A Machine Learning Approach to Optimizing Diabetes Healthcare Management Using SAS Analytic Suite

Taiwo Ajani
tajani@ferrum.edu
Computer Technology and Information Systems
Ferrum College
Ferrum, VA 24088, USA

George S. Habek
ghabek@ncsu.edu
Business Analytics, Poole College of Management
North Carolina State University
Raleigh, NC 27695, USA

Abstract

A typical healthcare business challenge was explored to demonstrate the effectiveness of data mining and machine learning techniques on large-scale medical and pharmacy claims data for about 70,000 patients newly diagnosed with type II diabetes. The business challenge was to move uncontrolled diabetic patient (H1AC > 7) to a controlled state (H1AC < 7). Two algorithms were explored for this purpose and the regression was observed to perform slightly better than decision tree. Regression model was subsequently used to score "new" data. Analyses revealed the drivers and probabilities of a patient being diagnosed as controlled. Obtained results can provide incentives for the business decision maker to explore interventional programs that could enhance the quality of treatment for the uncontrolled diabetic. The article provides an added value to business and the analytic literature by exploring and explaining predictive analytics and associated techniques from the perspective of the business.

Keywords: data mining, healthcare analytics, SAS suite, machine learning, regression, decision tree.

1. INTRODUCTION

The healthcare landscape is experiencing unprecedented growth in data complexity influenced by rapid technology changes, availability of mobile applications, newly discovered diseases and evolving legislation, However, rapid progress is being made in clinical and data analytics, with unprecedented opportunities for improving healthcare quality (Bates et al., 2014). Healthcare systems are interested in predicting who will become very sick, return to the emergency room (ER) or even die after a recent diagnosis or treatment. As a result, the sector is increasingly interested in exploring new techniques to manage complex data for pattern discovery and decision making (Alkhatib, Talaei-Khoel & Ghapanchi, 2015; Raghupathi, 2010).

Data mining has been proposed in several quarters to address existing data deluge with the proposition that anecdotal data can be trained, validated, modeled and used to score "new" data (Saini & Kohli, 2018; Alkhatib, Talaei-Khoel & Ghapanchi, 2015; Raghupathi & Raghupathi, 2014; Koh & Tan, 2005). Alkhatib (2015, p.3) described data mining as a "process by which data is gathered, analyzed and stored in order to

produce useful and quality information and knowledge". According to Srinivas et al. (2010), mining is an important step in a knowledge discovery process that consists of data cleaning, data integration, data selection, pattern recognition and knowledge presentation. In contrast to traditional data analysis, data mining is discovery driven and may accomplish class description, association, classification, clustering, prediction and time series analysis (Srinivas, Rani, & Govrdhan, 2010).

The purpose of data mining and analytics is to build models for determining an event and predicting the outcome of it. This allows decision makers an actionable information upon which interventional approaches can be developed and deployed. Several authors have described classification data mining techniques with illustrations of their applications to healthcare; these include: *Rule set classifiers*, *Neuro-Fuzzy* (Srinivas et al., 2010), *Bayesian Network Structure Discoveries, Decision Tree algorithms*, *Neural Network Architecture, Support Vector Machines* (Hachesu, et al. 2013). Others have used *Regression algorithms* in healthcare analytics. This case study uses SAS suite to select the better of decision tree (DT) and regression (R) models in analyzing drivers of a controlled diabetic.

Models: Decision Tree and Regression

According to Suryawanshi, DT uses a combination of mathematical and computational techniques to aid description and classification (2012), and to extract knowledge from datasets (Kaur & Wasan, 2004). It is a visual and analytical decision support tool that can provide a graphical representation of obtained knowledge in the form of a tree in a flow chart-like structure (Kaur et al., 2018; Kajabadi, 2009; Kaur & Wasan, 2004). Kajabadi et al. (2009) wrote that each non-leaf node denotes a test on an attribute, and each branch indicates an output of the test. As a result of the hierarchical arrangement of nodes and branches, DT is easily understandable and interpreted. Kaur and Wasan (2006) posited that DTs are best suited for data mining because they are inexpensive to construct, easy to interpret, easy to integrate with database systems and have comparable or better accuracy in many applications. Its reliability and established accuracy in clinical decision-making make it the technique of choice for this study (Sitar-Taut & Sitar-Taut, 2010).

Although the DT can be applied as a *continuous* analytic instrument, it is naturally suitable to address discrete (nominal) attributes, so the probability scores are grouped. The regression, (similar to partial least squares, and neural network algorithms) responds better to (numeric) continuous attributes, so the probability scores are more linear. Several authors have explored the use of regression models in healthcare data analysis and found them to be very suitable (Topuz et al., 2018; Rasella et al., 2014).

Analytical Tools

A plethora of analytical tools currently exist. The most common include R, Python, Hadoop, MongoDB (Saini & Kohli, 2018; Das, Pandey & Saxena, 2017), Stata, SPSS, Revolution Analytics, Alpine Data lab, and SAS amongst others. According to Acock (2005), SAS programs are versatile in the breadth and depth of their capability for data analysis and management although they have steep learning curves. While SPSS and Stata are easier to learn, they are less capable for analytics hence Acock concluded that SAS was best for power users. The SAS suites are offered to educational institutions for free. A study based on a sample of 1,139 articles drawn from three journals found that SAS was used in 42.6 percent of data analyses in health service research (Dembe, Patridge & Geist, 2011). Available tools including SAS suite provide incentives for researchers to mine data; develop actionable models; and help defuse the myth behind current *Big Data* conundrum.

It is common knowledge that data analytic techniques can improve healthcare practices and medical efficiency provided they are expertly applied (Murdoch & Detsky, 2013; Koh & Tan, 2005). What is not known is the preparedness of healthcare organizations to apply analytic tools to address business challenges. Another potential issue is users' ability to correctly pair analytical tools to business questions or challenge. This paper aims to demonstrate the use of SAS analytical techniques including DT and R to address a business challenge in diabetes healthcare.

2. DIABETES HEALTHCARE MANAGEMENT

Diabetes Healthcare

The American Diabetes Association (2018) estimated that the total costs of diagnosed diabetes rose by 26 percent from $245 billion in 2012 to $327 billion in 2017. The report also showed that people with diagnosed diabetes incur $16,752 average medical expenditures per year. About $9,601 of this is directly attributed to diabetes. As a result, diagnosed diabetics have medical expenditures approximately 2.3 times higher than what expenditures would be in the

absence of diabetes. Care for people diagnosed with diabetes account for 25% health care dollars in the U.S., and more than half of that expenditure is directly attributable to diabetes.

