TITANIC SURVIVAL REPORT

Hellen Mkumbo

Centennial College

BA 706

David Parent

July 15, 2022

Table of Contents

Background 4

Purpose 4

Problem definition 4

Methodology 4

Data Exploration 5

1.0. Missing Values 6

2.0. The relationship between the target variable and other variable 6

Sex vs Survival rate 6

Passenger class Vs Survival 7

Data Cleaning 9

Data Partition 9

Modelling Documentation 10

1.0 Model 10

1.1. Decision Tree 10

Maximal Tree 10

Optimal Tree 11

Probability Tree 11

1.2. Regression 12

Full Regression 12

Forward Regression 13

Backward Regression 13

Stepwise Regression 14

Interpretation of the regression 15

1.3. Neural Network 15

Full Neural Network 15

AutoNeural Network 16

Reduced Variable Set NN – stepwise 16

Assessment 17

Conclusion 17

Annex 1: Survival Prediction Diagram 18

References 19

Background

The sinking of the Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the widely considered "unsinkable" RMS Titanic sank after colliding with an iceberg. Unfortunately, there weren't enough lifeboats for everyone onboard, resulting in the death of 1502 out of 2224 passengers and crew.

While there was some element of luck involved in surviving, it seems some groups of people were more likely to survive than others.

Purpose

The purpose of this analysis is to predict what sorts of people were more likely to survive by using the passenger data (i.e name, age, gender, socio-economic class, etc) and trying to predict the classification of who Survived or deceased.

Problem definition

There are two sets of data. One (train.csv) is used to train our model and contains information on survival and death. One for testing (test.csv), which we will use to test our models, is missing information on survival and death.

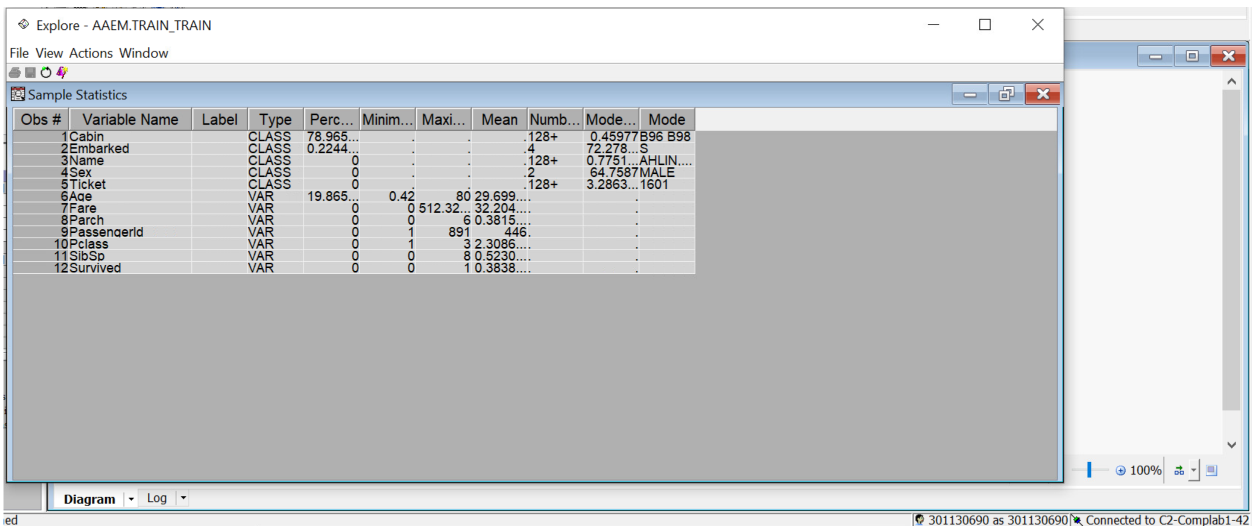
Methodology

Predictive Modeling was performed using SAS Enterprise Miner to analyze this dataset. In this analysis various types of models were used to evaluate the association between the variables. I am using Titanic dataset from Kaggle.com which contains a training and test dataset.

Data Exploration

In the initial stage, we'll conduct an exploratory data analysis for our problem. Both the train and test data are examined in the exploratory data analysis dataset to identify the characteristics that might affect the survival rate. By establishing a connection between each feature and survival, the data is thoroughly studied.

A subset of the passengers on board are represented by 891 records and 12 columns in the train dataset. The binary variable Survived, which is part of the train dataset, will be utilized as the target. Our Input variables will include name, age, fare, sex, embarked, cabin, ticket, pseudo class, #siblings/spouse, #parents/children.


