Predictive Models for Early Detection of Chronic Kidney Disease

Akhilesh Yadav Gaddam

ABSTRACT

Chronic Kidney Disease (CKD) is a major public health problem related to serious sequelae, including end-stage kidney failure, cardiovascular disease, premature death, and increased health care expenditures. Recent statistics show that 30 million American adults have CKD and a millions of others are at increased risks. Early detection can help prevent the progression of kidney disease to kidney failure. In this project, an attempt has been made to predict CKD on the basis of patient's records using logistic regression models. A screening tool is then created which can be used by healthcare practitioners to help with early diagnosis of CKD patients.

1. INTRODUCTION

Chronic kidney disease (CKD) is a condition characterized by a gradual loss of kidney function over time. The two main causes of chronic kidney disease are diabetes and high blood pressure, which are responsible for up to two-thirds of the cases. You may also have an increased risk for kidney disease if you are older or have a family history of kidney failure or belong to a population group that has a high rate of diabetes or high blood pressure, such as African Americans, Hispanic Americans, Asian, Pacific Islanders, and American Indians. The earlier kidney disease is detected, the better the chance of slowing or stopping its progression. An estimated 75% of the seven million Americans with moderate-to-severe chronic kidney disease are undiagnosed. Improved prediction models to identify high-risk subgroups for chronic kidney disease enhance the ability of health care providers to prevent or delay serious sequelae, including kidney failure, cardiovascular disease, and premature death. It is important that prediction modeling be done using appropriate statistical methods to achieve the highest accuracy, while avoiding overfitting and poor calibration. Logistic regression has been used to create our predictive model and R and SAS softwares to run the data. Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Here, 1 stands for the patient with CKD and 0 for the patient without CKD. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

2. LOGISTIC REGRESSION MODEL

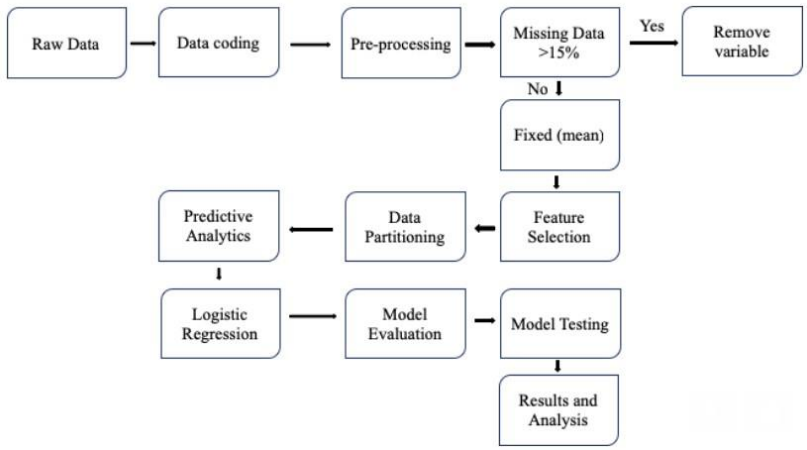
2.1 Exploring the data

The data set includes 8819 data points and 33 variables. Two dummy variables have been created for race group (Black, Hispanic/Others).

2.2 Cleaning the data

The raw data is first analyzed and then preprocessed to identify outliers and missing data. To deal with the missing data, mean imputation has been used so that the overall mean will not be affected. If more than 15% of the values in a row were missing, the row has been entirely removed. The data is then standardized.

2.3 Framework



There was a total of 6000 records. The data was partitioned into two – training data and testing data using SRS. It was split in such a way that the training data included 50% of CKD affected patients and the testing data set included the remaining 50% of patients with CKD. Training data set includes 504 data points (232 with CKD and 272 without CKD). Testing data set includes 5496 data points (232 with CKD and 5264 without CKD).

