A Machine Learning Based Look at Diabetes

April 14th, 2024

ISYE7406| Project Group 167

Victor Cerabone | vcerabone3@gatech.edu

Student ID Last 4: 4394

ABSTRACT:

The most accurate machine learning model is not always the best one for the job. Good but not great models are useful if the application and related tradeoffs are deeply understood. Minimizing the most severe inaccurate predictions is often key to success. This project explores the tuning and evaluation of classification algorithms designed to predict Type-2 Diabetes from various body measurements and demographic characteristics of individuals. Health assessment data from the CDC is the data source. The complex process of data mining and its challenges are also described. In short, machine learning models can aid in the fight against a prevalent health issue in the United States as a complement standard blood testing.

INTRODUCTION:

A machine learning model that can accurately predict Type-2 diabetes from simple measures would complement or even replace the current practice of blood testing to provide early detection or prevention of the disease. Examples of "simple measures" include a person's age, waist circumference, time spent sedentary, or time spent exercising. Blood testing devices like continuous glucose monitors or lab-based tests are the benchmark for accuracy; however, these methods are inconvenient and, in some cases, cost prohibitive compared to simple measures. This study explores the challenge of developing and validating machine learning models to predict whether an individual's fasting blood sugar level is indicative of diabetes without a blood test.

Why study diabetes? Diabetes is prevalent in the United States with over 10% of Americans having the disease. The rate is even higher for senior citizens for whom the incidence of diabetes is 30% (American Diabetes Association).

Data quality and data import issues were major challenges during this study. Unexpected output from several R libraries caused numerous "false starts" in the datamining flow. Variables that might increase the predictive power of the model were unusable due to a lack of documentation explaining the relationship between tables in the dataset. Finally, the lack of dataset documentation also led to incorrectly joining the data, which caused unrealistic zero error results for certain models. This portion of potentially usable data was discarded.

I focused my problem-solving efforts on the data prep and model evaluation stages using my approach of breaking down seemingly insurmountable issues into a collection of small problems. While I expected numerous iterations in the tuning and evaluation phase, I was surprised by the number of repetitions required to perform data prep work including data import, missing value cleanup, and experimentation with joins.

While this study produced models and insights, the primary value came in the form of personal growth. I increased my understanding of how to prepare data for modeling under real-world conditions. I extended my knowledge of training a "good enough but not perfect" model. I explored the significance of inaccurate predictions and their cost to the end user as part of model evaluation. Overall, I developed a strong confidence that I can execute a machine learning use case outside of the contrived examples with which I have previously worked in an academic setting.

DATA SOURCES:

The data for this study is from the Centers for Disease Control 's National Health and Nutrition Examination Survey [NHANES]. Four datafiles documenting demographics, physical activity, and body measurements were chosen from the vast collection of NHANES files. While there are statistical methods for variable selection available, I relied on existing diabetes research for variable selection (Centers for Disease Control and Prevention).

I performed inner joins on the four datafiles yielding a single dataset of 4438 observations and 67 columns. The data cleanup phase started with the removal of columns that contained more that 10% missing values [see Appendix I]. Next, I removed observations containing missing values bringing the observation count to 3961. Column selections were again refined using research on diabetes factors. Some variables were renamed for clarity. Finally, I created a binary response variable "abnormal_fbs" indicating diabetes. For a diabetic person abnormal_fbs is set to 1 because their fasting blood sugar was ≥ 126 mg/dl. For a non-diabetic individual abnormal_fbs was set to 0 because their fasting blood sugar was < 126 mg/dl. Figure 1 is a dataset example after cleanup.

Figure 1 – Example Data After Cleanup

abnormal_fbs	bmi	waist _circ	age	gender	race	work_act_vig	work_act_mod	min_sedentary
0	25.2	84.0	27	2	3	2	1	600
1	42.9	127.1	36	2	4	1	1	180
0	29.8	115.0	80	1	3	2	1	300
1	30.3	105.7	63	2	2	2	2	360

Figure 2 - Variable Name, Type and Description

Variable Name	Type	Discrete Values	Description
abnormal_fbs	Categorical Binary	2	Response
bmi	Continuous	383	Body mass index [kg/m^2]
waist _circ	Continuous	764	Waist circumference [cm]
age	Continuous	63	Age in years at screening
gender	Categorical Nominal	2	1=male, 2=female
race	Categorical Nominal	6	1=Mexican American, 2=Other Hispanic, 3=Non-Hispanic White, 4=Non-Hispanic Black, 6=Non-Hispanic Asian, 7=Other Race – Including Multi Racial
work_act_vig	Categorical Binary	4	Does your work involve vigorous-intensity activity that causes large increases in breathing or heart rate for at least 10 minutes continuously? 1=Yes, 2=No, 7=Refused, 9=Don't know
work_act_mod	Categorical Binary	3	Does your work involve moderate-intensity activity that causes small increases in breathing or heart rate for at least 10 minutes continuously? 1=Yes, 2=No, 7=Refused, 9=Don't know
min_sedentary	Continuous	47	How much time do usually spend sitting on a typical day [min]?

