MKTG 746 BIG DATA AND PREDICTIVE ANALYST

FINAL PROJECT REPORT
GROUP 1

Presented By:
Harleen Kaur (301302804)
Mahek Modi (301390323)
Mustufa Amodwala(301384802)

Data Mining and Predictive Modeling Project
Dataset Name:
Occupation and Outcome Analysis

This project explores a dataset on occupation and outcomes to build predictive models using SAS Enterprise Miner. We will analyze various factors influencing career outcomes through decision trees, logistic regression, and neural networks. Our goal is to identify key variables and create models that can accurately predict binary outcomes related to occupations. This report will take you through our thought process, methodology, and insights gained from the data mining journey.

Data Overview and Preparation

Data Source

The dataset was obtained from Kaggle, titled "Occupations and Outcome". It includes detailed information about various occupational factors and their associated outcomes, which are crucial for predictive modeling.

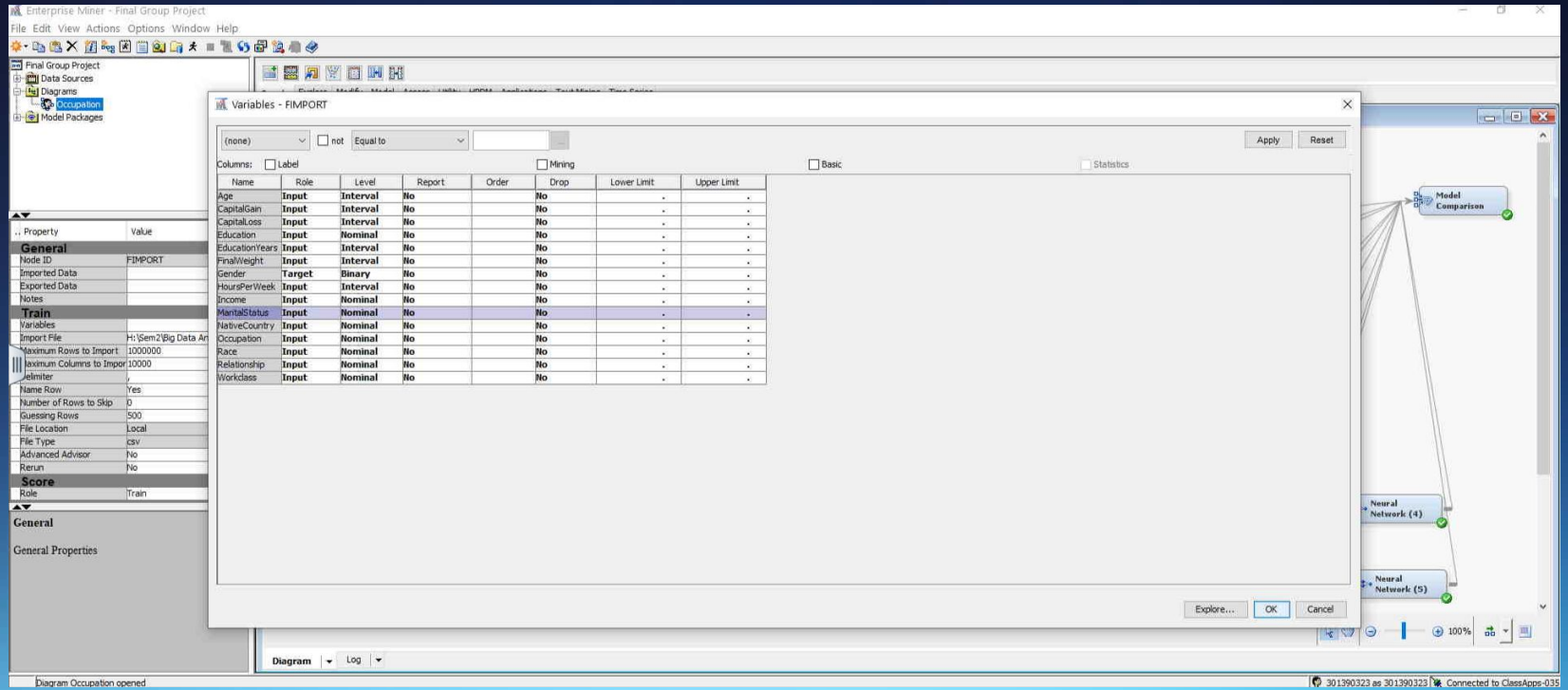
Data Splitting

The data was divided into training and validation sets to ensure the model is robust and can generalize well to new, unseen data.

Variable Selection

The initial analysis focused on identifying key variables that could significantly influence the target outcome. These variables were selected based on their importance and potential impact on the predictive model accuracy.

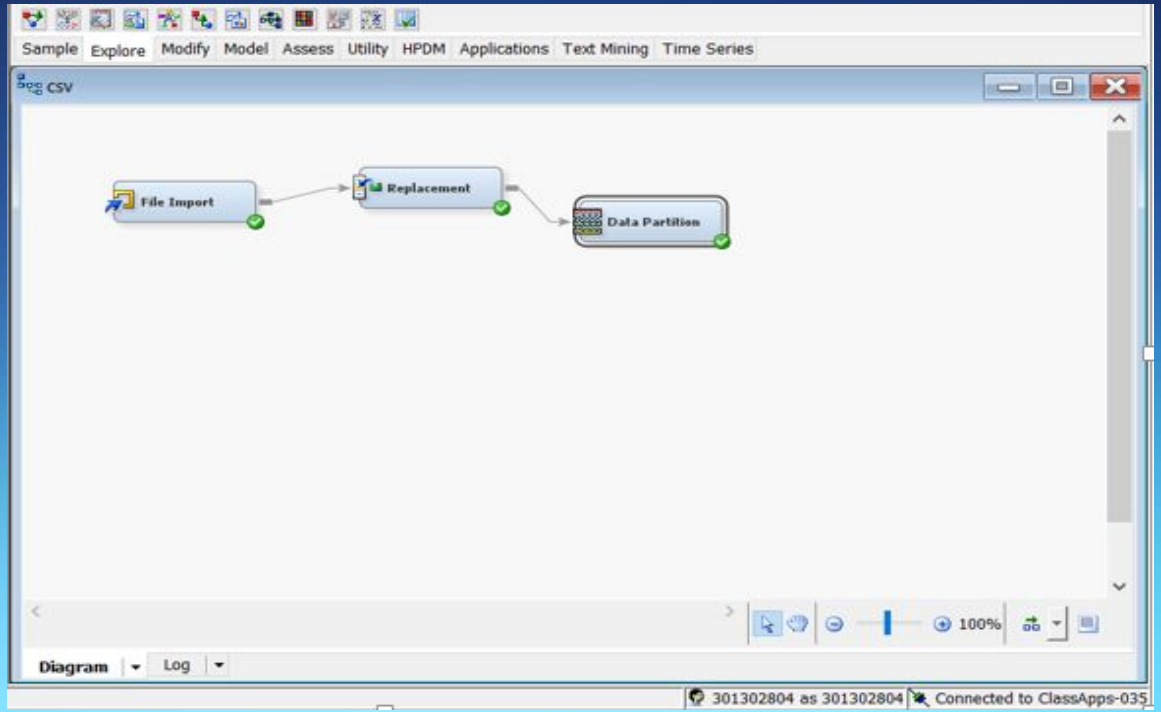
Variable Selection and Configuration in Predictive Modeling



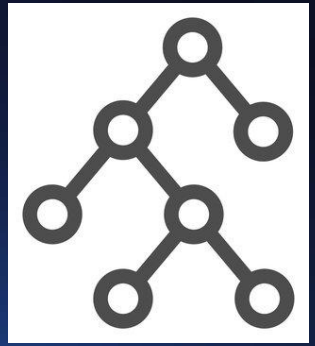
Dataset Name: Occupations and Outcome

Data Partition

The data set was partitioned where training was set at 70 and Validation was set at 30. This helps in getting better results as a part of it is used for training and then it gets validated to ensure performance accuracy.



Decision Tree Model: Maximal Tree

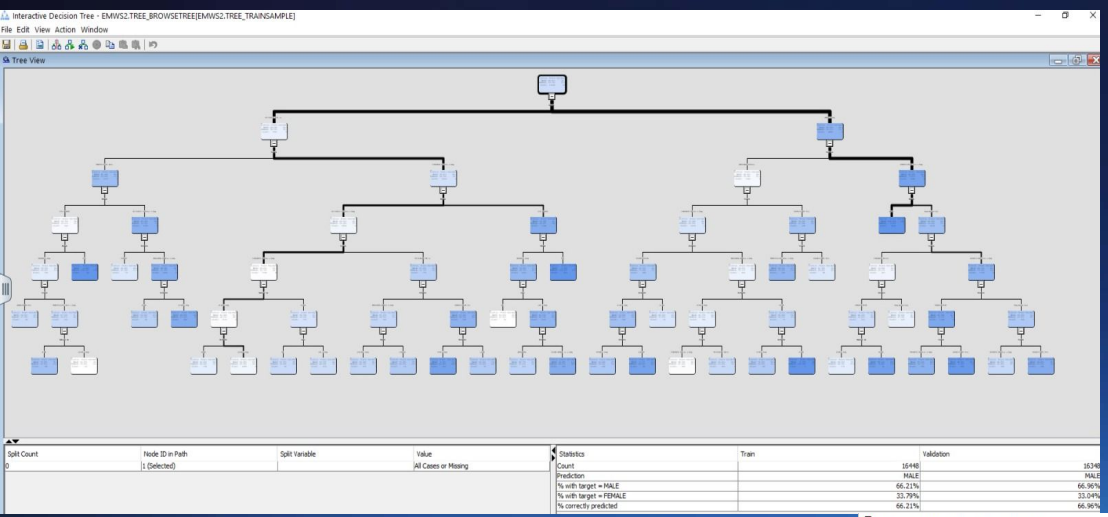


- The maximal tree represents the most complex version of the decision tree, where all possible splits in the data have been made. It captures every interaction and pattern within the dataset, making it a highly detailed model.