2.4 Stepwise Selection

The actual set of predictor variables used in the final regression model must be determined by analysis of the data. Determining this subset is called the variable selection problem. Since variable selection is needed, stepwise selection is performed. Initially, with stepwise selection six significant variables were observed. Keep the real life scenario in mind, three more variables have been included. The nine variables observed are age, diabetes, race group, hypertension, ckd, activity, CHF, PVD and CVD.

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-6.1698	0.8004	59.4156	<.0001
Activity	1	-0.1756	0.1795	0.9568	0.3280
Age	1	0.0811	0.00957	71.7783	<.0001
CHF	1	-0.0946	0.5680	0.0277	0.8678
CVD	1	0.8653	0.4173	4.2982	0.0382
Diabetes	1	1.0238	0.3360	9.2840	0.0023
Female	1	0.3353	0.2475	1.8355	0.1755
Hypertension	1	0.5540	0.2672	4.2977	0.0382
Insured	1	0.5395	0.4571	1.3928	0.2379
PVD	1	0.5615	0.4508	1.5513	0.2129

Change in Variable	Increase in Log Odds of Having CKD
10 units in Age	2.25
Unit change in CVD	2.376
Unit Change in Diabetes	2.784
Unit Change in Activity	0.839
Unit Change in CHF	0.91
If Female	1.398
Unit Change in Hypertension	1.74
If Insured	1.715
Unit Change in PCD	1.753

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	266.5108	9	<.0001
Score	216.6438	9	<.0001
Wald	136.8144	9	<.0001

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	697.514	449.004
SC	701.737	491.229
-2 Log L	695.514	429.004

The likelihood ratio chi-square of 266.5108 with a p-value of less than 0.001 tells us that our model as a whole fits significantly better than an empty model. The Score and Wald tests are asymptotically equivalent tests of the same hypothesis tested by the likelihood ratio test. These tests also indicate that the model is statistically significant.

The Akaike Information Criterion (AIC) is an estimator of the relative quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Thus, AIC provides a means for model selection. After running different models, ones with lower AIC were preferred.

2.5 Confusion Matrix

A confusion matrix is a table that is often used to describe the performance of a classification model a set of test data for which the true values are known. It allows the visualization of the performance of an algorithm i.e. it allows easy identification of confusion between classes e.g. one class is commonly mislabeled as the other. Most performance measures are computed from the confusion matrix.

Abbreviation	Name	Description
TP	True Positives	Number of correctly predicted cases of CKD
TN	True Negatives	Number of cases where it was correctly predicted that the patient does not have CKD
FP	False Positives	Number of incorrectly predicted cases of CKD
FN	False Negatives	Number of cases where it was incorrectly predicted that the patient does not have CKD

The confusion matrix is shown below. Positive classification is when the person has CKD and negative classification is when the person does not have CKD.

