1 Executive Summary

2 Introducing the California Teachers Study (CTS)

3 Objectives

4 Exploratory Data Analysis

5 Machine Learning Models

6 Conclusion

Predicting and Modeling Mortality Risk Baesd on CTS Data

Kiera (Xiaoyu) Zhu

8/9/2022

1 Executive Summary

The California Teachers Study (CTS) is an observational cohort study that follows and records health data from female teachers, administrators, school nurses, and other members of the California State Teachers Retirement System (CalSTRS). The data from the CTS has been used for extensive research into breast cancer while also giving insight into other cancers and diseases. The main focus of this explorational study is the OSHPD hospitalization records that contain hospitalization data for CTS participants.

The primary objectives of this study are as follows:

- 1. Develop the best fitting model that **predict the death risk within 1-month time window** based on baseline characteristics and hospitalization.
- 2. check if time window works after putting hospitalization characteristics into model.
- 3. Assess what **certain co-morbidities and death causation** are significant factors in predicting the risk of death.
- 4. Find out **temporal(seasonal)** trends related to death.

Finding 1: The XGBoot Model Is the Best Fitting Model.

Finding 2: Time Window Works In Random Forest Model._

Finding 3: Lifestyle Habbits Will Affect Your Health.

Finding 4: Seasonal Variable Has Some Influence But Not Obvious.

2 Introducing the California Teachers Study (CTS)

The California Teachers Study (CTS) is an observational cohort study established in 1995 that follows female teachers, administrators, school nurses, and other members of the California State Teachers Retirement System (CalSTRS). Members of the CalSTRS have provided information regarding their health and behaviors, and information regarding these patients have continued to be recorded and studied. The data provided has allowed for extensive research on breast cancer, and has given insight into other cancers and disease. This report will focus in particular on the OSHPD hospitalization records that contain hospitalization information from the participants of the CTS.

3 Objectives

The primary objective of this report is to develop the best fitting model to predict the short-term risk of death based on baseline characteristics and hospitalization. The hospitalization data (which includes diagnoses, co-morbidities, length of stay, discharge date) of California Teacher’s Study Participants from 2000 to 2015 will be used for the analysis. Key characteristics and possible co-variables such as age, ethnicity, height, and weight (taken from the California Teacher’s Study Questionnaire) will also be incorporated.

Machine learning will be used to build the best fitting prediction model that predicts the probability of death after prior in-patient hospitalization. In particular, the effect of certain time windows between hospital discharge and death, as well as the effect of co-morbidities and specific procedures will be examined to determine their significant in predicting short-term risk of death after in-patient hospitalization.

4 Exploratory Data Analysis

4.1 Basic Data Exploration

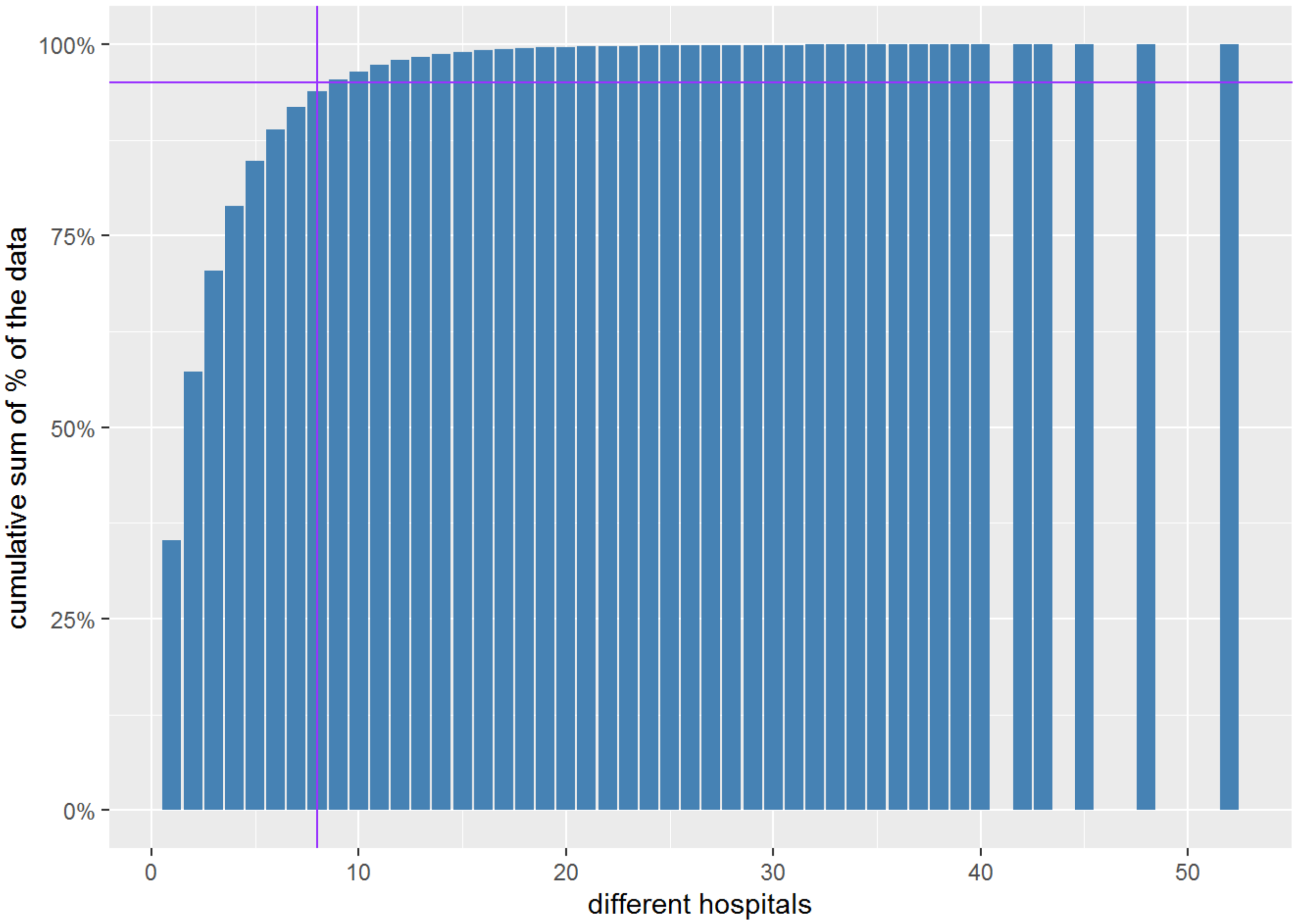
The data set consists of 154315 CTS participants and 164 variable. The primary outcome that will be examined in this project is death of the CTS participants.