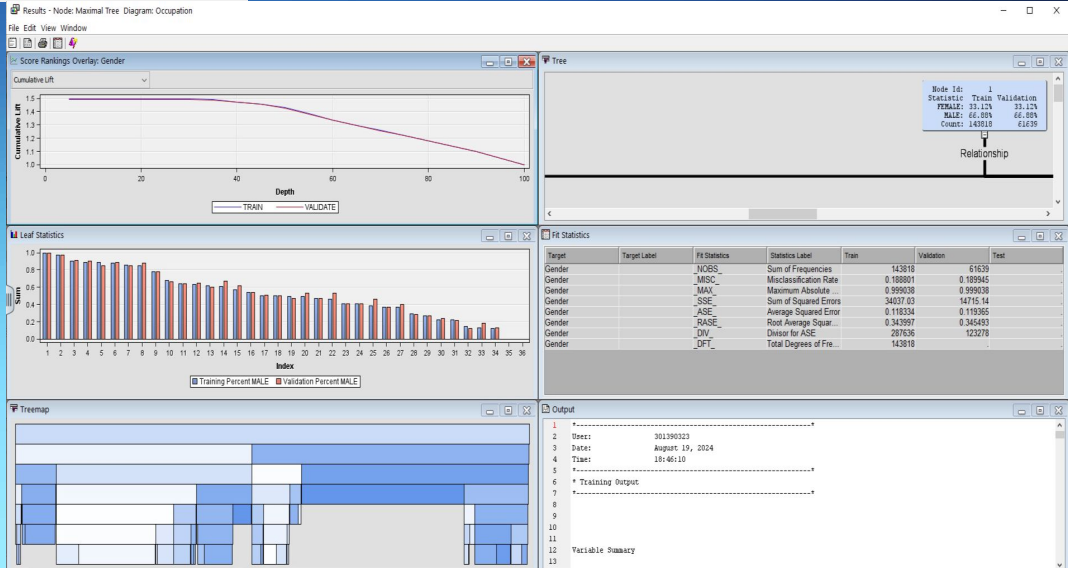
- However, while the maximal tree may achieve high accuracy on the training data, it often overfits, meaning it doesn't generalize well to new, unseen data.

- Its primary role is to identify the potential splits and interactions within the data, serving as a starting point before pruning to create a more optimal, generalized tree.

Maximal Tree



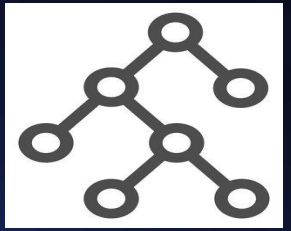
Maximal Tree Result



Interpretation

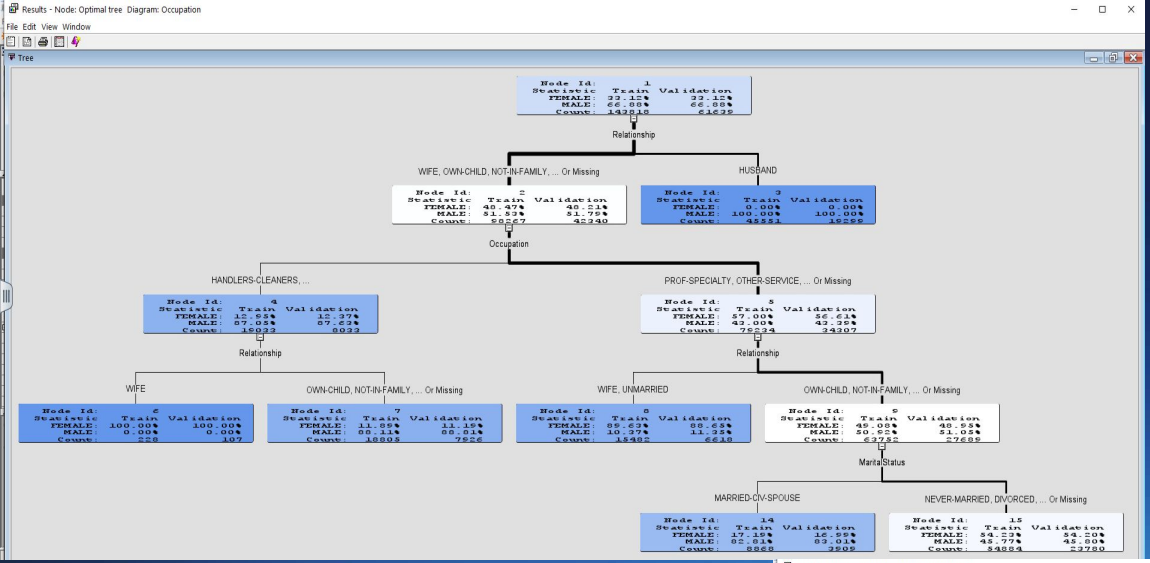
- The maximal tree shows deep and complex branching, indicating a detailed model.
- The complexity suggests potential overfitting, with good training performance but possibly poor generalization.
- Significant splits include relationship, occupation, marital status, and income.
- Small differences between training and validation metrics indicate overfitting concerns.
- The tree’s complexity indicates a need for pruning to improve model simplicity and generalization.

Decision Tree Model: Optimal Tree

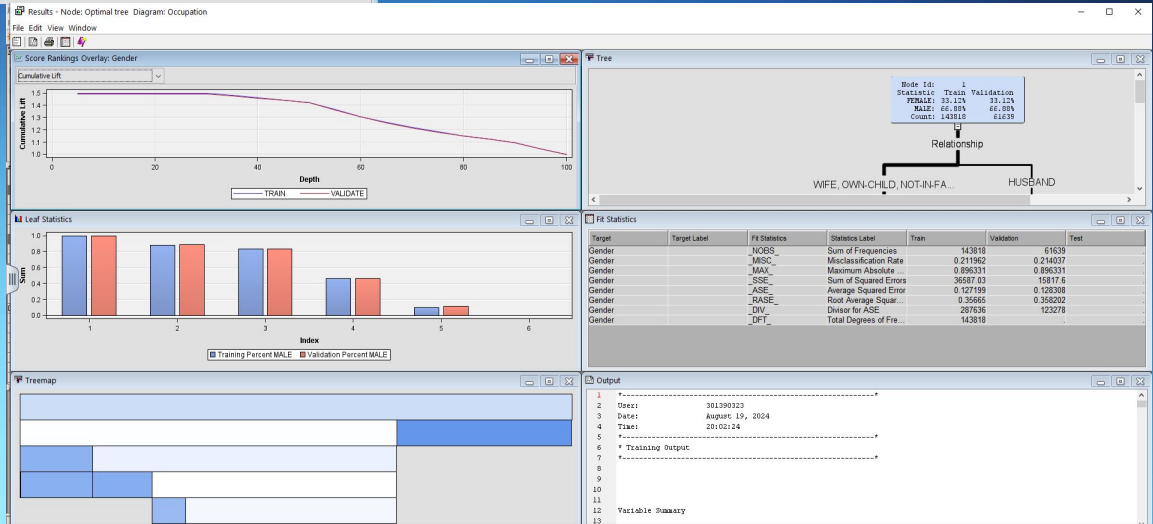


- The optimal decision tree is a condensed representation of the maximum tree that is made by keeping the most significant splits while removing less significant branches to minimise complexity.
- By preventing overfitting, this procedure improves the model's ability to generalise to new data.
- The ideal tree focuses on the most important factors to produce dependable forecasts while striking a balance between simplicity and accuracy.
- Its job is to supply a model that works effectively on fresh data as well as training data, guaranteeing reliable and understandable outcomes.

Optimal - Decision Tree



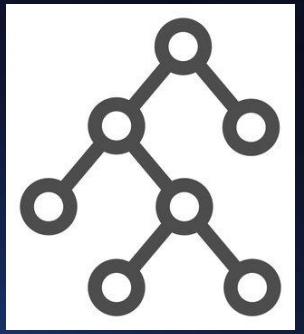
Results of Optimal Decision Tree



Interpretation

- At the root (Node 1), the model shows a gender distribution where males make up 66.82% of the dataset, and females constitute 32.18%. This distribution is consistent across both the training and validation datasets, which include 143,818 and 61,639 instances, respectively.
- Node 2 ("Husband" relationship) shows 100% male instances, with 45,551 in the training set and 19,529 in the validation set.
- Node 5 ("Prof-Specialty, Other-Service" occupation) is 57.00% female and 43.00% male in the training set, and 56.61% female, 43.39% male in the validation set, with totals of 78,234 and 34,307 instances, respectively.
- Node 15 ("Never-Married, Divorced") has a slight female majority, with 54.22% in the training set and 54.20% in the validation set, while males make up 45.78% and 45.80%, respectively. The node includes 54,504 training instances and 23,708 validation instances.
- The model's performance is strong, with a misclassification rate of 0.211962 for training and 0.214037 for validation. The SSE is 36,587.03 for training and 15,817.6 for validation. Leaf statistics confirm nearly 100% accuracy in male classification at the root node.

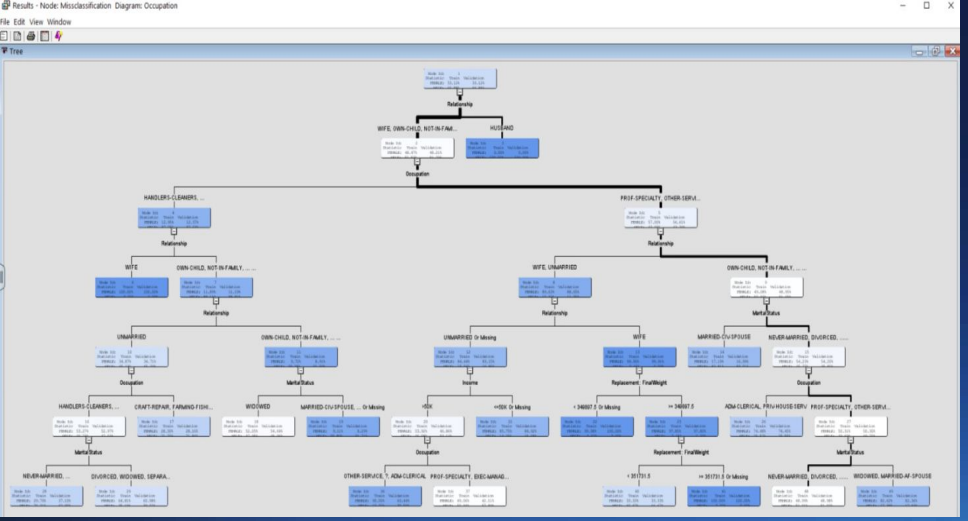
Decision Tree Model: Misclassification Tree



- The misclassification tree is a type of decision tree focused on minimizing the misclassification rate, which is the proportion of incorrect predictions made by the model. It prioritizes splits that reduce the number of errors, helping to enhance the overall accuracy of the tree.