Maguire and Dhar (2012) found that the top 10% of newly diagnosed type II diabetes patients account for 68% of healthcare utilization. The global health and economic cost is enormous with type II diabetes accounting for 90-95% of all diagnosed cases of diabetes (CDC, 2017). Diagnosed diabetes account for more than 20% of health care spending in the United States (US). Although it may be underreported, diabetes was found to be the seventh leading cause of death in the United States in 2013. It leads to microvascular and macrovascular complications which result in enormous morbidity, disability and mortality (Young et al., 2008); it is also the leading cause of kidney failure, lower-limb amputations, and adult-onset blindness (CDC, 2017).

Given the enormous costs, limited socioeconomic resources, and healthcare organizations' limited ability to invest in disease management initiatives for high-risk diabetic patients, it is important to develop models that predict which patients are at highest risk of adverse medical outcomes. Big data analytics can enhance healthcare quality per unit of spend through optimum utilization of available healthcare data (Kumari & Rani, 2018) existing in several silos including claims, clinical, geomapping, condition, and pharmacy amongst others.

Business Challenge and Goals

Patients with diabetes either don't make enough insulin (type I diabetes) or can't use insulin properly (type II diabetes). Insulin is needed to allow blood sugar (glucose) needed for energy to enter body cells; however, when the body is inadequate in insulin or unable to use it effectively, blood sugar builds up and can lead to several complications (CDC, 2017). The business challenge is to move uncontrolled type II diabetic patient (H1AC > 7) to a controlled state (H1AC < 7). In order to accomplish this, we will (a) develop models and determine the factors associated with a patient designated as controlled; (b) we will score new data to predict the probability of a patient being controlled.

The key opportunity in deploying advanced analytics is the insight it provides in developing interventional approach toward the patient care optimization. Traditionally, management of the diabetic population health relied on arcane reactive approaches that have seen the economic costs and management of diabetics increase every decade. For example, if we have an insurance provider that wants to minimize the dollar risk for the covered diabetic patients; however, once the patient is admitted to the ER or hospital due to conditions that require immediate medical care, it's basically too late – as the treatment must occur and the insurance company must cover their agreed upon portion. Basically, the insurance provider is being descriptive or "looking in the past". However, what if the insurance provider was able to establish specific drivers and a probability for why the patient may have to visit the ER *before* the event actually occurred? Results can inform interventional programs that could save the insurance company dollars at risk for that patient. This method is regarded as being prescriptive or "looking ahead of the curve". This example is the new perspective of the insurance provider.

From the patients' perspective: Suppose drivers can be established for what is causing a patient to be uncontrolled with regards to their diabetic management and the probability of such an event is obtained; this exercise helps the patient to be more proactive regarding their health. Perhaps the insurance provider can establish health coach programs for these patients that are deemed to have a high probability of being uncontrolled and reach out to them – as this would not only improve the patient's care but also minimize the dollars at risk for the insurance provider.

Despite the promises of data mining in healthcare however, there are several challenges that often need to be overcome. For instance, data often exist in dissimilar technology platforms; inferring knowledge from complex heterogeneous patients and data sources may be quite difficult and complicated especially for the untrained, so also is the ability to leverage the patient/data correlations in longitudinal records. This may further exacerbate the already computationally difficult task of analyzing clinical data.

Objectives

This paper demonstrates the use of SAS suite and data mining steps involved in leveraging massive data sets in providing timely patient intervention and personalized care that could benefit components of a healthcare system including provider, payer, patient and management. Srinivas et al. (2010) decried the lack of effective analysis tools to discover hidden relationships and trends in data. Through the use of SAS analytical software, this article introduces the healthcare stakeholder to the specific use of machine learning tools and the possible challenges and/or

techniques associated with using these tools in the healthcare domain.

3. DATA SOURCES

Available Data, Sources and Preparation
Advanced analytics cannot be performed, until data sources and attributes are identified, cleaned and prepared for analyses. Three distinct data sources were identified for this article as follows: (a) Clinical table containing lab results at the patients level (from various hospital labs in a .XLSX format); (b) Demographic table containing information such as the patients' age and gender (from the insurance provider in a .CSV delimited format); (c) Geo-mapping table containing information for mapping such as city and state (from the insurance provider in a .TAB delimited format).

4. SAS STUDIO ANALYTICAL DATA PREPARATION PROCESS

The analytical techniques and tools used in this case study include the following: SAS® 9.4 (BASE); SAS® Studio 3.71; SAS® Enterprise Miner (EM) 14.3. The analytical data preparation process can be very time consuming and is essential in ensuring the analytics conducted is not only accurate but answers the specific business question desired. SAS Studio is a point-and-click, menu- and wizard-driven tool that empowers users to analyze data and publish results.

SAS Studio Data Preparation Flow
This flow is divided into 7 sections (see Figure 2 in the Appendix) as described below: We created a folder and library in SAS studio (this is basically a Libname statement); and imported the three different files (excel, csv, tab format) to create respective SAS datasets. The three datasets were merged, using patient_ID and performing an inner join on the clinical dataset. Overall, we have approximately seventy (70) thousand unique patient records and approximately 130 variables. Based on industry standard for diabetes management, we used the clinical measure of Hemoglobin A1c for the business rule creation of the Target – what we used as our Y for the predictive model of a controlled diabetic. A simple SAS program was written for this task:

```
if HA1c < 7 Then Diabetes_Controlled = 1;
Else 0.
```

Figure 3: SAS code for creating the target

We performed a one-way frequency exploration on the newly created variable to assess the theoretical soundness for predictive modeling. Based on experience, the best practice is to have 80% 0's and 20% 1's for a binary target (Y). Our results yielded about 68% 0's and 32% 1's – which is deemed acceptable. We would not wish to have a 50%/50% split or more 1's than 0's as that would not theoretically ensure a strong predictive model. The goal of the modeling process is to establish the drivers that increase the percentage of 1's (uncontrolled) and optimize those results.

Subsequently, we created a 10% random sample from the final dataset with the target. This is basically a litmus test that will be used for scoring the new model at the end of the EM flow. Since this 10% dataset will have the actual targets, we can immediately inspect if the model was able to predict those records accurately - specifically the 1's. This is a very important step in the data mining process before scoring a new dataset ("blind file") where no target exists. The final piece of the flow de-duplicates those 10% records from the final dataset that will be brought into EM for data mining & predictive modeling. This creates 90% of the raw datasets that will be brought into the EM for modeling. A simple SAS program to execute this is shown below. Once preparation processes concluded in SAS Studio, data was exported into the EM.

```
data health.diabetes_to_model_90pct;
merge health.diabetes_merged (in=a)
health.diabetes_to_be_scored_10pct (in=b);
if a and b then delete;
run;
```

Figure 4: SAS code for deduplication

5. DATA MINING & PREDICTIVE ANALYTICS

SAS EM and the SEMMA Process
The SAS Enterprise Miner is a machine learning tool that helps to streamline the data mining process for the user to create accurate predictive and descriptive analytical models using vast amounts of data. The EM functionality depends on the SEMMA process. SEMMA is an acronym for Sample, Explore, Modify, Model, and Assess - all pertain to conducting data mining and predictive modeling tasks. The process creates a specific

step-by-step strategy for executing this analytical exercise. It also allows for several check points along the way in case certain steps need to be modified or adjusted.