Deceased	Frequency
No	70911
Yes	83404

Out of all of the participants in the data set, around 51.6% of participants with prior in-patient hospitalization records had died. The data set is relatively balanced with regards to the target outcome and will help reduce potential biases when building prediction models.

4.2 Dealing with One Participant_key with Multiple Hospital Visits

As shown in the follwoing graph, Many patients went to more than 1 hospitals, and about 95% of participants went to less than 8 hospitals. Although most of the basic characteristics remain unchanged for a patient, the hospitalization data from different hospitalization visits may differ. In this case, we only use the hospitalization data of the last visit for a patient for accuracy because we think it is the latest for most related to the patients’ death risk. Finally, we get unique data set with **48324** patients and corresponding variabels.



4.3 Variable Selection: Choose Only Important Variables

Especially because over 150 variables may be overwhelming to look through, a main focus will be on choosing primary variables and deleting duplicated and irrelevant variables.

```
# g-group: patients' personal information
# g1: Variables related to birth, death, age
g1= c('date_of_birth_dt', 'date_of_death_dt','age_at_baseline')
# g2: Variables related to demography and personal characteristics
g2 = c('ses_quartile_ind','height_q1','weight_q1','bmi_q1','participant_race','adop
ted', 'twin')
# g3: Variables related to medical history
g3= c('hysterectomy_ind', 'bilateral_mastectomy_ind','bilateral_oophorectomy_ind',
'preg_ever_q1', 'brca_selfsurvey', 'mammo_ever_q1',
      'endoca_self_q1', 'cervca_self_q1','ovryca_self_q1', 'lungca_self_q1', 'leuk_
self_q1', 'hodg_self_q1', 'colnca_self_q1', 'thyrca_self_q1', 'meln_self_q1', 'diab
_self_q1', "stroke_self_q1","hrtatk_self_q1","hbp_self_q1",
      'asthma_q3')
# g4: variables related to physical activity, smoking, alcohol usage and diet
g4= c('allex_hrs_q1', 'sit_hrs','sleep_hrs', 'vit_reg_no',
      'diet_plant', 'diet_highprotfat','alchl_analyscat', 'smoke_statcat' )
```

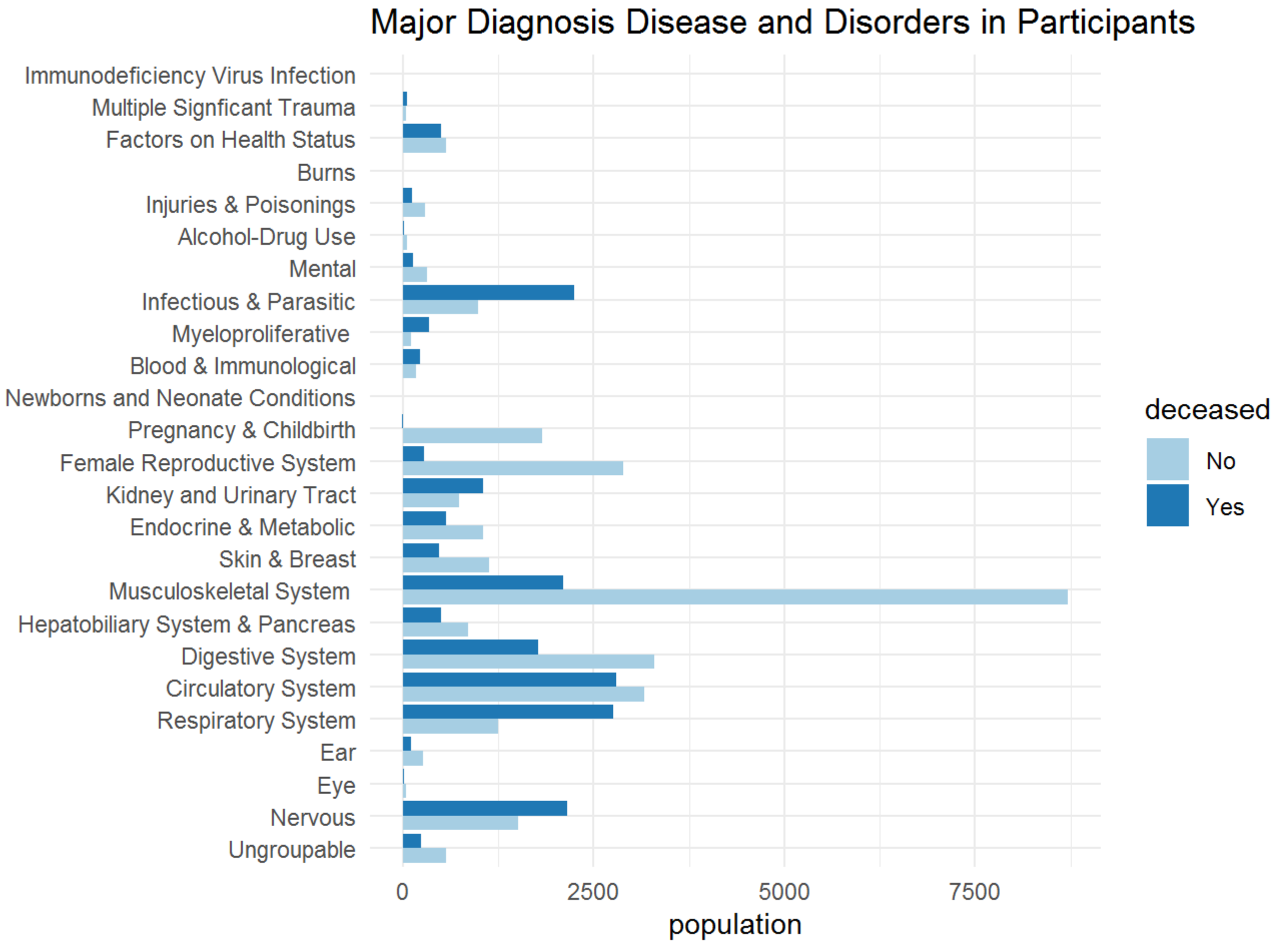
```
# h-group: hisptalized information
# h1: admission and discharge date to be transformed
h1 = c("admission_dt","discharge_dt")
# h2: visit information
h2 = c("admission_typ","length_of_stay_day_cnt","dnr_flg",'major_diag_cat_cde', "pa
tient_care_typ","patient_disposition_cde")
# h3: payer information
h3 = c("payer_cat_cde","total_charges_amt")
# h4~h5: diagnoses and procedures by ccs categories
h4 = c("diag_ccs_code1", "proc_ccs_code1")
```

```
# t-group: target variable
# t1: Indicator for deceased.0=No; 1=Yes.
t1= c("deceased")
```