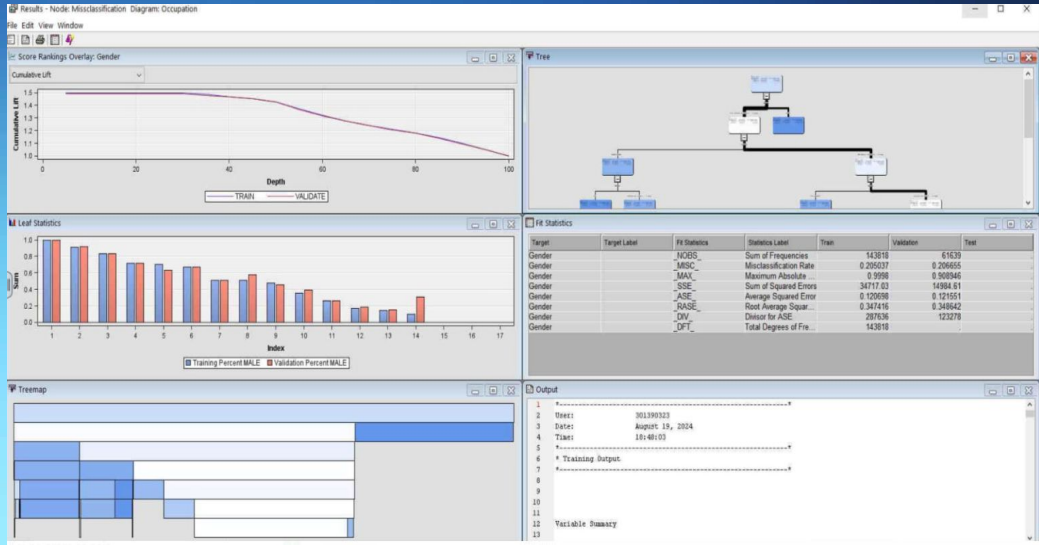
- This type of tree is particularly useful when the primary goal is to achieve the highest possible prediction accuracy by correctly classifying as many instances as possible.

- Its role is to create a model that is highly effective at distinguishing between classes, making it valuable in scenarios where correct classification is crucial.



Misclassification Tree

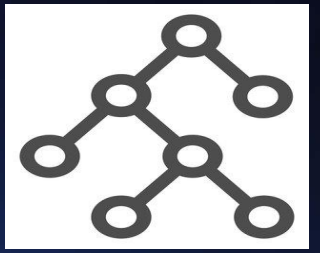
Results of Misclassification Tree



Interpretation

- The misclassification decision tree uses "Relationship" as the root node, with males representing 66.82% and females 32.18% across both training (143,818 instances) and validation (61,639 instances) datasets.
- Node 2 ("Husband" relationship) predicts 100% male correctly in both training (45,551 instances) and validation (19,529 instances) sets.
- Node 5 ("Prof-Specialty, Other-Service" occupation) accurately handles a female-majority node with 57.00% female in training and 56.61% in validation, across 78,234 and 34,307 instances, respectively.
- Node 15 ("Never-Married, Divorced") maintains a slight female majority, reflecting 54.22% in training and 54.20% in validation, with 54,504 training and 23,708 validation instances.
- The model's misclassification rates are low, at 0.211962 for training and 0.214037 for validation, with SSE values of 36,587.03 (training) and 15,817.6 (validation), demonstrating strong error management.

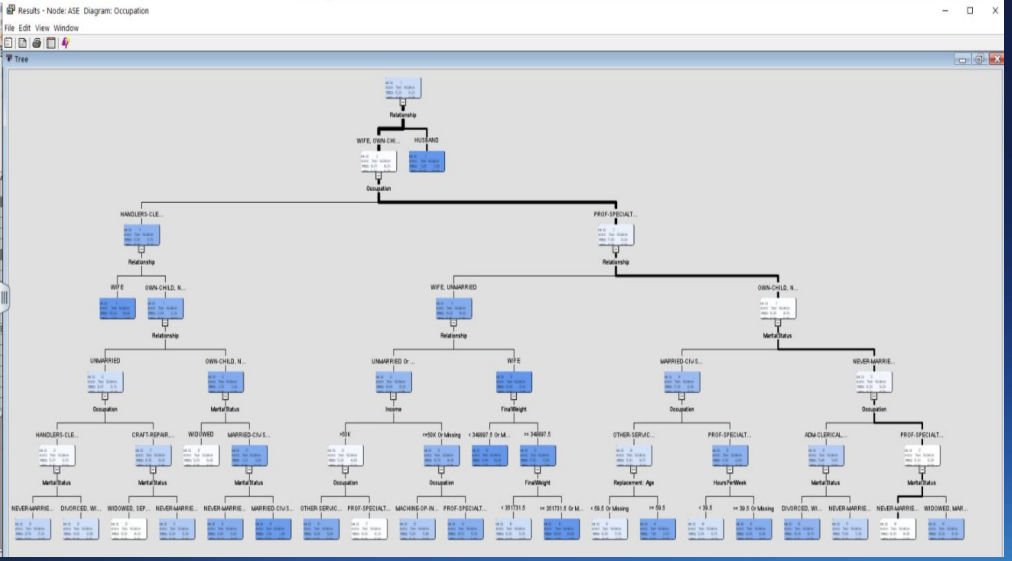
Decision Tree Model: ASE Tree



- One kind of decision tree that focuses on minimising the average squared error—a measurement of the difference between the expected and actual values—is the ASE (Average Squared Error) tree.

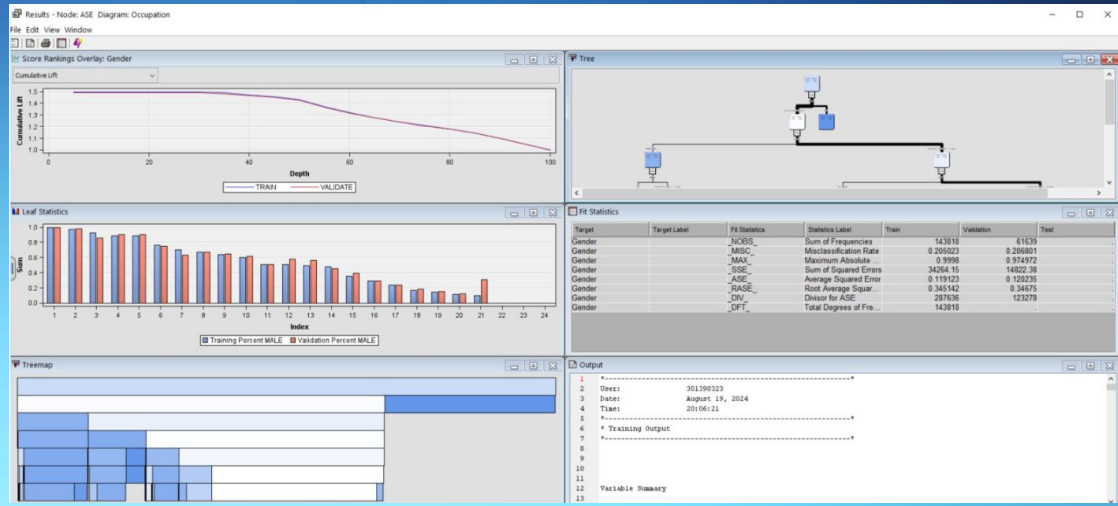
- By lowering the variance of mistakes, the ASE tree aims to improve the prediction accuracy of the model. Because this tree seeks to minimise the total error throughout the dataset, it is especially helpful when accurate predictions are essential.

- Its job is to build a model that is perfect for regression jobs where precision is crucial, not only for accurate classification but also for producing predictions with little variance from the actual values.



