Communicating Analytical Results and Interpreting your ML Models with SAS Viya – 5 Tips and Tricks that will make your life as data scientist easier

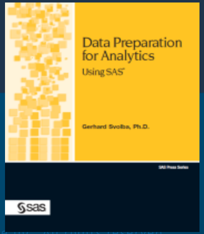
Gerhard Svolba

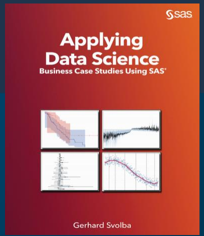
Analytic Solutions Architect

SAS Austria

Credits for Input to:
Martin Schütz, Tamara Fischer

Twitter: https://twitter.com/gsvolba
https://github.com/gerhard1050
https://www.linkedin.com/in/gerhardsvolba/

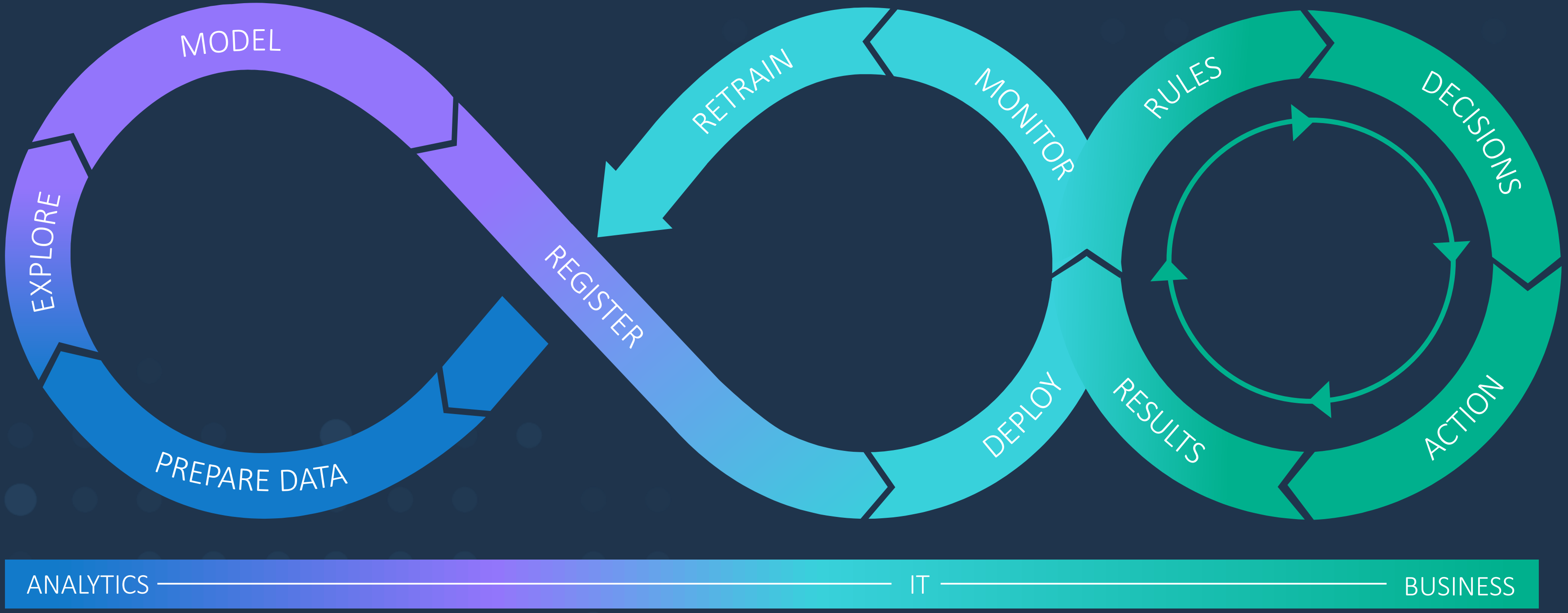




We (data scientists) want to communicate our results

- Acceptance of our results
- Better understanding – better usage in the business process
- Less „last minute“ misunderstandings

THE SAS DECISIONING PROCESS

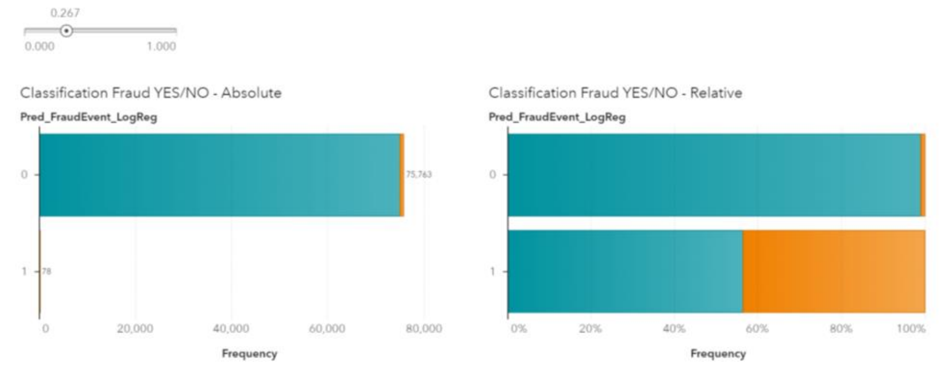


5 Tips (featuring SAS Visual Analytics, SAS Model Studio and SAS Coding)

1. Perform interactive cutoff analysis
2. Quantify the importance of explanatory variables
3. Turn on the model interpretability charts
4. Use a decision tree to „explain“
5. Display the (hidden) regression coefficient

Tip #1:

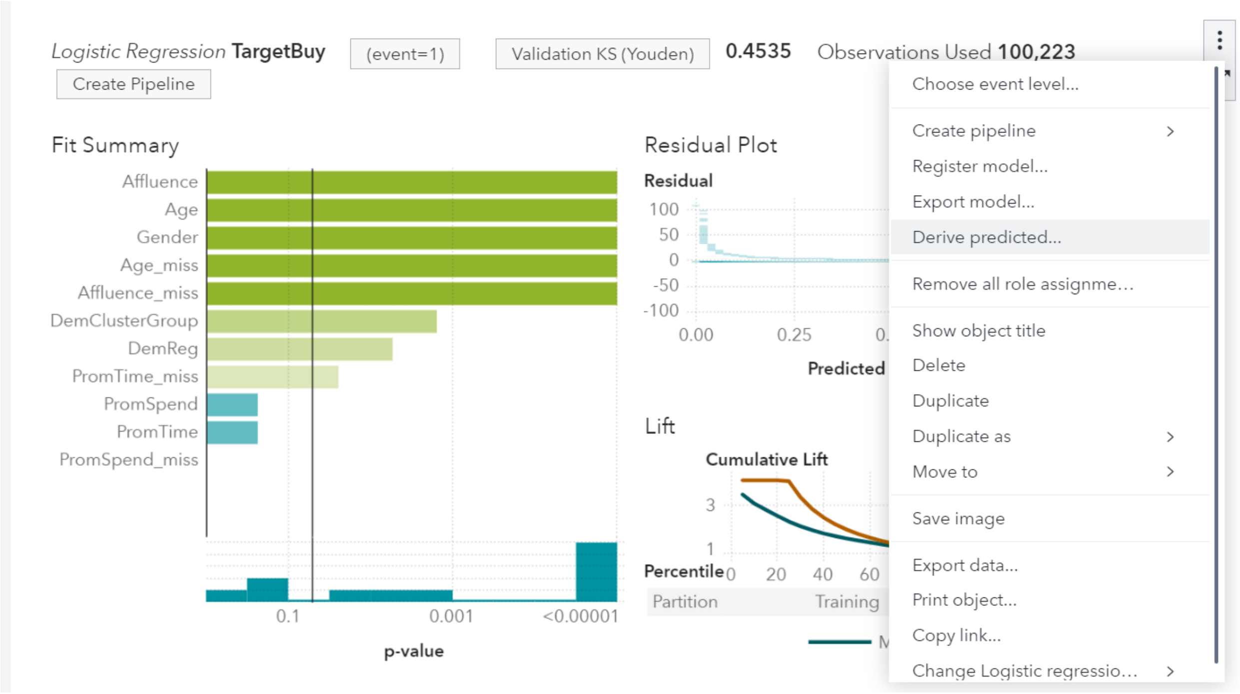
Perform interactive cutoff analysis to illustrate the consequences on the Good/Bad classification



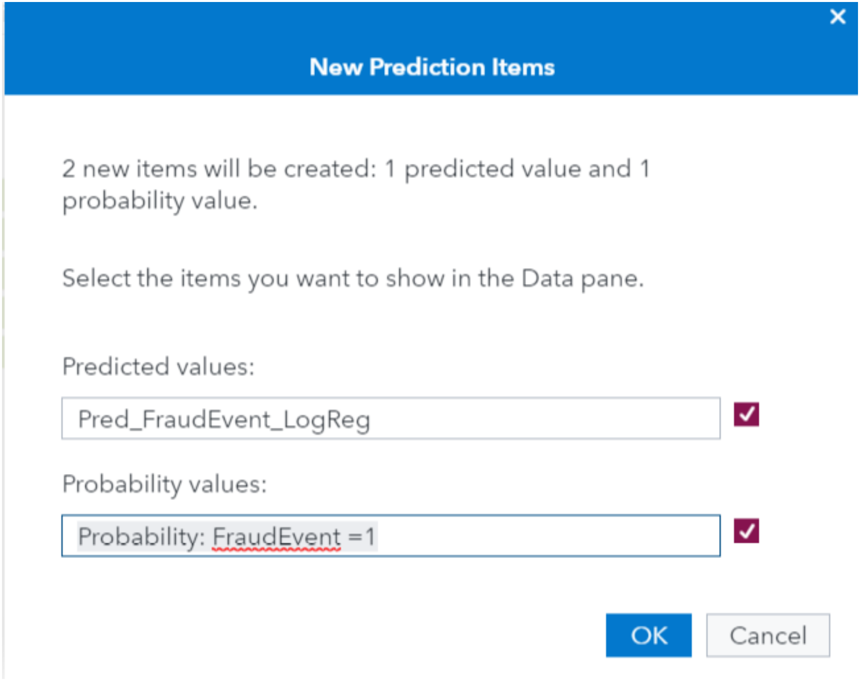
Illustrate the outcome (deliverable) of a predictive model

- A predictive model
- creates predictions.
- (In case of a binary classification task it outputs the probability the that event takes place.)
- You want to show this!
- And illustrate the consequences of different cutoff values for the business decision.