4.4 Primary Exposures

4.4.1 Major Diagnostic Categories

The Major Diagnostic Categories (MDC) are formed by dividing all possible principal diagnoses (from ICD-9) into 25 mutually exclusive diagnosis groupings. The diagnoses in each MDC correspond to a single organ system or etiology and, in general, are associated with a particular medical specialty.



4.4.2 Diagnosis CCS Code

The diagnosis CCS code contains information regarding the different diagnoses that the CTS participant received during hospitalization. In the data set, there are five different diagnosis ccs codes, with one of the five as the primary diagnosis. All five diagnoses CCS codes will be included in the prediction model, but the frequency of different primary diagnoses will be further examined.

Primary Diagonise	Frequency
Other	31511
Osteoarthritis	5125
Tuberculosis	2909
Acute cerebrovascular disease	1425
Fracture of neck of femur	1275
Cardiac dysrhythmias	1210
Pneumonia	1187
Spondylosis;intervertebral disc disorders;back problems	1031

Primary Diagonise	Frequency
Benign neoplasm of uterus	929
Undefined	862
Congestive heart failure; nonhypertensive	860

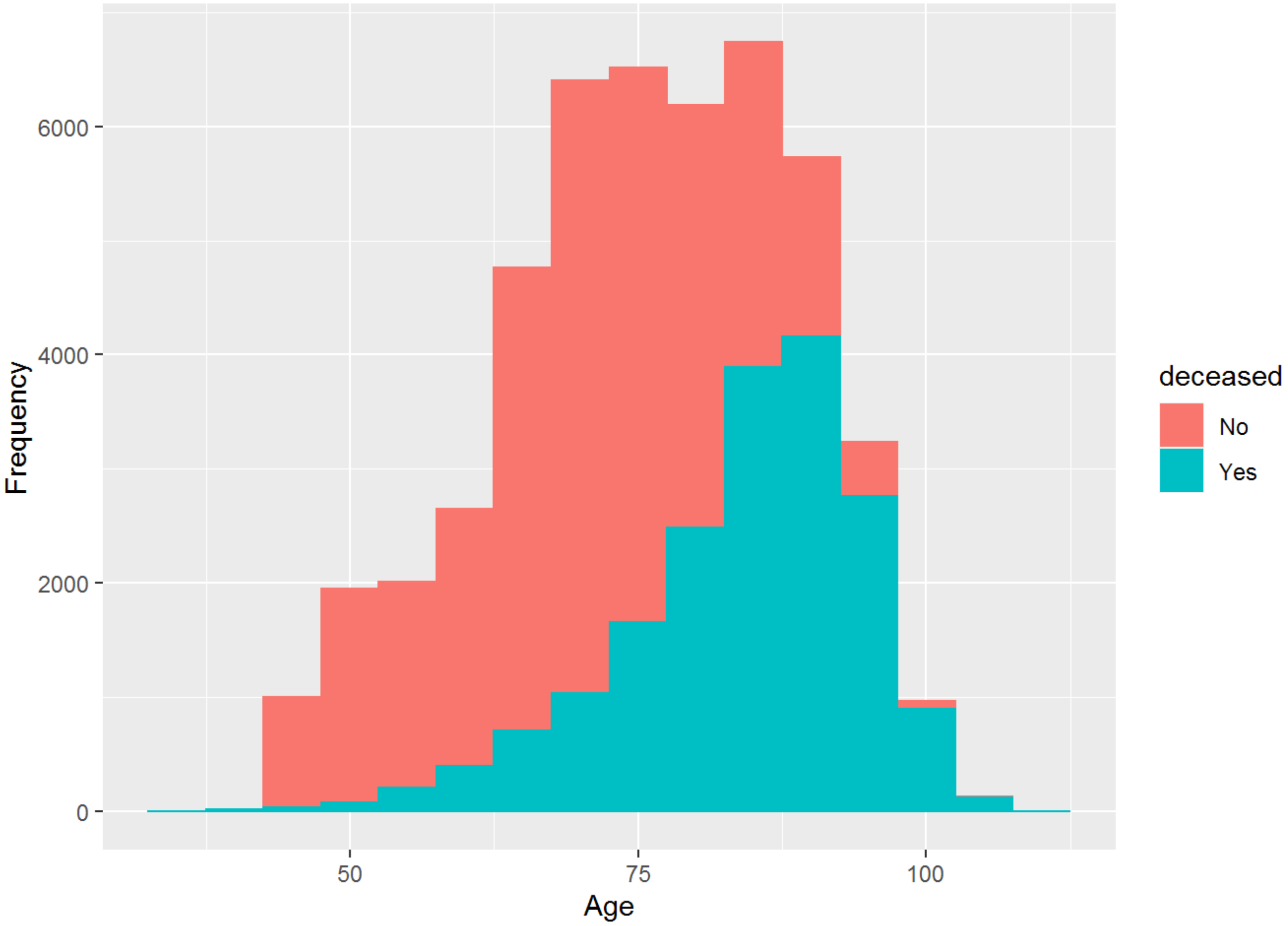
4.4.3 Procedure CSS Code

