Exploring Reliable, Rule-Based AI and Automated Machine Learning

WELCOME TO OUR WEBINAR

Mikhail Golovnya
Senior Advisory Data Scientist

David Peralta
Area Marketing Manager

WEB-AUDIO:
Please make sure you have your computer audio system activated and your speakers turned up.

QUESTIONS:
You can enter your questions at any time in the questions section.

Minitab

About Our Speakers:

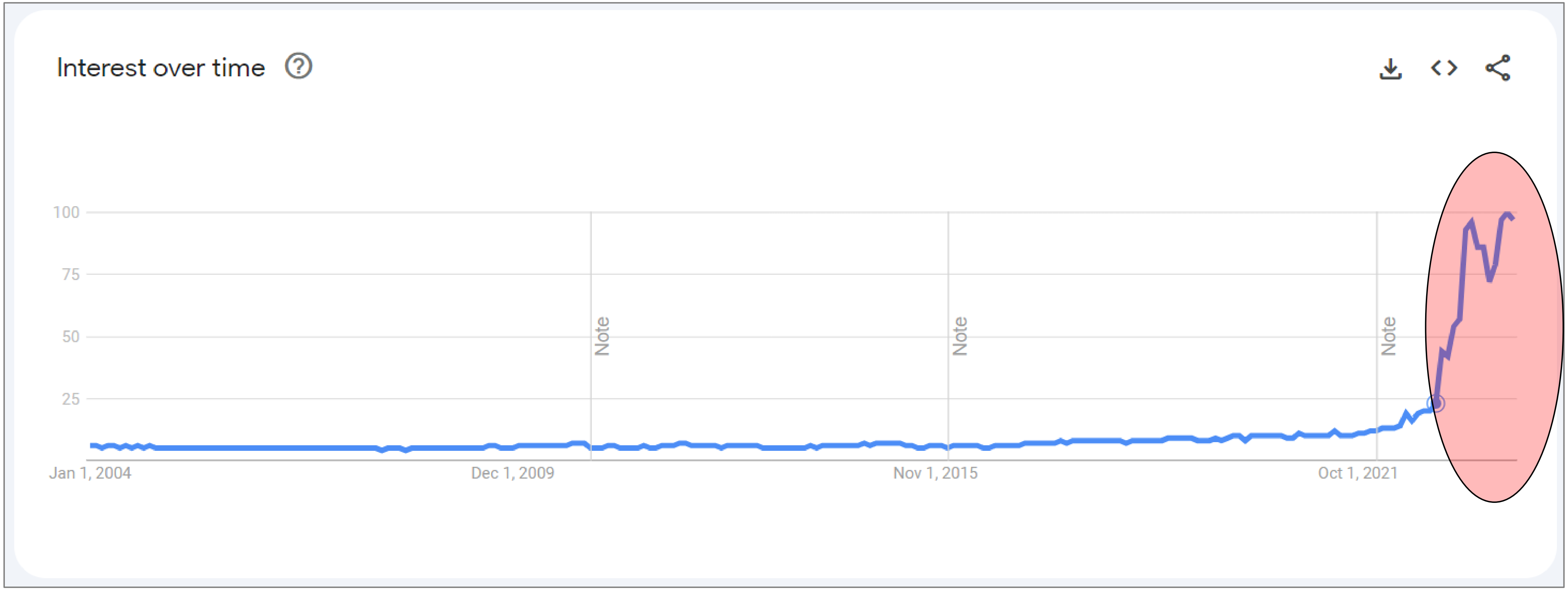
Mikhail Golovnya

Senior Advisory Data Scientist

Mikhail is a Senior Advisory Data Scientist at Minitab. He has been prototyping new machine learning algorithms and modeling automation for the past twenty years.

Mikhail has been a major contributor to Minitab's on-going search for technological improvements among the most important algorithms in Machine Learning.

Why is Everyone Suddenly Talking About AI?



Minitab

Chat GPT: Generative AI

How to prepare for evolving global AI legislation

Intel CEO Now Expects the Chip Shortage To Last Until 2024

- Generative AI has grabbed the world’s attention.
- Generative AI has great potential but also challenging realities:
  - Potential: Improved labor productivity
  - Challenging realities: hallucinating chatbots, hard-to-obtain GPU chips, potentially huge liabilities around copyright, concerns about bias and accuracy, impending global legislation.
- Generative AI has crowded out other types of AI techniques, some that have been with us for many years.
- This represents a major opportunity for us to highlight our recent investments.

Different Types of AI

- **Reactive Machines:**
  - These are basic rule-based systems that operate based on predefined rules.
- **Expert Systems:**
  - These are computer systems that mimic the decision-making ability of a human expert in a specific domain.
- **Machine Learning (ML) Systems:**
  - ML is a subset of AI that focuses on developing algorithms and models that enable computers to learn from data.
  - Types of ML systems include supervised learning, unsupervised learning, and reinforcement learning.
- **Neural Networks:**
  - Inspired by the human brain, neural networks are a key component of many AI systems.
- **Narrow AI (Weak AI):**
  - These AI systems are designed and trained for a specific task or a narrow set of tasks.
  - Examples include virtual personal assistants, image recognition software, and language translation services.
- **Limited Memory:**
  - These AI systems can learn from historical data to make better decisions.
  - Self-driving cars often use limited memory AI to navigate based on past experiences.
- **Self-aware AI:**
  - This refers to hypothetical AI systems with self-awareness and consciousness.
- **Theory of Mind:**
  - This is a more advanced form of AI that can understand human emotions, beliefs, intentions, and thoughts.
- **General AI (Strong AI):**
  - General AI systems can understand, learn, and apply knowledge across diverse domains.
  - They can perform any intellectual task that a human being can do.
- **Superintelligent AI:**
  - This is a theoretical AI that surpasses human intelligence in every aspect.
- **Robotics AI:**
  - AI is often integrated into robots to enable them to perceive, learn, and interact with the environment.

Minitab

Survey 1

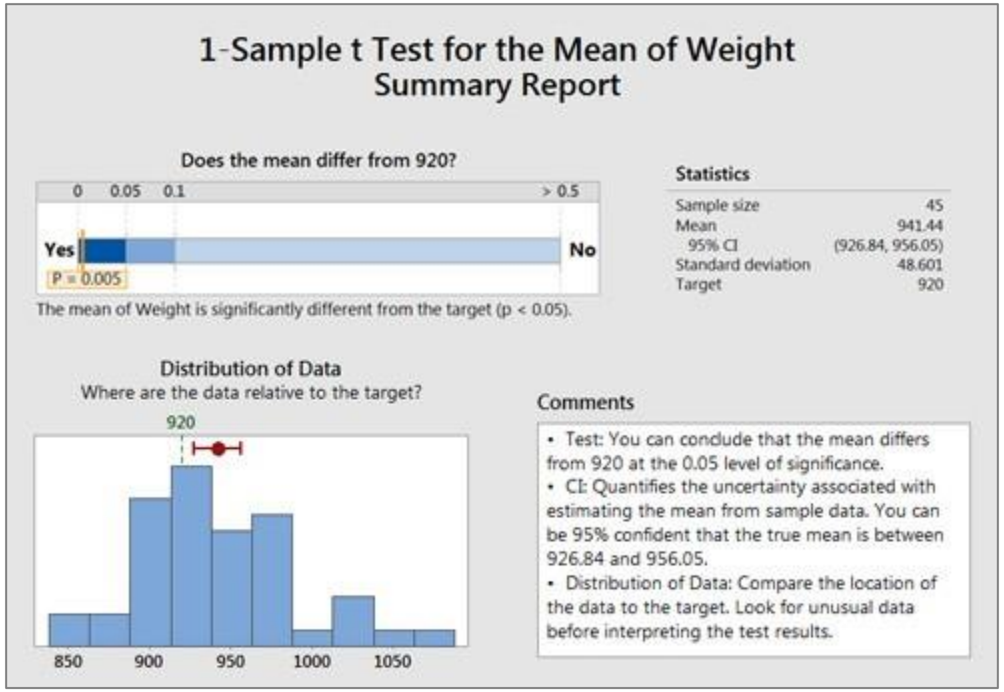
- What types of AI are you most likely to use in your work?
  - Rule-based/Expert Systems (i.e., Q/A Guides and Wizards)
  - Machine Learning Systems (i.e., Predictive Analytics automation)
  - Image/Language/Assistant Systems (i.e., Chat GPT, Watson)
  - Deep Learning Memory Networks (i.e., self-driven cars)
  - Self-Aware/Super-Intelligent Systems (i.e., Skynet)

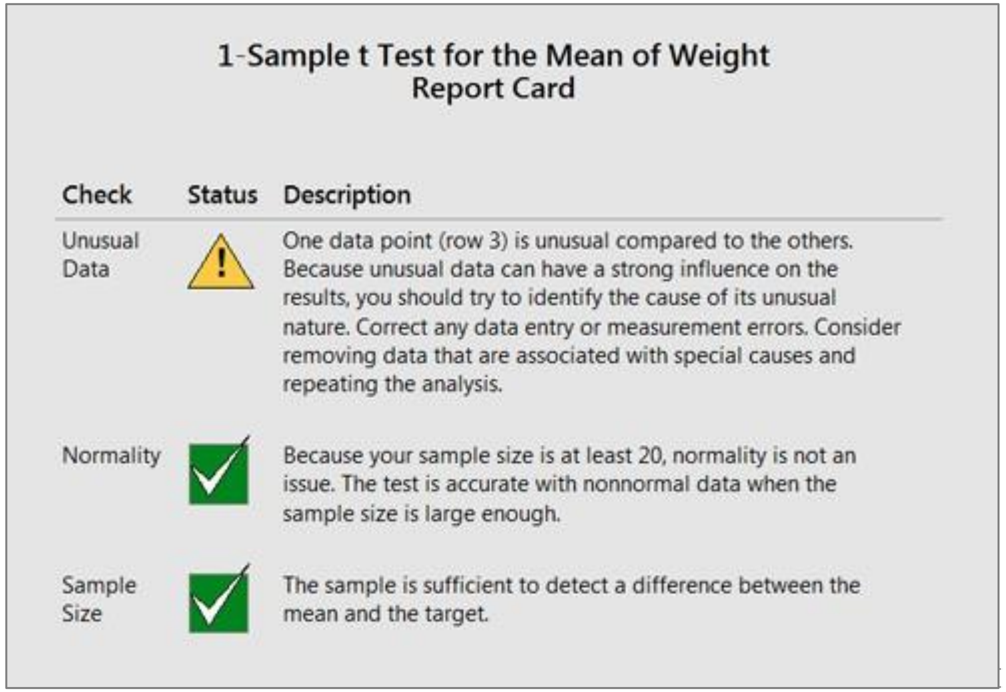
Types of AI that MSS Supports or Will Support

- **Reactive Machines:**
  - These are basic rule-based systems that operate based on predefined rules.
- **Expert Systems:**
  - These are computer systems that mimic the decision-making ability of a human expert in a specific domain.
- **Machine Learning (ML) Systems:**
  - ML is a subset of AI that focuses on developing algorithms and models that enable computers to learn from data.
  - Types of ML systems include supervised learning, unsupervised learning, and reinforcement learning.
- **Neural Networks:**
  - Inspired by the human brain, neural networks are a key component of many AI systems.
- **Narrow AI (Weak AI):**
  - These AI systems are designed and trained for a specific task or a narrow set of tasks.
  - Examples include virtual personal assistants, image recognition software, and language translation services.
- **Limited Memory:**
  - These AI systems can learn from historical data to make better decisions.
  - Self-driving cars often use limited memory AI to navigate based on past experiences.
- **Self-aware AI:**
  - This refers to hypothetical AI systems with self-awareness and consciousness.
- **Theory of Mind:**
  - This is a more advanced form of AI that can understand human emotions, beliefs, intentions, and thoughts.
- **General AI (Strong AI):**
  - General AI systems can understand, learn, and apply knowledge across diverse domains.
  - They can perform any intellectual task that a human being can do.
- **Superintelligent AI:**
  - This is a theoretical AI that surpasses human intelligence in every aspect.
- **Robotics AI:**
  - AI is often integrated into robots to enable them to perceive, learn, and interact with the environment.

MSS: Expert Guidance

- Can anyone think of an example of an 'expert machine' in Minitab that mimics the decision-making ability of a human expert in a specific domain?
- ***Assistant Menu***





Two Common Business Problems

- Anyone investing into data collection and management eventually faces the following common needs:

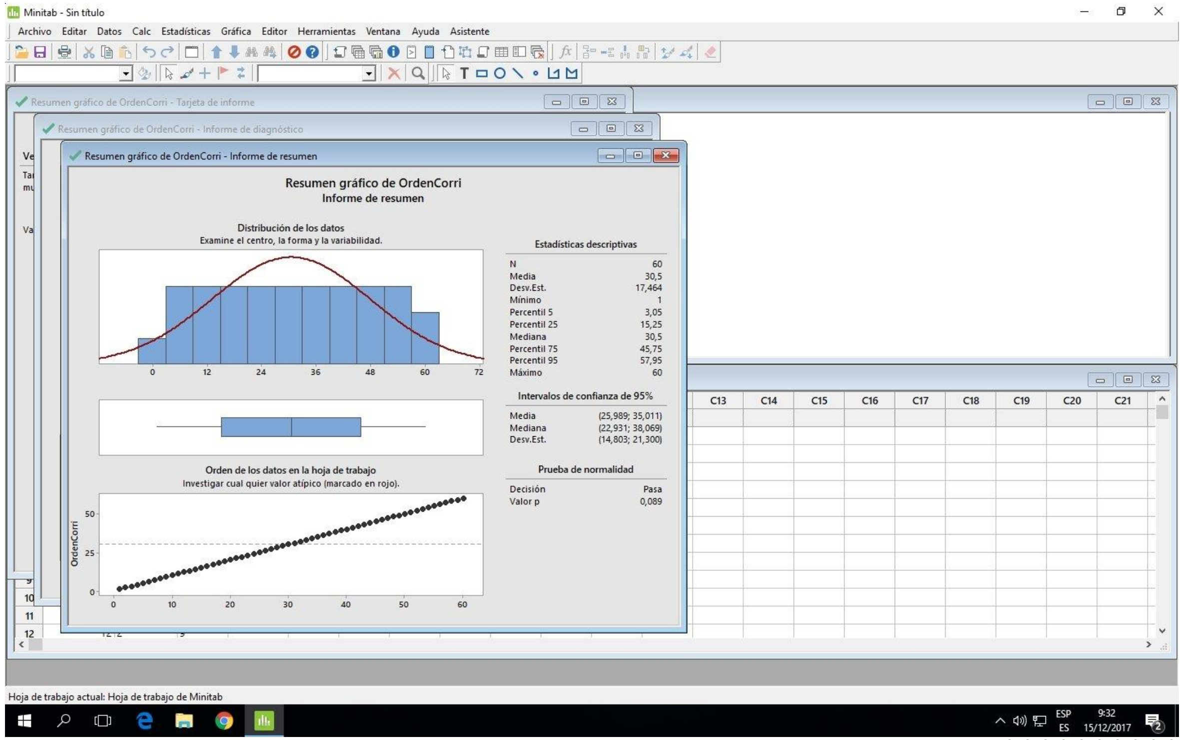
- **Problem 1**: Find the most accurate predictive model subject to some natural constraints

- **Problem 2**: Facing a very large number of attributes/features/variables in your data, identify which ones are most predictive of the outcome

- **AutoML AI** addresses the above problems (and more)!