Select „Derive Predicted“ in a predictive model created with SAS Visual Analytics



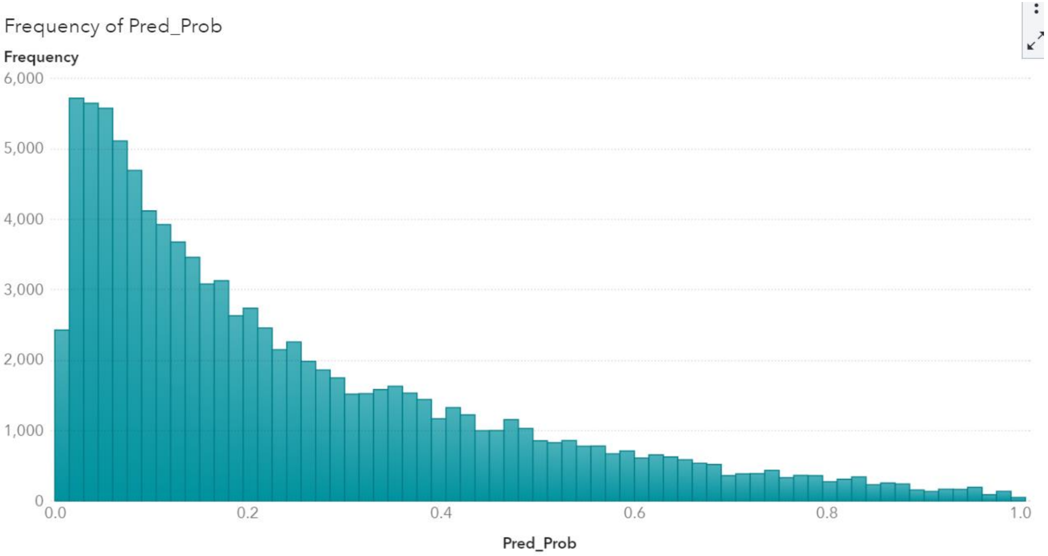
Name your output variables



You receive new variables in the data

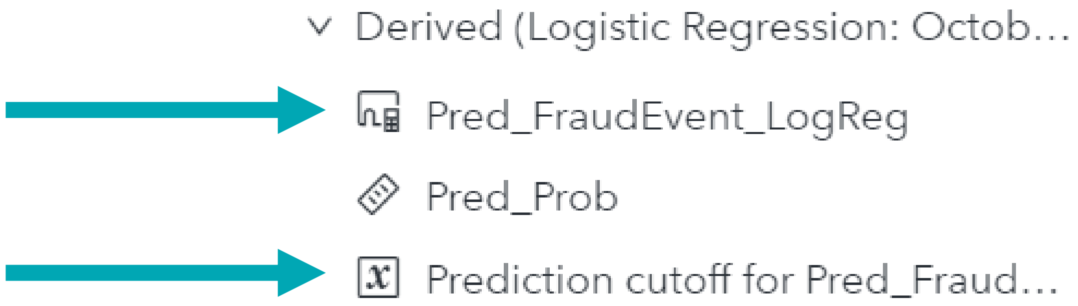
Derived (Logistic Regression: Octob…

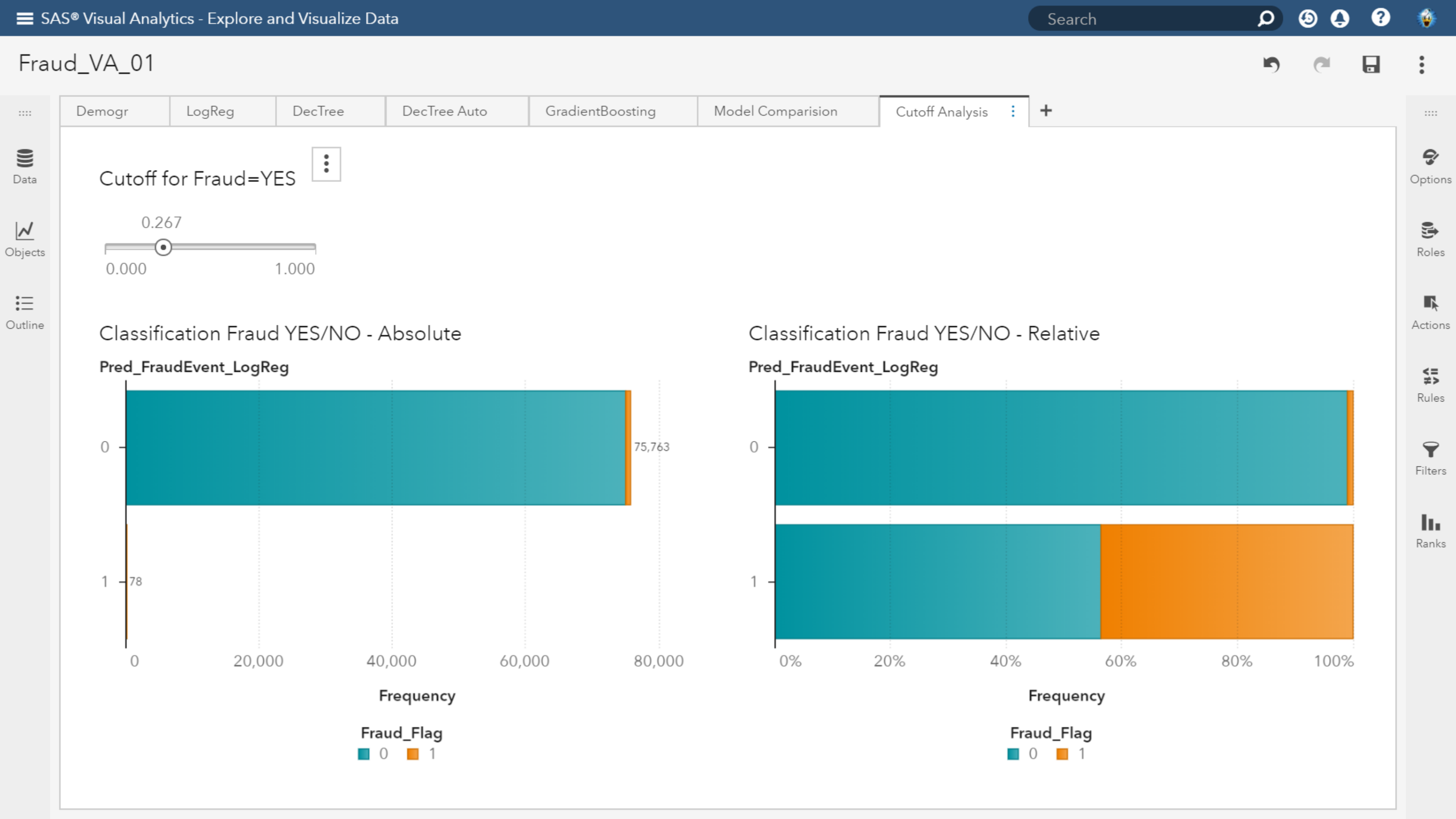
- Pred_FraudEvent_LogReg
- Pred_Prob
- Prediction cutoff for Pred_Fraud…