Explore - AAEM.TRAIN_TRAIN

File View Actions Window

Sample Statistics

Obs #	Variable Name	Label	Type	Perc...	Minim...	Maxi...	Mean	Numb...	Mode...	Mode
1	Cabin		CLASS	78.965...	.	.	.	128+	0.45977	B96 B98
2	Embarked		CLASS	0.2244...	.	.	.	4	72.278...	S
3	Name		CLASS	0	.	.	.	128+	0.7751...	AHLIN....
4	Sex		CLASS	0	.	.	.	2	64.7587	MALE
5	Ticket		CLASS	0	.	.	.	128+	3.2863...	1601
6	Age		VAR	19.865...	0.42	80	29.699....	.	.	
7	Fare		VAR	0	0	512.32...	32.204....	.	.	
8	Parch		VAR	0	0	6	0.3815....	.	.	
9	PassengerId		VAR	0	1	891	446	.	.	
10	Pclass		VAR	0	1	3	2.3086....	.	.	
11	SibSp		VAR	0	0	8	0.5230....	.	.	
12	Survived		VAR	0	0	1	0.3838....	.	.	

While the test dataset contains similar information with other 418 passengers on board and 11 columns. We are missing the Survived column in the test data set.

The only difference between train and test data is the Survived column which indicates if the passenger survived the disaster or not.

1.0. Missing Values

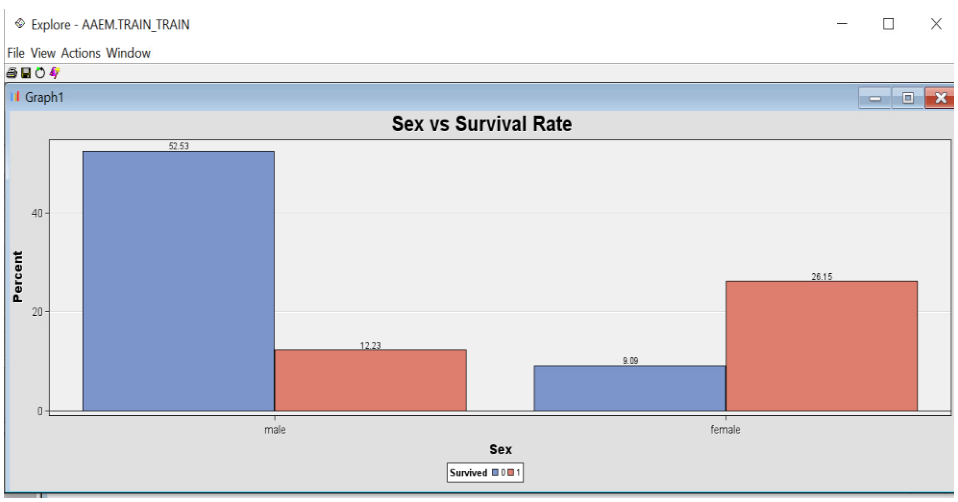
A short glance at the data reveals that there are some missing variables, and that they are all in the Age, cabin, and embarked features. Removing the feature is one option, while replacing the missing value with a fixed value or the mean is another.

2.0. The relationship between the target variable and other variable

Sex vs Survival rate

As we can see, much more females than males survived. Even more significant findings relate to passengers who died, where women make up a relatively small portion compared to men.

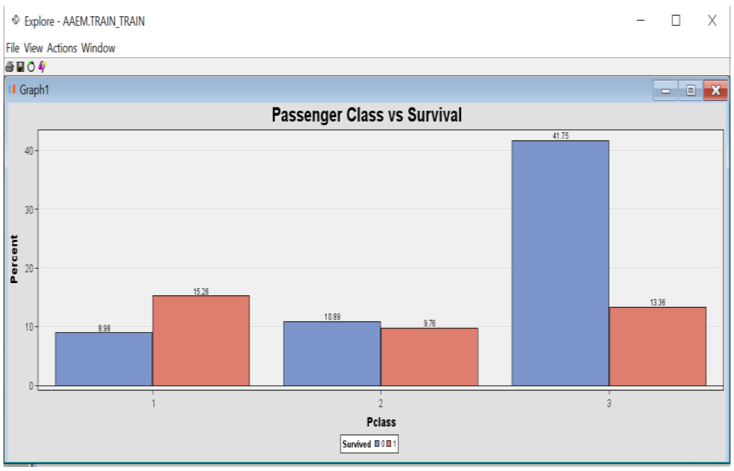
Exhibit 1: Sex vs Survival



Passenger class Vs Survival

Here, we can see that third-class passengers had a larger risk of passing away than first-class passengers, who had a better chance of surviving.

