Final Project

In this final project, I am not expecting you to be data mining experts; however I am looking for significant progress using the model building topics learned in this course. Please work independently!

Recall the earlier assignment:
Find an appropriate data set to use for your Final Project. This data set should have many observations and predictor variables. Each predictor variable may be any form (quantitative or qualitative). The response variable must be binary. Any application topic is appropriate…business, sports, medical, educational, government, engineering, …

I would like you to incorporate ALL that you have learned in this course. This includes ideas we have discussed, for example: over/under fitting, risk, average squared error, R^2 , sensitivity, specificity … I would like you to build the "best" model utilizing each of the following techniques:

- CHAID
- Exhaustive CHAID
- CART
- Logistic Regression
- SAS Enterprise Miner: Decision Tree
- SAS Enterprise Miner: Regression (Preferred but not mandatory)
- SAS Enterprise Miner: Neural Network (Optional)

After you have completed the above techniques, please provide your "best" model from all contending models. What were your steps? Be sure to clearly define how you determined your "best" model. Was your ultimate prediction goal? Was your ultimate goal to correctly predict the probability of success? Was your ultimate goal to describe how the response variable varies based on a combination of predictor variables? Be very clear and specific.

Initial Model Building steps:

- Import your data set into SPSS. Be sure to clearly define all variables in the Variable View (ie Label, Values, Measure …).
- Import your data set into SAS Enterprise Miner. Be sure to complete all the appropriate import procedures (naming variables, defining missing values …).

Preliminary Analysis:

- For a preliminary analysis, do some simple descriptive statistics, charts/graphs, and cross-tabs. Submit everything you find interesting…

Deliverables:

- Your written group report must have *at most* 20 pages, everything included. All figures and computer output must be numbered. Due Date: Wednesday, May 8th.
- Please prepare a 10 minute Power Point presentation of your procedures and most interesting findings. Due Date: Presentation emailed to me by Monday, April 29th 2:00pm. Presentations will occur on Monday (4/29) and Wednesday (5/1).

Applied Analytics and Decision Tree Models that Predict the Diabetic Cleveland Clinic Patients Users of the Personal Health Record System ~ Spring Semester 2019 ~ Dr. Fridline

Kimberly Schveder and Sonia Naeem

Table of Contents

Introduction and Preliminary Analysis:

A personal health record system is a system that was introduced in the last decade for patients to monitor their health status from their chosen hospital, for their benefit to monitoring their health and for understanding what is going on with their bodies and when they need to visit their physician. Accordingly, ". Access to effective and tailored patient education, electronic patient–provider communication, and the wealth of clinical information and web-based resources contained within modern PHRs could lead to improvements in chronic disease outcomes through improved patient-centered care and self-management." In this case, according to some notes on the data set, the type 2 diabetic patients (aged 18 to 75 years) were documented from July 2008 through June 2009 as either using the PHR system or not, along with a variety of health and demographic information that can be used to predict the data set that was thought to have been associated with whether the patients will monitor their health status. The main focus of this project is answering the question of which variables predict the probability of success and which model predicts the probability of success of the patient using the PHR system the best (most accurately and frequently, with the smallest variance).

LIST OF VARIABLES:

Name	Codes/Values	Abbreviation
Age	years	Age
Gender	1 = female; 0 = male	Female
Race	1 = Caucasian; 0 = other	Caucasian
Insurance Type	1 = Commercial; 0 = other	Insurance
Household Income	US Dollars in thousands	Income
Provider Engagement	percentage of the physician's patients from the study sample who logged in at least 1 day during the study	Engagement
PHR user group	1 = User; 0 = Nonuser	User
Frequency of PHR use	number of logins during study period	Logins
Dilated retinal eye exam	eye exam recorded within study period 1 = yes; 0 = no	Eye Exam
Pneumococcal Vaccination	documented lifetime vaccination 1 = yes; 0 = no	Pneumo Vaccine
Attention to Kidneys	Use of ACEi/ARB and/or test for microalbuminuria within study period 1 = yes; 0 = no	ACE ARB ALB
Attention to Feet	documented foot exam within study period 1 = yes; 0 = no	Foot Exam
Smoking Cessation	documented nonsmoker 1 = yes ; 0 = no	Nonsmoking
HbA1c Test	HbA1c value measured within study period 1 = yes; 0 = no	HBA1C Test
HbA1c value	%, last documented value within study period	HBA1C
Blood pressure	mmHg, last documented value within study period	SBP
Blood pressure	mmHg, last documented value within study period	DBP
Cholesterol	mg/dL, last documented value within study period	LDL

Figure 1: PHR dataset variables

It was found that, as seen in Figure 1, that the variables age, gender, race, insurance type, household income, smoking status, and HBA1C percentage (which is used to determine how much glucose is in a patient's blood) are all good predictors of PHR usage according to the output from all of the models shown and discussed below. As you might recognize, the patient's demographic information is more associated with whether they use the PHR system or not, as will be shown in the models in this project.

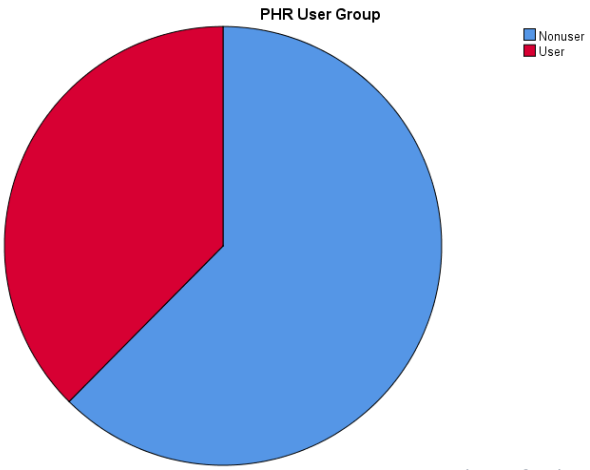


Figure 2: Pie chart of the response variable PHR User.

Right off the bat, we can see there are many more users than nonusers of the PHR system, as seen in the pie-chart in Figure 2. About two thirds of the diabetic patients use the Personal Health Record (PHR) system. About one third use this online system.

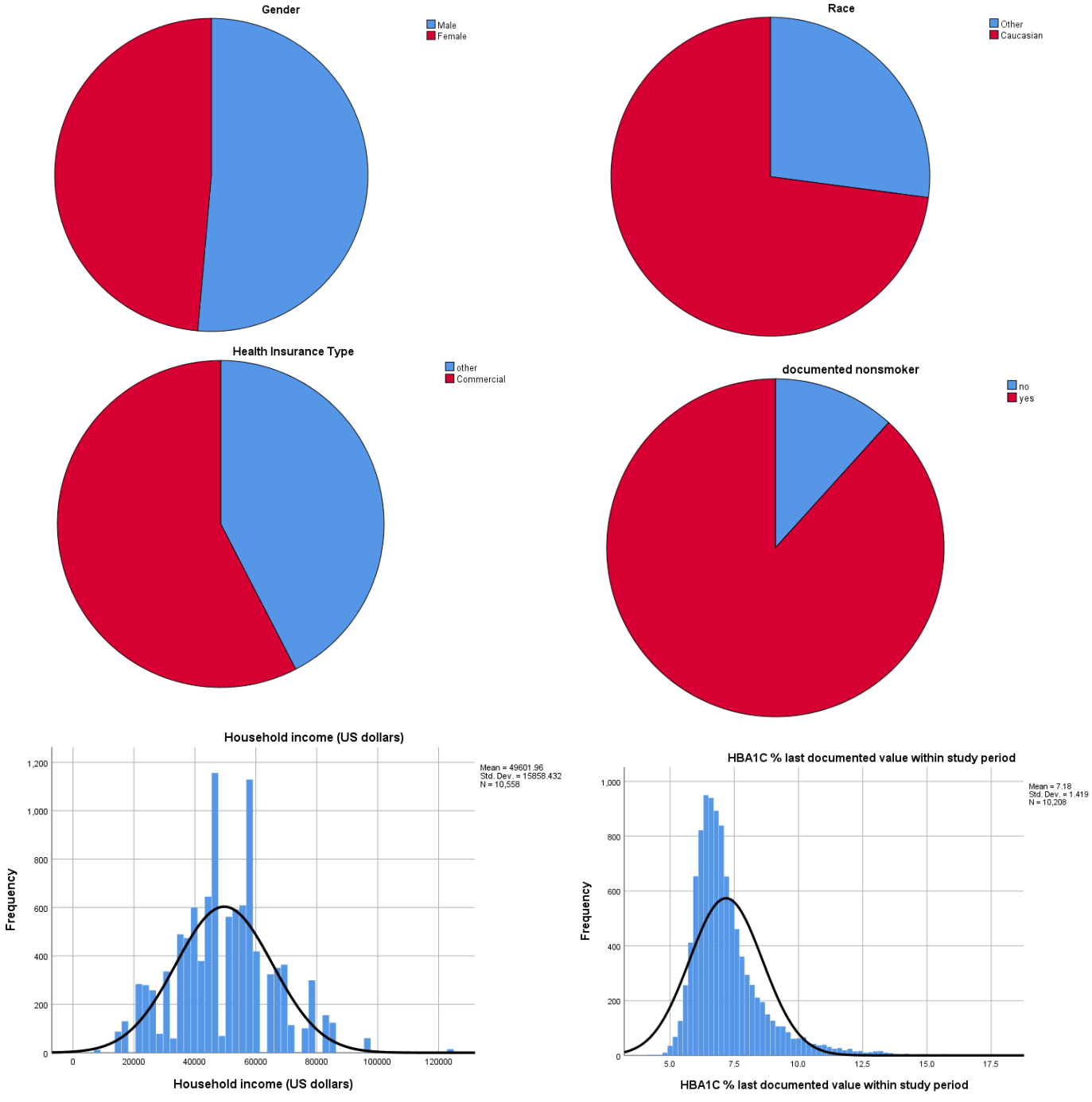


Figure 3: some Pie graphs and histograms of the(potential) predictor variables

There is about an equal amount of males and females in this dataset, according to the pie-chart of gender in Figure 3. There is an overwhelming majority of Caucasians than racial minorities in this dataset. A majority of the patients in this dataset have commercial health insurance. This is good. A majority of the patients in this dataset do not smoke, according to the nonsmoker pie-chart. The distribution of age is skewed left, so there are more people of older age in this data set (mean is about 61 years of age), according to the age histogram in Figure 3. The people in this data set are

mainly middle class (their average annual income is about $50,000), according to the income histogram. The HBA1C levels are right skewed, so there are some outliers with a lot of blood glucose, according to the histogram of HBA1C % in Figure 3. This is not surprising because all of them have type 2 diabetes, some more severe than others. Accordingly, "Hemoglobin A1c levels between 5.7% and 6.4% mean you have a higher chance of getting diabetes. Levels of 6.5% or higher mean you have diabetes." (Source: https://www.webmd.com/diabetes/guide/glycated-hemoglobin-test-hba1c) So, it makes sense that the mean is within the range that indicates that a person has diabetes.

CHAID and Exhaustive CHAID:

I found that the ECHAID output was exactly the same as the CHAID output, for both the training and testing data set. I have placed the SPSS code for the ECHAID in the appendix. Out of all of the potential independent variables, as listed in the Model summary for the CHAID algorithm in Figure 4, the independent variables included in the CHAID model to predict the PHR user group, were Race, Health Insurance type, Gender, HBA1C%, Household income. The validation was based on a split sample, with approximately 50% of the data in the training and 50% in the testing datasets, with the test variable being used as the splitting variable for the split sample. The max tree depth was set to be 3, and that was also what the result was. The minimum cases in the parent node was 100 and the minimum number of cases in the child nodes could 50. There was 16 nodes in total, 9 of them being terminal.