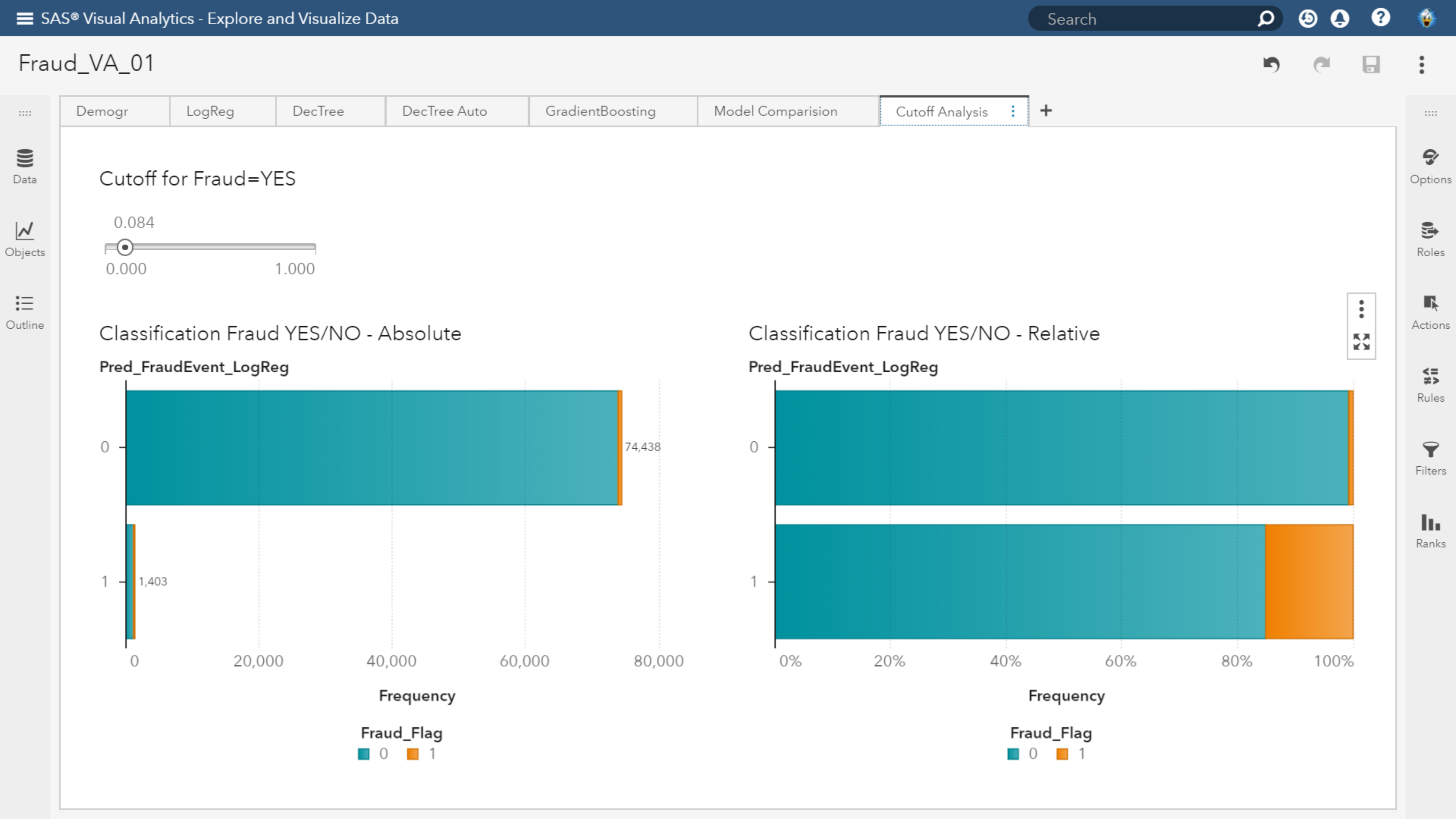


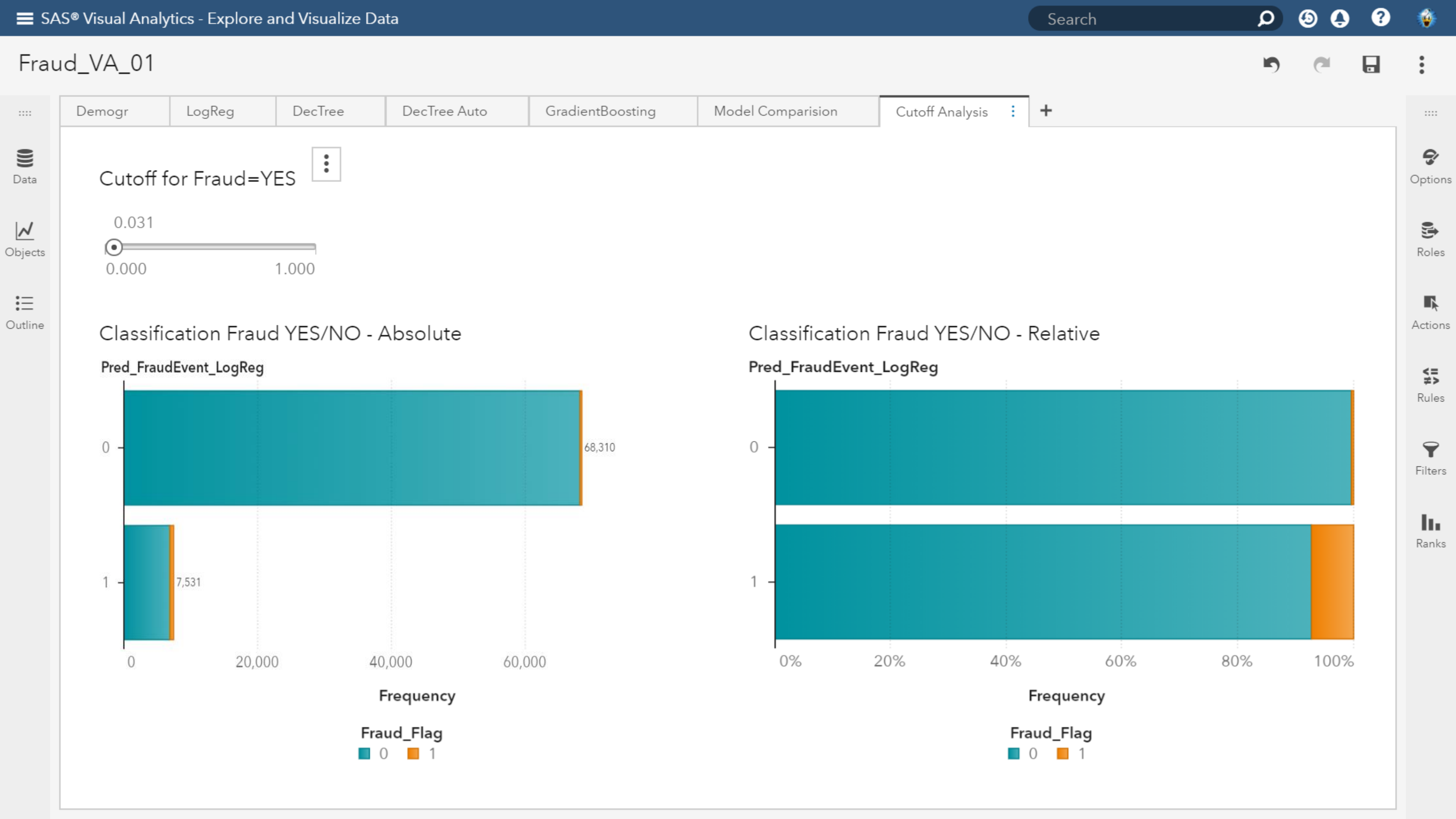
Also allows you to interactively „play“ with the cutoff point

- Important to illustrate the outcome of a predictive model
- What are the consequences of a certain cutoff point on
  - Number of customers, transactions flagged with YES
  - Expected false positives, false negatives, …



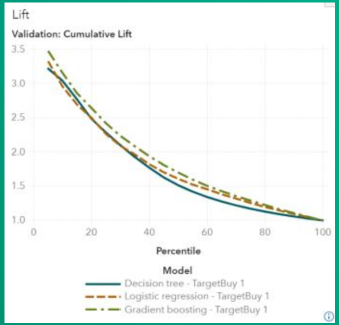

SAS® Visual Analytics - Explore and Visualize Data
Search
Fraud_VA_01
Demogr
LogReg
DecTree
DecTree Auto
GradientBoosting
Model Comparision
Cutoff Analysis
Data
Objects
Outline
Options
Roles
Actions
Rules
Filters
Ranks
Cutoff for Fraud=YES
0.267
0.000
1.000
Classification Fraud YES/NO - Absolute
Pred_FraudEvent_LogReg
0
1
75,763
78
0
20,000
40,000
60,000
80,000
Frequency
Fraud_Flag
0
1
Classification Fraud YES/NO - Relative
Pred_FraudEvent_LogReg
0
1
0%
20%
40%
60%
80%
100%
Frequency
Fraud_Flag
0
1


SAS® Visual Analytics - Explore and Visualize Data
Search
Fraud_VA_01
Demogr
LogReg
DecTree
DecTree Auto
GradientBoosting
Model Comparision
Cutoff Analysis
Data
Objects
Outline
Options
Roles
Actions
Rules
Filters
Ranks
Cutoff for Fraud=YES
0.084
0.000
1.000
Classification Fraud YES/NO - Absolute
Pred_FraudEvent_LogReg
0
1
74,438
1,403
0
20,000
40,000
60,000
80,000
Frequency
Fraud_Flag
0
1
Classification Fraud YES/NO - Relative
Pred_FraudEvent_LogReg
0
1
0%
20%
40%
60%
80%
100%
Frequency
Fraud_Flag
0
1

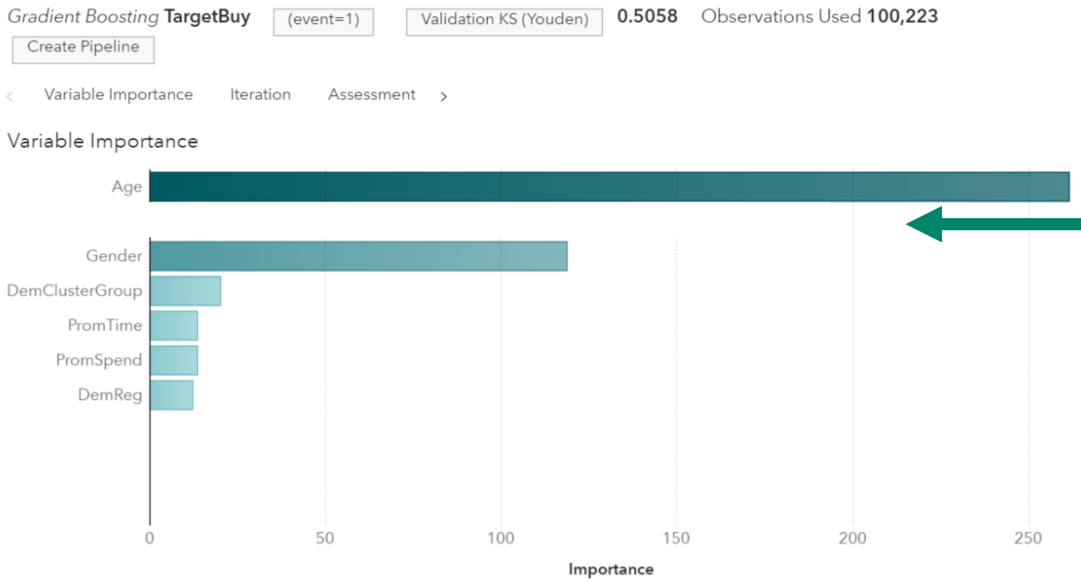

SAS® Visual Analytics - Explore and Visualize Data
Search
Fraud_VA_01
Demogr
LogReg
DecTree
DecTree Auto
GradientBoosting
Model Comparision
Cutoff Analysis
Data
Objects
Outline
Options
Roles
Actions
Rules
Filters
Ranks
Cutoff for Fraud=YES
0.031
0.000
1.000
Classification Fraud YES/NO - Absolute
Pred_FraudEvent_LogReg
0
1
68,310
7,531
0
20,000
40,000
60,000
Frequency
Fraud_Flag
0
1
Classification Fraud YES/NO - Relative
Pred_FraudEvent_LogReg
0
1
0%
20%
40%
60%
80%
100%
Frequency
Fraud_Flag
0
1

Tip #2:

Quantify the importance of explanatory variables in a predictive model with a business case



Variable importance chart in a gradient boosting model

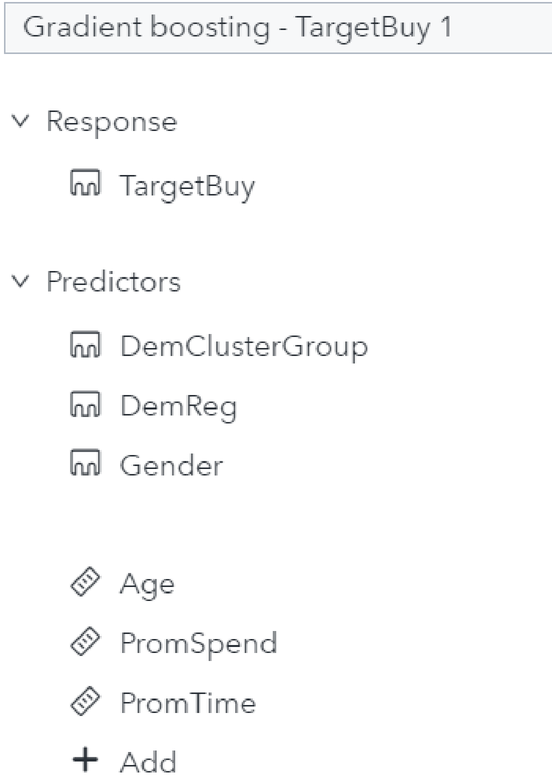


What happens, if we do not have variable „AFFLUENCE“ available?

What happens, if we do not have variable „AFFLUENCE“ available?

- Will other variables substitute the missing content?
- Will the model quality go down?

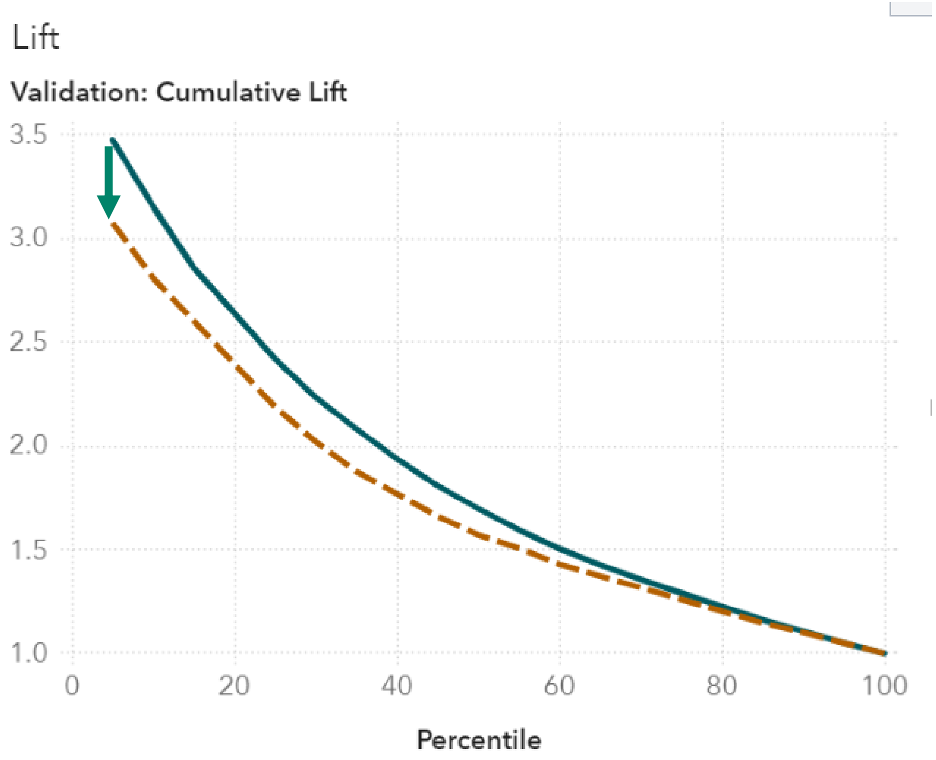
1. Create a copy of your model
2. Remove the variable of interest
3. Compare the old and the new model



Compare the old and the new model

- Lift drops from 3.47 to 3.07

- What does that mean in € ?



