Project 2 – Predictive Analytics

DSBA-6201 - Business Intelligence and Analytics

Group members:

Yawo Eklou

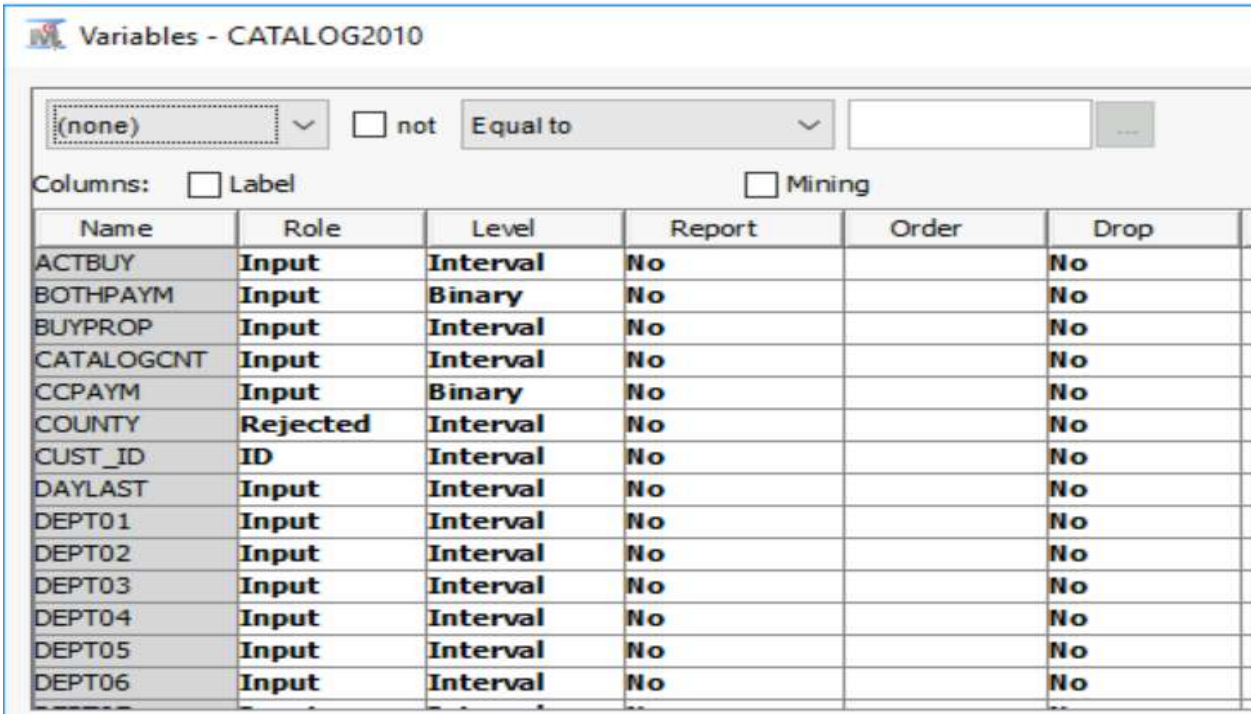
Marcie Matejka

May 10, 2021

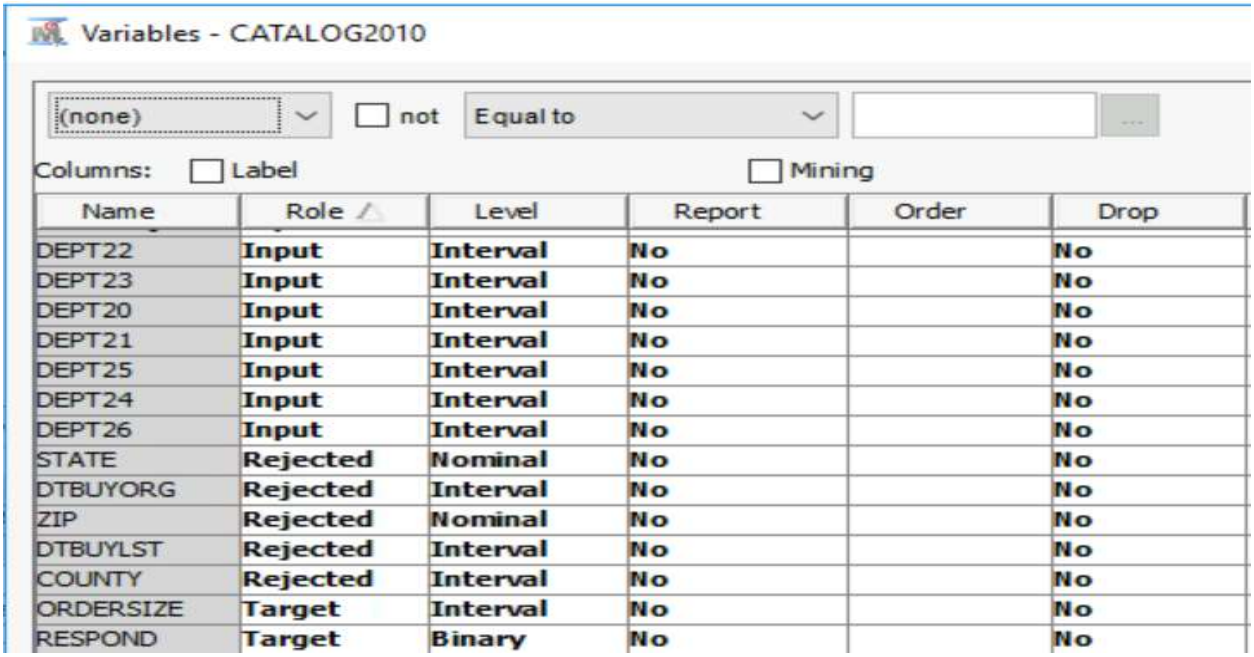
PART I: Predictive Analytics (EDA & RFM)

Completion of the initial tutorial with Catalog Data.

There are several variables that should be modified.

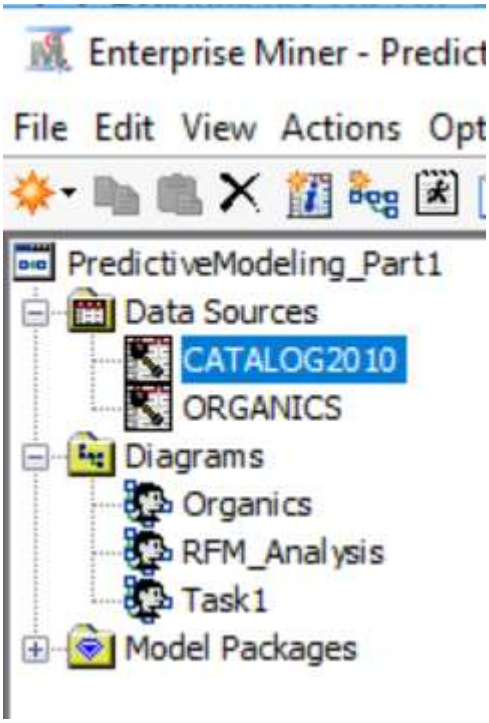


Name	Role	Level	Report	Order	Drop
ACTBUY	Input	Interval	No		No
BOTHPAYM	Input	Binary	No		No
BUYPROP	Input	Interval	No		No
CATALOGCNT	Input	Interval	No		No
CCPAYM	Input	Binary	No		No
COUNTY	Rejected	Interval	No		No
CUST_ID	ID	Interval	No		No
DAYLAST	Input	Interval	No		No
DEPT01	Input	Interval	No		No
DEPT02	Input	Interval	No		No
DEPT03	Input	Interval	No		No
DEPT04	Input	Interval	No		No
DEPT05	Input	Interval	No		No
DEPT06	Input	Interval	No		No

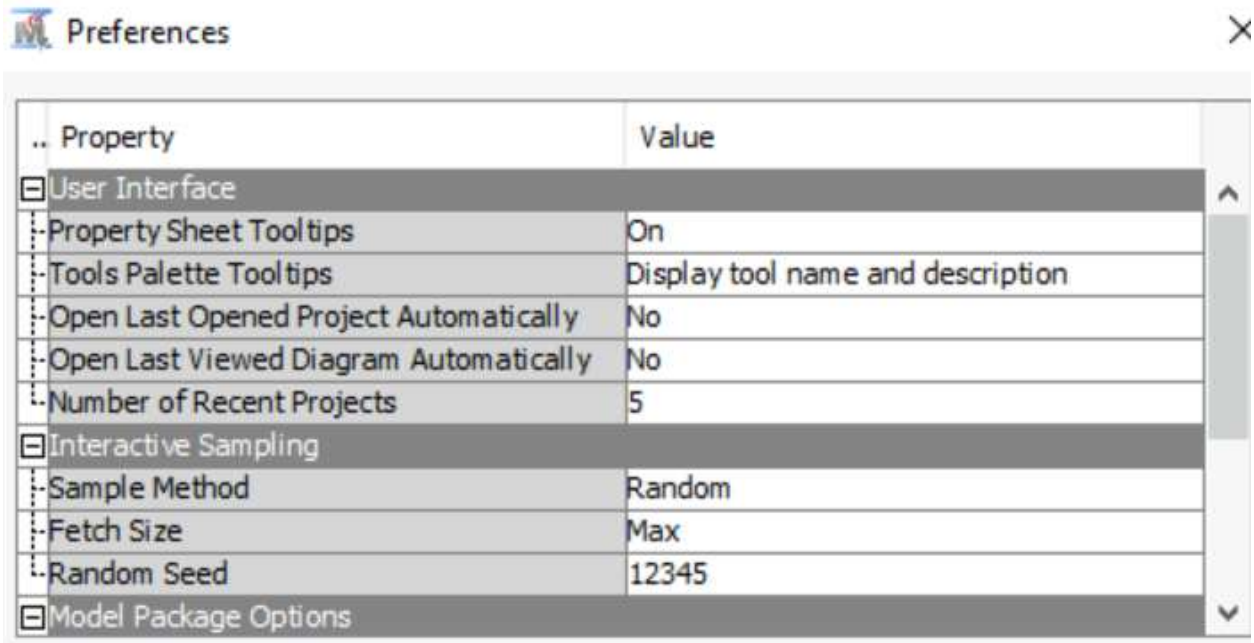


Name	Role	Level	Report	Order	Drop
DEPT22	Input	Interval	No		No
DEPT23	Input	Interval	No		No
DEPT20	Input	Interval	No		No
DEPT21	Input	Interval	No		No
DEPT25	Input	Interval	No		No
DEPT24	Input	Interval	No		No
DEPT26	Input	Interval	No		No
STATE	Rejected	Nominal	No		No
DTBUYORG	Rejected	Interval	No		No
ZIP	Rejected	Nominal	No		No
DTBUYLST	Rejected	Interval	No		No
COUNTY	Rejected	Interval	No		No
ORDERSIZE	Target	Interval	No		No
RESPOND	Target	Binary	No		No

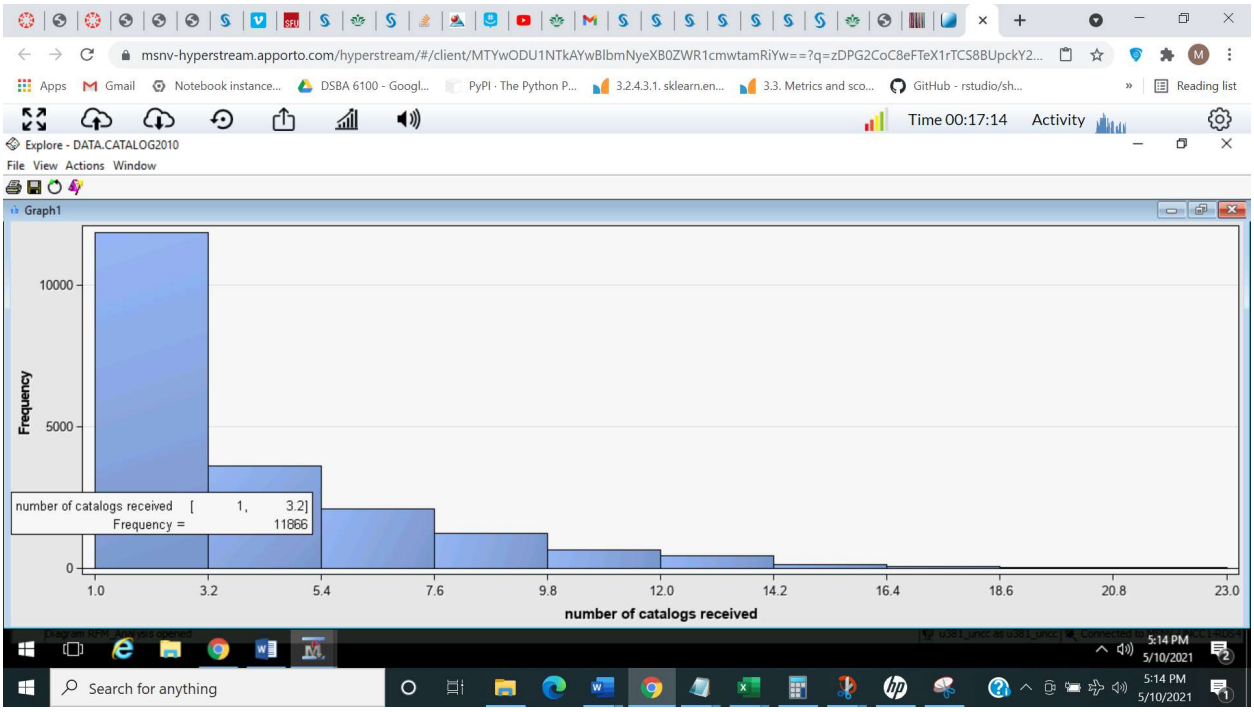
The CATALOG2010 data source is added to the Data Sources entry in the Project panel.



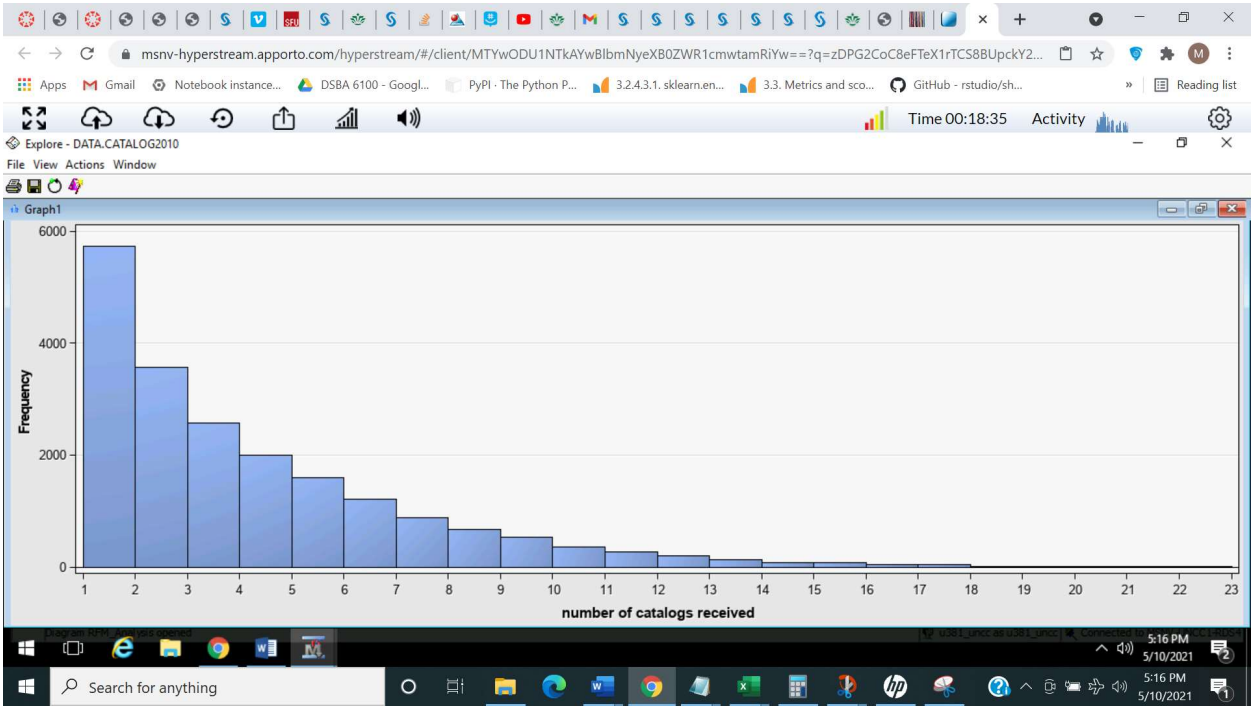
Follow the steps below to change the preference settings of SAS Enterprise Miner to use a random sample or all of the data source data in the Explore window



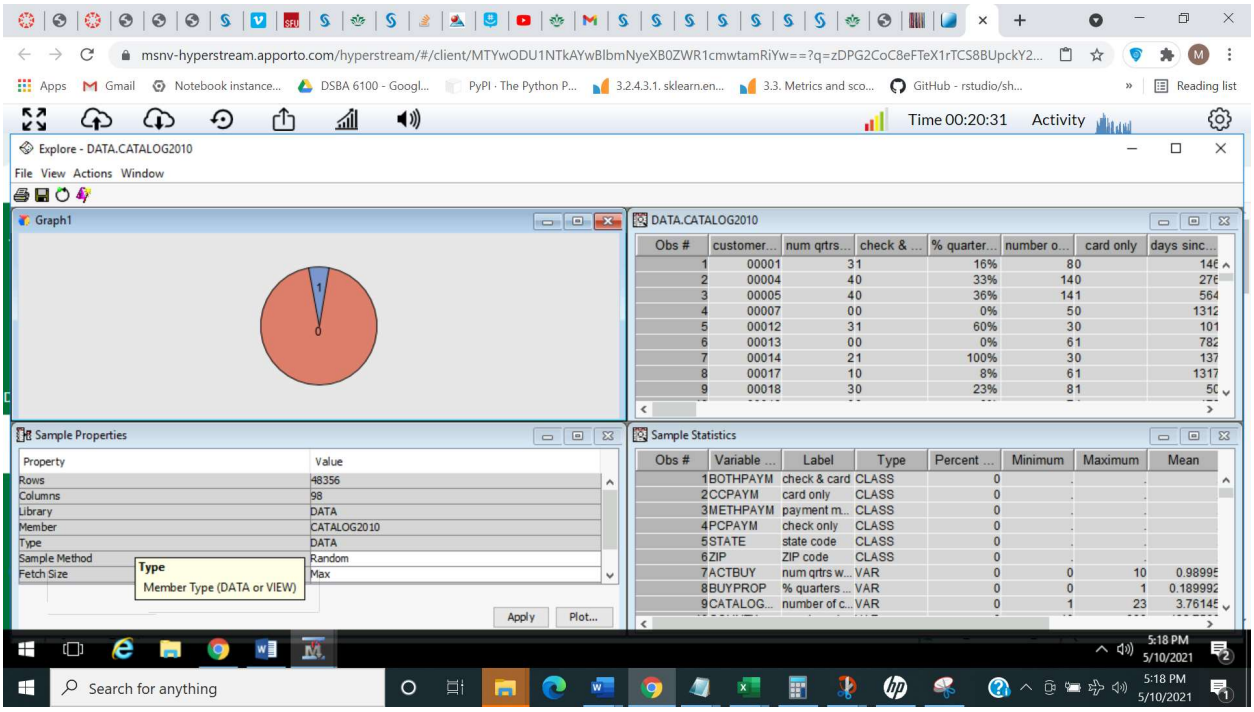
Creating a Histogram for a Single Variable



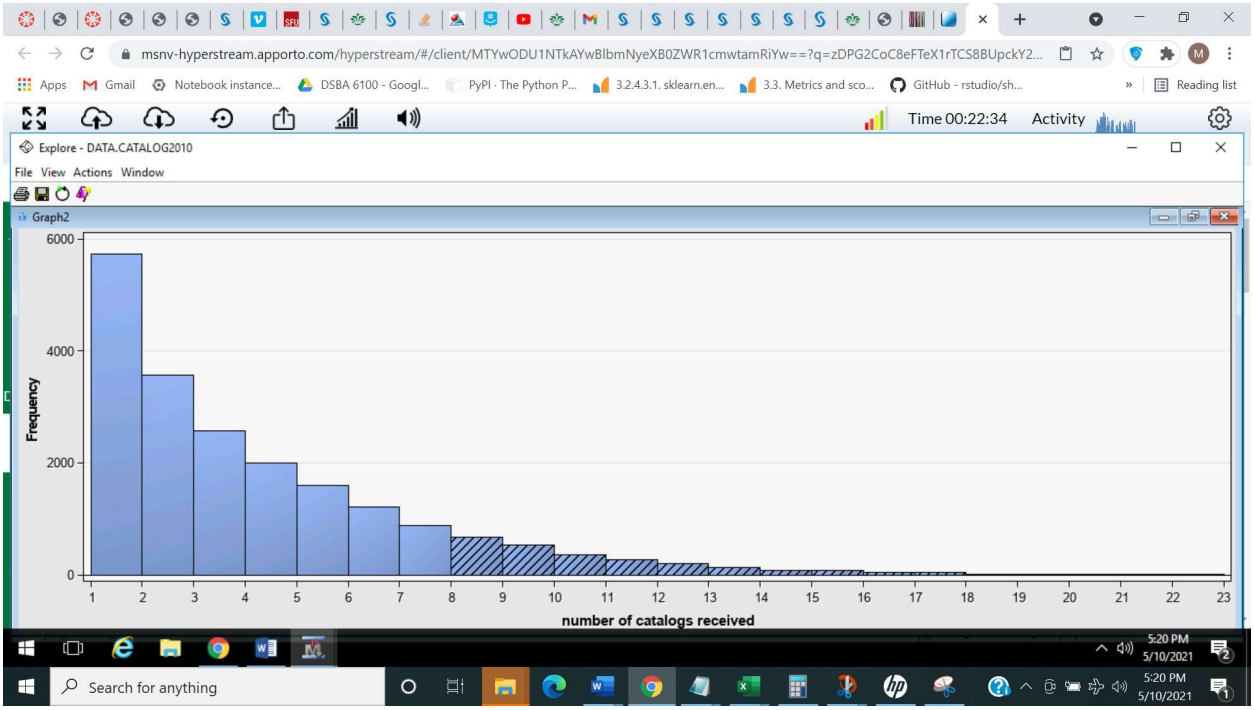
Changing the Graph Properties for a Histogram

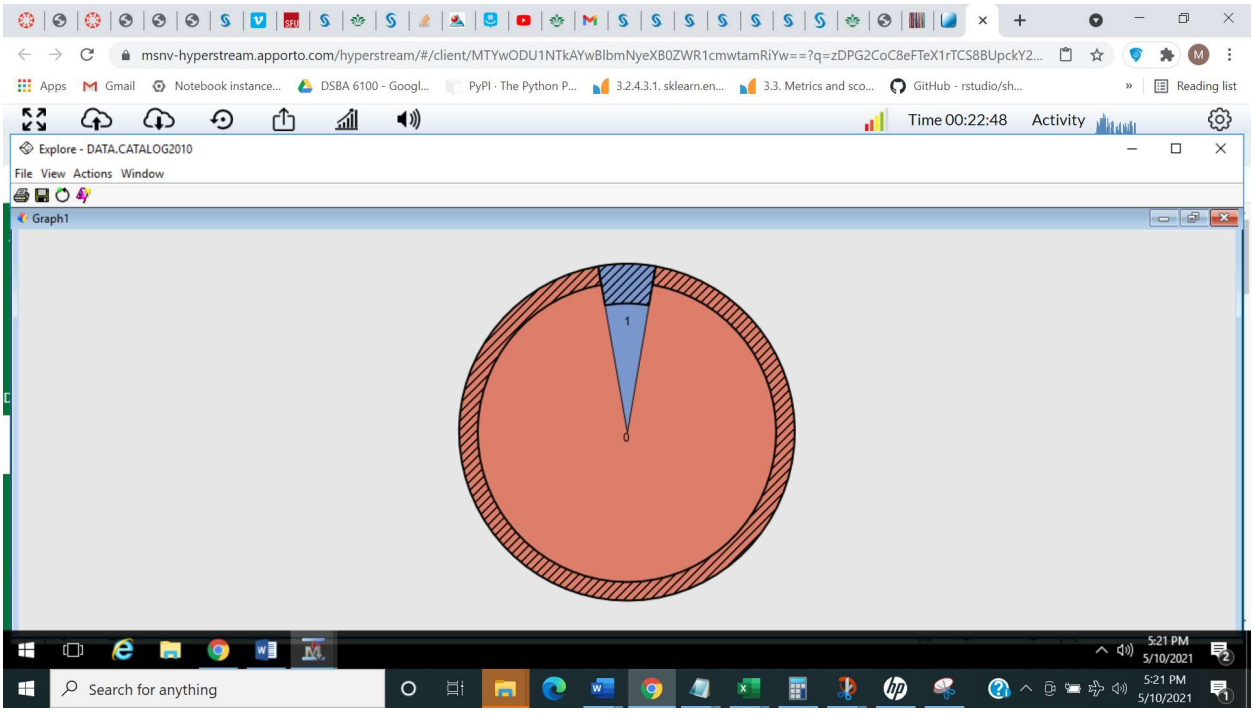


Adding Other Graphs

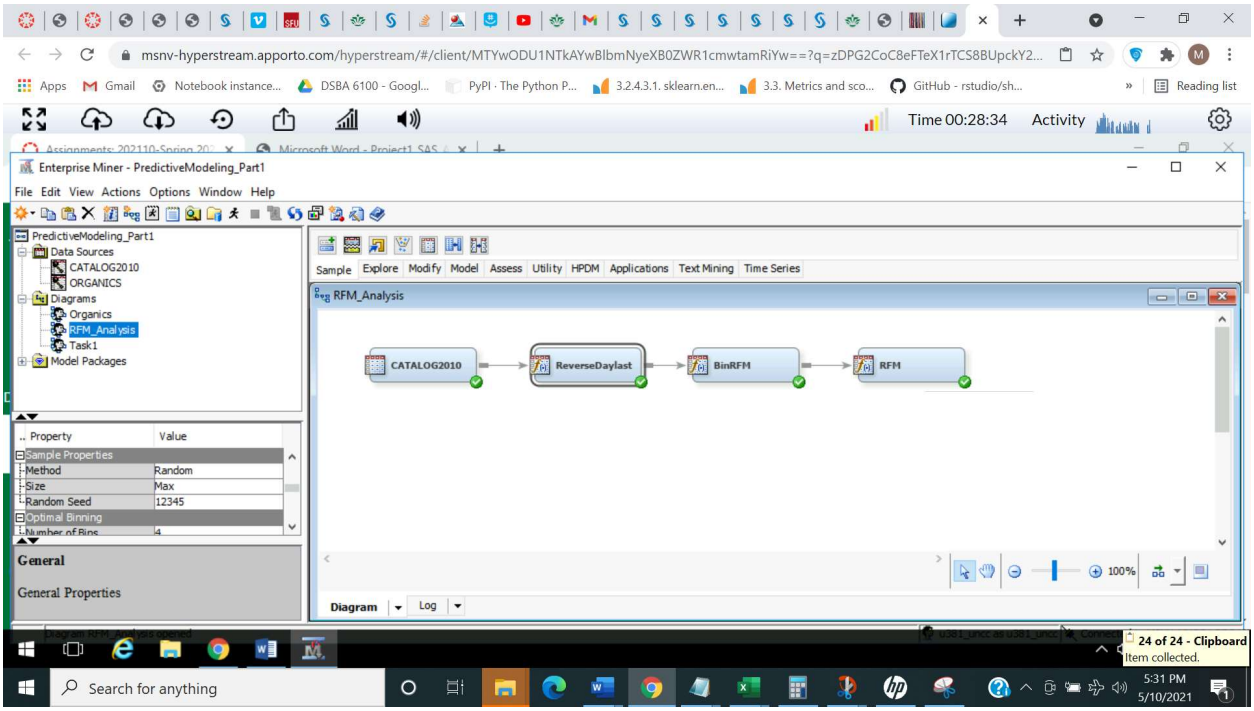


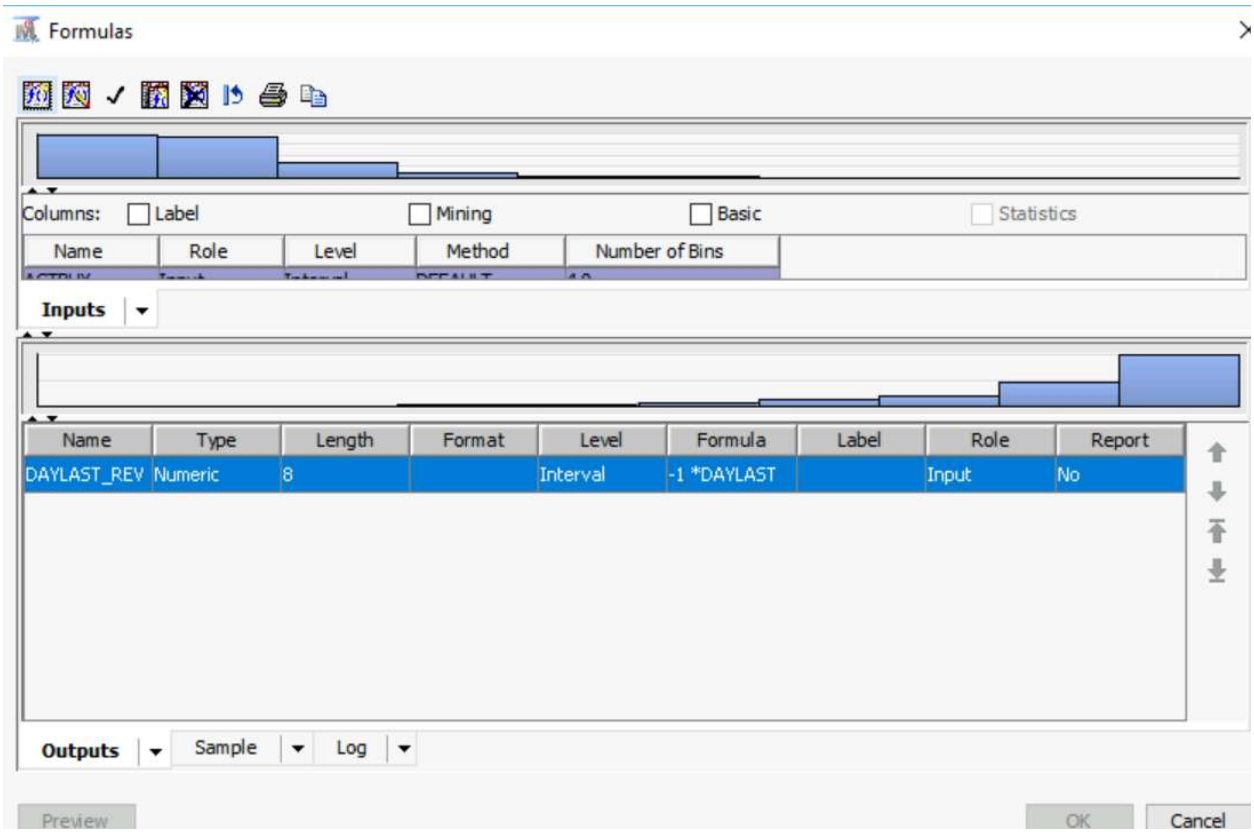
Exploring Variable Associations

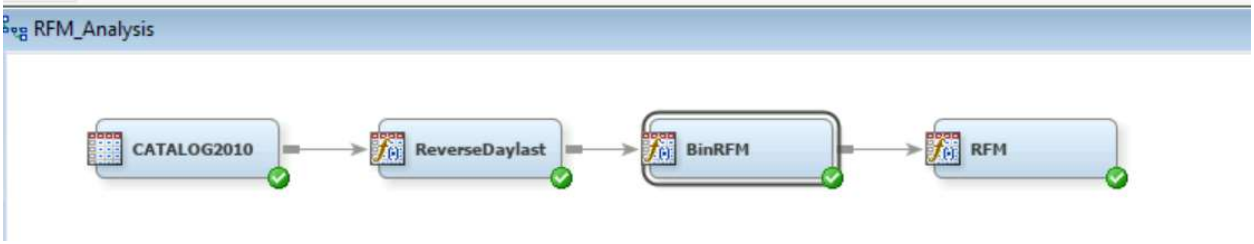


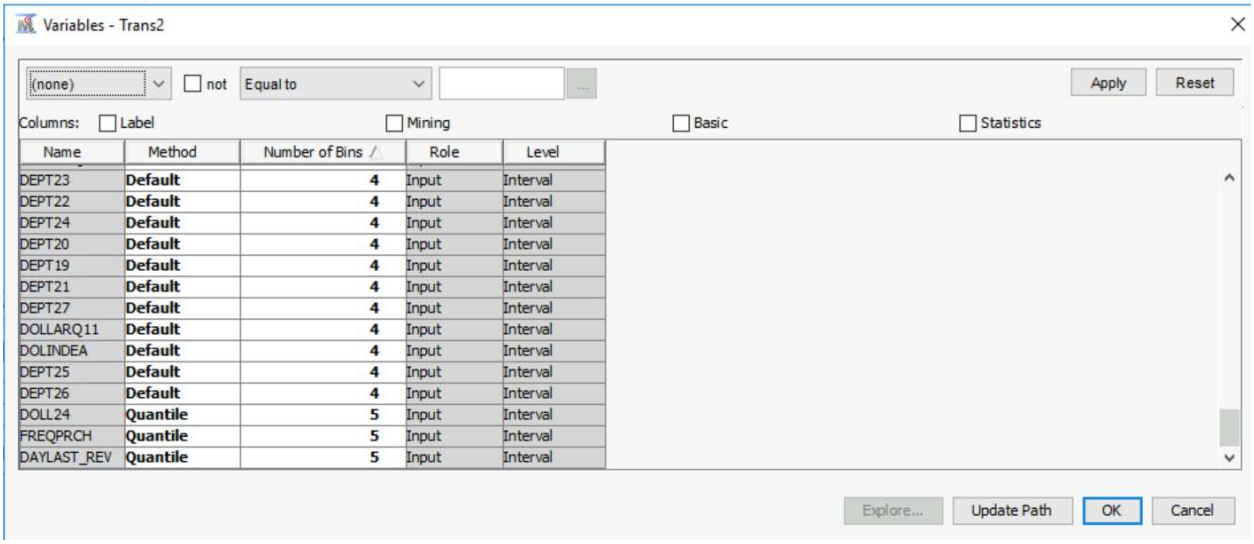


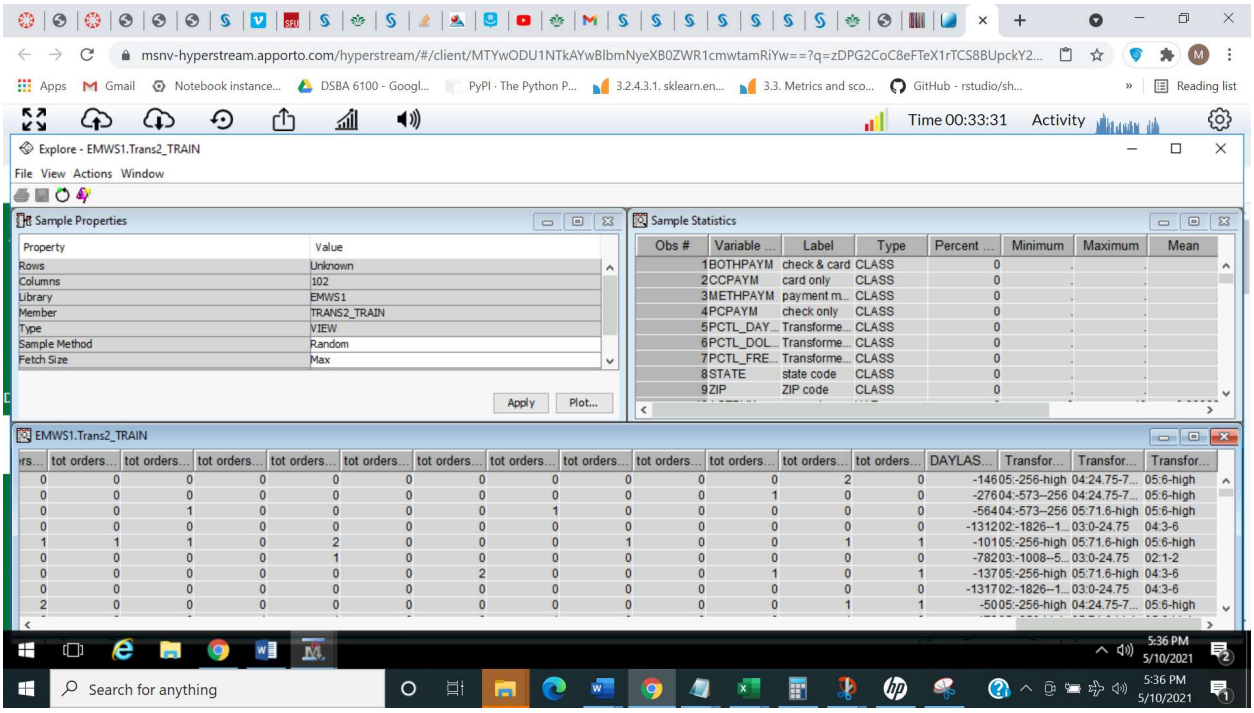
Catalog Case Study: Performing RFM Analysis of the CATALOG Data

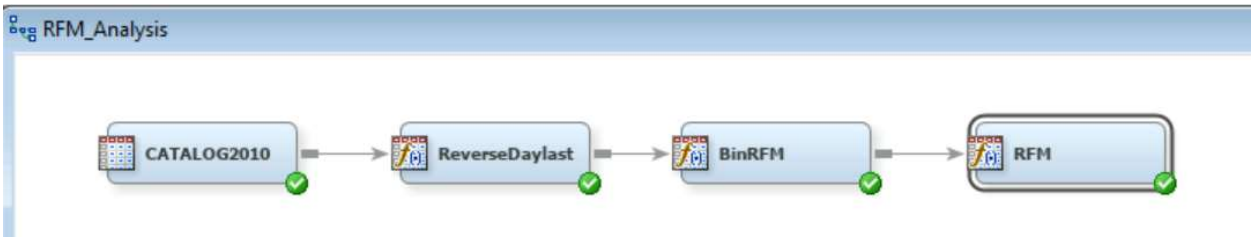


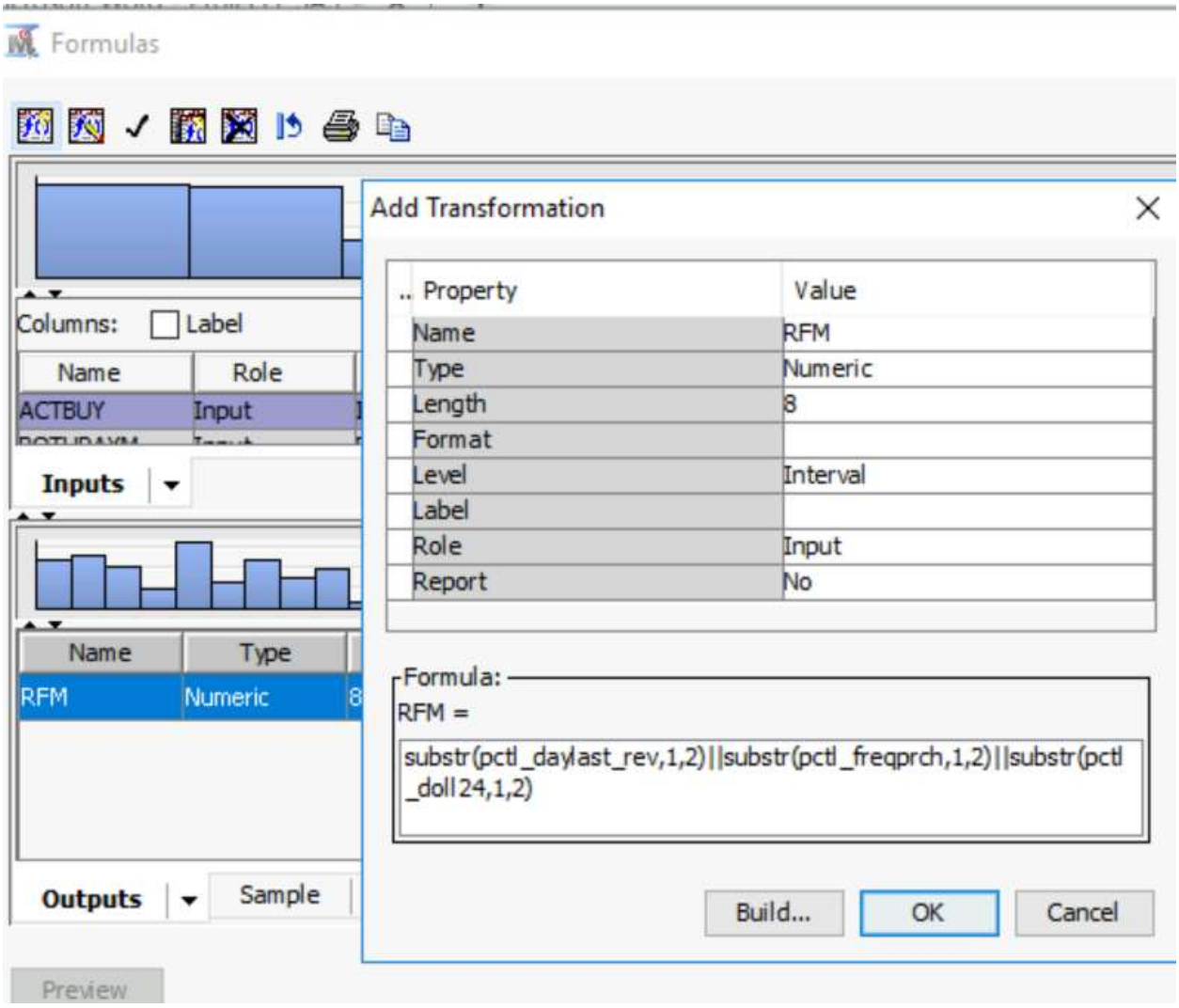

Formulas
Columns: Label Mining Basic Statistics
Name Role Level Method Number of Bins
Inputs
Name Type Length Format Level Formula Label Role Report
DAYLAST_REV Numeric 8 Interval -1 *DAYLAST Input No
Outputs Sample Log
Preview OK Cancel