Classification Tree

Model Summary

Specifications	Growing Method	CHAID
	Dependent Variable	PHR User Group
	Independent Variables	Years of Age, Gender, Race, Health Insurance Type, Household income (US dollars), Dilated retinal eye exam recorded within study period, documented lifetime Pneumococcal Vaccination, Use of ACEi/ARB and/or test for microalbuminuria on Kidneys within Study Period, documented foot exam within study period, documented nonsmoker, HbA1c value measured within study period, HBA1C % last documented value within study period, SBP (mmHg), last documented value within study period, DBP (mmHg), last documented value within study period, Bad Cholestrol (LDL, measured in mg/dL), last documented value within study period, BMI (in kg/m^2) last documented value within study period
	Validation	Split Sample
	Maximum Tree Depth	3
	Minimum Cases in Parent Node	100
	Minimum Cases in Child Node	50
Results	Independent Variables Included	Race, Health Insurance Type, Gender, HBA1C % last documented value within study period, Household income (US dollars)
	Number of Nodes	16
	Number of Terminal Nodes	9
	Depth	3

Figure 4

In node 0, we have 5333 total people in the training sample, as shown in Figure 5, of which 1989 or 37.3% are users of the PHR system. In the testing sample, we have similar results; there are 5413 total, of which 2047, or 37.8% are users of the PHR system. (Note that the testing tree has the same splits and nodes as the training). As seen in the training sample tree, the first split is by race. The percent of Caucasians who are users of the PHR system is 43%, as seen in node 1, with the count being 1671 out of the 3886 total (testing: 43.3%, or 1710 out of 3948 for this node). On the contrary, the percent of the "others" (racial minorities) who are users of the PHR system is 22%(23% in the testing), with the count being 318 out of 1447 total (337 out of 1465 in testing) in the sample, according to node 2. This is about half that of Caucasians. The next splits will likely explain why this is.

The next split for both race categories turned out to be by health insurance type. So, according to node 3, of the people who are Caucasian and who have "other" type of health insurance (such as that from the government and government-funded health insurance), 33% use the PHR system (33.5% in the testing sample). By contrast, the people who ae Caucasian who have commercial health insurance use the PHR system at a rate of 50.4% (same as in the testing sample tree), according to node 4. In node 5, we see that of those people who are racial minorities and who have "other" health insurance, 11.4% (14% in the testing sample tree) use the PHR system. But, as seen in node 6, the people who are racial minorities with the commercial type of health insurance, only 29.8% of them used the PHR (same as in the testing sample).

Finally, according to node 7, the people who are Caucasian, who have the other type of health insurance, and who are male, only 38.9% (36.6% in the testing sample) of them used the PHR. We see that the people who are Caucasian, who have the other type of health insurance, and who are female, only 26.9% (testing sample: 30.4%) of them used the PHR.

According to node 9, of those people who are considered Caucasians, who have "commercial" health insurance, and who have an HBA1C percentage last documented in the study period of less than or equal to 6.60% (same in testing), 59.6% (55.2% in training) use the PHR system (this is the highest rate). Of these people considered Caucasians, who have commercial health insurance, and who have an HBA1C percentage last documented in the study period between 6.60% and 7.90% (same as in the testing sample), 49.7% (51.9% in the testing sample) use the PHR system (node 10). Lastly, of these people considered Caucasians, who have commercial health insurance, and who have an HBA1C percentage last documented in the study period larger than 7.90%, 37.2% (40.4% in the testing sample) use the PHR system (node 11). Perhaps, the smaller this value, the more likely they will check their health status.

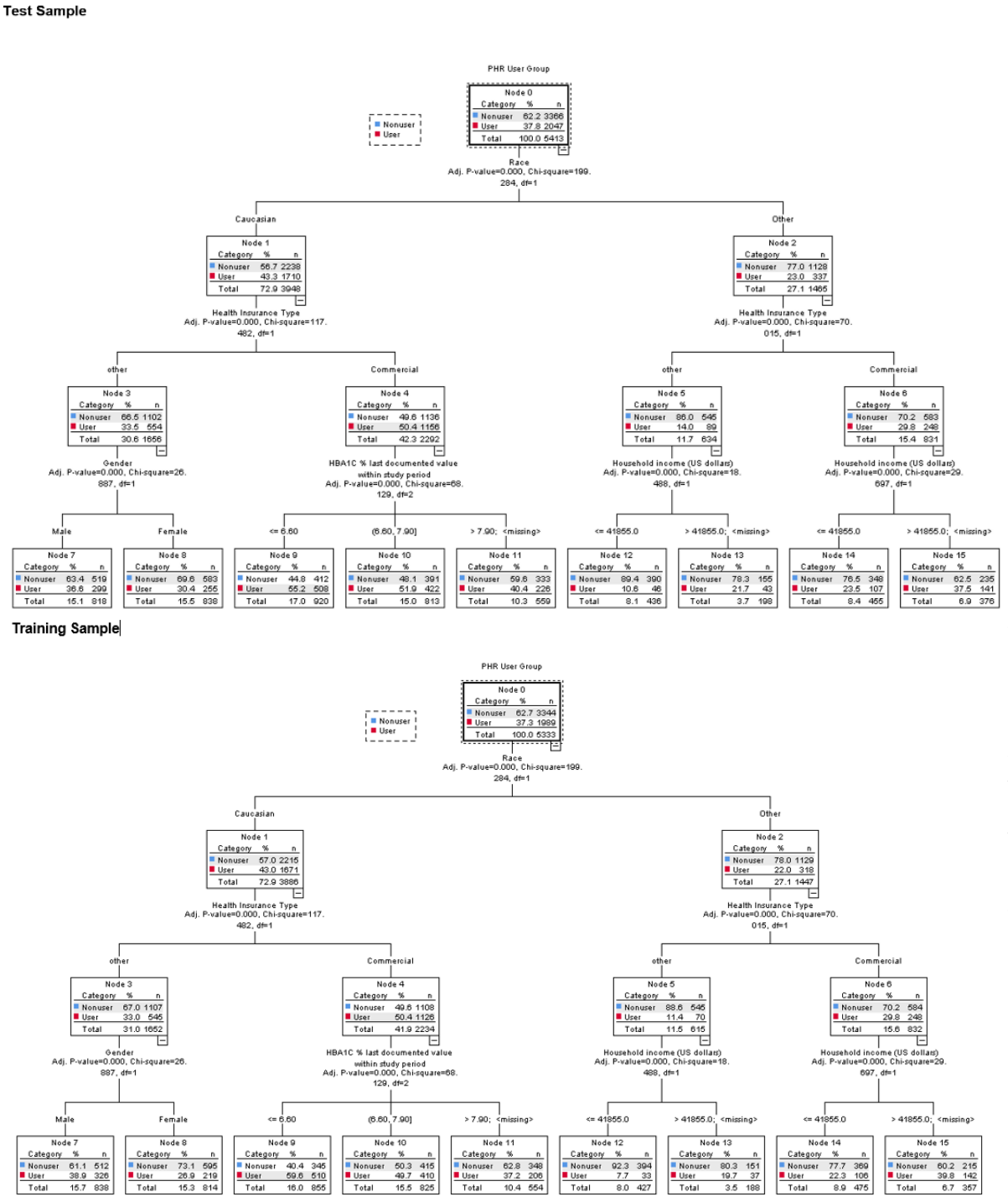


Figure 5: CHAID/ECHAID decision trees.

In node 12, those considered to be racial minorities, who have "other" health insurance, and who have a household income less than or equal to $41855, use the PHR system at a rate of 7.7% (10.6% in the testing). This is the lowest recorded percentage of users). Those considered to be racial minorities, who have "other" health insurance, and who have a household income greater than $41855 (or have missing information about income level), use the PHR system at a rate of 19.7% (21.7% in the testing sample) (see node 13). In node 14, those considered to be racial minorities, who have commercial health insurance, and who have a household income less than or equal to $41855, use the PHR system at a rate of 22.3% (23.5% in the testing sample). In node 15, those considered to be racial minorities, who have commercial health insurance, and who have a household income greater than $41855 (or have missing information about income level), use the PHR system at a rate of 39.8% (37.5% in the testing sample).

Next, we are looking at the Gains for Nodes in Figure 6. In the training sample, the gains for the nodes is shown above, for each of the terminal nodes. The nodes that are the most interesting, in my opinion, are nodes 9, 11, and 12. In node 9, of those people who are considered Caucasians, who have "commercial" health insurance, and who have an HBA1C percentage last documented in the study period of less than or equal to 6.60%, 59.6% (55.2% in the testing dataset) use the PHR system (this is the highest rate) = (510/855)*100% (in the testing dataset (508/920)*100%) use the PHR system. This is the highest rate of people using the PHR system. There are 855 patients in this node (920 for in the node in the testing dataset), which is 16.0% (17% in the testing) of the grand total number of observations in the training data set (5333) (testing data set (5413)). This is largest amount of people that fall in a terminal node than each of the other nodes in the training sample. The index of node 9 (159.9% for training; for testing 146%) tells us that this node has people who used the PHR system is at a rate of 59.9% (55.2% for testing) higher than the overall average rate of success of 37.3% in this terminal node (37.8% in testing).

Target Category: User

Gains for Nodes

Sample	Node	Node N	Node Percent	Gain N	Gain Percent	Response	Index
Training	9	855	16.0%	510	25.6%	59.6%	159.9%
	10	825	15.5%	410	20.6%	49.7%	133.2%
	15	357	6.7%	142	7.1%	39.8%	106.6%
	7	838	15.7%	326	16.4%	38.9%	104.3%
	11	554	10.4%	206	10.4%	37.2%	99.7%
	8	814	15.3%	219	11.0%	26.9%	72.1%
	14	475	8.9%	106	5.3%	22.3%	59.8%
	13	188	3.5%	37	1.9%	19.7%	52.8%
	12	427	8.0%	33	1.7%	7.7%	20.7%
Test	9	920	17.0%	508	24.8%	55.2%	146.0%
	10	813	15.0%	422	20.6%	51.9%	137.3%
	15	376	6.9%	141	6.9%	37.5%	99.2%
	7	818	15.1%	299	14.6%	36.6%	96.7%
	11	559	10.3%	226	11.0%	40.4%	106.9%
	8	838	15.5%	255	12.5%	30.4%	80.5%
	14	455	8.4%	107	5.2%	23.5%	62.2%
	13	198	3.7%	43	2.1%	21.7%	57.4%
	12	436	8.1%	46	2.2%	10.6%	27.9%

Growing Method: CHAID

Dependent Variable: PHR User Group

Figure 6

In node 11, of these people considered Caucasians, who have commercial health insurance, and who have an HBA1C percentage last documented in the study period larger than 7.90%, 37.2% = (206/554)*100% of them use the PHR system. This is the most average rate of people using the PHR system. There are 554 patients in this node, which is 10.4% of the grand total number of observations in the training data set (5333). The index of node 11 (99.7%) tells us that this node has people who used the PHR system is at a rate of 0.3% lower than the overall average rate of success of 37.3% in this terminal node.

In node 12, those considered to be racial minorities, who have "other" health insurance, and who have a household income less than or equal to $41855, use the PHR system at a rate of 7.7% = (33/427)*100% (7.9% = in the testing data) of them use the PHR system. This is the smallest rate of people using the PHR system. There are 427 (559 in the testing) in this node, which is 8% (10.3% n the testing) of the grand total number of observations in the training data set (5333) and (5413 in the testing data set). This is the second smallest number of people in a terminal node. The index of node 12 (20.7%) (27.9% in the testing data set) tells us that this node has people who used the PHR system is at a rate of about 80% lower than the overall average rate of success of 37.6% (37.8% in the testing) in this terminal node.

