Classification of patients suffering from HIV and drug abuse on the basis of severity of illness

Final Consulting Report

Healthcare Analytics Using Texas Hospital Inpatient Discharge Public Use Data File

Group 6

Jayshil Patel	jayshil@tamu.edu
Tanshi Arora	tanshi.arora@tamu.edu
Amrutha Mandadi	amrutha.mandadi66@tamu.edu

Contents

ABSTRACT

Healthcare is one of the largest and most complex industries that is continuously progressing towards new and better out-patient and in-patient delivery care. Analytics in healthcare provides a combination of financial and administrative data alongside information that can aid patient care efforts, better services, and improve existing procedures. Healthcare analytics is the process of deriving insights from raw data and finding patterns and correlations to make better healthcare decisions. A huge chunk of healthcare data is hence collected in order to analyze the potential areas that can be improved for better services, disease management and healthcare employee management. Current objective in healthcare is to have a healthy population over providing better treatment. Improving care before treatment is one of the best ways of introducing quality to healthcare.

There is an observed mismatch between the resources and the demand which is one of the clinical challenges faced today. One such resource is doctor's time that needs to be efficiently utilized to improve healthcare. Clinical efficiency can be improved indirectly by reducing the number of patients in the healthcare that will also result in better population health.
Healthcare industry has a varied patient group and it is required by them to service the needs of all kinds of patients. Healthcare for certain groups such as patients with disability, HIV, substance abuse, chronic diseases require special care and extended services. It is important to study and use analytics in this domain to be prepared and to provide the best service without any gap in the process.

Having more information about the patient suffering from HIV and drug abuse order will help the administration of the hospital to improve clinical efficiency. There are different units in hospitals such as Coronary Care Unit, Pediatric Unit, Rehabilitation Unit etc which are recommended to patients depending on their illness. The utmost goal of healthcare industry is to provide care and comfort to the patient while on treatment. On similar lines, as quoted in our data mining problem, patients with HIV or drug abuse condition would require extra care when

they come in for other illnesses. Hence, the issue in the healthcare industry to provide better services to specially categorized patients such as HIV or drug abuse patients can be improved by mining into data to predict better services to them. This could also help clinics and admins to understand the resource utilization and predict future allocation based on current utilization trend. It allows clinics to be prepared and to provide better care based on the illness of such patients considering their underlying medical conditions.

In this project, we are considering group of patients suffering from HIV or diagnosed with drug/substance abuse and how analysis of healthcare data can help us point to services that can be improved. The data mining problem is worked upon by following the data mining steps. We selected the data for HIV/drug abuse patients at an earlier stage to have a focused dataset to solve our business problem. The data mining problem is to classify a new patient into one of the specialized units in the hospital, that forms our target variable. As we classify patient based on certain parameters, they behave as input variables such as illness, ethnicity, age, type of admission etc. On identifying the required variable for our data mining problem, we selected appropriate data and performed dimension reduction and other visualization techniques to clean, transform and validate the dataset required for classification. Further, we partitioned the data to build a model using training set and to test the accuracy of the model using validation set. We identified 14 variables that were used as predictors for our target variable. These predictors were used to create models that helped us in our classifying the HIV or drug abuse patients.

Finally, the above steps were repeated to build different models which were compared with each other. Finally, on comparison we identified KNN to be the best model to classify our project data for HIV and drug abuse patients. We then validated the constructed model for its performance using the validation set.

INTRODUCTION

Healthcare is one of the largest and most complex industries that is continuously progressing towards new and better out-patient and in-patient delivery care. Analytics in healthcare provides a combination of financial and administrative data alongside information that can aid patient care efforts, better services, and improve existing procedures. Healthcare analytics is the process of deriving insights from raw data and finding patterns and correlations to make better healthcare decisions. A huge chunk of healthcare data is hence collected in order to analyze the potential areas that can be improved for better services, disease management and healthcare employee management.

There is an observed mismatch between the resources and the demand which is one of the clinical challenges faced today. One such resource is doctor's time that needs to be efficiently utilized to improve healthcare. Clinical efficiency can be improved indirectly by reducing the number of patients in the healthcare that will also result in better population health.
Healthcare industry has a varied patient group and it is required by them to service the needs of all kinds of patients. Healthcare for certain groups such as patients with disability, HIV, substance abuse, chronic diseases require special care and extended services. It is important to study and use analytics in this domain to be prepared and to provide the best service without any gap in the process.

Historical data is mostly used to understand how to improve accessibility and affordability towards health for the population. Advanced healthcare analytics can help in propelling business growth model towards better medications, less medical costs and improved health. Now-a-days, more and more data are collected on health which can be used to get better insights about predictive modelling, virtual health care methods and health-care education services. Healthcare data can also be used to predict diseases and prevent outbreaks in future. Healthcare analytics has developed beyond just analysis and reporting. In today's world healthcare is making predictions in order to provide a better healthcare world. Also, alternatives to underlying health problems are provided to shift the focus towards healthier living from earlier focus being on providing better healthcare services.

The current objective in healthcare is to have a healthy population over providing better treatment. Improving care before treatment is one of the best ways of introducing quality to healthcare. There is an observed mismatch between the resources and the demand which is one of the clinical challenges faced today. One such resource is doctor's time that needs to be efficiently utilized to improve healthcare. Clinical efficiency can be improved indirectly by reducing the number of patients in the healthcare that will also result in better population health.

Healthcare industry has a varied patient group and it is required by them to service the needs of all kinds of patients. Healthcare for certain groups such as patients with disability, HIV, substance abuse, chronic diseases require special care and extended services. It is important to study and use analytics in this domain to be prepared and to provide the best service without any gap in the process.

In this project, we are considering group of patients suffering from HIV or diagnosed with drug/substance abuse and how analysis of healthcare data can help us point to services that can improved. The characteristics of big data namely Volume, Velocity and Variety are found in the features of healthcare data sources. The data sources for healthcare system are broadly classified as Structured data, Semi-structured data and Unstructured data. Hence, the exploration of healthcare data to achieve valuable insights is a daunting task due to the enormous variety of data from various sources. To extract Value from the healthcare data, data must be collected, processed, analyzed and visualized efficiently to build decision support strategies for different issues in healthcare.

Thus, it is inevitable to have analytics for healthcare with the increase in the type of diseases, population and services being provided. Also, it becomes necessary to train the healthcare workers to learn to use analytics and stay abreast with technology to make the world healthier.

LITERATURE REVIEW

We referred to multiple papers to come up with the following literature review that are in-line with the problem in Healthcare based on our data mining problem.

1. Big Data Analytics for Healthcare Industry: Impact, Applications, and Tools

Today, the challenge in healthcare industry is to handle healthcare data that is growing both in volume and velocity. Most of this growing data generated by the system are in the form of hard copies that need to be digitized. Big data helps improve healthcare delivery and reduce its cost, while supporting advanced patient care. Healthcare analytics is used to predict and make decisions to improve healthcare services.

The impact of healthcare analysis is varied. There are different ways healthcare analytics can be utilized to improve healthcare as healthcare itself is a huge collaboration of various domains within itself. From patient's perspective, better service that can be provided is through healthy lifestyle options. These decisions can be made based on patient's daily lifestyle, diet, exercise and other activities. Analytics on healthcare data can also be utilized to make decisions about healthcare providers and predict better treatment options to patients. A pathway of right innovation is provided as an added advantage by the healthcare analytics that can help recognize new diseases and their respective treatments.

Among these different pathways that are defined under healthcare analytics, our data mining problem falls under "right care" pathway to be provided to the patients. The utmost goal of healthcare industry is to provide care and comfort to the patient while on treatment. On similar lines, as quoted in our data mining problem, patients with HIV or drug abuse condition would require extra care when they come in for other illnesses. Hence, the issue in the healthcare industry to provide better services to specially categorized patients such as HIV or drug abuse patients can be improved by mining into data to predict better services to them. This could also help clinics and admins to understand the resource utilization and predict future allocation based on current utilization trend. It allows clinics to be prepared and to provide better care based on the illness of such patients considering their underlying medical conditions. Data mining problem also address client's risk for retention in care failure before client falls out of care. Healthcare industry has always faced this issue where patients under special category or a condition fall out of care due to their underlying condition adding to their illness or disease. Hence, clinical efficiency is lost in such failures of retaining care for HIV or drug abuse patients.

2. Survey of Big data Analytics in Healthcare and Government

Healthcare industry generated large amounts of data and so does the government under its health sector. Hence the government also requires technology to manage and analyse these huge amounts of data to derive useful insights.

Analytics in healthcare has evolved lately for many reasons. One of it being improvement of healthcare and its quality of service. This can be done in various ways such as providing patient centric services, detecting diseases earlier, monitoring hospital's quality and improving treatment method. The data mining problem we have considered involves providing customer centric services where the underlying conditions of HIV and drug abuse patients are considered to provide better facilities and care at hospitals and clinic when they are admitted for other illnesses. The data from healthcare allows to analyse for such targeted groups based on which their treatment and medicines can be prescribed effectively. Healthcare analytics for the government is required to address basic needs to the population quickly. The analysis of healthcare data will help in predicting the percentage of population that will require immediate or subsidized healthcare.

Considering our data mining problem, the population of HIV and drug abuse patients can be analysed for their admissions to hospitals for other illnesses. This will help in the government sector to provide faster and better facilities to such patients and in planning the hospitalization process for their admission. Further, the population visiting a government healthcare are generally looking for subsidized or lower healthcare charges. As health of the population is most important for any nation, providing best care at affordable rates is always a trade-off. With this trade off, the chosen data mining problem addresses the issue of making special accommodations and providing better care for targeted groups such as HIV or drug abuse patients.

3. Big data analytics in healthcare: promise and potential

With the growing amount of data in healthcare, it is more required for digitization of this data. Healthcare organizations must acquire tools, technologies and infrastructure to effectively manage and utilize this humongous data. Generally, a patient visits the doctor and discusses their condition for which treatment and care decisions are made by the individual doctor. In order to make this process more efficient, analytics can be used where deep research on past cases and prediction of future treatments will help in making this decision more effective and proof-based.

Here we discuss about evidence-based medicine or treatment and patient profile analytics. As proof-based treatment or results are always compared better to trials, analytics helps in achieving this aspect in the healthcare industry. The promise to make health better is always risky and to cut down on this risk, various tools are adopted in make accurate decisions using the systems. Further, patient profiling is important in healthcare lately as it gives the leverage to the healthcare organizations in providing better care to the patients based on the group they fall into.

In our data mining problem, we focus on patient profiling by grouping patients with HIV and drug abuse condition and they would require special needs and care while being treated for the illness they have been admitted for. Due to their underlying condition, these groups could develop certain side-effects or illness physically or mentally that can be addressed if aware of. This way, it improves clinical efficiency as analysis and study would help in providing right treatment based on their condition and side-effects. This in turn cuts down the patient count for each doctor that increases their efficiency too. Further, admin of healthcare organizations can plan their resources based on the study in order to accommodate such patients as quickly as possible and at an affordable price.

4. Analysis of Research in Healthcare Data Analytics

Recent studies show that analytics in healthcare has shifted its focus from being a volume-based business to a quality-based business. The management and analysis of large complex data also now requires to be processed quickly and accurately. This will improve healthcare practice, changing individual life style and driving them into longer life, prevent diseases, illnesses and infections.

Data mining refers to gathering data and preparing them for analysis and prediction. The data is transformed and classified based on the output we are seeking that is required to be qualitative and accurate. While dealing with large data, it is important to understand how to store it and utilize it for prediction and analysis as retrieving this data at a good speed makes a difference in the efficiency of healthcare.

Data mining problem for HIV and drug abuse patients require to perform regular tests and check-ups whose data needs to be stored and compared to the past data for that individual.

To make the right choice for a patient, historical data is equally important to target the right care for their relapse. In order to do this, we have considered factors such as source_of_admission and type_of_illness from which the health of the patient can be understood. Also, as the patients have been diagnosed with HIV or drug abuse disorder, few of these illnesses could be due to their diagnosed disorders. Hence, this helps in preparing the required treatment and care for such patients to ensure smooth experience and better patient service.

The collection of data from patients about their disorder is a step closer to clinical efficiency, as it helps in sharing these facts to patients with similar condition. Data collected for a group of patients classified based on their disorder or disease can be useful in understanding the general treatment and diagnosis trend that can improve overall population health.

5. Exploring clinical care processes using data analytics

Achieving cost-effective healthcare is one of the main challenges faced today. The pattern in the care delivery in healthcare organizations are required to be explored that can produce optimal outcomes at a reduced cost or lower cost. The healthcare data is large spread across years and across varied patients. While studying such large amounts of data, we might lose out on essential factors hidden in it. Additionally, repetitive processing and movement of data across servers can also introduce missing values or wrong patterns of data into the dataset. It is important to address these issues in the dataset before transforming data for analysis and prediction. To add to this complexity, patients have large data tagged to them based on various tests and treatments they seek. There could also be introduction of noises or unwanted additional data that would mislead analysis or prediction.

Moving towards quality-based business in healthcare, this challenge of obtaining the right reliable dataset forms the first step. They have been many tools designed to handle such misleading data and with analytics in healthcare, it is achievable. Dirty data is filtered and cleaned for the required accurate data that is currently being done on real-time using different analytic tools. The data is spread across healthcare system that needs to be integrated to maintain consistency and to provide consistent and accurate outcomes.

Furthermore, algorithms are used to test the data for the outcome that is expected, and the right algorithm is then chosen for making predictions on data that is generated. The techniques discussed to overcome the problem in dataset is similarly applied to Texas Healthcare Information Collection (THCIC) data to address the data mining problem chosen for the project. On performing above techniques, a clean dataset is obtained to further classify HIV and drug abuse patients using factors such as their source of admission and the type pf illness for which they are admitted.

Thus, the main goal of the project is to improve clinical efficiency for a subset of patient population who suffer from HIV or who have been diagnosed with drug abuse disorder.

PROBLEM FORMULATION

The data mining problem is worked upon by following the data mining steps. We selected the data for HIV/drug abuse patients at an earlier stage to have a focused dataset to solve our business problem. The data mining problem is to classify a new patient into one of the specialized units in the hospital, that forms our target variable. As we classify patient based on certain parameters, they behave as input variables such as illness, ethnicity, age, type of admission etc. On identifying the required variable for our data mining problem, we selected appropriate data and performed dimension reduction and other visualization techniques to clean, transform and validate the dataset required for classification. Further, we partitioned the data to build a model using training set and to test the accuracy of the model using validation set.

The variables used in the data to classify patients are mostly categorical. Also, the target variable being categorical, the model built in this report is logistic. The additional advantage of logistic regression to easily learn from the training set makes it easy to implement.

1. Data Exploration

The THCIC data is cleaned and examined for the following variables that are required for our data mining problem. Based on the identified target variable, required input variables or predicted variables are selected in this part of data mining.

Predictor Variable: Predictor variables are mapped to the target variable through an empirical relationship. They can be categorical, continuous or integer. Predictions can be of three types: decisions, rankings and estimates.

1. *TYPE_OF_ADMISSION (Categorical Variable):*

   This field is used to indicate the type of admission of the patient. It consists of values such as “Emergency”, “Urgent”, “Elective”, “Newborn”, “Trauma Center”, “Information not available” and “Invalid”. The patient requiring immediate assistance will fall under the emergency or urgent category depending on the situation. This predictor will help us in getting the type of admission information about the patient. Since, we are classifying the patients suffering from HIV and drug abuse it will facilitate in giving us the clarity of the patient's condition when he/she was admitted.

2. *SOURCE_OF_ADMISSION (Categorical Variable):*

There can be various reasons for which the patient can be admitted. Sometimes, they are referred by a clinic, recommended by lawyers in the court or transferred from other hospitals. This field will source number which will specify their source of admission. This predictor will help our analysis of patients suffering from HIV and drug abuse by providing the details of their source. For example, if there are more patients recommended by court this might imply that people suffering from HIV and drug abuse are either in worst helpless condition or the ones which are referred from other hospitals may signify that there is still lack of medications available to cure them. After performing the thorough analysis some conclusion can be made.

3. *PAT_STATUS (Categorical Variable):*

This field indicates the status of the patient when he/she is about to leave the hospital. For example, this can contain, discharged, expired etc. depending on the situation. It will store numbers which will depict the status of the patient. This predictor will help us know that whether the patient was cured fully when he/she left or there was just little improvement. This can help us track the patients which were not fully cured.

4. *RACE (Categorical Variable):*

This field will store the information about the patient's race. It will include a code referring to a certain race. "1" refers to American Indian/Eskimo/Aleut, "2" refers to Asian or Pacific Islander, "3" refers to Black, "4" refers to White, "5" refers to Other and "`" refers to Invalid. There are certain races which are prone to specific type of diseases. If we come to know about that information, then patients arriving later of same race will be highly prone to that specific disease. Having this information can facilitate in providing cure to people of that race too.

5. *ETHNICITY (Categorical Variable):*

This field includes the ethnicity of the patient. It indicates that whether a patient is Hispanic or not. It includes code signifying that value. “1’ implies Hispanic Origin, “2” implies “Not of Hispanic Origin” and “`” implies Invalid. This predictor will be helpful for knowing the ethnicity of the patient. If there are more people from a certain ethnicity suffering from HIV and drug abuse. Then in future, precautions can be taken, or awareness can be spread to those specific ethnicity people about the HIV and drug abuse in order to avoid it. As hospital/clinic administration, programs can be organized for people around in the nearby areas.

6. *LENGTH_OF_STAY (Continuous Variable):*

This field indicates the duration for which the patient was admitted in the hospital. It is calculated by subtracting the day patient entered the hospital from when he/she leaves the hospital. The minimum length is 1 and maximum length is 9999. This predictor will help us know the duration for which the HIV and drug abuse patients stays in the hospital. Having an average number for this can help the hospital/clinic administration to plan accordingly. For example, like it can facilitate in the bed management that tentatively how long the patient will stay in the hospital or clinic.

7. *PAT_AGE (Categorical Variable):*

This field specifies the age range of the patient. We are using the range of 22 to 26. Mostly, the AIDS and drug abuse is common in the youth. Choosing this predictor and fixing it to a value of 22 to 26 will help our classification.

8. *FIRST_PAYMENT_SRC (Categorical Variable):*

This field indicates the source of the payment done by the patient. There are various options a patient can choose to pay like he can opt for Insurance, Medicaid etc. This field will include codes indicating the option which patient has opted for. This predictor will help us know the mode of payment chosen by the patient. This can facilitate in knowing the financial status of the most patients.

9. *TYPE_OF_BILL (Categorical Variable):*

This field includes the information about the claim data. It includes of three digits. The first indicates the type of facility, second shows the type of care and the third shows the sequence of the claim. This predictor will help the hospital/clinic administration know the details about the kind of facility used by the patients so that they can plan in the future to accordingly manage for that facility. This also gives information about the type of care undertaken by patients suffering from HIV and drug abuse. this can help the administration to plan for that care unit. Maybe they can plan inventory for that unit.

10. *TOTAL_CHARGES (Continuous Variable):*

This includes the total sum the patient must pay for the services he/she availed at the hospital. This predictor will help the hospital/clinic administration to know how much amount of money is charged by the patients suffering from HIV and drug abuse. This will help the administration for financial statements.

11. *PRINC_DIAG_CODE (Continuous Variable):*

This code indicates the patient's principal diagnosis when he/she arrived at the hospital. This predictor will facilitate us in having information about the patient's principal disease.

12. *RISK_MORTALITY (Categorical Variable):*

This field indicates that what are the chances of a patient to die. This value has been assigned considering All Patient Refined (APR) Diagnosis Related Group (DRG) from the 3M APR-DRG Grouper. This predictor will help us knowing that what would be the dying chance of the patient suffering from HIV and drug abuse.

13. *ILLNESS_SEVERITY (Categorical Variable):*

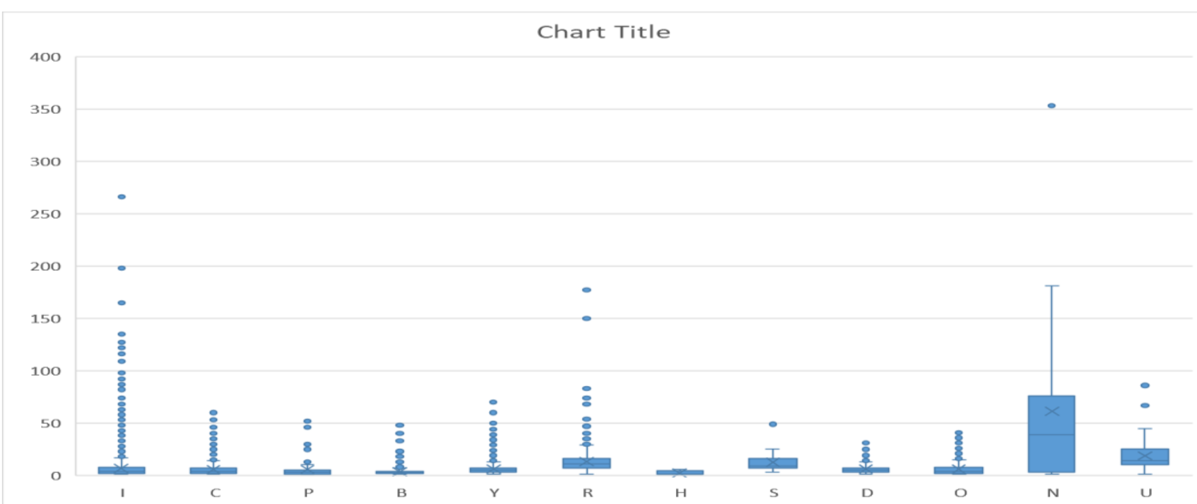
This field indicates that what is the level of severity of the illness of the patient. It includes four code:"1" indicating Minor "2" indicating Moderate "3" indicating Major and "4" indicating Extreme. This value has been assigned considering All Patient Refined (APR) Diagnosis Related Group (DRG) from the 3M APR-DRG Grouper. This predictor will help us know the suffering level of the patients having HIV and drug abuse. Classifying patients on this basis can help planning the medications.

Target Variable: The target variable is the one for which we need the output or classification hence in this case our target variable is SPEC_UNIT_1.

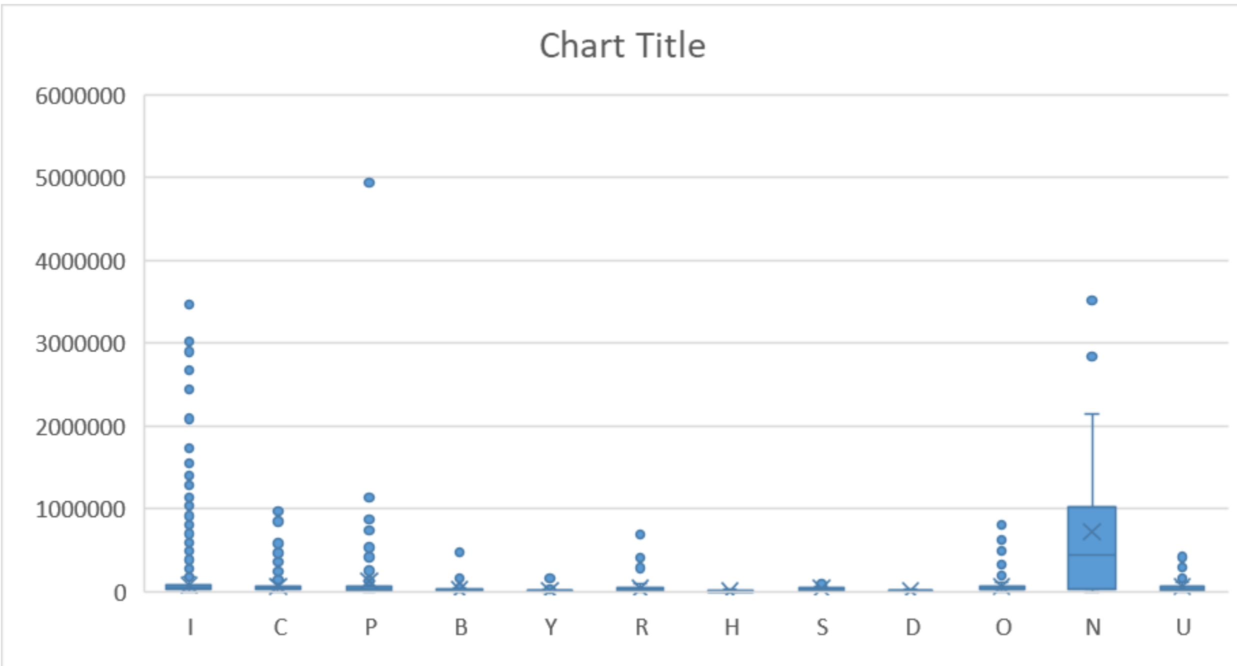
1. *SPEC_UNIT_1 (Categorical Variable):*

This field indicates the specialty units in which most days during stay occurred based on number of days by Type of Bill or Revenue Code. In order by number of days in the unit. It includes codes such as "C" which means Coronary Care Unit, "P" Pediatric Unit etc. With the analysis of this variable will help us identify the speciality unit which the patient was in.

Now we try to understand the relationship between predictor and target variables.

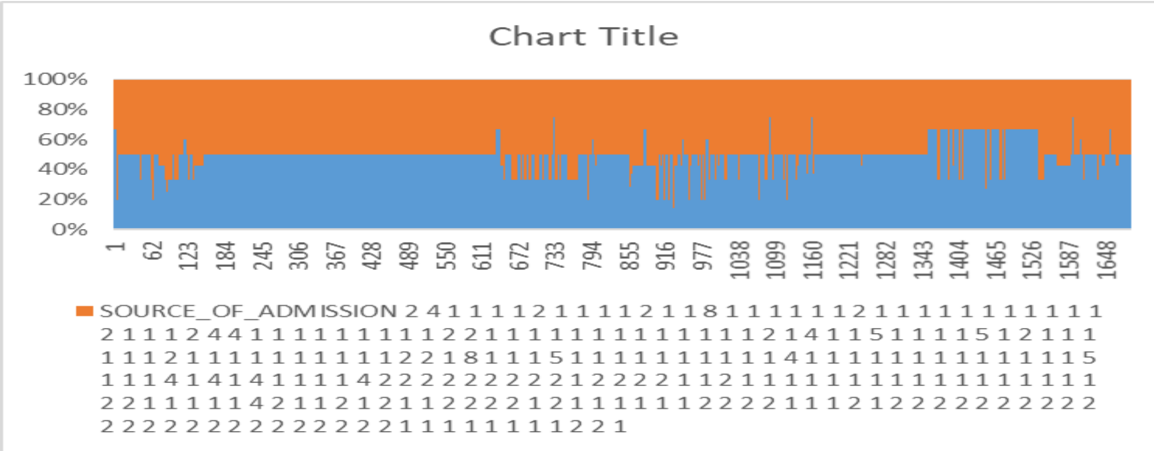


In the above graph we check the relation between predictor LENGTH_OF_STAY and SPEC_UNIT. As we see the box plot, we can identify the relation between the type of specialty unit and the length of stay. The values are distributed around the median quite uniformly for such a large data set. Even the number of outliers is very less. This shows that using length of stay as a predictor will enhance our analysis. Usage of such variables help us reduce average error in case of our predictions.