Figure 2 describes the variables. The variable "waist_circ" was removed from the dataset because it is highly correlated to "bmi", potentially causing model fit issues [See Appendix II].

All predictors except "min_sedentary" are useful for classification given separated median values when comparing the two response classes using statistical significance tests [see Appendix III]. The variable min_sedentary was kept in the dataset even though by itself it does not account for variation in response as it may provide an interaction affect.

Figure 3 shows that presence of diabetes is a relatively rare event [14.6%]. This imbalance will be addressed before training to avoid classification model bias towards the majority class.

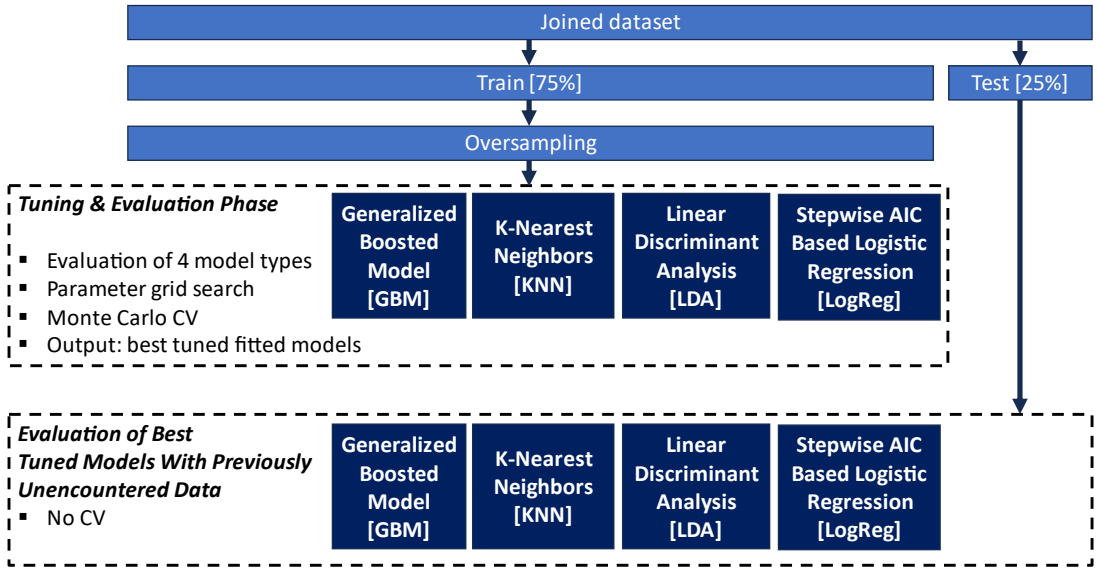
Figure 3 – Response Distribution

	abnormal_fbs = 0	abnormal_fbs = 1	Total
Count	3381	580	3961
Percent of total observations	85.4%	14.6%	100%

DATA MINING METHODOLOGY:

Figure 4 shows the overall methodology. There are 3961 observations [n] to begin this section.

Figure 4



First, I randomly split the data into Train [75%, n=2971] and Test [25%, n=990] sets. Next, I "up sampled" the train data to address the imbalance in the response. The bootstrapping-based up sampling method I used created a more balanced response but does not produce an exact split to avoid overfitting [see Figure 5].

Figure 5 – Response Distribution After Up-Sampling

	abnormal_fbs = 0	abnormal_fbs = 1	Total
Count	2524	2465	4989
Percent of total observations	50.6%	49.4%	100%

I used the Train data to tune parameters for the Generalized Boosted [GBM] and K-Nearest Neighbor [KNN] models and to evaluate them together with the non-tunable Linear Discrimination Analysis [LDA] and Logistic Regression [LogReg] models. I used grid search and Monte Carlo Cross Validation [CV] to tune and evaluate.

I made a final evaluation of models using only the Test data without CV. As the Test data is "new data" that was never touched during the development of the models, this evaluation helps understand whether the models are truly generalized, meaning they will perform well in production.

ANALYSIS AND RESULTS:

The R library carat is used for grid search and cross validation functions. Figure 6 summarizes model tuning and parameters. Figures 7, and 8 show the effect of parameter tuning on model performance.