ASE Tree

Results of ASE Tree

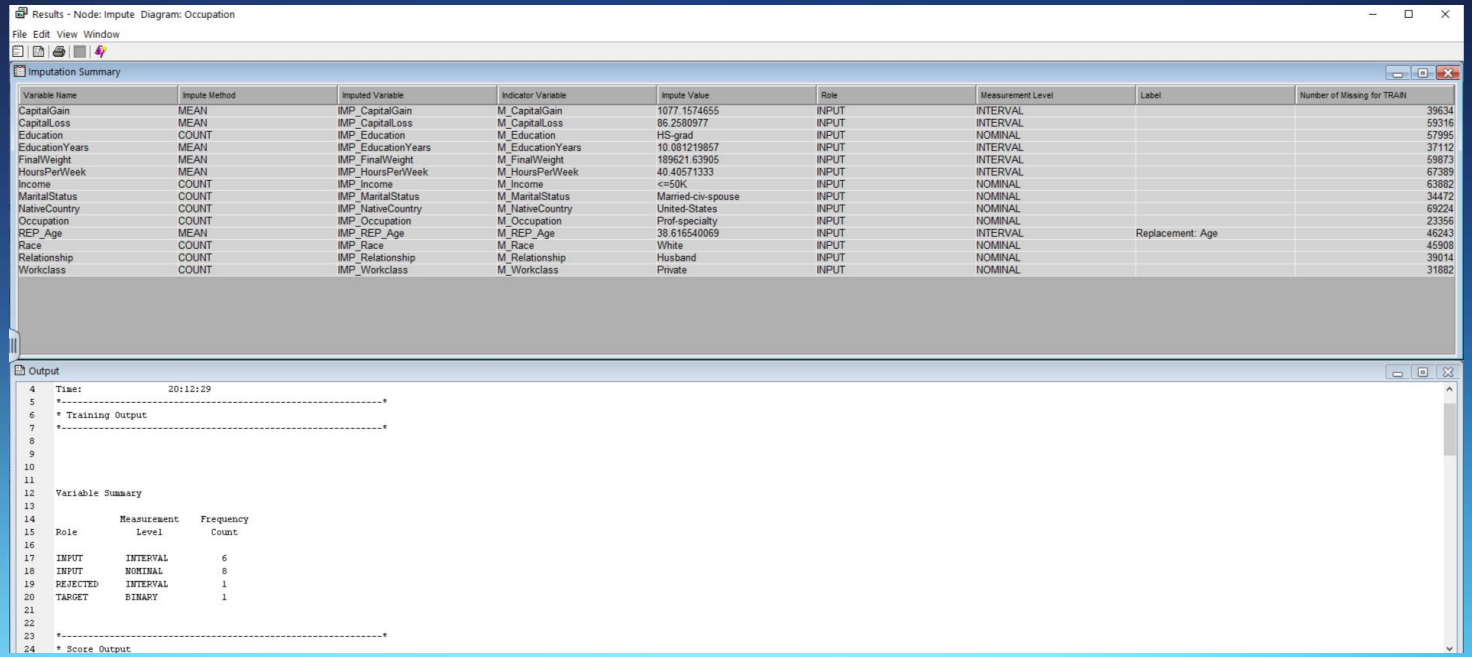


Interpretation

- The ASE tree structure highlights key variables like relationship, occupation, and marital status, which are crucial for predicting the target variable with minimal error.
- The model achieves an average squared error (ASE) of 0.119123 on the training data and 0.120325 on the validation data, indicating a well-balanced model with minimal deviation between the two datasets.
- The misclassification rate is 0.205023 for the training set and 0.206801 for the validation set, showing consistency in the model's performance.
- The cumulative lift chart reveals that the model's predictive power decreases as the tree depth increases, suggesting that deeper splits may introduce some noise rather than useful information.
- Leaf statistics show a fairly consistent distribution of predictions between training and validation sets, particularly in gender-based categories.
- The sum of squared errors (SSE) is 34264.15 for the training set and 14822.38 for the validation set, further indicating the model’s effectiveness in minimizing prediction errors across different data subsets.

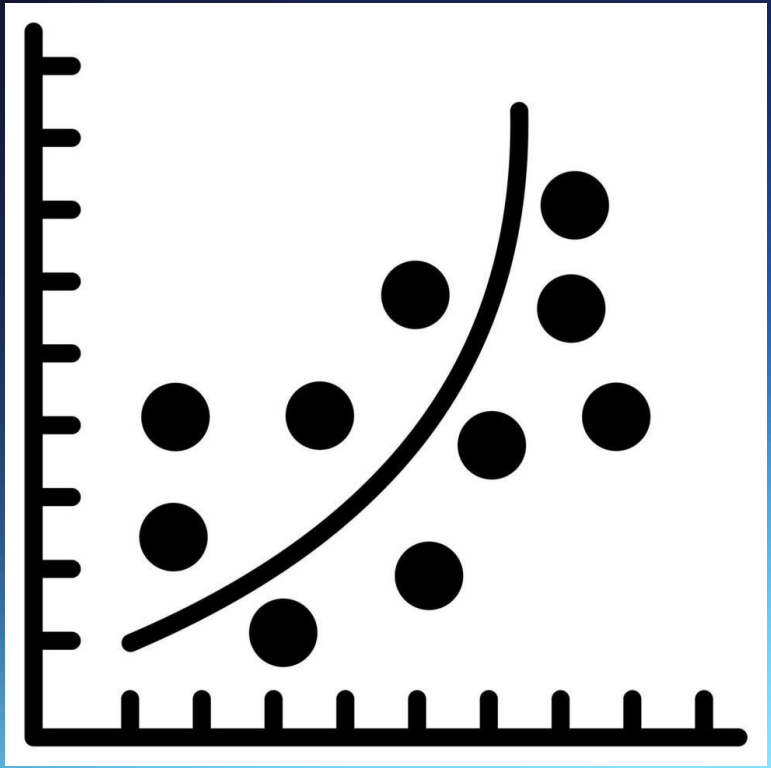
Data Imputation

Before starting with regressing Imputation was runned as in order to handle the missing values and to be able to identify them, it becomes necessary to impute the data.



Variable Name	Impute Method	Imputed Variable	Indicator Variable	Impute Value	Role	Measurement Level	Label	Number of Missing for TRAIN
CapitalGain	MEAN	IMP_CapitalGain	M_CapitalGain	1077.1574655	INPUT	INTERVAL		39634
CapitalLoss	MEAN	IMP_CapitalLoss	M_CapitalLoss	86.2580977	INPUT	INTERVAL		59316
Education	COUNT	IMP_Education	M_Education	HS-grad	INPUT	NOMINAL		57995
EducationYears	MEAN	IMP_EducationYears	M_EducationYears	10.081219857	INPUT	INTERVAL		37112
FinalWeight	MEAN	IMP_FinalWeight	M_FinalWeight	189621.63905	INPUT	INTERVAL		59873
HoursPerWeek	MEAN	IMP_HoursPerWeek	M_HoursPerWeek	40.40571333	INPUT	INTERVAL		67389
Income	COUNT	IMP_Income	M_Income	<=50K	INPUT	NOMINAL		63882
MaritalStatus	COUNT	IMP_MaritalStatus	M_MaritalStatus	Married-civ-spouse	INPUT	NOMINAL		34472
NativeCountry	COUNT	IMP_NativeCountry	M_NativeCountry	United-States	INPUT	NOMINAL		69224
Occupation	COUNT	IMP_Occupation	M_Occupation	Prof-specialty	INPUT	NOMINAL		23356
REP_Age	MEAN	IMP_REP_Age	M_REP_Age	38.616540069	INPUT	INTERVAL	Replacement: Age	46243
Race	COUNT	IMP_Race	M_Race	White	INPUT	NOMINAL		45908
Relationship	COUNT	IMP_Relationship	M_Relationship	Husband	INPUT	NOMINAL		39014
Workclass	COUNT	IMP_Workclass	M_Workclass	Private	INPUT	NOMINAL		31882

