# 1.0 SUMMARY

In this data mining report, we will be going to consider the Framingham dataset for our analysis to find out the factors contributing to the Cardio Vascular Disease (CVD). Using SAS Enterprise Miner we will conduct this analysis and try to compare different methods to mine the data. Also, the comparison between different Neural networks, Regression, Decision trees, and other research nodes, will be the key aspects of this report. The dataset consists of different parameters which might or not contribute to the disease and we will explicitly dig down to those factors.

Any disorder that affects the heart or blood vessels is referred to as cardiovascular disease (CVD). It is usually linked to the accumulation of fatty deposits in the arteries and an increased risk of blood clots. It's also been connected to damage to arteries in organs including the brain, heart, kidneys, and eyes. These are the reasons, why this exercise is important for society and mankind. This dataset has a lot of potentials and can be used for further research also.

# 2.0 BUSINESS PROBLEM

In the whole world, Cardiovascular Diseases (CVD) are rising every day and this is becoming one of the major issues for children to old aged people. This disease is affecting the whole spectrum of ages. This is the leading cause of death in the US. One person in the US dies every 36 seconds and in every 4 deaths, one is caused by CVDs (Prevention, Accessed March 12, 2020.). Between 2016 and 2017, the United States spent $363 billion on heart disease. This covers the costs of medical services, medications, and lost productivity as a result of mortality (Virani SS, 2021).

After all the research and improvements in the medical field, the rate of CVD is gradually increasing at a consistent rate. Our business problem is to find out relevant factors associated with the CVDs using different data mining models.

# 3.0 METHODOLOGY- DATA MINING APPROACH

In this report, we will be emphasizing the SEMMA model which was introduced by the SAS institute. SEMMA stands for Sample, Explore, Modify, Model, and Assess, and it refers to the steps in a data mining project. The SAS Institute views the process as a five-stage cycle:

1. Sample 🡪 This stage is the initial step and a small portion of the large dataset (sample) is being separated which contains significant information but is small enough to manipulate quickly. This step is optional and can be carried out depending on the dataset size.
2. Explore 🡪 This step consists of different data exploration to find out any significant anomalies or trends in the data.
3. Modify 🡪 This step modifies or transforms the data for better consumption in the models created later.
4. Model 🡪 This stage involves modeling the data by instructing the program to look for a combination of data that consistently predicts a desired outcome.
5. Asses 🡪 This step involves examining the data and judging how well it performs by evaluating and comparing the utility and reliability of the data mining discoveries.

Despite the fact that the SEMMA procedure is independent of the DM tool of choice, it is tied to the SAS Enterprise Miner software and purports to guide the user through the development of DM applications. SEMMA is a simple method that allows for the orderly and proper creation and management of DM projects. It, therefore, provides a framework for his conception, creation, and progress, assisting in the presentation of business solutions as well as the identification of de DM's business objectives.

# 4.0 DATA EXPLORATION

The nodes Graph Explore, Multiplot, and Stat Explorer is used to exploring data. The status variable is our goal variable. We're more interested in the number of people who have died so that we may analyze the ones who are still living and don't have CVD. We can observe the systolic (3%) and diastolic(3%) are leading the worth table using Stat Explorer. Observations, measurement levels, and data mining responsibilities are summarized.

|  |  |
| --- | --- |
|  |  |

Table 1 Variable worth

# 5.0 DATA PARTITION CREATION OF MODel SETS

|  |  |
| --- | --- |
|  | In this practice, we have partitioned the data using a data partition node and further used it for the train(40%), test(30%), and validation(30%). A random seed value has been inputted using the last 5 digits of the PNumber(81987). |

Table 2 Data splitting and setting a random seed(81987)

# 6.0 DATA MODELLING

Regression, Decision Trees, and Neural Networks are used to model the data. To increase the model's performance, we employed nodes like Gradient Boosting and Ensemble.

## 6.1 REGRESSION

Our target variable is binary which only can produce output as dead or alive. The normal linear regression model cannot work on the binary variables and because of that, we utilized the logistic regression model here.

## 6.2 MODIFICATION NODES

Data modification has been done using the nodes like, replacement, impute and filter. But the main modification in this scenario is imputed which replaced the nominal variable data with count and interval variables with mean values.

