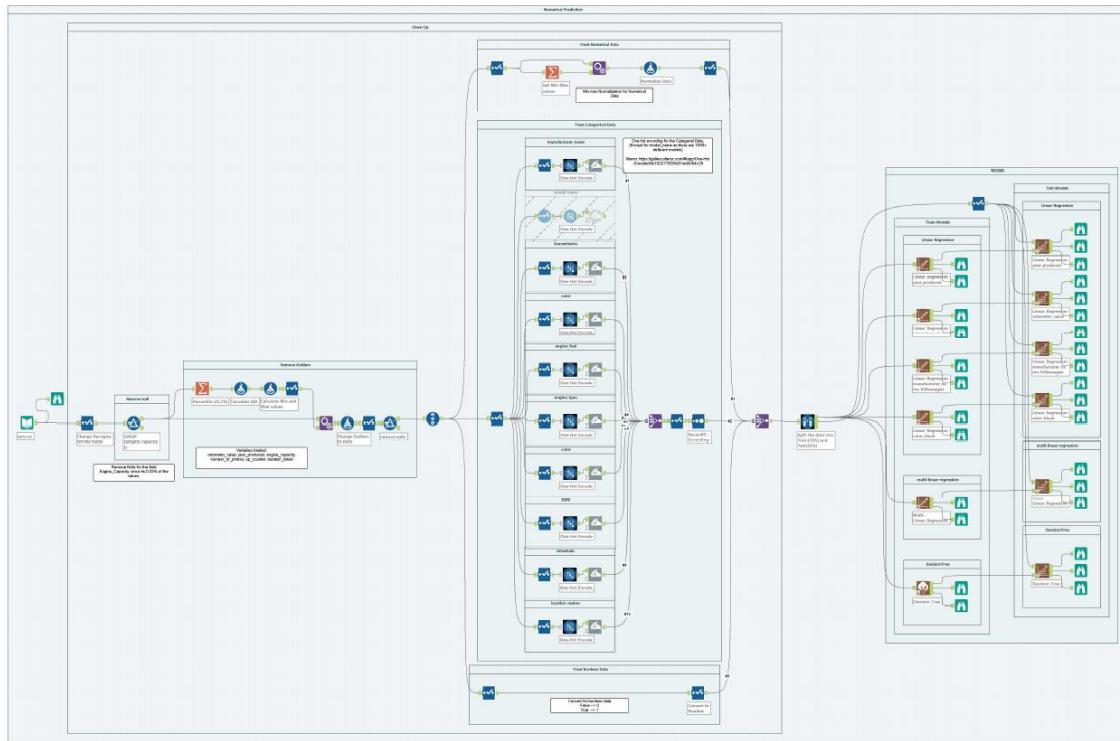


M6 Practical Challenge: Predictive Modeling with Alteryx

Table of Contents

Numerical Prediction.....	2
Imported Data:.....	2
Clean Up:	3
Models:.....	14
Best Model	20
Yes/No Prediction.....	21
Imported Data:.....	21
Clean Up:	21
Models:.....	27
Best Model	32

Numerical Prediction



Imported Data:

The imported values are:

Results - Select (12) - Output																		
	manufacturer_name	model_name	transmission	color	odometer_val...	year_produced	engine_fuel	engine_has_gas	engine_type	engine_capacity	body_type	has_warranty	state	drivetrain	price_usd	is_exchangeable	location_region	number_of_photo
1	Subaru	Outback	automatic	silver	190000	2010	gasoline	False	gasoline	2.5	universal	False	owned	all	10900	False	Минская обл.	9
2	Subaru	Outback	automatic	blue	280000	2002	gasoline	False	gasoline	3	universal	False	owned	all	5000	True	Минская обл.	12
3	Subaru	Forester	automatic	red	402000	2001	gasoline	False	gasoline	2.5	universal	False	owned	all	2800	True	Минская обл.	4
4	Subaru	Impreza	mechanical	blue	10000	1999	gasoline	False	gasoline	3	sedan	False	owned	all	9999	True	Минская обл.	9
5	Subaru	Legacy	automatic	black	280000	2001	gasoline	False	gasoline	2.5	universal	False	owned	all	213411	True	Гомельская обл.	14
6	Subaru	Outback	automatic	silver	132449	2011	gasoline	False	gasoline	2.5	universal	False	owned	all	14700	True	Минская обл.	20
7	Subaru	Forester	automatic	black	318280	1998	gasoline	False	gasoline	2.5	universal	False	owned	all	3000	True	Минская обл.	8
8	Subaru	Legacy	automatic	silver	350000	2004	gasoline	False	gasoline	2.5	sedan	False	owned	all	4500	False	Брестская обл.	7
9	Subaru	Outback	automatic	grey	179000	2010	gasoline	False	gasoline	2.5	universal	False	owned	all	12900	False	Минская обл.	17
10	Subaru	Forester	automatic	silver	571317	1999	gasoline	False	gasoline	2.5	universal	False	owned	all	4200	True	Минская обл.	8
11	Subaru	Forester	mechanical	other	280000	2003	gasoline	False	gasoline	2	suv	False	owned	all	6900	True	Минская обл.	14
12	Subaru	Tribeca	automatic	grey	256000	2008	gasoline	False	gasoline	3.6	suv	False	owned	all	8350	True	Минская обл.	18
13	Subaru	Forester	mechanical	other	321000	2002	gasoline	False	gasoline	2	suv	False	owned	all	4300	False	Минская обл.	13
14	Subaru	Justy	mechanical	red	49999	2001	gasoline	False	gasoline	1.3	hatchback	False	owned	all	1666	False	Гомельская обл.	8
15	Subaru	Outback	automatic	brown	154685	2011	gasoline	False	gasoline	2.5	universal	False	owned	all	8600	True	Минская обл.	24
16	Subaru	Outback	automatic	black	163219	2004	gasoline	False	gasoline	2	universal	False	owned	all	7300	True	Минская обл.	17
17	Subaru	Outback	automatic	other	318650	2005	gasoline	False	gasoline	3	universal	False	owned	all	7587.97	True	Минская обл.	7
18	Subaru	Impreza	mechanical	blue	191000	2005	gasoline	False	gasoline	2	sedan	False	owned	all	10950	False	Минская обл.	12
19	Subaru	Forester	automatic	silver	179000	2014	gasoline	False	gasoline	2	suv	False	owned	all	12700	False	Минская обл.	14
20	Subaru	Forester	automatic	black	159000	2013	gasoline	False	gasoline	2	suv	False	owned	all	16500	False	Минская обл.	6
21	Subaru	Outback	automatic	white	257498	2008	gasoline	False	gasoline	2.5	universal	False	owned	all	8700	False	Гомельская обл.	10
22	Subaru	Tribeca	automatic	silver	11402	2005	gasoline	False	gasoline	3	suv	False	owned	all	7500	False	Минская обл.	18
23	Subaru	Tribeca	automatic	black	180000	2006	gasoline	False	gasoline	3	suv	False	owned	all	8650	False	Минская обл.	30
24	Subaru	Forester	automatic	blue	240000	2001	gasoline	False	gasoline	2.5	universal	False	owned	all	3500	True	Минская обл.	5
25	Subaru	Legacy	automatic	green	249448	2002	gasoline	False	gasoline	2.5	universal	False	owned	all	3800	True	Могилевская обл.	12
26	Subaru	Tribeca	automatic	other	250000	2007	gasoline	False	gasoline	3	suv	False	owned	all	7200	True	Гомельская обл.	7
27	Subaru	Outback	automatic	green	417000	1997	gasoline	False	gasoline	2	universal	False	owned	front	1850	True	Минская обл.	18
28	Subaru	Outback	automatic	violet	377000	1999	gasoline	False	gasoline	2.5	universal	False	owned	all	3800	False	Брестская обл.	12
29	Subaru	Impreza	mechanical	black	300000	1999	gasoline	False	gasoline	2	universal	False	owned	all	3000	False	Минская обл.	6
30	Subaru	Legacy	mechanical	black	270000	2004	gasoline	False	gasoline	2	sedan	False	owned	all	6200	False	Витебская обл.	10
31	Subaru	Forester	mechanical	green	44444	1999	gasoline	False	gasoline	2	universal	False	owned	all	3700	False	Минская обл.	6
32	Subaru	Legacy	automatic	black	299000	1994	gas	True	gasoline	2.2	universal	False	owned	all	650	False	Гомельская обл.	9
33	Subaru	Impreza	mechanical	violet	340000	1993	gasoline	False	gasoline	1.6	universal	False	owned	all	2000	True	Могилевская обл.	4
34	Subaru	Legacy	mechanical	black	123456	1991	gasoline	False	gasoline	2.2	sedan	False	owned	all	800	False	Брестская обл.	6
35	Subaru	Impreza	mechanical	silver	118000	2011	gasoline	False	gasoline	2.5	sedan	False	owned	all	14950	True	Минская обл.	26
36	Subaru	Impreza	automatic	black	246113	2008	gasoline	False	gasoline	1.5	hatchback	False	owned	all	5350	False	Минская обл.	13
37	Subaru	Outback	automatic	green	128748	2016	gasoline	False	gasoline	3.6	universal	False	owned	all	17059.12	False	Гродненская обл.	24

Results - Select (12) - Output

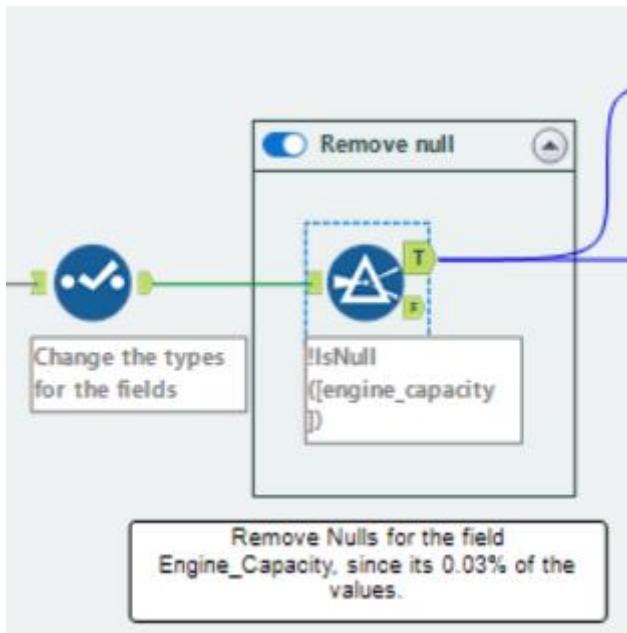
Cell Viewer • 3,868 of 38,511 records displayed [partial result] | ↑ ↓ |