RFM_Analysis
CATALOG2010
ReverseDaylast
BinRFM
RFM


Variables - Trans2
(none) not Equal to Apply Reset
Columns: Label Mining Basic Statistics
Name Method Number of Bins Role Level
DEPT23 Default 4 Input Interval
DEPT22 Default 4 Input Interval
DEPT24 Default 4 Input Interval
DEPT20 Default 4 Input Interval
DEPT19 Default 4 Input Interval
DEPT21 Default 4 Input Interval
DEPT27 Default 4 Input Interval
DOLLARQ11 Default 4 Input Interval
DOLINDEA Default 4 Input Interval
DEPT25 Default 4 Input Interval
DEPT26 Default 4 Input Interval
DOLL24 Quantile 5 Input Interval
FREQPRCH Quantile 5 Input Interval
DAYLAST_REV Quantile 5 Input Interval
Explore... Update Path OK Cancel

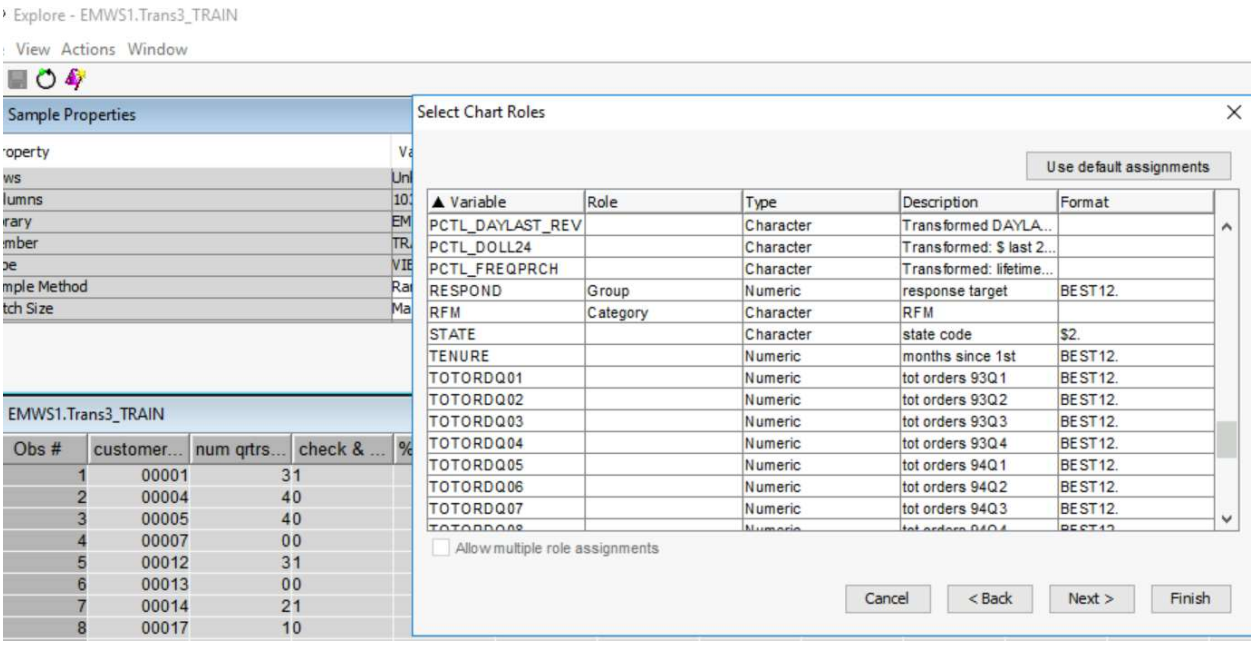

msnv-hyperstream.apporto.com/hyperstream/#/client/MTYwODU1NTkAYwBlbmNyeXB0ZWR1cmwtamRiYw==?q=zDPG2CoC8eFTeX1rTCS8BUpckY2...
Time 00:33:31
Activity
Explore - EMWS1.Trans2_TRAIN
File View Actions Window
Sample Properties
Property
Value
Rows
Unknown
Columns
102
Library
EMWS1
Member
TRANS2_TRAIN
Type
VIEW
Sample Method
Random
Fetch Size
Max
Apply
Plot...
Sample Statistics
Obs #
Variable ...
Label
Type
Percent ...
Minimum
Maximum
Mean
1 BOTHPAYM check & card CLASS 0
2 CCPAYM card only CLASS 0
3 METHPAYM payment m... CLASS 0
4 PCPAYM check only CLASS 0
5 PCTL_DAY... Transforme... CLASS 0
6 PCTL_DOL... Transforme... CLASS 0
7 PCTL_FRE... Transforme... CLASS 0
8 STATE state code CLASS 0
9 ZIP ZIP code CLASS 0
EMWS1.Trans2_TRAIN
Search for anything
5:36 PM
5/10/2021

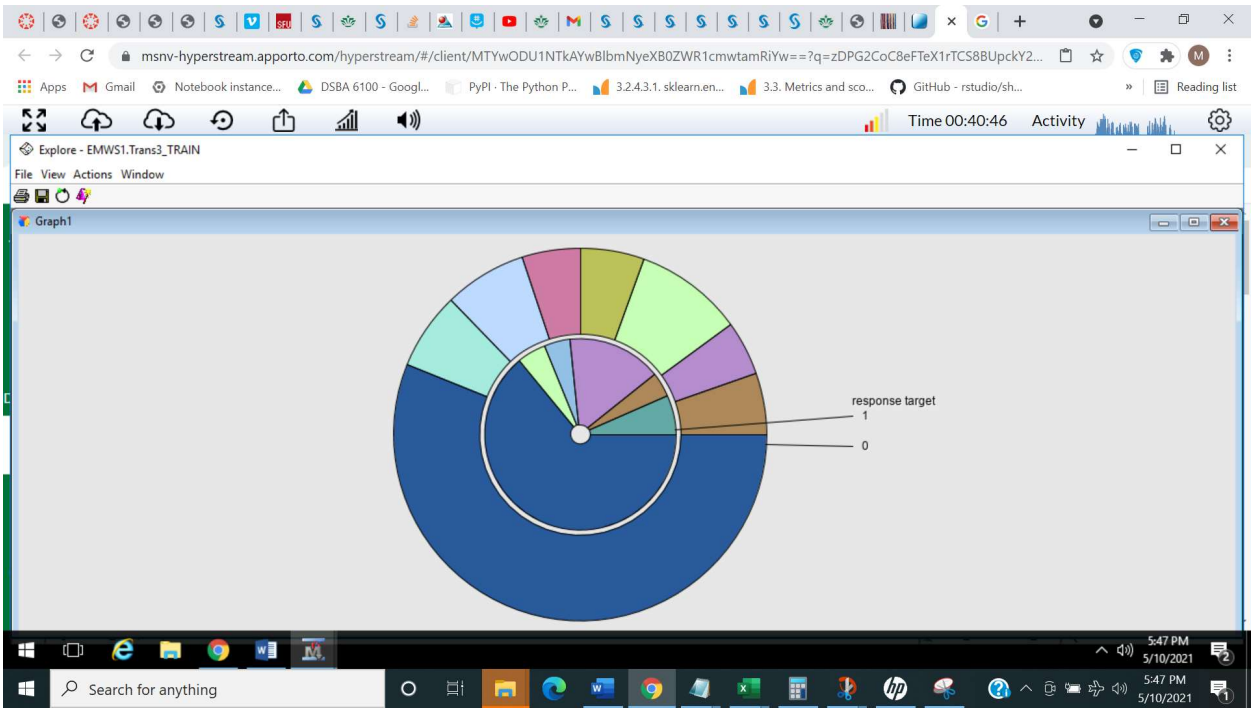

RFM_Analysis
CATALOG2010
ReverseDaylast
BinRFM
RFM

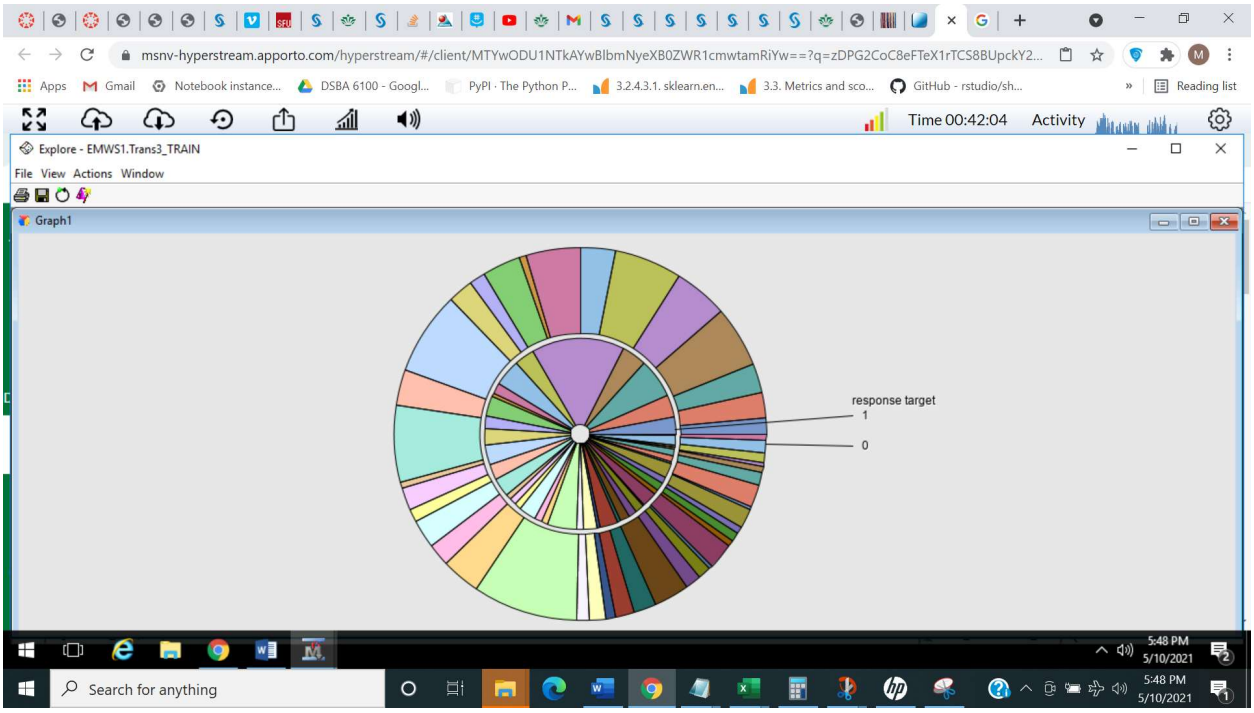

Formulas
Columns:
Label
Name
Role
ACTBUY
Input
Inputs
Name
Type
RFM
Numeric
Outputs
Sample
Preview
Add Transformation
Property
Value
Name
RFM
Type
Numeric
Length
8
Format
Level
Interval
Label
Role
Input
Report
No
Formula:
RFM =
substr(pctl_daylast_rev,1,2)||substr(pctl_freqprch,1,2)||substr(pctl_doll24,1,2)
Build...
OK
Cancel

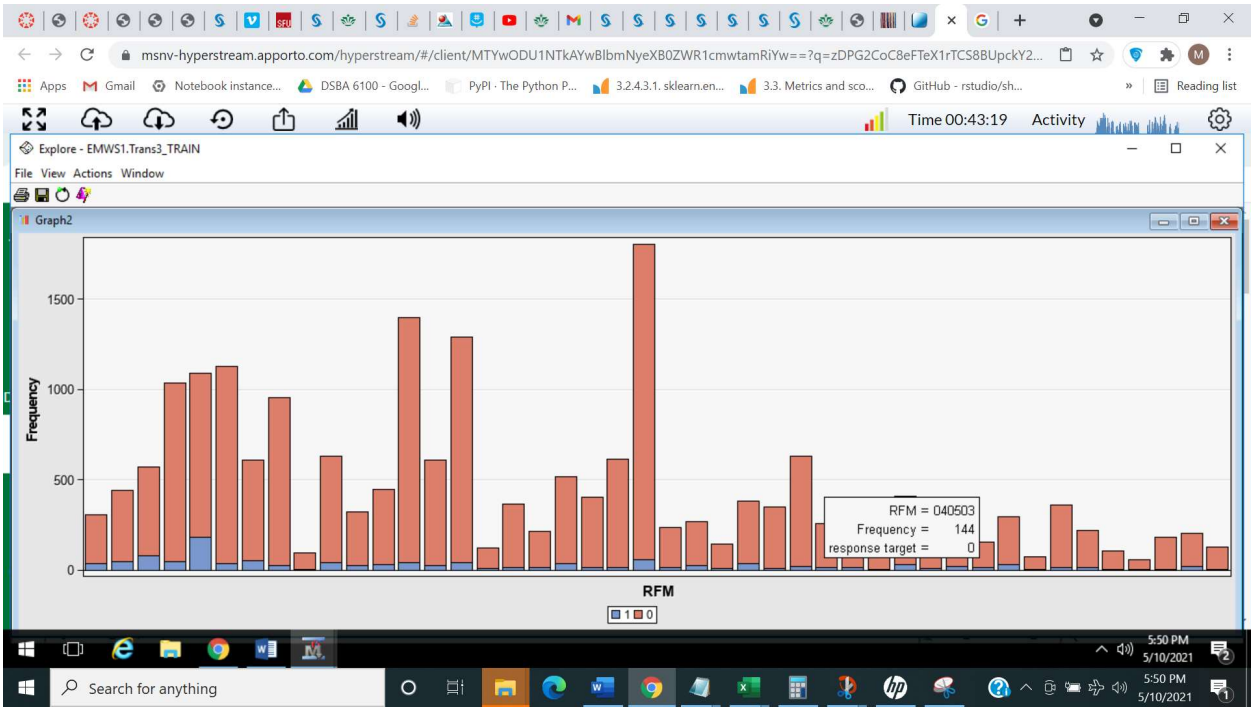
Catalog Case Study: Performing Graphical RFM Analysis

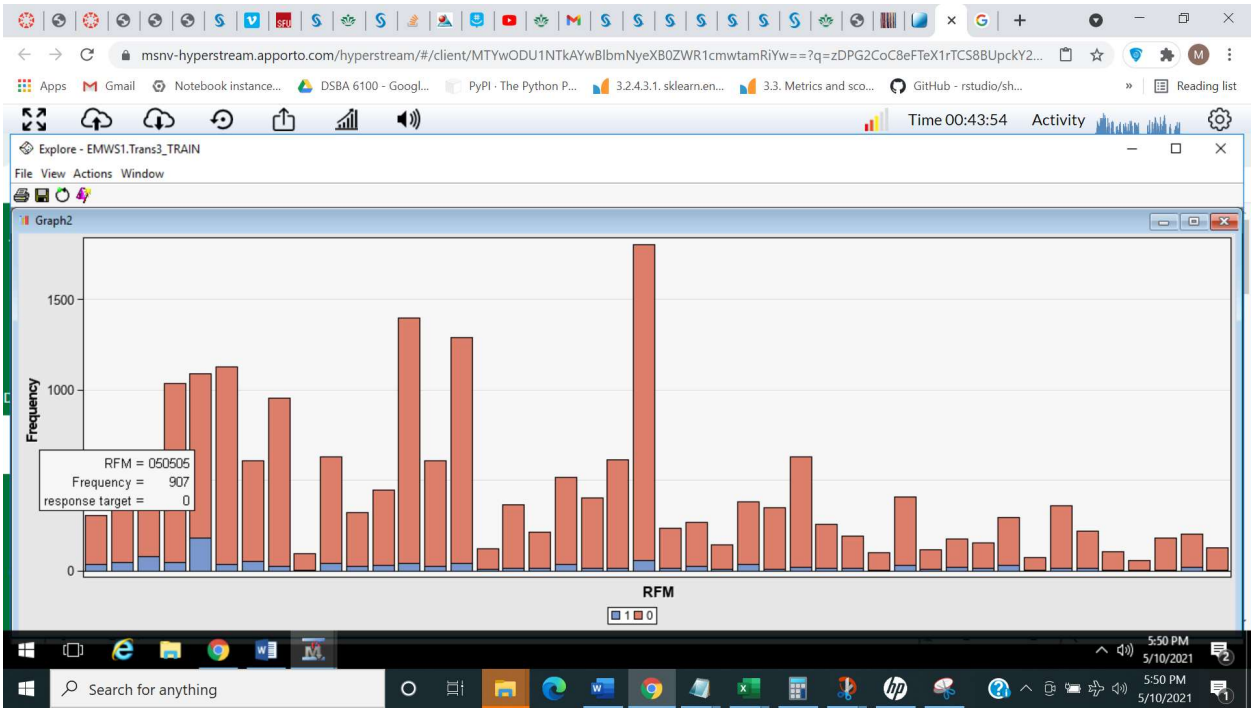
A Grouped Pie Chart

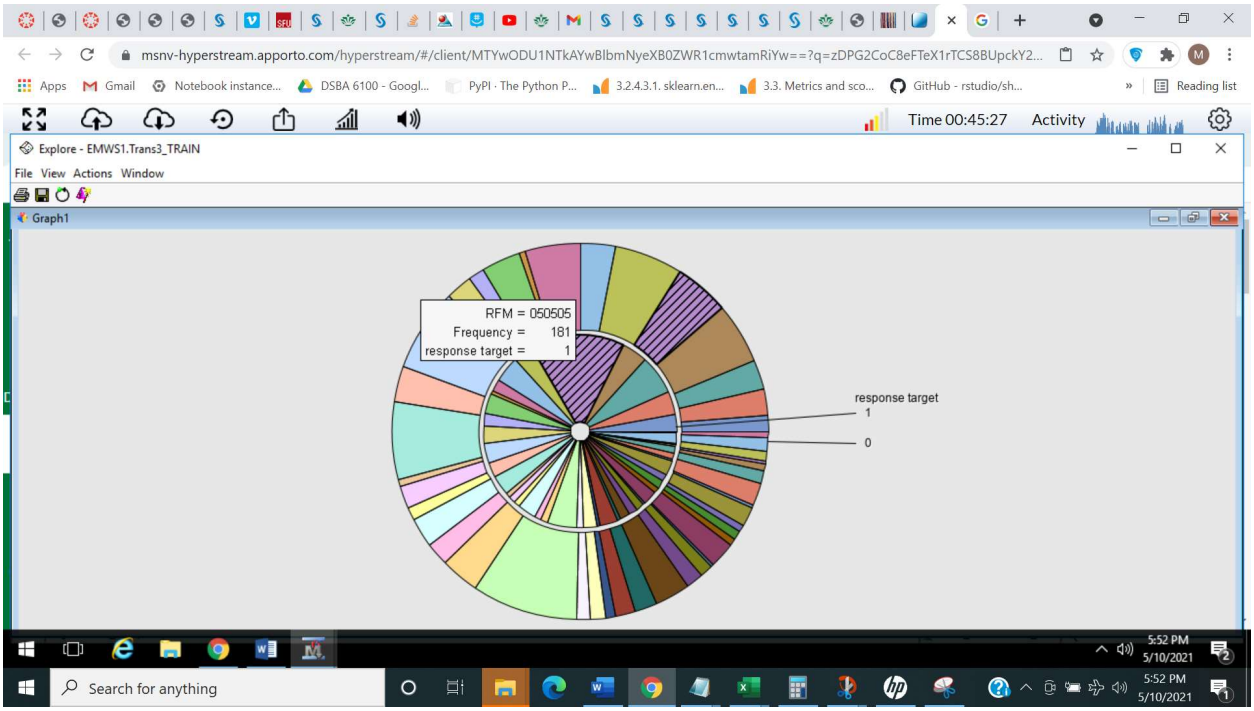


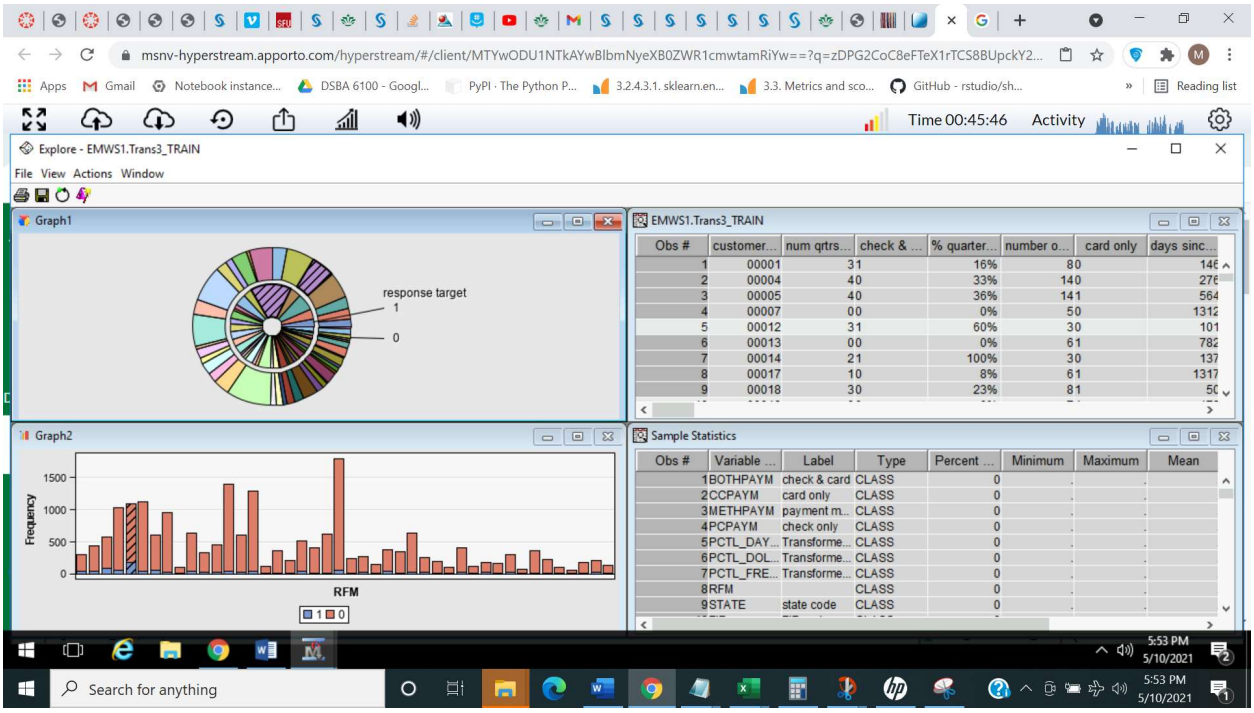


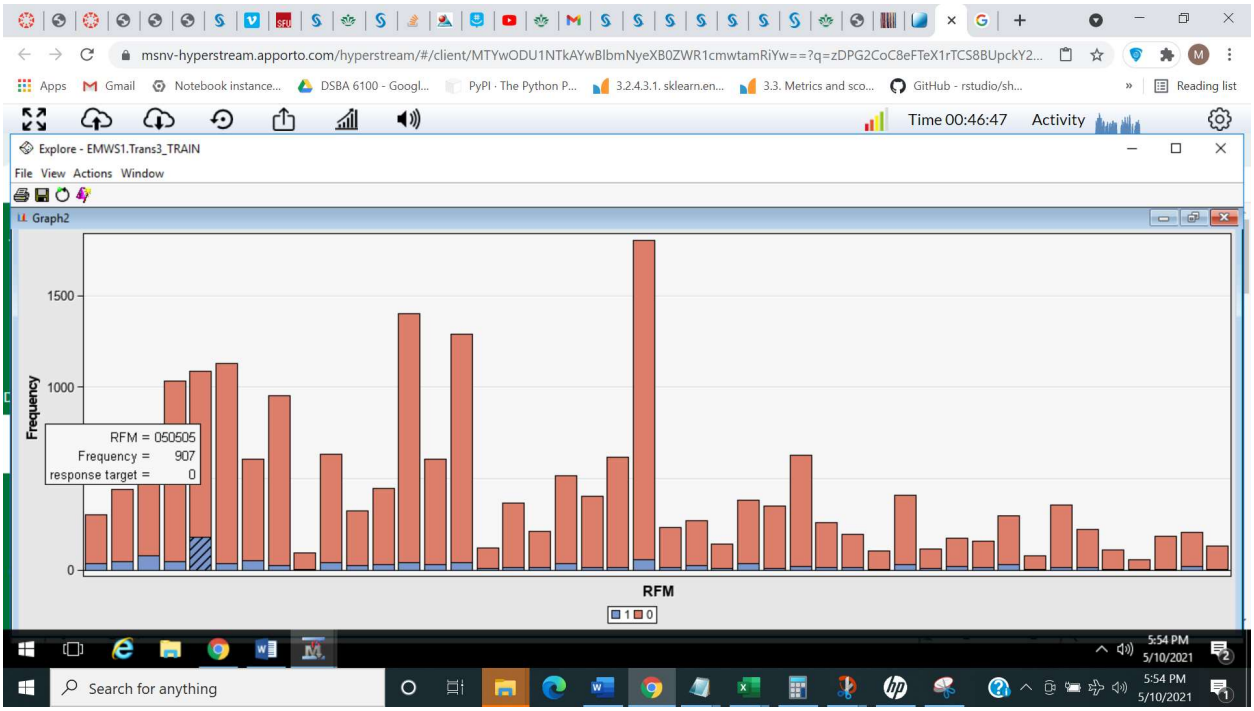

msnv-hyperstream.apporto.com/hyperstream/#/client/MTYwODU1NTkAYwBlbmNyeXB0ZWR1cmwtamRiYw==?q=zDPG2CoC8eFTeX1rTCS8BUpckY2...
Time 00:42:04
Activity
Explore - EMWS1.Trans3_TRAIN
File View Actions Window
Graph1
response target
1
0
Search for anything
5:48 PM
5/10/2021

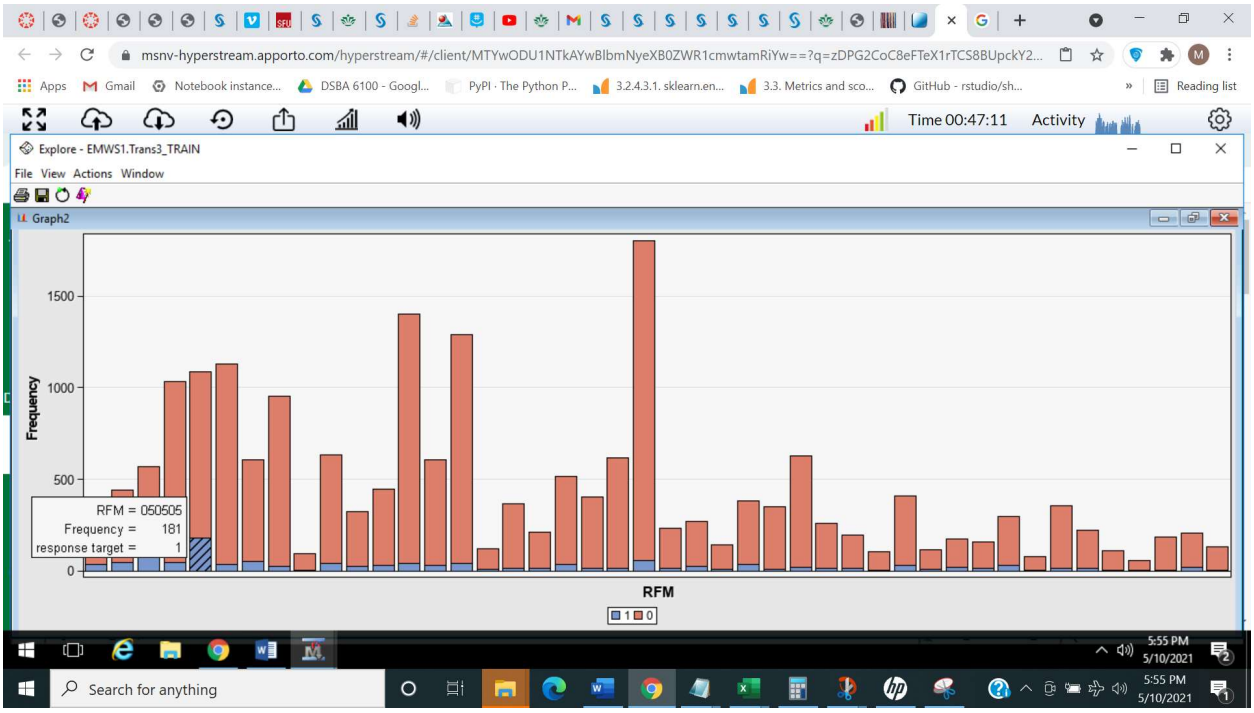

msnv-hyperstream.apporto.com/hyperstream/#/client/MTYwODU1NTkAYwBlbmNyeXB0ZWR1cmwtamRiYw==?q=zDPG2CoC8eFTeX1rTCS8BUpckY2...
Time 00:43:19
Activity
Explore - EMWS1.Trans3_TRAIN
File View Actions Window
Graph2
1500
1000
500
0
Frequency
RFM
RFM = 040503
Frequency = 144
response target = 0
1 0
Search for anything
5:50 PM
5/10/2021


msnv-hyperstream.apporto.com/hyperstream/#/client/MTYwODU1NTkAYwBlbmNyeXB0ZWR1cmwtamRiYw==?q=zDPG2CoC8eFTeX1rTCS8BUpckY2...
Time 00:43:54
Activity
Explore - EMWS1.Trans3_TRAIN
File View Actions Window
Graph2
Frequency
1500
1000
0
RFM = 050505
Frequency = 907
response target = 0
RFM
1 0
Search for anything
5:50 PM
5/10/2021


msnv-hyperstream.apporto.com/hyperstream/#/client/MTYwODU1NTkAYwBlbmNyeXB0ZWR1cmwtamRiYw==?q=zDPG2CoC8eFTeX1rTCS8BUpckY2...
Time 00:45:27
Activity
Explore - EMWS1.Trans3_TRAIN
File View Actions Window
Graph1
RFM = 050505
Frequency = 181
response target = 1
response target
1
0
Search for anything
5:52 PM
5/10/2021


msnv-hyperstream.apporto.com/hyperstream/#/client/MTYwODU1NTkAYwBlbmNyeXB0ZWR1cmwtamRiYw==?q=zDPG2CoC8eFTeX1rTCS8BUpckY2...
Time 00:45:46
Activity
Explore - EMWS1.Trans3_TRAIN
File View Actions Window
Graph1
response target
1
0
EMWS1.Trans3_TRAIN
Obs # customer... num qtrs... check & ... % quarter... number o... card only days sinc...
1 00001 31 16% 80 146
2 00004 40 33% 140 276
3 00005 40 36% 141 564
4 00007 00 0% 50 1312
5 00012 31 60% 30 101
6 00013 00 0% 61 782
7 00014 21 100% 30 137
8 00017 10 8% 61 1317
9 00018 30 23% 81 50
Graph2
Frequency
1500
1000
500
0
RFM
1 0
Sample Statistics
Obs # Variable ... Label Type Percent ... Minimum Maximum Mean
1 BOTHPAYM check & card CLASS 0
2 CCPAYM card only CLASS 0
3 METHPAYM payment m... CLASS 0
4 PCPAYM check only CLASS 0
5 PCTL_DAY... Transforme... CLASS 0
6 PCTL_DOL... Transforme... CLASS 0
7 PCTL_FRE... Transforme... CLASS 0
8 RFM CLASS 0
9 STATE state code CLASS 0
5:53 PM
5/10/2021
Search for anything
5:53 PM
5/10/2021

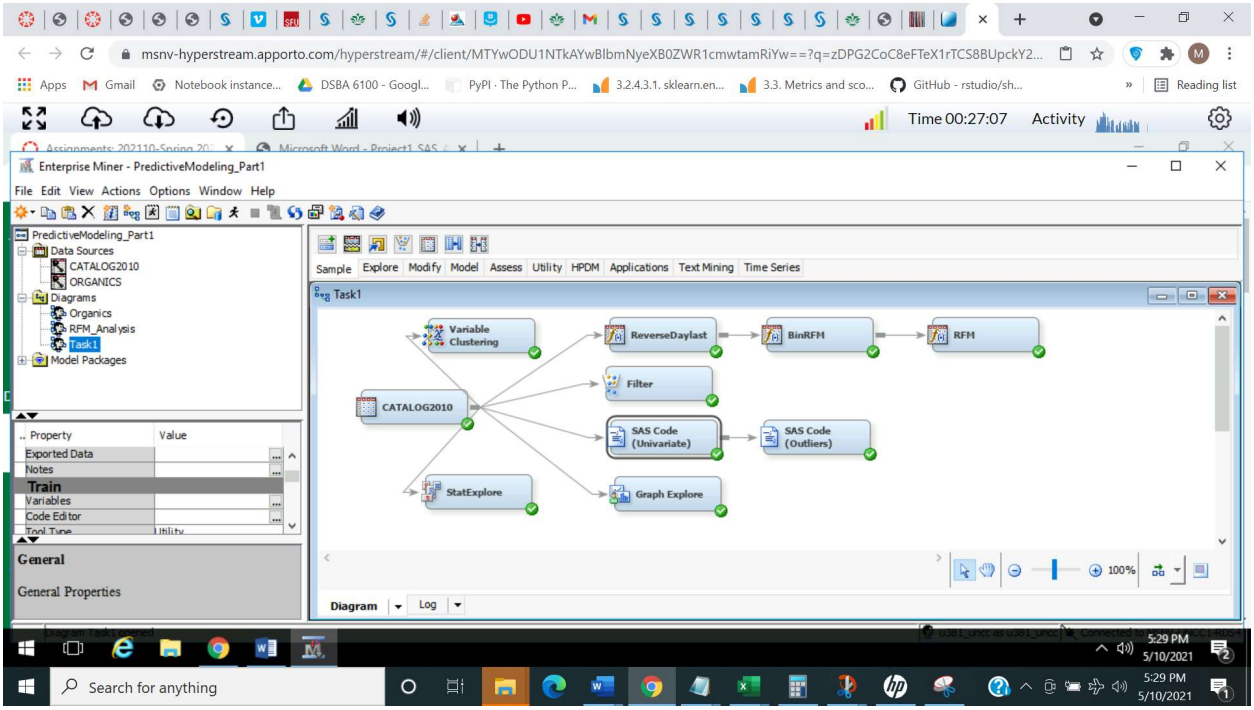

msnv-hyperstream.apporto.com/hyperstream/#/client/MTYwODU1NTkAYwBlbmNyeXB0ZWR1cmwtamRiYw==?q=zDPG2CoC8eFTeX1rTCS8BUpckY2...
Time 00:46:47
Activity
Explore - EMWS1.Trans3_TRAIN
File View Actions Window
Graph2
1500
1000
0
Frequency
RFM = 050505
Frequency = 907
response target = 0
RFM
1 0
5:54 PM
5/10/2021
Search for anything
5:54 PM
5/10/2021


msnv-hyperstream.apporto.com/hyperstream/#/client/MTYwODU1NTkAYwBlbmNyeXB0ZWR1cmwtamRiYw==?q=zDPG2CoC8eFTeX1rTCS8BUpckY2...
Time 00:47:11
Activity
Explore - EMWS1.Trans3_TRAIN
File View Actions Window
Graph2
Frequency
1500
1000
500
0
RFM = 050505
Frequency = 181
response target = 1
RFM
1 0
Search for anything
5:55 PM
5/10/2021

Task 1:

SAS Code (5points)

- **Univariate Analysis**
- **Outlier**
- **Plotting**



You need to perform additional analysis on various variables and make a report

1) You might want to study which variables are highly correlated. If you find such variables you can suggest dimension reduction by dropping one of the variables. **6points**

The following variables are highly correlated based a correlation of greater than 0.8. (http://www.sfu.ca/~dsignori/buec333/lecture%2016.pdf)

Variable 1	Variable 2	Correlation
days since last	months since last	0.999975251
months since last	days since last	0.999975251
avg $ net demand	total $ demand	0.993953254
total $ demand	avg $ net demand	0.993953254
avg $ demand	tot $ net demand	0.953178097
tot $ net demand	avg $ demand	0.953178097
tot units demand	total $ demand	0.881179133

total $ demand	tot units demand	0.881179133
avg $ net demand	tot units demand	0.877361971
tot units demand	avg $ net demand	0.877361971
lifetime orders	total $ demand	0.815401727
total $ demand	lifetime orders	0.815401727
avg $ net demand	lifetime orders	0.812388543
lifetime orders	avg $ net demand	0.812388543
lifetime orders	tot units demand	0.804471837
tot units demand	lifetime orders	0.804471837

Highly correlated variables may overly skew data analysis in specific directions. One remedy is to drop one of each pair of highly correlated variables. In the CATALOG2010 data, we would drop "months since last" and keep "days since last" because days is a more precise measure. Between variables "avg $ net demand" and "total $ demand", we would drop "avg $ net demand" because it is a comparison value and keep "total $ demand" because it is a final demand value rather than a value for comparison. For the same reason, we would keep "avg $ demand" instead of the comparison value of "tot $ net demand". When reviewing "tot unit demand" and "total $ demand", we have already indicated we will keep "total $ demand" so "tot unit demand" can be dropped. The next pair is "avg $ net demand" and "tot units demand". We have already indicated that we will drop both. If we must retain one, it would be "tot units demand" because it represents a final concrete value. When comparing variables "lifetime orders" and "total $ demand", we have already indicated that we will keep "total $ demand". Between "avg $net demand" and "lifetime orders", we can indicated that we will drop both of these. However, if we must retain one it will be "lifetime orders". In comparing "lifetime orders" and "tot units demand", we will drop "tot units demand" because it is more highly correlated with another variable – "total $ demand".

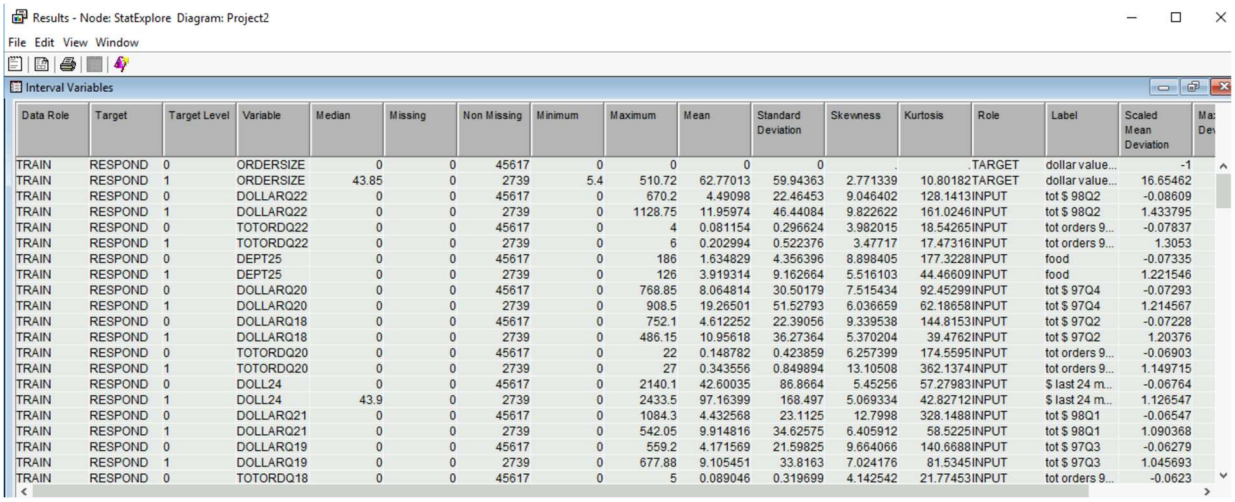
2) You can study in if there are outliers in your variables **6points**

After adding and running the filter node to the project, we have the following results:

Number Of Observations

Data Role	Filtered	Excluded	DATA
TRAIN	20607	27749	48356

With the StatExplore Node, we generated the Interval Variables table which displays a table of summary statistics for interval variables including mean and standard deviation.



