Non-Predictive Use of Decision Tree and Friends

How supervised machine learning models can help you beyond the usual task of prediction and classification

Gerhard Svolba, Data Scientist, SAS Austria

Data Scientist @SAS - Medium | LinkedIn | Github | SAS-Books | SAS Articles
Youtube: DataPreparation4DataScience | Data Science Use Cases

#5
Building an expert model

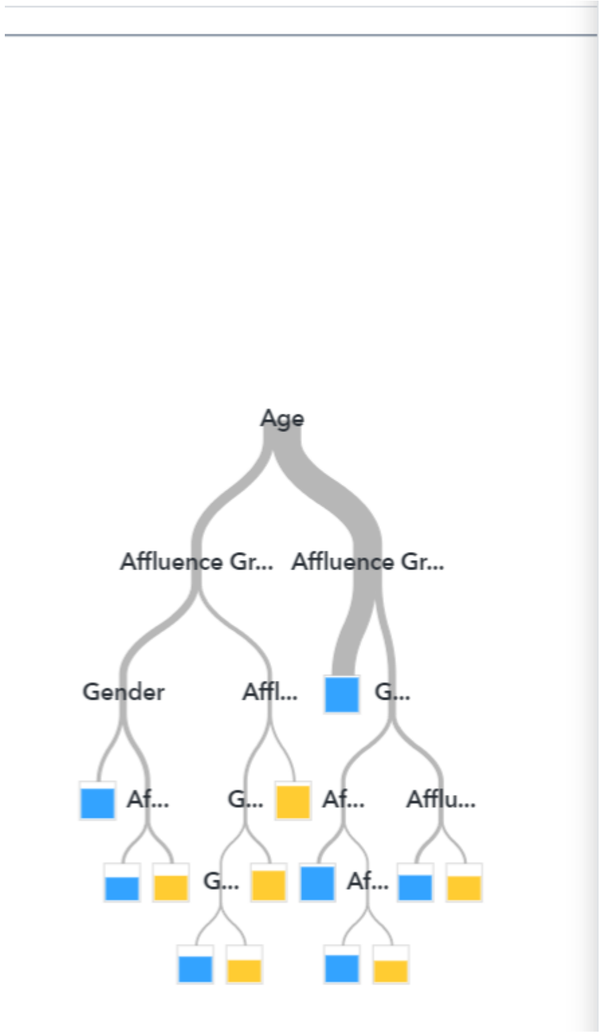
Basic Idea

- Use the “interactive tree” facilities for decision trees in SAS Visual Analytics to define a decision tree based on your business knowledge (expert model)

This answers the following questions

- How many responders do I have in individually defined business segments?
- What is the predictive power of my expert model?
- How does it compare to automated machine learning models?
- What is the “price” (in accuracy) that I have to pay for using my business segments/business rules, compared to using the analytical optimal split variables and split values

Step 1: Create a decision tree with the variables that you want to use in your expert model

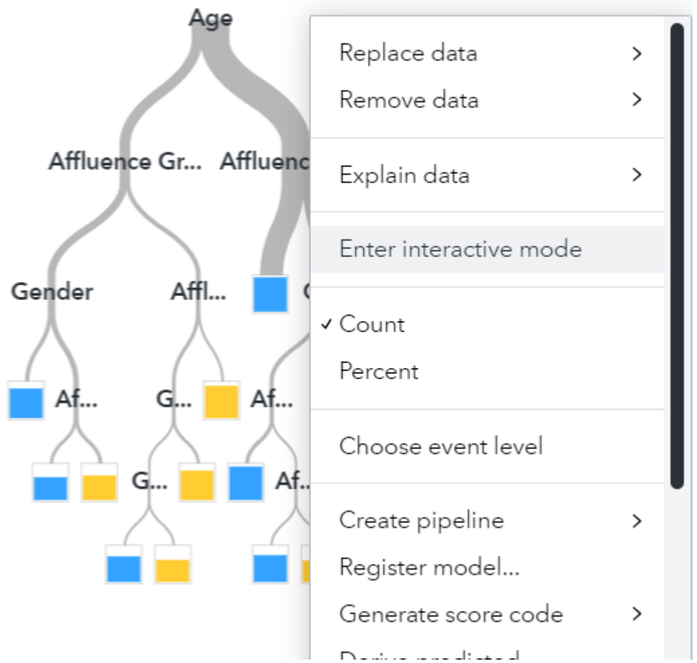


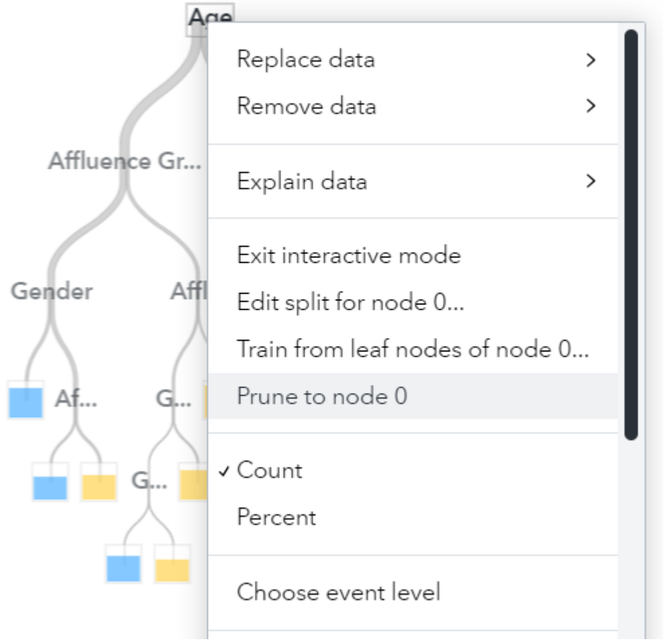
Decision tree - Organics Purchase Indicat...

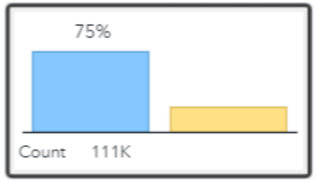
+ Assign data

- Response + Add
  - Organics Purchase Indicator
- Predictors + Add
  - Gender
  - Geographic Region
  - Loyalty Status
  - Age
  - Affluence Grade
  - Loyalty Card Tenure
  - Total Spend
- Partition ID + Add
- Frequency + Add
- Weight + Add

Step 2: Start the interactive mode and prune the tree to only retain the root node.