Search Data Metadata Actions × 000

	body_type	has_warranty	state	drivetrain	price_usd	is_exchangeable	location_region	number_of_photos	up_counter	feature_0	feature_1	feature_2	feature_3	feature_4	feature_5	feature_6	feature_7	feature_8	feature_9	duration_list...
1	universal	False	owned	all	10900	False	Минская обл.	9	13	False	True	True	True	False	True	True	True	True	16	
2	universal	False	owned	all	5000	True	Минская обл.	12	54	False	True	False	True	False	False	False	True	True	83	
3	suv	False	owned	all	2800	True	Минская обл.	4	72	False	True	False	False	False	False	False	True	True	151	
4	sedan	False	owned	all	9999	True	Минская обл.	9	42	True	False	86								
5	universal	False	owned	all	2134.11	True	Гомельская обл.	14	7	False	True	True	True	False	False	False	True	True	7	
6	universal	False	owned	all	14700	True	Минская обл.	20	56	False	True	False	False	True	False	True	True	True	67	
7	universal	False	owned	all	3000	True	Минская обл.	8	147	False	True	False	True	True	False	False	True	True	307	
8	sedan	False	owned	all	4500	False	Брестская обл.	7	29	False	True	True	False	False	False	False	False	True	73	
9	universal	False	owned	all	12900	False	Минская обл.	17	33	False	True	87								
10	universal	False	owned	all	4200	True	Минская обл.	8	11	False	True	True	False	True	False	False	False	True	43	
11	suv	False	owned	all	6900	True	Минская обл.	14	6	False	True	True	False	True	True	True	True	True	11	
12	suv	False	owned	all	8350	True	Минская обл.	18	61	False	True	True	True	True	False	False	True	True	80	
13	suv	False	owned	all	4300	False	Минская обл.	13	2	False	True	False	2							
14	hatchback	False	owned	all	1666	False	Гомельская обл.	8	94	True	False	230								
15	universal	False	owned	all	8600	True	Минская обл.	24	34	False	True	False	True	False	False	False	True	True	63	
16	universal	False	owned	all	7300	True	Минская обл.	17	22	False	True	True	False	True	False	False	False	True	35	
17	universal	False	owned	all	7587.97	True	Минская обл.	7	166	False	True	False	True	True	False	False	True	True	468	
18	sedan	False	owned	all	10950	False	Минская обл.	12	7	False	True	False	False	False	False	False	False	True	21	
19	suv	False	owned	all	12700	False	Минская обл.	14	14	False	False	True	16							
20	suv	False	owned	all	16500	False	Минская обл.	6	34	False	True	True	False	True	False	True	True	True	59	
21	universal	False	owned	all	8700	False	Гомельская обл.	10	11	False	True	True	False	True	True	True	True	True	13	
22	suv	False	owned	all	7500	False	Минская обл.	18	18	False	True	False	True	True	True	False	True	True	188	
23	suv	False	owned	all	8650	False	Минская обл.	30	25	False	True	False	True	True	False	False	True	True	37	
24	universal	False	owned	all	3500	True	Минская обл.	5	42	False	True	False	True	True	False	False	True	True	86	
25	universal	False	owned	all	3800	True	Минская обл.	12	44	False	True	False	True	True	False	False	True	True	67	
26	suv	False	owned	all	7200	True	Гомельская обл.	7	62	False	True	False	False	True	False	False	True	True	204	
27	universal	False	owned	front	1850	True	Минская обл.	18	23	False	True	False	False	True	False	False	False	True	71	
28	universal	False	owned	all	3800	False	Брестская обл.	12	10	False	True	False	True	True	False	False	False	True	21	
29	universal	False	owned	all	3000	False	Минская обл.	6	103	False	True	581								
30	sedan	False	owned	all	6200	False	Витебская обл.	10	21	False	True	True	False	False	False	False	False	True	48	
31	universal	False	owned	all	3700	False	Минская обл.	6	20	False	True	False	True	True	False	False	False	True	60	
32	universal	False	owned	all	650	False	Гомельская обл.	9	47	False	False	False	False	True	False	False	True	False	85	
33	universal	False	owned	all	2000	True	Могилевская обл.	4	5	False	True	91								
34	sedan	False	owned	all	800	False	Брестская обл.	6	16	False	True	True	20							
35	sedan	False	owned	all	14950	True	Минская обл.	26	132	True	False	174								
36	hatchback	False	owned	all	5350	False	Минская обл.	13	11	False	True	True	False	True	True	True	True	True	62	
37	universal	False	owned	all	17059.12	False	Гродненская обл.	24	11	False	True	False	True	True	True	True	True	True	15	
38	universal	False	owned	all	4700	False	Минская обл.	6	21	False	True	61								

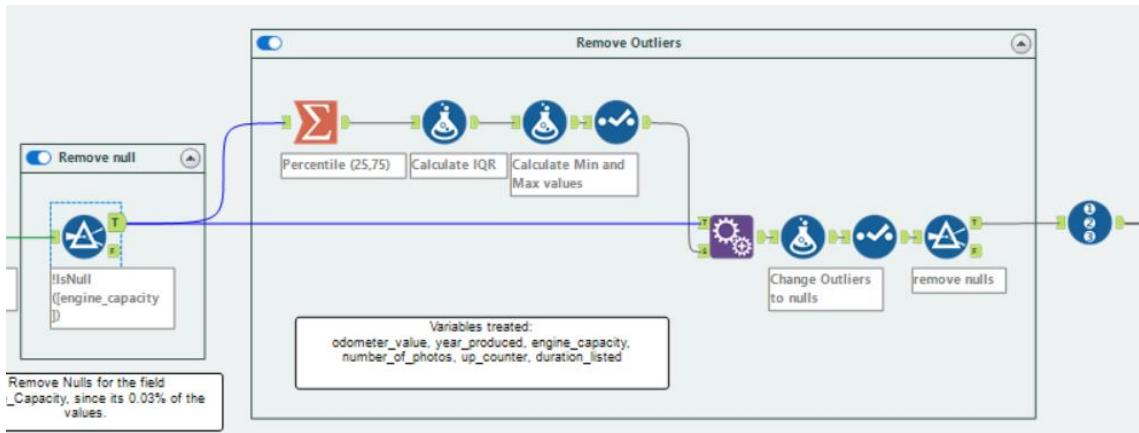
Clean Up:

Nulls:



First, we remove the nulls from Engine_Capacity.

Outliers:



Use Summarize tool to calculate the 25 and 75 percentiles. The Formula Tool to calculate the IQR, and then again the Formula tool to calculate the Upper and Lower limits.

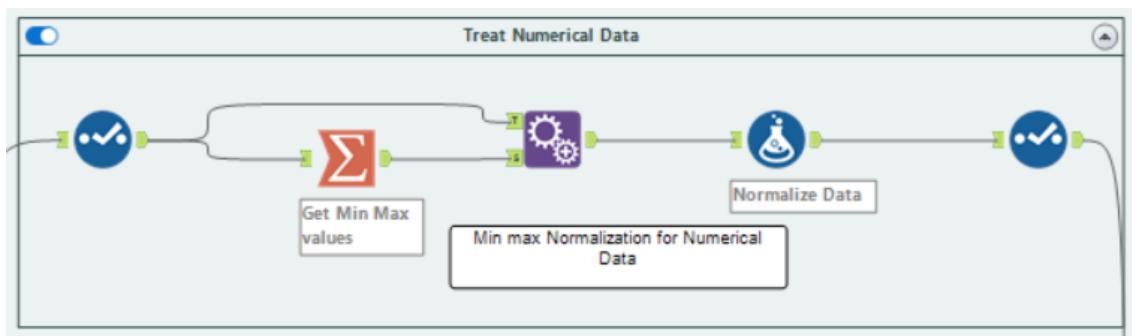
Field	Action	Output Field Name
odometer_value	Percentile	odometer_value_P05
odometer_value	Percentile	odometer_value_P75
year_produced	Percentile	year_produced_P25
year_produced	Percentile	year_produced_P75
engine_capacity	Percentile	engine_capacity_P25
engine_capacity	Percentile	engine_capacity_P75
number_of_photo...	Percentile	number_of_photos_P25
number_of_photo...	Percentile	number_of_photos_P75
up_counter	Percentile	up_counter_P25
up_counter	Percentile	up_counter_P75
duration_listed	Percentile	duration_listed_P25
duration_listed	Percentile	duration_listed_P75

Field	Action	Output Field Name
odometer_value	Formula	[odometer_value_P75]-[odometer_value_P25]
year_produced	Formula	[year_produced_P75]-[year_produced_P25]
engine_capacity	Formula	[engine_capacity_P75]-[engine_capacity_P25]
number_of_photos	Formula	[number_of_photos_P75]-[number_of_photos_P25]
up_counter	Formula	[up_counter_P75]-[up_counter_P25]
duration_listed	Formula	[duration_listed_P75]-[duration_listed_P25]

After that we use the Append Fields tool to append the Min and Max limits to the original dataframe. After that we use again the Formula tool to remove the values under the Min and above the Max thresholds.

The we use the Select tool, to deselect the Mins and Maxs fields, and a Filter tool to remove the nulls values

Treat Numerical Data:



Select tool is used to select only the Numerical Data. Then the Summarize tool is used to get the min and max values of each field. The Append Fields is used to append the min and max values to the dataframe

The screenshot shows three DataWrangler configuration windows:

- Select (34) - Configuration:** Shows a list of fields with checkboxes. The "RecordID" field is checked and set to type Int32, size 4, and renamed. Other fields like "manufacturer_name" and "model_name" are checked but left as V_WString.
- Summarize (37) - Configuration:** Shows a list of fields with checkboxes. The "RecordID" field is checked and set to type Int32. It also lists "odometer_value" and "year_produced" as Double fields.
- Append Fields (39) - Configuration:** Shows a list of input fields with checkboxes. The "RecordID" field is checked and set to type Int32, size 4. It lists various target fields such as "odometer_value", "year_produced", "engine_capacity", etc., along with their corresponding source fields (e.g., "Min_odometer_value", "Max_odometer_value").

Lastly the formula tool is used to apply the Min-Max algorithm. Then the Select tool is used to deselect the min, max fields.

The screenshot shows two DataWrangler configuration windows:

- Formula (40) - Configuration:** Shows two formulas being applied to numerical fields:
 - For "odometer_value": $\frac{([odometer_value]-[Min_odometer_value])}{([Max_odometer_value]-[Min_odometer_value])}$
 - For "year_produced": $\frac{([year_produced]-[Min_year_produced])}{([Max_year_produced]-[Min_year_produced])}$
- Select (41) - Configuration:** Shows a list of fields with checkboxes. Most fields are checked and set to their original types and sizes. The "RecordID" field is checked and set to type Int32, size 4. Some fields like "Min_odometer_value" and "Max_odometer_value" are unchecked.

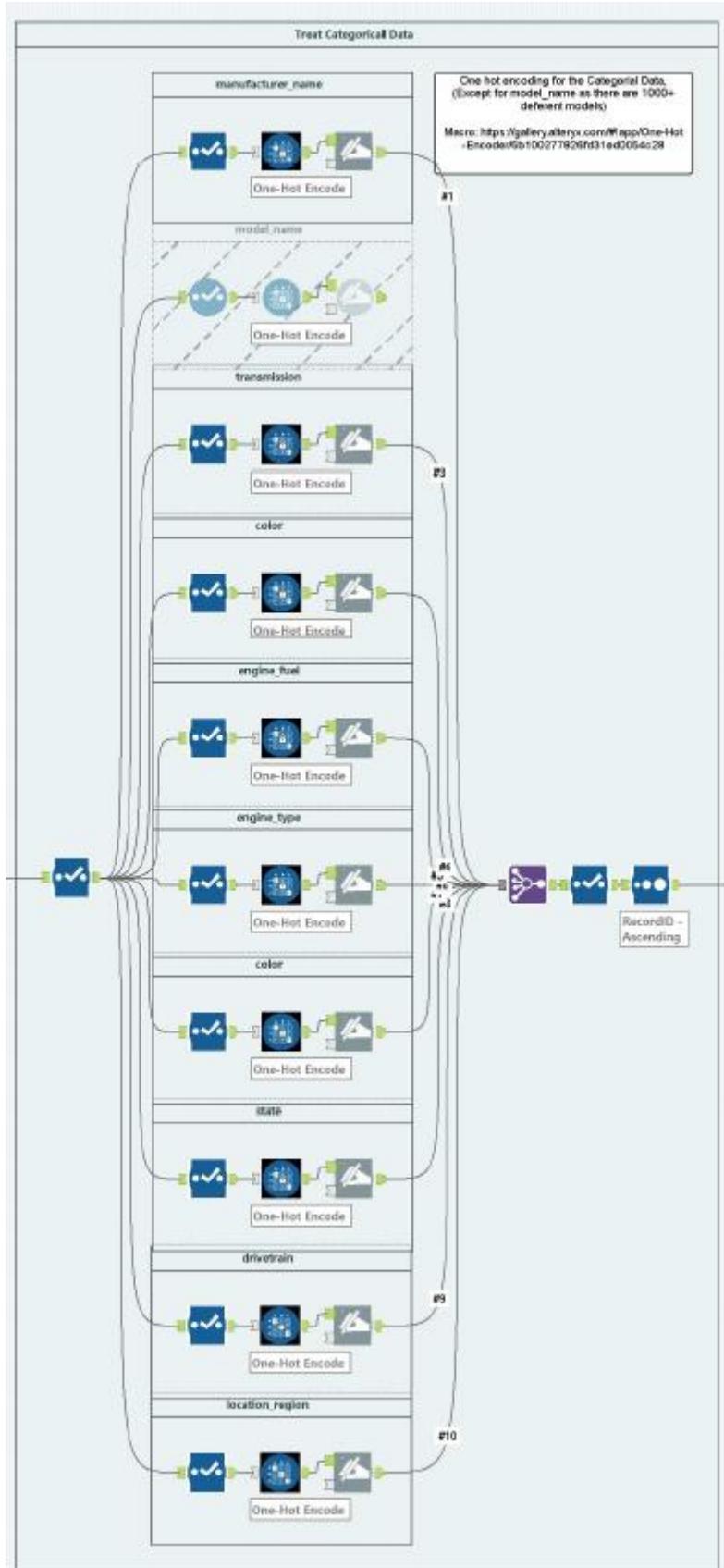
We have the next results:

Results - Select (41) - Output

8 of 8 Fields | Cell Viewer | * 15,382 of 34,905 records displayed(partial results) | ↑ ↓ |

Record	RecordID	odometer_value	year_produced	engine_capacity	price_usd	number_of_photos	up_counter	duration_listed
1	1	0.288316	0.790698	0.605263	10900	0.285714	0.169014	0.061303
2	2	0.440061	0.604651	0.736842	5000	0.392857	0.746479	0.318008
3	3	0.610015	0.581395	0.605263	2800	0.107143	1	0.578544
4	4	0.015175	0.534884	0.736842	9999	0.285714	0.577465	0.329502
5	5	0.424886	0.581395	0.605263	2134.11	0.464286	0.084507	0.02682
6	6	0.200985	0.813953	0.605263	14700	0.678571	0.774648	0.256705
7	7	0.531108	0.651163	0.605263	4500	0.214286	0.394366	0.279693
8	8	0.271624	0.790698	0.605263	12900	0.571429	0.450704	0.333333
9	9	0.866945	0.534884	0.605263	4200	0.25	0.140845	0.164751
10	10	0.424886	0.627907	0.473684	6900	0.464286	0.070423	0.042146
11	11	0.388467	0.744186	0.894737	8350	0.607143	0.84507	0.306513
12	12	0.487102	0.604651	0.473684	4300	0.428571	0.014085	0.007663
13	13	0.234727	0.813953	0.605263	8600	0.821429	0.464789	0.241379
14	14	0.247677	0.651163	0.473684	7300	0.571429	0.295775	0.1341
15	15	0.289833	0.674419	0.473684	10950	0.392857	0.084507	0.08046
16	16	0.271624	0.883721	0.473684	12700	0.464286	0.183099	0.061303
17	17	0.241275	0.860465	0.473684	16500	0.178571	0.464789	0.226054
18	18	0.390736	0.744186	0.605263	8700	0.321429	0.140845	0.049808
19	19	0.366316	0.674419	0.736842	7500	0.607143	0.239437	0.720307
20	20	0.364188	0.581395	0.605263	3500	0.142857	0.577465	0.329502
21	21	0.378525	0.604651	0.605263	3800	0.392857	0.605634	0.256705
22	22	0.379363	0.72093	0.736842	7200	0.214286	0.859155	0.781609
23	23	0.632777	0.488372	0.473684	1850	0.607143	0.309859	0.272031
24	24	0.572079	0.534884	0.605263	3800	0.392857	0.126761	0.08046
25	25	0.409712	0.651163	0.473684	6200	0.321429	0.28169	0.183908
26	26	0.674422	0.534884	0.473684	3700	0.178571	0.267606	0.229985
27	27	0.453718	0.418605	0.526316	650	0.285714	0.647887	0.32567
28	28	0.515933	0.395349	0.368421	2000	0.107143	0.056338	0.348659
29	29	0.187338	0.348837	0.526316	800	0.178571	0.211268	0.076628
30	30	0.373464	0.744186	0.342105	5350	0.428571	0.140845	0.237548
31	31	0.195369	0.930233	0.894737	17059.12	0.821429	0.140845	0.057471
32	32	0.151745	0.418605	0.473684	1700	0.178571	0.577465	0.348659
33	33	0.520167	0.534884	0.605263	3500	0.464286	0.549296	0.268199
34	34	0.591806	0.488372	0.473684	3000	0.25	0.380282	0.176245
35	35	0.326252	0.744186	0.605263	8100	0.321429	0	0.003831
36	36	0.437785	0.697674	0.605263	8500	0.857143	0.014085	0.007663
37	37	0.470401	0.581395	0.605263	3300	0.107143	0	0.003831
38	38	0.247712	0.677007	0.473684	4400	0.428571	0.889156	0.020214

Treat Categorical Data:



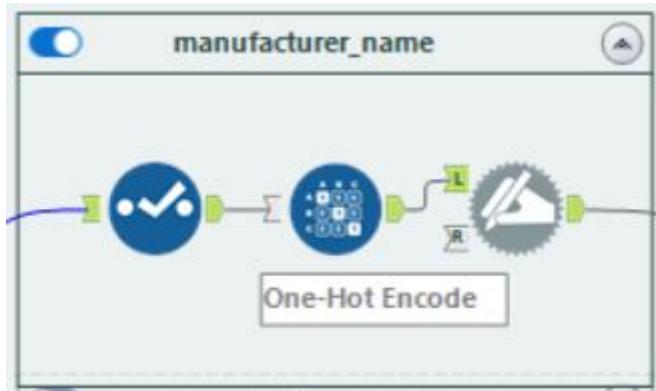
First, we use the Select tool to select only the Categorical variables:

Select (35) - Configuration

TIP: To reorder multiple rows: select, right-click and drag.

	Field	Type	Size	Rename	Description
<input checked="" type="checkbox"/>	RecordID	Int32	4		
<input checked="" type="checkbox"/>	manufacturer_name	V_WString	254		
<input checked="" type="checkbox"/>	model_name	V_WString	254		
<input checked="" type="checkbox"/>	transmission	V_WString	254		
<input checked="" type="checkbox"/>	color	V_WString	254		
<input type="checkbox"/>	odometer_value	Double	8		
<input type="checkbox"/>	year_produced	Double	8		
<input type="checkbox"/>	engine_fuel	V_WString	254		
<input type="checkbox"/>	engine_has_gas	Bool	1		
<input checked="" type="checkbox"/>	engine_type	V_WString	254		
<input type="checkbox"/>	engine_capacity	Double	8		
<input checked="" type="checkbox"/>	body_type	V_WString	254		
<input type="checkbox"/>	has_warranty	Bool	1		
<input checked="" type="checkbox"/>	state	V_WString	254		
<input checked="" type="checkbox"/>	drivetrain	V_WString	254		
<input type="checkbox"/>	price_usd	Double	8		
<input type="checkbox"/>	is_exchangeable	Bool	1		
<input checked="" type="checkbox"/>	location_region	V_WString	254		
<input type="checkbox"/>	number_of_photos	Double	8		
<input type="checkbox"/>	up_counter	Double	8		
<input type="checkbox"/>	feature_0	Bool	1		
<input type="checkbox"/>	feature_1	Bool	1		
<input type="checkbox"/>	feature_2	Bool	1		
<input type="checkbox"/>	feature_3	Bool	1		

After that each cell do the One-Hot Encoding to each variable, so I'm only going to explain the first one. The variable model_name, is ignore because it has more than a 1000 different values.



First the select tool is used to select only the RecordID and the manufacturer_name fields. Then we apply the One-Hot Encoding Macro (<https://gallery.alteryx.com/#!app/One-Hot-Encoder/5b100277826fd31ed0054c28>) And lastly we use the Dynamic Rename tool to add the original name of the field as Prefix

The screenshot shows three Dataiku configuration panels side-by-side:

- Select (49) - Configuration:** A table showing columns like RecordID, manufacturer_name, model_name, transmission, color, engine_fuel, engine_type, body_type, state, drivetrain, and location_region. The RecordID column is selected.
- One-Hot Encoder (48) - Configuration:** A panel titled "Questions" with "Choose Field: RecordID (String)" dropdown set to "RecordID (Int32)".
- Dynamic Rename (73) - Configuration:** A list of car manufacturers with checkboxes next to them. The list includes Acura, Alfa_Romeo, Audi, BMW, Buick, Cadillac, Chevrolet, Chrysler, Citroen, Dacia, Daewoo, Dodge, Fiat, Ford, Geely, Great_Wall, Honda, Hyundai, Infiniti, Iveco, Jaguar, Jeep, Kia, LADA, Lancia, Land_Rover, Lexus, Lifan, Lincoln, Mazda, and Mercedes_Benz. The "Rename Mode: Add Prefix/Suffix" dropdown is set to "Prefix".

We get the next result:

	Be...	manufacturer_name_Acura	manufacturer_name_Alfa_Romeo	manufacturer_name_Audi	manufacturer_name_BMW	manufacturer_name_Buick	manufacturer_name_Cadillac	manufacturer_name_Citroen	manufacturer_name_Chevrolet	manufacturer_name_C...
1	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0
100	0	0	0	0	0	0	0	0	0	0
1000	0	0	0	0	0	0	0	0	0	0
10...	0	0	0	0	0	0	0	0	0	0
10...	0	0	0	0	0	0	0	0	0	0
10...	0	0	0	0	0	0	0	0	0	0
10...	0	0	0	0	0	0	0	0	0	0
10...	0	0	0	0	0	0	0	0	0	0
10...	0	0	0	0	0	0	0	0	0	0
10...	0	0	0	0	0	0	0	0	0	0
10...	0	0	0	0	0	0	0	0	0	0
10...	0	0	0	0	0	0	0	0	0	0
10...	0	0	0	0	0	0	0	0	0	0
10...	0	0	0	0	0	0	0	0	0	0
10...	0	0	0	0	0	0	0	0	0	0
10...	0	0	0	0	0	0	0	0	0	0
10...	0	0	0	0	0	0	0	0	0	0
10...	0	0	0	0	0	0	0	0	0	0
10...	0	0	0	0	0	0	0	0	0	0
10...	0	0	0	0	0	0	0	0	0	0
10...	0	0	0	0	0	0	0	0	0	0
10...	0	0	0	0	0	0	0	0	0	0
10...	0	0	0	0	0	0	0	0	0	0
10...	0	0	0	0	0	0	0	0	0	0
10...	0	0	0	0	0	0	0	0	0	0
10...	0	0	0	0	0	0	0	0	0	0
1001	0	0	0	0	0	0	0	0	0	0
10...	0	0	0	0	0	0	0	0	0	0
10...	0	0	0	0	0	0	0	0	0	0
1002	0	0	0	0	0	0	0	0	0	0
10...	0	0	0	0	0	0	0	0	0	0
10...	0	0	0	0	0	0	0	0	0	0
10...	0	0	0	0	0	0	0	0	0	0
10...	0	0	0	0	0	0	0	0	0	0
10...	0	0	0	0	0	0	0	0	0	0
10...	0	0	0	0	0	0	0	0	0	0
10...	0	0	0	0	0	0	0	0	0	0
10...	0	0	0	0	0	0	0	0	0	0
10...	0	0	0	0	0	0	0	0	0	0
10...	0	0	0	0	0	0	0	0	0	0
10...	0	0	0	0	0	0	0	0	0	0
10...	0	0	0	0	0	0	0	0	0	0
10...	0	0	0	0	0	0	0	0	0	0
10...	0	0	0	0	0	0	0	0	0	0
10...	0	0	0	0	0	0	0	0	0	0
1003	0	0	0	0	0	0	0	0	0	0

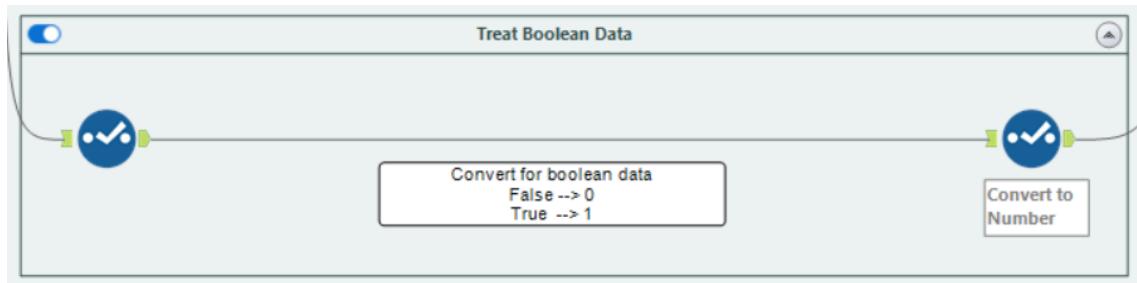
After we complete this process for all the Categorical variables, we Join them using a Join Multiple tool using the RecordID as keys. We then convert the RecordID to an Int, and use a Sort tool to order the Results

The first screenshot shows the 'Join Multiple' configuration with 'Join by Record Position' selected. It lists inputs #1 through #10 and defines a Cartesian join. The second screenshot shows the 'Select' configuration where a specific field, 'manufacturer_name_Acura', is highlighted. The third screenshot shows the 'Sort' configuration with 'Use Dictionary Order' checked.

