CHRONIC KIDNEY DISEASE

Objective

The objective of the case study is to **develop a method for identifying individuals at risk of having Chronic Kidney Disease** (*hereinafter referred to as "CKD")* and also to develop an **easy to use screening Tool** with simple Health Questionnaires to identify patients at risk of CKD and to help patients decide whether to be tested for CKD.

Background

CKD is a worldwide Public health problem with an increasing incidence, prevalence and high cost. Approximately 37 million American Adults have CKD and millions of others at increased risk. Early detection can help prevent or delay the progression of kidney disease to kidney failure[1]. With the decreasing awareness of CKD among patients, there is a need of accurate, convenient and easy to use screening tool which identifies the patient risk of having CKD

Executive Summary

This report provides insights as to which features are most important and likely indicators of CKD and how these features are jointly used to determine CKD cases for the 2819 patients. We have combined all the possible risk factors indicating CKD to measure the overall risk faced by study subjects. We have used two approaches to predict CKD one using a **classification model** and other using a **screening tool** which identifies the risk patients on a scale of 1 to 3.

The limitation of the data is that it is not a random sample of US adults therefore cannot be applied to US population in actual decision making.

Logistic regression model is used to predict the CKD cases where only a list of important variables is used to predict CKD. A screening tool was designed as well to identify high risk patients. In both methods variables like Age, Diabetes, Hypertension, Race group category like Hispanic, Cardiovascular disease, Peripheral Vascular Disease, Congestive Heart failure and Anaemia emerged as the most important variables leading to CKD. So, we determined **high-risk patients based on the prevalence of most of these symptoms/variables in a patient.**

This report also gives **recommendations** based on the screening tool answered by patients, we can classify them as either a no risk, moderate risk or a high-risk patient for CKD on a scale of 1 to 3.

Exploratory Data Analysis:

In our analysis, we develop a method for identifying individuals at risk of having CKD by developing **classification model using Logistic regression**. The dataset we have contains 8819 observations and 33 variables with some variables playing an important role in causing CKD.

Data cleaning and Missing value treatment:

- The dataset contains several numerical and categorical variables providing various information on patients medical information. Also, the data provided has many missing values and there were 2777 missing values for one or more features. Logistic regression cannot be performed in a data with missing value in it. And, removing the observation with missing value will also totally reduce the accuracy.

- Histogram of count of missing values against each variable show that income variable is more missing ***(Exhibit 1).*** The one practical reason behind the Income variable having more missing values is because **people might be little reluctant in revealing their income status whether it is above the median or not**. The distribution of missing values suggested that the data is not MCAR (Missing Completely at Random). As the below distribution is not uniform, mean-substitution will introduce bias into the data. So, performing Multiple imputation using MICE package

- In order to verify whether the imputed values are indeed credible values, density plot were plotted which indicate that this is the case (magenta represents imputed data and blue represent the actual non missing data distributions ***in Exhibit 2)***.

Exhibit 1

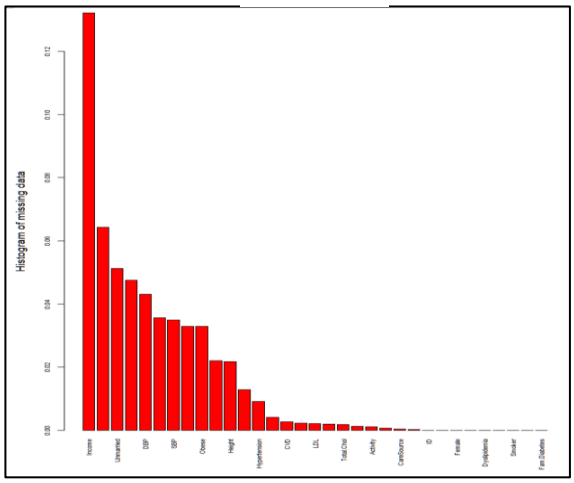
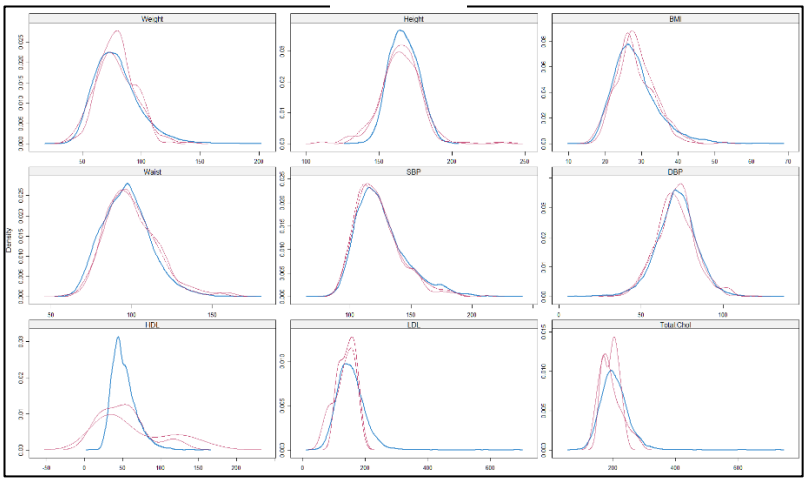


