International Journal of
Health Sciences
(IJHS)

Enhancing Healthcare Claims and Membership Data Quality: SPSS Modeler Predictive Analysis

International Journal of Health Sciences

ISSN: 2710-2564 (Online)

Vol. 7, Issue No.8, pp.51 - 63, 2024 www.carijournals

Enhancing Healthcare Claims and Membership Data Quality: SPSS Modeler Predictive Analysis

Dhivya Sudeep

Application Architect, Healthcare Client

Richmond, Virginia, USA

https://orcid.org/0009-0006-0005-0405

Accepted: 30th Oct, 2024, Received in Revised Form: 14th Nov, 2024, Published: 22nd Nov, 2024

Abstract

Purpose: This paper explores the use of SPSS Modeler predictive analysis to enhance healthcare claims and membership data quality. The analysis uses advanced analytical techniques and algorithms to identify discrepancies and improve data accuracy, improving decision-making and operational efficiencies within healthcare organizations. The paper also provides insights to optimize data integrity, streamline claims processing, and ultimately improve patient care outcomes by ensuring that accurate and reliable data can sustain all healthcare operations.

Methodology: This paper explores the use of advanced data mining and predictive analytics techniques to improve the identification of claims and membership data quality. The study aims to leverage supervised learning methods, including Neural Networks and the Auto Data Prep Modeling Option, and unsupervised learning methods, utilizing cutting-edge machine learning algorithms to train models capable of detecting and addressing data quality issues.

Findings: Data Quality of an application affects various factors of an organization including operations, decision making and Planning. It therefore becomes very important to make sure that the data being stored and used is of high quality. Data must be regularly monitored and cleaned to support more informed and effective healthcare decision-making. As per a research study published by MIT Sloan, poor data quality has made companies lose around 15% to 25% of their revenues [1]. Another study found that data scientist spends around 80% of their time cleaning and correcting data leaving them with only 20%of time to perform the actual analysis [2].

Unique Contributions to Theory, Practice, and Policy: By incorporating advanced predictive analytics techniques like supervised and unsupervised learning models within SPSS Modeler, the study enhances the ability to proactively address data quality issues, streamline operations, and ensure regulatory compliance. It can also help healthcare organizations by offering. Innovative perspectives on how data mining and predictive analysis can help reshare healthcare data governance, policy development and industry-wide standards for data quality.

Keywords: *SPSS Modeler, Predictive Analysis, Healthcare Data Quality, Neural Networks, Auto Data Prep Modeling*

I INTRODUCTION:

Data Quality is important especially in the healthcare sector where data quality can help with ensuring patient care and reducing cost. As healthcare organizations increasingly depend on data to drive decisions, data that is inconsistent, incorrect and incomplete can often lead to significant challenges.

One area where data quality issues are particularly impactful is in the claims and membership domain, which plays a central role in the financial and operational aspects of healthcare delivery. Claims data including billing information, diagnosis code and treatment details form the foundation of how reimbursement processes are defined. Membership data comprising of patient demographics data, insurance details and eligibility information is used to manage operational functions such as enrollment, coverage management and benefits processing. Poor data quality in the claims and membership area may result in payments delays, errors in reimbursements and compliance related violations. For example: incorrect patient identifiers or missing information may lead to denied claims or overpayments creating financial and operational restrictions. The accuracy of these datasets is critical to ensure providers are paid properly and patients receives the right benefits, and the healthcare organization remain complaint with healthcare regulations.

As the demand for regulation and compliance increase and the demand for data-driven insights grows, addressing data quality in the claims and membership areas has become more critical than ever. Healthcare organizations that fail to effectively manage data quality risk losing trust among patients, providers, and insurers, while also facing potential financial and legal consequences. Ensuring that data is accurate, consistent, and up to date is therefore a top priority for organizations looking to streamline operations, optimize reimbursement processes, and improve patient care.