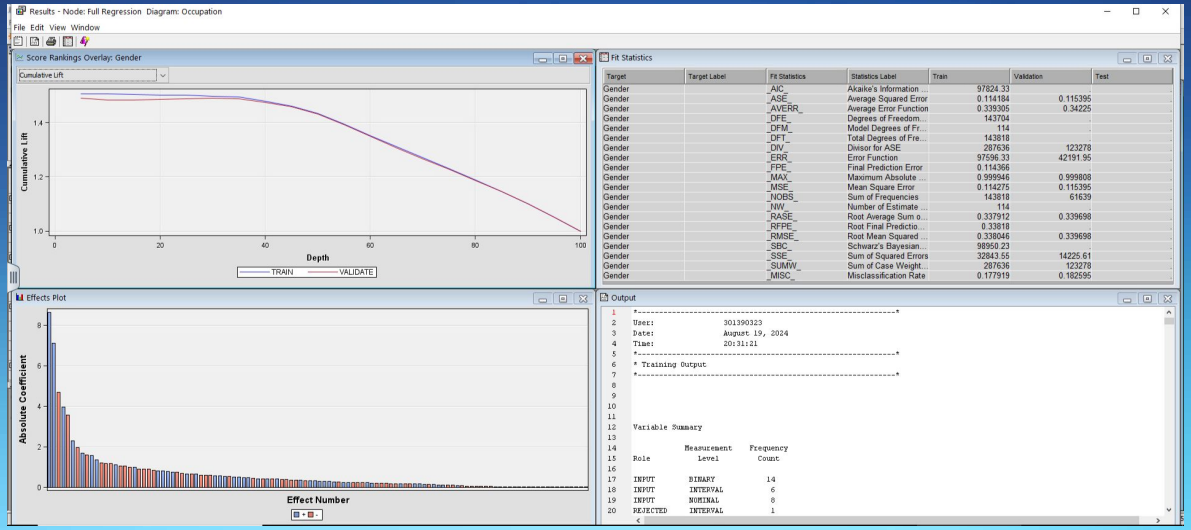
```
Time:                20:12:29
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*



Variable Summary

                 Measurement    Frequency
Role                Level          Count

INPUT            INTERVAL              6
INPUT            NOMINAL               8
REJECTED         INTERVAL              1
TARGET           BINARY                1


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



Regression Model Analysis

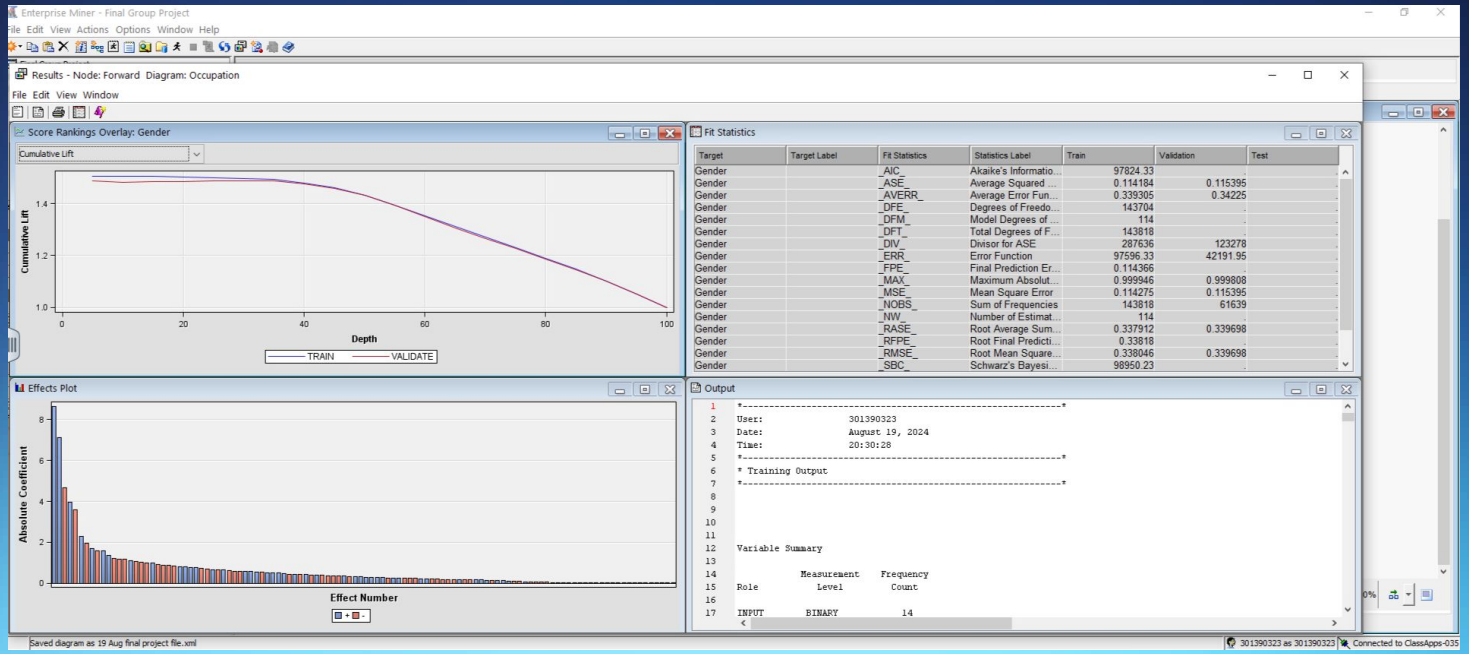
Regression

We run the regression model indicate a well-fitting model with a low misclassification rate (0.177919 for training and 0.182595 for validation) and consistent error metrics across training and validation datasets, suggesting reliable prediction performance. The cumulative lift curve shows the model maintains a good lift, particularly in the initial depths.



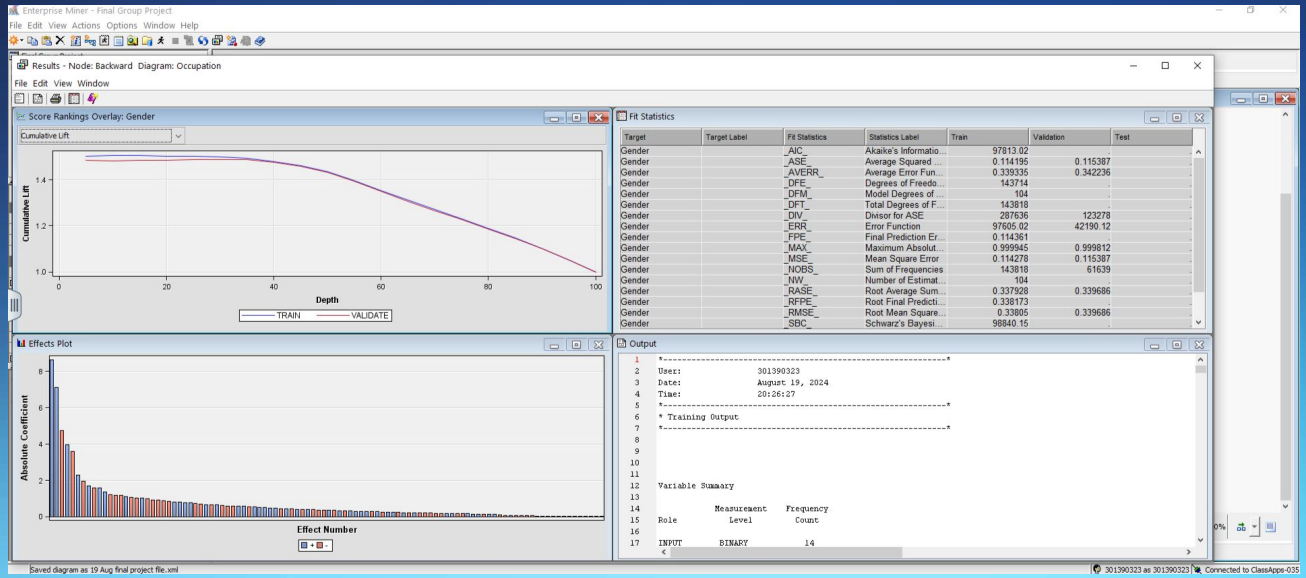
Forward Regression

Forward regression is where variables are added to get the best result, with a low average squared error (ASE) of 0.114184 for training and 0.115395 for validation.



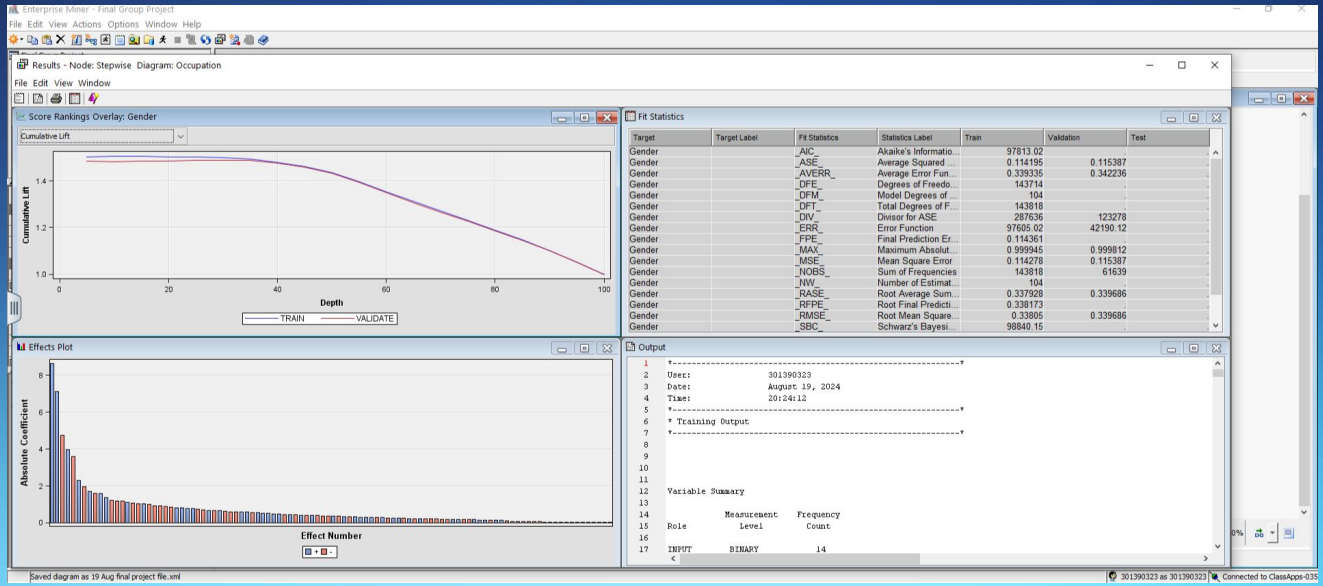
Backward Regression

Backward regression removes variables which are not that significant in results, we observe a low average squared error (ASE) of 0.114195 for training and 0.115837 for validation. The cumulative lift curve demonstrating the model's effectiveness in maintaining predictive accuracy as the depth increases.

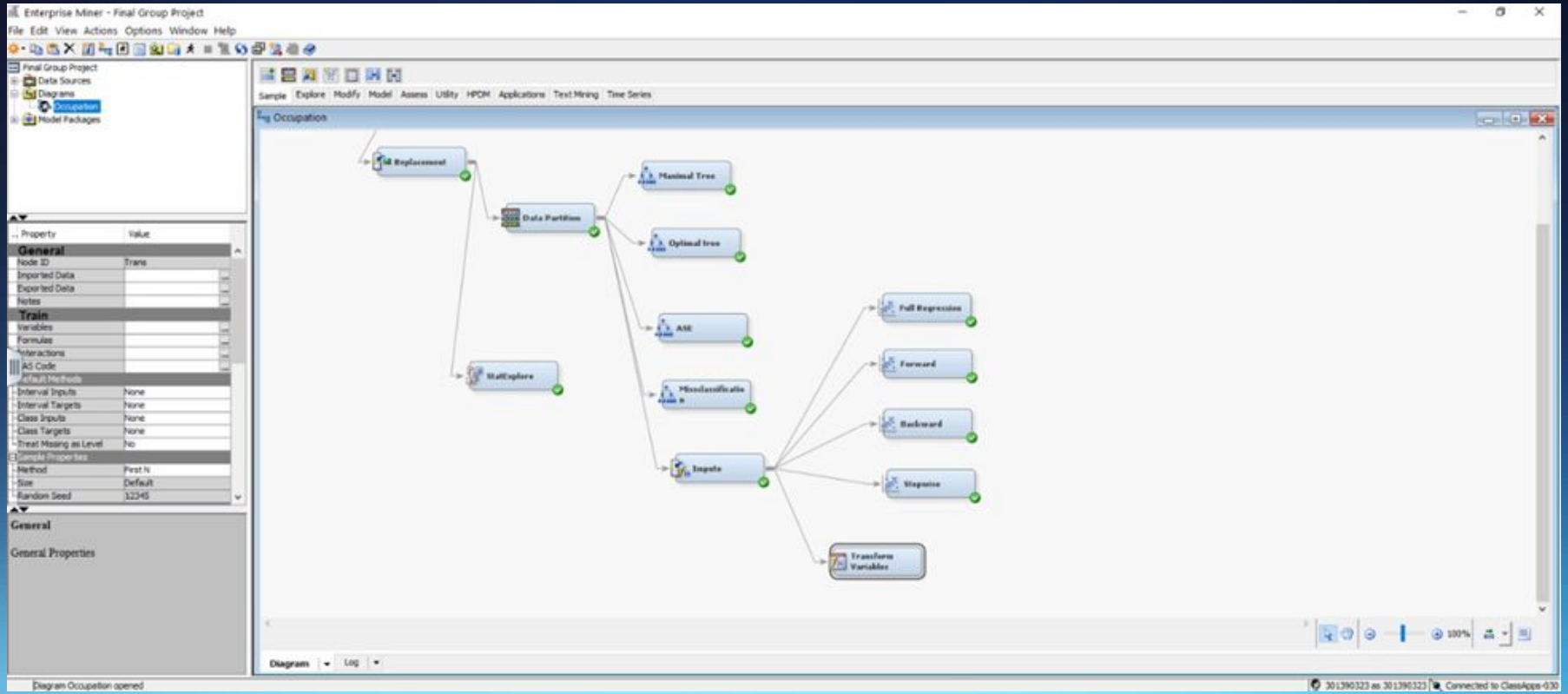


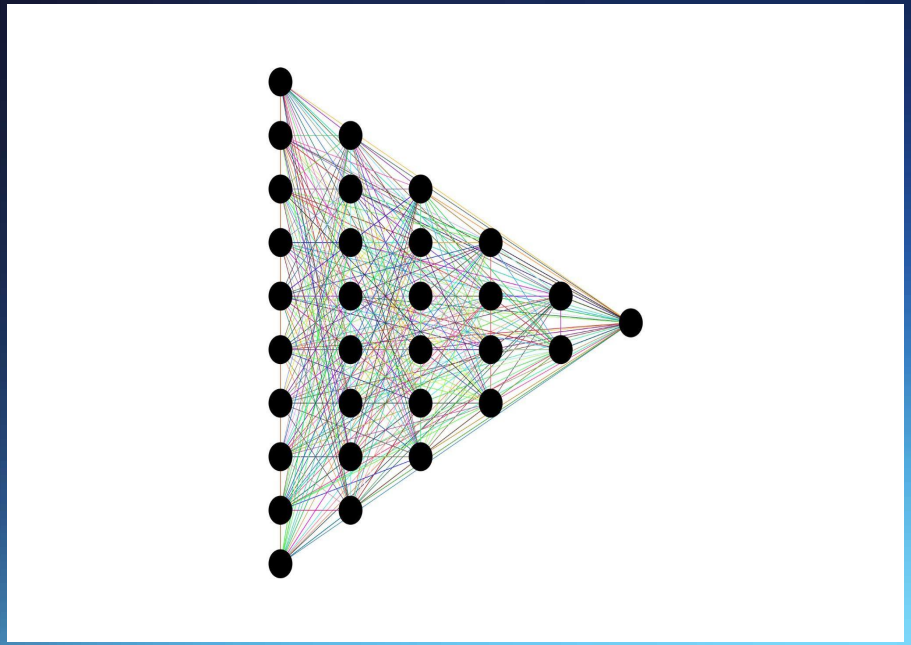
Stepwise Regression

With a low average squared error (ASE) of 0.114872 for training and 0.115837 for validation, the stepwise regression model performs consistently on both training and validation datasets