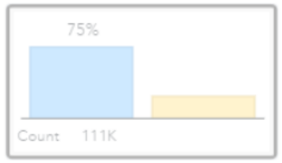




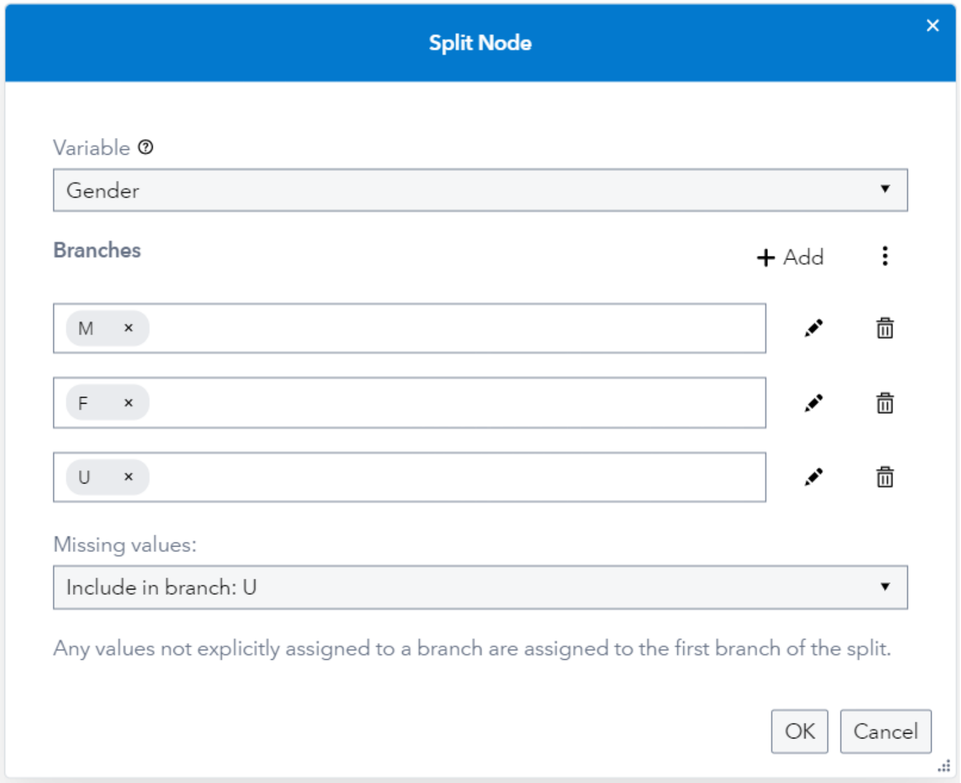


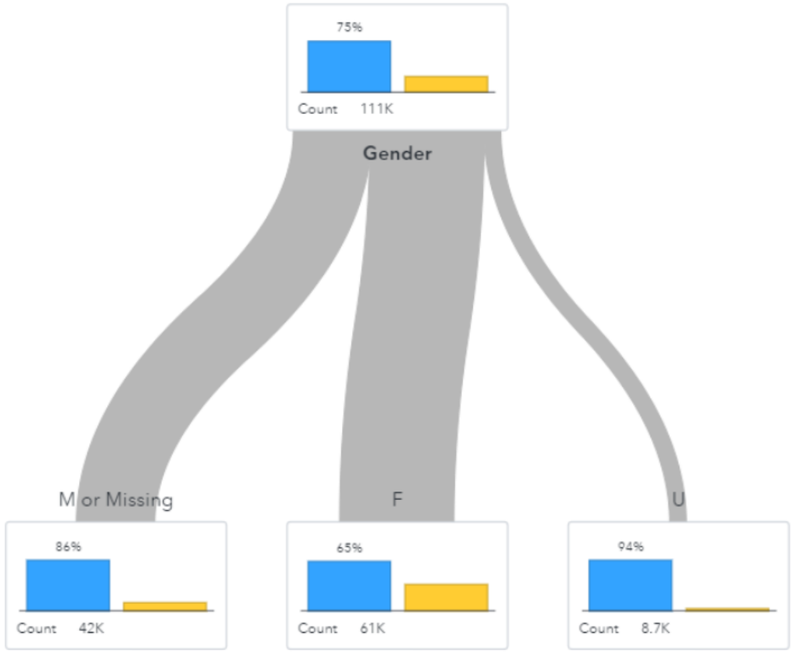
-(Pruned)-

Step 3: Select “Split node” to perform the splits of your choice



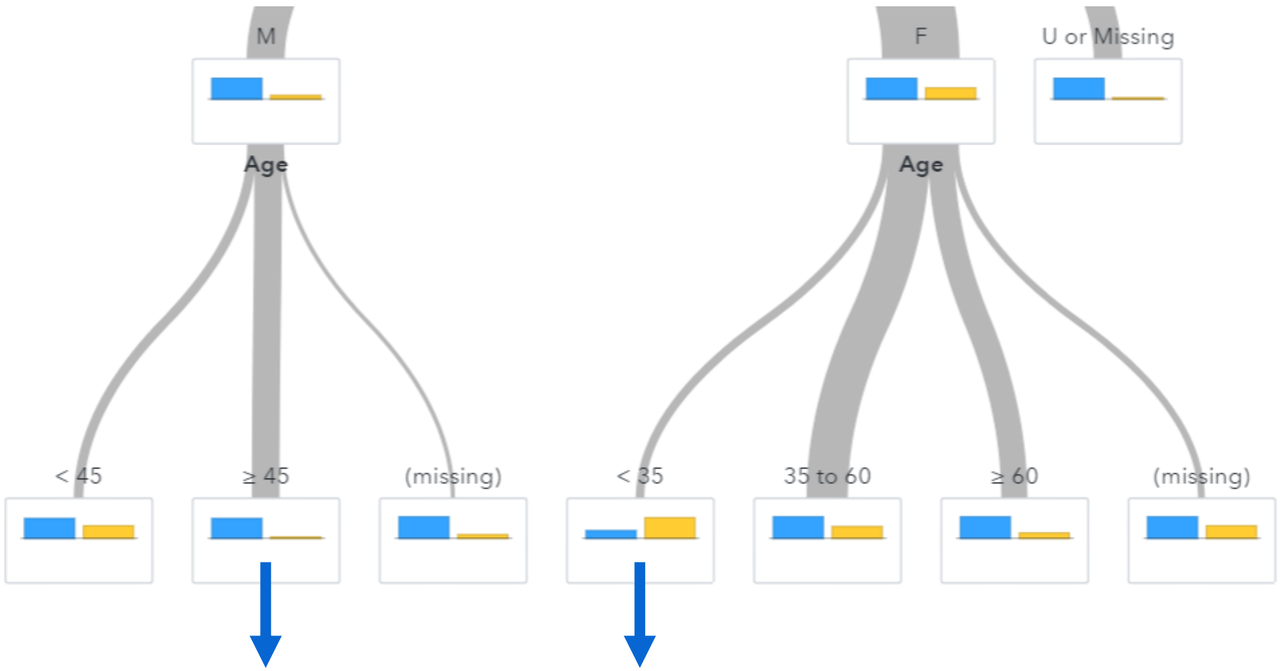
-(Pruned)-





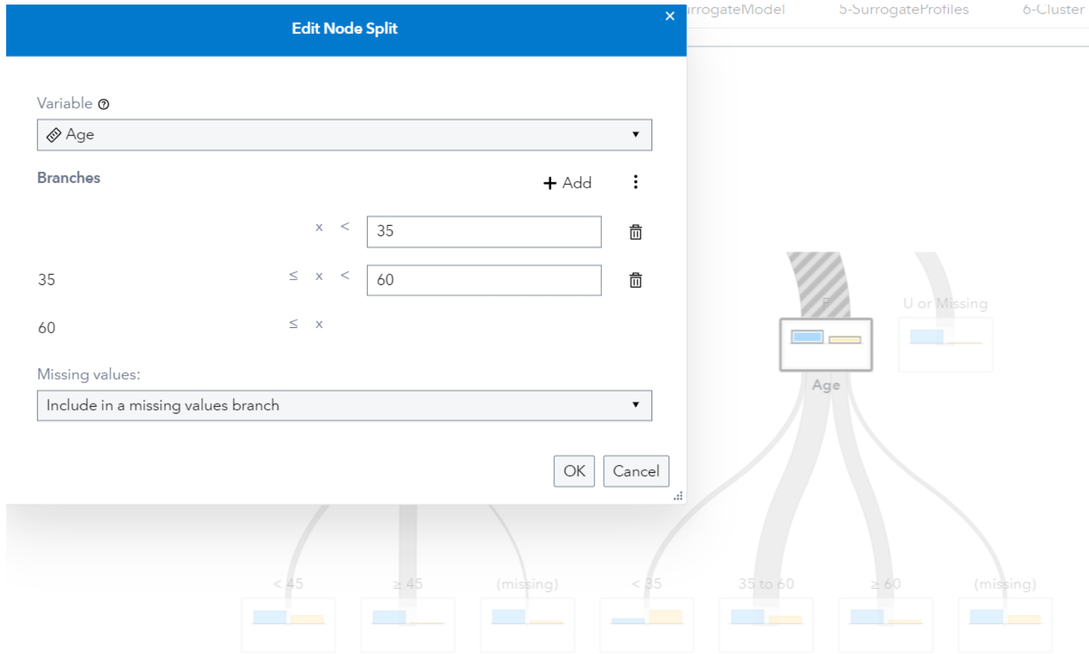
Organics Purchase Indicator
0 1

Step 4..n : Repeat this procedure until you have defined the tree according to your business rules

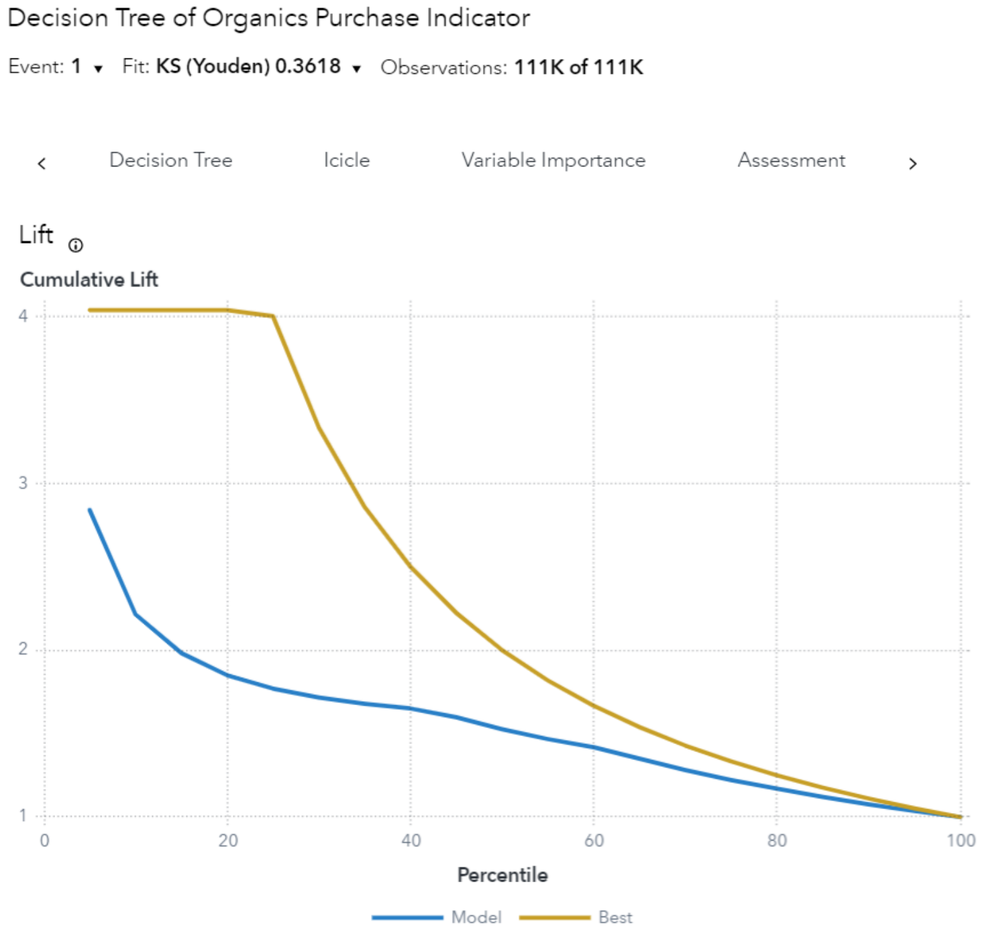


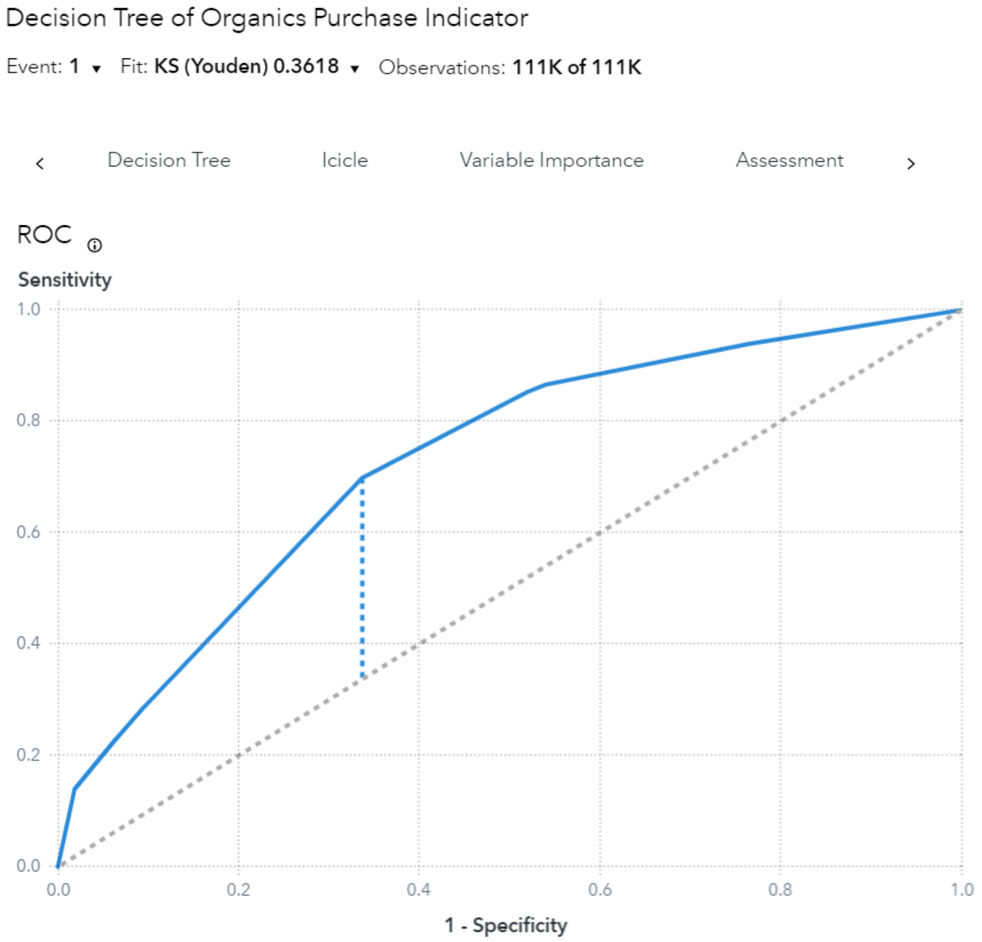
Node ID:	8
Gender:	M
Age:	≥ 45
Count:	20,810
0:	18,795 (90.32%)
1:	2,015 (9.68%)

Node ID:	4
Gender:	F
Age:	< 35
Count:	5,420
0:	1,565 (28.87%)
1:	3,855 (71.13%)



Step 5: (optionally) Review the lift chart or other assessment statistics to understand the predictive power of your expert model





Application Recommendations

- Preferred Method: Decision Tree
- Recommended SAS Tool: SAS Visual Analytics

#6
Calculate meaningful reference values (Check the plausibility of values and events)

Basic Idea

- Detecting anomalies and incorrect values
- Go beyond absolute values or averages per case
  → calculate expected values
- Use other attributes per person or per case to calculate individual (average) values
- Remove the “yes, … but …“ conversations
  - Yes, the values in region A are higher, but there are more patient with severe diseases
  - Yes, the values for these doctors are higher, but most of them are in specialization D, which usually has higher values
- Compare the actual value against the expected value and base your selection decision on these