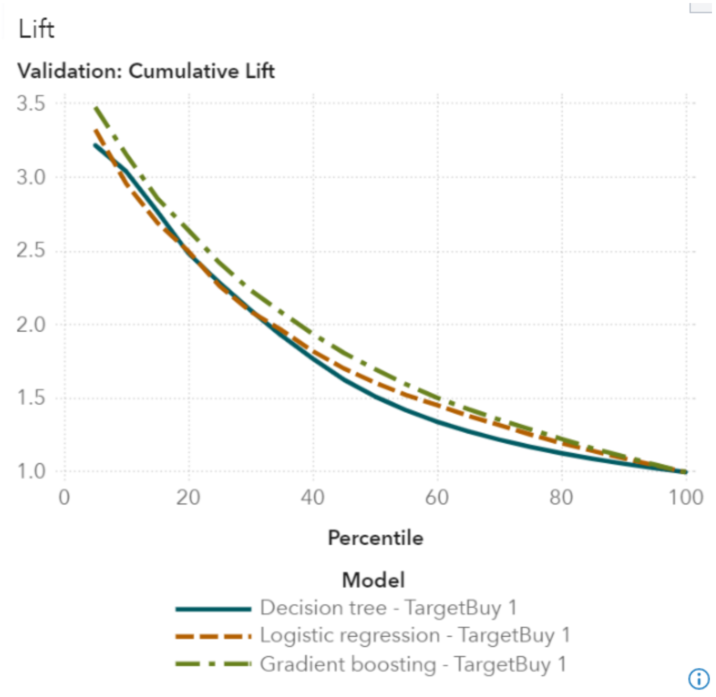
Calculating a business case

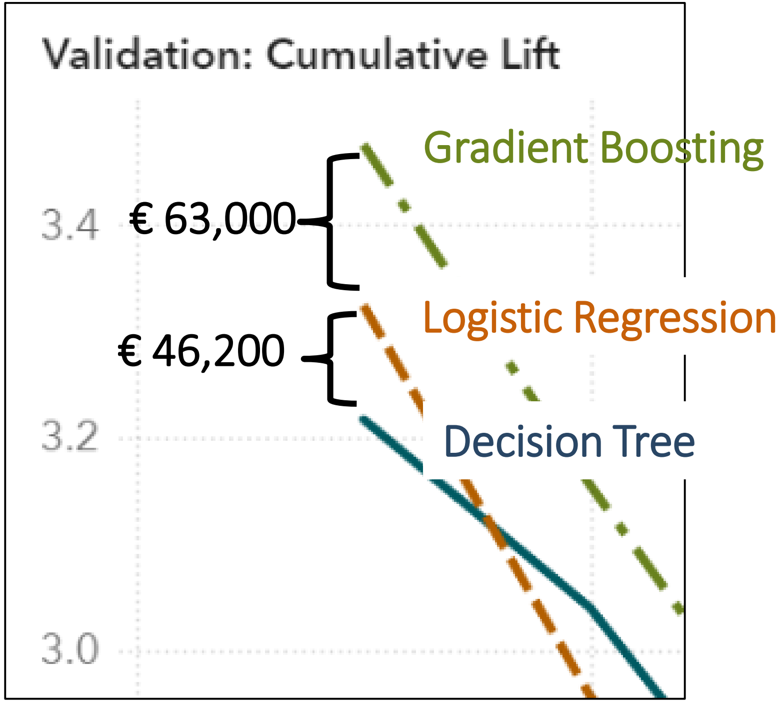
- Assume we have 2 Mio customers
- A campaign offer is sent to the top 5 % (100,000)
- A responding customer contributes a profit of € 35
- Assuming a baseline (autonomous) response of 12 %
  - A lift of 3.47 → 41.64 %
  - A lift of 3.07 → 36.84 %
- Not having variable AFFLUENCE costs us 4.8 % response
  - 100,000 * 4.8 % = 4800 missed responders * €35 = € 168,000

Quantify the effect of data quality on your business results

- Part III contains simulation case studies for data availability, data quantity, data correctness and data completeness
- Illustrated in € ($) values based on a business case study
- http://support.sas.com/svolba
- https://github.com/gerhard1050/Data-Quality-for-Data-Science-Using-SAS

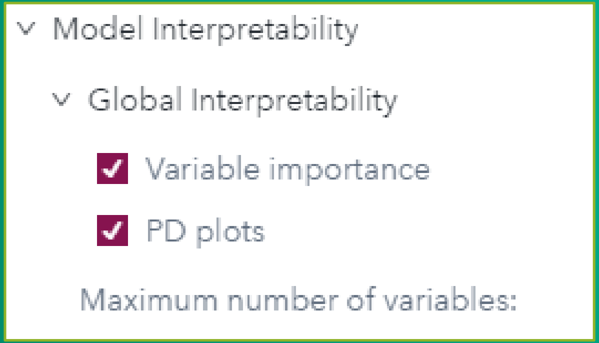
How much does it cost to use a simpler (better explainable) model for my predictions

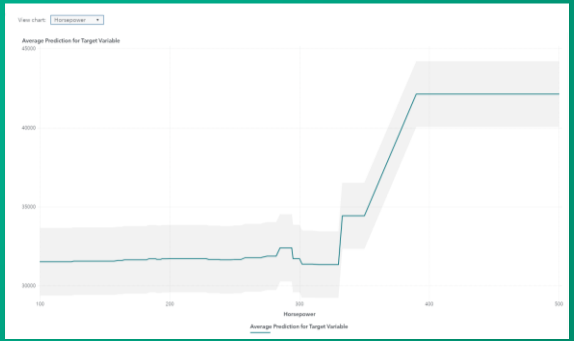




Tip #3:

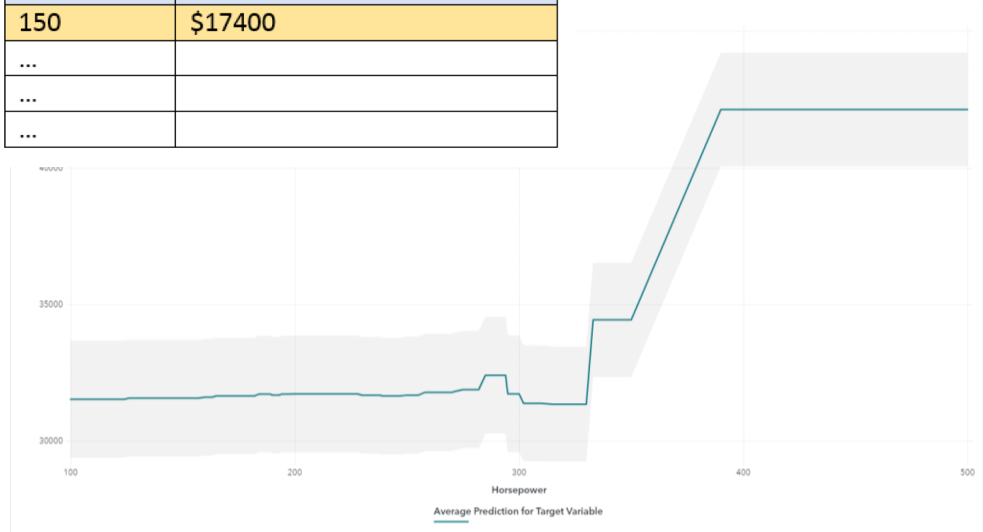
Turn on the model interpretability charts in SAS Model Studio

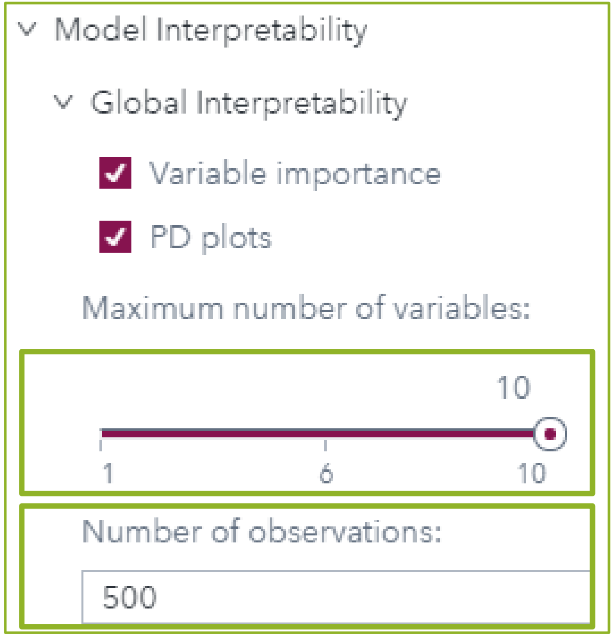




Partial Dependency Plot (PD)

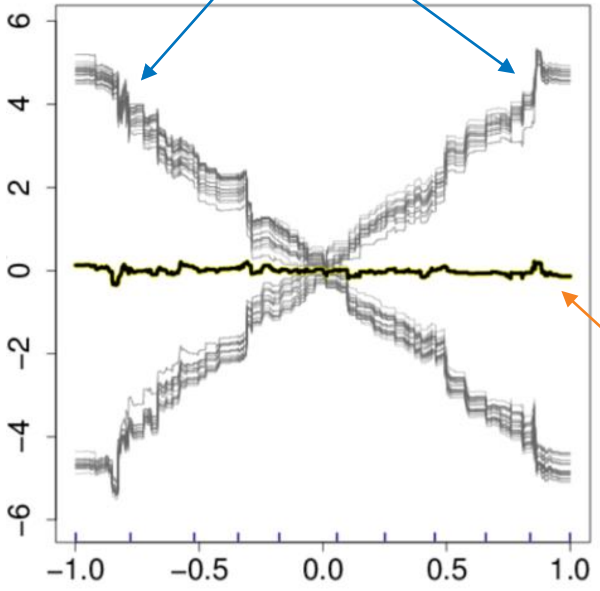
Horsepower	Average Predicted MSRP
100	$15400
150	$17400
…	
…	
…	



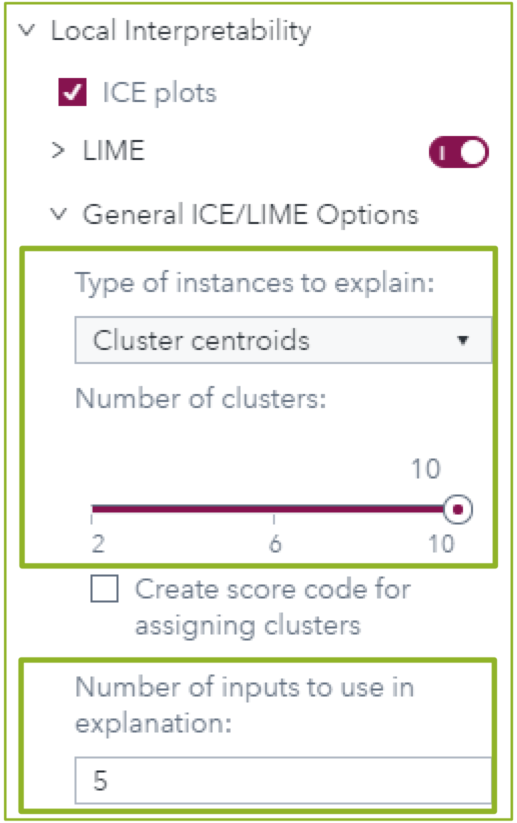


Individual Conditional Expectation (ICE)