The procedure CSS code functions similarly to the diagnoses CCS code mentioned previously. The procedure CCS code contains information regarding the procedure that the participant underwent during hospitalization. There were five procedure CCS codes given in the dataset, and all five will be included in the model. However, just as with the diagnoses CCS codes, the different primary procedures will be further examined.

Primary Diagonise	Frequency
Other	19902
Alive	14452
Arthroplasty Knee	3084
Hip Replacement	2471
Hysterectomy	1612
Undefined	1455
Respiratory Intubation	1453
Blood Transfusion	1003
Fracture Treatment	906
Spinal Fusion	823
Other procedures to assist delivery	705
Physical Therapy	458

4.4.4 Age

Age is one of the baseline covariates in regards to patient health. Therefore, **age_at_death** is created as new variable. The age for each participant will be calculated. For participants that have died, the age in which they passed away will be calulated. For participants that are still alive, their age at the end of 2015 will be caculated, as this study will only be looking at hospitalization data until the year 2015.



4.5 Data Processing in Detail

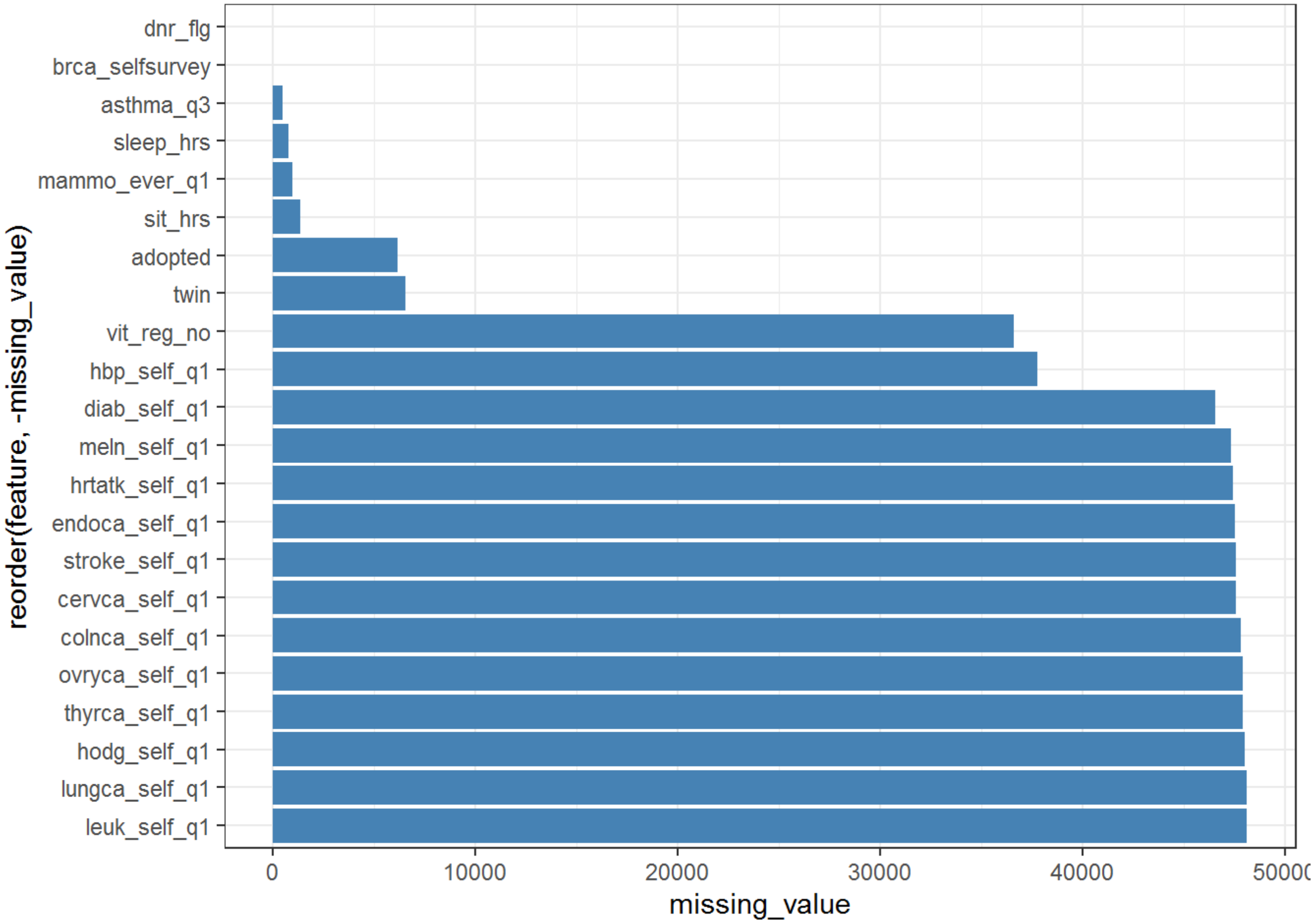
4.5.1 Admission Date Later Than Death Date