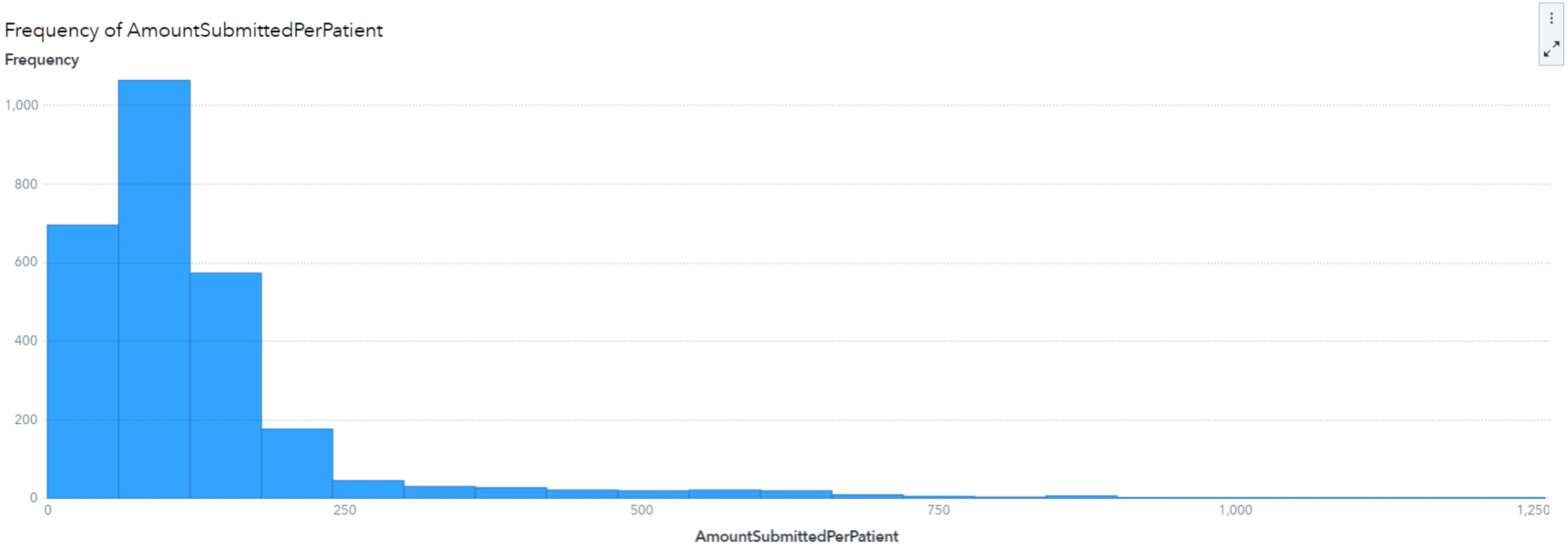
Data from medical service - Doctors send their invoices to public health insurance

Absolute Value

Average per Patient

MedUnitID ▲	NumberOfPatients	ContractSince	Specialisation	AdditionalTraining2	AdditionalTraining1	Region	AmountSubmitted	AmountSubmittedPerPatient
1780186	440	1993	Specialisation 6	NO	NO	4	6491	14.75
2051592	510	1991	Specialisation 1	YES	NO	4	60071	117.79
2718167	350	1992	Specialisation 2	YES	NO	7	17441	49.83
2723173	390	1994	Specialisation 2	NO	YES	15	59971	153.77
2885798	470	1990	Specialisation 1	NO	NO	7	53251	113.30
3116230	380	1984	Specialisation 2	YES	NO	7	68831	181.13
3136701	470	1997	Specialisation 5	NO	NO	12	37801	80.43
4747725	430	1997	Specialisation 5	NO	NO	7	39751	92.44
4747725	430	1991	Specialisation 3	NO	NO	7	14491	33.70
5103700	450	1986	Specialisation 5	NO	NO	1	291976	648.84
5268940	360	1996	Specialisation 5	NO	NO	4	43641	121.23
5919898	480	1992	Specialisation 6	NO	NO	6	40441	84.25
5941567	540	1995	Specialisation 3	NO	NO	7	233047	431.57
6889317	410	1991	Specialisation 1	YES	NO	7	86541	211.08
6993836	500	1993	Specialisation 3	YES	NO	1	54601	109.20
7884893	410	1992	Specialisation 3	YES	NO	4	59021	143.95
7945239	400	1991	Specialisation 3	NO	NO	9	64781	161.95
7994071	400	1991	Specialisation 3	NO	NO	7	40811	102.03
8166151	500	1979	Specialisation 5	NO	YES	9	69081	138.16
8201951	470	1991	Specialisation 1	NO	NO	7	91061	193.75
8467576	360	1989	Specialisation 3	YES	NO	4	51191	142.20
8467576	390	1996	Specialisation 3	NO	NO	3	52071	133.52
8563621	530	1988	Specialisation 6	YES	NO	12	51891	97.91
8565171	410	1995	Specialisation 3	YES	NO	12	131064	319.67
8649966	390	1996	Specialisation 2	YES	NO	7	11865	30.42
8764588	440	1992	Specialisation 4	NO	YES	8	43381	98.59

Trying to choose limits for raising an alert, triggering a request for documents, …



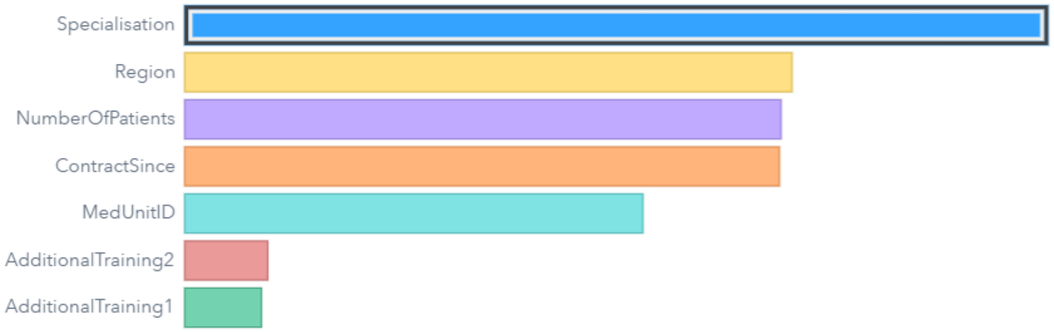
Step 1: Understand the relationships of explanatory variables on the submitted values (e.g. with the “Automated Explainer”)

What are the characteristics of AmountSubmitted?

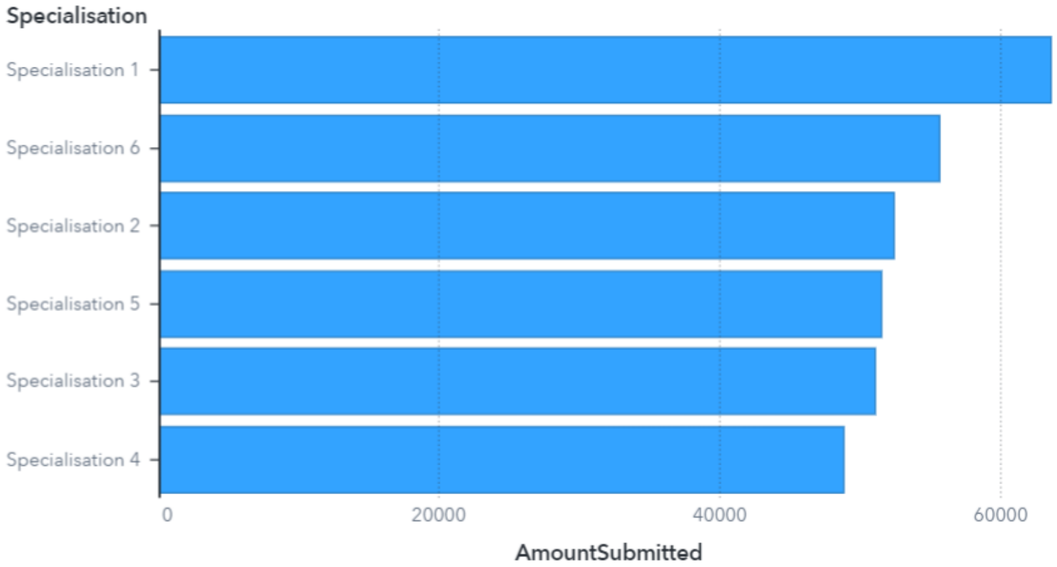
AmountSubmitted ranges from 314 to 446K. Average AmountSubmitted is 53K. Most cases (the middle 80%) have an AmountSubmitted between 15K and 81K. Region best differentiates th e highest (top 10%) and the lowest (bottom 10%) AmountSubmitted cases. The three most related factors are Specialisation, Region, and NumberOfPatien

ts. There are 189 cases that might be outliers, with AmountSubmitted greater than or equal to 105K.

What factors are most related to AmountSubmitted?

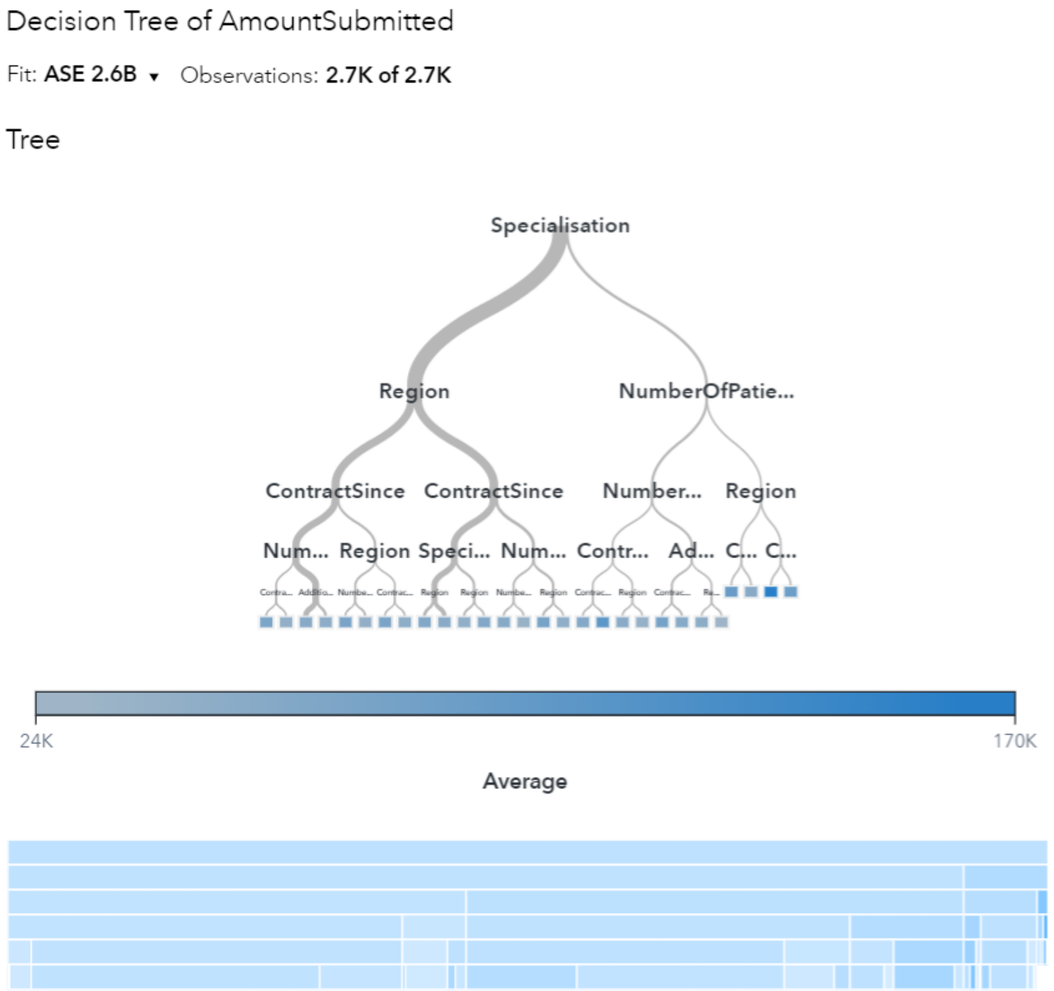


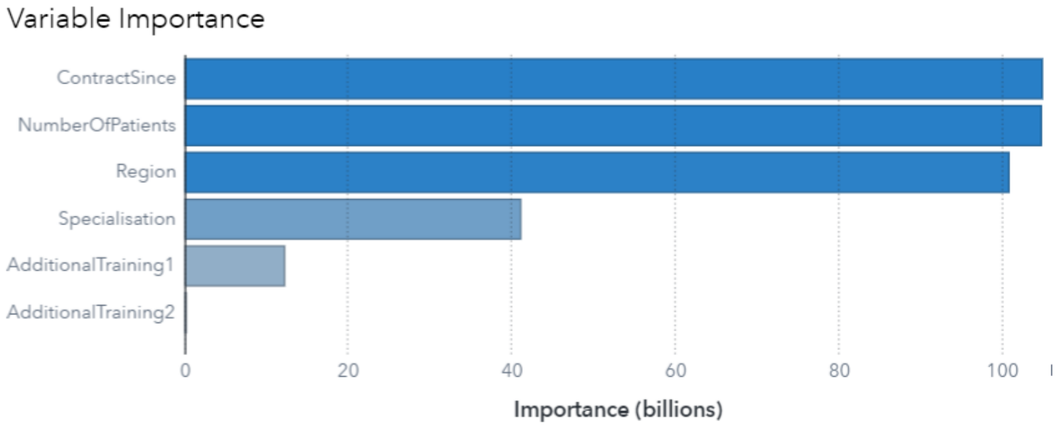
What is the relationship between AmountSubmitted and Specialisation?