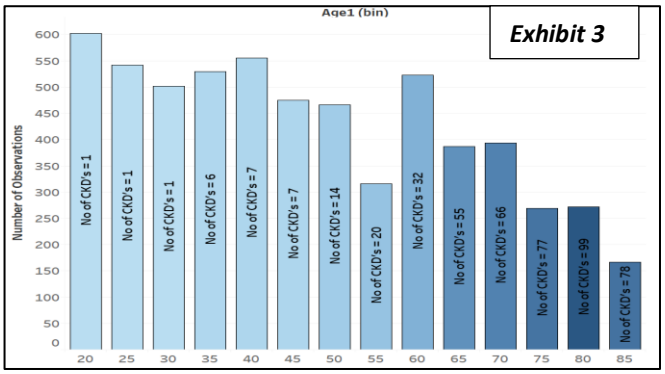
Exhibit 2



Using Correlation to explain for important variables

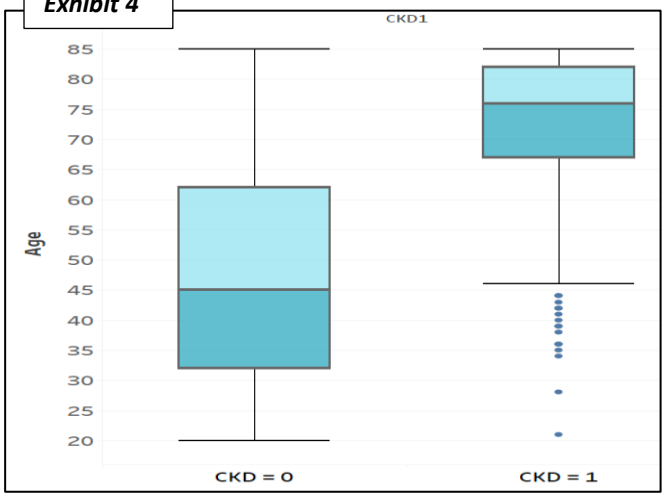
- **AGE**: Distribution of age is right skewed, As the age increases, there is a high risk to have CKD. It can be seen from the below graph ***(Exhibit 3 & 4)*** that people **above 60 years** of age are prone to have CKD even though the number of observations is very less. **Positive high correlation (0.37)** with CKD.

Exhibit 3



- **Hypertension:** If a person is having any stage 1 – 4 of High Blood pressure, they are more prone to have CKD. It is the second highest corelated variable with **Positive correlation (0.23)** with CKD.

