Sixu Meng

Prof. Ajay Anand, PhD

Final Capstone Project Report (URMC-oncology)

05 May 2019

PROJECT INTRODUCTION

Polypharmacy, the simultaneous use of multiple drugs to treat a single ailment or condition, is extremely common in older adults patients.[1] This way of treatment has been considered inappropriate in most situations. For this purpose, the Polypharmacy team assigned phase one of this project to us. The phase one was designing an automatic data cleaning algorithm to clean the data collected by the polypharmacy log forms in GAP dataset.

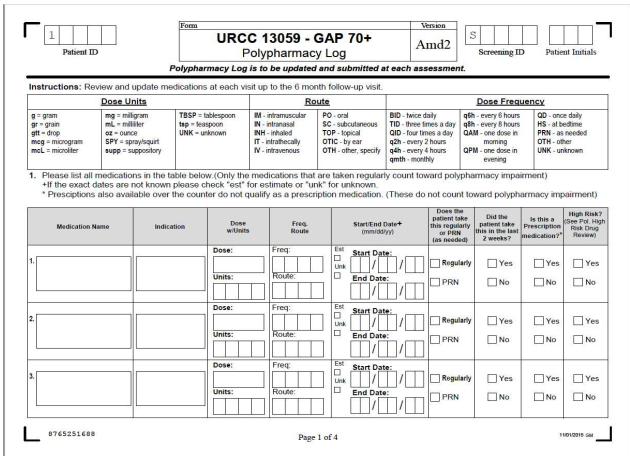
Polypharmacy also increases the risk of death and other bad things, and one of them is physical fall. Falls are common, under-recognized events in the lives of older cancer adults are the leading cause of traumatic mortality in this patient group. [2] Almost one in three older adults fall each year,[3] and about 10% of these falls result in injuries, [4] including, fractures, traumatic brain injuries, internal bleeding and so on; Besides, falls are normally related to fear of falling, compromised physical performance, declined cognitive function. Therefore, phase two of this project was building a model that successfully predict the falling among the ageing cancer population for both better diagnosis and prognosis.

Phase One: Automatic Data Cleaning Algorithm Design

DATA PROVENANCE

The Dataset is collected by the NCI Community Oncology Research Program which brings cancer research studies and results to patients in a variety of community setting across the United States by medical log forms. After manually filling the forms, the information was

scanned into the system. For this reason, our object one is to design an algorithm which can automatically correct the misspelt input medication names.



Patient ID

Form: URCC 13059 - GAP 70+ Polypharmacy Log

Version: Amd2

Screening ID

Patient Initials

Polypharmacy Log is to be updated and submitted at each assessment.

Instructions: Review and update medications at each visit up to the 6 month follow-up visit.

Dose Units			Route		Dose Frequency		
g = gram	mg = milligram	TBSP = tablespoon	IM - intramuscular	PO - oral	BID - twice daily	q6h - every 6 hours	QD - once daily
gr = grain	mL = milliliter	tsp = teaspoon	IN - intranasal	SC - subcutaneous	TID - three times a day	q8h - every 8 hours	HS - at bedtime
gtt = drop	oz = ounce	UNK = unknown	INH - inhaled	TOP - topical	QID - four times a day	QAM - one dose in morning	PRN - as needed
mcg = microgram	SPY = spray/squirt		IT - intrathecally	OTIC - by ear	q2h - every 2 hours	QPM - one dose in evening	OTH - other
mcL = microliter	supp = suppository		IV - intravenous	OTH - other, specify	q4h - every 4 hours		UNK - unknown
					qmth - monthly		

1. Please list all medications in the table below (Only the medications that are taken regularly count toward polypharmacy impairment)
   +If the exact dates are not known please check "est" for estimate or "unk" for unknown.
   * Prescriptions also available over the counter do not qualify as a prescription medication. (These do not count toward polypharmacy impairment)