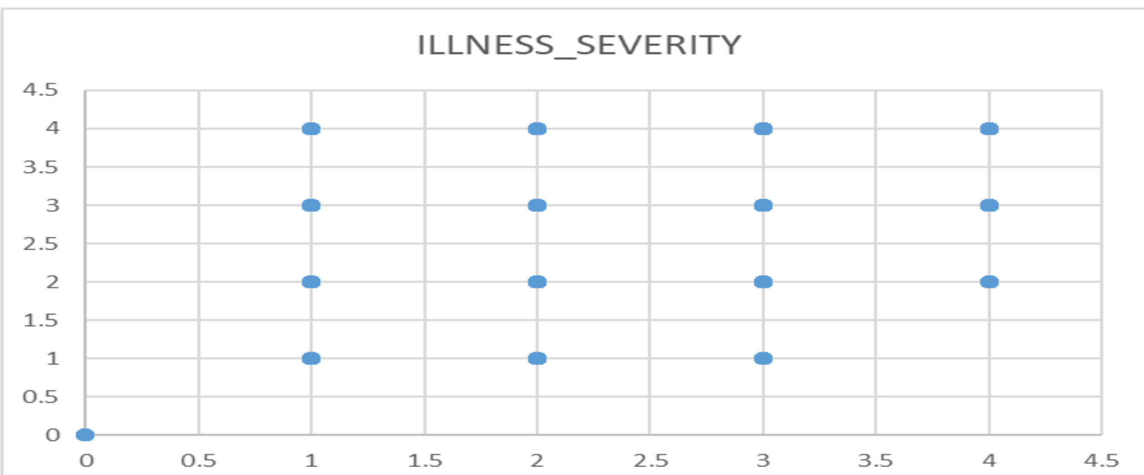


In this graph, we create a box plot to understand the relation between SPEC_UNIT and TOTAL_CHARGES. As we know, that in general cases our charge or cost of treatment depends on the disease. Also, the specialty unit is specific for specific type of disease. Hence, the cost and specialty unit are directly linked to each other. This assumption of ours is proven true by the graph above where the distribution box for the cost is similar for similar specialty units. Even in this case the number of outliers is less.

Apart from this let us understand the relation between multiple predictors as well.



This is a stacked bar plot for variables SOURCE_OF_ADMISSION and TYPE_OF_ADMISSION. Similar columns tend to capture similar information which can lead to multi collinearity. This must be avoided for which we can use just one column instead of both.



It is the similar case with variables ILLNESS_SEVERITY and RISK_MORTALITY. They capture similar information hence we can eliminate either of these.

Apart from this we have a lot of categorical variables that need to be transformed to perform further analysis for upcoming reports.

2. *Data Identification*

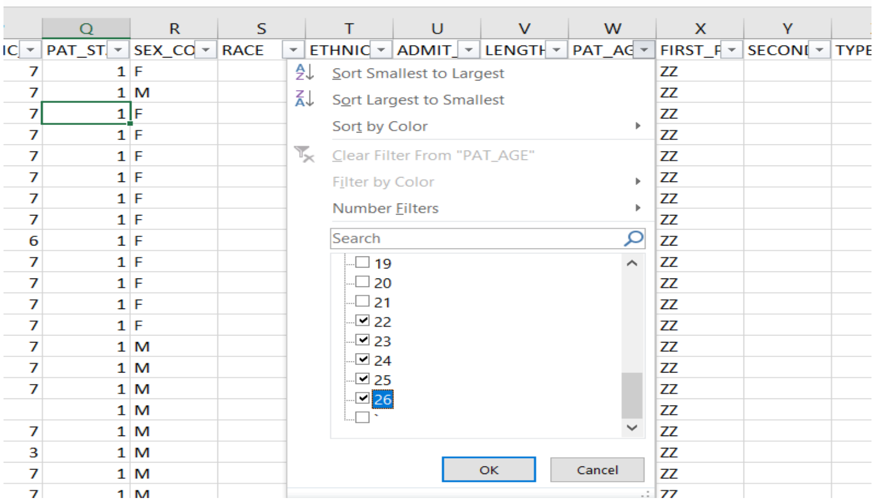
We have identified 13 variables that can be used as predictors for our target variable. These predictors can be used to create models that will help us in our predictions. Once we analyse the given data, we need to first identify the data that is useful for us. Hence, we need to reduce our data from 194 columns to 15 columns including the target variable and ID. This reduction in the number of columns is done as the other columns are of no value while classifying the Specialty units, they just increase the volume of data for our business case.

	RECORD_I	DISCHARG	THCIC_ID	PROVIDER	TYPE_OF_	SOURCE_(	SPEC_UNI	SPEC_UNI	SPEC_UNI	SPEC_UNI	SPEC_UNI	PAT_STATI	PAT_ZIP	PAT_COUI	COUNTY	PUBLIC_H	PAT_STATI	SEX_CODE	RACE	ETHNICITY	ADMIT_W	LENGTH_(	PAT_AGE	FII
2	2.2E+11	2013Q2	100	Austin Stat	2	8						TX	78644	US	55	7	1	F	4	1	1	2	11	ZZ
3	1.2E+11	2013Q1	100	Austin Stat	2	8						TX	77807	US	41	7	1	M	4	2	2	2	11	ZZ
4	2.2E+11	2013Q2	100	Austin Stat	2	8						TX	76705	US	309	7	1	F	3	2	3	1	6	ZZ
5	2.2E+11	2013Q2	100	Austin Stat	2	8						TX	76501	US	27	7	1	F	4	2	4	5	11	ZZ
6	2.2E+11	2013Q2	100	Austin Stat	2	8						TX	76554	US	27	7	1	F	4	2	1	2	10	ZZ
7	1.2E+11	2013Q1	100	Austin Stat	2	8						TX	77802	US	41	7	1	F	4	2	4	5	11	ZZ
8	2.2E+11	2013Q2	100	Austin Stat	2	8						TX	77802	US	41	7	1	F	4	2	2	2	12	ZZ
9	1.2E+11	2013Q1	100	Austin Stat	2	8						TX	78611	US	53	7	1	F	4	2	5	3	11	ZZ
10	1.2E+11	2013Q1	100	Austin Stat	2	8						TX	77539	US	167	6	1	F	4	2	1	1	7	ZZ

Thus, we select those 15 variables and create another sheet of data that does not have any unnecessary columns.

RECORD_ID	TYPE_OF_ADMISSION	SOURCE_OF_ADMISSION	SPEC_UNIT_1	PAT_STATUS	RACE	ETHNICITY	LENGTH_OF_STAY	PAT_AGE	FIRST_PAYMENT_SRC	TYPE_OF_BILL	TOTAL_CHARGES	PRINC_DIAG_CODE	RISK_MORTALITY	ILLNESS_SEVERITY
1.20134E+11	2	8		1	4	2	3	23	ZZ	111	1362	30480	1	1
2.2013E+11	2	8		1	4	2	3	23	ZZ	111	454	2929	1	1
2.2013E+11	2	8		1	4	2	7	23	ZZ	111	3178	2929	1	1
2.2013E+11	2	8		1	4	2	3	23	ZZ	111	1362	30570	1	1
2.2013E+11	2	8		1	4	2	11	23	ZZ	111	4994	30480	1	1
2.2013E+11	2	8		1	4	2	14	24	ZZ	111	6356	30480	1	1
1.20134E+11	2	8		1	4	2	3	23	ZZ	111	1362	30500	1	1
2.2013E+11	2	8		1	4	2	9	23	ZZ	111	4086	30570	1	1
1.20134E+11	2	8		1	3	2	5	23	ZZ	111	2270	29284	1	1
2.2013E+11	2	8		1	3	2	3	24	ZZ	111	1362	30590	1	1
2.2013E+11	2	8		1	3	2	3	24	ZZ	111	1362	30590	1	1
2.2013E+11	2	8		1	3	2	9	23	ZZ	111	4086	30590	1	1
1.20134E+11	2	8		1	3	2	13	23	ZZ	111	5902	30520	1	1
2.2013E+11	2	8		1	3	2	2	23	ZZ	111	908	29284	1	1
1.20134E+11	2	8		1	4	2	6	23	ZZ	111	2724	2929	1	1
1.20134E+11	2	8		1	4	2	1	23	ZZ	111	454	30480	1	1
1.20134E+11	2	8		1	4	2	5	23	ZZ	111	2270	30590	1	1
2.2013E+11	2	8		1	4	2	5	23	ZZ	111	2270	29284	1	1
2.2013E+11	2	8		1	4	2	5	23	ZZ	111	2270	2929	1	1
1.20134E+11	2	8		1	4	2	8	24	ZZ	111	3632	2929	1	1
2.2013E+11	2	8		1	4	2	21	23	ZZ	111	9534	2929	1	1

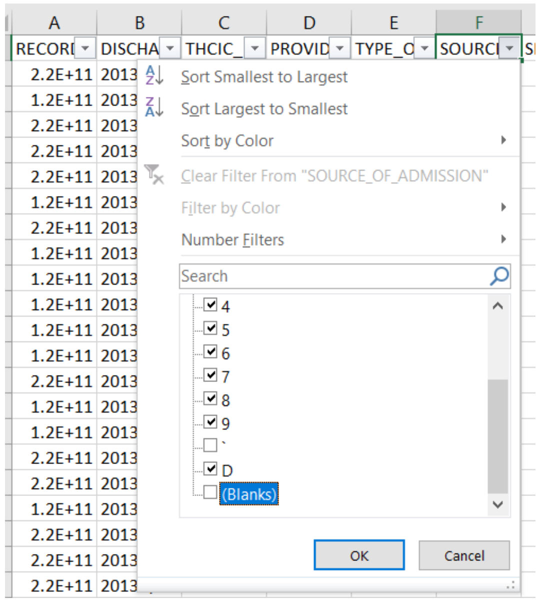
Once we are done with columns, we need to work on reducing the number of rows by identifying the data that we need. Here, we need the data for patients who are 'HIV and drug/alcohol use patients. As we analyse the column PAT_AGE we see that the values 22-26 for that variable specifies these patients. Hence, using this criterion we filter out the data and utilize it for further tasks. Hence after this step our data reduces from 719,371 rows to 52,146 rows. This is done as our classification is based on a subset of patients having a specific condition associated with them. If we include other data, it can lead to wrong predictions as well.



IMPLEMENTATION

1. Data Cleaning

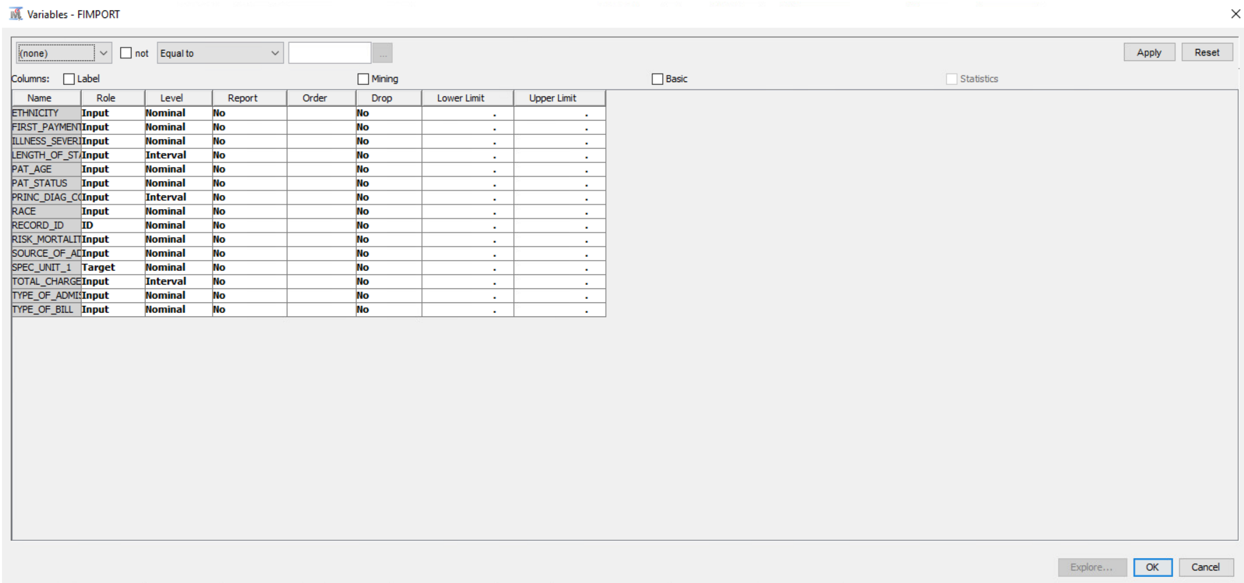
In this step we must identify the anomalies present in our data. For this we need to check each variable that is involved in our analysis. Hence, we observe that there are not many trash values in each column, but there are blanks and the symbol [‘] which need to be filtered out and removed.



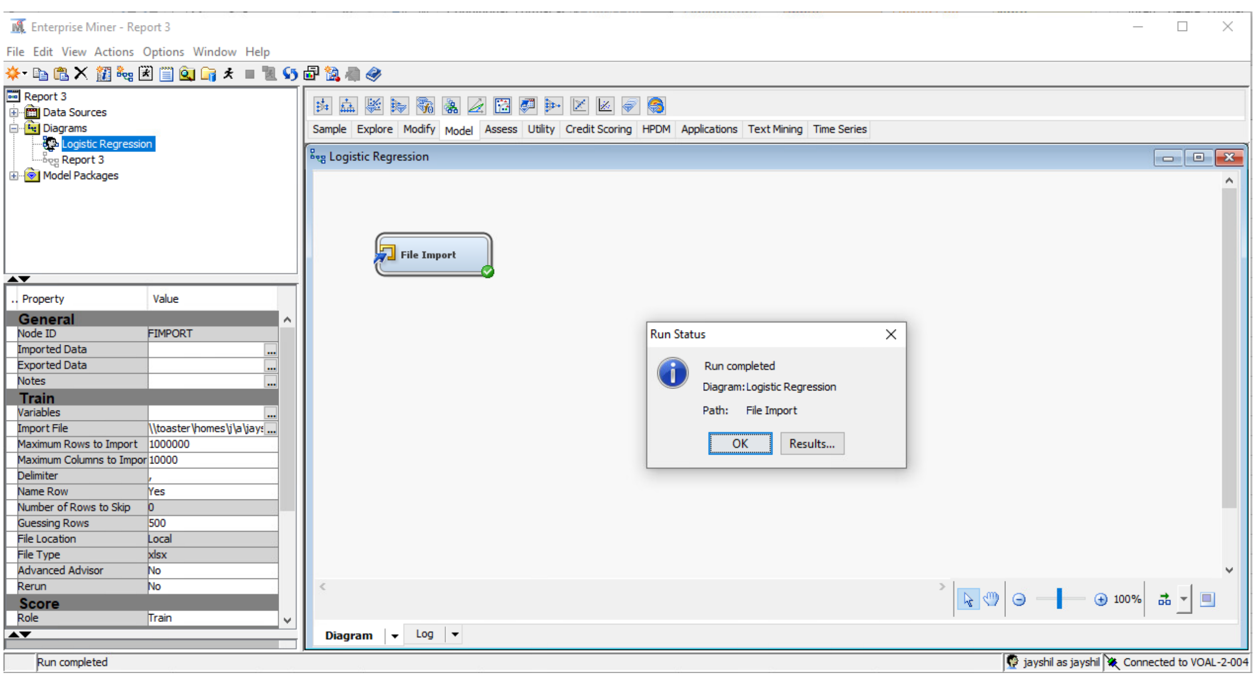
This step is needed because having blanks and symbols will cause hindrance in our classification as they add errors to our analysis. Also, null values specify nothing which means assuming them as zero also is not correct. After we are done with this step for all variables, our number of rows reduce from 52,146 rows to 32,529 rows.

2. *Data Import*

The primary step for running our analysis is to import the data in to SAS Enterprise Miner. In this case, we need to identify the variables that are predictors for our analysis and the variable that is the target variable. Since we have already identified those in our data exploration step, all we need to do is to set them as they are in this step using the 'Edit variables' option in the SAS Enterprise Miner from the 'File Import' node.



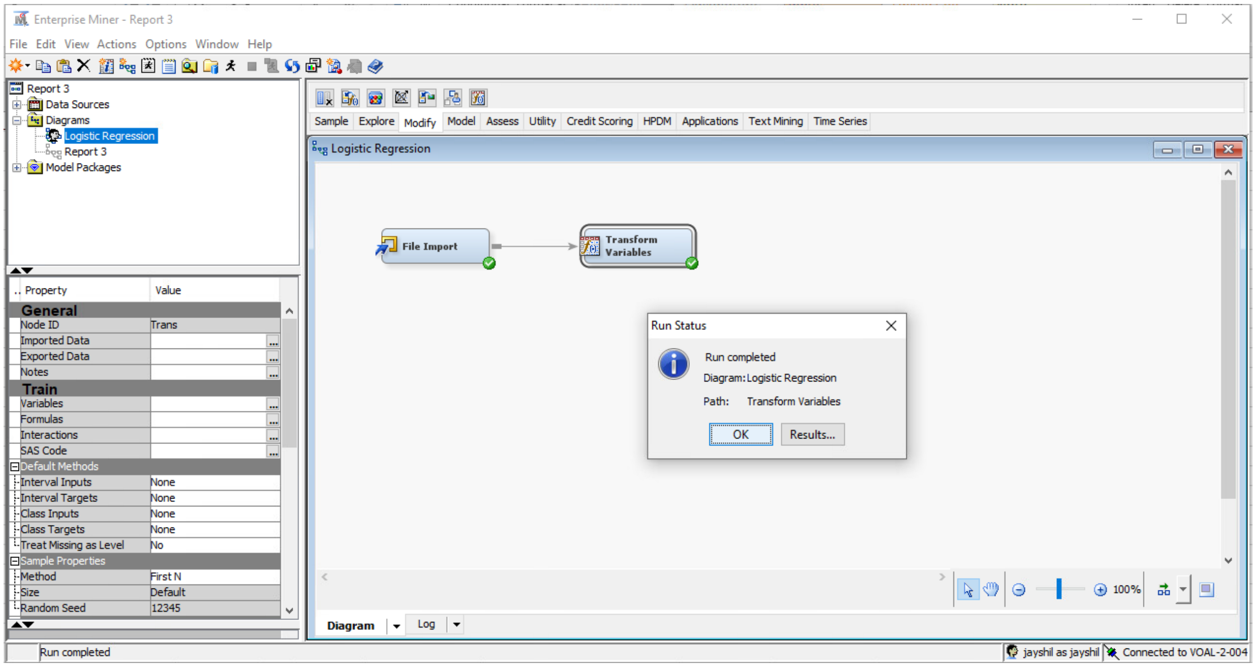
After the node is run successfully, we get the dialog box as shown below which indicates the file has been successfully imported. When we click on results, we can view the summary of the data that has been imported on to the tool.



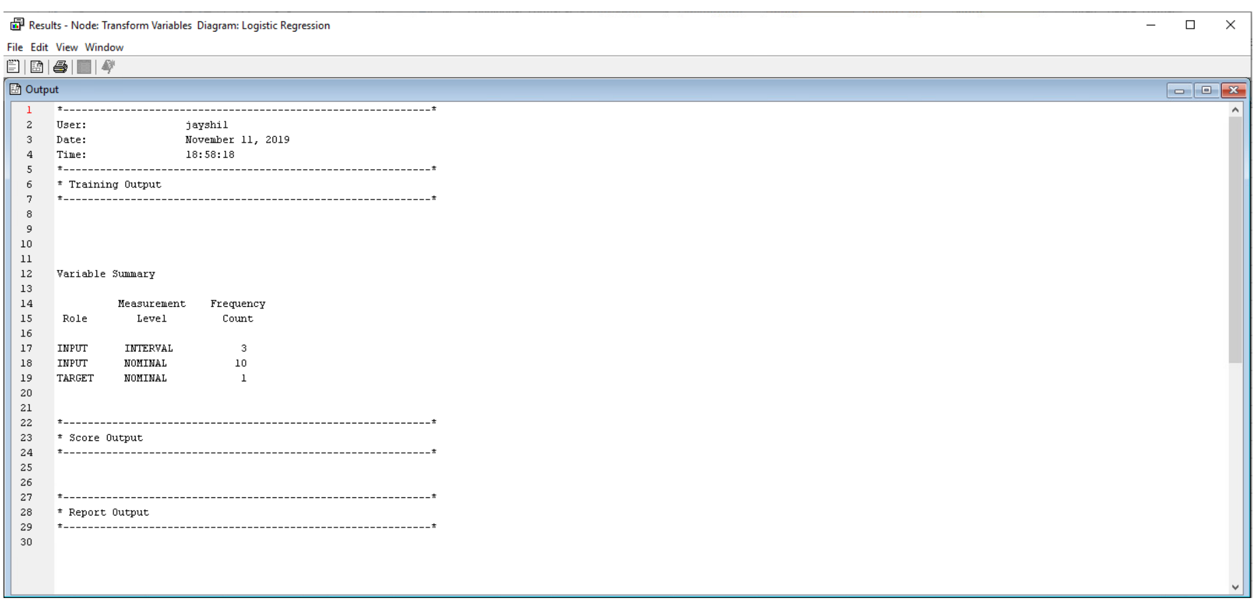
3. Fixing Problems with The Data

As we have seen above, we have already cleaned up the data by removing invalid data (nulls or symbol or any other aberration), after which we imported the files in the tool, now we need to transform it for our use. The data taken into consideration has many limitations on how it can be manipulated and assessed. This is mainly because the data type of each variables differs. The data therefore first needs to be checked if it is numerical or text, continuous, integer or categorical. This will help us realize what sort of operations need to be performed on them to transform them into a usable format. Say for example, one of the fields is categorical and has strings stored in them. These strings will cause an error if used directly in the logistic regression algorithm as it can only take numeric values. Thus, we need to transform these variables into a form which can be understood by the algorithm, which leads to the creation of dummy variables. These dummy variables are a way of bridging this gap between the data and the model.

Hence, in our data we see that other than 3 interval variables which continuous numeric values. These variables can stay as is since they will be interpreted by the algorithm in a normal manner. Other than these 3, we have all the other 10 variables as categorical. All these variables are necessary for our analysis since they can help us classify our target variable which is again a categorical field. Thus, these variables need to be manipulated in such a way that they can be useful as predictor variables. For this, we will add another node to our diagram which is the 'Transform variables' node. This node takes our clean data as an input and creates dummy variables for all the variables that are categorical. Creating of dummy variables creates n-1 columns for 1 variable which has n categories. This is a transformation of text value to numeric form which is identifiable by the algorithm. Hence, after we run this node, all the 10 variables will have dummy variables created for them. We can click on results to see the output of this node.



We can click on results to see the output of this node. Hence we see in the output below that 3 input variables identified as interval, 10 variables identified as categorical (nominal) and 1 target variable that is categorical (nominal).



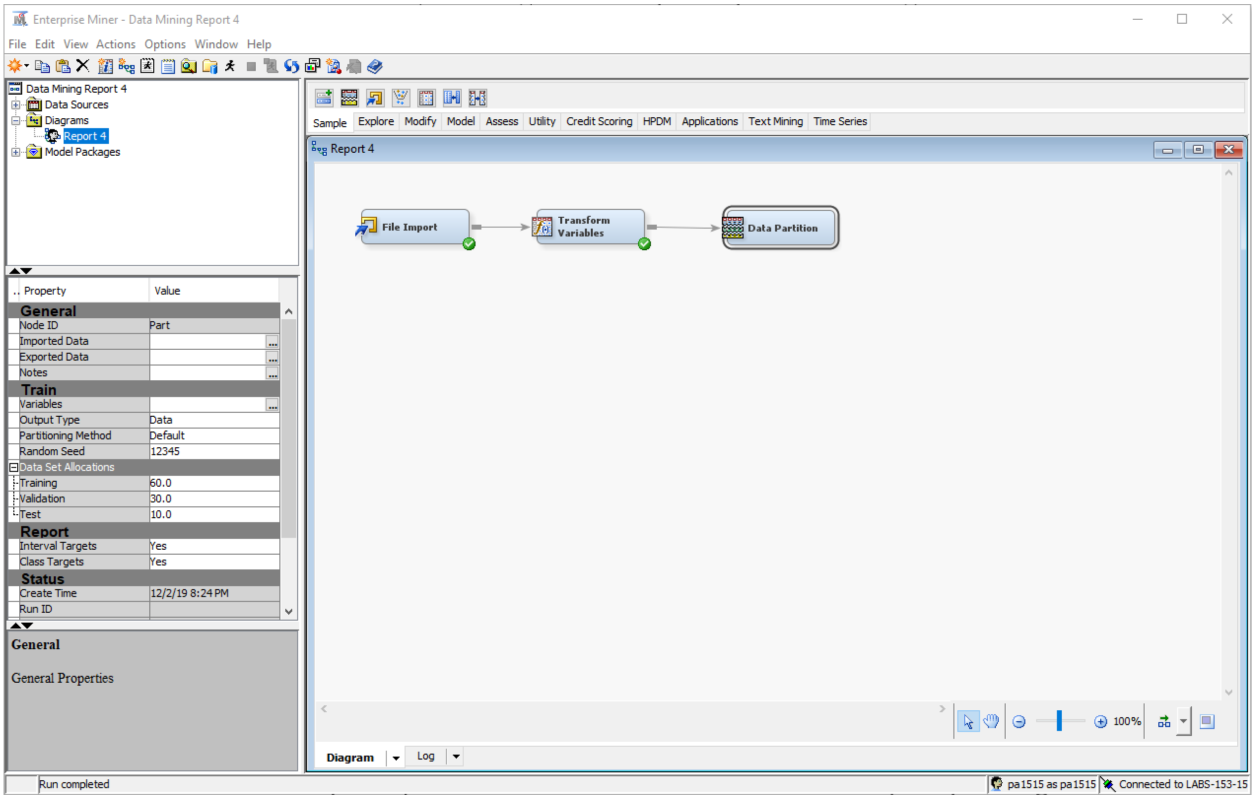

```
*------------------------------------------------------------*
User:                jayshil
Date:                November 11, 2019
Time:                18:58:18
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*




Variable Summary

          Measurement    Frequency
 Role        Level         Count

INPUT      INTERVAL          3
INPUT      NOMINAL          10
TARGET     NOMINAL           1


*------------------------------------------------------------*
* Score Output
*------------------------------------------------------------*


*------------------------------------------------------------*
* Report Output
*------------------------------------------------------------*
```

4. Creating A Model Set

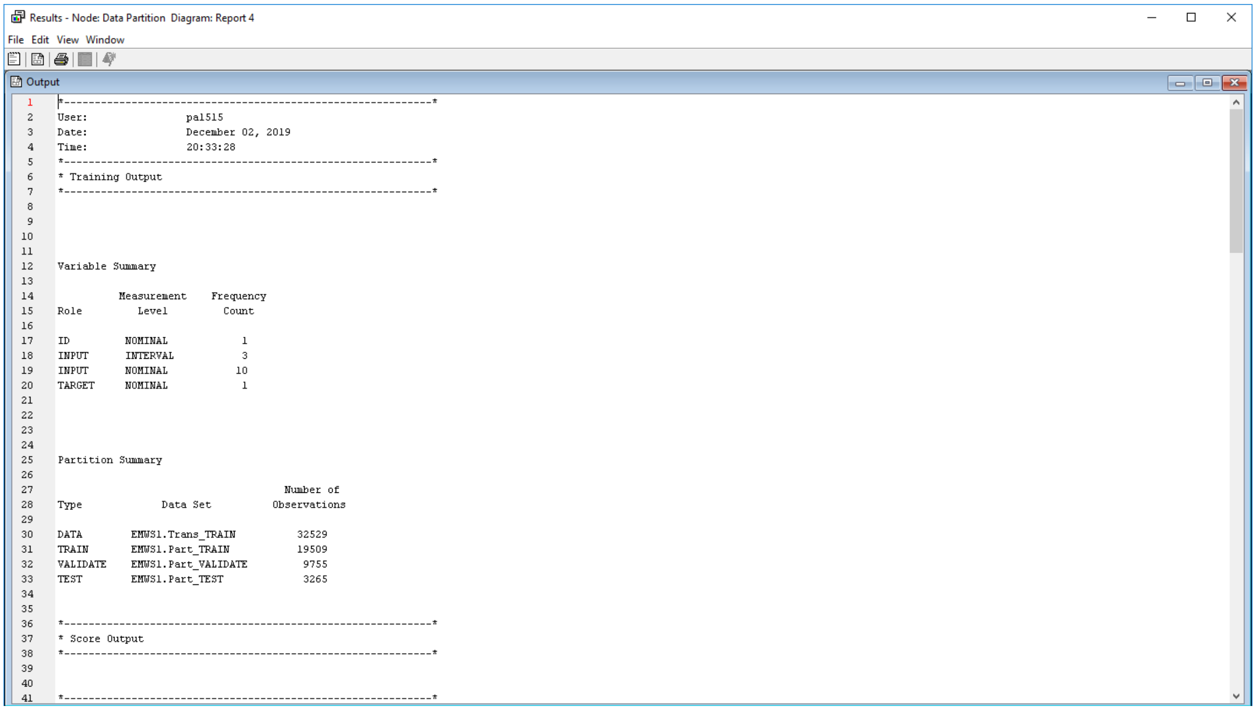
In the case of supervised algorithm, it means that the algorithm learns through the data and then apply the pattern it understood to the data for which we need to classify our outcomes. The data that we have for this contains 32,529 rows in the table. This data can be split into three parts where one part can be used by the algorithm to learn the pattern of the data, the second part can be used to validate this pattern that the algorithm learned and check if it gives us appropriate results that is with error rate which is tolerable for our use and the third part of the data can be used to test the algorithm. This third dataset is helpful in the case where there are multiple algorithms run and we need to select the best possible model and then run it on this test data set. When an algorithm learns from a training data set, and it applies this model to validation data set it might seem accurate, but this might not necessarily mean that it is correct, hence to learn the characteristic of a data set wholly we need a separate test data. Hence, the three parts of our data will be training data, validation data and test data.

