URMC Caregiver Project Report

Chenwei Wu, Haosong Rao, Xuening Zhang, Meizhu Wang
Team 16, DSC383W, Prof. Anand
{cwu59, hrao, xzhang97}@u.rochester.edu, mwang74@ur.rochester.edu
University of Rochester
Sponsor: URMC Geriatric Oncology Team lead by Dr. Xu & Dr. Ramsdale

Tables of Contents

1. INTRODUCTION

This project report illustrates how 1) the drug names with generics and classes were labeled, and 2) how the Machine Learning techniques were used to predict the quality of life (QOL) of caregivers of older patients with advanced cancer. This was conducted as the final project of DSC383W at the University of Rochester in Spring 2020.

For part 1, we designed an algorithm that cleans the raw dataset of drug names and labels drugs with generics and classes using the existing reference dataset. Drugs name labeling can consume a lot of time if done manually and computer algorithms can speed up the process. However, most of the drug names do not have direct matches in the reference data and spelling errors exist a lot. We were able to overcome these challenges using fuzzy matching and association rules generation and develop an automation algorithm that can be potentially extended to other datasets.

For part 2, we developed predictive models for caregiver outcomes. Caregivers are people who help another person who can no longer be able to perform daily tasks necessary for everyday survival alone. Old people with cancers depends on caregivers to support them at every stage of the illness trajectory [1]. However, taking care of others itself can also adversely impact caregivers' emotional and physical well-being, which is a phenomenon known as the caregiver burnout. Caregivers might face significant level of stress when they do not meet the expectations. About 40% to 70% of caregivers suffer from depression, while many others also have anxiety and distress as a result of pressure [2]. These long-term stresses are harmful for health. Hence it is crucial for clinicians to identify and predict what specific groups of caregivers will have lower QOL and potentially prevent the frustration.

In this clinical setting, six QOL outcome variables for caregivers were derived from two questionnaires: one with Generalized Anxiety Disorder-7, Distress Thermometer, and Patient Health, the other with Short-Form Health (SF12) survey. Previous literature has used logistic regression models for prediction and analysis of the contributing factors to the deterioration of caregiver QOL. We implemented, fine-tuned, and evaluated multiple machine learning models for each of the 6 outcome variables and compared their results with Logistic Regression and saw a significant improvement. The major challenge here is to deal with the data imbalance issue and optimize model performance. A new technique called Local Interpretable Model-agnostic Explanation (LIME) was also used to study the positive or negative effects of our predictors on the caregiver QOL outcome variables.

2. DATA PREPROCESSING

2.1 Dataset for Labelling Cancer Treatments

There is one reference dataset that contains both drug names and their labeled classes and generics, and another raw dataset with only drug names and our task is to label them with proper class and generics.

The reference dataset contains 85 rows of data, each of which represents the cancer treatments/drugs (from 1 to 3 drugs for each patient) for patients, while the raw dataset has 541 rows (number of patients) of drug names that need to be labeled. Out of these 541 rows, there are

368 non-empty rows and here we simply ignore the empty ones because there are no drugs to be labeled for those patients.

2.2 Dataset for Caregiver QOL Prediction

This dataset is provided by Prof. Huiwen Xu and contains information of 541 patients and their 414 caregivers. Each row represents one caregiver Figure 2.2.1 shows the data dictionary.

1	id	Num	8			Patient ID
2	StudyArm	Char	7	$7.	$7.	Studyarm: GA, Control
3	Age	Num	8	11.	11.	How old are you?
4	Gender	Char	1	$1.	$1.	Gender: 1=male, 0=female
5	gradecat	Num	8			grade by category: 1= < high school, 2=HS, 3= >HS
6	incomecat	Num	8			Income: 1= <=50000, 2= >50000, 3=Decline to answer run
7	racecat	Num	8			Race by category: 1=Non-Hispanic White, 2=Black, 3=others
8	Living	Char	1	$1.	$1.	1=Independent Living (More than 1 story), 2=Independent Living (1 story), 3=Others
9	relationship_cat	Num	8			Relationship: 1=Spouse and Cohabiting partner, 2=Son/Daughter, 3=Others
10	GI_Lung	Num	8			Cancer type (1-GI, 2-Lung, 3-Other) new-coded
11	TTC3	Num	8	11.	11.	Cancer stage: 1=stage 3, 2=stage4, 3=other
12	TTC5	Num	8	NOYESF.	11.	Chemotherapy
13	TTC6a	Num	8	NOYESF.	11.	Monoclonal antibodies
14	TTC6b	Num	8	NOYESF.	11.	Hormonal treatment
15	TTC6c	Num	8	NOYESF.	11.	oral cancer treatment
16	TTC6d	Num	8	NOYESF.	11.	Radiation therapy

17	distress	Num	8			Patient distress (>=4)
18	ImpairedPolypharmacy	Num	8	NOYESF.	11.	Impaired Polypharmacy
19	cognition	Num	8	NOYESF.		Impaired cognition
20	nutrition	Num	8	NOYESF.		Impaired nutrition
21	phy_performance	Num	8	NOYESF.		Impaired physicial performance
22	function	Num	8	NOYESF.		Impaired function status
23	ImpairedCom	Num	8	NOYESF.	11.	Impaired Comorbidity
24	psychological	Num	8	NOYESF.		Impaired psychological
25	ImpairedMS	Num	8	NOYESF.	11.	Impaired Medical Social Support
26	calcimpairedCom	Num	8			CG comorbidity
27	cgdistress	Num	8			Caregiver Impaired distress (>=4)
28	cggad7	Num	8			Caregiver impaired GAD7 (>=5)
29	cgphq2	Num	8			Caregiver impaired PHQ2 (>=2)
30	SF12total	Num	8			SF12 total score
31	AGG_PHYS	Num	8			PHYSICAL HEALTH TOTAL SCORE - SF12
32	AGG_MENT	Num	8			MENTAL HEALTH TOTAL SCORE - SF12