Gain is equal to the number that use the PHR, in each node, divided into the total number that use the PHR (5333). If I were to select half of the available customers, according to the graph of percentile vs. gain in Figure 7, we would have about 64% of the users of the PHR. (Just to note, the percentile sorts the file from most likely to least likely to use the PHR.) The straight line, tells us, by the way, that if a half of the available customers were selected, about 50% would show up in the file. As same in testing dataset, if I were to select half of the available customers, according to the graph of percentile vs. gain, we would have about 64% of the users of the PHR. The straight line, tells us, by the way, that if a half of the available customers were selected, about 50% would show up in the file.

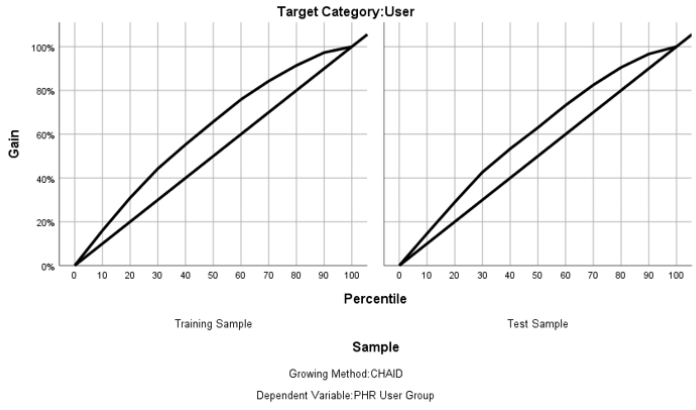


Figure 7

The response rate is the rate at which patients use the PHR. Percentile tells us that the file is sorted by most likely to use PHR to least likely to use the PHR system. According to the graph in Figure 7, of percentile vs. response, the top 10% of customers have an overall PHR user rate of about 60%; the top 50% of the patients had PHR user rate of about 50%; and all of the available patients in the training data set had an overall rate of about 37% of using the PHR online system.

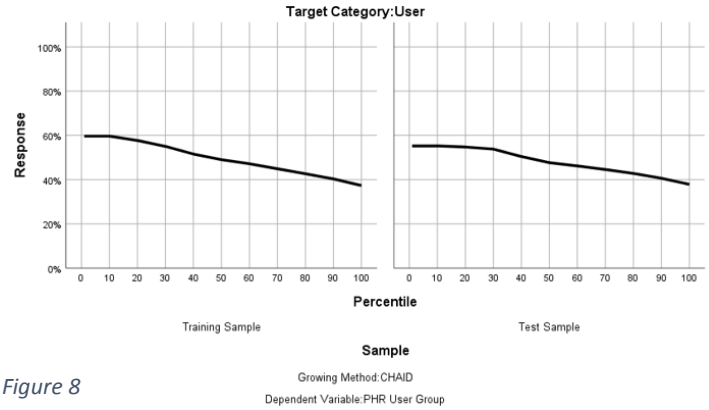


Figure 8

In the testing dataset, In Figure 8, according to the graph of percentile vs. response, the top 10% of customers have an overall PHR user rate of about 60%, the top 50% of the patients had PHR user rate of about 50%; and all of the available patients in the testing data set had an overall rate of about 37% of using the PHR online system.

The index, or the lift, is equal to the node response rate divided into the overall response rate. For example, as seen in Figure 9, the top 50% of patients have an index of about 130%. In other words, the top 50% of these customers are 1.30 times more likely to use the PHR system than the overall 100% of the patients.

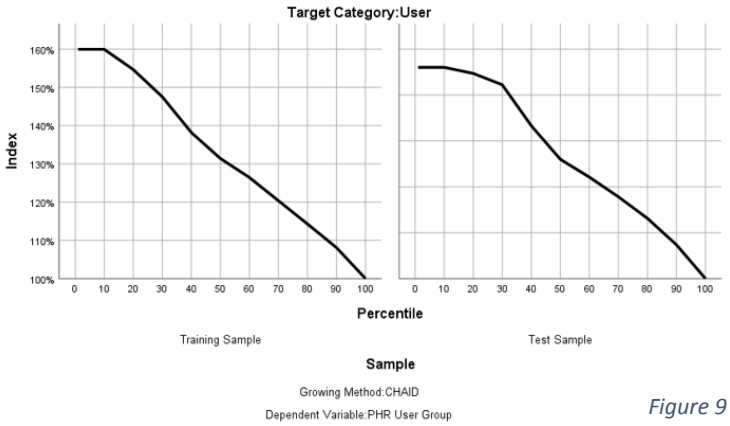


Figure 9

For testing dataset, for example, as seen in Figure 9, the top 50% of patients have an index of about 130%. In other words, the top 50% of these customers are 1.30 times more likely to use the PHR system than the overall 100% of the patients.

Risk

Sample	Estimate	Std. Error
Training	.342	.006
Test	.360	.007

Growing Method: CHAID

Dependent Variable: PHR User Group

Figure 10

Risk is defined as the number of those predicted incorrectly divided into the total sample size (for the training dataset, this is 5333. For the testing, it's 5413). Here, the risk for the training dataset is 0.342 = (345+1479)/5333. The percentage of the time that we predicted incorrectly is 34.2%. The standard error of risk is close to zero; it's 0.006, as seen in the Risk table in Figure 10. The risk for the testing dataset is 0.360 = (412+1539)/5413. The percentage of the time that we predicted incorrectly is 36.0%. The standard error of risk is close to zero, is 0.007.

Specificity is the percentage of patients who didn't use the PHR, who were correctly predicted. Sensitivity is the percentage of patients who used the PHR system, who were correctly predicted. For the training dataset, the specificity is 89.7% = (2999/(2999+345)) and the sensitivity is 25.6%, in Figure 11. So most of those who did not use the PHR data set were correctly predicted, but most of those who did use PHR were not correctly predicted.

Classification

Sample	Observed	Predicted: Nonuser	Predicted: User	Percent Correct
Training	Nonuser	2999	345	89.7%
	User	1479	510	25.6%
	Overall Percentage	84.0%	16.0%	65.8%
Test	Nonuser	2954	412	87.8%
	User	1539	508	24.8%
	Overall Percentage	83.0%	17.0%	64.0%

Growing Method: CHAID

Dependent Variable: PHR User Group

Figure 11

For the testing data, in this table of those patients who didn't use the PHR we correctly predicted 87.8% = (2954/(412+2954))x100% and of those patients who used the PHR system we correctly predicted 24.8%. Overall, we predict 64.0% of cases correctly, according to Figure 11. Specificity focusing on the actual of those patients who didn't use the PHR system. So 87.8% of those patients who didn't use the PHR system were correctly predicted. Sensitivity basically talking about row percentages, 24.8% of those who did use PHR were correctly predicted.

Approximately 50% of the cases (SAMPLE)	Terminal Node Identifier	Mean
Testing data	7	.2325
	8	.2129
	9	.2492
	10	.2501
	11	.2419
	12	.0952
	13	.1704
	14	.1800
	15	.2349
	Total	.2184
Training data	7	.2377
	8	.1967
	9	.2407
	10	.2500
	11	.2336
	12	.0713
	13	.1581
	14	.1734
	15	.2395
	Total	.2116

Figure 12

The Average Squared Error (or ASE) gives an idea of how well the model predicted the observations are actually true or false, or of the data's successfulness or not. It is found by subtracting the probability of success from each target value (which are all going to be 0s or 1s) in each of the observations, squaring each of these differences, and then dividing the sum of the differences by the total number of observations in the dataset. We want this measurement to be as small as possible, as that will tell us that there is little error. We can also use this measure for comparing models from datasets. ASE is used in part of the validation and can be used to detect overfitting. We can also compare nodes.

As seen above in Figure 12, for the training data, the ASE for node 10 (the "mean" of the squared errors) is the largest of all of the nodes, with ASE = 0.25. So, the model doesn't predict the best for those people considered Caucasians, who have commercial health insurance, and who have an HBA1C percentage last documented in the study period between 6.60% and 7.90%. On the contrary, node 12 has an ASE of 0.0713, the smallest of all of the nodes. The model predicts most accurately for those considered to be racial minorities, who have "other" health insurance, and who have a household income less than or equal to $41855. This happens to be the people in the node with the smallest response rate (or rate of success).

Same as in testing dataset, in Figure 12, the ASE for node 10 (the "mean" of the squared errors) is the largest of all of the nodes, with ASE = 0.2501. On the contrary, node 12 has an ASE of 0.0952, the smallest of all of the nodes. Because there is little difference in the ASE for the training and testing datasets, with their total model ASE measures being 0.2116 and 0.2184, respectively, the model fits well.

