MGMT 571 Homework #2- Ian Bach

1). The target variable in Regression node of SAS enterprise miner

(a) could be a binary variable.

(b) could be an interval variable.

(c) could be an ordinal variable.

(d) all the above.

Answer: D

2). With regards to Transform Variables node, which is NOT true?

(a) It helps stabilize variances.

(b) It helps handle nonlinearity.

(c) The optimal binning maximizes the relationship to the target.

(d) We usually transform variables before we impute missing values.

Answer: D

3). With regards to linear regression, which is NOT true?

(a) R2 measures the proportion of variability in the predictors that can be explained using the target variable.

(b) R2 is always non-decreasing with the addition of a new predictor, whether the new predictor is useful or not.

(c) Residual plots display the residual values on the y-axis and fitted values on the x-axis.

(d) Linear regression is sensitive to outliers.

Answer: A

4). Given the following four training instances: (x1 = −1, y1 = 0),(x2 = 0, y2 = 1),(x3 = 1, y3 = 2),(x4 = 2, y4 = 3). Which of the following parameter (β0, β1) best model this data using the ordinary least squares regression y = β0 + β1x? Show your work.

(a) (1,1)

(b) (0.7,0.6)

(c) (0.6,0.6)

(d) (0.6,0.7)

Calculating the Means of X and Y

Future x = -1 + 0 + 1 + 2 / 4 = 0.5

Future y = 0 + 1 + 2 + 3 / 4 = 1.5

Calculate β1

Numerator

1. (-1-0.5)(0.1.5) = 2.25
2. (0-0.5)(1-1.5) = 0.25
3. (1-0.5)(2-1.5) = 0.25
4. (2-0.5)(3-1.5) = 2.25

= 5

Denominator:

1. (1-0.5)^2 = 2.25
2. (0-0.5)^2 = 0.25
3. (1-0.5)^2 = 0.25
4. (2-0.5)^2 = 2.25

= 5

Calculate β0

1.5 – (1)(0.5) = 1.5 – 0.5 = 1

Best model parameters are (1,1)

5). Describe two approaches to detect outliers of the class variable and two approaches to detect outliers of the interval variable, respectively.

- Detecting Outliers in Class Variables
- Detecting Outliers in Interval Variables

Detecting Outliers in Class Variables:

Class variables represent distinct categories or classes including outliers in their context typically refer to unexpected categories. One approach to detecting outlies for class variables if frequency – based detection where you can analyze the frequency of each category. Categories with very low counts can signal outliers especially if they don't logically fit within the dataset context. For example, if the dataset primarily contains yes or no responses, a small group of maybe responses might be an outlier.

Another approach is to use cross-tabulation with other variables. By examining the distribution of the class variable in relation to other variables, we can identify outliers that appear in unexpected combinations or contexts. If a class appears alongside another variable in a way that rarely or never occurs within the training data, it may be an indication of an outlier.

Detecting Outliers in Interval Variables:

Interval variables (continuous numerical values) can have outliers in the form of values that are unusually high or low compared to the rest of the data. One common approach to detect outliers in interval variables is using the standard deviation and z-score method. This involves calculating the z score for each observation, which indicates how many standard deviations a value is from the mean. Typically, values with z-scores above 3 or below -3 are considered outliers, though this threshold can be adjusted based on the specific distribution of the data.

6). Describe three approaches to handle missing data. Which node does SAS enterprise miner handle missing data?