Adjusting the metadata for the input data node

There are four (4) roles of concern: (a) ID, (b) Target, (c) Input, (d) Rejected. Each role have different levels that were addressed including binary, interval, nominal, ordinal and unary.

Data Exploration to assess missing values

Certain models like regression and neural networks cannot have missing values for any variable otherwise those records will be deleted and made unavailable for modeling. Therefore, if any of these models are used, it is important that the practitioner make the necessary modifications by first exploring the data using StatExplore node in EM. Henceforth, the impute node can be activated as explained later in this chapter (Note that the best practice is to partition data before imputation is performed). In addition, there are two other pieces of output that offer business value: (a) a bar chart showing only the correlations between the nominal (categorical) inputs against the target (Y = controlled) in descending order (most associated to least associated); (b) a bar chart showing the correlations between all of the variables (nominal and interval) against the target. Figure 5 (see Appendix) shows three pieces of information which provide business value:

First, the upper left window shows how all the categorical variables correlated with the target in descending order. We observed that the number of adverse events is highly associated with a controlled diabetic. The bottom window shows all the variables (categorical and continuous) and their association with the target in descending order. The upper right window shows descriptive statistics output, which allowed us to inspect one critical column – Missing. This statistic is essential for modeling, because if a variable is missing, then the entire record is omitted from the model. Therefore, possible imputation methods may need to be applied to the categorical and continuous variables, respectively.

Data Partitioning

An important step prior to model building is to divide the data into training and validation datasets; best practice suggests 70:30 respectively. Note that the dataset being partitioned is the 90% raw dataset that was previously created in the SAS Studio. Furthermore, data partitioning needs to account for the distribution of 0's and 1's within the target – this is called stratification. The software handles this automatically. This ensures that both datasets (training and validation) will have almost the exact percentage of 0's and 1's within the target. Figure 6 provides a summary statistics for class targets.

Summary Statistics for Class Targets

Data=DATA

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
Controlled	0	0	42455	67.9900	
Controlled	1	1	19988	32.0100	

Data=TRAIN

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
Controlled	0	0	29719	67.9913	
Controlled	1	1	13991	32.0087	

Data=VALIDATE

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
Controlled	0	0	12736	67.9870	
Controlled	1	1	5997	32.0130	

Figure 6: Summary Statistics for Class Targets

Data Imputation

There are many techniques to use for imputation. It is best practice to use the DT method for nominal variables (this is due to the fact that DTs handle missing values in the growth of the trees). For the interval variables, the median is recommended (Figure 7) because it is more robust and less sensitive to outliers than other measurement such as the mean.

Property	Value
General	
Node ID	Impt
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Nonmissing Variables	No
Missing Cutoff	50.0
Class Variables	
Default Input Method	Tree
Default Target Method	None
Normalize Values	Yes
Interval Variables	
Default Input Method	Median
Default Target Method	None
Default Constant Value	

Figure 7: EM Property Selection for Imputation

For the categorical (nominal) variables, there are a variety of techniques to choose from. The Tree replaces missing class variable values with values

that are estimated by analyzing each input as a target. The remaining input and rejected variables are used as predictors. The imputed value for each input variable is based on the other input variables, hence this imputation technique is more accurate than simply using the variable *Mean* or *Median* to replace the missing tree values. For the continuous variables, there are also a variety of techniques to choose from. A preferred method is the distribution technique, which replaces values that are calculated based on the random percentiles of the variable's distribution. The assignment of values is based on the probability distribution of the non-missing observations. This imputation method typically does not change the distribution of the data very much. Another common technique is the *Median*, which replaces missing interval variable values with the 50th percentile. The *Median* is less sensitive to extreme values than the *Mean* or *Midrange*. Therefore, the *Median* is preferable when you want to impute missing values for variables that have skewed distributions because it allows you to select for each variable.

Data Transformation

An important modification technique that helps normalize the data is to adjust left or right skewness in a variable (the heaviness of the tail in terms of kurtosis). There are many mathematical techniques to do this adjustment. The best practice which we also use in this study is the Log(x).

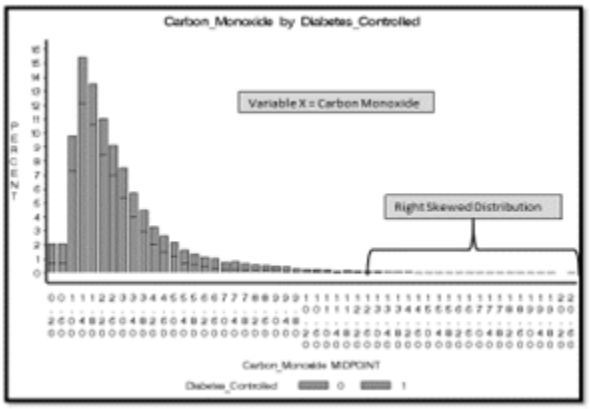


Figure 8a: Data Normalization Using the Log(x)

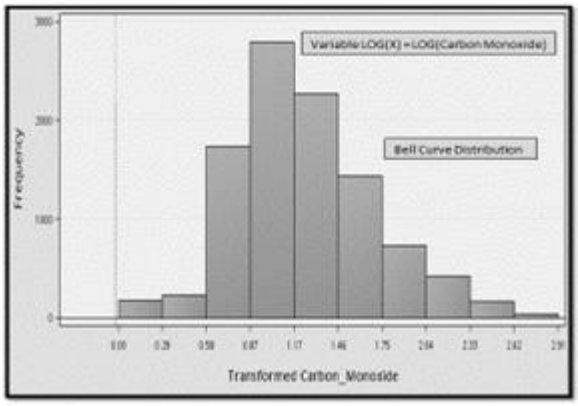


Figure 8b: Data Normalization Using the Log(x)

It is highly recommended to not use the overall method for variables, but rather, select specific transformation techniques for each variable. The overall method may tend to be dangerous as it is usually not the case that all categorical variables should have the same mathematical transformation applied, as in continuous variables. Figure 9 illustrates the second and recommended way, which selects specific techniques *a la carte* for the variables.