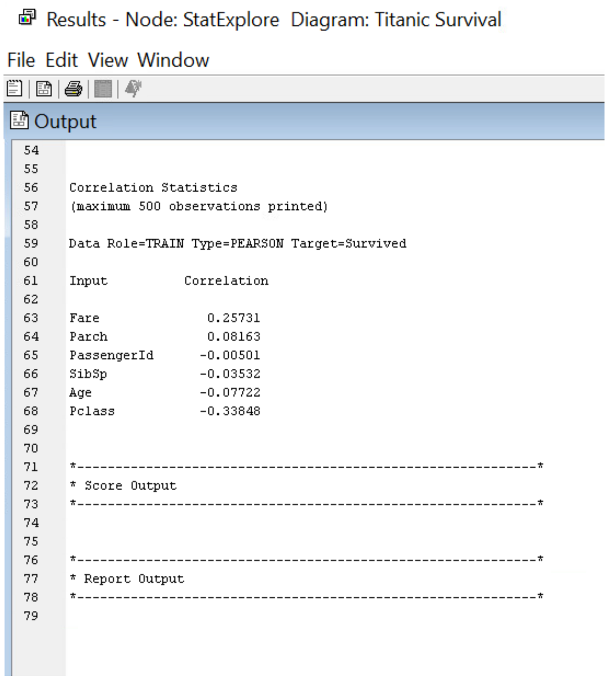
Exhibit 2: Passenger class Vs Survival



The models place a lot of importance on these and other relationships between the variables and the survival rate.

We might be able to determine which variables are crucial based on the correlation. As demonstrated in Exhibit 3, the variables fare and parch have a positive relationship with the target variable (survive), while passengerId, SibSp, age, and pclass have a negative relationship with the target variable.

