

Exploring SAS[®] Enterprise Miner[™] Special Collection



LAB REPORT ASSIGNMENT

P2862667

De Montfort University Leicester



Department Of Data Analytics

Project work on

“PREDICTION OF BANK DATA SETS”

Project Submitted By

PRIYADHARSHINI MANICKAM(P2862667)

Under the guidance and supervision of

Ali Khan

19APRIL 2024

CONTENTS

Abstract.....	2
Business Problem.....	2
Methodology - SEMMA framework.....	2
Data Exploration.....	3
Data partition creation of model sets.....	3
Data Modification.....	4
Data Modelling.....	4
Regression.....	4
Development of models.....	4
Model performance.....	4
Chosen Regression equation.....	4
Neural Network.....	5
Development of models.....	5
Model performance.....	5
Overfitting analysis.....	5
Neural network architecture of best model.....	5
Final weights results discussion of best model.....	5
Decision Tree.....	6
Development of models.....	6
Performance of models.....	6
Critical path of best model.....	6
Target path of interest.....	6
Overfitting analysis.....	6
Analysis of the best model.....	7
Conclusion.....	8

Abstract

In today's dynamic banking landscape, harnessing the power of data mining techniques has become imperative for uncovering valuable insights and enhancing decision-making processes. This study explores the application of data mining methodologies to analyze bank datasets, aiming to extract actionable knowledge and drive strategic initiatives. The primary objective of this research is to employ SAS Enterprise Miner to delve into the vast reservoirs of banking data and unearth hidden patterns, trends, and correlations. By leveraging advanced analytics, including classification algorithms, clustering methods, and predictive modeling techniques, we seek to address critical challenges faced by banks, such as customer churn prediction, fraud detection, and targeted marketing. Direct marketing campaigns represent targeted promotional efforts where companies communicate directly with potential customers to promote products or services. Unlike traditional mass advertising, direct marketing focuses on reaching specific individuals or groups through various channels such as email, direct mail, social media, and phone calls. These campaigns often employ personalized messages tailored to the preferences and behaviours of the target audience, aiming to elicit a direct response or engagement. Through careful data analysis and segmentation, companies can refine their messaging and targeting strategies to maximize effectiveness and return on investment. Direct marketing campaigns enable businesses to establish direct connections with consumers, build brand awareness, drive sales, and foster long-term customer relationships. However, they also require meticulous planning, creative execution, and adherence to privacy regulations to ensure ethical and successful outcomes. As a data miner, you've been tasked by a Portuguese banking institution to analyze the BANK dataset, which comprises direct marketing campaigns primarily conducted through phone calls. These campaigns often necessitated multiple contacts with the same client to ascertain whether they would subscribe ('yes') or not ('no') to the bank term deposit. Your objective is to compile a technical report outlining your findings, including the best model for identifying the factors contributing to a subscription to the bank term deposit. The dataset holds 4,521 records with 17 attributes such as Age, Job, Marital Education, Default, Balance, Housing, Loan, Contact, Day_of_week, Month, Duration, Campaign, P days, Previous, P outcome, Y.

Business Problem

A common business problem that can be addressed using SAS Enterprise Miner is customer churn prediction for a telecommunications company. Customer churn, or attrition, refers to the phenomenon where customers stop doing business with a company. Predicting churn is crucial for businesses, as acquiring new customers is often more expensive than retaining existing ones.

- Data Preparation
- Feature Engineering
- Model Building
- Model Evaluation
- Deployment
- Monitoring and Optimization

Methodology - SEMMA framework

The SEMMA framework, which stands for Sample, Explore, Modify, Model, and Assess, is a methodology commonly used in data mining and predictive analytics. Originally developed by SAS, it provides a structured approach to solving business problems using data.

- Sample: In the Sample phase, the focus is on data collection and sampling.
- Explore: Once the data is collected, the Explore phase involves exploratory data analysis (EDA) to gain insights and understand the characteristics of the data.
- Modify: In the Modify phase, the focus shifts to data preprocessing and feature engineering.
- Model: The Model phase involves building predictive models using machine learning algorithms or statistical techniques.
- Assess: In the Assess phase, the performance of the predictive models is evaluated using relevant metrics and validation techniques.

Data Exploration

Data exploration is a crucial step in the data analysis process that involves understanding the structure, content, and relationships within a dataset. It aims to gain insights, detect patterns, identify anomalies, and formulate hypotheses that can guide subsequent analysis and decision-making. Here are some common techniques and approaches used in data exploration:

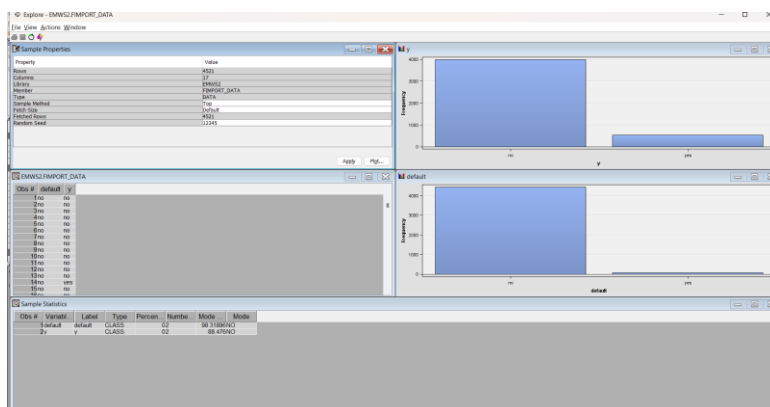
- Importing Data
- Summary Statistics
- Data Grid View
- Data Visualization
- Correlation Analysis
- Variable Selection
- Data Quality Analysis
- Interactive Exploration
- Model Comparison