Here in this case, we have divided our data into three parts where 60% of the data is training data, 30% of data is validation data and the last part of 10% of data is test data set on which the model is applied to verify if the algorithm is working accurately. Here what happens is that the algorithm will use the training data set and understand how the target variable varies in accordance with changes in predictor variables. Then it will apply this pattern on our validation data set and check how correctly it is able to classify the records, we can check error percentage for this. After this it again applies this pattern to a separate test data to confirm if the model is working properly.

Thus, below we see that after transforming the data, I've added a 'Data Partition' node, where I've also specified the 60:30:10 ratio for training, validation and test datasets. I run this node and my data partition are created.



I run this node and my data partition are created. We can check the output by clicking on results button. Thus, below we see the output where 19509 rows are assigned as training data, 9755 rows are assigned as validation data and 3265 rows are assigned as test data. Now data is ready to build a model using an algorithm.

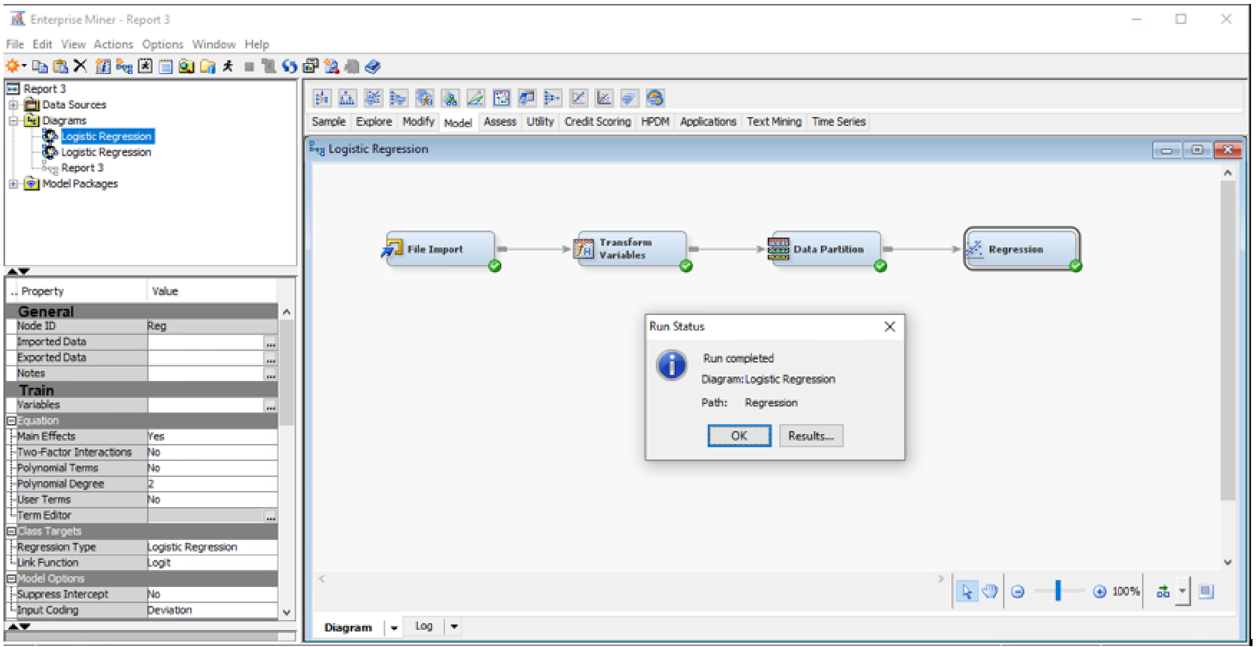

Results - Node: Data Partition Diagram: Report 4
File Edit View Window
Output

```
*------------------------------------------------------------*
User:                    pal515
Date:                    December 02, 2019
Time:                    20:33:28
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*




Variable Summary

          Measurement    Frequency
Role         Level         Count

ID         NOMINAL           1
INPUT      INTERVAL          3
INPUT      NOMINAL          10
TARGET     NOMINAL           1




Partition Summary

                                       Number of
Type              Data Set           Observations

DATA        EMWS1.Trans_TRAIN           32529
TRAIN       EMWS1.Part_TRAIN            19509
VALIDATE    EMWS1.Part_VALIDATE          9755
TEST        EMWS1.Part_TEST              3265


*------------------------------------------------------------*
* Score Output
*------------------------------------------------------------*


*------------------------------------------------------------*
```

5. Building the Model

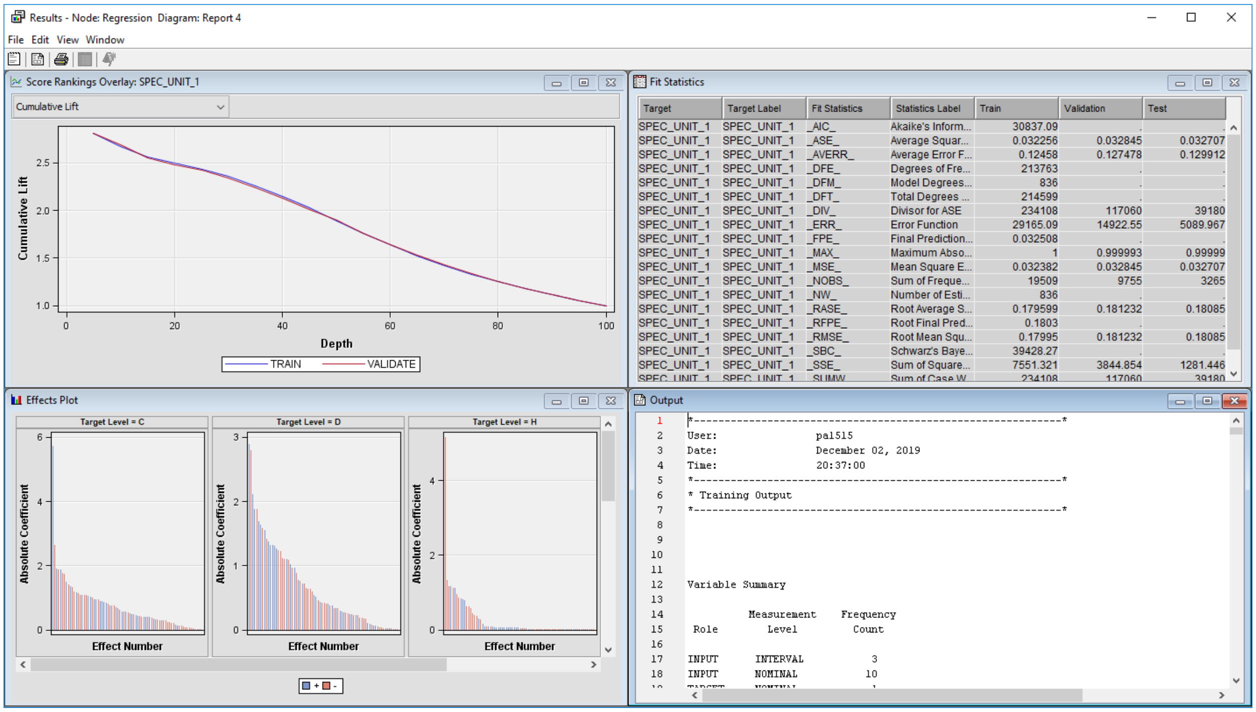
Here we have to try out different models that might suit our data. Hence, for this task I have decided to try out 3 models on the same data and then compare the results to check which one will work better for me. Good model doesn't really mean that the prediction has to be 100% accurate as this will mean overfitting where even the noise is considered as a signal but instead we need to compare it using the error in prediction which will give us an insight on how good the model is.

5.1 REGRESSION:

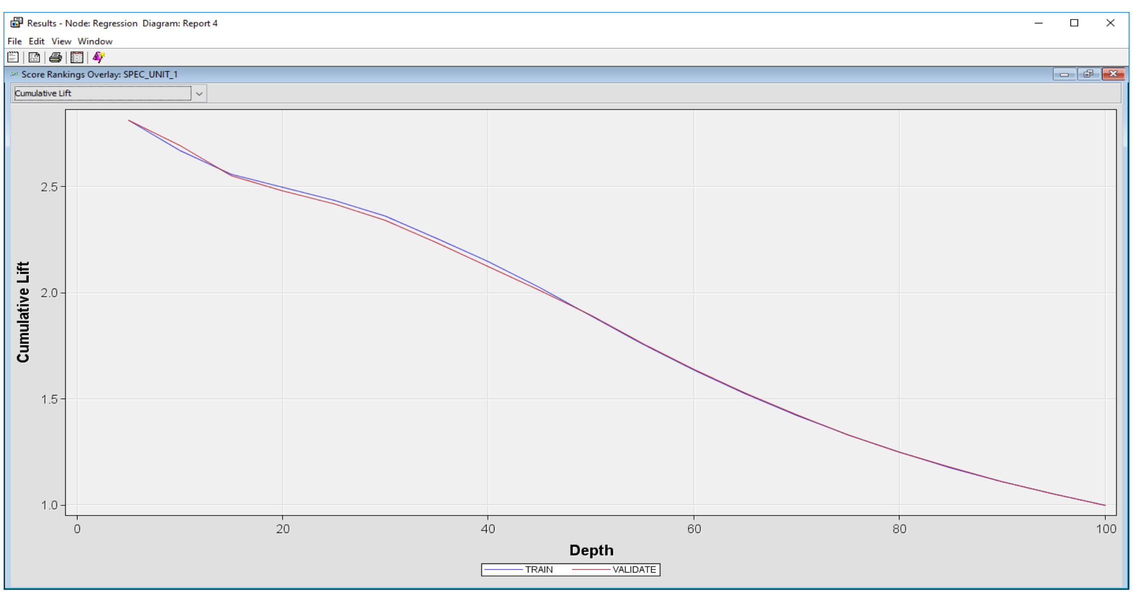


Assessing the Model

As we click on the results button, we will be able to see the result that are produced by this algorithm. Hence, below we see the output that has been generated.



Lift Chart: As per the below output the lift chart here shows that the training and validation data set show somewhat similar trend in terms of classification. There are places where there is error or deviation, but this seems to be tolerable.



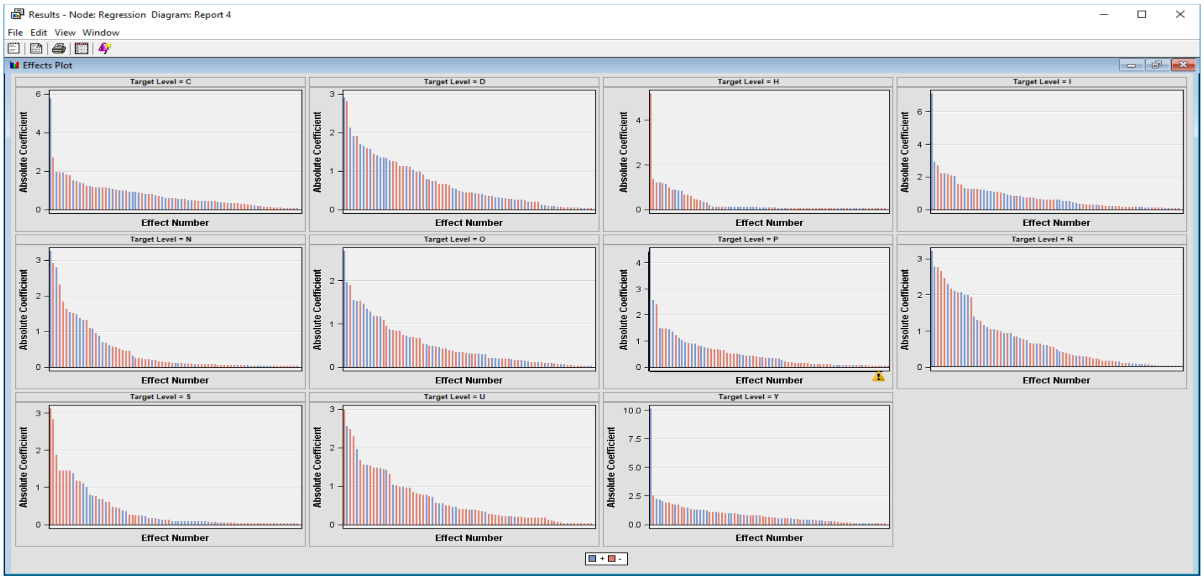
Fit Statistics: As we look at the fit statistics, we mainly check for the Root Mean Squared Error to check our outcome. Hence, we see that the value for training data is 0.17995, for validation data it is 0.181232 and for the test data it is 0.18085. Thus, we can say that the model has performed consistently with all the datasets indicating that our result has no issues in terms of data pattern being analysed. Also, the low value of the errors indicates that even the classification has been accurate.

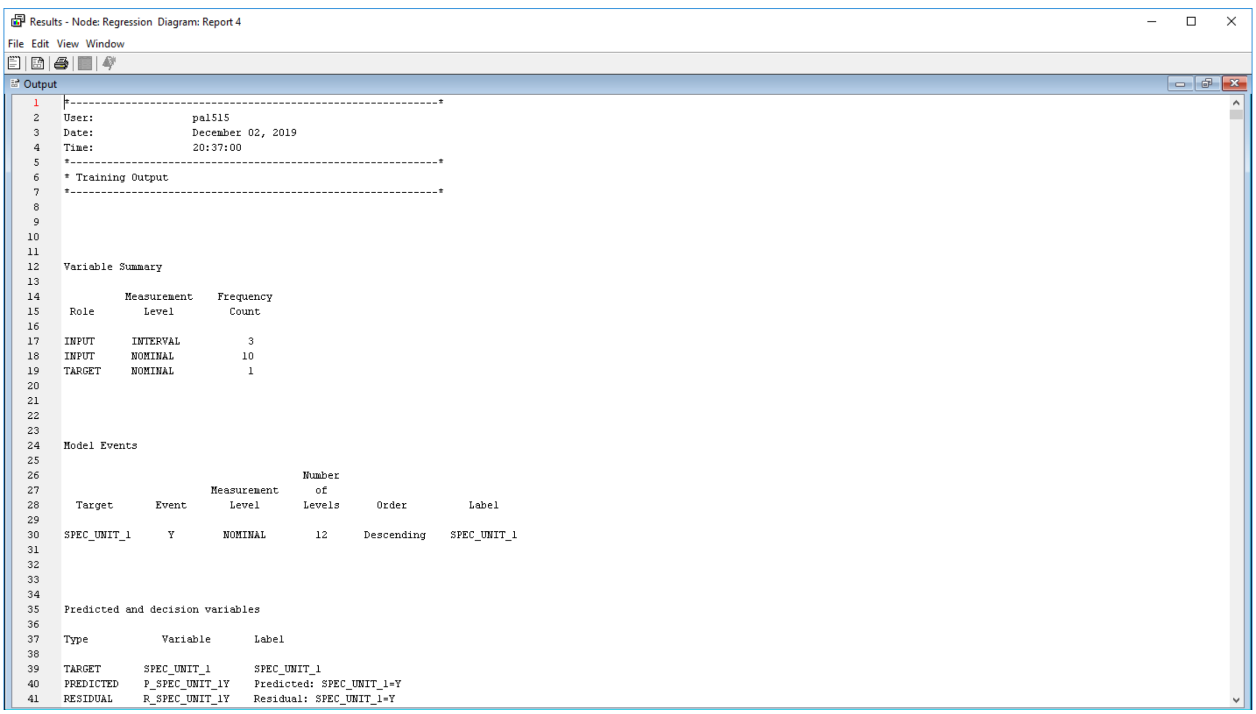
Results - Node: Regression Diagram: Report 4

File Edit View Window

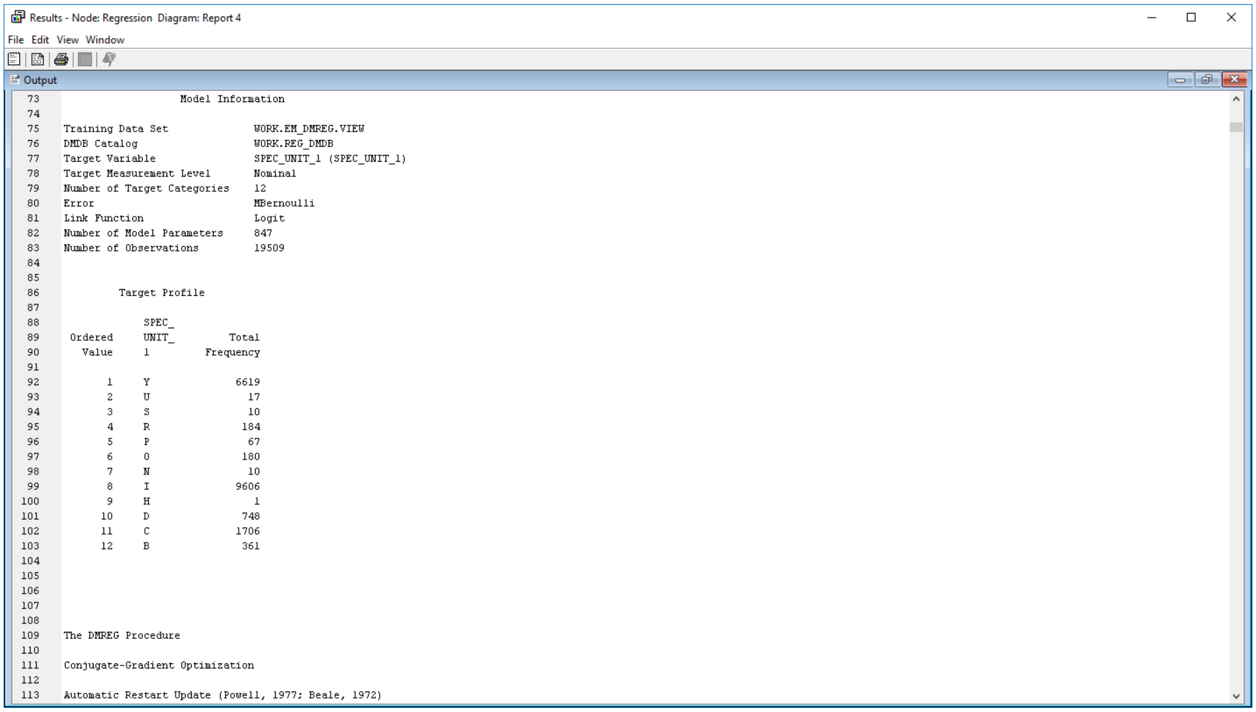
Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
SPEC_UNIT_1	SPEC_UNIT_1	_AIC_	Akaike's Information Criterion	30837.09	.	.
SPEC_UNIT_1	SPEC_UNIT_1	_ASE_	Average Squared Error	0.032256	0.032845	0.032707
SPEC_UNIT_1	SPEC_UNIT_1	_AVERR_	Average Error Function	0.12458	0.127478	0.129912
SPEC_UNIT_1	SPEC_UNIT_1	_DFE_	Degrees of Freedom for Error	213763	.	.
SPEC_UNIT_1	SPEC_UNIT_1	_DFM_	Model Degrees of Freedom	836	.	.
SPEC_UNIT_1	SPEC_UNIT_1	_DFT_	Total Degrees of Freedom	214599	.	.
SPEC_UNIT_1	SPEC_UNIT_1	_DIV_	Divisor for ASE	234108	117060	39180
SPEC_UNIT_1	SPEC_UNIT_1	_ERR_	Error Function	29165.09	14922.55	5089.967
SPEC_UNIT_1	SPEC_UNIT_1	_FPE_	Final Prediction Error	0.032508	.	.
SPEC_UNIT_1	SPEC_UNIT_1	_MAX_	Maximum Absolute Error	1	0.999993	0.99999
SPEC_UNIT_1	SPEC_UNIT_1	_MSE_	Mean Square Error	0.032382	0.032845	0.032707
SPEC_UNIT_1	SPEC_UNIT_1	_NOBS_	Sum of Frequencies	19509	9755	3265
SPEC_UNIT_1	SPEC_UNIT_1	_NW_	Number of Estimate Weights	836	.	.
SPEC_UNIT_1	SPEC_UNIT_1	_RASE_	Root Average Sum of Squares	0.179599	0.181232	0.18085
SPEC_UNIT_1	SPEC_UNIT_1	_RFPE_	Root Final Prediction Error	0.1803	.	.
SPEC_UNIT_1	SPEC_UNIT_1	_RMSE_	Root Mean Squared Error	0.17995	0.181232	0.18085
SPEC_UNIT_1	SPEC_UNIT_1	_SBC_	Schwarz's Bayesian Criterion	39428.27	.	.
SPEC_UNIT_1	SPEC_UNIT_1	_SSE_	Sum of Squared Errors	7551.321	3844.854	1281.446
SPEC_UNIT_1	SPEC_UNIT_1	_SUMW_	Sum of Case Weights Times Freq	234108	117060	39180
SPEC_UNIT_1	SPEC_UNIT_1	_MISC_	Misclassification Rate	0.262033	0.266017	0.261868

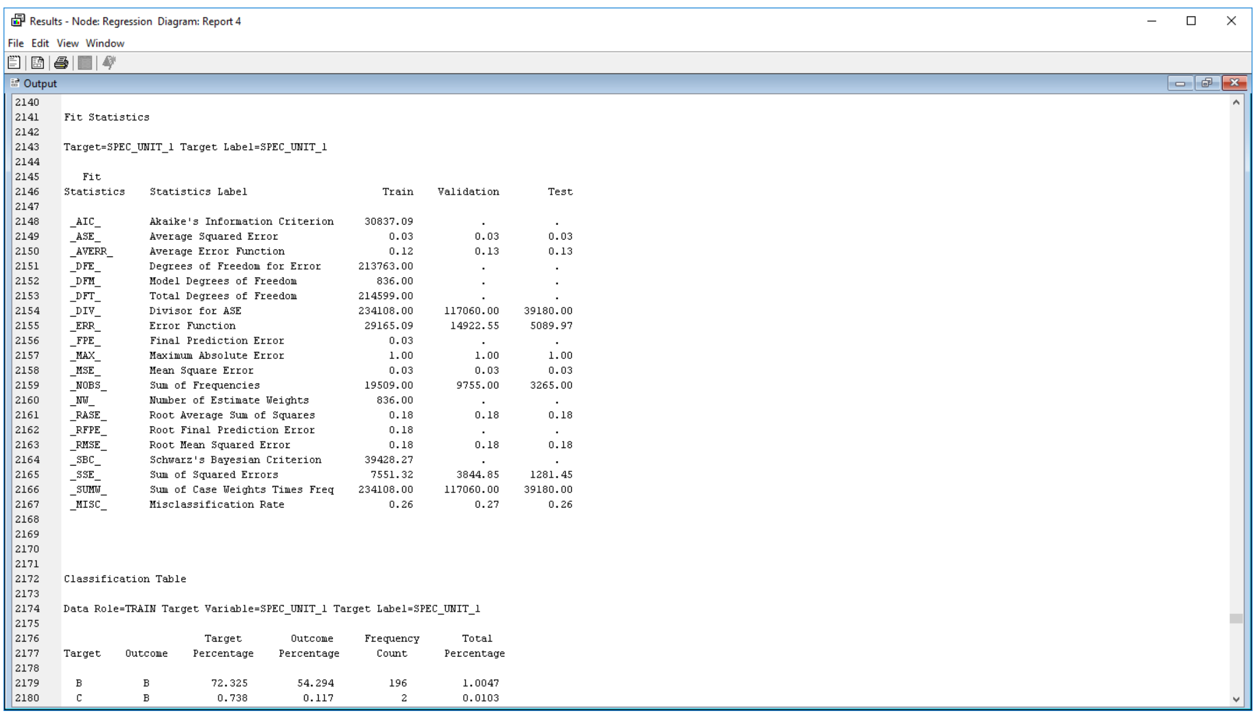





```
*------------------------------------------------------------*
User:               pa1515
Date:               December 02, 2019
Time:               20:37:00
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*




Variable Summary

          Measurement    Frequency
 Role        Level         Count

INPUT      INTERVAL          3
INPUT      NOMINAL          10
TARGET     NOMINAL           1




Model Events

                                                  Number
                              Measurement           of
  Target        Event            Level            Levels        Order          Label

SPEC_UNIT_1       Y             NOMINAL             12       Descending    SPEC_UNIT_1




Predicted and decision variables

Type            Variable        Label

TARGET          SPEC_UNIT_1     SPEC_UNIT_1
PREDICTED       P_SPEC_UNIT_1Y  Predicted: SPEC_UNIT_1=Y
RESIDUAL        R_SPEC_UNIT_1Y  Residual: SPEC_UNIT_1=Y
```




```
                    Model Information

Training Data Set             WORK.EM_DMREG.VIEW
DMDB Catalog                  WORK.REG_DMDB
Target Variable               SPEC_UNIT_1 (SPEC_UNIT_1)
Target Measurement Level      Nominal
Number of Target Categories   12
Error                         MBernoulli
Link Function                 Logit
Number of Model Parameters    847
Number of Observations        19509


         Target Profile

             SPEC_
 Ordered     UNIT_        Total
   Value     1        Frequency

       1     Y             6619
       2     U               17
       3     S               10
       4     R              184
       5     P               67
       6     O              180
       7     N               10
       8     I             9606
       9     H                1
      10     D              748
      11     C             1706
      12     B              361





The DMREG Procedure

Conjugate-Gradient Optimization

Automatic Restart Update (Powell, 1977; Beale, 1972)
```




```

Fit Statistics

Target=SPEC_UNIT_1 Target Label=SPEC_UNIT_1

   Fit
Statistics    Statistics Label                      Train    Validation        Test

 _AIC_        Akaike's Information Criterion     30837.09             .           .
 _ASE_        Average Squared Error                  0.03          0.03        0.03
 _AVERR_      Average Error Function                 0.12          0.13        0.13
 _DFE_        Degrees of Freedom for Error      213763.00             .           .
 _DFM_        Model Degrees of Freedom             836.00             .           .
 _DFT_        Total Degrees of Freedom          214599.00             .           .
 _DIV_        Divisor for ASE                   234108.00     117060.00    39180.00
 _ERR_        Error Function                     29165.09      14922.55     5089.97
 _FPE_        Final Prediction Error                 0.03             .           .
 _MAX_        Maximum Absolute Error                 1.00          1.00        1.00
 _MSE_        Mean Square Error                      0.03          0.03        0.03
 _NOBS_       Sum of Frequencies                 19509.00       9755.00     3265.00
 _NW_         Number of Estimate Weights           836.00             .           .
 _RASE_       Root Average Sum of Squares            0.18          0.18        0.18
 _RFPE_       Root Final Prediction Error            0.18             .           .
 _RMSE_       Root Mean Squared Error                0.18          0.18        0.18
 _SBC_        Schwarz's Bayesian Criterion       39428.27             .           .
 _SSE_        Sum of Squared Errors               7551.32       3844.85     1281.45
 _SUMW_       Sum of Case Weights Times Freq    234108.00     117060.00    39180.00
 _MISC_       Misclassification Rate                 0.26          0.27        0.26




Classification Table

Data Role=TRAIN Target Variable=SPEC_UNIT_1 Target Label=SPEC_UNIT_1

                        Target       Outcome    Frequency       Total
Target     Outcome    Percentage    Percentage      Count    Percentage

  B           B         72.325        54.294        196        1.0047
  C           B          0.738         0.117          2        0.0103
```

