ECON245 G1: Applied Healthcare Analytics
Final Report

AY2022-2023 Term 1

Group 1

Group members:

Clarice Liu Shen Xin (01396978)

Ieysaa Bin-Suhayl (01472991)

Liz Baiju (01389580)

Meshalini R Vikneswaran (01399577)

Rhea Singhania (01390995)

Table of Contents

1.0 Introduction

Our group aims to study the factors that affect the intrinsic demand for General Health Examinations (GHE), in the absence of cost influence. We sought to find a model which could best predict how existing individual demographic characteristics (eg. attitudes, status, medical experiences etc.) would influence the uptake of a GHE exam in Vietnam. Through this, we hope to better understand what additional measures the government would need to take to increase GHE uptake across all demographic groups in Vietnam.

GHE's have been observed to be associated with increased chronic disease recognition and treatment, risk factor control, preventive service uptake, and most importantly, improved patient-reported outcomes (Liss, Uchida, Wilkes, Radakrishnan, Linder, 2021). Given that Vietnam will be an aged society by 2035 (The World Bank, 2021), it is crucial for the country to focus on early disease recognition and management, because it can help lower the country's healthcare burden in the long run, and ensure sustainability of the healthcare system.

Currently, the Vietnamese government has focused efforts on making health care affordable for all people (Le, Kubo, Fujino, Pham, Matsuda, 2010), thus improving the Vietnamese health financing system greatly. This is perfectly in line with what the project article revealed, which was that finances was one of Vietnamese's biggest concerns and barriers to GHE uptake.

Interestingly, based on the project article, a number of people still remain sceptical about the value of periodic GHE, either finding them costly and without benefit or questioning their quality. Vietnamese patients are sometimes even sceptical of health professionals' expertise. Many have therefore suggested replacing GHE with more effective healthcare solutions.

To better understand this situation in Vietnam, current literature has studied the effect of patient's attitudes to the time, cost and quality of the medical service provided with regards to the frequency of GHEs, but reached conflicting results. An additional approach seeking to investigate behaviour and attitudes toward GHEs in Vietnam was also used by the study.

Since the current study's main insight was that cost concerns remain one of the main factors affecting peoples' uptake of GHE, our group was interested in exploring how much other underlying demographic factors contributed to Vietnamese's willingness to take up GHE, in the absence of cost concerns.

To eliminate cost influence, our group created a dependent variable "wouldtake_GHE" to replace UseMon, where all individuals in the study would have hypothetically received cash vouchers for GHE usage. The original variable UseMon looked at which option respondents would choose if provided cash for having GHEs: allsoon = use all the money to have a GHE soon, partly = use part of the money for a GHE and save the rest, later = take the money and have a GHE later. To make our new "wouldtake_GHE" variable binary, UseMon responses stating "allsoon" and "partly" were combined to return a value of "1" under "wouldtake_GHE", for these factors showed some definitive desire to take up GHE. The response "later" would return a value of "0", for this does not indicate commitment to take up GHE. Individuals are now faced with the choice to either use the vouchers for GHE, or not.

1.1 Understanding Vietnam and the environment

Vietnam is a country which retains strong traditional and cultural beliefs. The cultural context of Vietnam is thus important to understand to guide the direction of our analysis.

Through the various stages of Vietnam's growth and development, a strong culture of familism and higher family obligation values (Park, 2004) remains ingrained in the people. Stronger sense of responsibility towards family could thus potentially affect how an individual values their health, and thus uptake of GHE. Different people with different extents of family responsibilities could thus be anticipated to behave differently. On top of this, it is also worth noting that strong familial ties would mean that the family unit forms a large part of their environment and ideas they are exposed to. Experiences of family members could thus likely greatly influence their own perception with things too.

Considering the fact that Vietnam has also grown from being one of the world's poorest nations to a middle-income economy in one generation, this has also prompted consideration with regards to how educational levels might affect uptake of GHE. With GDP per capita increasing 3.6 times between 2002 and 2021, and

poverty rates declining from 14 in 2010 to 3.8 percent in 2020 (The world bank, 2022), general educational levels in Vietnam should be expected to rise. Intuitively, it can be understood that higher educational levels would allow individuals to better understand the value of healthcare, and thus take up GHE. However, given that there are still significant educational gaps which need to be improved on (Tran, Yang, 2022), it is not certain how strongly education would affect uptake of GHE in Vietnam. This is thus another factor that would be interesting to consider while exploring the data.

Vietnam has also started moving towards the usage of telehealth in recent years (Kiet, 2020) (The voice of Vietnam, 2020). It is hence also important to keep in mind how the model of healthcare deliverables is changing with time, and how the Vietnamese people may react to it. Given that educational levels can vary greatly within Vietnam, it is very likely that Vietnameses acceptance of technology usage would also differ. Thus, if the latest healthcare model continues to lean towards the employment of technology, the GHE uptake may vary according to educational levels of individuals too. Different educational levels may thus indirectly affect people's GHE uptake in this manner.

Overall, our group aims to gain better insight regarding which non-monetary reasons could be attributed to poor GHE uptake in Vietnam, while considering the unique underlying features of this population. Understanding how different people with demographic features react towards GHE in terms of uptake could help identify which subgroups the Vietnamese government may need to focus more efforts on, and thus establish the likely direction in which a future solution for this situation in Vietnam would take.

2.0 Data

This dataset was collected to understand the behaviours and tendencies that influence the attendance of the public for general health examinations in Vietnam. The intent behind the survey was to investigate the exact factors (Individual Characteristics or general perception) that potentially affect the demand for these GHE's.

For the given study, participants were chosen at random. Participants came from a wide variety of locations such as secondary schools, hospitals, companies, government agencies and randomly selected households in Hanoi, including Hospital 125 Thai Thinh (Dong Da District) and Vietnam-Germany Hospital (Hoan Kiem

District). Participants hence were representative of a variety of different characteristics and behaviour. Thus the sampling population would be a good representation of the general population without selection bias.

When considering potential issues with non respondents, we observe that the current dataset has quite a large number of missing observations from RecExam, RecPerExam, CHPerc and SuitExer. Omitting them however would cause the sample size to decrease drastically, and much information would be lost.

Another potential issue is the age demographic for their sample. The age group that had the largest number of respondents were aged 18-29 years old, followed by a significantly smaller but still sizeable group aged 30-39 years old. There were extremely few respondents aged above 50 years of age, and extremely few aged below 18 years of age (Appendix Item 2.1). The sample obtained is not representative of the general population, and hence the results obtained would not necessarily be generalizable across the Vietnam population.

To clean the data, our group has used data wrangling as well as feature engineering. This was mainly to remove the variables that were not relevant to our research question, and including them would not give us relevant information of interest, and could even potentially lead to misleading results.

Firstly, given that some age groups had quite a scarce number of respondents, we decided to reduce the number of age groups since those age groups would not be very significant on their own. We have changed the age groups from 5 factor levels to 3. Given the extremely few respondents aged <18 and >=50, we have thus combined those <18 and 18-29 to "<29", and those >=50 and 40-49 to ">=40" (Appendix Item 2.2)

Secondly, we have also reduced the factor levels for Jobstt (job status) from 6 to 3. We have grouped them into broader categories based on the implication of each status. Those with job status of "other","housewife" and "retirer" were grouped as "Unemployed" since all of them would have no active income and thus would likely have similar behaviour/rationale behind financial decisions. We also grouped those with "unstable" and "stable" jobs into "Employed", for they have active income.

Thirdly, we also changed Edu levels from 4 to 3, combining PostGrad (post-graduate) and Grad (college/university) into “MinGrad”.

Fourth, we have changed the scale variables (ImpressInfo, Reliability, Respon, Assurance, Empathy, PopularInfo, AttractInfo, SuffInfo, Tangibles) from 5 factor levels to 2. These variables originally had a score of any number between 1-5, but given that these are subjective numbers assigned by an individual, the exact value itself may not be very telling of how important people actually view it to be. Our focus was to find out if an individual had a good or bad opinion overall, while staying away from being too specific as that could give us misleading results. We hence grouped any value ≤ 2 together and named it “BelowAvg”, and “AboveAvg” otherwise. Similarly, we have changed the factor levels for CHPerc from 3 to 2 for the same reasons, combining “good" and “quite” (quite good) into just “good”.

Next, to tackle the large numbers of missing observations but still retain the most data we could, we first extracted out all the variables of relevance to us (non-cost related variables), before omitting those with missing observations. We selected a total of 38 non cost-related variables, namely "wouldtake_GHE", "Age_gr", "Sex", "Jobstt", "HealthIns", "BMI", "MaritalStt", "Edu" , "Wsttime", "Wstmon", "DiscDisease", "Lessbelqual", "NotImp", "HthyPriority", "Habit", "FlwHealth", "PerTrmt", "AcqTrmt", "StabHthStt" , "StChoise", "MedCabinet", "ExpCare", "ExamTools", "Tangibles", "Assurance", "Reliability", "Respon", "Empathy", "CHPerc", "SuitFreq", "SuffInfo", "AttractInfo", "ImpressInfo", "PopularInfo", "UseIT", "AfterIT", "Tooluseskills", and "EvalExer". If omission were to be done before the selection of relevant variables, the number of observations would drop drastically from 1638 to 1088, losing 550 observations. However with the extraction of relevant variables, omission of NA values only reduced our observations from 1638 to 1439, losing only 199 observations. The omission of the NA values also does not introduce bias or skew our data (the median values of the other variables do not change by much). This step hence allowed us to retain more information, and allow for LASSO and ridge analysis to be carried out.

With this expansive data collected, we are able to effectively analyse the more influential variables on the uptake of GHE, in the absence of cost concerns.

3.0 Analyses

3.1 Base Model

Our initial impression of the data and its variables, leads us to broadly bin the predictor variables into three general categories: (i) Personal characteristics of the respondent, (ii) Opinions/Perceptions of health service in Vietnam, and (iii) Attitudes towards personal health.

At an initial juncture we believe that all three categories could potentially affect a person's propensity to take up a GHE (if it were hypothetically, free of cost). Since our response variable "wouldtake_GHE" can have only 2 types of values: 0 (if they would not take the GHE) and 1 (if they would take the GHE), it is a binary outcome variable and logistic regression is the most appropriate statistical model to use. Thus, our base model is composed of our response variable "wouldtake_GHE", run in a logistic regression against all variables in the dataset without any interaction terms.

It is evident that this base model suffers in interpretability due to the large number of categorical variables and their factor-levels. Moreover, we identify that the variables in our base model could potentially be correlated to each other. For e.g., Education of the respondent could be correlated to their perception of mass media related to public health in Vietnam. Therefore, we start off our analysis with regularisation techniques (LASSO and Ridge) to build our initial two models.

Our aim is to test the predictive power and goodness-of-fit of various models. Logistic, Ridge, and LASSO are three types of regressions we have chosen to apply for our model analysis. We carried out all three regression approaches to identify which is the most appropriate for our research question.

3.2 Ridge regression

Primarily, a ridge regression serves as a shrinkage approach, i.e. we fit a model containing all predictors (in our case, 37) and then use the ridge technique to regularise the coefficient estimates towards zero. We also want our model to be useful in prediction in the future, i.e. we want to avoid overfitting our model on the data we currently have. Recognising this potential issue, regularisation techniques such as ridge (and lasso, which we also use) helps reduce high variance and improves model usability. Shrinkage of the coefficients is

achieved by penalising the regression with a penalty term called **L2-norm** (the sum of the squared coefficients).

The subsequent benefits of this approach are a) reduction in variables (improving our model's interpretability and application), and b) dealing with any multicollinearity in our model.

Our the outcome of base logistic regression model generates multiple coefficients for all the variables (a problem that is compounded when you consider that the majority of predictor variables in our analysis are factor variables, and hence will produce [(number of factor levels) -1] coefficients for each factor variable). Thus, we explore applying a ridge regression to our base model.

Data selection:

As mentioned earlier, we choose a subset of the original dataset (47 variables) and drop 9 variables (AffCost, Height, Weight, RecPerExam, RecExam, SuitFreq, ComSubsidy, Age). We dropped "height" and "weight" because we use "BMI " as a proxy instead. "Height" and "weight" are also naturally, highly correlated with "BMI ". Similarly we dropped "Age" as we instead use "Age_gr". The rest were dropped solely because they are not relevant to our analysis as they are directly related to the cost of GHE or cost-related demand of GHE. Hence, any analysis of these variables would not yield beneficial inference. We also centre and scale the BMI variable as it is numeric.

We initially approach the ridge analysis by looking for indicators of multicollinearity in our model.