AutoML in Minitab

- **Automatic discovery of the best predictive model** (regression or classification) among the following families of models (more models to be added in the future):
  - Classical multiple linear or logistic regression
  - Classification and Regression Trees (CART)
  - Random Forests (RF)
  - Stochastic Gradient Boosting (TN)
- **Automatic discovery of the key predictors** for the stochastic gradient boosting (TN) models
- We will illustrate the first task (best model discovery) using a delinquency prediction dataset and the second task (discover key predictors) using Word Bank / United Nations dataset



Survey 2

- How many Predictive Analytics algorithms do you routinely use?
  - None
  - One or Two
  - A handful of favorites
  - As many as I can get my hands on

Delinquency Prediction Dataset

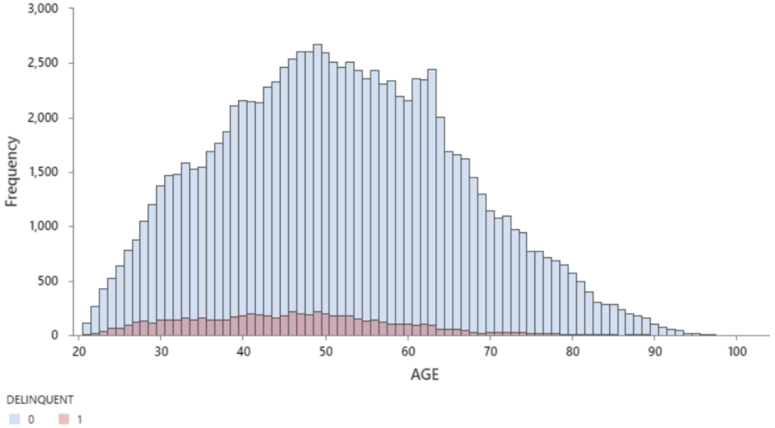
- **Delinquency Prediction** in Banking (Kaggle)
  - Predict who will experience at least 90-days past due or other delinquency within the next 2 years (about 6% of the accounts)
  - 108,376 instances and 6 predictors
  - Binary response variable
  - The original raw data available at https://www.kaggle.com/c/GiveMeSomeCredit
  - In this presentation we use a processed subset of the original data

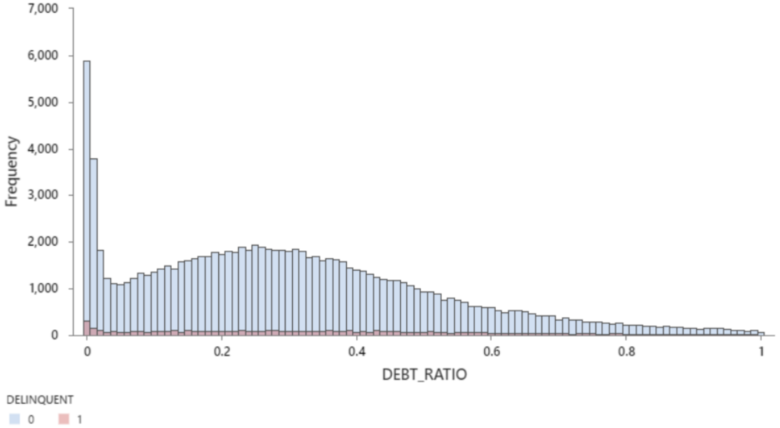
Minitab

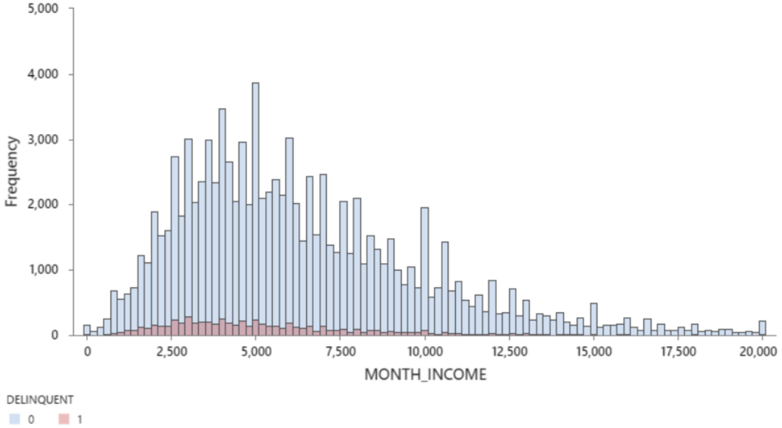
Variables of Interest

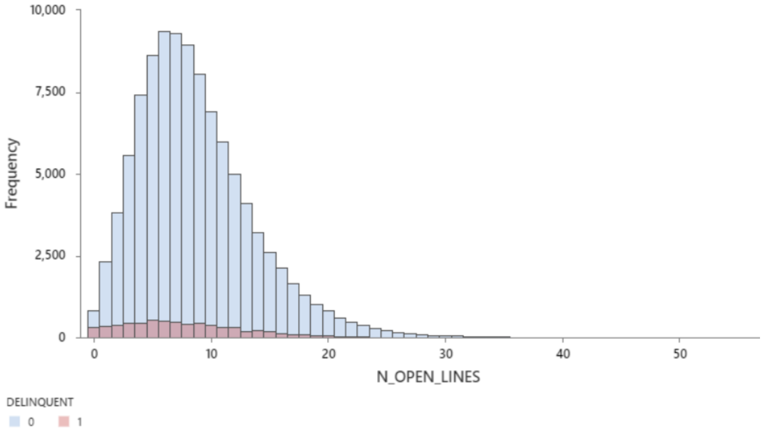
VARIABLE	DESCRIPTION
DELINQUENT	Person experienced 90 days past due delinquency or worse
AGE	Age of borrower in years
DEBT_RATIO	Monthly debt payments, alimony, living costs divided by monthly gross income
MONTH_INCOME	Monthly income
N_OPEN_LINES	Number of open loans (mortgages, car loans, credit cards, etc.)
N_MORTGAGES	Number of mortgage and real estate loans
N_DEPENDENTS	Number of dependents in family excluding yourself

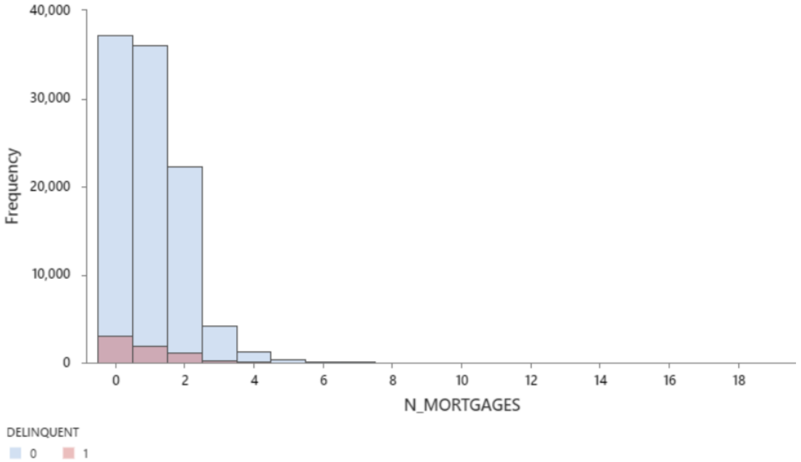
Descriptive Stage -1

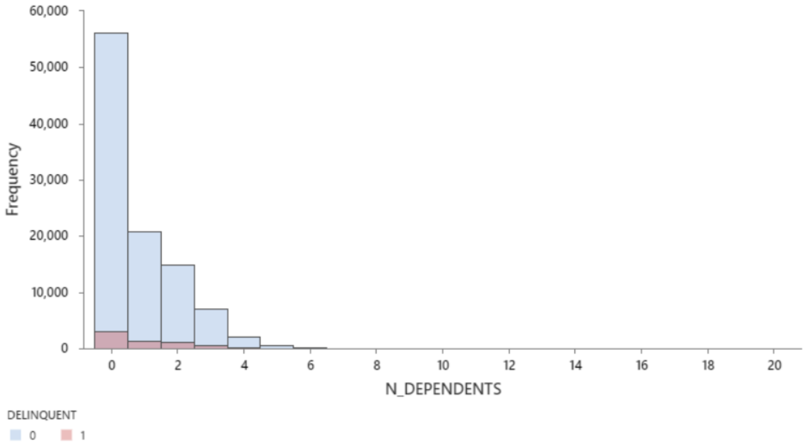




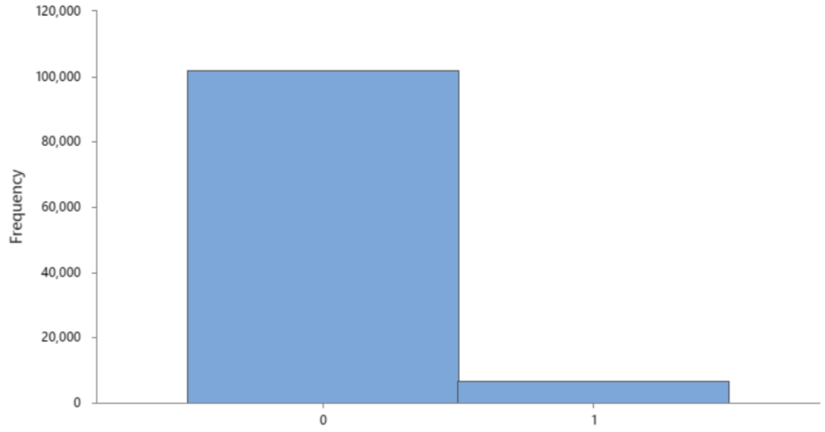


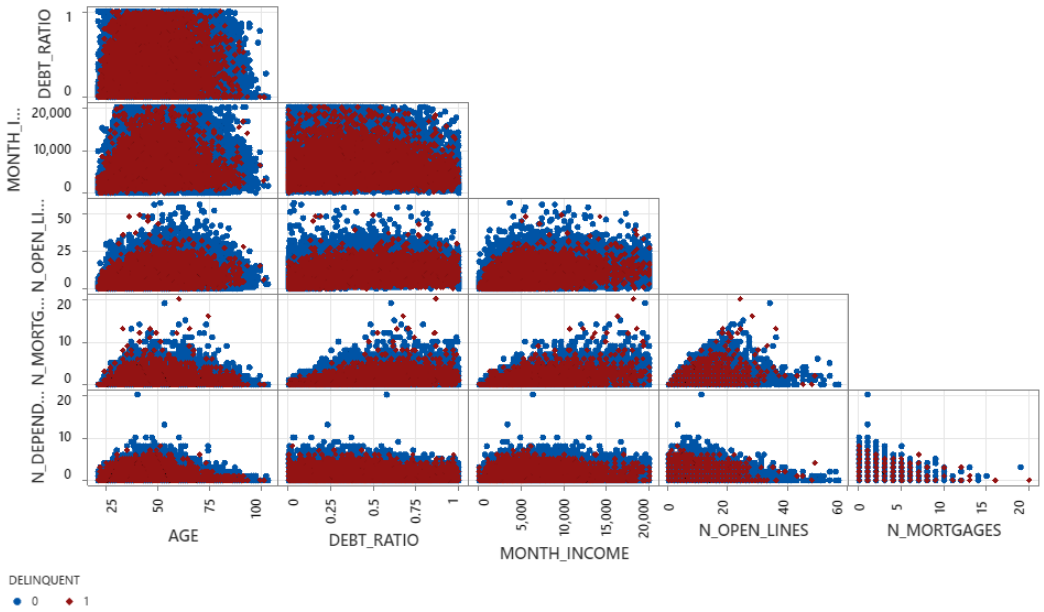


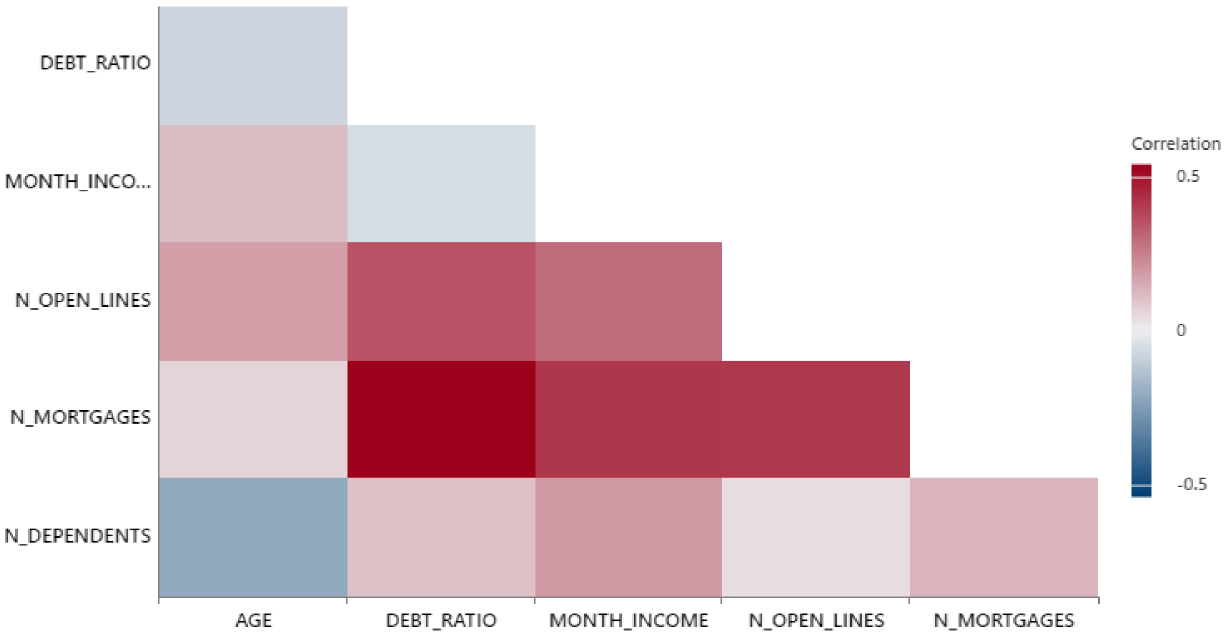




Descriptive Stage -2







Baseline Model (Logistic Regression)

Method

Link function	Logit
Rows used	108376
Test set fraction	50.0%

Response Information

Variable	Value	Training Count	Test Count	
DELINQUENT	1	3281	3283	(Event)
	0	50907	50905	
	Total	54188	54188	

Model Summary

Deviance R-Sq	Deviance R-Sq(adj)	AIC	AICc	BIC	Area Under ROC Curve	Test Deviance R-Sq	Test Area Under ROC Curve
3.48%	3.45%	23914.57	23914.57	23976.87	0.6534	3.60%	0.6549

Regression Equation

P(1) = exp(Y')/(1 + exp(Y'))

Y' = -1.6474 - 0.02268 AGE + 1.264 DEBT_RATIO - 0.000045 MONTH_INCOME - 0.01682 N_OPEN_LINES - 0.1169 N_MORTGAGES + 0.1238 N_DEPENDENTS