Figure 2.2.1: Data Dictionary for patients' and caregivers' information

Columns 3 to 9 are demographic information about older patients. All the older patients completed the Geriatric Assessment (GA). GA is a set of tools to evaluate a patient's fitness for cancer treatment and can help to develop a treatment plan [2]. It has a variety of domains that assess a patient's overall health condition, including physical function, comorbidity and polypharmacy, nutrition, cognitive function, social support, and psychological status. Columns 18 to 25 are the eight domains used for patients. The patient who has each of the symptoms is categorized as 1, otherwise 0;

Columns 27 to 32 are the six outcome variables for caregivers' evaluations of QOL. There are three dummy variables and three categorical variables:

- Caregiver Distress (*cgdistress*): ≥ 4 on distress thermometer is categorized as 1, otherwise 0
- Caregiver Anxiety (*cgphq2*): ≥ 5 on Generalized Anxiety Disorder - 7 is categorized as 1, otherwise 0
- Caregiver Depression (*cggad7*): ≥ 2 on Patient Health Questionnaire - 2 is categorized as 1, otherwise 0
- SF12total(continuous): Total score of SF12, which is a 12-question survey to provide a generic measure of the overall health-related quality of life [3]. It can be decomposed into two scales below, and higher scores indicate higher levels of health.
- AGG_PHYS(continuous): Total physical component score
- AGG_MENT(continuous): Total mental component score

There are few missing values in the predictors (variable 1-26) and the outcome variables. Here missing values in the outcome variables simply mean the patient did not take the survey. Thus we removed those 34 rows with missing data.

Since all the categorical attributes were previously labelled as integer values, which may result in poor performance, we applied one-hot encoding instead. It is a technique that removes the integer encoded variable and substitutes a new binary variable for each unique integer value. We also dropped the first column value to avoid multicollinearity. For example, "gradecat" has 3 categories and therefore encoded into 2 binary variables. When a patient belongs to category 1, the two binary variables are 0; either of the two binary variables is 1 when the patient belongs to that category.

Due to fact that outcome variables have correlations with each other, we can not include any other as a regressor during our prediction. Thus, we made 6 different datasets for 6 outcome variables, each using variable 1-26 as regressors and 1 of the 6 QOL variables mentioned above as outcome variable. Here our task is to build classification for the 3 categoricals and regression models for the 3 continuous. Later in Part 4, we will talk about why and how we transformed the regression tasks also into classifications.

3. EXPLORATORY ANALYSIS

3.1 Exploring Types of Labeling Tasks for Cancer Treatment/Drug

With data exploratory analysis, we broke down the labeling task for the cancer treatments/drugs into four types of mini tasks: 1) Direct matches of cancer treatments; 2) Generation of rules of associations; 3) Fuzzy text matching; 4) Manual labeling with help of clinicians.

Type 1: Some of the drug records (100 rows) in the raw dataset can be directly matched with the reference dataset. For example, one of the drugs we need to label in the raw dataset is Folfox-6, while there is also a row of "Folfox-6" in the reference data. We will directly match Type 1 data.

Type 2: Some of the drug records (170 rows) in the raw datasets do not have direct matches but their labels can be inferred by some rules of associations. For example, we have "Abraxane and Gemzar" and "Abraxane and Carboplatin" labeled but need to get the label of "Abraxane". We need to infer the classes and generics for Type 2 data.

Type 3: Some of the drug records (58 rows) have spelling errors. For example, "Docetaxel" is sometimes misspelled as "Docetaxill". We need to apply some fuzzy text-matching for Type 3.

Type 4: The rest of the drug records (40 rows) doesn't have any information we can refer to, like "R CROP". We need to label Type 4 with the help of clinicians Amita and Mustafa.

3.2 Exploring the Distribution of Information of Patients and Caregivers

Before building predictive models, it is necessary to take a closer look at the data. This section provides analysis for both patients and caregivers. Also, the issue of imbalance data is present and the solution will be addressed in the next part of the report.

3.2.1 Distribution of Impairments of Patients

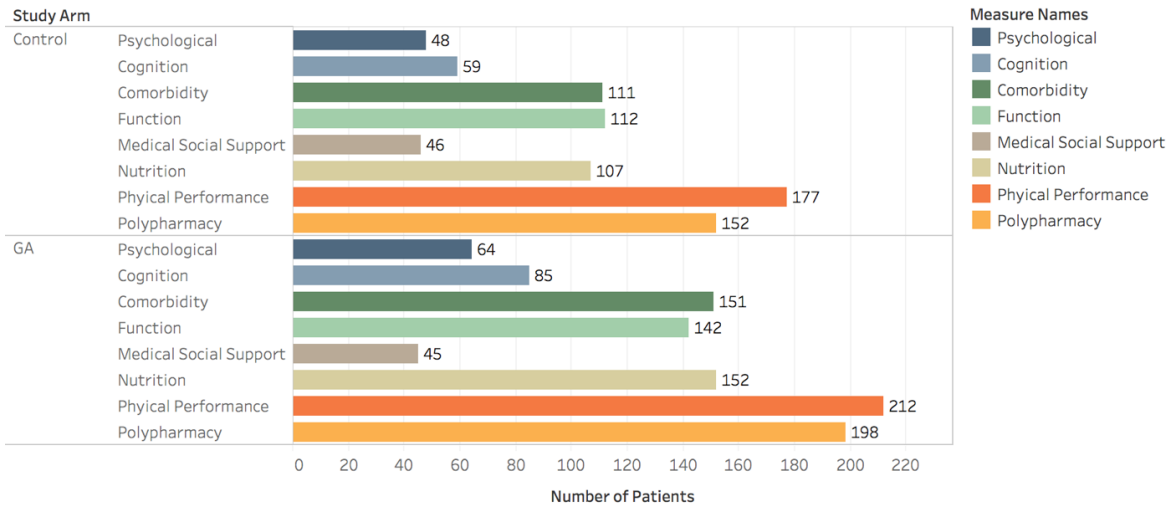


Figure 3.1.1: Distribution of Impairments

Figure 3.1.1 describes the prevalence of eight impairments identified by older patients. When a patient meets the cutoff point in the GA of one measure, he/she is considered as impaired. In both study arms where patients receive the same treatment within one arm, majority of patients are experiencing Physical Performance and Polypharmacy impairments. While medical social support takes the least proportion.