II METHODOLOGY:

SPSS Modeler is a data mining and predictive analytics platform that is designed to help businesses and organizations analyze vast amounts of data, identify hidden patterns, and make data-driven decisions. Unlike the traditional SPSS Statistics tool which focuses more on the traditional method of statistical analysis, the SPSS Modeler is designed specifically for advanced data mining, machine learning and predictive modeling. It is a powerful tool for both business analysts and data scientists to transform data into actionable insights, with a focus on automation, ease of use, and scalability.[3]

In the below Case study, we are going to take an elaborate look at how SPSS Modeler can be used to identify Healthcare Data Quality issues using supervised and unsupervised learning models.

Supervised learning models uses machine learning approach where the model is trained using labeled data. This technique works best when the issues are already known. The data can be trained with the known results to create a trained environment. The Common algorithms in supervised

International Journal of Health Sciences
ISSN: 2710-2564 (Online)
Vol. 7, Issue No.8, pp.51 - 63, 2024 www.carijournals

learning include decision trees, logistic regression, support vector machines (SVM), random forests, and neural networks.

A. Scenario 1 - Duplicate PCP Issue:

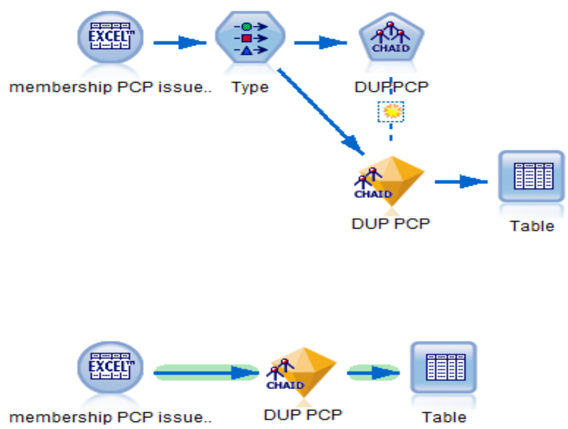


Figure 1: Supervised learning model for duplicate PCP Issue

Figure 1 shows two streams, the first stream shows the training and Model creation where the training data set has the target field which will train the model to identify good and bad records. The second stream is the already created model executed with a test dataset that does not have the target field. The model when executed will create the desired target field with the score. The model here is trained to identify if the same member has a duplicate PCP for a particular coverage period.

The Model output is displayed below - the model has two output fields $R-DUP-PCP which shows if the member has a duplicate PCP or not and the field $RC-DUP-PCP shows the scoring for each record which shows the confidence score of the model in producing the result.

Output Field Score

MEMBER ID	GRP	SUB GROUP	PCP EFF DATE	PCP TERM DATE	PROVIDER ID	$R-DUP PCP	$RC-DUP PCP
A12345678	L26578	123456	2024-01-01	2024-10-31	786534289	N	1.000
A12345678	L26578	123456	2024-11-01	2024-12-31	372537852	N	1.000
A67252465	S87654	836437	2024-01-01	2024-12-31	326859872	Y	1.000
A67252465	S87654	836437	2024-01-06	2024-12-31	189279263	Y	1.000
A67819236	I67544	314256	2024-01-01	2024-10-31	273873765	N	1.000
A67819236	I67544	314256	2024-11-01	2024-12-31	362537578	N	1.000

Figure 2: Scenario 1 Model output

B. Scenario 2 - Finalized claims having negative paid amounts:

International Journal of Health Sciences

ISSN: 2710-2564 (Online)

Vol. 7, Issue No.8, pp.51 - 63, 2024 www.carijournals

We will now look at another scenario which helps to identify incorrect negative paid amounts calculated for a claim under certain conditions. To identify if the claim having negative paid amount is a valid record or not, we do not have any field in our history data to confirm this. (Assuming this issue has never occurred in the past, but source sends negative paid amounts for finalized claims).

To identify this issue, we built the below model based on neural networks which worked accurately for our use case.