Figure 9: Variable Properties for Transformation Node

Building the models

It is important to note that although there are 14 modeling nodes (Figure 10), there are several different modeling techniques available within the property settings that result in much more than 14 algorithms. In this example, we will discuss two types of models (Decision Tree and Regression). Although 2 models are used in this study, there are several variations of techniques within the properties that can form different

algorithms for our task of establishing drivers of a controlled diabetic.

Figure 10: SAS Enterprise Miner – SEMMA – Model Tab

For the Decision Tree node, we leave the default properties as is, and they can be viewed as "smart starts" tested and developed by SAS R&D for a good starting point in dial settings for the model. However for Regression node, and model, we adjust a couple of settings. The default type of regression is Logistic, which is what we need for our example, since we have a binary target of a controlled diabetic (0/1). We select the stepwise model selection method as it is a hybrid of a forward and backward method and conduct a variable selection-like process to choose the drivers of the model. In addition, the criterion we desire is the validation misclassification rate, since we wish to accurately classify a patient of being controlled, hence, we also wish to minimize the misclassification rate. Furthermore, we select the dataset that is designed to verify the model's development – the validation data set (see Figure 11).

Property	Value
General	
Node ID	Reg
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Equation	
Main Effects	Yes
Two-Factor Interactions	No
Polynomial Terms	No
Polynomial Degree	2
User Terms	No
Term Editor	
Class Targets	
Regression Type	Logistic Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	Stepwise
Selection Criterion	Validation Misclassification
Use Selection Defaults	Yes

Figure 11: Property Values for Model Selection

Model Comparison
Once we build our models, an important aspect is to make a head-to-head comparison. The Model Comparison node is used for this assessment and can be found within the Assess tab of SAS EM as indicated below.

Figure 12: SAS Enterprise Miner – SEMMA – Assess Tab

SAS EM Model Comparison Flow
In order for EM to select the "best" model, we need to adjust some property settings. Figure 13 (see Appendix) shows the property setting adjustment for Model Comparison Flow.

Property	Value
General	
Node ID	MdlComp
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Assessment Reports	
Number of Bins	20
ROC Chart	Yes
Recompute	No
Model Selection	
Selection Data	Default
Selection Statistic	Misclassification Rate
HP Selection Statistic	Default
SAS Viya Selection Statistic	
Selection Table	Validation
Selection Depth	10

Figure 13: Property Setting Adjustment for Model Comparison Flow

SAS Enterprise Miner – Model Comparison Statistics
The goal is to accurately classify a controlled diabetic from the model, hence, we select the misclassification rate, and also use the validation data set as the file to use for selection. The next step is very important, as we must assess the model from a theoretical standpoint, before proceeding to look at the individual model results. Although Figure 14 (see Appendix) indicates that both models are very similar in misclassification rate, the regression model (0.202477) is slightly better and is selected over DT (0.20253). Overall, the models show approximately 80% chance of accurately classifying a controlled diabetic patient.

6. RESULTS AND DISCUSSION

SAS Enterprise Miner – Model Comparison Results – Cumulative Lift
Another important theoretical checkpoint is the measurement of the cumulative lift vs. the depth on the graph (See Figure 15 in the Appendix). *What does this graph mean?* Let's imagine that we have a file of 1,000 patients where we desire to develop the probability of being a controlled diabetic. We then score that file with a winning model from EM, and sort the file in descending

order by deciles or depth. Now it makes sense in this graph that the cumulative lift line is decreasing down to 1 when the depth reaches 100% of the scored file. Also, the cumulative lift is highest at the smallest depth, say 5% of the scored file in descending order. *We now ask, what is Cumulative Lift?* The idea of lift refers to the predictive power or accuracy of the model at various deciles of a scored file. We would expect the predictive power to be very high when we are not deep into the scored file, because there is not a lot of variability among the observations. However, the deeper you get into that scored file, the more observations are found, which increases the variability, and makes the predictive power more difficult to achieve, until you finally score the entire file, and are left with no lift, or a lift of 1. Based on experience, we have found that the "sweet" spot of this graph is actually at the 20th depth or decile. Basically, assessing the cumulative lift value should be done at that point for true predictive power. Finally, it is best practice, to look at the 20th decile, a cumulative lift of at least 2 to deem a satisfactory model has been established. We notice, in our models, our cumulative lift is around 2.1, which meets our minimum criterion.

SAS EM – Model Comparison Results – Cumulative % Captured Response

To explain Figure 16 (see Appendix), let's use our example of scoring a patient file of 1,000 records to develop the probability of a controlled diabetic. Based on simple random chance theory, we would expect to obtain 20% of the events (Controlled Diabetics or 1's) as we score the top 20% of the 1,000 observation patient file. Also, we would also expect to obtain 40% of the events as we score the top 40% of the file, and so forth. Recall, from the previous graph, our best practice minimum cumulative lift is 2. If we notice in Figure 16, we are capturing about 40% of the controlled diabetics at the 20th decile, which actually is a cumulative lift of 2 (40% / 20th decile = 2). So think of the random chance theory as a diagonal line drawn from (0,0; 20,20; 40,40; etc.). Therefore, we desire to achieve, in any model, at least 40% of the events of interest at the 20th decile. We need to always be above that imaginary diagonal line for our model to be "good". Finally, this point is the true point of predictive power, and, we can state that if we are provided a file of 1,000 patients whose probability of being a controlled diabetic is desired, we can assume that within the top 20% (200 patients), we can obtain 40% of the controlled diabetics (80 patients). This becomes very powerful, if going deep into a scored file is costly from a health outreach and coaching perspective.

SAS Enterprise Miner – Model Comparison Results – Fit Statistics

The Output shown in Figure 17 (see Appendix) depicts a summary of stepwise selection of models and shows that the Regression model was selected over the DT model. In theory, since all the models are equal, anyone can be used for scoring. However, it is important to assess the drivers of the model used for scoring to ensure it makes sense to the business needs. In addition, choosing a model that is easy to explain from a business perspective is important. The Regression model below shows the factors driving a controlled diabetic below:

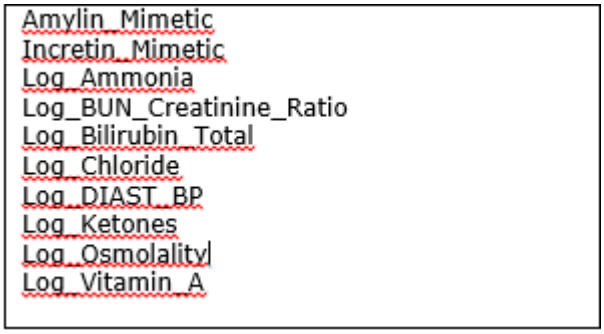


Figure 18: Drivers for controlled diabetic

Note that many of the main drivers selected by the model above, have been transformed using the log(x) technique.