Imputation Using Mean, Median, or Mode:

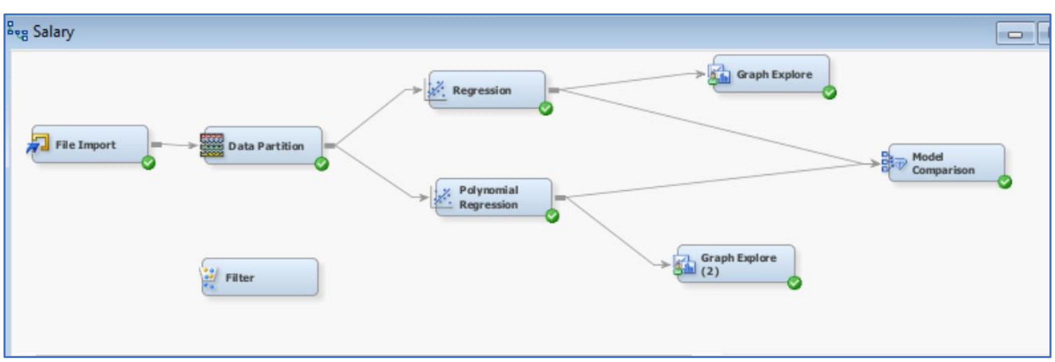
A straightforward approach to handling missing data is to fill in missing values with the mean, median, or mode of the non-missing values within the variable. This method is simple and preserves the sample size, which can be useful when the missingness is random and not related to the outcome variable. However, it can introduce bias if the missing data has a systematic pattern, and it may reduce variability in the data.

Predictive Model-Based Imputation:

Another approach is to use predictive models to estimate missing values based on other available variables. For instance, regression models, k-nearest neighbors, or machine learning models can predict the missing values by treating as dependent variables and using related variables to estimate them. This approach can be more accurate than mean or median imputation as it uses additional information, but it can be computationally intensive and may require assumptions.

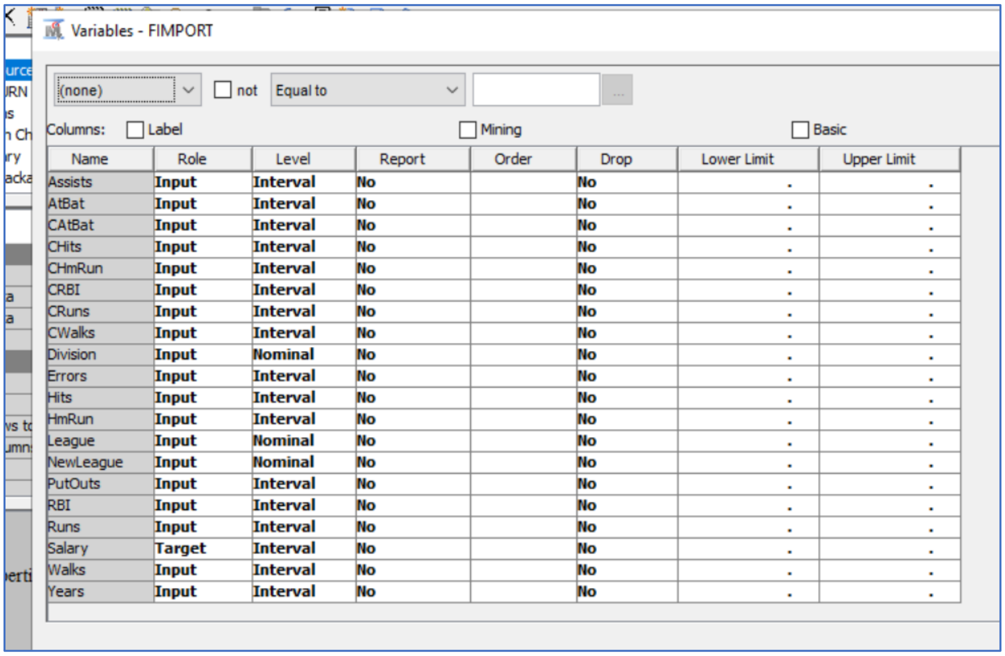
Multiple Imputation:

Multiple imputation is a robust approach where missing values are filled in multiple times to create several complete datasets, each with slightly different imputed values. Each dataset is then analyzed separately, and the results are combined to produce estimates that reflect the uncertainty due to the missing data. This method is effective because it accounts for variability and potential bias in the imputation process providing more accurate estimates than single imputation methods.