Results - Node: StatExplore Diagram: Project2

File Edit View Window

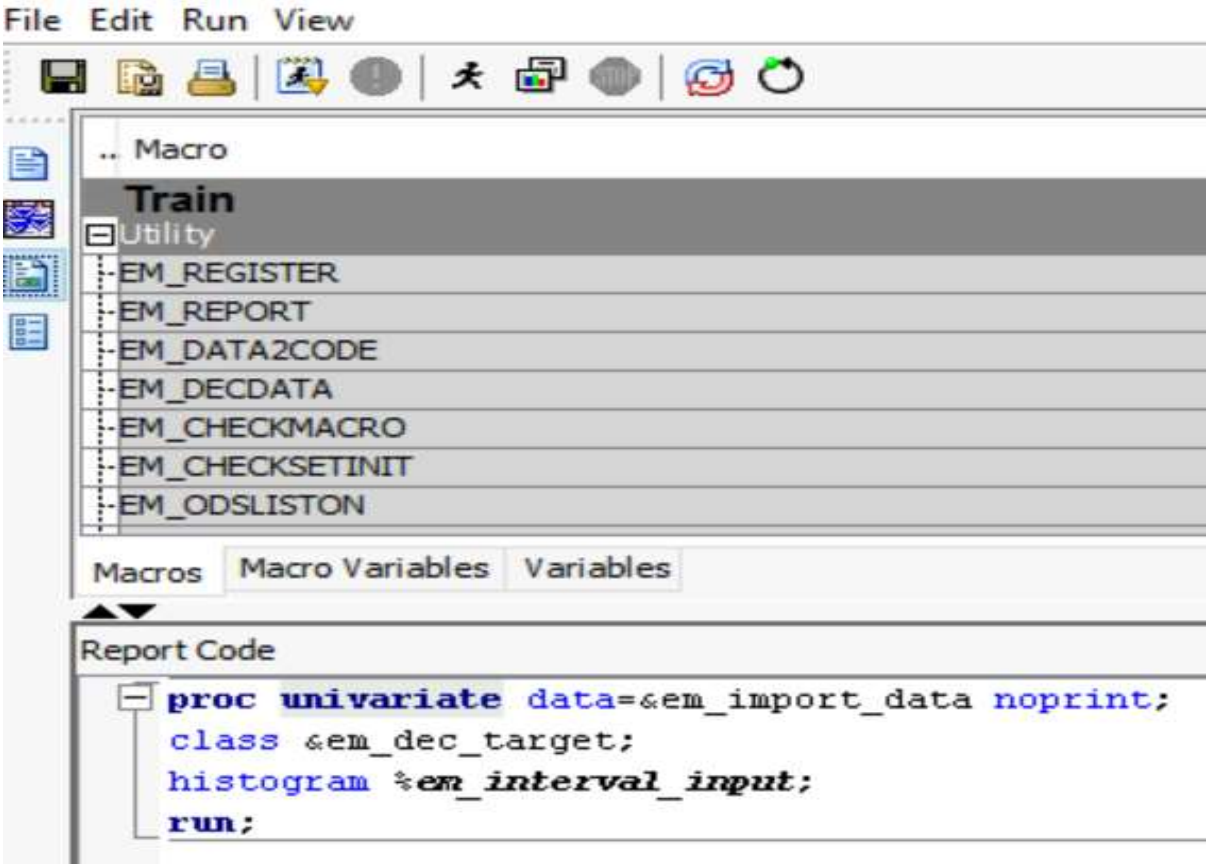
Interval Variables

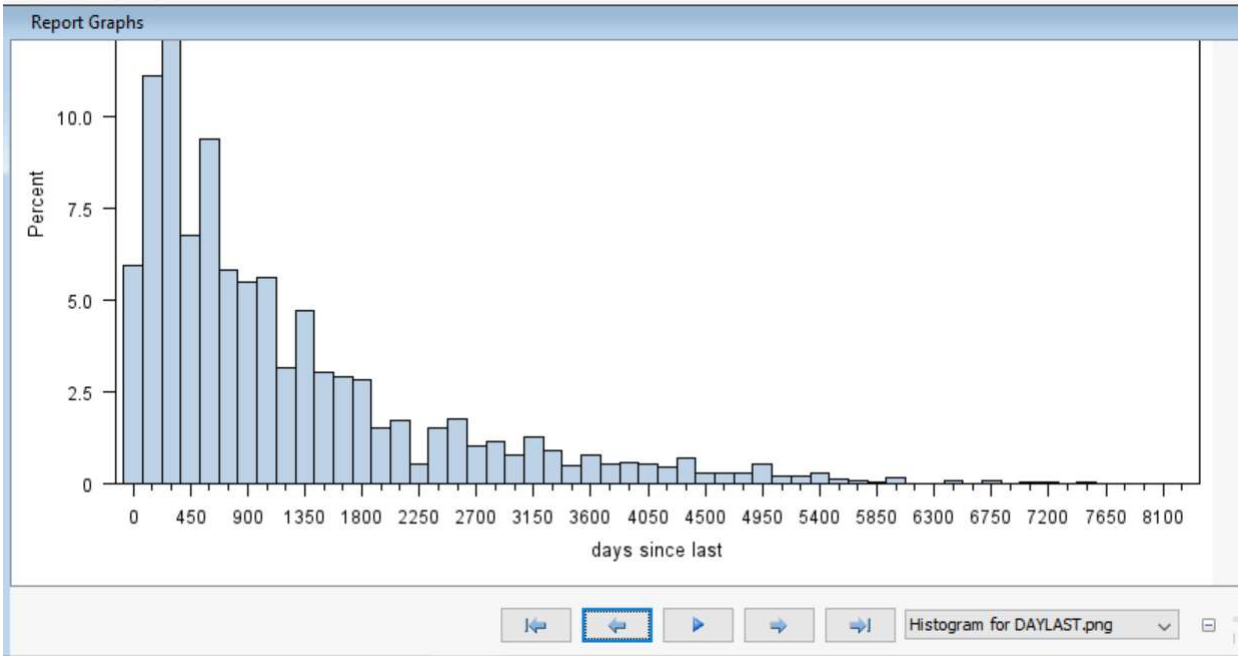
Data Role	Target	Target Level	Variable	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Role	Label	Scaled Mean Deviation	Ma: Dev
TRAIN	RESPOND	0	ORDERSIZE	0	0	45617	0	0	0	0	.	.	TARGET	dollar value...	-1	
TRAIN	RESPOND	1	ORDERSIZE	43.85	0	2739	5.4	510.72	62.77013	59.94363	2.771339	10.80182	TARGET	dollar value...	16.65462	
TRAIN	RESPOND	0	DOLLARQ22	0	0	45617	0	670.2	4.49098	22.46453	9.046402	128.1413	INPUT	tot $ 98Q2	-0.08609	
TRAIN	RESPOND	1	DOLLARQ22	0	0	2739	0	1128.75	11.95974	46.44084	9.822622	161.0246	INPUT	tot $ 98Q2	1.433795	
TRAIN	RESPOND	0	TOTORDQ22	0	0	45617	0	4	0.081154	0.296624	3.982015	18.54265	INPUT	tot orders 9...	-0.07837	
TRAIN	RESPOND	1	TOTORDQ22	0	0	2739	0	6	0.202994	0.522376	3.47717	17.47316	INPUT	tot orders 9...	1.3053	
TRAIN	RESPOND	0	DEPT25	0	0	45617	0	186	1.634829	4.356396	8.898405	177.3228	INPUT	food	-0.07335	
TRAIN	RESPOND	1	DEPT25	0	0	2739	0	126	3.919314	9.162664	5.516103	44.46609	INPUT	food	1.221546	
TRAIN	RESPOND	0	DOLLARQ20	0	0	45617	0	768.85	8.064814	30.50179	7.515434	92.45299	INPUT	tot $ 97Q4	-0.07293	
TRAIN	RESPOND	1	DOLLARQ20	0	0	2739	0	908.5	19.26501	51.52793	6.036659	62.18658	INPUT	tot $ 97Q4	1.214567	
TRAIN	RESPOND	0	DOLLARQ18	0	0	45617	0	752.1	4.612252	22.39056	9.339538	144.8153	INPUT	tot $ 97Q2	-0.07228	
TRAIN	RESPOND	1	DOLLARQ18	0	0	2739	0	486.15	10.95618	36.27364	5.370204	39.4762	INPUT	tot $ 97Q2	1.20376	
TRAIN	RESPOND	0	TOTORDQ20	0	0	45617	0	22	0.148782	0.423859	6.257399	174.5595	INPUT	tot orders 9...	-0.06903	
TRAIN	RESPOND	1	TOTORDQ20	0	0	2739	0	27	0.343556	0.849894	13.10508	362.1374	INPUT	tot orders 9...	1.149715	
TRAIN	RESPOND	0	DOLL24	0	0	45617	0	2140.1	42.60035	86.8664	5.45256	57.27983	INPUT	$ last 24 m...	-0.06764	
TRAIN	RESPOND	1	DOLL24	43.9	0	2739	0	2433.5	97.16399	168.497	5.069334	42.82712	INPUT	$ last 24 m...	1.126547	
TRAIN	RESPOND	0	DOLLARQ21	0	0	45617	0	1084.3	4.432568	23.1125	12.7998	328.1488	INPUT	tot $ 98Q1	-0.06547	
TRAIN	RESPOND	1	DOLLARQ21	0	0	2739	0	542.05	9.914816	34.62575	6.405912	58.5225	INPUT	tot $ 98Q1	1.090368	
TRAIN	RESPOND	0	DOLLARQ19	0	0	45617	0	559.2	4.171569	21.59825	9.664066	140.6688	INPUT	tot $ 97Q3	-0.06279	
TRAIN	RESPOND	1	DOLLARQ19	0	0	2739	0	677.88	9.105451	33.8163	7.024176	81.5345	INPUT	tot $ 97Q3	1.045693	
TRAIN	RESPOND	0	TOTORDQ18	0	0	45617	0	5	0.089046	0.319699	4.142542	21.77453	INPUT	tot orders 9...	-0.0623	

We moved this table into Excel to calculate the Q1 & Q3, IQR and outlier values for each variable.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	
1	Data Role	Target	Target Lev	Variable	Median	Missing	Non Missi	Minimum	Maximum	Mean	Standard I	Q1	Q3	IRQ	Q1 - 1.5(IQR)	Q3 + 1.5(IQR)	Skewness	K
2	TRAIN	RESPOND	0	ACTBUY	1	0	45617	0	10	0.951597	1.1202	-0.1686	2.071797	2.2404	-3.529202145	5.43239613	1.864435	5
3	TRAIN	RESPOND	1	ACTBUY	1	0	2739	0	10	1.52647	1.611549	-0.08508	3.138018	3.223098	-4.919725601	7.97266463	1.559801	2
4	TRAIN	RESPOND	0	BUYPROP	0.1	0	45617	0	1	0.18395	0.253858	-0.06991	0.437808	0.507716	-0.831481204	1.199381951	1.91111	3
5	TRAIN	RESPOND	1	BUYPROP	0.181818	0	2739	0	1	0.265676	0.293457	-0.02778	0.559133	0.586913	-0.908149761	1.439502674	1.278582	0
6	TRAIN	RESPOND	0	CATALOGCNT	3	0	45617	1	24	3.678453	3.052072	0.626381	6.730525	6.104144	-8.529834304	15.88674072	1.591797	2
7	TRAIN	RESPOND	1	CATALOGCNT	4	0	2739	1	27	5.220884	3.99266	1.228224	9.213543	7.985319	-10.74975524	21.19152231	1.260401	1
8	TRAIN	RESPOND	0	DAYLAST	782	0	45617	0	8265	1204.939	1236.789	-31.8496	2441.728	2473.578	-3742.216042	6152.094552	1.741862	3
9	TRAIN	RESPOND	1	DAYLAST	362	0	2739	0	7859	759.7401	981.9968	-222.257	1741.737	1963.994	-3168.24702	4687.727122	2.517462	
10	TRAIN	RESPOND	0	DEPT01	0	0	45617	0	59	0.469386	1.686947	-1.21756	2.156333	3.373893	-6.278399852	7.217172678	7.020023	9
11	TRAIN	RESPOND	1	DEPT01	0	0	2739	0	49	0.917853	3.032121	-2.11427	3.949975	6.064243	-11.21063275	13.04633921	6.733124	6
12	TRAIN	RESPOND	0	DEPT02	0	0	45617	0	24	0.275928	1.101996	-0.82607	1.377924	2.203993	-4.132057982	4.68391365	5.496759	4
13	TRAIN	RESPOND	1	DEPT02	0	0	2739	0	19	0.564074	1.76376	-1.19969	2.327834	3.52752	-6.490965289	7.619114249	4.44541	2
14	TRAIN	RESPOND	0	DEPT03	0	0	45617	0	60	1.025692	2.662534	-1.63684	3.688227	5.325069	-9.624445181	11.67582953	4.713144	3
15	TRAIN	RESPOND	1	DEPT03	0	0	2739	0	54	2.085433	4.697419	-2.61199	6.782852	9.394838	-16.70424309	20.87510837	4.338569	2
16	TRAIN	RESPOND	0	DEPT04	0	0	45617	0	45	0.645921	2.200775	-1.55485	2.846696	4.401549	-8.157176645	9.449019599	6.159116	6
17	TRAIN	RESPOND	1	DEPT04	0	0	2739	0	47	1.396495	3.791373	-2.39488	5.187868	7.582746	-13.76899695	16.56198709	4.582803	2
18	TRAIN	RESPOND	0	DEPT05	0	0	45617	0	28	0.520552	1.479894	-0.95934	2.000446	2.959788	-5.399024703	6.4401278	3.882806	
19	TRAIN	RESPOND	1	DEPT05	0	0	2739	0	19	0.874407	2.014991	-1.14058	2.889398	4.029982	-7.185556981	8.934370416	3.080609	1
20	TRAIN	RESPOND	0	DEPT06	0	0	45617	0	32	0.825591	1.917107	-1.09152	2.742698	3.834213	-6.842835614	8.494018287	3.328319	1
21	TRAIN	RESPOND	1	DEPT06	0	0	2739	0	29	1.241329	2.520296	-1.27897	3.761625	5.040592	-8.839855472	11.32251338	3.423343	2

We compared this info to the Univariate Histogram for each variable generated from SAS Node Code.





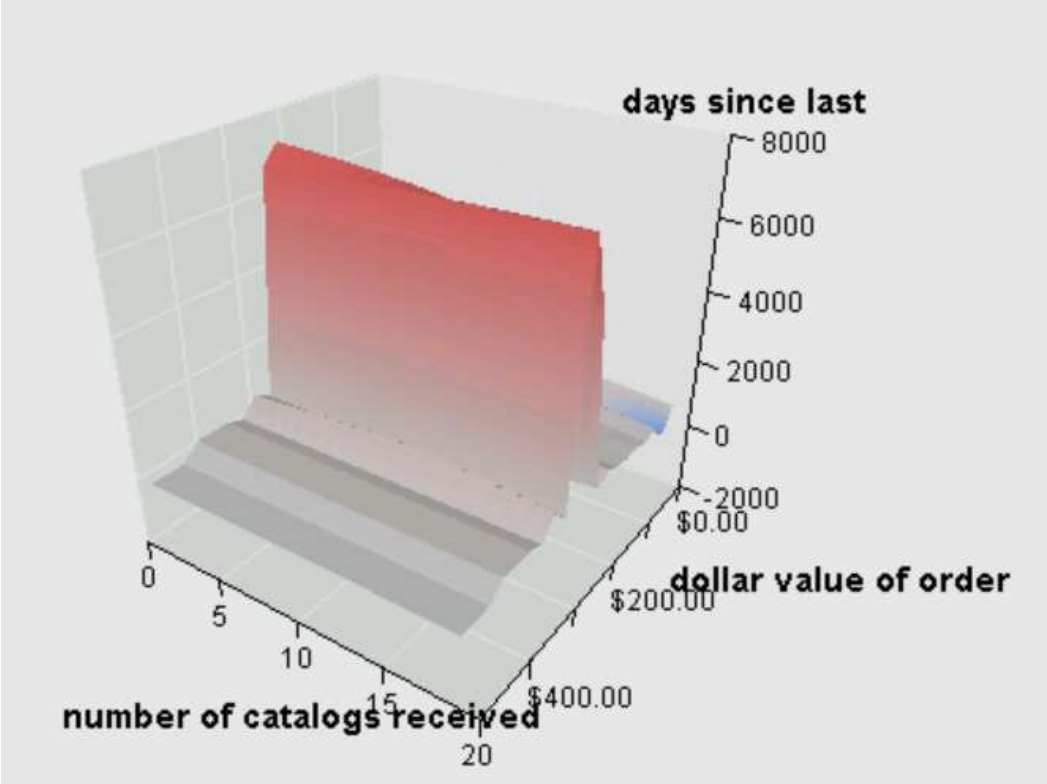
Using the table and histograms for each variable, we were able to determine which variables have outliers that need to be removed.

IF ACTBUY > 7.97266463 then delete;
IF DAYLAST > 6152.094552 then delete;
IF DEPT05 > 8.934370416 then delete;
IF DEPT07 > 1.999186007 then delete;
IF DEPT10 > 11.19664554 then delete;
IF DEPT11 > 4.347513557 then delete;
IF DEPT15 > 6.712963144 then delete;
IF DEPT16 > 6.355999274 then delete;
IF DEPT18 > 3.775612457 then delete;
IF DEPT19 > 3.556790332 then delete;
IF DEPT20 > 1.63582488 then delete;
IF DEPT21 > 1.21526434 then delete;
IF DEPT27 > 8.91633293 then delete;
IF DOLINDEA > 199.4817483 then delete;
IF DOLLARQ01 > 104.7142525 then delete;
IF DOLLARQ04 > 112.6889834 then delete;
IF DOLLARQ05 > 119.9368083 then delete;
IF DOLLARQ06 > 108.8219948 then delete;
IF DOLNETDA > 191.7379656 then delete;
IF MONLAST > 202.0299063 then delete;
IF TOTORDQ01 > 1.847747185 then delete;
IF TOTORDQ03 > 1.373660894 then delete;
IF TOTORDQ05 > 1.958658868 then delete;
IF TOTORDQ09 > 1.813416952 then delete;
IF TOTORDQ10 > 1.86231867 then delete;
IF TOTORDQ12 > 2.48326094 then delete;
IF TOTORDQ13 > 1.985993095 then delete;
IF TOTORDQ14 > 1.784962061 then delete;
IF TOTORDQ16 > 2.658483902 then delete;
IF TOTORDQ17 > 2.261503116 then delete;
IF TOTORDQ19 > 1.964440261 then delete;
IF UNITSLAP > 105.2575375 then delete;
IF UNTLANPO > 12.00516903 then delete;

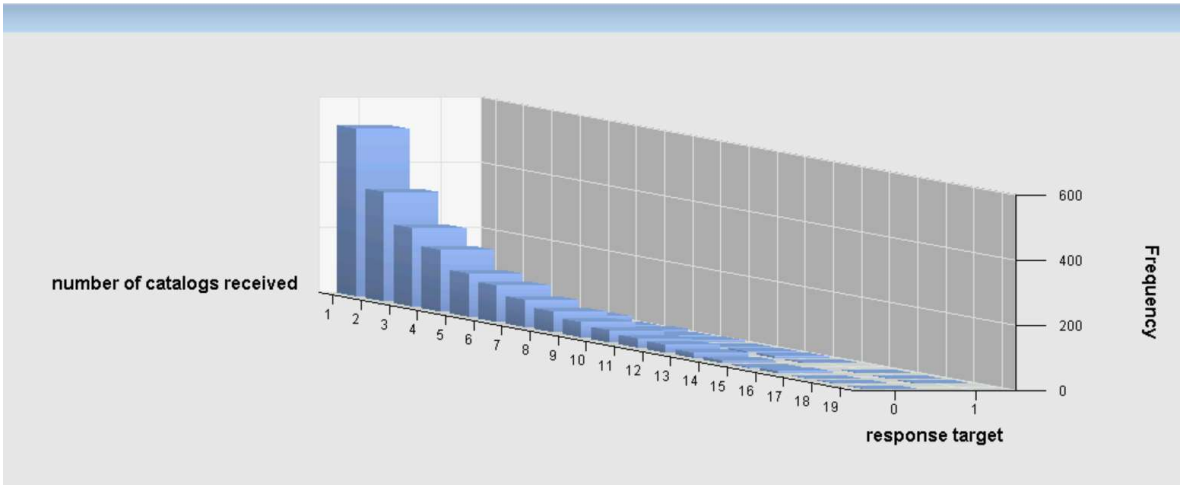
Report Code

```
data ExportedCatalog2010; set &EM_IMPORT_DATA;
IF ACTBUY > 7.97266463 then delete;
IF DAYLAST > 6152.094552 then delete;
IF DEPT05 > 8.934370416 then delete;
IF DEPT07 > 1.999186007 then delete;
IF DEPT10 > 11.19664554 then delete;
IF DEPT11 > 4.347513557 then delete;
IF DEPT15 > 6.712963144 then delete;
```

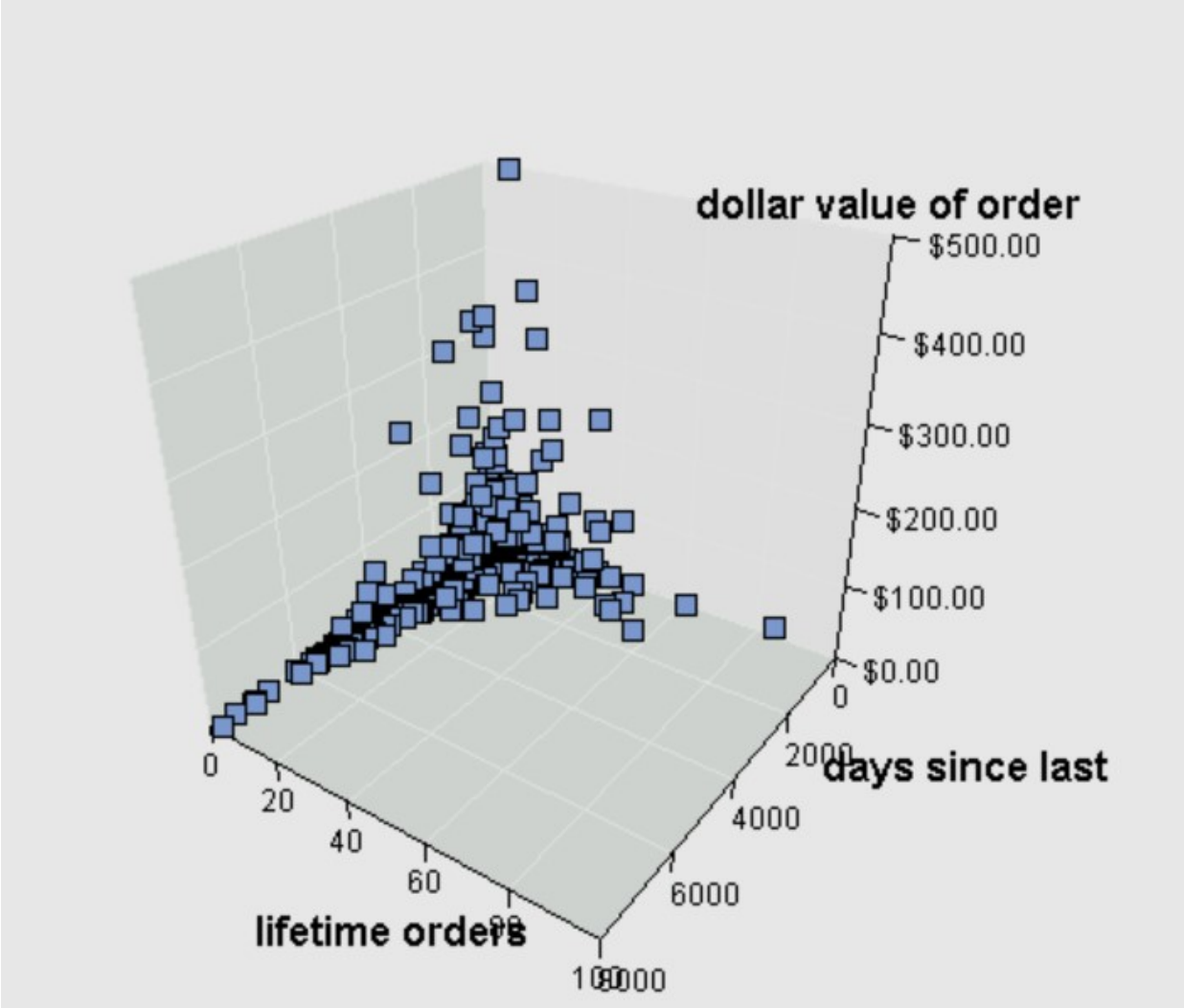
3) You can make 3-D plots to get a better sense of how independent variables affect the dependent variables **6points**



This 3D surface area graph shows the relationship between Number of catalogs received, days since last purchase and the target variable of dollar value of order. When the days since last order increases part 6000, then dollar value of the order approaches $200.00.



This bar chart with both a category and series variable indicates that most customers received 1 catalog. The number of catalogs received does not increase the response target.



This scatter plot in 3 dimensions indicates that the dollar value of the order does not increase

with the number of lifetime order or the days since last order. Customers with lifetime orders > 30 tend to have ordered in the last 2000 days. The highest dollar value orders are from customers in the lower half of lifetime number of order and the lower half of days since last order. This informs us that order dollar value does not increase with the number of lifetime orders or the number of days since last order.

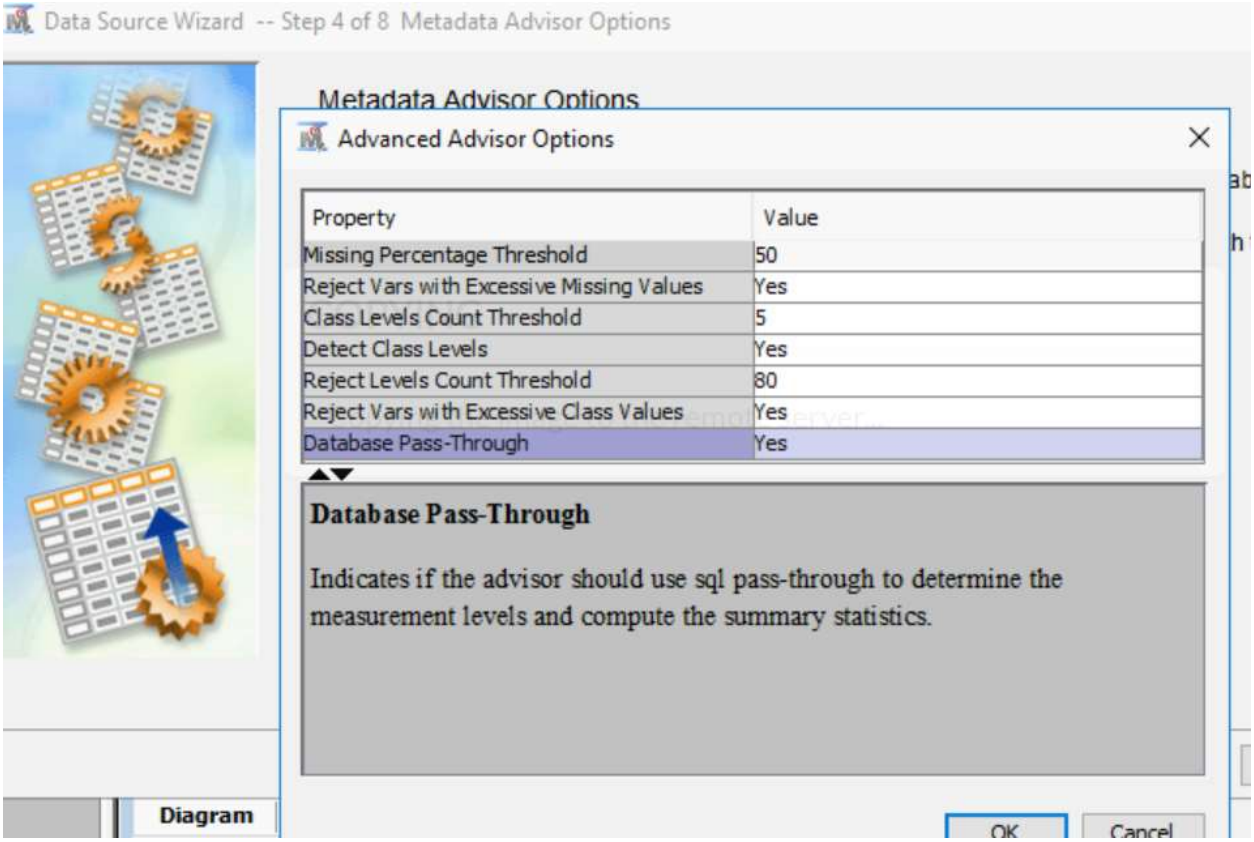
Task 2: 10 points

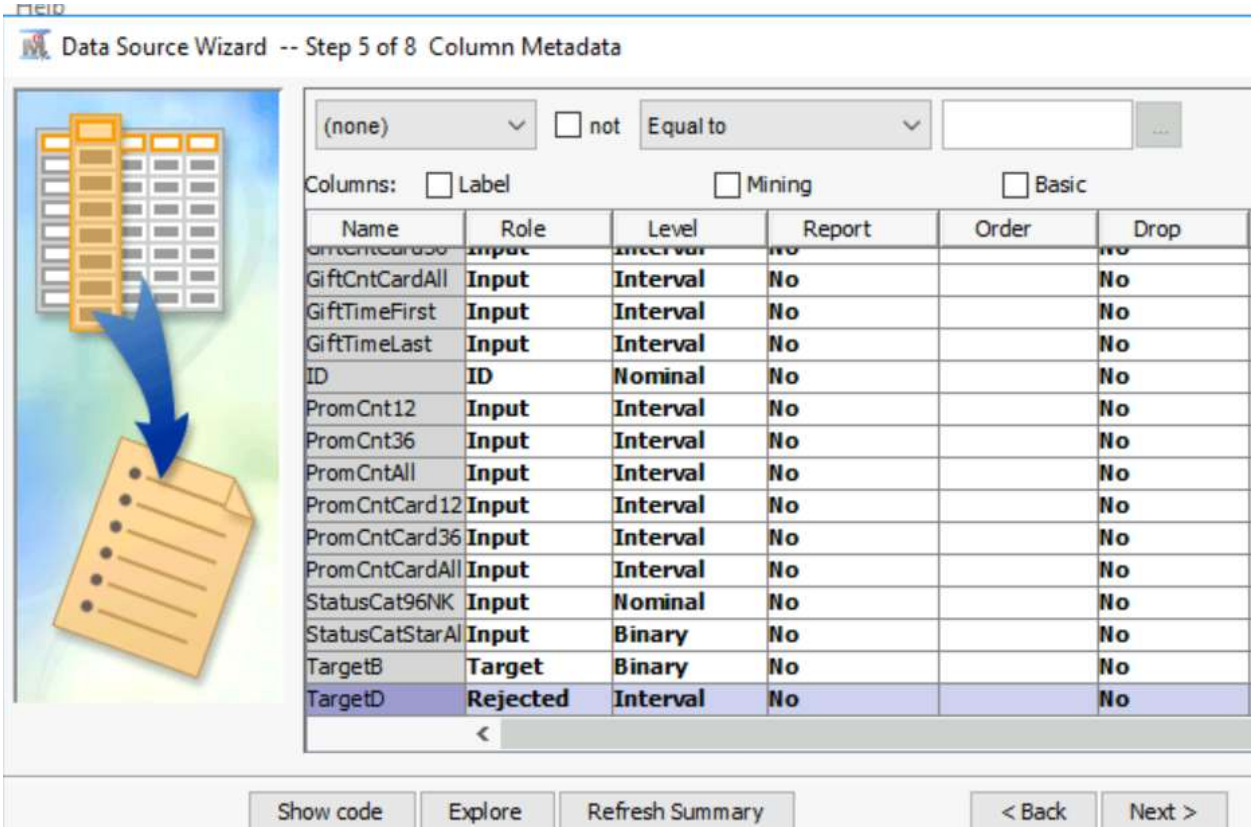
A national veterans' organization seeks to better target its solicitations for donation. By soliciting only, the most likely donors, less money is spent on solicitation efforts and more money is available for charitable concerns. Solicitations involve sending a small gift to an individual and including a request for a donation. Gifts to donors include mailing labels and greeting cards.

The organization has more than 3.5 million individuals in its mailing database. These individuals are classified by their response behaviors to previous solicitation efforts. Of particular interest is the class of individuals identified as *lapsing donors*. These individuals made their most recent donation between 12 and 24 months ago. The organization seeks to rank its lapsing donors based on their responses to a greeting card mailing sent in June of 1997. (The charity calls this the 97NK Campaign.) With this ranking, a decision can be made to either solicit or ignore a lapsing individual in the June 1998 campaign.

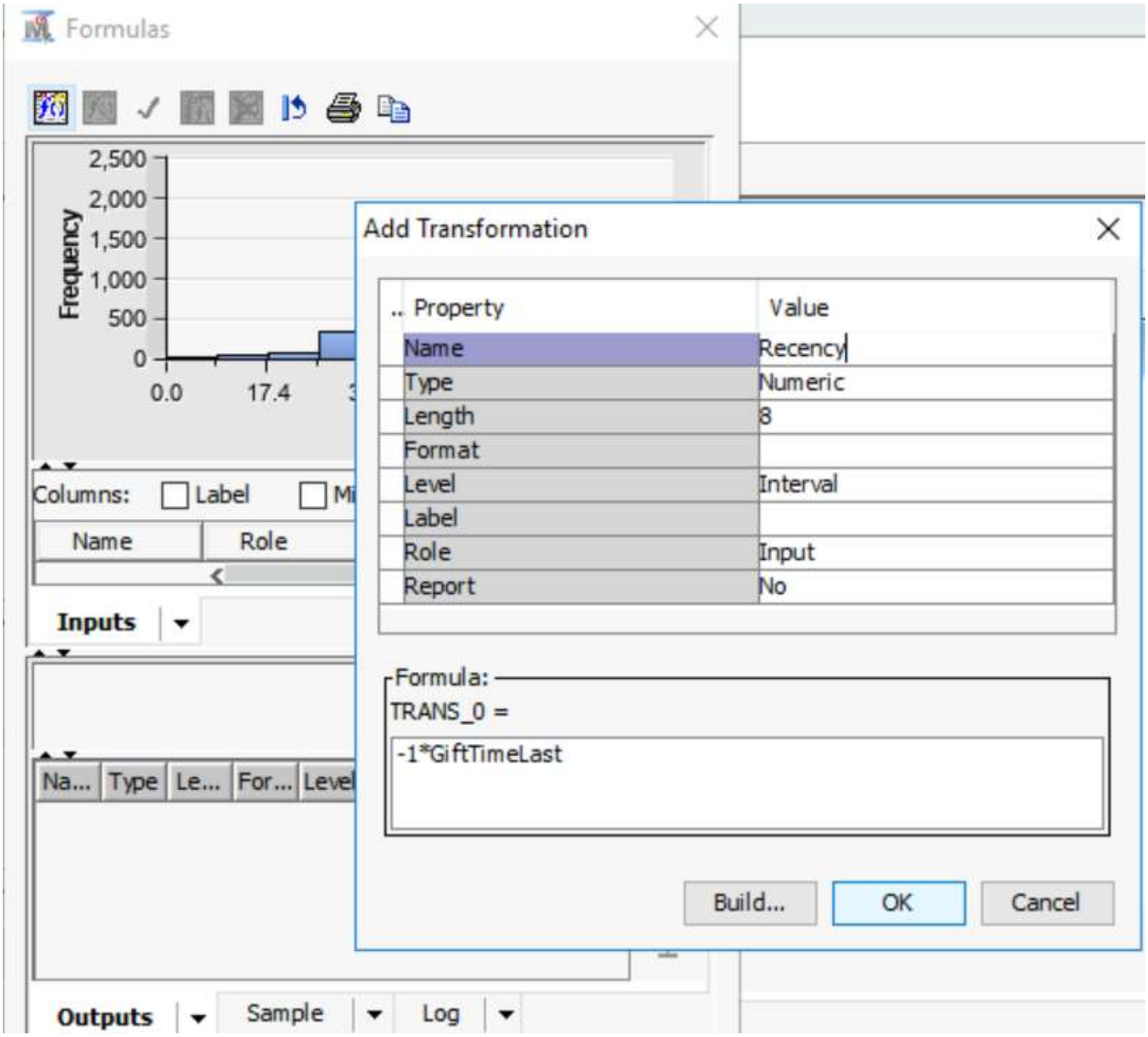
1. RFM Analysis of Charity Direct Mail Data

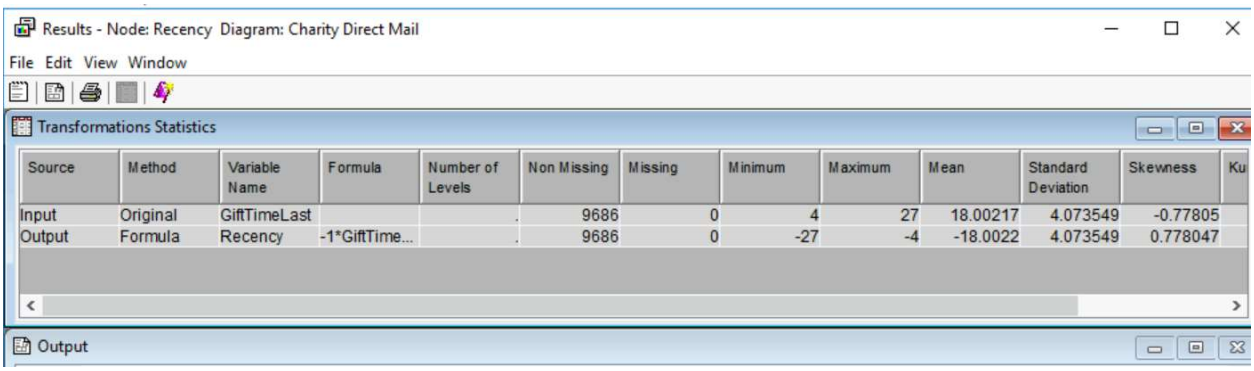
   a) Define PVA97NK as a data source in SAS Enterprise Miner. Use the Advanced Metadata Advisor options to customize the following:
   • Change the Class Levels Count Threshold from 20 to 5.
   • Change the Reject Levels Count Threshold from 20 to 80.
   • Reject the variable TargetD


Data Source Wizard -- Step 4 of 8 Metadata Advisor Options
Metadata Advisor Options
Advanced Advisor Options
Property
Value
Missing Percentage Threshold
50
Reject Vars with Excessive Missing Values
Yes
Class Levels Count Threshold
5
Detect Class Levels
Yes
Reject Levels Count Threshold
80
Reject Vars with Excessive Class Values
Yes
Database Pass-Through
Yes
Database Pass-Through
Indicates if the advisor should use sql pass-through to determine the measurement levels and compute the summary statistics.
Diagram
OK
Cancel

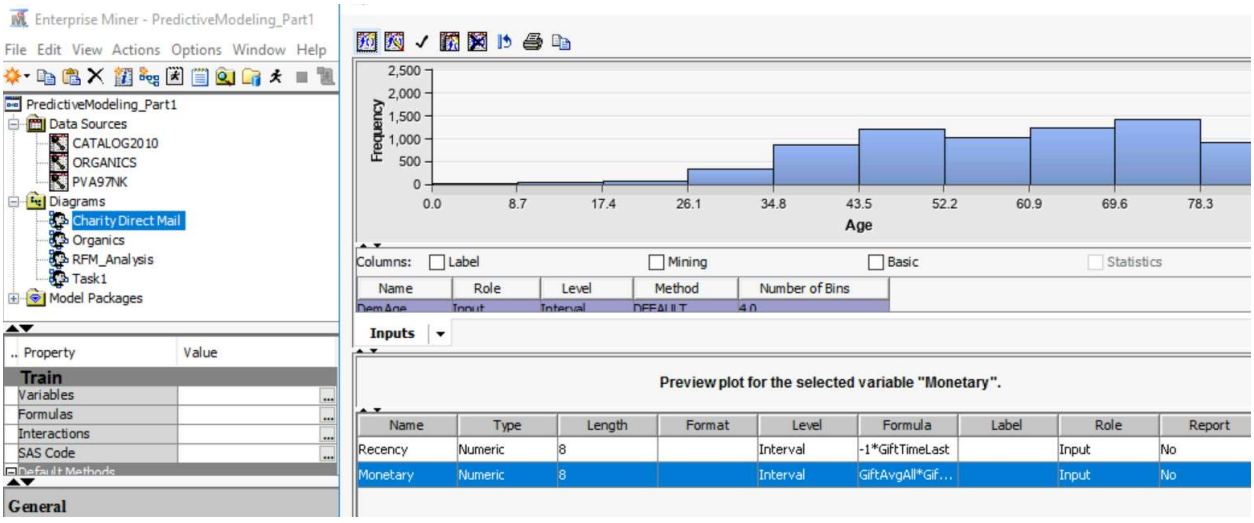

Data Source Wizard -- Step 5 of 8 Column Metadata
(none)
not
Equal to
Columns:
Label
Mining
Basic
Name
Role
Level
Report
Order
Drop
GiftCntCardAll Input Interval No No
GiftTimeFirst Input Interval No No
GiftTimeLast Input Interval No No
ID ID Nominal No No
PromCnt12 Input Interval No No
PromCnt36 Input Interval No No
PromCntAll Input Interval No No
PromCntCard12 Input Interval No No
PromCntCard36 Input Interval No No
PromCntCardAll Input Interval No No
StatusCat96NK Input Nominal No No
StatusCatStarAl Input Binary No No
TargetB Target Binary No No
TargetD Rejected Interval No No
Show code
Explore
Refresh Summary
< Back
Next >

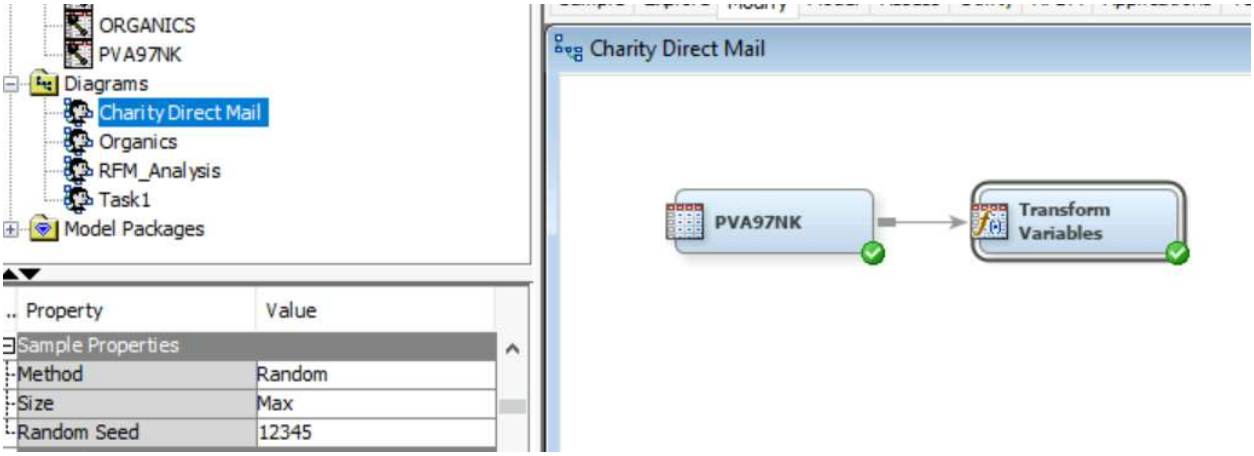
b) Create a new diagram and transform the R, F, and M variables as described previously to create four bins of each variable. Concatenate them to create an RFM variable.