Figure 6 - Parameter Tuning Details

Model	Best Parameters	Parameter Discussion
GBM gbm Library & Method	Shrinkage = 0.5 interaction.depth = 7 ntree = 1000	The parameter shrinkage is the algorithm learning rate. Small learning rates mean smaller adjustments leading to less over fitting in trained models encountering new data. Small learning rates encourage learning from tree to tree during building, while larger rates give more focus to each individual tree. Large values often mean better trained models but poor performance when encountering new data. Values of 0.01, 0.25,0.5, 0.75 and 1 were evaluated. I was surprised that a moderate value of 0.5 provided the best accuracy. Although a shrinkage rate of 0.01 was favorable in my previous work, this moderate value of 0.5 may raise overfitting concerns. The interaction.depth parameter controls the max number of tree splits. Higher values here allow better fitting in the training stage but risk overfitting with new data. Lower values could "tune out" the complex relationships a tree model usually excels at describing. Values 4,5,6,7, and 8 were evaluated based on guidance [Hastie, T., Tibshirani, R., & Friedman, J. H.]. In theory more trees equate to a greater "ensemble" effect and increased accuracy; however, compute time increases proportionally to the number of trees. I used 500 trees as a starting point and increasing the number of trees in the forest [ntree] had little impact on accuracy. 500 and 10000 trees were considered.
KNN kknn Library & Method	K= 3	The parameter K determines how many neighbor data points are used to consider which class a datapoint should belong to. Even values of K generally should be avoided to avoided ties in voting. The values 3,5,7,9, and 11 were considered.
LDA mass Library, lda method	NA	No tuning
LogReg stats Library, glm Method	NA	No tuning