There may be misfilling happened that some data’s admission dates are later than death date. So we delete these invalid data. There are 48178 records now.

4.5.2 Missing Data

Looking at character value counts, the most missing variables are self-filling questions, which are not filled mostly because they didn’t have the disease. So we label them and set NA to “B” or “N”. Looking at numeric variables, the most missing variable are self-filling basic information. For example, diet_highprotfat and diet_plant, which produce NA because excel problem. So we are going to fill these NA with mean of each variable. And for questionnare variable like preg_ever_q1, I just set NAs to 0. .

char missing map | num missing map



4.6 Turn Variable into Factors

This is one complicated process in the study. I check every single variable selected. For those are meant to be factors with two categories, I create the categories for them. And for multiple categories, I turn them all into different factor levels. And for sleep_hrs and sit_hrs these two variables which were char variables, I turn them into numeric variables by giving them matching values.

```
dim(data)[2]
```

```
## [1] 52
```

```
sum(sapply(data,is.factor))
```

```
## [1] 36
```

```
sum(sapply(data,is.numeric))
```

```
## [1] 12
```

4.7 Create new variables

As the objectives mentioned above, we should build some specific variables for further study. So discharge_to_death, discharge_to_death_cat, death_1mon, death_6mon, death_1yr, death_2yr, death_up5yr, season_death,date_of_death_dt are created.

4.7.1 Time Window After Death

One of the aims of this research project is to predict the probability of death within a certain time window after hospitalization and to assess whether the time window itself is a significant factor in assessing the risk of death. Therefore, a new variable will be created to count the number days between discharge from the hospital and the date of death.

```
data %>%
  group_by(discharge_to_death_cat) %>%
  summarise(n = n())
```

```
## # A tibble: 7 x 2
##   discharge_to_death_cat     n
##   <chr>                  <int>
## 1 1 month                 7933
## 2 1 yr                    1184
## 3 2 yrs                   1184
## 4 5 yrs                   1708
## 5 5 yrs+                  1099
## 6 6 months                3036
## 7 <NA>                   32034
```

4.7.2 Temporal Trends Variable

I also create seasonality of death (season_death), to see if there’s seasonal pattern related to death.

```
## # A tibble: 5 x 2
##   season_death     n
##   <fct>        <int>
## 1 Alive        33508
## 2 Fall          3930
## 3 Spring        4064
## 4 Summer        3798
## 5 Winter        2878
```