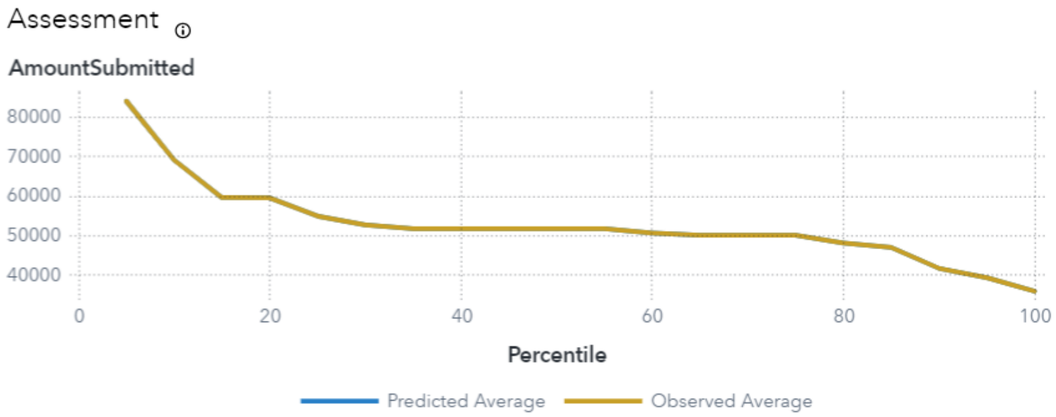


When Specialisation is Specialisation 1, the average of AmountSubmitted is a high value. When Specialisation is Specialisation 2, Specialisation 5, Specialisation 3 or Specialisation 4 , the average of AmountSubmitted is a low value. The most common Specialisation value is Specialisation5.

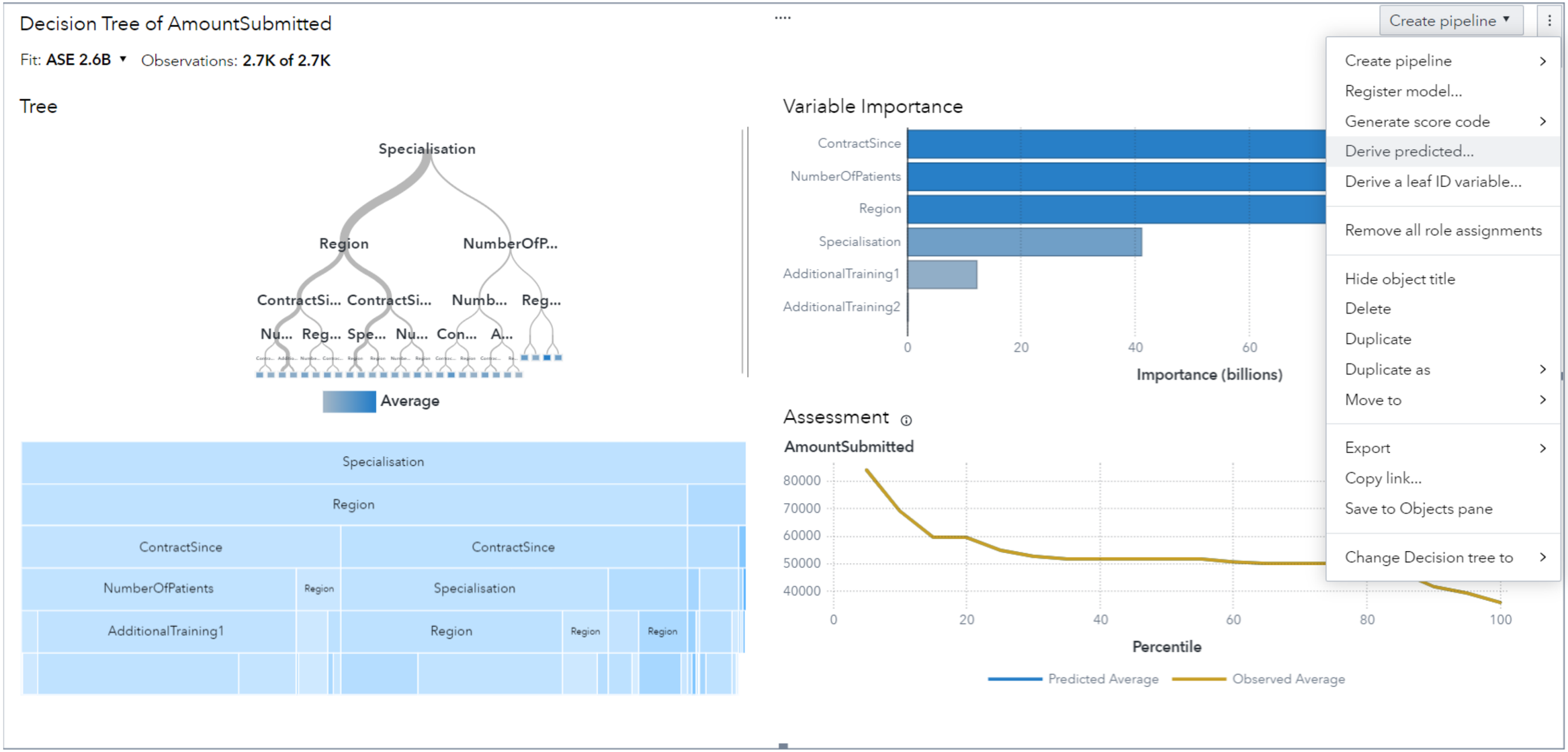
Step 2: Create a decision tree (or other model) to model the relationship between the available features and the submitted amount







Step 3: Derive the predicted values from these models and calculate the difference “actual-expected”



Step 4: Use this variable for the selection of cases for further investigation. E.g. by providing filters: Specialization 2 and Difference > 108,000

DifferenceFromExpectation

108,792 to 386,869

-157,089 386,869

Specialisation

Specialisation 1 | Specialisation 2 | **Specialisation 3** | Specialisation 4 | Specialisation 5 | Specialisation 6

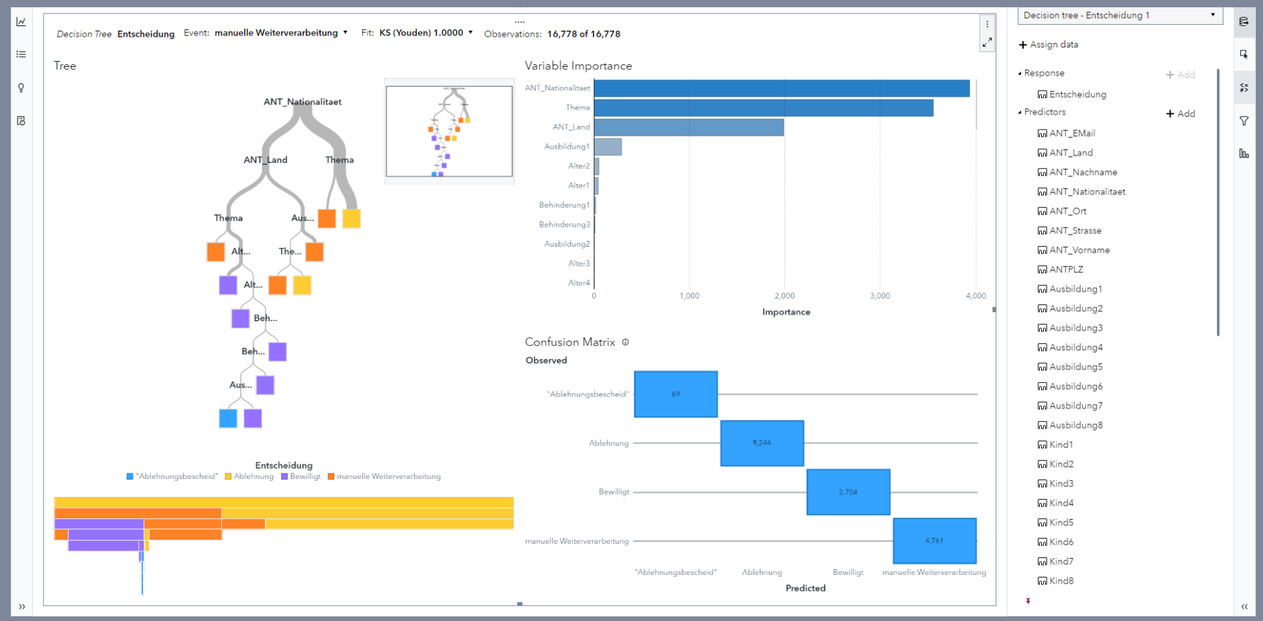
Absolute Value | Average per Patient | Expected Value | Difference Actual-Expect

MedUnitID ▲	NumberOfPatients	ContractSince	Specialisation	AdditionalTraining2	AdditionalTraining1	Region	AmountSubmitted	AmountSubmittedPerPatient	ExpectedAmountSubmitted	DifferenceFromExpectation
5941567	540	1995	Specialisation 3	NO	NO	7	233047	431.57	51,699	181,348
22460980	580	1987	Specialisation 3	NO	NO	10	257415	443.82	51,699	205,716
65426727	250	1993	Specialisation 3	YES	NO	3	298361	1,193.44	59,627	238,734
79590937	540	1989	Specialisation 3	NO	NO	9	307595	569.62	50,102	257,493
85768488	400	1990	Specialisation 3	NO	NO	7	250653	626.63	51,699	198,954
118841702	430	1993	Specialisation 3	NO	NO	9	197456	459.20	50,102	147,354
315755344	420	1996	Specialisation 3	NO	NO	6	254994	607.13	72,176	182,818
322553627	590	1989	Specialisation 3	NO	NO	9	290385	492.18	50,102	240,283
388940769	390	1998	Specialisation 3	YES	NO	6	268606	688.73	72,176	196,430
412968231	260	1988	Specialisation 3	YES	NO	9	190900	734.23	50,102	140,798
457780259	330	1994	Specialisation 3	NO	NO	1	222857	675.32	51,699	171,158
484814016	320	1985	Specialisation 3	NO	NO	3	355817	1,111.93	59,627	296,190
521214018	490	1985	Specialisation 3	YES	NO	7	255425	521.28	51,699	203,726
587383476	390	1996	Specialisation 3	NO	NO	1	189610	486.18	51,699	137,911
602299917	490	1990	Specialisation 3	YES	NO	1	267183	545.27	51,699	215,484
612724512	410	1995	Specialisation 3	NO	NO	6	270021	658.59	59,627	210,394
680069619	500	1997	Specialisation 3	NO	NO	12	252489	504.98	72,176	180,313
910285542	420	1994	Specialisation 3	NO	NO	1	300067	714.45	51,699	248,368
927875675	640	1996	Specialisation 3	NO	NO	12	264077	412.62	72,176	191,901

Application Recommendations

- Preferred Method: Decision Tree, Linear and Poisson Regression
- Recommended SAS Tool: SAS Code and SAS Visual Analytics