Figure 7 - Boosted Model Tuning Results

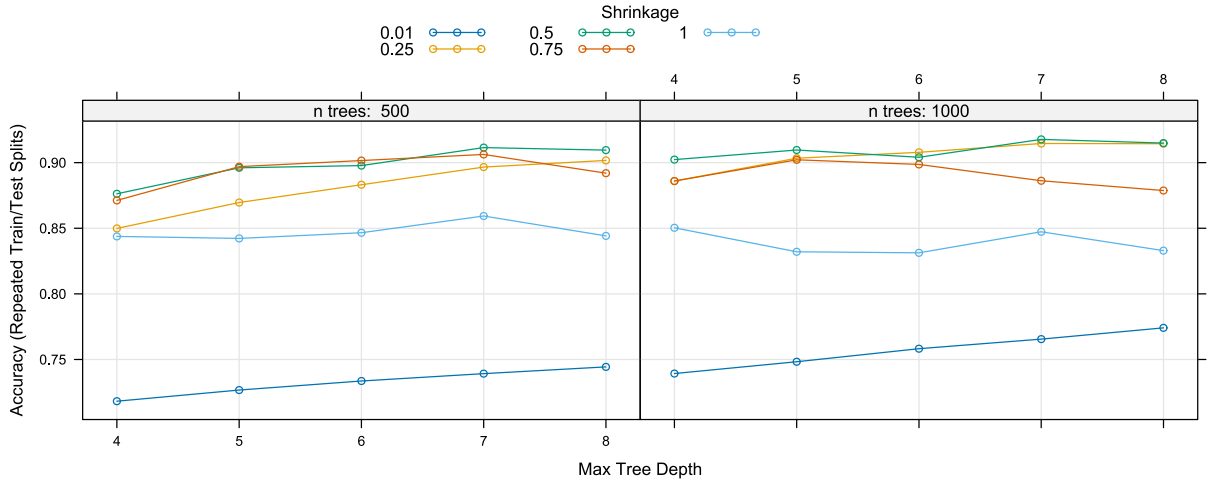


Figure 8 - KNN Model Tuning Results

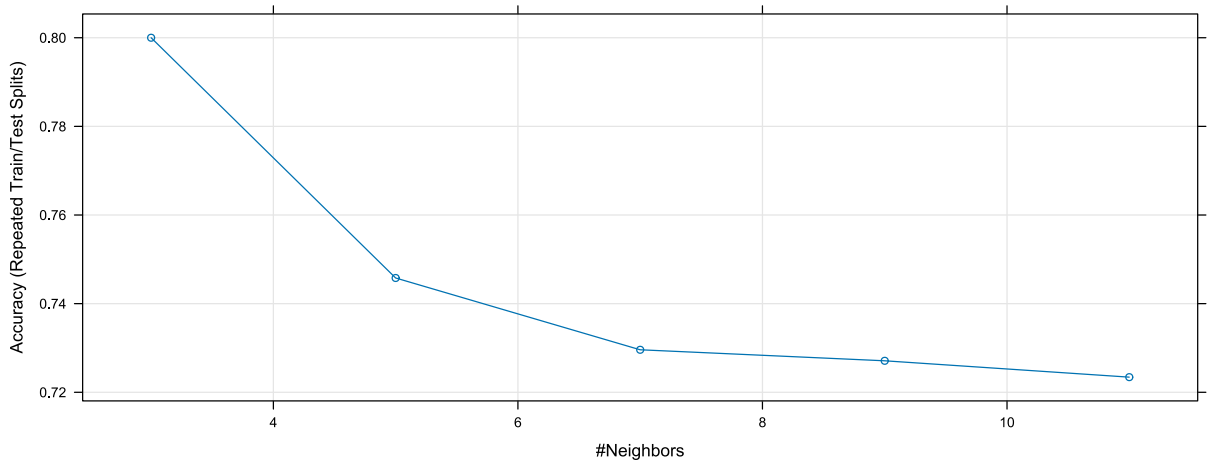
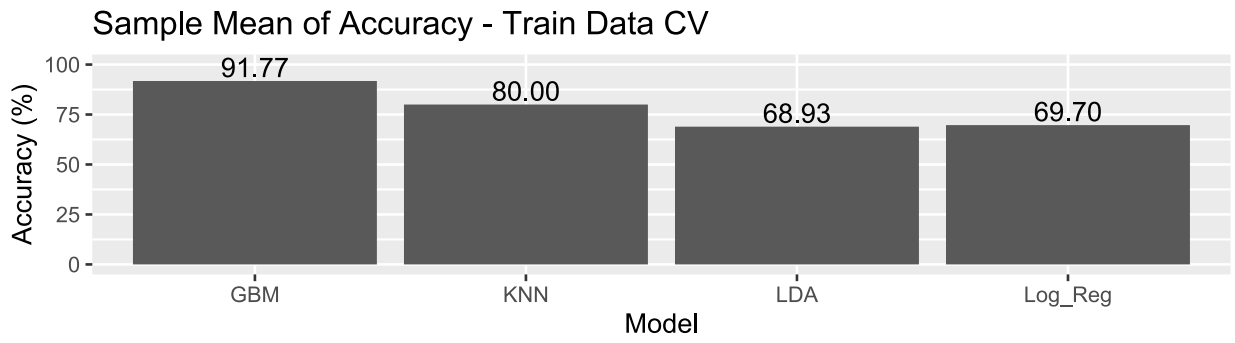


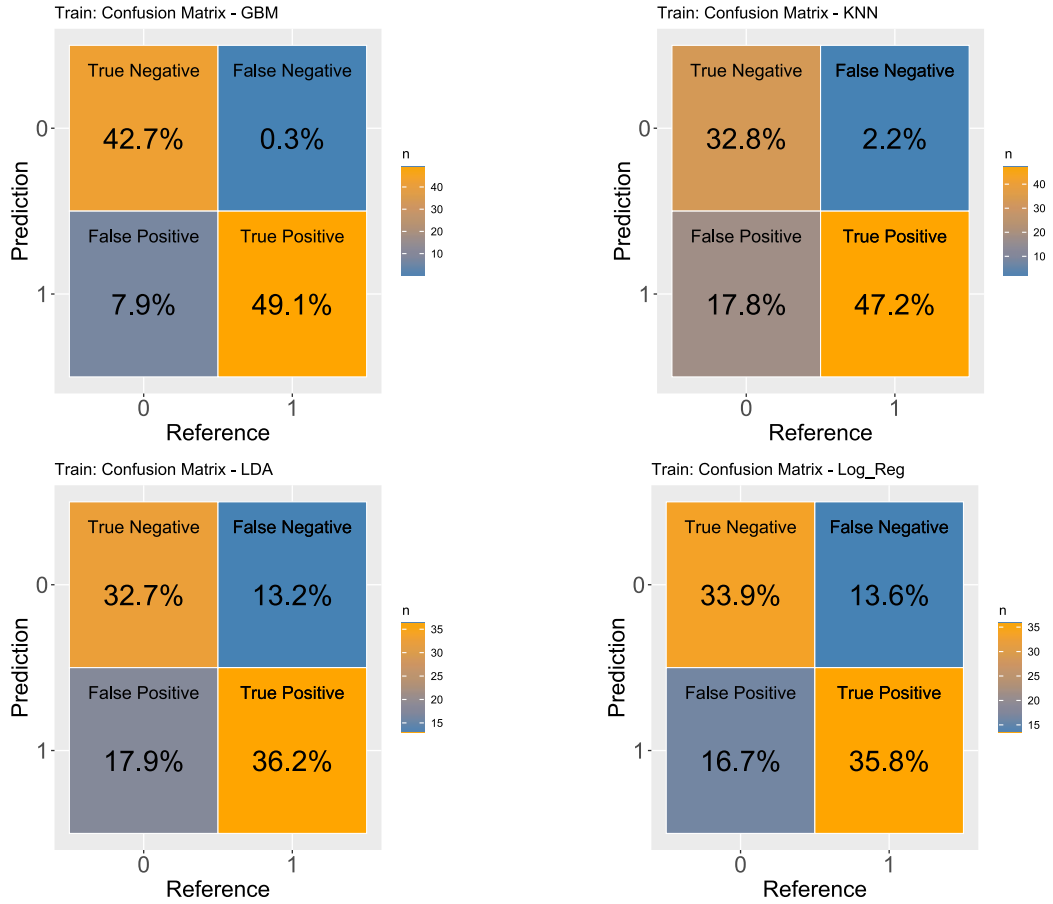
Figure 9 shows that the GBM model has the highest accuracy in training compared to its peers. The KNN model has second-highest accuracy. Statistical significance testing confirms both the accuracy ranks for GBM and KNN; however, there is not enough statistical evidence to determine the rank of LDA relative to LogReg [see Appendix IV].

