BSTA 478 - Team Presentation

Prepared by:

Alain Euksuzian

Morgane Rahuba-Pigeon

Wesley Imbayarwo

Table Of Contents

Goal Of The Study

The goal of this study is to determine which of the explanatory variables influences the number of smokers using various predictive modeling techniques to determine future trends.

Motivation Of The Study

Smoking is a significant public health concern that is responsible for causing numerous chronic diseases and millions of deaths worldwide annually. Understanding the complex nature of smoking behavior and its determinants is crucial in developing tailored interventions and policies to reduce the individual and societal harms associated with tobacco use. This study can provide insights into these determinants and inform the development of effective interventions and policies.

Data Description

Dependant Variable (Y):

- Smoker (yes/no)

Explanatory Variables (X):

- Age
- Sex
- Body Mass Index (BMI)
- Number of Children
- Yearly Medical Insurance Cost in USD
- Region

Source

- Kaggle

Data Description: Pre-processing

Original Dataset Sample:

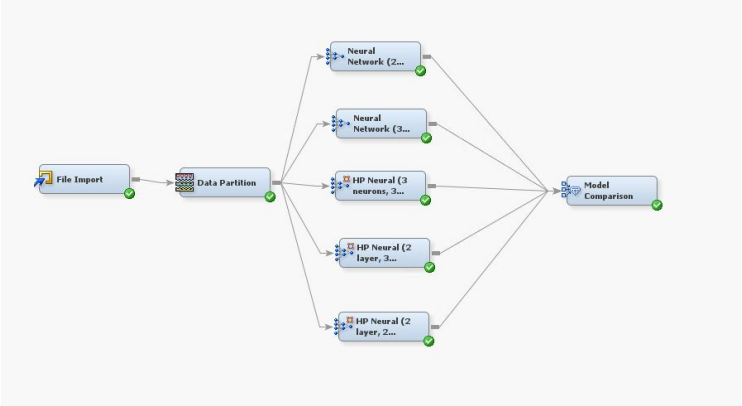
age	sex	bmi	children	smoker	region	charges
19	female	27.9	0	yes	southwest	16884.92
18	male	33.77	1	no	southeast	1725.552
28	male	33	3	no	southeast	4449.462
33	male	22.705	0	no	northwest	21984.47
32	male	28.88	0	no	northwest	3866.855
31	female	25.74	0	no	southeast	3756.622
46	female	33.44	1	no	southeast	8240.59
37	female	27.74	3	no	northwest	7281.506
37	male	29.83	2	no	northeast	6406.411

Processed Dataset Sample:

age	sexValue	bmi	children	smokerValue	regionValue	charges
19	0	27.9	0	1	1	16884.924
18	1	33.77	1	0	2	1725.5523
28	1	33	3	0	2	4449.462
33	1	22.705	0	0	3	21984.471
32	1	28.88	0	0	3	3866.8552
31	0	25.74	0	0	2	3756.6216
46	0	33.44	1	0	2	8240.5896
37	0	27.74	3	0	3	7281.5056

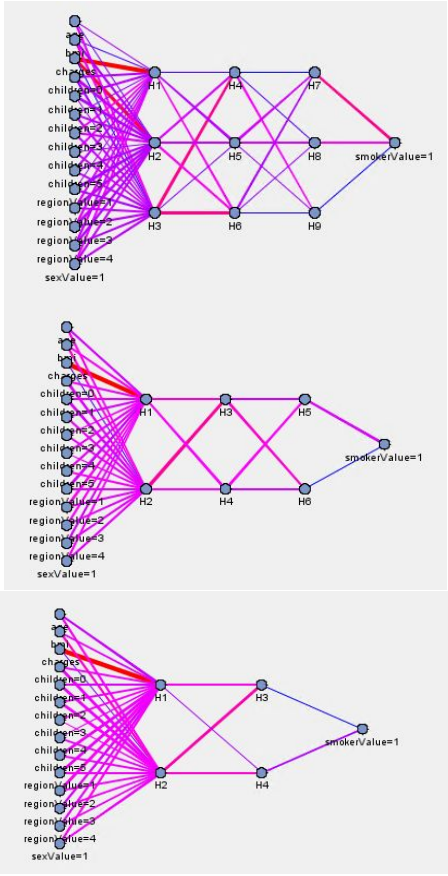
Predictive Models: Neural Network

SAS Flowchart:



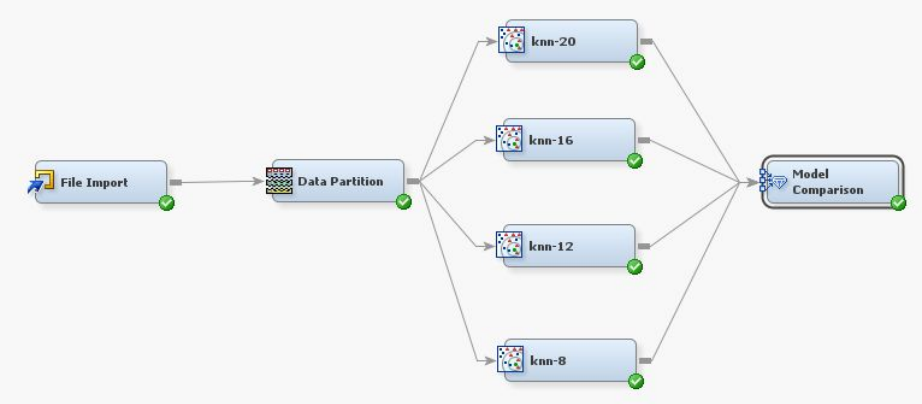
Error Terms:

Model	RMSE	Missclasification
3 Hidden Units	0.133929	0.021536
2 Hidden Units	0.149484	0.029026
3 Layer, 3 Neurons	0.115406	0.023408
3 Layer, 2 Neurons	0.080684	0.008427
2 Layer, 2 Neurons	0.143271	0.026217



Predictive Models: KNN

SAS Flowchart:



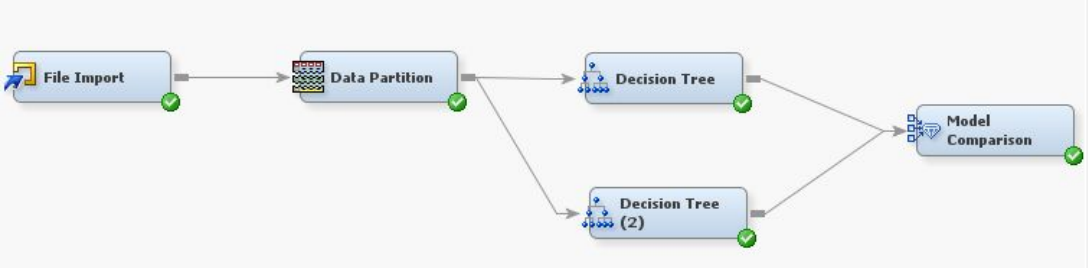
Model Comparison:

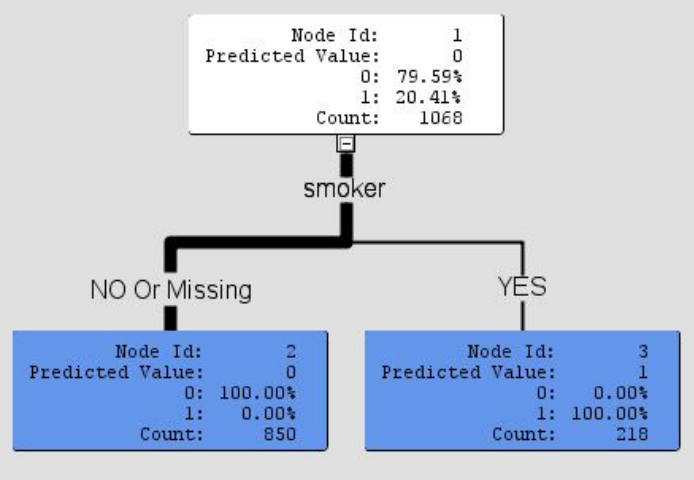
Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Test: Misclassification Rate	Train: Number of Estimated Weights	Train: Sum of Frequencies	Train: Sum of Case Weights Times Freq	Train: Total Degrees of Freedom	Train: Model Degrees of Freedom	Train: Degrees of Freedom for Error	Train: Average Squared Error	Train: Root Average Squared Error
Y	MBR4	MBR4	knn-12	smokerV...		0.055556	3	1068	2136	1068	3	1065	0.041985	0.204903
	MBR3	MBR3	knn-8	smokerV...		0.059259	3	1068	2136	1068	3	1065	0.04174	0.204303
	MBR2	MBR2	knn-16	smokerV...		0.059259	3	1068	2136	1068	3	1065	0.043547	0.208678
	MBR	MBR	knn-20	smokerV...		0.059259	3	1068	2136	1068	3	1065	0.045059	0.21227