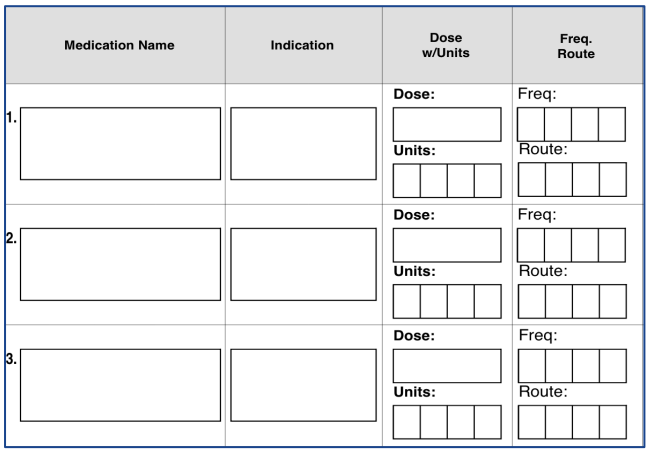
	Medication Name	Indication	Dose w/Units	Freq. Route	Start/End Date+ (mm/dd/yy)	Does the patient take this regularly or PRN (as needed)	Did the patient take this in the last 2 weeks?	Is this a Prescription medication?*	High Risk? (See Pol. High Risk Drug Review)
1.			Dose: Units:	Freq: Route:	Est. Unk. Start Date: End Date:	Regularly PRN	Yes No	Yes No	Yes No
2.			Dose: Units:	Freq: Route:	Est. Unk. Start Date: End Date:	Regularly PRN	Yes No	Yes No	Yes No
3.			Dose: Units:	Freq: Route:	Est. Unk. Start Date: End Date:	Regularly PRN	Yes No	Yes No	Yes No

8765251688

Page 1 of 4

11/01/2015 GM

Figure 1.1 Form for collecting polypharmacy log of patients



	Medication Name	Indication	Dose w/Units	Freq. Route
1.			Dose: Units:	Freq: Route:
2.			Dose: Units:	Freq: Route:
3.			Dose: Units:	Freq: Route:

Figure 1.2 Zoom in form of figure 1.1

DATA PREPARATION

Firstly, we corrected those misspelt medication names with a medication names dictionary. Secondly, we added the generic of each medication to our patients' records. Finally, we sent the number of generics each patient is taking to polypharmacy team for future clinical diagnosis.

Data Exploratory Analysis

The target dataset has two columns: patient ID and medication names. We also received three dictionaries for identifying the misspelt names and for the corresponding relation between medication names and generic names. One medication name might correspond to two or more generic names. Also, we have a dataset of 60 cancer patients, which has been cleaned by the polypharmacy team as a sample dataset for previewing purpose. We performed our data cleaning algorithm on this 60-patients-dataset to test the accuracy afterwards.

There are three types of error input in the target dataset: misspelt medication name, wrongly put the generic name into medication column, wrongly put generic name into medication column plus misspelt the generic name.

ALGORITHM DESIGN

A Python library named difflib [5] was used for words correction by alphabet frequency pattern. We apply this Algorithm only if target data's medication words cannot be found in the dictionaries.

Because of the second and third type of error mentioned in the Data Exploratory Analysis paragraph, we first started the matching on generic names. The "Directly Matching" means exact word found in generic column of the dictionary. The "Algorithm Correction" means misspelling detected but successfully corrected by the algorithm. The "Unknown Value" means not found or matched and will be sent for medication matching.

Directly matching	Algorithm correction	Unknown value
2703	2541	2423

Figure 1.3 Summary of Algorithm correction of generic name

Secondly, we run the matching algorithm on the medication name to fix the first type of error. The "Total Correction Number" represents the total number of words corrected so far. The "Directed matching" means exact word found in the medication column of the dictionary. The "Algorithm Correction" means the number of words corrected by the algorithm on medication names matching. The "Unknown Value" means not found or matched by both steps. The "Warning" means the first letter of the original words is not the same as the algorithm generated words.

Because of the limitation from the number of medication names contained in the dictionary, the unknown value will need to be sent back to polypharmacy team for manually cleaning. However, the automatic algorithm correction has successfully and efficiently cleaned the 87.5% of the dataset which can save our clinician's time and other resources.

ACCURACY TESTING

	Total Correction Number	Directed matching	Algorithm correction	Unknow value	Warnings
Times	6709	998	467	958	114

Figure 1.4 Summary of Algorithm correction of medication matching

Algorithm accuracy testing was performed at 0.9 cutoff rate on the 60-patient-dataset having 496 rows. According to the summary report, 192 rows were exactly matched with the sample's answer. However, 83 rows generated not only contains the sample answers but also have other generic names included. Since one medication can belong to multiple generic, we

	testing_misspelled_names	Correction	new_generic_list	ans	new_mark
0	AMLODIPINE	AMLODIPINE	[AMLODIPINE]	[AMLODIPINE]	Gen_Found
1	ASPIRIN	ASPIRIN	[ASPIRIN]	[ASPIRIN]	Gen_Found
2	FINASTERIDE	FINASTERIDE	[FINASTERIDE]	[FINASTERIDE]	Gen_Found
3	GABAPENTIN	GABAPENTIN	[GABAPENTIN]	[GABAPENTIN]	Gen_Found
4	GLIMEPIRIDE	GLIMEPIRIDE	[GLIMEPIRIDE]	[GLIMEPIRIDE]	Gen_Found

Figure 1.5 Part of the result of testing

considered these rows as correct matching with more detailed result. Thus, the algorithm correction accuracy is 94.295% of the 298 corrections.