3.2.2. Distribution of Caregiver Emotional Health Outcomes

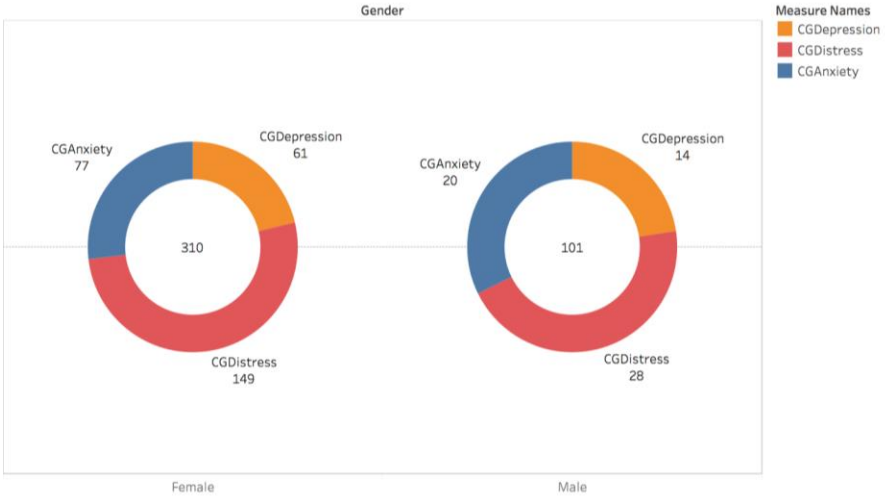


Figure 3.2.2: Distribution of Emotional Health Outcomes by Gender

Figure 3.2.2 provides a donut-chart overview of three categorical outcome variables. The number of female patients is three times the male patients. Caregiver Depression is the most serious mental health problem because more than 40% of them report it, especially female ones. We want to explore more about each of them accordingly.

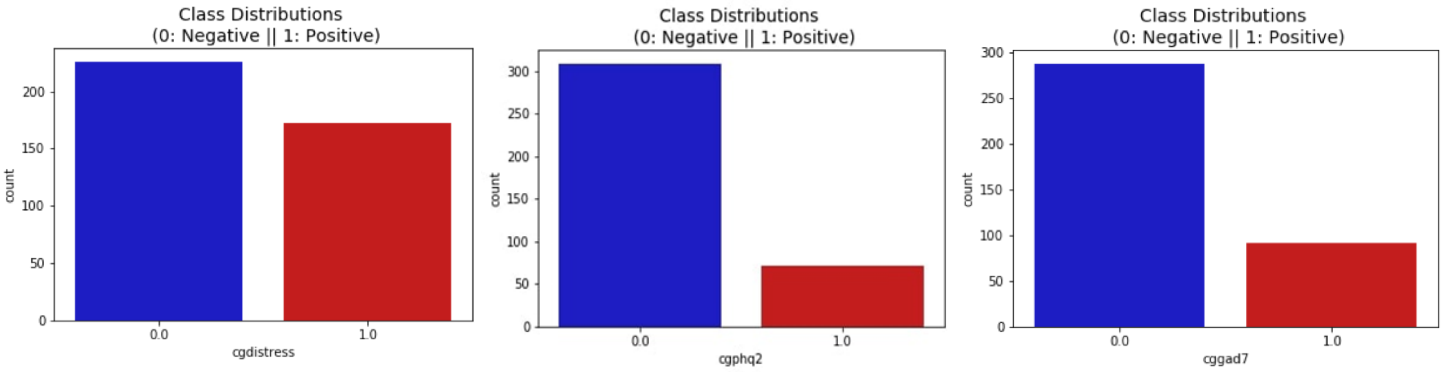


Figure 3.2.3: Distribution of three categorical outcome variables

After a series of preprocessing procedures (drop missing values, one-hot encoding), the class distributions of three emotional outcomes are shown in Figure 3.2.3. *cgdistress* has 226 zeros (distress thermometer<4) and 172 ones (distress thermometer≥ 4), *Cggad7* has 300 zeros (anxiety disorder<5) and 97 ones (anxiety disorder≥5), *Cgphq2* has 309 zeros (Patient Health Questionnaire<2) and 71 ones (Patient Health Questionnaire>=2). It is obvious that the graphs of *cgphq2* and *cggad7* are not equally distributed, which means the problem of data imbalance exists. Training models on imbalanced data can lead to false sense of performance because the minority class, caregivers who have the symptoms in this case, would be ignored.

3.3 Dimensionality Reduction with Principal Component Analysis

We also used Principal Component Analysis (PCA) for the purpose of exploring the outcome variable distribution. PCA is a technique that reduces high dimensional data to lower dimensions and therefore enables visualization. We reduced our data to two dimensions and Figure 3.3.1 is an example of our *cgdistress* data projected on a 2D plane. This can tell us there is no obvious linearly separable relationship between the two classes and hence we can guess that Machine Learning models can perform better than Logistic. Also, we learn that we'd better use non-linear kernels for models like Support Vectors Machine.