SAS EM – Regression Scoring Results

Now that a solid predictive model has been established for the business question, the next step is to take the winning algorithm to score a patient file i.e. the remaining 10% raw data originally randomly selected and separated as previously indicated. Essentially, we are using the 10% datasets to verify the regression algorithm and develop the probability of controlled diabetic. The DT is a discrete algorithm, so the probability scores are grouped. However, the Regression, (similar to partial least squares, and neural network algorithms) is continuous, so the probability scores are more linear. The Figure below shows the first 10 patients sorted on their probability of being controlled (highest to lowest). This indicates that the model is very useful not only in assessing the distribution of probabilities but more importantly to decide on an appropriate cut-off for assigning which patients exhibit the probability of a given event (for example, being a controlled diabetic).

Patient_ID	Controlled	Probability for level 1 of Co...
1.303634424...	1.0	0.8170103596362283
6.927410991...	1.0	0.7924117380593262
9.876233050...	1.0	0.7897975478471009
7.797040216...	1.0	0.789405552385899
8.413553671...	1.0	0.7840471775353454
4.682905296...	1.0	0.7829350415489733
5.986654475...	1.0	0.7825617247187416
1.320028338...	1.0	0.7821573306515491
6.495120331...	0.0	0.7819917349673773
1.063950438...	0.0	0.7818783693274509

Figure 19: The first ten (10) Patients Sorted by Probability of Being controlled (highest to lowest)

Thus, any patient possessing a score greater than the cut-off score (often decided by the business) is deemed to be a "1"; otherwise, they are tagged as a "0". In this example, the probability of a patient being a controlled diabetic is calculated. It would be desirable for a majority of patients to be at the high end of the score spectrum, as that would mean lower risk of being uncontrolled and as a result would imply lower medical risk. The lower probability implies that the patients are uncontrolled and need better care management. The desire would be to move the population towards the higher end scores - being more controlled with their diabetes.

7. CONCLUSION

This paper demonstrates *big data* promise and potential in healthcare specifically using SAS analytical tool. Data preparation is critical before any data mining and predictive modeling can be executed. Figure 20 (see Appendix) shows the final flow of the algorithm described above, which were used in developing the model. In our specific example, the goal is to manage population health and gain better insight into the patients. Transitioning patients from being uncontrolled with their diabetes to a more controlled state is essential in minimizing risk and also optimizing care. The drivers of a controlled diabetic appear to be chemical factors each of which can be influenced with the right therapeutic intervention in uncontrolled diabetics. Given new patients records with corresponding diagnosis, data mining and analytic techniques demonstrated are able to determine patients at risk or the probabilities of a patient going/returning to the ER. This can cause enormous stress to the patients, health and socio-economic systems in terms of costs, morbidity and death. Hence early interventional initiatives can target such patients and potentially move them to a more desirable state resulting in better business decision making and dollar savings for health systems.

8 REFERENCES

Acock, A.C. (2005). SAS, Stata, SPSS: A Comparison. *Journal of Marriage and Family*. Retrieved September 21, 2005 from: https://onlinelibrary.wiley.com/doi/full/10.1111/j.1741-3737.2005.00196.x

Alkhatib, M.A., Talaei-Khoel, A & Ghapanchi, A. H. (2015). Analysis of Research in Healthcare Data Analytics. *Australasian Conference on Information Systems, 2015 Sydney.*

American Diabetes Association (2018). The Cost of Diabetes (March 22, 2018). Retrieved September 24, 2018 from: http://www.diabetes.org/advocacy/news-events/cost-of-diabetes.html

Bates, D.W., Saria, S., Ohno-Machado, L., Shah, A. & Escobar, G. (2014). Big Data In Health Care: Using Analytics To Identify And Manage High-Risk And High-Cost Patients. *Health Affairs 33, No. 7*(2014): 1123-1131

Center for Disease Control (CDC). 2017 https://www.cdc.gov/chronicdisease/resources/publications/aag/pdf/2016/diabetes-aag.pdf

Das, D., Pandey, R. & Saxena, A. (2017). Disease prediction using Hadoop with Python. Retrieved September 22, 2018 from: https://www.researchgate.net/publication/322714853_Disease_prediction_using_Hadoop_with_Python

Dembe, A.E., Patridge, J.S., & Geist, L.C. (2011). Statistical Software Applications Used in Health Services Research: Analysis of Published Studies in the US. *BMC Health Services Research 11*:252. Retrieved September 21, 2018 from: https://bmchealthservres.biomedcentral.com/track/pdf/10.1186/1472-6963-11-252

Hachesu, P.R., Ahmadi, M., Alizadeh, S. & Sadoughi, F. (2013). Use of Data Mining Techniques to Determine and Predict Length of Stay of Cardiac Patients. *Healthc Inform Res. 2013 Jun;19*(2):121-129. English. Published online June 30, 2013. https://doi.org/10.4258/hir.2013.19.2.121

Kajabadi A, Saraee M.H., Asgari S. (2009). Data mining cardiovascular risk factors; *Proceedings of International Conference on Application of Information and*

Communication Technologies; 2009 Oct 14-16; Baku, Azerbaijan. pp. 1-5

Kaur, A., & Arora, J. (2018). Heart Disease Prediction Using Data Mining Techniques: A Survey. Intl. J. of Advanced research in Computer Science vol 9(2). Retrieved September 23, 2018 from: http://www.ijarcs.info/index.php/Ijarcs/article/view/5872

Kaur, H. & Wasan, S.K. (2006). Empirical Study on Applications of Data Mining Techniques in Healthcare. *Journal of Computer Science 2 (2): 194-200, 2006*

Koh, H.C. & Tan, G. (2005). Data Mining Applications in Healthcare. *J Healthc Inf Manag. 2005 Spring; 19*(2):64-72.

Kumari, S. & Rani K.S. (2018), Big Data Analytics for Healthcare System (February 7, 2018). *IADS International Conference on Computing, Communications & Data Engineering* (CCODE) 7-8 February. SSRN: https://ssrn.com/abstract=3168338 or http://dx.doi.org/10.2139/ssrn.3168338

Maguire, J. & Dhar, V. (2012). Comparative effectiveness for oral anti-diabetic treatments among newly diagnosed type 2 diabetics: data-driven predictive analytics in healthcare. *Health Systems July 2013, Volume 2, Issue 2*, pp 73–92 Retrieved 12/31/2017 from: https://link.springer.com/article/10.1057/hs.2012.20

Murdoch, T.B. & Detsky, A.S. (2013). The Inevitable Application of Big Data to Health Care. *JAMA 2013; 309*(13):1351-1352. doi:10.1001/jama.2013.393

Raghupathi, W. & Raghupathi, V. (2014). Big Data Analytics in healthcare: promise and potential. *Health Information Science and Systems 2*(1) page 3.