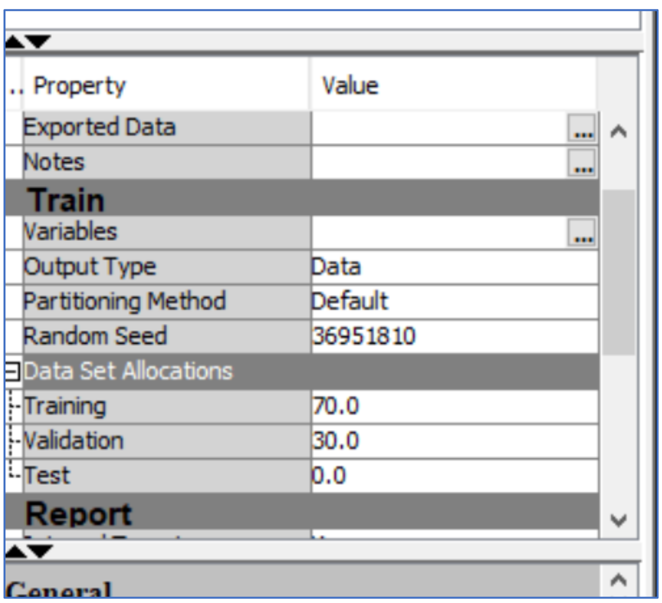


7).

(a) Split data sets into 70% training and 30% validation using Data Partition node. Set the "Random Seed" of Data Partition node using your Purdue ID number.



Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Assists	Input	Interval	No		No	.	.
AtBat	Input	Interval	No		No	.	.
CAtBat	Input	Interval	No		No	.	.
CHits	Input	Interval	No		No	.	.
CHmRun	Input	Interval	No		No	.	.
CRBI	Input	Interval	No		No	.	.
CRuns	Input	Interval	No		No	.	.
CWalks	Input	Interval	No		No	.	.
Division	Input	Nominal	No		No	.	.
Errors	Input	Interval	No		No	.	.
Hits	Input	Interval	No		No	.	.
HmRun	Input	Interval	No		No	.	.
League	Input	Nominal	No		No	.	.
NewLeague	Input	Nominal	No		No	.	.
PutOuts	Input	Interval	No		No	.	.
RBI	Input	Interval	No		No	.	.
Runs	Input	Interval	No		No	.	.
Salary	Target	Interval	No		No	.	.
Walks	Input	Interval	No		No	.	.
Years	Input	Interval	No		No	.	.



Property	Value
Exported Data	...
Notes	...
Train	
Variables	...
Output Type	Data
Partitioning Method	Default
Random Seed	36951810
Data Set Allocations	
Training	70.0
Validation	30.0
Test	0.0
Report	

(b) Fit a linear regression model to predict Salary using the rest 19 predictors. Perform model check and report your findings.

Property	Value
Train	
Variables	...
Equation	
Main Effects	Yes
Two-Factor Interactions	No
Polynomial Terms	No
Polynomial Degree	2
User Terms	No
Term Editor	...
Class Targets	
Regression Type	Linear Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	