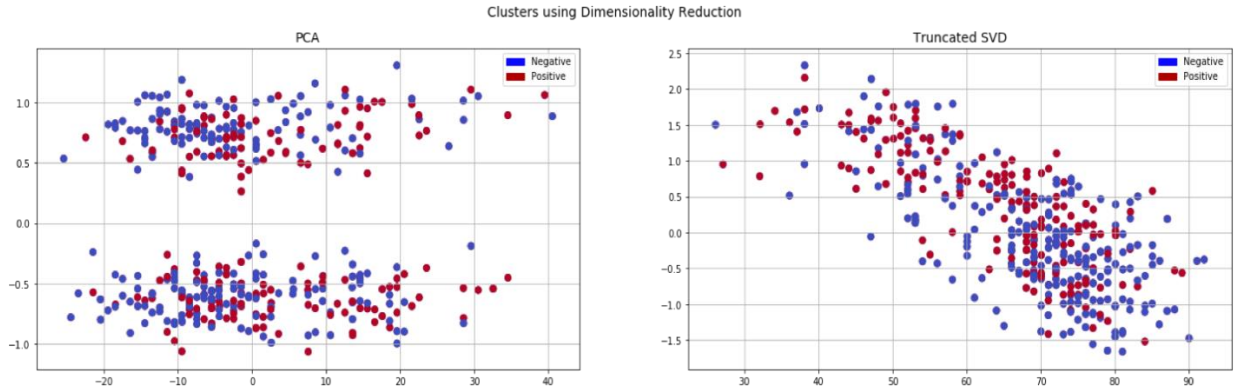


Figure 3.3.1: Dimensionality Reduction with PCA

3.4 Spearman Correlation Heatmap

Correlation measures the direction and strength of two variable's tendency to vary together and helps us to see what regressors are probably good for classification. There are two major types of

correlation analysis: Pearson and Spearman. Pearson measures linear relationship between two variables, while Spearman correlation measures the monotonic relation between two variables, in which the two variables vary together not at a constant rate. Spearman correlation measures on the ranked data that is more suitable to analyze our data, which are all categorical. Figure 3.4.1 is the formula of spearman correlation.

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

Figure. 3.4.1 Formula of Spearman Correlation

Shown in Figure 3.4.2 is the correlation heatmap. Here we can see patient distress, age and gender are the most significant ones and might be useful in the models.

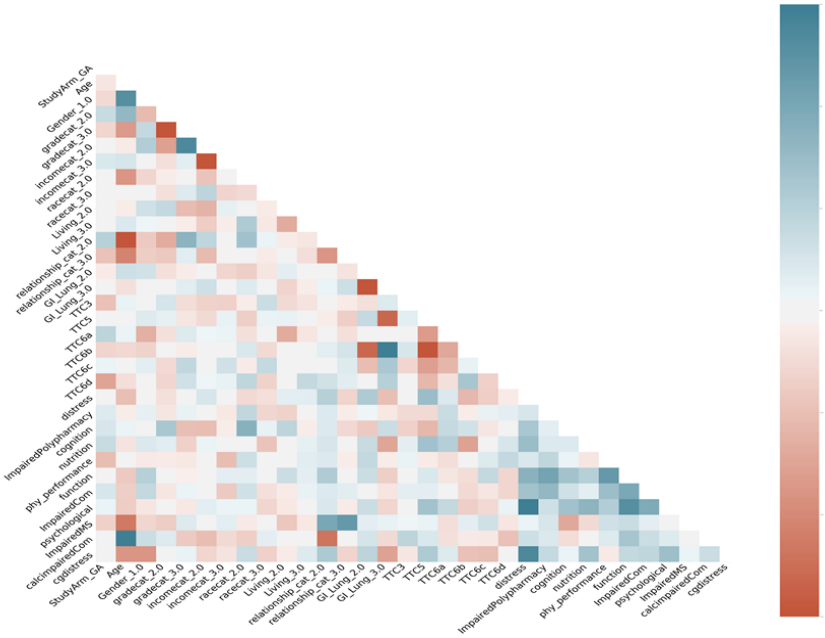


Figure 3.4.2 Spearman Correlation Heatmap

4.Model Design and Implementation

4.1 Labelling Cancer Treatments

For each of the four tasks discussed above, we designed approaches to find appropriate labels.

Type 1: We looped the raw and reference datasets and directly matched the drugs with labels.

Type 2: We scanned through the reference datasets and collapsed drug records iteratively with intersections down using Python's set operations. Recall the example from part 3, which is to infer the class and generic of Abraxane using records of "Abraxane and Gemzar" and "Abraxane and Carboplatin". We transformed the two records into sets and found their intersections, which is "Abraxane" and did the same intersection operation to their generics and classes to single out the labels for "Abraxane". We then add the "Abraxane" into the dataset and in the next iteration we can break "Abraxane and Gemzar" into "Abraxane" and "Gemzar", "Abraxane and Carboplatin" into "Abraxane" and "Carboplatin". And we add these into the datasets. The loop continues until there are no records we can break down and that is the minimal rules of association we can find. Using singled out drug labels we can complete the type 2 task with ease.

Type 3: We calculated the Levenshtein Distance between drug names and fuzzy matched those with high similarity via the Fuzzywuzzy package. Levenshtein distance is defined in figure 4.1.1:

$$\mathrm{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0, \\ \min \begin{cases} \mathrm{lev}_{a,b}(i-1,j) + 1 \\ \mathrm{lev}_{a,b}(i,j-1) + 1 \\ \mathrm{lev}_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} \end{cases} & \text{otherwise.} \end{cases}$$

Figure 4.1.1 Formula of Levenshtein Distance between word a and b

Where 1(ai≠bj) denotes 0 when a=b and 1 otherwise. It is important to note that the "lev" rows on the minimum above correspond to a deletion, an insertion, and a substitution in that order. Then we calculated the Levenshtein similarity ratio, defined in figure 4.1.2, and matched drugs names with a similarity ratio higher than 0.87 (threshold we found).

$$\frac{(|a| + |b|) - lev_{a,b}(i,j)}{|a| + |b|}$$

Figure 4.1.2 Formula of Levenshtein Similarity Ratio

Type 4: We asked for help from clinicians and they manually labeled the data.

