learning class to handle the problem in our own way with the reference of the leader boards' reports in the milestones on the kaggle data. The best score we obtain up to this point is 0.446664.

1 Member contributions

Our team has 3 members. The principal executors in the project are Loc DOAN and Ha PHAM. Van NGUYEN takes part in group discussions and tests code.

Firstly we worked together to find out a plan for project: find references, explore data, preprocessing data, training models. Secondly, each member taked responsibility on coding a particular part:

- Loc DOAN: preprocess data
- Ha PHAM: train models

Thirdly, every part is reviewed by another members in the group to ensure that everyone can follow the project procedure and correct errors. Finally all members participant in writing report.

2 Problem

There are four data sets in Release 3 provided in the Heritage Health prize: Members Data (Table 2), Claims Data (Table 3), Lab Count Data (Table 4), Drug Count Data (Table 5) and Outcome Data (Table 6) in Appendix A. In the dataset, the basis information of each member is provided in Members Data, Claims Data, Lab Count Data, and Drug Count Data. The claim data is available for 3 years: Y1, Y2, Y3. There are DaysInHospital (DIH) data for Y2 and Y3.

The goal is to predict how many days a patient will spend in a hospital in Y3. In order to build a prediction model, it is necessary to train a model based on member information in Y1 and his/her DIH in Y2.

How to evaluate a model to be more accuracy in prediction DIH in Y3 than another one? An entry's predictive accuracy will be judged by comparing (i) the predicted number of days a member will spend in the hospital reflected in the entry and (ii) the actual number of days a member spent in the hospital. Prediction accuracy of a model is evaluated based on $RMSLE$ (Root Mean Squared Logarithmic Error):

$$RMSLE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}[\log(p_i+1) - \log(a_i+1)]^2}$$

where

- i is a member;
- n is the total number of members;
- p_i is the predicted number of days spent in hospital for member i in the test period;
- a_i is the actual number of days spent in hospital for member i in the test period.

RMSLE provides square root of mean square error between log scale of prediction and actual DaysInHospital in one year. The smaller RMSLE is, the more accuracy the prediction model is considered to be.

According to our point of view, this is a regression problem, and thus the typical accuracy measure for a regression problem is Root Mean Square Error ($RMSE$). In order to explore existing algorithms for minimizing error, we transform the number of days spent in hospital into log scale, that is, considering $\log(\mathrm{DIH}+1)$ in instead of DIH then $RMSLE$ becomes $RMSE$. In addition, regression training model directly on DIH can provide a very big negative prediction, which can cause error on $RMSLE$. Hence all our model trains dataset on log scale of DIH, i.e. $\log(\mathrm{DIH}+1)$. Such log scaled transform is also used in almost Milestones, for instance [1, 2] and some others. At the end, all negative prediction should be truncated for convenience.

3 Preprocessing data

Raw information of a member contains categorical values (Sex, AgeAtFirstClaims, ProviderID, VendorID, PCP, Year, Speciality, PrimaryConditionGroup, ProcedureGroup), numerical values (PayDelay, LengthOfStay, DSFS, CharlsonIndex, SupLOS) and a lot of missing values in Claims Dataset (Figure 1).

	Missing Values	% of Total Values
LengthOfStay	2597392	97.3
DSFS	52770	2.0
Vendor	24856	0.9
ProviderID	16264	0.6
PrimaryConditionGroup	11410	0.4
Specialty	8405	0.3
PlaceSvc	7632	0.3
PCP	7492	0.3
ProcedureGroup	3675	0.1

Figure 1: Summary of missing value in Claims Dataset

Each member can have one or more claims per year (Figure 2).

		no_Providers	no_Vendors	no_PCPs	no_PlaceSvcs	no_Specialities	no_PrimaryConditionGroups	no_ProcedureGroups
MemberID	**Year**							
4	**Y2**	1	1	1	1	1	1	1
210	**Y1**	4	4	2	3	3	4	5
	Y2	3	3	1	2	3	2	3
	Y3	2	2	1	2	2	2	2
3197	**Y1**	3	3	1	2	2	2	2

Figure 2: Counting claim information per Year and MemberID

It is necessary to summarize information of each member into an observation data. Thanks to [1] for giving us such a detail guide to process from raw data to a useful one. The basis idea is to group information of each patient by MemberID and Year. Given a MemberID and a Year, we need to idenfity the following information of the corresponding patient

- Sex
- Age
- Number of claims
- Which and how many Provider/Vendor/PlaceSvc/Speciality/PrimaryConditionGroup/ProcedureGroups are used?