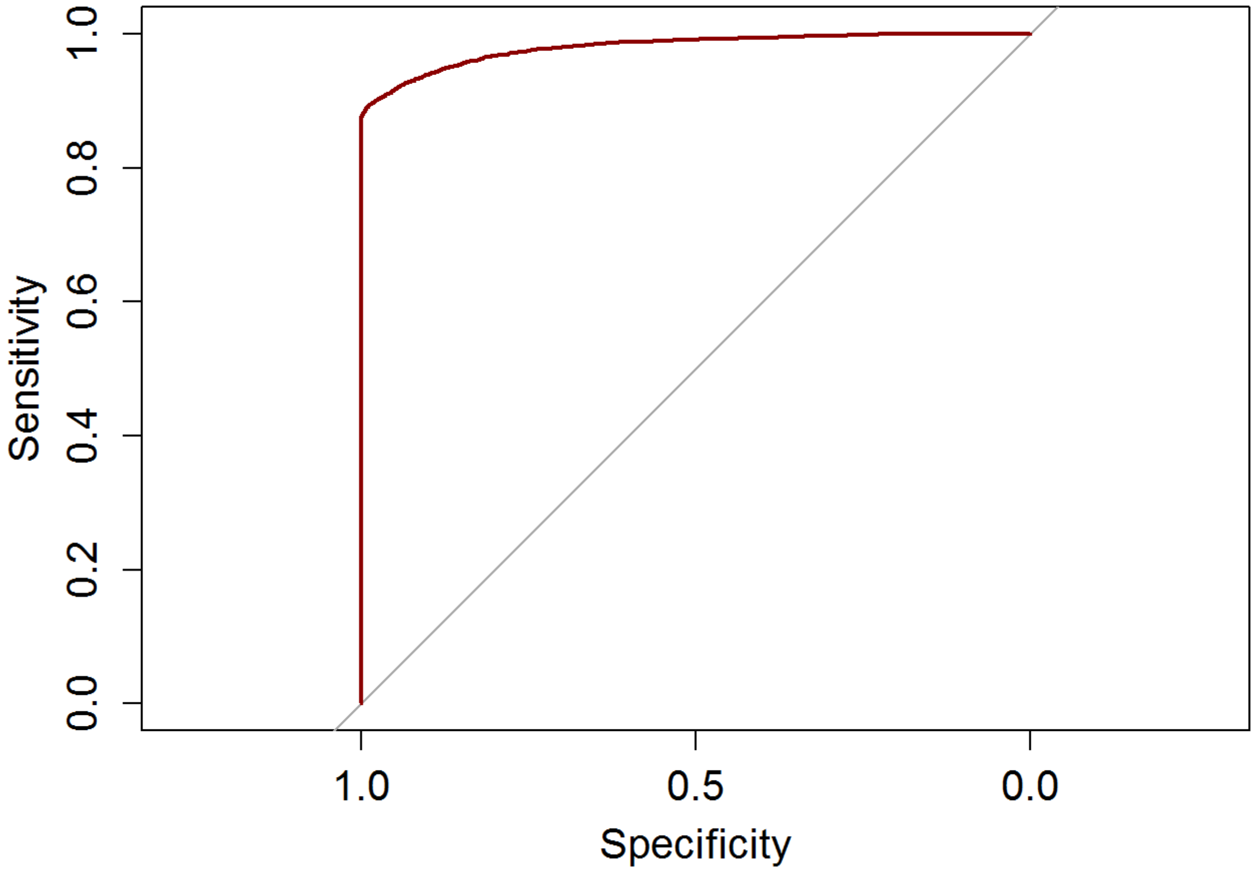
5 Machine Learning Models

After cleaning the dataset, I build a series of model to meet the objectives above. Mainly it’s two parts, which are Feature Selection part and Making Predicting Model part. For feature selection, the standard 75:25 training and testing split would be used for more training dataset to see variable importance. And the key for classification is variable deceased. For model-building, the standard 70:30 training and testing split would be used for more testing dataset to test model performance. And we only keep death_1mon as certaion time window and key for classification.

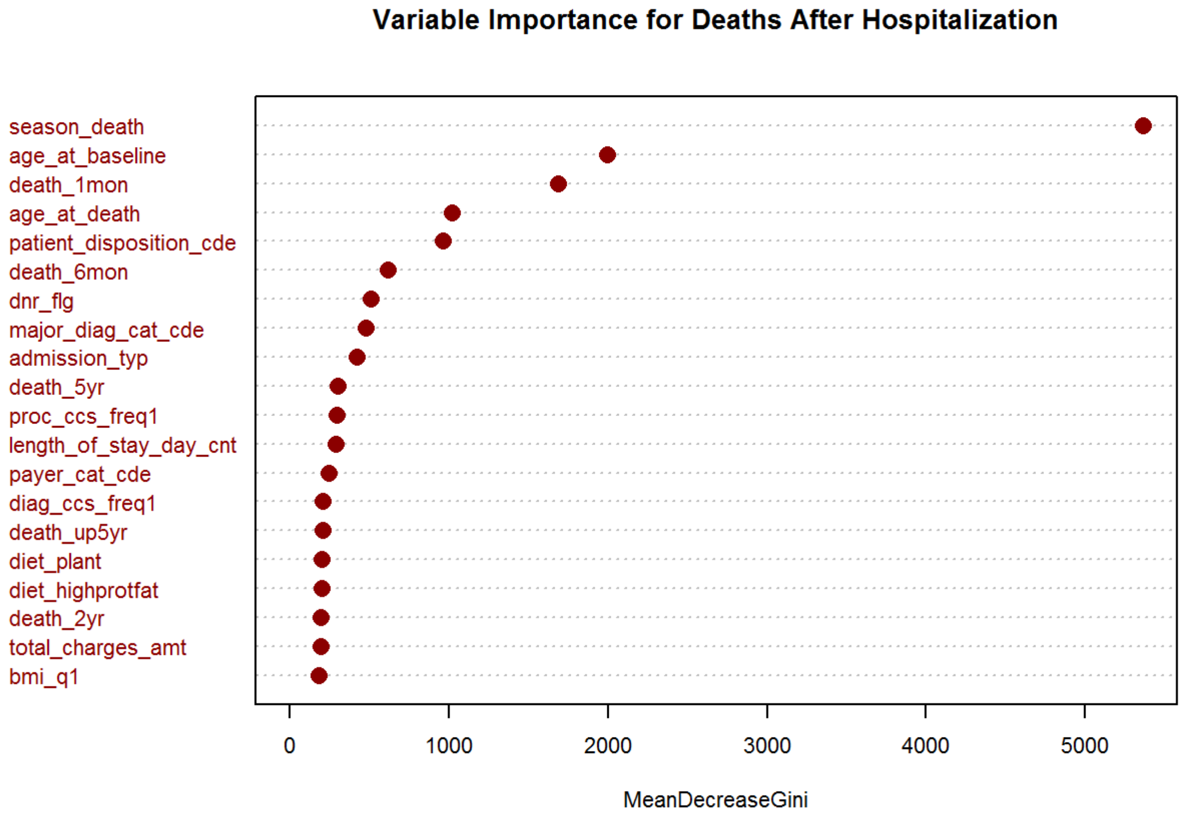
5.1 Feature selection: use Random Forest Model

Random forests will be used to create decision trees that will predict the short-term risk of death in CTS participants with prior hospitalization records. In total, there are 54 features that will be used to create the random forests. These features include our outcome of interest, whether the participant died after being discharged, and hospitalization data such as diagnoses and procedure codes. Key demographic data such as age and race are also included. Also, information regarding certain lifestyle behaviors, such as alcohol and tobacco use, physical exercise, and diet was also incorporated.