During our exploratory data analysis, we ran a logistic regression on the whole dataset and the resultant output displayed extremely high standard errors associated with each variable's coefficient estimate (Appendix 3.2.1), which indicated potential correlation between our predictor variables. However when instead, we run our base logistic regression (dropping the irrelevant variables), the standard errors significantly shrunk to values <1 for all variables. This seems to indicate that our reduced dataset with the fewer variables has less correlation amongst the predictors.

Further, the VIF graph of our ridge regression shows a near constant value seemingly independent of lambda, and well below the industry standard of VIF = 10. This is another indicator of reduced correlation in our dataset (Appendix of 3.2.2).

Consequently, the VIF graph does not help us choose a suitable value for lambda to use for our ridge model. We use a 10-fold cross-validation as an alternative way to determine a suitable lambda value. Using cross-validation we run multiple iterations of fitting the regression on the training sets and validating it against a test set. We run multiple iterations of this on a hundred lambda values. The lambda values correspond to different error percentages of the trained model on the test set. Our aim is to select the lambda value that produces the lowest error rate. Once we obtain the lambda value (Appendix 3.2.3), we fit the model to our entire dataset and obtain the coefficients of the ridge model.

We also note that ridge regression does not *select* variables, it only shrinks them *towards* 0. So our model still contained all the predictor variables albeit showing differing levels of effect on our response variable.

3.3 Lasso

Our next analysis approach is using a second regularisation method - LASSO (Least Absolute Shrinkage and Selection Operator). Similar to a ridge regression, a LASSO also uses a *cost function* that it penalises. Where ridge uses the squares of the coefficients, LASSO uses their magnitude. The major difference between the two dimension shrinkage methods is that: where ridge only shrinks coefficients to near 0 values, LASSO can eliminate them completely (i.e. shrink coefficient values completely to 0). Thus, LASSO can be used as a variable selection approach.

Once again, we use the cross-validation approach to select the lambda used in our LASSO model. Selecting, from a range of lambda values, the best lambda is selected as the one that produces the smallest error rate when the training set data fitted model is run with the validation set in multiple iterations (Appendix 3.2.4). We use this lambda to run the regression on the entire dataset and obtain the coefficients of the variables retained after LASSO. Our model significantly reduces in terms of predictor variables with LASSO, and inferences from important variables in our dataset can now be drawn.

Furthering our understanding of the effect of the predictor variables on our response variable, we want to understand the effect of interaction terms on our dependent variable. We also want to implement contextual understanding of the status quo of Vietnamese people and its public health space into our model, and explore if we can further refine it.

3.3 Logistic regression - Domain knowledge

Model Selection

The next set of models uses expert and domain knowledge to come up with a strong model to investigate. Using our domain knowledge, we took into account the methods of data collection, context, and other important factors to understand which specific independent variables influence the predictor variable the most. Through the research conducted we found that a combination of personal characteristics and perceptions tend to influence our dependent variable, with an emphasis placed on the former. As shown in the study," (Vuong, 2017) nearly 52% of participants who were reluctant to attend GHEs stated *Wsttime (Waste of time)* as one of the most common reasons. Whereas "among the 81% of participants who were prepared to attend GHEs, the main reason given was *HthyPriority* (Health was a priority)" (Vuong, 2017) . We decided to select a combination of factors we felt most influenced the demand for GHE's. These factors were then selected to be the independent variables in our base regression model. The chosen independent variables are: *Edu, Sex, BMI,, HealthIns, Wsttime, Wstmon, FlwHealth, PerTrmt, ExamTools, CHPers, Habit, StabHthStt, MedCabinet, Tooluseskills, ExamTools, SuffInfo, ImpressInfo and UseIT.*

The following is the logistic regression model with all the chosen predictor variables:

Model 3

$$log(Would\_take\_GHE) = a + b(MaritalStt) + c(Edu) + d(Sex) + e(BMI) + f(HealthIns) + g(Wsttime) + h(Wstmon) + i(HthyPriority) + j(FlwHealth) + k(PerTrmt) + l(ExamTools) + m(CHPerc) + n(Habit) + o(StabHthStt) + p(MedCabinet) + q(Tooluseskills) + r(ExamTools) + s(SuffInfo) + t(ImpressInfo) + u(UseIT)$$

Here the coefficients (a,b,c…,t) of the predictors is the log odds ratio. Using the exponential (np.exp()) of this coefficient, we can obtain the odds ratio. For categorical independent variables, "the odds ratio compares

the odds of the event occurring for each category of the predictor relative to the reference category, given that all other variables remain constant" (Dickson, 2020).

Considering some interesting relationships between the variables in our dataset (Appendix Item 3.3.1), we worked upon the above logistic regression to model the relevant interaction terms. The next few sets of models take into account a combination of the independent variables in the base logistic model with different interaction terms.

Model 4

$$log(Would\_take\_GHE) = a + b(MaritalStt) + c(Edu) + d(Sex) + e(BMI) + f(HealthIns) + g(HthyPriority) + h(PerTrmt) + i(MedCabinet) + j(ExamTools) + k(Tooluseskills) + l(UseIT) + m(Edu*UseIT) + n(MedCabinet*Tooluseskills)$$

Model 5

$$log(Would\_take\_GHE) = a + b(MaritalStt) + c(Edu) + d(Sex) + e(BMI) + f(HealthIns) + g(\text{PerTrmt}) + h(\text{ExamTools}) + i(\text{Habit}) + j(\text{StabHthStt}) + k(Tooluseskills) + l(UseIT) + m(CHPerc) + n(\text{HthyPriority}) + o(\text{CHPerc*StabHthStt}) + p(\text{Edu*UseIT}) + q(\text{MaritalStt*Sex}) + r(\text{MaritalStt*HthyPriority}) + s(\text{NotImp*HthyPriority})$$

Model 6

$$log(Would\_take\_GHE) = a + b(MaritalStt) + c(Edu) + d(Sex) + e(BMI) + f(HealthIns) + g(\text{PerTrmt}) + h(\text{ExamTools}) + i(\text{Habit}) + j(\text{StabHthStt}) + k(\text{Tooluseskills}) + l(\text{AcqTrmt}) + m(\text{UseIT}) + n(\text{Wsttime}) + o(\text{Wstmon}) + p(\text{CHPerc}) + q(\text{HthyPriority}) + r(\text{NotImp}) + s(\text{Wsttime*Wstmon}) + t(\text{CHPerc*StabHthStt}) + u(\text{Edu*UseIT}) + v(\text{MaritalStt*Sex}) + w(\text{MaritalStt*HthyPriority}) + x(\text{NotImp*HthyPriority})$$

4.0 Results

4.1 Model Predictive Power : Model Performance

For the purpose of our report, the model predictive power or the model performance is measured in terms of its accuracy to predict the occurrence of our Y variable(would_take_GHE). A more accurate model is associated with a higher performance. By plotting the ROC (Receiver Operating Characteristic) curve, we can obtain our AUC value (the area under the ROC curve). Using the pROC package, we plot the true positive rate against the false positive. Generally, a model with good predictive ability tends to have an AUC value that is close to 1. In order to find the model with the strongest predictive power, we plotted the ROC of all 5 Models and derived the corresponding AUC, using this as our metric for comparison.

4.2 Model Interpretability

Model interpretability gives us insight into the relationship between the independent and dependent variables. A model with higher interpretability helps us understand why exactly the independent variables predict our dependent attribute. One of our objectives through this analysis is to generate useful and relevant inferences and methods that can be scaled into tangible actions (such as targeted media campaigns, etc). In order to objectively assess our models in terms of interpretability as well as remove possible subjective bias in our evaluation, we have outlined three criteria that constitute model interpretability. Along with testing the model predictive power, we will also apply this interpretability framework to all our models.

Objective measure for interpretability :

1) **Simplicity of model:** This is captured by the number of independent variables within the model. The higher the number of predictor terms in the final model, the lower its simplicity score.
2) **Comprehensibility:** This refers to the ability to understand a model and explain the resulting outcomes in ways that are understandable to the layman. We chose this criteria in order to ensure scalability and usefulness of the model. We want to include variables that are practical and easier to collect.
3) **Accuracy of measurement:** This is captured by the measure of reliability/accuracy of the survey data collected for the variables included in the final model. Here, numeric or dichotomous answers that are more specific than subjective rating or opinion answers are preferred in our chosen final model.

4.3 Trade off between predictability and interpretability

Although both predictability and interpretability are important indicators when choosing our best model, the complication arises because as model predictability increases so does the complexity of a model, at the expense of interpretability. While recognising the importance of predictive power of our model, we also want to maintain a high measure of interpretability of the final model. Therefore our chosen final model attempts to strike a balance between these two measures.

4.4 Choosing Our Best Model

4.4.1 Regularisation Approach

1) **Ridge (Model 1)**

   a) **Output**

   Using cross-validation (Appendix Item 3.2.3), we obtain the lambda that corresponds to the least error. This lambda value (best lambda = 0.1158208) is used to fit the final ridge model with the coefficients as seen in the output (Appendix Item 4.4.1.1) The confusion matrix showed a 78.18% accuracy rate in prediction.

   b) **Inference**

   I. Through coefficient shrinkage done by ridge regression we see many of the rating-related variables (such as Empathy, Tangibles, AttractInfo, etc) have extremely low coefficient values.

   II. Habit, HealthIns, Wstmon are a few of the variables that have comparatively high coefficients in the model. Wstmon has a positive coefficient which implies that people who think GHE's are a waste of money, are very likely to take a GHE if the GHE costs no money (captured as woultake_GHE = 1).

   c) **Predictive Power**

   Using (Appendix Item 4.4.1.2), we were able to plot the ROC graph for this model. The AUC value we obtained equals 0.7084371.

   d) Interpretability

   I. **Simplicity**: This model still retains all 37 variables albeit some variables' coefficients are near zero.

   II. **Comprehensibility:** The model suffers in comprehensibility as a good understanding of statistics is necessary to draw inference from the ridge model.

   III. **Accuracy of Measurement:** Has a mix of objective and subjective variables

2) **Lasso (Model 2)**

a) Output

Using cross-validation (Appendix Item 3.2.4), we obtain the lambda that corresponds to the least error. Using this lambda value (bestlambda = 0.01105565), the final fitted LASSO model on the whole dataset (analysis5) shows that it cuts down on the variables from 37 to 19 predictors (Appendix Item 4.4.1.3). The retained variables are: Sex, HealthIns, Wstmon, Lessbelqual, NotImp, HlthyPriority, Habit, FlwHealth, PerTrmt, AcqTrmt, StabHthStt, ExamTools, CHPerc, SuitFreq, ImpressInfo, UseIT, Tooluseskills, EvalExer, BMI. The confusion matrix showed a 77.83% accuracy rate in prediction.

b) Inference:

I. Ceteris paribus, a male is less likely to to take up a GHE as compared to a female. Additionally, respondents with prior history of medical treatment have a higher probability of taking GHE than those who don't.

II. Attitude towards health: Low frequency of exercise has a negative effect on probability of GHE uptake. Interestingly, we see that if one has a medical cabinet at their home, or if they regularly take medical measurements then they are unlikely to take up GHEs'. This may be because they do not find unique value in GHEs beyond basic treatments.

III. A positive perception of the state of public/personal health seems to reduce the likelihood of taking up a GHE. As inferred from the CHPerc and StabHthStt, it could be hypothesized that GHE is not necessary if one feels generally healthy.

c) Predictive Power

Using (Appendix Item 4.4.1.4), we were able to plot the ROC graph for this model. The AUC value we obtained equals 0.701265.

d) Interpretability

I. **Simplicity**: This model is considerably simpler as it reduces from 37 terms to 19.

II. **Comprehensibility:** A majority of the variables are scalable and easy to collect/reproduce. (e.g. Sex, BMI, HealthIns, etc)

III. **Accuracy of Measurement:** A majority of the variables are binary/numerical in nature

4.4.2 Logistic Regression with Interaction terms

1) **Model 4**

   a) **Output**

   Using the glm() function and specifying the parameter to be family = binomial (Appendix Item 4.4.2.1), we obtained the AIC value of 1456.6. The Akaike information criterion (AIC) tells us how well the model fits the data. Both the interaction terms, (Edu*UseIT) and (MedCabinet*Tooluseskills) are shown to be statistically significant after running the logistic regression.

   b) **Inference**

   Since the interaction terms are shown to be statistically significant we can infer:

   I. An individual's level of education (Edu**)** affects their willingness to use IT in order to detect health problems (UseIT), and hence influencing their uptake of GHE

   II. A respondent keeping a medical cabinet(MedCabinet) at home could depend upon whether or not they have the skills to use basic medical equipment (Tooluseskil), ultimately influencing the GHE attendance as well.

   c) **Predictive power**

   Using (Appendix Item 4.4.2.2), we were able to plot the ROC graph for this model. The AUC value we obtained equals 0.682.

   d) **Interpretability**

   I. **Simplicity**: This model has 11 independent variables and 2 interaction terms

   II. **Comprehensibility:** The model is relatively straightforward, the independent variables are mostly based on individual characteristics. Could be interpreted by someone with no expertise in statistics.