We get the following data:

This screenshot displays the final output of the Talend job. It shows a table with 101 columns and 101 rows of data. The columns represent various fields from the joined datasets, including manufacturer names, location regions, and technical specifications like transmission type and engine fuel type. The data is presented in a grid format with many zeros and some binary values.

Treat Boolean Data:



We first select only the Boolean fields using the Select tool. Then we change the types of the variables to Int using the Select tool.

Select (36) - Configuration

TIP: To reorder multiple rows: select, right-click and drag.

	Field	Type	Size	Rename	Description
<input checked="" type="checkbox"/>	RecordID	Int32	4		
<input type="checkbox"/>	manufacturer_name	V_WString	254		
<input type="checkbox"/>	model_name	V_WString	254		
<input type="checkbox"/>	transmission	V_WString	254		
<input type="checkbox"/>	color	V_WString	254		
<input type="checkbox"/>	odometer_value	Double	8		
<input type="checkbox"/>	year_produced	Double	8		
<input type="checkbox"/>	engine_fuel	V_WString	254		
<input checked="" type="checkbox"/>	engine_has_gas	Bool	1		
<input type="checkbox"/>	engine_type	V_WString	254		
<input type="checkbox"/>	engine_capacity	Double	8		
<input type="checkbox"/>	body_type	V_WString	254		
<input checked="" type="checkbox"/>	has_warranty	Bool	1		
<input type="checkbox"/>	state	V_WString	254		
<input type="checkbox"/>	drivetrain	V_WString	254		
<input type="checkbox"/>	price_usd	Double	8		
<input checked="" type="checkbox"/>	is_exchangeable	Bool	1		
<input type="checkbox"/>	location_region	V_WString	254		
<input type="checkbox"/>	number_of_photos	Double	8		
<input type="checkbox"/>	up_counter	Double	8		
<input checked="" type="checkbox"/>	feature_0	Bool	1		
<input checked="" type="checkbox"/>	feature_1	Bool	1		
<input checked="" type="checkbox"/>	feature_2	Bool	1		
<input type="checkbox"/>	feature_3	Bool	1		

Options: Use commas as decimal separators (String/Numeric conversions only)

Select (122) - Configuration

TIP: To reorder multiple rows: select, right-click and drag.

	Field	Type	Size	Rename	Description
<input checked="" type="checkbox"/>	RecordID	Int32	4		
<input checked="" type="checkbox"/>	engine_has_gas	Int32	4		
<input checked="" type="checkbox"/>	has_warranty	Int32	4		
<input checked="" type="checkbox"/>	is_exchangeable	Int32	4		
<input checked="" type="checkbox"/>	feature_0	Int32	4		
<input checked="" type="checkbox"/>	feature_1	Int32	4		
<input checked="" type="checkbox"/>	feature_2	Int32	4		
<input checked="" type="checkbox"/>	feature_3	Int32	4		
<input checked="" type="checkbox"/>	feature_4	Int32	4		
<input checked="" type="checkbox"/>	feature_5	Int32	4		
<input checked="" type="checkbox"/>	feature_6	Int32	4		
<input checked="" type="checkbox"/>	feature_7	Int32	4		
<input checked="" type="checkbox"/>	feature_8	Int32	4		
<input checked="" type="checkbox"/>	feature_9	Int32	4		
<input checked="" type="checkbox"/>	*Unknown	Unknown	0		Dynamic or

Options: Use commas as decimal separators (String/Numeric conversions only)

Getting the next results:

Results - Select (122) - Output

14 of 14 Fields | Cell Viewer | 14,876 of 34,905 records displayed(partial results)

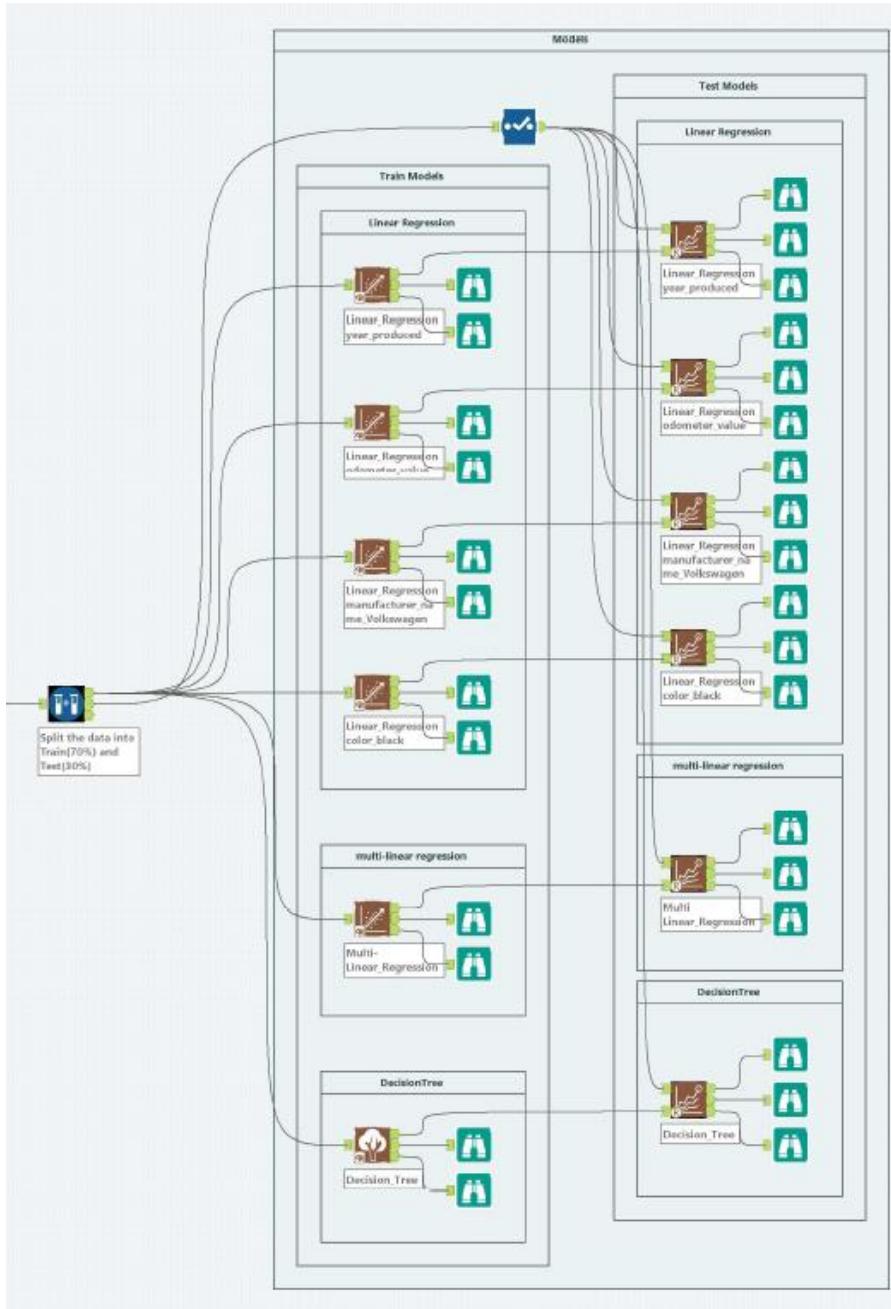
Record	RecordID	engine_has_gas	has_warranty	is_exchangeable	feature_0	feature_1	feature_2	feature_3	feature_4	feature_5	feature_6	feature_7	feature_8	feature_9
1	1	0	0	0	0	1	1	1	0	1	0	1	1	1
2	2	0	0	1	0	1	0	0	1	1	0	0	0	1
3	3	0	0	1	0	1	0	0	0	0	0	0	1	1
4	4	0	0	1	1	0	0	0	0	0	0	0	0	0
5	5	0	0	1	0	1	0	1	1	0	0	0	0	1
6	6	0	0	1	0	1	0	0	0	1	0	1	1	1
7	7	0	0	0	0	1	1	0	0	0	0	0	0	1
8	8	0	0	0	0	1	1	1	1	1	1	1	1	1
9	9	0	0	1	0	1	1	0	0	1	0	0	0	1
10	10	0	0	1	0	1	0	0	1	0	1	0	1	1
11	11	0	0	1	0	1	1	1	1	1	0	0	1	1
12	12	0	0	0	0	1	0	0	0	0	0	0	0	0
13	13	0	0	1	0	1	0	1	0	0	0	0	1	1
14	14	0	0	1	0	1	1	0	1	1	0	0	0	1
15	15	0	0	0	0	1	0	0	0	0	0	0	0	1
16	16	0	0	0	0	0	1	1	0	1	1	1	1	1
17	17	0	0	0	0	1	1	0	1	1	0	1	1	1
18	18	0	0	0	0	1	1	0	1	1	1	0	1	1
19	19	0	0	0	0	1	0	1	1	1	1	0	1	1
20	20	0	0	1	0	1	0	0	1	0	0	0	1	1
21	21	0	0	1	0	1	0	0	0	0	0	0	1	0
22	22	0	0	1	0	1	0	0	1	1	0	0	1	1
23	23	0	0	1	0	1	0	0	0	1	0	0	0	1
24	24	0	0	0	0	1	0	0	1	1	0	0	0	1
25	25	0	0	0	0	0	1	1	0	0	0	0	0	1
26	26	0	0	0	0	1	0	0	1	1	0	0	0	1
27	27	1	0	0	0	0	0	0	0	1	0	0	0	1
28	28	0	0	1	0	0	0	0	0	0	0	0	0	1
29	29	0	0	0	0	0	0	0	0	0	0	0	1	1
30	30	0	0	0	0	1	1	1	0	1	1	1	0	1
31	31	0	0	0	0	1	0	1	1	1	1	1	1	1
32	32	0	0	1	0	0	0	1	0	0	0	0	0	1
33	33	0	0	1	1	0	0	0	0	0	0	0	0	0
34	34	0	0	1	0	1	0	0	0	0	0	0	0	1
35	35	0	0	0	0	0	0	0	0	0	0	0	1	1
36	36	1	0	0	0	0	0	1	1	0	1	0	0	1
37	37	0	0	1	0	1	0	0	0	1	0	0	0	1
38	39	n	n	n	n	n	n	n	n	n	n	n	n	n

Lastly we Join the outputs from the previous 3 treatment (Numerical, Categorical and Boolean) using a Join Multiple Tool

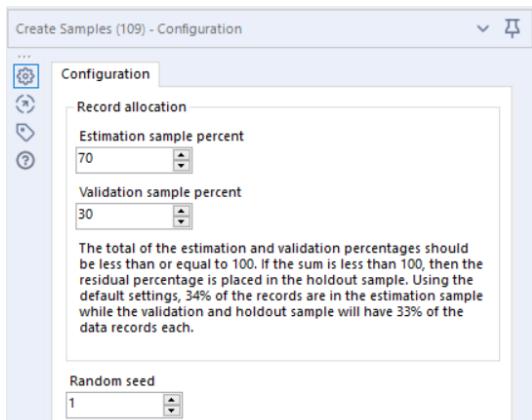
The screenshot shows the 'Join Multiple (123) - Configuration' dialog. At the top, there are two radio button options: 'Join by Record Position' (unchecked) and 'Join by Specific Fields' (checked). Below this is a grid where 'Input_#1' (RecordID), 'Input_#2' (RecordID), and 'Input_#3' (RecordID) are joined together. A note below the grid states: 'Error on multidimensional joins of more than 16 Records'. There is also a checkbox 'Only Output Records that Join from All Inputs' which is unchecked. At the bottom, there is a table listing fields from 'Input_#2' with their types: Int32, Byte, Byte.