Results - Node: Regression Diagram: Report 4

File Edit View Window

Output

Classification Table

Data Role=TRAIN Target Variable=SPEC_UNIT_1 Target Label=SPEC_UNIT_1

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
B	B	72.325	54.294	196	1.0047
C	B	0.738	0.117	2	0.0103
D	B	1.476	0.535	4	0.0205
I	B	19.557	0.552	53	0.2717
O	B	1.107	1.667	3	0.0154
P	B	0.369	1.493	1	0.0051
U	B	0.369	5.882	1	0.0051
Y	B	4.059	0.166	11	0.0564
C	C	50.000	0.234	4	0.0205
I	C	50.000	0.042	4	0.0205
B	D	0.704	0.277	1	0.0051
C	D	1.408	0.117	2	0.0103
D	D	49.296	9.358	70	0.3588
I	D	9.859	0.146	14	0.0718
R	D	7.746	5.978	11	0.0564
Y	D	30.986	0.665	44	0.2255
H	H	100.000	100.000	1	0.0051
B	I	0.763	23.546	85	0.4357
C	I	13.851	90.445	1543	7.9092
D	I	0.682	10.160	76	0.3896
I	I	75.296	87.320	8388	42.9955
N	I	0.054	60.000	6	0.0308
O	I	1.293	80.000	144	0.7381
P	I	0.153	25.373	17	0.0871
R	I	0.135	8.152	15	0.0769
U	I	0.018	11.765	2	0.0103
Y	I	7.756	13.053	864	4.4287
I	N	25.000	0.010	1	0.0051
N	N	75.000	30.000	3	0.0154
O	O	100.000	0.556	1	0.0051
I	P	33.333	0.052	5	0.0256
P	P	53.333	11.940	8	0.0410

Results - Node: Regression Diagram: Report 4

File Edit View Window

Output

Data Role=VALIDATE Target Variable=SPEC_UNIT_1 Target Label=SPEC_UNIT_1

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
B	B	74.783	47.514	86	0.8816
C	B	0.870	0.117	1	0.0103
D	B	0.870	0.268	1	0.0103
I	B	17.391	0.416	20	0.2050
P	B	0.870	2.941	1	0.0103
Y	B	5.217	0.181	6	0.0615
D	C	25.000	0.268	1	0.0103
I	C	50.000	0.042	2	0.0205
U	C	25.000	14.286	1	0.0103
B	D	2.703	1.105	2	0.0205
C	D	1.351	0.117	1	0.0103
D	D	48.649	9.651	36	0.3690
I	D	8.108	0.125	6	0.0615
R	D	6.757	5.376	5	0.0513
Y	D	32.432	0.725	24	0.2460
I	H	100.000	0.021	1	0.0103
B	I	0.773	23.757	43	0.4408
C	I	13.736	89.566	764	7.8319
D	I	0.629	9.383	35	0.3588
H	I	0.018	100.000	1	0.0103
I	I	75.279	87.193	4187	42.9216
N	I	0.036	33.333	2	0.0205
O	I	1.348	82.418	75	0.7688
P	I	0.216	35.294	12	0.1230
R	I	0.144	8.602	8	0.0820
U	I	0.072	57.143	4	0.0410
Y	I	7.749	13.021	431	4.4182
I	N	25.000	0.021	1	0.0103
N	N	75.000	50.000	3	0.0308
I	P	45.455	0.104	5	0.0513
P	P	27.273	8.824	3	0.0308
R	P	9.091	1.075	1	0.0103
Y	P	18.182	0.060	2	0.0205
C	R	9.091	0.117	1	0.0103
D	R	9.091	0.268	1	0.0103
I	R	9.091	0.021	1	0.0103

Classification Table: As we analyse the event classification table, we see that for the training data the false negative is 927, true negative is 10708, false positive is 2182 and true positive is 5692.

Now, for validation data these numbers are false negative being 469, true negative being 5317, false positive being 1128 and true positive being 2841.

Results - Node: Regression Diagram: Report 4

File Edit View Window

Output

Event Classification Table

Data Role=TRAIN Target=SPEC_UNIT_1 Target Label=SPEC_UNIT_1

False Negative	True Negative	False Positive	True Positive
927	10708	2182	5692

Data Role=VALIDATE Target=SPEC_UNIT_1 Target Label=SPEC_UNIT_1

False Negative	True Negative	False Positive	True Positive
469	5317	1128	2841

Assessment Score Rankings

Data Role=TRAIN Target Variable=SPEC_UNIT_1 Target Label=SPEC_UNIT_1

Depth	Gain	Lift	Cumulative Lift	% Response	Cumulative % Response	Number of Observations	Mean Posterior Probability
5	181.153	2.81153	2.81153	95.3893	95.3893	976	0.94526
10	167.096	2.53025	2.67096	85.8462	90.6202	975	0.86307
15	155.672	2.32834	2.55672	78.9959	86.7441	976	0.78809
20	149.874	2.32469	2.49874	78.8718	84.7770	975	0.72768
25	143.564	2.18339	2.43564	74.0779	82.6363	976	0.67066
30	136.277	1.99820	2.36277	67.7949	80.1640	975	0.60846
35	125.599	1.61565	2.25599	54.8156	76.5412	976	0.53494
40	114.787	1.39058	2.14787	47.1795	72.8729	975	0.45671
45	102.493	1.04187	2.02493	35.3484	68.7016	976	0.37902
50	89.173	0.69227	1.89173	23.4872	64.1825	975	0.29795
55	75.829	0.42322	1.75829	14.3590	59.6552	975	0.21599
60	63.586	0.28991	1.63586	9.8361	55.5015	976	0.13901

Results - Node: Regression Diagram: Report 4

File Edit View Window

Output

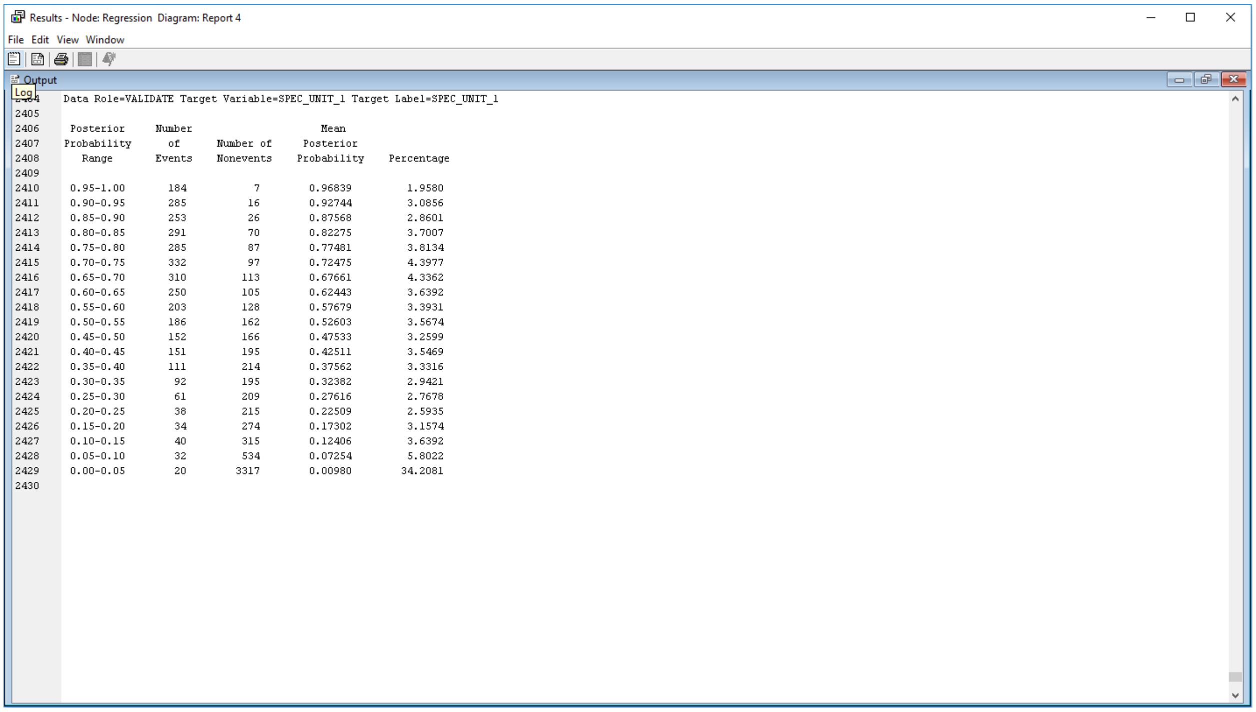
Assessment Score Distribution

Data Role=TRAIN Target Variable=SPEC_UNIT_1 Target Label=SPEC_UNIT_1

Posterior Probability Range	Number of Events	Number of Nonevents	Mean Posterior Probability	Percentage
0.95-1.00	409	10	0.96885	2.1477
0.90-0.95	584	41	0.92484	3.2036
0.85-0.90	487	77	0.87465	2.8910
0.80-0.85	529	111	0.82427	3.2805
0.75-0.80	633	178	0.77433	4.1571
0.70-0.75	645	168	0.72445	4.1673
0.65-0.70	647	223	0.67552	4.4595
0.60-0.65	508	248	0.62516	3.8751
0.55-0.60	427	229	0.57550	3.3626
0.50-0.55	331	287	0.52533	3.1678
0.45-0.50	299	329	0.47541	3.2190
0.40-0.45	273	361	0.42419	3.2498
0.35-0.40	227	398	0.37501	3.2036
0.30-0.35	164	412	0.32482	2.9525
0.25-0.30	117	494	0.27536	3.1319
0.20-0.25	90	482	0.22506	2.9320
0.15-0.20	64	583	0.17494	3.3164
0.10-0.15	69	643	0.12426	3.6496
0.05-0.10	62	969	0.07308	5.2847
0.00-0.05	54	6647	0.01002	34.3482

Data Role=VALIDATE Target Variable=SPEC_UNIT_1 Target Label=SPEC_UNIT_1

Posterior Probability Range	Number of Events	Number of Nonevents	Mean Posterior Probability	Percentage
0.95-1.00	184	7	0.96839	1.9580
0.90-0.95	285	16	0.92744	3.0856


Results - Node: Regression Diagram: Report 4

File Edit View Window

Output

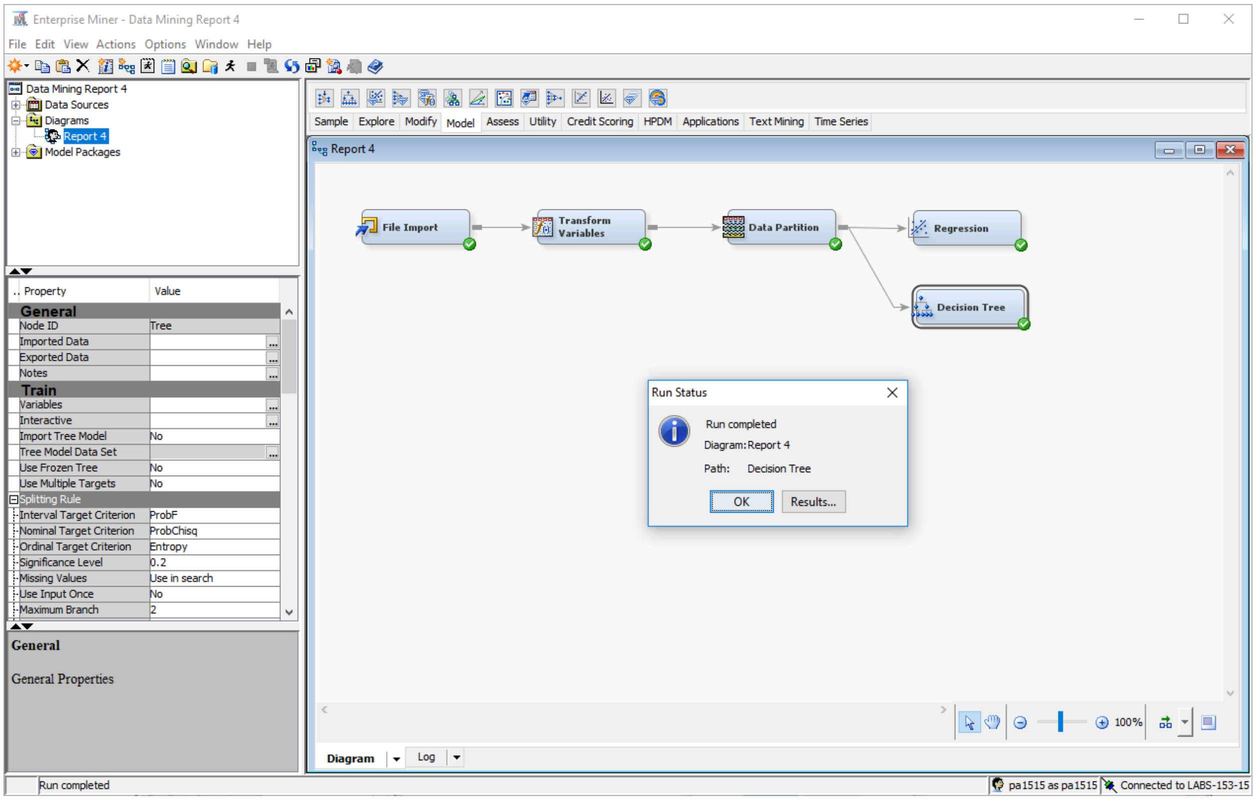
Log

Data Role=VALIDATE Target Variable=SPEC_UNIT_1 Target Label=SPEC_UNIT_1

Posterior Probability Range	Number of Events	Number of Nonevents	Mean Posterior Probability	Percentage
0.95-1.00	184	7	0.96839	1.9580
0.90-0.95	285	16	0.92744	3.0856
0.85-0.90	253	26	0.87568	2.8601
0.80-0.85	291	70	0.82275	3.7007
0.75-0.80	285	87	0.77481	3.8134
0.70-0.75	332	97	0.72475	4.3977
0.65-0.70	310	113	0.67661	4.3362
0.60-0.65	250	105	0.62443	3.6392
0.55-0.60	203	128	0.57679	3.3931
0.50-0.55	186	162	0.52603	3.5674
0.45-0.50	152	166	0.47533	3.2599
0.40-0.45	151	195	0.42511	3.5469
0.35-0.40	111	214	0.37562	3.3316
0.30-0.35	92	195	0.32382	2.9421
0.25-0.30	61	209	0.27616	2.7678
0.20-0.25	38	215	0.22509	2.5935
0.15-0.20	34	274	0.17302	3.1574
0.10-0.15	40	315	0.12406	3.6392
0.05-0.10	32	534	0.07254	5.8022
0.00-0.05	20	3317	0.00980	34.2081

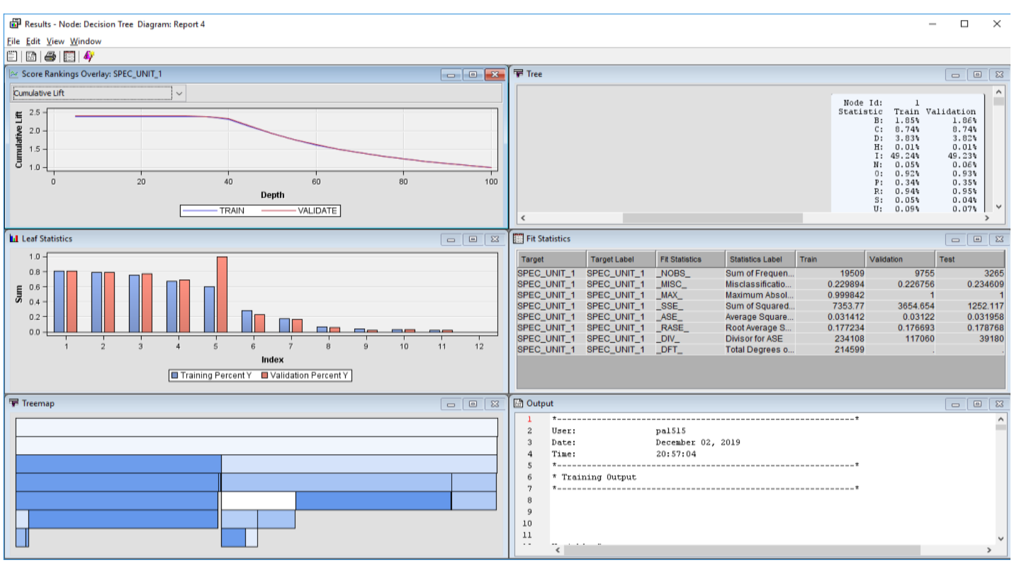
5.2 Decision Tree:

This method simplifies the breakdown of the data in the form a tree structure which can be used or building classification models. It can be used for categorical and numerical predictors. It includes decision nodes and leaf nodes. In case of numerical data, if the value mentioned in the node. From the training data it understands the nuances of the data and predicts the classes or values for the validation dataset.

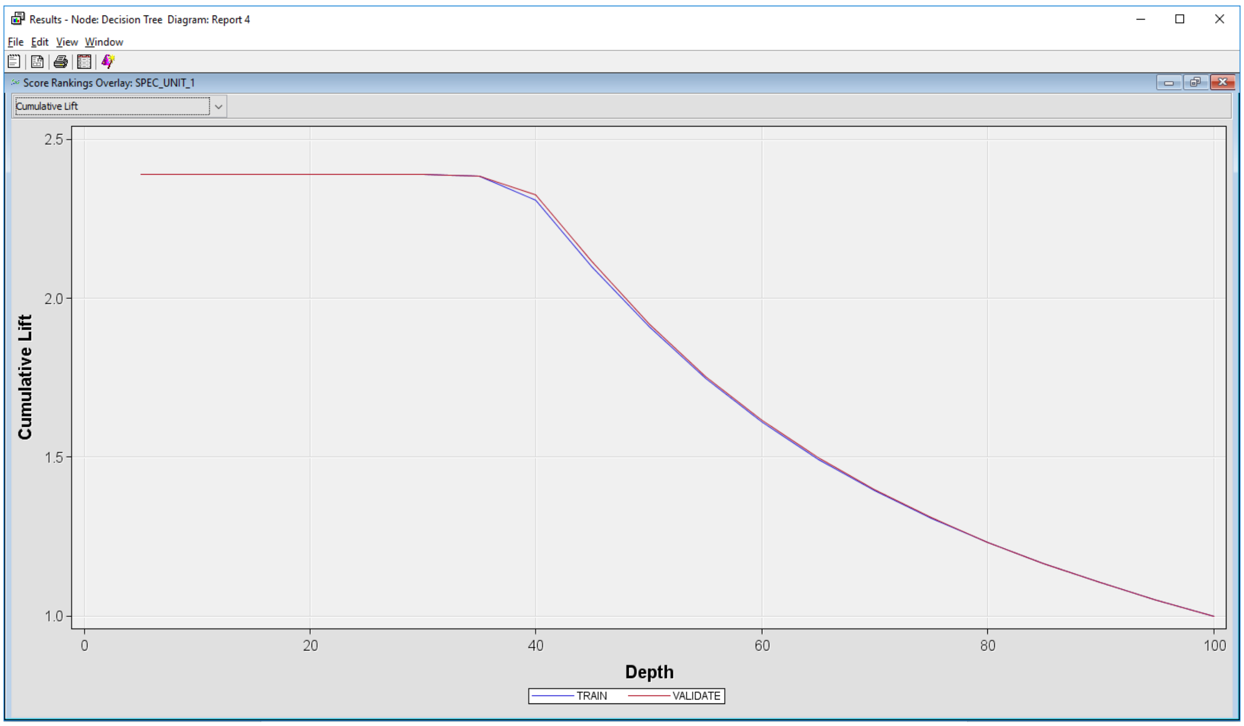


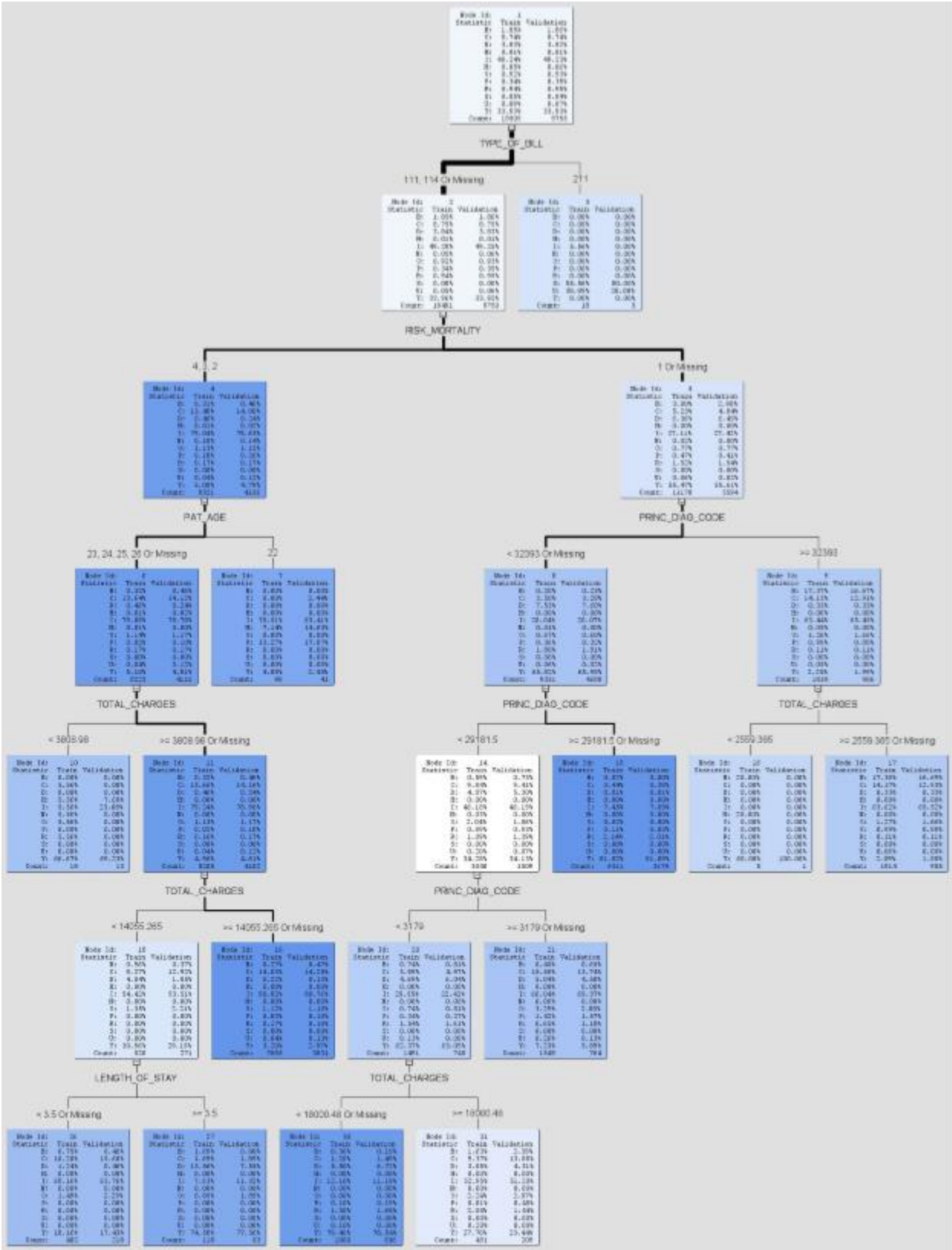
Assessing The Model

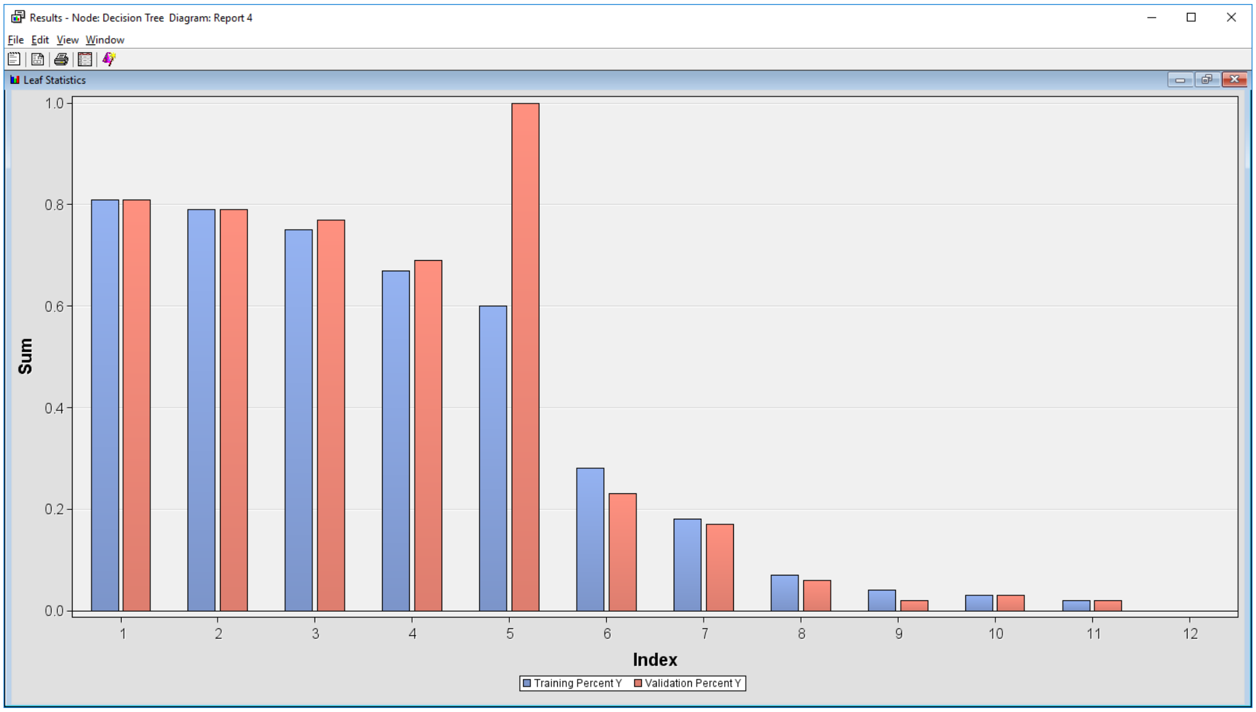
As we click on the results button, we will be able to see the result that are produced by this algorithm. Hence, below we see the output that has been generated.