Phase Two: Machine Learning-Based Falling Rate Prediction Model in Aging Cancer Patients

BACKGROUND INTRODUCTION

For the second part of this project, we have developed a hypothesis on the falling problem in older cancer patients. Falls are one of the leading causes of injuries in the ageing people; 30% ageing people over 65 are the victim of falls (Todd and Skelton, 2004)[6]; 10% of falls lead to serious damage (Goldacre et al., 2002)[7]. Previous researches outline the correlation between falling and both physical and cognitive function in ageing people. Our question developed on this topic is: Can we use cognitive performance, physical performance, cancer type, biomarker, social support and demographic information to predict how likely a patient is going to fall.

In the literature reviewing process, we found that well-studied predictors of falls include: age, dementia, physical performance, frailty, visual and musculoskeletal disorder, and medications such as psychotropic drugs. (Lastrucci et al., 2017)[8] and the measurements include Frontal Assessment Battery (Kataoka, 2015)[9], Short Physical Performance Battery(Pua, 2018), Fried Frailty Index Components (Sharma et al., 2019)[10], etc. However, to the best of our knowledge, no classification model of falling has been published up to the date using multiple predictors.

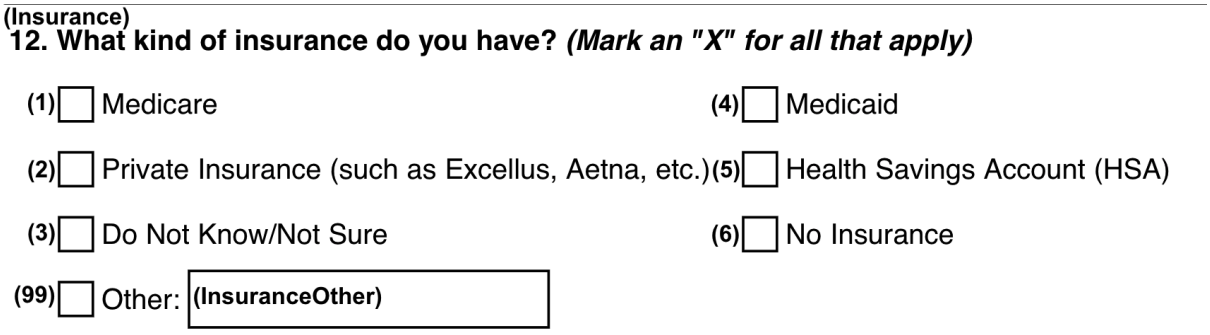
Whether falling frequency differences between different age and gender group? Whether patients taking psychotropic meds show a higher falling rate compared to the control patients. Is there any dimorphic effect of psychotropic medication on patients? We certainly hope these questions can be answered by the end of this project.

DATA PROVENANCE

All participants (N = 755) were recruited from Wilmot Cancer Center and Highland Hospital by geriatric polypharmacy team. Patients needed to be over age 75 and diagnosed with

certain types of cancers to be qualified for this study. Demographic information, financial and social status, and clinical data were collected for data analysis. Major clinical features related to the falls, including IADL, SBBS, MiniCog, GDS were also collected in this dataset. There are 77 features in total, which contains both numerical and categorical data. 35 features were initially selected based on less than 15 missing value.

For the numerical data, we select the mean to fill up the continuous missing data and median to fill up the discrete missing data


(Insurance)
12. What kind of insurance do you have? *(Mark an "X" for all that apply)*

(1) ☐ Medicare
(4) ☐ Medicaid
(2) ☐ Private Insurance (such as Excellus, Aetna, etc.)
(5) ☐ Health Savings Account (HSA)
(3) ☐ Do Not Know/Not Sure
(6) ☐ No Insurance
(99) ☐ Other: (InsuranceOther)

Figure 2.1 Question from GAP log from on insurance

For categorical data, we distributed the data by classes frequency. Take insurance as an example, patients are allowed to select multiple answers on the log form. A classes frequency histogram was created for visualization. Top five classes were selected and others were merged into another class.