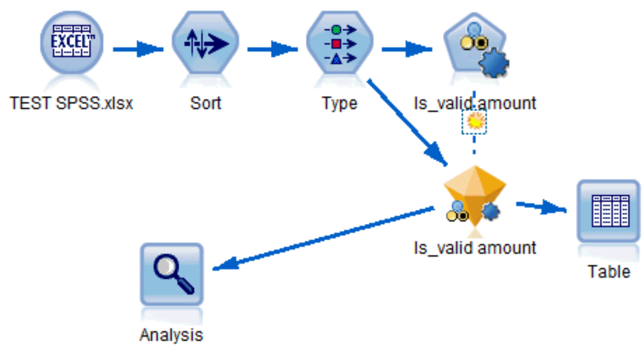
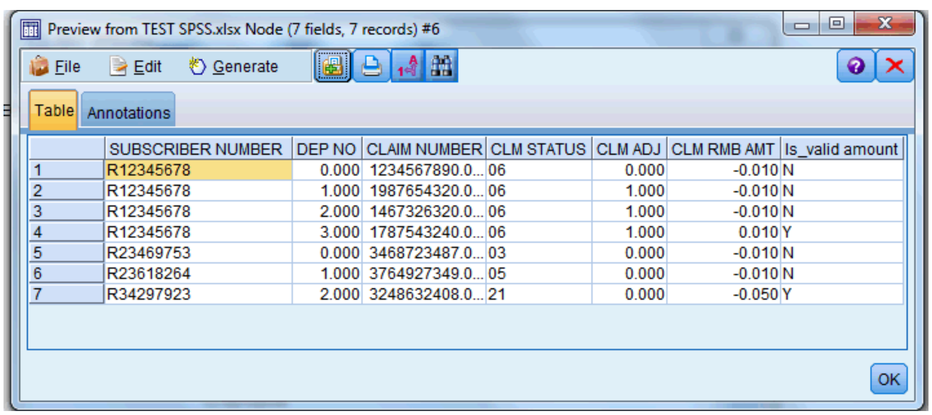


Figure 3: Supervised learning model for negative claim amount

In the above model, TEST.SPSS.xlsx is the Input file which was designed to train the Model. As seen below, we have used the below fields Subscriber, Claim, Dependent Number, claim status code, Claim Adjustment, Claim Reimbursement amount and Is_Valid amount.



Preview from TEST SPSS.xlsx Node (7 fields, 7 records) #6

	SUBSCRIBER NUMBER	DEP NO	CLAIM NUMBER	CLM STATUS	CLM ADJ	CLM RMB AMT	Is_valid amount
1	R12345678	0.000	1234567890.0...	06	0.000	-0.010	N
2	R12345678	1.000	1987654320.0...	06	1.000	-0.010	N
3	R12345678	2.000	1467326320.0...	06	1.000	-0.010	N
4	R12345678	3.000	1787543240.0...	06	1.000	0.010	Y
5	R23469753	0.000	3468723487.0...	03	0.000	-0.010	N
6	R23618264	1.000	3764927349.0...	05	0.000	-0.010	N
7	R34297923	2.000	3248632408.0...	21	0.000	-0.050	Y

Figure 4: TEST.SPSS Input data set

Is_valid amount field is created manually to train the model to identify valid and invalid records. When Claim Status is 06, 03, 05(Finalized claims) the reimbursement amount cannot be negative. Claim Status 21 is a reversal and so having a negative amount is a valid scenario. Based on the scenario, we have given a Y/N for the field and have mentioned the Type for this field as Output.

Below is the output of the model that was run using the TEST.SPSS model.

International Journal of Health Sciences

ISSN: 2710-2564 (Online)

Vol. 7, Issue No.8, pp.51 - 63, 2024

	SUBSCRIBER NUMBER	DEP NO	CLAIM NUMBER	CLM STATUS	CLM ADJ	CLM RMB AMT	Is_valid amount	$XF-Is_valid amount	$XFC-Is_valid amount
1	R12345678	0.000	1234567890.0...	06	0.000	-0.010	N	N	0.992
2	R12345678	2.000	1467326320.0...	06	1.000	-0.010	N	N	0.992
3	R12345678	3.000	1787543240.0...	06	1.000	0.010	Y	Y	0.982
4	R12345678	1.000	1987654320.0...	06	1.000	-0.010	N	N	0.992
5	R34297923	2.000	3248632408.0...	21	0.000	-0.050	Y	Y	0.988
6	R23469753	0.000	3468723487.0...	03	0.000	-0.010	N	N	1.000
7	R23618264	1.000	3764927349.0...	05	0.000	-0.010	N	N	1.000

Figure 5: Model output using TEST.SPSS Input

The model generated two fields at the end, one is $XF-Is_Valid amount and $XFC-Is_valid amount.