Coefficients

Term	Coef	SE Coef	Z-Value	P-Value	VIF
Constant	-1.6474	0.0776	-21.22	0.000	
AGE	-0.02268	0.00142	-16.01	0.000	1.11
DEBT_RATIO	1.264	0.100	12.58	0.000	1.79
MONTH_INCOME	-0.000045	0.000007	-6.22	0.000	1.62
N_OPEN_LINES	-0.01682	0.00451	-3.73	0.000	1.40
N_MORTGAGES	-0.1169	0.0269	-4.35	0.000	2.00
N_DEPENDENTS	0.1238	0.0152	8.16	0.000	1.09

Algorithm Discovery

Model Selection

Best Model within Type	Average -Loglikelihood	Area Under ROC Curve	Misclassification Rate
TreeNet®*	0.2140	0.6997	0.0608
Random Forests®	0.2871	0.6717	0.0606
CART®	0.2357	0.6700	0.3705
Logistic Regression	0.2212	0.6496	0.0606

** Best model across all model types with maximum area under ROC curve. Output for the best model follows.*

Hyperparameters for Best TreeNet® Model

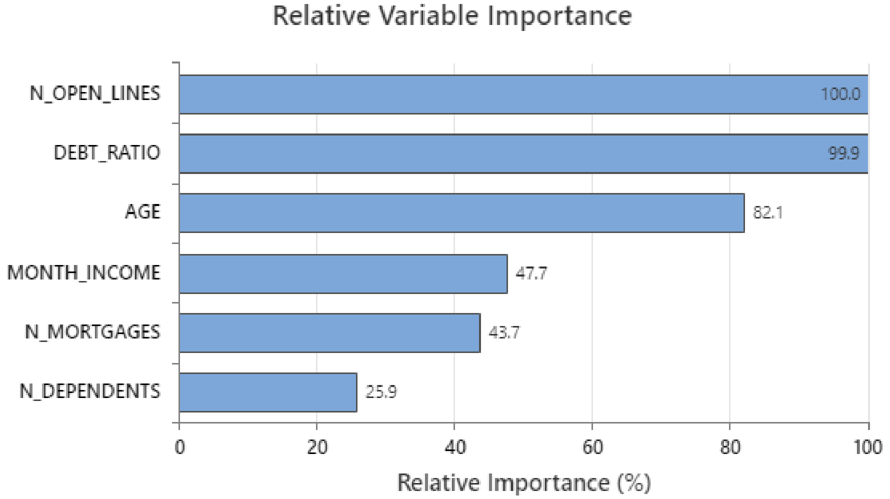
Number of trees grown	300
Optimal number of trees	161
Learning rate	0.1
Subsample fraction	0.7
Maximum terminal nodes per tree	6
Number of predictors selected for node splitting	Total number of predictors = 6

Problem 1 solved!
TN gave 5-point
boost in accuracy!

Interaction Discovery

Method

Criterion for selecting optimal number of trees	Maximum area under ROC curve
Model validation	50/50% training/test sets
Learning rate	0.03
Subsample selection method	Completely random
Subsample fraction	0.5
Maximum terminal nodes per tree	6
Minimum terminal node size	10
Number of predictors selected for node splitting	Total number of predictors = 6
Rows used	108376



Model Summary

Total predictors	6
Important predictors	6
Number of trees grown	300
Optimal number of trees	298

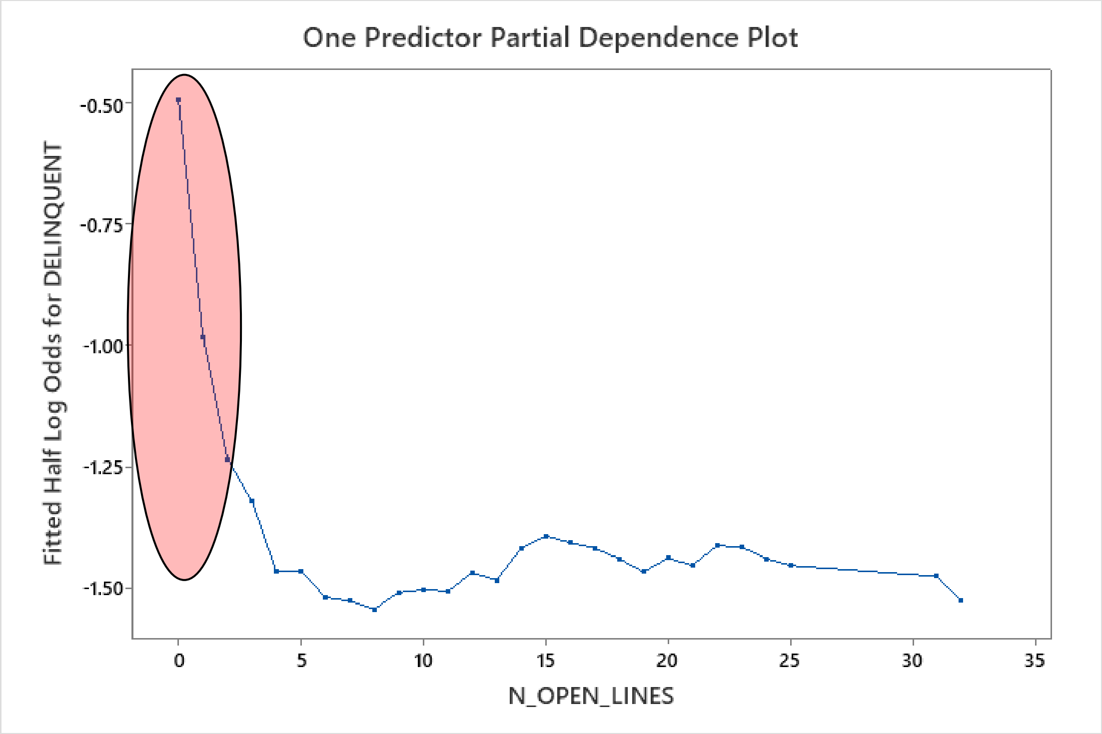
Statistics	Training	Test
Average -loglikelihood	0.2080	0.2137
Area under ROC curve	0.7285	0.7000
95% CI	(0.7198, 0.7372)	(0.6911, 0.7089)
Lift	3.1779	2.6508
Misclassification rate	0.0603	0.0608

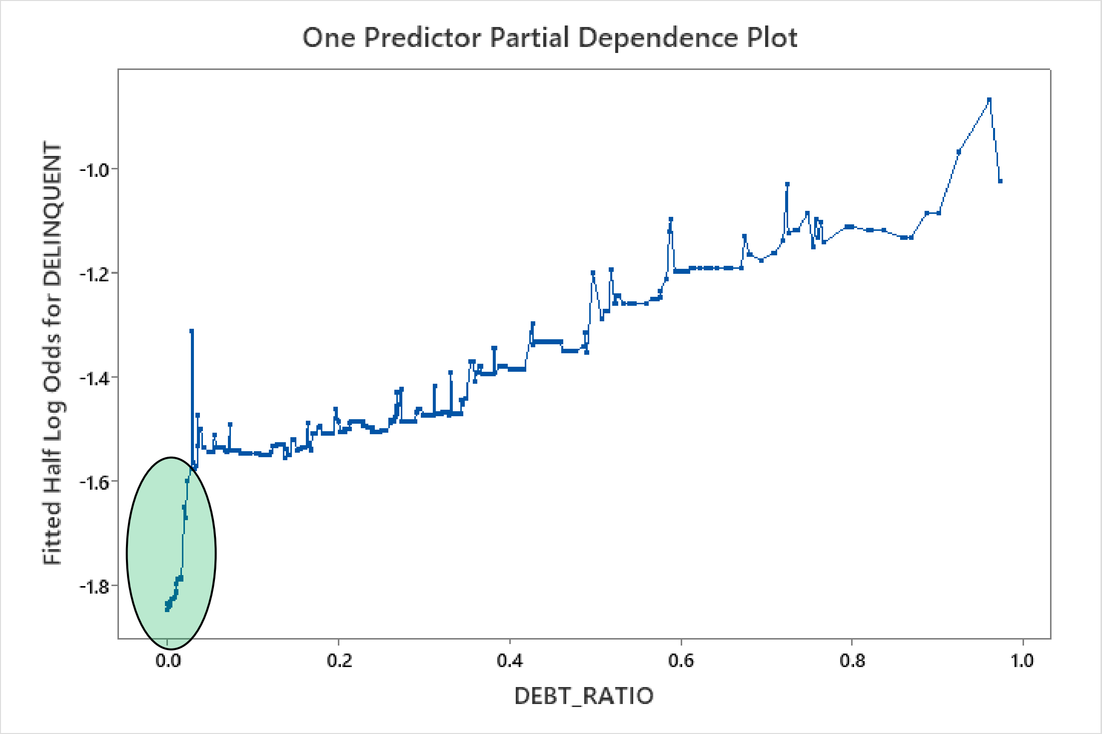
Model Summary

Total predictors	6
Important predictors	6
Number of trees grown	300
Optimal number of trees	298

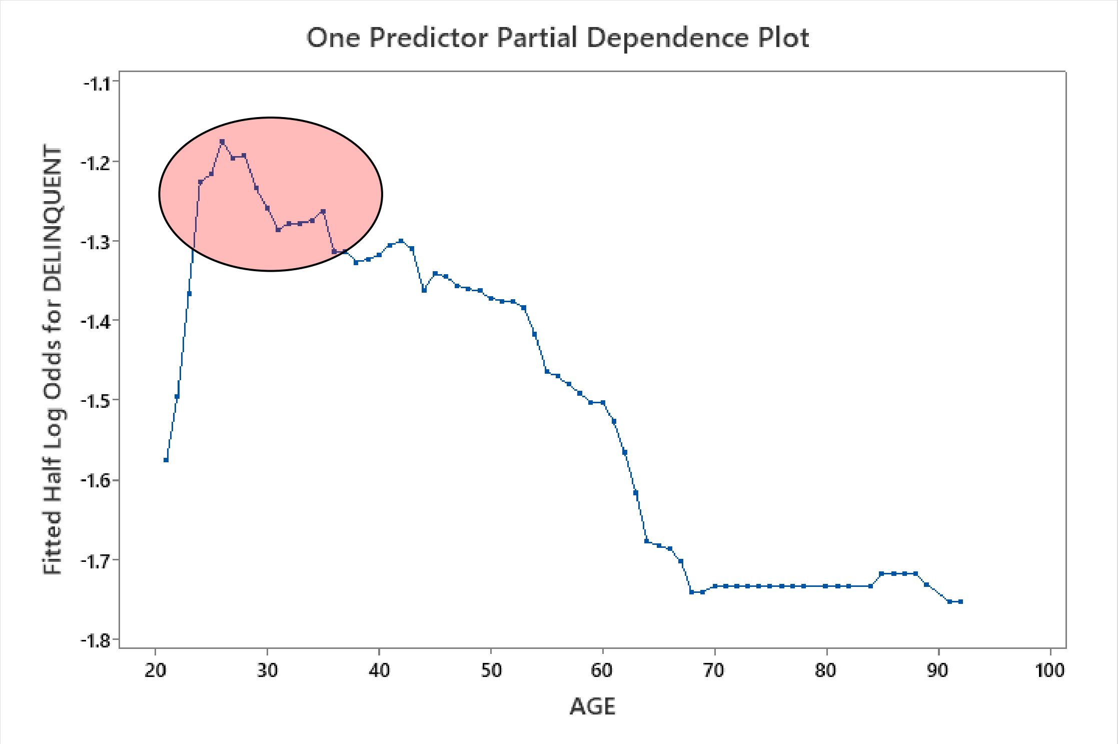
Statistics	Training	Test
Average -loglikelihood	0.2120	0.2150
Area under ROC curve	0.7159	0.6921
95% CI	(0.7072, 0.7247)	(0.6831, 0.7011)
Lift	2.7937	2.5411
Misclassification rate	0.0605	0.0606