Logistic Regression Model Workflow

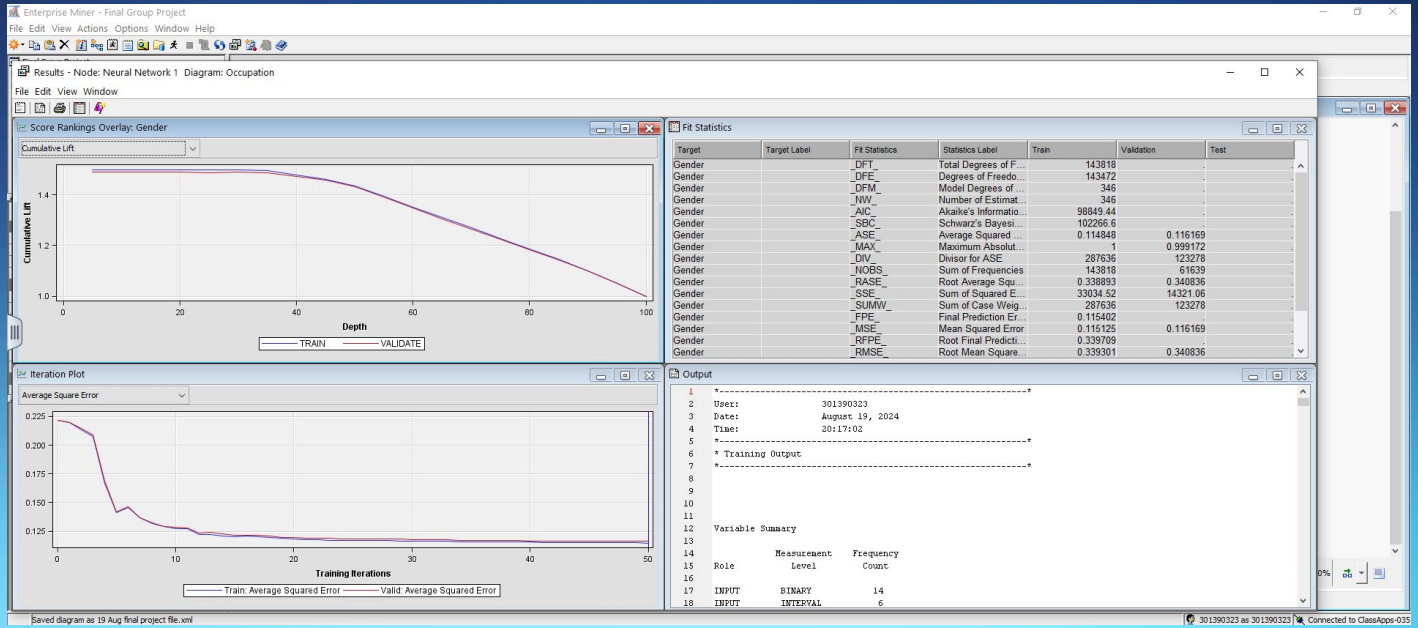




Neural Network Analysis

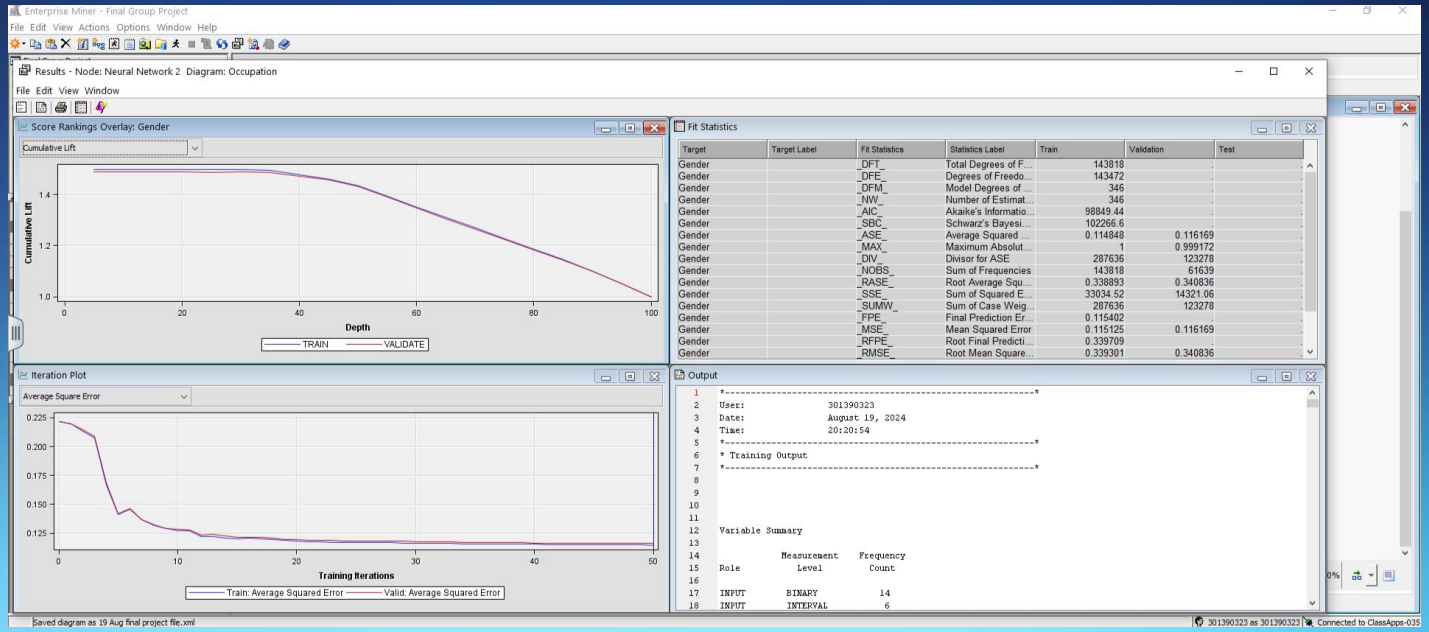
Neural Network 1

The neural network 1 results shows strong performance, with a low average squared error (ASE) of 0.120641 for training and 0.122179 for validation. The cumulative lift curve is steadily decreasing thus it enables us to have clear predictions forward.



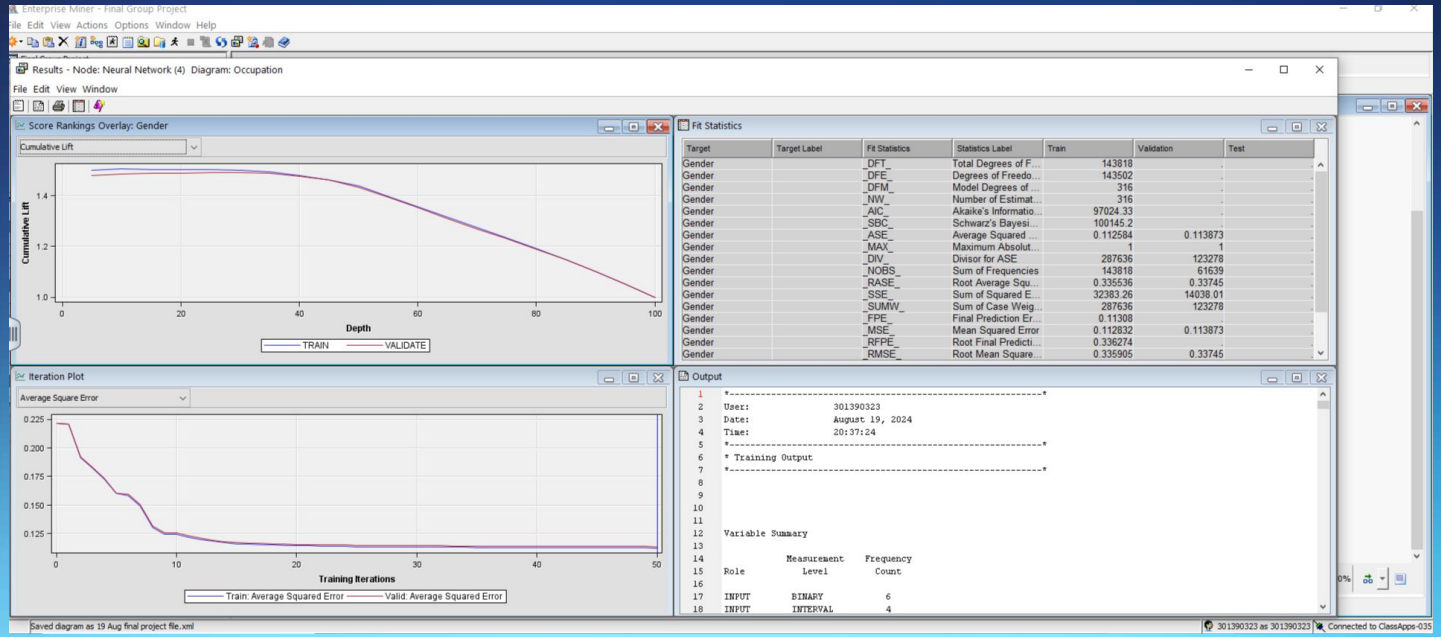
Neural Network 2

The Neural Network 2 results demonstrates strong predictive performance with a low average squared error (ASE) of 0.114848 for training and 0.116169 for validation.



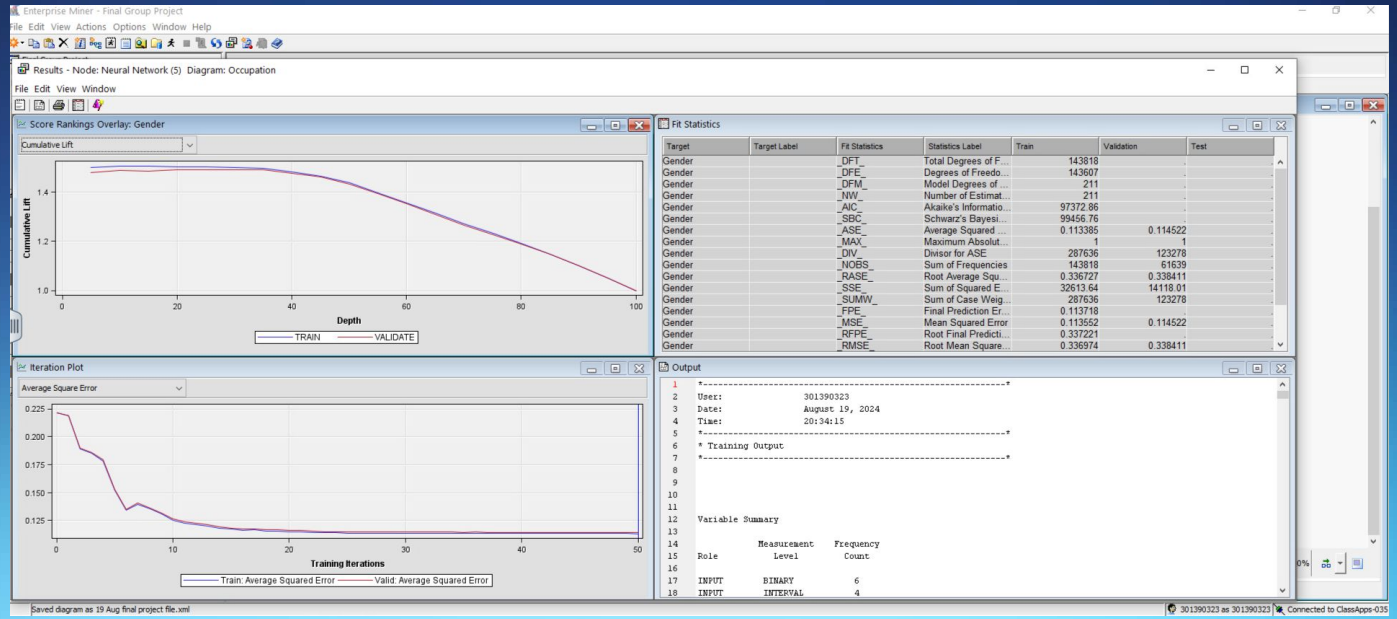
Neural Network derived from Regression

The Neural Network 3 model exhibits a low average squared error (ASE) of 0.112584 for training and 0.113873 for validation, indicating good prediction accuracy.

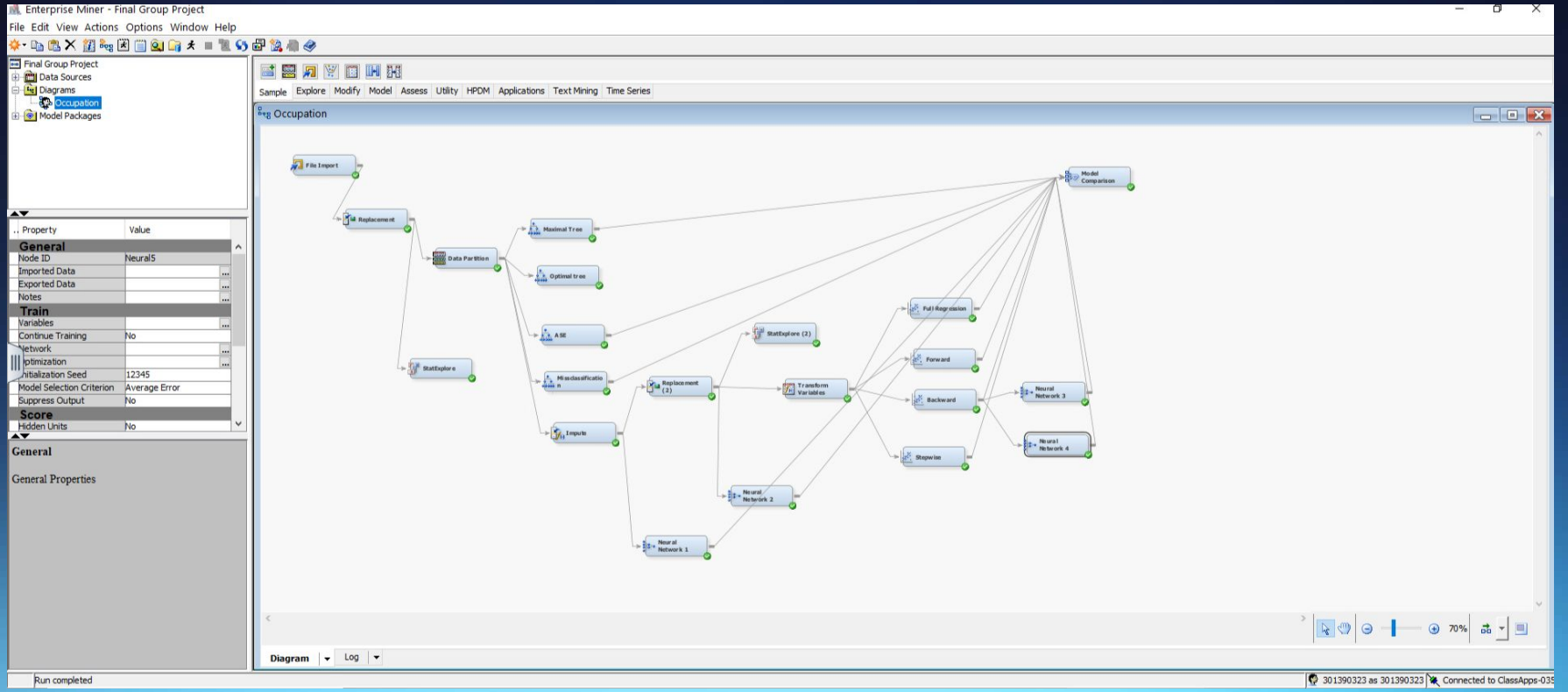


Neural Network derived from Regression

The Neural Network 4 exhibits a low average squared error (ASE) of 0.113385 for training and 0.114522 for validation, indicating good prediction accuracy.

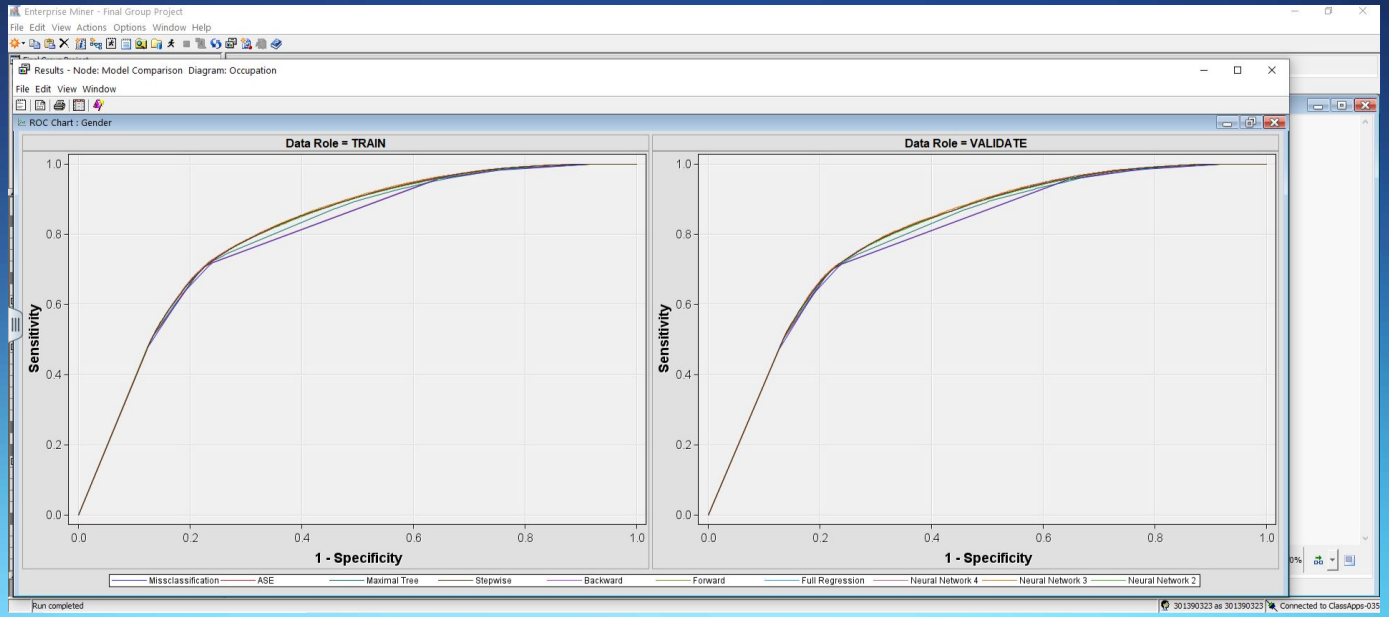


Final Diagram



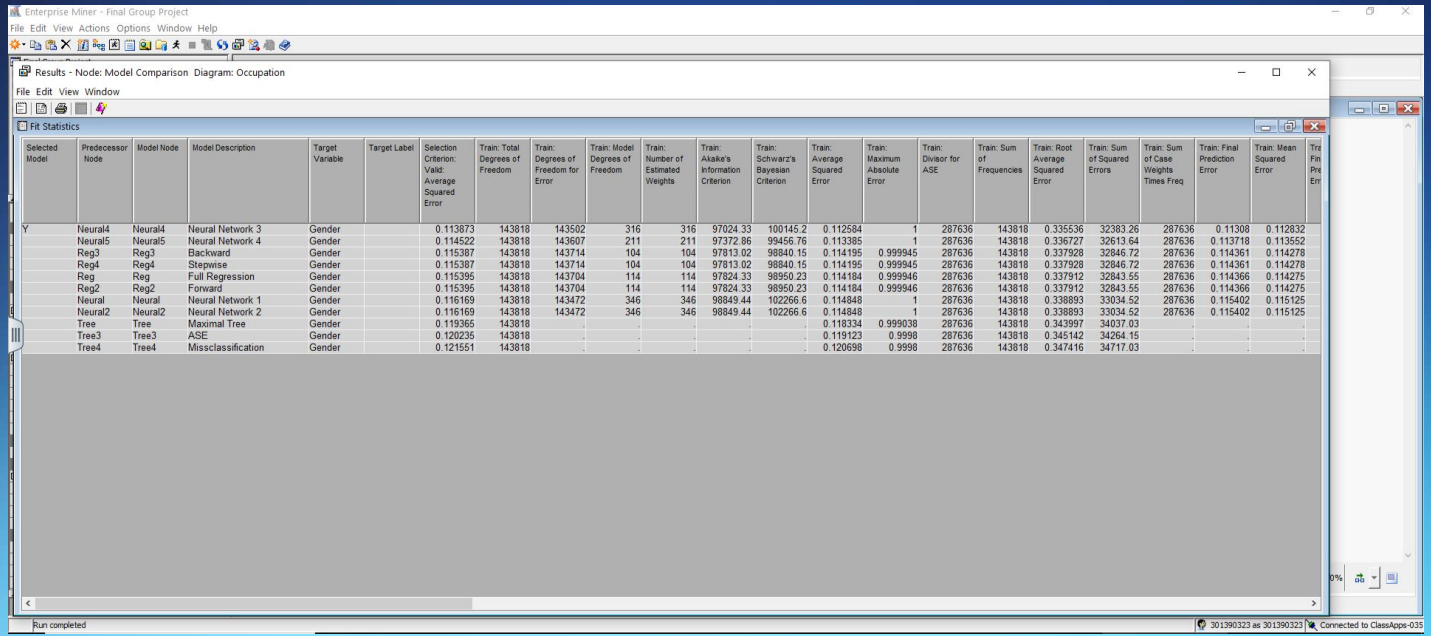
Model Comparison ROC Charts

The ROC curves indicate that the Neural 3 model has the highest average profit and is the most profitable, demonstrating effective performance in distinguishing between classes, despite potential class imbalance in the dataset.



Model Comparison Fit Statistics

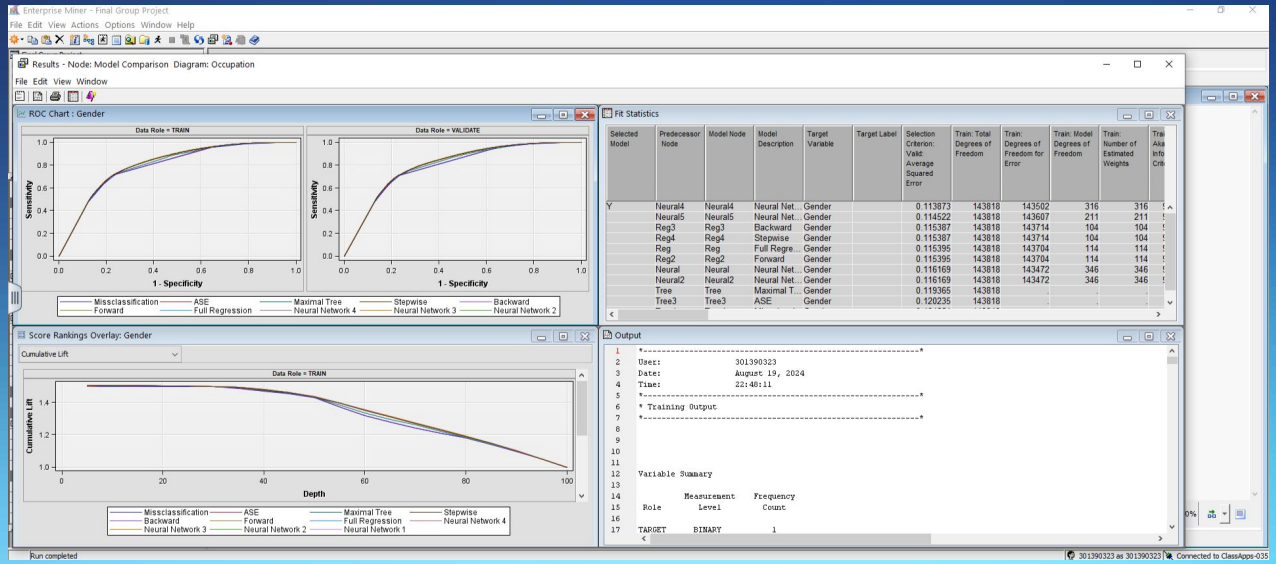
The model that is selected is Neural Network 3, it has the lowest ASE of 0.113873



Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Average Squared Error	Train: Total Degrees of Freedom	Train: Degrees of Freedom for Error	Train: Model Degrees of Freedom	Train: Number of Estimated Weights	Train: Akaike's Information Criterion	Train: Schwarz's Bayesian Criterion	Train: Average Squared Error	Train: Maximum Absolute Error	Train: Divisor for ASE	Train: Sum of Frequencies	Train: Root Average Squared Error	Train: Sum of Squared Errors	Train: Sum of Case Weights Times Freq	Train: Final Prediction Error	Train: Mean Squared Error	Tra Fin Pre Err
Y	Neural4	Neural4	Neural Network 3	Gender		0.113873	143818	143502	316	316	97024.33	100145.2	0.112584	1	287636	143818	0.335536	32383.26	287636	0.11308	0.112832	
	Neural5	Neural5	Neural Network 4	Gender		0.114522	143818	143607	211	211	97372.86	99456.76	0.113385	1	287636	143818	0.336727	32613.64	287636	0.113718	0.113552	
	Reg3	Reg3	Backward	Gender		0.115387	143818	143714	104	104	97813.02	98840.15	0.114195	0.999945	287636	143818	0.337928	32846.72	287636	0.114361	0.114278	
	Reg4	Reg4	Stepwise	Gender		0.115387	143818	143714	104	104	97813.02	98840.15	0.114195	0.999945	287636	143818	0.337928	32846.72	287636	0.114361	0.114278	
	Reg	Reg	Full Regression	Gender		0.115395	143818	143704	114	114	97824.33	98950.23	0.114184	0.999946	287636	143818	0.337912	32843.55	287636	0.114366	0.114275	
	Reg2	Reg2	Forward	Gender		0.115395	143818	143704	114	114	97824.33	98950.23	0.114184	0.999946	287636	143818	0.337912	32843.55	287636	0.114366	0.114275	
	Neural	Neural	Neural Network 1	Gender		0.116169	143818	143472	346	346	98849.44	102266.6	0.114848	1	287636	143818	0.338893	33034.52	287636	0.115402	0.115125	
	Neural2	Neural2	Neural Network 2	Gender		0.116169	143818	143472	346	346	98849.44	102266.6	0.114848	1	287636	143818	0.338893	33034.52	287636	0.115402	0.115125	