$XF-Is_valid amount is the output value predicted based on the predictors specified (Claim Status code and Claim Reimbursement amount)

$XFC-Is_valid amount is the Confidence Level of the Output generated.

If we see fields Is_valid amount and $XF-Is-valid amount - the values are exactly same which means, the model has predicted the values correctly.

The next step is to test with real-time data to see if the Model works fine.

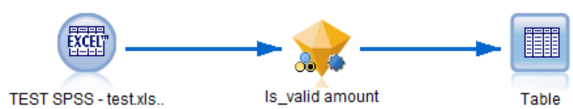


Figure 6: Real-time data testing model

We created a new test data which does not have the field Is_valid amount.

Preview from TEST SPSS - test.xlsx Node (6 fields, 7 records) #3

File Edit Generate

Table Annotations

	SUBSCRIBER NUMBER	DEP NO	CLAIM NUMBER	CLM STATUS	CLM ADJ	CLM RMB AMT
1	R32476923	0.000	1276732469.0...	21	0.000	-0.010
2	R32476923	1.000	1236581423.0...	06	1.000	-0.010
3	R32476923	2.000	1232187354.0...	06	1.000	-0.010
4	R32476923	3.000	4375875345.0...	03	1.000	0.010
5	R23475923	0.000	4752375475.0...	03	0.000	-0.010
6	R23475923	1.000	3658259473.0...	05	0.000	-0.010
7	R23475923	2.000	3794529754.0...	21	0.000	-0.050

OK

International Journal of Health Sciences

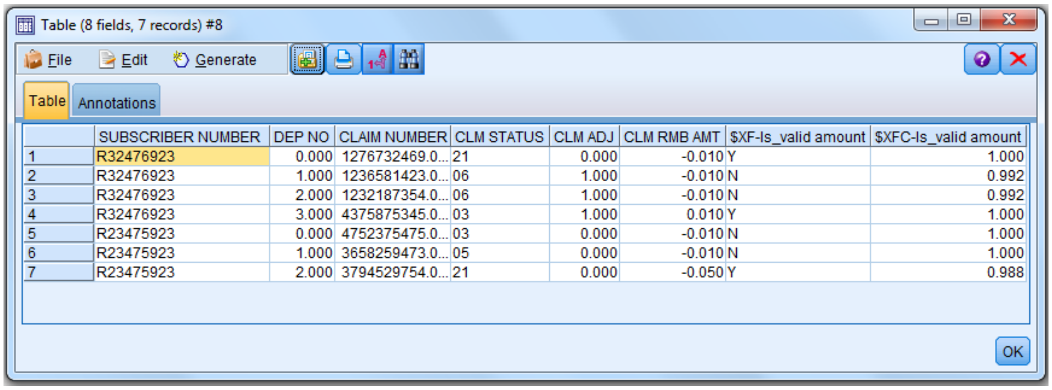
ISSN: 2710-2564 (Online)

Vol. 7, Issue No.8, pp.51 - 63, 2024

Figure 7: TEST_SPSS-test Input dataset

The Model Nugget "Is_valid amount" shown in Figure 6 is being copied from the trained model.