CART

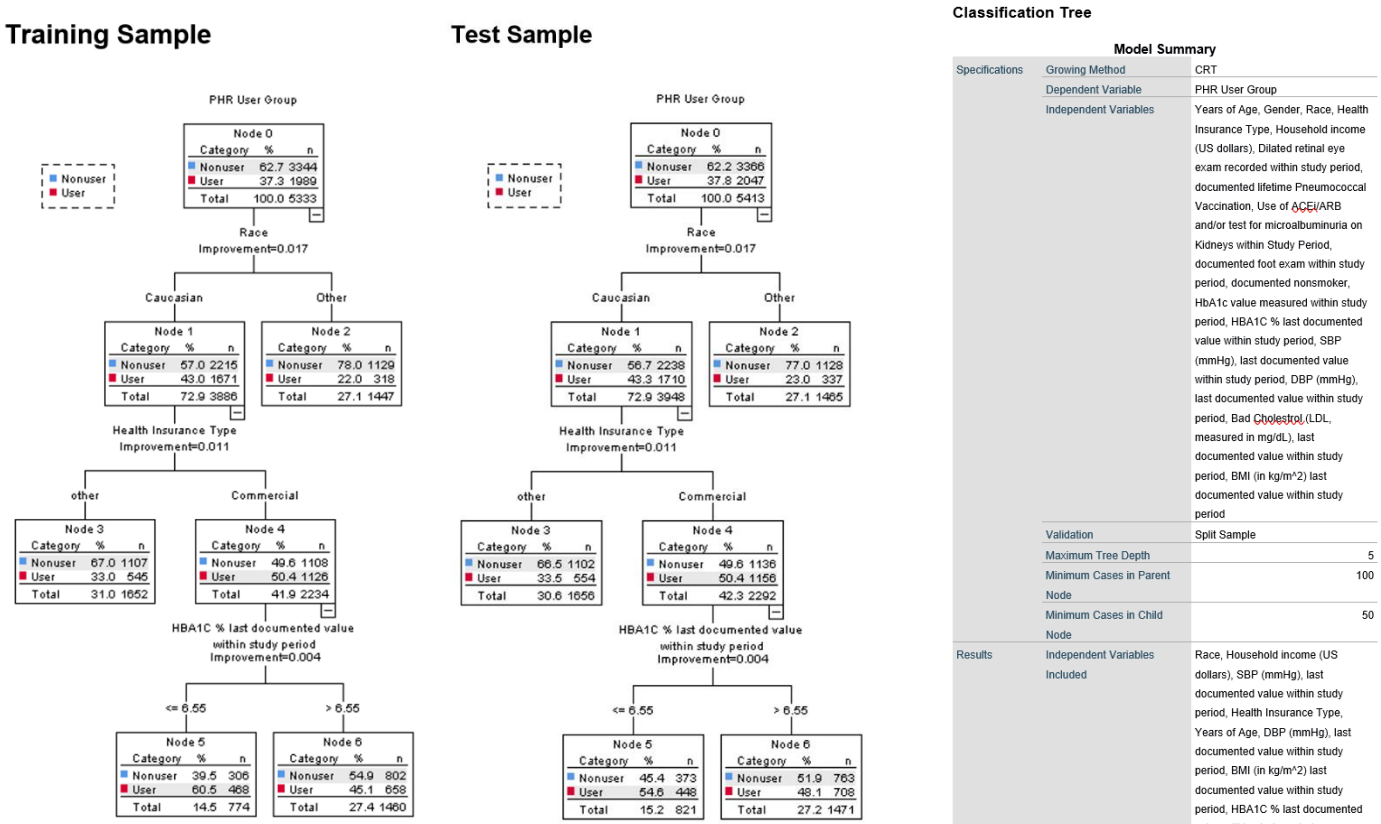


Figure 14: CART decision trees

Classification Tree

Model Summary

Specifications	Growing Method	CRT
	Dependent Variable	PHR User Group
	Independent Variables	Years of Age, Gender, Race, Health Insurance Type, Household income (US dollars), Dilated retinal eye exam recorded within study period, documented lifetime Pneumococcal Vaccination, Use of ACEI/ARB and/or test for microalbuminuria on Kidneys within Study Period, documented foot exam within study period, documented nonsmoker, HbA1c value measured within study period, HBA1C % last documented value within study period, SBP (mmHg), last documented value within study period, DBP (mmHg), last documented value within study period, Bad Cholestrol (LDL, measured in mg/dL), last documented value within study period, BMI (in kg/m^2) last documented value within study period
	Validation	Split Sample
	Maximum Tree Depth	5
	Minimum Cases in Parent Node	100
	Minimum Cases in Child Node	50
Results	Independent Variables Included	Race, Household income (US dollars), SBP (mmHg), last documented value within study period, Health Insurance Type, Years of Age, DBP (mmHg), last documented value within study period, BMI (in kg/m^2) last documented value within study period, HBA1C % last documented value within study period
	Number of Nodes	7
	Number of Terminal Nodes	4
	Depth	3

Figure 13

In our model summary of the CART, in Figure 13, we used the same set of variables as above with the CHAID and Exhaustive CHAID. For the CART model, the algorithm ended up selecting the variables Race, Household income (US dollars), SBP (mmHg), last documented value within study period, Health Insurance Type, Years of Age, DBP (mmHg), last documented value within study period, BMI (in kg/m^2) last documented value within study period, HBA1C % last documented value within study period to predict whether the patients use the PHR system or not. There are only 3 levels, with 7 nodes, 4 of which are terminal nodes. The validation (or the splitting of the data set into the training and testing data) was done by the split sample method, once again.

Our CART algorithm produced a very "cute" tree, as shown in Figure 14; it is very concise, but still tells us a lot of information about the patients and the factors that went into patients choosing to use the PHR system or not. The training and testing decision are very similar and have the same levels and splits (see Figure 14). So, the first node of the training data set, tells us that the response rate, or the rate of success, is 37.3%. So, 37.3% of the overall set of patients, ended up using the PHR system. The first split is by race, which creates the first level of this data set. Those who are Caucasian use the PHR system at a rate of 43% (see Node 1). Those who are "other" (or rather, minorities), use the PHR system at about half the rate of the people who are Caucasian; the rate of success for minorities is 22% (see node 2). This is the

first terminal node. Next, we have another split from node 1 into nodes 3 and 4, by the variable health insurance type. For the people who are Caucasian and who have the “other” type of health insurance (such as government-funded), the rate of usage of the PHR system is 33% (node 3). In node 4, for the people who are Caucasian and who have the “commercial” type of health insurance, the rate of usage of the PHR system is 50.4%. This latter node leads into the next and final split, by the variable HBA1C % last documented. For the people who are Caucasian, who have the “commercial” type of health insurance, and who have a value of HBA1C of <= 6.55, the rate of usage of the PHR system is 60.5% (node 5). For the people who are Caucasian, who have the “commercial” type of health insurance, and who have a value of HBA1C of > 6.55, the rate of usage of the PHR system is 45.1% (node 6).

The first node of the testing data set, tells us that the rate of success, as seen in Figure 14, is 37.8%, this is the overall set of patients, ended up using the PHR system. The first split is by race, patients who are Caucasian, use the PHR system at a rate of 43.3% in Node 1, and those who are “other” (or rather, minorities), use the PHR system, their rate of success for minorities is 23.0% in node 2, which is the first terminal node. We have another split from node 1 into nodes 3 and 4, by the variable health insurance type. For the people who are Caucasian and who have the “other” type of health insurance (such as government-funded), the rate of usage of the PHR system is 33.6% in node 3, and in node 4 for the people who are Caucasian and who have the “commercial” type of health insurance, the rate of usage of the PHR system is 50.4%. This node split into the variable HBA1C % last documented. For the people who are Caucasian, who have the “commercial” type of health insurance, and who have a value of HBA1C of <= 6.55, the rate of usage of the PHR system is 54.6% in node 5. For the people who are Caucasian, who have the “commercial” type of health insurance, and who have a value of HBA1C of > 6.55, the rate of usage of the PHR system is 48.1% in node 6.

The terminal nodes 5, 6, and 2 were found to be the most interesting, as shown in Figure 15. In the training data set CART model, node 5 has the people who are Caucasian, who have the “commercial” type of health insurance, and who have a value of HBA1C of <= 6.55. These people have the highest rate of response (of using the PHR system) of all of the terminal nodes, which is 60.5%. There are 774 people in this node (or 14.5% of the total number of people in the training data set (5333). This is the smallest among of people in a node in the training data set. The gain for node 5 is 23.5%. So, there are 468 patients who used the PHR system out of the 774 in this node, which amounts to this percentage/proportion (468/774=.235). Next, the index for node 5 is 162.1%. That is, people in this node are 1.621 times more likely to use the PHR system than the overall average response of the people (37.3%).

Target Category: User

Gains for Nodes

		Node		Gain			
Sample	Node	N	Percent	N	Percent	Response	Index
Training	5	774	14.5%	468	23.5%	60.5%	162.1%
	6	1460	27.4%	658	33.1%	45.1%	120.8%
	3	1652	31.0%	545	27.4%	33.0%	88.5%
	2	1447	27.1%	318	16.0%	22.0%	58.9%
Test	5	821	15.2%	448	21.9%	54.6%	144.3%
	6	1471	27.2%	708	34.6%	48.1%	127.3%
	3	1656	30.6%	554	27.1%	33.5%	88.5%
	2	1465	27.1%	337	16.5%	23.0%	60.8%

Growing Method: CRT

Dependent Variable: PHR User Group

Figure 15

In the training data set CART model, node 2 has the people who are “other” (or rather, minorities) and who use the PHR system at about half the rate of the people who are Caucasian. These people have the lowest rate of response (of using the PHR system) of all the terminal nodes. There are 1447 people in this node (or 27.1% of the total number of people in the training data set (5333), giving us 45.1% as the response rate. The gain for this node is 16% = 318/1447. There are 318 patients who used the PHR system out of the 1447 in this node.

Lastly, the index for node 2 is 58.9%. So, people in this node are (1-.589=) 0.411 times less likely to use the PHR system than the overall average response of the people (37.3%). In the training data set CART model, node 6 has the people who are Caucasian, who have the “commercial” type of health insurance, and who have a value of HBA1C of > 6.55. These people have the most average rate of response (of using the PHR system), 45.1%, of all of the terminal nodes. There are 1460 people in this node (or 27.4% of the total number of people in the training data set (5333). The gain for node 5 is 33.1%. So, there are 658 patients who used the PHR system out of the 1460 in this node, which amounts to this percentage/proportion (658/1460=.235). Lastly, the index for node 5 is 120.8%. That is people in this node are 1.208 times more likely to use the PHR system than the overall average response of the people (37.3%).

In testing dataset, node 5 has the highest rate of response (of using the PHR system) is 144.3%, as shown in Figure 15. Gain is 21.9% and index is 144.3% that means people in this node are 1.443 times more likely to use the PHR system than the overall average response of the people (37.8%). Node 2 has the lowest rate of response (of using the PHR system) is 23.0%. The gain is 16.5% and index is 60.8% that means people in this node are 0.608 times less likely

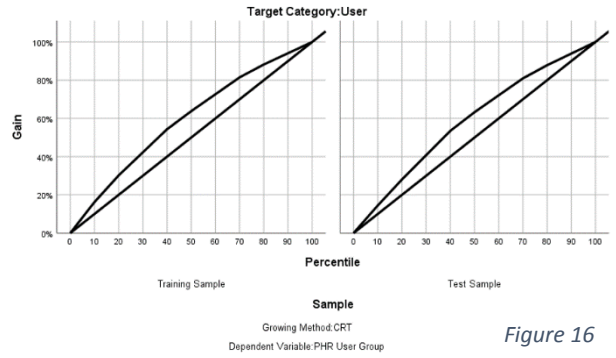


Figure 16

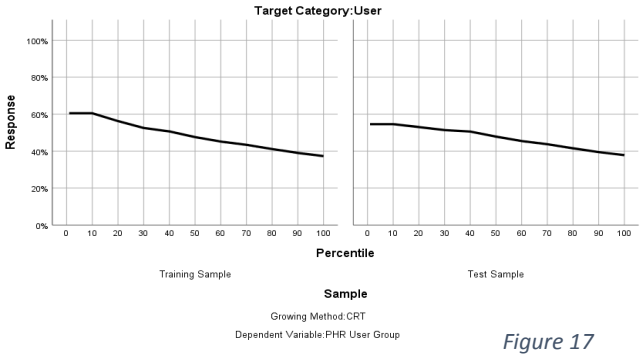


Figure 17

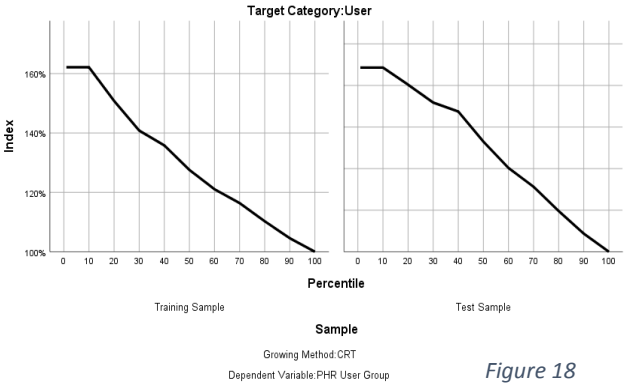


Figure 18

to use the PHR system than the overall average response of the people (37.8%). Node 6 has the average rate of response (of using the PHR system) is 48.1%. Gain is 34.6% and index is 127.3% that means people in this node are 1.273 times more likely to use the PHR system than the overall average response of the people (37.8%).

The percentile (sorted from the likelihood that they responded (from most likely to least likely to use the PHR system) vs. gain, as described above in the interpretations, is shown in Figure 16. If we randomly select 30% of the patients, according to the straight line in both of the plots, then 30% respond with using the PHR system. According to the curved line, if we pick the top 30% of patients, we will get about 41% of them to respond with using the PHR system.

The percentile (sorted from the likelihood that they responded (from most likely to least likely to use the PHR system) vs. response rate, is in Figure 17. If we randomly select the top 30% of the patients, according to both plots (for the training and testing sample), then 50% respond with using the PHR system.

The percentile (sorted from the likelihood that they responded (from most likely to least likely to use the PHR system) vs. index is in Figure 18. If we randomly select the top 30% of the patients, according to both plots (for the training and testing sample), then their index will be about 140%. This tells us that the top 30% of patients will use the PHR system about 1.4 times that the overall response rate.