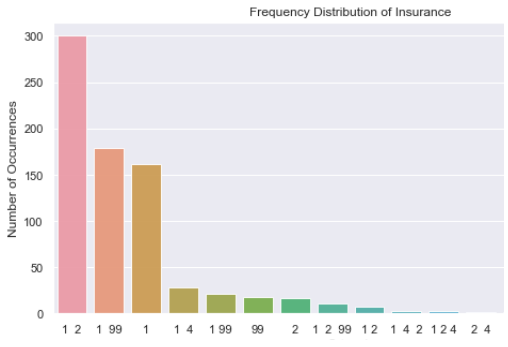


Figure 2.2 Distribution of patient's insurance selection

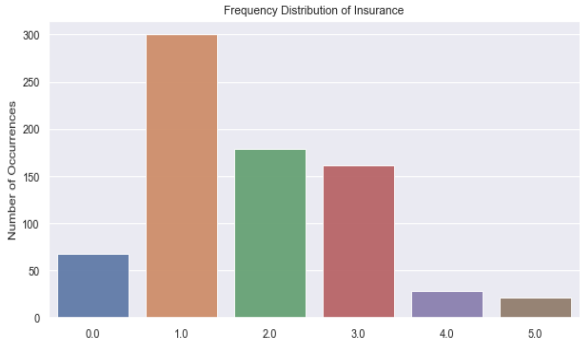


Figure 2.3 Distribution of insurance selection after merging

DATA PREPROCESSING

The dimensionality of this dataset is very high since we selected 35 features initially. We performed feature selection not only reducing the number of features but also improving the prediction result.

Feature Selection

RFECV (Recursively feature elimination using cross validation)[11] was firstly introduced in our feature selection. It is a method which is widely used in data analysis with a type of machine learning model. The model can help with classification on the training dataset using cross-validation. Furthermore, this algorithm can rank the feature importance by returning the number of features needed at best accuracy score.

```
from sklearn.feature_selection import RFECV
from sklearn.ensemble import RandomForestClassifier

# The "accuracay" scoring is proportional to the number of corr
clf_rf_4 = RandomForestClassifier()
rfecv = RFECV(estimator=clf_rf_4, step=1, cv=5,scoring='accurac
rfecv = rfecv.fit(X_train, y_train)

print('Optimal number of features :', rfecv.n_features_)
print('Best features :', X_train.columns[rfecv.support_])
```

```
Optimal number of features : 13
Best features : Index(['Grade', 'Marital', 'Live', 'Health',
'Age', 'Feel', 'Insurance',
       'Income', 'Living', 'FH2', 'FH3', 'KPS', 'cancertyp
e'],
      dtype='object')
```

Figure 2.4 Algorithm and output from RFECV

Figure 2.5 Graph of number of features vs. AUC

The AUC score goes up and down between 0.75 and 0.80 as the number of features selected increases to 8 and the optimal number of features is 13.