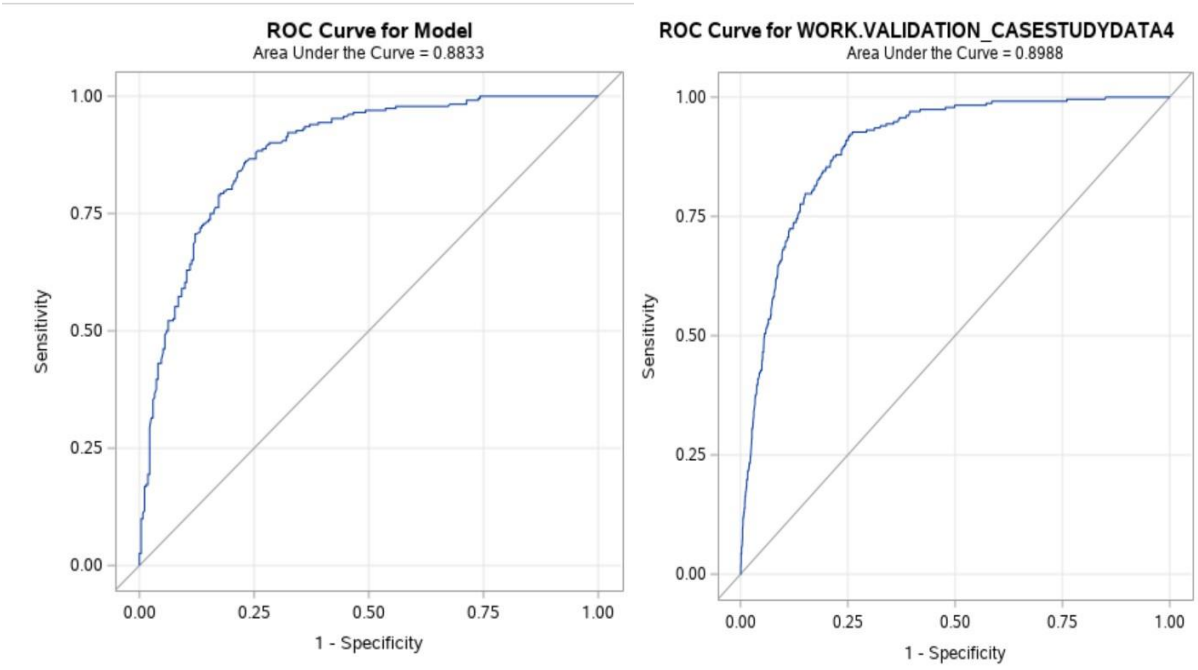
Outcome	Predicted Negative	Predicted Positive
Actual Negative	4157	1107
Actual Positive	31	201

2.5 Accuracy, Sensitivity and Specificity

Three performance metrics are used to evaluate the analytics models: accuracy, sensitivity, and specificity. Definitions of the three metrics along with their descriptions are shown in the table below.

Metric	Description	Equation
Accuracy	Measures the ability of the model to correctly predict the class label of new and unseen data	(TP + TN) /(TP+TN+FP+FN)
Sensitivity	Measures the proportion of positives (or Yes's) that are correctly identified as such	TP/(TP+FN)
Specificity	Measures the proportion of negatives (or No's) that are correctly identified as such	TN/(TN+FP)

AUC - ROC curve is a performance measurement for classification problem at various thresholds settings. ROC is a probability curve and AUC represents degree or measure of separability. It tells how much model is capable of distinguishing between classes. Higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s. By analogy, higher the AUC, better the model is at distinguishing between patients with CKD and without CKD. The ROC curve is plotted with TPR against the FPR where TPR is on y-axis and FPR is on the x-axis. After implementing our model on validation data set, the ROC curve on the left shows that the prediction rate of our model is 89.88%



More models were created after increasing the significant variables which delivered better prediction rate but were not chosen due to overfitting.

3. SCREENING TOOL

A screening tool has been created after interpreting the results of our logistic regression model. It is a checklist or questionnaire which can be used by healthcare professionals in predicting the presence of CKD in patients. The table below depicts the weightage of each significant variable.

Variable	Weightage
Age	0 if it is less than 40 20 if it is more than 40 and less than 60 40 if it is more than 60
Gender	0 if it is male 5 if it is female
Insurance	0 if not insured 5 if insured
Physical activity	Range: 0-5 0 being highly physically active 5 being very less physically active
Presence of CVD	0 if not present 10 if present
Presence of PVD	0 if not present 5 if present
Presence of Diabetes	0 if not present 20 if present
Presence of Hypertension	0 if not present 15 if present
Presence of CHF	0 if not present 15 if present

4. CONCLUSION

Our model estimated the probability of chronic kidney disease using the β coefficients with 9 recognized and suspected risk factors for chronic kidney disease. Subgroups were identified ranging from virtually no probability to very high probability. Any individual's risk can be estimated as the probability that individuals with the same 9 specific characteristics have chronic kidney disease using the formula:

$p_i = \frac{1}{1+exp_{exp}(\beta(^{0}\beta+{}_{0}+{}^{\beta 1}\beta^{X}{}_{1}X^{i})_{i})}$. Further research is needed to simultaneously assess the role of multiple risk factors and to validate this model in other populations.