4.2 Predictive Models for Caregiver QOL

4.2.1 Classification for three categorical variables (cgdistress, cggad7, cgphq2)

We started with variable *cgdistress*. After data cleaning in part 2, we separated data into train and test sets with a 4:1 ratio. Then we selected models from Logistic Regression, KNN Classifier(KNN), Random Forest(RF), AdaBoost(Ada), Gradient Boost(GB), and Support Vector Machine(SVM) by grid searching with 5 folds Cross-Validation, using SKLearn package. Here are some explanations of terminologies: KNN classifier assigns data to the class most common among its k nearest neighbors. Random forests classifier is an ensemble method that constructs multiple decision trees then classifies according to the mode of the classes of the individual trees. Both adaboost and gradient boost learn from previous mistakes: adaboost learns by increasing the weight of misclassified data points while gradient boost learns directly on residual errors made by the previous prediction. SVM divides the data points of two separate categories by a clear gap that is as wide as possible.

Grid Search searches for the best hyperparameters based on certain scoring metrics and exhaustively generates combinations from a grid of pre-specified parameter values. K-Fold Cross-validation (CV) estimates the performance of a model, by randomly shuffling the data and dividing them into K groups. It sets each one of the K groups as validation, uses the remaining K-1 groups as train sets, and then fit and evaluate the model iteratively. Figure 4.2.1.1 represents the grid search for three hyperparameters of a random forest classifier, in which each point represents a combination and the brightness stands for the cross-validation score. As shown, CV score decreases when the hyperparameters are been set too large, thus, blindly going after more complicated model will result in overfitting.

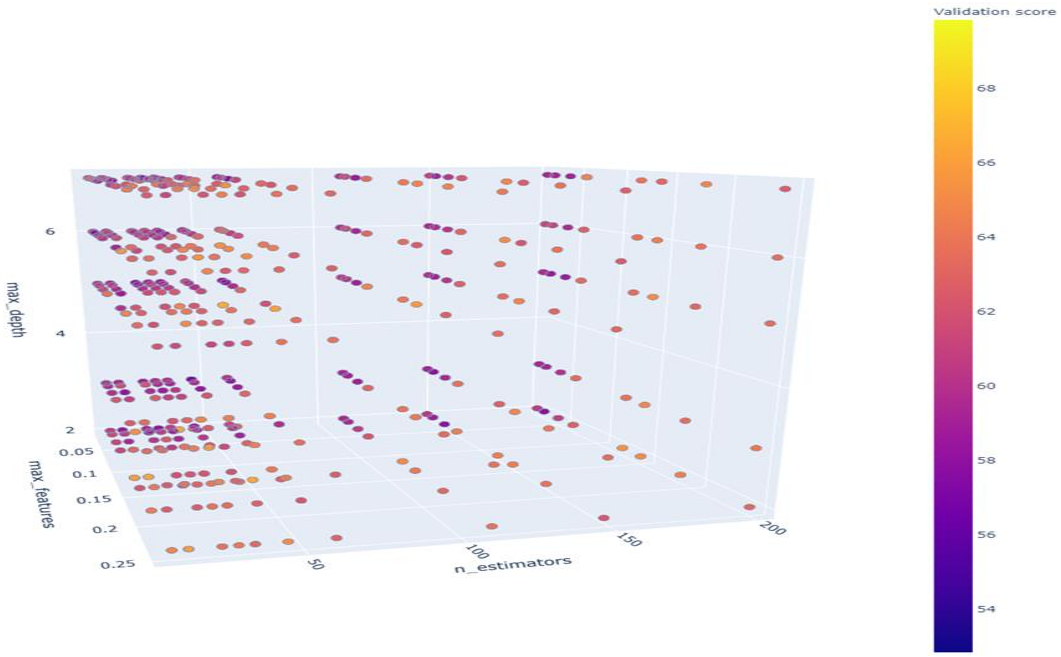


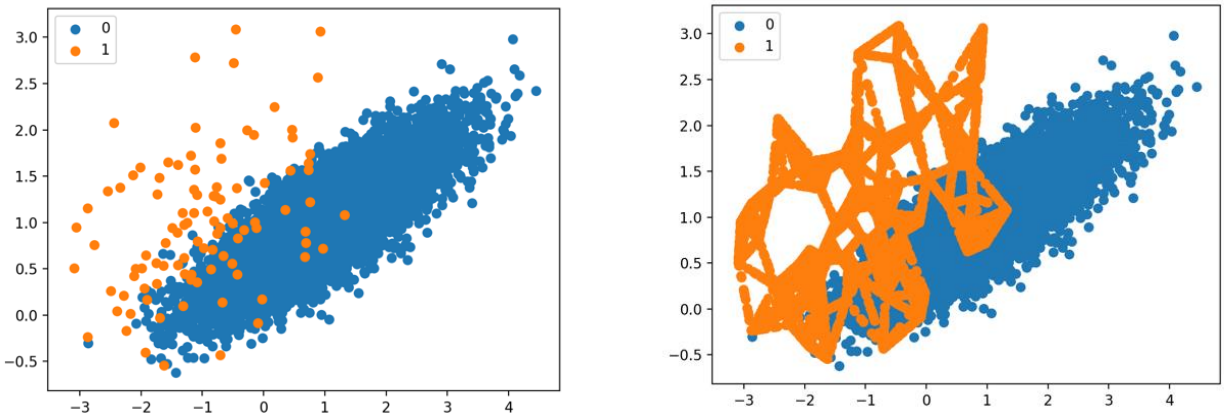
Figure. 4.2.1.1 3D Representation of Grid Search

We performed grid search over many combinations of hyperparameters, as recorded in the table shown in Table 4.2.1.2. The results will be discussed in part 5.

Table 4.2.1.2 Table of Hyperparameters used in our Grid Search

	Logistic	**KNN**	**RF**	**Ada**	**GB**	**SVM**
1	C	Leaf Size	max_features	n_estimators	n_estimators	C
2	L1/L2 Penalty	n_neighbors	max_depth	Learning rate	max_depth	gamma
3		metric	min_samples_leaf		learning_rate	kernel
4			min_samples_split			
5			n_estimators			

Figure 4.2.1.3 Example of SMOTE (No units because it's an arbitrary example)



From the exploratory analysis, we found that variable *cggad7* and *cgphq2* have serious data imbalance, unlike *cgdistress*. Thus, we came up with two different approaches that gave us the

same satisfying results in mitigating imbalance: Synthetic Minority Over-sampling Technique (SMOTE) and class weights. SMOTE, which we implemented via the imblearn package, oversamples the minority classes by creating synthetic data points close in the feature space. Here in figure 4.2.1.3 we can see the minority class significantly increased after using SMOTE to create synthetic data. Assigning proper weights, which equals the ratio of minority to majority class, can also make the models stress more on minority class samples and have the same effects.