III. **Accuracy of Measurement:** Out of the 11 independent variables, all 11 are either numeric/ dichotomous answers

2) Model 5

a) **Output**

Using (Appendix Item 4.4.2.3), we obtained the AIC value of 1432.1. We can see that the AIC value for this model is lower than that of model 4. This could imply that model 5 better fits the data. The interaction terms (Edu*UseIT),(CHPerc*StabHthStt) and (MaritalStt*Sex) are shown to be statistically significant after running the logistic regression.

b) Inference

From the interaction terms are shown to be statistically significant we can infer:

I. As explained in model 4, education and willingness to use IT in order to detect health problems could be correlated.

II. If a respondent and their family all enjoyed good health StabHthStt, then they had a positive opinion on public health(CHPerc), impacting their decision to uptake GHE.

III. There seems to be a correlation between the uptake of GHE and the participants MaritalStt and Sex.

c) Predictive power

Using (Appendix Item Fig 4.4.2.4), we were able to plot the ROC graph for this model. The AUC value we obtained equals 0.710 (2nd highest overall)

d) Interpretability

I. **Simplicity**: This model has 13 independent variables and 5 interaction terms

II. **Comprehensibility:** Slightly more complex that model 4 but could still be interpreted by someone with no expertise in statistics.

III. **Accuracy of Measurement:** Out of the 13 independent variables, 9 are either numeric/ dichotomous answers and only 1 is an opinion answer.

3) Model 6

a) **Output :** Using (Appendix Item Fig 4.4.2.5), we obtained the AIC value of 1429.3. We can see that this model has the best AIC relative to the other two models. This could imply that

model 6 best fits the data. The interaction terms (Edu*UseIT),(CHPerc*StabHthStt) and (MaritalStt*Sex) are shown to be statistically significant after running the logistic regression.

b) Inference

Given that model 5 and model 6 have the same significant interaction terms, we can infer the same conclusions from model 5

c) **Predictive power**

Using (Appendix Item Fig 4.4.2.6), we were able to plot the ROC graph for this model. The AUC value we obtained equals 0.717 (highest overall)

d) Interpretability

I. **Simplicity**: This model has 17 independent variables and 6 interaction terms

II. **Comprehensibility:** Model 6 is the most complex out of all the model, the numerous independent variables and interaction term make it more difficult to understand, therefore someone with no knowledge of statistics might face trouble.

III. **Accuracy of Measurement:** Out of the 17 independent variables, 17 are either numeric/ dichotomous answers and only 1 is an opinion answer.

5.0 Conclusion:

Considering both interpretability as well as its predictive power, we choose Model 5 as the best model. The predictive power of Model 5 is almost equal to the highest AUC among all models. The small loss in predictive power (compared to Model 6) is more than compensated in its dominance in interpretation. Both the LASSO and Ridge models are lower in predictive power when compared to Models 5 and 6. Model 4 is arguably the simplest, but has the lowest predictive power. Thus, Model 5 is the best balance of high predictive power and high interpretability. Through our analysis, we ascertain that Model 5 will be the most useful in predicting how existing characteristics (eg. attitudes, status, medical experiences etc.) would influence the uptake of a GHE exam in Vietnam, and consequently improve understanding what additional measures the government would need to take to increase GHE uptake.

6.0 Discussion:

6.1 Implications from results

Model_5

Under model 5, we saw that the variable "PerTrmtyes" (whether a respondent receives long-term medical treatment) showed a positive significance against would_take_GHE. Intuitively, we would think that someone who is receiving long-term medical treatment would be suffering from a long term medical condition. Thus, having regular health checkups would be especially important to them in detecting any potential relapses, and ensuring that their current condition is stable and well managed. This could help save future medical costs due to early detection.

In addition, we find that people who are unwilling to use IT to detect health problems have a negative correlation with would_take_GHE. The study indicates that people in Vietnam tend to find it more reliable to seek thorough and detailed opinions through health examinations after having dealt with the health problem themselves using IT technology. To increase GHE uptake in Vietnam, the government should focus their efforts on convincing people to accept IT usage in healthcare through allowing them to experience the benefit firsthand (The voice of Vietnam, 2020), for convincing is always best done through experience.

Another interesting thing to note is our study showed a negative correlation between being an unmarried male and would_take_GHE. Intuitively, considering the strong culture of familism in Vietnam, this could be because unmarried men do not have as heavy family responsibilities as compared to married men. Without children or wives to care for, they become less cautious of their health due to lighter responsibilities. Thus, the government should focus on increasing GHE uptake amongst unmarried men as they are more likely to fall through the cracks in terms of healthcare due to the lack of self incentive.

"Tooluseskill", "HthyPriority" and "ChPerc" also, had very clear reasoning for their distinct significances. Tooluseskill had a negative impact on would_take_GHE, since having the skills to use basic medical equipment would mean you are not reliant on attending GHEs for minor health issues. While HthyPriority has a positive effect on GHE attendance, since having a higher priority of personal health would mean

wanting regular health checkups. CHPerc, has a negative impact on would_take_GHE, since people who generally have a good perception on their public health would not find it a need to go for GHE.

6.2 How results compare to existing findings

Previous findings suggest that GHE attendance could be improved by raising the budget for healthcare schemes. Our analysis, however, took a deeper look into non-money related factors. Hence, we managed to find various significant underlying variables that concurrently contributed to GHE uptake. Our findings thus contributed an extra layer of depth to the existing report in terms of understanding GHE attendance variation.

6.3 Broader context of Results

In a broader context, the significance of personal characteristics, attitudes towards personal health, and the general perception of Vietnam health service quality, are serious factors for the Vietnam health system to face. It would not be sufficient to solve this by only providing subsidies and increased government spending in healthcare, but it is also necessary to adjust government policies accordingly to take into greater account the attitude and characteristics of personal health of the population in Vietnam.

6.4 How the study could have been improved

The biggest drawback to the study was the biases incurred. For example, in order to choose between regression models, multiple measures that try to balance a mix of having a low bias and a low variance can be used, such as the AIC value. This, though, creates regularisation bias, which leads to the overlooking of inferences.

Moreover, we combined response categories to make their variables binary, we assumed there is not much difference between the similar categories regarding the impact against the dependent variable. Although most of the chosen variables were not statistically significant, we lost potential inferences from the originally more specific categories.

Furthermore, our variable of interest, wouldtake_GHE, was derived from combining the “allsoon” and “partly” categories so that the variable would become binary. Considering this variable is the focus of our

analyses and we have assumed the same correlation within “allsoon” as within “partly”, there lies a critical possibility that a direct reporting bias exists.

There is also the possibility that Omitted Variable Bias exists as we excluded money related variables. Although this reduced our standard errors, the drawbacks still need to be acknowledged.

All things considered, if respondents were given more explicit questions with a reduced scale, there would have been much lower biases and especially a lower measurement bias.

7.0 Appendix

Fig 1.1.1 (Proportion of respondents from the respective age groups)

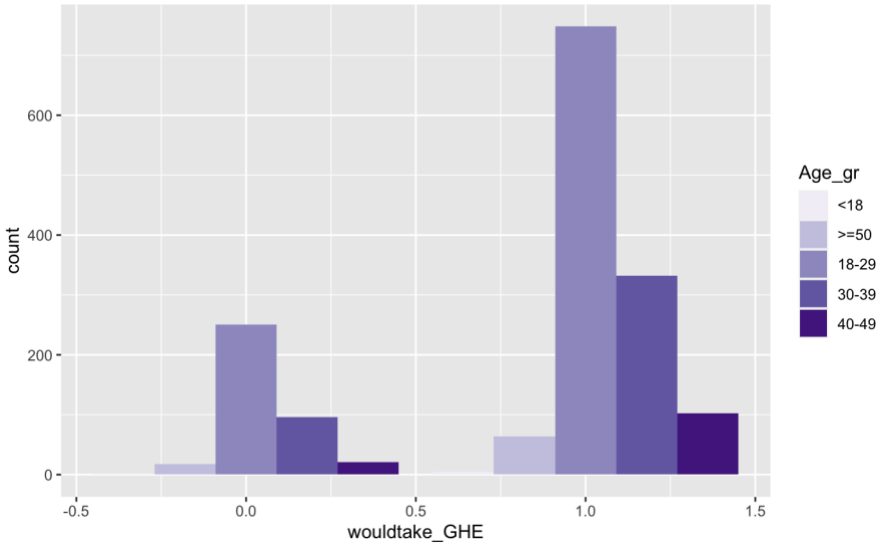


Fig. 1.1.2 (Proportion of respondents from the new respective age groups)

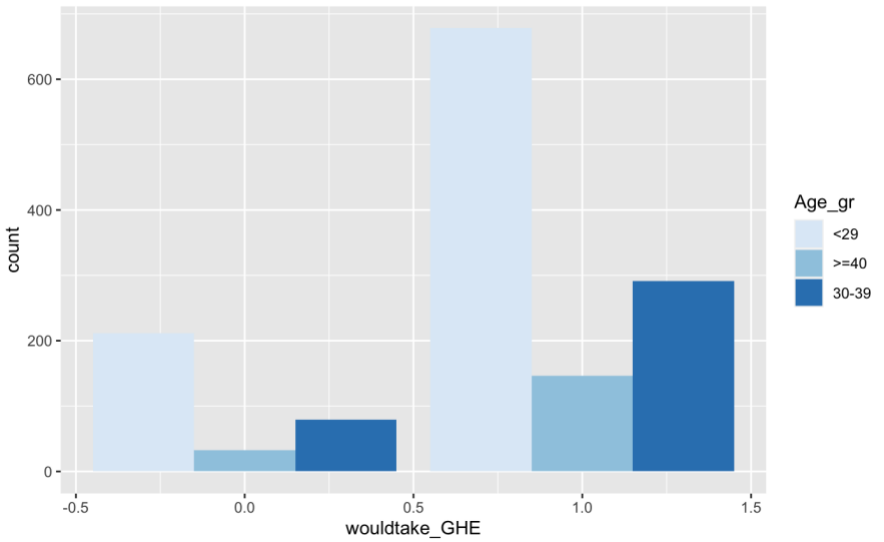
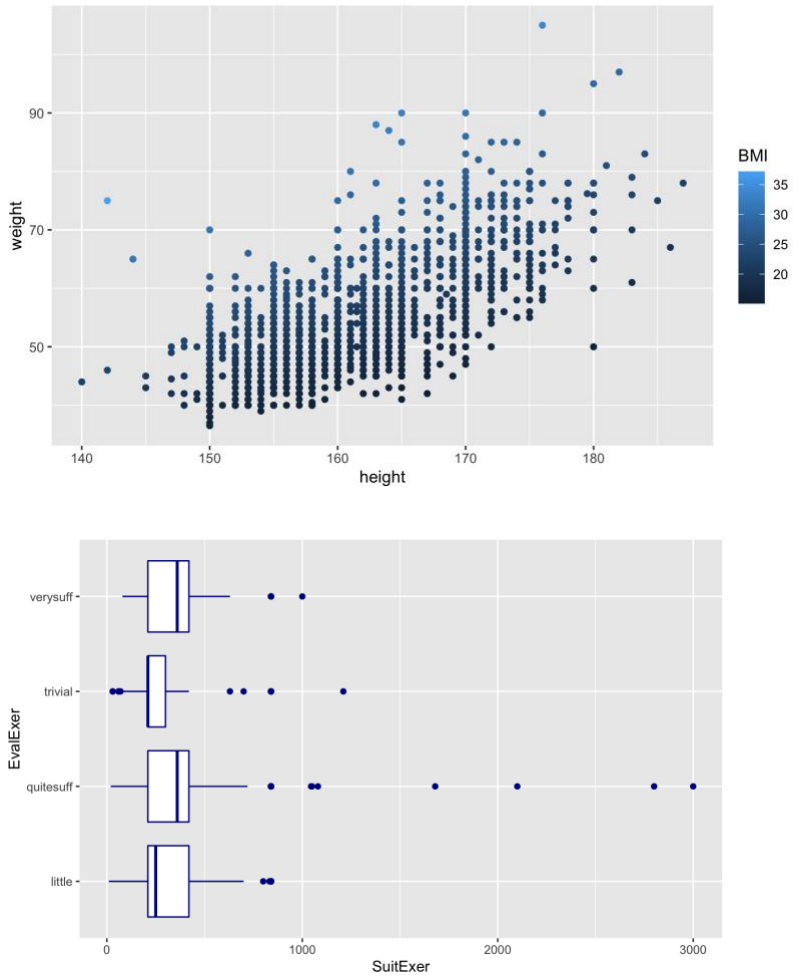