Raghupathi W. (2010). "Data Mining in Health Care. In Healthcare Informatics: Improving Efficiency and Productivity." Edited by Kudyba S. Taylor & Francis; 2010:211–223

Rasella, D., Harhay, M.O., Pamponet, M.L., Aquino, R., & Barreto, M.L. (2014). Impact of primary health care on mortality from heart and cerebrovascular diseases in Brazil: a nationwide analysis of longitudinal data *BMJ 2014*;349 :g4014

Saini S., & Kohli S. (2018) Healthcare Data Analysis Using R and MongoDB. In: Aggarwal V., Bhatnagar V., Mishra D. (eds) Big Data Analytics. *Advances in Intelligent Systems and Computing*, vol 654. Springer, Singapore

Sitar-Taut, D.A. & Sitar-Taut, A.V. (2010). Overview on how data mining tools may support cardiovascular disease prediction. *J Appl Comput Sci 2010;4*(8):57–62.

Srinivas, K., Rani, B.K., & Govrdhan, A. (2010). Application for Data Mining Techniques in Healthcare and Prediction of Heart Attacks. *Intl. J. on Computer Sc. and Engineering (IJCSE) vol 2* (2) pg. 250-255

Suryawanshi, R.D. & Thakore, D.M. (2012). Classification techniques of datamining to identify class of the text with fuzzy logic; *Proceedings of 2012 International Conference on Information and Computer Applications*; 2012 Feb 17-18; Hong Kong. pp. 263-267

Topuz, K., Zengul, F.D., Dag, A., Almehmi, A., & Yildirim, M.B. (2018). Predicting graft survival among kidney transplant recipients: A Bayesian decision support model. *Decision Support Systems, Volume 106*: Pages 97-109, https://doi.org/10.1016/j.dss.2017.12.004.

Young, B. A., Lin, E., Von Korff, M., Simon, G., Ciechanowski, P., Ludman, E. J. … Katon, W. J. (2008). Diabetes Complications Severity Index and Risk of Mortality, Hospitalization, and Healthcare Utilization. *The American Journal of Managed Care*, *14*(1), 15–23.

Appendices and Annexures

Patient_ID	AGE	GENDER	ADVERSE_EVENT START_DATE	ADVERSE_EVENT STOP_DATE	BUN_Creatinine Ratio	HEMOGLOBIN A1c	BMI	HYPERTENSION	RENAL DISEASE	City	State
85348444	73	F			2.10	4.24	23.12	0	0	Deer Lake	PENNSYLVANIA
507587021	82	F	April 04, 2010	April 12, 2010	53.58	11.49	24.82	0	0	Lake Park	NORTH CAROLINA
561197284	76	F			75.52	0.16	28.70	0	0	Altoona	ALABAMA
618214598	69	M	November 03, 2007	November 06, 2007	1.75	0.02	27.95	0	0	Lithonia	GEORGIA
743515800	90	F	July 25, 2011	August 03, 2011	49.33	6.68	25.47	0	0	Beersheba Springs	TENNESSEE
911507067	75	F	May 06, 2011	May 15, 2011	46.84	2.10	22.54	1	0	Madawaska	MAINE
1009556938	82	M	July 03, 2011	July 06, 2011	8.96	7.35	29.28	0	0	Herron	MONTANA
1319853858	75	F	August 02, 2007	August 11, 2007	18.28	20.59	23.70	0	0	Linden	MICHIGAN