After running the random forest, the resulting AUC value is 0.9803, indicating that the random forest model is able to accurately predict death in participants with prior hospitalization records.



```
## Area under the curve: 0.9803
```



5.2 Best Fitting Model

Random forest, bagging and XGBoost models were selected to perform classification. Results in **table 1** showed that the test AUCs of the best model of each type were very similar. All these three model, I set **death_1mon** as certain time windom. Especially as time goes on, like when a participant is over five years out from a hospitalization, it may be harder to attribute death to their hospitalization.

For the Random Forest models, the test AUC is 0.969, which denotes great performance in distinguishing between classes. Using training and testing sets with a 70:30 split and a balanced Random Forest model. And by viewing the ROC curve, it is evident that a trade-off exists between specificity and sensitivity.

For the Bagging model, the results is very similar with random forest model but more flexible. Bagging fits multiple models based on different datasets. The test AUC is 0.969. And both bagging and random forest are suitable for imbalanced data like CTS.

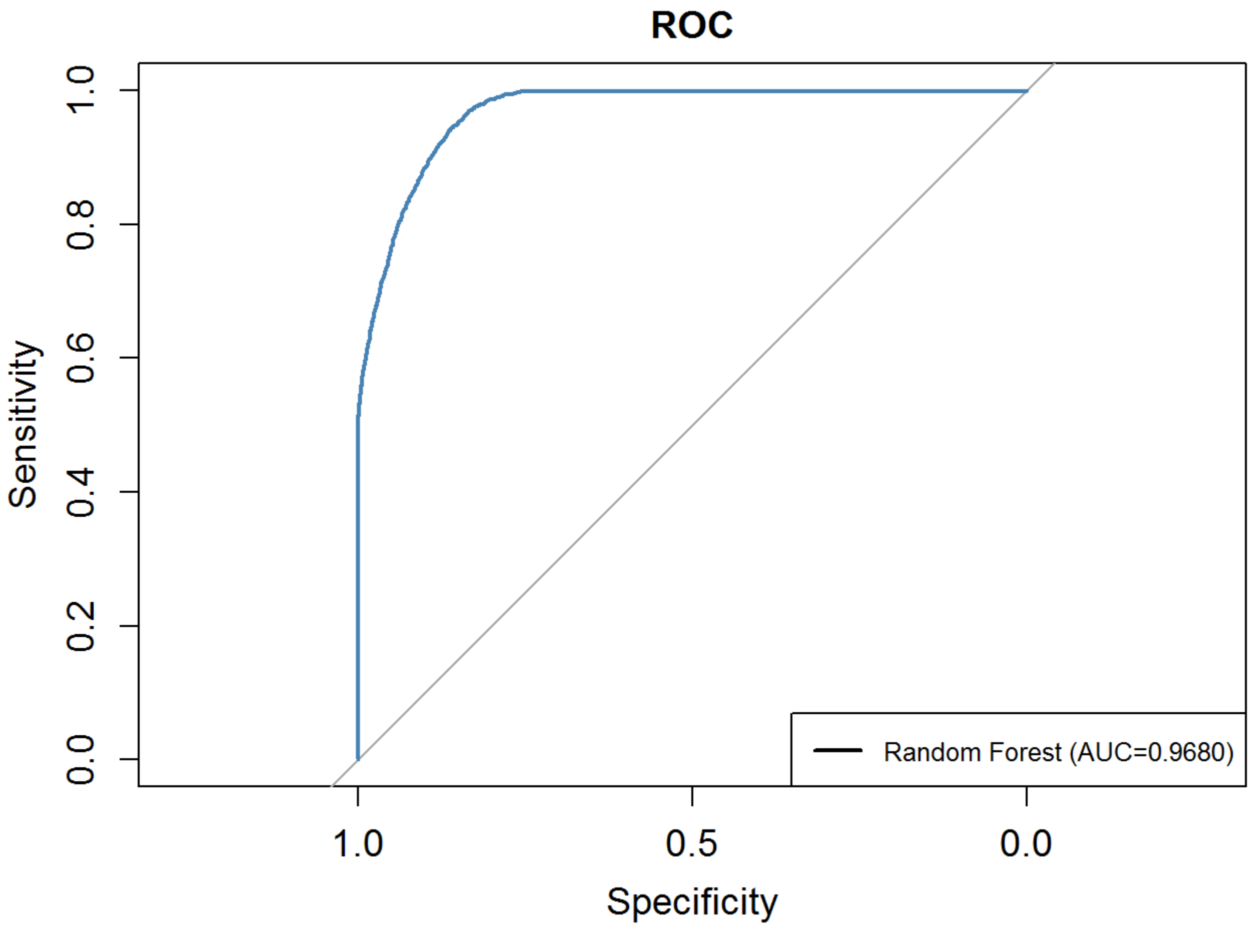
For XGBoost, or extreme gradient boosting model, its test AUC is highest, which is 0.972. This model was able to include the most features and handle factor variables with many levels. That’s why its performance is the best.