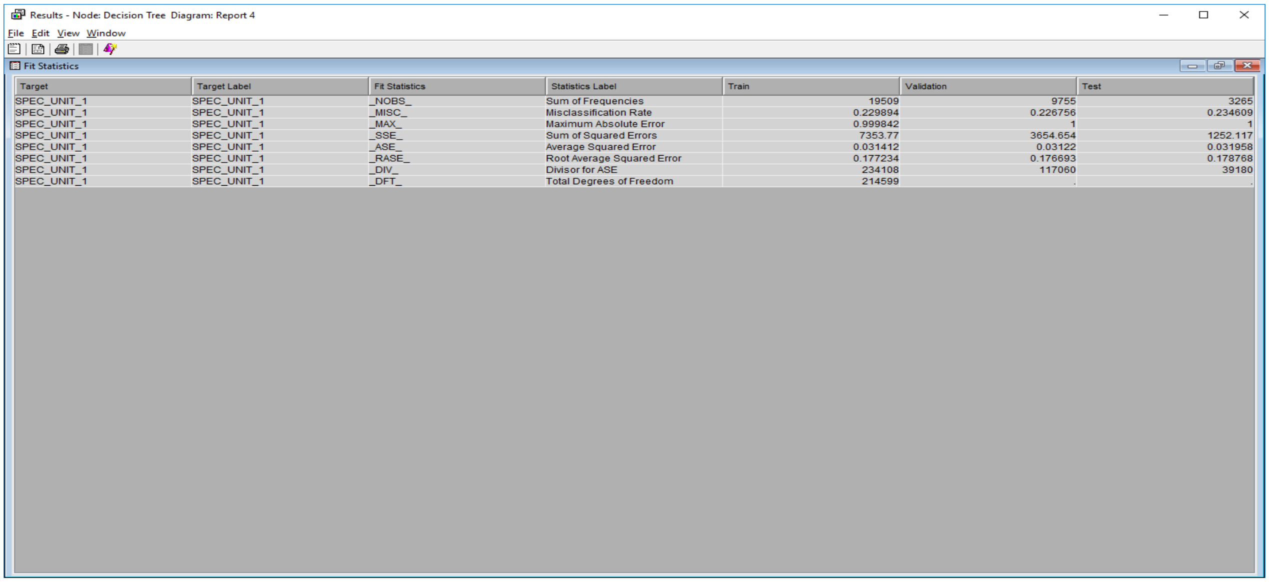
Lift Chart: As per the below output the lift chart here shows that the training and validation data set show somewhat similar trend in terms of classification. There are places where there is error or deviation, but this seems to be tolerable.



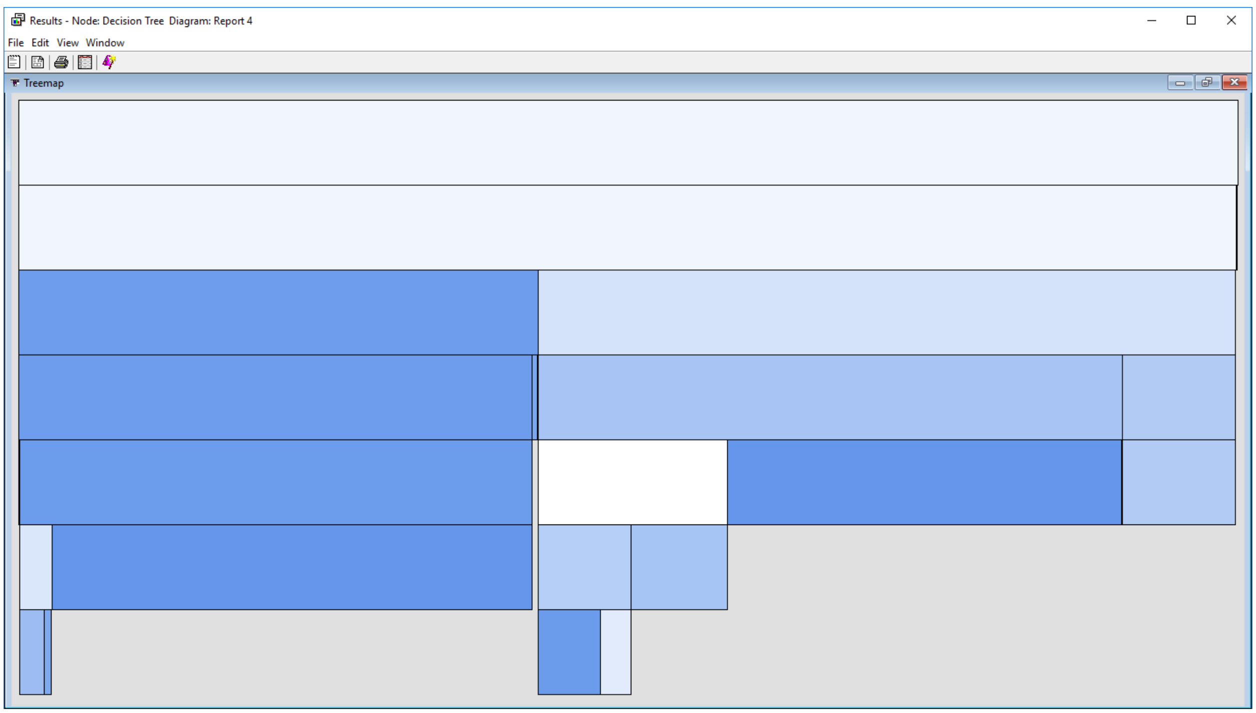

TYPE_OF_BILL
111, 114 Or Missing
211
RISK_MORTALITY
4, 3, 2
1 Or Missing
PAT_AGE
PRINC_DIAG_CODE
22
TOTAL_CHARGES
PRINC_DIAG_CODE
TOTAL_CHARGES
TOTAL_CHARGES
PRINC_DIAG_CODE
LENGTH_OF_STAY
TOTAL_CHARGES



Fit Statistics: As we take a look at the fit statistics, we mainly check for the Root Average Squared Error to check our outcome. Hence, we see that the value for training data is 0.177234, for validation data it is 0.176693 and for the test data it is 0.178768. Thus we can say that the model has performed fairly consistently with all the datasets indicating that our result has no issues in terms of data pattern being analyzed. Also, the low value of the errors indicate that even the classification has been fairly accurate.



Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
SPEC_UNIT_1	SPEC_UNIT_1	_NOBS_	Sum of Frequencies	19509	9755	3265
SPEC_UNIT_1	SPEC_UNIT_1	_MISC_	Misclassification Rate	0.229894	0.226756	0.234609
SPEC_UNIT_1	SPEC_UNIT_1	_MAX_	Maximum Absolute Error	0.999842	1	1
SPEC_UNIT_1	SPEC_UNIT_1	_SSE_	Sum of Squared Errors	7353.77	3654.654	1252.117
SPEC_UNIT_1	SPEC_UNIT_1	_ASE_	Average Squared Error	0.031412	0.03122	0.031958
SPEC_UNIT_1	SPEC_UNIT_1	_RASE_	Root Average Squared Error	0.177234	0.176693	0.178768
SPEC_UNIT_1	SPEC_UNIT_1	_DIV_	Divisor for ASE	234108	117060	39180
SPEC_UNIT_1	SPEC_UNIT_1	_DFT_	Total Degrees of Freedom	214599	.	.



Results - Node: Decision Tree Diagram: Report 4

File Edit View Window

Output

```
*------------------------------------------------------------*
User:                pal515
Date:                December 02, 2019
Time:                20:57:04
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*




Variable Summary

          Measurement    Frequency
Role         Level         Count

ID         INTERVAL          1
ID         NOMINAL           1
INPUT      INTERVAL          3
INPUT      NOMINAL          10
TARGET     NOMINAL           1




Model Events

                                          Number
                           Measurement      of
  Target       Event          Level       Levels      Order          Label

SPEC_UNIT_1      Y          NOMINAL         12      Descending    SPEC_UNIT_1




Predicted and decision variables

Type            Variable        Label

TARGET       SPEC_UNIT_1       SPEC_UNIT_1
```

Results - Node: Decision Tree Diagram: Report 4

File Edit View Window

Output

```


Variable Importance

                                                                                     Ratio of
                                       Number of                                    Validation
                                       Splitting                      Validation   to Training
Variable Name      Label                 Rules          Importance    Importance    Importance

RISK_MORTALITY     RISK_MORTALITY          1              1.0000        1.0000        1.0000
PRINC_DIAG_CODE    PRINC_DIAG_CODE         3              0.9113        0.9273        1.0176
TOTAL_CHARGES      TOTAL_CHARGES           4              0.3072        0.3050        0.9929
LENGTH_OF_STAY     LENGTH_OF_STAY          1              0.1591        0.1470        0.9242
TYPE_OF_BILL       TYPE_OF_BILL            1              0.0740        0.0610        0.8254
PAT_AGE            PAT_AGE                 1              0.0403        0.0462        1.1459



Tree Leaf Report

                                 Training
Node              Training        Percent     Validation     Validation
 Id      Depth   Observations       Y        Observations     Percent Y

 19        5         7685          0.03          3831           0.03
 15        4         6311          0.81          3179           0.81
 17        4         1814          0.02           905           0.02
 21        5         1549          0.07           764           0.06
 30        6         1000          0.79           536           0.79
 31        6          491          0.28           209           0.23
 26        6          402          0.18           218           0.17
 27        6          118          0.75            53           0.77
  7        3           98          0.04            41           0.02
  3        1           18          0.00             5           0.00
 10        4           18          0.67            13           0.69
 16        4            5          0.60             1           1.00




Fit Statistics
```

Results - Node: Decision Tree Diagram: Report 4

File Edit View Window

Output

```

Fit Statistics

Target=SPEC_UNIT_1 Target Label=SPEC_UNIT_1

   Fit
Statistics    Statistics Label                 Train    Validation        Test

 _NOBS_       Sum of Frequencies            19509.00       9755.00     3265.00
 _MISC_       Misclassification Rate            0.23          0.23        0.23
 _MAX_        Maximum Absolute Error            1.00          1.00        1.00
 _SSE_        Sum of Squared Errors          7353.77       3654.65     1252.12
 _ASE_        Average Squared Error             0.03          0.03        0.03
 _RASE_       Root Average Squared Error        0.18          0.18        0.18
 _DIV_        Divisor for ASE              234108.00     117060.00    39180.00
 _DFT_        Total Degrees of Freedom     214599.00           .           .




Classification Table

Data Role=TRAIN Target Variable=SPEC_UNIT_1 Target Label=SPEC_UNIT_1

                       Target       Outcome     Frequency       Total
Target    Outcome    Percentage    Percentage     Count       Percentage

  B          I         2.9404       98.061         354          1.8145
  C          I        13.8134       97.479        1663          8.5243
  D          I         1.0051       16.176         121          0.6202
  I          I        74.7903       93.733        9004         46.1531
  N          I         0.0665       80.000           8          0.0410
  O          I         1.4702       98.333         177          0.9073
  P          I         0.4901       88.060          59          0.3024
  R          I         0.2907       19.022          35          0.1794
  U          I         0.0748       52.941           9          0.0461
  Y          I         5.0586        9.201         609          3.1216
  I          S         5.5556        0.010           1          0.0051
  S          S        55.5556      100.000          10          0.0513
  U          S        38.8889       41.176           7          0.0359
  B          Y         0.0939        1.939           7          0.0359
```

Classification Table: As we analyse the event classification table we see that for the training data the false negative is 609, true negative is 11448, false positive is 1442 and true positive is 6010. Now, for validation data these numbers are false negative being 260, true negative being 5713, false positive being 732 and true positive being 3050.

Results - Node: Decision Tree Diagram: Report 4

File Edit View Window

Output

Data Role=VALIDATE Target Variable=SPEC_UNIT_1 Target Label=SPEC_UNIT_1

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
B	I	3.0161	99.448	180	1.8452
C	I	13.9410	97.538	832	8.5290
D	I	0.8881	14.209	53	0.5433
I	I	75.2178	93.482	4489	46.0174
N	I	0.1005	100.000	6	0.0615
O	I	1.5080	98.901	90	0.9226
P	I	0.5362	94.118	32	0.3280
R	I	0.3351	21.505	20	0.2050
U	I	0.1005	85.714	6	0.0615
Y	I	4.3566	7.855	260	2.6653
S	S	80.0000	100.000	4	0.0410
U	S	20.0000	14.286	1	0.0103
B	Y	0.0264	0.552	1	0.0103
C	Y	0.5553	2.462	21	0.2153
D	Y	8.4611	85.791	320	3.2804
H	Y	0.0264	100.000	1	0.0103
I	Y	8.2760	6.518	313	3.2086
O	Y	0.0264	1.099	1	0.0103
P	Y	0.0529	5.882	2	0.0205
R	Y	1.9302	78.495	73	0.7483
Y	Y	80.6452	92.145	3050	31.2660

Event Classification Table

Data Role=TRAIN Target=SPEC_UNIT_1 Target Label=SPEC_UNIT_1

False Negative	True Negative	False Positive	True Positive
609	11448	1442	6010

Results - Node: Decision Tree Diagram: Report 4

File Edit View Window

Output

D	Y	8.4611	85.791	320	3.2804
H	Y	0.0264	100.000	1	0.0103
I	Y	8.2760	6.518	313	3.2086
O	Y	0.0264	1.099	1	0.0103
P	Y	0.0529	5.882	2	0.0205
R	Y	1.9302	78.495	73	0.7483
Y	Y	80.6452	92.145	3050	31.2660

Event Classification Table

Data Role=TRAIN Target=SPEC_UNIT_1 Target Label=SPEC_UNIT_1

False Negative	True Negative	False Positive	True Positive
609	11448	1442	6010

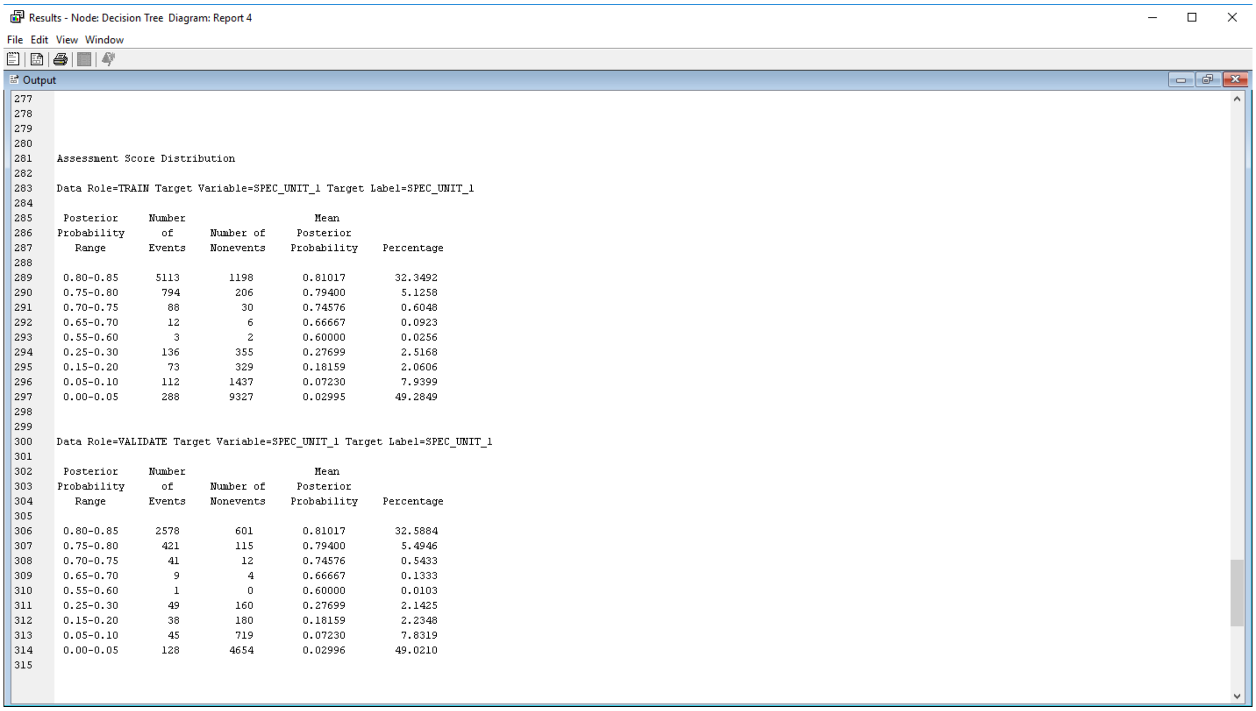
Data Role=VALIDATE Target=SPEC_UNIT_1 Target Label=SPEC_UNIT_1

False Negative	True Negative	False Positive	True Positive
260	5713	732	3050

Assessment Score Rankings

Data Role=TRAIN Target Variable=SPEC_UNIT_1 Target Label=SPEC_UNIT_1

Depth	Gain	Lift	Cumulative Lift	% Response	Cumulative % Response	Number of Observations	Mean Posterior Probability
5	138.792	2.38792	2.38792	81.0173	81.0173	976	0.81017
10	138.792	2.38792	2.38792	81.0173	81.0173	975	0.81017



Results - Node: Decision Tree Diagram: Report 4

Assessment Score Distribution

Data Role=TRAIN Target Variable=SPEC_UNIT_1 Target Label=SPEC_UNIT_1

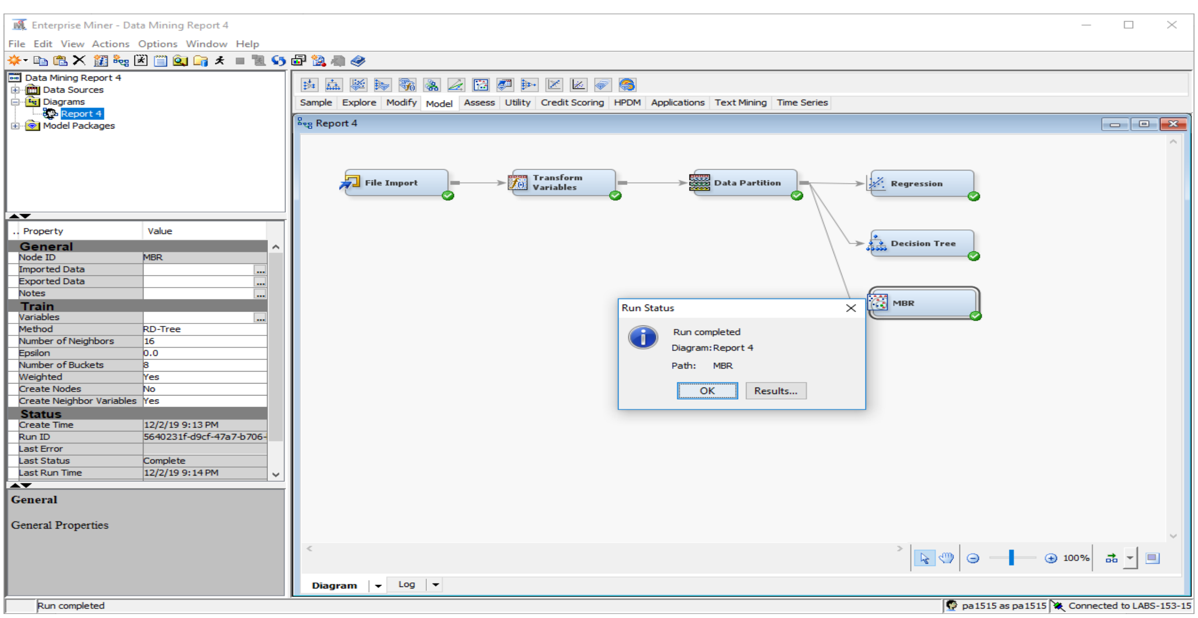
Posterior Probability Range	Number of Events	Number of Nonevents	Mean Posterior Probability	Percentage
0.80-0.85	5113	1198	0.81017	32.3492
0.75-0.80	794	206	0.79400	5.1258
0.70-0.75	88	30	0.74576	0.6048
0.65-0.70	12	6	0.66667	0.0923
0.55-0.60	3	2	0.60000	0.0256
0.25-0.30	136	355	0.27699	2.5168
0.15-0.20	73	329	0.18159	2.0606
0.05-0.10	112	1437	0.07230	7.9399
0.00-0.05	288	9327	0.02995	49.2849

Data Role=VALIDATE Target Variable=SPEC_UNIT_1 Target Label=SPEC_UNIT_1

Posterior Probability Range	Number of Events	Number of Nonevents	Mean Posterior Probability	Percentage
0.80-0.85	2578	601	0.81017	32.5884
0.75-0.80	421	115	0.79400	5.4946
0.70-0.75	41	12	0.74576	0.5433
0.65-0.70	9	4	0.66667	0.1333
0.55-0.60	1	0	0.60000	0.0103
0.25-0.30	49	160	0.27699	2.1425
0.15-0.20	38	180	0.18159	2.2348
0.05-0.10	45	719	0.07230	7.8319
0.00-0.05	128	4654	0.02996	49.0210

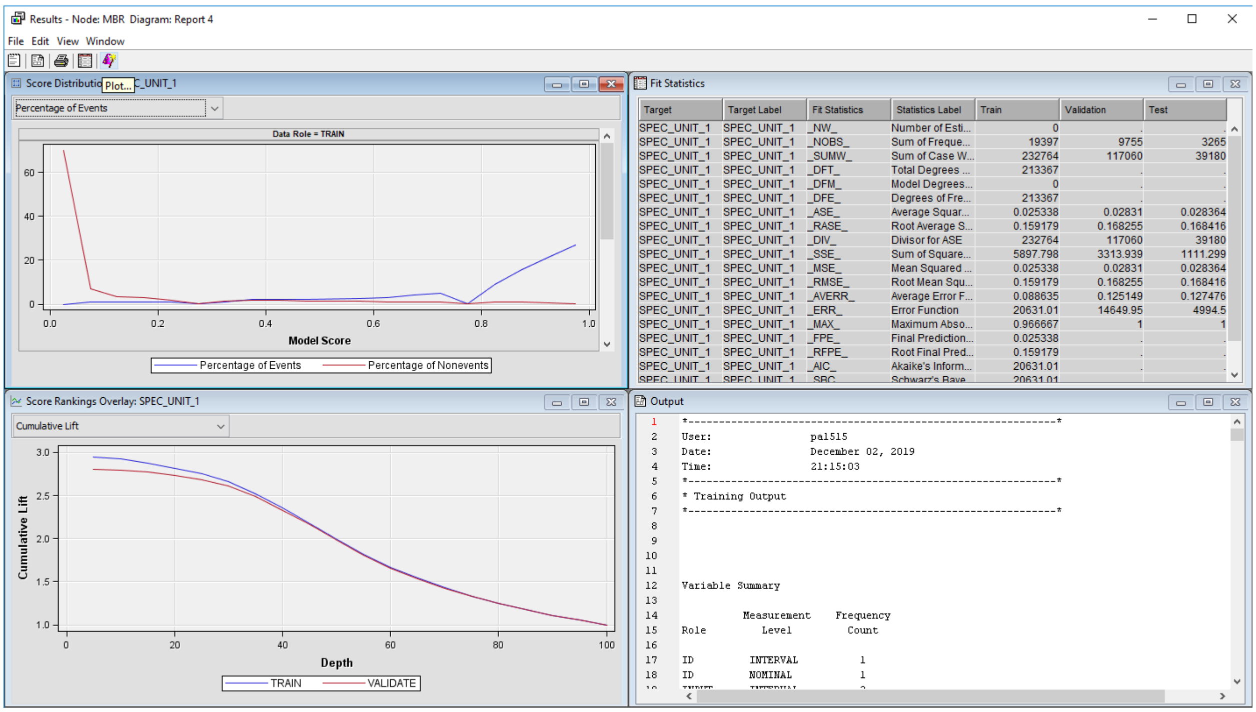
5.3 *KNN*

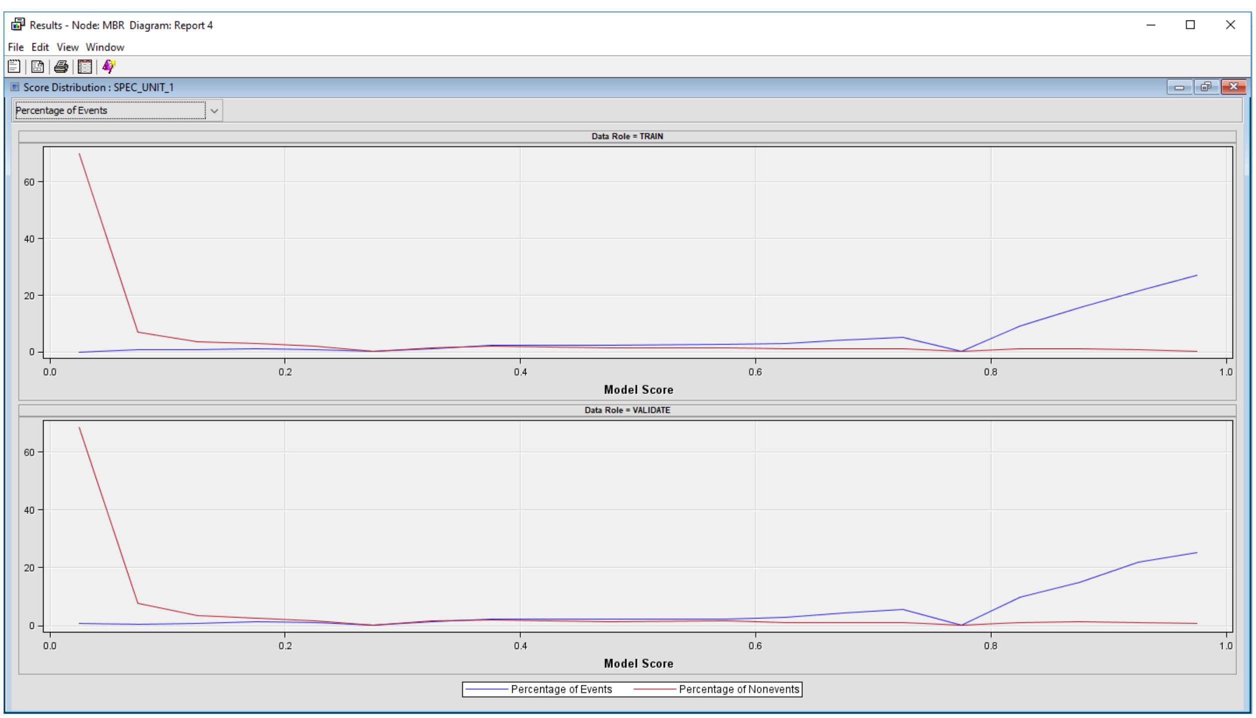
In the K nearest neighbour method the algorithm assigns similar records to each other. This keeps on continuing until all the records are classified. Here, we can define k as per our observation on how much accuracy we can acquire.