For both the training and testing datasets, the risk is the value of the overall misclassifiation rate of the respective datasets. Notice that, for the training data set, in the risk table in Figure 19, that the overall percentage of correct predicts (in the classification table in Figure 20), 65.7%, is 100%-65.7%=34.3%, the response rate for the training data CART model. For testing dataset, that the overall percentage of correct predicts (in the classification table), 63.6% which is 100-63.3%=36.4% the response rate for the testing data CART model.

Risk

Sample	Estimate	Std. Error
Training	.343	.006
Test	.364	.007

Growing Method: CRT

Dependent Variable: PHR User Group

Figure 19

Classification

Sample	Observed	Predicted: Nonuser	Predicted: User	Percent Correct
Training	Nonuser	3038	306	90.8%
	User	1521	468	23.5%
	Overall Percentage	85.5%	14.5%	65.7%
Test	Nonuser	2993	373	88.9%
	User	1599	448	21.9%
	Overall Percentage	84.8%	15.2%	63.6%

Growing Method: CRT

Dependent Variable: PHR User Group

Figure 20

Recall that specificity is the percentage of patients who didn't use the PHR, who were correctly predicted. Sensitivity is the percentage of patients who used the PHR system, who were correctly predicted. The classification table in Figure 20 tells us a lot of information about specificity and sensitivity of the two data sets. First, with specificity of the training dataset, which is the percentage of patients who were correctly predicted to not use the PHR system, is 90.8% = (3038/(3038+306))*100%. Now, the sensitivity of the training dataset, which is the percentage of patients who were correctly predicted to use the PHR system, is 23.5% = (1521/(1521+468))*100%. In testing dataset, specificity is (2993/(2993+373))*100% = 88.9% and sensitivity is 21.9% and the overall percentage is 63.6%.

The Average Squared Error (or ASE) for the training and testing data sets are 0.2175 and 0.2223, respectively, as shown in Figure 21. Thus, because these values are so close, the CART model built fits well on the data (it doesn't over fit the data). Also, note that these ASE values are consistent with the CHAID training and testing overall ASEs.

As seen above, for the training data, the ASE for node 2 (the "mean" of the squared errors) is the smallest of all of the nodes, with ASE = 0.1772. So, the model predicts the best for those people considered those who are "other" (or

Summarize

ASE for Terminal Nodes CART algorithm

Mean Approximately 50% of the cases (SAMPLE)	Terminal Node Identifier	squared error, CART
testingdata	2	.1772
	3	.2226
	5	.2514
	6	.2506
	Total	.2223
trainingdata	2	.1715
	3	.2211
	5	.2390
	6	.2476
	Total	.2175

Figure 21

rather, minorities) and who use the PHR system at about half the rate of the people who are Caucasian. On the contrary, node 5 has an ASE of 0.2514, the largest of all the nodes. The model does not predict most accurately for those considered to Caucasian, who have the "commercial" type of health insurance, and who have a value of HBA1C of <= 6.55.

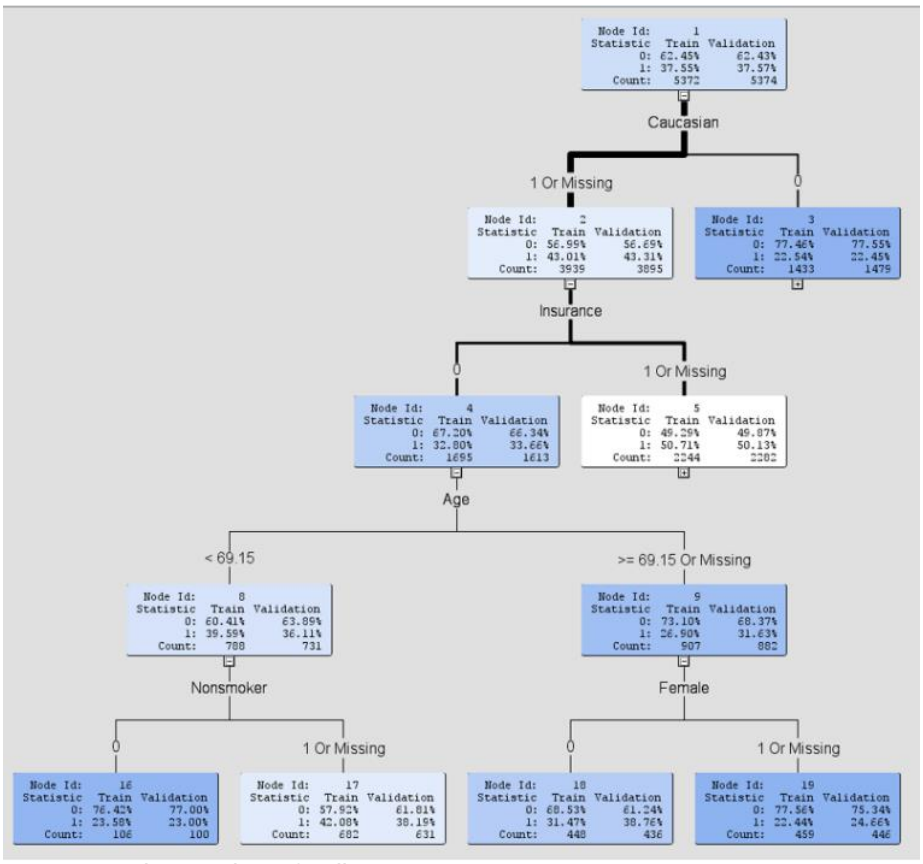


Figure 22

The number of splits were 11 and the number of terminal nodes being 12. There are also 4 levels.

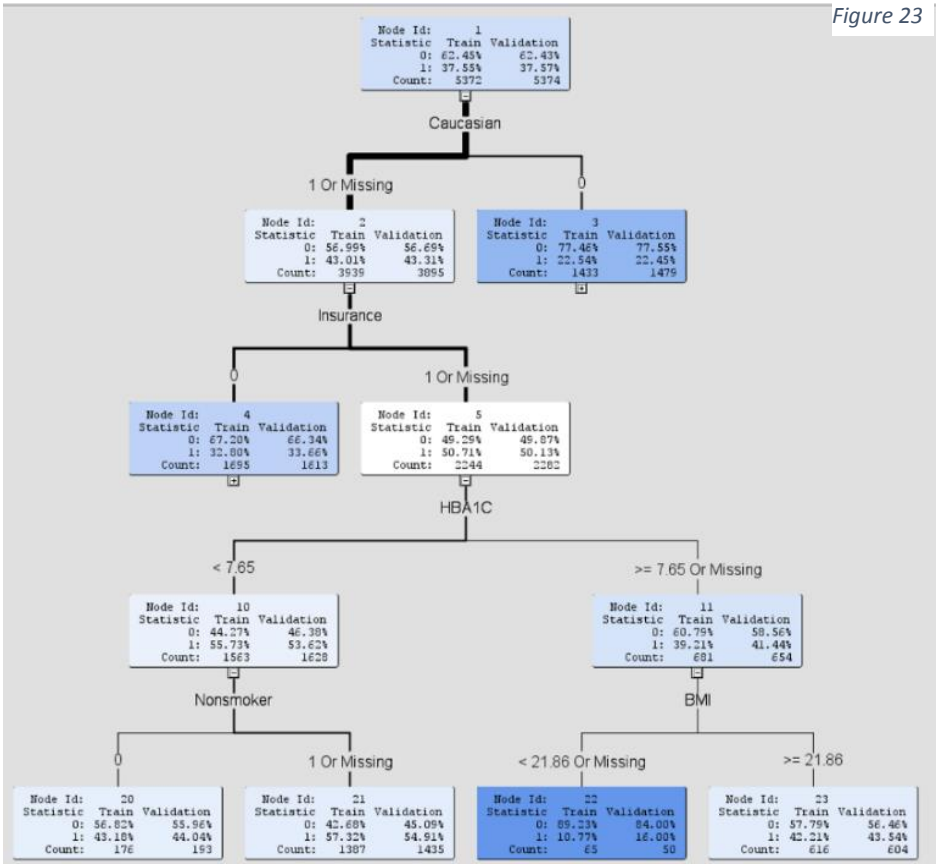
First, node 16, in Figure 22 tells us that, of the people who are smokers, whose age is less than 69.15 years, whose insurance type is 0, or "other", and who is Caucasian or missing a recorded race identity, the response rate of using the PHR is 23.58% for the training and 23% for the validation (testing). In addition, of these people, 76.42% (training; 77% validation) of them did not use the PHR system. Node 17 tells us that, of the people who are not smokers, whose age is less than 69.15 years, whose insurance type is 0, or "other", and who is Caucasian or missing a recorded race identity, the response rate of using the PHR is 42% for the training and 38.19% for the validation (testing). In addition, of these people, 57.92% (training; 61.81% validation) of them did not use the PHR system. Next, Node 18 tells us that, of the people who are male, whose age is greater than or equal to 69.15 years, whose insurance type is 0, or "other",

In testing dataset, the ASE for node 2 (the "mean" of the squared errors) is the smallest of all of the nodes, with ASE = 0.1715, as shown in Figure 21. So, the model predicts the best for those people considered those who are "other" (or rather, minorities) and who use the PHR system at about half the rate of the people who are Caucasian. Node 6 has an ASE of 0.2476, the largest of all of the nodes . The model does not predict most accurately for those considered to Caucasian, who have the "commercial" type of health insurance, and who have a value of HBA1C of <= 6.55. The nodes 5 and 6 are different from training and testing because the two node ASEs are so close together, they are basically the same in this measure; they are both the largest ASE, and this is consistent with the training and testing datasets.

SAS Enterprise Miner Decision Tree Output:

The SAS Enterprise Miner Decision Tree Output is a combination of the CHAID and CART algorithms. This gave the most splits and terminal nodes.

Figure 23



and who is Caucasian or missing a recorded race identity, the response rate of using the PHR is 31.47% for the training and 38.76% for the validation (testing). In addition, of these people, 68.53% (training; 61.24% validation) of them did not use the PHR system.

Node 19 tells us that, of the people who are female (or have a missing gender record), whose age is greater than or equal to 69.15 years, whose insurance type is 0, or "other", and who is Caucasian or missing a recorded race identity, the response rate of using the PHR is 22.44% for the training and 24.66% for the validation (testing). In addition, of these people, 77.56% (training; 75.34% validation) of them did not use the PHR system.

Node 20, in Figure 23, tells us that, of the people who are smokers, whose HBA1C level is less than 7.65, whose insurance type is 1, or "commercial" (or missing an insurance indicator in their record, and who are Caucasian or missing a recorded race identity, the response rate of using the PHR is 43.18% for the training and 44.04% for the validation (testing). In addition, of these people, 56.82% (training; 55.96% validation) of them did not use the PHR system.