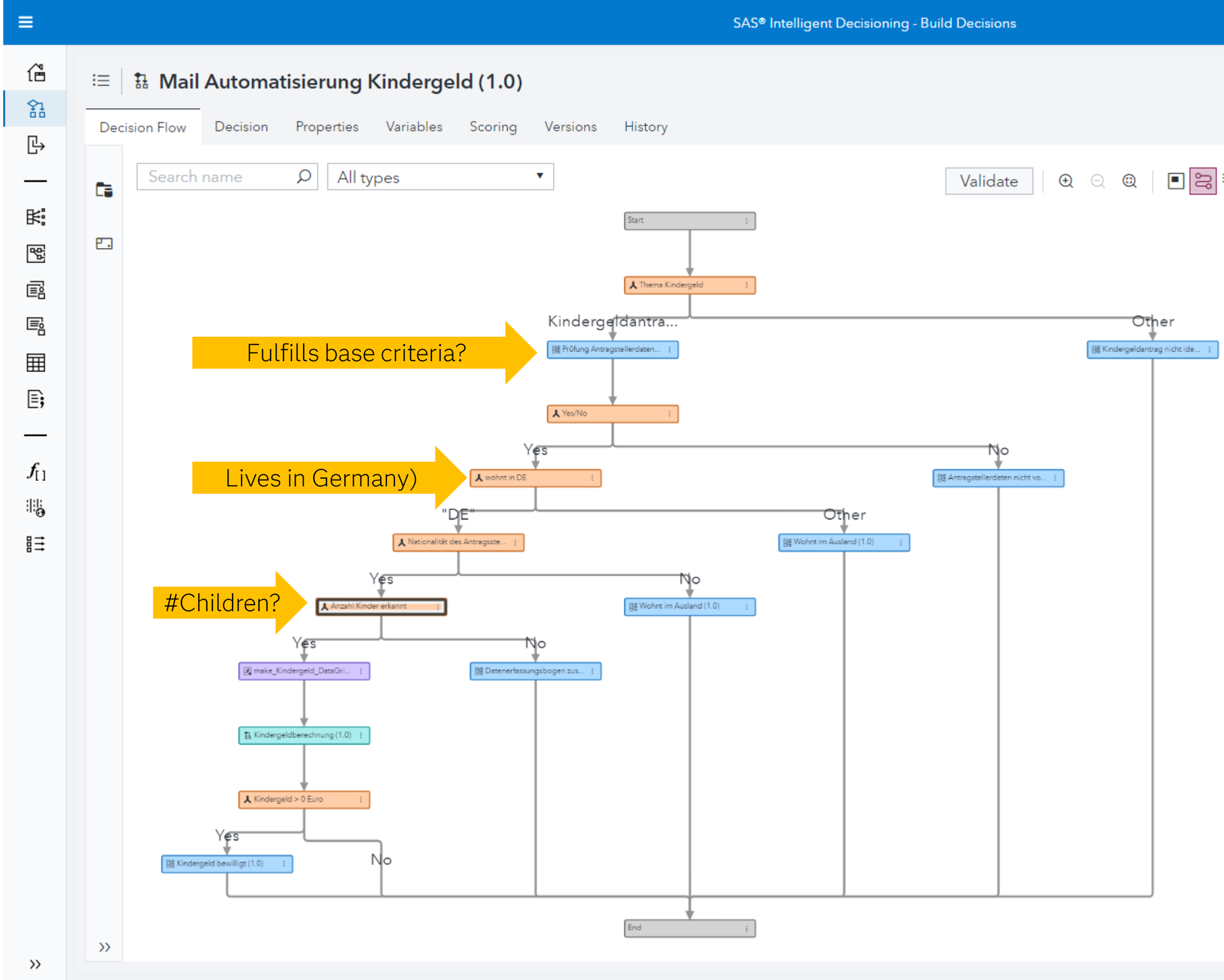
#7
Validating/Auditing existing business rules



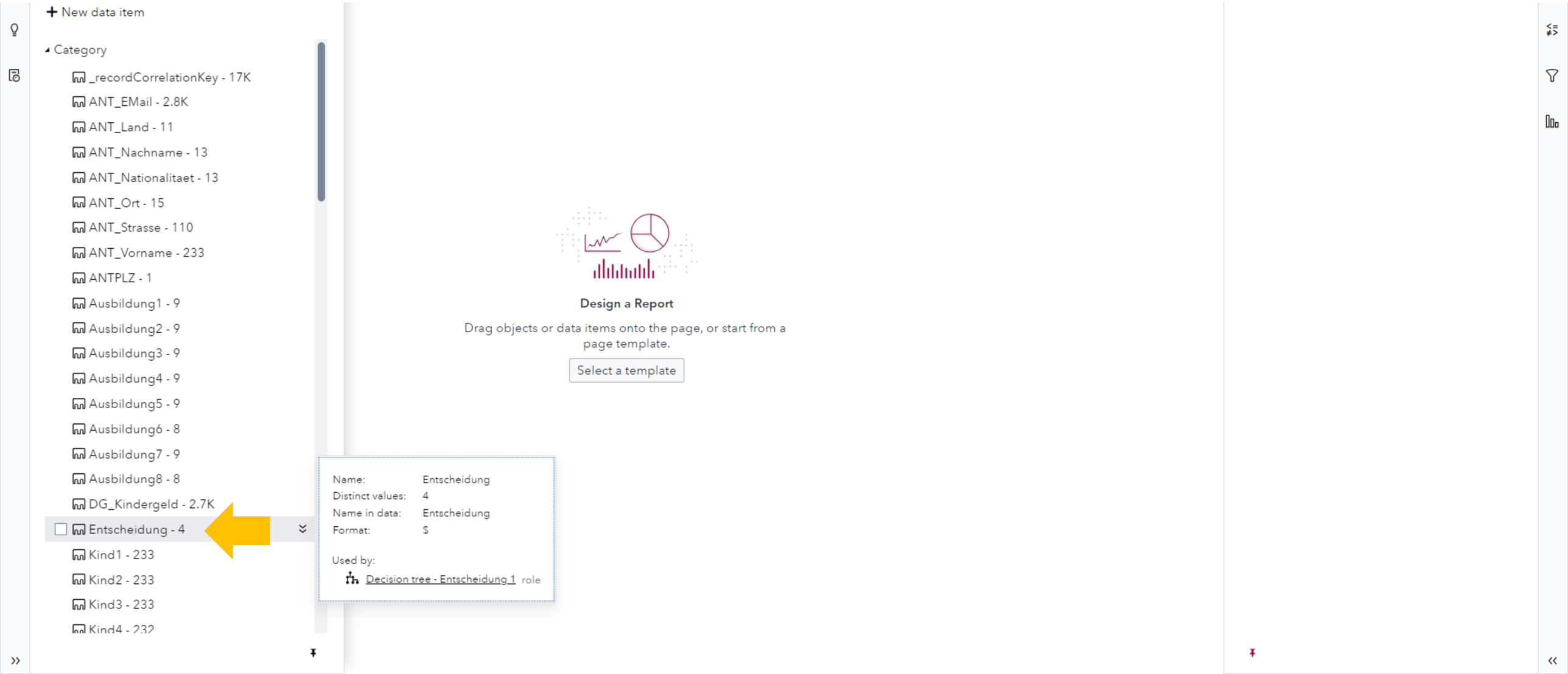
Credits to Ulrich Reincke, SAS Germany

Basic Idea

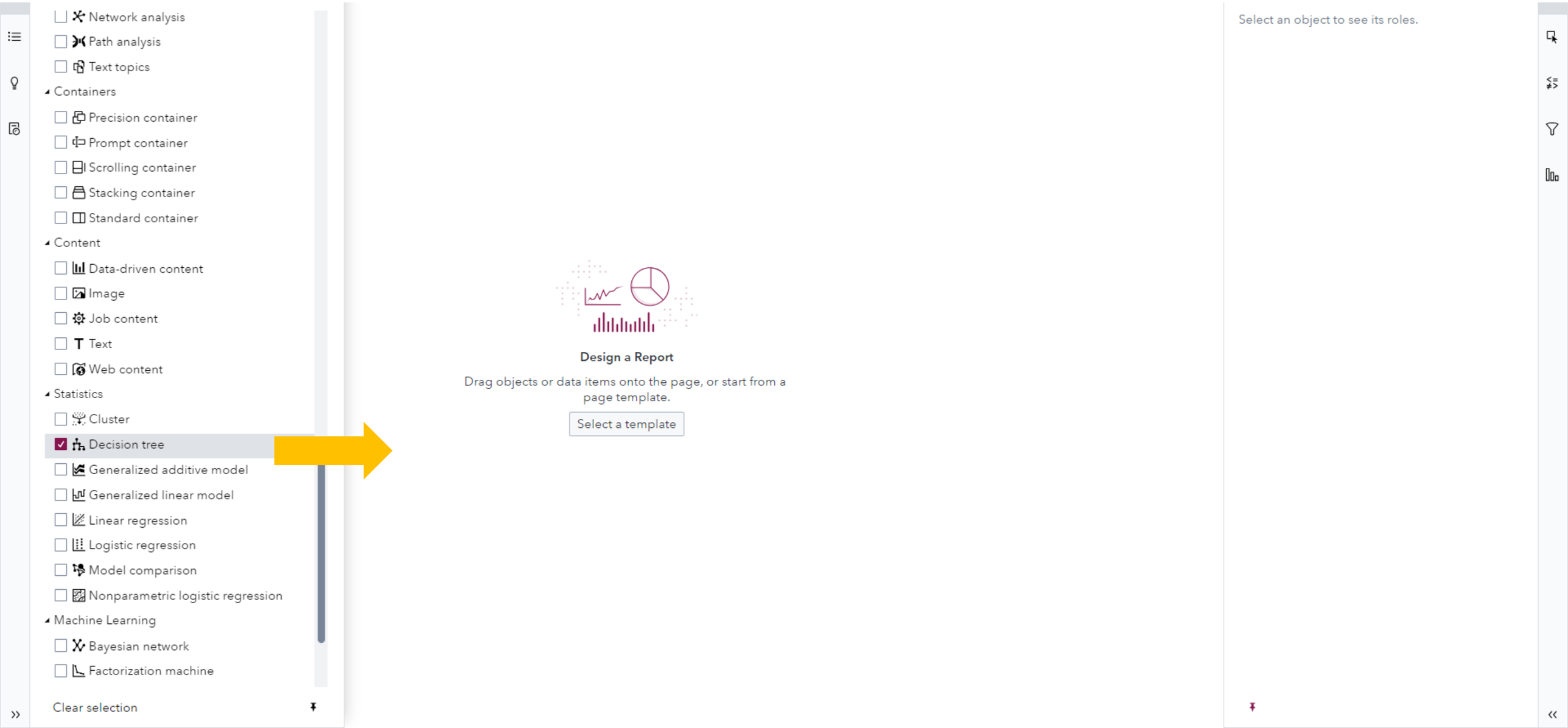
- Rules for the granting of child benefits are used in the business process
- The application data and the outcome (decision) are documented
- Use a decision tree to explain the outcome based on the application data
  - (Should) reflect the rules that are in place
  - Shows where overrides and exceptions have been made