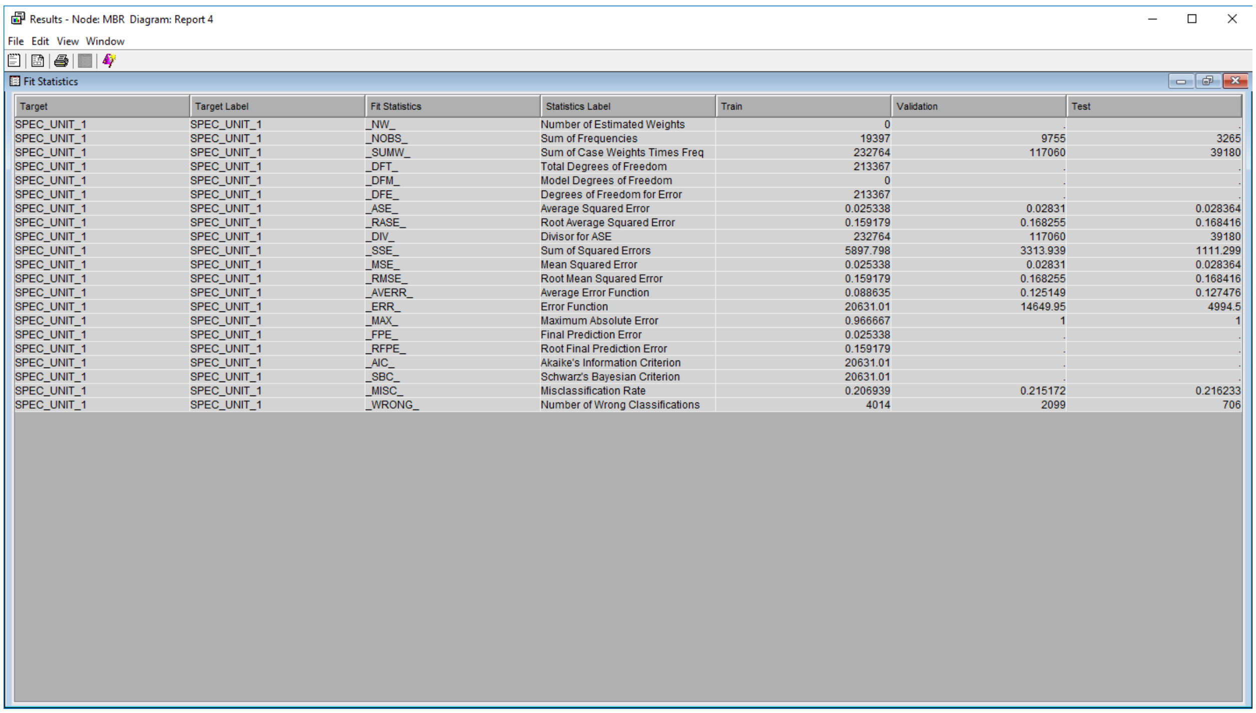
Assessing the Model:

As we click on the results button, we will be able to see the result that are produced by this algorithm. Hence, below we see the output that has been generated.



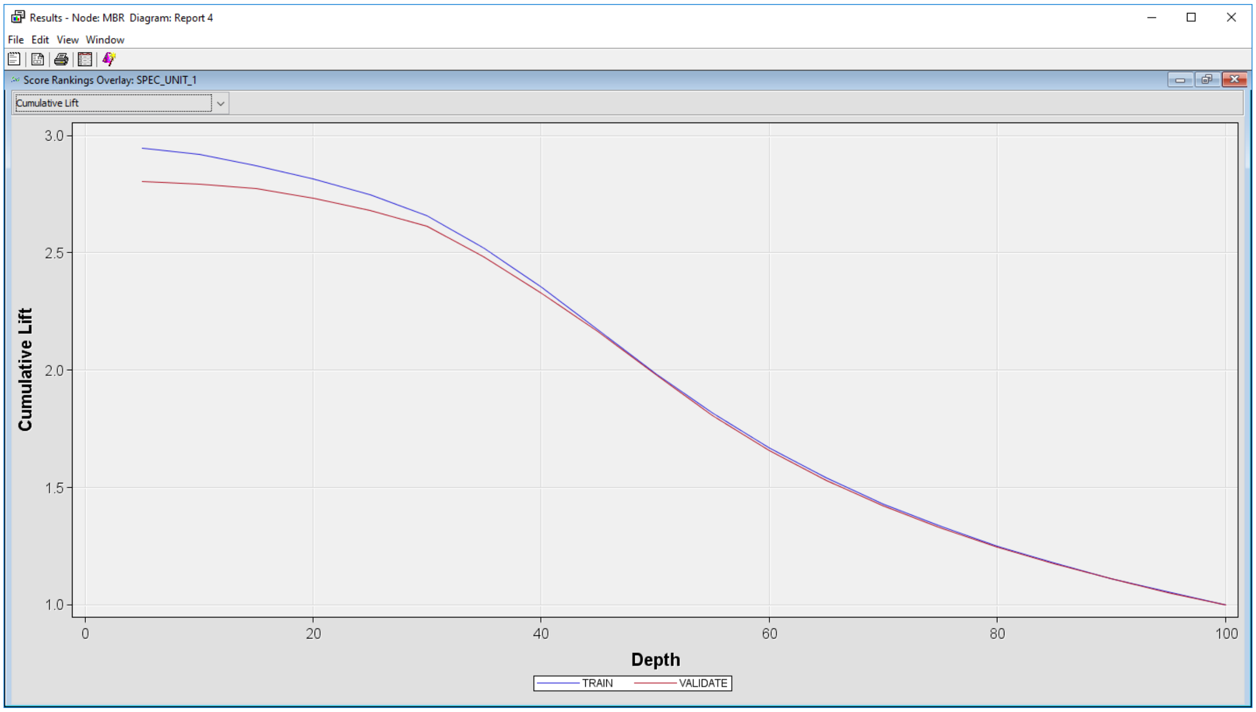


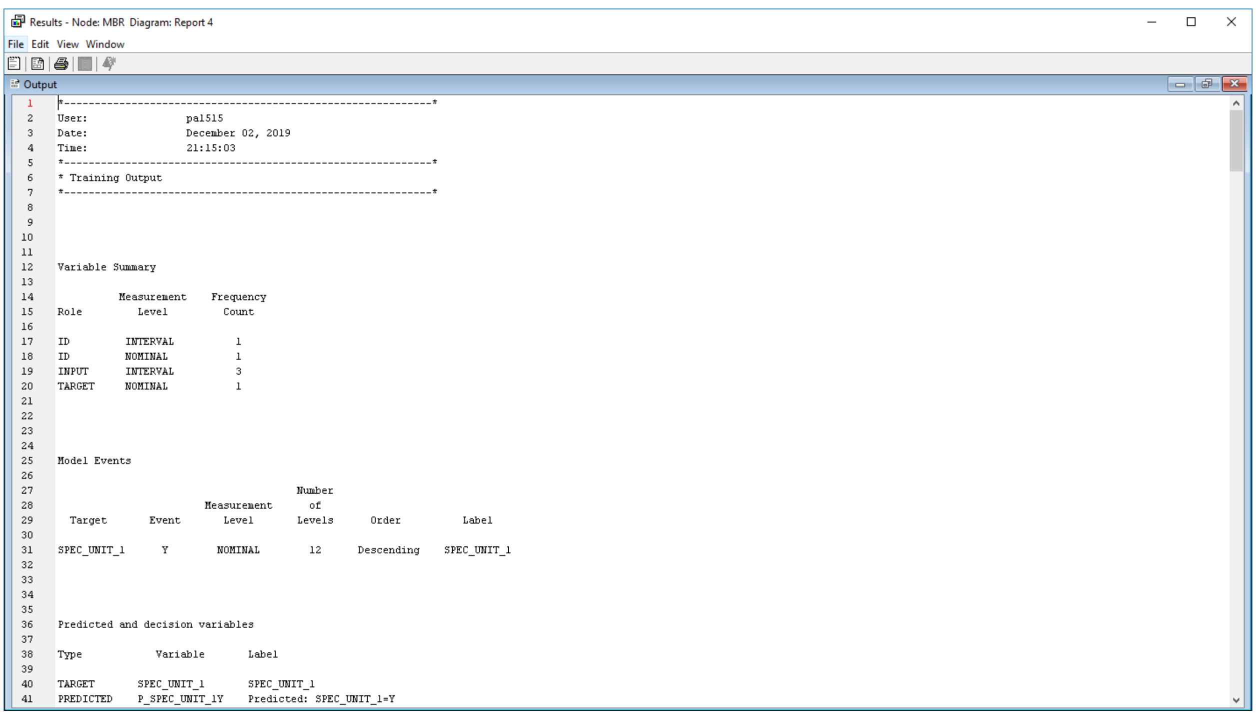
Fit Statistics: As we take a look at the fit statistics, we mainly check for the Root Mean Squared Error to check our outcome. Hence, we see that the value for training data is 0.159179, for validation data it is 0.168255 and for the test data it is 0.168416. Thus we can say that the model has performed fairly consistently with all the datasets indicating that our result has no issues in terms of data pattern being analysed. Also, the low value of the errors indicate that even the classification has been fairly accurate.



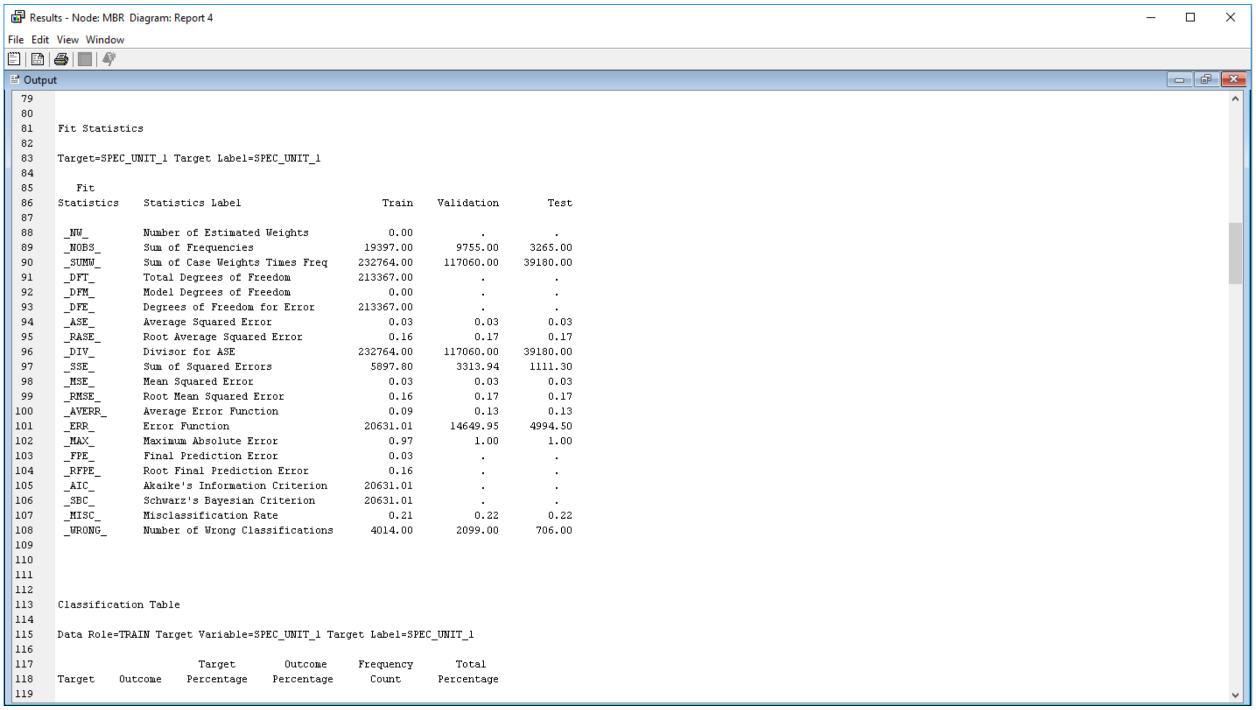
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
SPEC_UNIT_1	SPEC_UNIT_1	_NW_	Number of Estimated Weights	0	.	.
SPEC_UNIT_1	SPEC_UNIT_1	_NOBS_	Sum of Frequencies	19397	9755	3265
SPEC_UNIT_1	SPEC_UNIT_1	_SUMW_	Sum of Case Weights Times Freq	232764	117060	39180
SPEC_UNIT_1	SPEC_UNIT_1	_DFT_	Total Degrees of Freedom	213367	.	.
SPEC_UNIT_1	SPEC_UNIT_1	_DFM_	Model Degrees of Freedom	0	.	.
SPEC_UNIT_1	SPEC_UNIT_1	_DFE_	Degrees of Freedom for Error	213367	.	.
SPEC_UNIT_1	SPEC_UNIT_1	_ASE_	Average Squared Error	0.025338	0.02831	0.028364
SPEC_UNIT_1	SPEC_UNIT_1	_RASE_	Root Average Squared Error	0.159179	0.168255	0.168416
SPEC_UNIT_1	SPEC_UNIT_1	_DIV_	Divisor for ASE	232764	117060	39180
SPEC_UNIT_1	SPEC_UNIT_1	_SSE_	Sum of Squared Errors	5897.798	3313.939	1111.299
SPEC_UNIT_1	SPEC_UNIT_1	_MSE_	Mean Squared Error	0.025338	0.02831	0.028364
SPEC_UNIT_1	SPEC_UNIT_1	_RMSE_	Root Mean Squared Error	0.159179	0.168255	0.168416
SPEC_UNIT_1	SPEC_UNIT_1	_AVERR_	Average Error Function	0.088635	0.125149	0.127476
SPEC_UNIT_1	SPEC_UNIT_1	_ERR_	Error Function	20631.01	14649.95	4994.5
SPEC_UNIT_1	SPEC_UNIT_1	_MAX_	Maximum Absolute Error	0.966667	1	1
SPEC_UNIT_1	SPEC_UNIT_1	_FPE_	Final Prediction Error	0.025338	.	.
SPEC_UNIT_1	SPEC_UNIT_1	_RFPE_	Root Final Prediction Error	0.159179	.	.
SPEC_UNIT_1	SPEC_UNIT_1	_AIC_	Akaike's Information Criterion	20631.01	.	.
SPEC_UNIT_1	SPEC_UNIT_1	_SBC_	Schwarz's Bayesian Criterion	20631.01	.	.
SPEC_UNIT_1	SPEC_UNIT_1	_MISC_	Misclassification Rate	0.206939	0.215172	0.216233
SPEC_UNIT_1	SPEC_UNIT_1	_WRONG_	Number of Wrong Classifications	4014	2099	706

Lift Chart: As per the below output the lift chart here shows that the training and validation data set show somewhat similar trend in terms of classification. There are places where there is error or deviation but this seems to be tolerable.






```
*------------------------------------------------------------*
User:                pa1515
Date:                December 02, 2019
Time:                21:15:03
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*




Variable Summary

          Measurement    Frequency
Role         Level          Count

ID         INTERVAL           1
ID         NOMINAL            1
INPUT      INTERVAL           3
TARGET     NOMINAL            1




Model Events

                                            Number
                        Measurement           of
  Target       Event       Level            Levels      Order          Label

SPEC_UNIT_1      Y        NOMINAL            12       Descending    SPEC_UNIT_1




Predicted and decision variables

Type            Variable       Label

TARGET       SPEC_UNIT_1      SPEC_UNIT_1
PREDICTED    P_SPEC_UNIT_1Y   Predicted: SPEC_UNIT_1=Y
```



Fit Statistics

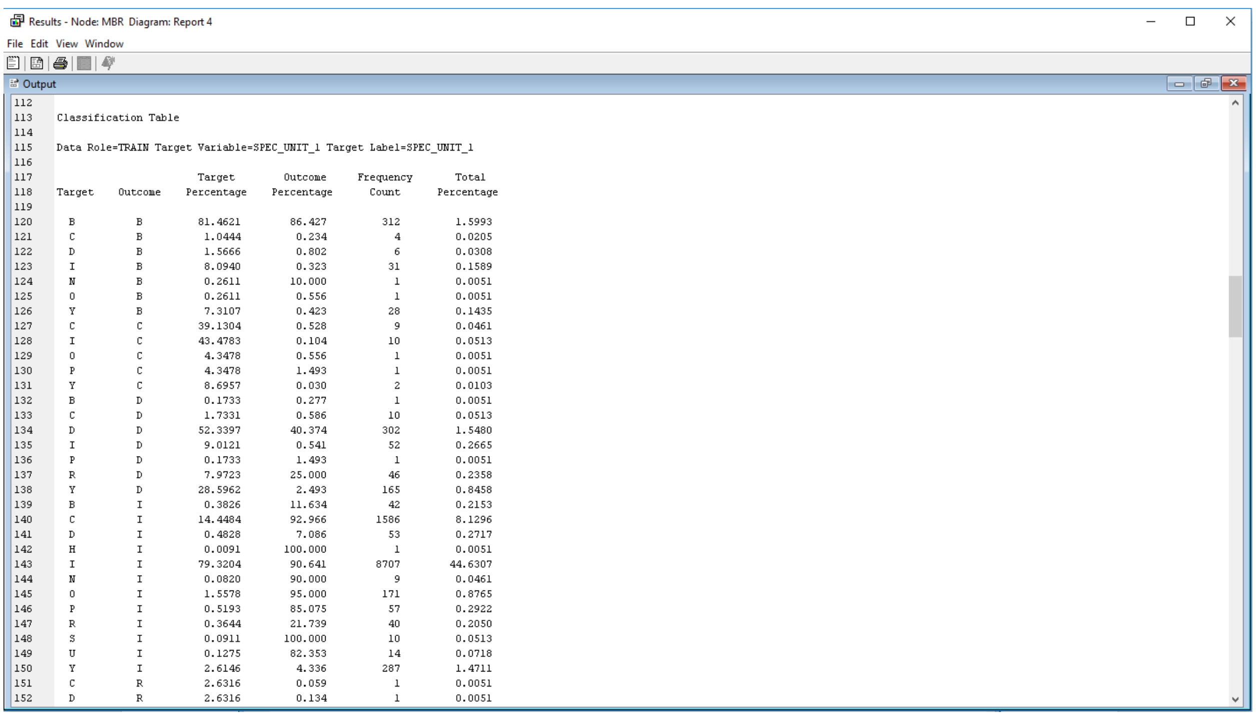
Target=SPEC_UNIT_1 Target Label=SPEC_UNIT_1

Fit Statistics	Statistics Label	Train	Validation	Test
_NW_	Number of Estimated Weights	0.00	.	.
NOBS	Sum of Frequencies	19397.00	9755.00	3265.00
SUMW	Sum of Case Weights Times Freq	232764.00	117060.00	39180.00
DFT	Total Degrees of Freedom	213367.00	.	.
DFM	Model Degrees of Freedom	0.00	.	.
DFE	Degrees of Freedom for Error	213367.00	.	.
ASE	Average Squared Error	0.03	0.03	0.03
RASE	Root Average Squared Error	0.16	0.17	0.17
DIV	Divisor for ASE	232764.00	117060.00	39180.00
SSE	Sum of Squared Errors	5897.80	3313.94	1111.30
MSE	Mean Squared Error	0.03	0.03	0.03
RMSE	Root Mean Squared Error	0.16	0.17	0.17
AVERR	Average Error Function	0.09	0.13	0.13
ERR	Error Function	20631.01	14649.95	4994.50
MAX	Maximum Absolute Error	0.97	1.00	1.00
FPE	Final Prediction Error	0.03	.	.
RFPE	Root Final Prediction Error	0.16	.	.
AIC	Akaike's Information Criterion	20631.01	.	.
SBC	Schwarz's Bayesian Criterion	20631.01	.	.
MISC	Misclassification Rate	0.21	0.22	0.22
WRONG	Number of Wrong Classifications	4014.00	2099.00	706.00

Classification Table

Data Role=TRAIN Target Variable=SPEC_UNIT_1 Target Label=SPEC_UNIT_1

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage



Classification Table

Data Role=TRAIN Target Variable=SPEC_UNIT_1 Target Label=SPEC_UNIT_1

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
B	B	81.4621	86.427	312	1.5993
C	B	1.0444	0.234	4	0.0205
D	B	1.5666	0.802	6	0.0308
I	B	8.0940	0.323	31	0.1589
N	B	0.2611	10.000	1	0.0051
O	B	0.2611	0.556	1	0.0051
Y	B	7.3107	0.423	28	0.1435
C	C	39.1304	0.528	9	0.0461
I	C	43.4783	0.104	10	0.0513
O	C	4.3478	0.556	1	0.0051
P	C	4.3478	1.493	1	0.0051
Y	C	8.6957	0.030	2	0.0103
B	D	0.1733	0.277	1	0.0051
C	D	1.7331	0.586	10	0.0513
D	D	52.3397	40.374	302	1.5480
I	D	9.0121	0.541	52	0.2665
P	D	0.1733	1.493	1	0.0051
R	D	7.9723	25.000	46	0.2358
Y	D	28.5962	2.493	165	0.8458
B	I	0.3826	11.634	42	0.2153
C	I	14.4484	92.966	1586	8.1296
D	I	0.4828	7.086	53	0.2717
H	I	0.0091	100.000	1	0.0051
I	I	79.3204	90.641	8707	44.6307
N	I	0.0820	90.000	9	0.0461
O	I	1.5578	95.000	171	0.8765
P	I	0.5193	85.075	57	0.2922
R	I	0.3644	21.739	40	0.2050
S	I	0.0911	100.000	10	0.0513
U	I	0.1275	82.353	14	0.0718
Y	I	2.6146	4.336	287	1.4711
C	R	2.6316	0.059	1	0.0051
D	R	2.6316	0.134	1	0.0051

Results - Node: MBR Diagram: Report 4

File Edit View Window

Output

Data Role=VALIDATE Target Variable=SPEC_UNIT_1 Target Label=SPEC_UNIT_1

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
B	B	80.6283	85.083	154	1.5787
C	B	0.5236	0.117	1	0.0103
D	B	1.5707	0.804	3	0.0308
I	B	11.5183	0.458	22	0.2255
Y	B	5.7592	0.332	11	0.1128
C	C	37.5000	0.352	3	0.0308
D	C	12.5000	0.268	1	0.0103
I	C	50.0000	0.083	4	0.0410
C	D	1.0676	0.352	3	0.0308
D	D	47.3310	35.657	133	1.3634
I	D	7.4733	0.437	21	0.2153
P	D	0.3559	2.941	1	0.0103
R	D	8.1851	24.731	23	0.2358
Y	D	35.5872	3.021	100	1.0251
B	I	0.4401	13.260	24	0.2460
C	I	14.5608	93.083	794	8.1394
D	I	0.4034	5.898	22	0.2255
H	I	0.0183	100.000	1	0.0103
I	I	79.0391	89.754	4310	44.1825
N	I	0.1100	100.000	6	0.0615
O	I	1.6321	97.802	89	0.9124
P	I	0.5868	94.118	32	0.3280
R	I	0.3668	21.505	20	0.2050
S	I	0.0734	100.000	4	0.0410
U	I	0.1284	100.000	7	0.0718
Y	I	2.6407	4.350	144	1.4762
C	R	12.5000	0.234	2	0.0205
I	R	6.2500	0.021	1	0.0103
R	R	43.7500	7.527	7	0.0718
Y	R	37.5000	0.181	6	0.0615
B	Y	0.0788	1.657	3	0.0308
C	Y	1.3137	5.862	50	0.5126
D	Y	5.6227	57.373	214	2.1937
I	Y	11.6658	9.246	444	4.5515
O	Y	0.0525	2.198	2	0.0205

Classification Table: As we analyze the event classification table we see that for the training data the false negative is 492, true negative is 11504, false positive is 1386 and true positive is 6125. Now, for validation data these numbers are false negative being 261, true negative being 5688, false positive being 757 and true positive being 3049.