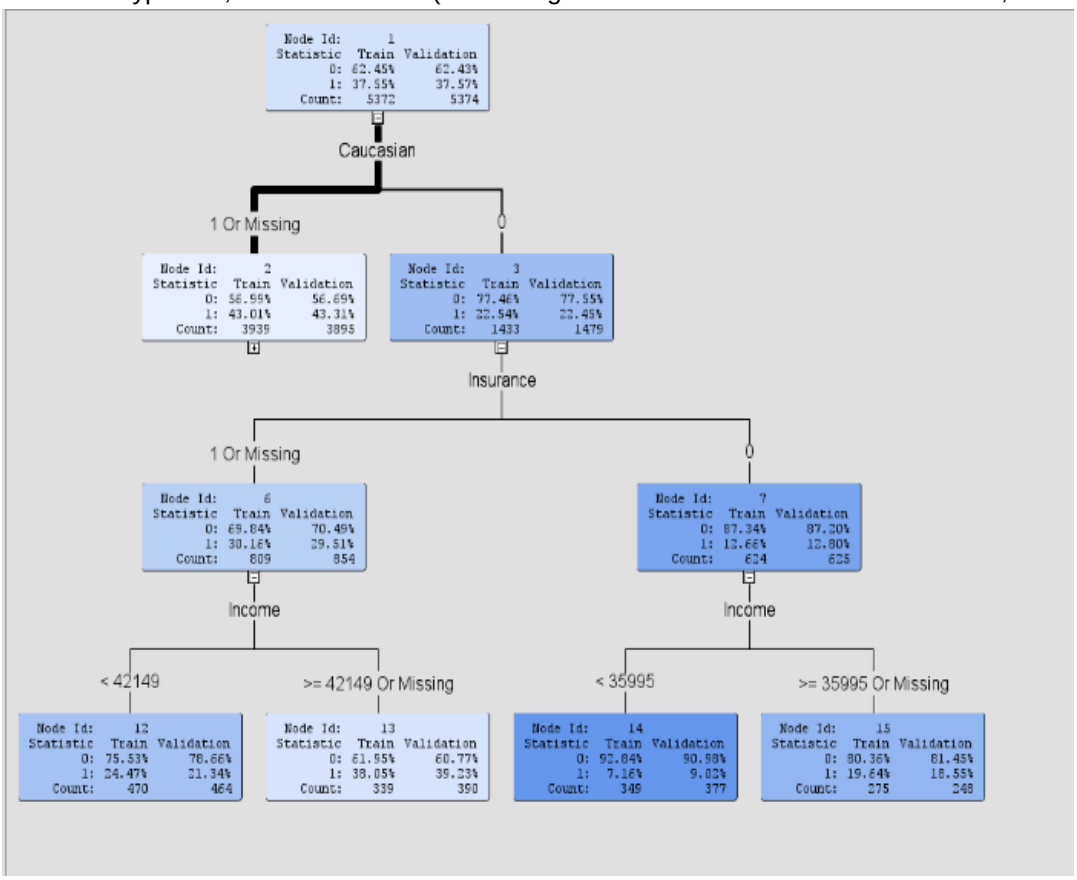


Figure 24

Node 21 tells us that, of the people who are not smokers, whose HBA1C level is less than 7.65, whose insurance type is 1, or "commercial" (or missing an insurance indicator in their record, and who are Caucasian or missing a recorded race identity, the response rate of using the PHR is 57.32% for the training and 54.91% for the validation (testing). In addition, of these people, 42.68% (training; 45.09% validation) of them did not use the PHR system.

Node 22 tells us that, of the people whose BMI < 21.86 or missing, whose HBA1C level is greater than or equal to 7.65 (or missing), whose insurance type is 1, or "commercial" (or missing an insurance indicator in their record, and who are Caucasian or missing a recorded race identity, the response rate of using the PHR is 10.77% for the training and 16% for the validation (testing). In addition, of these people, 89.23% (training; 84% validation) of them did not use the PHR system.

Node 23 tells us that, of the people whose BMI ≥ 21.86 or missing, whose HBA1C level is greater than or equal to 7.65 (or missing), whose insurance type is 1, or "commercial" (or missing an insurance indicator in their record, and who are Caucasian or missing a recorded race identity, the response rate of using the PHR is 42.21% for the training and 43.54% for the validation (testing). In addition, of these people, 57.79% (training; 56.46% validation) of them did not use the PHR system.

Node 12, in Figure 24, tells us that, of the people whose income< 42149, whose insurance type is 1, or "commercial" (or missing an insurance indicator in their record), and who are not Caucasian, the response rate of using the PHR is 24.47% for the training and 21.34% for the validation (testing). In addition, of these people, 75.53% (training; 78.66% validation) of them did not use the PHR system.

Node 13 tells us that, of the people whose income ≥ 42149 or missing, whose insurance type is 1, or "commercial" (or missing an insurance indicator in their record), and who are not Caucasian, the response rate of using the PHR is 38.05% for the training and 39.23% for the validation (testing). In addition, of these people, 61.95% (training; 60.77% validation) of them did not use the PHR system.

Node 14 tells us that, of the people whose income< 35995, whose insurance type is 0, or "other", and who are not Caucasian, the response rate of using the PHR is 7.16% for the training and 9.02% for the validation (testing) (this is the lowest response rate). In addition, of these people, 92.84% (training; 90.58% validation) of them did not use the PHR system.

Node 15 tells us that, of the people whose income ≥ 35995, whose insurance type is 0, or "other", and who are not Caucasian, the response rate of using the PHR is 19.84% for the training and 18.55% for the validation (testing) (this is the lowest response rate). In addition, of these people, 80.36% (training; 81.45% validation) of them did not use the PHR system.

Fit Statistics

Target=User Target Label=' '

Fit

Statistics	Statistics Label	Train	Validation
ASE	Average Squared Error	0.21	0.22
MISC	Misclassification Rate	0.34	0.35

Figure 25

As discussed above, the average squared error (ASE) can be used in taking a look at overfitting and looking at the error in the model. We want the value to be small, and the overall training and validation ASEs to be close together. For the training and validation (testing) data sets, the ASE is 0.21 and 0.22, both displayed n Figure 25, respectively. These are small (close to 1 and less than 0.5), close to one another (so this decision tree model is good), and consistant with the ASE values of the other models (CART and CHAID) above.

Classification Table

Data Role=TRAIN Target Variable=User Target Label=' '

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	69.3350	82.3547	2763	51.4334
1	0	30.6650	60.5850	1222	22.7476
0	1	42.6820	17.6453	592	11.0201
1	1	57.3180	39.4150	795	14.7990

Data Role=VALIDATE Target Variable=User Target Label=' '

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	68.7484	80.7154	2708	50.3908
1	0	31.2516	60.9708	1231	22.9066
0	1	45.0871	19.2846	647	12.0394
1	1	54.9129	39.0292	788	14.6632

Figure 26

Next, is the misclassification rate, or the risk, which is equal to 0.34 = (1222+592)/(2763+1222+592+795) = (22.75+11.02)/(100) for the training and 0.35 = (1231+647)/(2708+1231+647+788) = (22.91+12.04)/(100) for the validation data sets, as shown in Figure 25, with the numbers in the calculations coming from Figure 26.This is also consistant with the other two decision tree models. These are similar values for risk as the other decision trees (CART and CHAID) from above.

Taking a look at the classification table for the training and validation datasets, in Figure 26, I am assuming that the target is the observed and the outcome is the predicted. Here, we have the target precetage vs the outcome percentages. Notice the differences and errors. Next, the specificity, which is the percentage of times that the model correctly predicts the nonusers of the PHR system, it's 82.35% = (2763/(2763+592))*100%. Now, the sensitivity of the training dataset, which is the percentage of times that the model correctly predicts the users of the PHR system, is 39.42% = (795/(795+1222))*100%. These measures of specificity and sensitivity are different from the CHAID and CART output (in this case, sensitivity is larger in this decision tree than in the other two, which is good). These will be compared in the conclusion. In testing dataset, specificity is (2708/(2708+647))*100% = 80.7% and sensitivity is =39.03 = (788/(788+1231))*100%, and the overall percentage response rate is for training is = (2763+795)/(592+2763+795+1222) = 0.66232315711,and testing is (2708+788)/(647+2708+788+1231) = 0.65053963528. These are consistent with the other decision tree response rates.

All of the highest and lowest rated nodes are consistent for all of these decision tree models. Those who do not have as many resources and who have the other type of health insurance (such as Medicaid), and who are not white, are the least likely to monitor their health status online, than those who are not like them. This is saddening; if you don't know, Cleveland and surrounding areas are the most racially and socioeconomically stratified in terms of resources, to these people may not have as many resources, such as internet access. This was in the decade, so people were not as connected as we are now.

Logistic Regression:

The variables that were used to build the logistic regression model were age, gender, race, insurance type, income, HBA1C percentage, and smoking status because, as noted in the introduction, these variables all in some way were associated with predicting PHR usage in the decision tree models in the first few models of this project.

The multicollinearity assumption of logistic regression: There does not appear to be correlations between the two numerical variables age, income, and HBA1C, according to this correlation matrix because all of the absolute values of the bivariate correlations between each of the combinations of these variables are less than 0.5, as shown in Figure 27. *Imputations for the missing values of the numeric variables:* I replaced the missing values in the numerical variables that will be included in the model by placing the mean of

Correlations

		Years of Age	Household income (US dollars)	HBA1C % last documented value within study period
Years of Age	Pearson Correlation	1	.050**	-.175**
	Sig. (2-tailed)		.000	.000
	N	10746	10558	10208
Household income (US dollars)	Pearson Correlation	.050**	1	-.086**
	Sig. (2-tailed)	.000		.000
	N	10558	10558	10032
HBA1C % last documented value within study period	Pearson Correlation	-.175**	-.086**	1
	Sig. (2-tailed)	.000	.000	
	N	10208	10032	10208

**. Correlation is significant at the 0.01 level (2-tailed).

Figure 27

Result Variables

	Result Variable	N of Replaced Missing Values	Case Number of Non-Missing Values First	Case Number of Non-Missing Values Last	N of Valid Cases	Creating Function
1	Age_1	0	1	10746	10746	SMEAN(Age)
2	Income_1	188	1	10746	10746	SMEAN (Income)
3	HBA1C_1	538	1	10746	10746	SMEAN (HBA1C)

Figure 28

the known observations for each variable in for each of the missing values for the variable, as shown in the Result variables table in Figure 28. So, we will include some form of these numeric variables in the model. I used the mean of all of the known observations in the variable to input for the whole data set (this should be fine because the mean is just an *estimate* for the missing observations. The model was then split by the test variable into a training and testing dataset for the logistic regression model verification. I fit a model on the training dataset, then used that model to fit the testing data set.

The selected cases are the cases from the training data and the unselected are testing data. First, for the training dataset, 5333 of the observations were selected to build the logistic regression model on, as shown in Figure 29. The other 5413 unselected cases were saved for later analysis. The SPSS output, however, automatically fit the logistic regression model that was from the training data onto the testing data, as you will see in the classification table below, so that the risk (misclassification rate), the correct classification rate, and the sensitivity and specificity can be analyzed in the two data sets and compared to each other. Also, as you can see in the case processing summary, there were no missing cases, after I imputed the variables with missing values.

Logistic Regression

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	5333	49.6
	Missing Cases	0	.0
	Total	5333	49.6
Unselected Cases		5413	50.4
Total		10746	100.0

a. If weight is in effect, see classification table for the total number of cases.

Figure 29

Now, taking a look at the Model Summary comparison statistics in Figure 30, both the Cox & Snell R square (which takes into account the log likelihood of the final model vs. the log likelihood of the baseline model) and Nagelkerke R square (a modification of the Cox & Snell) statistics increase as the steps increase (as more variables are added to the model) from 0.039 (step 1) to 0.098 (step 6) and from 0.053 (step 1) to 0.133 (step 6), for each of the respective measures. The -2 log likelihood ratio also decreases as the number of variables added to the model increases from 6834.745 to 6497.442 in steps 1 and 6, respectively. The variables that were added in each step added significance to the model, and we should go with the model 9the last one) with the largest Cox & Snell R2 measurements and smallest log likelihood. However, these measures are substitutes (or Pseudo R2 measures) two of the true R2 in a least squares regression model, and are not as easy to interpret as other measurements of model evaluation. The specificity, sensitivity, risk, and ASE are good measurements, and will be discussed later.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	6834.745[a]	.039	.053
2	6650.377[a]	.071	.097
3	6594.471[a]	.081	.111
4	6555.172[a]	.088	.120
5	6528.596[a]	.092	.126
6	6497.442[a]	.098	.133

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

Figure 30

The logit of this model is logit($\hat{\pi}$) = -0.214 - 0.02*age_1 +0.000*income_1 – 0.172*HBA1C + 0.85*race+0.614*Insurance+0.533*nonsmoker, according to Figure 31. All of these terms are significant at the 5% level (but the constant term is not), according to the Wald tests (which tests the null hypothesis of the true coefficient of a predictor variable to be equal to zero vs. the alternative that it is not equal to zero. It used the chi-square test statistic) for each, as seen in the variables in the equation table above. So, we conclude that each of the predictor variable coefficients do not equal zero, at the 5% level of significance. The standard errors of each of the coefficients of the predictor variables are also relatively small, when compared to the coefficients.