Model Roles- Inputs, output, and rejected variables:

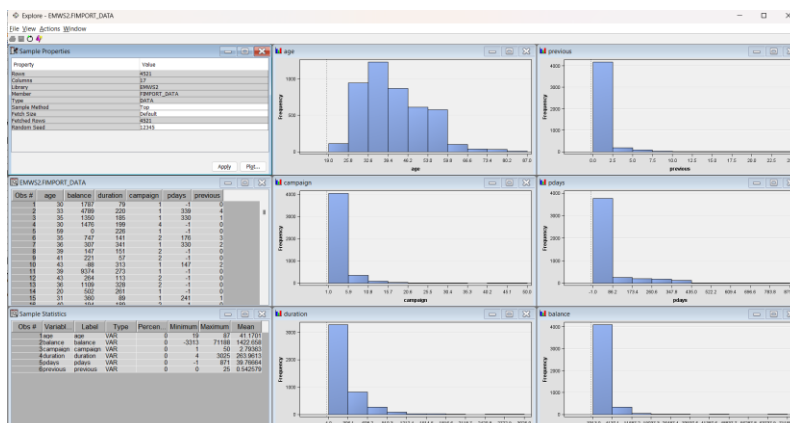
Inputs – Click on statistics to see the present missing values of maximum number of the maximum percentage one was in well is 21. If it is of a 50 that you were rejected by over 50, you were rejected.

Data types:

1. **Binary classification**, the term "binary" refers to a classification problem where there are only two possible outcomes or classes. These outcomes are typically represented as 0 and 1, or as negative (0) and positive (1).



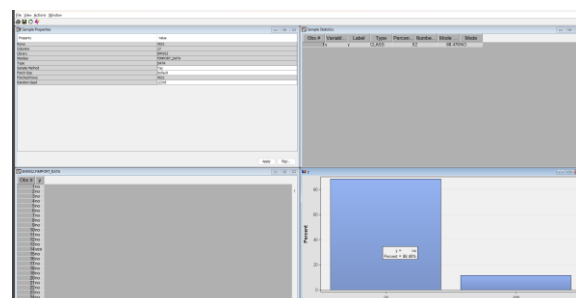
2. **Interval**, refers to a type of variable measurement level that represents continuous numerical values. Interval is six variables and six values.



[illegible][illegible]

We know that input, targets, and the data rules because the percentage is greater than 86.60 which means higher than 50. So it's rejected.

We check the target variable and explore and change to the percentage Y =no Percent=88.48%, Y=yes Percent=11.52% it's an Imbalanced dataset.



Missing data

Go to edit variables and click explore to see the white empty plots otherwise click on the statistics to see the percentage of missing values

Variance and standard deviation

Go to edit variables and explore click on the statistics to see the variance and standard deviation.

Formula :

The population standard deviation formula is given as:

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (X_i - \mu)^2}$$

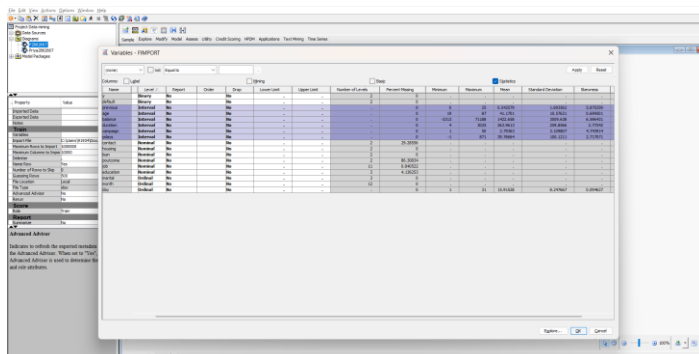
Here,

- σ = Population standard deviation symbol
- μ = Population mean
- N = total number of observations

Skewness

Go to edit variables select the Interval and check the skewness of the variable's columns .

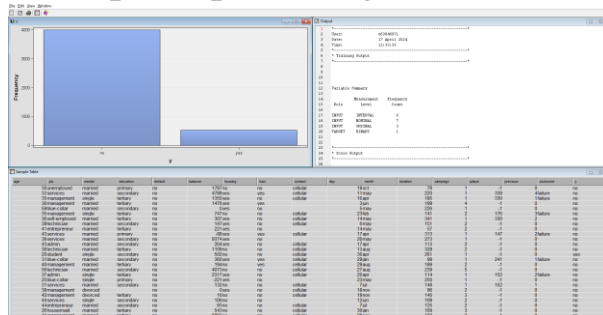
Now we can see the above results for Missing data, Variance Standard deviation and Skewness.