Same as for cgdistress, we performed Grid Search with cross validation for multiple models for *cggad7* and *cgphq2*. By creating a pipeline process, we integrate SMOTE in the beginning of slicing out each cross-validation fold, making sure that the model is trained on a balanced data set and validated on the imbalanced real data. The evaluation was also done on a test set containing only original data. The Smoted Cross-Validation is demonstrated in Figure 4.2.1.4.

Data imbalance also means accuracy is no more a feasible way to evaluate the model in grid search. Accuracy is skewed when there are a lot of true negatives, while in our case true positives (whether the caregiver is impaired) are of more clinical importance. Thus, we should instead use the F1 score, as illustrated in Figure 4.2.1.5.

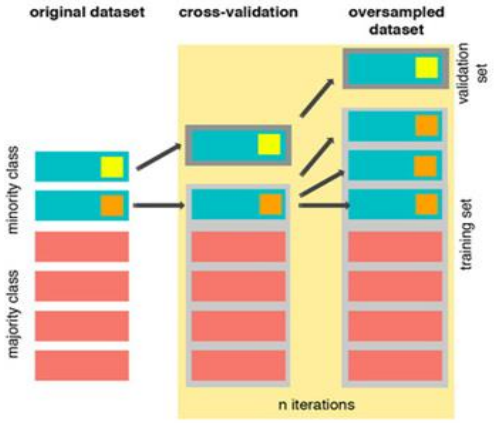


Figure 4.2.1.4 Pipeline of SMOTE-Cross-Validation

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$

$$accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

Figure 4.2.1.5 Formulas of F1 and accuracy

Finally, we reset the decision threshold probability value p by choosing the point closest to the top left corner of the ROC space to maximize the performance.

These methods significantly improved the performance of the model, and results will be shown in part 5.

4.2.2 Transforming Regression into Classification (SF12, AGG_MENT, AGG_PHYS)

After overcoming the data imbalance issues in the classifications, we met other challenges in the regression task. Due to the all categorical nature and lack of predicting power of the predictors, we found it hard to reduce RMSE (9-10) during the regression task. We consulted Prof Xu and discretized the continuous outcome variable, cutting by the first 25 quantiles. Here we will use SF12 as an example in Figure 4.2.2.1:

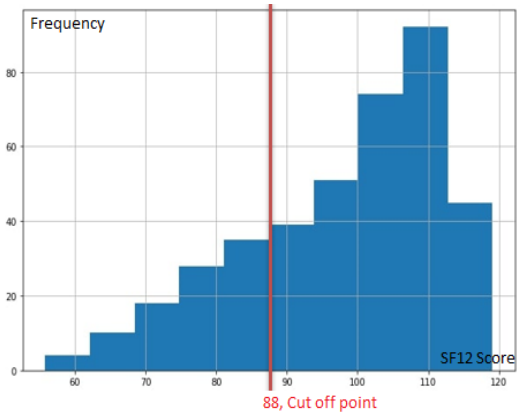


Figure 4.2.2.1 Distribution of original SF12 and cutoff line

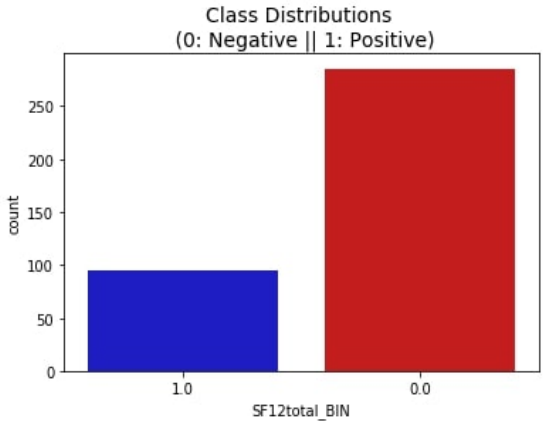


Figure 4.2.2.2 Class Distributions of SF12 after categorization

Here all the caregivers will score lower than 88 will be given class of 1 (Impaired health, positive) while those above 88 will have class 0. (Better health, negative)

This leads us to a similar imbalanced classification like before. Now we perform the exact same procedure in 4.2.1 and results will be discussed in part 5.

5. Model Performance

We will first list a table (Table 5.1) of final best model performance for each variable, then showcase improvement in results with *SF12* and *cgphq2* as more specific examples. For *cgdistress*, we give overall accuracy, while providing the test accuracy on both classes for the rest due to imbalance.

Table 5.1 Performance for all models for all outcomes (Test) vs Logistic

	Cgdistress (RF)	**Cgphq2 (SVM)**	**SF12 (RF)**	**Cggad7 (SVM)**	**AggMent (GB)**	**AggPhys (Ada)**
AUC	0.71	0.71	0.81	0.63	0.69	0.73
Accuracy Class 0	0.69(Overall)	0.66	0.73	0.66	0.69	0.70
Accuracy Class 1	~	0.78	0.71	0.57	0.62	0.75
AUC(LR)	0.63	0.62	0.74	0.61	0.56	0.63
Acc Class 0(LR)	0.61(Overall)	0.57	0.93	0.62	0.62	0.56
Acc Class 1(LR)	~	0.56	0.3	0.52	0.5	0.58

Obviously, our revised Machine Learning models (Smote/Class weight, F1 search) is better than Logistic Regression:

1) SF12: Revised RF (smoted, f1) is the best model of all

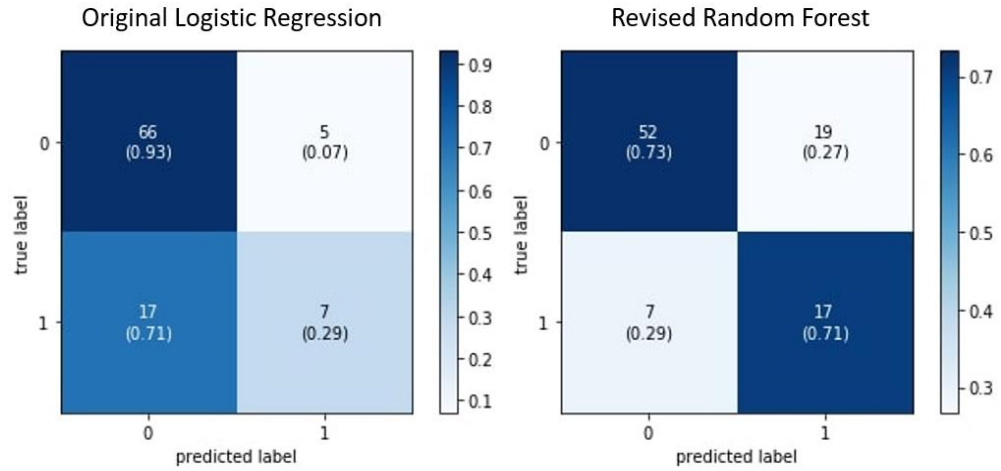


Figure 5.2 Confusion matrix for SF12 Comparison

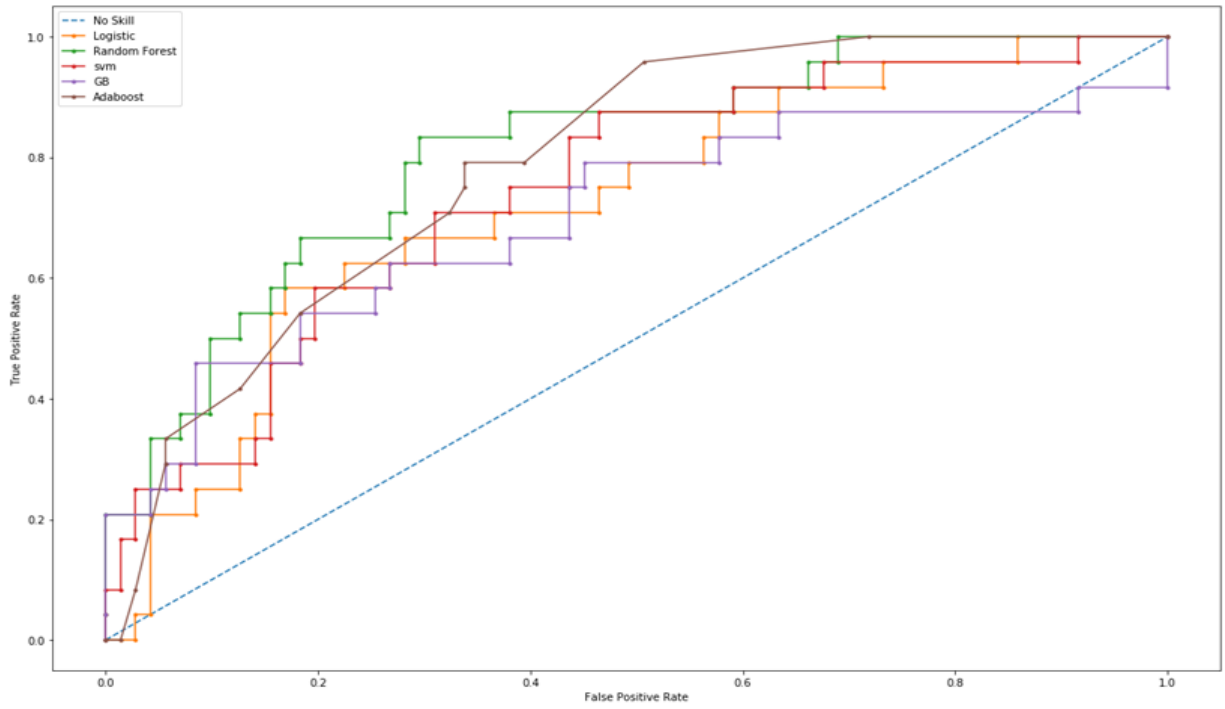


Figure 5.3 ROC Curve Comparison

2) Cgphq2: Revised SVM (smoted, F1) better than Plain Logistic (unsmoted, accuracy) and Revised Logistic (smoted, F1)