Results - Node: MBR Diagram: Report 4

File Edit View Window

Output

Event Classification Table

Data Role=TRAIN Target=SPEC_UNIT_1 Target Label=SPEC_UNIT_1

False Negative	True Negative	False Positive	True Positive
494	11504	1386	6125

Data Role=VALIDATE Target=SPEC_UNIT_1 Target Label=SPEC_UNIT_1

False Negative	True Negative	False Positive	True Positive
261	5688	757	3049

Assessment Score Rankings

Data Role=TRAIN Target Variable=SPEC_UNIT_1 Target Label=SPEC_UNIT_1

Depth	Gain	Lift	Cumulative Lift	% Response	Cumulative % Response	Number of Observations	Mean Posterior Probability
5	194.403	2.94403	2.94403	99.8849	99.8849	976	1.00000
10	192.110	2.89814	2.92110	98.3280	99.1068	975	0.98743
15	187.011	2.76819	2.87011	93.9189	97.3769	976	0.93750
20	181.326	2.64257	2.81326	89.6570	95.4479	975	0.89901
25	174.791	2.48668	2.74791	84.3681	93.2310	976	0.85109
30	165.801	2.20820	2.65801	74.9195	90.1807	975	0.76669
35	151.959	1.68950	2.51959	57.3212	85.4844	976	0.62232
40	135.346	1.18988	2.35346	40.3703	79.8480	975	0.44339
45	117.160	0.71751	2.17160	24.3435	73.6781	976	0.26965
50	98.679	0.32247	1.98679	10.9408	67.4076	975	0.12137

Results - Node: MBR Diagram: Report 4

File Edit View Window

Output

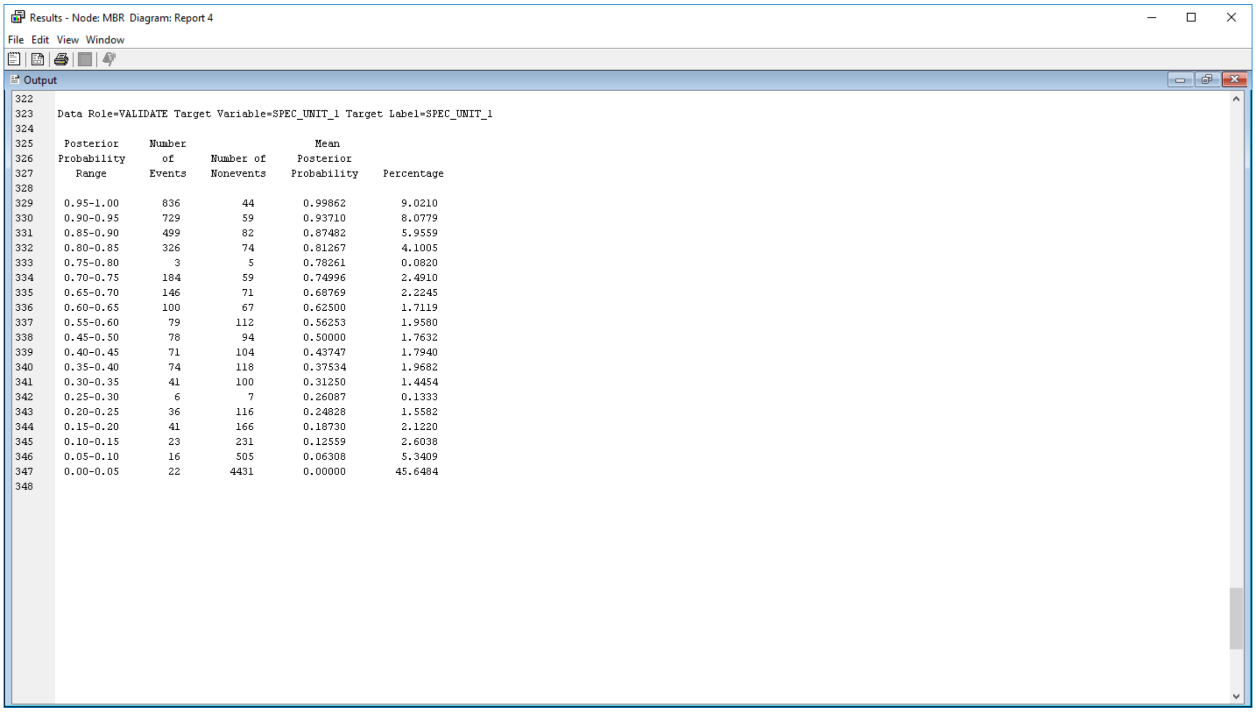
Assessment Score Distribution

Data Role=TRAIN Target Variable=SPEC_UNIT_1 Target Label=SPEC_UNIT_1

Posterior Probability Range	Number of Events	Number of Nonevents	Mean Posterior Probability	Percentage
0.95-1.00	1788	8	0.99857	9.2060
0.90-0.95	1425	97	0.93683	7.8015
0.85-0.90	1038	149	0.87487	6.0844
0.80-0.85	597	152	0.81268	3.8393
0.75-0.80	8	5	0.78261	0.0666
0.70-0.75	339	123	0.74988	2.3681
0.65-0.70	279	146	0.68786	2.1785
0.60-0.65	198	138	0.62500	1.7223
0.55-0.60	164	171	0.56252	1.7172
0.45-0.50	151	198	0.50000	1.7889
0.40-0.45	148	229	0.43740	1.9324
0.35-0.40	148	241	0.37508	1.9940
0.30-0.35	74	195	0.31244	1.3789
0.25-0.30	13	16	0.26087	0.1486
0.20-0.25	60	249	0.24873	1.5839
0.15-0.20	71	363	0.18681	2.2246
0.10-0.15	58	458	0.12545	2.6449
0.05-0.10	60	907	0.06262	4.9567
0.00-0.05	0	9045	0.00000	46.3632

Data Role=VALIDATE Target Variable=SPEC_UNIT_1 Target Label=SPEC_UNIT_1

Posterior Probability Range	Number of Events	Number of Nonevents	Mean Posterior Probability	Percentage
0.95-1.00	836	44	0.99862	9.0210
0.90-0.95	729	59	0.93710	8.0779
0.85-0.90	499	82	0.87482	5.9559
0.80-0.85	326	74	0.81267	4.1005

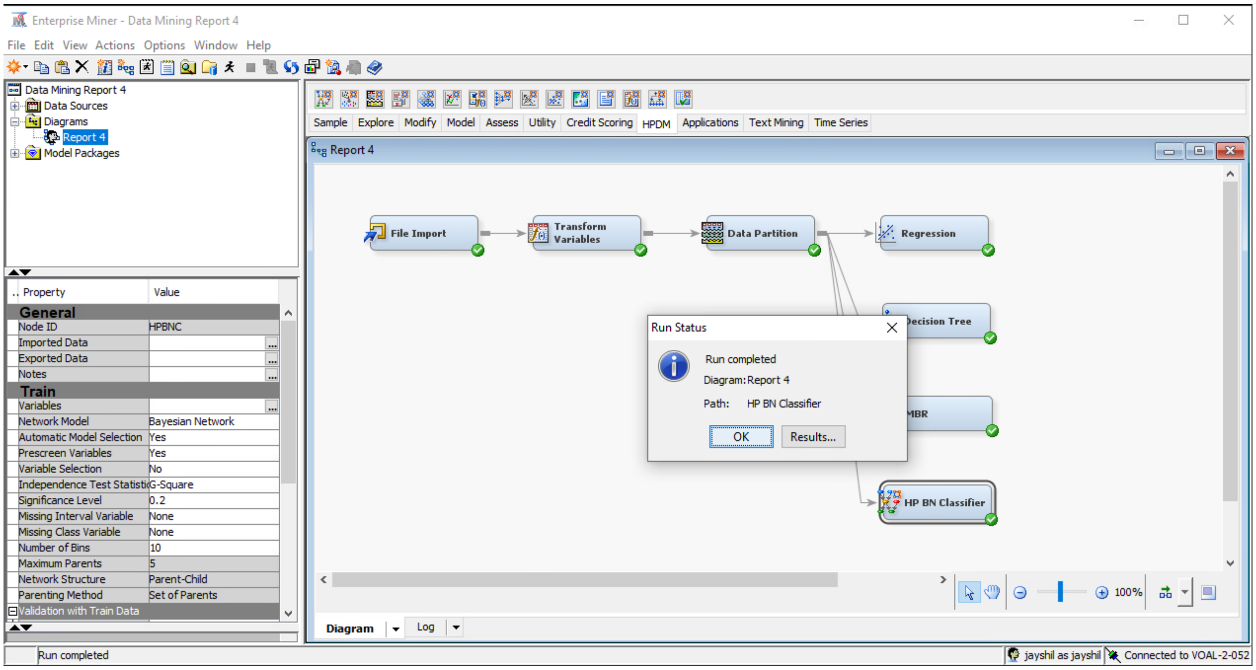

Results - Node: MBR Diagram: Report 4

Data Role=VALIDATE Target Variable=SPEC_UNIT_1 Target Label=SPEC_UNIT_1

Posterior Probability Range	Number of Events	Number of Nonevents	Mean Posterior Probability	Percentage
0.95-1.00	836	44	0.99862	9.0210
0.90-0.95	729	59	0.93710	8.0779
0.85-0.90	499	82	0.87482	5.9559
0.80-0.85	326	74	0.81267	4.1005
0.75-0.80	3	5	0.78261	0.0820
0.70-0.75	184	59	0.74996	2.4910
0.65-0.70	146	71	0.68769	2.2245
0.60-0.65	100	67	0.62500	1.7119
0.55-0.60	79	112	0.56253	1.9580
0.45-0.50	78	94	0.50000	1.7632
0.40-0.45	71	104	0.43747	1.7940
0.35-0.40	74	118	0.37534	1.9682
0.30-0.35	41	100	0.31250	1.4454
0.25-0.30	6	7	0.26087	0.1333
0.20-0.25	36	116	0.24828	1.5582
0.15-0.20	41	166	0.18730	2.1220
0.10-0.15	23	231	0.12559	2.6038
0.05-0.10	16	505	0.06308	5.3409
0.00-0.05	22	4431	0.00000	45.6484

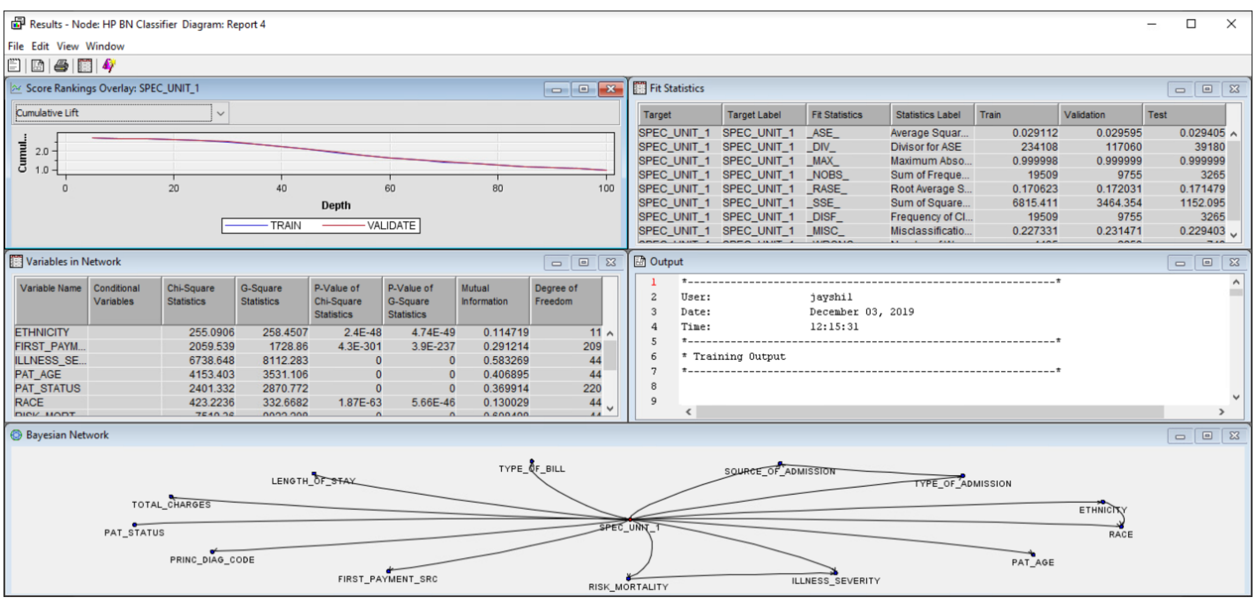
5.4 Naïve Bayes:

In Naïve Bayes method we use class probabilities for identifying the class of the data. Mostly, the benchmark for allocating a class for a record is the cut-off probability. It is used for categorical predictors but numerical predictors can be used if they are binned or converted to categorical predictors.

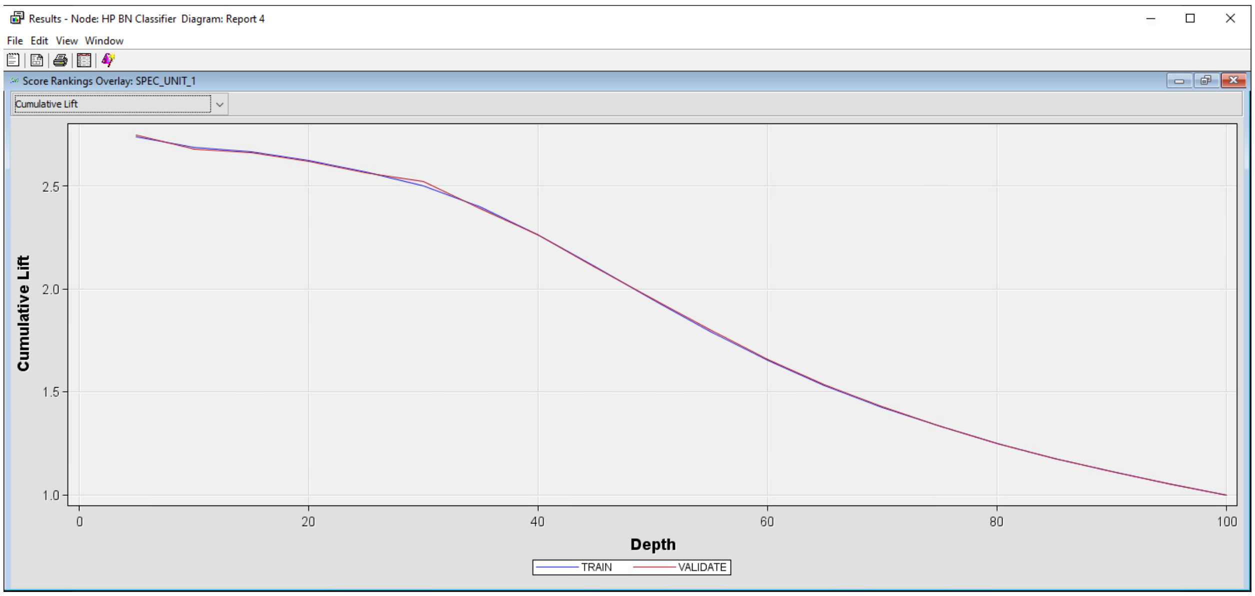


Assessing the Model:

As we click on the results button, we will be able to see the result that are produced by this algorithm. Hence, below we see the output that has been generated.



Lift Chart: As per the below output the lift chart here shows that the training and validation data set show somewhat similar trend in terms of classification. There are places where there is error or deviation but this seems to be tolerable.



Fit Statistics: As we take a look at the fit statistics, we mainly check for the Root Mean Squared Error to check our outcome. Hence, we see that the value for training data is 0.170523, for validation data it is 0.172031 and for the test data it is 0.171479. Thus we can say that the model has performed fairly consistently with all the datasets indicating that our result has no issues in terms of data pattern being analysed. Also, the low value of the errors indicate that even the classification has been fairly accurate.

Results - Node: HP BN Classifier Diagram: Report 4

File Edit View Window

Fit Statistics

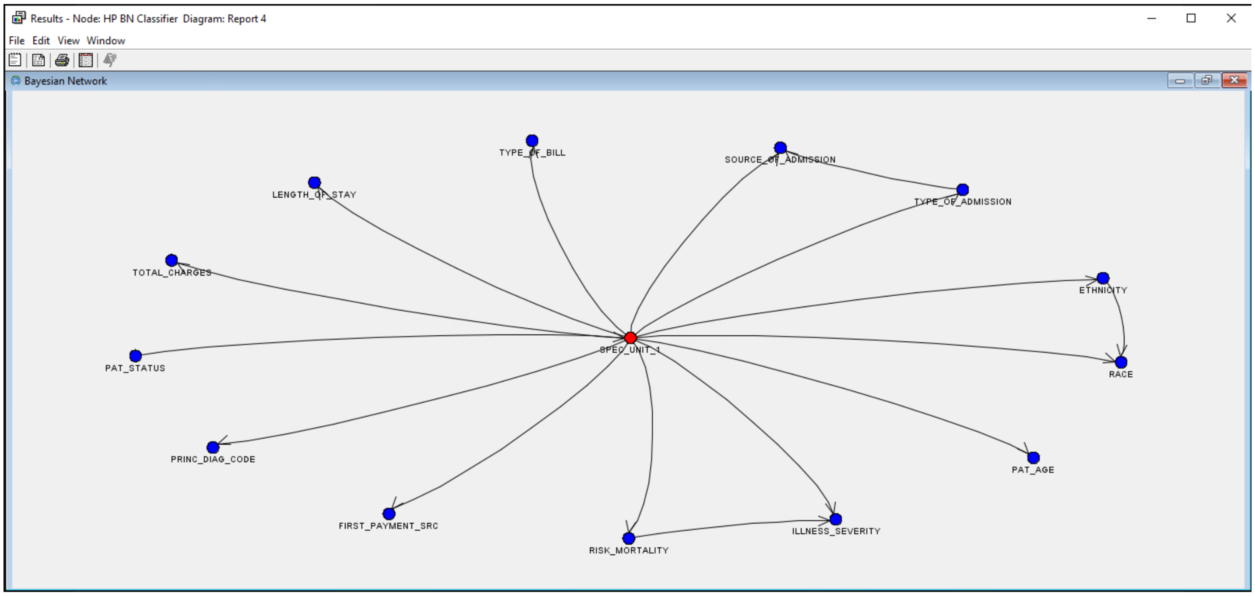
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
SPEC_UNIT_1	SPEC_UNIT_1	_ASE_	Average Squared Error	0.029112	0.029595	0.029405
SPEC_UNIT_1	SPEC_UNIT_1	_DIV_	Divisor for ASE	234108	117060	39180
SPEC_UNIT_1	SPEC_UNIT_1	_MAX_	Maximum Absolute Error	0.999998	0.999999	0.999999
SPEC_UNIT_1	SPEC_UNIT_1	_NOBS_	Sum of Frequencies	19509	9755	3265
SPEC_UNIT_1	SPEC_UNIT_1	_RASE_	Root Average Squared Error	0.170623	0.172031	0.171479
SPEC_UNIT_1	SPEC_UNIT_1	_SSE_	Sum of Squared Errors	6815.411	3464.354	1152.095
SPEC_UNIT_1	SPEC_UNIT_1	_DISF_	Frequency of Classified Cases	19509	9755	3265
SPEC_UNIT_1	SPEC_UNIT_1	_MISC_	Misclassification Rate	0.227331	0.231471	0.229403
SPEC_UNIT_1	SPEC_UNIT_1	_WRONG_	Number of Wrong Classifications	4435	2258	749

Results - Node: HP BN Classifier Diagram: Report 4

File Edit View Window

Variables in Network

Variable Name	Conditional Variables	Chi-Square Statistics	G-Square Statistics	P-Value of Chi-Square Statistics	P-Value of G-Square Statistics	Mutual Information	Degree of Freedom
ETHNICITY		255.0906	258.4507	2.4E-48	4.74E-49	0.114719	11
FIRST_PAYMENT_SRC		2059.539	1728.86	4.3E-301	3.9E-237	0.291214	209
ILLNESS_SEVERITY		6738.648	8112.283	0	0	0.583269	44
PAT_AGE		4153.403	3531.106	0	0	0.406895	44
PAT_STATUS		2401.332	2870.772	0	0	0.369914	220
RACE		423.2236	332.6682	1.87E-63	5.66E-46	0.130029	44
RISK_MORTALITY		7519.36	9022.208	0	0	0.608498	44
SOURCE_OF_ADMISSION		4922.456	4488.313	0	0	0.45334	88
TYPE_OF_ADMISSION		5028.235	4875.264	0	0	0.470234	44
TYPE_OF_BILL		14355.67	276.2271	0	3.21E-31	0.118571	55
LENGTH_OF_STAY		2297.255	240.2885	0	1.11E-18	0.11064	77
PRINC_DIAG_CODE		27614.13	15763.28	0	0	0.744481	99
TOTAL_CHARGES		1678.61	209.4116	1.7E-299	3.45E-14	0.103328	77

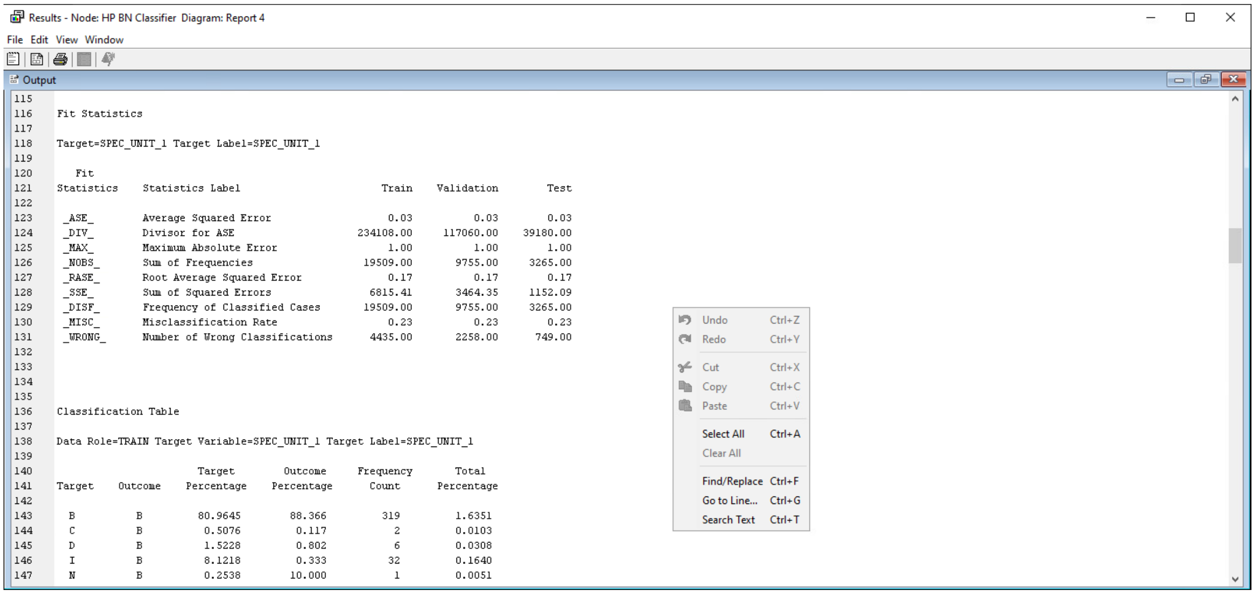


Results - Node: HP BN Classifier Diagram: Report 4

File Edit View Window

Output

```
*------------------------------------------------------------*
User:                jayshil
Date:                December 03, 2019
Time:                12:15:31
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*




Variable Summary

         Measurement    Frequency
Role        Level         Count

ID         INTERVAL         1
ID         NOMINAL          1
INPUT      INTERVAL         3
INPUT      NOMINAL         10
TARGET     NOMINAL          1




Model Events

                                        Number
                       Measurement        of
  Target       Event      Level         Levels      Order          Label

SPEC_UNIT_1      Y       NOMINAL          12      Descending    SPEC_UNIT_1

```



Fit Statistics

Target=SPEC_UNIT_1 Target Label=SPEC_UNIT_1

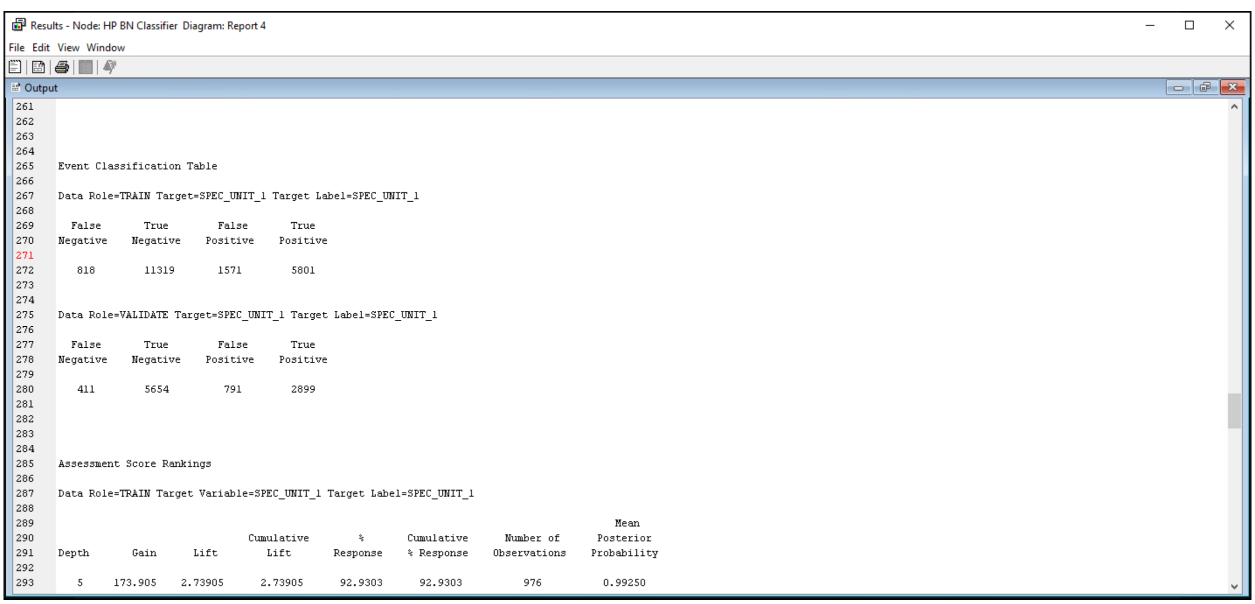
Fit Statistics	Statistics Label	Train	Validation	Test
ASE	Average Squared Error	0.03	0.03	0.03
DIV	Divisor for ASE	234108.00	117060.00	39180.00
MAX	Maximum Absolute Error	1.00	1.00	1.00
NOBS	Sum of Frequencies	19509.00	9755.00	3265.00
RASE	Root Average Squared Error	0.17	0.17	0.17
SSE	Sum of Squared Errors	6815.41	3464.35	1152.09
DISF	Frequency of Classified Cases	19509.00	9755.00	3265.00
MISC	Misclassification Rate	0.23	0.23	0.23
WRONG	Number of Wrong Classifications	4435.00	2258.00	749.00

Classification Table

Data Role=TRAIN Target Variable=SPEC_UNIT_1 Target Label=SPEC_UNIT_1

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
B	B	80.9645	88.366	319	1.6351
C	B	0.5076	0.117	2	0.0103
D	B	1.5228	0.802	6	0.0308
I	B	8.1218	0.333	32	0.1640
N	B	0.2538	10.000	1	0.0051

Classification Table: As we analyse the event classification table we see that for the training data the false negative is 818, true negative is 11319, false positive is 1571 and true positive is 5801. Now, for validation data these numbers are false negative being 411, true negative being 5654, false positive being 791 and true positive being 2899.