Exhibit 3: Correlation Statistics



Results - Node: StatExplore Diagram: Titanic Survival

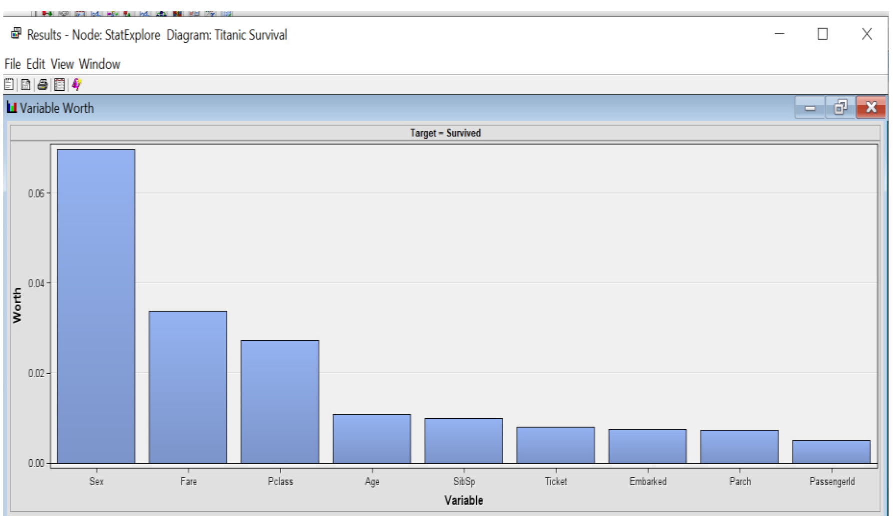
File Edit View Window

Output

```
Correlation Statistics
(maximum 500 observations printed)

Data Role=TRAIN Type=PEARSON Target=Survived

Input          Correlation

Fare              0.25731
Parch             0.08163
PassengerId      -0.00501
SibSp            -0.03532
Age              -0.07722
Pclass           -0.33848


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


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

In the variable worth plot below shows that sex, fare, passenger class, age variables seem to have highest worth with the targeted variable Survived.

Exhibit 4: Variable Worth



Data Cleaning

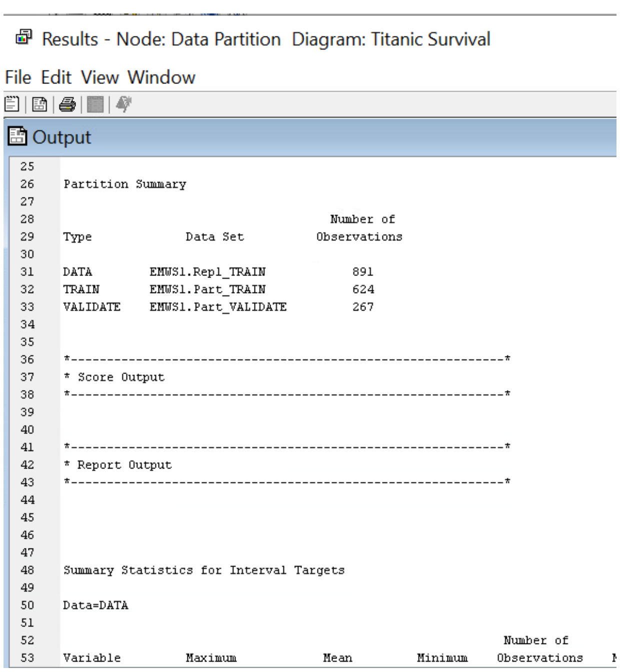
We impute values to utilize as replacements for missing values in the input data for regression and neural network models. To enhance performance with regression and neural network models, we also modified the input data.

In Decision Tree models, the replacement is accomplished by substituting random samples for the missing values.

Data Partition

The input data must be divided into train and validation datasets. We do this action in order to obtain the most accurate assessment of the model's performance. We must divide the data into 70 percent training and 30 percent validation because there are only 891 passenger cases in total.

Exhibit 5: Partition Summary of the Input data



Modelling Documentation

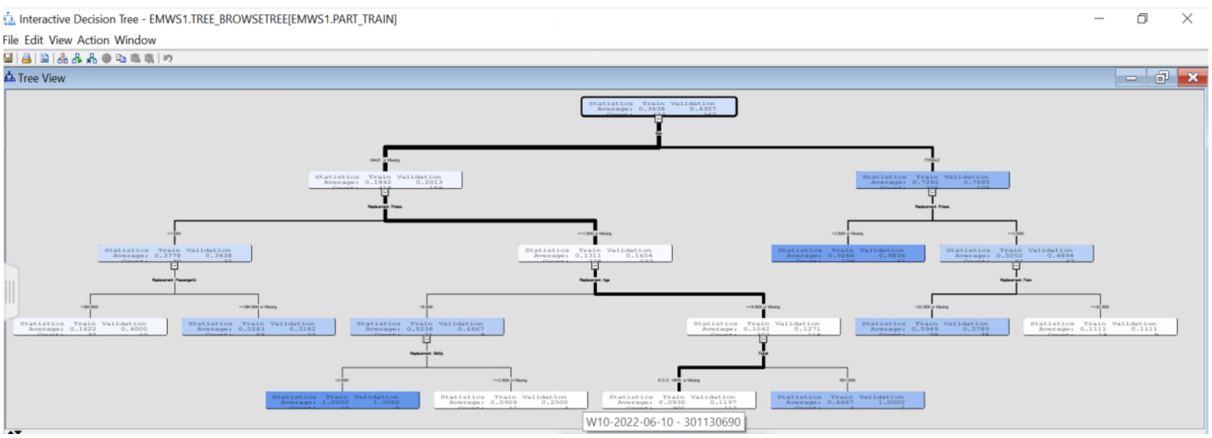
1.0 Model

1.1. Decision Tree

Maximal Tree

The subsequent model favored men from lower social classes by selecting the variable pclass as a reliable predictor in both cases. Age and fare are also shown to be predictive, with older men having a higher death rate than younger men. Splits on tickets, cabins, siblings/spouses rarely occurred.