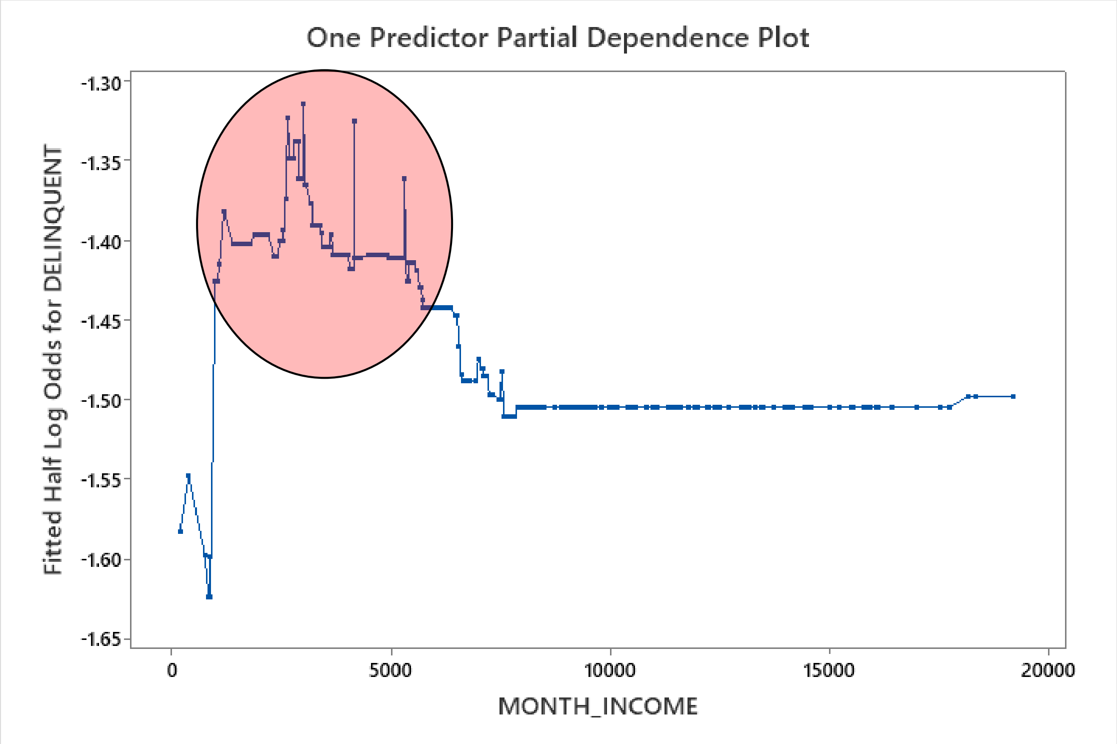
Model Insights - 1



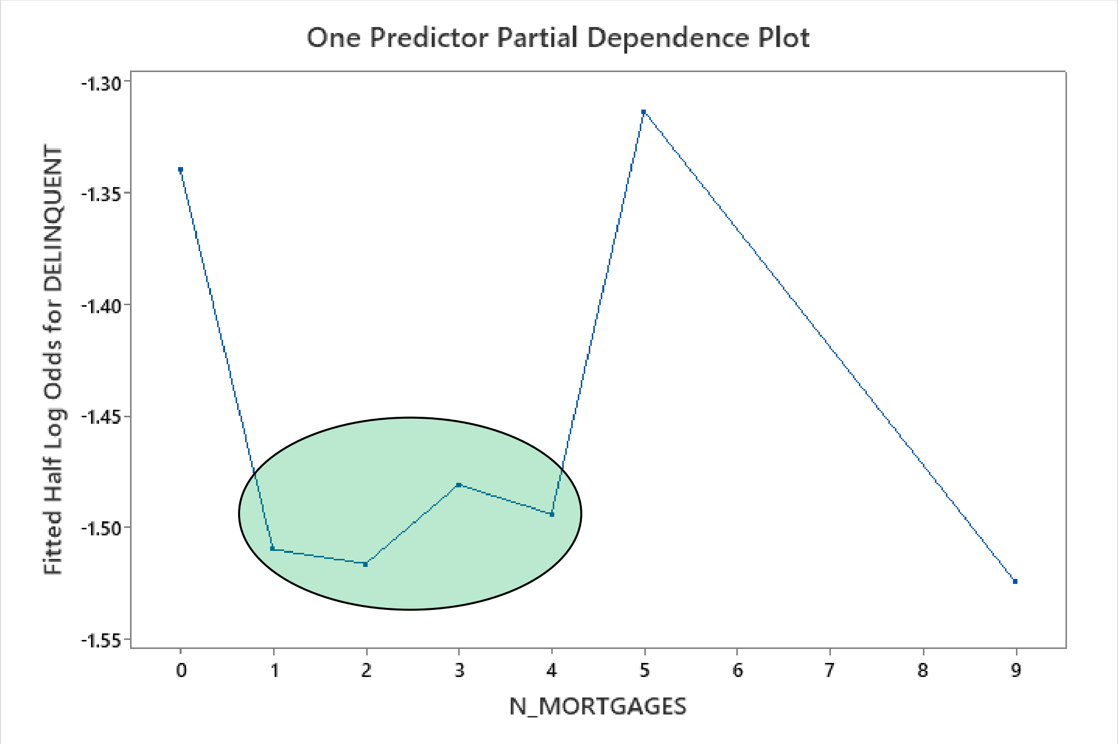


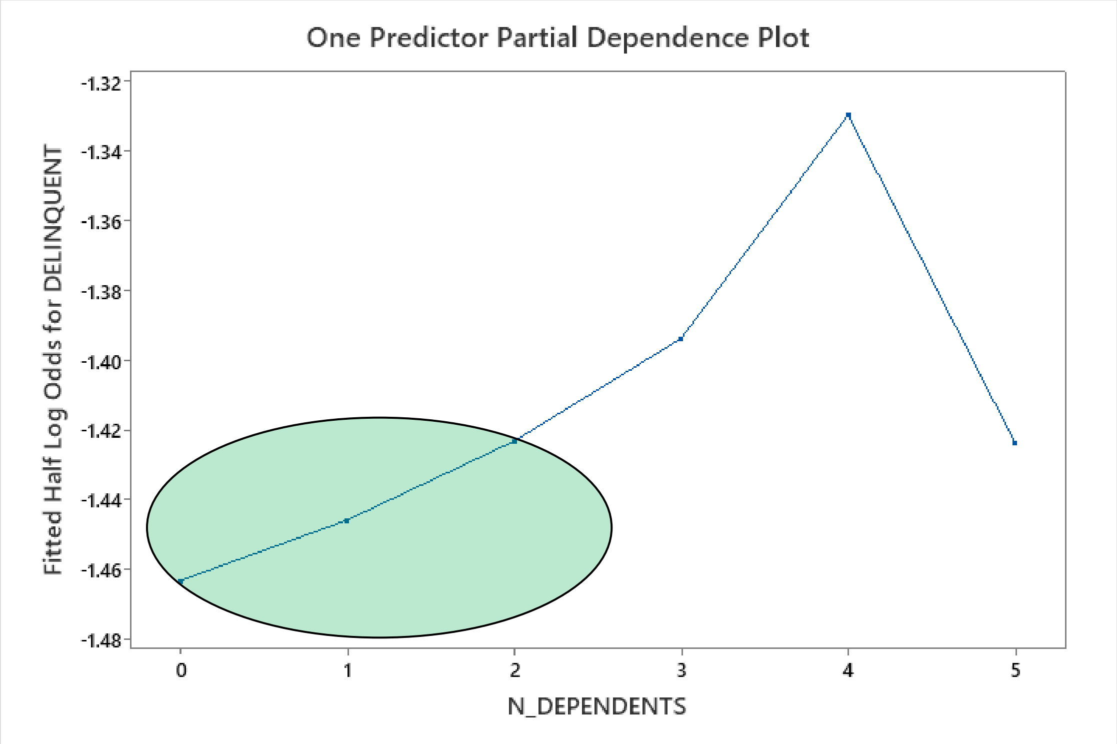
Model Insights - 2





Model Insights - 3





Survey 3

- How many variables do your datasets usually have?
  - Single digits
  - Tens
  - Hundreds
  - Thousands

World Hunger Dataset

- Understanding the key drivers associated with world hunger is crucial for setting up international policies and efforts to provide necessary relief
- Both World Bank and the United Nations provide useful public datasets to researchers
- The datasets include various annual socio/economic indicators for different world countries/regions covering the period from 2000 to 2018 (pre-covid)
- Our focus of interest is which socio/economic indicators are predictive of the prevalence of undernourishment (hunger) within a region
- After data preparation and merging, our final modeling dataset has **98 variables** and 4780 observations

Minitab

Descriptive Statistics - 1

Statistics

Variable	N	N*	Mean
TEST	4780	0	0.10000
CLEAN_FUEL	4717	63	61.902
ELECTRICITY	4756	24	78.123
NR_DEPLETION	4391	389	4.586
FOREST_DEPLETION	4431	349	1.2658
AG_LAND	4488	292	38.499
NR_VALUE_ADDED	4501	279	11.650
H2O_WITHDRAW	717	4063	126.5
INCOME_GROWTH	106	4674	2.530
PREV_DEATH	916	3864	24.577
CHILD_WORK	267	4513	21.09
CO2_EMISSIONS	4503	277	4.3065
CORRUPTION_CONTROL	3606	1174	-0.0722
DISASTERS	165	4615	1.191
EASE_BUSINESS	185	4595	96.69
ELECTRICITY_COAL	2895	1885	19.454
ENERGY_IMPORTS	2781	1999	-34.45
ENERGY_INTENSITY	3738	1042	5.5245
ENERGY_USE	2848	1932	2291.9
FERT_RATE	4617	163	2.9782
FOOD_PRODUCTION	4445	335	78.054
FOREST_AREA	4755	25	32.770
FOSSIL_CONSUMPTION	2849	1931	65.453
GDP_GROWTH	4679	101	3.841
GHG_NET_EMISSIONS	448	4332	-43.83
GINI_INDEX	1393	3387	37.386
GOV_EFFECTIVENESS	3601	1179	-0.0709
EDUCATE_EXPENSE	3069	1711	14.761
HOSPITAL_BEDS	2679	2101	3.5652
INCOME_SHARE	1390	3390	6.6955
INTERNET_USE	4590	190	30.749
LABOR_PARTICIPATION	4480	300	67.271
LIFE_EXPECTANCY	4605	175	69.388
LITERACY_RATE	1503	3277	80.038
MAMMALS_THREAT	239	4541	162.2
METHANE_EMISSIONS	4503	277	1.7154
MORTALITY_RATE	4760	20	41.361
NET_MIGRATION	916	3864	-443204
NOX_EMISSIONS	4492	288	0.5369
PATENTS	2292	2488	49478
SAFE_H2O	2691	2089	67.877
SAFE_SANITATION	3090	1690	51.106
AIR_POLLUTION	2320	2460	30.688
POLITICAL_STABILITY	3632	1148	-0.0591
POPULATION_AGED	4572	208	7.5819
POPULATION_DENSITY	4769	11	229.2
POVERTY_RATIO	856	3924	24.210
POPULATION_OVWT	3978	802	43.178
POPULATION_HUNGER	3857	923	11.446
WOMEN_PARLIAMENT	4565	215	17.679
F2M_LABOR	4480	300	69.676
REGULATORY_QUALITY	3601	1179	-0.0784
ELECTRICITY_RENEW	3828	952	30.210
ENERGY_RENEW	4365	415	32.424
RD_EXPENDITURE	2157	2623	1.0492
RULE_OF_LAW	3645	1135	-0.0744
SCHOOL_ENROLL_P	4000	780	102.67
SCHOOL_ENROLL_PS	3463	1317	0.96971
SCIENCE_ARTICLES	4541	239	60980
LEGAL_RIGHTS	1617	3163	5.1047
PROTECTED_AREAS	899	3881	12.648
UNEMPLOYMENT	4480	300	7.8170
UNMET_CONTRACEPTION	500	4280	19.429
ACCOUNTABILITY	3657	1123	-0.0478

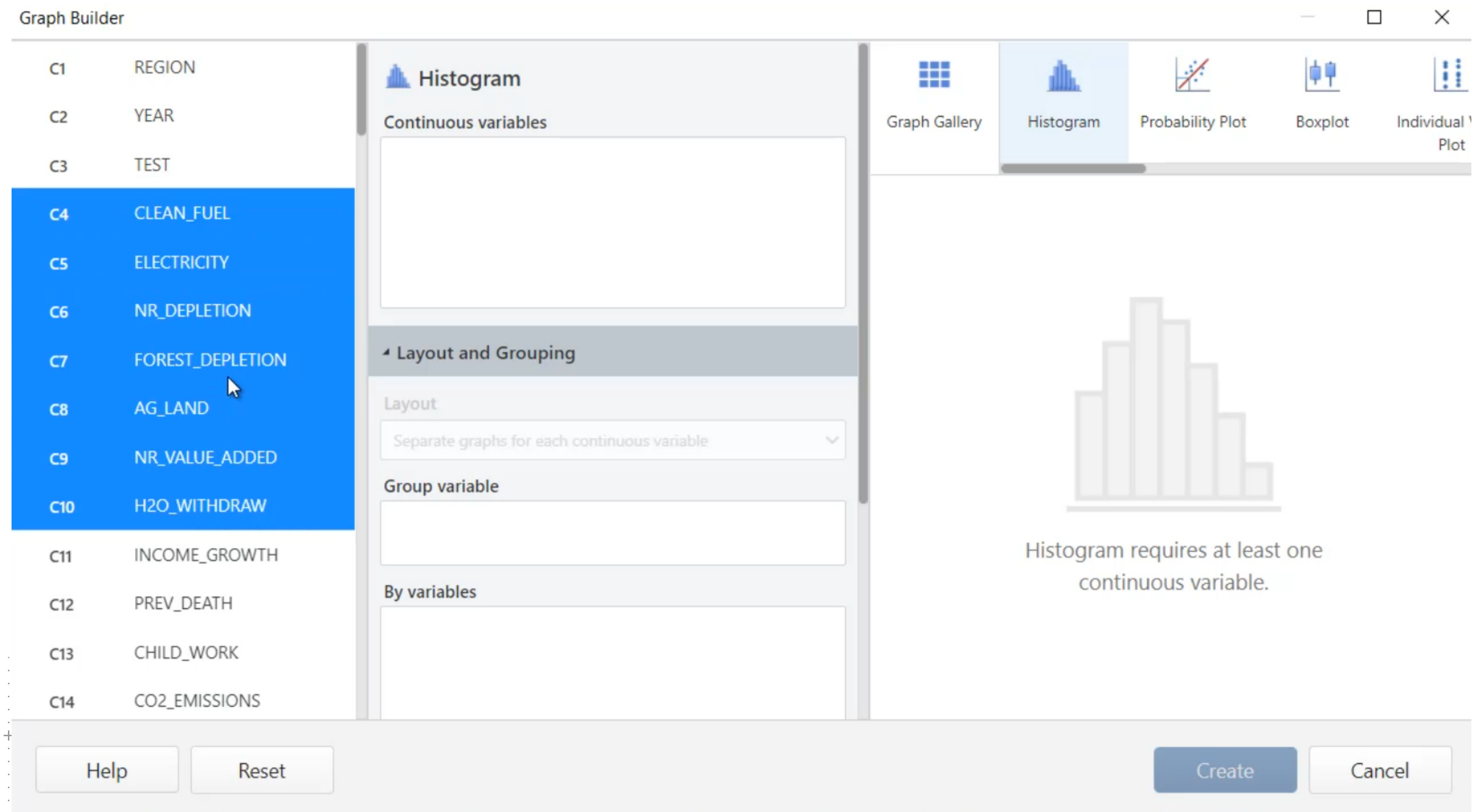
Descriptive Statistics - 2

- We are challenged by the sheer number of available variables
- The following descriptors may significantly impact follow-up analyses
  - Categorical values
  - Missing values
  - Extreme values
  - Dependencies
- **Graph Builder** – MSS Rule-based AI machinery – is designed to address this challenge quickly and efficiently!