- PayDelay/DSFS/DrugCount/LabCount/CharlsonIndex/LengthOfStay: choose a feature to represent the corresponding information (max, min, average, standard deviation, range)
- Construct feature to indicate whether LengthOfStay is available. In the case LengthOfStay is not available, create feature to indicate whether it is supprimed for security.

All features derived from raw data are listed in Figure 7. Some models use all features in Table 7 (in the dataset *train_data_hhp.csv* and *test_data_hhp.csv*) and some others do not use ProviderID and VendorID (in the dataset *train_data_hhp*1.*csv* and *test_data_hhp*1.*csv*).

4 Candidate models

Underlying algorithms used in our model are Linear Regression, XGboost (regression), LGBM (regression) RandomForest (regression) and Neutral Network (regression).

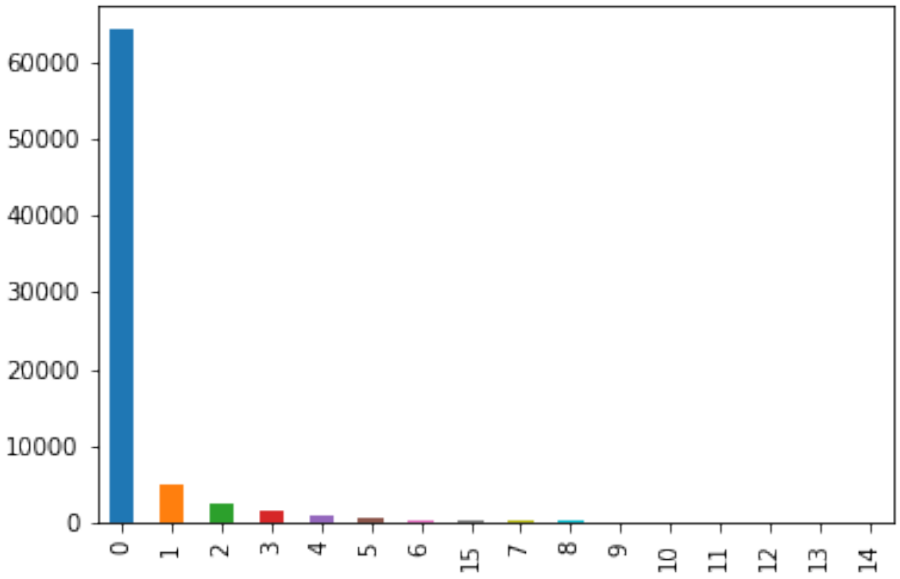

Figure 3: DaysInHospital Year 2

Figure 3 shows that the target DIH is imbalanced. Over 85% of claims has 0 DIH. To deal with imbalanced problem, we have tried to resample data by some techniques such as over and under resampling (Smoteenn), over resampling (Smote, Adasyn) but they do not work well in this case.

Another approach is tuning parameters (for example, $n\_estimator$, max_depth in LGBM, XGB, RF). For neutral network regressor, prediction is very sensitive with numbers of hidden layers and nodes. We have not found yet any solution to choose good parameters. Consequently, prediction accuracy in all algorithms is not high (see Figure 4).

Almost prediction tends closely to 0. One can see disagreement in prediction between these algorithms in Figure 5.

Correlations between prediction by different base algorithms are quite hight and varies in large range between 0.46 (LGBM and Random Forest) and 0.96 (Neutral Network and XGB) (see Figure 6).