The model was then executed and below is the output of the model shown in Figure 6



	SUBSCRIBER NUMBER	DEP NO	CLAIM NUMBER	CLM STATUS	CLM ADJ	CLM RMB AMT	$XF-Is_valid amount	$XFC-Is_valid amount
1	R32476923	0.000	1276732469.0...	21	0.000	-0.010	Y	1.000
2	R32476923	1.000	1236581423.0...	06	1.000	-0.010	N	0.992
3	R32476923	2.000	1232187354.0...	06	1.000	-0.010	N	0.992
4	R32476923	3.000	4375875345.0...	03	1.000	0.010	Y	1.000
5	R23475923	0.000	4752375475.0...	03	0.000	-0.010	N	1.000
6	R23475923	1.000	3658259473.0...	05	0.000	-0.010	N	1.000
7	R23475923	2.000	3794529754.0...	21	0.000	-0.050	Y	0.988

Figure 8: Model output shown in figure 6

If we see the output fields $XF-Is_valid amount and $XFC-Is_valid amount, the values are coming correctly.

All finalized status claims having a negative amount is reported as N and the rest as Y.

Now that we confirm the Model is working fine, we can publish this model to a DB2 scoring adapter. Once the Model is published to scoring adapter and an SQL call can be written to the Model to fetch probability scores.

For example:

If probability score > 0.7

"Valid Record. No action required"

Else

If Probability score is < 0.5

"Invalid Amount for Finalized status code"

Inform source team about the data issue and request for a corrected file.

C. Scenario 3 - Validating accurate value for patient Sex field:

In the next scenario, we will explore a situation where we are uncertain about which modeling technique to apply and how the Auto Data Prep Modeling option can assist in determining the appropriate approach.

International Journal of Health Sciences

ISSN: 2710-2564 (Online)

Vol. 7, Issue No.8, pp.51 - 63, 2024 www.carijournals

To do this, let's take an example to validate patient sex field for incorrect values coming from source. The acceptable value for sex is "Male/Female" which is the value further used for downstream analysis but many source systems send "M/F" instead of "Male/Female".

Here to train the model, we will use the Sex field from the Target table as an output variable in the Training Model as it would have the correct value. We are using two Input files here, one that has acceptable values for SEX("Male/Female") and one that has incorrect values for SEX("M/F").

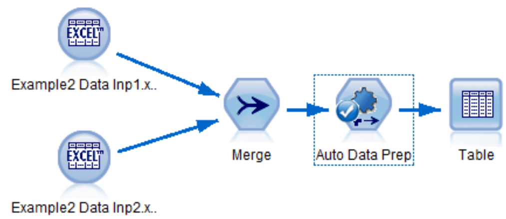


Figure 9: Auto Data Prep Model

Below are the contents from Input 1 having incorrect values for Sex. Here Inp_sex is used as an Input which has incorrect values (wrong format) coming from source.

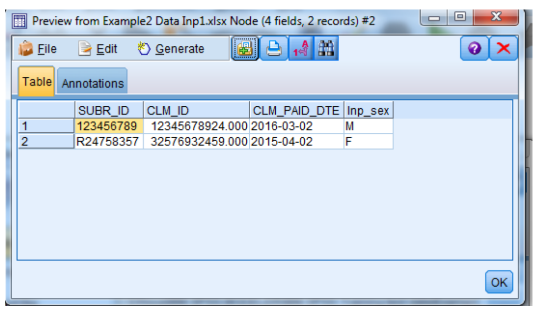


Figure 10: Input file with Incorrect Sex value

Below are the contents from Input 2 having accurate values for sex. This file has history data taken from target table and Out_sex is the correct value for Sex.