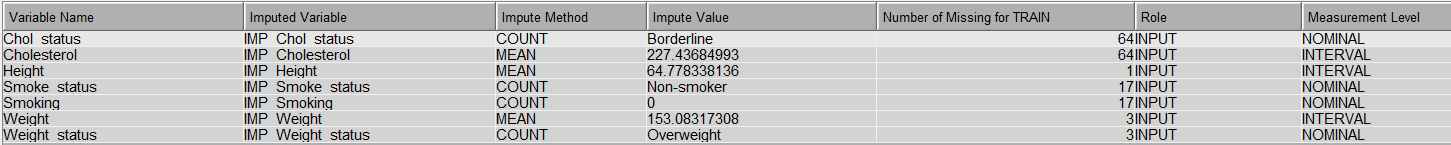


Table 3 Impute node summary

Chart, bar chart

Description automatically generatedTransformation node is used for data transformation. It has been observed that variable Systolic has high positive skewness of 1.48. To decrease the skewness and make the distribution more normal we have used log transformation on Systolic.

## 6.3 SELECTION METHODS

Three selection methods are used to find out the best and most appropriate regression equation. We have used Forward selection, Backward elimination, Stepwise selection method. As discussed above we have transformed Systolic and hence we are performing regression after transforming the variable.

## 6.4 MODEL PERFORMANCE

In search of the best regression model. We have used Regression directly from data portioning node [Regression (Missing)], we have used imputation and used regression [Regression (imputed)], we have then used filter and used regression [Regression (Filtered)], and finally we have used transformation and used regression [Regression (Transform)]. After comparison of all models, we finally come to a point where we can say Regression via transformation node is more statistically significant.

In search of the best model, we have then used selection methods. However, after analysis on the regression models we get to know stepwise, forward, and backward models are performing exactly same. All has same misclassification rate and have equal statistical significance.

We can consider any one model out of these three we have considered Stepwise selection method for our further analysis. The amount of evidence to say that this is a null model is incredibly week p value is less than 0.05 and chi-squared value is massive i.e.., 316. For more information see the image\_1 in **Appendix-1.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Non-Cu. % Res. Scope** | **Cu. Lift @ 10% depth** | **True -** | **False -** | **True +** | **False +** |
| **Regression T STEP** | **38** | **1.98** | **84.12** | **15.87** | **41.20** | **58.79** |
| Regression TB | 38 | 1.98 | 84.14 | 15.87 | 41.20 | 58.79 |
| Regression TF | 38 | 1.98 | 84.14 | 15.87 | 41.20 | 58.79 |
| Regression (MISSING) | 40 | 1.98 | 86.20 | 13.79 | 36.51 | 63.48 |
| Regression (IMPUTED) | 44 | 1.96 | 85.06 | 14.93 | 39.02 | 60.97 |
| Regression (TRANSFORM) | 42 | 1.93 | 83.71 | 16.28 | 39.36 | 60.63 |
| Regression (FILTERED) | 44 | 1.96 | 85.06 | 14.93 | 39.02 | 60.97 |

Note: All the values in the tables are of the validation set.

The cumulative lift to stepwise regression is around 2.15 on the validation set. If we talk about non-cumulative lift the point at which the model performance dips below 1 that is if we pass 38% of data, we can say our model performance is worse than random. We are getting 38% scope on regression stepwise. To get the best model we have used Gradient Boosting on Regression T STEP and we see observe that the model performance dips majorly. For details, please check **Appendix 1.**

Considering non-cumulative lift, after we pass 28% data our model performs worst then random. Cumulative lift is around 1.90 on the validation set at 10% depth. Observing the Target variable if we see at DEAD, we have 26.29 % True Positives and 73.70% False Positive. Clearly the model performance dips.

## 6.5 CHOSEN REGRESSION EQUATION

Below is the equation for regression stepwise. This is the chosen regression equation with the intercept of -17.8954.

Logit P = -17.8954 + IMP\_Cholestrol \* (0.00569) + IMP\_Height \* (-0.0789) + IMP\_Weight\_status Normal \* (-0.3383) + IMP\_Weight\_status Overweight \* (-0.3845) + Log\_Systolic \* (10.1038) + Sex Female \* (-0.5611)

# 7.0 NEURAL NETWORKS

A neural network is a collection of algorithms that attempt to detect underlying patterns in a sample of data using a technique that resembles the human brain. We have used various type of Neural Networks (NN) in our model.

## 7.1 DEVELOPMENT OF MODELS

In search of best NN model we tried all the possible ways by which we can enhance our model. We first compared the basic models. We get to know that NN impute is the best model with cumulative lift of 2.05 at 5% depth on the validation set, with 40.36 True positive values. If we talk about non-cumulative lift the point at which the model performance dips below 1 that is if we pass 42% of data, we can say our model performance is worse than random. We are getting 42% scope on Imputed Neural Network (NN IMPUTE).