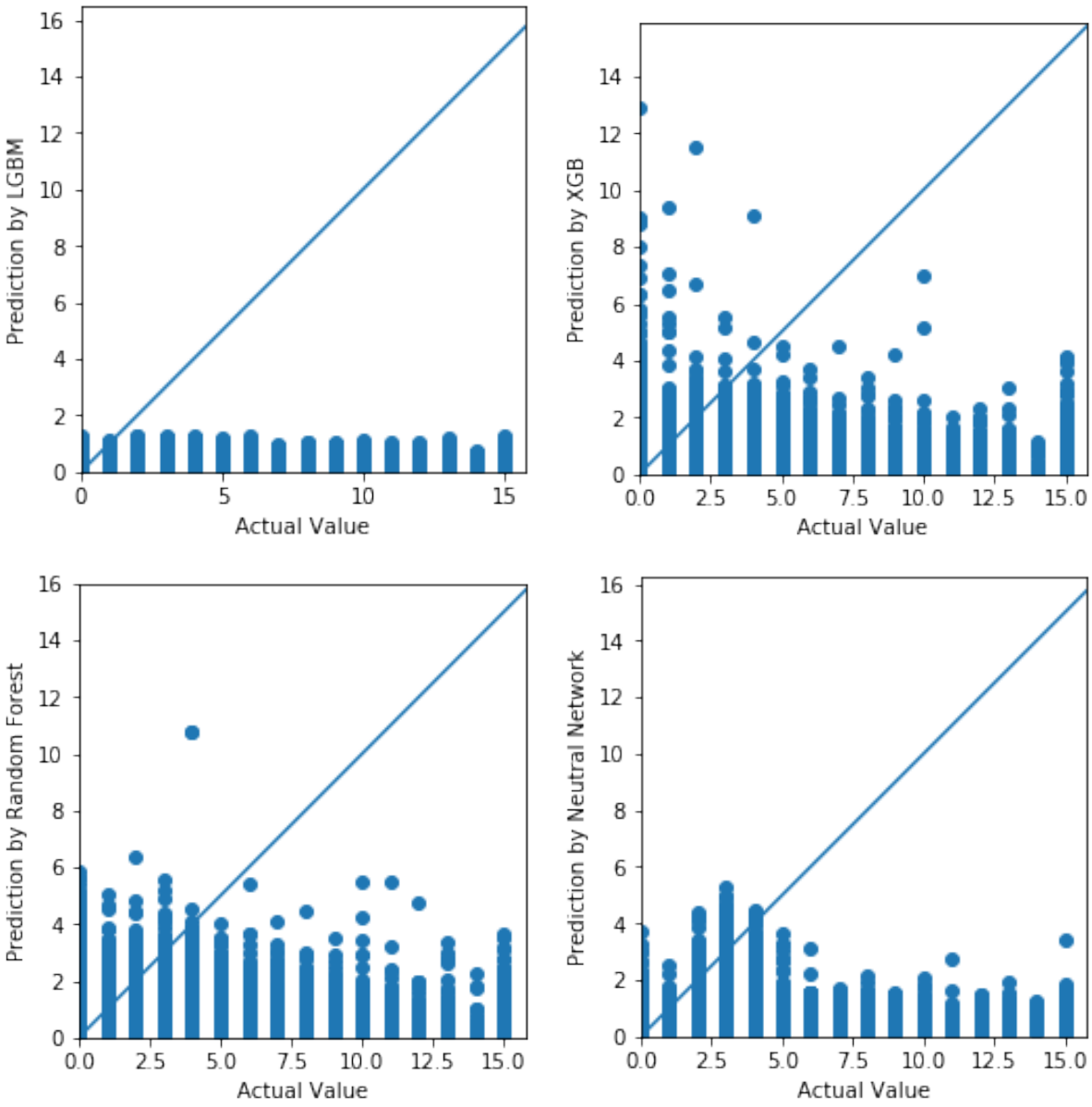

Figure 4: Comparison between actual and prediction DIH in Year 3

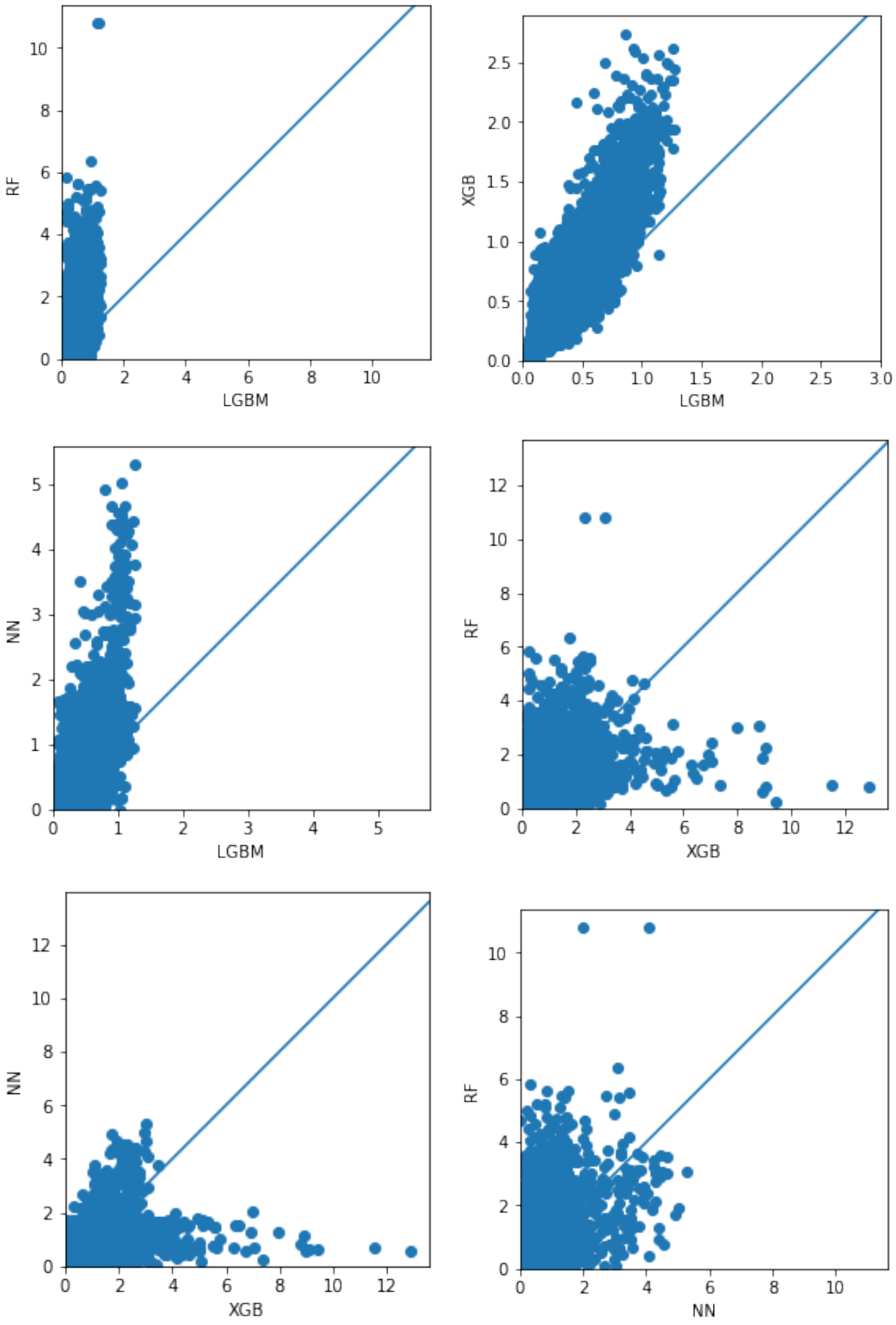

Figure 5: Comparison predictions between algorithms

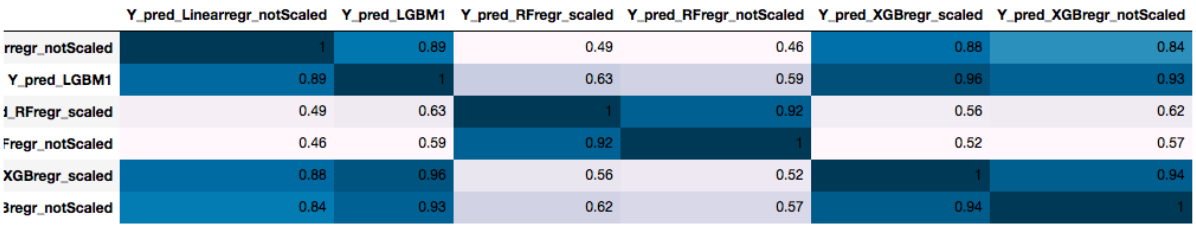

Figure 6: Heat map of base predictions

5 Result

Accuracy score by each model is displayed in Table 1. LGBM has the best performance with score at 0.446664 by tuning parameters n- estimator and depth-max. The worst one is Random Forest with score at 0.482489. Random Forest is also the most overfitting model. Tuning parameters on Random Forest has not been done.