Variable	Mean	Std. Deviation	Variance	Skewness	Kurtosis
Age	35.214	12.158	147.816	-.054	2.940
Sex	1.500	.500	.250	.000	-.000
Education	12.500	2.500	6.250	.000	-.000

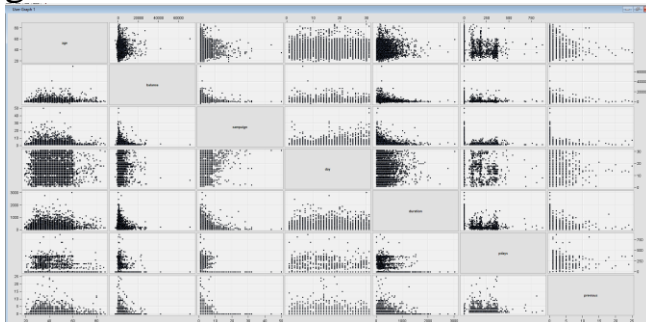
Outliers

Create ~Graph Explore ~right click to run the code for the result



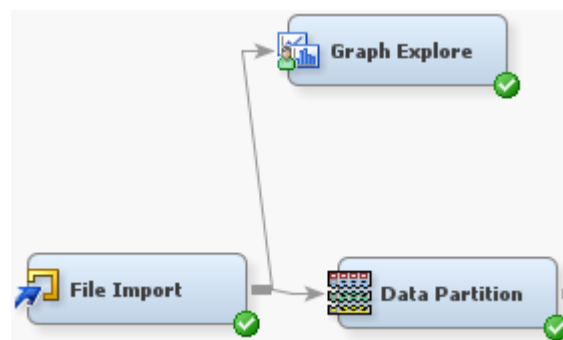
Multicollinearity

Create ~Graph Explore ~ click to run the code for the result and go to Plot click the matrix to see the multicollinearity plots.



The selected sampling method

Create ~Data Partition and connect to File Import.



Development of models

Identifying the missing values, outliers, skewed distributions, or variables that may need the transformation.

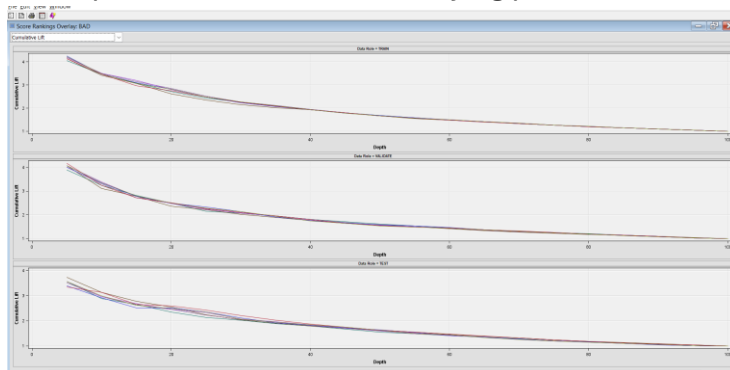
Model performance

Table X: Regression models performance

MODEL VARIATIONS	ROC INDEX	CU. LIFT	SCOPE %	TRUE - %	FALS E - %	TRUE + %	FALS E + %
Reg-Imp-Default	0.859	3.29	20	1147	108	47	53
Reg-Imp-Backward	0.859	3.68	20	1159	105	50	41
Reg-Imp-TR-Backward-R2	0.859	3.29	20	1147	108	47	53
Reg-Imp-TR-Default	0.859	3.32	20	1159	105	50	41
Reg-Imp-TR-Backward-Chi2	0.827	3.16	20	1153	110	45	47
Reg-Imp-TR-Backward	0.827	3.32	20	1159	105	50	41
Reg-Imp-Backward-R2	0.826	3.16	20	1153	110	45	47
Reg-Imp-Backward-Chi2	0.826	3.32	20	1159	105	50	41

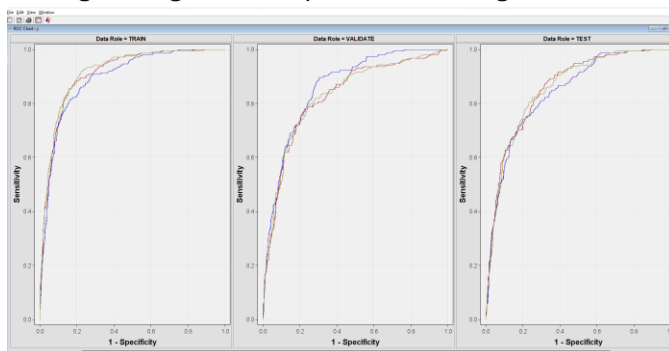
Example, cumulative lift diagram

The cumulative lift diagram provides a visual representation of the performance of the predictive model in identifying positive cases compared to random selection.



ROC curves diagram

curve provides a graphical representation of the performance of the predictive model in distinguishing between positive and negative cases across different classification thresholds.



Chosen Regression equation

First, we need to add all the variables otherwise. I'll just write the significant variables or the simplified regression model, and you will include only the significant variables. Because there are many variables that can be very long equations. So just simplify the equation and include only the significant variables. Not significant is significant. So we can simplify the equation because with regression it's a nominal ordinal. Categorical and general indicator variable variables will be encoded.