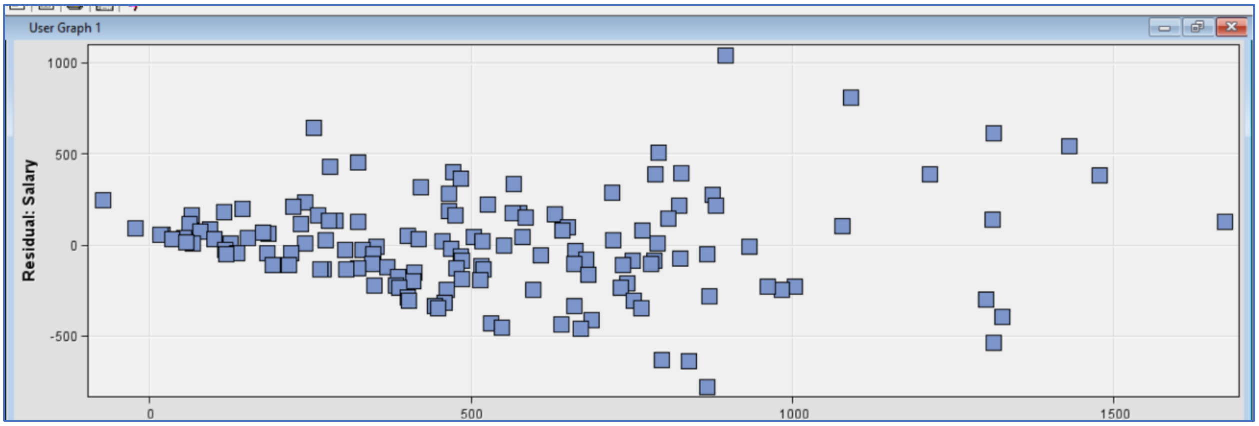
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	19	19062061	1003266	11.72	<.0001
Error	164	14044390	85637		
Corrected Total	183	33106451			

Model Fit Statistics

R-Square	0.5758	Adj R-Sq	0.5266
AIC	2108.6748	BIC	2115.5231
SBC	2172.9735	C(p)	20.0000

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-20.5023	98.6785	-0.21	0.8357
Assists	1	-0.1129	0.2447	-0.46	0.6451
AtBat	1	-1.2341	0.7029	-1.76	0.0810
CAtBat	1	-0.0702	0.1553	-0.45	0.6517
CHits	1	-0.0580	0.7973	-0.07	0.9421
CHmRun	1	0.1835	1.8409	0.10	0.9207
CRBI	1	0.2568	0.8130	0.32	0.7525
CRuns	1	1.4669	0.8544	1.72	0.0879
CWalks	1	-0.6879	0.3810	-1.81	0.0728
Division E	1	59.3783	22.7848	2.61	0.0100
Errors	1	-1.2174	4.7794	-0.25	0.7993
Hits	1	6.6315	2.6316	2.52	0.0127
HmRun	1	-2.0177	7.3762	-0.27	0.7848
League A	1	-21.1684	44.8391	-0.47	0.6375
NewLeague A	1	-0.3086	45.3870	-0.01	0.9946
PutOuts	1	0.1419	0.0841	1.69	0.0935
RBI	1	1.5187	3.0480	0.50	0.6190
Runs	1	-3.9178	3.2842	-1.19	0.2346
Walks	1	6.3844	2.2021	2.90	0.0043



The linear regression model fitted to predict Salary using 19 predictors explains approximately 57.6% of the variance, as indicated by an R-Square of 0.5758 and an Adjusted R-Square of 0.5266. The model is statistically significant, with an F-statistic of 11.72 and a p-value below 0.0001, confirming that at least one predictor significantly impacts Salary. The fit statistics, including an Akaike's Information Criterion (AIC) of 2108.6748 and Schwarz's Bayesian Criterion (SBC) of 2172.9735, support this model as a viable predictive tool. However, the residual plot shows some variability around zero, suggesting potential issues and influential outliers. Although the model provides a reasonable fit, it may benefit from further refinement, such as addressing outliers or exploring interaction effects among predictors.