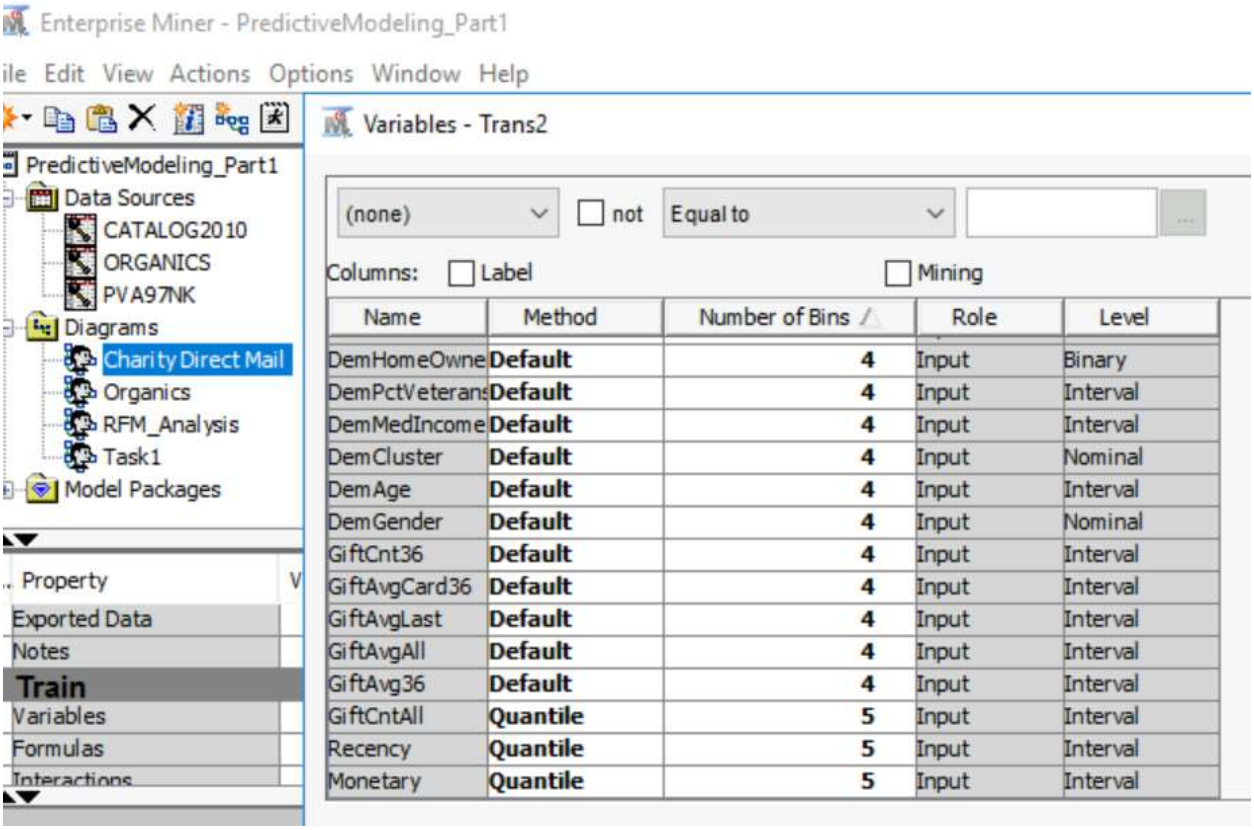


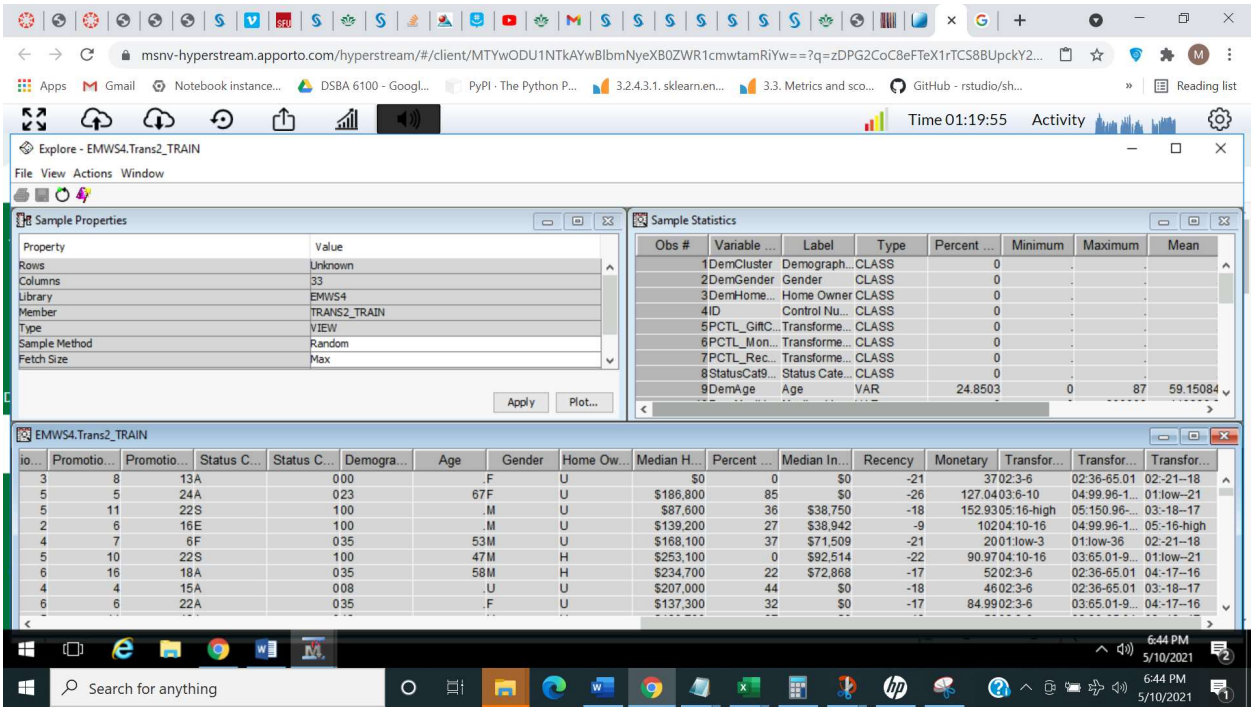


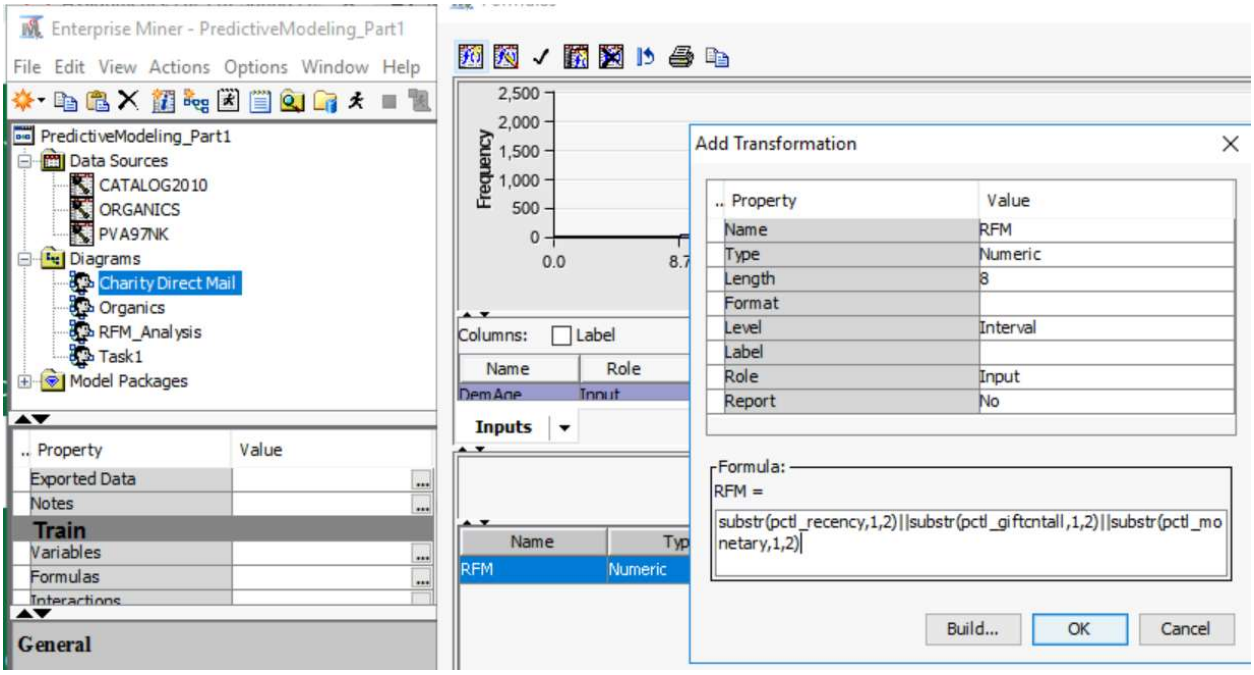
Source	Method	Variable Name	Formula	Number of Levels	Non Missing	Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Ku
Input	Original	GiftTimeLast		.	9686	0	4	27	18.00217	4.073549	-0.77805	
Output	Formula	Recency	-1*GiftTime...	.	9686	0	-27	-4	-18.0022	4.073549	0.778047	

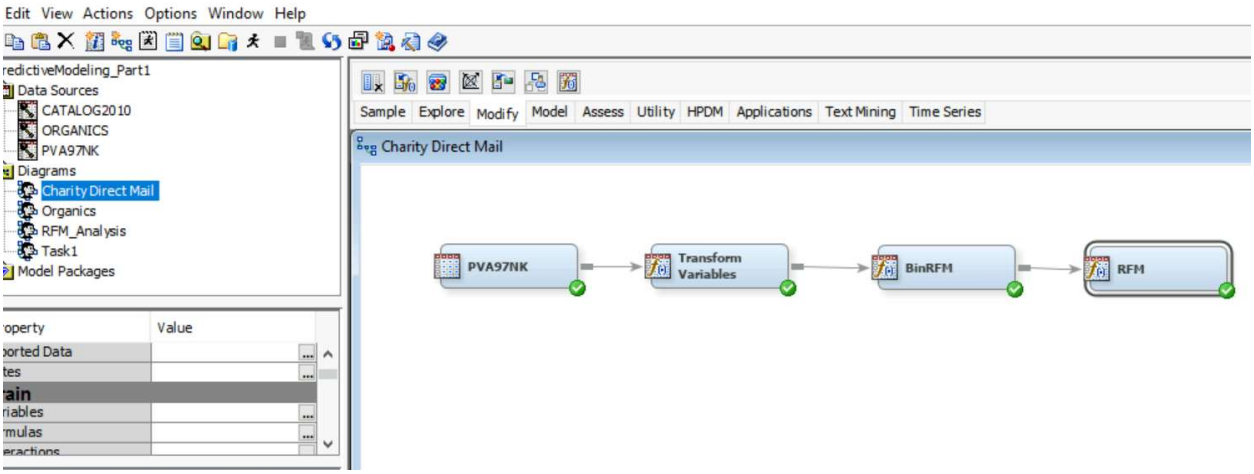

Enterprise Miner - PredictiveModeling_Part1
File Edit View Actions Options Window Help
PredictiveModeling_Part1
Data Sources
CATALOG2010
ORGANICS
PVA97NK
Diagrams
Charity Direct Mail
Organics
RFM_Analysis
Task1
Model Packages
Property
Value
Train
Variables
Formulas
Interactions
SAS Code
General
Frequency
Age
Columns:
Label
Mining
Basic
Statistics
Name
Role
Level
Method
Number of Bins
Inputs
Preview plot for the selected variable "Monetary".
Name
Type
Length
Format
Level
Formula
Label
Role
Report
Recency
Numeric
Interval
-1*GiftTimeLast
Input
No
Monetary
Numeric
Interval
GiftAvgAll*Gif...
Input
No


ORGANICS
PVA97NK
Diagrams
Charity Direct Mail
Organics
RFM_Analysis
Task1
Model Packages
Property
Value
Sample Properties
Method
Random
Size
Max
Random Seed
12345
Charity Direct Mail
PVA97NK
Transform Variables

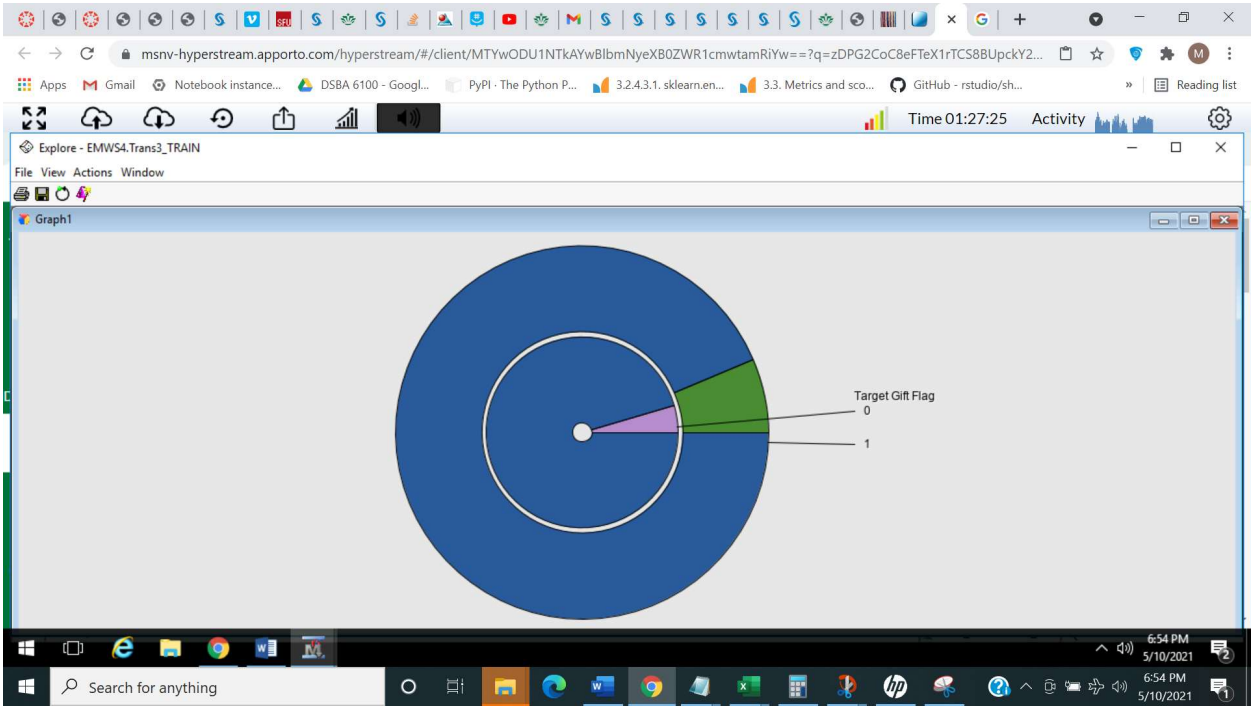

Enterprise Miner - PredictiveModeling_Part1
ile Edit View Actions Options Window Help
PredictiveModeling_Part1
Data Sources
CATALOG2010
ORGANICS
PVA97NK
Diagrams
Charity Direct Mail
Organics
RFM_Analysis
Task1
Model Packages
Property
Exported Data
Notes
Train
Variables
Formulas
Interactions
Variables - Trans2
(none)
not
Equal to
Columns:
Label
Mining
Name
Method
Number of Bins
Role
Level
DemHomeOwne Default 4 Input Binary
DemPctVeterans Default 4 Input Interval
DemMedIncome Default 4 Input Interval
DemCluster Default 4 Input Nominal
DemAge Default 4 Input Interval
DemGender Default 4 Input Nominal
GiftCnt36 Default 4 Input Interval
GiftAvgCard36 Default 4 Input Interval
GiftAvgLast Default 4 Input Interval
GiftAvgAll Default 4 Input Interval
GiftAvg36 Default 4 Input Interval
GiftCntAll Quantile 5 Input Interval
Recency Quantile 5 Input Interval
Monetary Quantile 5 Input Interval

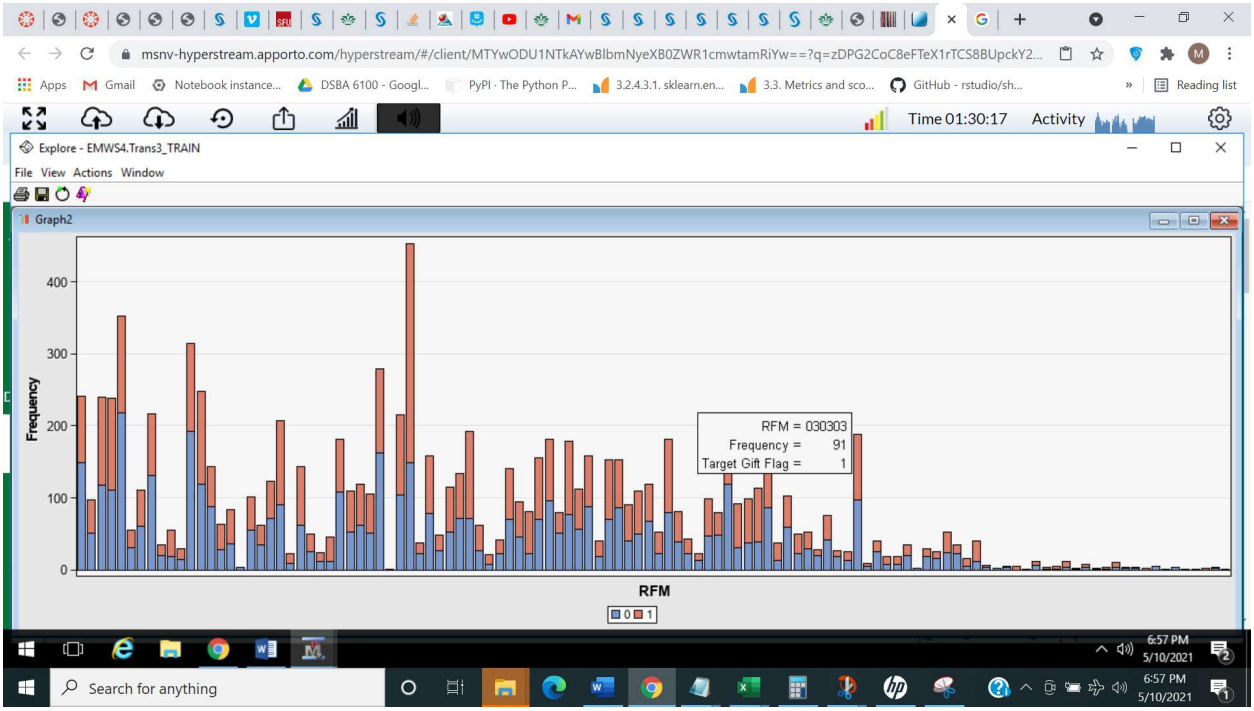

msnv-hyperstream.apporto.com/hyperstream/#/client/MTYwODU1NTkAYwBlbmNyeXB0ZWR1cmwtamRiYw==?q=zDPG2CoC8eFTeX1rTCS8BUpckY2...
Time 01:19:55
Activity
Explore - EMWS4.Trans2_TRAIN
File View Actions Window
Sample Properties
Property
Value
Rows
Unknown
Columns
33
Library
EMWS4
Member
TRANS2_TRAIN
Type
VIEW
Sample Method
Random
Fetch Size
Max
Apply
Plot...
Sample Statistics
Obs #
Variable ...
Label
Type
Percent ...
Minimum
Maximum
Mean
1 DemCluster Demograph... CLASS 0
2 DemGender Gender CLASS 0
3 DemHome... Home Owner CLASS 0
4 ID Control Nu... CLASS 0
5 PCTL_GiftC... Transforme... CLASS 0
6 PCTL_Mon... Transforme... CLASS 0
7 PCTL_Rec... Transforme... CLASS 0
8 StatusCat9... Status Cate... CLASS 0
9 DemAge Age VAR 24.8503 0 87 59.15084
EMWS4.Trans2_TRAIN
Search for anything
6:44 PM
5/10/2021


Enterprise Miner - PredictiveModeling_Part1
File Edit View Actions Options Window Help
PredictiveModeling_Part1
Data Sources
CATALOG2010
ORGANICS
PVA97NK
Diagrams
Charity Direct Mail
Organics
RFM_Analysis
Task1
Model Packages
Property
Value
Exported Data
Notes
Train
Variables
Formulas
Interactions
General
Frequency
2,500
2,000
1,500
1,000
500
0
0.0
Columns:
Label
Name
Role
Inputs
Add Transformation
Property
Value
Name
RFM
Type
Numeric
Length
8
Format
Level
Interval
Label
Role
Input
Report
No
Formula:
RFM =
substr(pctl_recency,1,2)||substr(pctl_giftcntall,1,2)||substr(pctl_monetary,1,2)
Build...
OK
Cancel


Edit View Actions Options Window Help
Data Sources
CATALOG2010
ORGANICS
PVA97NK
Diagrams
Charity Direct Mail
Organics
RFM_Analysis
Task1
Model Packages
Sample
Explore
Modify
Model
Assess
Utility
HPDM
Applications
Text Mining
Time Series
Charity Direct Mail
PVA97NK
Transform Variables
BinRFM
RFM

c) Explore the data and perform graphical RFM analysis using a grouped pie chart and a stacked bar chart.





d) Calculate response rate for 040404 and 030303 group?

040404 = 54/51+54 = 51.4%

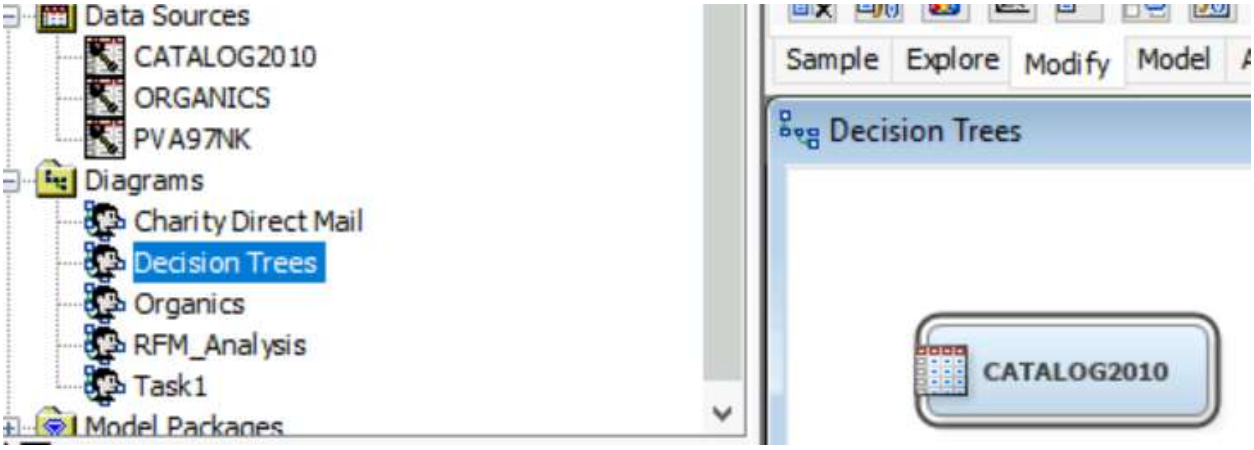
030303 = 91/97+91 = 48.4%

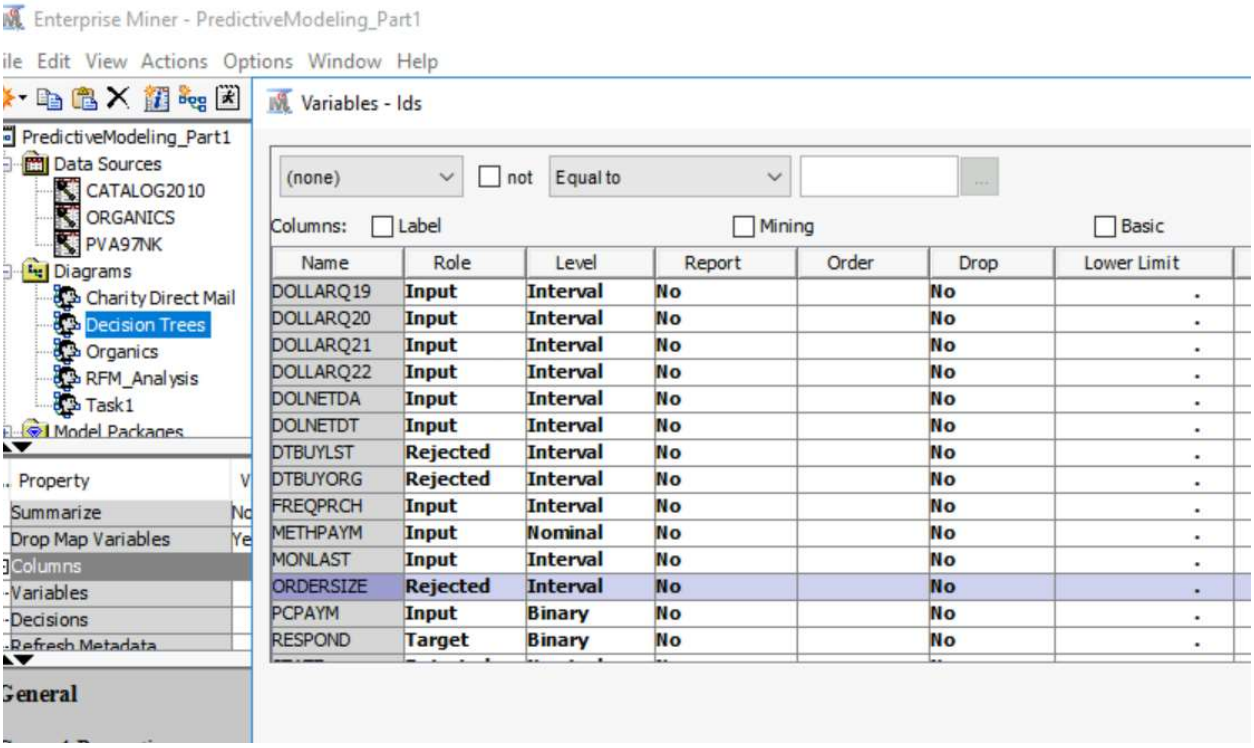
e) Each promotional mailing (request for a gift) costs $2.3, and the average donation is about $21 What is the break-even response rate for this promotion? Do any RFM cells exceed this response rate? Remember to account for the fact that in the population, 95% of mailings are not responded to, while this sample is oversampled to 50% responders and 50% non-responders.

Break-even is 10.95%. Many RFM cells exceed that rate including 040404 & 030303.

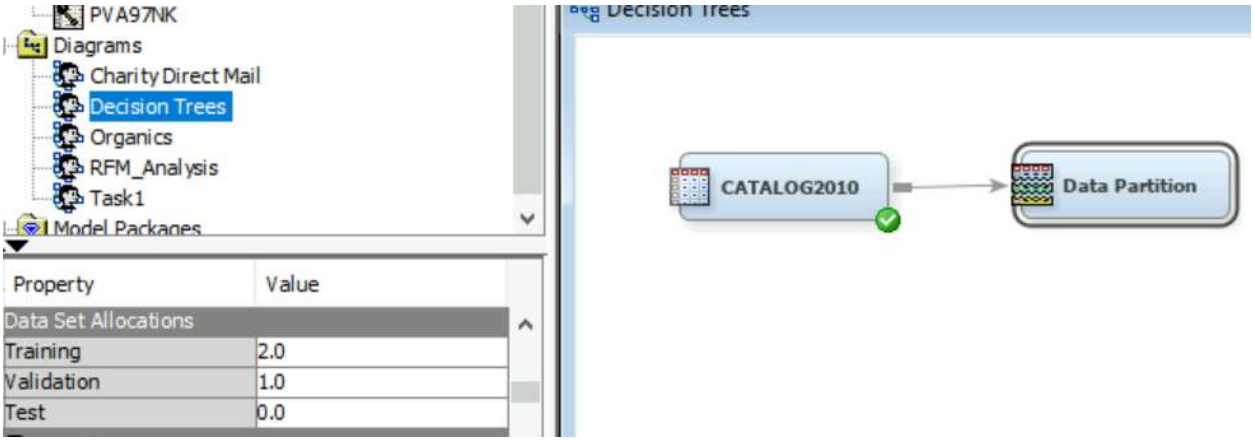
Part 2: Predictive Modeling: Decision Trees

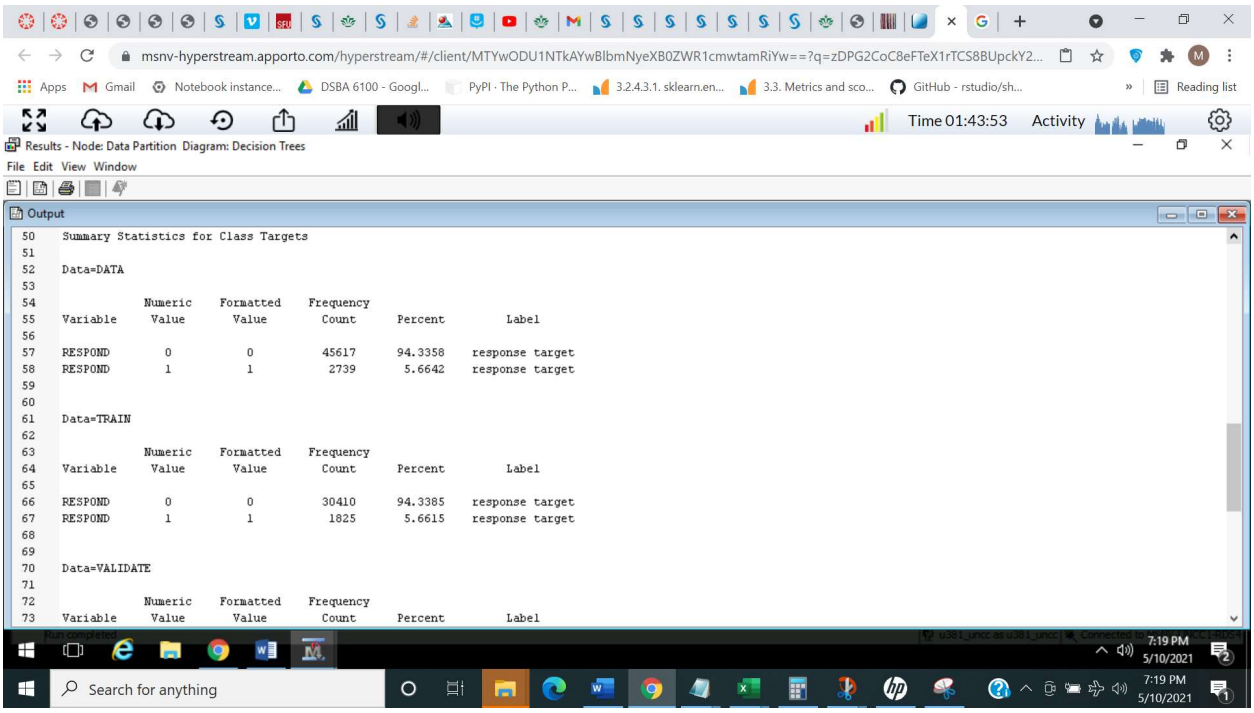
CATALOG CASE STUDY: CONSTRUCTING A DECISION TREE PREDICTIVE MODEL

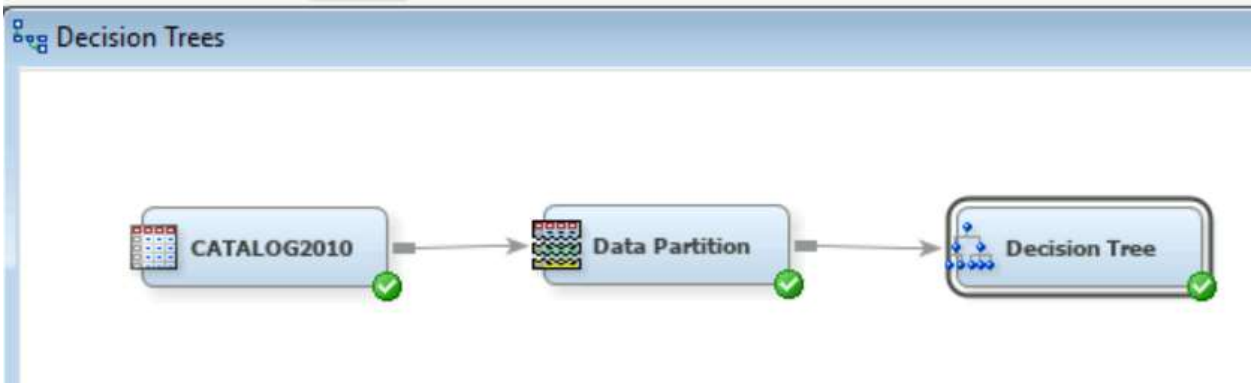


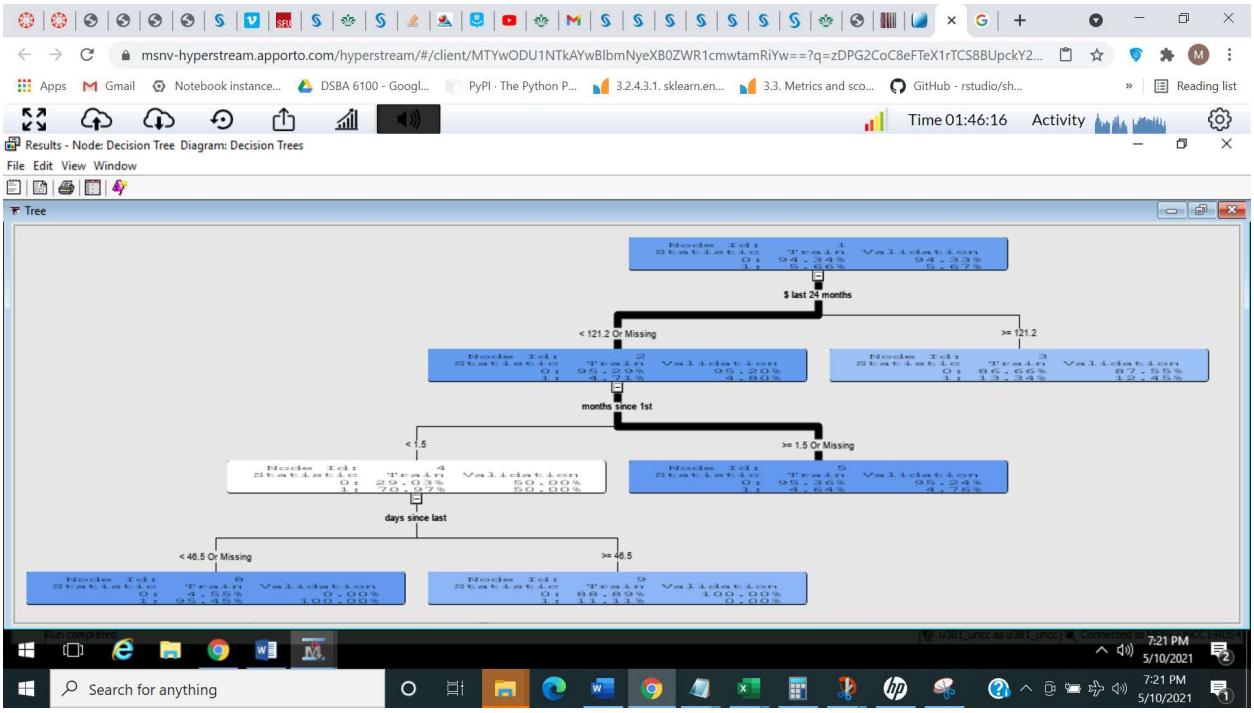


Name	Role	Level	Report	Order	Drop	Lower Limit
DOLLARQ19	Input	Interval	No		No	.
DOLLARQ20	Input	Interval	No		No	.
DOLLARQ21	Input	Interval	No		No	.
DOLLARQ22	Input	Interval	No		No	.
DOLNETDA	Input	Interval	No		No	.
DOLNETDT	Input	Interval	No		No	.
DTBUYLST	Rejected	Interval	No		No	.
DTBUYORG	Rejected	Interval	No		No	.
FREQPRCH	Input	Interval	No		No	.
METHPAYM	Input	Nominal	No		No	.
MONLAST	Input	Interval	No		No	.
ORDERSIZE	Rejected	Interval	No		No	.
PCPAYM	Input	Binary	No		No	.
RESPOND	Target	Binary	No		No	.