Fig 2.3 Correlation between: i. height, weight and BMI and ii. SuitExer and EvalExer



2.4 Chi-Sq Tests: To show correlation between variables

```r
chisq.test(analysis$SuitExer, analysis$EvalExer)
```

```
Warning in chisq.test(analysis$SuitExer, analysis$Ev
  Chi-squared approximation may be incorrect

        Pearson's Chi-squared test

data:  analysis$SuitExer and analysis$EvalExer
X-squared = 292.17, df = 210, p-value = 0.0001518
```

```r
chisq.test(analysis$height, analysis$BMI)
```

```
Warning in chisq.test(analysis$height, analysis$BMI)
  Chi-squared approximation may be incorrect

        Pearson's Chi-squared test

data:  analysis$height and analysis$BMI
X-squared = 74860, df = 27370, p-value < 2.2e-16
```

```r
chisq.test(analysis$RecExam, analysis$RecPerExam)
```

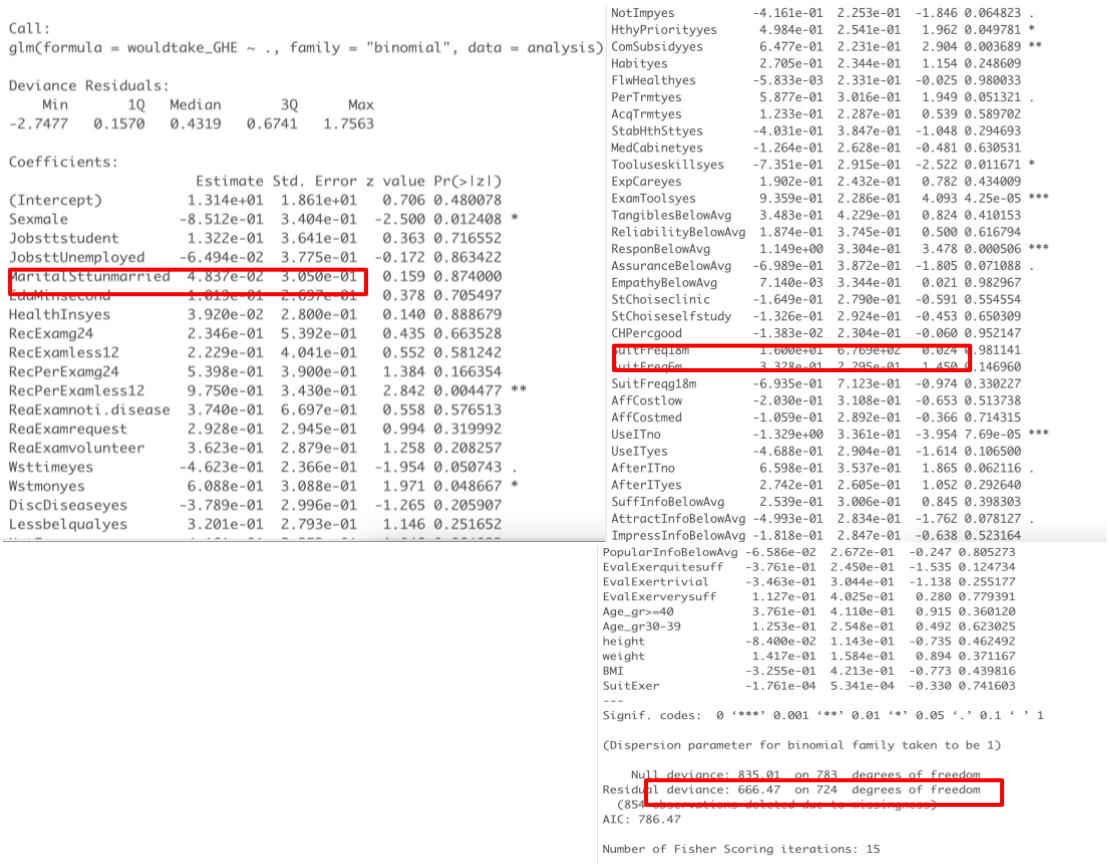
```
        Pearson's Chi-squared test

data:  analysis$RecExam and analysis$RecPerExam
X-squared = 466.85, df = 4, p-value < 2.2e-16
```

```r
chisq.test(analysis$RecPerExam, analysis$wouldtake_GHE)
```

```
        Pearson's Chi-squared test

data:  analysis$RecPerExam and analysis$wouldtake_GHE
X-squared = 28.8, df = 2, p-value = 5.575e-07
```

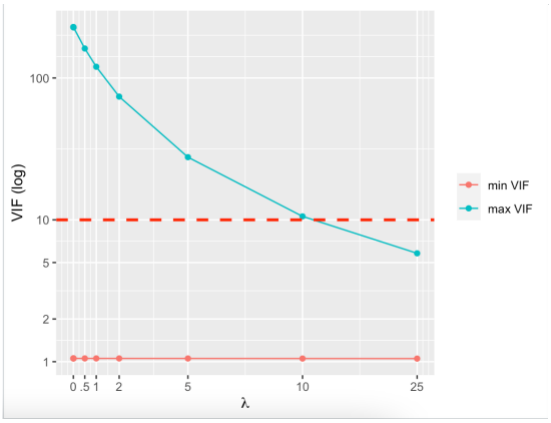
3.2.1. Regression Results of Whole Dataset - High Std Errors

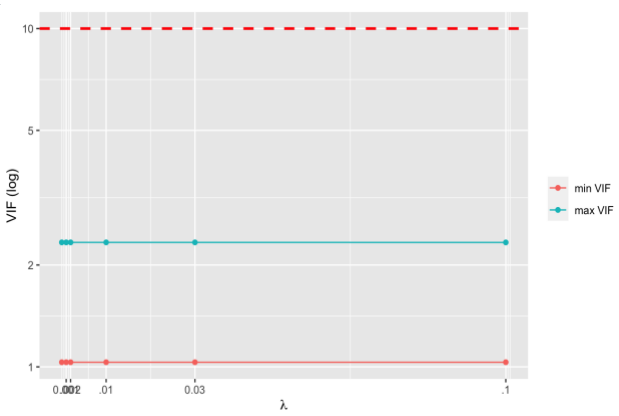
#running on the whole dataset to show the std error
dt_ch <- glm(wouldtake_GHE~. , data=analysis, family="binomial")
summary(dt_ch)



```
Call:
glm(formula = wouldtake_GHE ~ ., family = "binomial", data = analysis)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7477   0.1570   0.4319   0.6741   1.7563

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          1.314e+01  1.861e+01   0.706 0.480078
Sexmale             -8.512e-01  3.404e-01  -2.500 0.012408 *
Jobsttstudent        1.322e-01  3.641e-01   0.363 0.716552
JobsttUnemployed    -6.494e-02  3.775e-01  -0.172 0.863422
MaritalSttunmarried  4.837e-02  3.050e-01   0.159 0.874000
EduMinSecond         1.019e-01  2.697e-01   0.378 0.705497
HealthInsyes         3.920e-02  2.800e-01   0.140 0.888679
RecExamg24           2.346e-01  5.392e-01   0.435 0.663528
RecExamless12        2.229e-01  4.041e-01   0.552 0.581242
RecPerExamg24        5.398e-01  3.900e-01   1.384 0.166354
RecPerExamless12     9.750e-01  3.430e-01   2.842 0.004477 **
ReaExamnoti.disease  3.740e-01  6.697e-01   0.558 0.576513
ReaExamrequest       2.928e-01  2.945e-01   0.994 0.319992
ReaExamvolunteer     3.623e-01  2.879e-01   1.258 0.208257
Wsttimeyes          -4.623e-01  2.366e-01  -1.954 0.050743 .
Wstmonyes            6.088e-01  3.088e-01   1.971 0.048667 *
DiscDiseaseyes      -3.789e-01  2.996e-01  -1.265 0.205907
Lessbelqualyes       3.201e-01  2.793e-01   1.146 0.251652
NotImpyes           -4.161e-01  2.253e-01  -1.846 0.064823 .
HthyPriorityyes      4.984e-01  2.541e-01   1.962 0.049781 *
ComSubsidyyes        6.477e-01  2.231e-01   2.904 0.003689 **
Habityes             2.705e-01  2.344e-01   1.154 0.248609
FlwHealthyes        -5.833e-03  2.331e-01  -0.025 0.980033
PerTrmtyes           5.877e-01  3.016e-01   1.949 0.051321 .
AcqTrmtyes           1.233e-01  2.287e-01   0.539 0.589702
StabHthSttyes       -4.031e-01  3.847e-01  -1.048 0.294693
MedCabinetyes       -1.264e-01  2.628e-01  -0.481 0.630531
Tooluseskillsyes    -7.351e-01  2.915e-01  -2.522 0.011671 *
ExpCareyes           1.902e-01  2.432e-01   0.782 0.434009
ExamToolsyes         9.359e-01  2.286e-01   4.093 4.25e-05 ***
TangiblesBelowAvg    3.483e-01  4.229e-01   0.824 0.410153
ReliabilityBelowAvg  1.874e-01  3.745e-01   0.500 0.616794
ResponBelowAvg       1.149e+00  3.304e-01   3.478 0.000506 ***
AssuranceBelowAvg   -6.989e-01  3.872e-01  -1.805 0.071088 .
EmpathyBelowAvg      7.140e-03  3.344e-01   0.021 0.982967
StChoiseclinic      -1.649e-01  2.790e-01  -0.591 0.554554
StChoiseselfstudy   -1.326e-01  2.924e-01  -0.453 0.650309
CHPercgood          -1.383e-02  2.304e-01  -0.060 0.952147
SuitFreq18m          1.600e+01  6.769e+02   0.024 0.981141
SuitFreq6m           3.328e-01  2.295e-01   1.450 0.146960
SuitFreqg18m        -6.935e-01  7.123e-01  -0.974 0.330227
AffCostlow          -2.030e-01  3.108e-01  -0.653 0.513738
AffCostmed          -1.059e-01  2.892e-01  -0.366 0.714315
UseITno             -1.329e+00  3.361e-01  -3.954 7.69e-05 ***
UseITyes            -4.688e-01  2.904e-01  -1.614 0.106500
AfterITno            6.598e-01  3.537e-01   1.865 0.062116 .
AfterITyes           2.742e-01  2.605e-01   1.052 0.292640
SuffInfoBelowAvg     2.539e-01  3.006e-01   0.845 0.398303
AttractInfoBelowAvg -4.993e-01  2.834e-01  -1.762 0.078127 .
ImpressInfoBelowAvg -1.818e-01  2.847e-01  -0.638 0.523164
PopularInfoBelowAvg -6.586e-02  2.672e-01  -0.247 0.805273
EvalExerquitesuff   -3.761e-01  2.450e-01  -1.535 0.124734
EvalExertrivial     -3.463e-01  3.044e-01  -1.138 0.255177
EvalExerverysuff     1.127e-01  4.025e-01   0.280 0.779391
Age_gr>=40           3.761e-01  4.110e-01   0.915 0.360120
Age_gr30-39          1.253e-01  2.548e-01   0.492 0.623025
height              -8.400e-02  1.143e-01  -0.735 0.462492
weight               1.417e-01  1.584e-01   0.894 0.371167
BMI                 -3.255e-01  4.213e-01  -0.773 0.439816
SuitExer            -1.761e-04  5.341e-04  -0.330 0.741603
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 835.01  on 783  degrees of freedom
Residual deviance: 666.47  on 724  degrees of freedom
  (854 observations deleted due to missingness)
AIC: 786.47

Number of Fisher Scoring iterations: 15
```

3.2.2 (VIF graphs for whole dataset vs sub-dataset(preferred))





3.2.3 (Cross-Validation on Ridge to Find Best Lambda)

set.seed(42)
cvrr.out <- cv.glmnet(x, y, alpha=0, family ="binomial") #obtain the 10-fold cross validation
bestlam_rr <- cvrr.out$lambda.min #identify the lambda for smallest CV error
bestlam_rr

```
> bestlam_rr <- cvrr.out$lambda.min  #identify the lambda for smallest CV error
> bestlam_rr
[1] 0.09615644
>
```