Train: Average S

Predictive Models: Decision Trees

SAS Flowchart:



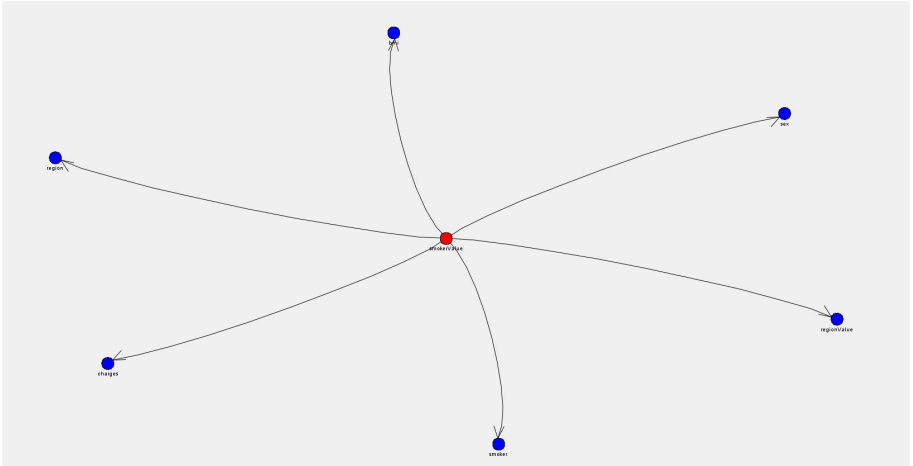


Fit Statistics:

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
smokerValue		NOBS	Sum of Frequencies	1068	.	270
smokerValue		MISC	Misclassification Rate	0	.	0
smokerValue		MAX	Maximum Absolute Error	0	.	0
smokerValue		SSE	Sum of Squared Errors	0	.	0
smokerValue		ASE	Average Squared Error	0	.	0
smokerValue		RASE	Root Average Squared Error	0	.	0
smokerValue		DIV	Divisor for ASE	2136	.	540
smokerValue		DFT	Total Degrees of Freedom	1068	.	.

Predictive Models: Naive Bayes

SAS Flowchart:

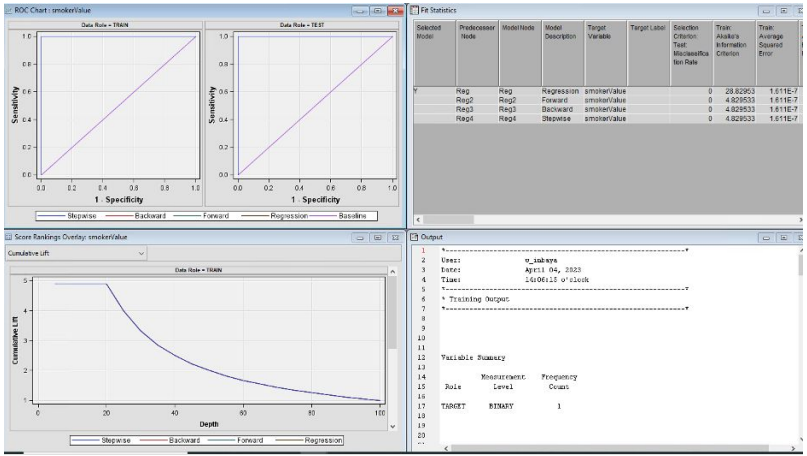


Fit Statistics:

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
smokerValue		ASE	Average Squared Error	0.010364		0.010283
smokerValue		DIV	Divisor for ASE	2136		540
smokerValue		MAX	Maximum Absolute Error	0.424132		0.364739
smokerValue		NOBS	Sum of Frequencies	1068		270
smokerValue		RASE	Root Average Squared Error	0.101806		0.101404
smokerValue		SSE	Sum of Squared Errors	22.13852		5.55272
smokerValue		DISF	Frequency of Classified Cases	1068		270
smokerValue		MISC	Misclassification Rate	0		0
smokerValue		WRONG	Number of Wrong Classifications	0		0

Predictive Models: Logistic Regression

Results:



Fit Statistics:

Fit Statistics

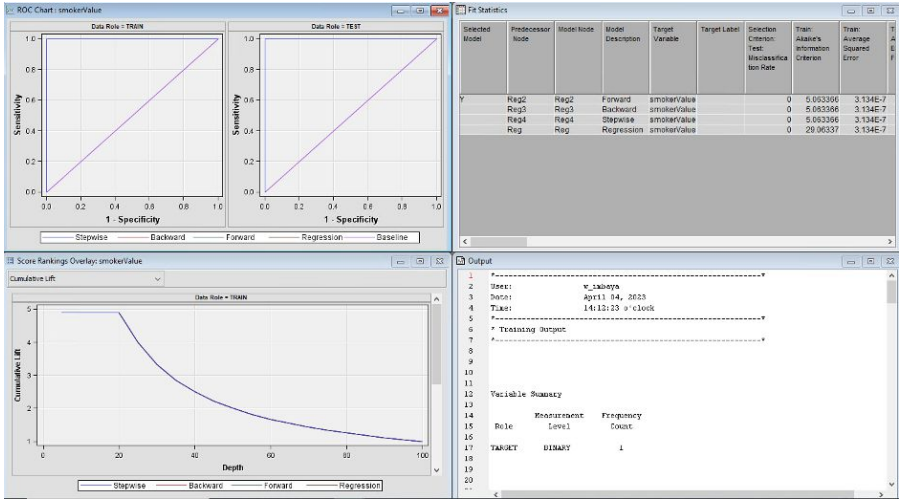
Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Test: Misclassification Rate	Train: Akaike's Information Criterion	Train: Average Squared Error
Y	Reg	Reg	Regression	smokerValue		0	28.82953	1.611E-7
	Reg2	Reg2	Forward	smokerValue		0	4.829533	1.611E-7
	Reg3	Reg3	Backward	smokerValue		0	4.829533	1.611E-7
	Reg4	Reg4	Stepwise	smokerValue		0	4.829533	1.611E-7

Fit Statistics

Test: Root Mean Square Error	Test: Sum of Square Errors	Test: Sum of Case Weights Times Freq	Test: Misclassification Rate	Test: Lower 95% Conf. Limit for TMISC	Test: Upper 95% Conf. Limit for TMISC	Train: Roc Index	Train: Gini Coefficient	Train: Kolmogorov-Smirnov Statistic
.0004008	8.673E-5	540	0	0	0.01357	1	1	1
.0004008	8.673E-5	540	0	0	0.01357	1	1	1
.0004008	8.673E-5	540	0	0	0.01357	1	1	1
.0004008	8.673E-5	540	0	0	0.01357	1	1	1

Predictive Models: Probit Regression

Results:



Fit Statistics:

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Test: Misclassification Rate	Train: Akaike's Information Criterion	Train: Average Squared Error	T A E F
Y	Reg2	Reg2	Forward	smokerValue		0	5.063366	3.134E-7	
	Reg3	Reg3	Backward	smokerValue		0	5.063366	3.134E-7	
	Reg4	Reg4	Stepwise	smokerValue		0	5.063366	3.134E-7	
	Reg	Reg	Regression	smokerValue		0	29.06337	3.134E-7	

Fit Statistics

Train: Root Mean Squared Error	Train: Schwarz's Bayesian Criterion	Train: Sum of Squared Errors	Train: Sum of Case Weights Times Freq	Train: Misclassification Rate	Test: Average Squared Error	Test: Average Error Function	Test: Divisor for TASE	Test: Error Function	T M A E
.0005604	15.01045	.0006695	2136	0	3.163E-7	.0004999	540	0.269959	
.0005604	15.01045	.0006695	2136	0	3.163E-7	.0004999	540	0.269959	
.0005604	15.01045	.0006695	2136	0	3.163E-7	.0004999	540	0.269959	
.0005636	98.69297	.0006695	2136	0	3.163E-7	.0004999	540	0.269959	

Results & Conclusion

Model Comparison:

Model	RMSE	Missclasification
Neural Network	0.080684	0.008427
KNN	0.204903	0.05555
Decision Tree	0	n/a
Naïve Bayes	0.1	0
Logistic Regression	0.0004	0
Probit Regression	0.0005	0

Model Comparison:

- **As Decision tree has the lowest RMSE and can be used for both categorical and numerical data, we conclude it is the best model**

Sources

- Tiwari, M.K., Singh, S.K., & Singh, S.K. (2015). Decision Tree Based Insurance Forecasting Using Data Mining Techniques. International Journal of Computer Science and Information Technologies, 6(1), 1-5.

- Predicting Insurance Losses with Decision Trees by David W. Aha