Subgroups of Individual Observations

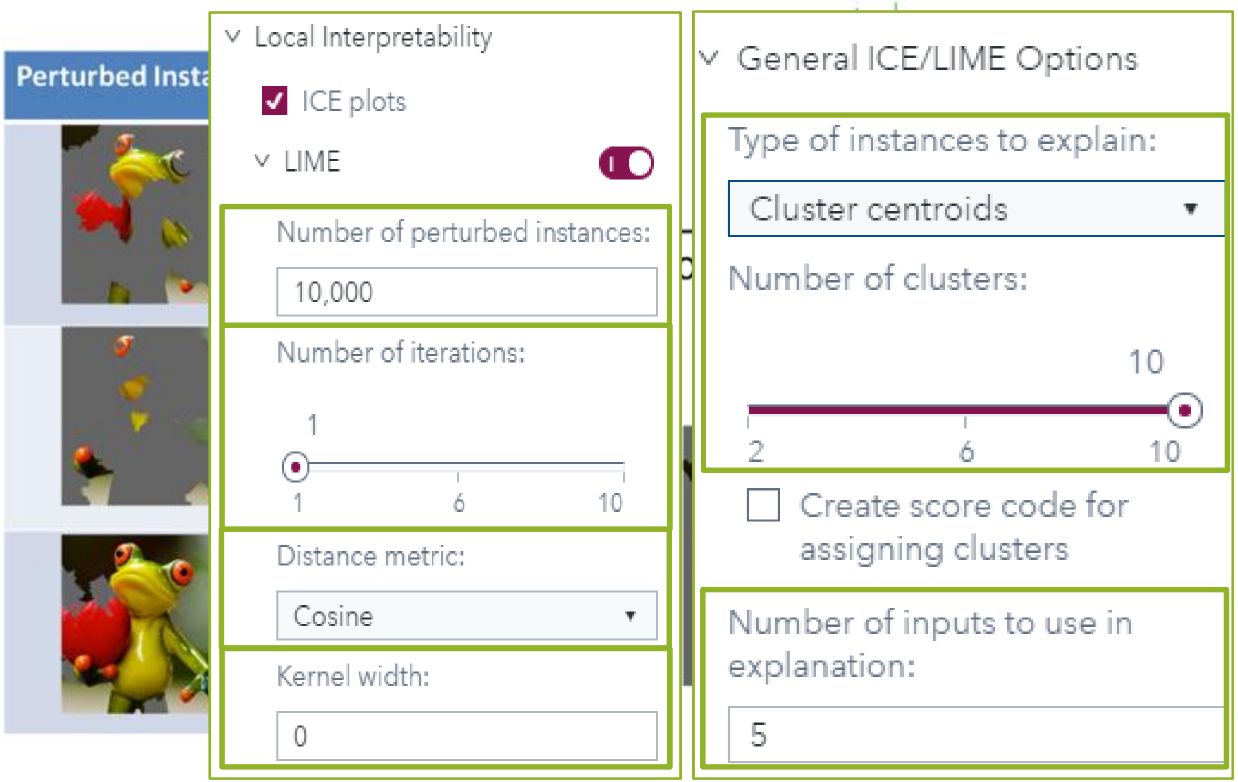


PD Funktion



Local Interpretable Model-Agnostic Explanation (LIME)

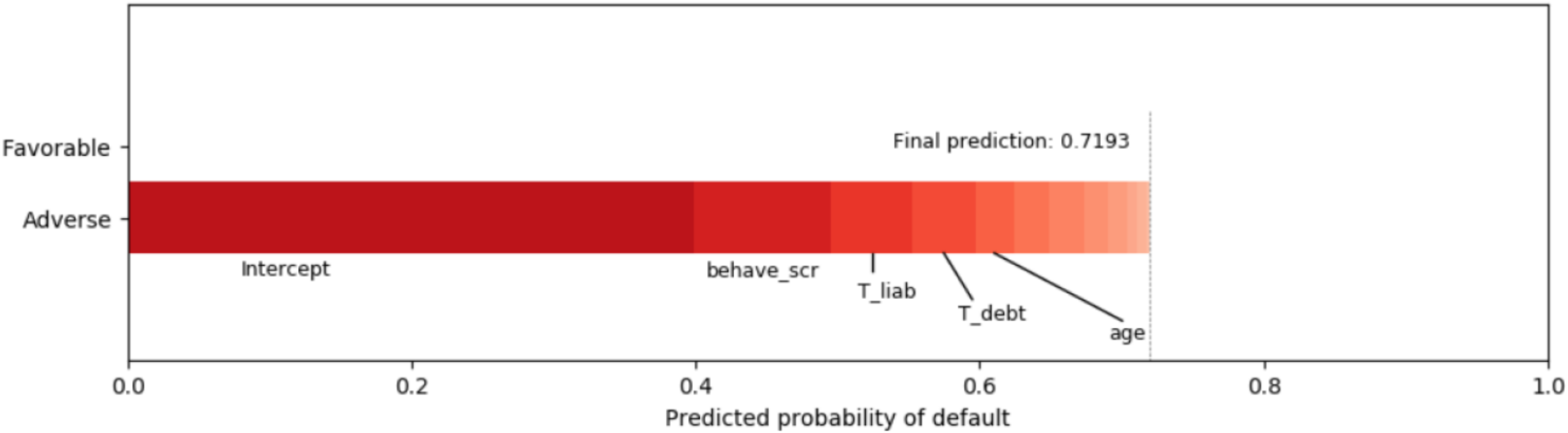
Original Image
P(tree frog) = 0.54



SHAP (SHapley Additive exPlanations)

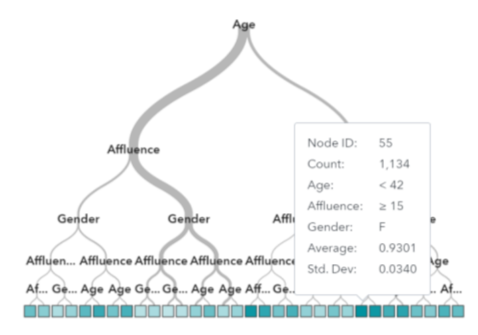
(using CAS-Action „linearExplainer“)

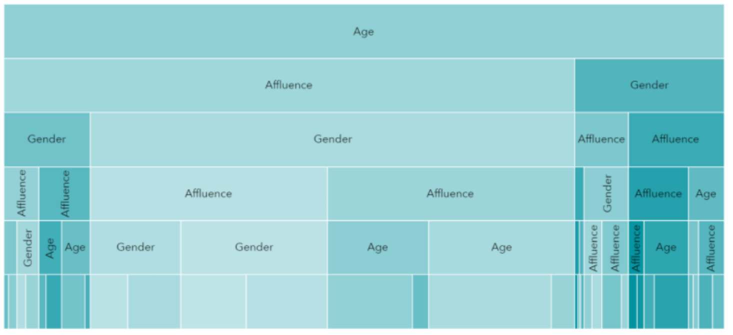
- based on game theory's Shapley values:
  - method for assigning payouts to players (depending on their contribution to the total payout)
  - Shapley values explain how to fairly distribute the payout among the players



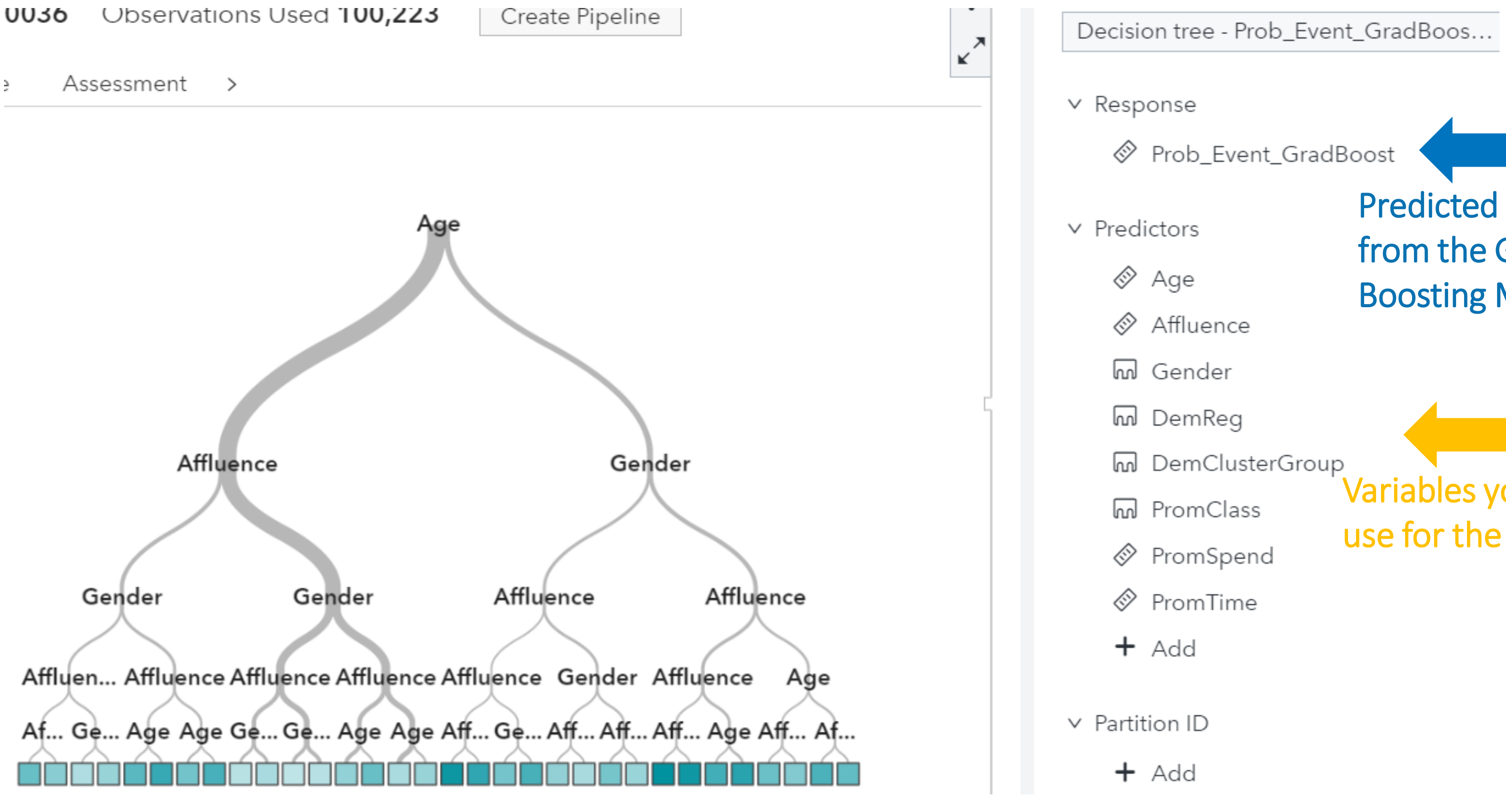
Tip #4:

Use a decision tree to „explain“ why customers received a high/low predicted probability





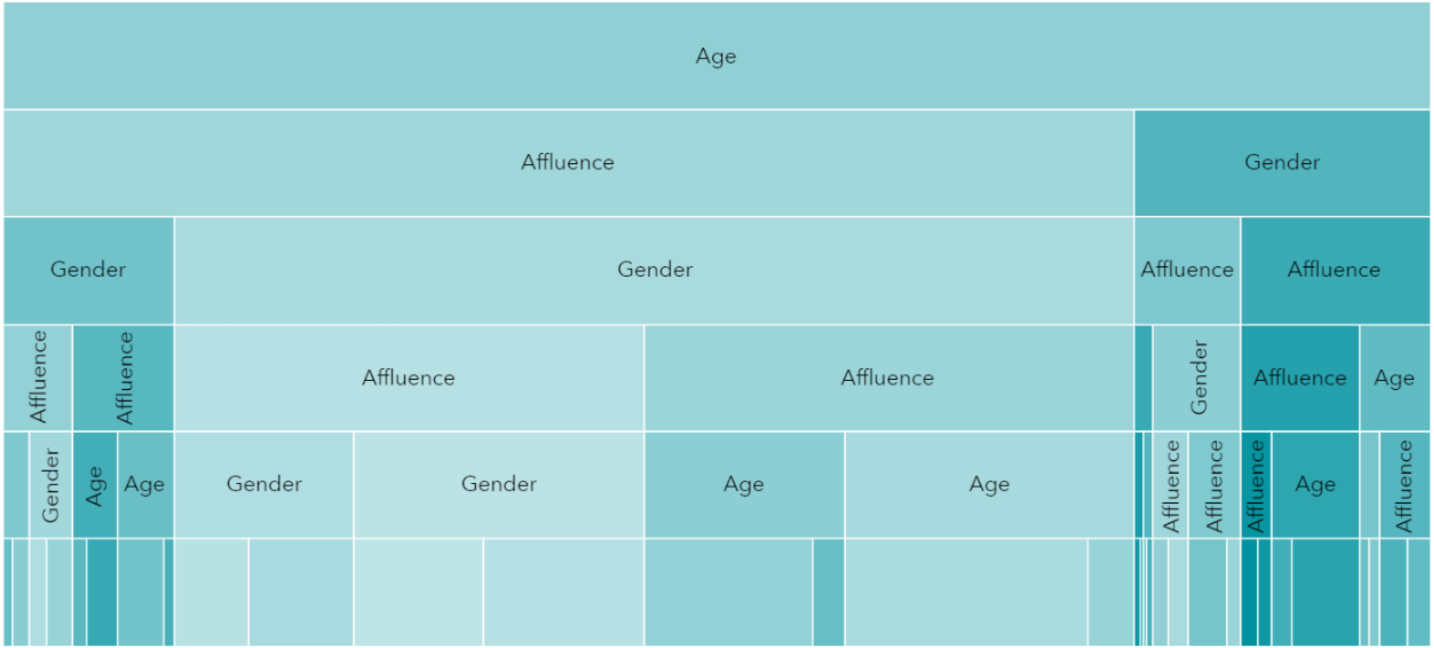
General Idea



Predicted Probability from the Gradient Boosting Model

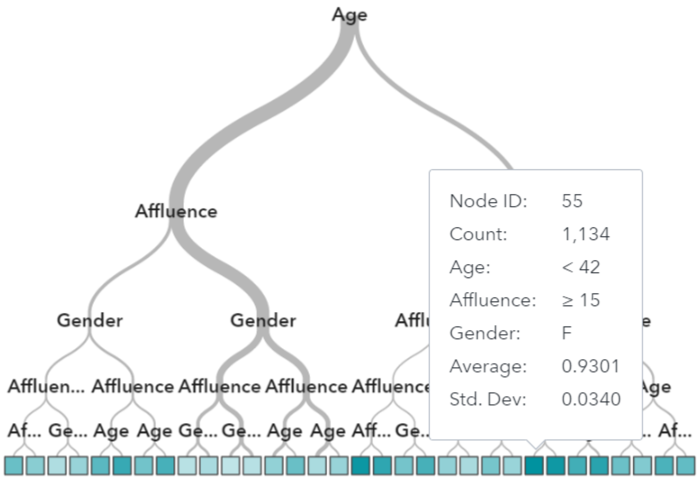
Variables you want to use for the explanation

Decision Tree creates segments with high/low predicted probability

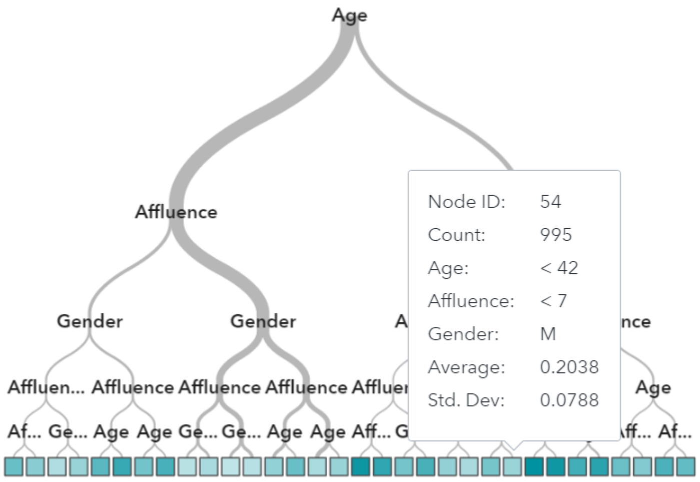


You can interpret the segments

Young affluent ladies →
PredictedProb = 93 %



Young non-affluent men→
PredictedProb = 20 %

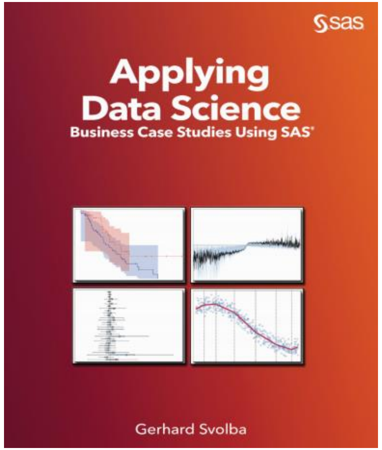


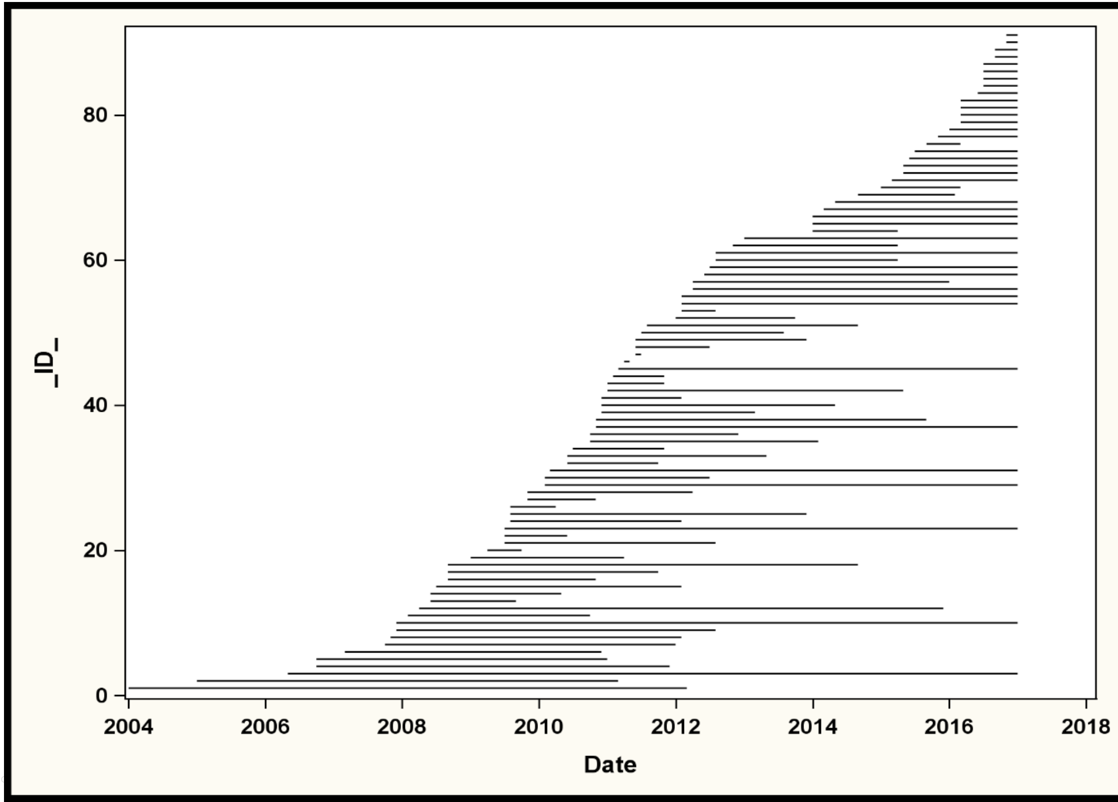
Tip #5:

Display the (hidden) regression coefficient of the reference category

Survival analysis performed for employee headcount data

- Observe Careers per Employee
  - Different length
  - „Left company“ or „censored“





How long will Gerhard still stay in our company?

Given certain risk factors, what is the expected survival in 6 months and the probability to resign within the next 6 months.

EmpNo	Department	Gender	TechKnowH...	_T_	EM_SURVFCST	EM_SURVEVENT	T_FCST
1003	TECH_SUPPORT	M	YES	128	0.240	0.000	134
1010	TECH_SUPPORT	M	YES	109	0.240	0.011	115
1023	SALES_ENGINEER	M	YES	90	0.108	0.313	96
1029	TECH_SUPPORT	M	YES	83	0.386	0.133	89
1031	TECH_SUPPORT	F	YES	82	0.177	0.219	88
1037	ADMINSTRATION	M	NO	74	0.471	0.066	80
1045	ADMINSTRATION	M	NO	70	0.494	0.053	76
1054	TECH_SUPPORT	F	YES	59	0.316	0.102	65
1055	SALES_ENGINEER	M	YES	59	0.313	0.103	65

Modeling the survival with the PHREG Procedure

```
proc phreg data=employees;
 CLASS department gender TechKnowHow / PARAM=reference REF=first;
 MODEL Duration*Status(1)= department gender TechKnowHow / SELECTION=stepwise;
run;
```

Class Level Information					
Class	Value	Design Variables			
Department	ADMINSTRATION	0	0	0	0
	MARKETING	1	0	0	0
	SALES_ENGINEER	0	1	0	0
	SALES_REP	0	0	1	0
	TECH_SUPPORT	0	0	0	1
Gender	F	0			
	M	1			
TechKnowHow	NO	0			
	YES	1			

Parameter		DF	Parameter Estimate
Department	MARKETING	1	-0.50141
Department	SALES_ENGINEER	1	1.47708
Department	SALES_REP	1	1.28348
Department	TECH_SUPPORT	1	1.00944
TechKnowHow	YES	1	-1.26948

ADMINISTRATION = ?