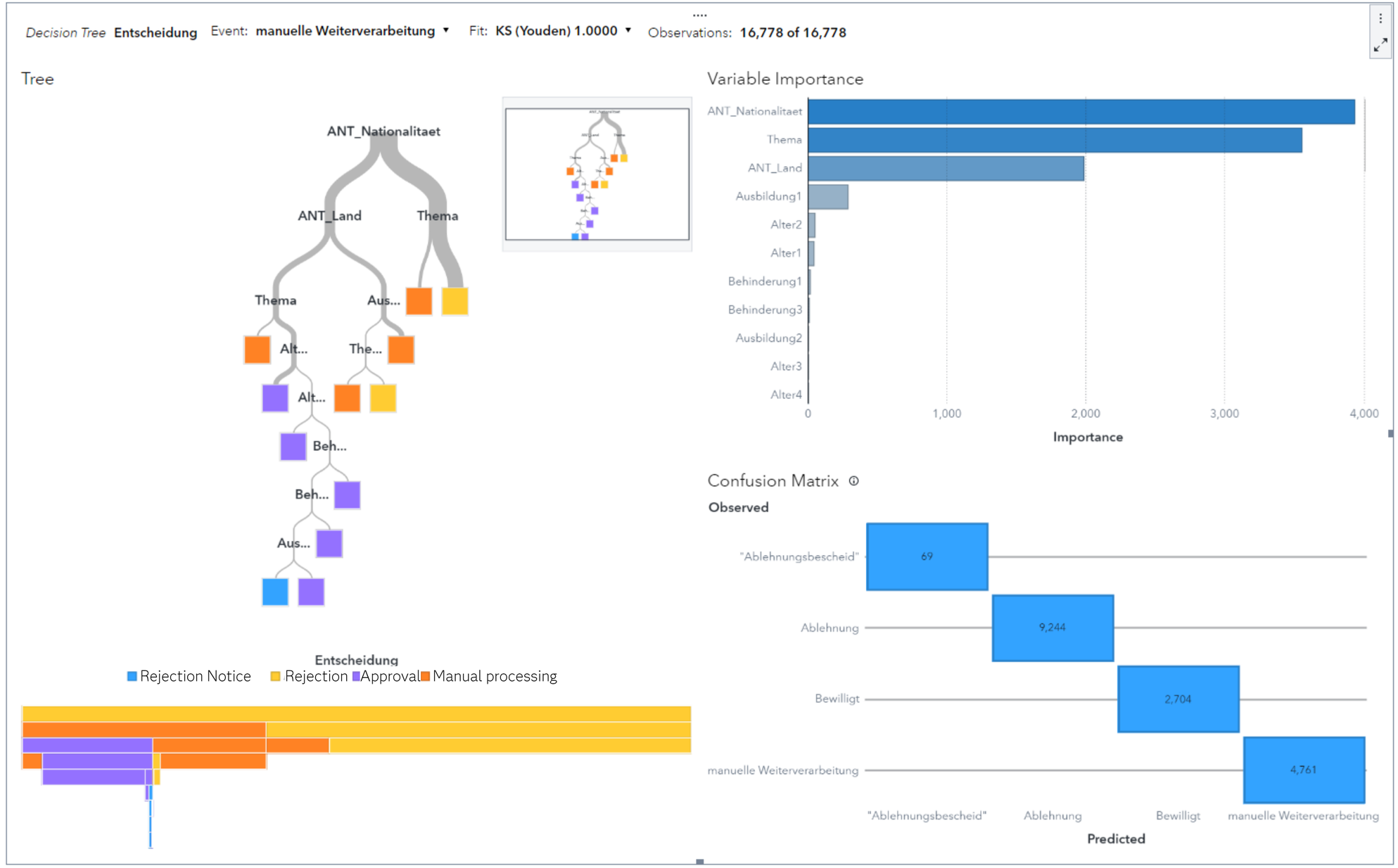
Explain the outcome variable using a decision tree

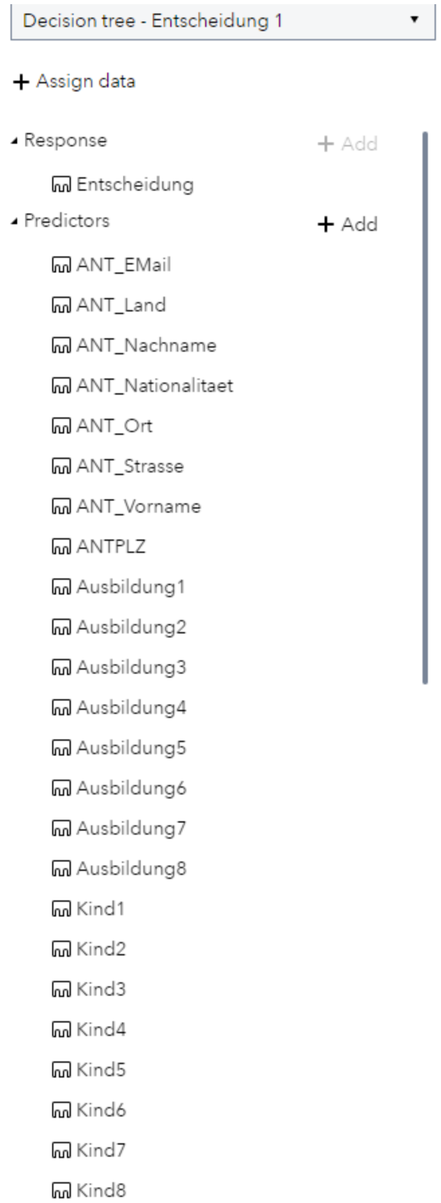


Explain the outcome variable using a decision tree

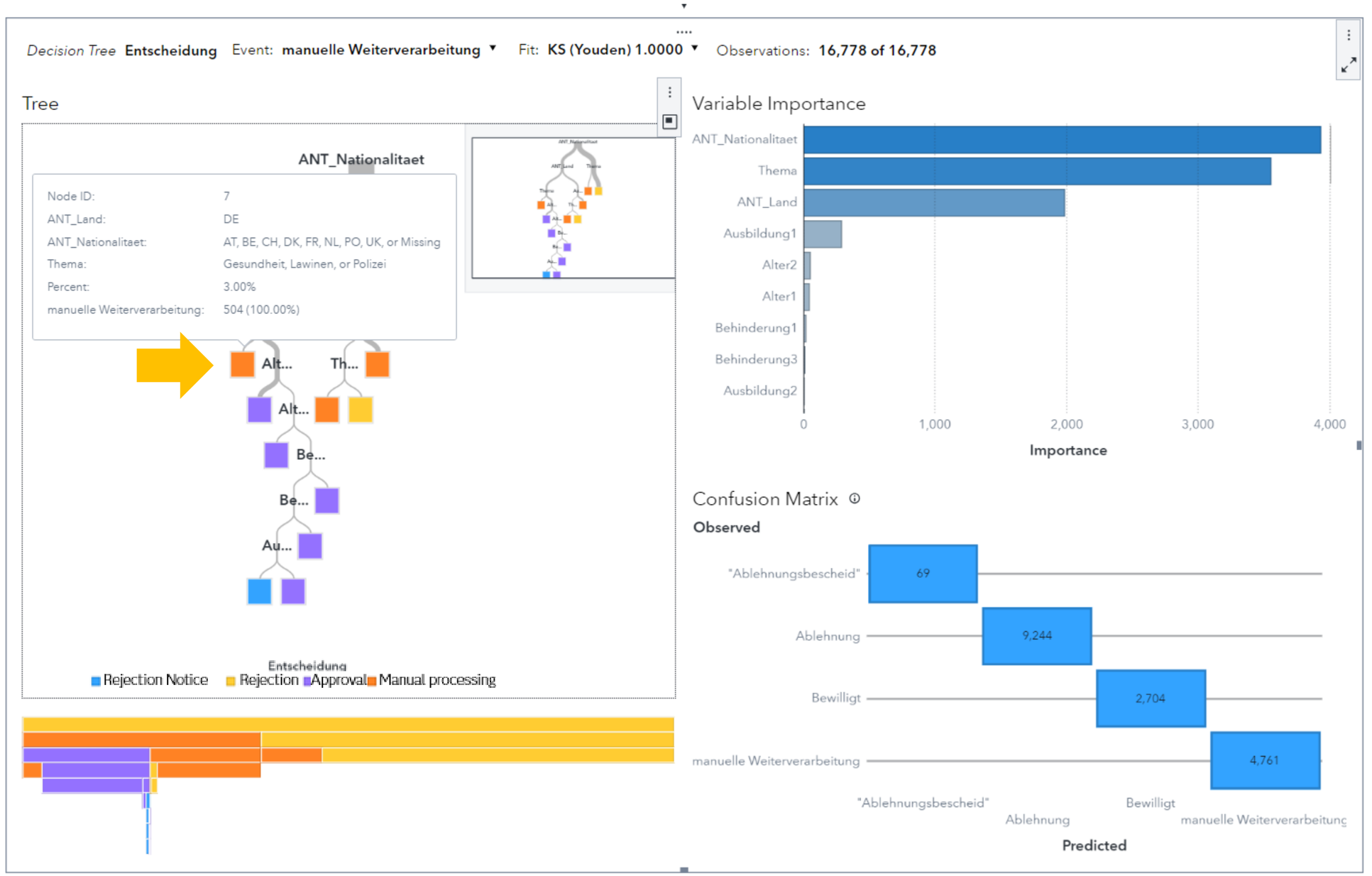


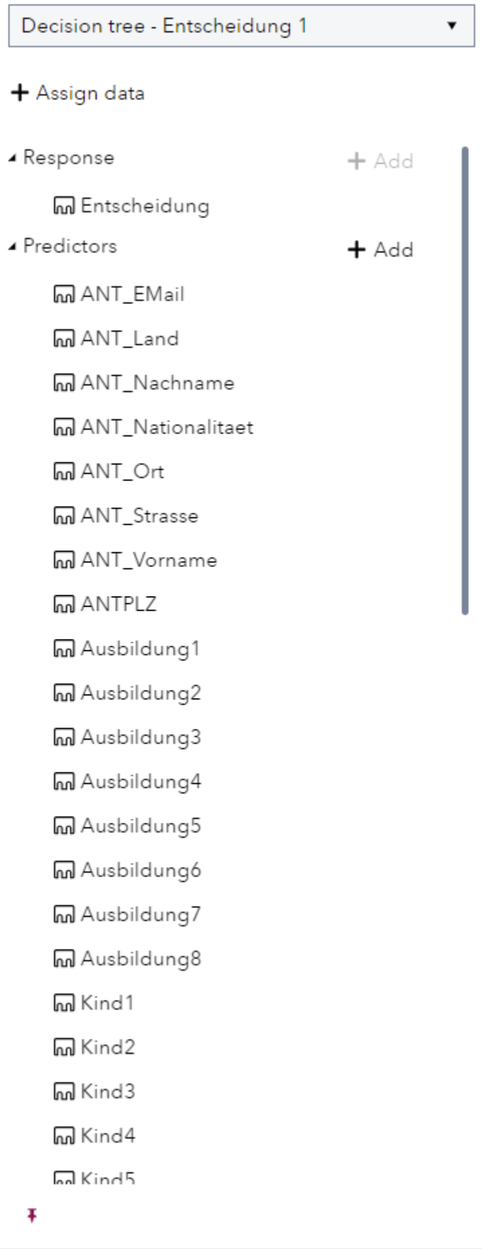
Decision Tree that mirrors the operational process