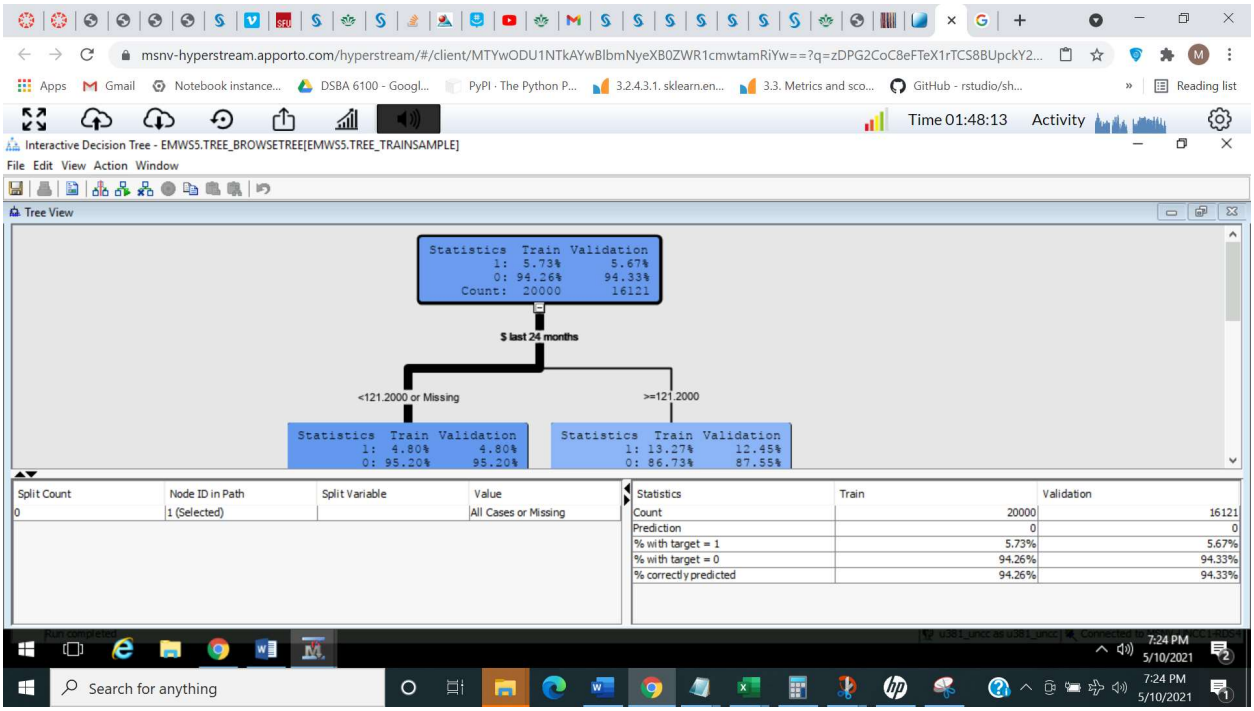

PVA97NK
Diagrams
Charity Direct Mail
Decision Trees
Organics
RFM_Analysis
Task1
Model Packages
Property
Value
Data Set Allocations
Training
2.0
Validation
1.0
Test
0.0
Decision Trees
CATALOG2010
Data Partition

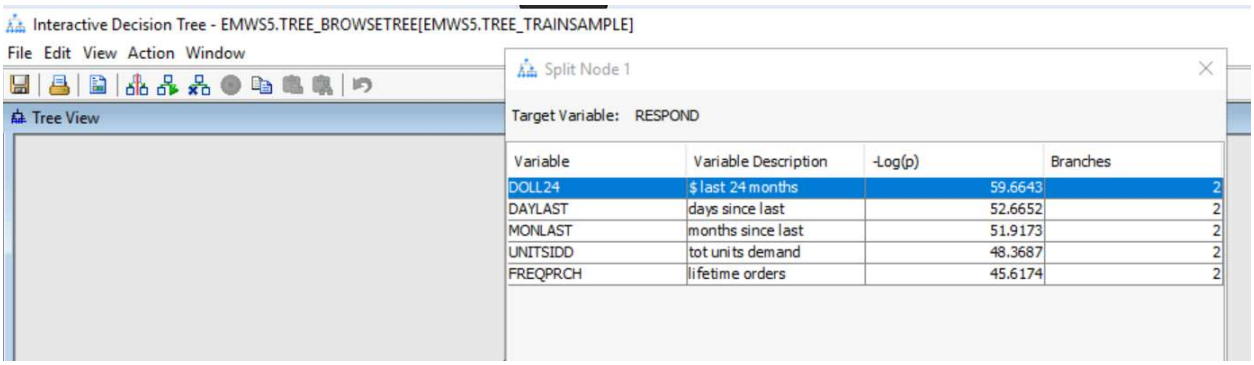

msnv-hyperstream.apporto.com/hyperstream/#/client/MTYwODU1NTkAYwBlbmNyeXB0ZWR1cmwtamRiYw==?q=zDPG2CoC8eFTeX1rTCS8BUpckY2...
Time 01:43:53
Activity
Results - Node: Data Partition Diagram: Decision Trees
File Edit View Window
Output
Summary Statistics for Class Targets
Data=DATA
Numeric Formatted Frequency
Variable Value Value Count Percent Label
RESPOND 0 0 45617 94.3358 response target
RESPOND 1 1 2739 5.6642 response target
Data=TRAIN
Numeric Formatted Frequency
Variable Value Value Count Percent Label
RESPOND 0 0 30410 94.3385 response target
RESPOND 1 1 1825 5.6615 response target
Data=VALIDATE
Numeric Formatted Frequency
Variable Value Value Count Percent Label
Search for anything
7:19 PM
5/10/2021

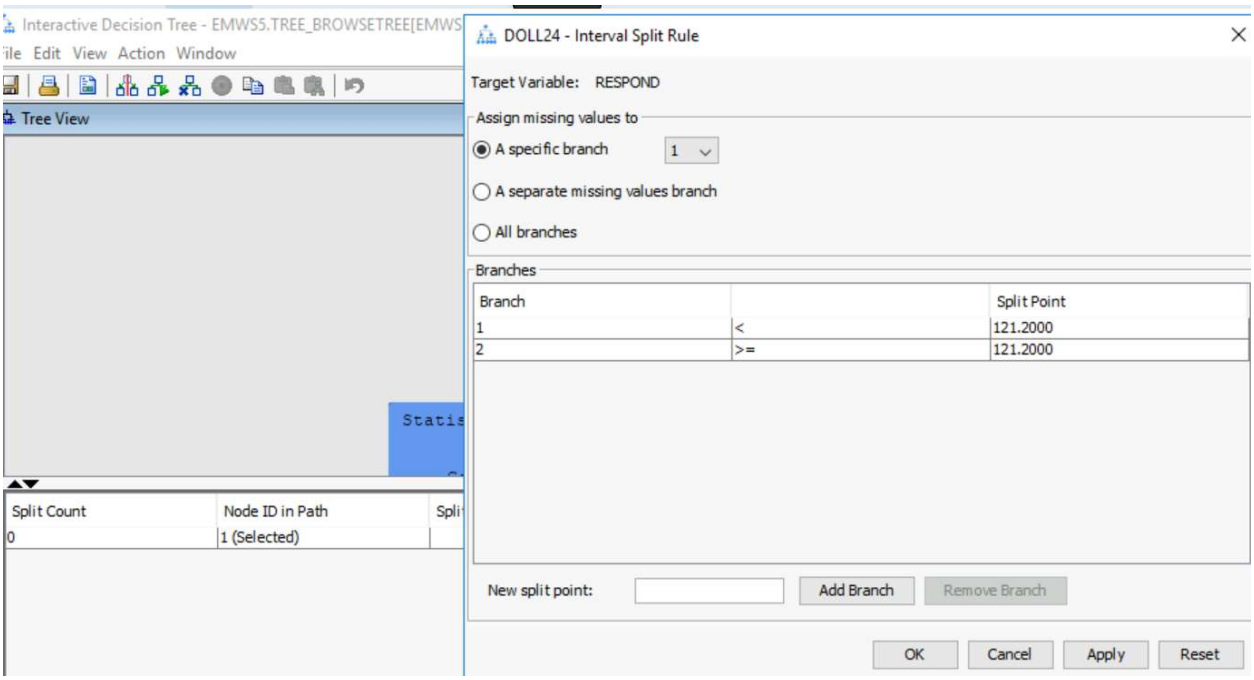

Decision Trees
CATALOG2010
Data Partition
Decision Tree

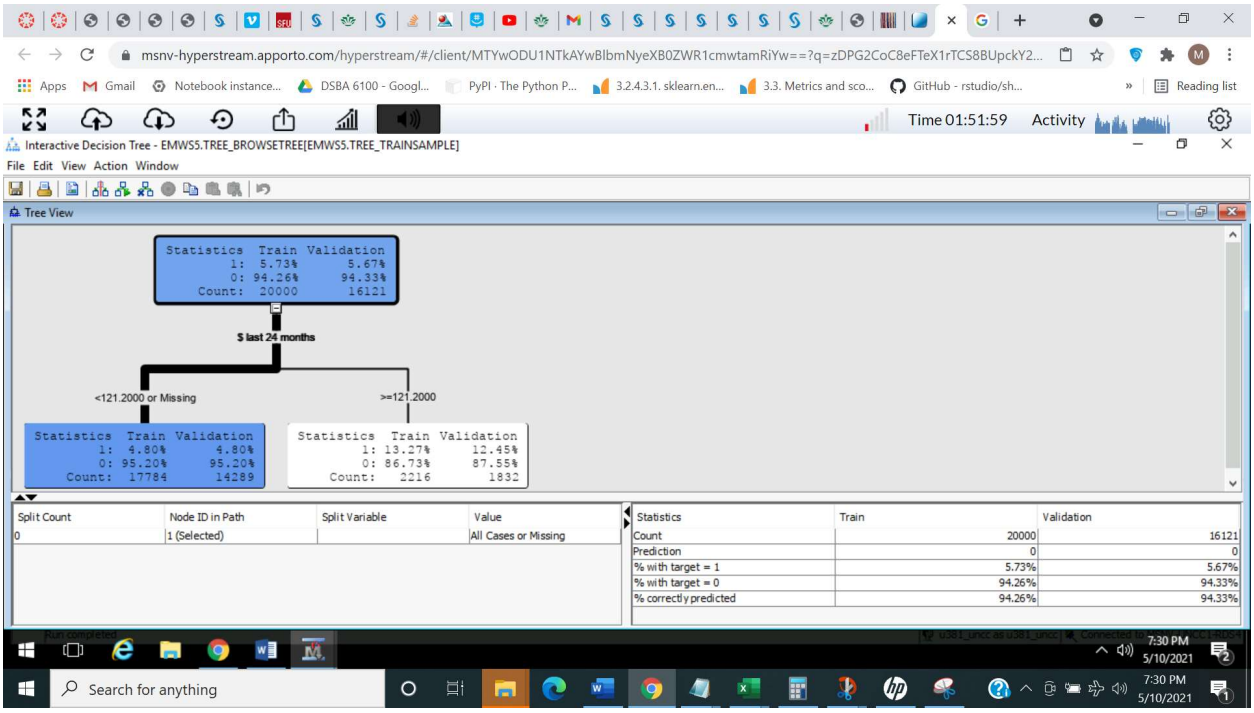


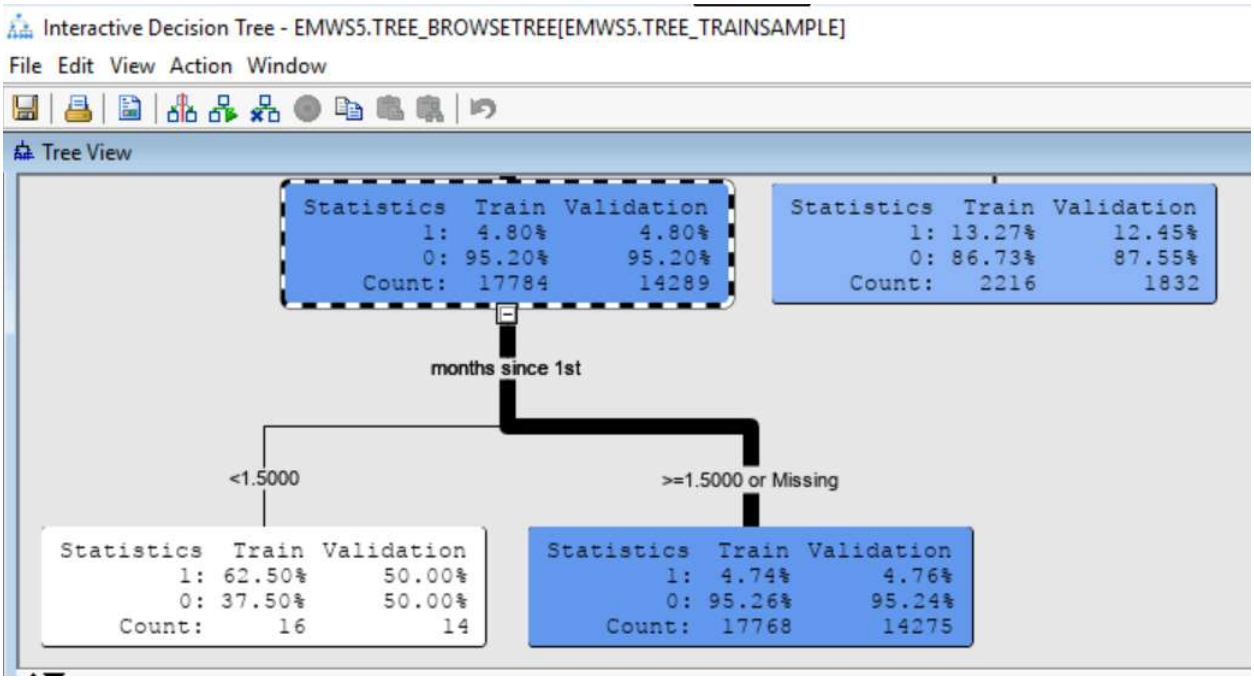
Interactive Decision Tree (Self-Study)

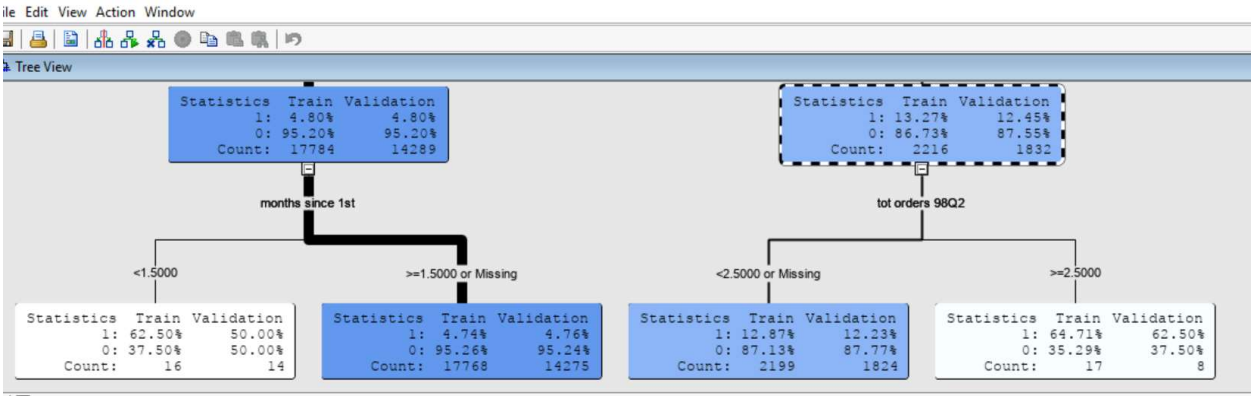


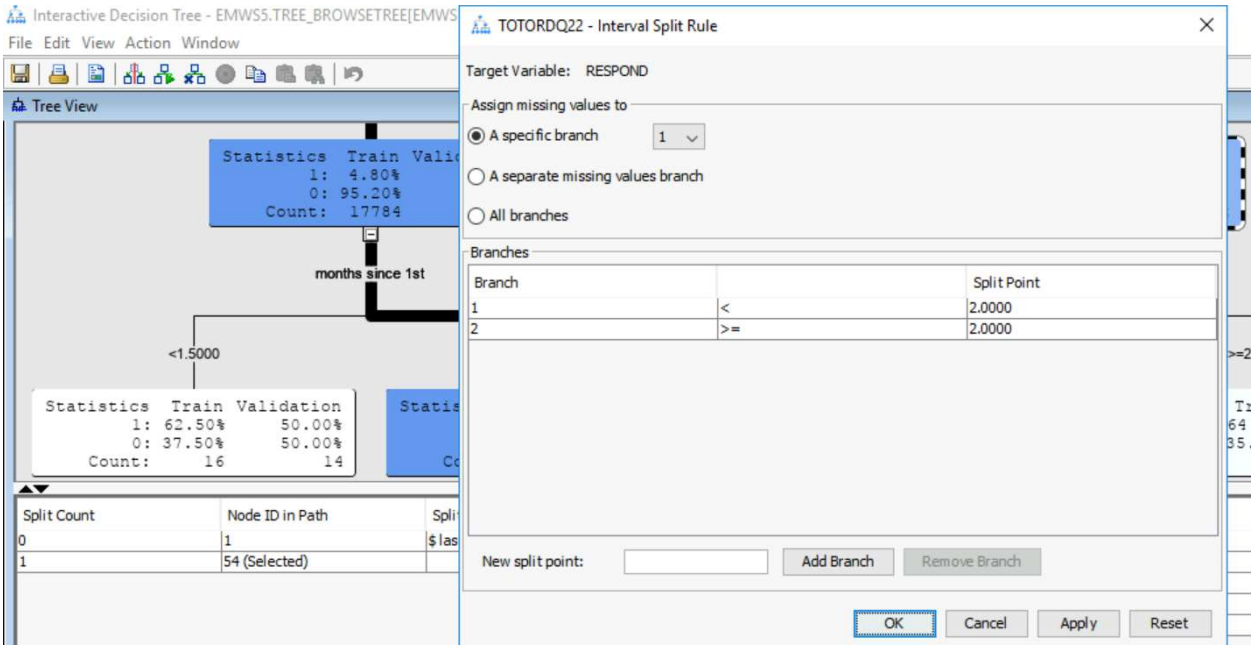

Interactive Decision Tree - EMWS5.TREE_BROWSETREE[EMWS5.TREE_TRAINSAMPLE]
File Edit View Action Window
Tree View
Split Node 1
Target Variable: RESPOND
Variable | Variable Description | -Log(p) | Branches
DOLL24 | $ last 24 months | 59.6643 | 2
DAYLAST | days since last | 52.6652 | 2
MONLAST | months since last | 51.9173 | 2
UNITSIDD | tot units demand | 48.3687 | 2
FREQPRCH | lifetime orders | 45.6174 | 2

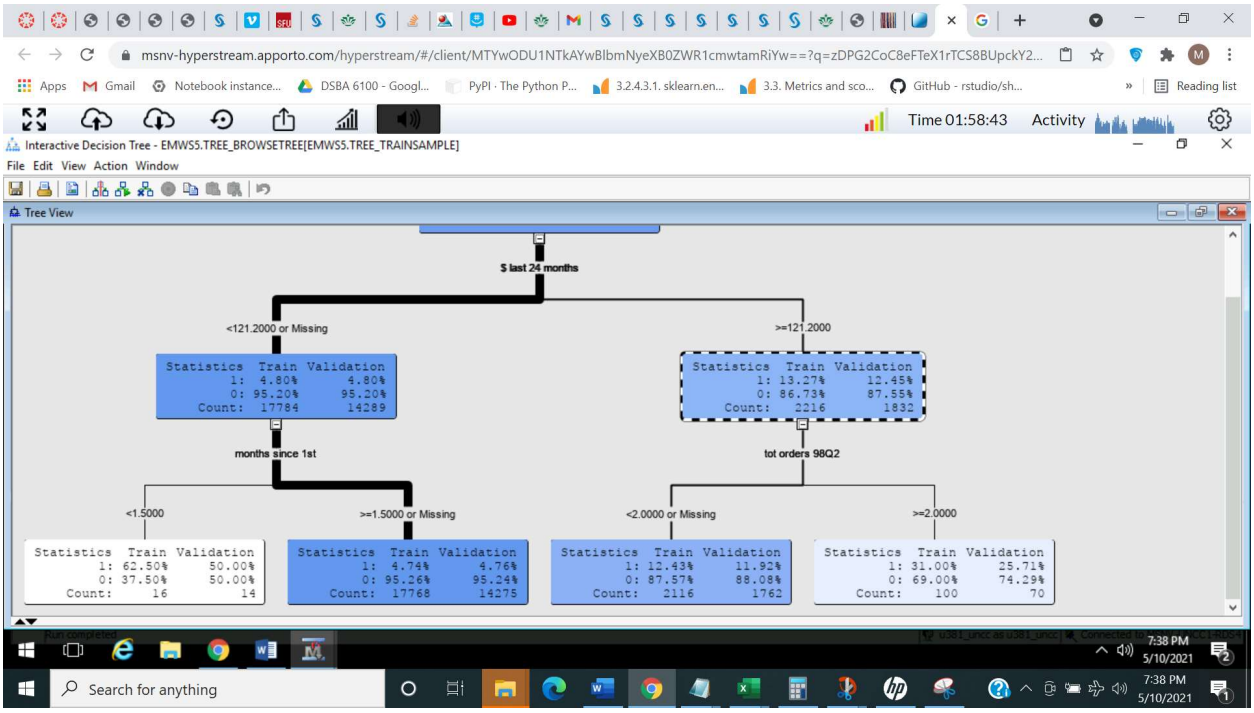

Interactive Decision Tree - EMWS5.TREE_BROWSETREE[EMWS
File Edit View Action Window
Tree View
Split Count | Node ID in Path
0 | 1 (Selected)
DOLL24 - Interval Split Rule
Target Variable: RESPOND
Assign missing values to
A specific branch
1
A separate missing values branch
All branches
Branches
Branch | Split Point
1 | < | 121.2000
2 | >= | 121.2000
New split point:
Add Branch
Remove Branch
OK
Cancel
Apply
Reset

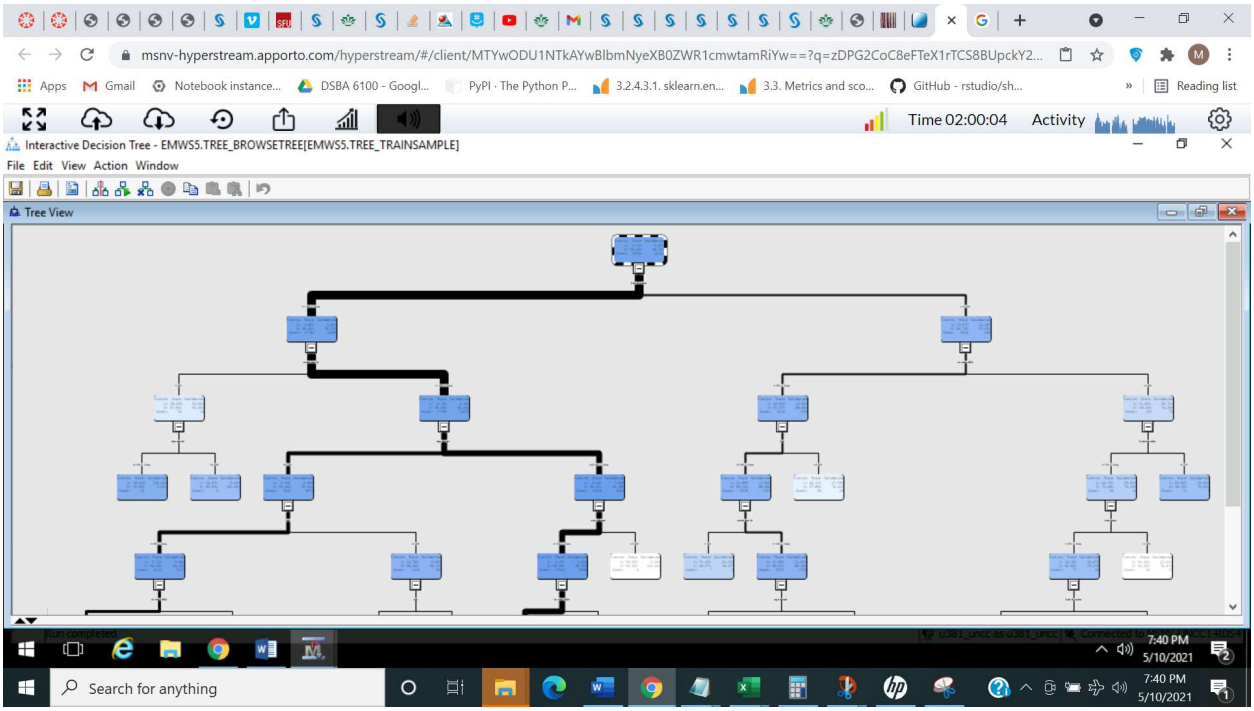

Interactive Decision Tree - EMWS5.TREE_BROWSETREE[EMWS5.TREE_TRAINSAMPLE]
File Edit View Action Window
Tree View
Statistics Train Validation
1: 5.73% 5.67%
0: 94.26% 94.33%
Count: 20000 16121
$ last 24 months
<121.2000 or Missing
>=121.2000
Statistics Train Validation
1: 4.80% 4.80%
0: 95.20% 95.20%
Count: 17784 14289
Statistics Train Validation
1: 13.27% 12.45%
0: 86.73% 87.55%
Count: 2216 1832
Split Count
Node ID in Path
Split Variable
Value
0
1 (Selected)
All Cases or Missing
Statistics
Train
Validation
Count
20000
16121
Prediction
0
0
% with target = 1
5.73%
5.67%
% with target = 0
94.26%
94.33%
% correctly predicted
94.26%
94.33%
Search for anything
7:30 PM
5/10/2021

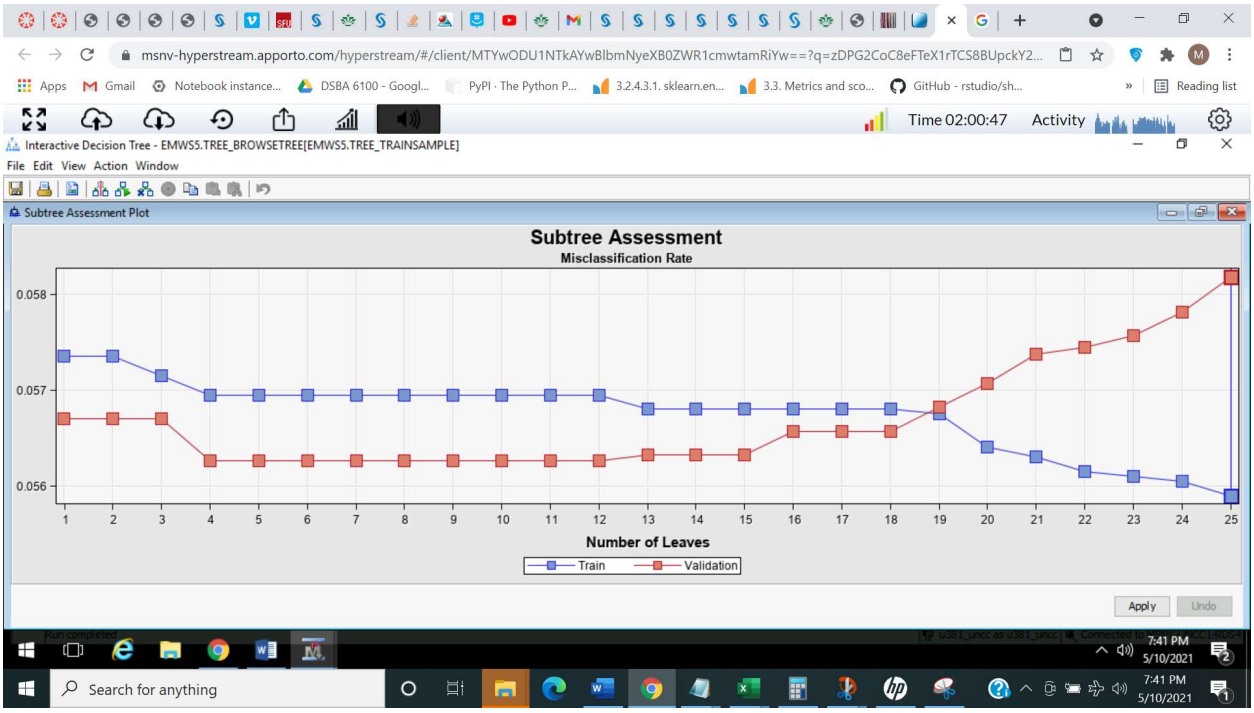

Interactive Decision Tree - EMWS5.TREE_BROWSETREE[EMWS5.TREE_TRAINSAMPLE]
File Edit View Action Window
Tree View
Statistics Train Validation
1: 4.80% 4.80%
0: 95.20% 95.20%
Count: 17784 14289
Statistics Train Validation
1: 13.27% 12.45%
0: 86.73% 87.55%
Count: 2216 1832
months since 1st
<1.5000
>=1.5000 or Missing
Statistics Train Validation
1: 62.50% 50.00%
0: 37.50% 50.00%
Count: 16 14
Statistics Train Validation
1: 4.74% 4.76%
0: 95.26% 95.24%
Count: 17768 14275

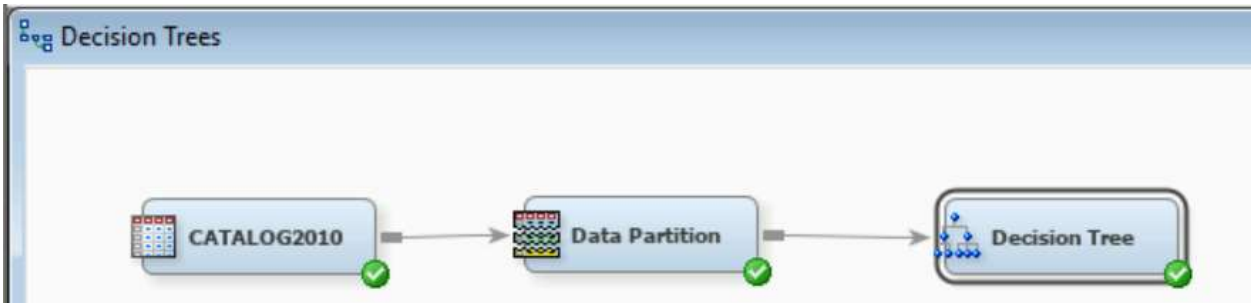

ile Edit View Action Window
Tree View
Statistics Train Validation
1: 4.80% 4.80%
0: 95.20% 95.20%
Count: 17784 14289
months since 1st
<1.5000
>=1.5000 or Missing
Statistics Train Validation
1: 62.50% 50.00%
0: 37.50% 50.00%
Count: 16 14
Statistics Train Validation
1: 4.74% 4.76%
0: 95.26% 95.24%
Count: 17768 14275
Statistics Train Validation
1: 13.27% 12.45%
0: 86.73% 87.55%
Count: 2216 1832
tot orders 98Q2
<2.5000 or Missing
>=2.5000
Statistics Train Validation
1: 12.87% 12.23%
0: 87.13% 87.77%
Count: 2199 1824
Statistics Train Validation
1: 64.71% 62.50%
0: 35.29% 37.50%
Count: 17 8

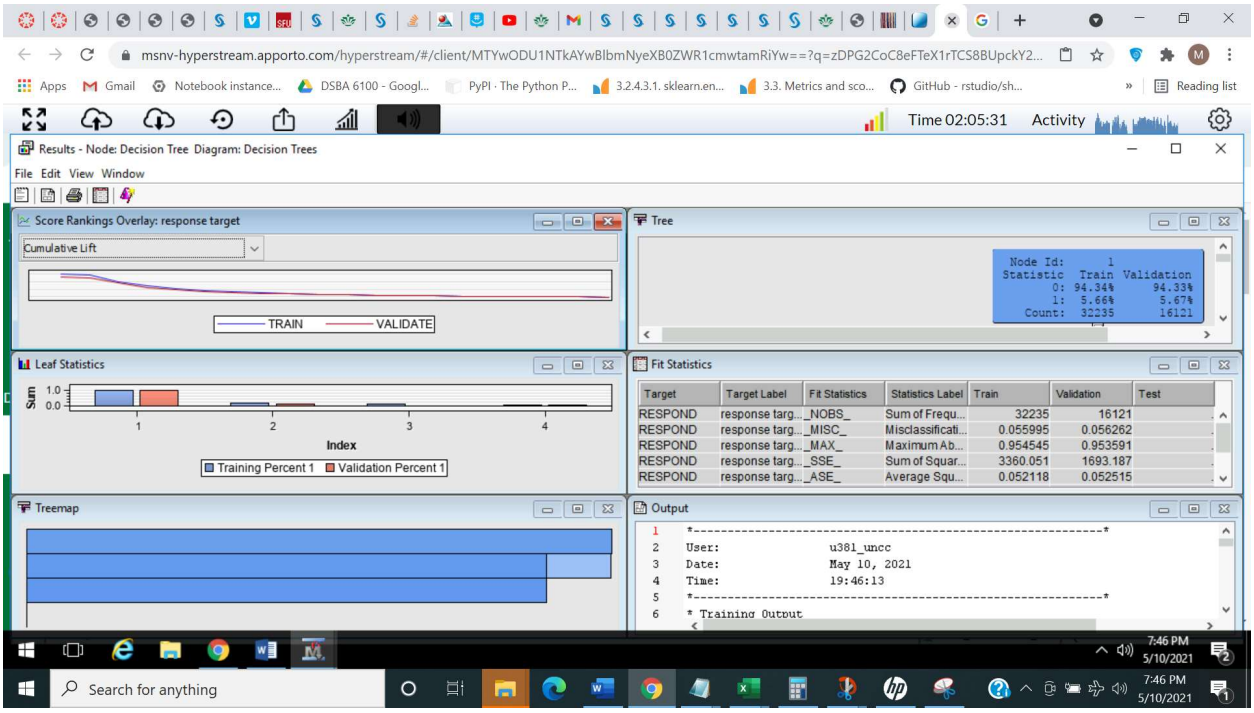

Interactive Decision Tree - EMWS5.TREE_BROWSETREE[EMWS
File Edit View Action Window
Tree View
Statistics Train Vali
1: 4.80%
0: 95.20%
Count: 17784
months since 1st
<1.5000
Statistics Train Validation
1: 62.50% 50.00%
0: 37.50% 50.00%
Count: 16 14
Split Count
Node ID in Path
0
1
1
54 (Selected)
TOTORDQ22 - Interval Split Rule
Target Variable: RESPOND
Assign missing values to
A specific branch
1
A separate missing values branch
All branches
Branches
Branch
Split Point
1
<
2.0000
2
>=
2.0000
New split point:
Add Branch
Remove Branch
OK
Cancel
Apply
Reset

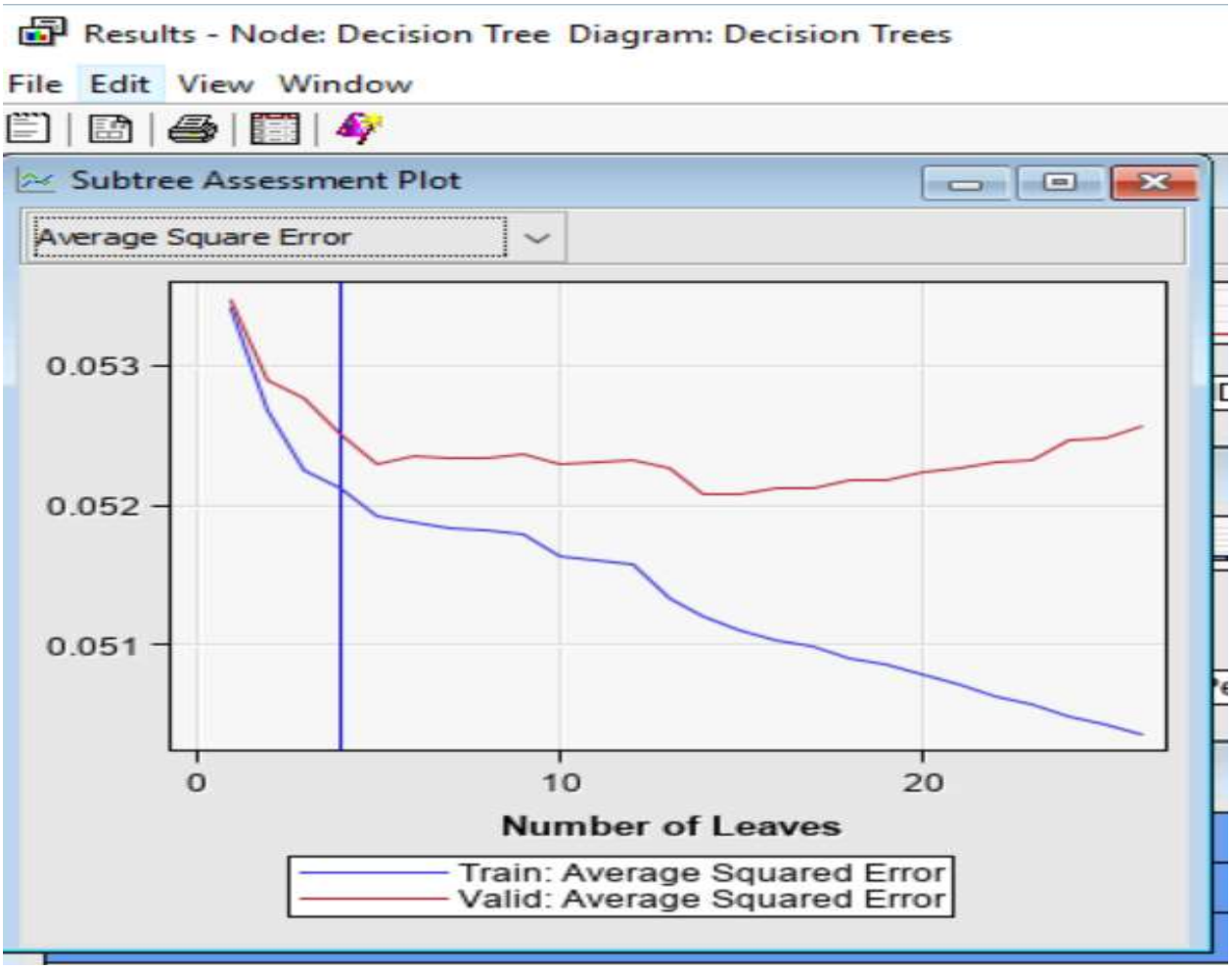

msnv-hyperstream.apporto.com/hyperstream/#/client/MTYwODU1NTkAYwBlbmNyeXB0ZWR1cmwtamRiYw==?q=zDPG2CoC8eFTeX1rTCS8BUpckY2...
Time 01:58:43
Activity
Interactive Decision Tree - EMWS5.TREE_BROWSETREE[EMWS5.TREE_TRAINSAMPLE]
File Edit View Action Window
Tree View
$ last 24 months
<121.2000 or Missing
>=121.2000
Statistics Train Validation
1: 4.80% 4.80%
0: 95.20% 95.20%
Count: 17784 14289
Statistics Train Validation
1: 13.27% 12.45%
0: 86.73% 87.55%
Count: 2216 1832
months since 1st
tot orders 98Q2
<1.5000
>=1.5000 or Missing
<2.0000 or Missing
>=2.0000
Statistics Train Validation
1: 62.50% 50.00%
0: 37.50% 50.00%
Count: 16 14
Statistics Train Validation
1: 4.74% 4.76%
0: 95.26% 95.24%
Count: 17768 14275
Statistics Train Validation
1: 12.43% 11.92%
0: 87.57% 88.08%
Count: 2116 1762
Statistics Train Validation
1: 31.00% 25.71%
0: 69.00% 74.29%
Count: 100 70
Search for anything
7:38 PM
5/10/2021

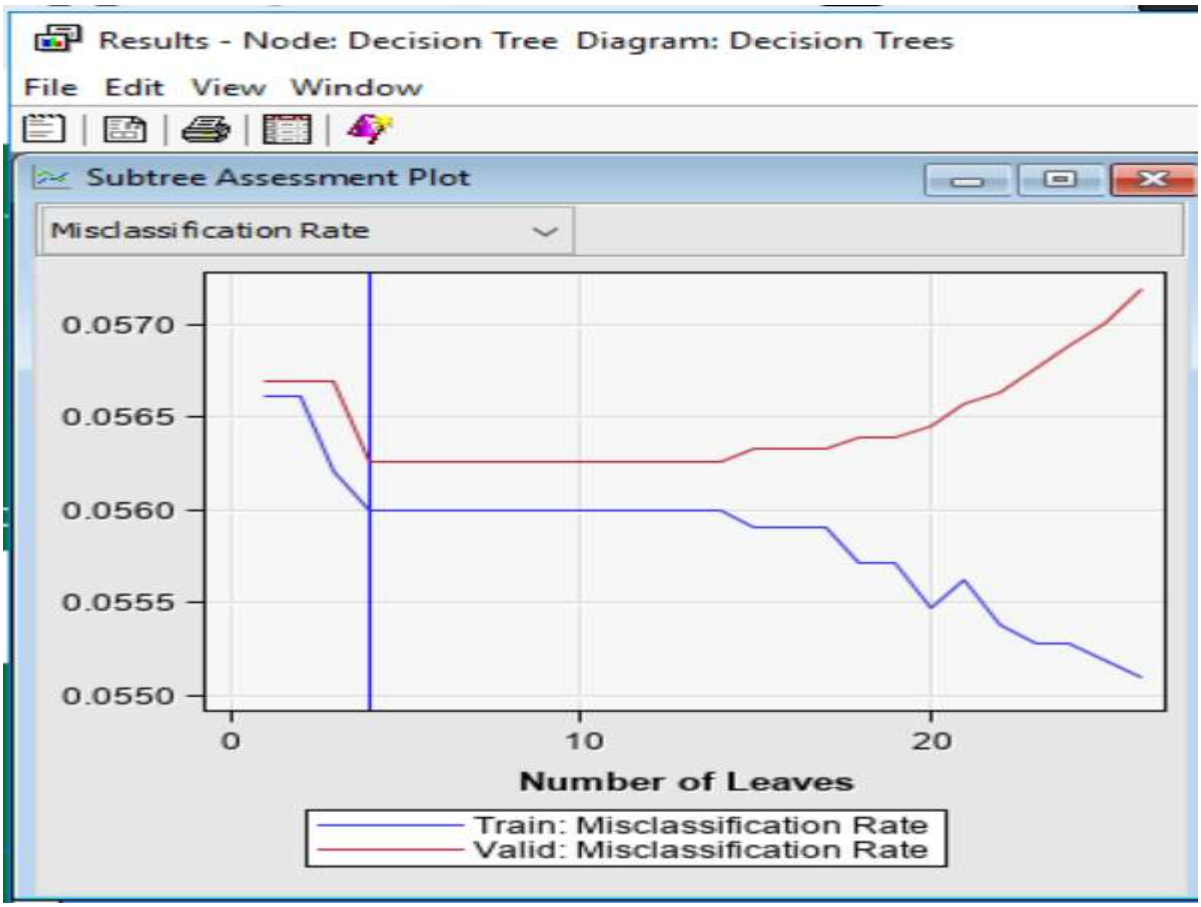

msnv-hyperstream.apporto.com/hyperstream/#/client/MTYwODU1NTkAYwBlbmNyeXB0ZWR1cmwtamRiYw==?q=zDPG2CoC8eFTeX1rTCS8BUpckY2...
Time 02:00:04
Activity
Interactive Decision Tree - EMWS5.TREE_BROWSETREE[EMWS5.TREE_TRAINSAMPLE]
File Edit View Action Window
Tree View
Search for anything
7:40 PM
5/10/2021