Figure 9 – Train Data Accuracy



Accuracy is a useful aggregate for overall ranking; however, the confusion matrices shown in Figure 10 are a better basis for understanding classification behavior.

Figure 10 – Train Data Confusion Matrices

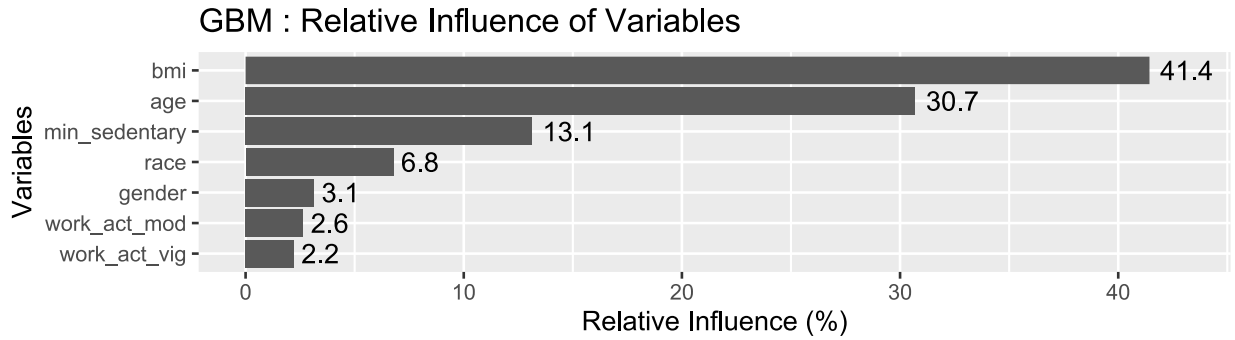


Most important when considering the application of classification models is an understanding of the "cost" of false negatives [FN] and false positives [FP]. In practice, the cost of incorrect predictions is not necessarily monetary. The concept of cost is only used to weight the two types of incorrect predictions. The most severe incorrect prediction is an FN in which a diabetic individual is predicted not to have the disease. A FN individual will not get the medical care they need because of model error. GBM has the lowest rate of FNs at 0.3%.

FPs represent non-diabetic individuals predicted to have the disease. The FP cost includes inconvenience, medical expenses, and anxiety related to imagining they have a disorder they do not. While FPs are undesirable, the cost of these erroneous results is not nearly as severe to individuals as FNs. GBM also has the lowest rate for FPs at 7.9%.

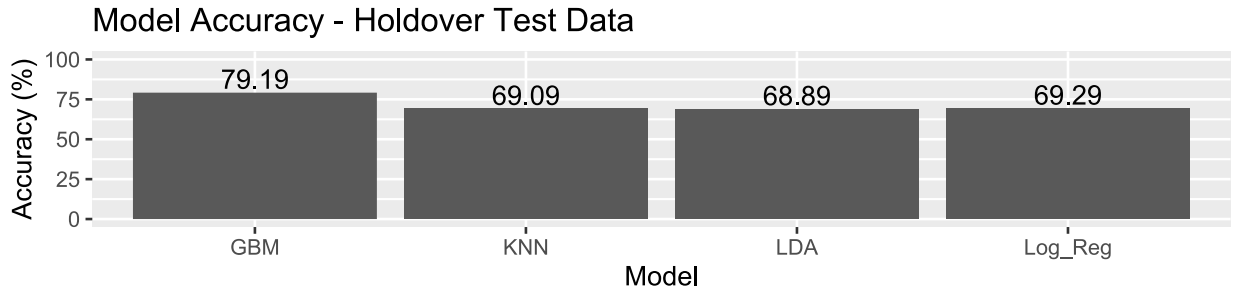
One nice feature of this GBM algorithm its ranking of variable significance. Figure 11 shows that BMI, age, and time sedentary are the top three most influential predictors of diabetes.