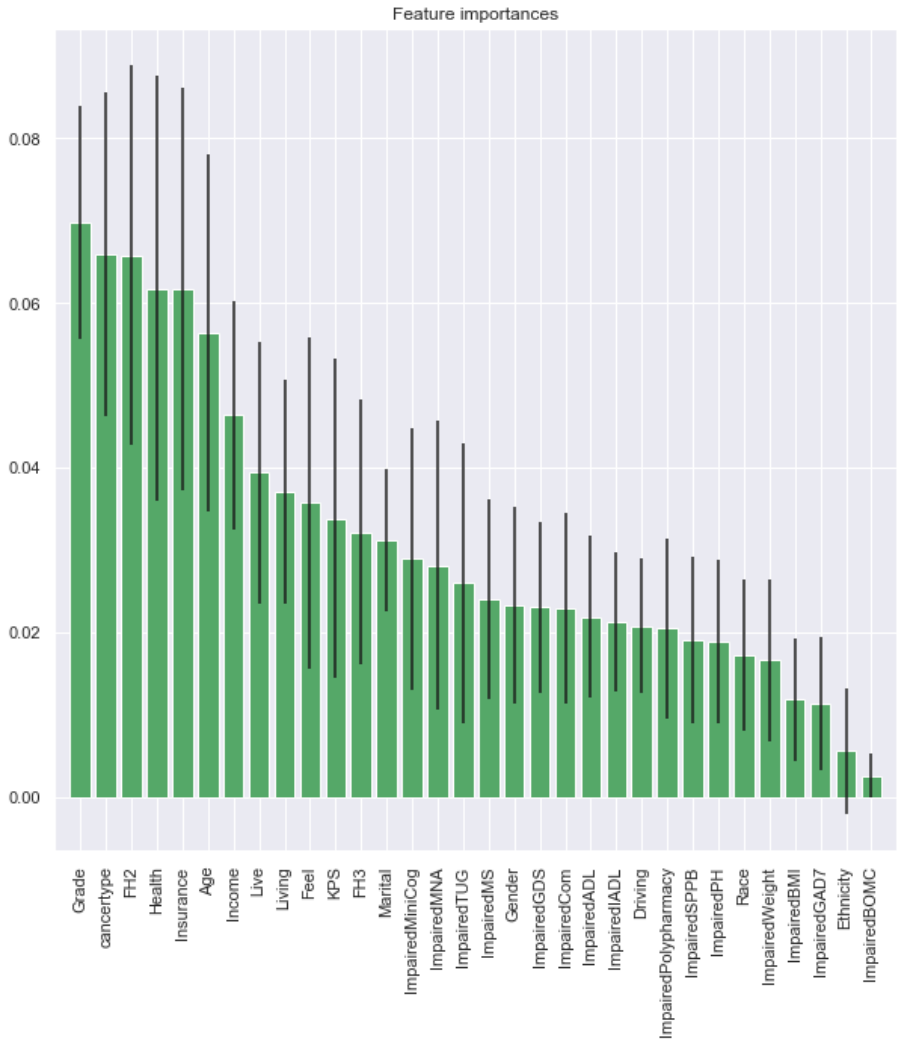


Figure 2.6 Graph of features importance

We also implemented the Tree-based feature selection algorithm which is a feature elimination algorithm based on the feature importance. Feature_importances, a parameter from Sklearn package, is used in this classifier. The top features were selected by multiple times of selection based on these two methods above.

In addition to that, the confusion matrix was also used for feature elimination. In Figure 2.7, we can see that FH2 and FH3 have a correlation of 0.5. We consider there might be some correlation with this ratio in the clinical area. Thus, we keep one of them in our further modeling.

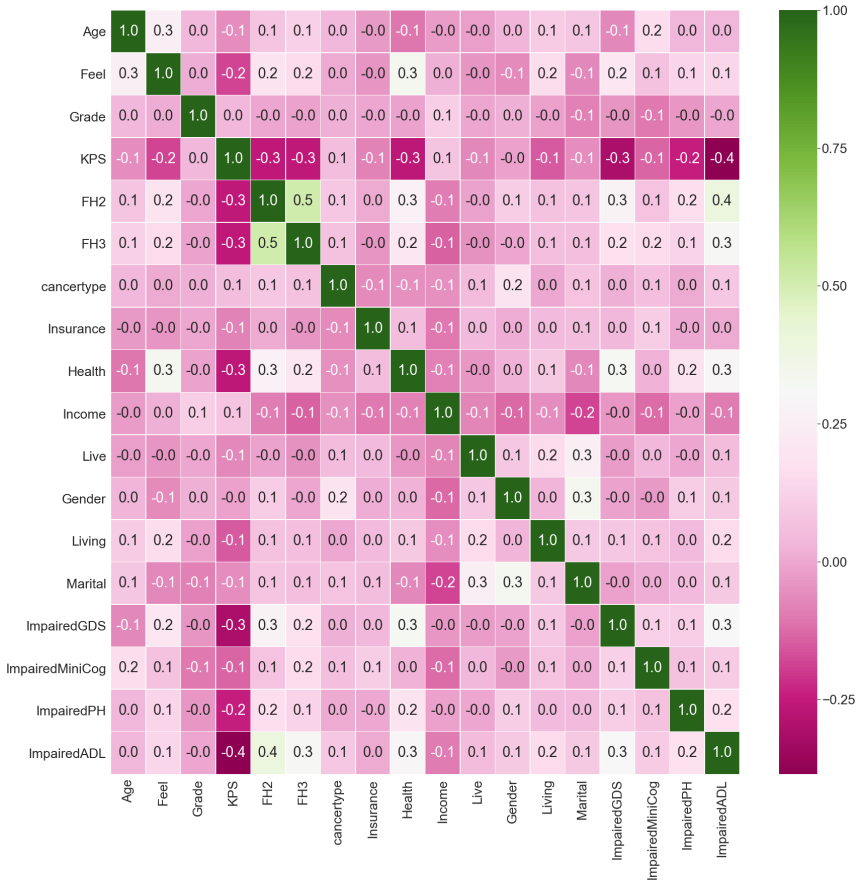


Figure 2.7 Part of the feature's confusion matrix

Data Resampling

The dataset is imbalanced on the falling result. We performed several re-sampling techniques which balanced the training set for a reasonable result.