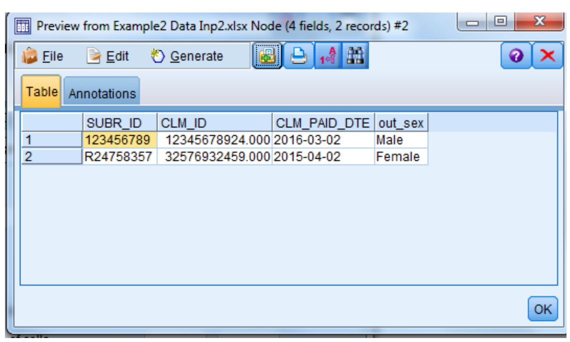


Figure 11: Input File from Target table having correct Sex value

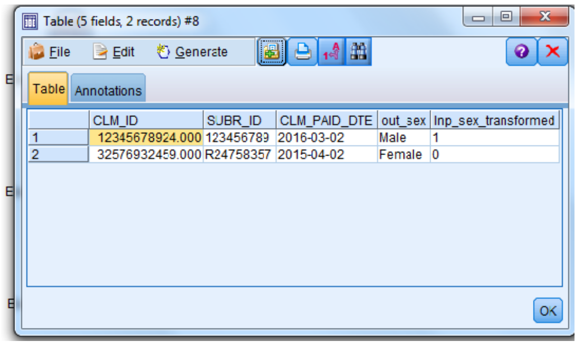
Here Inp_sex is specified as Input and Out_sex is specified as Target value.

International Journal of Health Sciences

ISSN: 2710-2564 (Online)

Vol. 7, Issue No.8, pp.51 - 63, 2024

Below is the output when the model is executed:



	CLM_ID	SJBR_ID	CLM_PAID_DTE	out_sex	Inp_sex_transformed
1	12345678924.000	123456789	2016-03-02	Male	1
2	32576932459.000	R24758357	2015-04-02	Female	0

Figure 12: Output of Model created in Figure 9

A new field Inp_sex_transformed is generated which says that the Input field is just transformed to a new value, and it has assigned 1 for Male and 0 for Female.

Now, let's test this Model with an input file that has few records with value "M/F" and few records as "Male/Female".

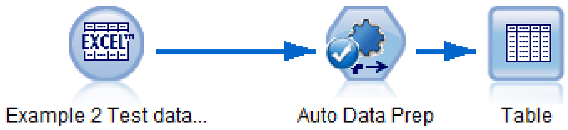


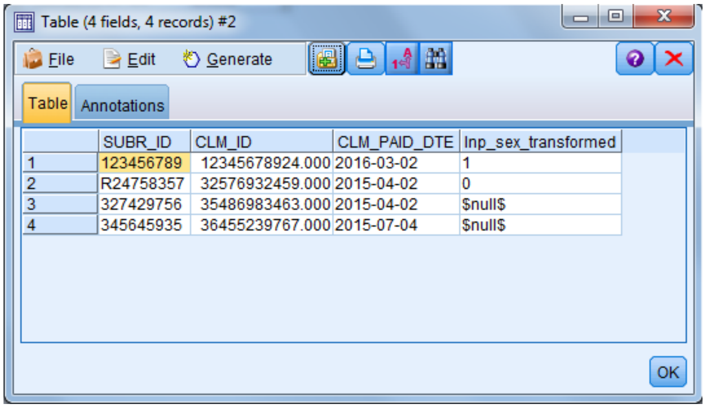
Figure 13: Testing Auto Data Prep Model with trained model

Below is the input file used for this testing

	SUBR_ID	CLM_ID	CLM_PAID_DTE	Inp_sex
1	123456789	12345678924.000	2016-03-02	M
2	R24758357	32576932459.000	2015-04-02	F
3	327429756	35486983463.000	2015-04-02	Male
4	345645935	36455239767.000	2015-07-04	Female

Figure 14: Input file for Auto Data Prep Model

Below is the output of the model



	SUBR_ID	CLM_ID	CLM_PAID_DTE	Inp_sex_transformed
1	123456789	12345678924.000	2016-03-02	1
2	R24758357	32576932459.000	2015-04-02	0
3	327429756	35486983463.000	2015-04-02	$null$
4	345645935	36455239767.000	2015-07-04	$null$

Figure 15: Auto Prep Model output results

The output field "Inp_sex_transformed" has returned three values "0, 1 and '$null$'

Whenever the value is 0 – the output field in the table can be auto changed to Male, when 1, it can be changed to Female and for Null – no action required as it is the correct value

Unsupervised learning models uses machine learning approach where the model is trained on data that does not have labeled outcomes. Instead, unsupervised learning focuses on identifying patterns or relationships within the data itself. Unsupervised learning techniques works best for cases where the data is new or still unexplored, tools provide algorithms to identify anomalies using unlabeled data.