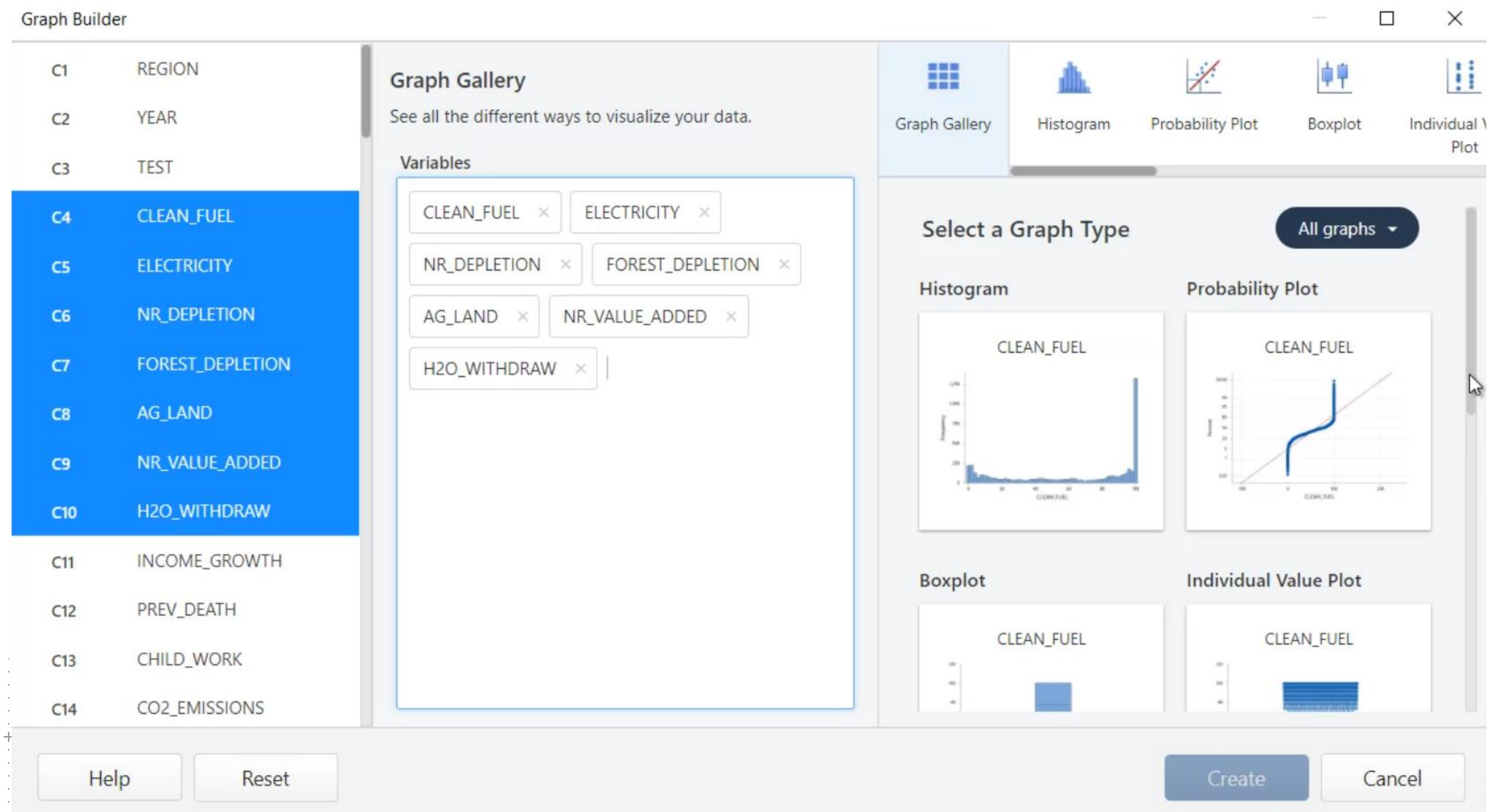
Statistics

Variable	N	N*	Mean
211_PREV_UNDERNOURISH	2643	2137	10.743
211_POP_UNDERNOURISH	1843	2937	15.39
212_PREV_MFOOD_INSEC	472	4308	27.03
212_POP_MFOOD_INSEC	472	4308	37948
212_PREV_SFOOD_INSEC	475	4305	9.788
212_POP_SFOOD_INSEC	475	4305	13437
221_PROP_STUNTED	597	4183	26.439
221_POP_STUNTED	2740	2040	2857
222_PROP_WASTED	552	4228	6.390
222_POP_WASTED	552	4228	452.6
222_PROP_OVERWEIGHT	575	4205	6.416
222_POP_OVERWEIGHT	2740	2040	510.7
223_PROP_ANEM	3340	1440	27.722
223_PROP_ANEM_NP	3340	1440	27.418
223_PROP_ANEM_PR	3340	1440	32.501
231_PROD_LSFP	105	4675	572.2
231_PROD_SSFP	110	4670	52.95
232_INCOME_LSFP	73	4707	6498
232_INCOME_SSFP	78	4702	1276
251_BREEDS_GENETIC	290	4490	1.614
251_PLANT_GENETIC	963	3817	76199
252_PROP_BREEDS_RISK	1808	2972	47.195
2A1_AG_VALUE_ADDED	3197	1583	12.067
2A1_AG_ORIENTATION_INDEX	2305	2475	0.5676
2A1_AG_GOV_EXPENDITURE	2305	2475	2.9547
2A2_TOT_FLOWS	2423	2357	77.60
2B1_AG_EXPORT_SUBSIDIES	455	4325	119.2
2C1_FOOD_PRICE_ANOMALIES	1528	3252	-0.1515

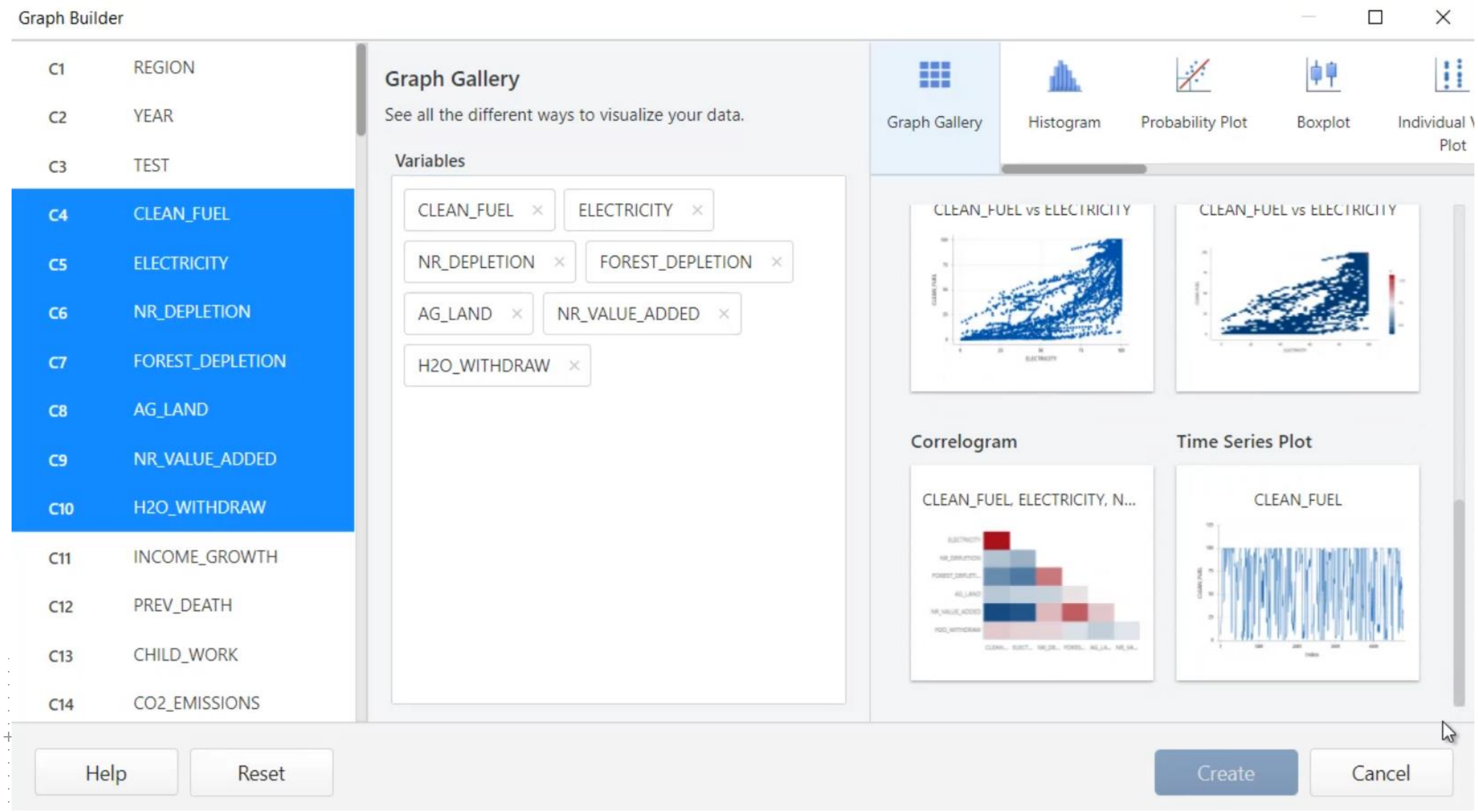
MSS Graph Builder – Rule-Based Automation



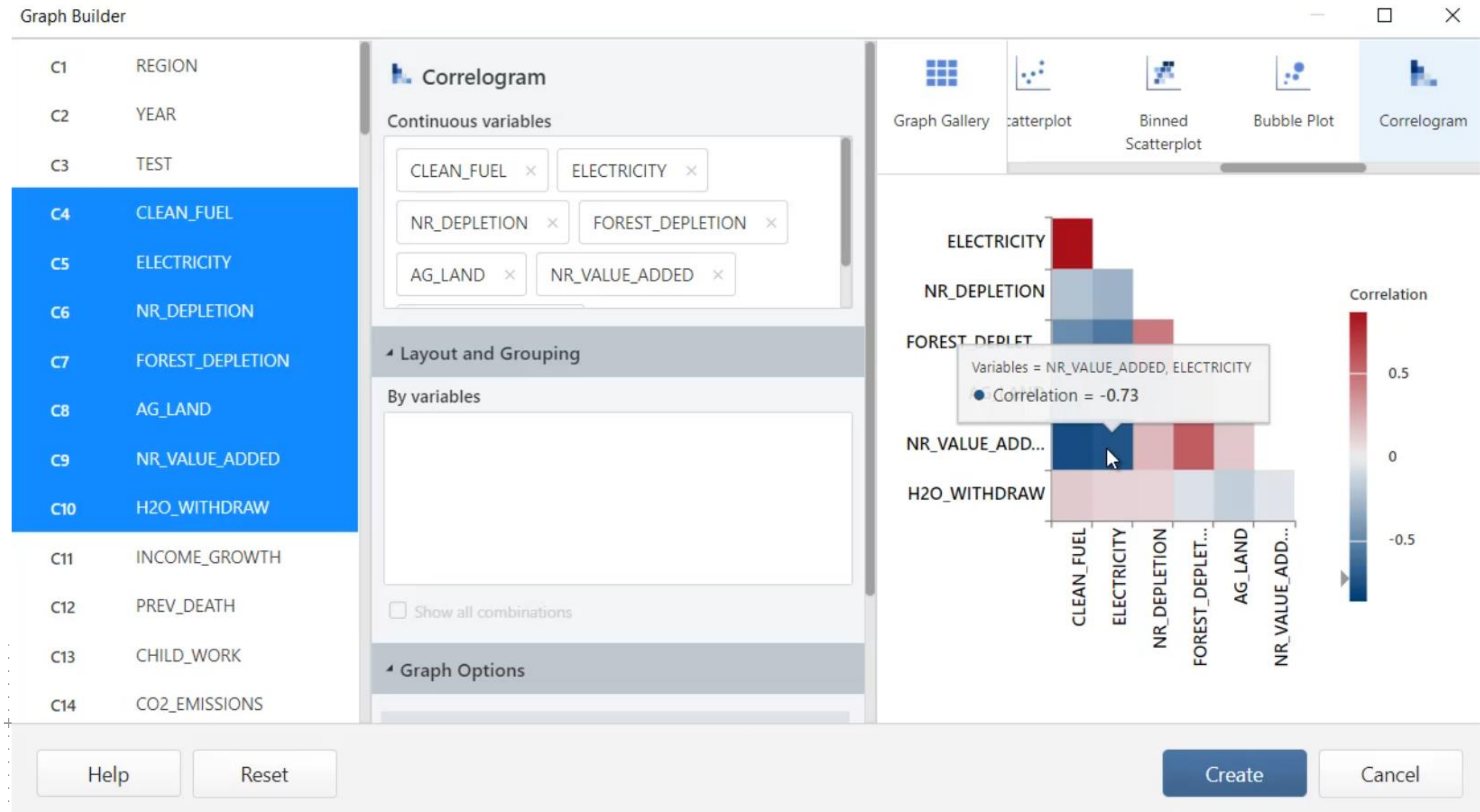
MSS Graph Builder – Rule-Based Automation



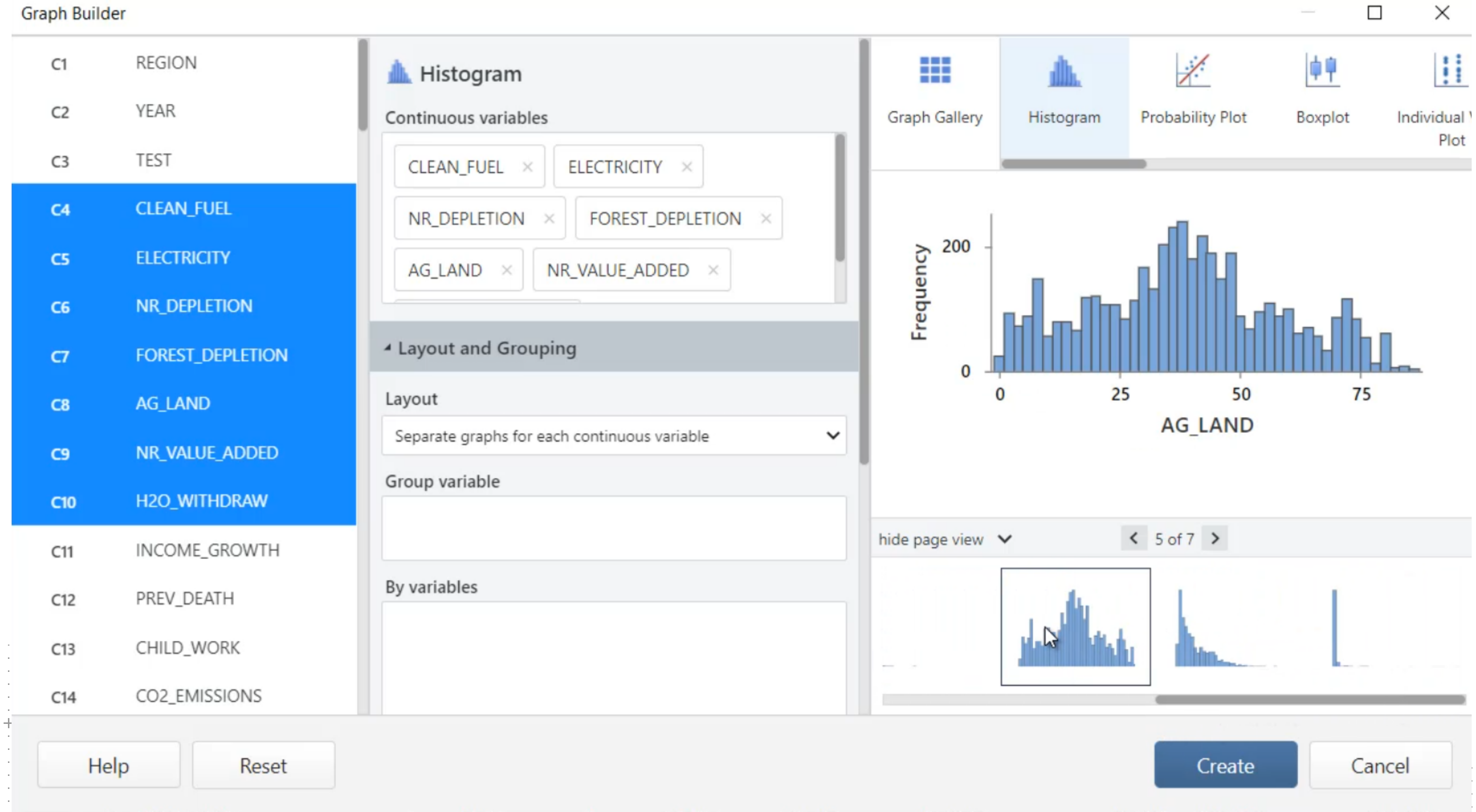
MSS Graph Builder – Rule-Based Automation