SMOTE is an over-sampling technique which is one of the most commonly used resampling algorithms for clinical data. SMOTE consists of synthesizing elements for the minority class, based on those already existing data. It will randomly pick a point from the minority class and computing the k-nearest neighbors for this point. In the end, The synthetic points are added between the chosen point and its neighbors.

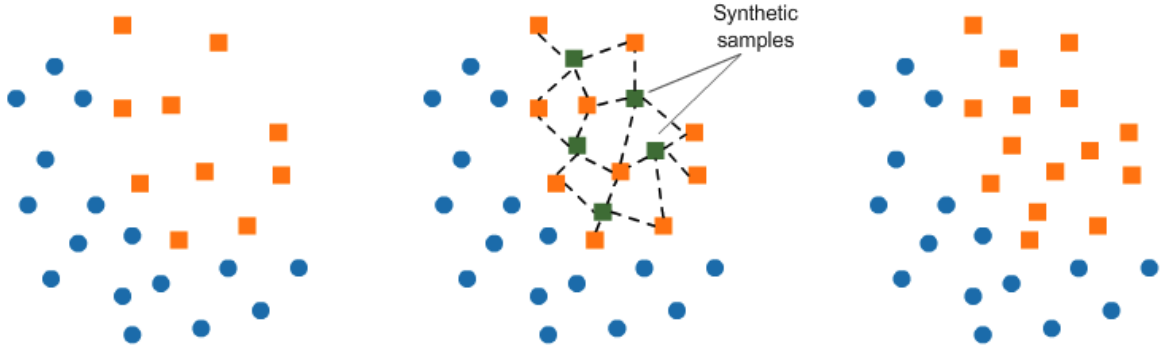


Figure 2.8 Illustration of SMOTE

The under-sampling technique we picked to use with SMOTE is called Tomek. The Tomek Links will remove unwanted overplay between classes where majority class links are removed until all minimally distanced nearest neighbor are of same class. After using Tomek Links, it will be easier for machine learning algorithms to distinguish the difference between class A and class B.

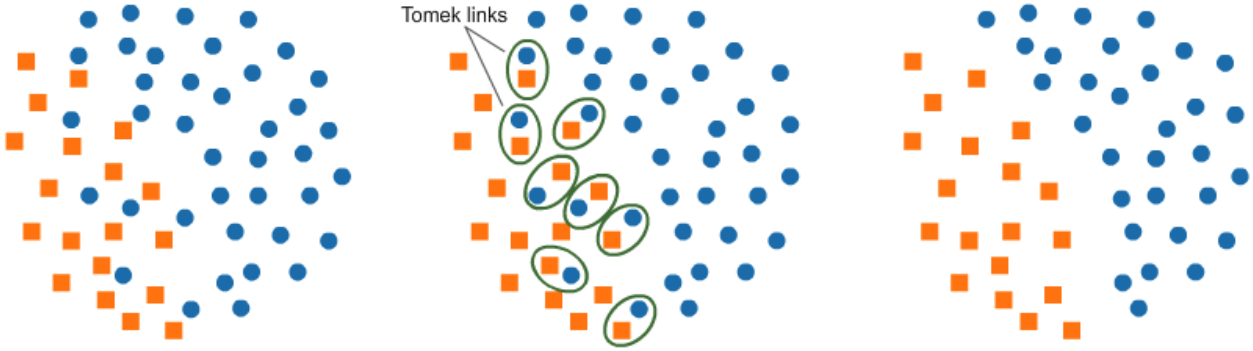


Figure 2.9 Illustration of Tomek

In the end, we performed SMOTE first to generate more samples in minority class, and then, we use Tomek Links to clean overlaps between majority and minority classes.

In addition, we implement another technique for purely undersampling called Instance Hardness Threshold. Instance hardness threshold uses an estimator to estimate the instance hardness of the samples. It reduces less important data points in the dataset to perform under-

sampling.