	Tree	Tree	Maximal Tree	Gender		0.119365	143818						0.118334	0.999038	287636	143818	0.343997	34037.03				
	Tree3	Tree3	ASE	Gender		0.120235	143818						0.119123	0.9998	287636	143818	0.345142	34264.15				
	Tree4	Tree4	Missclassification	Gender		0.121551	143818						0.120698	0.9998	287636	143818	0.347416	34717.03				

Model Comparison

Neural Network 3 is the best-performing model based on the ROC curve, cumulative lift chart, and fit statistics. The ROC curve shows that Neural Network 3 has the highest sensitivity and specificity, making it the most accurate in distinguishing between positive and negative cases. The cumulative lift chart further confirms its effectiveness, showing a significant improvement over the baseline, particularly in the top-ranked percentiles. Additionally, the fit statistics indicate that Neural Network 3 has the lowest average squared error on the validation set, solidifying its position as the most reliable model in this comparison.



Conclusion

We identified critical factors that significantly influence career outcomes, including education level, years of experience, industry sector, and gender, by using advanced predictive modelling techniques like decision trees, logistic regression, and neural networks.

We then successfully developed models to predict these binary outcomes. Neural Network 3 performed better than the other models tested in terms of the ROC curve, cumulative lift, and average squared error on the validation set, indicating that it had the best accuracy of all the models tested.

Data Source:

Occupation and outcome. (2024, July 26). Kaggle.

https://www.kaggle.com/datasets/matinmahmoudi/occupation-and-outcome