Exhibit 4

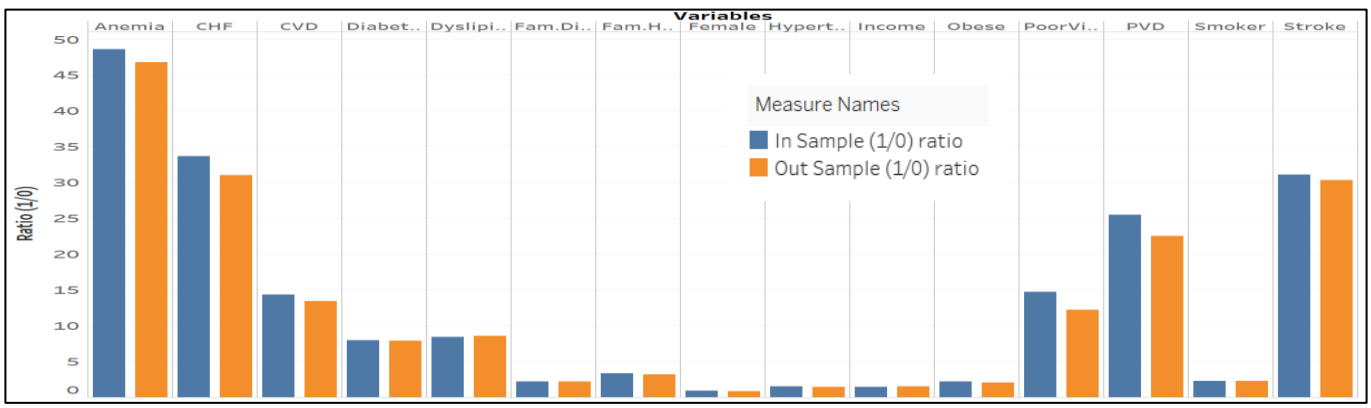


With the above correlation method, we cannot conclude for an importance of a variable. There are many other variables which are highly correlated with one or more other independent variables leading to Multicollinearity. By correlation matrix, we can conclude the set of variables which are multicollinear to each other and using one variable in each pair of correlated variables will avoid the model being undermined by statistical significance of independent variable.

Analysing in sample and out sample data:

In-sample data of 6000 observations and out-sample data of 2819 observations were analyzed to find whether in-sample and out-sample have any similarities and deviations ***(Exhibit 5)***. It was found that the distribution of 1/0 ratio for every parameter in focus appears to be similar for both In-sample and Out-Sample Data.

Exhibit 5



Logistic Regression model

The following steps were performed to develop a Logistic regression Model:

a) Variable selection
b) Logistic regression model construction and validation
c) Out-sample data prediction

Variable Selection:

Variable selection is mainly performed to select "**best" subset of predictors**. The main goal is to perform a better prediction using logistic regression and explain the relationships in the data. Starting it with Stepwise procedures which is combination of both backward eliminations and forward selection, to find the most important variables will addresses the situation where variables are added or removed early in the process.

Further we performed, the **variance inflation factor (VIF)**. Each predictor in our model will have a VIF value. If the VIF is equal to 1 there is no multicollinearity among factors, but if the VIF is greater than 1, the predictors may be moderately correlated. A rule of thumb commonly used in practice is **if a VIF is > 10, it means high multicollinearity**. In our case, with values around 1, we are in good shape, and can proceed with our regression.

And those variables are selected based on results obtained from Stepwise, Correlation matrix and VIF.

Logistic regression model construction and validation:

The main objective of developing logistic regression model is to predict the likelihood of getting CKD for out sample data (2819 observations). After building a model, we need to determine the accuracy of this model on predicting the outcome for out sample data which is not used to build the model. In other words, we need to estimate the prediction error. So, the basic step is to

1. Build the model on a training data set
2. Apply the model on in sample data with newly created test set.
3. Compute the prediction error and accuracy.

- First step is to divide the in-sample data into **Training (65%) and Testing (35%) sets** with both having equal proportion of 0's and 1's for the CKD. One important information to look at while partitioning the dataset, we are building a model on a fraction of the data set only, possibly leaving out some interesting information about data, **leading to higher bias.** To overcome that **k-fold cross-validation method** is used to evaluate the model performance on different subset of the training data and then calculate the average prediction error rate.
- Four Logistic regression models were built on the training data and validated against the test data.
- Results of all models were shown to differentiate the results and accuracy by removing certain variables.
  a) **Model 1**: All Variables
  b) **Model 2 (selected through above variable selection):** Age, Racegrp, HDL, LDL, SBP, DBP, PVD, Activity, Fam.Hyper, Diabetes, CVD, CHF, Anemia
  c) **Model 3 (selected through stepwise elimination)**: Age, Female, Unmarried, Weight, BMI, Obese, Waist, HDL, LDL, Dyslipidemia, PVD, Activity, Smoker, Hypertension, Diabetes, CVD, Anemia, CHF.
  d) Model 4 contains the variables from Model 2 but some variables like SBP & DBP were combined to include as hypertension and CVD was replaced by stroke as the latter had a higher correlation with CKD.
  Model 4: Age, Racegrp, HDL, LDL, PVD, Activity, Hypertension, Fam.Hyper, Diabetes, Stroke, CHF, Anemia.

There are many critical metrics which is used to measure the performance of the model. Among the four models, model 1 & 3 were not considered due to obvious multicollinearity issues. Model 2 & 4 were selected based on the following three factors:

AIC Value:

It is a relative measure of model fit and is one of the measures used to compare different models built on the same dataset. **Lower the AIC value, better is the model.** From the below AIC values, we can say that Model 4 is a better model, but since AIC is the not the only criteria to decide, let's proceed to further steps to decide.