MODEL DEVELOPMENT

Traditional statistical methods have limitations when dealing with medical data. Machine Learning can introduce more predictive power than traditional statistical tools like

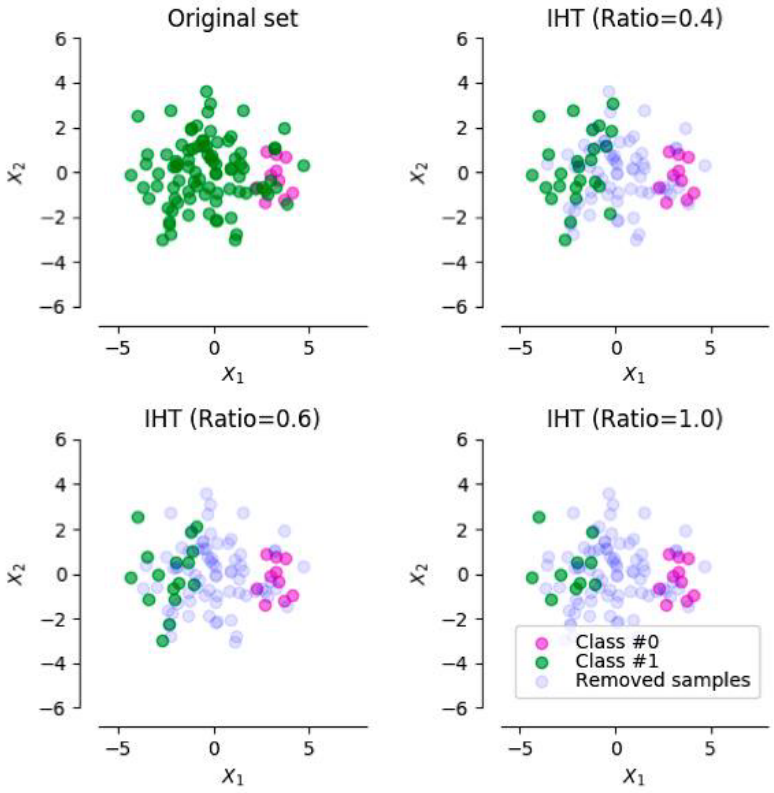


Figure 2.10 Illustration of Instance Hardness Threshold

regression and more capability of accommodating non-linear dataset. Machine Learning technique has Better performance for high-dimensional dataset compared to regression and most importantly Machine Learning supports a large volume, and a wide type of clinical research data such as Genome Data, MRI Data, CT Data, Lab Test, and even medical chart text data.

In the medical area, people care a lot about true-positive and false-negative rather than false-positive and true-negative. High precision when predicting that patients are going to fall is preferred rather than high precision when predicting that patients are not going to fall.

We built our model based on Three types of machine learning model: Logistic

$$\text{Recall} = \frac{tp}{tp + fn}$$

Figure 2.11 Formula of recall

Regression[12], Random Forest[13], and MLP neural network[14].

PERFORMANCE AND RESULTS

The performance was tested on 100 times of resampling and training-testing split.

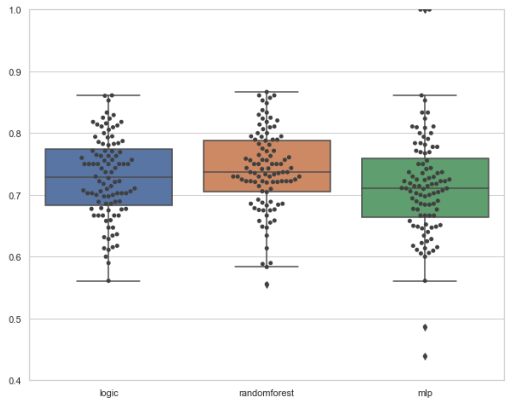


Figure 2.12 box-plot of recall ratio

	logic	randomforest	mlp
count	100.000000	100.000000	100.000000
mean	0.728111	0.739756	0.708105
std	0.065841	0.064784	0.113084
min	0.560976	0.555556	0.000000
25%	0.683929	0.705548	0.664474
50%	0.728501	0.736842	0.710819
75%	0.773693	0.788278	0.758527
max	0.861111	0.866667	1.000000

Figure 2.13 Data summary of recall ratio

The box-plot above reported the three machine learning model's recall ratio of the prediction to testing dataset. In other words, we recorded the probability of predicting falling patient among the actual falling patient. The data frame on the right shows the standard deviation of each model's recall ratio. The random forest has the best overall ratio with the lowest standard deviation and the highest mean value.