MSS Graph Builder – Rule-Based Automation



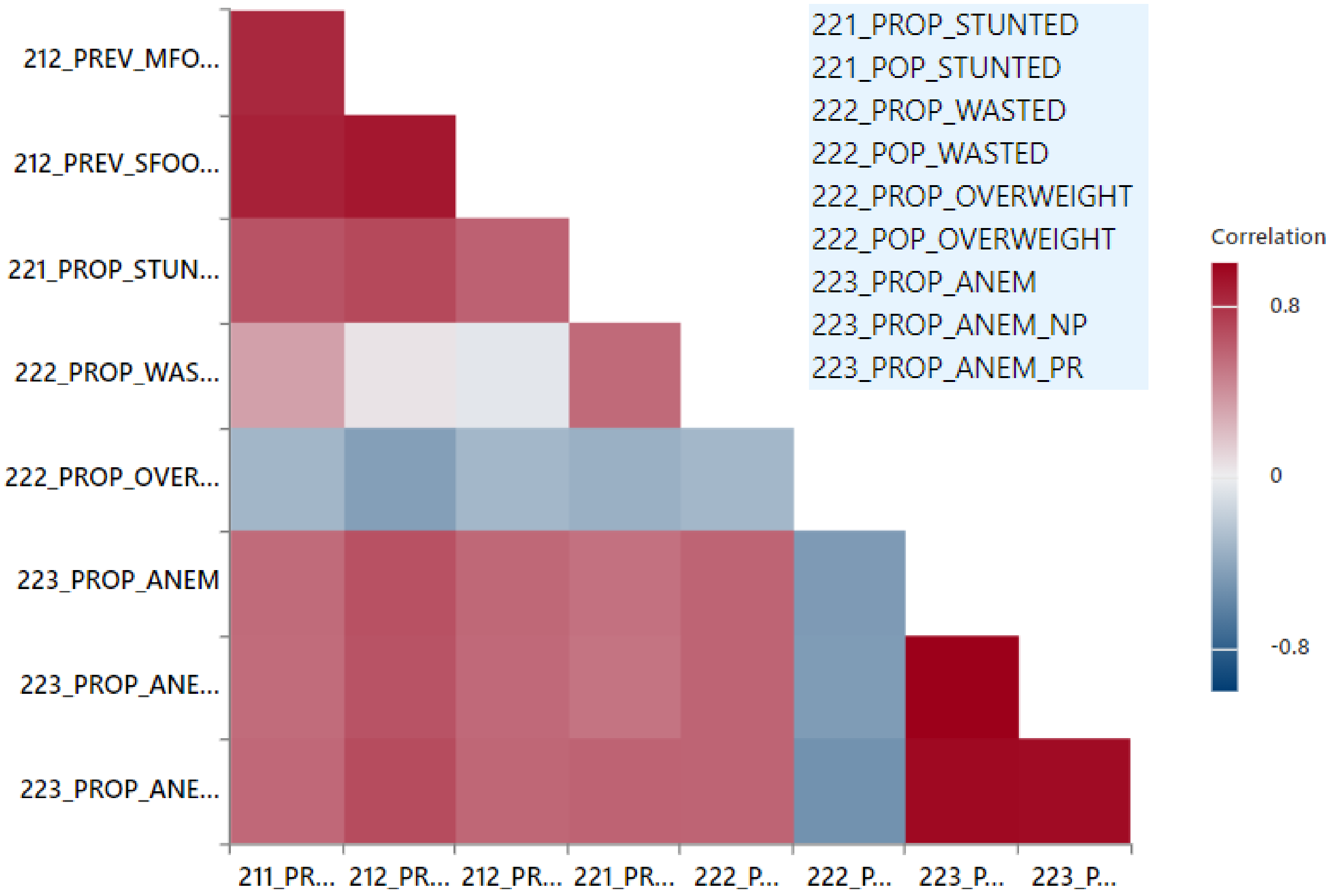
MSS Graph Builder – Rule-Based Automation



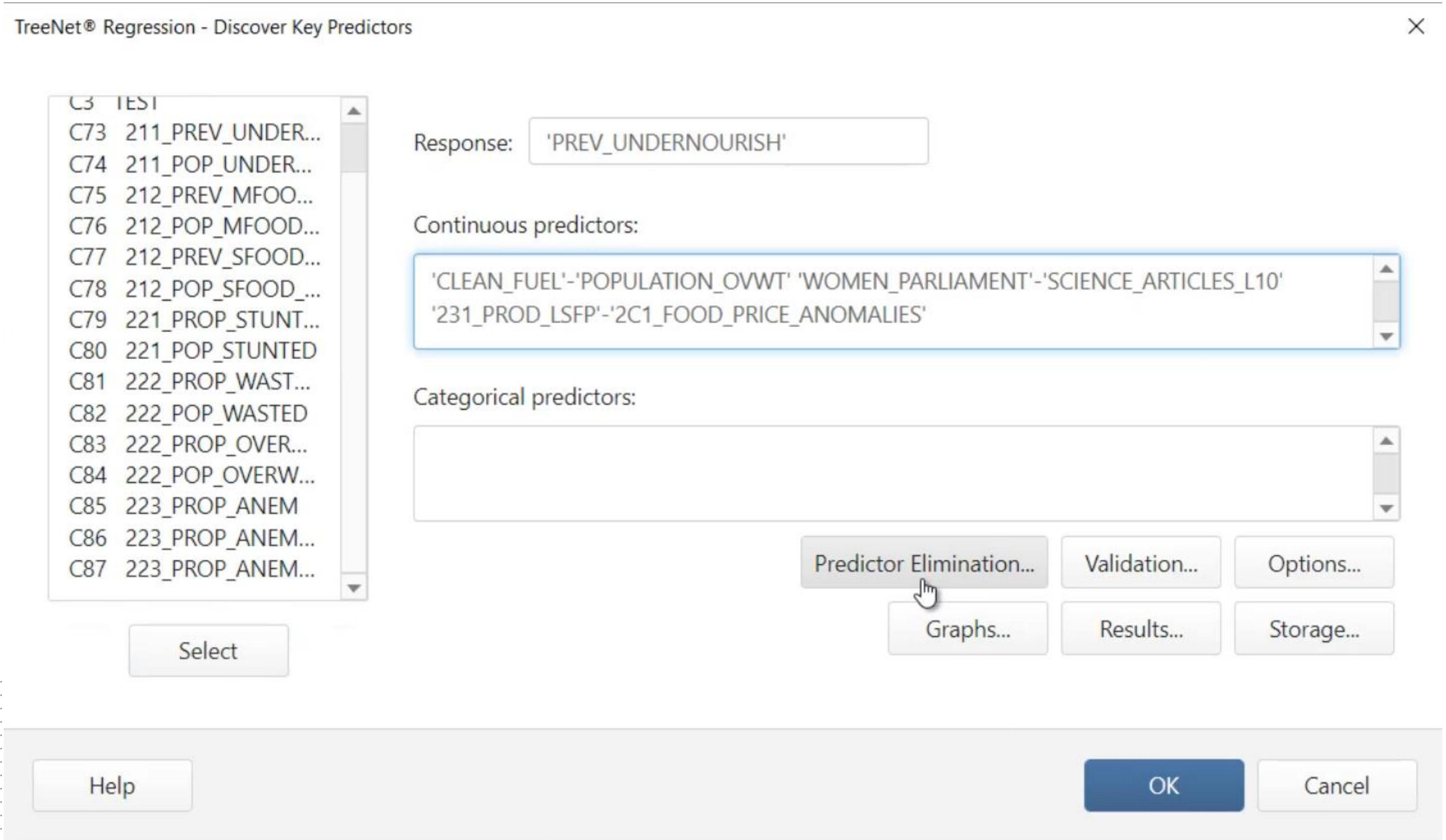
Minitab

Eliminating Dependencies

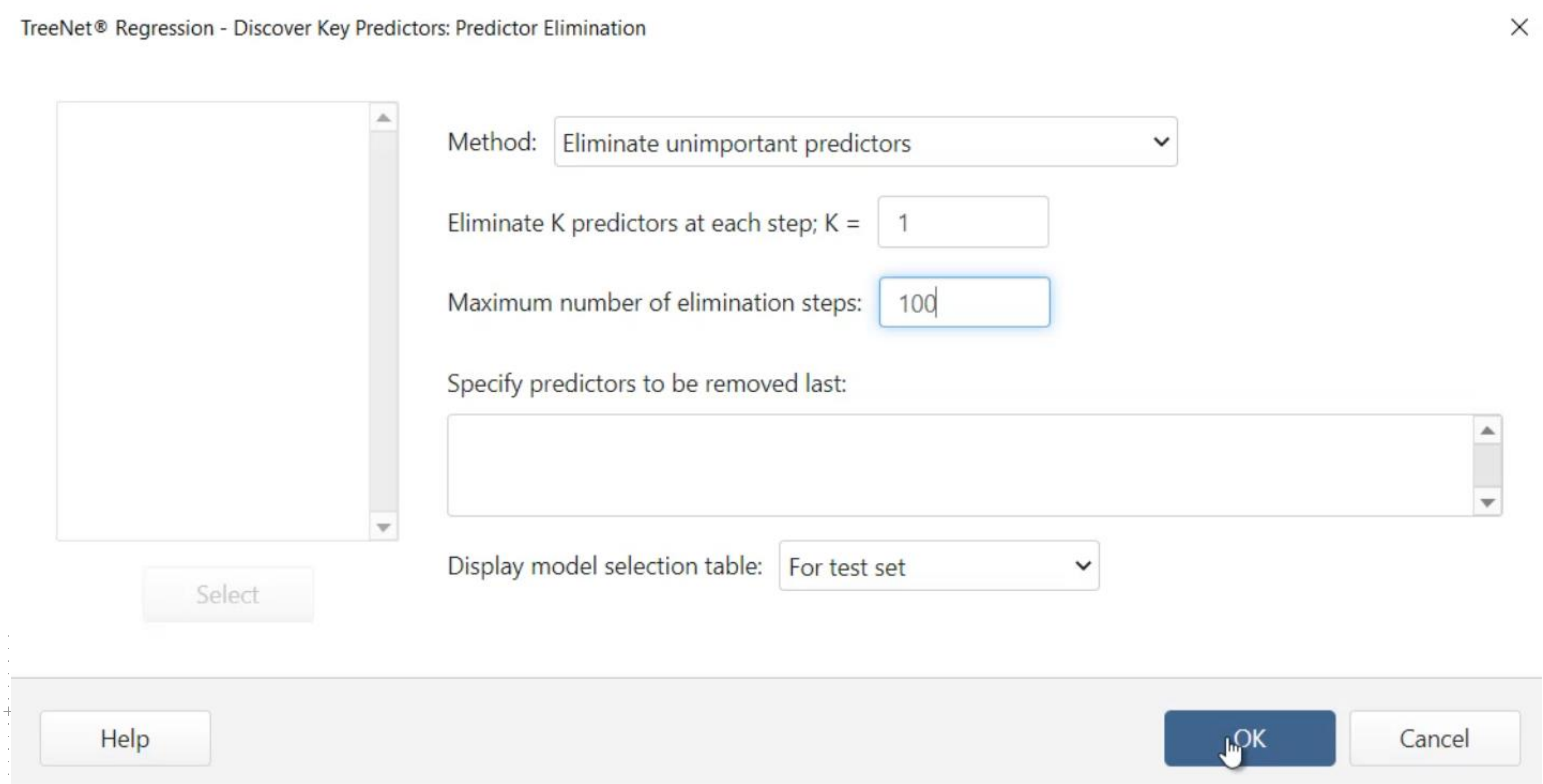
- **Prevalence of Undernourishment** is the response variable of interest
- Using Graph Builder and common sense, we have decided to drop UN indicators 2.1.1, 2.1.2, 2.2.1, 2.2.2, 2.2.3 because they are the **direct effects** of undernourishment and not its causes
- We kept all remaining indicators, including the ones largely missing



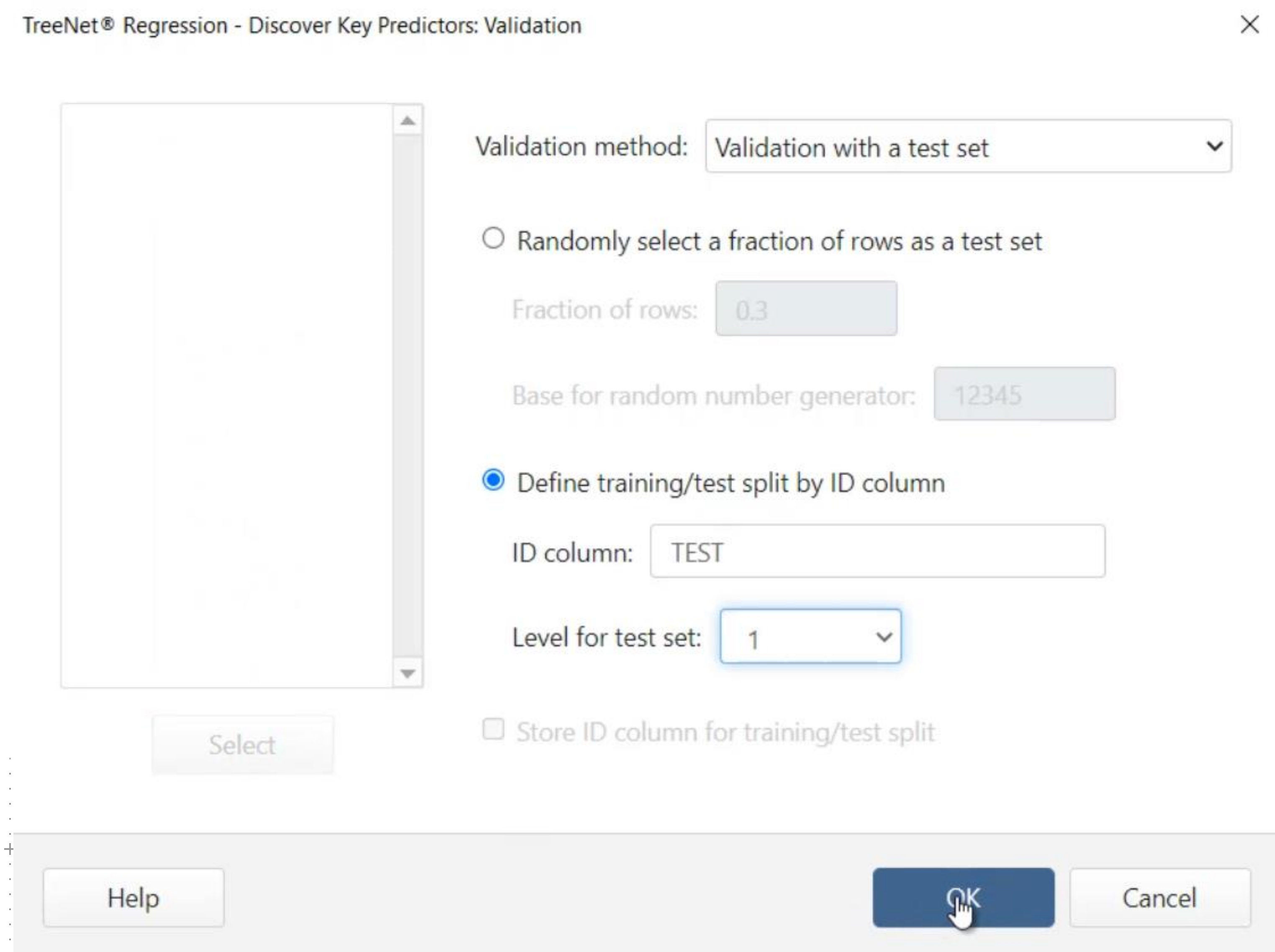
Setting Up Predictor Discovery



Setting Up Predictor Discovery

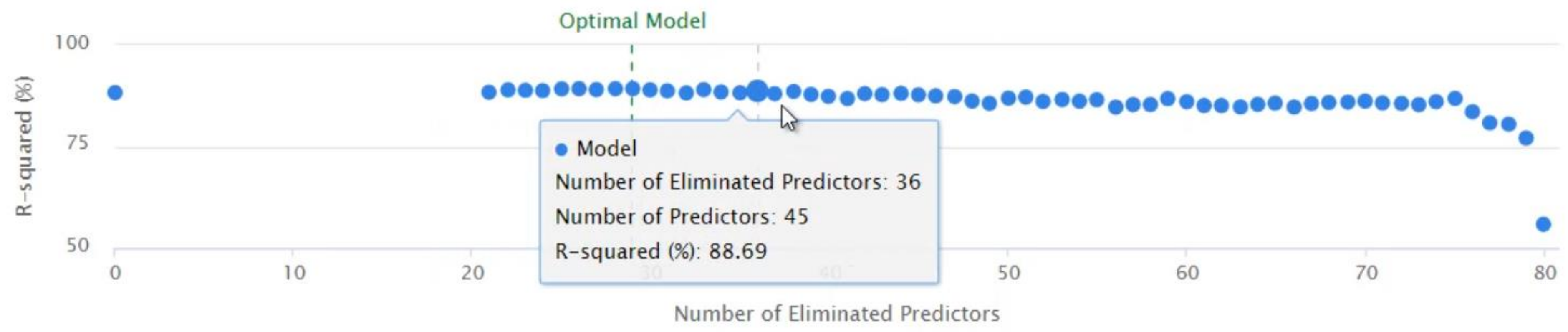


Setting Up Predictor Discovery



Final Model Selection

Select an Alternative Model



Model	Optimal Number of Trees	R-squared (%)	Number of Predictors	Eliminated Predictors
1	300	88.25	81	None
2	300	88.35	60	2B1_AG_EXPORT_SUBSIDIES, 251_PLANT_GENETIC, 251_BREEDS_GENETIC, 232_INCOME_SSFP, 232_INCOME_LSFP, 231_PROD_SSFP, 231_PROD_LSFP, SCIENCE_ARTICLES_L10, POP_DENSITY_L10, DROUGHT_INDEX, RAINF