TARGET	OUTCOME	TARGET PERCENTAGE	OUTCOME PERCENTAGE	FREQUENCY COUNT	TOTAL PERCENTAGE
NO	NO	91.6930	96.583	1159	85.53
YES	NO	8.3070	67.741	105	7.7491
NO	YES	45.054	3.4167	41	3.0258
YES	YES	54.945	32.258	50	3.6900

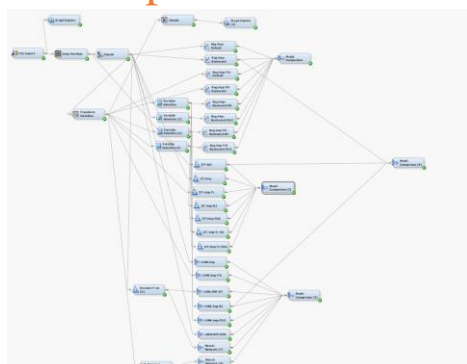
DATA ROLE=VALIDATE TARGET=Y TARGET LABEL=Y

False	True	False	True
Negative	Negative	Positive	Positive
108	1147	53	47

Decision Tree

The general form of this modeling approach. Once the relationship is extracted, then one or more decision rules that describe the relationships between inputs and targets can be derived. Rules can be selected and used to display the decision tree, which provides a means to visually examine and describe the tree-like network of relationships that characterize the input and target values. Decision rules can predict the values of new or unseen observations that contain values for the inputs, but that might not contain values for the targets.

Development of models

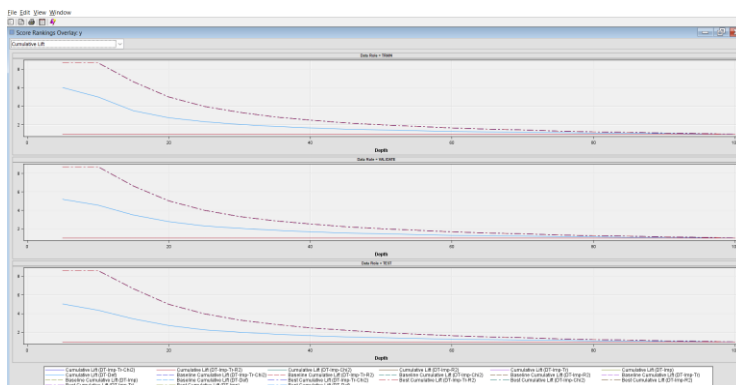


Performance of models

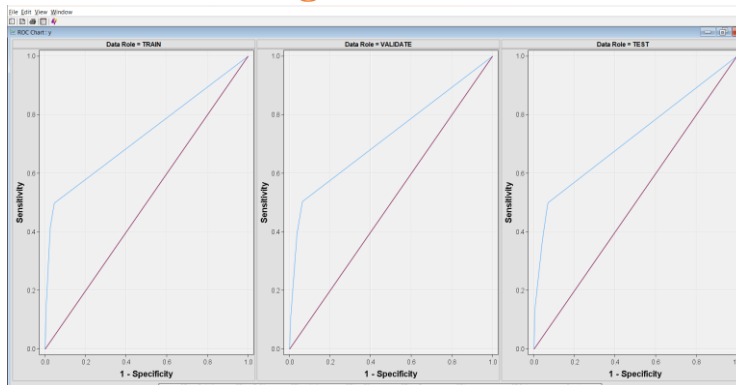
Table X: Decision tree models performance

MODEL VARIATIONS	ROC INDEX	CU. LIFT	SCOPE %	TRUE - %	FALS E - %	TRUE + %	FALS E + %
DT-Def	0.723	2.75	20	1154	94	61	46
DT-Imp	0.5	1	20	1200	155	0	0
DT-Imp-chi2	0.5	1	20	1200	155	0	0
DT-Imp-R2	0.5	1	20	1200	155	0	0
DT-Imp-Tr	0.5	1	20	1200	155	0	0
DT-Imp-Tr-chi2	0.5	1	20	1200	155	0	0
DT-Imp-Tr-R2	0.5	1	20	1200	155	0	0

Example,
cumulative lift diagram

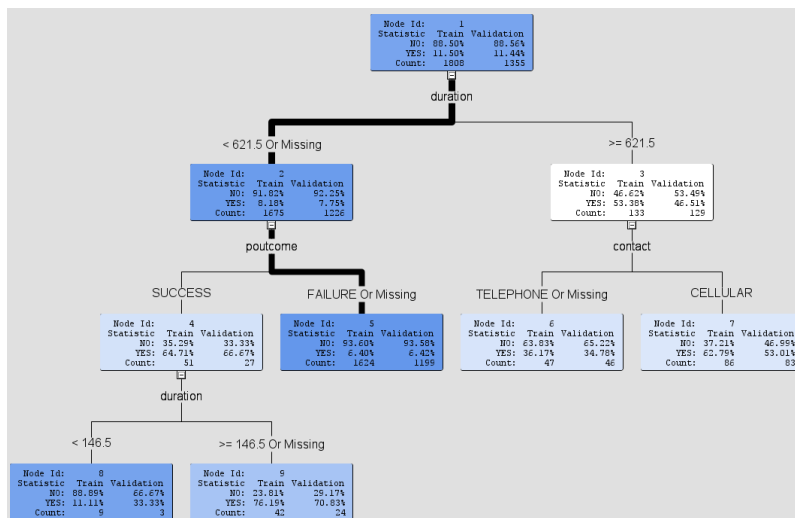


ROC curves diagram



Critical path of best model

The critical one ,the thick line .So these represent that critical path .So we can see ID Number five.