Figure 11 – GBM Variable Influence



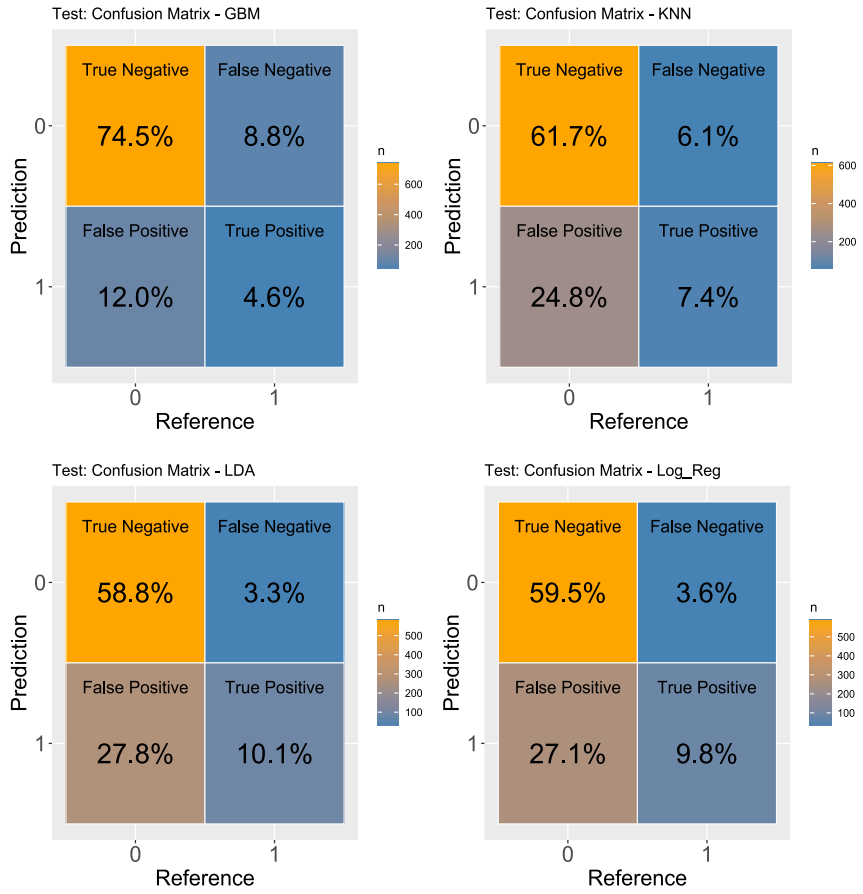
Moving on to the Holdover Test data performance as shown in Figure 12, the GBM and KNN models again rank first and second respectively; however, a deeper look at classification behavior is required to understand whether accuracy should decide which model is best for the application.

Figure 12 – Test Data Accuracy



The confusion matrices in Figure 13 provide that deeper look. GBM does have the highest accuracy and an excellent balance between FN [8.8%] and FP [12.0%], however the goal is to minimize FNs given their severe repercussions. The LDA or LogReg models have the lowest FN rates with 3.3% and 3.6% respectively. I was surprised by the dramatic change in the results compared to the training phase. Now, I understand that the GBM and KNN models were over fit during training given this "reversal" in classification behavior. Further evidence of over fitting for GBM and KNN is the reduced differentiation from peers on the accuracy metric that is observed in the test phase.

Figure 13 – Test Confusion Matrices



CONCLUSIONS:

The LDA and Logistic Regression models could be deployed as a compliment but not a replacement for blood testing. These two models seem reasonably generalized with a good balance between bias and variance; however, more evaluation is required on additional, new, unencountered data. Potential for improvement of all models exists in finding additional predictors in NHANES, implementation of imputation for missing data, and resolution of the join issues.

LESSONS LEARNED:

A growth mindset believes there is value in learning from mistakes. I had my share of failure and growth in this study.

The XPT formatted data from NHANES was initially simple to import into an R data frame using the haven or SASxport libraries. Unfortunately, these tools produced R data unusable for boxplots or modeling. These two libraries produced data frames where each column contained a list object rather than a vector object. Once I implemented the foreign library (R Core Team) columns did contain vectors and methods that expect a vector worked as anticipated.

This was the largest data cleanup task I have performed to date and the "multidimensionality" of the empty values was especially challenging. My standard approach has been to remove all rows with empty values. In this case, however, that approach would have removed all the data! I developed a new approach of first removing columns with empty value counts greater than 10% and then deleted empty rows afterward. While this method removed 26 columns and potential predictors, only 11% of the observations were removed which is close to my 10% target.