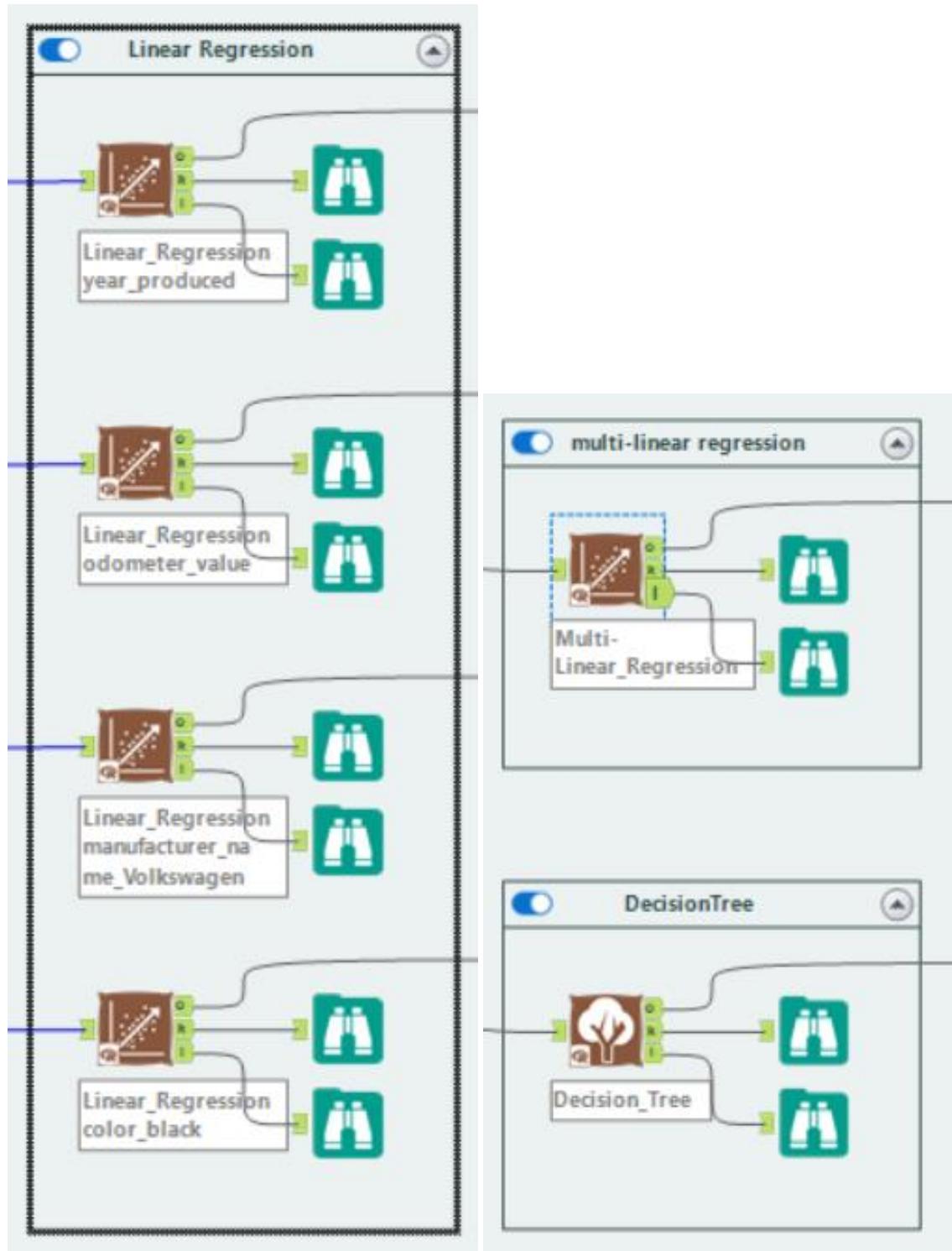
Input	Field	Type
Input_#2	RecordID	Int32
Input_#2	manufacturer_name_Acura	Byte
Input_#2	manufacturer_name_Alfa_Romeo	Byte
Input_#2	manufacturer_name_Audi	Byte
Input_#2	manufacturer_name_BMW	Byte
Input_#2	manufacturer_name_Buick	Byte
Input_#2	manufacturer_name_Cadillac	Byte
Input_#2	manufacturer_name_Chery	Byte
Input_#2	manufacturer_name_Chevrolet	Byte
Input_#2	manufacturer_name_Chrysler	Byte
Input_#2	manufacturer_name_Citroen	Byte

Models:



First, we Split the data into Train (70%) and Test (30%) and train the models:





Models Train:

Linear regression:

The image displays four separate windows of a 'Linear Regression' configuration tool, each showing a 'Setup' screen with various parameters for training a linear regression model.

- Top Left (Model name: Linear_Regression_year_produced):** Target variable is 'price_usd'. Predictor variables selected include 'odometer_value', 'year_produced' (checked), 'engine_capacity', 'number_of_photos', 'up_counter', 'duration_listed', 'manufacturer_name_Acura', 'manufacturer_name_Alfa_Romeo', 'manufacturer_name_Audi', and 'manufacturer_name_BMW'.
- Top Right (Model name: Linear_Regression_odometer_value):** Target variable is 'price_usd'. Predictor variables selected include 'odometer_value' (checked), 'year_produced', 'engine_capacity', 'number_of_photos', 'up_counter', 'duration_listed', 'manufacturer_name_Acura', 'manufacturer_name_Alfa_Romeo', 'manufacturer_name_Audi', and 'manufacturer_name_BMW'.
- Bottom Left (Model name: Linear_Regression_manufacturer_name_Volkswagen):** Target variable is 'price_usd'. Predictor variables selected include 'manufacturer_name_SsangYong', 'manufacturer_name_Subaru', 'manufacturer_name_Suzuki', 'manufacturer_name_Toyota', 'manufacturer_name_Volkswagen' (checked), 'manufacturer_name_Volvo', 'manufacturer_name_BA3', 'manufacturer_name_ГАЗ', 'manufacturer_name_ЗАЗ', and 'manufacturer_name_Москвич'.
- Bottom Right (Model name: Linear_Regression_color_black):** Target variable is 'price_usd'. Predictor variables selected include 'location_region_Минская_обл.' (checked), 'location_region_Могилевская_обл.', 'transmission_automatic', 'transmission_mechanical', 'color_black', 'color_blue', 'color_brown', 'color_green', 'color_grey', and 'color_orange'.

Multi-linear regression & Decision tree:

The image shows two side-by-side configuration panels. The left panel is for 'Linear Regression (148) - Configuration' and the right panel is for 'Decision Tree (144) - Configuration'. Both panels have a 'Setup' tab selected. In the 'Model name' section, 'Multi-Linear_Regression' is entered for linear regression and 'Decision_Tree' is entered for the decision tree. Under 'Select the target variable', both have 'price_usd' chosen. In the 'Select the predictor variables' section, both lists include 'odometer_value', 'year_produced', 'engine_capacity', 'number_of_photos', 'up_counter', 'duration_listed', and manufacturer names for Acura, Alfa Romeo, Audi, and BMW. Each list has a 'Selected' count of 119 and a 'Fields' count of 119. Buttons for 'Customize' and a back arrow are at the bottom of each list.

Results:

Linear regression:

Record Report																
1 Report for Linear Model Linear_Regression_year_produced																
2	Basic Summary															
3	Call: lm(formula = price_usd ~ year_produced, data = the.data)															
4	Residuals:															
5	<table border="1"> <thead> <tr> <th>Min</th><th>1Q</th><th>Median</th><th>3Q</th><th>Max</th></tr> </thead> <tbody> <tr> <td>-11501</td><td>-2252</td><td>-808</td><td>1263</td><td>37901</td></tr> </tbody> </table>	Min	1Q	Median	3Q	Max	-11501	-2252	-808	1263	37901					
Min	1Q	Median	3Q	Max												
-11501	-2252	-808	1263	37901												
6	Coefficients:															
7	<table border="1"> <thead> <tr> <th></th><th>Estimate</th><th>Std. Error</th><th>t value</th><th>Pr(> t)</th></tr> </thead> <tbody> <tr> <td>(Intercept)</td><td>-9080</td><td>93.76</td><td>-96.85</td><td>< 2.2e-16 ***</td></tr> <tr> <td>year_produced</td><td>24613</td><td>143.47</td><td>171.55</td><td>< 2.2e-16 ***</td></tr> </tbody> </table> <p>Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1</p>		Estimate	Std. Error	t value	Pr(> t)	(Intercept)	-9080	93.76	-96.85	< 2.2e-16 ***	year_produced	24613	143.47	171.55	< 2.2e-16 ***
	Estimate	Std. Error	t value	Pr(> t)												
(Intercept)	-9080	93.76	-96.85	< 2.2e-16 ***												
year_produced	24613	143.47	171.55	< 2.2e-16 ***												
8	<p>Residual standard error: 4082.1 on 24432 degrees of freedom Multiple R-squared: 0.5464, Adjusted R-Squared: 0.5464 F-statistic: 29431 on 1 and 24432 degrees of freedom (DF), p-value < 2.2e-16</p>															
Record Report																
1 Report for Linear Model Linear_Regression_odometer_value																
2	Basic Summary															
3	Call: lm(formula = price_usd ~ odometer_value, data = the.data)															
4	Residuals:															
5	<table border="1"> <thead> <tr> <th>Min</th><th>1Q</th><th>Median</th><th>3Q</th><th>Max</th></tr> </thead> <tbody> <tr> <td>-11959</td><td>-3194</td><td>-908</td><td>2095</td><td>40801</td></tr> </tbody> </table>	Min	1Q	Median	3Q	Max	-11959	-3194	-908	2095	40801					
Min	1Q	Median	3Q	Max												
-11959	-3194	-908	2095	40801												
6	Coefficients:															
7	<table border="1"> <thead> <tr> <th></th><th>Estimate</th><th>Std. Error</th><th>t value</th><th>Pr(> t)</th></tr> </thead> <tbody> <tr> <td>(Intercept)</td><td>11960</td><td>78.75</td><td>151.86</td><td>< 2.2e-16 ***</td></tr> <tr> <td>odometer_value</td><td>-15082</td><td>190.80</td><td>-79.05</td><td>< 2.2e-16 ***</td></tr> </tbody> </table> <p>Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1</p>		Estimate	Std. Error	t value	Pr(> t)	(Intercept)	11960	78.75	151.86	< 2.2e-16 ***	odometer_value	-15082	190.80	-79.05	< 2.2e-16 ***
	Estimate	Std. Error	t value	Pr(> t)												
(Intercept)	11960	78.75	151.86	< 2.2e-16 ***												
odometer_value	-15082	190.80	-79.05	< 2.2e-16 ***												
8	<p>Residual standard error: 5408.8 on 24432 degrees of freedom Multiple R-squared: 0.2037, Adjusted R-Squared: 0.2036 F-statistic: 6248 on 1 and 24432 degrees of freedom (DF), p-value < 2.2e-16</p>															

Record Report

1 Report for Linear Model

Linear_Regression_manufacturer_name_Volkswagen

2 Basic Summary

3 Call:

```
lm(formula = price_usd ~ manufacturer_name_Volkswagen, data = the.data)
```

4 Residuals:

Min	1Q	Median	3Q	Max
-6377	-4278	-1778	2209	43622

5 Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	6377.59	41.17	154.899	< 2.2e-16 ***
manufacturer_name_Volkswagen	-86.56	122.44	-0.707	0.47957

Significance codes: 0 ***. 0.001 **. 0.01 *. 0.05 . 0.1 ' '

6 Residual standard error: 6061 on 24432 degrees of freedom
Multiple R-squared: 2.046e-05, Adjusted R-Squared: -2.047e-05
F-statistic: 0.4998 on 1 and 24432 degrees of freedom (DF), p-value 0.4796