Model	Dataset	Parameters	RMSLE
LGBM	without ProviderID and VendorID scaled input	boosting-type = 'dart', metric = 'rmse', n-estimators = 4500 num-leaves = 30 max-depth = 6, reg-alpha = 0.436193, reg-lambda = 0.479169, colsample-bytree = 0.508716, min-split-gain = 0.024766	0.446772
Linear Regression	not scale input without ProviderID, VendorID		0.4517539
XGBoost	scale input	default	0.448326
XGBoost	not scale input	default	0.448341071
Random forest	scaled input	default	0.482489
Random forest	not scaled input	default	0.488098
Neutral network	not scale input without ProviderID and VendorID	4 layers, 15, 10, 5 nodes in hidden layers	0.452469
Median			0.447077

Table 1: Result

All models are tuned manually by using hold - out dataset for validation. The computation cost is less than cross validation with RandomSearch or GridSearch. Furthermore, we get a better score with hold - out dataset.

In each model, we remove some features based on feature importance level and correlation coefficient between the features. Firstly, we remove all the feature with 0 important level in the model. Secondly, in the subgroup with high correlation (in this case, it is counter groups which contains the features No-Claims, No-Providers, No - Vendors, ...), we keep only the feature with highest importance level.

We've tried to ensemble by several methods such as linear regression, combinatorics, XGB but the aggregate result is worse than LGBM. The best ensemble prediction is based on median of base predictions at score 0.447077. It is still less accurate than LGBM prediction. Hence we propose the prediction obtained by LGBM as the final result.

One can see that the accuracy is not so good. To improve the score, one of a perspective is to deal with imbalanced data. Moreover, every model is train on full dataset up to now. Another approach is that one can divide full dataset into subset as in almost Milestone, for example, subset1 with information on sex and age, subset2 with information on DrugCount, subset3 on LabCount, subset4 on LengthOfStay, subset4 on PayDelays ... Hence ensemble works on base prediction which is trained on set of several subsets. Furthermore, AdminissionRisk, one factor obtained by domain knownledge in [1] Study the impact of AdminissionRisk on DIH is still not studied carefully. Several features have not created and considered in our preprocessing data such as Lab and Drug Velocity, pairwise combination

between PrimaryConditionGroup, Speciality and ProcedureGroup, allocation for highest probability of all Provider and Vendor, Primary Care Physical as in [1]. It should be studied more in the future. Concerning on missing data, they are typically replaced by 0 or -1. One of a remained problem is to study some method to replace missing data by more reasonable values.

References

[1] Phil Brierley, David Vogel, and Randy Axelrod. Our milestone 1 solution for heritage health prize. *How we did it, Team 'Market Makers'*.

[2] Willem Mestron. My milestone 1 solution for heritage health prize.

References

[1] Phil Brierley, David Vogel, and Randy Axelrod: *Our milestone 1 Solution for Heritage Health Prize*, How we did it - Team "Market Makers", 2011.

[2] Willem Mestron: *My milestone 1 Solution for Heritage Health Prize*, 2011.

Appendices

A Raw Data

The raw datasets are provided in Table 2, 3, 4, 5.

MemberID	Member pseudonym.
AgeAtFirstClaim	Age in years at the time of the first claim's date of service computed from the date of birth; Generalized into ten year age intervals.
Sex	Biological sex of member: M = Male; F=Female.