Further to dig even more deeper we have used combination of {Principal Component Analysis (PCA), Variable Selection Methods (VS), Variable Selection Methods with CHI- Squared R, Decision Trees} with impute node, transform node, and default data portioning node to identify the best Neural Network Model out of all these models shown in above figure 2. We have used number of hidden units as 3 in all the Neural Networks except in NN 1 NODE.

Note: We have created (Imputed DUMMY) node by keeping the Missing cutoff of 50.0 setting the indicator variable type to unique and changing the role to input

## 7.2 MODEL PERFORMANCE

Below table describes model performance of all the models. Note the table below describes the results on the validation set.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Non-Cu. % Res Scope** | **Cu. Lift**  **@10% depth** | **True -** | **False -** | **True +** | **False +** |
| NN IMPUTE | 42 | 1.86 | 81.84 | 18.15 | 40.36 | 59.36 |
| NN MISSING | 30 | 1.93 | 86.41 | 13.58 | 37.01 | 62.98 |
| NN 1 NODE | 42 | 1.93 | 84.75 | 15.24 | 39.53 | 60.46 |
| NN TR VS CHI-R | 34 | 1.91 | 87.75 | 12.24 | 36.34 | 64.65 |
| NN IMP VS CHI-R | 42 | 1.86 | 86.51 | 13.48 | 37.68 | 62.31 |
| NN PCA VS | 48 | 1.83 | 84.54 | 15.45 | 36.85 | 63.14 |
| NN PCA | 48 | 1.83 | 83.19 | 16.80 | 37.85 | 62.14 |
| **NN VS DEFAULT** | **39** | **1.98** | **82.78** | **17.21** | **42.71** | **57.28** |
| NN DT | 29 | 1.91 | 84.95 | 15.04 | 39.53 | 60.46 |
| NN VS DUMMY | 24 | 1.91 | 82.46 | 17.53 | 39.19 | 60.80 |
| DM NEURAL | 39 | 1.88 | 86.30 | 13.69 | 37.68 | 62.31 |

Note: All the values in the tables are of the validation set.

The best model we get is from Neural Network using Variable Selection in default settings (NN VS DEFAULT). Note that default settings refers that the Variable Selection node is connected to impute node and impute and Variable selection node both are on default settings.

Using NN VS DEFAULT we have the cumulative lift of 2.11 at 5% depth with 42.71 True Positive values on the validation set. If we talk about non-cumulative lift the point at which the model performance dips below 1 that is if we pass 39% of data, we can say our model performance is worse than random.

## 7.3 NEURAL NETWORK ARCHITECTURE OF BEST MODEL

Chart

Description automatically generatedFigure 3 depicts Neural Network Architecture of our Best Model that is NN VS DEFAULT.

Table

Description automatically generated with medium confidenceBelow figure 4 is about the table of weights of our best model. Darker the color of more relevance the variable holds in the model.

Figure 4

# 8.0 DECISION TREES

A decision tree is a class discriminator that separates the training data set recursively until each partition is wholly or primarily composed of samples from one class.

## 8.1 DEVELOPMENT OF MODELS

We have used DEFAULT, PRUNED, THREE WAY and MANUAL Decision Trees. We have compared all the Trees in search of best model.

As discussed above in 7.1 (DEVELOPMENT OF MODELS) we have used imputed decision tree with neural network in search of best model.

Decision trees are never imputed as it has negative effect on the output. However, if we are using Neural Network after decision tree we can use impute as Neural Network throws away the model containing missing data.

Here dummy variables are created for more information read note in 7.1

## 8.2 PERFORMANCE OF MODELS

Comparing all the four models we get to know that Decision Tree manual is giving the best performance.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Non-Cu. % Res. Scope** | **Cu. Lift**  **@10% depth** | **True -** | **False -** | **True +** | **False +** |
| Decision Tree(DEFAULT) | 22 | 1.88 | 92.94 | 7.05 | 26.46 | 75.53 |
| Tree(PRUNED) | 22 | 1.88 | 92.94 | 7.05 | 26.46 | 75.53 |
| Decision Tree THREE WAY | 43 | 1.64 | 90.04 | 9.95 | 27.47 | 72.52 |
| **Decision Tree MANAUAL** | **24** | **1.85** | **86.61** | **13.38** | **33.33** | **66.66** |

Note: All the values in the tables are of the validation set.