ADMINISTRATION = 0

(the reference category)

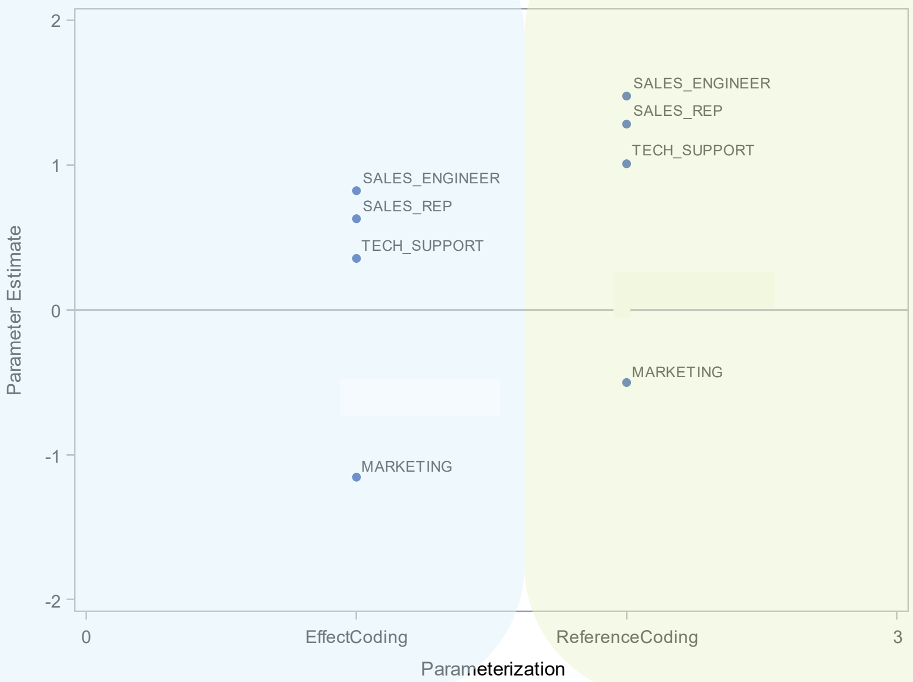
Comparing the EFFECT and REFERENCE Coding

```
proc phreg data=employees;
 CLASS department gender TechKnowHow / PARAM=effect REF=first;
 MODEL Duration*Status(1)= department gender TechKnowHow / SELECTION=stepwise;
run;
```

Class Level Information					
Class	Value	Design Variables			
Department	ADMINSTRATION	-1	-1	-1	-1
	MARKETING	1	0	0	0
	SALES_ENGINEER	0	1	0	0
	SALES_REP	0	0	1	0
	TECH_SUPPORT	0	0	0	1
Gender	F	-1			
	M	1			
TechKnowHow	NO	-1			
	YES	1			

Parameter		DF	Parameter Estimate
Department	MARKETING	1	-1.15513
Department	SALES_ENGINEER	1	0.82336
Department	SALES_REP	1	0.62976
Department	TECH_SUPPORT	1	0.35572
TechKnowHow	YES	1	-0.63474

ADMINISTRATION = -0.654




```
proc phreg data=employees;
 CLASS department gender TechKnowHow / PARAM=reference REF=first;
 MODEL Duration*Status(1)= department gender TechKnowHow / SELECTION=stepwise;
run;
```

Class Level Information					
Class	Value	Design Variables			
Department	ADMINSTRATION	0	0	0	0
	MARKETING	1	0	0	0
	SALES_ENGINEER	0	1	0	0
	SALES_REP	0	0	1	0
	TECH_SUPPORT	0	0	0	1
Gender	F	0			
	M	1			
TechKnowHow	NO	0			
	YES	1			

Parameter		DF	Parameter Estimate
Department	MARKETING	1	-0.50141
Department	SALES_ENGINEER	1	1.47708
Department	SALES_REP	1	1.28348
Department	TECH_SUPPORT	1	1.00944
TechKnowHow	YES	1	-1.26948

ADMINISTRATION = 0

How can we calculate the the (hidden) value of the reference category in effect coding

Analysis of Maximum Likelihood Estimates							
Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
Department	MARKETING	1	-1.15513	0.47794	5.8414	0.0157	0.606
Department	SALES_ENGINEER	1	0.82336	0.52244	2.4838	0.1150	4.380
Department	SALES_REP	1	0.62976	0.29224	4.6436	0.0312	3.609
Department	TECH_SUPPORT	1	0.35572	0.29940	1.4117	0.2348	2.744
TechKnowHow	YES	1	-0.63474	0.27370	5.3781	0.0204	0.281

$$-[\underbrace{(-1.155)+0.823+0.630+0.356}] = -0.654$$

```
PROC PHREG DATA=Employees;
 CLASS department gender TechKnowHow / PARAM=effect REF=first;
 MODEL Duration*Status(1)= department gender TechKnowHow /
                          SELECTION=stepwise;
RUN;
```

How can we automate this calculation?

https://github.com/gerhard1050/Applying-Data-Science-Using-SAS

Macro Parameters

The following parameters can be specified with the macro.

ParmEst

The name of the data set that contains the ParameterEstimates, created with the ODS OUTPUT statement. Default = ParameterEstimates.

ClassLevels

The name of the data set that contains the ClassLevelInfo, created with the ODS OUTPUT statement. Default = ClassLevelInfo.

OutputDS

The name of the data set that shall contain the output data set. Default = _ParmEst_XT_.

```
%CALC_REFERENCE_CATEGORY( ParmEst     = ParameterEstimates1,
                          ClassLevels = ClassLevelInfo1);
```

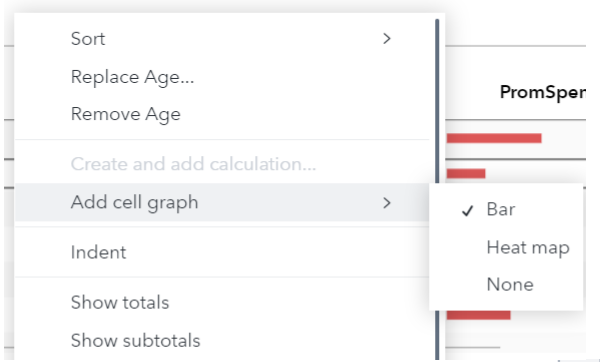
The output format of `ParmEst` and `ClassLevels` varies between different regression procedures in SAS. Please contact the author (Email: sastools.by.gerhard@gmx.net) in case your output file does not match the requirements of the macro.

Bonus Tip:

Use sparklines and bars to illustrate properties of customer segments and clusters

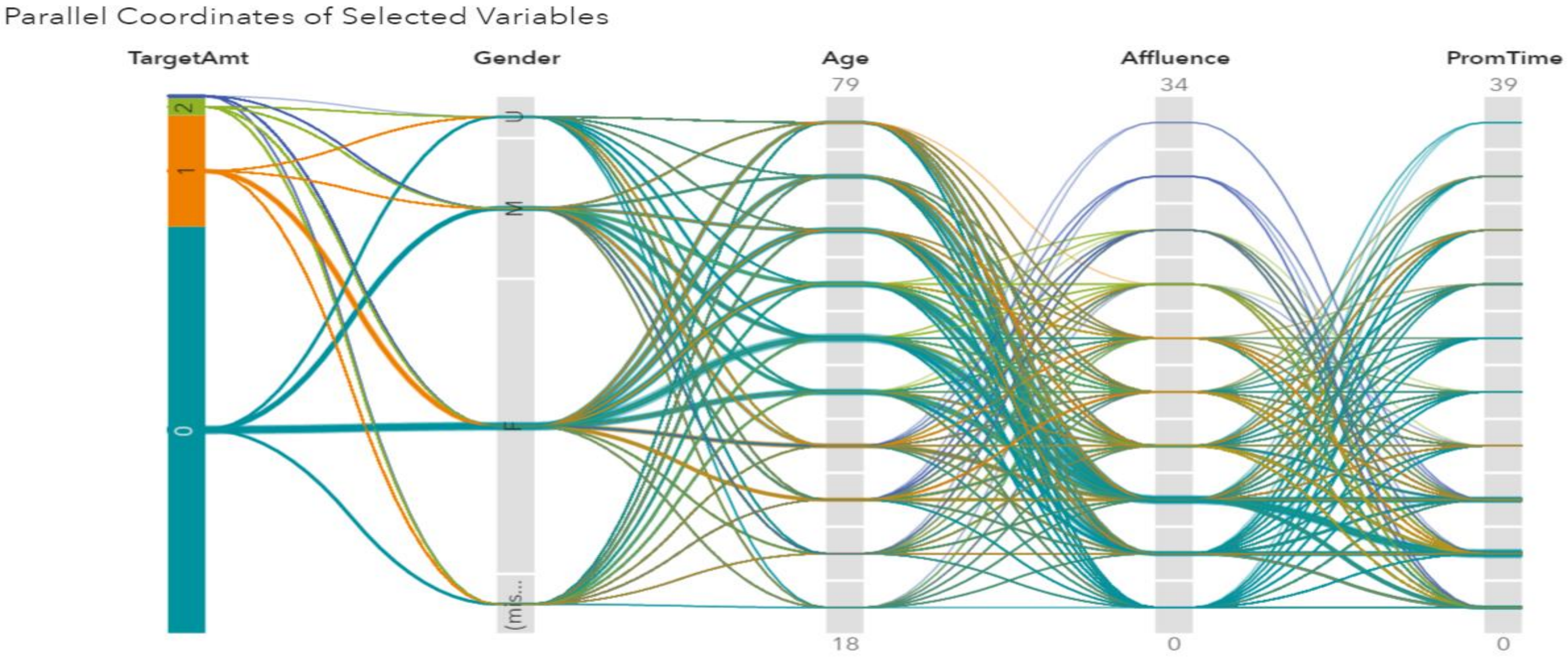
Cluster Profiling

- Create a crosstab in SAS Visual Analytics
- Add barcharts to illustrate the values