Go to the view model nodes. So this is correct. Go to the view model nodes.so this is the F then role for the critical path or decision true for the critical path.

Node = 5

if poutcome IS ONE OF: FAILURE or MISSING

AND duration < 621.5 or MISSING

then

Tree Node Identifier = 5

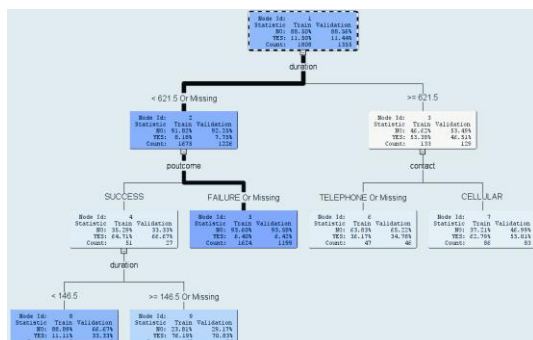
Number of Observations = 1624

Predicted: y=yes = 0.06

Predicted: y=no = 0.94

Target path of interest

That's mean when the output equal to one.so here the output is equal to one. There's also no silver cheque when one you have highest percentage when class equal to one.



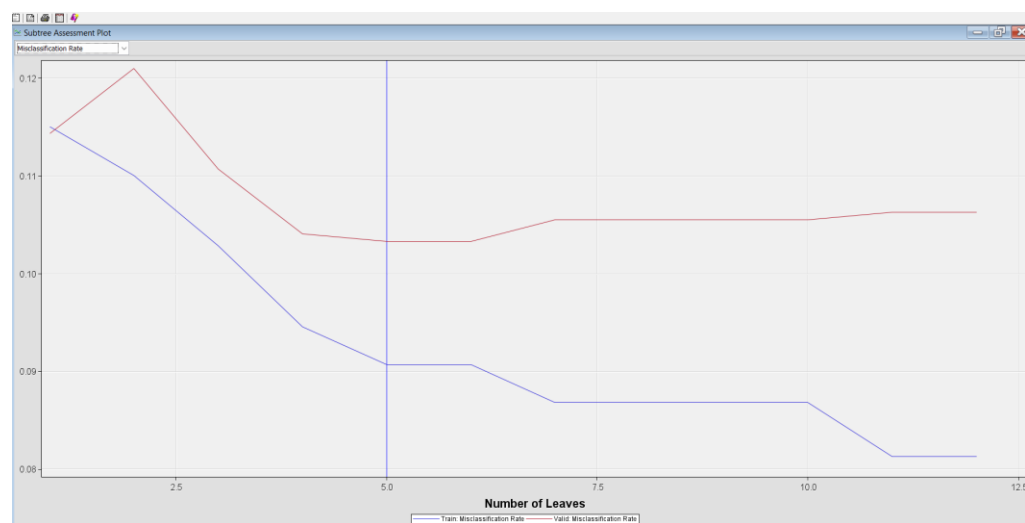
Overfitting analysis

Now will go to view. And I'll say it's good to include the variable input of variable importance.

Variable Importance				
Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance
duration	duration	2	1.0000	1.0000
outcome	outcome	1	0.7704	0.7444
contact	contact	1	0.2758	0.2111
balance	balance	0	0.0000	0.0000
housing	housing	0	0.0000	0.0000
age	age	0	0.0000	0.0000
previous	previous	0	0.0000	0.0000
pdays	pdays	0	0.0000	0.0000
education	education	0	0.0000	0.0000
default	default	0	0.0000	0.0000
job	job	0	0.0000	0.0000
loan	loan	0	0.0000	0.0000
campaign	campaign	0	0.0000	0.0000
marital	marital	0	0.0000	0.0000
day	day	0	0.0000	0.0000
month	month	0	0.0000	0.0000

Above variables are important variables.

So we can use the misclassification rate and can see here after I think after this ,there is no improvement in the validation errors.



Neural Network

The Neural Network node enables you to construct, train, and validate multilayer feed-forward neural networks. By default, the Neural Network node automatically constructs a multilayer feed-forward network that has one hidden layer consisting of three neurons. In general, each input is fully connected to the first hidden layer, each hidden layer is fully connected to the next hidden layer, and the last hidden layer is fully connected to the output. The Neural Network node supports many variations of this general form.

Development of models

Import and clean your dataset. Split data into training, validation, and testing sets. Add and configure the Neural Network node. Train the model using the training data. Assess model performance on the validation set. Check performance on the testing set to ensure generalization. Fine-tune hyperparameters for better performance. Consider ensemble methods for further improvement.