Exhibit 6: Maximal Decision Tree

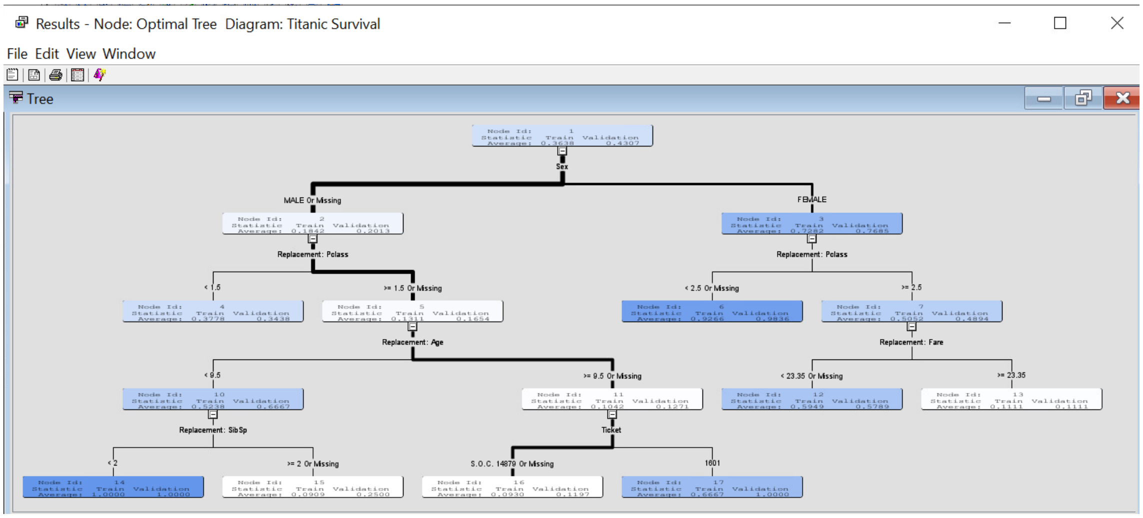


Even though the majority of the fit improvement happens over the first few splits, the maximal, fifteen-leaf tree seems to produce a lower misclassification rate than any of its simpler predecessors, according to the plot for training data. The maximum tree appears to be the preferable choice for associating predictions with cases, according to the plot using training data. This figure is deceptive, though, when merely considering the outcomes from the training set of data.

Optimal Tree

The optimal tree shown in the exhibit 7 the model has prune some selected variables in the model.

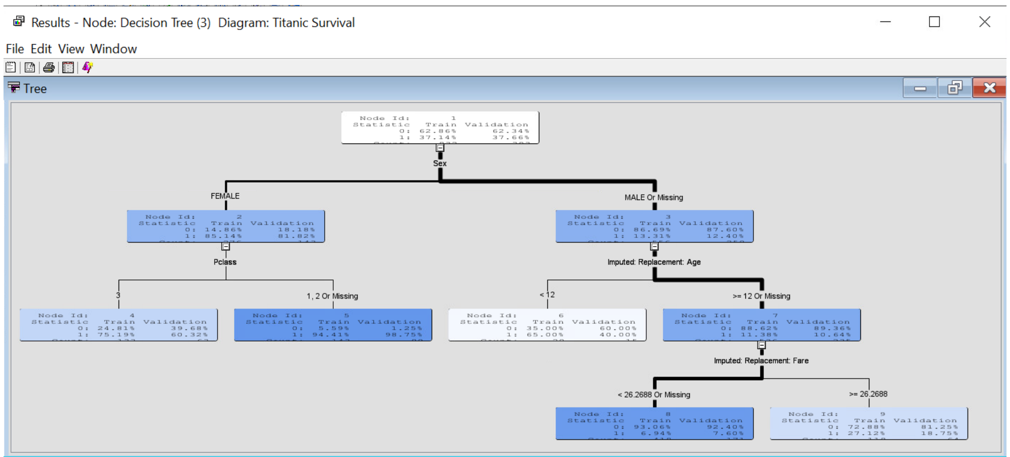
Exhibit 7: Optimal Decision Tree



Probability Tree

In the probability tree shown in the exhibit 8, it shows that the model has eliminated other variables from the model leaving Sex, Pclass, Age, and Fare as the selected variable in the model used.

Exhibit 8: Probability Decision Tree

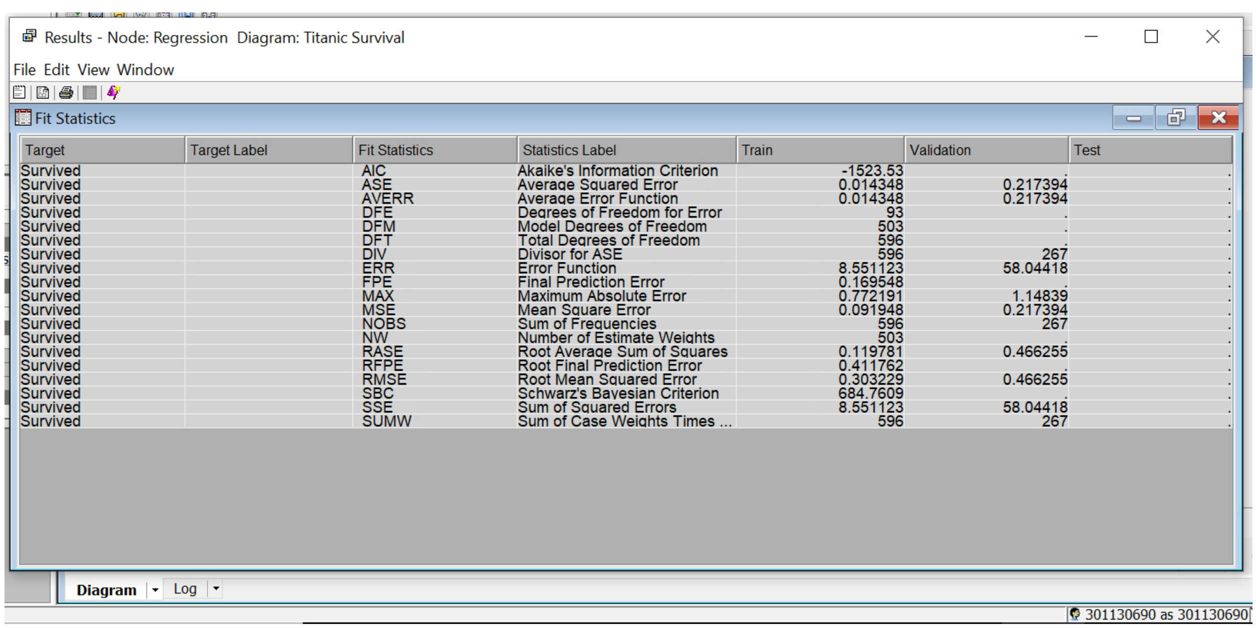


1.2. Regression

Full Regression

Logistic regression is the best technique for building a model for a binary variable. In our case, the target variable is survived. Age, Pclass, and Sex variables were chosen for the variable selection process and will be used in the model.

Exhibit 9: Regression Summary

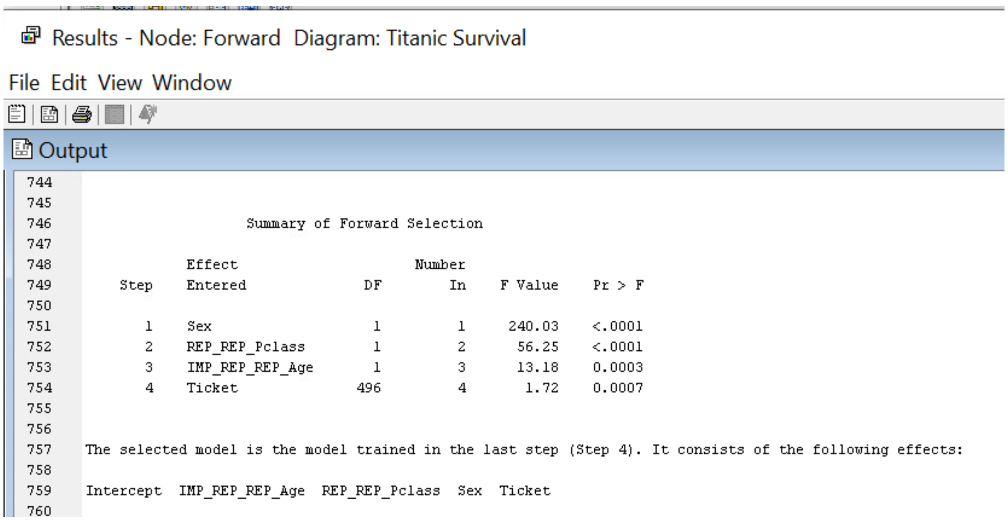


Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Survived		AIC	Akaike's Information Criterion	-1523.53	.	.
Survived		ASE	Average Squared Error	0.014348	0.217394	.
Survived		AVERR	Average Error Function	0.014348	0.217394	.
Survived		DFE	Degrees of Freedom for Error	93	.	.
Survived		DFM	Model Degrees of Freedom	503	.	.