<u>(c) Fit a nonlinear regression with polynomial basis. Perform model check and report</u>

<u>your findings.</u>

Train	
Variables	...
Equation	
Main Effects	Yes
Two-Factor Interactions	No
Polynomial Terms	Yes
Polynomial Degree	2
User Terms	No
Term Editor	...
Class Targets	
Regression Type	Logistic Regression
Link Function	Logit

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	155	32742415	211241	16.25	<.0001
Error	28	364036	13001		
Corrected Total	183	33106451			

Model Fit Statistics

R-Square	0.9890	Adj R-Sq	0.9281
AIC	1708.5733	BIC	3386.7774
SBC	2210.1033	C(p)	156.0000

The nonlinear regression model with a polynomial basis was fitted to predict Salary using a polynomial degree of 2, allowing for quadratic relationships among the predictors. This model achieved an impressive R-Square of 0.9890 and an Adjusted R-Square of 0.9281, indicating that it explains approximately 99% of the variance in Salary, a significant improvement over the linear model. The F-statistic of 16.25 with a p-value of <0.0001 confirms the overall significance of the model.

However, while the high R-Square suggests a strong fit, the residual plot shows some patterns and larger residuals at the extremes, potentially indicating overfitting or sensitivity to certain observations. The model's fit statistics, such as the Akaike Information Criterion (AIC) of 1708.5733 and Schwarz’s Bayesian Criterion (SBC) of 2210.1033, should be compared with the linear model's values to determine if the complexity of this model is justified. Overall, although the nonlinear model captures more variance in Salary, the residual patterns suggest that the model might be overly complex, which could affect its generalizability to new data.

(d) Compare the average squared error of all regression models on the validation set.

For each model, explore different options in the Property panel. What's your best model?

Train: Akaike's Information Criterion	2108.67	1708.57
Train: Average Squared Error	76328.21	1978.46
Train: Average Error Function	76328.21	1978.46
Selection Criterion: Valid: Average Squared Error	147621.58	1583283.17
Train: Degrees of Freedom for Error	164.00	28.00
Train: Model Degrees of Freedom	20.00	156.00
Train: Total Degrees of Freedom	184.00	184.00
Train: Divisor for ASE	184.00	184.00
Train: Error Function	14044390.21	364036.00
Train: Final Prediction Error	94944.84	24024.11
Train: Maximum Absolute Error	1079.67	170.12
Train: Mean Square Error	85636.53	13001.29
Train: Sum of Frequencies	184.00	184.00
Train: Number of Estimate Weights	20.00	156.00
Train: Root Average Sum of Squares	276.28	44.48
Train: Root Final Prediction Error	308.13	155.00
Train: Root Mean Squared Error	292.64	114.02
Train: Schwarz's Bayesian Criterion	2172.97	2210.10
Train: Sum of Squared Errors	14044390.21	364036.00
Train: Sum of Case Weights Times Freq	184.00	184.00

Statistics	Reg	Reg2
Valid: Average Squared Error	147621.58	1583283.17
Valid: Average Error Function	147621.58	1583283.17
Valid: Divisor for VASE	79.00	79.00
Valid: Error Function	11662105.06	125079370.75
Valid: Maximum Absolute Error	2100.73	6026.77
Valid: Mean Square Error	147621.58	1583283.17
Valid: Sum of Frequencies	79.00	79.00
Valid: Root Average Squared Error	384.22	1258.29
Valid: Root Mean Square Error	384.22	1258.29
Valid: Sum of Square Errors	11662105.06	125079370.75
Valid: Sum of Case Weights Times Freq	79.00	79.00

In comparing the linear and nonlinear regression models, the linear regression model (Reg) outperforms the polynomial regression model (Reg2) based on validation set metrics. The Validation Average Squared Error (ASE) for the linear model is 147,621.58, significantly lower than the 1,583,283.17 ASE for the polynomial model. Similarly, the Validation Root Mean Square Error (RMSE) for the linear model is 384.22, compared to 1,258.29 for the polynomial model. These metrics indicate that the linear model provides more accurate predictions for Salary on the validation set. Adjustments could potentially improve model performance, such as using variable selection methods or experimenting with interaction terms in the linear model, or lowering the polynomial degree in the nonlinear model to reduce overfitting. However, based on the current results, the linear regression model is the best option, balancing simplicity and accuracy without the overfitting observed in the polynomial model.