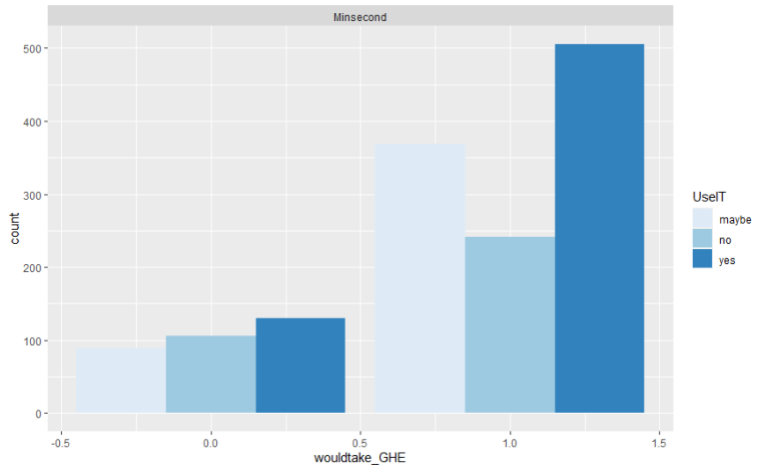
3.2.4 (Cross-Validation on LASSO to Find Best Lambda)

set.seed(1805)
cv.out <- cv.glmnet(x, y, alpha=1, family = "binomial")
plot(cv.out)
bestlam <- cv.out$lambda.min #finding best lambda using cv
bestlam

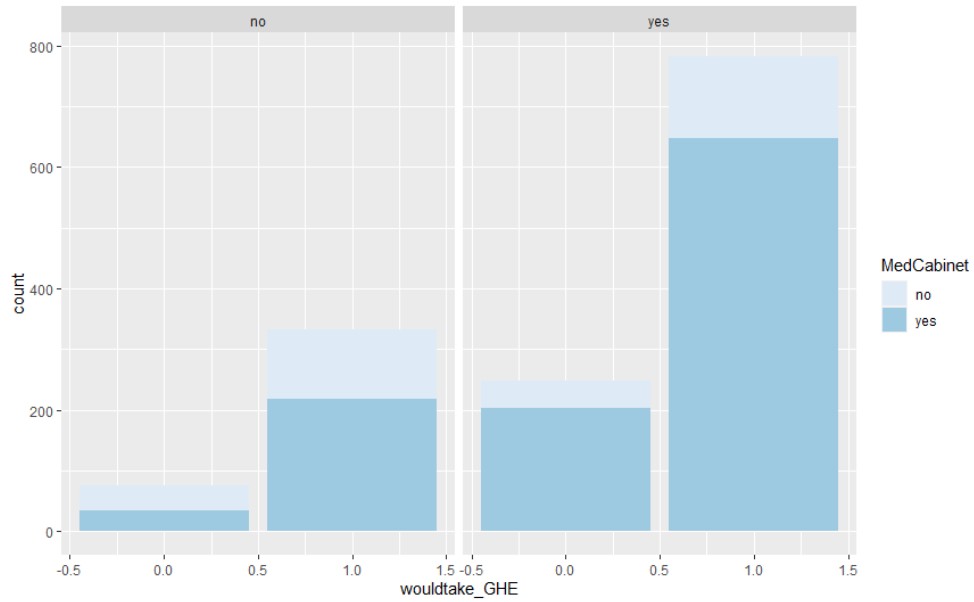
```
> bestlam <- cv.out$lambda.min #finding best lambda using cv
> bestlam
[1] 0.01105565
>
```

3.3.1 (Interesting Relationships between Predictors)

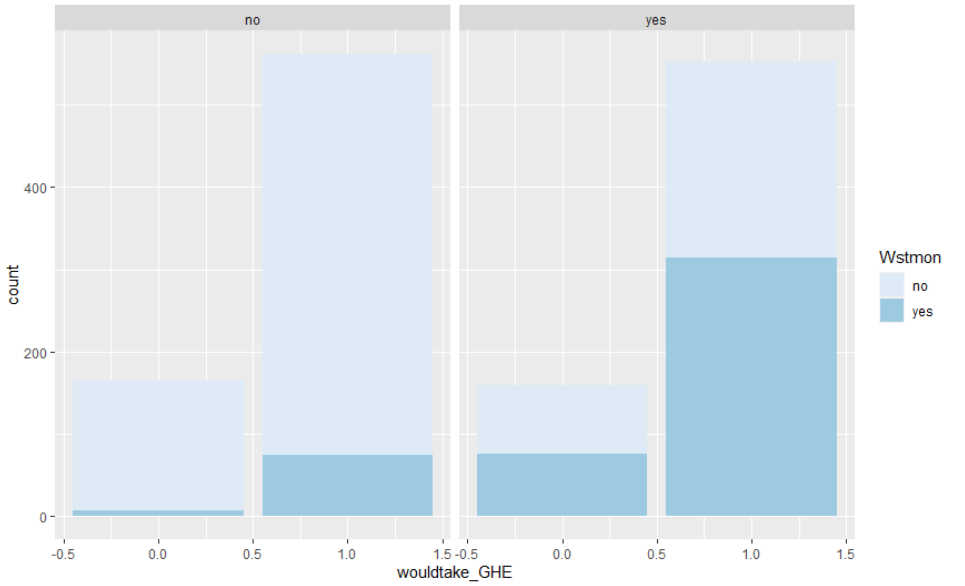
Interaction term 1 : Edu and UseIT



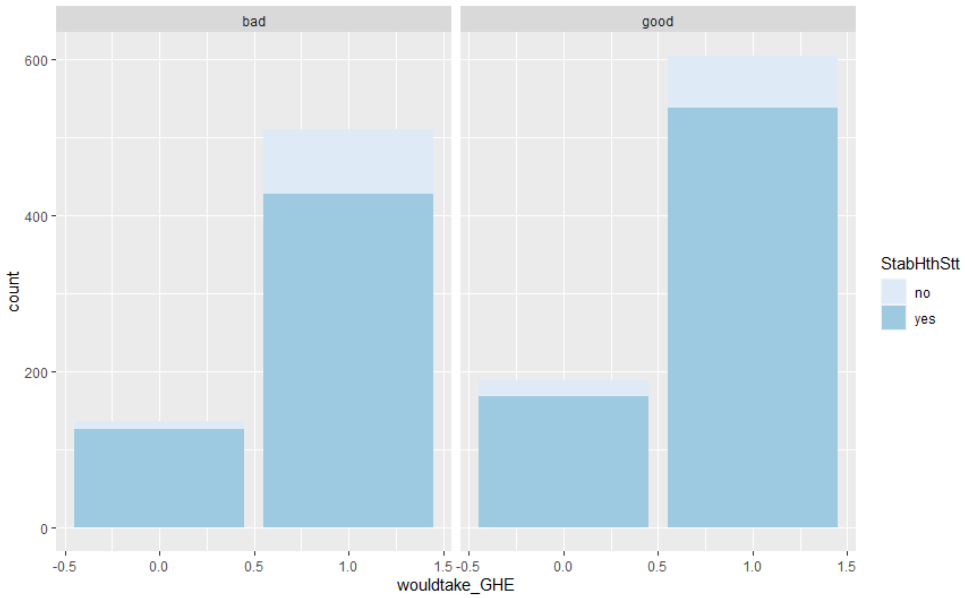
Interaction term 2 : MedCabinet and Tooluseskills



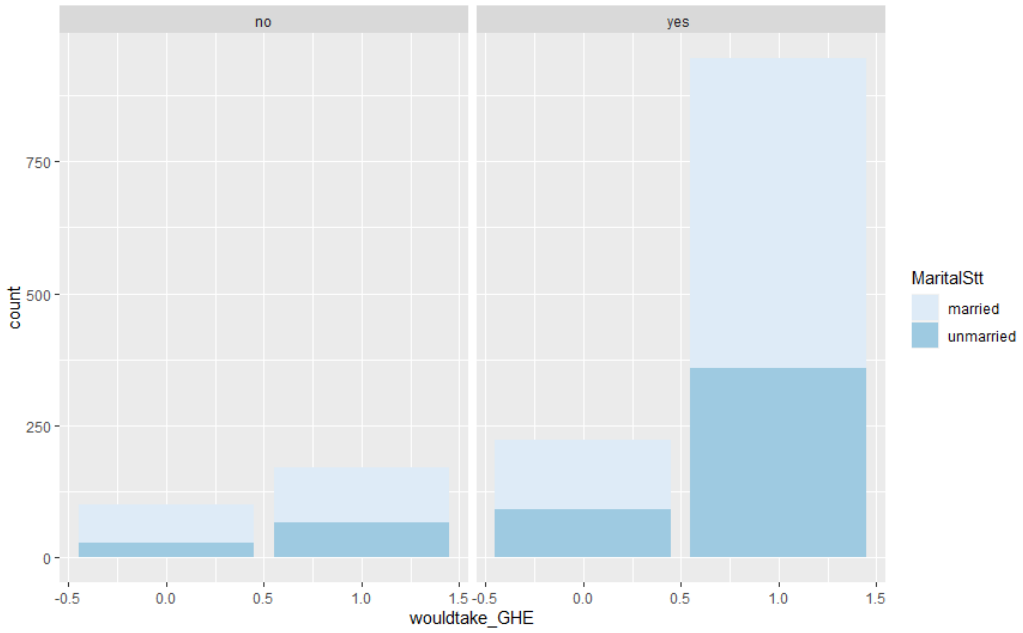
Interaction term 3: Wsttime and Wstmon



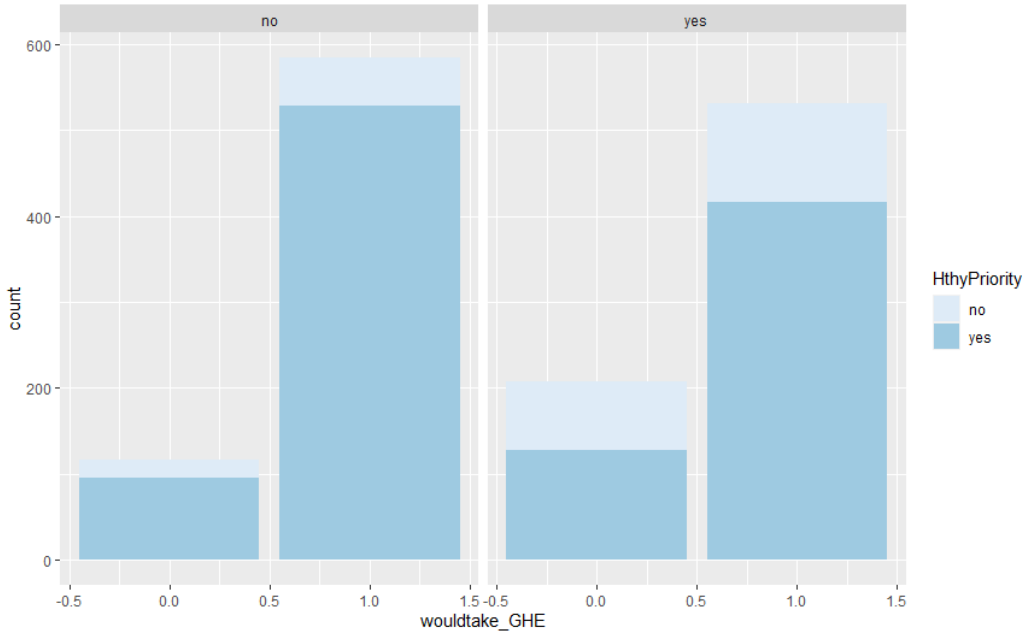
Interaction term 4: CHPerc and StabHthStt



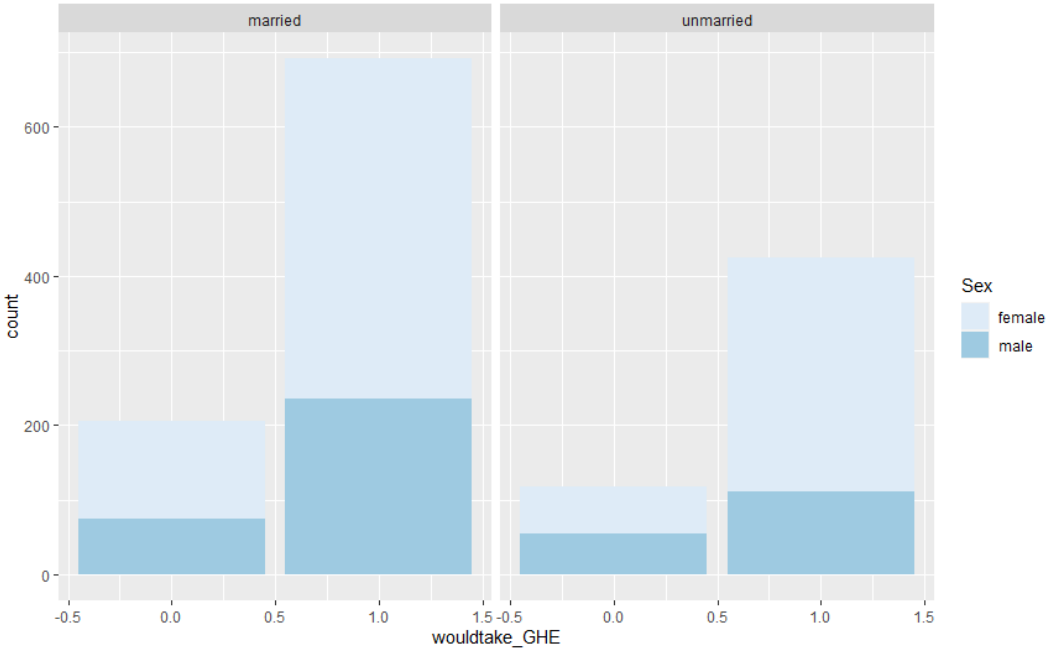
Interaction term 5: MaritalStt and HthyPriority