Record Report

1 Report for Linear Model

Linear_Regression_color_black

2 Basic Summary

3 Call:

```
lm(formula = price_usd ~ color_black, data = the.data)
```

4 Residuals:

Min	1Q	Median	3Q	Max
-8636	-4007	-1707	2143	44193

5 Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	5807	42.34	137.15	< 2.2e-16 ***
color_black	2929	96.73	30.28	< 2.2e-16 ***

Significance codes: 0 ***. 0.001 **. 0.01 *. 0.05 . 0.1 ' '

6 Residual standard error: 5950.5 on 24432 degrees of freedom
Multiple R-squared: 0.03617, Adjusted R-Squared: 0.03613
F-statistic: 917 on 1 and 24432 degrees of freedom (DF), p-value < 2.2e-16

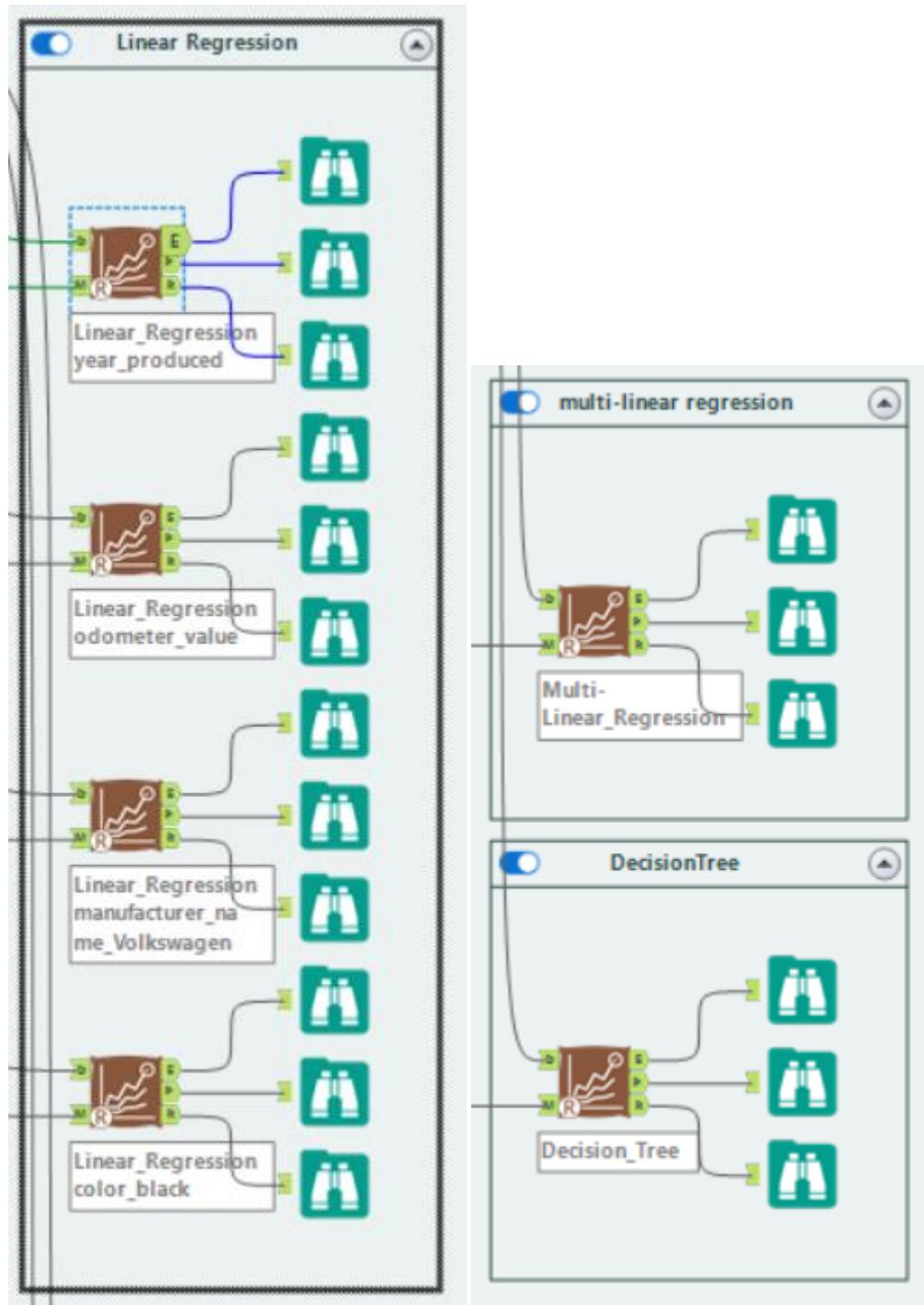
Multi-linear regression:

	R SQUARED 0.788		ADJUSTED R SQUARED 0.787
	MEAN ABSOLUTE ERROR 1788.925		MEAN ABSOLUTE PERCENT ERROR 1.086
	MEAN SQUARED ERROR 7784112.278		ROOT MEAN SQUARED ERROR 2790.002
	F-STATISTIC 837.66 on 108 and 24325 degrees of freedom		RESIDUAL STANDARD ERROR 2796.246 on 24315 degrees of freedom

Decision tree:

	R SQUARED 0.906		ADJUSTED R SQUARED 0.906
	MEAN ABSOLUTE ERROR 1217.199		MEAN ABSOLUTE PERCENT ERROR 0.733
	MEAN SQUARED ERROR 3440538.56		ROOT MEAN SQUARED ERROR 1854.869

Test Models:



Results:

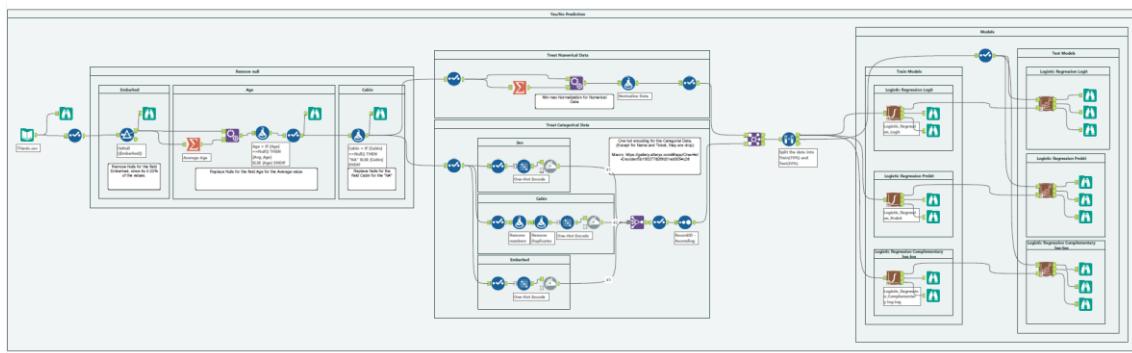
Fit and error measures					
Model	Correlation	RMSE	MAE	MPE	MAPE
Linear_Regression_year_produced	0.7348	4,242.4608	2,729.2587	-21.5418	153.4939
Linear_Regression_odometer_value	0.4656	5,535.6134	3,792.1623	-252.2018	285.4971

Model	Correlation	RMSE	MAE	MPE	MAPE
Linear_Regression_manufacturer_name_Volkswagen	0.0031	6,253.4248	4,424.6812	-294.9812	323.8264
Model	Correlation	RMSE	MAE	MPE	MAPE
Linear_Regression_color_black	0.1871	6,143.0067	4,281.5543	-301.5833	329.6564
Model	Correlation	RMSE	MAE	MPE	MAPE
Multi-Linear_Regression	0.8812	2,956.0920	1,841.0487	15.5772	73.1644
Model	Correlation	RMSE	MAE	MPE	MAPE
Decision_Tree	0.9336	2,241.3461	1,357.4195	-50.2233	68.0776

Best Model

We can see from the result that the model with the best RMSE and MAE is the **Decision Tree**, follow closely by the Multi-Linear Regression. With way worst results we find the different Linear regression Models. We can see the same results if we look at the R2 variable, we find that the Decision Tree have the highest one, follow by the Multi-Linear Regression.

Yes/No Prediction

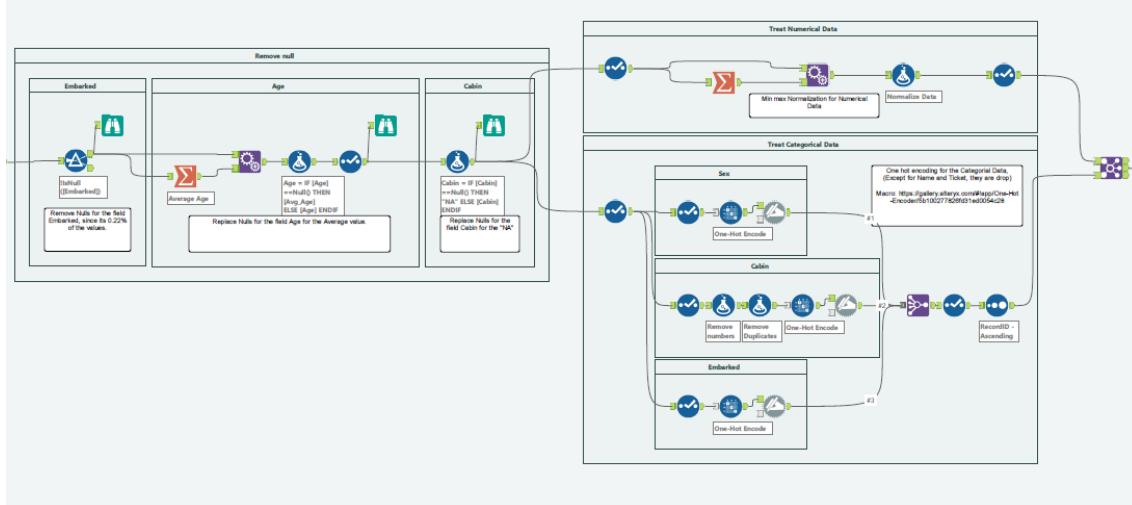


Imported Data:

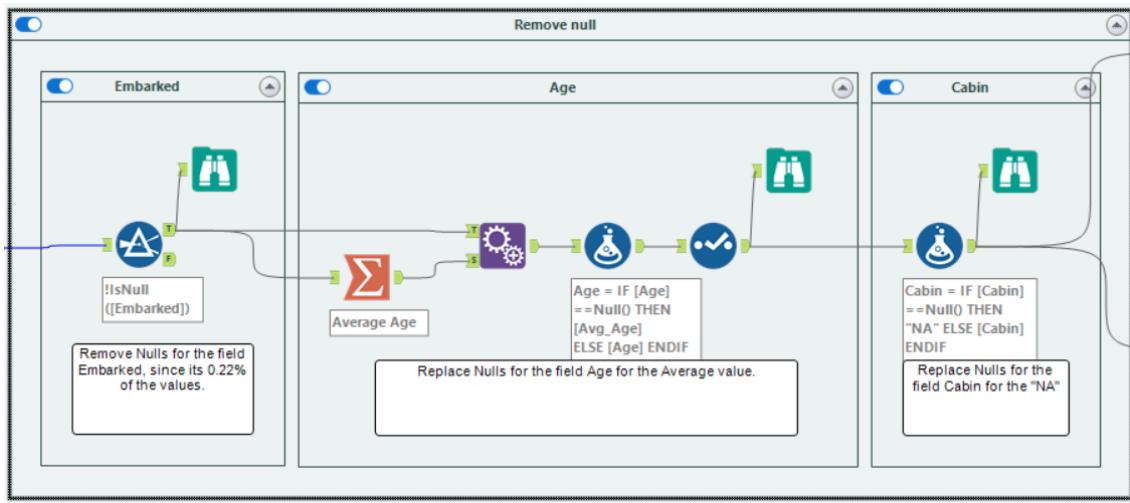
The imported values are:

Record	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1 1	0	3	Braund, Mr. Owen Harris	male	22	1	0	0	A/5 21171	7.25	[Null]	S
2 2	1	1	Cummings, Mrs. John Bradley (Florence Briggs Tha...	female	38	1	0	0	PC 17599	71.2833	C85	C
3 3	1	1	Heikkinen, Miss. Laina	female	26	0	0	0	STON/O2. 3101282	7.925	[Null]	S
4 4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	0	113803	53.1	C123	S
5 5	0	3	Allen, Mr. William Henry	male	35	0	0	0	373450	8.05	[Null]	S
6 6	0	3	Moran, Mr. James	male	[Null]	0	0	0	330877	8.4583	[Null]	Q
7 7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	0	17463	51.8625	E46	S
8 8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	1	349909	21.075	[Null]	S
9 9	1	3	Johnson, Mrs. Oscar W. (Elisabeth Vilhelmina Berg)	female	27	0	2	0	347742	11.1333	[Null]	S
10 10	1	2	Nasser, Mrs. Nicholas (Adele Achen)	female	14	1	0	0	237736	30.0708	[Null]	C
11 11	1	3	Sandstrom, Miss. Marguerite Rut	female	4	1	1	0	PP 9549	16.7	G6	S
12 12	1	1	Bonnevie, Miss. Sophie	female	58	0	0	0	133333	26.55	C103	S
13 13	0	3	Edwards, Mr. William Henry	male	20	0	0	0	A/5.21151	8.05	[Null]	S
14 14	0	3	Andersson, Mr. Anders John	male	39	1	5	0	347082	31.275	[Null]	S
15 15	0	3	Vestrom, Miss. Hilda Amanda Adolffina	female	14	0	0	0	350406	7.9542	[Null]	S
16 16	1	3	Hewlett, Mrs. Mary D (Kingcome)	female	55	0	0	0	248706	19.5	[Null]	S
17 17	0	3	Rice, Master. Eugene	male	2	4	1	0	382452	29.125	[Null]	Q
18 18	1	2	Williams, Mr. Charlie Eugene	male	[Null]	0	0	0	244373	13	[Null]	S
19 19	0	3	Vander Plank, Mrs. Julius (Emelia Maria Vandem...	female	31	1	0	0	345763	18	[Null]	S
20 20	1	3	Masseyman, Mrs. Fatima	female	[Null]	0	0	0	2649	7.225	[Null]	C
21 21	0	2	Fynney, Mr. Joseph J	male	35	0	0	0	239965	26	[Null]	S
22 22	1	2	Beebe, Mr. Lawrence	male	34	0	0	0	246968	13	D56	S
23 23	1	3	McGowan, Miss. Anna "Annie"	female	15	0	0	0	330923	8.0292	[Null]	Q
24 24	1	1	Sloper, Mr. William Thompson	male	28	0	0	0	113788	35.5	A6	S
25 25	0	3	Palsson, Miss. Torborg Danira	female	8	3	1	1	349909	21.075	[Null]	S
26 26	1	3	Asplund, Mrs. Carl Oscar (Selma Augusta Emilia J...	female	38	1	5	0	347077	31.3875	[Null]	S
27 27	0	3	Emir, Mr. Farred Chehab	male	[Null]	0	0	0	2631	7.225	[Null]	C
28 28	0	1	Fortune, Mr. Charles Alexander	male	19	3	2	0	19950	263	C23 C25 C27	S
29 29	1	3	O'Dwyer, Miss. Ellen "Nellie"	female	[Null]	0	0	0	330959	7.8792	[Null]	Q
30 30	0	3	Todoroff, Mr. Lazio	male	[Null]	0	0	0	349216	7.8958	[Null]	S
31 31	0	1	Uruchurtu, Don. Manuel E	male	40	0	0	0	PC 17601	27.7208	[Null]	C
32 32	1	1	Spencer, Mrs. William Augustus (Marie Eugenie)	female	[Null]	1	0	0	PC 17569	146.5208	B7B	C
33 33	1	3	Glynn, Mrs. Mary Agatha	female	[Null]	0	0	0	335677	7.75	[Null]	Q
34 34	0	3	Weirich, Mr. Edward H	male	68	0	0	0	345579	10.5	[Null]	S
35 35	0	1	Meyer, Mr. Edgar Joseph	male	28	1	0	0	PC 17604	8.2708	[Null]	C
36 36	0	1	Holmeson, Mr. Alexander Oskar	male	42	1	0	0	113789	53	[Null]	S
37 37	1	3	Mamee, Mr. Harry	male	[Null]	0	0	0	2677	7.2292	[Null]	C
38 38	A	3	Flournoy, Mr. Ernest Charles	male	A	A	A	A	A	A	A	A

Clean Up:



Nulls:



First, we remove the Nulls from Embarked using a Filter tool.

For Age we calculate the Average Using the Summarize tool, then using the Append Field we add it to the original Dataframe, and with the formula tool we change the nulls for the Average. Then the Select tool is used to deselect the Average field

Field	Type
PassengerId	V_String
Survived	V_String
Pclass	Double
Name	V_String
Sex	V_String
Age	Double
SibSp	Double
Parch	Double
Ticket	V_String

Action	Output Field Name
Age	Avg_Age

Input	Field	Type	Size	Rename
Target	PassengerId	V_String	254	
Target	Survived	V_String	254	
Target	Pclass	Double	8	
Target	Name	V_String	254	
Target	Sex	V_String	254	
Target	Age	Double	8	
Target	SibSp	Double	8	
Target	Parch	Double	8	
Target	Ticket	V_String	254	
Target	Fare	Double	8	
Target	Cabin	V_String	254	
Target	Embarked	V_String	254	
Source	Avg_Age	Double	8	
	*Unknown	Unknown	0	

Formula (11) - Configuration

Output Column	Data Preview
Age	22

```
fx IF [Age]==Null() THEN [Avg_Age]
X ELSE [Age] ENDIF
```

Data type: Double Size: 8

Select (17) - Configuration

Field	Type	Size	Rename	Description
PassengerId	V_String	254		
Survived	V_String	254		
Pclass	Double	8		
Name	V_String	254		
Sex	V_String	254		
Age	Double	8		
SibSp	Double	8		
Parch	Double	8		
Ticket	V_String	254		
Fare	Double	8		
Cabin	V_String	254		
Embarked	V_String	254		
Avg_Age	Double	8		
*Unknown	Unknown	0		Dynamic or...

Use commas as decimal separators (String/Numeric conversions only)

For Cabin, we replace the nulls with the constant NA

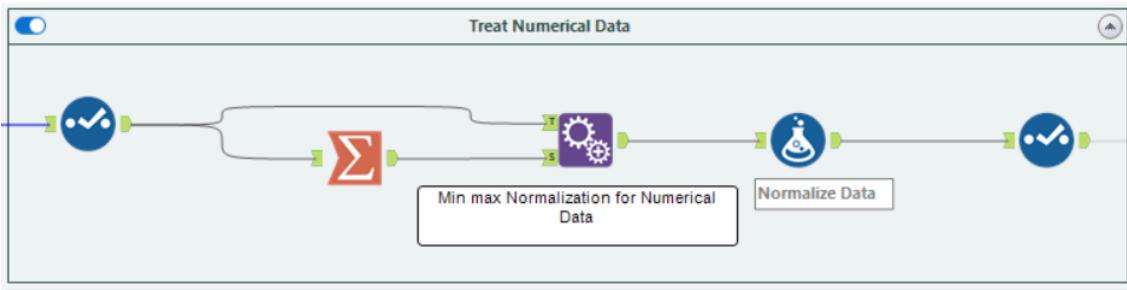
Formula (19) - Configuration

Output Column	Data Preview
Cabin	NA

```
fx IF [Cabin]==Null() THEN "NA" ELSE [Cabin] ENDIF
```

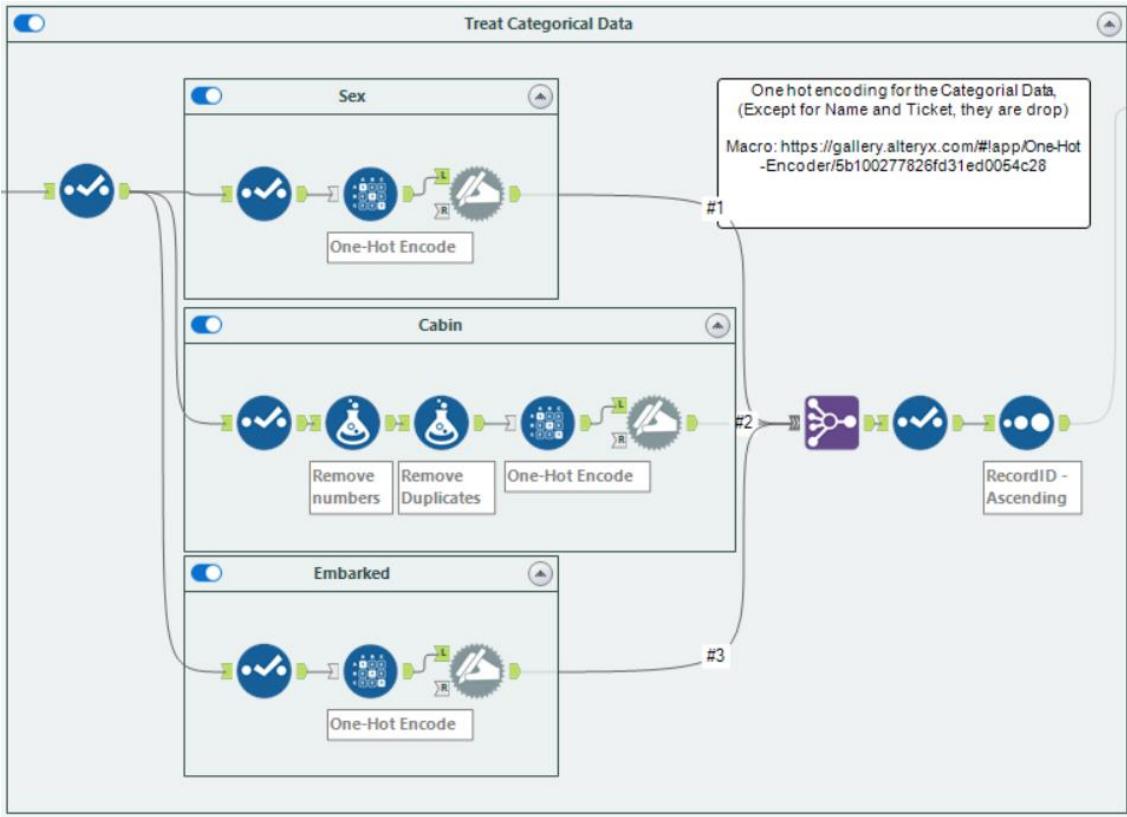
Data type: V_String Size: 254

Treat Numerical Data:



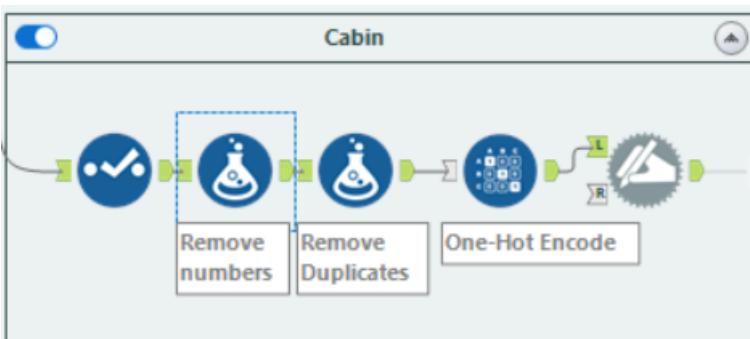
For Normalizing the Numerical Data, we used the same process used in the Numerical Predictor.

Treat Categorical Data:



For the Categorical Data, the variables Name and Ticket are drop because they don't bring any useful information. For the Sex and Embarked the same process from the Numerical Predictor is used.

The only difference is in Cabin, where we prepare the data before using the One-Hot encoding.



First we use the select tool to select only the PassengerID and the Cabin fields. Then the Formula tool is used to remove the numbers for the cabins, and a second Formula tool is used to delete duplicate letters, then we perform the One-Hot Encoding like before.

The first screenshot shows the 'Select (47) - Configuration' tool. It lists fields: PassengerId (selected), Name, Sex, Ticket, Cabin (selected), Embarked, and *Unknown. The 'Type' column shows various types like Int32, V_String, and Unknown. The 'Size' column shows sizes like 4, 254, and 0. The 'Rename' and 'Description' columns are empty.