Event Classification Table

Data Role=TRAIN Target=SPEC_UNIT_1 Target Label=SPEC_UNIT_1

False Negative	True Negative	False Positive	True Positive
818	11319	1571	5801

Data Role=VALIDATE Target=SPEC_UNIT_1 Target Label=SPEC_UNIT_1

False Negative	True Negative	False Positive	True Positive
411	5654	791	2899

Assessment Score Rankings

Data Role=TRAIN Target Variable=SPEC_UNIT_1 Target Label=SPEC_UNIT_1

Depth	Gain	Lift	Cumulative Lift	% Response	Cumulative % Response	Number of Observations	Mean Posterior Probability
5	173.905	2.73905	2.73905	92.9303	92.9303	976	0.99250

Results - Node: HP BN Classifier Diagram: Report 4

File Edit View Window

Output

Assessment Score Distribution

Data Role=TRAIN Target Variable=SPEC_UNIT_1 Target Label=SPEC_UNIT_1

Posterior Probability Range	Number of Events	Number of Nonevents	Mean Posterior Probability	Percentage
0.95-1.00	2324	231	0.97839	13.0965
0.90-0.95	1264	223	0.92623	7.6221
0.85-0.90	716	180	0.87816	4.5928
0.80-0.85	359	121	0.82707	2.4604
0.75-0.80	271	112	0.77807	1.9632
0.70-0.75	197	92	0.72777	1.4814
0.65-0.70	194	105	0.67295	1.5326
0.60-0.65	119	118	0.62730	1.2148
0.55-0.60	119	91	0.57492	1.0764
0.50-0.55	93	95	0.52759	0.9637
0.45-0.50	110	156	0.47401	1.3635
0.40-0.45	112	117	0.42501	1.1738
0.35-0.40	76	99	0.37348	0.8970
0.30-0.35	98	212	0.32473	1.5890
0.25-0.30	139	320	0.27714	2.3528
0.20-0.25	92	252	0.22227	1.7633
0.15-0.20	64	264	0.17322	1.6813
0.10-0.15	69	415	0.12214	2.4809
0.05-0.10	82	515	0.07420	3.0601
0.00-0.05	121	9172	0.00439	47.6344

Data Role=VALIDATE Target Variable=SPEC_UNIT_1 Target Label=SPEC_UNIT_1

Results - Node: HP BN Classifier Diagram: Report 4

File Edit View Window

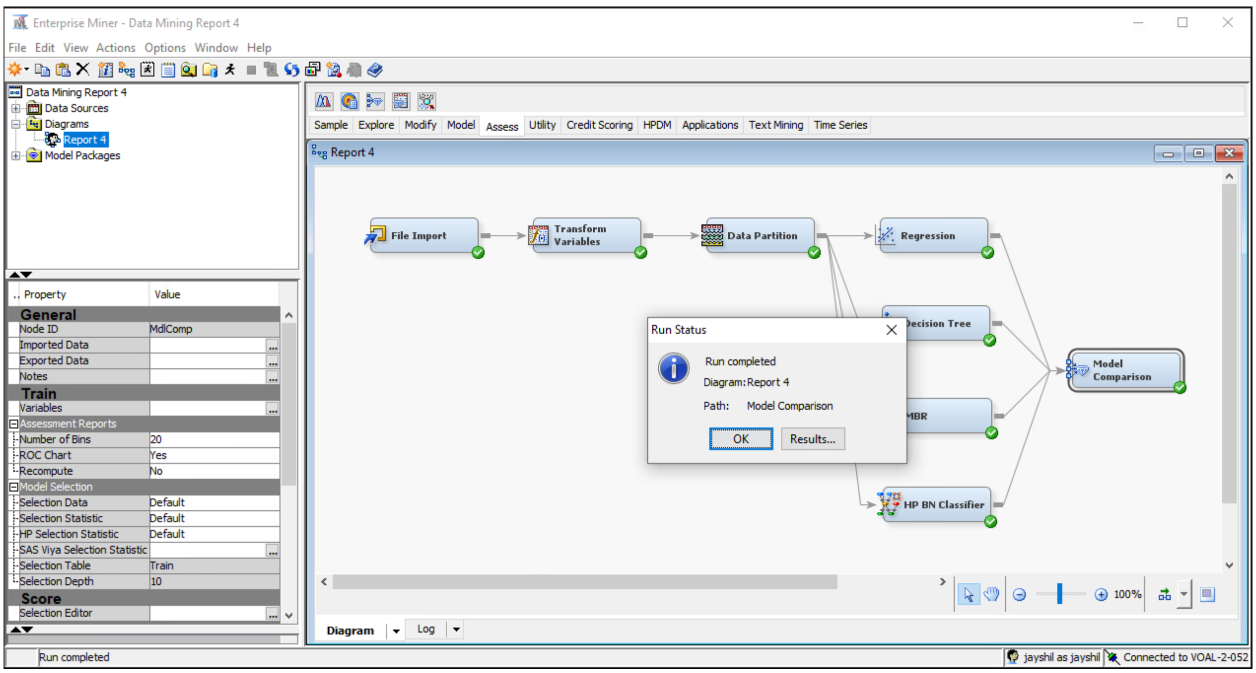
Output

Data Role=VALIDATE Target Variable=SPEC_UNIT_1 Target Label=SPEC_UNIT_1

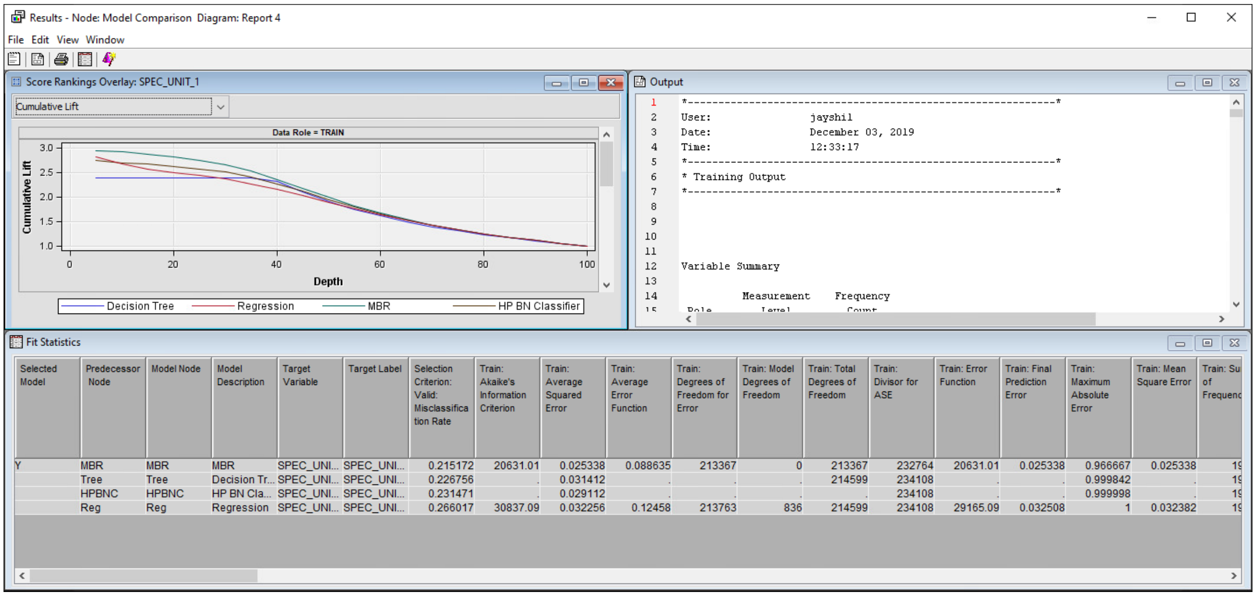
Posterior Probability Range	Number of Events	Number of Nonevents	Mean Posterior Probability	Percentage
0.95-1.00	1074	107	0.97890	12.1066
0.90-0.95	681	113	0.92623	8.1394
0.85-0.90	367	96	0.87872	4.7463
0.80-0.85	191	50	0.82746	2.4705
0.75-0.80	157	42	0.77712	2.0400
0.70-0.75	111	61	0.72707	1.7632
0.65-0.70	79	61	0.67681	1.4352
0.60-0.65	55	58	0.62565	1.1584
0.55-0.60	45	40	0.57268	0.8713
0.50-0.55	48	62	0.52721	1.1276
0.45-0.50	78	69	0.47314	1.5069
0.40-0.45	59	77	0.42597	1.3942
0.35-0.40	34	53	0.37163	0.8919
0.30-0.35	61	109	0.32497	1.7427
0.25-0.30	70	170	0.27612	2.4603
0.20-0.25	43	128	0.22294	1.7529
0.15-0.20	33	125	0.17453	1.6197
0.10-0.15	33	200	0.12042	2.3885
0.05-0.10	39	233	0.07504	2.7883
0.00-0.05	52	4591	0.00421	47.5961

6. Model Comparison:

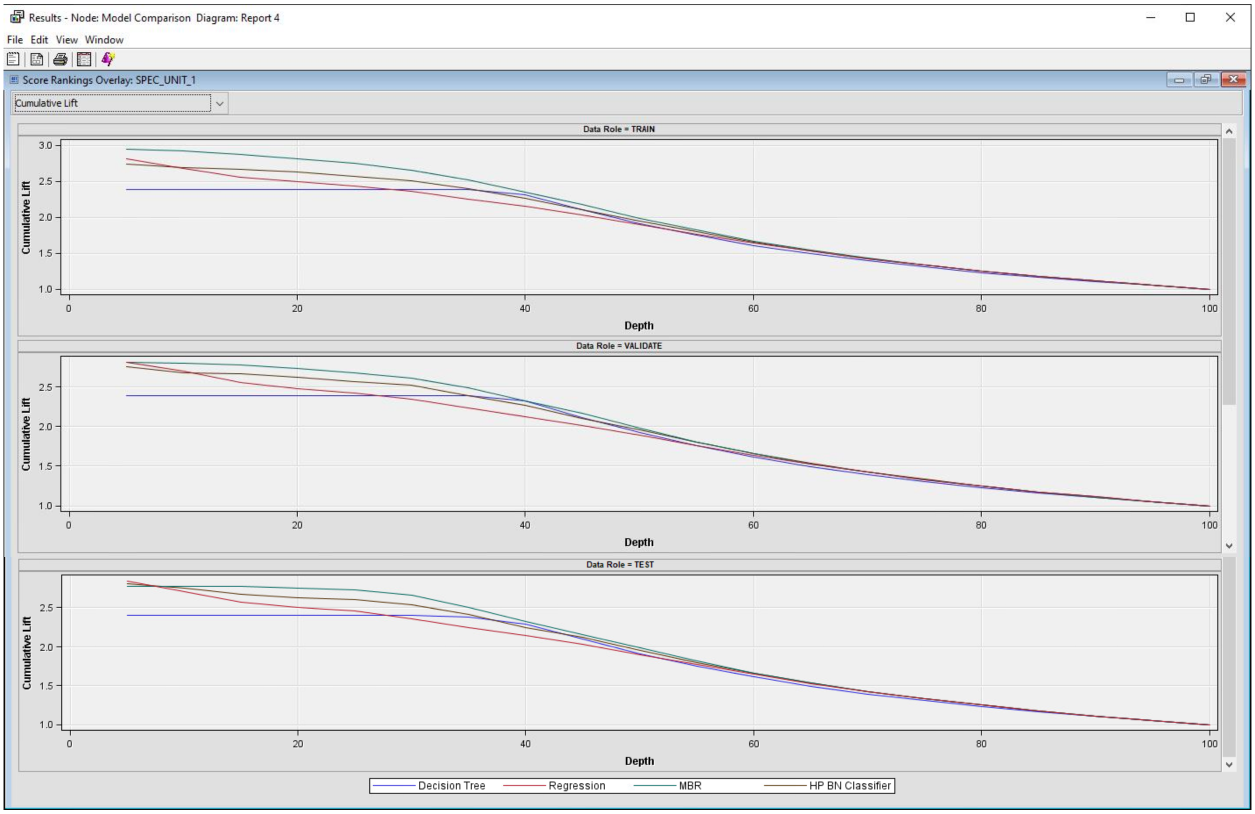
As we ran 4 different models on the same dataset, we can see that each of them gives us slightly different performance in terms of classifying the data. This tells us that the type of data, number of variables and number of observations affect the model to a great extent. Hence, we ran a model comparison node to compare all these 4 models and find the one that most suits our data.



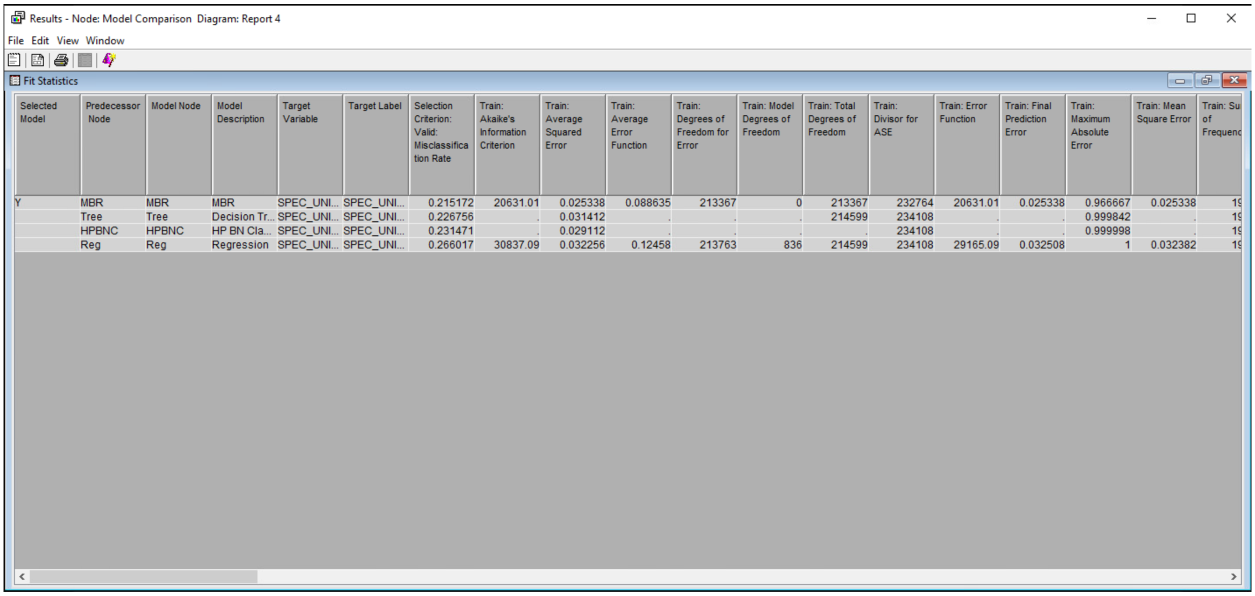
As we see below, we can see the outputs combined that help us to compare all the models that we want to check.



Hence as we see the lift chart, they all start at different points with a little but of variation for each but as it goes towards the end they tend to combine giving us similar results, however we can still differentiate which model performs better based on certain statistics. Here we also notice the trend for cumulative lift is almost similar for all training, validation and test data sets which indicates the models performed comparatively well for all the data sets as there are no major portion that can indicate any overfitting or any errors.



As we see below, in the fit statistics, the model comparison node has selected k Nearest Neighbour as the best model for this particular case of classification. The criteria that it has used as a parameter is the Misclassification rate which is lowest for k Nearest Neighbour i.e. 0.215172. All the others have a higher misclassification rate than this which means that they are more prone to error while classifying new records. Also, if we look at other parameters such as Average Squared Error, even that is very low for this K Nearest Neighbour than the rest of the models. Hence, we select k Nearest Neighbour is the model that is best for this scenario.

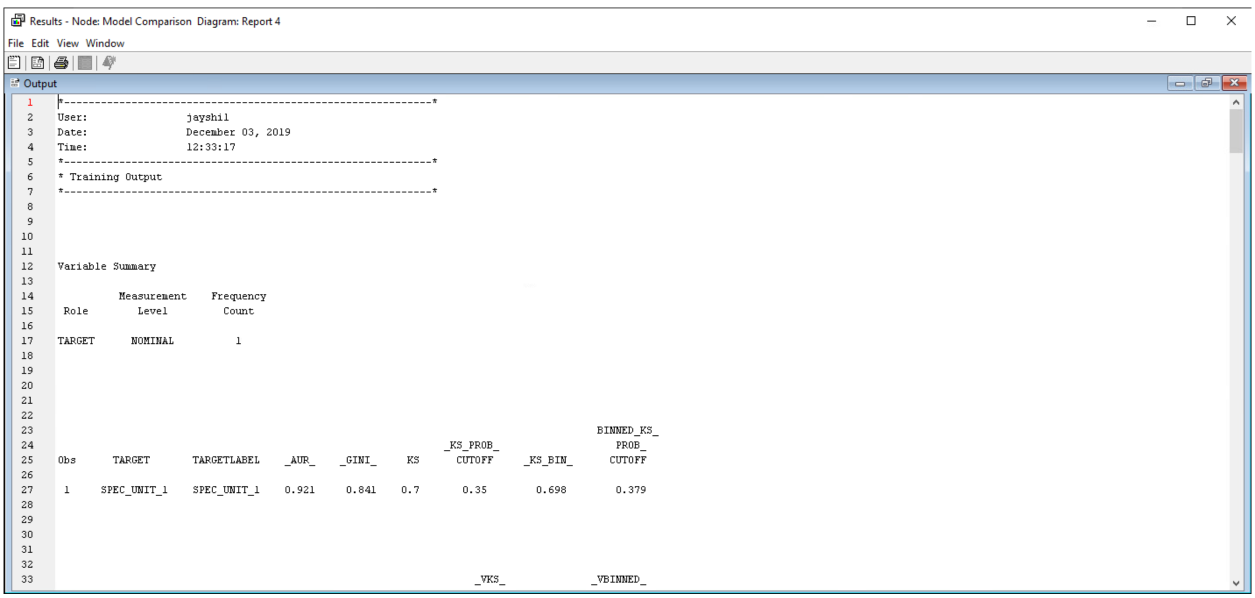


Results - Node: Model Comparison Diagram: Report 4

File Edit View Window

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Train: Akaike's Information Criterion	Train: Average Squared Error	Train: Average Error Function	Train: Degrees of Freedom for Error	Train: Model Degrees of Freedom	Train: Total Degrees of Freedom	Train: Divisor for ASE	Train: Error Function	Train: Final Prediction Error	Train: Maximum Absolute Error	Train: Mean Square Error	Train: Su of Frequenc
Y	MBR	MBR	MBR	SPEC_UNI...	SPEC_UNI...	0.215172	20631.01	0.025338	0.088635	213367	0	213367	232764	20631.01	0.025338	0.966667	0.025338	19
	Tree	Tree	Decision Tr...	SPEC_UNI...	SPEC_UNI...	0.226756	.	0.031412	.	.	.	214599	234108	.	.	0.999842	.	19
	HPBNC	HPBNC	HP BN Cla...	SPEC_UNI...	SPEC_UNI...	0.231471	.	0.029112	.	.	.	.	234108	.	.	0.999998	.	19
	Reg	Reg	Regression	SPEC_UNI...	SPEC_UNI...	0.266017	30837.09	0.032256	0.12458	213763	836	214599	234108	29165.09	0.032508	1	0.032382	19

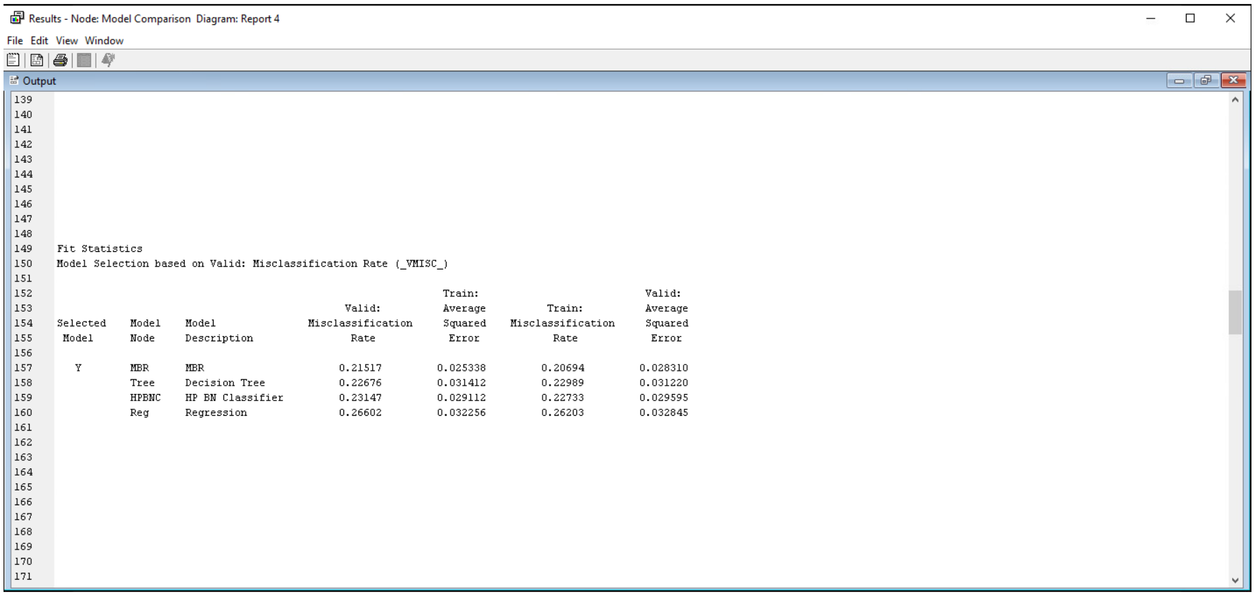


Results - Node: Model Comparison Diagram: Report 4

File Edit View Window

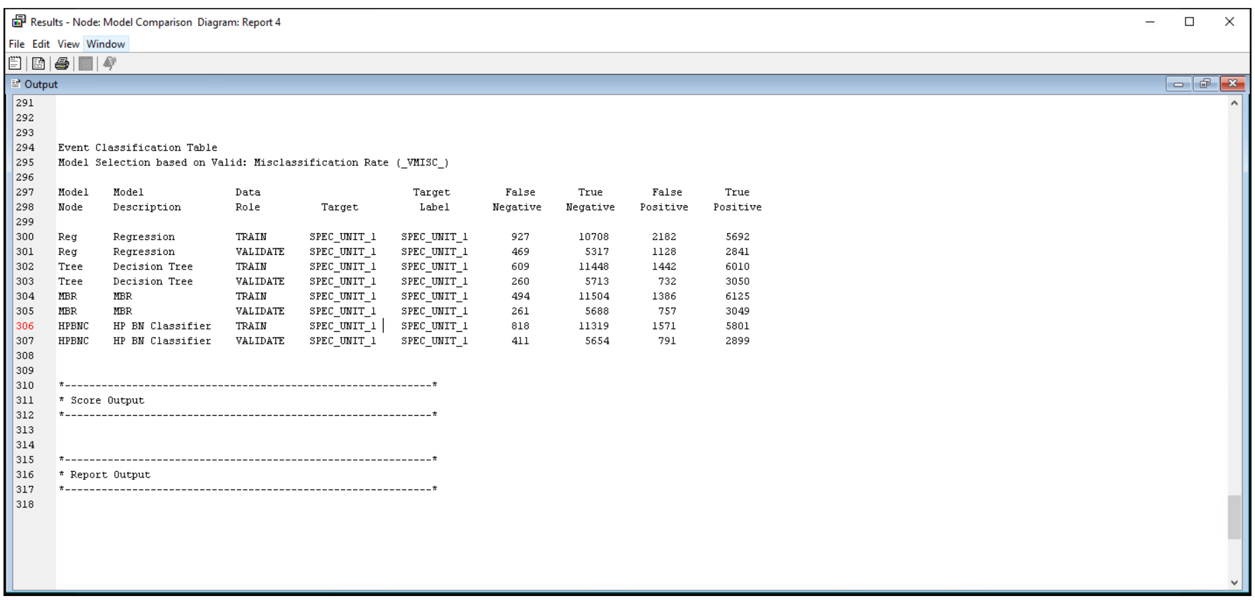
Output

```
*------------------------------------------------------------*
User:                jayshil
Date:                December 03, 2019
Time:                12:33:17
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*




Variable Summary

          Measurement    Frequency
 Role        Level         Count

TARGET      NOMINAL          1





                                                                              BINNED_KS_
                                                         _KS_PROB_               PROB_
Obs      TARGET       TARGETLABEL    _AUR_    _GINI_    KS     CUTOFF     _KS_BIN_    CUTOFF

 1     SPEC_UNIT_1    SPEC_UNIT_1    0.921     0.841    0.7     0.35        0.698      0.379






                                                               _VKS_              _VBINNED_
```



Fit Statistics
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Selected Model	Model Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
Y	MBR	MBR	0.21517	0.025338	0.20694	0.028310
	Tree	Decision Tree	0.22676	0.031412	0.22989	0.031220
	HPBNC	HP BN Classifier	0.23147	0.029112	0.22733	0.029595
	Reg	Regression	0.26602	0.032256	0.26203	0.032845



Event Classification Table
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Model Node	Model Description	Data Role	Target	Target Label	False Negative	True Negative	False Positive	True Positive
Reg	Regression	TRAIN	SPEC_UNIT_1	SPEC_UNIT_1	927	10708	2182	5692
Reg	Regression	VALIDATE	SPEC_UNIT_1	SPEC_UNIT_1	469	5317	1128	2841
Tree	Decision Tree	TRAIN	SPEC_UNIT_1	SPEC_UNIT_1	609	11448	1442	6010
Tree	Decision Tree	VALIDATE	SPEC_UNIT_1	SPEC_UNIT_1	260	5713	732	3050
MBR	MBR	TRAIN	SPEC_UNIT_1	SPEC_UNIT_1	494	11504	1386	6125
MBR	MBR	VALIDATE	SPEC_UNIT_1	SPEC_UNIT_1	261	5688	757	3049
HPBNC	HP BN Classifier	TRAIN	SPEC_UNIT_1	SPEC_UNIT_1	818	11319	1571	5801
HPBNC	HP BN Classifier	VALIDATE	SPEC_UNIT_1	SPEC_UNIT_1	411	5654	791	2899

```
*------------------------------------------------------------*
* Score Output
*------------------------------------------------------------*

*------------------------------------------------------------*
* Report Output
*------------------------------------------------------------*
```

Also, as we see below the event classification table, the training and validation data sets have performed similarly for k Nearest Neighbour. Apart from that even the False Negative and False Positive are low which indicates that the misclassification was lower than the other models indicating higher accuracy in terms of classification. This fact is supported by the low Misclassification rate as seen in fit statistics.

Conclusion

We started our data mining project by identifying a crucial healthcare issue of "Classifying patients suffering from HIV and drug abuse on the basis of severity of illness using factors such as spec_unit and source of admission". Then we went on to analyse the healthcare data where we were able to identify the variables that actually did help us to understand the relation between different variables. This we did using different graphs such bar graphs, box plots and scatter plots. This step actually helped us to understand if there was any relation between predictor and target variables or not and also between predictor variables which helped us avoid multicollinearity. This way we were able to identify the actual variables that helped us in our data mining problem. This step helped us to reduce the volume of data immensely. After that we removed the unnecessary columns and cleaned the data of trash values which were blank or symbol or invalid. After cleaning the data, we transformed the categorical variables to dummy variables that were useable by the algorithm for our analysis. After creating dummies, we partitioned the data into training, validation and test data sets. This was done so that we can train our data, validate the model and then finally test it with the test data set to check if the model did not over fit. For this, we used 4 different models that are Logistic Regression, Decision Tree, k Nearest Neighbour and Naïve Bayes. After that we compared the output of all the 4 models using a Model Comparison node that helped us identify the best model for our data. Hence, we concluded that k Nearest Neighbour was the best model as it had lowest Misclassification Rate among all the models. This was also supported by the fact that it has lowest Average Squared Error.

REFERENCES

- https://www.izenda.com/data-analytics-healthcare-industry/
- https://www.sisense.com/glossary/healthcare-analytics-basics/
- https://healthinformatics.uic.edu/blog/how-health-care-analytics-improves-patient-care/
- https://www.dataversity.net/data-analytics-important-healthcare/#
- https://www.healthcareitnews.com/news/here-are-6-major-issues-facing-healthcare-2019-according-pwc
- http://www.ihi.org/resources/Pages/IHIWhitePapers/ReducingHospitalMortalityRatesPart2.aspx