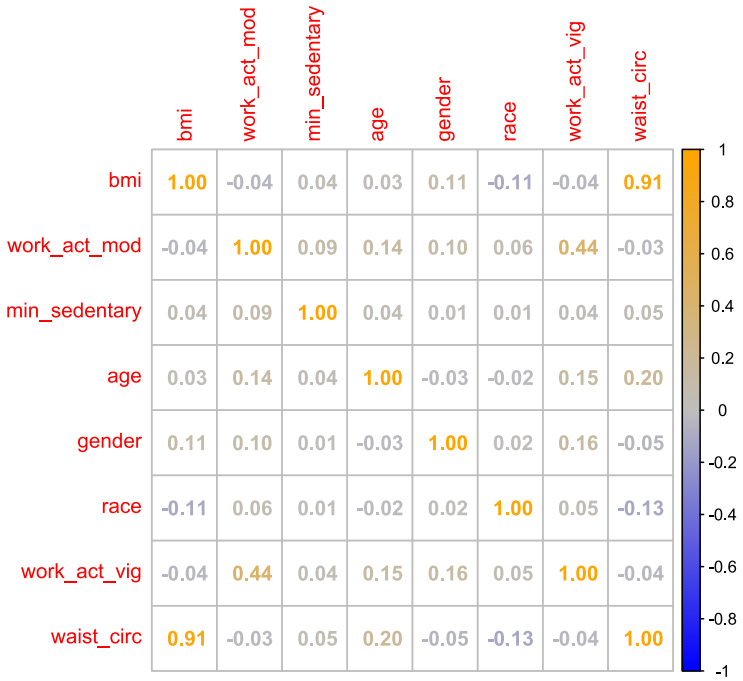
The biggest lesson came from training models, understanding how they fit unbiased data and interpreting their usability for classification. Early in my machine learning studies I heard that model accuracy does not have to be perfect for usefulness; however, I never understood that concept in practice until now. The progressions through the tasks associated with this project will prove to be a major benefit to me professionally.

APPENDIX:

I. **Missing Data:** Where missing data per column is above 444 [10% of observations] column deletion, rather than imputation or other missing value method, was used. Note that column names are from the original data.

SDDSRVYR	RIDSTATR	RIAGENDR	RIDAGEYR	RIDAGEMN	RIDRETH1	RIDRETH3	RIDEXMON
0	0	0	0	4438	0	0	0
DMDBORN4	DMDYRUSZ	DMDEDUC2	DMDMARTZ	RIDEXPRG	SIALANG	SIAPROXY	SIAINTRP
0	3129	186	186	3568	0	0	0
FIALANG	FIAPROXY	FIAINTRP	MIALANG	MIAPROXY	MIAINTRP	AIALANGA	WTINTPRP
271	271	271	261	261	261	1014	0
WTMECPRP	SDMVPSU	SDMVSTRA	INDFMPIR	LBXGLU	PAQ605	PAQ610	PAD615
0	0	0	630	265	0	3324	3332
PAQ620	PAQ625	PAD630	PAQ635	PAQ640	PAD645	PAQ650	PAQ655
0	2527	2537	0	3395	3402	0	3350
PAD660	PAQ665	PAQ670	PAD675	PAD680	BMDSTATS	BMXWT	BMIWT
3351	0	2637	2642	8	0	78	4294
BMXRECUM	BMIRECUM	BMXHEAD	BMIHEAD	BMXHT	BMIHT	BMXBMI	BMDBMIC
4438	4438	4438	4438	83	4369	90	4260
BMDBMIC	BMXLEG	BMILEG	BMXARML	BMIARML	BMXARMC	BMIARMC	BMXWAIST
4260	261	4254	192	4324	191	4324	237
BMIWAIST	BMXHIP	BMIHIP					
4278	234	4283					

II. **Correlation Matrix:** The variable waist_cir was removed from the dataset as its correlation coefficient to bmi was above 0.9.