Model performance

Table X: Neural network models performance

MODEL VARIATION S	ROC INDEX	CU. LIFT	SCOPE %	TRUE - %	FALSE - %	TRUE + %	FALSE + %
ANN-Imp-R2	0.862	3.38	20	1148	99	56	52
ANN-Imp-Chi2	0.860	3.37	20	1148	99	56	52
ANN-Imp	0.859	1.25	20	1200	155	0	0
ANN-Imp-TR	0.859	3.51	20	1136	102	53	64
ANN-IMP-6HN	0.840	3.32	20	1142	97	58	58
Neural Network	0.710	3.19	20	1148	107	48	52
Neural Network	0.575	1.63	20	1200	155	0	0
ANN-IMP-DT	0.566	2.09	20	1176	144	11	24

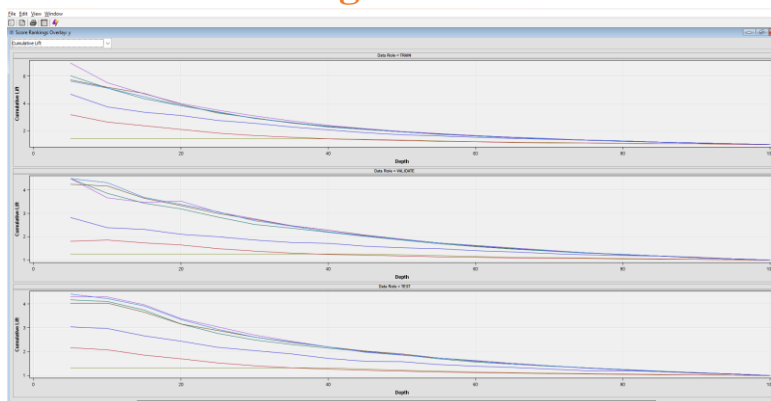
Fit Statistics

Model Selection based on Valid: Roc Index (_VAUR_)

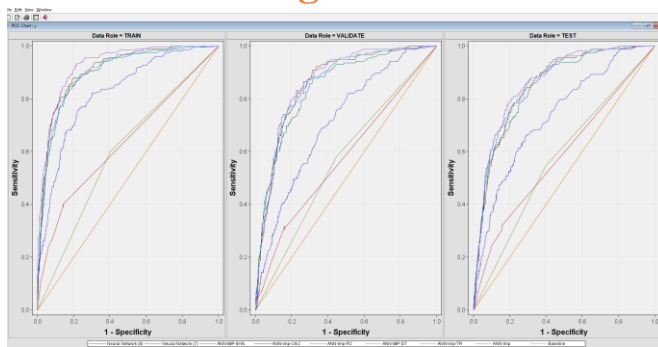
		Train:		Valid:		Valid:	
		Valid: Average	Train: Average	Valid: Squared	Misclassification	Squared	Misclassification
Selected	Model	Roc	Squared	Misclassification	Squared	Misclassification	
Model	Node	Model Description	Index	Error	Rate	Error	Rate
Y	Neural4	ANN-Imp-R2	0.862	0.06109	0.08518	0.08373	0.12251
	Neural5	ANN-Imp-Chi2	0.860	0.06651	0.09679	0.08265	0.11439
	Neural	ANN-Imp	0.859	0.06886	0.09956	0.08286	0.11144
	Neural2	ANN-Imp-TR	0.859	0.06886	0.09956	0.08286	0.11144
	Neural6	ANN-IMP-6HN	0.840	0.06944	0.09679	0.08613	0.11734
	Neural8	Neural Network (8)	0.710	0.08084	0.10288	0.10335	0.12399
	Neural7	Neural Network (7)	0.575	0.09616	0.11504	0.10133	0.11439
	Neural3	ANN-IMP-DT	0.566	0.10015	0.11504	0.10080	0.11439

Example,

cumulative lift diagram



ROC curves diagram



Overfitting analysis

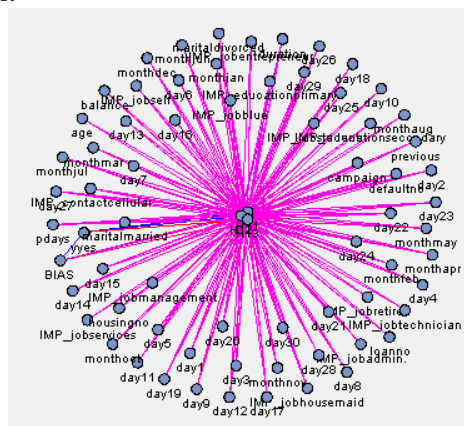
Here's a concise overview of overfitting analysis in neural networks .Split your dataset into training and validation sets. Train your neural network on the training data. Evaluate the model's performance on the validation set. Track metrics like loss and accuracy on both training and validation sets. Compare performance between training and validation sets. If the model performs significantly better on the training set than on the validation set, it might be overfitting.

[illegible]

Neural network architecture of best model

ANN is a computational system consisting of many interconnected units called **artificial neurons**. The connection between artificial neurons can transmit a signal from one neuron to another. So, there are multiple possibilities for connecting the neurons based on which the **architecture** we are going to adopt for a specific solution.