Survived		DFT	Total Degrees of Freedom	596	.	.
Survived		DIV	Divisor for ASE	596	267	.
Survived		ERR	Error Function	8.551123	58.04418	.
Survived		FPE	Final Prediction Error	0.169548	.	.
Survived		MAX	Maximum Absolute Error	0.772191	1.14839	.
Survived		MSE	Mean Square Error	0.091948	0.217394	.
Survived		NOBS	Sum of Frequencies	596	267	.
Survived		NW	Number of Estimate Weights	503	.	.
Survived		RASE	Root Average Sum of Squares	0.119781	0.466255	.
Survived		RFPE	Root Final Prediction Error	0.411762	.	.
Survived		RMSE	Root Mean Squared Error	0.303229	0.466255	.
Survived		SBC	Schwarz's Bayesian Criterion	684.7609	.	.
Survived		SSE	Sum of Squared Errors	8.551123	58.04418	.
Survived		SUMW	Sum of Case Weights Times ...	596	267	.

Forward Regression

As demonstrated in Exhibit 10, the model entered four effects—Sex, Pclass, Age, and Ticket—into the forward regression results when choosing the model developed in the last phase (step 4). The four effects are all present in the model.

Exhibit 10: Forward Regression Results



Summary of Forward Selection

Step	Effect Entered	DF	Number In	F Value	Pr > F
1	Sex	1	1	240.03	<.0001
2	REP_REP_Pclass	1	2	56.25	<.0001
3	IMP_REP_REP_Age	1	3	13.18	0.0003
4	Ticket	496	4	1.72	0.0007

The selected model is the model trained in the last step (Step 4). It consists of the following effects:

Intercept IMP_REP_REP_Age REP_REP_Pclass Sex Ticket

Backward Regression

The outcome of backward regression indicates that the model chosen is the model from the final stage, as seen in exhibit 11. Age, Sex, and Ticket are selected in the model as model effects while fare, Pclass, PassengerId, SibSp, Embarked, and Parch are eliminated from the model.

Exhibit 11: Backward Regression Results

Results - Node: Backward Diagram: Titanic Survival

File Edit View Window

Output