At this stage, a practical clinic prediction tool was brought up according to these methods above. The prediction tool was designed to be capable of sending a probability ratio

cancertype	Health	Grade	Age	FH2	Live	KPS	Living	Driving	Feel	Marital	ImpairedMNA	ImpairedMiniCog	ImpairedIADL	ImpairedGDS	Falling	Probability
2	2.0	7.0	1	1.0	1.0	3	1.0	1.0	2	4.0	0	0	0	0	0	0.242482
2	5.0	1.0	0	3.0	2.0	2	2.0	0.0	1	5.0	1	0	1	1	1	0.932660
2	3.0	7.0	0	2.0	1.0	3	2.0	1.0	2	2.0	0	0	1	0	1	0.559271
6	4.0	2.0	0	1.0	1.0	2	2.0	1.0	2	6.0	1	0	1	0	0	0.448995
1	5.0	5.0	1	2.0	1.0	3	1.0	1.0	2	2.0	1	1	1	1	1	0.823329

Figure 2.14 Data Frame demo of falling probability ratio on patient

of falling by analyzing each patient's data. Logistic Regression CV[15] and its parameter predict_proba, which returns the patient's probability ratio of belonging to each class, was implemented. In this binary classification task, 50% was used as the threshold for classification. However, a higher threshold will be needed for more certainty in the clinically study.

CONCLUSION

To the best of our knowledge, this is the first study that used machine learning algorithms like random forest and MLP for predicting the falling rate among the ageing cancer population. Meanwhile, our model performance is much better compared to most of the previous studies using other predictive models like regression model.

Our study shows that machine learning has large predictive power and potential in the future clinical study. Meanwhile, we would want to optimize the model by further improving its specificity in the future.

In this capstone project, I have engaged in algorithm development and presentation preparation for every step. We were coached by Professor Ajay who has guided us from day one. I appreciate his wisdom and problem-solving technique that helped us achieved this goal. I have collaborated with a professional clinical team and three best data science students. I sincerely felt that it is a great pleasure and to work with Dr.Ramsdale and her team and my appreciation of each's efforts and contributions to this project is beyond my words.

WORK CITED

1. URMC-oncology presentation slides on Blackboard. URL: https://learn.rochester.edu/webapps/blackboard/execute/content/file?cmd=view&content_id=_1768437_1&course_id=_38661_1&framesetWrapped=true (Accessed 3/8/2019)
2. CDC. Fatalities and Injuries from Falls Among Older Adults -- United States, 1993-2003 and 2001-2005. Morbidity and Mortality Weekly Report. 2006; 55(45):1221–4. [PubMed: 17108890]
3. Ganz DA, Bao Y, Shekelle PG, Rubenstein LZ. Will my patient fall? JAMA: The Journal of the American Medical Association. 2007; 297(1):77–86.
4. Tinetti ME, Williams CS. Falls, injuries due to falls, and the risk of admission to a nursing home. N Engl J Med. 1997; 337(18):1279–84. [PubMed: 9345078]
5. Wes McKinney. Data Structures for Statistical Computing in Python, Proceedings of the 9th Python in Science Conference, 51-56 (2010). URL: https://docs.python.org/2/library/difflib.html
6. Skelton, D. A., et al. "Prevention of Falls Network Europe: a Thematic Network Aimed at Introducing Good Practice in Effective Falls Prevention across Europe." *European Journal of Ageing*, vol. 1, no. 1, 2004, pp. 89–94., doi:10.1007/s10433-004-0008-z.
7. Tinetti, Mary E. "Factors Associated with Serious Injury During Falls by Ambulatory Nursing Home Residents." Journal of the American Geriatrics Society, vol. 35, no. 7, 1987, pp. 644–648., doi:10.1111/j.1532-5415.1987.tb04341.x.
8. Lastrucci, Vieri, et al. "Identification of Fall Predictors in the Active Elderly Population from the Routine Medical Records of General Practitioners." Primary Health Care Research & Development, vol. 19, no. 02, 2017, pp. 131–139., doi:10.1017/s146342361700055x.
9. Kataoka, Hiroshi, and Satoshi Ueno. "Low FAB Score as a Predictor of Future Falling in Patients with Parkinson's Disease: a 2.5-Year Prospective Study." Journal of Neurology, vol. 262, no. 9, 2015, pp. 2049–2055., doi:10.1007/s00415-015-7814-4.
10. Sharma, Anjali, et al. "Frailty as a Predictor of Falls in HIV-Infected and Uninfected Women." Antiviral Therapy, 2019, doi:10.3851/imp3286.
11. Sklearn Package RFECV. URL: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFECV.html
12. Sklearn Package LogisticRegressionClassifier. URL: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
13. Sklearn Package RandomForestClassifier. URL: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
14. Sklearn Package neural_network.MLPClassifier. URL: https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html
15. Sklearn Package Logistic Regression CV. URL: https://scikit-

learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegressionCV.html