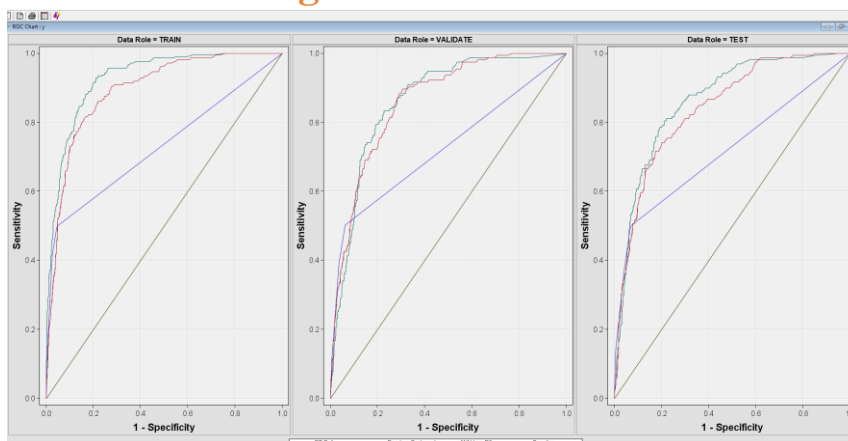
- There may be just two layers of neuron in the network – the input and output layer.
- There can be one or more intermediate **‘hidden’** layers of a neuron.
- The neurons may be connected with all neurons in the next layer and so on.



Selected Node	Predecessor Nodes	Node Index	Model Description	Target Variable	Target Label	Selection Criterion: Wald Test Index	Train Average Squared Error	Train Average Squared Function	Train Degrees of Freedom for Error	Train Model Degrees of Freedom	Test Model Degrees of Freedom	Train Deviation for AIC	Train Error Function	Train Final Prediction Error	Train Maximum Absolute Residual	Train Mean Square Error	Train Sum of Squares	Train Number of Outliers	Train Train Average Squared Error	Train Test Average Squared Error	Train Test Mean Squared Error	Train Test Schwartz's Akaike	
Y	Neural4 Res2	Neural4 Res2	AHN Imp-R2 D-Net Train Imp-Backward D-Net	v	v	0.862	967.116	0.061086	0.20053	1687	121	1808	3616	725.116	0.069556	0.996475	0.065474	1808	16	0.247168	0.264303	0.256879	1632
						0.858	933.5203	0.072645	0.248316	1792	150	1808	3616	901.5203	0.07489	0.996675	0.074302	1808	16	0.271178	0.273789	0.272895	1021
						0.823		0.079559				1808	3616		0.959591							0.275002	

Table X Summary results of the best performing models

Example,
cumulative lift diagram



Discussion on the breadth of areas of application and research in data mining.

Can you identify any emerging trends or research directions in predictive modeling beyond the topics covered in the module?

Predictive modeling tries to find good rules (models) for guessing (predicting) the values of one or more variables in a data set from the values of other variables in the data set. After a good rule has been found, it can be applied to new data sets (scoring) that might or might not contain the variable or variables that are being predicted. The various methods that find prediction rules go by different names in different areas of research, such as regression, function mapping, classification, discriminant analysis, pattern recognition, concept learning, supervised learning, and so on. Because SAS Enterprise Miner is intended especially for the analysis of large data sets, all of the predictive modeling nodes are designed to work with separate training, validation, and test sets. The Data Partition node provides a convenient way to split a single data set into the three subsets, using simple random sampling, stratified random sampling, or user defined sampling. Each predictive modeling node also enables you to specify a fourth scoring data set that is not required to contain the target variable. These four different uses for data sets are called the roles of the data sets. For the training, validation and test sets, the predictive modeling nodes can produce two output data sets: one containing the original data plus scores (predicted values, residuals, classification results, and so on), the other containing various statistics pertaining to the fit of the model (the error function, misclassification rate, and so on). For scoring sets, only the output data set containing scores can be produced. SAS Enterprise Miner provides a number of tools for predictive modeling. Three of these tools are the Regression node, the Decision Tree node, and the Neural Network node. The methods used in these nodes come from several areas of research, including statistics, pattern recognition, and machine learning. These different areas use different terminology, so before discussing predictive modeling methods, it will be helpful to clarify the terms used in SAS Enterprise Miner. The following list of terms is in logical, not alphabetical order. A more extensive alphabetical glossary can be found in the Glossary.

Conclusion

My journey with SAS Enterprise Miner has been instrumental in expanding my skill set and enhancing my understanding of predictive modeling. By reflecting on my experiences, identifying challenges, and embracing opportunities for growth, I am better equipped to leverage data mining techniques effectively in future endeavors.

Recommendation

Having demonstrated proficiency in SAS Enterprise Miner, I commend your dedication to mastering data mining techniques and leveraging the platform's capabilities effectively. To further enhance your skills and advance your expertise in predictive analytics, I recommend the following steps:

- ❖ Continuous learning
- ❖ Hands-on projects
- ❖ Specialization tracks
- ❖ Community Engagement
- ❖ Mentorship and collaboration.

My Reflections on the process – What did I learn from this exercise?

Understanding of Data Mining Workflow: Using SAS Enterprise Miner has provided me with a deeper understanding of the data mining workflow, from data preprocessing to model deployment. I've learned the importance of each step in the process and how they collectively contribute to building effective predictive models.

Hands-On Experience with Data Preparation: Working with real-world data in SAS Enterprise Miner has given me hands-on experience with data preparation techniques such as data cleaning, transformation, and feature engineering. I've learned how to handle missing values, outliers, and other data quality issues to ensure the accuracy and reliability of my models.