Interaction term 6: NotImp and HthyPriority



Interaction term 7: Marital Status and Sex



4.4.1.1 (Ridge Coefficients and Confusion Matrix)

rr_model <- glmnet(x,y,alpha=0, family = "binomial", lambda = bestlam_rr) #fit on whole data
rr_model

rr.coef <- predict(rr_model, type="coefficients", s=bestlam_rr)

rr.coef

```
> rr.coef
47 x 1 sparse Matrix of class "dgCMatrix"
                              s1
(Intercept)           1.251719411
Age_gr>=40            0.099155199
Age_gr30-39           0.093915845
Sexmale              -0.273170861
Jobsttstudent         0.047566780
JobsttUnemployed     -0.013791939
HealthInsyes          0.155102881
MaritalSttunmarried   0.002667828
EduMinsecond         -0.086112299
Wsttimeyes           -0.083082997
Wstmonyes             0.214232161
DiscDiseaseyes        0.065880316
Lessbelqualyes        0.170468576
NotImpyes            -0.349237625
HthyPriorityyes       0.427406045
Habityes              0.107469780
FlwHealthyes          0.152151068
PerTrmtyes            0.274114065
AcqTrmtyes            0.081736172
StabHthSttyes        -0.170654184
StChoiseclinic       -0.069789375
StChoiseselfstudy    -0.019575847
MedCabinetyes         0.086239282
ExpCareyes           -0.037847809
ExamToolsyes          0.137680274
TangiblesBelowAvg    -0.020886309
AssuranceBelowAvg    -0.076408163
ReliabilityBelowAvg   0.044164176
ResponBelowAvg        0.094886132
EmpathyBelowAvg       0.008990866
CHPercgood           -0.198441653
SuitFreq18m           0.140228833
SuitFreq6m            0.042932429
SuitFreqg18m         -0.437377379
SuffInfoBelowAvg     -0.009903187
AttractInfoBelowAvg  -0.054710612
ImpressInfoBelowAvg  -0.088391338
PopularInfoBelowAvg  -0.049716380
UseITno              -0.261050601
UseITyes              0.042637627
AfterITno            -0.081181814
AfterITyes           -0.057018680
Tooluseskillsyes     -0.181835484
EvalExerquitesuff    -0.071031457
EvalExertrivial      -0.129776408
EvalExerverysuff      0.046290080
stdz_bmi              0.036594807
```

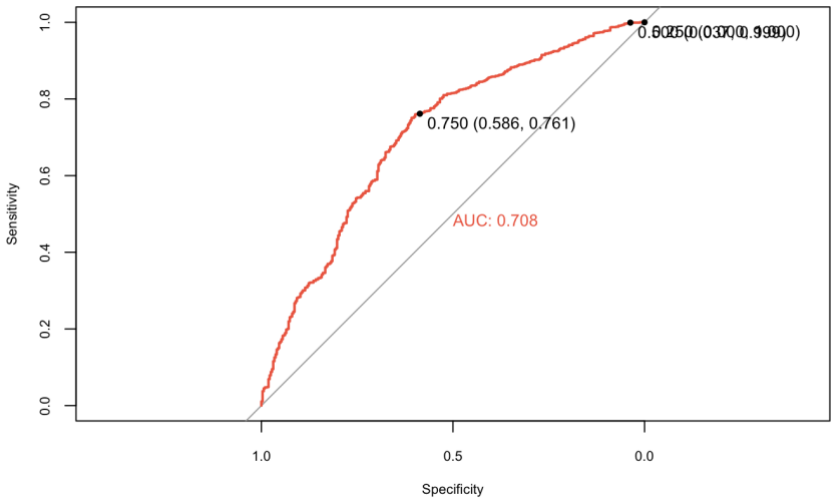
assess.glmnet(lasso_model, newx = x, newy = y)
confusion.glmnet(lasso_model, newx = x, newy = y)

```
> assess.glmnet(rr_model, newx = x, newy = y)
$deviance
       s0
0.9789683
attr(,"measure")
[1] "Binomial Deviance"

$class
       s0
0.2175122
attr(,"measure")
[1] "Misclassification Error"

$auc
[1] 0.707842
attr(,"measure")
[1] "AUC"
```

```
> confusion.glmnet(rr_model, newx = x, newy = y)
          True
Predicted    0    1 Total
    0       12    1    13
    1      312 1114  1426
    Total  324 1115  1439

 Percent Correct:  0.7825
```

Fig 4.4.1.2 (ROC and AUC for Ridge)

#roc curve
problasso = predict(lasso_model,type=c("response"), newx = x, lambda = bestlam)
par(cex.lab=0.75,cex.axis=0.75, cex = 0.90) #global command to set font sizes
# calculate and draw ROC
ROCr = roc(y==1 ~ problasso, data = analysis5)
plot(ROCr,print.thres=c(0.25,0.5,0.75),legacy.axes=F, print.auc = TRUE,col="Coral2")



4.4.1.3 (LASSO Coefficients and Confusion Matrix)

lasso_model <- glmnet(x,y, alpha=1, lambda=bestlam, family="binomial") #fitting model on whole dataset

lasso.coef <- predict(lasso_model, type="coefficients", s=bestlam) #predicting model on whole dataset
lasso.coef
lasso.coef[lasso.coef!=0]

```
> lasso.coef <- predict(lasso_model, type="coefficients", s=bestlam)
> lasso.coef
47 x 1 sparse Matrix of class "dgCMatrix"
                              s1
(Intercept)          1.144207685
Age_gr>=40           .
Age_gr30-39          .
Sexmale             -0.293109960
Jobsttstudent        .
JobsttUnemployed     .
HealthInsyes         0.123208147
MaritalSttunmarried  .
EduMinsecond         .
Wsttimeyes           .
Wstmonyes            0.230516485
DiscDiseaseyes       .
Lessbelqualyes       0.161987474
NotImpyes           -0.451334379
HthyPrioorityyes     0.571620096
Habityes             0.060187961
FlwHealthyes         0.107492695
PerTrmtyes           0.340273917
AcqTrmtyes           0.015273789
Tooluseskillsyes    -0.131423592
EvalExerquitesuff    .
EvalExertrivial     -0.056618505
EvalExerverysuff     .
stdz_bmi             0.006498675
StabHthSttyes       -0.083306006
StChoiseclinic       .
StChoiseselfstudy    .
MedCabinetyes        .
ExpCareyes           .
ExamToolsyes         0.108225934
TangiblesBelowAvg    .
AssuranceBelowAvg    .
ReliabilityBelowAvg  .
ResponBelowAvg       .
EmpathyBelowAvg      .
CHPercgood          -0.188060903
SuitFreq18m          .
SuitFreq6m           .
SuitFreqg18m        -0.481741143
SuffInfoBelowAvg     .
AttractInfoBelowAvg  .
ImpressInfoBelowAvg -0.067595557
PopularInfoBelowAvg  .
UseITno             -0.358309849
UseITyes             .
AfterITno            .
AfterITyes           .
```

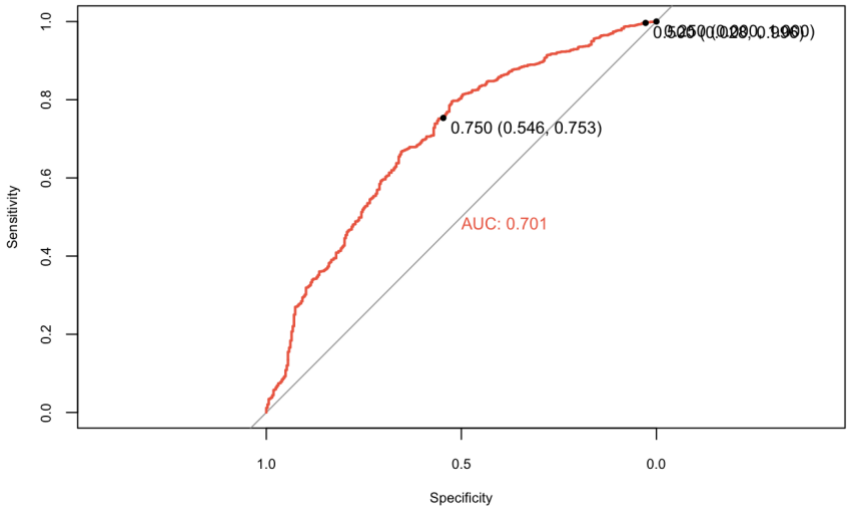
assess.glmnet(lasso_model, newx = x, newy = y)
confusion.glmnet(lasso_model, newx = x, newy = y)

```
> assess.glmnet(lasso_model, newx = x, newy = y)
$deviance
       s0
0.9839595
attr(,"measure")
[1] "Binomial Deviance"

$class
       s0
0.2216817
attr(,"measure")
[1] "Misclassification Error"

$auc
[1] 0.701265
attr(,"measure")
[1] "AUC"
```

```
> confusion.glmnet(lasso_model, newx = x, newy = y)
          True
Predicted    0    1 Total
    0        9    4    13
    1      315 1111  1426
    Total  324 1115  1439

 Percent Correct:  0.7783
```

4.4.1.4 (ROC and AUC for LASSO)

```
#roc curve
problasso = predict(lasso_model,type=c("response"), newx = x, lambda = bestlam)
par(cex.lab=0.75,cex.axis=0.75, cex = 0.90) #global command to set font sizes
# calculate and draw ROC
ROCr = roc(y==1 ~ problasso, data = analysis5)
plot(ROCr,print.thres=c(0.25,0.5,0.75),legacy.axes=F, print.auc = TRUE,col="Coral2")
```



4.4.2.1 (Model 4 Output)

```
x1 <- model.matrix(wouldtake_GHE~., analysis2)[,-1]
y1 <- analysis2$wouldtake_GHE

model_4 <- glm(wouldtake_GHE ~ MaritalStt+Edu +Sex +BMI + HealthIns +HthyPriority  + PerTrmt
+MedCabinet+ ExamTools+ Tooluseskills + UseIT+(Edu*UseIT)+ (MedCabinet*Tooluseskills),
data=analysis2, family="binomial")
summary(model_4) #AIC: 1456.6
```

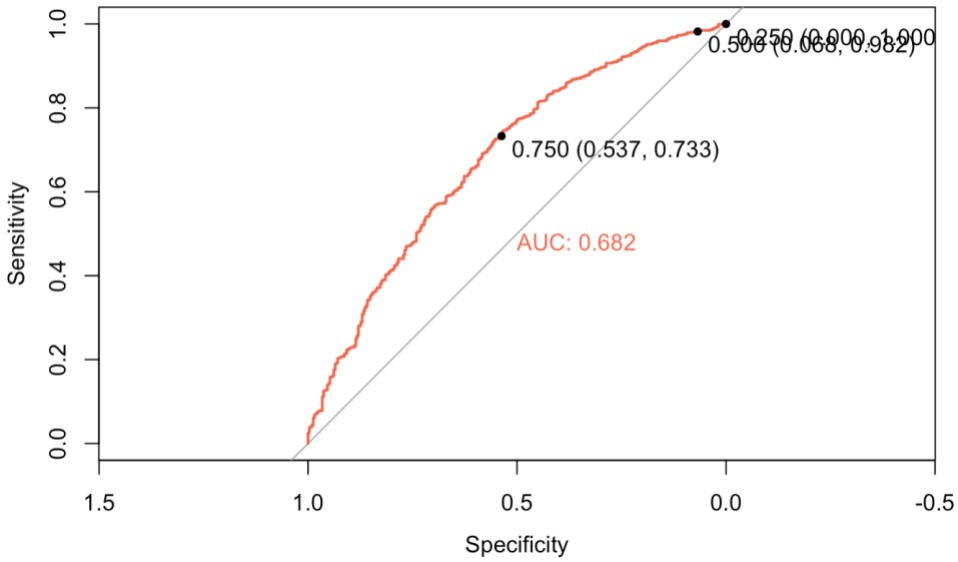
#(output on next page)

```
Call:
glm(formula = wouldtake_GHE ~ MaritalStt + Edu + Sex + BMI +
    HealthIns + HthyPriority + PerTrmt + MedCabinet + ExamTools +
    Tooluseskills + UseIT + (Edu * UseIT) + (MedCabinet * Tooluseskills),
    family = "binomial", data = analysis2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3786   0.3645   0.5778   0.7174   1.4293

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -0.69516    0.63509  -1.095 0.273697
MaritalSttunmarried           -0.07559    0.15170  -0.498 0.618291
EduMinsecond                   0.54766    0.36578   1.497 0.134326
Sexmale                       -0.54671    0.14959  -3.655 0.000257 ***
BMI                            0.04394    0.02777   1.582 0.113660
HealthInsyes                   0.18069    0.16634   1.086 0.277360
HthyPriorityyes                0.89043    0.15369   5.794 6.88e-09 ***
PerTrmtyes                     0.57800    0.17869   3.235 0.001218 **
MedCabinetyes                  0.74309    0.27018   2.750 0.005953 **
ExamToolsyes                   0.34376    0.14211   2.419 0.015565 *
Tooluseskillsyes               0.21994    0.26220   0.839 0.401570
UseITno                       -0.52320    0.20404  -2.564 0.010342 *
UseITyes                       0.16824    0.17864   0.942 0.346303
EduMinsecond:UseITno          -0.34099    0.44462  -0.767 0.443128
EduMinsecond:UseITyes         -1.33023    0.43249  -3.076 0.002100 **
MedCabinetyes:Tooluseskillsyes -0.89230    0.33830  -2.638 0.008350 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1535.0  on 1438  degrees of freedom
Residual deviance: 1424.6  on 1423  degrees of freedom
AIC: 1456.6

Number of Fisher Scoring iterations: 4
```