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 6[f]	SMEAN(Age)	-.020	.004	32.878	1	.000	.980	.973	.987
	SMEAN(Income)	.000	.000	38.944	1	.000	1.000	1.000	1.000
	SMEAN(HBA1C)	-.172	.024	53.196	1	.000	.842	.804	.882
	Race	.850	.078	118.086	1	.000	2.340	2.007	2.727
	Health Insurance Type	.614	.074	69.432	1	.000	1.849	1.600	2.136
	documented nonsmoker	.533	.098	29.778	1	.000	1.704	1.407	2.064
	Constant	-.214	.326	.434	1	.510	.807		

a. Variable(s) entered on step 1: Race.
b. Variable(s) entered on step 2: Health Insurance Type.
c. Variable(s) entered on step 3: SMEAN(HBA1C).
d. Variable(s) entered on step 4: SMEAN(Income).
e. Variable(s) entered on step 5: SMEAN(Age).
f. Variable(s) entered on step 6: documented nonsmoker.

Figure 31

The age coefficient interpretation is as follows: the adjusted odds ratio is $e^{-0.02} = 0.98$, according to Figure 31. While adjusted for the other variables in the model, we predict that the odds of a diabetic patient using the PHR system decreases by ((0.98-1)= -0.02) 2% for each additional year of age. The 95% confidence interval for the odds of age is, for every increase in one year of a diabetic patient's age, the predicted odds of using the PHR system decreases by ([(0.973-1) to (0.987-1)] or 0.027 to 0.013) 2.7% to 1.3% for each additional year of age.

The income coefficient interpretation is as follows: the adjusted odds ratio is $e^0 = 1$. While adjusted for the other variables in the model, we predict that the odds of a diabetic patient using the PHR system is nonchanging for each additional unit of income (in US dollars). This result turned out to be statistically significant due to the large sample size. It is debatably practically significant. It is fine to include in the model here, because it was associated with PHR in two of the decision trees above. But, in further research, a logistic regression model would be performed without income within it. The 95% CI is (1,1) which is redundant to interpretation because the interpretation would be the same as the adjusted odds ratio interpretation. It is left out here.

The HBA1C coefficient interpretation is as follows: the adjusted odds ratio is $e^{-0.172} = 0.842$. While adjusted for the other variables in the model, we predict that the odds of a diabetic patient using the PHR system decreases by (((0.842-1) = -0.158) 15.8% for each additional unit of HBA1C. The 95% confidence interval for the odds of HBA1C is, for every increase in one unit of a diabetic patient's age, the predicted odds of using the PHR system decreases by ([(0.804-1) to (0.882-1)]) or -0.192 to -0.118) 19.2% to 11.8% for each additional unit of HBA1C.

Note that that Race, Health Insurance type, and nonsmoker all have positive predicted odds (0.85, 0.614, and 0.533, respectively), and odds ratios (2.34, 1.849, and 1.704 respectively), and corresponding odds ratio confidence intervals ((2.007, 2.727), (1.6,2.136), (1.407, 2.064), respectively) that are greater than one, according to Figure 31. So, all of these variables are positively associated with the patients that use the PHR system. All of these variables are binary, with 1 denoting Caucasian, commercial health insurance, and 1 for being a nonsmoker, respectively.

For patients who are Caucasian, the adjusted odds ratio is $e^{0.85} = 2.34$. While adjusting for the other variables on the model, the predicted odds of using the PHR system is 2.34 times greater for those who are Caucasian than it is for those who are not. The 95% confidence interval for the odds of race is, for patients who are Caucasian, the predicted odds of using the PHR system is 2.007 to 2.727 times greater than those who are not Caucasian.

For patients who have commercial health insurance, the adjusted odds ratio is $e^{0.614} = 1.849$. While adjusting for the other variables on the model, the predicted odds of using the PHR system is 1.849 times greater for those who have commercial health insurance than it is for those who do not. The 95% confidence interval for the odds of health, insurance type is, for patients who have commercial health insurance, the predicted odds of using the PHR system is 1.60 to 2.136 times greater than those who have another type of health insurance.

For patients who do not smoke, the adjusted odds ratio is $e^{0.533} = 1.704$. While adjusting for the other variables on the model, the predicted odds of using the PHR system is 1.704 times greater for those who do not smoke than it is for those do not. The 95% confidence interval for the odds of nonsmoking status is, for patients who do not smoke, the predicted odds of using the PHR system is 1.407 to 2.064 times greater than those who do smoke.

Contingency Table for Hosmer and Lemeshow Test

		PHR User Group = Nonuser		PHR User Group = User		
		Observed	Expected	Observed	Expected	Total
Step 6	1	486	468.707	47	64.293	533
	2	421	425.069	112	107.931	533
	3	387	391.063	146	141.937	533
	4	369	370.267	164	162.733	533
	5	330	351.628	203	181.372	533
	6	332	329.755	201	203.245	533
	7	298	298.246	235	234.754	533
	8	263	266.580	270	266.420	533
	9	238	240.409	295	292.591	533
	10	220	202.276	316	333.724	536

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
6	12.239	8	.141

Figure 32: Hosmer and Lemeshow test output

The Hosmer and Lemeshow statistics are meant to measure a lack of fit. The observations are partitioned into 10 bins, or equally sized groups (in this case, about 533 are in each group, plus or minus 1 or 2), as shown in Figure 32. These are called decile groups. In the contingency table for Hosmer and Lemeshow test is a classification table of the observed and expected of the people that fall into each bin, in each category of the response variable (diabetes in this case). The cases of observed and expected in each category are sorted by the predicted probability $\hat{\pi}$ (from most likely to occur to least likely to occur). In the part of this contingency table where PHR system was used, in bin 1 there are 47 patients who used the PHR, and this was farily close to its expected value of 64.293. In bin 1 where the patients did not use PHR, there are 486, this is very close to the expected value of this bin and category, which is 468.707. This is seen in all of the bins that the observed is close to its corresponding expected value. This will be talked about in the hypothesis testing; The sum of the (10 bins times 2 response categories =) 20 squared differences between the observed and expected values, that are then divided by each of the corresponding expected values, gives just a small chi-square test statistic to not reject the null hypothesis of the Hosmer Lemeshow test.

The Hosmer and Lemeshow Test tests whether the model is a good fit or not. The null hypothesis is that the model is a good fit and the alternative is that it is not. Because the p-value (= 0.141), according to Figure 32, is not significant at any reasonable significance level, we fail to reject the null. Because the test is a chi-square test, then this conclusion tells us that there is no difference between the observed and the expected. This implies that the model estimates fit the data at an acceptable level.

This classification table in Figure 33 has the output for the training and testing datasets. The rate of correct classification is 61.8% for the training data set, and 69.2% for the testing dataset. The overall risk estimate = (621+1417)/(1927+1368+621+1417) = 0.3821 = 1 – 0.618 = 1 – response rate (rate of misclassification) for training. Similarly, for testing, it is = (631+1429)/(631+1429+1937+1416) = 0.3805= 1 – 0.619. This is seen in the classification table, where the cut value was set to 0.35, close to the percentage of people who were recorded to have used the PHR system in the dataset as a whole, according to the pie-chart of PHR users in Figure 2 above. The model did about the same job of predicting diabetes and no diabetes of the patients in the data set, as the specificity (57.6% = (1927/(1927+1417))*100%); the percentage that the model correctly predicted that patients do not use the PHR system) and sensitivity (68.8% = (1368/(1368+621))*100%) the percentage that the model correctly predicted that patients did use the PHR system). These values of specificity and sensitivity are both are close and rather high (above 50%). Also, note that when the model from the training dataset was used on the testing dataset, the values of the specificity and sensitivity were very close together (note that the model, so there is no overfitting. So, the model predicts correctly which category patients fall in well.

Classification Table[a]

Observed			Predicted: Selected Cases[b] PHR User Group: Nonuser	Selected Cases[b] PHR User Group: User	Selected Cases[b] Percentage Correct	Unselected Cases[c] PHR User Group: Nonuser	Unselected Cases[c] PHR User Group: User	Unselected Cases[c] Percentage Correct
Step 6	PHR User Group	Nonuser	1927	1417	57.6	1937	1429	57.5
		User	621	1368	68.8	631	1416	69.2
	Overall Percentage				61.8			61.9

a. The cut value is .350

b. Selected cases Approximately 50% of the cases (SAMPLE) EQ 1

c. Unselected cases Approximately 50% of the cases (SAMPLE) NE 1

Figure 33

Case Summaries (ASE) for Training and Testing

squared_error

Approximately 50% of the cases (SAMPLE)	N	Mean
testingdata	5413	.2148
trainingdata	5333	.2116
Total	10746	.2132

Figure 34

The overall model ASEs for the training and testing datasets are very close together, being 0.2148 and 0.2116, respectively. So, there is little over an underfitting after fitting the logit and predicted probabilities on both of the datasets because the ASEs are so close together for the training and testing dataset. Also, the total ASE (0.2132) is consistent and similar to the other models (the decisions trees) total ASE measures. So, this is another good model that predicts the PHR users.

Neural Network Model (SAS Enterprise Miner):

This is the procedure flow chart in SAS Enterprise Miner that was completed, as seen in Figure 35. I first did a replacement step to the PHR data, then did a data partition, then imputed the missing values of the numeric variables income, HBA1C, and Age the means of the respective variables. Lastly, I completed a neural network with the variables age, HBA1C, income, health insurance type, nonsmoker, race, and gender because these variables all came out to be associated with the response varianle PHR at least somewhere in the decision tree models previously shown.

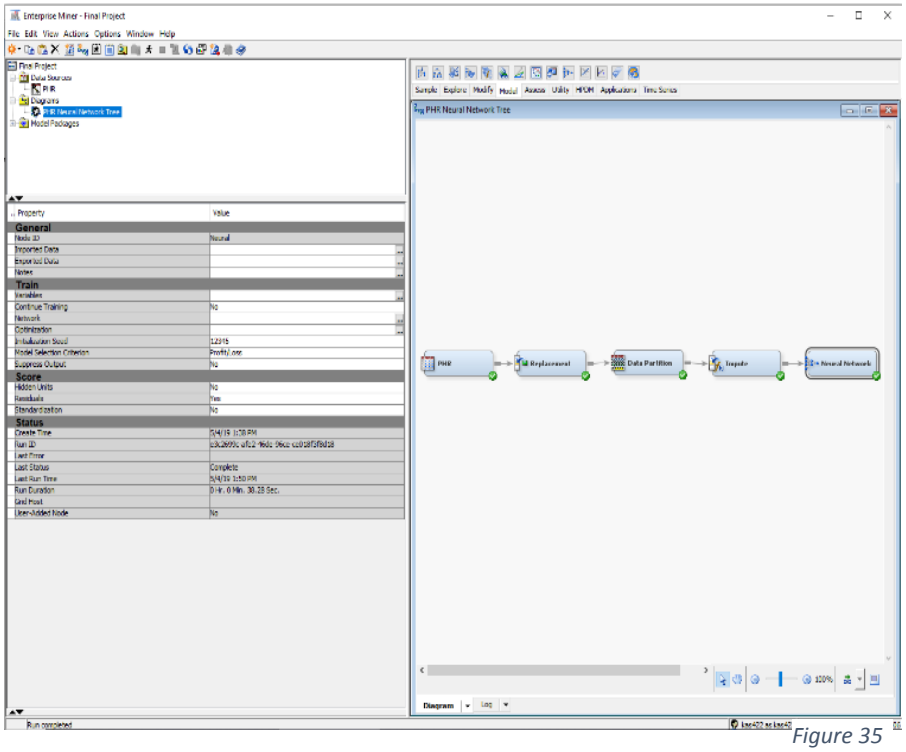


Figure 35

The average squared error and misclassification are similar to the values observed from the other models (see Figure 36), both being about 0.21 (for training and for testing) and about 0.35 (for both training and

testing, respectively). Notice that the model contains 28 weights (_NW_: number of weights).