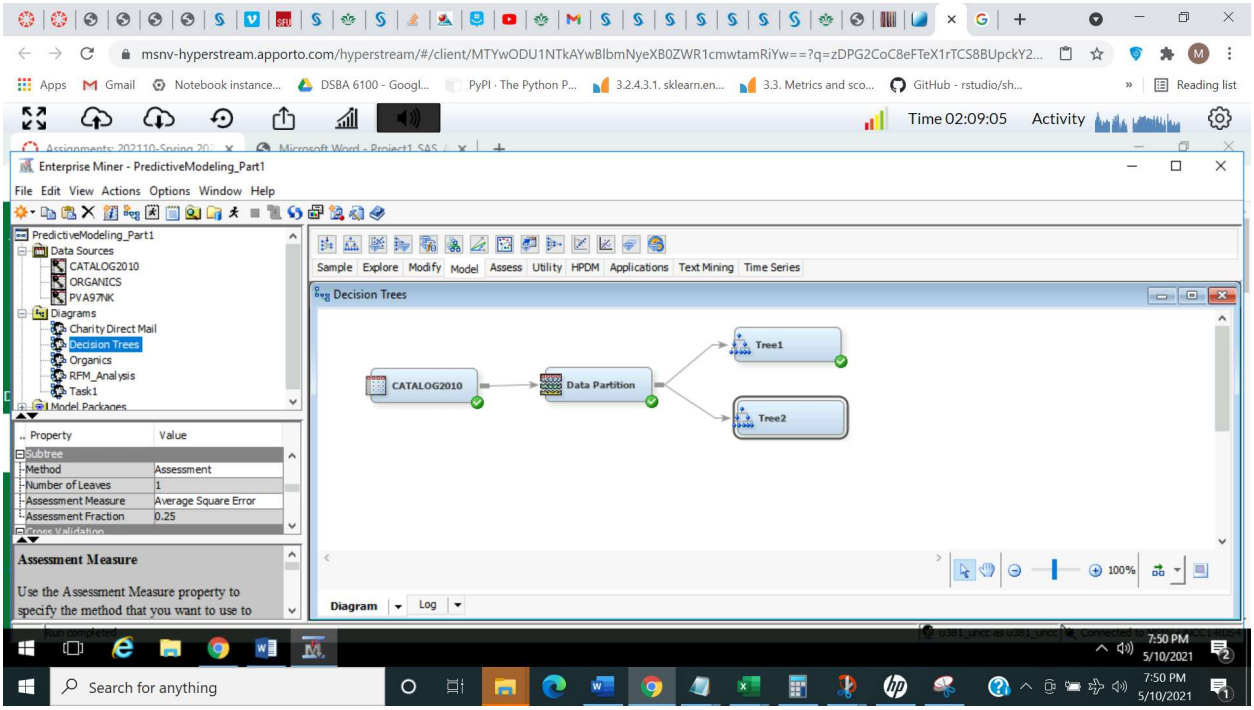

msnv-hyperstream.apporto.com/hyperstream/#/client/MTYwODU1NTkAYwBlbmNyeXB0ZWR1cmwtamRiYw==?q=zDPG2CoC8eFTeX1rTCS8BUpckY2...
Time 02:00:47
Activity
Interactive Decision Tree - EMWS5.TREE_BROWSETREE[EMWS5.TREE_TRAINSAMPLE]
File Edit View Action Window
Subtree Assessment Plot
Subtree Assessment
Misclassification Rate
0.058
0.057
0.056
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
Number of Leaves
Train
Validation
Apply
Undo
Search for anything
7:41 PM
5/10/2021

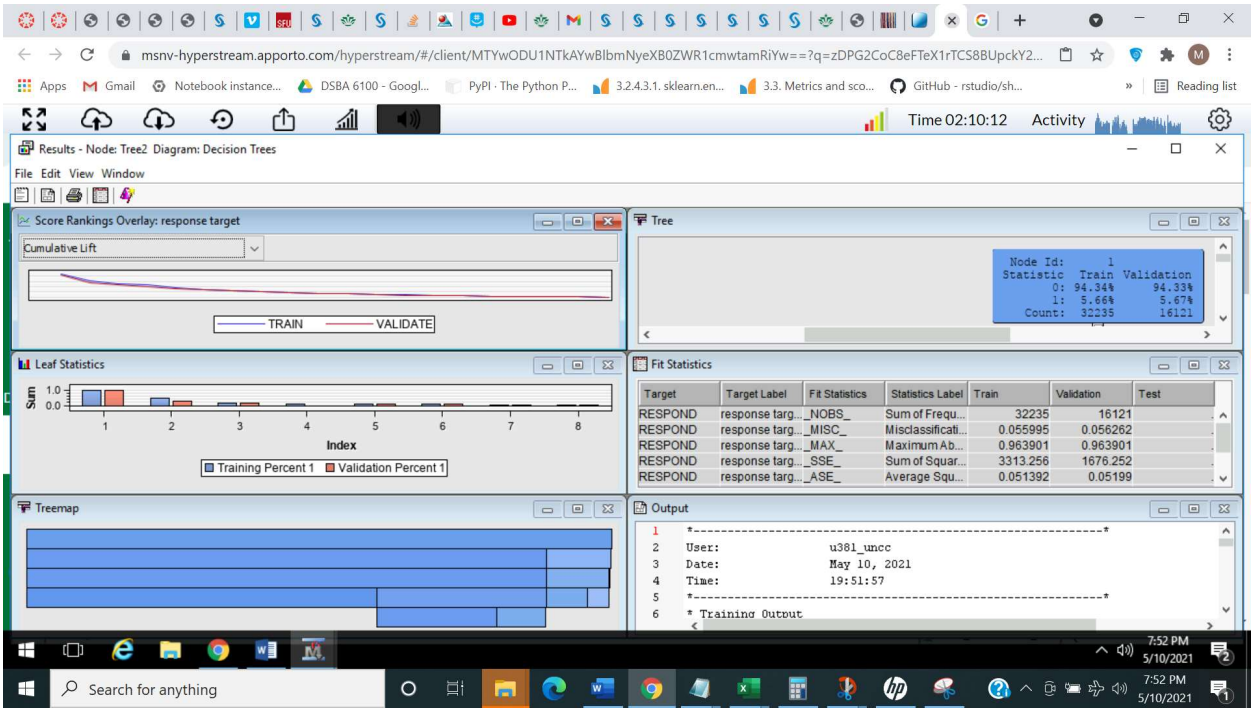

Decision Trees
CATALOG2010
Data Partition
Decision Tree

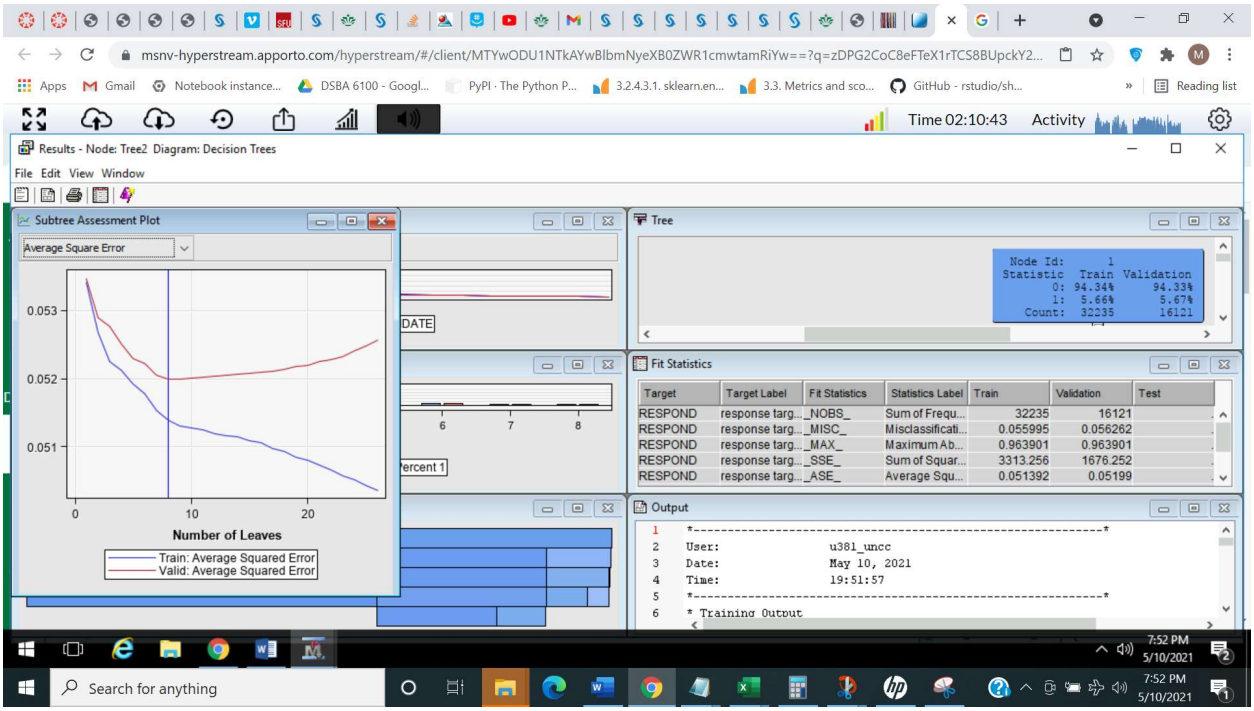

Results - Node: Decision Tree Diagram: Decision Trees
File Edit View Window
Score Rankings Overlay: response target
Cumulative Lift
TRAIN
VALIDATE
Tree
Node Id: 1
Statistic Train Validation
0: 94.34% 94.33%
1: 5.66% 5.67%
Count: 32235 16121
Leaf Statistics
Sum
Index
Training Percent 1
Validation Percent 1
Fit Statistics
Target Target Label Fit Statistics Statistics Label Train Validation Test
RESPOND response targ... _NOBS_ Sum of Frequ... 32235 16121
RESPOND response targ... _MISC_ Misclassificati... 0.055995 0.056262
RESPOND response targ... _MAX_ Maximum Ab... 0.954545 0.953591
RESPOND response targ... _SSE_ Sum of Squar... 3360.051 1693.187
RESPOND response targ... _ASE_ Average Squ... 0.052118 0.052515
Treemap
Output
User: u381_uncc
Date: May 10, 2021
Time: 19:46:13
* Training Output
Time 02:05:31
Activity
Search for anything
7:46 PM
5/10/2021

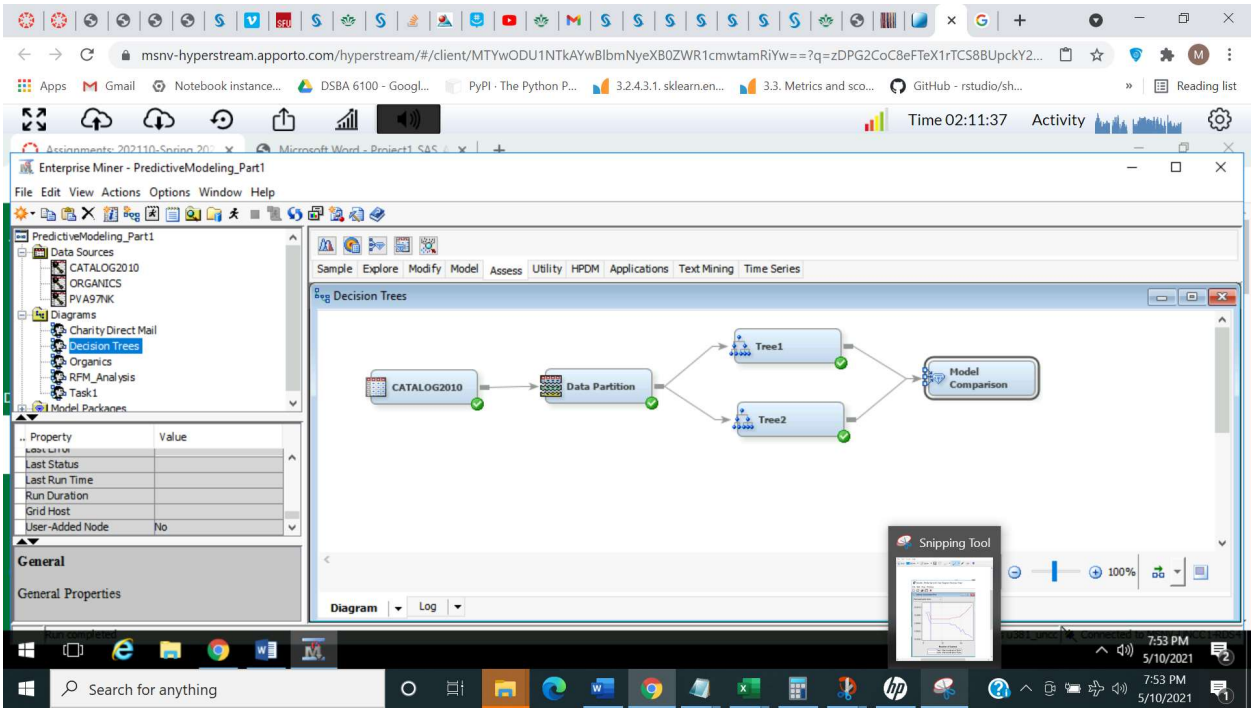

Results - Node: Decision Tree Diagram: Decision Trees
File Edit View Window
Subtree Assessment Plot
Average Square Error
0.053
0.052
0.051
0
10
20
Number of Leaves
Train: Average Squared Error
Valid: Average Squared Error

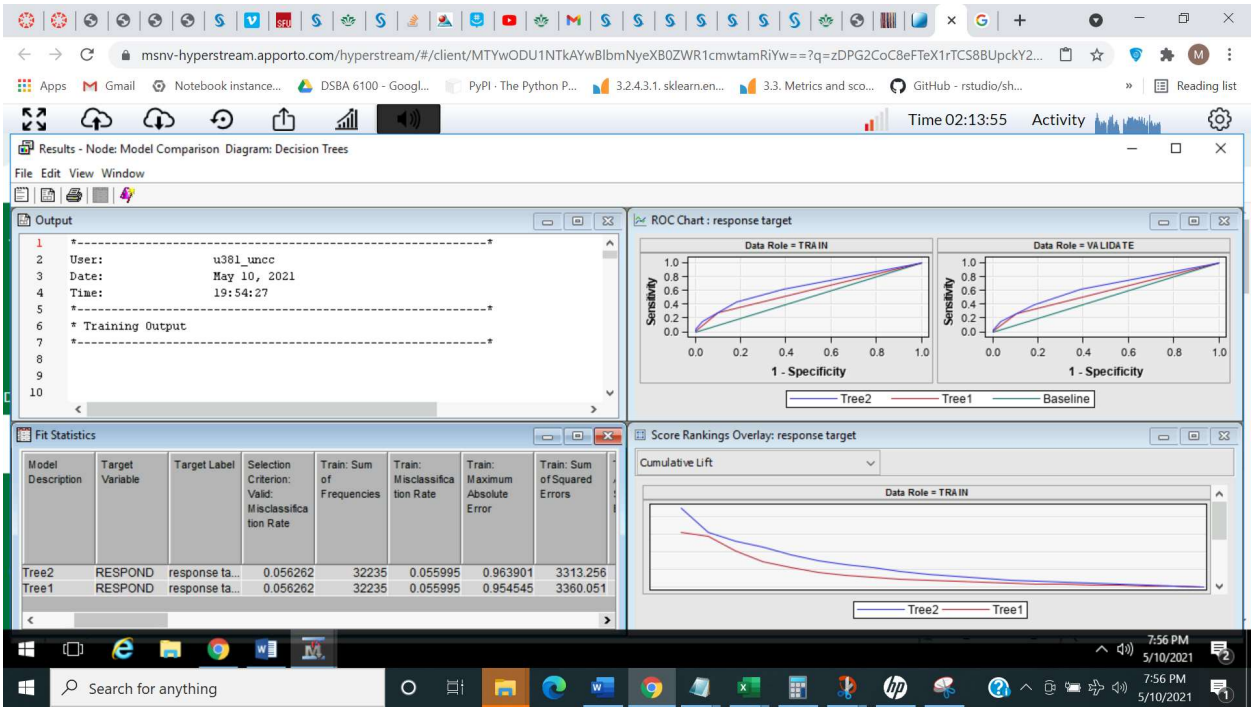

Results - Node: Decision Tree Diagram: Decision Trees
File Edit View Window
Subtree Assessment Plot
Misclassification Rate
0.0570
0.0565
0.0560
0.0555
0.0550
0
10
20
Number of Leaves
Train: Misclassification Rate
Valid: Misclassification Rate


msnv-hyperstream.apporto.com/hyperstream/#/client/MTYwODU1NTkAYwBlbmNyeXB0ZWR1cmwtamRiYw==?q=zDPG2CoC8eFTeX1rTCS8BUpckY2...
Time 02:09:05
Activity
Enterprise Miner - PredictiveModeling_Part1
File Edit View Actions Options Window Help
PredictiveModeling_Part1
Data Sources
CATALOG2010
ORGANICS
PVA97NK
Diagrams
Charity Direct Mail
Decision Trees
Organics
RFM_Analysis
Task1
Model Packages
Sample Explore Modify Model Assess Utility HPDM Applications Text Mining Time Series
Decision Trees
CATALOG2010
Data Partition
Tree1
Tree2
Property
Value
Subtree
Method
Assessment
Number of Leaves
1
Assessment Measure
Average Square Error
Assessment Fraction
0.25
Assessment Measure
Use the Assessment Measure property to specify the method that you want to use to
Diagram
Log
100%
7:50 PM
5/10/2021
Search for anything


msnv-hyperstream.apporto.com/hyperstream/#/client/MTYwODU1NTkAYwBlbmNyeXB0ZWR1cmwtamRiYw==?q=zDPG2CoC8eFTeX1rTCS8BUpckY2...
Time 02:10:12
Activity
Results - Node: Tree2 Diagram: Decision Trees
File Edit View Window
Score Rankings Overlay: response target
Cumulative Lift
TRAIN
VALIDATE
Tree
Node Id: 1
Statistic Train Validation
0: 94.34% 94.33%
1: 5.66% 5.67%
Count: 32235 16121
Leaf Statistics
Sum
Index
Training Percent 1
Validation Percent 1
Fit Statistics
Target Target Label Fit Statistics Statistics Label Train Validation Test
RESPOND response targ... _NOBS_ Sum of Frequ... 32235 16121
RESPOND response targ... _MISC_ Misclassificati... 0.055995 0.056262
RESPOND response targ... _MAX_ Maximum Ab... 0.963901 0.963901
RESPOND response targ... _SSE_ Sum of Squar... 3313.256 1676.252
RESPOND response targ... _ASE_ Average Squ... 0.051392 0.05199
Treemap
Output
1 *--*
2 User: u381_uncc
3 Date: May 10, 2021
4 Time: 19:51:57
5 *--*
6 * Training Output
7:52 PM
5/10/2021
Search for anything

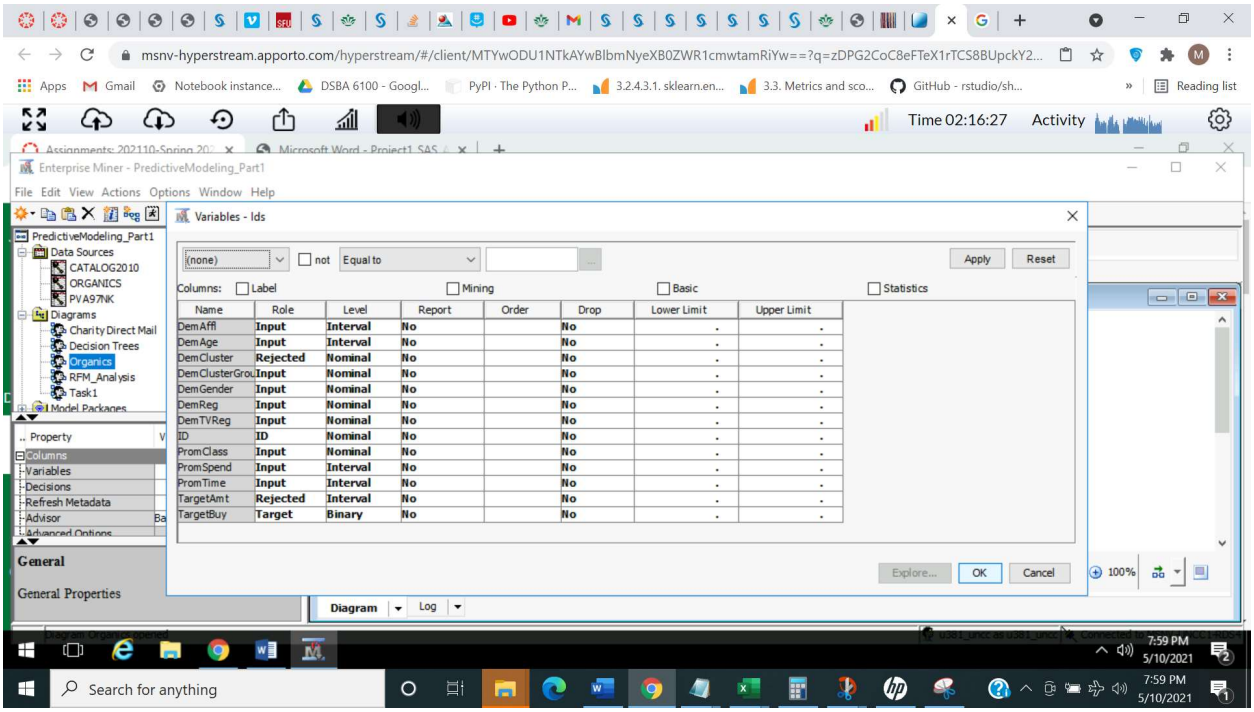

msnv-hyperstream.apporto.com/hyperstream/#/client/MTYwODU1NTkAYwBlbmNyeXB0ZWR1cmwtamRiYw==?q=zDPG2CoC8eFTeX1rTCS8BUpckY2...
Time 02:10:43
Activity
Results - Node: Tree2 Diagram: Decision Trees
File Edit View Window
Subtree Assessment Plot
Average Square Error
0.053
0.052
0.051
0
10
20
Number of Leaves
Train: Average Squared Error
Valid: Average Squared Error
Tree
Node Id: 1
Statistic Train Validation
0: 94.34% 94.33%
1: 5.66% 5.67%
Count: 32235 16121
Fit Statistics
Target Target Label Fit Statistics Statistics Label Train Validation Test
RESPOND response targ... _NOBS_ Sum of Frequ... 32235 16121
RESPOND response targ... _MISC_ Misclassificati... 0.055995 0.056262
RESPOND response targ... _MAX_ Maximum Ab... 0.963901 0.963901
RESPOND response targ... _SSE_ Sum of Squar... 3313.256 1676.252
RESPOND response targ... _ASE_ Average Squ... 0.051392 0.05199
Output
1 *--*
2 User: u381_uncc
3 Date: May 10, 2021
4 Time: 19:51:57
5 *--*
6 * Training Output
7:52 PM
5/10/2021
Search for anything


Enterprise Miner - PredictiveModeling_Part1
File Edit View Actions Options Window Help
PredictiveModeling_Part1
Data Sources
CATALOG2010
ORGANICS
PVA97NK
Diagrams
Charity Direct Mail
Decision Trees
Organics
RFM_Analysis
Task1
Model Packages
Sample Explore Modify Model Assess Utility HPDM Applications Text Mining Time Series
Decision Trees
CATALOG2010
Data Partition
Tree1
Tree2
Model Comparison
Property Value
Last Status
Last Run Time
Run Duration
Grid Host
User-Added Node No
General
General Properties
Snipping Tool
Diagram Log
Time 02:11:37
Activity
Search for anything
7:53 PM
5/10/2021

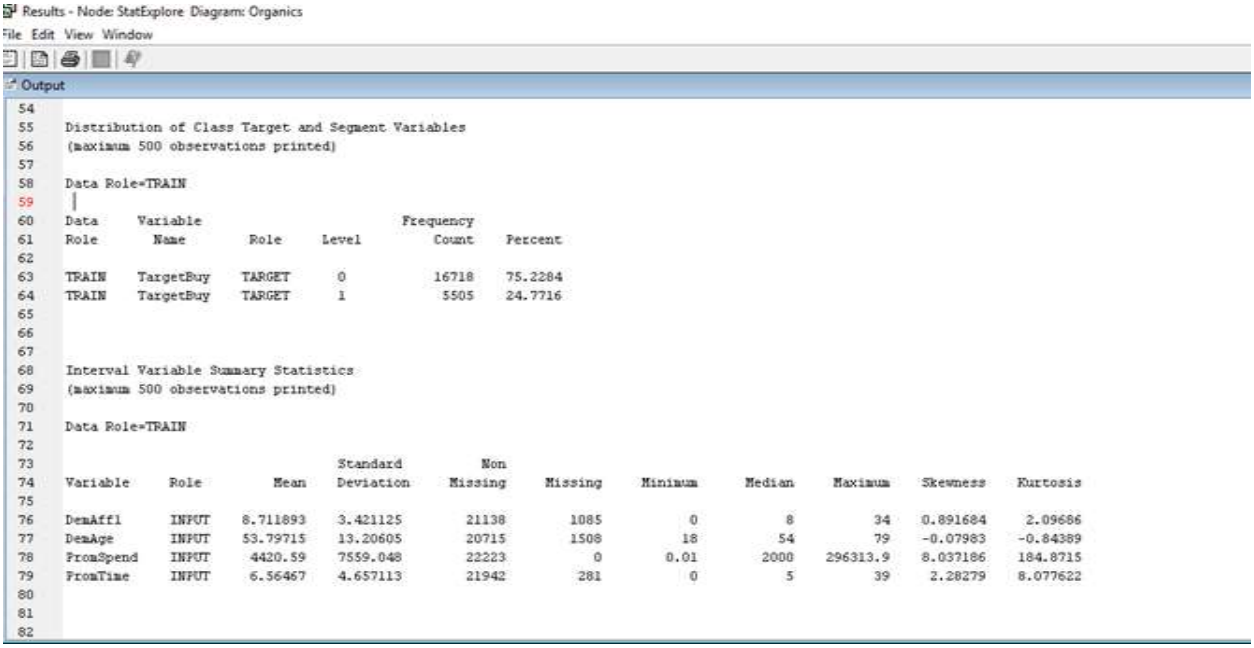

Results - Node: Model Comparison Diagram: Decision Trees
File Edit View Window
Output
User: u381_uncc
Date: May 10, 2021
Time: 19:54:27
* Training Output
ROC Chart : response target
Data Role = TRAIN
Data Role = VALIDATE
Sensitivity
1 - Specificity
Tree2 Tree1 Baseline
Fit Statistics
Model Description Target Variable Target Label Selection Criterion: Valid: Misclassification Rate Train: Sum of Frequencies Train: Misclassification Rate Train: Maximum Absolute Error Train: Sum of Squared Errors
Tree2 RESPOND response ta... 0.056262 32235 0.055995 0.963901 3313.256
Tree1 RESPOND response ta... 0.056262 32235 0.055995 0.954545 3360.051
Score Rankings Overlay: response target
Cumulative Lift
Data Role = TRAIN
Tree2 Tree1
Time 02:13:55
Activity
Search for anything
7:56 PM
5/10/2021

Task 3 Organic Dataset Decision Tree

1. Set the roles for the analysis variables as shown above.

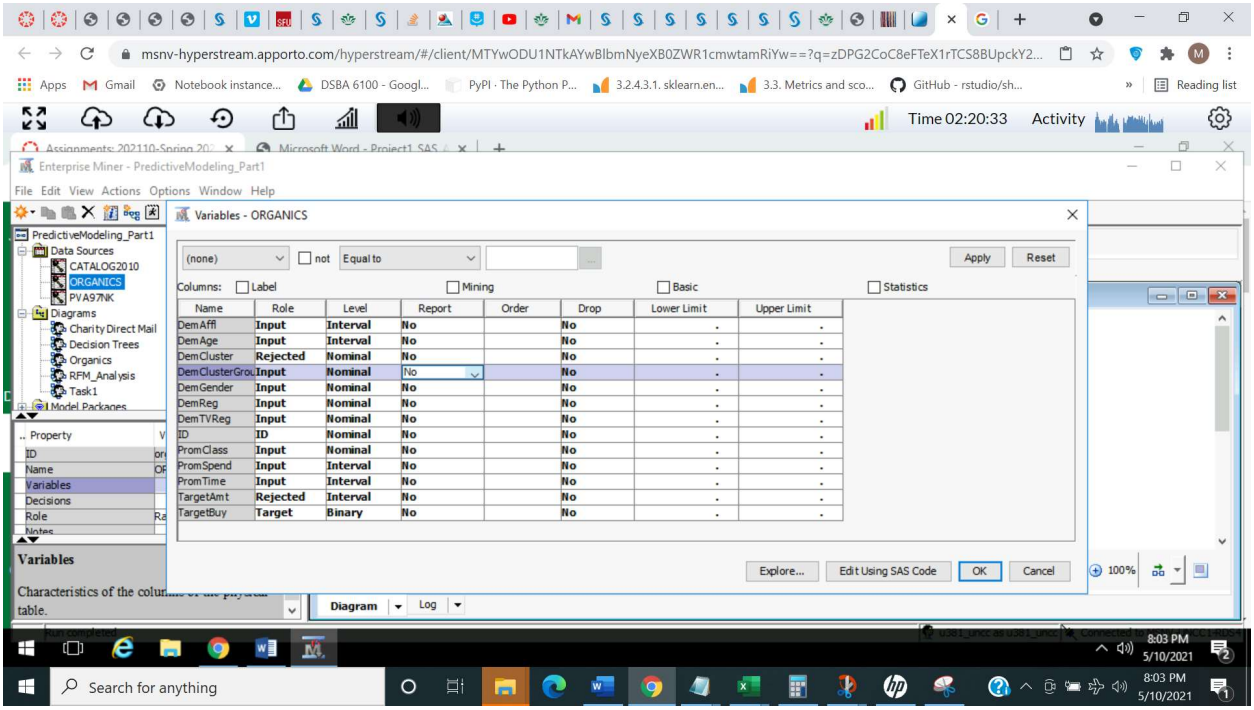


2. Examine the distribution of the target variable. What is the proportion of individuals who purchased organic products?



The proportion of individuals who purchased organics products is 24.7716 %.

3. The variable **DemClusterGroup** contains collapsed levels of the variable **DemCluster**. Presume that, based on previous experience, you believe that **DemClusterGroup** is sufficient for this type of modeling effort. Set the model role for **DemCluster** to Rejected.



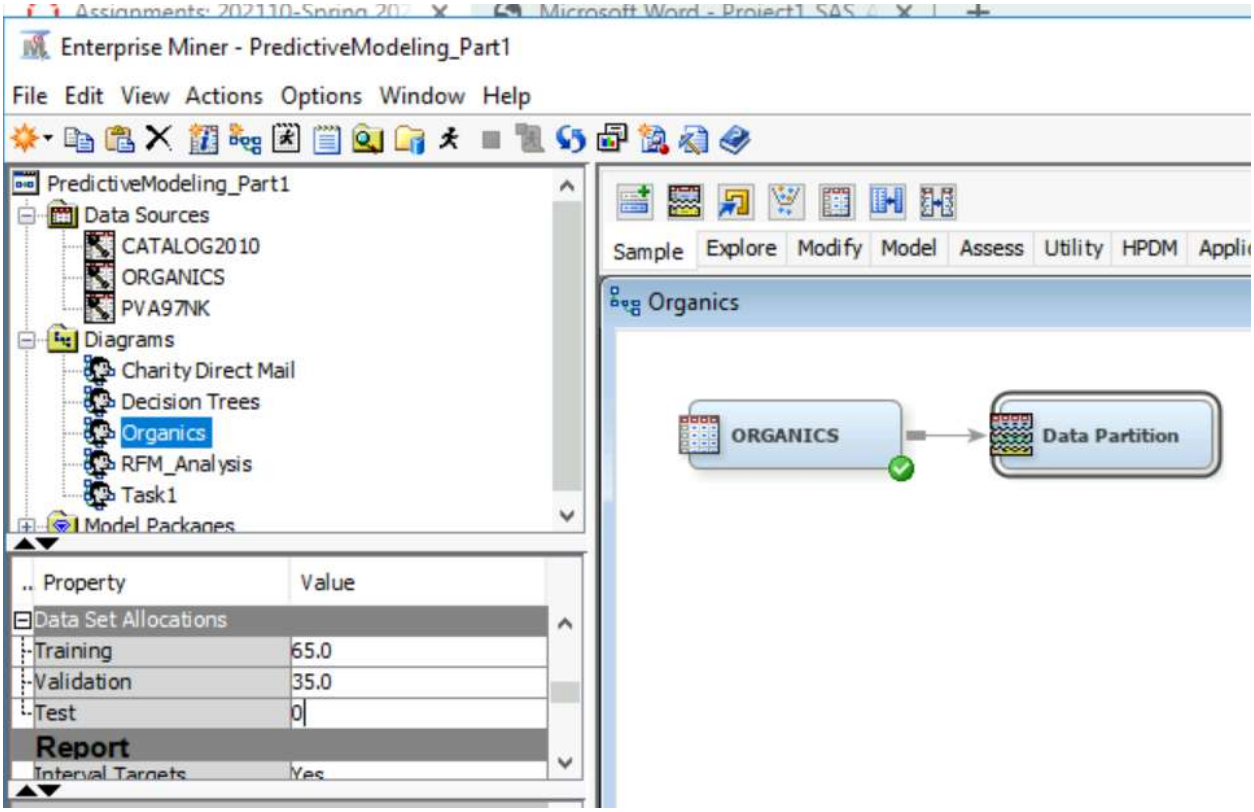
4. As noted above, only **TargetBuy** is used for this analysis, and it should have a role of **Target**. Can **TargetAmt** be used as an input for a model that is used to predict **TargetBuy**? Why or why not?

   No, **TargetAmt** can not be used to predict **TargetBuy** because any amount purchased indicates a buy so it is a redundant variable. Adding **TargetAmt** to our existing variables will skew data incorrectly because of its 100% correlation with the **TargetBuy Target** role variable.

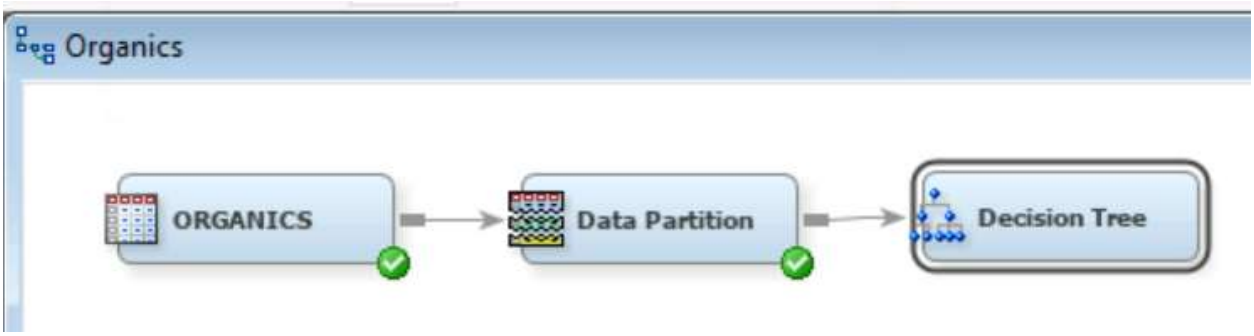
5. Finish the **ORGANICS** data source definition.

c) Add the **ORGANICS** data source to the Organics diagram workspace.

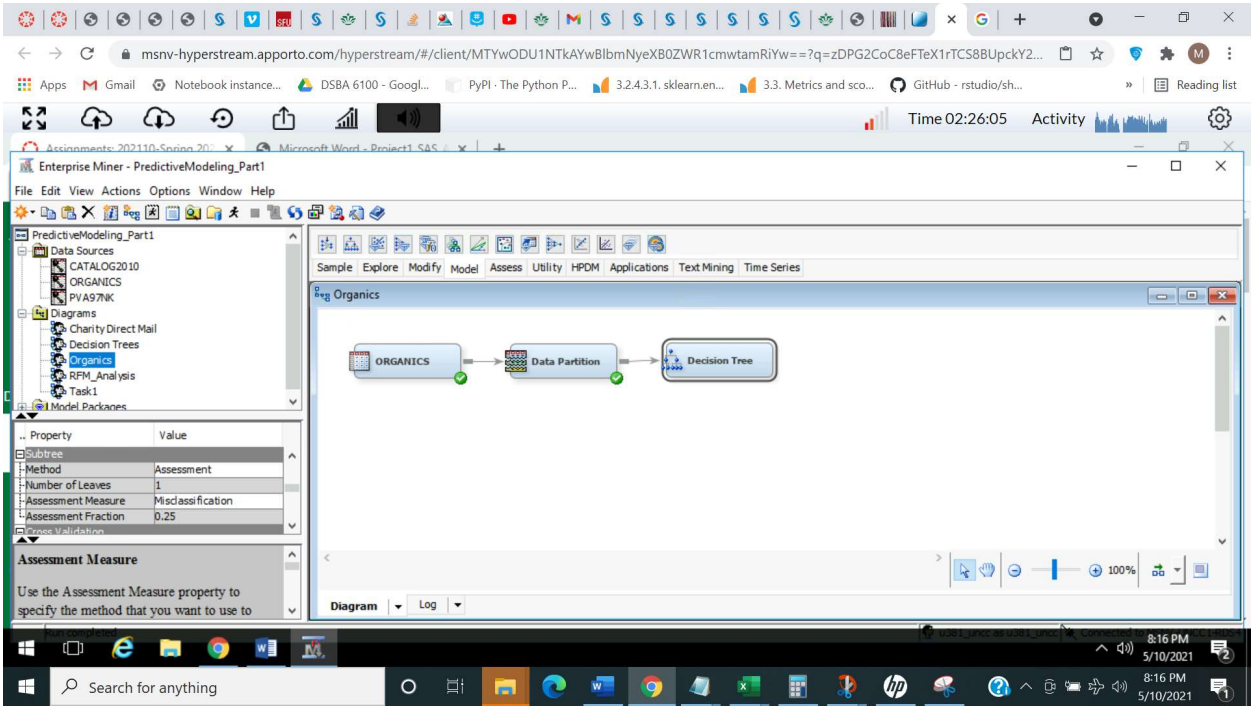
d) Add a **Data Partition** node to the diagram and connect it to the **Data Source** node. Assign 65% of the data for training and 35% for validation.



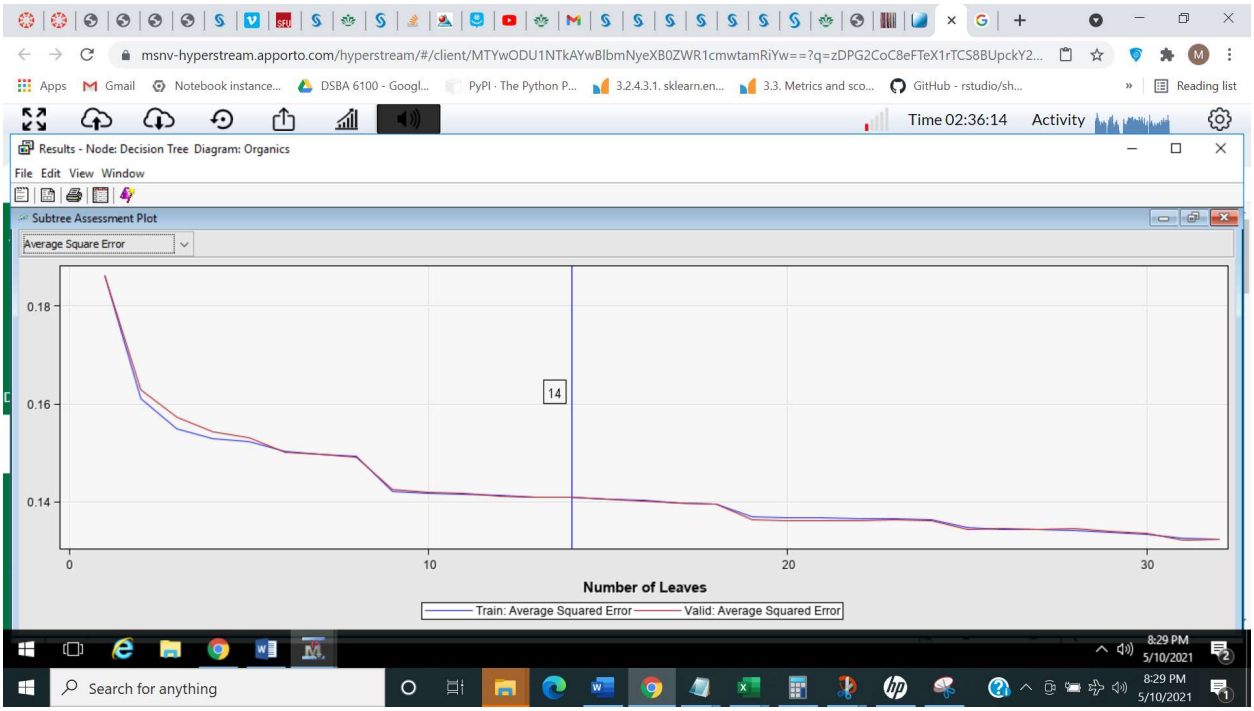
e) Add a **Decision Tree** node to the workspace and connect it to the **Data Partition** node.