4.4.2.2 (Model 4 ROC and AUC)

```
library(pROC)
prob = predict(model_4 ,type=c("response"),analysis2) # find predicted probabilities
par(cex.lab=1,cex.axis=1, cex = 1) #global command to set font sizes
# calculate and draw ROC
ROC = roc(y1==1 ~ prob, data =analysis2)
plot(ROC,print.thres=c(0.25,0.5,0.75),legacy.axes=F, print.auc = TRUE,col="Coral2")
```

#Graph : ROC and AUC for model 4



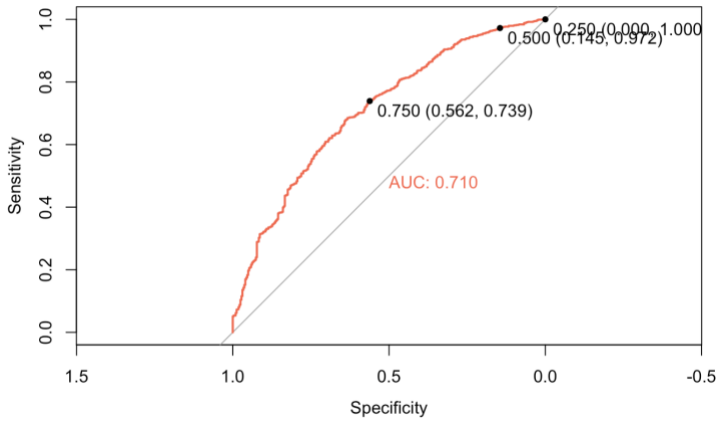
4.4.2.3 (Model 5 Output)

```
x1 <- model.matrix(wouldtake_GHE~., analysis2)[,-1]
y1 <- analysis2$wouldtake_GHE

model_5 <- glm(wouldtake_GHE ~ MaritalStt+Edu +Sex +BMI + HealthIns + PerTrmt + ExamTools  +
Habit + StabHthStt+ Tooluseskills + UseIT+ CHPerc
+HthyPriority+(CHPerc*StabHthStt)+(Edu*UseIT)+(MaritalStt*Sex)+(NotImp*HthyPriority)+
(MaritalStt*HthyPriority), data=analysis2, family="binomial")
summary(model_5) #AIC: 1432.1
```

```
Call:
glm(formula = wouldtake_GHE ~ MaritalStt + Edu + Sex + BMI +
    HealthIns + PerTrmt + ExamTools + Habit + StabHthStt + Tooluseskills +
    UseIT + CHPerc + HthyPriority + (CHPerc * StabHthStt) + (Edu *
    UseIT) + (MaritalStt * Sex) + (NotImp * HthyPriority) + (MaritalStt *
    HthyPriority), family = "binomial", data = analysis2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3587   0.3260   0.5461   0.7106   1.6005

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                        1.04610    0.78352   1.335 0.181833
MaritalSttunmarried                0.76172    0.31797   2.396 0.016594 *
EduMinsecond                       0.43083    0.37072   1.162 0.245179
Sexmale                           -0.29151    0.18490  -1.577 0.114891
BMI                                0.04453    0.02831   1.573 0.115684
HealthInsyes                       0.16083    0.17089   0.941 0.346649
PerTrmtyes                         0.54574    0.18952   2.880 0.003982 **
ExamToolsyes                       0.30641    0.14339   2.137 0.032608 *
Habityes                           0.23553    0.14340   1.643 0.100480
StabHthSttyes                     -0.84781    0.37795  -2.243 0.024883 *
Tooluseskillsyes                  -0.37958    0.16619  -2.284 0.022373 *
UseITno                           -0.70371    0.21073  -3.339 0.000839 ***
UseITyes                           0.12280    0.18016   0.682 0.495472
CHPercgood                        -1.37897    0.45936  -3.002 0.002683 **
HthyPriorityyes                    0.75302    0.30044   2.506 0.012196 *
NotImpyes                         -0.75108    0.31319  -2.398 0.016478 *
StabHthSttyes:CHPercgood           1.12761    0.47562   2.371 0.017749 *
EduMinsecond:UseITno              -0.26794    0.45436  -0.590 0.555383
EduMinsecond:UseITyes             -1.28475    0.43834  -2.931 0.003379 **
MaritalSttunmarried:Sexmale       -0.73732    0.28590  -2.579 0.009912 **
HthyPriorityyes :NotImpyes         0.22609    0.35143   0.643 0.520000
MaritalSttunmarried:HthyPriorityyes -0.57843    0.33542  -1.725 0.084616 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1535.0  on 1438  degrees of freedom
Residual deviance: 1388.1  on 1417  degrees of freedom
AIC: 1432.1
```

4.4.2.4 (Model 5 ROC and AUC)

```
library(pROC)
prob = predict(model_5 ,type=c("response"),analysis2) # find predicted probabilities
par(cex.lab=1,cex.axis=1, cex = 1) #global command to set font sizes
# calculate and draw ROC
ROC = roc(y1==1 ~ prob, data =analysis2)
plot(ROC,print.thres=c(0.25,0.5,0.75),legacy.axes=F, print.auc = TRUE,col="Coral2")
#Graph : ROC and AUC  for model 5
```



4.4.2.5 (Model 6 Output)

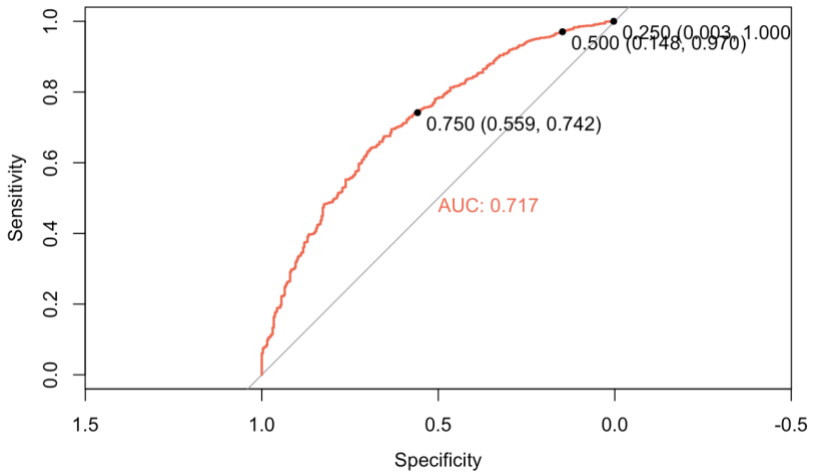
```
x1 <- model.matrix(wouldtake_GHE~., analysis2)[,-1]
y1 <- analysis2$wouldtake_GHE

model_6 <- glm(wouldtake_GHE ~ MaritalStt+Edu +Sex +BMI + HealthIns + PerTrmt + ExamTools  +
Habit + StabHthStt+ Tooluseskills + AcqTrmt+ UseIT+ Wsttime+ Wstmon+
+CHPerc+HthyPriority+NotImp+ (Wsttime*Wstmon)
+(CHPerc*StabHthStt)+(Edu*UseIT)+(MaritalStt*Sex)+(MaritalStt*HthyPriority)+(NotImp*HthyPriority),
data=analysis2, family="binomial")
summary(model_6) #AIC: 1429.3
```

```
Call:
glm(formula = wouldtake_GHE ~ MaritalStt + Edu + Sex + BMI +
    HealthIns + PerTrmt + ExamTools + Habit + StabHthStt + Tooluseskills +
    AcqTrmt + UseIT + Wsttime + Wstmon + +CHPerc + HthyPriority +
    NotImp + (Wsttime * Wstmon) + (CHPerc * StabHthStt) + (Edu *
    UseIT) + (MaritalStt * Sex) + (MaritalStt * HthyPriority) +
    (NotImp * HthyPriority), family = "binomial", data = analysis2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3888   0.2912   0.5365   0.7108   1.6187

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                        0.98169    0.80576   1.218  0.22309
MaritalSttunmarried                0.66266    0.32362   2.048  0.04060 *
EduMinsecond                       0.44729    0.37240   1.201  0.22971
Sexmale                           -0.28547    0.18565  -1.538  0.12413
BMI                                0.04284    0.02856   1.500  0.13369
HealthInsyes                       0.16078    0.17184   0.936  0.34948
PerTrmtyes                         0.51733    0.19112   2.707  0.00679 **
ExamToolsyes                       0.33107    0.14457   2.290  0.02202 *
Habityes                           0.22289    0.14697   1.517  0.12938
StabHthSttyes                     -0.76866    0.38667  -1.988  0.04682 *
Tooluseskillsyes                  -0.34547    0.16799  -2.056  0.03974 *
AcqTrmtyes                         0.07384    0.14648   0.504  0.61420
UseITno                           -0.67240    0.21278  -3.160  0.00158 **
UseITyes                           0.15289    0.18127   0.843  0.39899
Wsttimeyes                        -0.19224    0.17566  -1.094  0.27379
Wstmonyes                          0.98819    0.42297   2.336  0.01948 *
CHPercgood                        -1.36052    0.46161  -2.947  0.00321 **
HthyPrioorityyes                   0.70815    0.30502   2.322  0.02025 *
NotImpyes                         -0.80186    0.32119  -2.497  0.01254 *
Wsttimeyes:Wstmonyes              -0.61509    0.46409  -1.325  0.18505
StabHthSttyes:CHPercgood           1.07473    0.47802   2.248  0.02456 *
EduMinsecond:UseITno              -0.27712    0.45628  -0.607  0.54362
EduMinsecond:UseITyes             -1.32180    0.44045  -3.001  0.00269 **
MaritalSttunmarried:Sexmale       -0.68936    0.28872  -2.388  0.01696 *
MaritalSttunmarried:HthyPrioorityyes -0.58134    0.33910  -1.714  0.08646 .
HthyPrioorityyes :NotImpyes        0.27211    0.36239   0.751  0.45272
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1535.0  on 1438  degrees of freedom
Residual deviance: 1377.3  on 1413  degrees of freedom
AIC: 1429.3

Number of Fisher Scoring iterations: 5
```

4.4.2.6 (Model 6 ROC and AUC)

```
library(pROC)
prob = predict(model_6 ,type=c("response"),analysis2) # find predicted probabilities
par(cex.lab=1,cex.axis=1, cex = 1) #global command to set font sizes
# calculate and draw ROC
ROC = roc(y1==1 ~ prob, data =analysis2)
plot(ROC,print.thres=c(0.25,0.5,0.75),legacy.axes=F, print.auc = TRUE,col="Coral2")
```