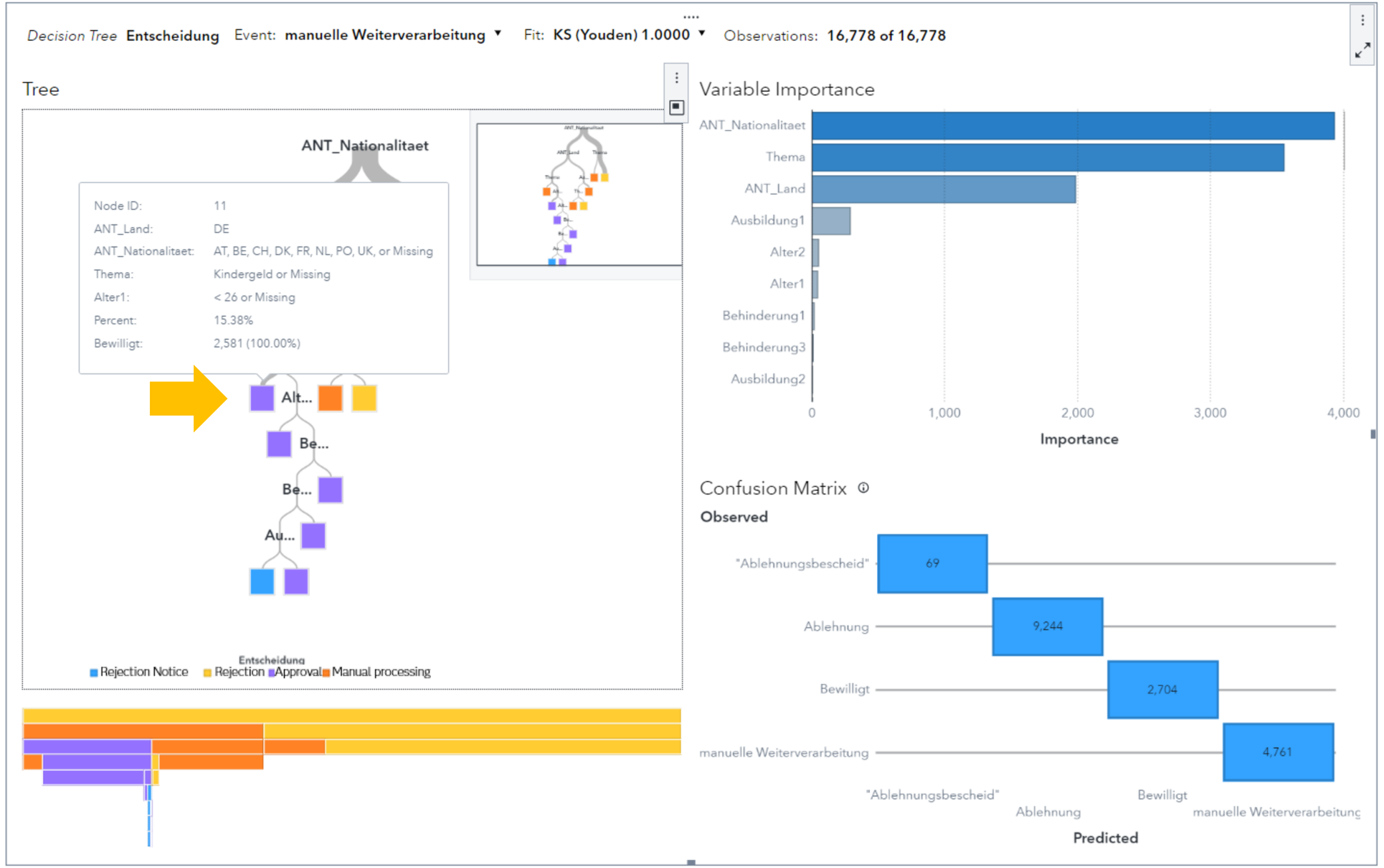


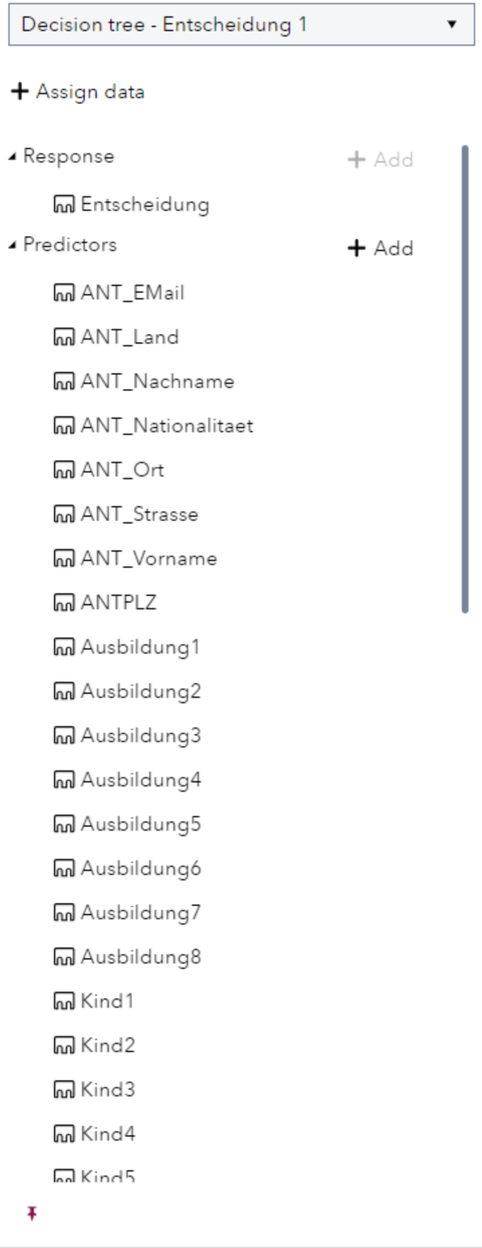
Reviewing different paths in the decision tree



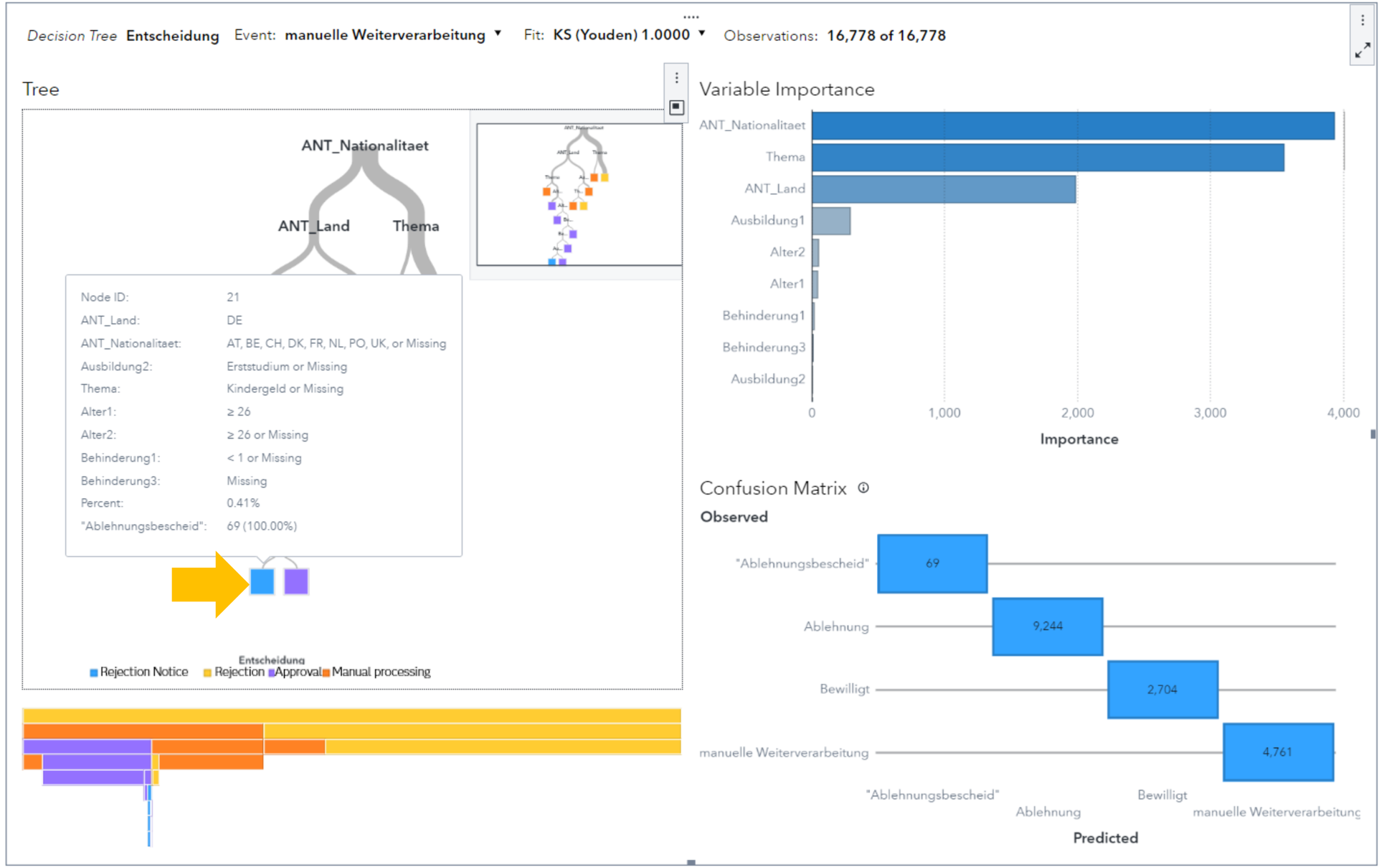


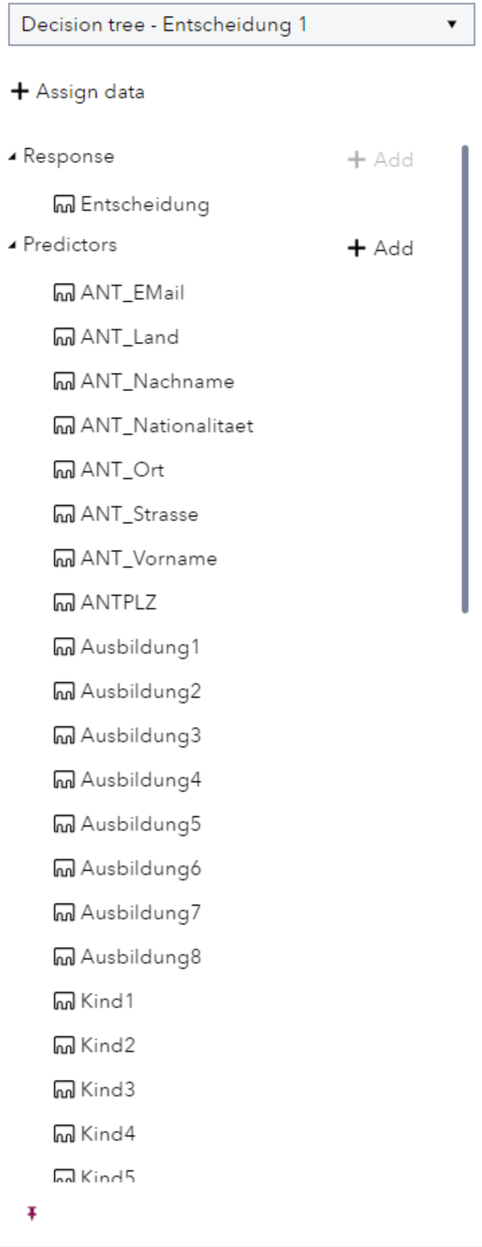
Reviewing different paths in the decision tree