Figure 1: A Portion of SAS Merged Dataset using Patient_ID as the Primary Key

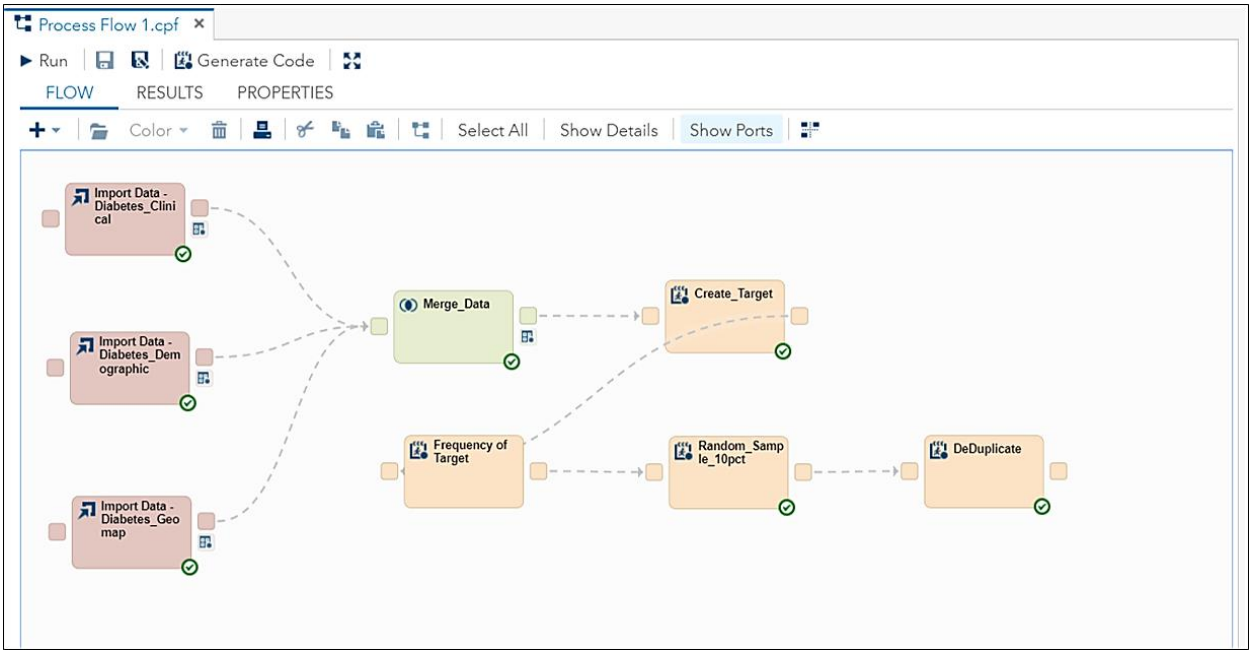


Figure 2: SAS Studio Data Preparation Flow

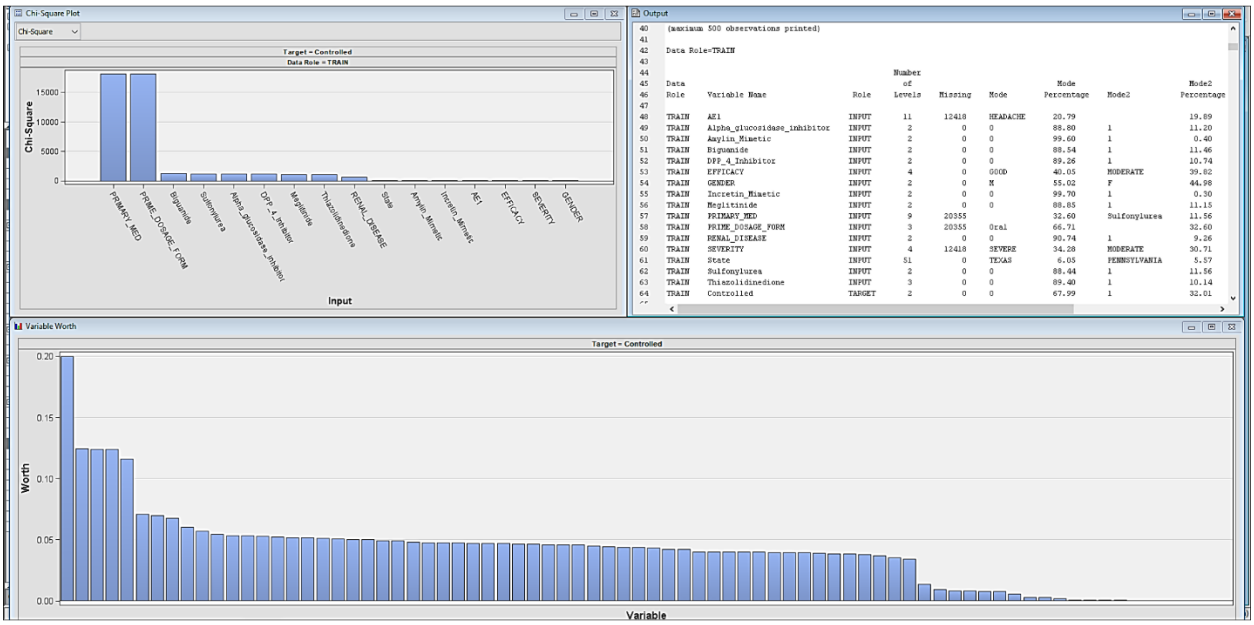
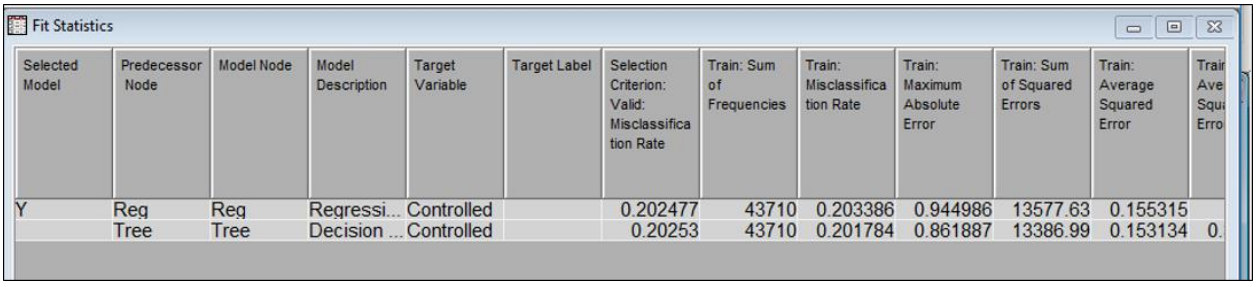


Figure 5: Result output to help Determine the Extent of Missing Data



Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Train: Sum of Frequencies	Train: Misclassification Rate	Train: Maximum Absolute Error	Train: Sum of Squared Errors	Train: Average Squared Error	Trai Ave Squ Erro
Y	Reg	Reg	Regressi...	Controlled		0.202477	43710	0.203386	0.944986	13577.63	0.155315	
	Tree	Tree	Decision ...	Controlled		0.20253	43710	0.201784	0.861887	13386.99	0.153134	0.