#Graph : ROC and AUC for model 5



Item 9.1 All Codes:

```
#Data wrangling + Feature engineering
datatest <- read.csv(file = "vietnamhs.csv")
str(datatest)
summary(datatest)

#changing UseMon to binary
library(dplyr)
datatest <- datatest |> mutate (wouldtake_GHE = ifelse( UseMon != "later", 1, 0))
datatest

#feature engineering:
#1.age groups from 5 lvls to 3 lvls
#unique(datatest$Age_gr)
datatest$Age_gr <- ifelse(datatest$Age_gr %in% c("<18", "18-29"), "<29", datatest$Age_gr)
datatest$Age_gr <- ifelse(datatest$Age_gr %in% c(">=50", "40-49"), ">=40", datatest$Age_gr)
unique(datatest$Age_gr)

#2. changing Jobstt levels from 6 to 3

#unique(datatest$Jobstt)

datatest$Jobstt <- ifelse(datatest$Jobstt %in% c("other","housewife","retirer"), "Unemployed",
datatest$Jobstt)
datatest$Jobstt <- ifelse(datatest$Jobstt %in% c("unstable","stable"), "Employed", datatest$Jobstt)

unique(datatest$Jobstt)

#2.changing Edu levels from 4 to 3

#unique(datatest$Edu)
datatest$Edu <- ifelse(datatest$Edu %in% c("Grad","PostGrad"), "MinGrad", “Minsecond”)
datatest$MaritalStt <- ifelse(datatest$MaritalStt %in% c("other", "unmarried"), "unmarried","married")

#unique(datatest$CHPerc)
datatest$CHPerc <- ifelse(datatest$CHPerc %in% c("good","quite"), "good", datatest$CHPerc)

#changing all scale variables from 5 factor levels to 2

datatest$ImpressInfo <- ifelse(datatest$ImpressInfo <= 2, "BelowAvg","AboveAvg")
datatest$Reliability <- ifelse(datatest$Reliability <= 2, "BelowAvg","AboveAvg")
datatest$Respon <- ifelse(datatest$Respon <= 2, "BelowAvg","AboveAvg")
datatest$Assurance <- ifelse(datatest$Assurance <= 2, "BelowAvg","AboveAvg")
datatest$Empathy <- ifelse(datatest$Empathy <= 2, "BelowAvg","AboveAvg")
datatest$PopularInfo <- ifelse(datatest$PopularInfo <= 2, "BelowAvg","AboveAvg")
datatest$AttractInfo <- ifelse(datatest$AttractInfo <= 2, "BelowAvg","AboveAvg")
datatest$SuffInfo <- ifelse(datatest$SuffInfo <= 2, "BelowAvg","AboveAvg")
```

```
datatest$Tangibles <- ifelse(datatest$Tangibles <= 2, "BelowAvg","AboveAvg")

#changing chr and appropriate num classes to factor levels

library(dplyr)
datatest_ftr <- datatest |>
  select( -c("id", "date","Age", "height","weight","BMI", "SuitExer", "wouldtake_GHE","UseMon")) #take
out the variables that should be numeric, chr and not factor
str(datatest_ftr)

#changing chr to factor
datatest_ftr[sapply(datatest_ftr, is.character)] <- lapply(datatest_ftr[sapply(datatest_ftr,
is.character)],as.factor)

#adding back numerical variables to our dataframe

datatest_num <- datatest |>
  select( c( "height","weight","BMI","wouldtake_GHE"))

analysis <- cbind(datatest_ftr, datatest_num)

#check our new dataset
str(analysis)

#checking the variable names
names(analysis)

#choosing non-cost related variables into our dataset
analysis1 <- analysis |>
  select("wouldtake_GHE", "Age_gr", "Sex", "Jobstt", "HealthIns", "BMI", "MaritalStt", "Edu" , "Wsttime",
"Wstmon", "DiscDisease", "Lessbelqual", "NotImp", "HthyPriority", "Habit", "FlwHealth", "PerTrmt",
"AcqTrmt", "StabHthStt" , "StChoise", "MedCabinet", "ExpCare", "ExamTools", "Tangibles", "Assurance",
"Reliability", "Respon", "Empathy", "CHPerc", "SuitFreq", "SuffInfo", "AttractInfo", "ImpressInfo",
"PopularInfo", "UseIT", "AfterIT", "Tooluseskills", "EvalExer")

#dataset with chosen variables without NA values
analysis2 <- na.omit(analysis1)
```

#Base Model

```
#running logistic regression with our 1st model (base model)
model_1 <- glm(wouldtake_GHE~. ,  data=analysis2, family="binomial")
summary(model_1 )
```

#Regularisation methods

```
#standardising numerical variable BMI - only for LASSO and ridge analysis
```

```
analysis5 <- analysis2 |>
  mutate(stdz_bmi = (BMI - mean(BMI)) / sd(BMI)) |>
  select(-BMI)

#creating matrix and training/testing sets for regularisation

library(glmnet)
x <- model.matrix(wouldtake_GHE~., analysis2)[,-1]
y <- analysis2$wouldtake_GHE
grid <- c(0, 0.001, 0.002, 0.01, 0.03, 0.1, 0.5, 1, 5, 10, 15, 20, 25, 30)

set.seed(1)
train <- sample(1:nrow(x), nrow(x)/2) # get half of data as training set
test <- (-train)  # the rest as testing set
y.test <- y[test]


#ridge
cvrr.out <- cv.glmnet(x, y, alpha=0, family ="binomial") #obtain the 10-fold cross validation

bestlam_rr <- cvrr.out$lambda.min  #identify the lambda for smallest CV error
bestlam_rr

rr_model <- glmnet(x,y,alpha=0, family = "binomial", lambda = bestlam_rr) #fit on whole data
rr_model

rr.coef <- predict(rr_model, type="coefficients", s=bestlam_rr)
rr.coef


assess.glmnet(rr_model, newx = x, newy = y)
confusion.glmnet(rr_model, newx = x, newy = y)

probrr = predict(rr_model,type=c("response"), newx = x, lambda = bestlam_rr)

library(pROC)
par(cex.lab=0.75,cex.axis=0.75, cex = 0.90) #global command to set font sizes
# calculate and draw ROC
ROCr = roc(y==1 ~ probrr, data = analysis5)
plot(ROCr,print.thres=c(0.25,0.5,0.75),legacy.axes=F, print.auc = TRUE,col="Coral2")
```

```
#Lasso

lasso.mod <- glmnet(x, y, alpha=1, lambda=grid) #explore how lasso selects variables with incr. values of lambda
coef(lasso.mod,s= grid)

set.seed(1805)
cv.out <- cv.glmnet(x, y, alpha=1, family = "binomial")
plot(cv.out)

bestlam <- cv.out$lambda.min #finding best lambda using cv
bestlam

lasso_model <- glmnet(x,y, alpha=1, lambda=bestlam, family="binomial") #fitting model on whole dataset

lasso.coef <- predict(lasso_model, type="coefficients", s=bestlam) #predicting model on whole dataset
lasso.coef
lasso.coef[lasso.coef!=0]

assess.glmnet(lasso_model, newx = x, newy = y)
confusion.glmnet(lasso_model, newx = x, newy = y)

#roc curve
problasso = predict(lasso_model,type=c("response"), newx = x, lambda = bestlam)
par(cex.lab=0.75,cex.axis=0.75, cex = 0.90) #global command to set font sizes
# calculate and draw ROC
ROCr = roc(y==1 ~ problasso, data = analysis5)
plot(ROCr,print.thres=c(0.25,0.5,0.75),legacy.axes=F, print.auc = TRUE,col="Coral2")
```

#Logistic Models

```
x1 <- model.matrix(wouldtake_GHE~., analysis2)[,-1]
y1 <- analysis2$wouldtake_GHE

#model4
model_4 <- glm(wouldtake_GHE ~ MaritalStt+Edu +Sex +BMI + HealthIns +HthyPriority  + PerTrmt
+MedCabinet+ ExamTools+ Tooluseskills + UseIT+(Edu*UseIT)+ (MedCabinet*Tooluseskills),
data=analysis2, family="binomial")
summary(model_4) #AIC: 1456.6

prob = predict(model_5 ,type=c("response"),analysis2) # find predicted probabilities
par(cex.lab=1,cex.axis=1, cex = 1) #global command to set font sizes
# calculate and draw ROC
ROC = roc(y1==1 ~ prob, data =analysis2)
plot(ROC,print.thres=c(0.25,0.5,0.75),legacy.axes=F, print.auc = TRUE,col="Coral2")
```

```
#model5
model_5 <- glm(wouldtake_GHE ~ MaritalStt+Edu +Sex +BMI + HealthIns + PerTrmt + ExamTools  +
Habit + StabHthStt+ Tooluseskills + UseIT+ CHPerc
+HthyPriority+(CHPerc*StabHthStt)+(Edu*UseIT)+(MaritalStt*Sex)+(NotImp*HthyPriority)+
(MaritalStt*HthyPriority), data=analysis2, family="binomial")
summary(model_5) #AIC: 1432.1

prob = predict(model_5 ,type=c("response"),analysis2) # find predicted probabilities
par(cex.lab=1,cex.axis=1, cex = 1) #global command to set font sizes
# calculate and draw ROC
ROC = roc(y1==1 ~ prob, data =analysis2)
plot(ROC,print.thres=c(0.25,0.5,0.75),legacy.axes=F, print.auc = TRUE,col="Coral2")

#model6
model_6 <- glm(wouldtake_GHE ~ MaritalStt+Edu +Sex +BMI + HealthIns + PerTrmt + ExamTools  +
Habit + StabHthStt+ Tooluseskills + AcqTrmt+ UseIT+ Wsttime+ Wstmon+
+CHPerc+HthyPriority+NotImp+ (Wsttime*Wstmon)
+(CHPerc*StabHthStt)+(Edu*UseIT)+(MaritalStt*Sex)+(MaritalStt*HthyPriority)+(NotImp*HthyPriority),
data=analysis2, family="binomial")
summary(model_6) #AIC: 1429.3

library(pROC)
prob = predict(model_6 ,type=c("response"),analysis2) # find predicted probabilities
par(cex.lab=1,cex.axis=1, cex = 1) #global command to set font sizes
# calculate and draw ROC
ROC = roc(y1==1 ~ prob, data =analysis2)
plot(ROC,print.thres=c(0.25,0.5,0.75),legacy.axes=F, print.auc = TRUE,col="Coral2")
```

8.0 References

Dickson, L. (2022, September 17). *How to interpret the odds ratio with categorical variables in logistic ...* How to Interpret the Odds Ratio with Categorical Variables in Logistic Regression. Retrieved November 13, 2022, from https://towardsdatascience.com/how-to-interpret-the-odds-ratio-with-categorical-variables-in-logistic-regression-5bb38e3fc6a8

Kiet, A. (2020, September 26). *Telehealth in Vietnam connects over 1,000 health centers*. hanoitimes.vn. Retrieved November 12, 2022, from https://hanoitimes.vn/telehealth-in-vietnam-connects-over-1000-health-centers-314319.html

Krogsbøll, L. T., Jørgensen, K. J., & Gøtzsche, P. C. (2019, January 31). *General Health checks in adults for reducing morbidity and mortality from disease*. The Cochrane database of systematic reviews. Retrieved November 12, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6353639/

Le, D.-C., Kubo, T., Fujino, Y., Pham, T.-M., & Matsuda, S. (2012, July 13). *Health Care System in Vietnam: Current situation and challenges*. Asian Pacific Journal of Disease Management. Retrieved November 12, 2022, from https://www.jstage.jst.go.jp/article/apjdm/4/2/4_23/_article

Liss, D. T., Uchida, T., Wilkes, C. L., Radakrishnan, A., & Linder, J. A. (2021, June 8). *General Health Checks in adult primary care: A Review*. JAMA. Retrieved November 12, 2022, from https://pubmed.ncbi.nlm.nih.gov/34100866/

Park, H. (2015, March 26). *Cultural values and family experiences in diverse ecological contexts: Implications for social change*. eScholarship, University of California. Retrieved November 12, 2022, from https://escholarship.org/uc/item/0rh785px

The Voice of Vietnam. (2020, December 29).*Telehealth project proves effective in Vietnam*. Retrieved November 12, 2022, from https://vovworld.vn/en-US/society/telehealth-project-proves-effective-in-vietnam-936039.vov

Tran, A. T. B., & Yang, J. (2022, September 7). *Reducing gaps in education remains important in Vietnam as new school year kicks in*. World Bank Blogs. Retrieved November 12, 2022, from https://blogs.worldbank.org/eastasiapacific/reducing-gaps-education-remains-important-vietnam-new-school-year-kicks

Vuong, Q.-H. (2017). Survey data on Vietnamese propensity to attend periodic general health examinations. *Scientific Data*, *4*(1). https://doi.org/10.1038/sdata.2017.142

World Bank. (2022, November 7).*The World Bank In Vietnam.* Retrieved November 12, 2022, from https://www.worldbank.org/en/country/vietnam/overview

World Bank Group. (2021, October 12). *Vietnam: Adapting to an aging society*. World Bank. Retrieved November 12, 2022, from https://www.worldbank.org/en/country/vietnam/publication/vietnam-adapting-to-an-aging-society