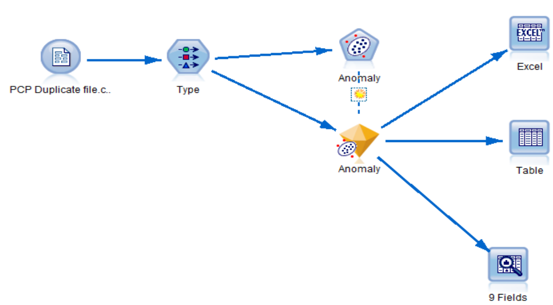


Figure 16: Unsupervised Learning Model

The stream above is an unsupervised model stream where there is no target field to train the Model. It uses machine learning techniques to identify anomalies giving us the opportunity to explore data that might be incorrect. With Unsupervised learning, it is possible to create larger and more complex models than with the supervised models. Un-supervised learning can be used for bridging the gap between input and output observations.

Shown below is how the output of an unsupervised model (figure 9) would look like giving us an opportunity to identify data issues that are not known and thereby helping to enhance data quality at a large scale.

International Journal of Health Sciences
ISSN: 2710-2564 (Online)
Vol. 7, Issue No.8, pp.51 - 63, 2024

Field	Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
PROD_OFRG_KEY		Nominal	--	--	--	--	--	17	30
MBR_PROD_ENROLLMNT_EF...		Continuous	2012-01-08	2015-01-10	--	--	--	--	30
MBR_PROD_ENROLLMNT_T...		Flag	--	--	--	--	--	1	30
MBRSHP_SOR_CD		Continuous	802	820	811.333	6.915	-0.529	--	30
SBSCRBR_CD		Nominal	--	--	--	--	--	5	30
MBR_SQNC_NBR		Continuous	10	50	29.333	13.113	-0.066	--	30
GRP		Nominal	--	--	--	--	--	4	30
SUB_GRP		Nominal	--	--	--	--	--	6	30
PCP_EFCTV_DT		Continuous	2012-01-08	2016-01-01	--	--	--	--	26
PCP_TRMNTN_DT		Nominal	--	--	--	--	--	10	30
PROV_ID		Continuous	1530144	76548441	31292862.100	34145957.504	0.475	--	30
$O-Anomaly		Flag	--	--	--	--	--	2	30

Figure 17: Unsupervised Model output

III. FINDINGS:

Data Mining and predictive analytics has many applications across the Healthcare Industry to improve Data quality. Below are some of the use cases and findings of using supervised and unsupervised learning models.

A. Data Quality Management:

SPSS Modeler provides robust tools for cleaning and transforming claims and membership data which is crucial for maintaining high quality accurate data. It automatically handles missing values, data exception and data inconsistency ensuring the health data maintained is of high quality. SPSS Modeler can help standardize and normalize data coming from different source systems to create consistent and useable datasets that can be used for downstream analysis. For example: different systems use different coding practices for procedures and diagnosis codes, SPSS can help aggregate these data from different sources and ensure they are converted into a single acceptable format which makes it easy for downstream analysis.

B. Machine Learning and Model Building:

SPSS Modeler uses advanced machine learning algorithms to build predictive models for claims approval. It can help with automatic prediction of whether a claim will be approved or

International Journal of Health Sciences

ISSN: 2710-2564 (Online)

Vol. 7, Issue No.8, pp.51 - 63, 2024 www.carijournals

denied thereby streamlining claims processing and reducing manual review workload. Custom predictive models can be built that can help with predicting high-cost claims, identifying underutilized services or to forecast resource needs for claims management teams.

C. Predictive analysis for claims processing:

- SPSS Modeler presents a wide variety of machine learning algorithms, including decision trees, neural networks, support vector machines (SVM), logistics regression, and random forests to develop predictive models to identify fraudulent claims by recognizing suspicious patterns such as unusually high billing amounts, inconsistent diagnosis, or duplicative claims. Identifying fraudulent claims early can help reduce financial loss for the healthcare organization.
- The SPSS predictive models can also help to predict claims that are likely to be denied or require additional investigation by analyzing historical data on claims processing.
- The SPSS predictive models can also help predict the cost of healthcare claims based on the diagnosis made, treatment provided, and patient demographics, which can help healthcare providers manage costs more efficiently and ensure accurate budgeting for claims reimbursements.