If we talk about non-cumulative lift the point at which the model performance dips below 1 that is if we pass 24% of data, we can say our model performance is worse than random. With True Positive value of 33.33, cumulative lift of 2.04 at 5% depth on the validation set. We are considering Decision Tree Manual for our further analysis as it has the best performance.

## 8.3 CRITICAL PATH

Diagram

Description automatically generatedThe path which is outlined with thick black bold line is the critical path. Shown in figure 5.

We can clearly notice that number of people who are Alive and have Systolic levels less then 151 or missing are 61.76 % on the validation set. The path also suggests the number of people who have Systolic levels less than 151 or missing and are Alive is approximately 764. However, this is not our business problem. Our business problem is about identifying the target path.

## 8.4 TARGET PATH OF INTEREST

Nodes 43, 45 and 46 are in our target path.

Nodes 46 and 45 are of highest significance with 89 % and 70 % respectively with prediction of status= Dead. Everything in pink is for identifying critical path.

A picture containing table

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated with medium confidence

## 8.5 OVERFITTING AND LIMITATIONS

Graphical user interface, chart, application

Description automatically generatedThere are limitations while using Decision Trees. Sometimes the model gets overfitted resulting in bad performance of the model. In our analysis, Default, Pruned, and Three way decision trees are over fitted. However, training and validation performs descent in Manual Decision Tree.

# 9.0 RESEARCH NODES

To find the best model, we have used 9 research nodes in our analysis. RESEARCH 1 - MBR, RESEARCH 2 - LARS, RESEARCH 3 - PLS, RESEARCH 4 - Rule, RESEARCH 5 - Boost, RESEARCH 6 - HPSVM, RESEARCH 7 - HPBNC, RESEARCH 8 - HPDMForest, RESEARCH 9 – HPNNA.

## 9.1 THEORY

(Please write this little)

## 9.2 SETTINGS

Default setting on all the Research nodes seems to work fine for us. Hence, we have used the default settings in all the nodes.

## 9.3 ANALYSIS

RESEARCH node 3, node1 and node 7 are performing the best If we talk in terms of getting more true positive values. However, the best performance is given by node 7 and is more statistically significant.

## 9.4 PERFORMANCE

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Non-Cu. % Res. Scope** | **Cu. Lift**  **@10% depth** | **True -** | **False -** | **True +** | **False +** |
| RESEARCH 1 | 29 | 1.77 | 81.74 | 18.25 | 36.85 | 63.14 |
| RESEARCH 2 | 44 | 1.93 | 88.27 | 11.72 | 35.34 | 64.65 |
| RESEARCH 3 | 38 | 1.89 | 87.03 | 12.96 | 38.02 | 61.97 |
| RESEARCH 4 | 56 | 1.83 | 92.84 | 7.15 | 26.63 | 73.36 |
| RESEARCH 5 | 19 | 1.83 | 92.73 | 7.26 | 27.63 | 72.36 |
| RESEARCH 6 | 29 | 1.89 | 92.21 | 7.78 | 26.80 | 73.19 |
| **RESEARCH 7** | **44** | **1.83** | **74.37** | **25.62** | **49.24** | **50.75** |
| RESEARCH 8 | 38 | 1.83 | 89.21 | 10.78 | 29.48 | 70.51 |
| RESEARCH 9 | 29 | 1.89 | 87.55 | 12.44 | 35.51 | 64.48 |

Note: All the values in the tables are of the validation set.

RESEARCH 7 – HP BN Classifier node gives the highest performance 49.24 True Positive values it’s a massive improvement as compared from where we started. If we talk about the scope of the model, it is 44% which is not at all bad.

## 9.5 ANALYSIS OF THE BEST MODEL

Table summary results of the best performing models.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Non-Cu. % Res. Scope** | **Cu. Lift**  **@10% depth** | **True -** | **False -** | **True +** | **False +** |
| DT MANUAL | 24 | 1.85 | 86.61 | 13.38 | 33..33 | 66.66 |
| REGRESSION T STEP | 38 | 1.98 | 84.12 | 15.87 | 41.20 | 58.79 |
| NN VS DEFAULT | 39 | 1.98 | 82.78 | 17.21 | 42.71 | 52.28 |
| RESEARCH NODE 7 | 44 | 1.83 | 74.37 | 25.62 | 49.24 | 50.75 |
| **ENSEMBLE** | **42** | **1.78** | **69.19** | **30.80** | **57.11** | **42.88** |

Chart

Description automatically generatedCumulative lift chart of the best performing models

Chart, line chart

Description automatically generatedNon- cumulative lift chart of the best performing models