	AIC	AUC	Accuracy	Specificity	Sensitivity
Model 2	1471.3	89.38%	92.47%	0.154	0.9892
Model 4	1463.5	89.68%	92.52%	0.154	0.9897

Confusion Matrix:

It is the one way of describing the **performance of a classification model**. It gives insight not only into the errors being made by your classifier but more importantly the types of errors that are being made. Though the true positives for both the models is same, the false positives are less in Model 4. Also, as we want to diagnose a severe disease, we prefer **higher sensitivity over high specificity.**

Model 2 Confusion Matrix

Outcome	Predicted: 0	Predicted: 1
Actual: 0	1916	137
Actual: 1	21	25

Model 4 Confusion Matrix

Outcome	Predicted: 0	Predicted: 1
Actual: 0	1917	137
Actual: 1	20	25

AUC - ROC curve:

It is a performance measurement for classification problem at various thresholds settings. **ROC is a probability curve and AUC represent degree or measure of separability**. Higher the AUC, better the model is at predicting 0's as 0's and 1's as 1's i.e. the model is at identifying the patients with and without CKD.

Though the ROC curve for both the models look similar, the area under ROC curve for Model 4 is a little higher compared to Model 2.

Exhibit 6

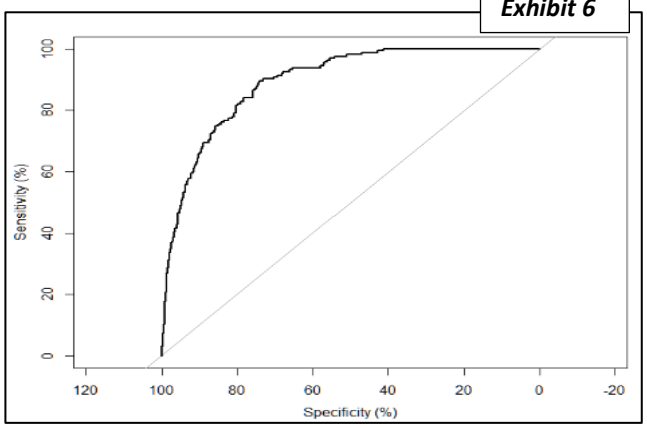
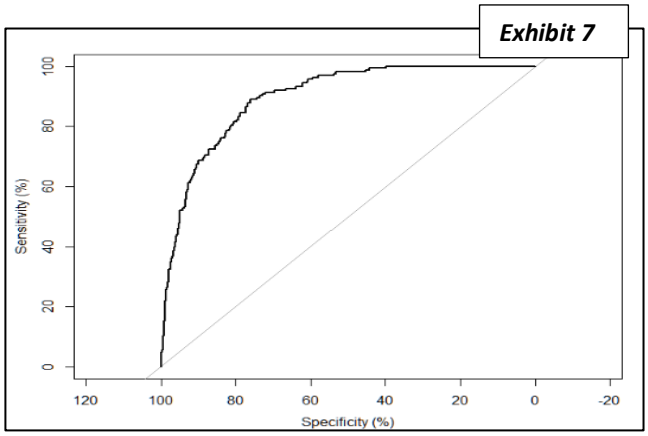


Exhibit 7



Interpreting ROC curves:

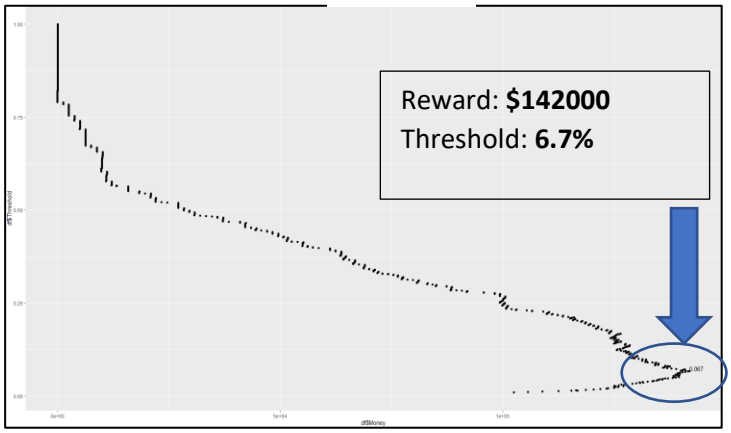
Model has AUC near to the 1 which means it has good measure of separability. A poor model has AUC near to the 0 which means it has worst measure of separability. And when AUC is 0.5, it means model has no class separation capacity. In our results, **AUC of 89.68% indicates the probability is 89% that a randomly chose person with CKD will be ranked higher than a randomly chosen person without CKD.**

From the above three factors, we can conclude that **Model 4** is the best among all the four models mentioned above.