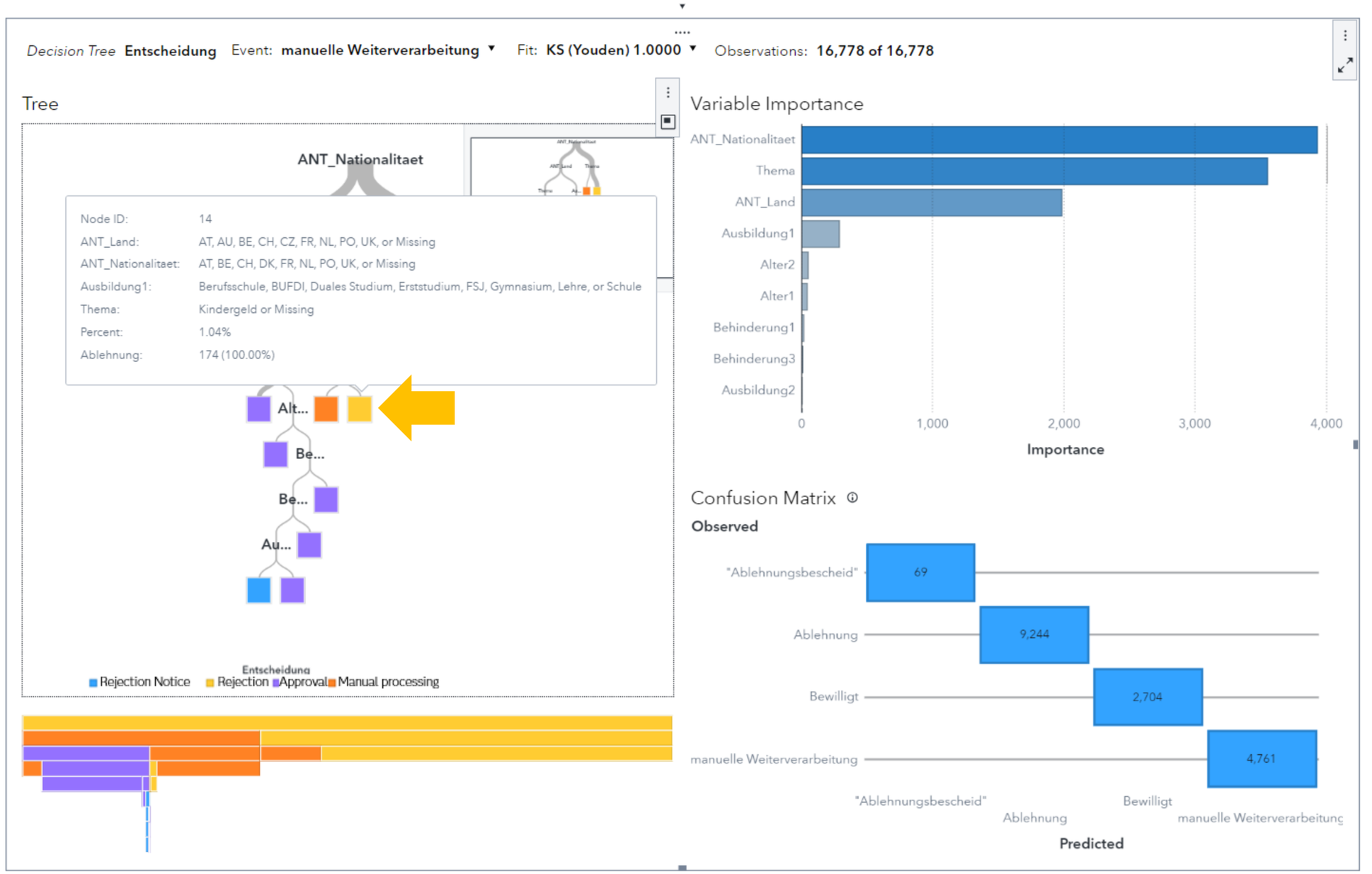


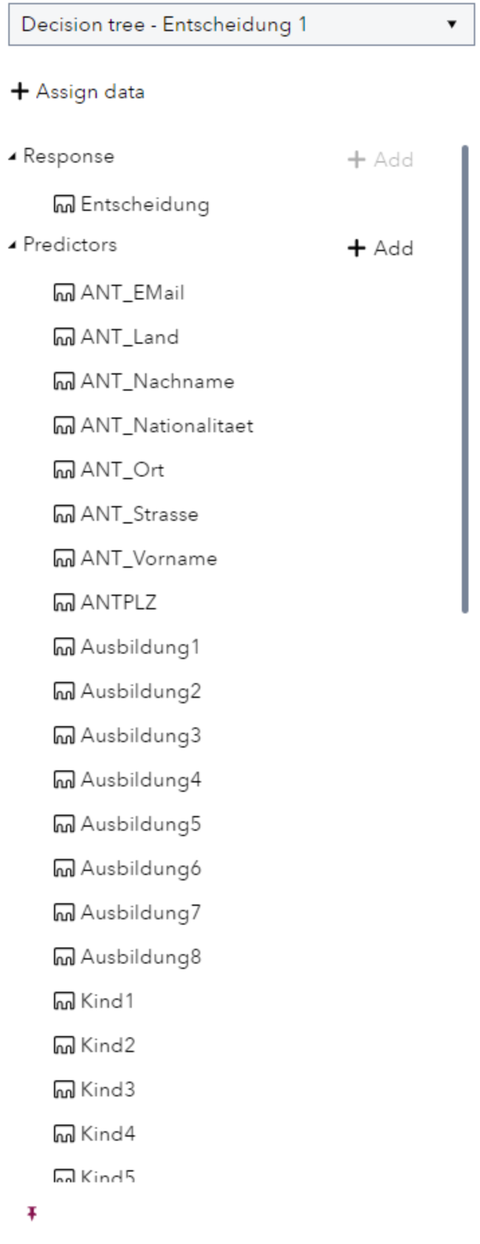
Reviewing different paths in the decision tree



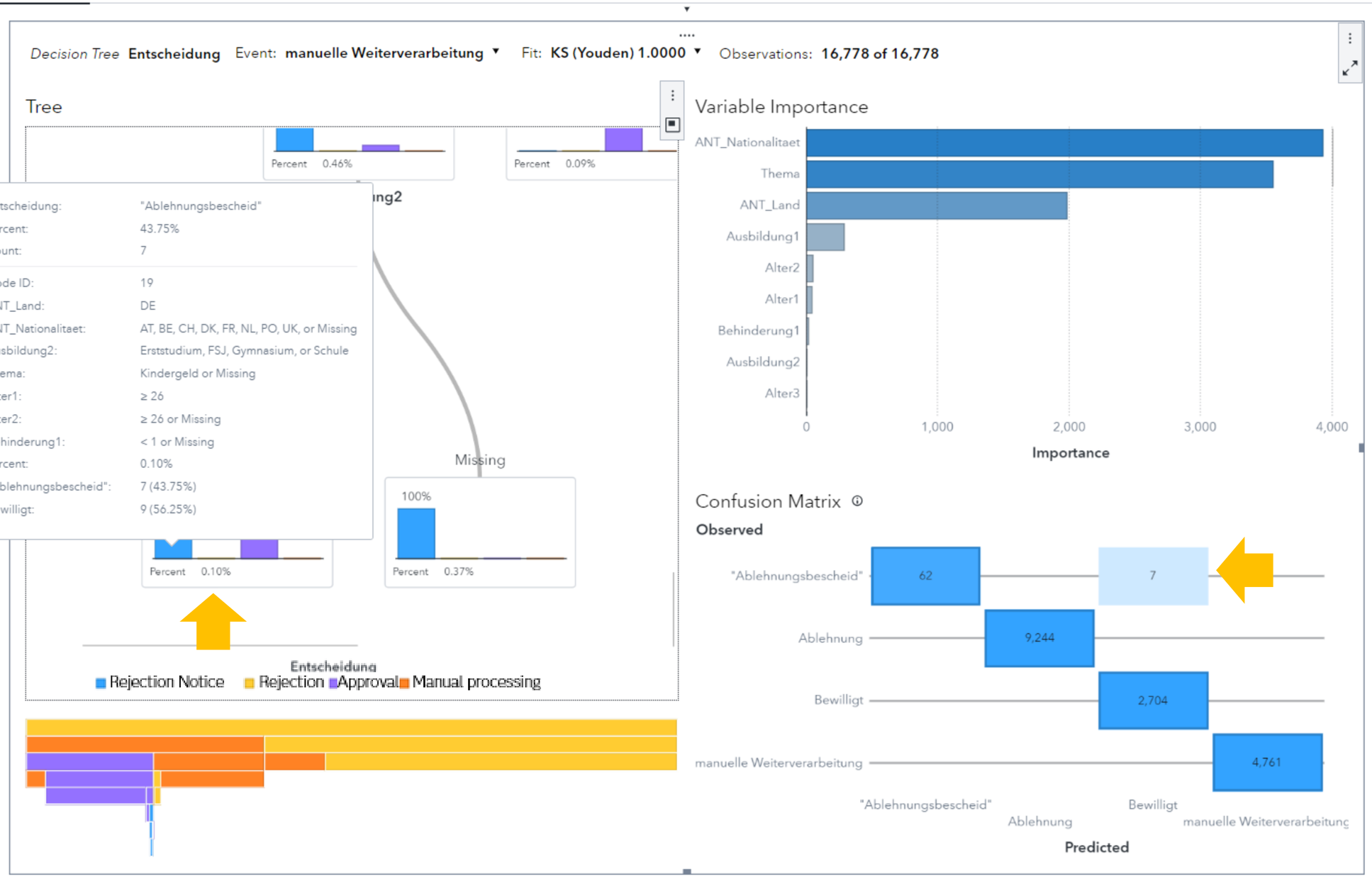


Reviewing different paths in the decision tree





Observe the confusion matrix to detect overrides and exceptions



Decision tree - Entscheidung 1

+ Assign data

Response + Add

Entscheidung

Predictors + Add

ANT_EMail
ANT_Land
ANT_Nachname
ANT_Nationalitaet
ANT_Ort
ANT_Strasse
ANT_Vorname
ANTPLZ
Ausbildung1
Ausbildung2
Ausbildung3
Ausbildung4
Ausbildung5
Ausbildung6
Ausbildung7
Ausbildung8
Kind1
Kind2
Kind3
Kind4
Kind5

Application Recommendations

- Preferred Method: Decision Tree
- Recommended SAS Tool: SAS Visual Analytics

- This is not limited to rules built and executed with SAS Intelligent Decisioning.
- It can be applied to any documented rules from your operational business.

Summary

- Supervised machine learning models help data scientist and business expert also beyond the classic application “prediction of values and probabilities”
- Using these methods, allows you to improve the status or the data quality, to better understand the relationship in your analysis data and enables you to present and interpret your results to business audiences
- These examples have been successfully tested and applied in many customer projects
- SAS Visual Analytics and SAS Code in SAS Studio are optimal to perform these analyses

HIMYM

How I Made Your Model with SAS® Viya

Season 1: The NoCode/LowCode Experience

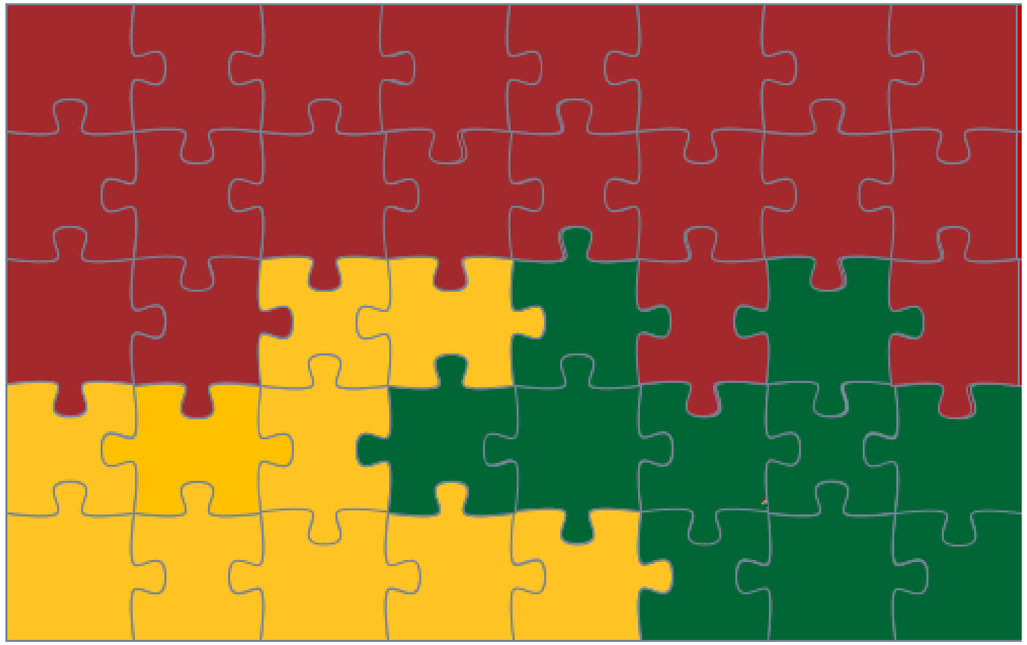
Episode 15: Profile the properties of your clusters

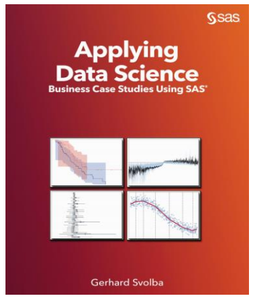
Author: Gerhard Svolba, Data Scientist, SAS

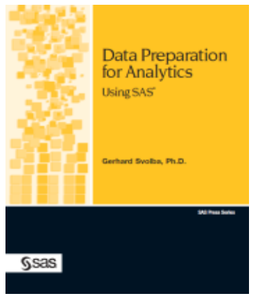
Data Preparation for Data Science

- Data Assembly
- Data Quality for Analytics
- Feature Generation

Gerhard Svolba,
Data Scientist @SAS
mailto:sastools.by.gerhard@gmx.net
gerhard.svolba@sas.com







Articles and Blogs — Medium, LinkedIn

Webinars — YouTube

Tipps & Tricks — SAS COMMUNITIES

Macros & Downloads