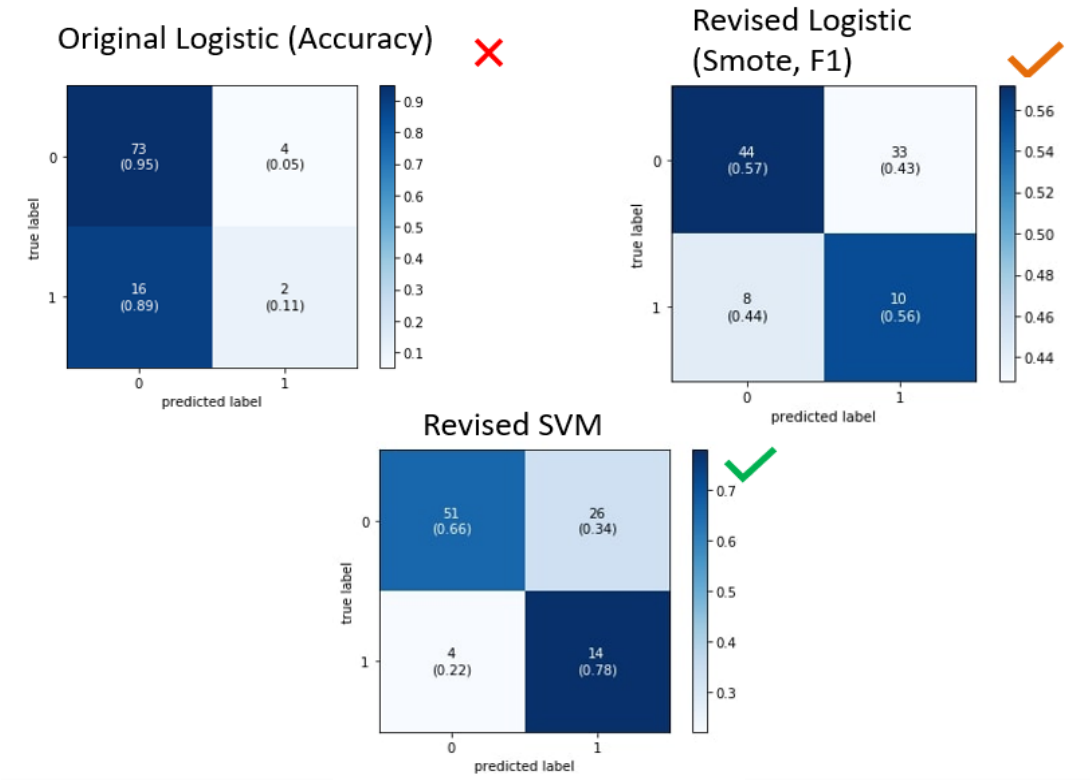


Figure 5.4 Confusion matrix for Cgphq2 Comparison

From the visualizations above, we can tell that smote/class weight and F1 based search will provide a big tradeoff between majority and minority class accuracies for imbalanced data. This is what we need in this project because true positives are far more significant. Also, machine learning models tend to overperform logistic regressions.

6.CONCLUSIONS

To better interpret model results and to help clinicians identify factors that might lead to distress and take preventive measures, we used LIME (Local Interpretable Model-agnostic Explanation) to understand the black-box models. LIME provides a sample-wise interpretation of positive and negative factors contributing to the final classification results. It modifies a single data sample by tweaking the feature values and observes the resulting impact on the output. Here is an example of using LIME on our radial basis kernel SVM model on *cgphq2*(depression). Since the radial basis kernel is not linear, it can't be illustrated traditionally with coefficients or feature weights.

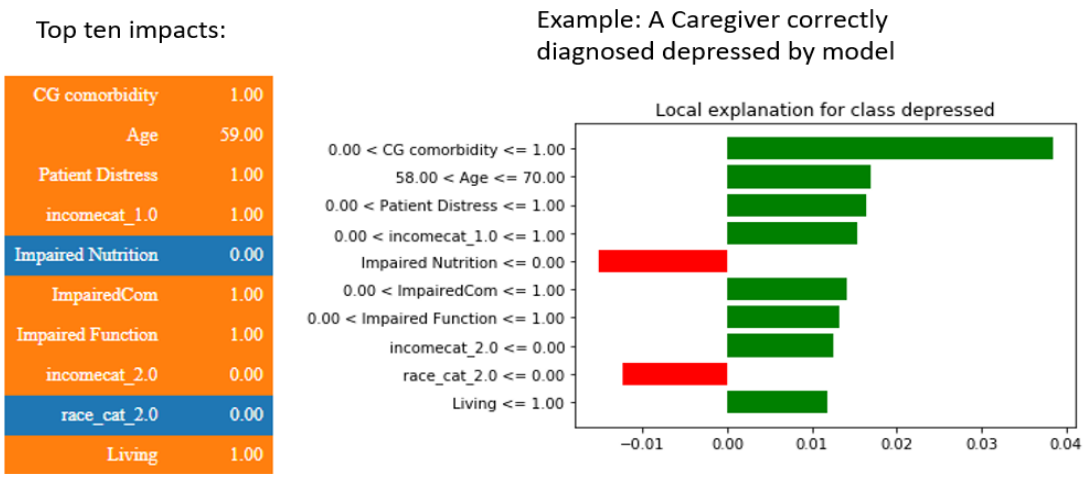


Figure 6.1 Top ten impact factors and their contributions to the caregiver's depression

As we can see in Figure 6.1, factors like caregiver comorbidity (1 is true,0 is false), low income (category 1 is low) and patient distress (1 true) contribute a lot to the caregiver's depression (green is positive contribution). Not having impaired nutrition (0 false) is preventing depression (red is negative contribution). Hence, we can infer that we need to pay more attention to caregivers with bad nutrition, low income… and provide relevant assistance.

In conclusion, in part 1 our algorithm can automate the labeling process for drugs. In part 2, our model generally outperforms logistic regression, as we can see the data has no linear separability from the PCA. Models like SVM with radial basis kernel(rbf) performs well a lot, probably due to the small data size and the advantage of rbf in dealing with high dimensional data. We can also interpret our model results frankly with the help of LIME. This can be of a lot of importance for clinicians to help caregivers at risk.

In the future, we can generalize our labeling algorithm to help clinicians with more data like this. Also, as suggested by Professor Xu, we can further enhance our predictive models using cross-model ensemble techniques like superlearner.

7.REFERENCES

[1]: Spatuzzi, Roberta, et al. "Does Family Caregiver Burden Differ Between Elderly and Younger Caregivers in Supporting Dying Patients with Cancer? An Italian Study - Roberta Spatuzzi, Maria Velia Giulietti, Marcello Ricciuti, Fabiana Merico, Francesca Romito, Giorgio Reggiardo, Loredana Birgolotti, Paolo Fabbietti, Letizia Raucci, Gerardo Rosati, Domenico Bilancia, Anna Vespa." *SAGE Journals*, journals.sagepub.com/doi/10.1177/1049909119890840.

[2]: Magnuson A, Allore H, Cohen HJ, et al. Geriatric assessment with management in cancer care: current evidence and potential mechanisms for future research. J Geriatr Oncol

[3]: Ingber, Ron. "Caregiver Stress Syndrome." *Caregiver.com*, 18 Dec. 2018, caregiver.com/articles/caregiver-stress-syndrome/.