Figure 14: Model Comparison Flow

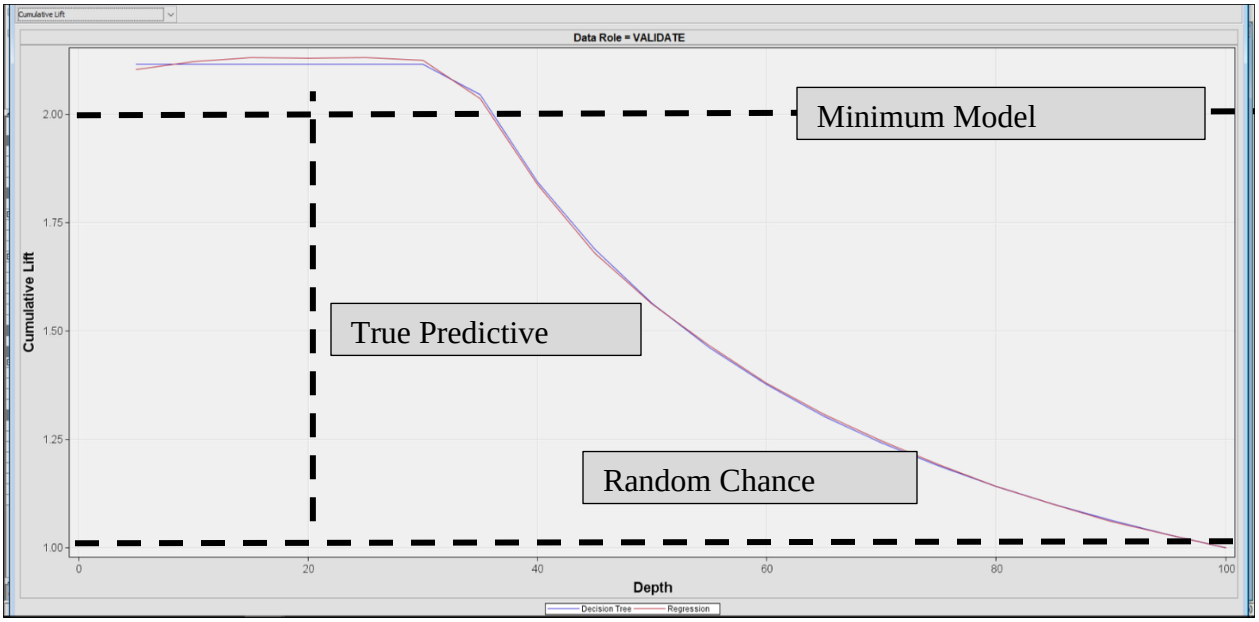


Figure 15: Model Comparison Results

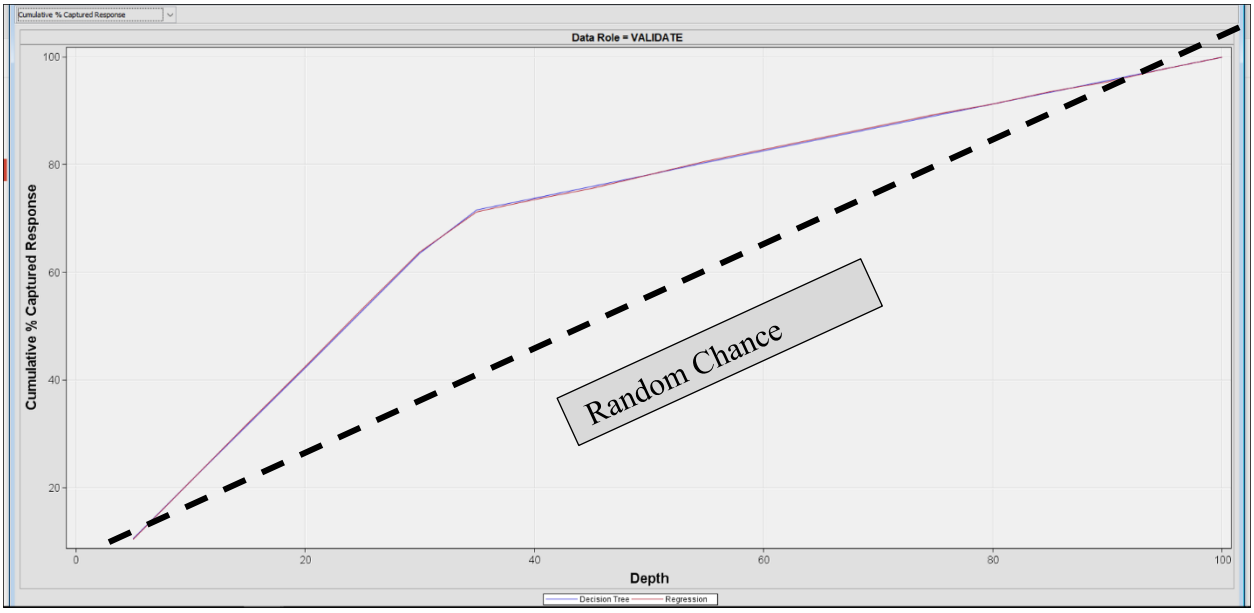


Figure 16: Model Comparison Results – Cumulative % Captured Response

Summary of Stepwise Selection

Step	Effect Entered	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq	Validation Misclassification Rate
1	LOG_Ketones	1	1	11907.8190		<.0001	0.2033
2	LOG_Bilirubin_Total	1	2	133.4161		<.0001	0.2033
3	LOG_Chloride	1	3	24.9012		<.0001	0.2032
4	LOG_DIAST_BP	1	4	21.3564		<.0001	0.2032
5	Amylin_Mimetic	1	5	19.1112		<.0001	0.2029
6	Incretin_Mimetic	1	6	19.4706		<.0001	0.2027
7	LOG_Osmolality	1	7	16.6654		<.0001	0.2026
8	LOG_Ammonia	1	8	14.6214		0.0001	0.2026
9	LOG_Vitamin_A	1	9	11.4710		0.0007	0.2026
10	LOG_BUN_Creatinine_Ratio	1	10	10.7005		0.0011	0.2025
11	LOG_Urine_PH	1	11	9.5332		0.0020	0.2026
12	LOG_Carbon_Monoxide	1	12	8.7770		0.0031	0.2026
13	LOG_Cholesterol	1	13	7.8737		0.0050	0.2027
14	LOG_Pyruvic_Acid	1	14	7.0667		0.0079	0.2027
15	LOG_Prostate_Specific_Antigen	1	15	6.1928		0.0128	0.2027
16	LOG_Blood_Volume	1	16	4.8873		0.0271	0.2028
17	LOG_Lactate	1	17	4.1612		0.0414	0.2029
18	LOG_MCV	1	18	3.9101		0.0480	0.2029

The selected model, based on the misclassification rate for the validation data, is the model trained in Step 10. It consists of the following effects:

Intercept Amylin_Mimetic Incretin_Mimetic LOG_Ammonia LOG_BUN_Creatinine_Ratio LOG_Bilirubin_Total LOG_Chloride LOG_DIAST_BP LOG_Ketones LOG_Osmolality LOG_Vitamin_A

Figure 17: Summary of Stepwise Selection

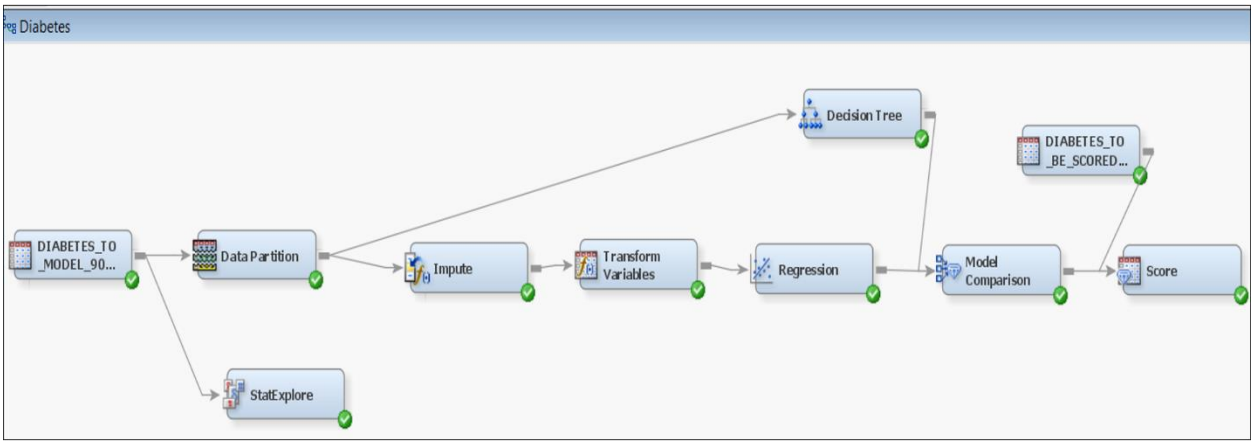


Figure 20: The Final Model Flow