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
User		_DFT_	Total Degrees of Free...	5372		
User	Target	_DFE_	Degrees of Freedom f...	5344		
User		_DFM_	Model Degrees of Fre...	28		
User		_NW_	Number of Estimated ...	28		
User		_AIC_	Akaike's Information C...	6542.179		
User		_SBC_	Schwarz's Bayesian C...	6726.669		
User		_ASE_	Average Squared Error	0.20943	0.212897	
User		_MAX_	Maximum Absolute Err...	0.940934	0.966628	
User		_DIV_	Divisor for ASE	10744	10748	
User		_NOBS_	Sum of Frequencies	5372	5374	
User		_RASE_	Root Average Squared...	0.457635	0.461408	
User		_SSE_	Sum of Squared Errors	2250.117	2286.22	
User		_SUMW_	Sum of Case Weights ...	10744	10748	
User		_FPE_	Final Prediction Error	0.211625		
User		_MSE_	Mean Squared Error	0.210527	0.212897	
User		_RFPE_	Root Final Prediction ...	0.460027		
User		_RMSE_	Root Mean Squared E...	0.458833	0.461408	
User		_AVERR_	Average Error Function	0.603702	0.611916	
User		_ERR_	Error Function	6486.179	6576.872	
User		_MISC_	Misclassification Rate	0.344937	0.353182	
User		_WRONG_	Number of Wrong Cla...	1853	1898	

Figure 36

The model did not have any trouble converging, according to Figure 37, because there is not an error here and it says that the convergence criterion was satisfied. So, we can continue with this model.

Optimization Results

Iterations	3 Function Calls	7	
Jacobian Calls	5 Active Constraints	0	
Objective Function	0.6030257755 Max Abs Gradient Element		0.0015452322
Lambda	0.001947405 Actual Over Pred Change		0.1166866657
Radius	0.1285215688		

Convergence criterion (FCONV=0.0001) satisfied.

NOTE: At least one element of the gradient is greater than 1e-3.

Figure 37

The initial values for the neural network weights are shown in Figure 38. This model does not show any signs of being overfitted. Notice that all of the variables mentioned above are in some way in this model. According to the literature and from what Dr. Fridline has told me, the neural network procedure is a "black box" procedure. What this means is that this model does not have any parameter interpretations necessary. So. I the analysis of the fit measures will be followed, to see how well this model predicts whether a patient uses the PHR health record system or not.

Optimization Start Parameter Estimates

N	Parameter	Estimate	Gradient Objective Function
1	Age_H11	0.449936	0.000014365
2	IMP_HBA1C_H11	0.626486	0.000070964
3	IMP_Income_H11	0.399520	0.000064969
4	Age_H12	0.106823	-0.001540
5	IMP_HBA1C_H12	-0.073033	0.001385
6	IMP_Income_H12	0.045826	-0.000641
7	Age_H13	0.574236	0.000355
8	IMP_HBA1C_H13	0.188019	-0.001018
9	IMP_Income_H13	-0.171582	0.000696
10	Caucasian0_H11	-0.577257	-0.000067543
11	Female0_H11	0.759319	0.000130
12	Insurance0_H11	-1.548109	0.000063518
13	Nonsmoker0_H11	-1.016937	0.000630
14	Caucasian0_H12	-0.191237	-0.001447
15	Female0_H12	-0.209835	0.005326
16	Insurance0_H12	0.044465	-0.004925
17	Nonsmoker0_H12	-0.255211	0.002206
18	Caucasian0_H13	0.172898	-0.001301
19	Female0_H13	-0.222086	-0.001640
20	Insurance0_H13	0.202668	0.002467
21	Nonsmoker0_H13	-0.177635	-0.003482
22	BIAS_H11	2.126908	-0.000114
23	BIAS_H12	0.441710	0.002541
24	BIAS_H13	-0.567888	0.001641
25	H11_User1	0.615443	-0.001514
26	H12_User1	1.677684	-0.001480
27	H13_User1	-1.367532	-0.001157

Figure 38

The iteration plot in Figure 39 shows the average squared error versus optimization iteration. The training and validation average squared error occurs stay constant and fairly close to each other (notice the scaling on the ASE y-axis. There is not vertical line that displays any divergence. We have a nice number of weights in the fitted neural network model, so the model performance was good.

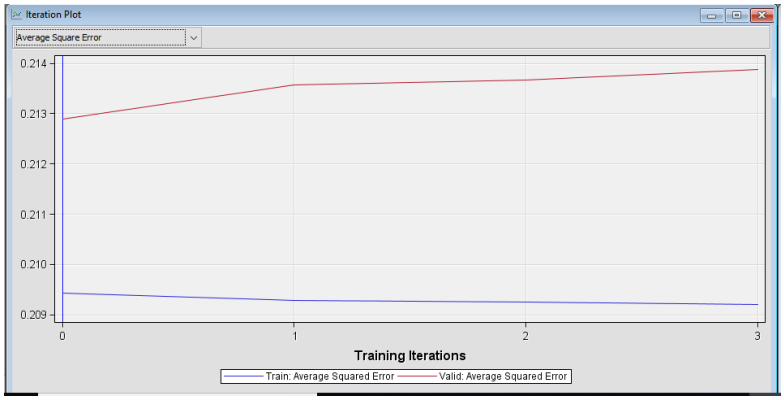


Figure 39

Taking a look at the classification table in Figure 40 for the training and validation datasets, I am assuming that the target is the observed and the outcome is the predicted. It shows the target percentage and the actual outcome percentage. (Notice that there is some error here between these values.) Next, the specificity, which is the percentage of times that the model correctly predicts the nonusers of the PHR system, it's 79.6423% = (2672/(2672+683))*100%, for the training dataset. Now, the sensitivity of the training dataset, which is the percentage of times that the model correctly predicts the users of the PHR system, is 41.99% = (847/(847+1170))*100%. These measures of specificity and sensitivity are different from the CHAID and CART output (in this case, sensitivity is larger in this decision tree than in the other two, which is good), but similar to the decision tree output from SAS. These will be compared in the conclusion. In testing dataset, specificity is (2618/(2618+737))*100% = 78.03% and sensitivity is =39.03% = (858/(858+1161))*100%.The overall correct percentage response rate is for training is = (2672+847)/(2672+1170+683+847) = 0.66,and testing is (2618+858)/(2618+1161+737+858) = 0.65. These are consistent with the other models.

Classification Table

Data Role=TRAIN Target Variable=User Target Label=' '

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	69.5471	79.6423	2672	49.7394
1	0	30.4529	58.0069	1170	21.7796
0	1	44.6405	20.3577	683	12.7141
1	1	55.3595	41.9931	847	15.7669

Data Role=VALIDATE Target Variable=User Target Label=' '

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	69.2776	78.0328	2618	48.7160
1	0	30.7224	57.5037	1161	21.6040
0	1	46.2069	21.9672	737	13.7142
1	1	53.7931	42.4963	858	15.9658

Figure 40

Conclusions:

Out of all of the models, the chosen model was logistic regression. Because, will all of the other measures (risk, ASE, correct classification rate) being about the same amount, as shown in Figure 41, logistic regression has the best combination of sensitivity and specificity, as they are both around 60%, plus or minus a few decimals. This is the best balance of correctly predicting PHR users and nonusers. Note that the cut-off for logistic regression model was 0.35, while the cut-offs for the CHAID, ECHAID, CART, and SAS Decision Trees were set to the default cut-off value of 0.5. In reality, any of these models could be good. In further research, we would look at how the sensitivity and specificity change in the other models, is the cut-off value was set to the approximate response rate of 0.35, as was in the logistic regression model. Out of the set of these models, the logistic regression model is the best one.

Model Type	Training/Testing	Misclassification Rate (A.K.A. Risk)	Correct Classification Rate	Sensitivity	Specificity	ASE
CHAID/ECHAID	Training	0.342	0.658	0.256	0.897	0.2116
	Testing	0.360	0.64	0.248	0.878	0.2184
CART	Training	0.343	0.657	0.235	0.908	0.2175
	Testing	0.364	0.636	0.219	0.889	0.2223
SAS Decision Tree	Training	0.34	0.66	0.3942	0.8235	0.21
	Testing	0.35	0.65	0.3903	0.807	0.22
Logistic Regression	Training	0.382	0.618	0.686	0.576	0.2116
	Testing	0.381	0.619	0.692	0.575	0.2148
Neural Network	Training	0.34	0.66	0.4199	0.7964	0.21
	Testing	0.35	0.65	0.3903	0.7803	0.21

Figure 41

Sources:

- Dr. Fridline's Applied Analytics and Decision Trees Lecture Notes, Spring 2019
- The dataset represents data from the study by Tenforde et al. "The Association Between Personal Health Record Use and Diabetes Quality Measures". J Gen Intern Med 2012; 27: 420-24.
- The SAS Enterprise Miner Manual

Appendix:

SPSS code/syntax for ECHAID procedure.

```
GET
  FILE='E:\Grad School 2018-2019\Spring 2019\Applied Analytics and Decision Trees\Projects\PHR
(3)\PHR (2)\PHR\PHR\phr.sav'.
DATASET NAME DataSet1 WINDOW=FRONT.
* Decision Tree.
TREE User [n] BY Caucasian [n] Insurance [n] Income [s] EyeExam [n] PneumoVaccine [n] ACEARBALB [n]
    FootExam [n] Nonsmoker [n] HBA1CTest [n] HBA1C [s] SBP [s] DBP [s] LDL [s] BMI [s] Age [s]
Female
    [n]
  /TREE DISPLAY=TOPDOWN NODES=STATISTICS BRANCHSTATISTICS=YES NODEDEFS=YES SCALE=AUTO
  /DEPCATEGORIES USEVALUES=[0 1] TARGET=[1]
  /PRINT MODELSUMMARY CLASSIFICATION RISK
  /GAIN CATEGORYTABLE=YES TYPE=[NODE] SORT=DESCENDING CUMULATIVE=NO
  /PLOT GAIN INDEX RESPONSE INCREMENT=10
  /RULES NODES=TERMINAL SYNTAX=INTERNAL TYPE=SCORING
  /SAVE NODEID PREDVAL PREDPROB
```

```
/METHOD TYPE=EXHAUSTIVECHAID
/GROWTHLIMIT MAXDEPTH=AUTO MINPARENTSIZE=100 MINCHILDSIZE=50
/VALIDATION TYPE=SPLITSAMPLE(test) OUTPUT=BOTHSAMPLES
/CHAID ALPHASPLIT=0.05 SPLITMERGED=NO CHISQUARE=PEARSON CONVERGE=0.001 MAXITERATIONS=100
  ADJUST=BONFERRONI INTERVALS=10
/COSTS EQUAL
/MISSING NOMINALMISSING=MISSING.
```