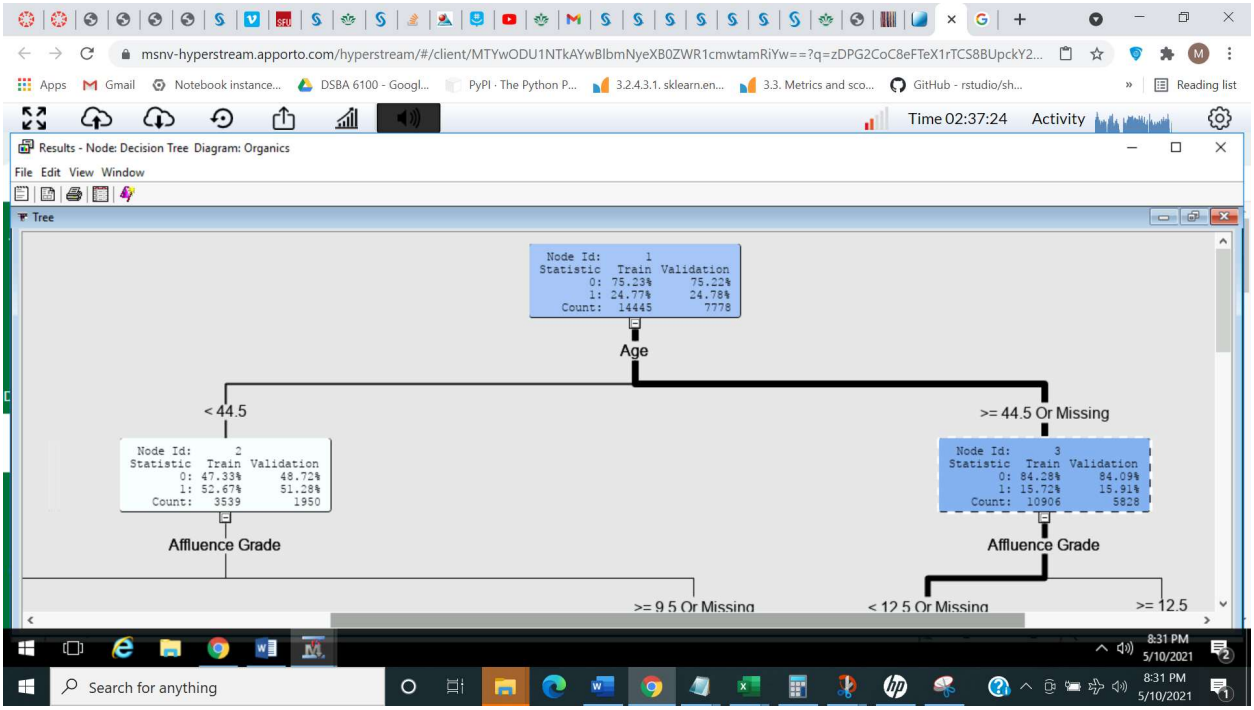
f) Create a **decision tree** model autonomously. Use **Misclassification** as the model assessment statistic.

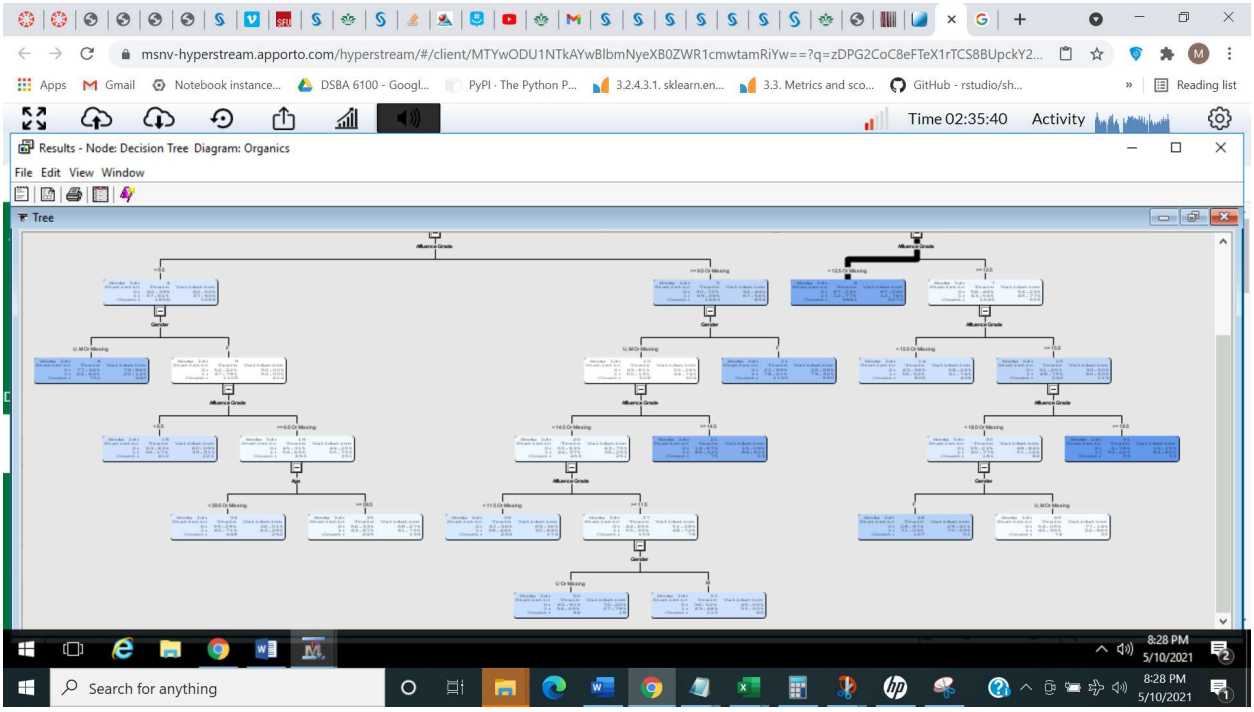


1) How many leaves are in the optimal tree? 14 leaves



2) Which variables were used for the first split? Age < 44.5 and Age >= 44.5 or Age is missing





3) What were the competing splits for this first split?

Affluence Grade (Importance=0.7806) and Gender (Importance=0.4090) were also competing for the first split.

Results - Node: Decision Tree Diagram: Organics

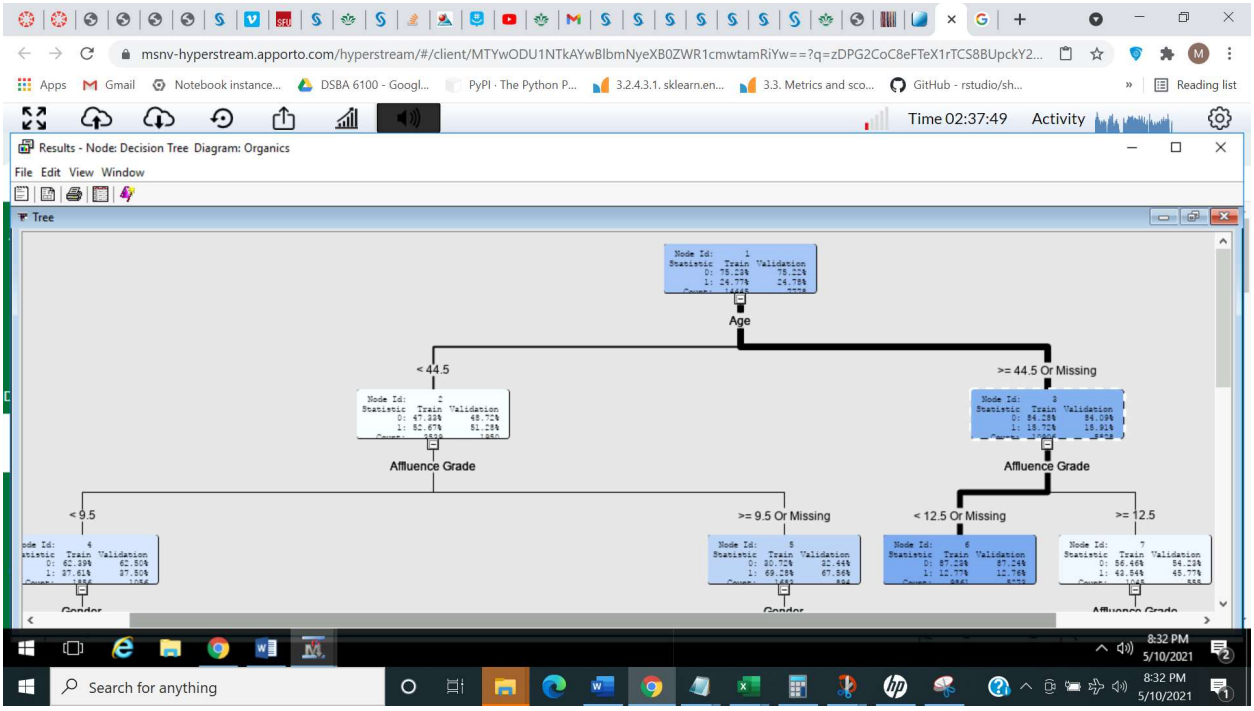
File Edit View Window

Output

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
DemAge	Age	2	1.0000	1.0000	1.0000
DemAffl	Affluence Grade	7	0.7806	0.7929	1.0158
DemGender	Gender	4	0.4090	0.5184	1.2674

4) Which variables were used for the second split for all branches from first split?

Affluence Grade < 9.5, Affluence Grade >=9.5 or Missing, Affluence Grade < 12.5 or Missing, Affluence Grade >= 12.5

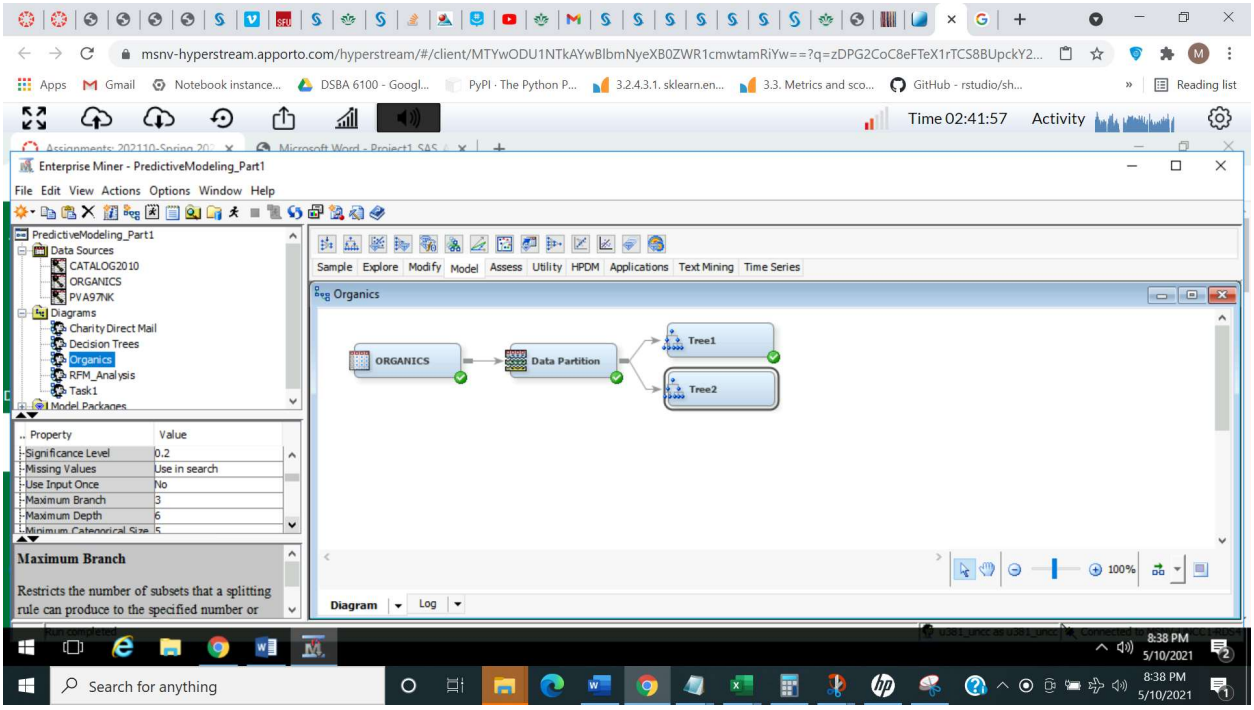


5) Discuss the results and provide your insights?

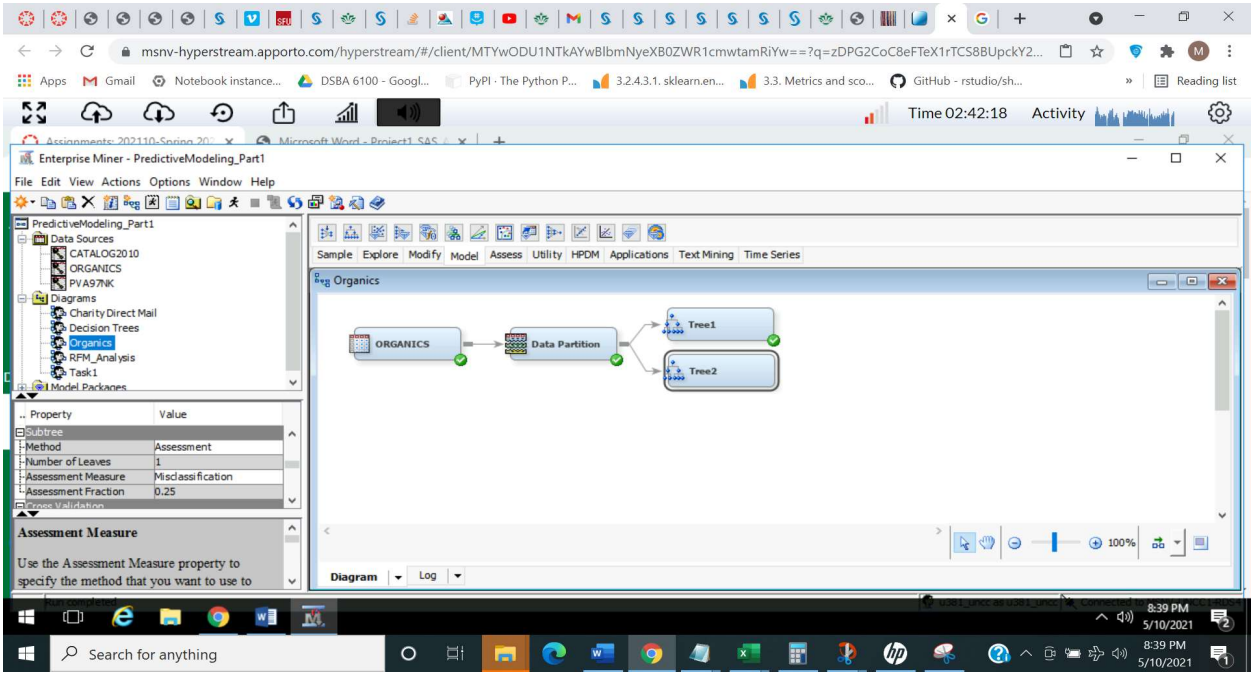
I see that Affluence Grade is split multiple times with different values for example in Node 7, Node 15 and Node 20. This is a complex model with 51 total nodes and nodes are reused to spilt as needed.

Add a second **Decision Tree** node to the diagram and connect it to the **Data Partition** node.

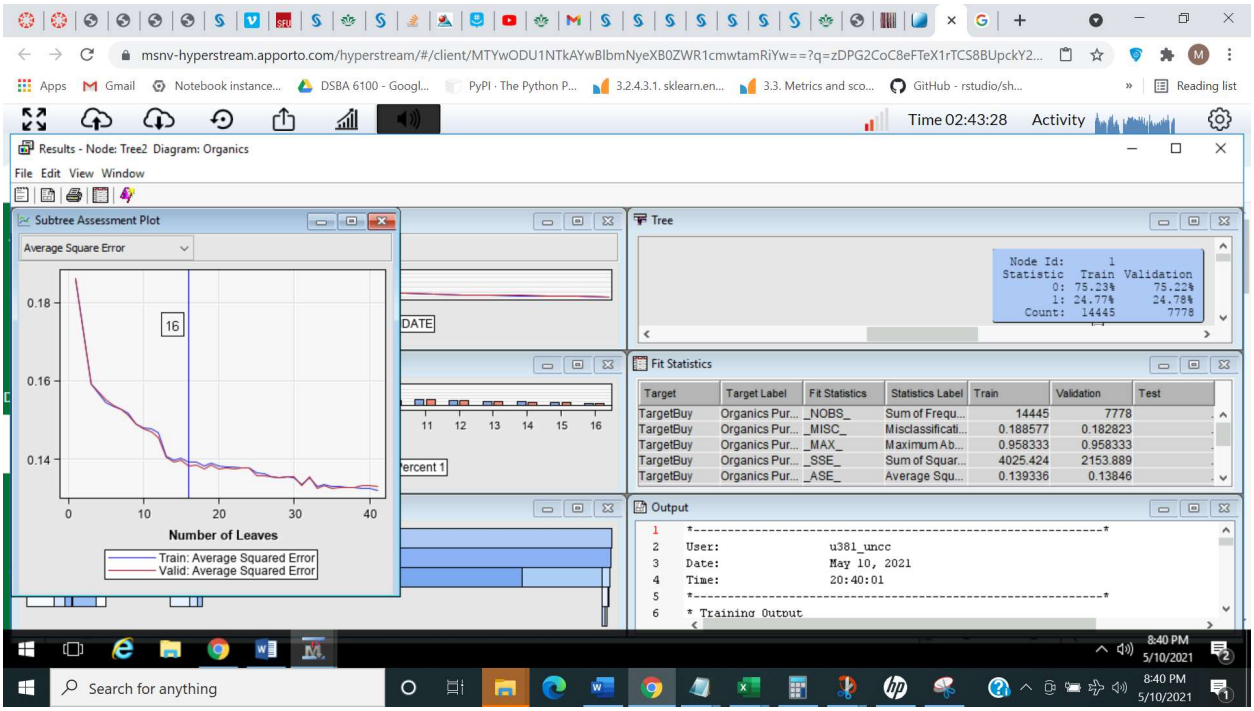
1) In the Properties panel of the new Decision Tree node, change the maximum number of branches from a node to 3 to allow for three-way splits.



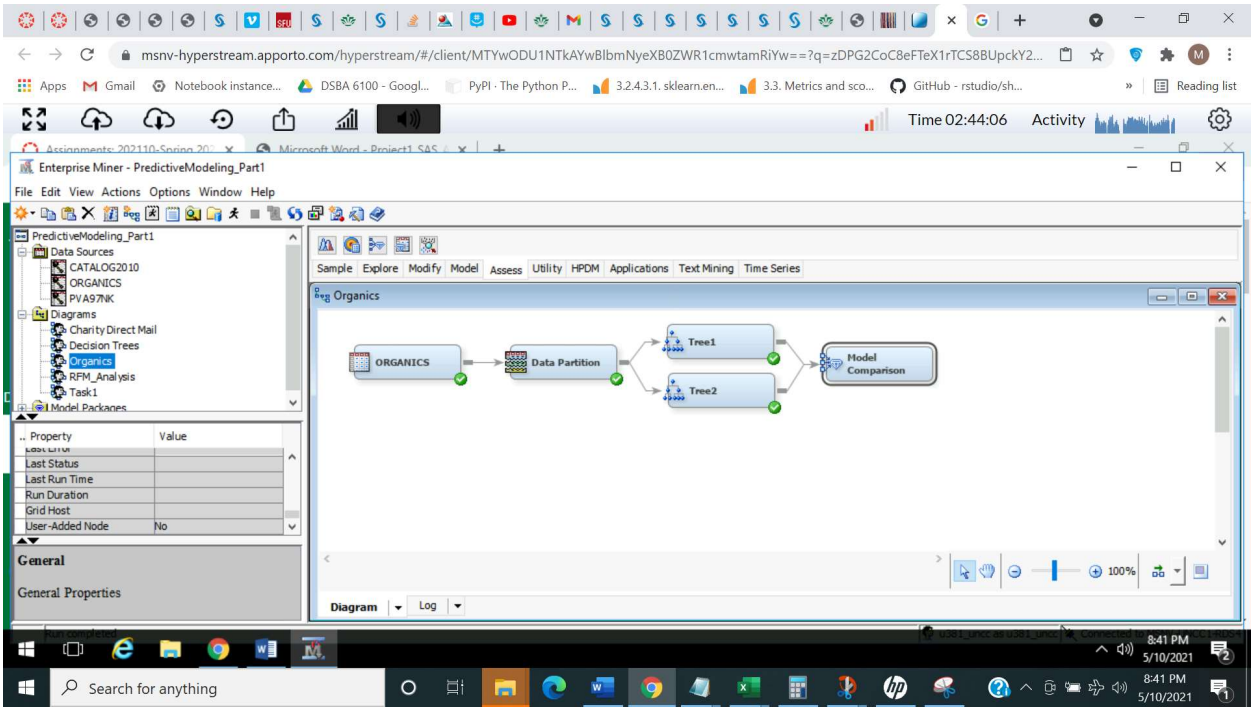
2) Create a decision tree model using **Misclassification** as the model assessment statistic.



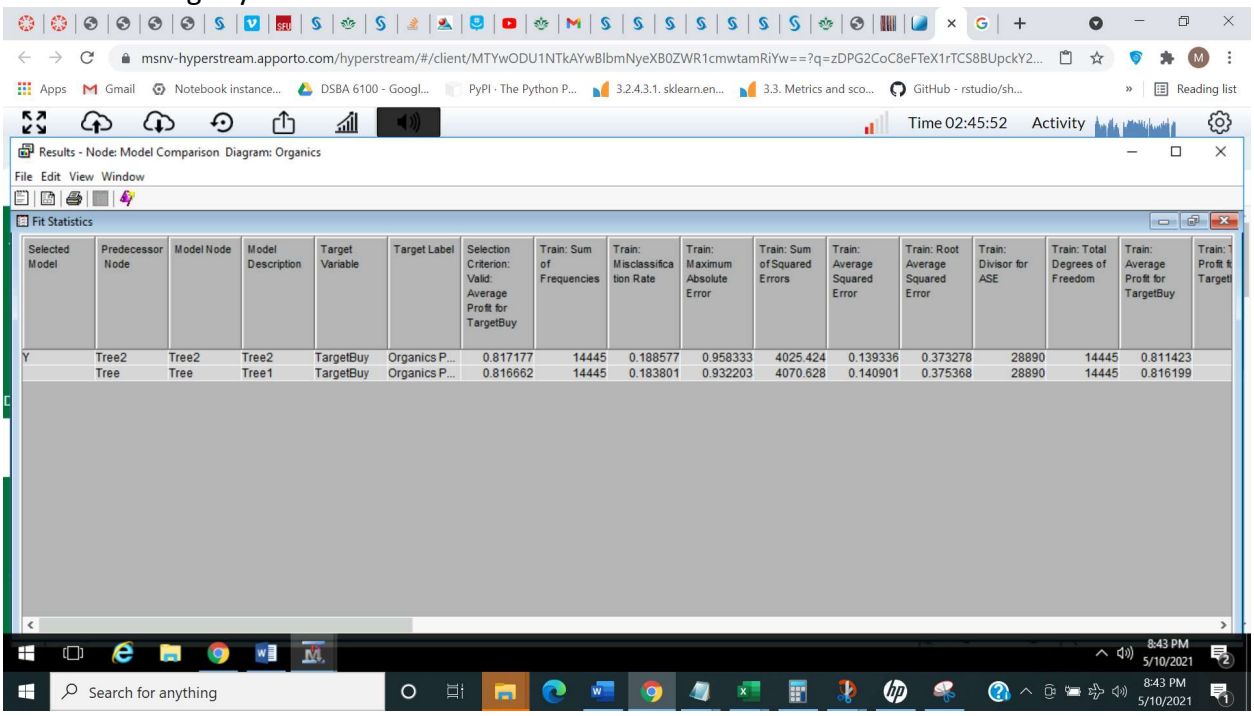
3) How many leaves are in the optimal tree? 16 in Tree2



h) Based on **Misclassification rate**, which of the decision tree models appears to be better?

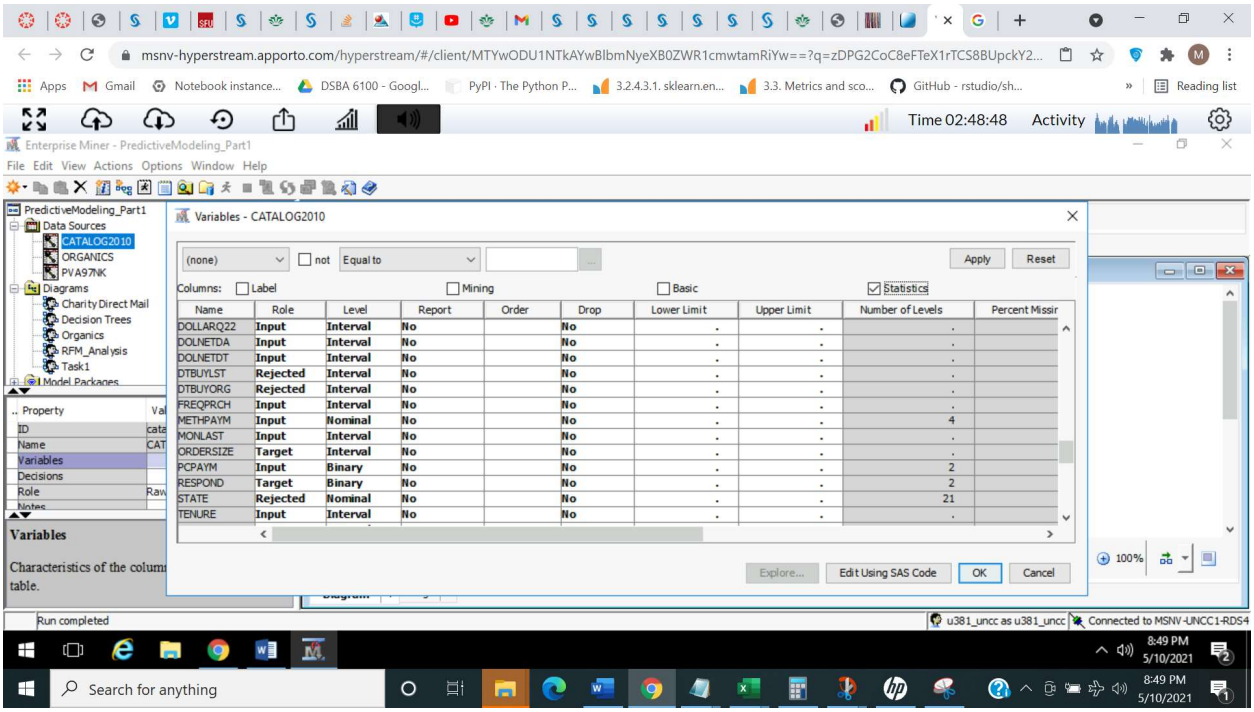


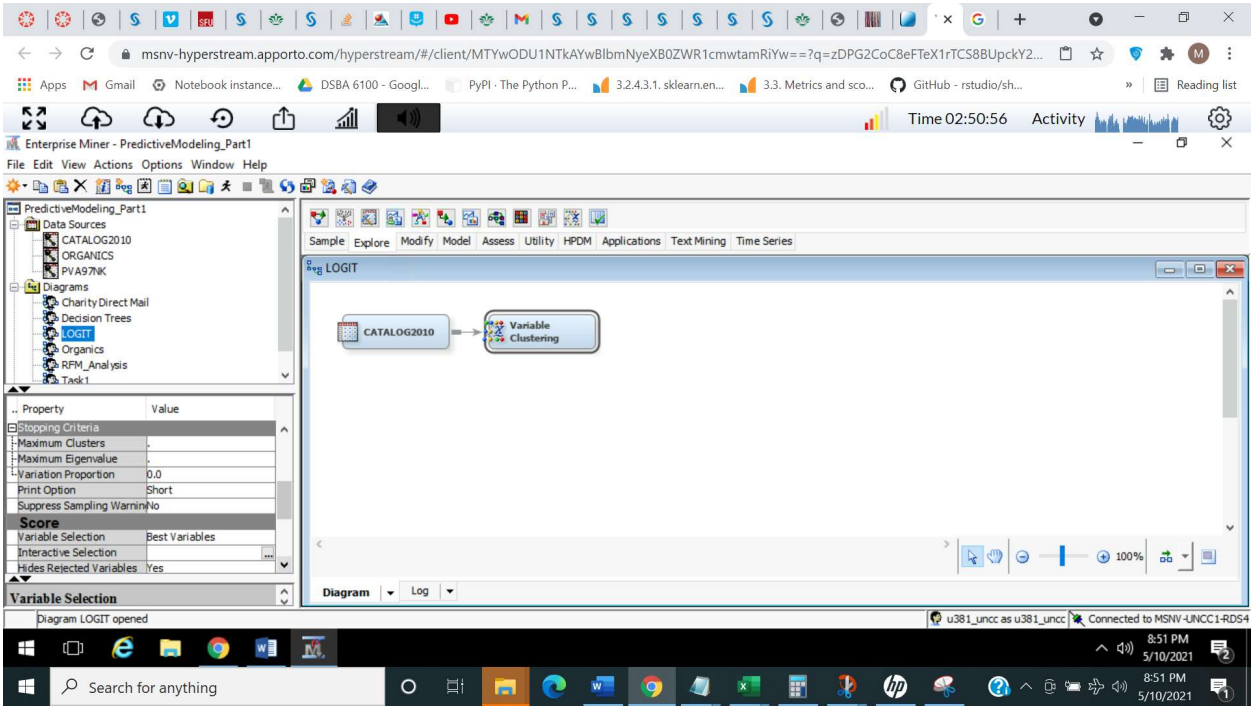
Tree2 has a slightly better misclassification rate of 0.188577 versus Tree1 0.183801.

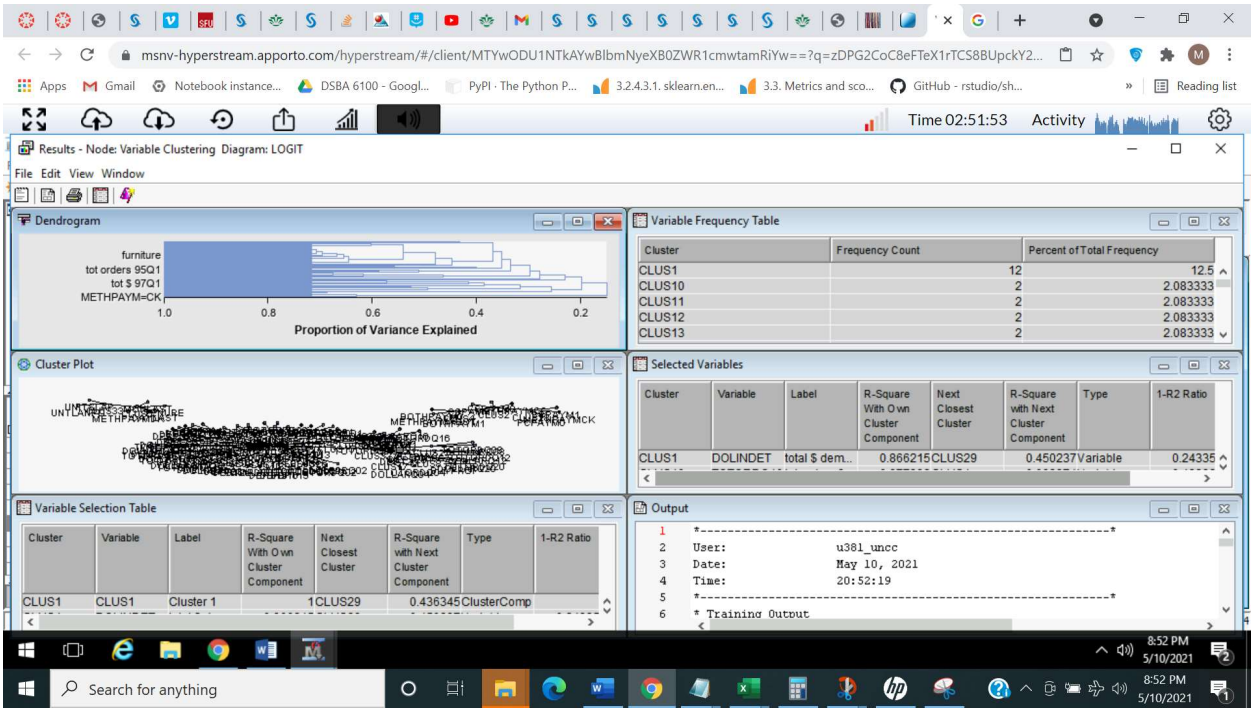


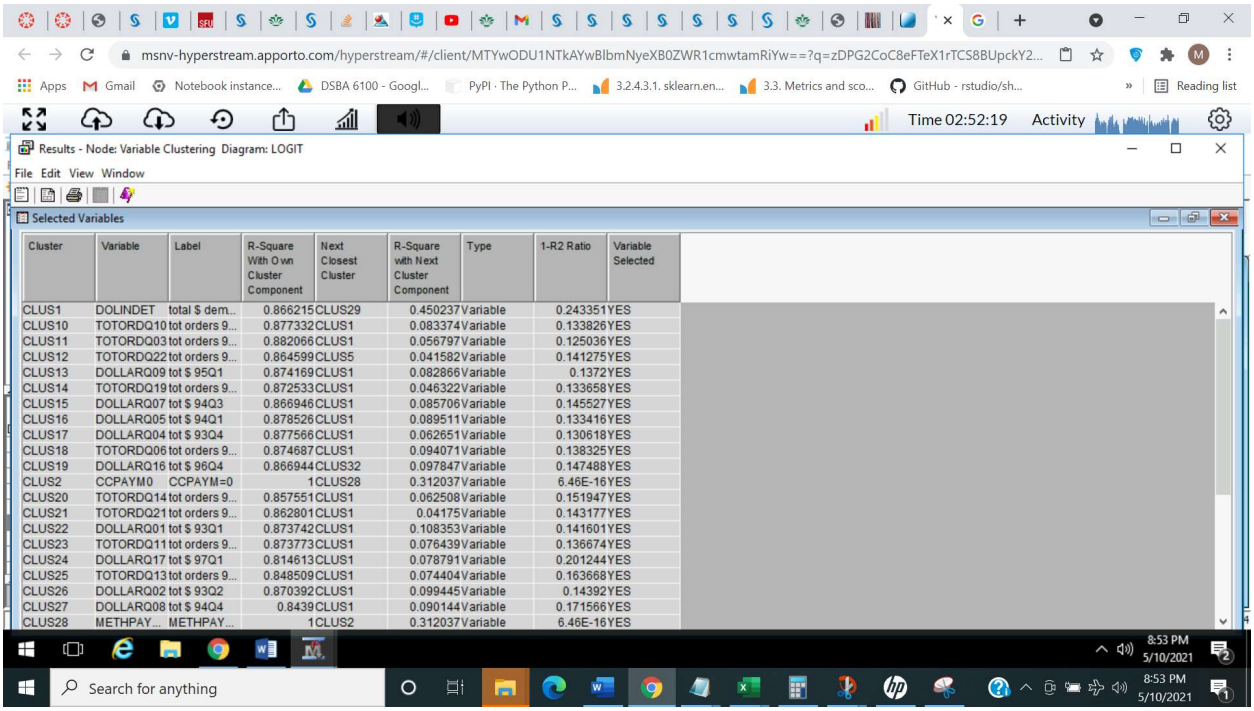
Part 3-Predictive Modeling (LOGIT)