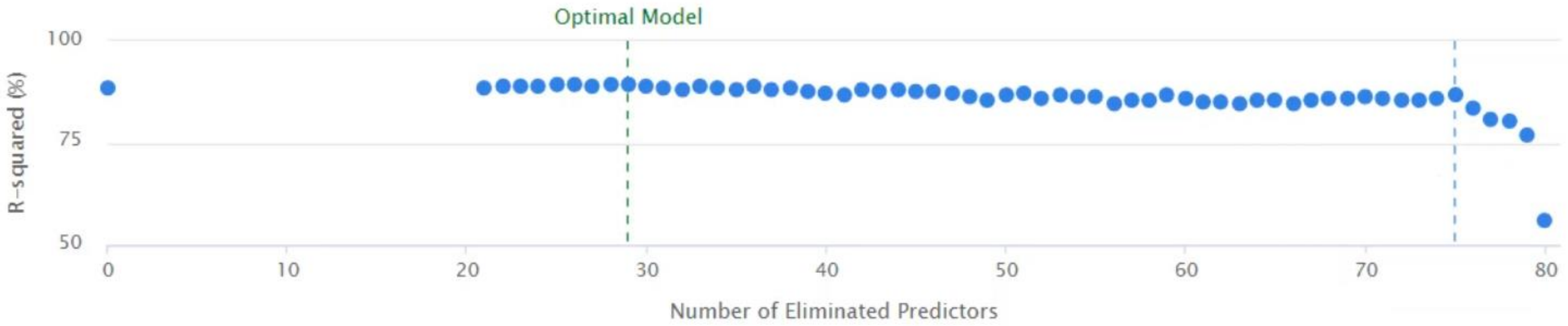
Help

Display Results

Cancel

Final Model Selection

Select an Alternative Model



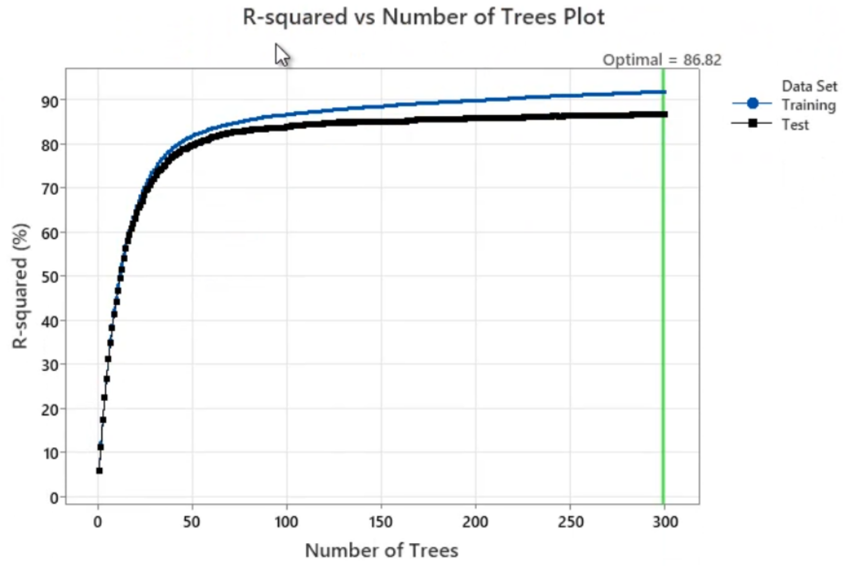
Model	Optimal Number of Trees	R-squared (%)	Number of Predictors	Eliminated Predictors
55	300	86.01	7	NR_VALUE_ADDED
56	299	86.82	6	SCIENCE_ARTICLES
57	300	83.53	5	RULE_OF_LAW
58	242	80.81	4	F2M_LABOR
59	278	80.47	3	CO2_EMISSIONS

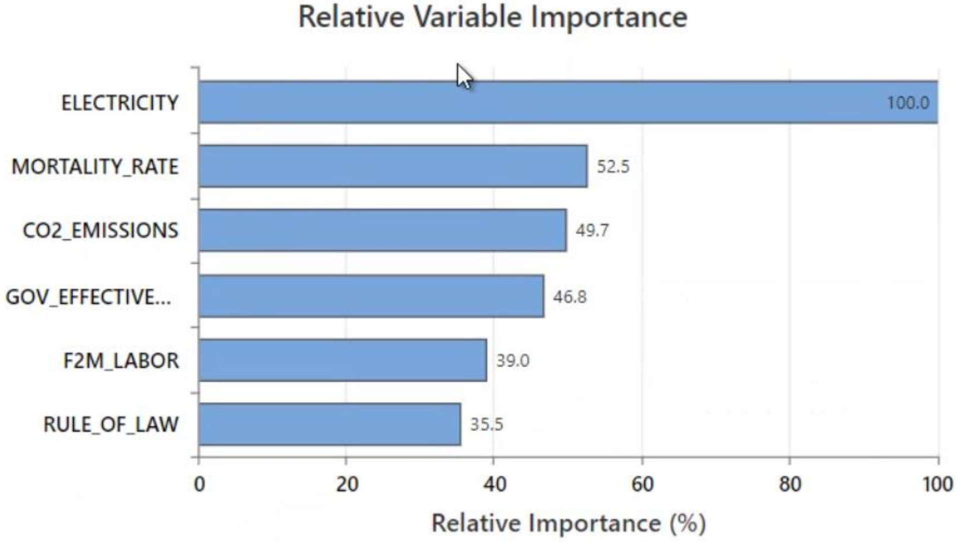
Help

Display Results

Cancel

Final Model



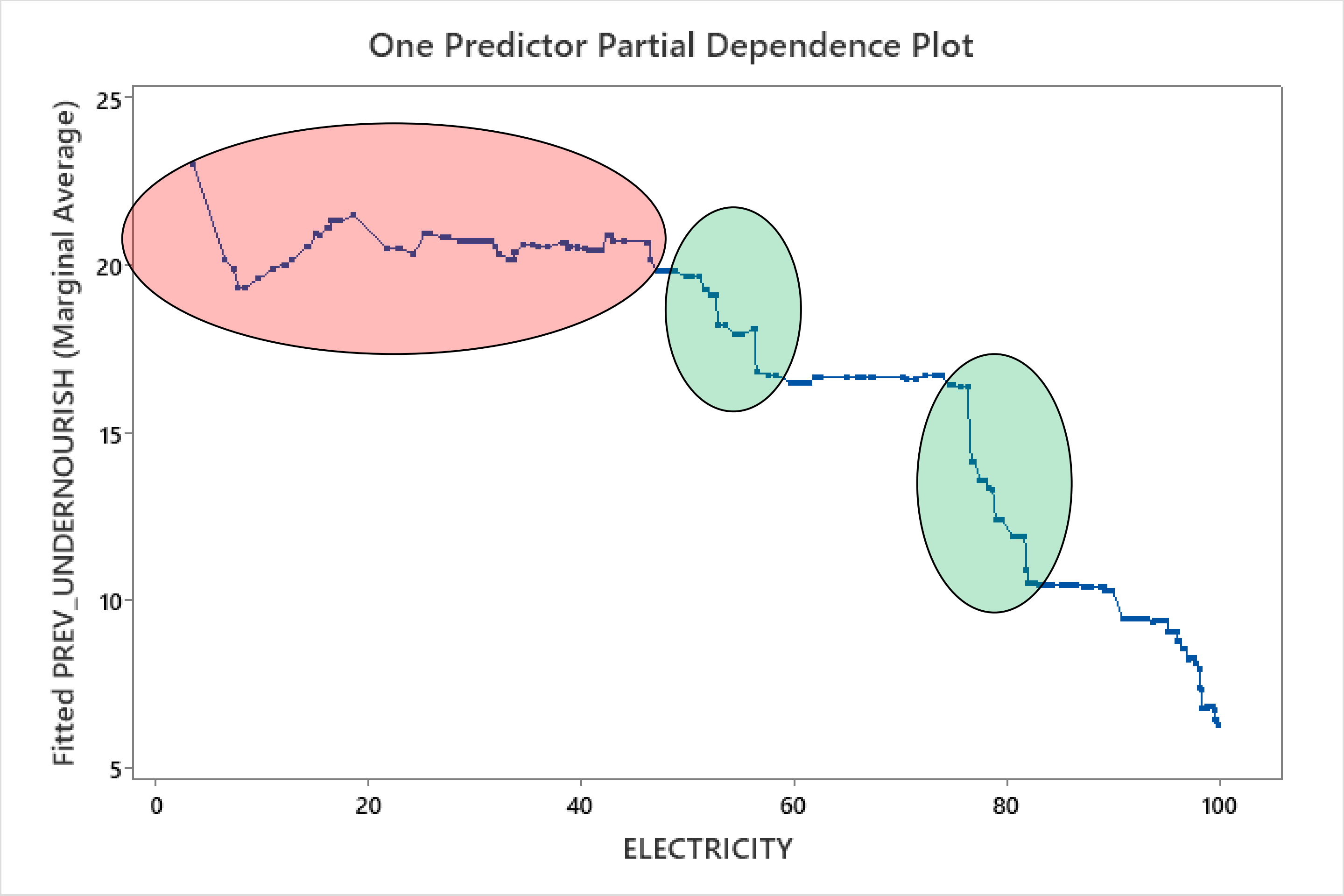


Model Summary

Total predictors	6
Important predictors	6
Number of trees grown	300
Optimal number of trees	299

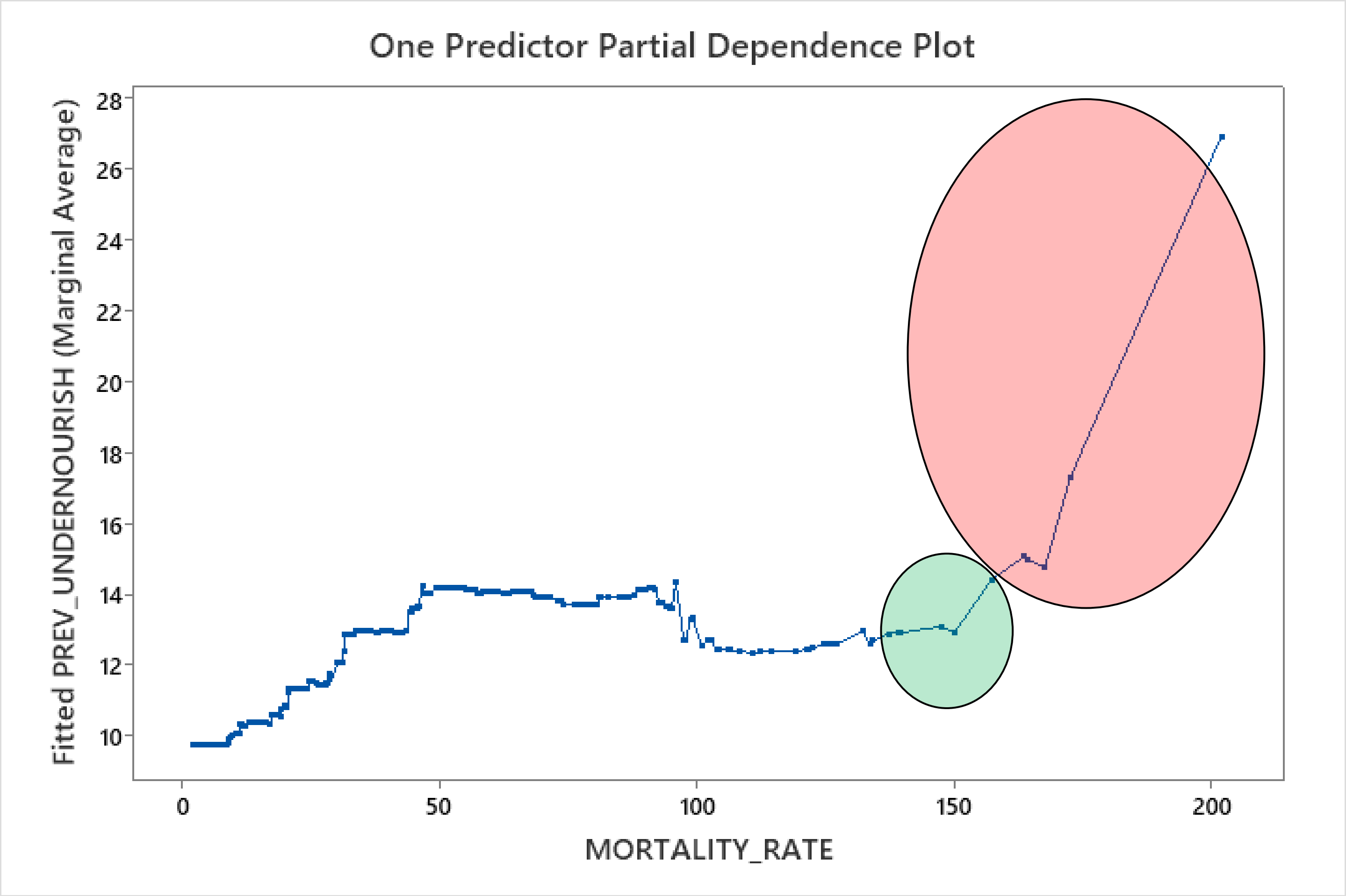
Statistics	Training	Test
R-squared	91.85%	86.82%
Root mean squared error (RMSE)	3.1675	3.8099
Mean squared error (MSE)	10.0328	14.5156
Mean absolute deviation (MAD)	2.0640	2.4547
Mean absolute percent error (MAPE)	0.2435	0.2850