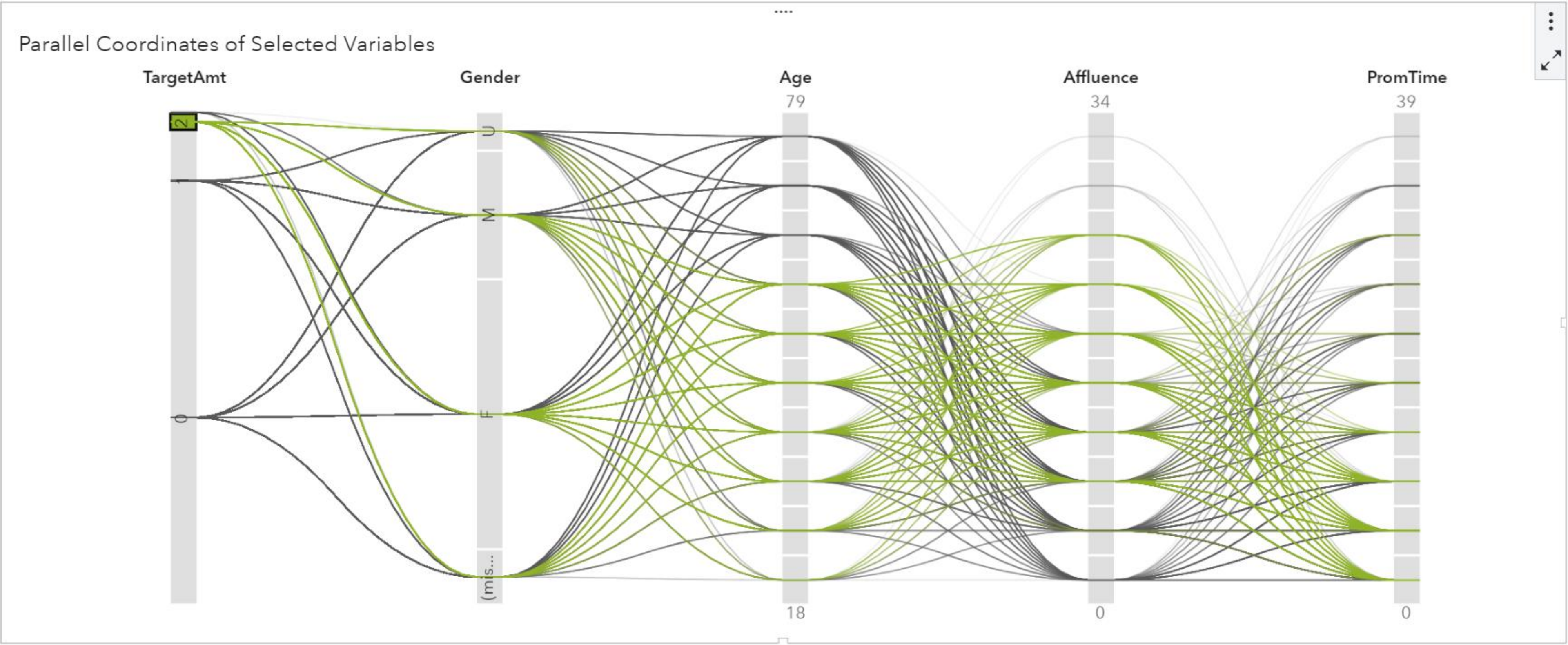


ClusterID	Frequency	Frequency Percent	Age	Female	Affluence	PromSpend	PromTime	Response%
5	19,463	19.42%	63	0.55	11	6071	6	21.60%
4	29,902	29.84%	42	0.56	8	2475	6	28.52%
3	9,459	9.44%	42	0.67	15	2588	6	66.48%
2	24,192	24.14%	63	0.50	6	5798	6	7.71%
1	4,727	4.72%	67	0.51	8	7525	21	14.13%

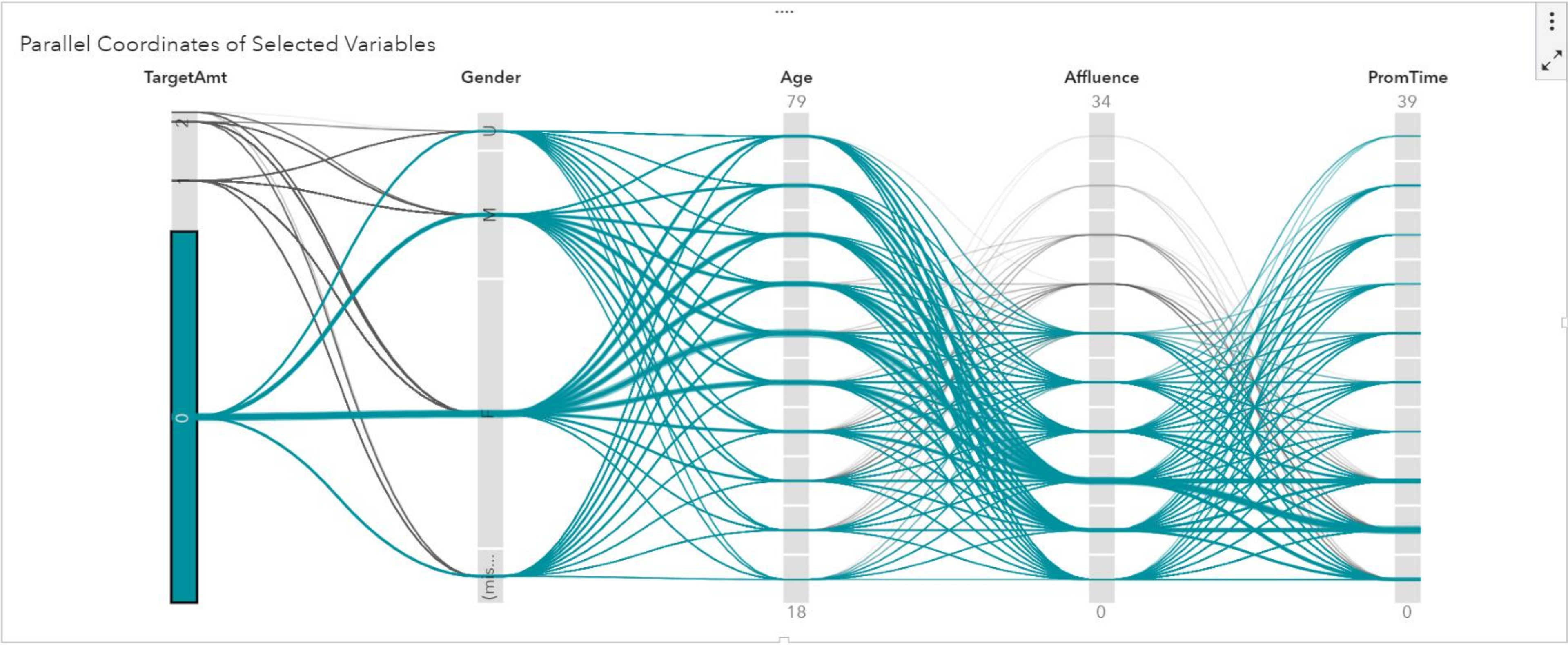
Use a Parallel Coordinate plot to illustrate cluster features

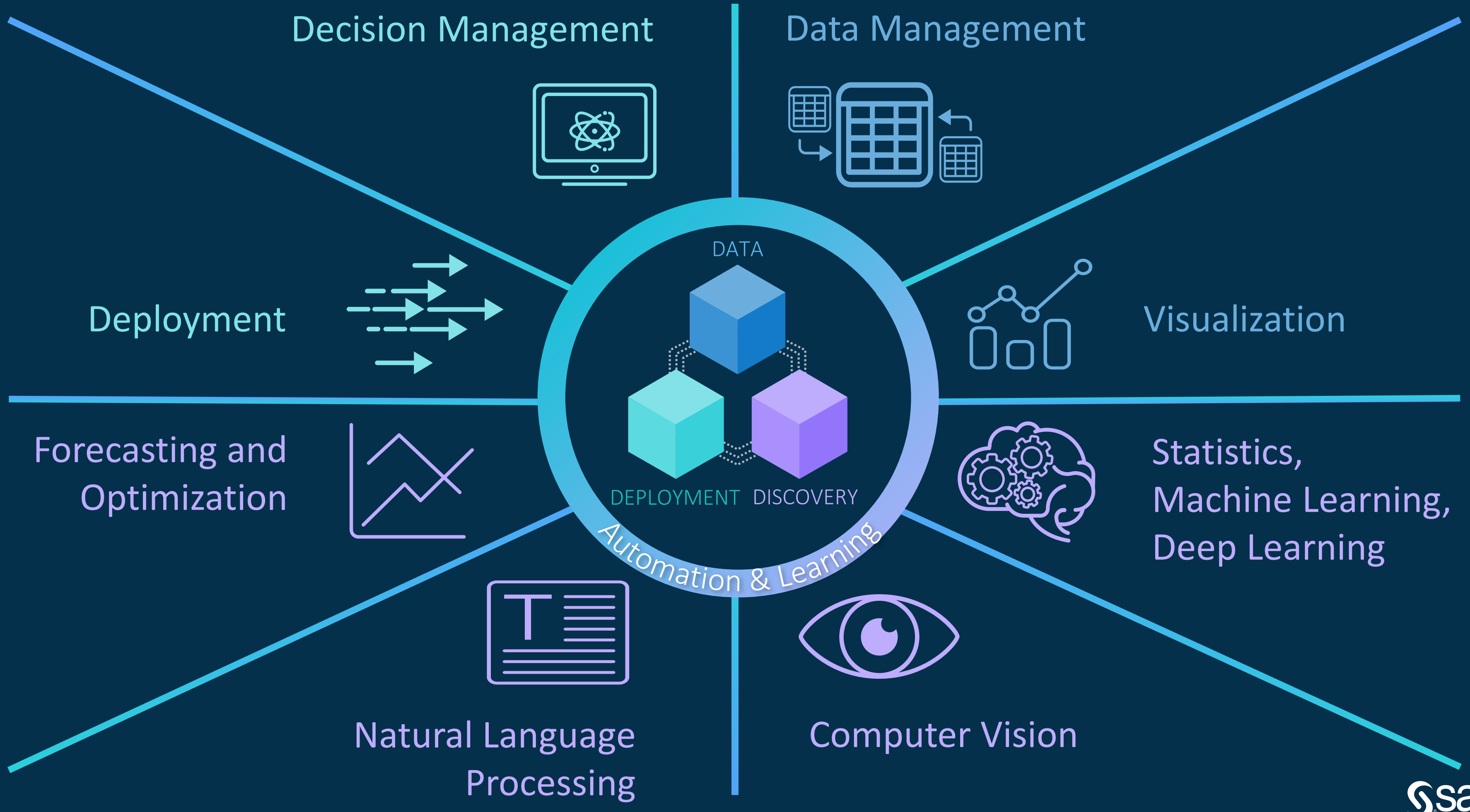


More affluent customers buy more often



Non-Buyers are older and less affluent




Decision Management
Data Management
Deployment
Visualization
Forecasting and Optimization
Statistics, Machine Learning, Deep Learning
Natural Language Processing
Computer Vision
DATA
DEPLOYMENT
DISCOVERY
Automation & Learning

Conclusion

- Communicating model results has many dimensions
- SAS Viya offers you a broad range of tools and methods to illustrate your findings
- Machine learning models that are understood are likely to have a higher business impact

Communicating Analytical Results and Interpreting your ML Models with SAS Viya – 5 Tips and Tricks that will make your life as data scientist easier

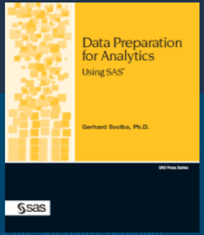
Gerhard Svolba

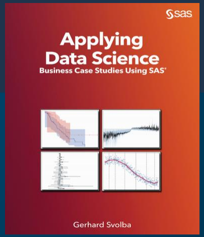
Analytic Solutions Architect

SAS Austria

Credits for Input to:
Martin Schütz, Tamara Fischer

Twitter: https://twitter.com/gsvolba
https://github.com/gerhard1050
https://www.linkedin.com/in/gerhardsvolba/





THE POWER TO KNOW.