III. Statistically significant difference of predictor medians is found at 95% confidence.

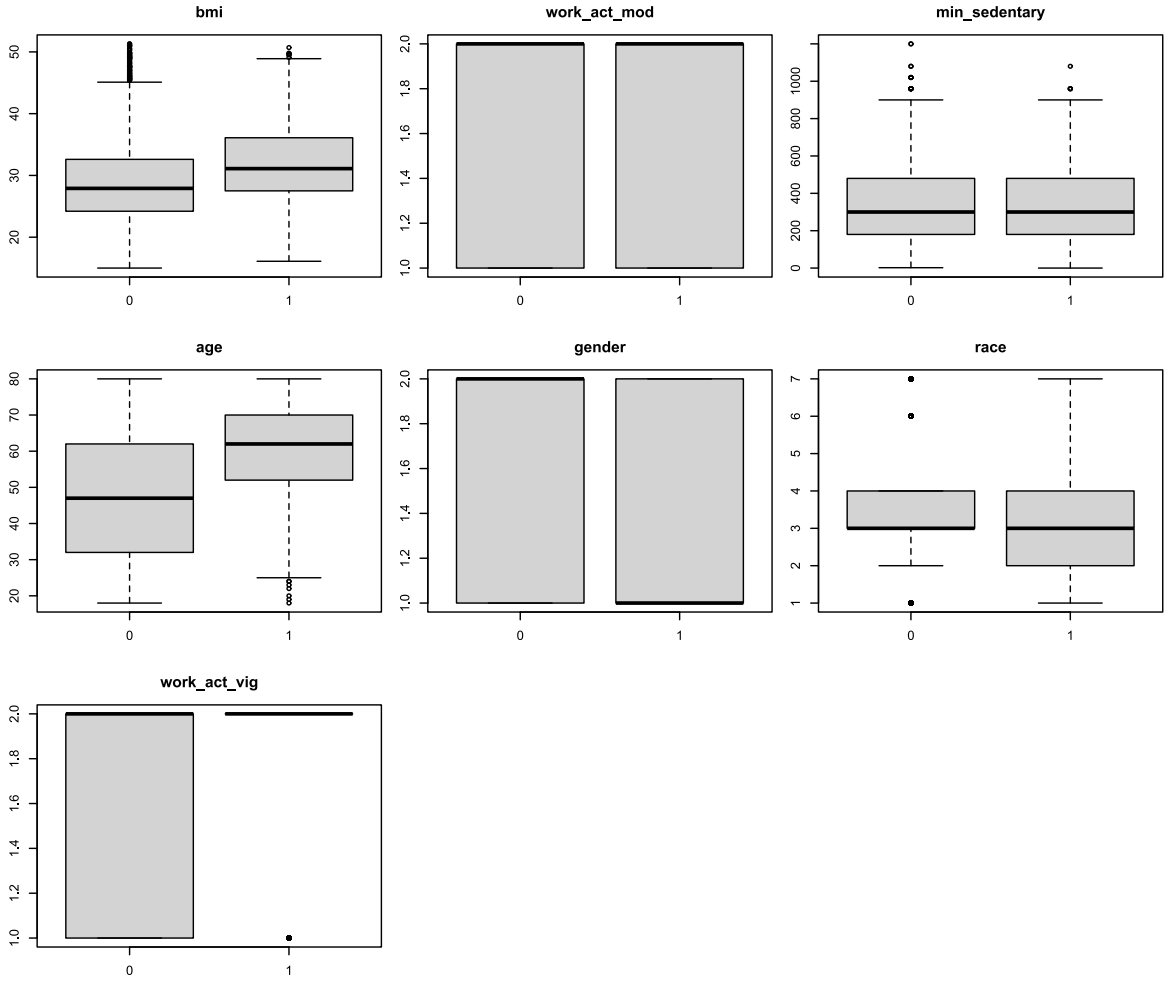
Rationale: p-values ≤ 0.05 for the Shapiro test [normality] coupled with p-value ≤ 0.05 for the Kruskal-Wallis test [difference in distributions and medians]. No outlier removal was performed on the training or testing data; however, outliers beyond ± 3 σ were removed from the boxplot for the sake of visual clarity.

	bmi	work_act_mod	min_sedentary	age	gender	race	work_act_vig
Shapiro p-values	0.00	0.00	0.00	0.00	0.00	0.00	0.00

H_0: the data is normally distributed. H_A: evidence exists that the data is not normally distributed.

	bmi	work_act_mod	min_sedentary	age	gender	race	work_act_vig
Kruskal Wallis p-values	0.00	0.00	0.52	0.00	0.04	0.000	0.00

H_0: the compared groups come from the same distribution and have the same median.
H_A: evidence exists that the groups have different distributions & have different medians.



IV. Statistically significant difference of error between GBM and peers and KNN peers is given. Statistically significant difference of error KDA and LogReg is not given.

Rationale: p-values ≤ 0.05 on either the T or Wilcox test indicate evidence that the Sample Means of Testing Error are different at a 95% confidence.

GBM vs peers	KNN	LDA	LogReg
T-Test	0.000	0.000	0.000
Wilcox Test	0.002	0.002	0.002
KNN vs peers	LogReg	LDA	Best KNN
T-Test	0.000	0.000	0.000
Wilcox Test	0.002	0.002	0.002
LDA vs peers	GBM	KNN	LogReg
T-Test	0.000	0.000	0.99
Wilcox Test	0.002	0.002	0.160

H0: the compared Sample Means of Testing error are the same.

HA: there is evidence that the compared Sample Means of Testing error are different.

BIBLIOGRAPHY AND CREDITS:

I References

American Diabetes Association. "About Diabetes - Statistics." Diabetes.org, https://diabetes.org/about-diabetes/statistics/about-diabetes#:~:text=Prevalence%3A%20In%202021%2C%2038.4%20million,of%20the%20population%2C%20had%20diabetes.

Centers for Disease Control and Prevention. [n.d.]. Diabetes: Risk factors for type 2 diabetes. Retrieved from https://www.cdc.gov/diabetes/basics/risk-factors.html

R Core Team. [n.d.]. foreign: Read Data Stored by Minitab, S, SAS, SPSS, Stata, Systat, Weka, dBase, ... [R package]. Retrieved from https://cran.r-project.org/web/packages/foreign/index.html

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer.4

II Data Source:

https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?Cycle=2017-2020