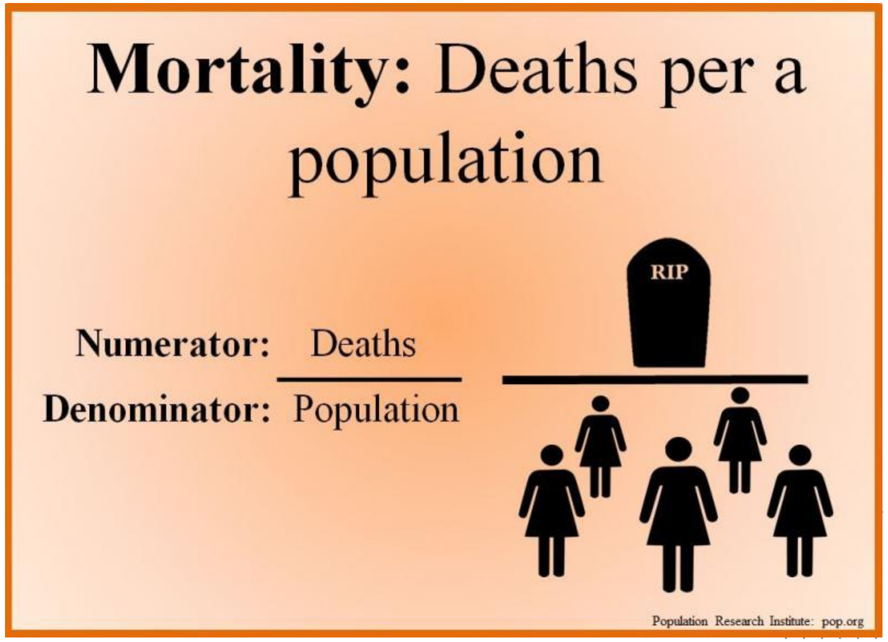
Model Insights - 1



- 60% and 80% electrification provide major markers associated with significant reduction in the prevalence of undernourishment

Model Insights - 2

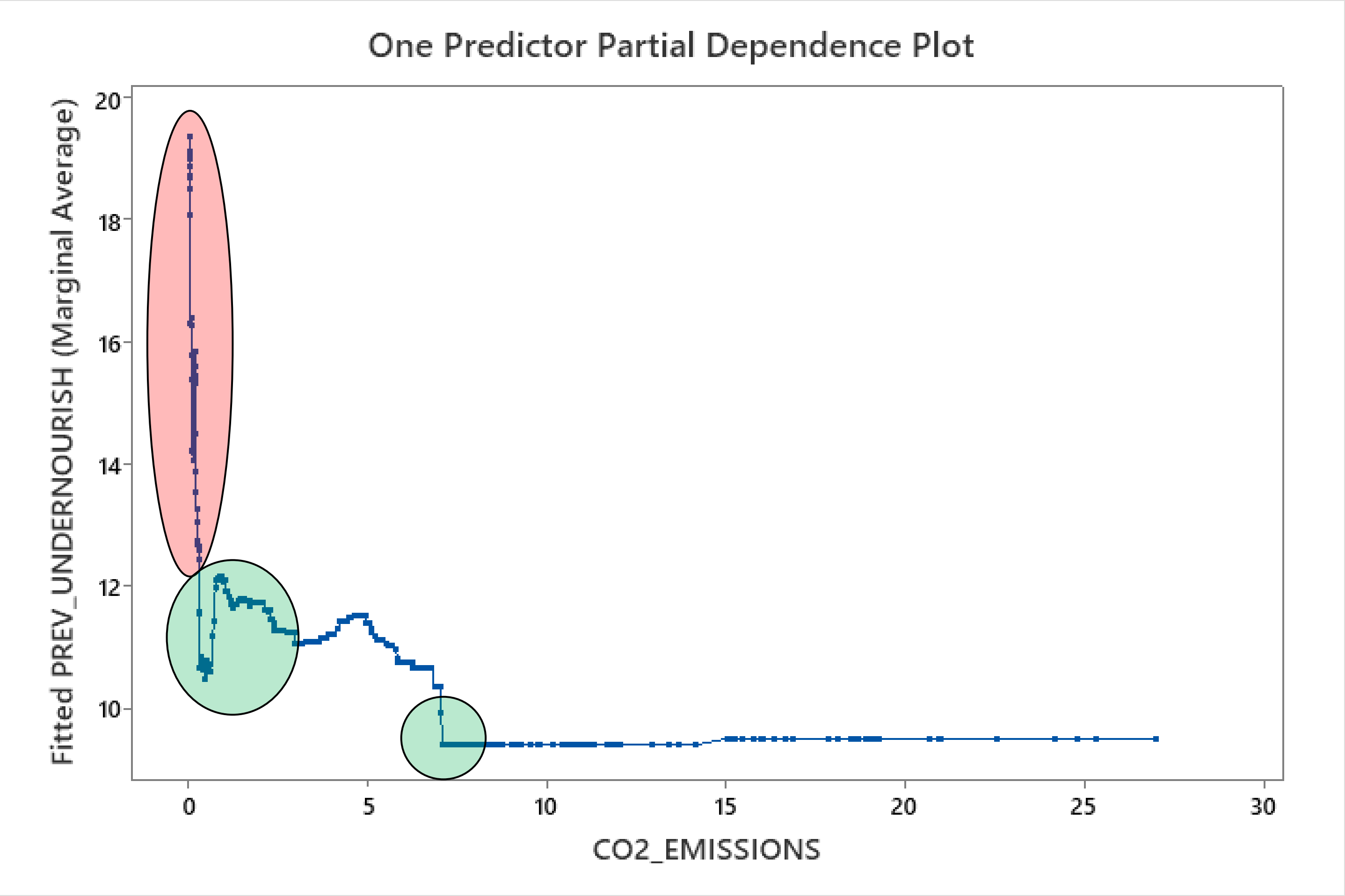




- Reducing mortality rate to below 150 is another substantial marker

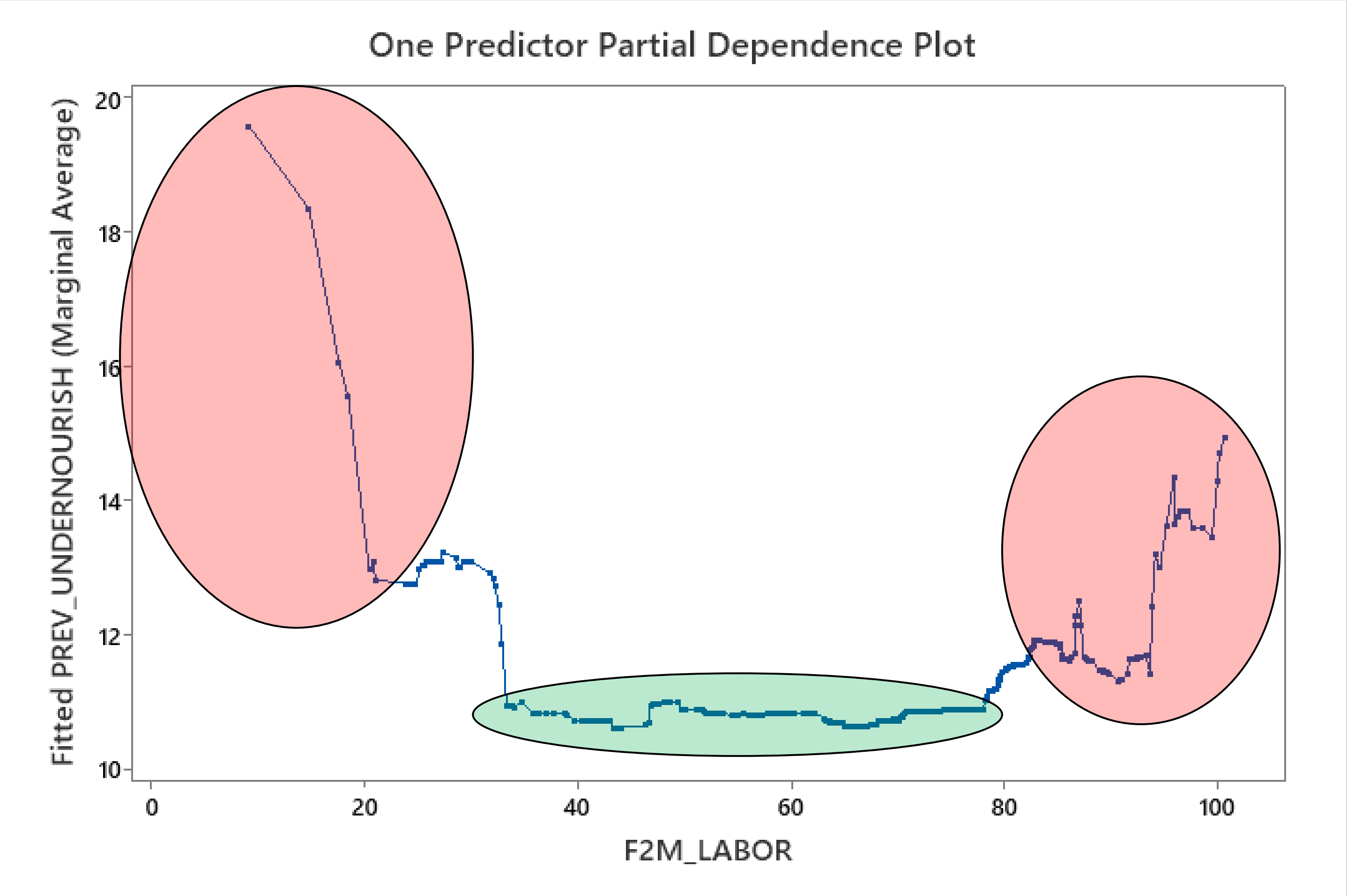
Minitab

Model Insights - 3



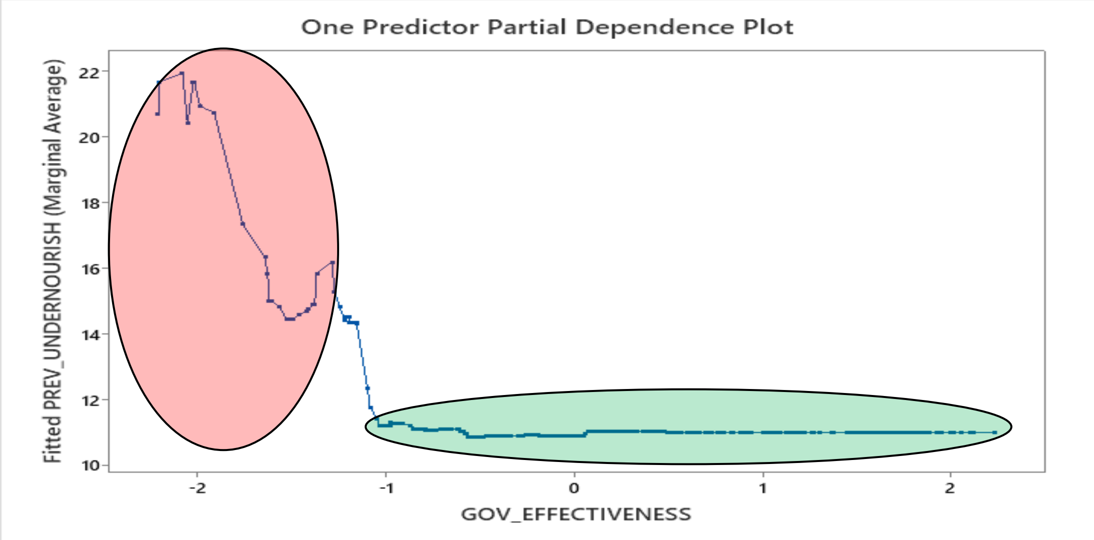
- Introduce minimal level of industrialization while allowing for manageable levels of CO_2 emissions (small footprint)

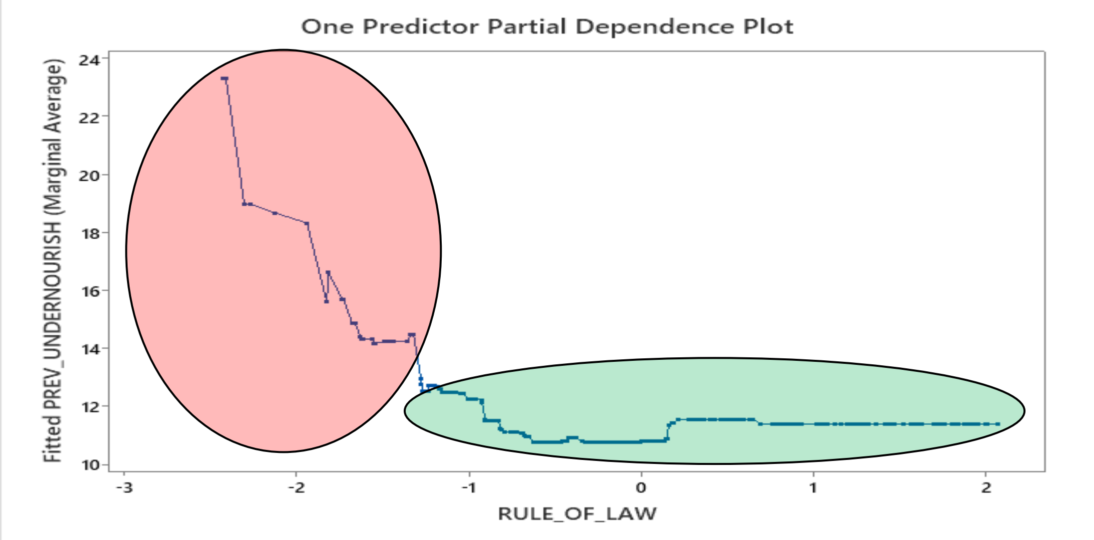
Model Insights - 4



- Develop policies aimed at reasonable balance between male and female labor: avoid both extremes

Model Insights - 5





- Develop local policies aimed at increasing government effectiveness and promote the rule of law

Problem 2 solved!
AutoML has identified
6 relevant predictors
out of the initial 97

Conclusion

- Graph Builder **Rule-Based AI** feature allows to quickly explore individual variables and their mutual dependencies
- Discover Best Model **AutoML AI** feature allows to quickly zero in on the most accurate model that suits your data
- Discover Key Predictors **AutoML AI** feature allows to identify which variables to focus on
- **AutoML AI** saves a huge amount of manual modeling effort and will give you a real advantage over your competitors!

Q&A

Minitab

Upcoming In-Person Events

Dates and Location in the US

- Rosemont, IL – June 18th
- Columbus, OH – August 15th
- Dallas, TX – September 10th
- Anaheim, CA – Date TBD

Minitab
EXCHANGE

Minitab

You have data. We have solutions. Imagine the possibilities.

At Minitab, we help customers around the world leverage the power of data analysis to gain insights and make a significant impact on their organizations. By unlocking the value of data, Minitab enables organizations to improve performance, develop life changing innovations and meet their commitments of delivering high quality products and services and outstanding customer satisfaction.

Minitab Connect® | Minitab® | SPM® | Minitab Model Ops®

Minitab Workspace® | Minitab Engage® | Education Hub™ | Real-Time SPC™

Minitab

thank you

Gracias	ευχαριστώ	Danke	Grazie	благодаря
Hvala	Obrigado	Kiitos	شكراً	Tak
Ahsante	Teşekkürler	متشكرم	Salamat Po	감사합니다
Cảm ơn	شكريه	Terima Kasih	Dank u Wel	Děkuji
நன்றி	Köszönöm	ありがとうございます	ขอบคุณครับ	Dziękuję
谢谢	Tack	Mulţumesc	спасибо	Merci
תודה	多謝晒	дякую	Ďakujem	