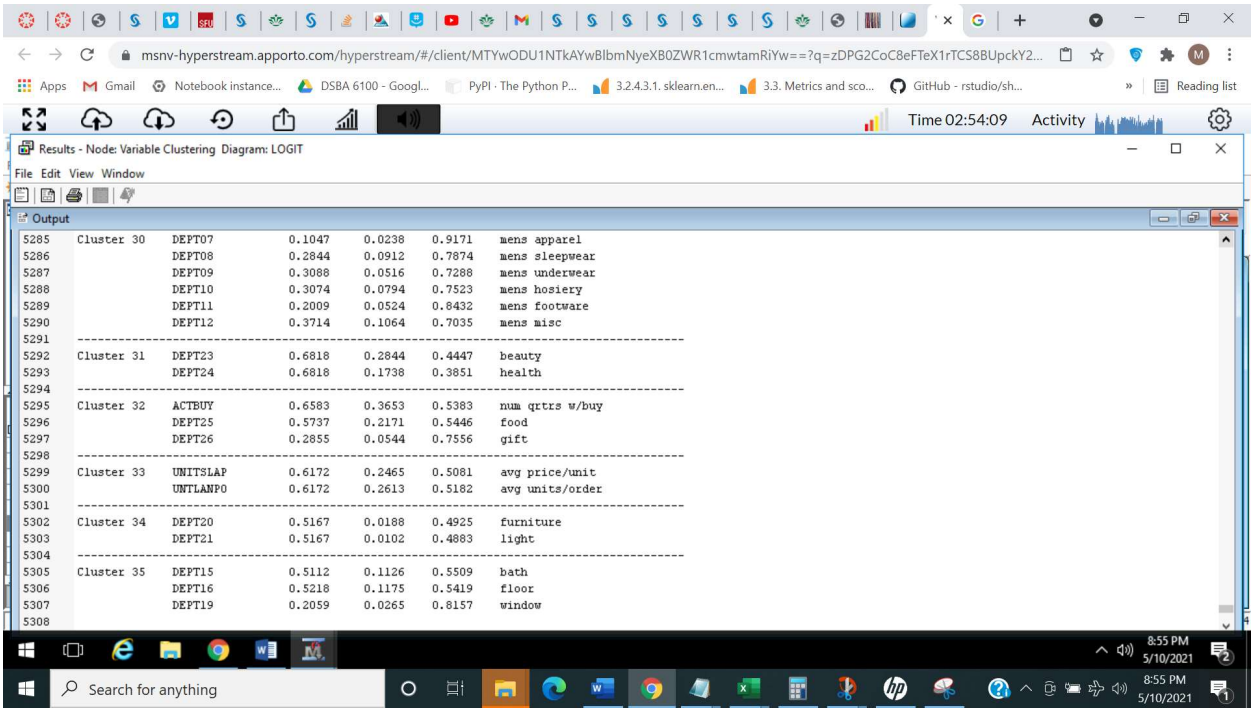
CATALOG CASE STUDY: FITTING A LOGISTIC REGRESSION MODEL

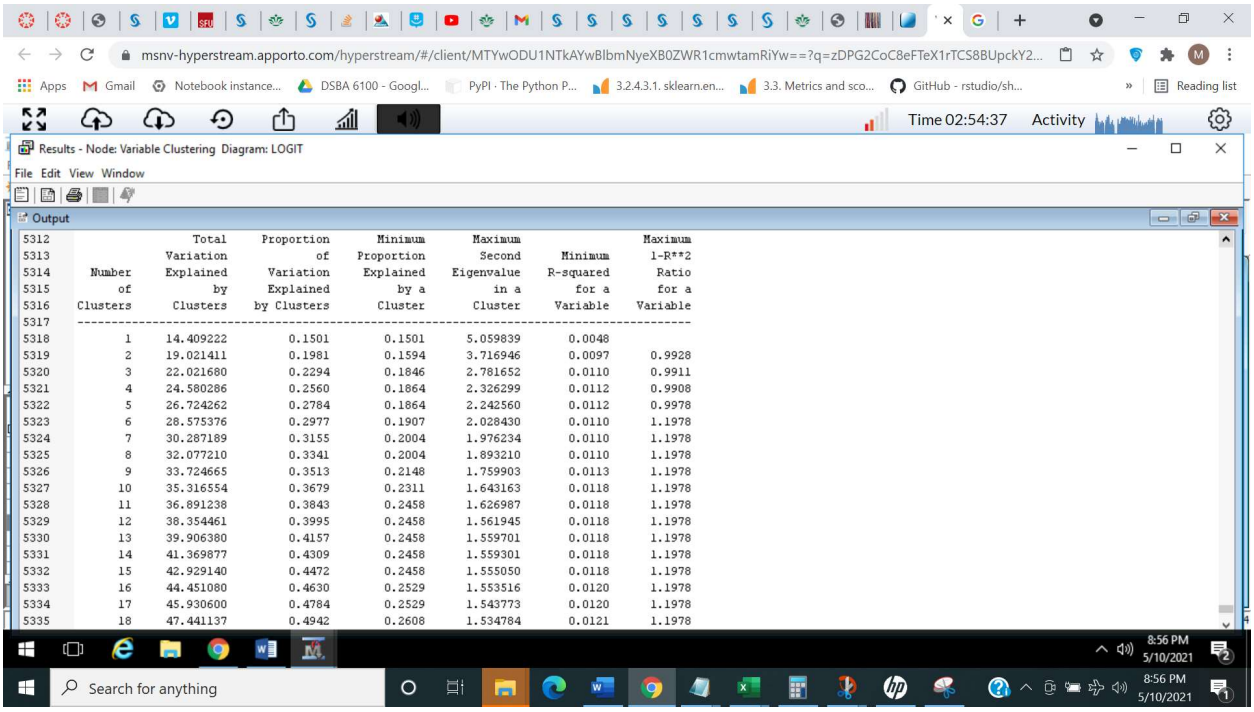


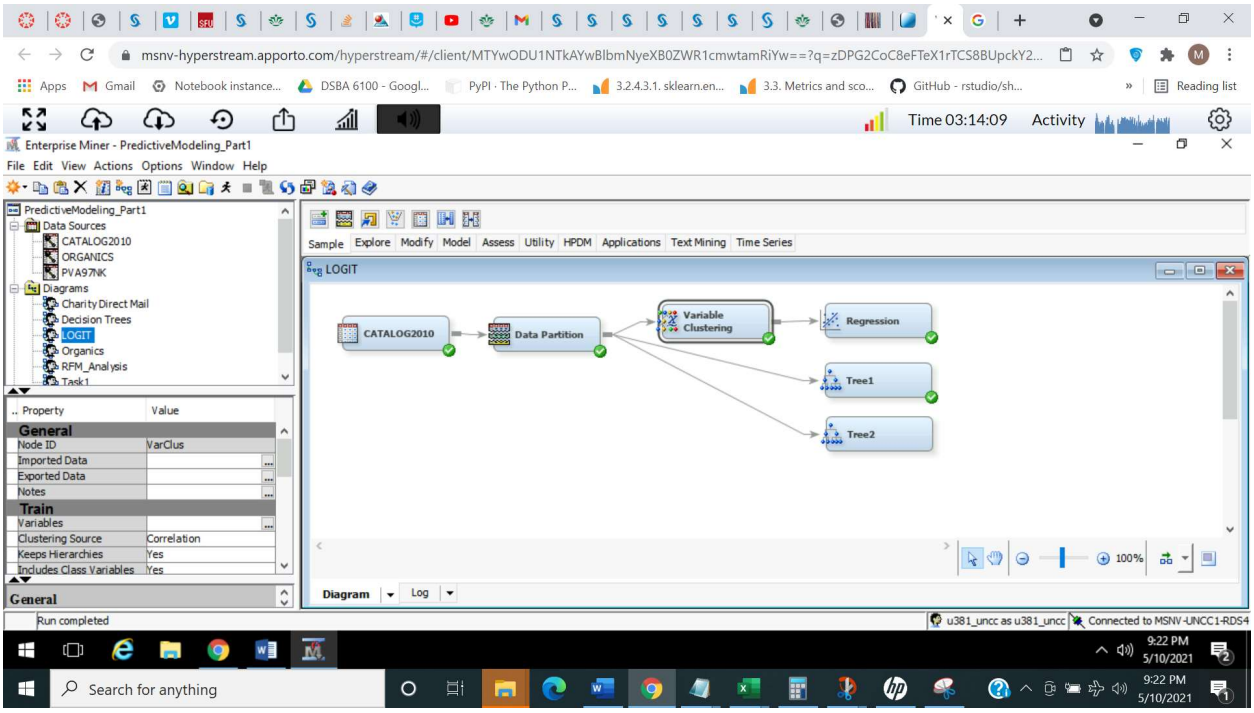


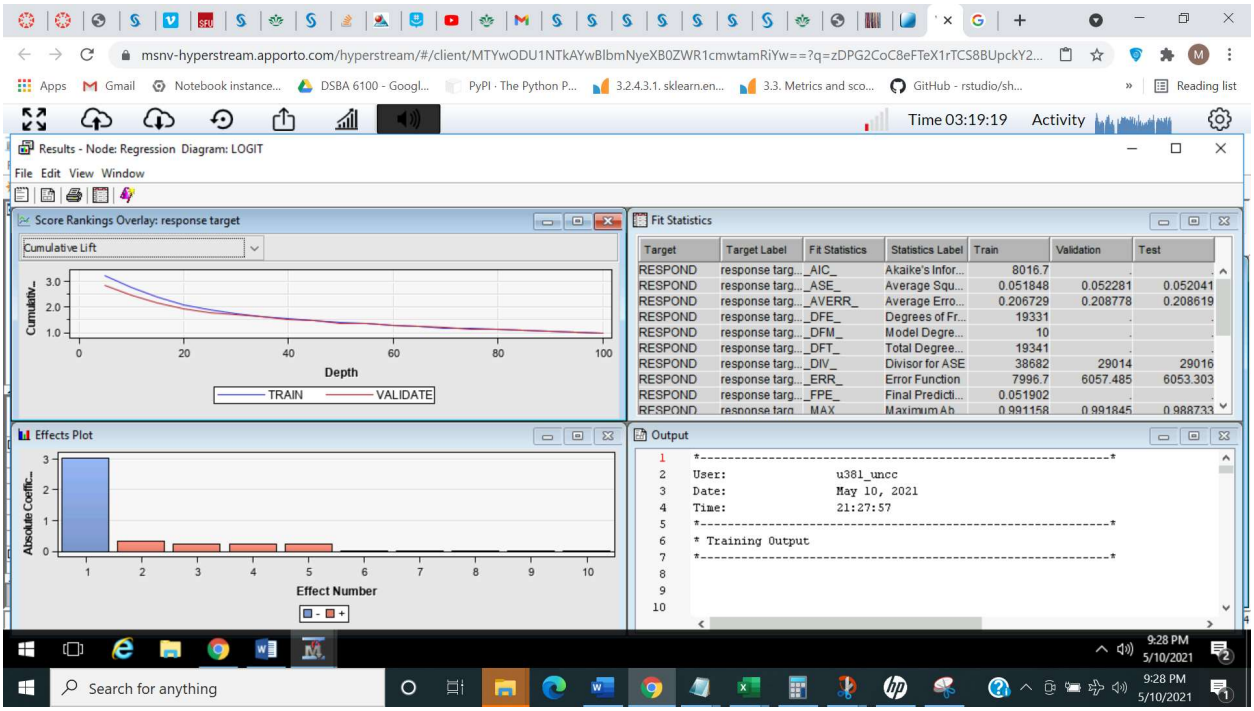

msnv-hyperstream.apporto.com/hyperstream/#/client/MTYwODU1NTkAYwBlbmNyeXB0ZWR1cmwtamRiYw==?q=zDPG2CoC8eFTeX1rTCS8BUpckY2...
Time 02:51:53
Activity
Results - Node: Variable Clustering Diagram: LOGIT
File Edit View Window
Dendrogram
furniture
tot orders 95Q1
tot $ 97Q1
METHPAYM=CK
1.0 0.8 0.6 0.4 0.2
Proportion of Variance Explained
Variable Frequency Table
Cluster | Frequency Count | Percent of Total Frequency
CLUS1 | 12 | 12.5
CLUS10 | 2 | 2.083333
CLUS11 | 2 | 2.083333
CLUS12 | 2 | 2.083333
CLUS13 | 2 | 2.083333
Cluster Plot
Selected Variables
Cluster | Variable | Label | R-Square With Own Cluster Component | Next Closest Cluster | R-Square with Next Cluster Component | Type | 1-R2 Ratio
CLUS1 | DOLINDET | total $ dem... | 0.866215 | CLUS29 | 0.450237 | Variable | 0.24335
Variable Selection Table
Cluster | Variable | Label | R-Square With Own Cluster Component | Next Closest Cluster | R-Square with Next Cluster Component | Type | 1-R2 Ratio
CLUS1 | CLUS1 | Cluster 1 | 1 | CLUS29 | 0.436345 | ClusterComp
Output
1 *--*
2 User: u381_uncc
3 Date: May 10, 2021
4 Time: 20:52:19
5 *--*
6 * Training Output
8:52 PM
5/10/2021
Search for anything
8:52 PM
5/10/2021

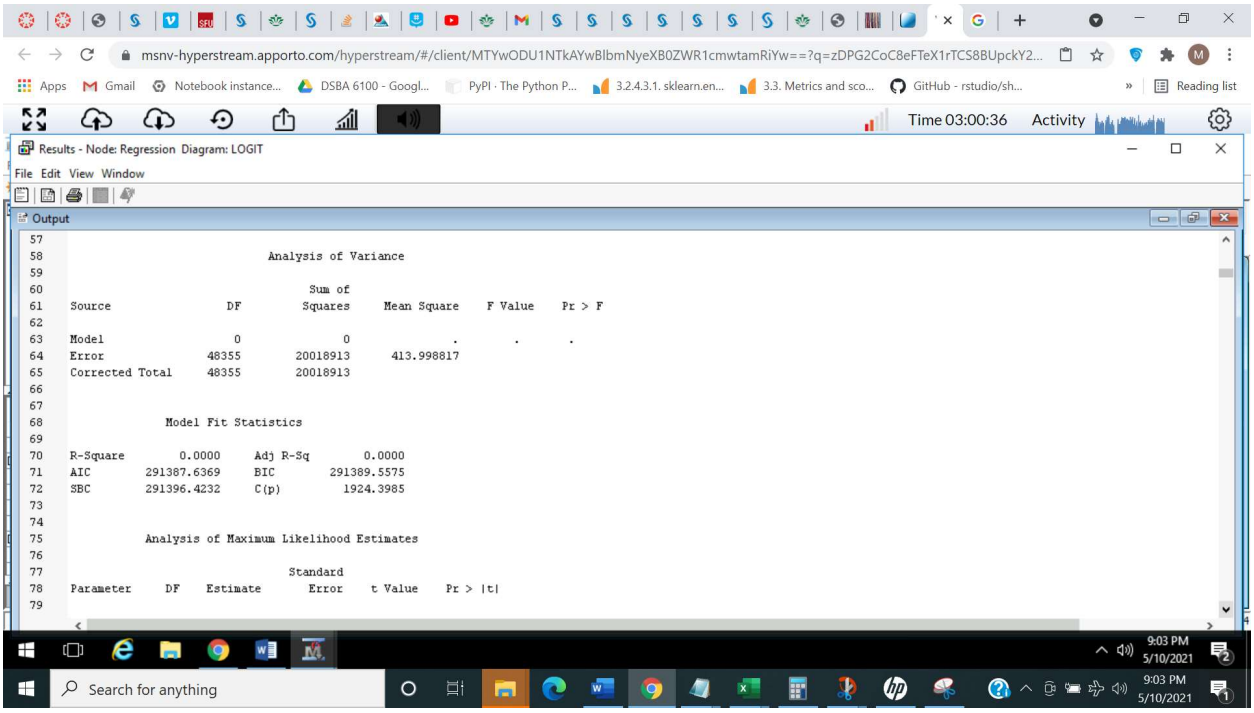

msnv-hyperstream.apporto.com/hyperstream/#/client/MTYwODU1NTkAYwBlbmNyeXB0ZWR1cmwtamRiYw==?q=zDPG2CoC8eFTeX1rTCS8BUpckY2...
Time 02:52:19
Activity
Results - Node: Variable Clustering Diagram: LOGIT
File Edit View Window
Selected Variables
Cluster | Variable | Label | R-Square With Own Cluster Component | Next Closest Cluster | R-Square with Next Cluster Component | Type | 1-R2 Ratio | Variable Selected
CLUS1 | DOLINDET | total $ dem... | 0.866215 | CLUS29 | 0.450237 | Variable | 0.243351 | YES
CLUS10 | TOTORDQ10 | tot orders 9... | 0.877332 | CLUS1 | 0.083374 | Variable | 0.133826 | YES
CLUS11 | TOTORDQ03 | tot orders 9... | 0.882066 | CLUS1 | 0.056797 | Variable | 0.125036 | YES
CLUS12 | TOTORDQ22 | tot orders 9... | 0.864599 | CLUS5 | 0.041582 | Variable | 0.141275 | YES
CLUS13 | DOLLARQ09 | tot $ 95Q1 | 0.874169 | CLUS1 | 0.082866 | Variable | 0.1372 | YES
CLUS14 | TOTORDQ19 | tot orders 9... | 0.872533 | CLUS1 | 0.046322 | Variable | 0.133658 | YES
CLUS15 | DOLLARQ07 | tot $ 94Q3 | 0.866946 | CLUS1 | 0.085706 | Variable | 0.145527 | YES
CLUS16 | DOLLARQ05 | tot $ 94Q1 | 0.878526 | CLUS1 | 0.089511 | Variable | 0.133416 | YES
CLUS17 | DOLLARQ04 | tot $ 93Q4 | 0.877566 | CLUS1 | 0.062651 | Variable | 0.130618 | YES
CLUS18 | TOTORDQ06 | tot orders 9... | 0.874687 | CLUS1 | 0.094071 | Variable | 0.138325 | YES
CLUS19 | DOLLARQ16 | tot $ 96Q4 | 0.866944 | CLUS32 | 0.097847 | Variable | 0.147488 | YES
CLUS2 | CCPAYM0 | CCPAYM=0 | 1 | CLUS28 | 0.312037 | Variable | 6.46E-16 | YES
CLUS20 | TOTORDQ14 | tot orders 9... | 0.857551 | CLUS1 | 0.062508 | Variable | 0.151947 | YES
CLUS21 | TOTORDQ21 | tot orders 9... | 0.862801 | CLUS1 | 0.04175 | Variable | 0.143177 | YES
CLUS22 | DOLLARQ01 | tot $ 93Q1 | 0.873742 | CLUS1 | 0.108353 | Variable | 0.141601 | YES
CLUS23 | TOTORDQ11 | tot orders 9... | 0.873773 | CLUS1 | 0.076439 | Variable | 0.136674 | YES
CLUS24 | DOLLARQ17 | tot $ 97Q1 | 0.814613 | CLUS1 | 0.078791 | Variable | 0.201244 | YES
CLUS25 | TOTORDQ13 | tot orders 9... | 0.848509 | CLUS1 | 0.074404 | Variable | 0.163668 | YES
CLUS26 | DOLLARQ02 | tot $ 93Q2 | 0.870392 | CLUS1 | 0.099445 | Variable | 0.14392 | YES
CLUS27 | DOLLARQ08 | tot $ 94Q4 | 0.8439 | CLUS1 | 0.090144 | Variable | 0.171566 | YES
CLUS28 | METHPAY... | METHPAY... | 1 | CLUS2 | 0.312037 | Variable | 6.46E-16 | YES
8:53 PM
5/10/2021
Search for anything
8:53 PM
5/10/2021

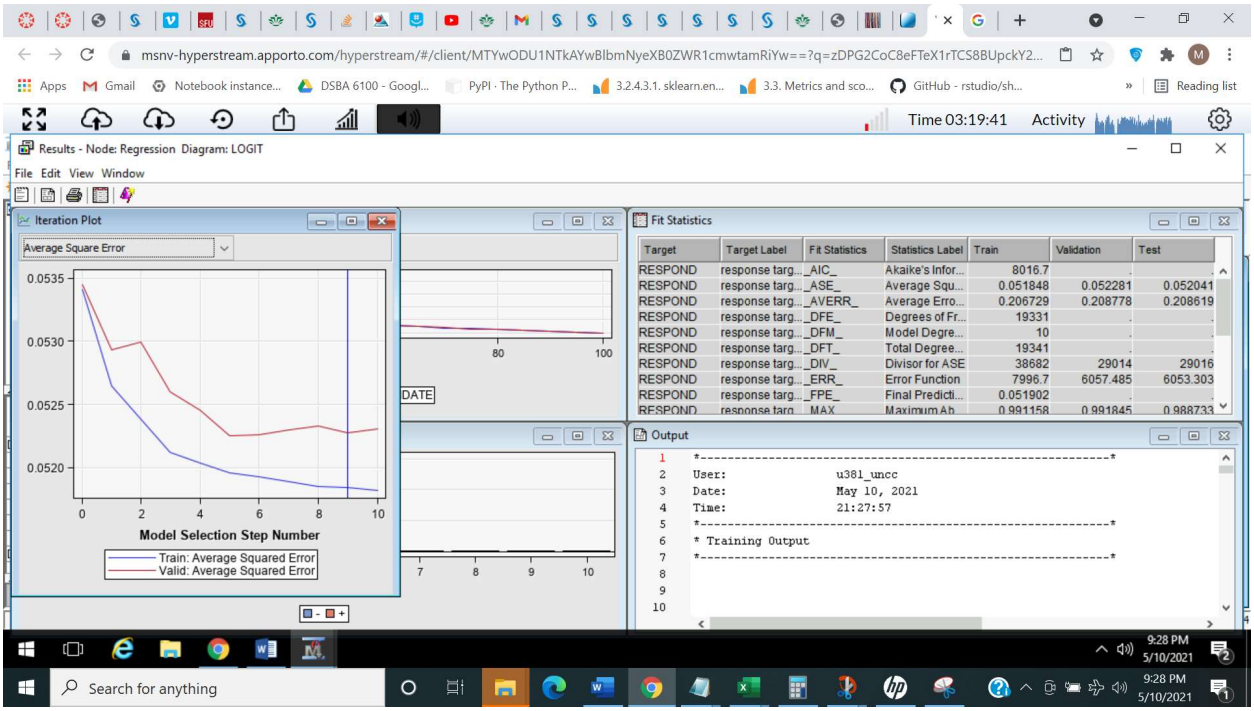

msnv-hyperstream.apporto.com/hyperstream/#/client/MTYwODU1NTkAYwBlbmNyeXB0ZWR1cmwtamRiYw==?q=zDPG2CoC8eFTeX1rTCS8BUpckY2...
Time 02:54:09
Activity
Results - Node: Variable Clustering Diagram: LOGIT
File Edit View Window
Output
5285 Cluster 30 DEPT07 0.1047 0.0238 0.9171 mens apparel
5286 DEPT08 0.2844 0.0912 0.7874 mens sleepwear
5287 DEPT09 0.3088 0.0516 0.7288 mens underwear
5288 DEPT10 0.3074 0.0794 0.7523 mens hosiery
5289 DEPT11 0.2009 0.0524 0.8432 mens footware
5290 DEPT12 0.3714 0.1064 0.7035 mens misc
5291 ---
5292 Cluster 31 DEPT23 0.6818 0.2844 0.4447 beauty
5293 DEPT24 0.6818 0.1738 0.3851 health
5294 ---
5295 Cluster 32 ACTBUY 0.6583 0.3653 0.5383 num qrtrs w/buy
5296 DEPT25 0.5737 0.2171 0.5446 food
5297 DEPT26 0.2855 0.0544 0.7556 gift
5298 ---
5299 Cluster 33 UNITSLAP 0.6172 0.2465 0.5081 avg price/unit
5300 UNTLANP0 0.6172 0.2613 0.5182 avg units/order
5301 ---
5302 Cluster 34 DEPT20 0.5167 0.0188 0.4925 furniture
5303 DEPT21 0.5167 0.0102 0.4883 light
5304 ---
5305 Cluster 35 DEPT15 0.5112 0.1126 0.5509 bath
5306 DEPT16 0.5218 0.1175 0.5419 floor
5307 DEPT19 0.2059 0.0265 0.8157 window
5308
8:55 PM
5/10/2021
Search for anything
8:55 PM
5/10/2021

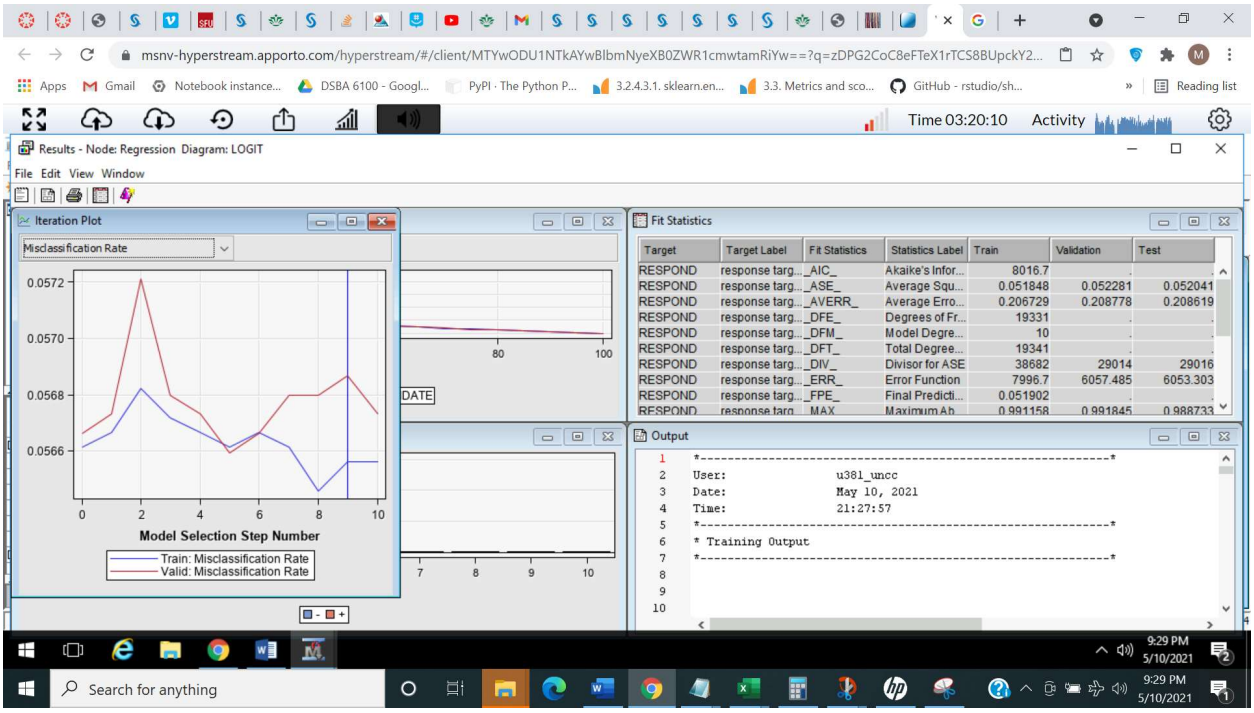

msnv-hyperstream.apporto.com/hyperstream/#/client/MTYwODU1NTkAYwBlbmNyeXB0ZWR1cmwtamRiYw==?q=zDPG2CoC8eFTeX1rTCS8BUpckY2...
Time 02:54:37
Activity
Results - Node: Variable Clustering Diagram: LOGIT
File Edit View Window
Output
5312 Total Proportion Minimum Maximum Maximum
5313 Variation of Proportion Second Minimum 1-R**2
5314 Number Explained Variation Explained Eigenvalue R-squared Ratio
5315 of by Explained by a in a for a for a
5316 Clusters Clusters by Clusters Cluster Cluster Variable Variable
5317 ---
5318 1 14.409222 0.1501 0.1501 5.059839 0.0048
5319 2 19.021411 0.1981 0.1594 3.716946 0.0097 0.9928
5320 3 22.021680 0.2294 0.1846 2.781652 0.0110 0.9911
5321 4 24.580286 0.2560 0.1864 2.326299 0.0112 0.9908
5322 5 26.724262 0.2784 0.1864 2.242560 0.0112 0.9978
5323 6 28.575376 0.2977 0.1907 2.028430 0.0110 1.1978
5324 7 30.287189 0.3155 0.2004 1.976234 0.0110 1.1978
5325 8 32.077210 0.3341 0.2004 1.893210 0.0110 1.1978
5326 9 33.724665 0.3513 0.2148 1.759903 0.0113 1.1978
5327 10 35.316554 0.3679 0.2311 1.643163 0.0118 1.1978
5328 11 36.891238 0.3843 0.2458 1.626987 0.0118 1.1978
5329 12 38.354461 0.3995 0.2458 1.561945 0.0118 1.1978
5330 13 39.906380 0.4157 0.2458 1.559701 0.0118 1.1978
5331 14 41.369877 0.4309 0.2458 1.559301 0.0118 1.1978
5332 15 42.929140 0.4472 0.2458 1.555050 0.0118 1.1978
5333 16 44.451080 0.4630 0.2529 1.553516 0.0120 1.1978
5334 17 45.930600 0.4784 0.2529 1.543773 0.0120 1.1978
5335 18 47.441137 0.4942 0.2608 1.534784 0.0121 1.1978
8:56 PM
5/10/2021
Search for anything
8:56 PM
5/10/2021

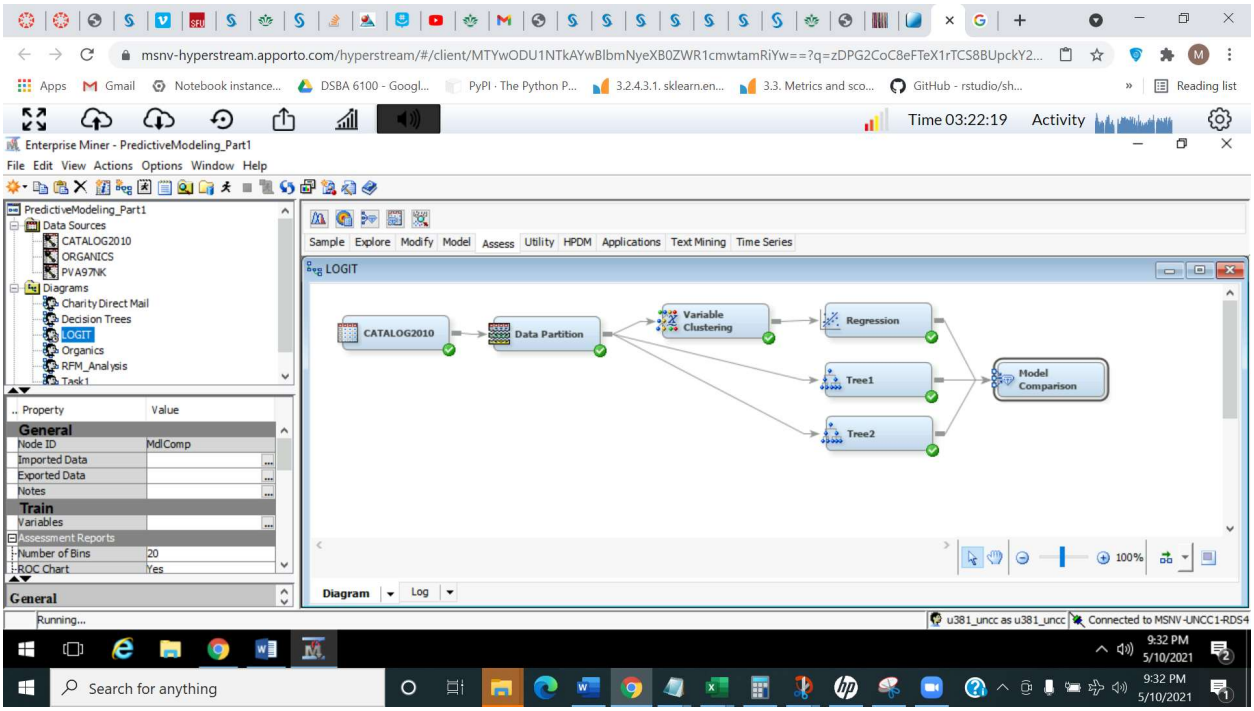

Enterprise Miner - PredictiveModeling_Part1
File Edit View Actions Options Window Help
PredictiveModeling_Part1
Data Sources
CATALOG2010
ORGANICS
PVA97NK
Diagrams
Charity Direct Mail
Decision Trees
LOGIT
Organics
RFM_Analysis
Task1
Sample Explore Modify Model Assess Utility HPDM Applications Text Mining Time Series
LOGIT
CATALOG2010
Data Partition
Variable Clustering
Regression
Tree1
Tree2
Property Value
General
Node ID VarClus
Imported Data
Exported Data
Notes
Train
Variables
Clustering Source Correlation
Keeps Hierarchies Yes
Includes Class Variables Yes
Diagram Log
Run completed
u381_uncc as u381_uncc
Connected to MSNV-UNCC1-RDS4
Time 03:14:09
Activity
9:22 PM
5/10/2021
Search for anything

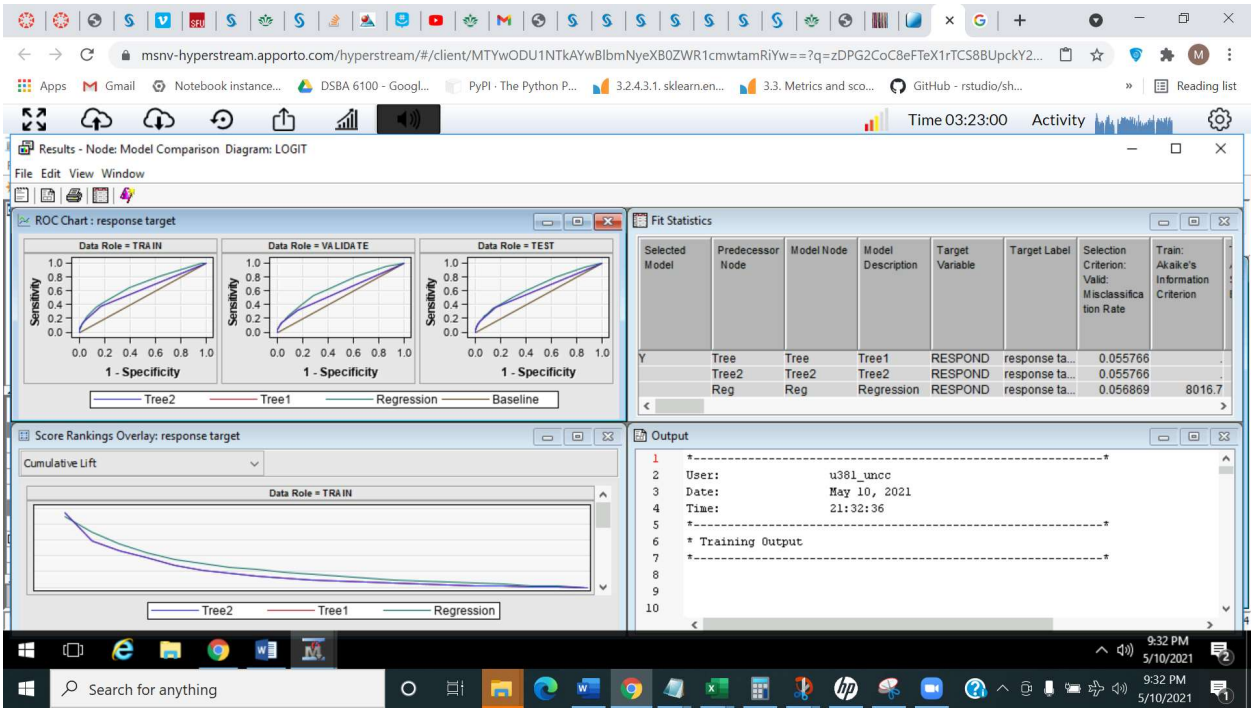

Results - Node: Regression Diagram: LOGIT
File Edit View Window
Score Rankings Overlay: response target
Cumulative Lift
Cumulative...
Depth
TRAIN
VALIDATE
Fit Statistics
Target Target Label Fit Statistics Statistics Label Train Validation Test
RESPOND response targ... _AIC_ Akaike's Infor... 8016.7
RESPOND response targ... _ASE_ Average Squ... 0.051848 0.052281 0.052041
RESPOND response targ... _AVERR_ Average Erro... 0.206729 0.208778 0.208619
RESPOND response targ... _DFE_ Degrees of Fr... 19331
RESPOND response targ... _DFM_ Model Degre... 10
RESPOND response targ... _DFT_ Total Degree... 19341
RESPOND response targ... _DIV_ Divisor for ASE 38682 29014 29016
RESPOND response targ... _ERR_ Error Function 7996.7 6057.485 6053.303
RESPOND response targ... _FPE_ Final Predicti... 0.051902
RESPOND response targ... _MAX_ Maximum Ab... 0.991158 0.991845 0.988733
Effects Plot
Absolute Coeffic...
Effect Number
Output
User: u381_uncc
Date: May 10, 2021
Time: 21:27:57
* Training Output
Time 03:19:19
Activity
9:28 PM
5/10/2021
Search for anything

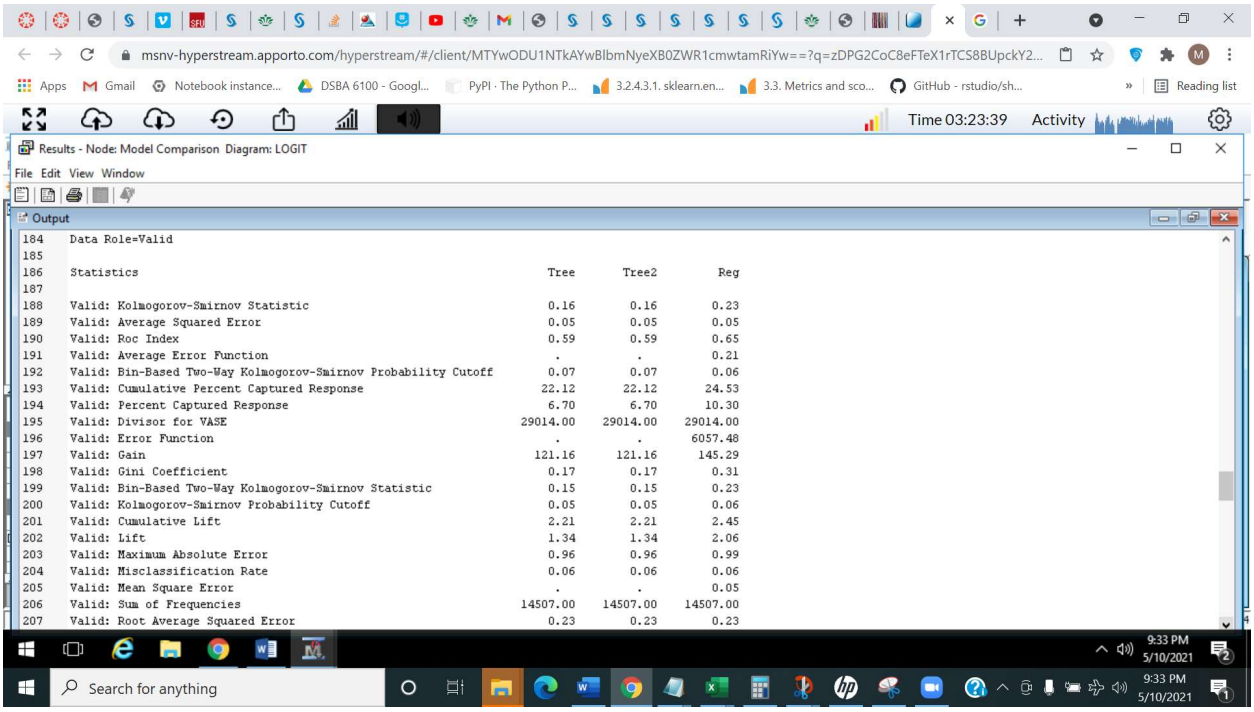

msnv-hyperstream.apporto.com/hyperstream/#/client/MTYwODU1NTkAYwBlbmNyeXB0ZWR1cmwtamRiYw==?q=zDPG2CoC8eFTeX1rTCS8BUpckY2...
Time 03:00:36
Activity
Results - Node: Regression Diagram: LOGIT
File Edit View Window
Output
Analysis of Variance
Sum of
Source DF Squares Mean Square F Value Pr > F
Model 0 0 . . .
Error 48355 20018913 413.998817
Corrected Total 48355 20018913
Model Fit Statistics
R-Square 0.0000 Adj R-Sq 0.0000
AIC 291387.6369 BIC 291389.5575
SBC 291396.4232 C(p) 1924.3985
Analysis of Maximum Likelihood Estimates
Parameter DF Estimate Standard Error t Value Pr > |t|
Search for anything
9:03 PM
5/10/2021


msnv-hyperstream.apporto.com/hyperstream/#/client/MTYwODU1NTkAYwBlbmNyeXB0ZWR1cmwtamRiYw==?q=zDPG2CoC8eFTeX1rTCS8BUpckY2...
Time 03:19:41
Activity
Results - Node: Regression Diagram: LOGIT
File Edit View Window
Iteration Plot
Average Square Error
0.0535
0.0530
0.0525
0.0520
0 2 4 6 8 10
Model Selection Step Number
Train: Average Squared Error
Valid: Average Squared Error
Fit Statistics
Target Target Label Fit Statistics Statistics Label Train Validation Test
RESPOND response targ... _AIC_ Akaike's Infor... 8016.7
RESPOND response targ... _ASE_ Average Squ... 0.051848 0.052281 0.052041
RESPOND response targ... _AVERR_ Average Erro... 0.206729 0.208778 0.208619
RESPOND response targ... _DFE_ Degrees of Fr... 19331
RESPOND response targ... _DFM_ Model Degre... 10
RESPOND response targ... _DFT_ Total Degree... 19341
RESPOND response targ... _DIV_ Divisor for ASE 38682 29014 29016
RESPOND response targ... _ERR_ Error Function 7996.7 6057.485 6053.303
RESPOND response targ... _FPE_ Final Predicti... 0.051902
RESPOND response targ... _MAX_ Maximum Ab... 0.991158 0.991845 0.988733
Output
User: u381_uncc
Date: May 10, 2021
Time: 21:27:57
* Training Output
Search for anything
9:28 PM
5/10/2021


msnv-hyperstream.apporto.com/hyperstream/#/client/MTYwODU1NTkAYwBlbmNyeXB0ZWR1cmwtamRiYw==?q=zDPG2CoC8eFTeX1rTCS8BUpckY2...
Time 03:20:10
Activity
Results - Node: Regression Diagram: LOGIT
File Edit View Window
Iteration Plot
Misclassification Rate
0.0572
0.0570
0.0568
0.0566
0 2 4 6 8 10
Model Selection Step Number
Train: Misclassification Rate
Valid: Misclassification Rate
DATE
Fit Statistics
Target | Target Label | Fit Statistics | Statistics Label | Train | Validation | Test
RESPOND | response targ... | _AIC_ | Akaike's Infor... | 8016.7 | |
RESPOND | response targ... | _ASE_ | Average Squ... | 0.051848 | 0.052281 | 0.052041
RESPOND | response targ... | _AVERR_ | Average Erro... | 0.206729 | 0.208778 | 0.208619
RESPOND | response targ... | _DFE_ | Degrees of Fr... | 19331 | |
RESPOND | response targ... | _DFM_ | Model Degre... | 10 | |
RESPOND | response targ... | _DFT_ | Total Degree... | 19341 | |
RESPOND | response targ... | _DIV_ | Divisor for ASE | 38682 | 29014 | 29016
RESPOND | response targ... | _ERR_ | Error Function | 7996.7 | 6057.485 | 6053.303
RESPOND | response targ... | _FPE_ | Final Predicti... | 0.051902 | |
Output
User: u381_uncc
Date: May 10, 2021
Time: 21:27:57
* Training Output
Search for anything
9:29 PM
5/10/2021


msnv-hyperstream.apporto.com/hyperstream/#/client/MTYwODU1NTkAYwBlbmNyeXB0ZWR1cmwtamRiYw==?q=zDPG2CoC8eFTeX1rTCS8BUpckY2...
Time 03:22:19
Activity
Enterprise Miner - PredictiveModeling_Part1
File Edit View Actions Options Window Help
PredictiveModeling_Part1
Data Sources
CATALOG2010
ORGANICS
PVA97NK
Diagrams
Charity Direct Mail
Decision Trees
LOGIT
Organics
RFM_Analysis
Task1
Property | Value
General
Node ID MdlComp
Imported Data
Exported Data
Notes
Train
Variables
Assessment Reports
Number of Bins 20
ROC Chart Yes
Sample Explore Modify Model Assess Utility HPDM Applications Text Mining Time Series
LOGIT
CATALOG2010
Data Partition
Variable Clustering
Regression
Tree1
Tree2
Model Comparison
100%
Diagram
Log
Running...
u381_uncc as u381_uncc
Connected to MSNV-UNCC1-RDS4
Search for anything
9:32 PM
5/10/2021


msnv-hyperstream.apporto.com/hyperstream/#/client/MTYwODU1NTkAYwBlbmNyeXB0ZWR1cmwtamRiYw==?q=zDPG2CoC8eFTeX1rTCS8BUpckY2...
Time 03:23:00
Activity
Results - Node: Model Comparison Diagram: LOGIT
File Edit View Window
ROC Chart : response target
Data Role = TRAIN
Data Role = VALIDATE
Data Role = TEST
Sensitivity
1 - Specificity
Tree2
Tree1
Regression
Baseline
Fit Statistics
Selected Model
Predecessor Node
Model Node
Model Description
Target Variable
Target Label
Selection Criterion: Valid: Misclassification Rate
Train: Akaike's Information Criterion
Y Tree Tree Tree1 RESPOND response ta... 0.055766
Tree2 Tree2 Tree2 RESPOND response ta... 0.055766
Reg Reg Regression RESPOND response ta... 0.056869 8016.7
Score Rankings Overlay: response target
Cumulative Lift
Data Role = TRAIN
Output
User: u381_uncc
Date: May 10, 2021
Time: 21:32:36
* Training Output
9:32 PM
5/10/2021
Search for anything


msnv-hyperstream.apporto.com/hyperstream/#/client/MTYwODU1NTkAYwBlbmNyeXB0ZWR1cmwtamRiYw==?q=zDPG2CoC8eFTeX1rTCS8BUpckY2...
Time 03:23:39
Activity
Results - Node: Model Comparison Diagram: LOGIT
File Edit View Window
Output
Data Role=Valid
Statistics Tree Tree2 Reg
Valid: Kolmogorov-Smirnov Statistic 0.16 0.16 0.23
Valid: Average Squared Error 0.05 0.05 0.05
Valid: Roc Index 0.59 0.59 0.65
Valid: Average Error Function . . 0.21
Valid: Bin-Based Two-Way Kolmogorov-Smirnov Probability Cutoff 0.07 0.07 0.06
Valid: Cumulative Percent Captured Response 22.12 22.12 24.53
Valid: Percent Captured Response 6.70 6.70 10.30
Valid: Divisor for VASE 29014.00 29014.00 29014.00
Valid: Error Function . . 6057.48
Valid: Gain 121.16 121.16 145.29
Valid: Gini Coefficient 0.17 0.17 0.31
Valid: Bin-Based Two-Way Kolmogorov-Smirnov Statistic 0.15 0.15 0.23
Valid: Kolmogorov-Smirnov Probability Cutoff 0.05 0.05 0.06
Valid: Cumulative Lift 2.21 2.21 2.45
Valid: Lift 1.34 1.34 2.06
Valid: Maximum Absolute Error 0.96 0.96 0.99
Valid: Misclassification Rate 0.06 0.06 0.06
Valid: Mean Square Error . . 0.05
Valid: Sum of Frequencies 14507.00 14507.00 14507.00
Valid: Root Average Squared Error 0.23 0.23 0.23
9:33 PM
5/10/2021
Search for anything