Exploration of Model Building Techniques: SAS Enterprise Miner offers a wide range of model building techniques, from traditional statistical models to advanced machine learning algorithms. I've had the opportunity to explore different modeling approaches and understand their strengths, weaknesses, and suitability for various types of data and predictive tasks.

Evaluation and Interpretation of Model Results: Evaluating and interpreting model results is a critical aspect of the data mining process. Using SAS Enterprise Miner, I've learned how to assess model performance using metrics such as accuracy, precision, recall, and ROC curves. I've also gained insights into interpreting model outputs and making informed decisions based on the results.

Appreciation for Model Deployment and Monitoring: Building predictive models is only the first step; deploying and monitoring models in production is equally important. Through this exercise, I've gained an appreciation for the challenges and considerations involved in deploying models into operational environments and monitoring their performance over time.

Overall, my experience with SAS Enterprise Miner has been incredibly valuable, providing me with practical skills, knowledge, and confidence in leveraging data mining techniques to extract insights and make informed decisions from data. I look forward to applying these learnings in future projects and continuing to explore the exciting field of predictive analytics.

References and Bibliography

Larsen, K. 2010. "Net Lift Models: Optimizing the Impact of Your Marketing Efforts." SAS Course Notes, SAS Institute Inc., Cary, NC.

Lo, Victory S.Y. 2002. "The True Lift Model — A Novel Data Mining Approach to Response Modeling in Database Marketing." SIGKDD Explorations 4 (2): 78–86.

Bishop, C. M. 1995. Neural Networks for Pattern Recognition. Oxford: Oxford University Press.

Appendix

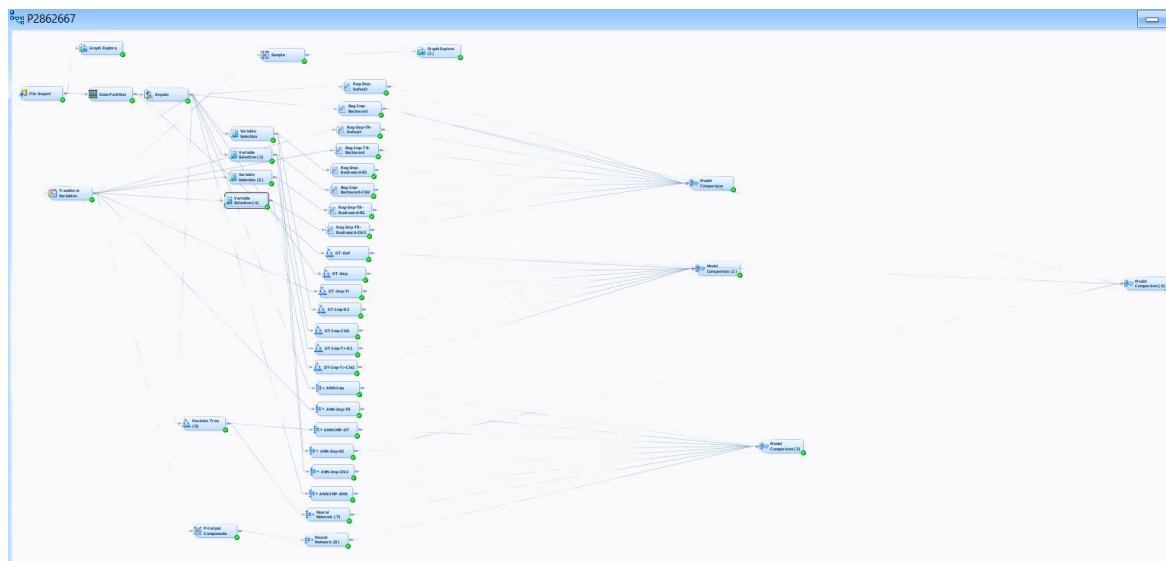
In this appendix, I reflect on the roles involved in data mining and the workflow diagram utilized throughout the process. This reflection captures my understanding, insights gained, and areas for further exploration in the field of data mining.

Data Mining Roles

As a data analyst, I have taken on the responsibility of understanding business objectives, exploring data, selecting appropriate models, and interpreting results to address business problems. This role requires a combination of technical expertise, domain knowledge, and analytical skills to derive actionable insights from data.

Workflow diagram

The workflow diagram serves as a visual representation of the data mining process, outlining the sequence of steps from data collection to model deployment. Through the workflow diagram, I have gained a holistic view of the data mining lifecycle and the interdependencies between different stages of the process.



FIRST AND FOREMOST, I THANK GOD ALMIGHTY FOR GIVING ME THE STRENGTH, COURAGE AND PERSEVERANCE TO PURSUE MY TASK SUCCESSFULLY THROUGHOUT THE PREPARATION OF THIS PROJECT WORK. I WISH TO THANK **ALI KHAN ASSISTANT PROFESSOR**, DEPARTMENT OF DATA ANALYTICS, FOR HER VALUABLE GUIDANCE AND RELENTLESS SUPPORT IN THE EXECUTION OF THE PROJECT.

I THANK ALL FACULTY MEMBERS OF THE DEPARTMENT OF DATA ANALYTICS FOR THEIR SUGGESTIONS DURING THE COURSE OF MY PROJECT WORK.