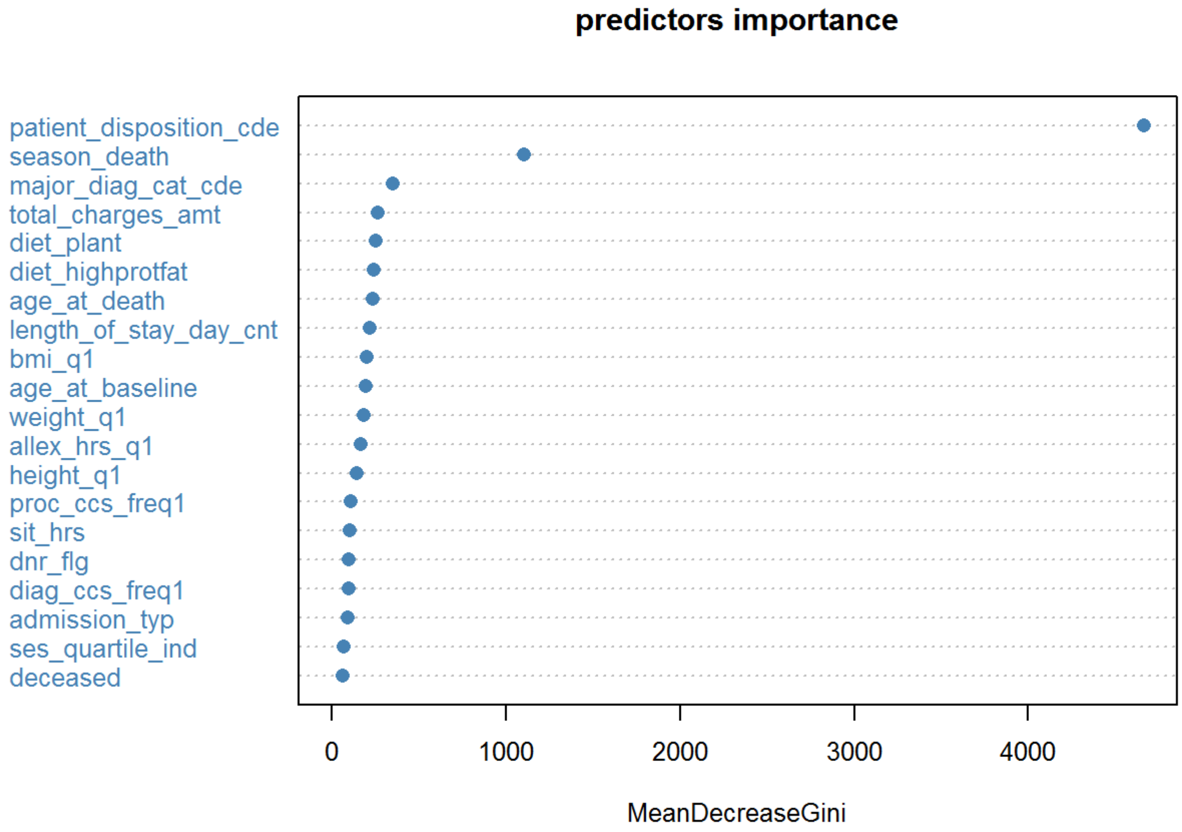
```
#performance metric table of 1-month time window

table <-matrix(c(rf_auc,bg_auc,xgb_tune_auc,rf_sen,bg_sen,xgb_tune_sen,rf_spec,bg_spec,xgb_tune_spec),ncol= 3,nrow=3,dimnames=list(c("Random Forest","Bagging","XGBoost"),c("AUC","Sensitivity","Specificity")))

# table<-readRDS("table1.rds")
kable<- kable(table,caption="Table 1: Performance matrix of different classification")
kable_styling(kable)
```

Table 1: Performance matrix of different classification

	AUC	Sensitivity	Specificity
Random Forest	0.9680	0.8413	0.9369
Bagging	0.9705	0.8724	0.9348
XGBoost	0.9716	0.8387	0.9401

Random Forest Bagging XBGoot





6 Conclusion

Finding 1: The XGBoot Model Is the Best Fitting Model.

We can see in table 1, the XGBoot model has best performance. To help balance the XGBoot model, the scale_pos_weight was assigned the standard value of the number of the majority class divided by the minority class, which ended up being roughly 30 in both cases. And, balanced accuracy was calculated here.

Finding 2: Time Window Works After Putting Other Variables In._ For feature selection part, we can see most time window variables(death1mon, death3mon, death1yr ec.) have great effect on prediction the risk of death. So maybe we should focus on patient just out of hospital and keep following up their condition to decrease mortality rate.

Finding 3: Lifestyle Habbits Will Affect Your Health. Many of the important variables in each of the models are similar. The ones with most overlap, which can help inform later model creation, between the two RF models created are:

- Patient disposition codes
- Major diagnosis code
- Cancer diagnoses
- Age
- Medication count
- BMI

- Diet For example, diet_plant(factor score for high plant-based pattern) and diet_highprotfat (factor score for high protein/fat pattern) reveal that daily diet really matter to these patients.

Finding 4: Seasonal Variable Has Some Influence But Not Obvious. From three classification models we can see that season_death (what season the patients were on when they died) show some feature importance but they're not significant. But this study only focus on death season. Other research team can focus on admission season or birth season for further study if interested.

Perception

Finally, the whole research exploratory project comes to end. I really learned a lot from this practicum. First, My R coding ability gets stupendous improvement. The whole project is coded with R language. From simple pipeline to self-make function in random forest, I feel more familiar with R right now. Second, my capability of processing variables increases a lot. CTS dataset has 100,000+ patients and 150+ variables, it's a great challenge to deal with it. I really enjoy check every single variable and transform it carefully. Last but least, I learn how to build XGBoot model in this study, it's queit a effective tool in machine learning especially this huge dataset. And I flexiblly use Random Forest model and Bagging model to compare their performance. Also I did used Logistic Model and QDA model to test the data, but didn't put them in the report for time limit. Thank you all.