```
Ticket            W.E.P. 5734          1     -0.1169      0.3055      -0.38      0.7030
Ticket            W/C 14208            1     -0.2220      0.3029      -0.73      0.4654


NOTE: No (additional) effects met the 0.05 significance level for removal from the model.


                  Summary of Backward Elimination

           Effect                       Number
    Step   Removed                DF        In    F Value    Pr > F

       1   IMP_REP_REP_Fare        1         8       0.00    0.9610
       2   REP_REP_Pclass          1         7       0.16    0.6881
       3   REP_REP_PassengerId     1         6       0.20    0.6523
       4   IMP_REP_REP_SibSp       1         5       0.33    0.5680
       5   IMP_Embarked            2         4       0.54    0.5843
       6   IMP_REP_REP_Parch       1         3       2.19    0.1419


The selected model is the model trained in the last step (Step 6). It consists of the following effects:

Intercept  IMP_REP_REP_Age  Sex  Ticket
```

Stepwise Regression

The best subset of variables for the Model can be chosen using a number of stopping rules provided by stepwise regression. The results of running the model stepwise are displayed in exhibit 12. The model is trained in the final stage using the variables chosen through stepwise regression (step 5). Passenger Class, Sex, Age, and Ticket are the four total effects evaluated via stepwise regression.

Exhibit 12: Stepwise Logistic Regression Results

Results - Node: Stepwise Diagram: Titanic Survival

File Edit View Window

Output

```
Ticket            W./C. 6608           1     -0.5774      0.1806      -3.20      0.0019
Ticket            W.E.P. 5734          1     -0.1169      0.3055      -0.38      0.7030
Ticket            W/C 14208            1     -0.2220      0.3029      -0.73      0.4654


NOTE: No (additional) effects met the 0.05 significance level for entry into the model.


                            Summary of Stepwise Selection

                         Effect                          Number
    Step    Entered            Removed             DF        In    F Value    Pr > F

       1    Sex                                     1         1     240.03    <.0001
       2    REP_REP_Pclass                          1         2      56.25    <.0001
       3    IMP_REP_REP_Age                         1         3      13.18    0.0003
       4    Ticket                                496         4       1.72    0.0007
       5                       REP_REP_Pclass       1         3       0.54    0.4654


The selected model is the model trained in the last step (Step 5). It consists of the following effects:

Intercept  IMP_REP_REP_Age  Sex  Ticket
```

Interpretation of the regression

All groups of women survive more frequently than all classes of males at every age. Compared to adult men, young boys have better survival chances. This demonstrates that women and children were prioritized for rescue.

1.3.Neural Network

Full Neural Network

The average squared error vs optimization iteration is displayed on the iteration plot. The vertical blue line shows a significant divergence in training and validation average squared error towards iteration 24.

Exhibit 13: Iteration Plot

AutoNeural Network

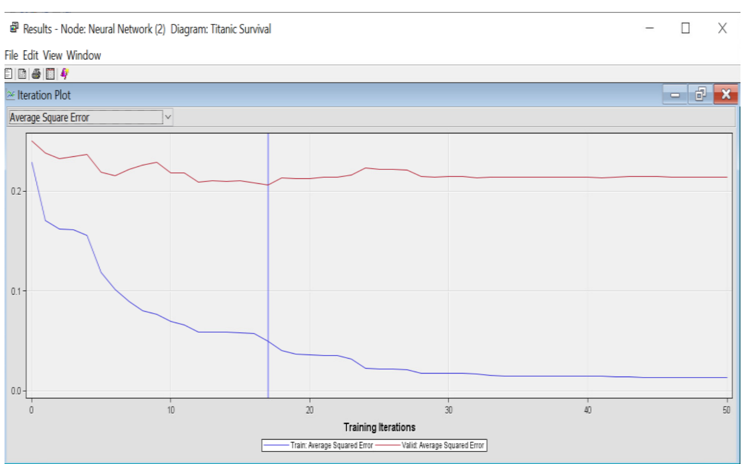
Iteration charts for the AutoNeural and Neural Network nodes are different. The final fit statistic vs the quantity of hidden neural network units is shown on the iteration plot of the AutoNeural node.

Exhibit 14: Iteration Plot

Reduced Variable Set NN – stepwise

In this plot shown in the exhibit 15 shows that the Validation Average Squared Error curve is above and very close to the edge.

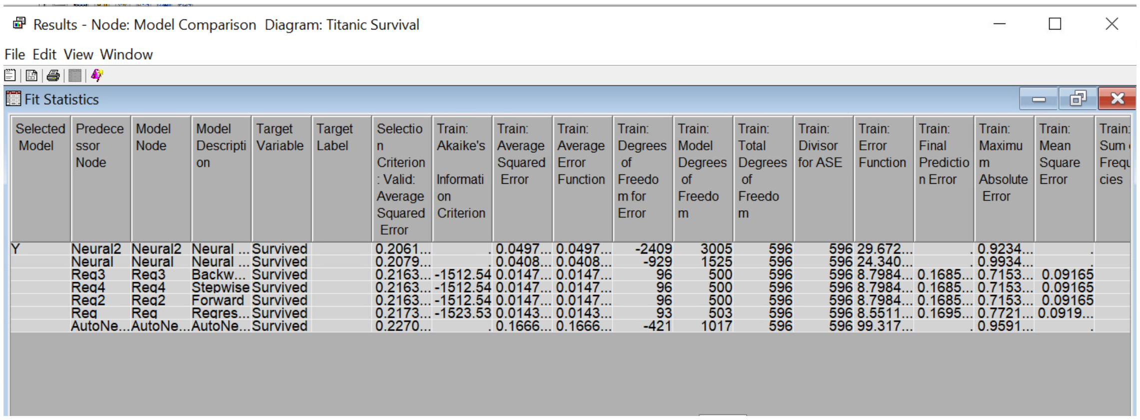
Exhibit 15: Iteration Plot



Assessment

As shown in the table below, the Neural Network 2 which was connected with the stepwise tree to reduce variable is selected as the best model of all the predictive models used for the Titanic disaster survival.

Exhibit 16: Fit Statistics

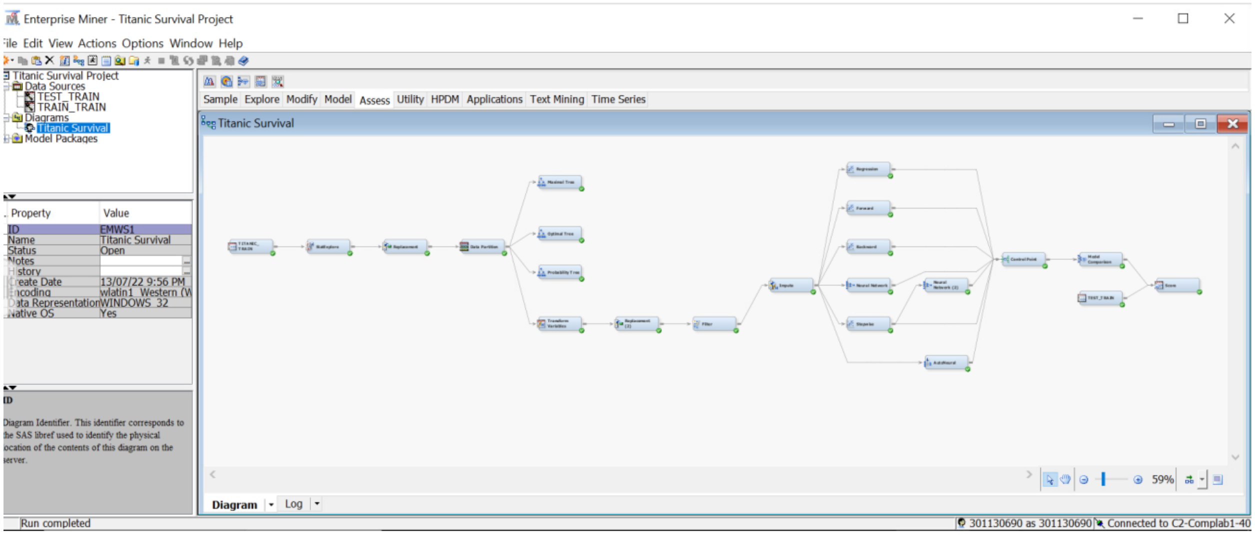


Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Average Squared Error	Train: Akaike's Information Criterion	Train: Average Squared Error	Train: Average Error Function	Train: Degrees of Freedom for Error	Train: Model Degrees of Freedom	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: Sum of Frequencies
Y	Neural2	Neural2	Neural ...	Survived		0.2061...	.	0.0497...	0.0497...	-2409	3005	596	596	29.672...	.	0.9234...	.	
	Neural	Neural	Neural ...	Survived		0.2079...	.	0.0408...	0.0408...	-929	1525	596	596	24.340...	.	0.9934...	.	
	Reg3	Reg3	Backw...	Survived		0.2163...	-1512.54	0.0147...	0.0147...	96	500	596	596	8.7984...	0.1685...	0.7153...	0.09165	
	Reg4	Reg4	Stepwise	Survived		0.2163...	-1512.54	0.0147...	0.0147...	96	500	596	596	8.7984...	0.1685...	0.7153...	0.09165	
	Reg2	Reg2	Forward	Survived		0.2163...	-1512.54	0.0147...	0.0147...	96	500	596	596	8.7984...	0.1685...	0.7153...	0.09165	
	Reg	Reg	Regres...	Survived		0.2173...	-1523.53	0.0143...	0.0143...	93	503	596	596	8.5511...	0.1695...	0.7721...	0.0919...	
	AutoNe...	AutoNe...	AutoNe...	Survived		0.2270...	.	0.1666...	0.1666...	-421	1017	596	596	99.317...	.	0.9591...	.	

Conclusion

In general, the model predicted survivors with Sex variable this is Miss, Mrs., Master. Females traveling in first, second class and pockets within 3rd class. Males with younger age with less than 3 siblings. Mr., Rev., and Others were not predicted to survive.

Annex 1: Survival Prediction Diagram



References

Titanic - Machine Learning from Disaster. (n.d.). Retrieved from Kaggle: https://www.kaggle.com/mashimo/a-very-simple-logistic-regression-model