Table 2: Members Data

MemberID	Member pseudonym.
ProviderID	Member pseudonym.
Vendor	Member pseudonym.
PCP	Primary care physician pseudonym.
Year	Year in which the claim was made: Y1; Y2; Y3.
Specialty	Generalized specialty.
PlaceSvc	Generalized place of service.
	Member pseudonym.
PayDelay	Number of days delay between the date of service (the date the actual procedure was performed or service provided) and date of payment. Values above 161 days (the 95% percentile) are top-coded as "162+".
LengthOfStay	Length of stay (discharge date – admission date + 1), generalized to: days up to six days; (1-2] weeks; (2-4] weeks; (4-8] weeks; (8-12 weeks]; (12-26] weeks; more than 26 weeks (26+ weeks).
DSFS	Days since first claim, computed from the first claim for that member for
	each year, generalized to: [0-1] month, (1-2] months, (2-3] months, (3-4]
	months, (4-5] months, (5-6] months, (6-7] months, (7-8] months, (8-9]
	months, (9-10] months, (10-11] months, (11-12] months.
PrimaryConditionGroup	Broad diagnostic categories, based on the relative similarity of diseases and
	mortality rates, that generalize the primary diagnosis codes (ICD-9-CM) [2].
CharlsonIndex	A measure of the affect diseases have on overall illness, grouped by
	significance, that generalizes additional diagnoses. Scores greater than zero
	are carried forward (for up to a year) in subsequent claims with a comorbidity
	score of zero [1,4].
ProcedureGroup	Broad categories of procedures, grouped according to the hierarchical
	structure defined by the Current Procedural Terminology (CPT) [3].
SupLOS	Indicates if the NULL value for the LengthOfStay variable is due to
	suppression done during the de-identification process. A value of 1 indicates
	that suppression was applied.

Table 3: Claims (level 2) Data

MemberID	2-Member pseudonym
Year	Year in which the drug prescription was filled: Y1; Y2; Y3.
DSFS	Days since first service (or claim), computed from the first claim for that member for each year, generalized to: [0-1] month, (1-2] months, (2-3] months, (3-4] months, (4-5] months, (5-6] months, (6-7] months, (7-8] months, (8-9] months, (9-10] months, (10-11] months, (11-12] months.
LabCount	Count of unique laboratory and pathology tests by DSFS. Values above 9, the 95 % percentile after excluding counts of zero, are top-coded as "10+".

Table 4: LabCount

MemberID	Member pseudonym.
Year	Year in which the drug prescription was filled: Y1; Y2; Y3.
DSFS	Days since first service (or claim), computed from the first claim for that member for each year, generalized to: [0-1] month, (1-2] months, (2-3] months, (3-4] months, (4-5] months, (5-6] months, (6-7] months, (7-8] months, (8-9] months, (9-10] months, (10-11] months, (11-12] months.
DrugCount	Count of unique prescription drugs filled by DSFS. No count is provided if prescriptions were filled before DSFS zero. Values above 6, the 95% percentile after excluding counts of zero, are top-coded as "7+".

Table 5: DrugCount

MemberID	Member pseudonym.
DaysInHospital Y2	Days in hospital, the main outcome, for members with claims in Y1. Values above 14 days (the 99% percentile) are top-coded as "15+".
DaysInHospital Y3	Days in hospital, the main outcome, for members with claims in Y1. Values above 14 days (the 99% percentile) are top-coded as "15+".
ClaimedTruncated	Members with truncated claims in the year prior to the main outcome are assigned a value of 1, and 0 otherwise.

Table 6: Outcome Data

B Features

Full set of features is listed in Table 7.

Variable	Number of columns	Description
Age	10	0/1
Sex	3	0/1
NoClaims	1	Count
ClaimsTruncated	1	0/1 (not used)
Speciality	13	0/1, Count
PlaceSvc	9	Count
PrimaryConditionGroup	46	Count
ProcedureGroup	18	Count
SupLOS	1	0/1
LengthOfStayUnknow	1	0/1
LengthOfStayKnow	1	0/1
ProviderID	46	Count
VendorID	42	Count
DSFS min	1	Number
DSFS max	1	Number
DSFS ave	1	Number
DSFS range	1	Number
DSFS std	1	Number
CharlsonIndex min	1	Number
CharlsonIndex max	1	Number
CharlsonIndex ave	1	Number
CharlsonIndex range	1	Number
CharlsonIndex std	1	Number
NoDrug	1	Count
DrugNull	1	0/1
Drugs min	1	Number
Drugs max	1	Number
Drugs ave	1	Number
Drugs range	1	Number
Drugs std	1	Number
NoLabs	1	Number
LabNull	1	0/1
Labs min	1	Number
Labs max	1	Number
Labs ave	1	Number
Labs range	1	Number
Labs std	1	Number

Table 7: List of features