Threshold value by maximizing Reward:

The next step after concluding the best model is to decide the threshold value by developing a function with the **objective of maximizing the Reward** (1300 * TP – 100 *FP) which will perform a grid search on all the possible threshold values and plot **Reward vs Probability Threshold** for models 2 & 4 ***(Exhibit 8)***. From the grid search function, model 4 had better maximum **reward $142000 at cut-off 6.7%.** This further proves that Model 4 is the best among those two models.

Exhibit 8



Out Sample Data Prediction:

With the above findings, Model 4 is used to predict the results for the out-sample data set (2819 patients) with a cut-off probability value of 6.7%.

Limitations:

1. The interpretations are limited to the set of variables provided in the dataset and there might be other factors which may influence the risk of CKD.
2. Model is developed taking the threshold to maximize the reward function which might not be best parameter.
3. Variable selected might not be the best combination and **more subject matter research is needed** in it.

4. Race group white is **overrepresented in our dataset** and our model trains accordingly and identifies the Hispanic to be important feature although **black and Hispanic have equally likely chances of having CKD according to research findings.**

5. The model is built newly on the training dataset and accuracy is decided based on ROC and AIC values. However, as we have additive data in future, we will have a better predictive model that will be trained on additive dataset.

Why Logistic Regression?

Logistic Regression is used because the output is very **informative** when compared to other classification models like Decision trees and Random forest. It not only gives measure of how relevant a predictor is based on coefficient size but also **tells about the direction of association**. (positive or negative). So, in our case study it is more useful as it tells that variables like Age, Diabetes, CHF and Hypertension are very relevant and they influence CKD positively.

Why not other Models?

Decision tree or random forest cannot be used for predicting CKD cases because it is inconsistent and gives **different results with additional modified data**. It cannot be trusted due to inaccuracy as **learning does not improve** on modifying the dataset. Also, they only take the variable importance into account without any information about the direction of association with CKD. But it is useful for screening tool as we want to classify the patients on a risk scale of CKD i.e. checking the risk level of patient.

Conclusion

Findings:

Age is the most important variable with the highest p-value and correlation with CKD. It is a positive indicator of CKD. Although anyone at any age can develop CKD, it is mostly seen from the data, that people greater than 65 years are more likely to get CKD.

Another finding is that most of the **heart related diseases** like CHF, PVD, CVD and Stroke were found to be positive influential factors indicating CKD increasing the risk level of CKD when **combined with higher aged people** and it is also explained by odd ratio insights and interpretation part.

Diabetes and Hypertension contributed majorly in CKD cases. Both these variables have the highest correlation with CKD after age and have very significant p-values. We also get the same variables for screening tool which **validates our findings.**

Insights:

Estimate, we get for the model summary are **logit(probability)** which is **log (Odds).**

	Estimate	Std. Error
Activity	-0.241	0.084
Hypertension	0.623	0.135
CHF	0.596	0.198
PVD	0.493	0.174
Stroke	0.401	0.194

1. Estimate of Hypertension (0.62) can be interpreted by keeping other variables constant, Odds of CKD patient with Hypertension over odds of CKD patient without Hypertension is Exp (0.623) = 1.85. In terms of percent change, we can say **odds of Hypertension patient having CKD are 85% higher than odds of non-hypertension patient having CKD.**
2. The **odds of CHF patient having CKD is 81% higher than odds of non-CHF patient having CKD** [Exp (0.596) = 1.81]. Similarly, we can say for PVD and stroke (63% and 49% respectively) as well, which leads to the conclusion that **heart related diseases** are increasing the risk level of CKD.
3. Estimate of activity (-0.241) is nothing but difference in log (odds). It can be interpreted as a **odds of person who is Standing or walking a lot having CKD is 21.5% lower than odds of person who mostly sits** [Exp (-0.241) = 0.785]. For every increase in activity, odds of getting CKD is 21.5% reduced.

Recommendation:

Based on our screening tool design we have identified the **3 category buckets** to identify patients as no risk, moderate risk and high risk of developing CKD. We can provide below recommendation to people falling on the respective risk scale.

Category Buckets	Risk Level
1	No Risk of having CKD
2	Moderate Risk of having CKD
3	High Risk of having CKD

Also, we would recommend and encourage people to be **proactive as it helps reduce stress, manage weight, and maintain blood pressure and blood glucose goals.** Maintaining these will lower the chances of the person developing CKD.