D. Data aggregation and unstructured data processing:

SPSS Modeler enables the aggregation and integration of claims data from multiple sources like Electronic Health Records (EHR), billing data, insurance data into a single coherent dataset. This allows for a complete view of patient claims data providing insights across the entire patient ecosystem. SPSS Modelers capability of natural language processing can help analyze unstructured data like medical records, clinical notes and physician documentation to extract valuable documentation for claims processing and membership management.

E. Advanced analytics for Member Management:

SPSS Modeler can analyze demographics, medical history and claims data to determine potential member profiles with health-related risk. The analysis data can be used to identify members that might be at a risk of requiring high-cost care, enabling preventive and proactive care management. The prediction models can also help with patient sentiment analysis based on patient language, speech pattern and emotional tone to identify if the patient suffers from anxiety, depression or any kind of mental health issue or is on the verge of experiencing the same, thereby allowing for more personalized care for patients.

V RECOMMENDATION:

Based on the findings from the SPSS Modeler predictive analysis, the following recommendations are proposed to further enhance the quality of healthcare claims and membership data. Predictive models need to be refined to suit each organizations data quality needs and the models created should not be a one-time solution, establishing a continuous monitoring system that can continue to identify new potential errors can help keeping the data quality high over time. Healthcare data is often stored in different sources including medical health records, insurance data and billing systems. It is important to combine the data coming from different sources into a common standard format which can help with improving the overall data quality. Using SPSS Modeler to merge the data sources will provide a comprehensive dataset for analysis. It is critical to train the Healthcare staff about the importance of accurate data collection and how poor data can impact patient care and other administrative processes. Providing training to how to collect data as per data entry standards, commonly done errors, and how predictive tools can be used to improve data management practices. The predictive models developed in the study showed examples of how SPSS Modeler can be used to successfully identify fraudulent claims and membership discrepancies. To further elevate this, models can be fine-tuned just for fraud detection by incorporating additional features like claim frequency patterns, patient history and provider behavior analysis. As Healthcare policies progress, the predictive models created should be adaptable to these changes. Updating the models regularly based on changes in data patterns or healthcare regulations will help to ensure the system remains effective over time.

VI CONCLUSION:

To conclude, SPSS Modeler presents a powerful approach in diagnosing and correcting claim and membership data quality issues. By using advanced modeling techniques such as neural networks based supervised learning, auto data prep modeling, when we are unsure on what modeling technique to use and unsupervised learning for large and complex datasets, healthcare organizations can identify hidden data errors, outliers and issues that may go unnoticed. This enables accurate data that can be presented to take better health decisions, optimize operational processes and enhance overall quality of data at a large scale. Ultimately, applying SPSS Modeler tools improves the accuracy of claims and membership data to supports better outcomes for both businesses and their customers.

REFERENCES:

[1] Thomas C. Redman, 2016, https://sloanreview.mit.edu/article/seizing-opportunity-in-data-quality/

[2] https://www.talend.com/resources/machine-learning-data-quality/#:~:text=As%20per%20a%20research%20study%20published%20in,their%20revenues%20due%20to%20poor%20data%20quality.

[3] Driving Innovation and Efficiency: The Key to Success with Information Observability Adoption in Enterprises - Latest Tech News | Gadgets | Opinions | Reviews. https://metatec.net/driving-innovation-and-efficiency-the-key-to-success-with-information-observability-adoption-in-enterprises/

[4] IBM SPSS Neural Networks 20 – User Guide, © Copyright IBM Corporation 1989, 2011.

[5] Ahmed Al-Imam, December 2018

https://www.researchgate.net/publication/329029699_A_Gateway_Towards_Machine_Learning_Predictive_Analytics_and_Neural_Networks_in_IBM-SPSS_SPSS_v24