The second screenshot shows the 'Formula (58) - Configuration' tool. It has an output column 'Cabin' set to NA. The formula is 'REGEX_Replace([Cabin], '\d+', '')'. The 'Data type' is V_String and 'Size' is 254.

The third screenshot shows the 'Formula (59) - Configuration' tool. It has an output column 'Cabin' set to NA. The formula is 'regex_replace([Cabin], "\b(\w+) (\?=.*\b\1 \?)", "")'. The 'Data type' is V_String and 'Size' is 254.

As on the Numerical Predictor, we Join all the categorical variables using the Multiple Join, then the Select tool changes the type of RecordID to int, lastly the Sort tool is used to order.

The first screenshot shows the 'Join Multiple (60) - Configuration' tool. It is set to 'Join by Specific Fields' and has three inputs: Input #1 (RecordID), Input #2 (RecordID), and Input #3 (RecordID). The 'Cartesian Join' checkbox is checked. There is an error message: 'Error on multidimensional joins of more than 16 Records'.

The second screenshot shows the 'Select (61) - Configuration' tool. It lists fields: RecordID (selected), Sex_female, Sex_male, Cabin_A, Cabin_B, Cabin_C, Cabin_D, Cabin_E, Cabin_F, Cabin_G, Cabin_H, Cabin_I, Cabin_J, Cabin_K, Cabin_L, Cabin_M, Cabin_N, Cabin_O, Cabin_P, Cabin_Q, Cabin_R, Cabin_S, Cabin_T, Embarked_C, Embarked_G, Embarked_Q, Embarked_S, and *Unknown. The 'Type' column shows Byte for most fields and Int32 for RecordID. The 'Size' column shows 1 for most fields and 4 for RecordID. The 'Rename' and 'Description' columns are empty.

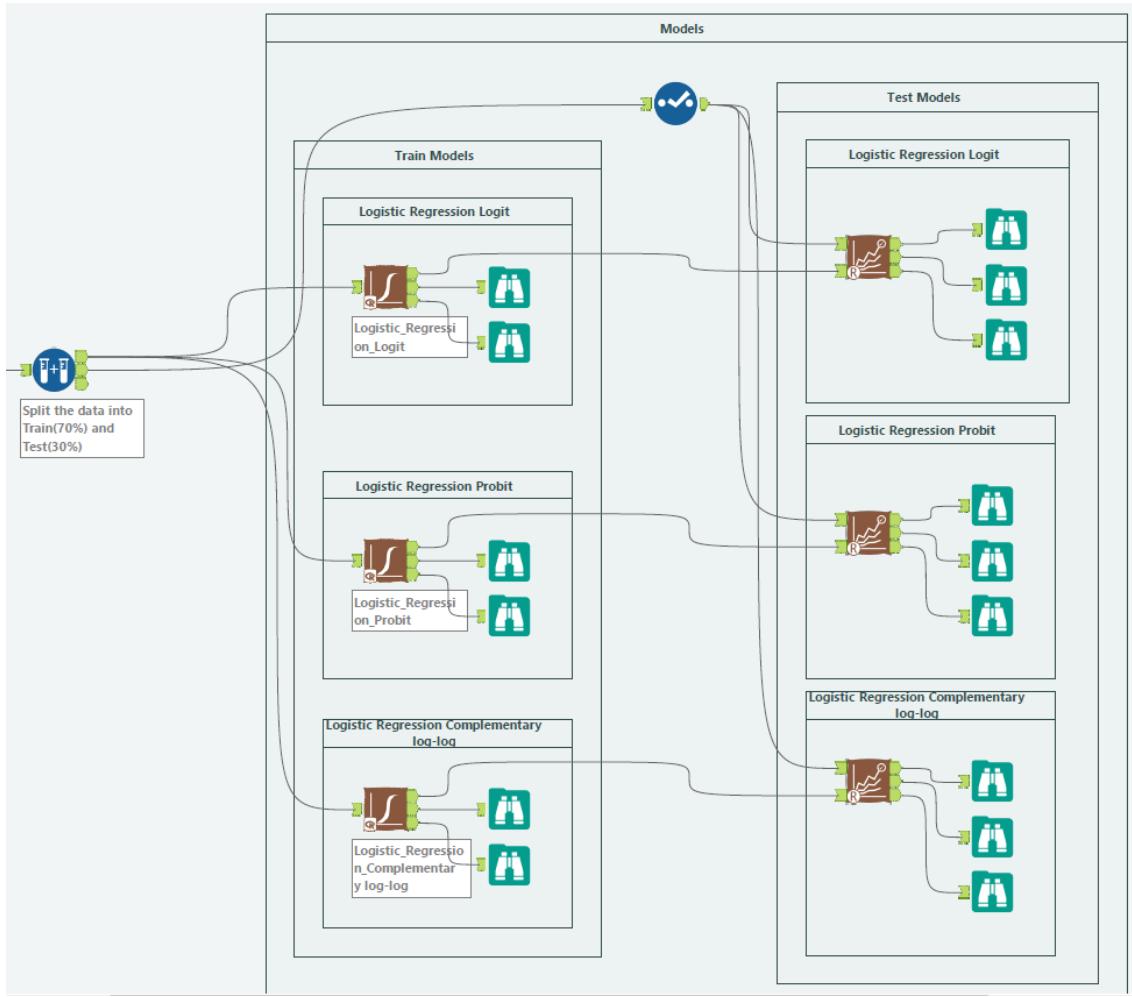
The third screenshot shows the 'Sort (62) - Configuration' tool. It has a single field 'RecordID' selected with 'Order' set to 'Ascending'.

Finally a Join tool is used to join the outputs from the Numerical and Categorical treatments

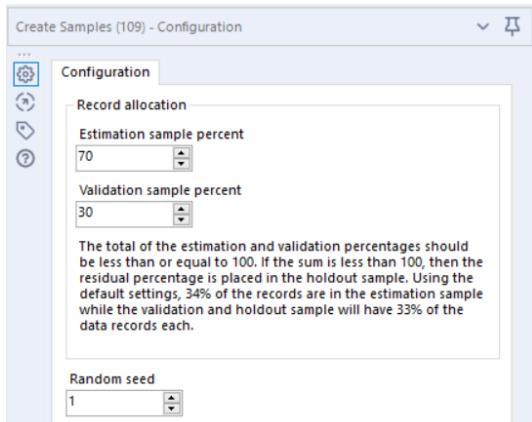
The dialog box shows the configuration for a 'Join (63)' operation. The top section is titled 'Join by Specific Fields' with a radio button. It displays a mapping between 'Left' input (PassengerId) and 'Right' input (RecordID). Below this is a table with columns: Input, Field, Type, Size, and Rename.

Input	Field	Type	Size	Rename
Left	PassengerId	Int32	4	
Left	Survived	V_String	254	
Left	Pclass	Double	8	
Left	Age	Double	8	
Left	SibSp	Double	8	
Left	Parch	Double	8	
Left	Fare	Double	8	
Right	RecordID	Int32	4	
Right	Sex_female	Byte	1	
Right	Sex_male	Byte	1	
Right	Cabin_A	Byte	1	
Right	Cabin_B	Byte	1	
Right	Cabin_C	Byte	1	
Right	Cabin_D	Byte	1	
Right	Cabin_E	Byte	1	
Right	Cabin_F	Byte	1	
Right	Cabin_F_E	Byte	1	
Right	Cabin_G	Byte	1	

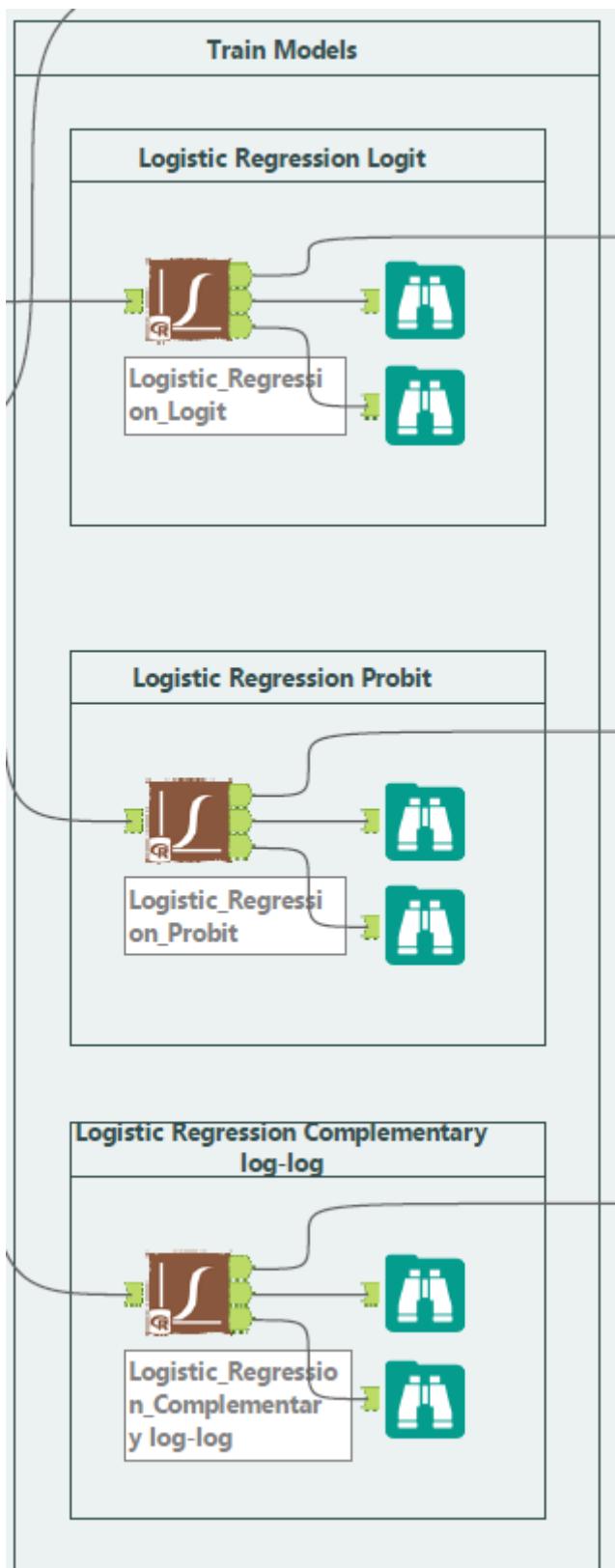
Models:



First, we Split the data into Train (70%) and Test (30%) and train the models:



Models Train:



Logistic Regression Logit:

The screenshot shows the configuration interface for a Logistic Regression Logit model. The left pane, titled 'Setup', includes fields for 'Type model name' (set to 'Logistic_Regression_Logit'), 'Select target variable' (set to 'Survived'), and 'Select predictor variables' (a list of 22 fields including Pclass, Age, SibSp, Parch, Fare, Sex_female, Sex_male, Cabin_A, Cabin_B, Cabin_C). The right pane, titled 'Customize', contains tabs for 'Model', 'Cross-validation', and 'Plots'. Under the 'Model' tab, there are options for 'Use sampling weights in model estimation (optional)', 'Use regularized regression', and 'Enter positive class for target variable (optional)'. The 'Select model type' dropdown is set to 'logit'.

Logistic Regression Probit:

The screenshot shows the configuration interface for a Logistic Regression Probit model. The left pane, titled 'Setup', includes fields for 'Type model name' (set to 'Logistic_Regression_Probit'), 'Select target variable' (set to 'Survived'), and 'Select predictor variables' (a list of 22 fields including Pclass, Age, SibSp, Parch, Fare, Sex_female, Sex_male, Cabin_A, Cabin_B, Cabin_C). The right pane, titled 'Customize', contains tabs for 'Model', 'Cross-validation', and 'Plots'. Under the 'Model' tab, there are options for 'Use sampling weights in model estimation (optional)', 'Use regularized regression', and 'Enter positive class for target variable (optional)'. The 'Select model type' dropdown is set to 'probit'.

Logistic Regression Complementary log-log:

The image shows two side-by-side windows for configuring a Logistic Regression model named "Logistic_Regression_Complementary log-log".

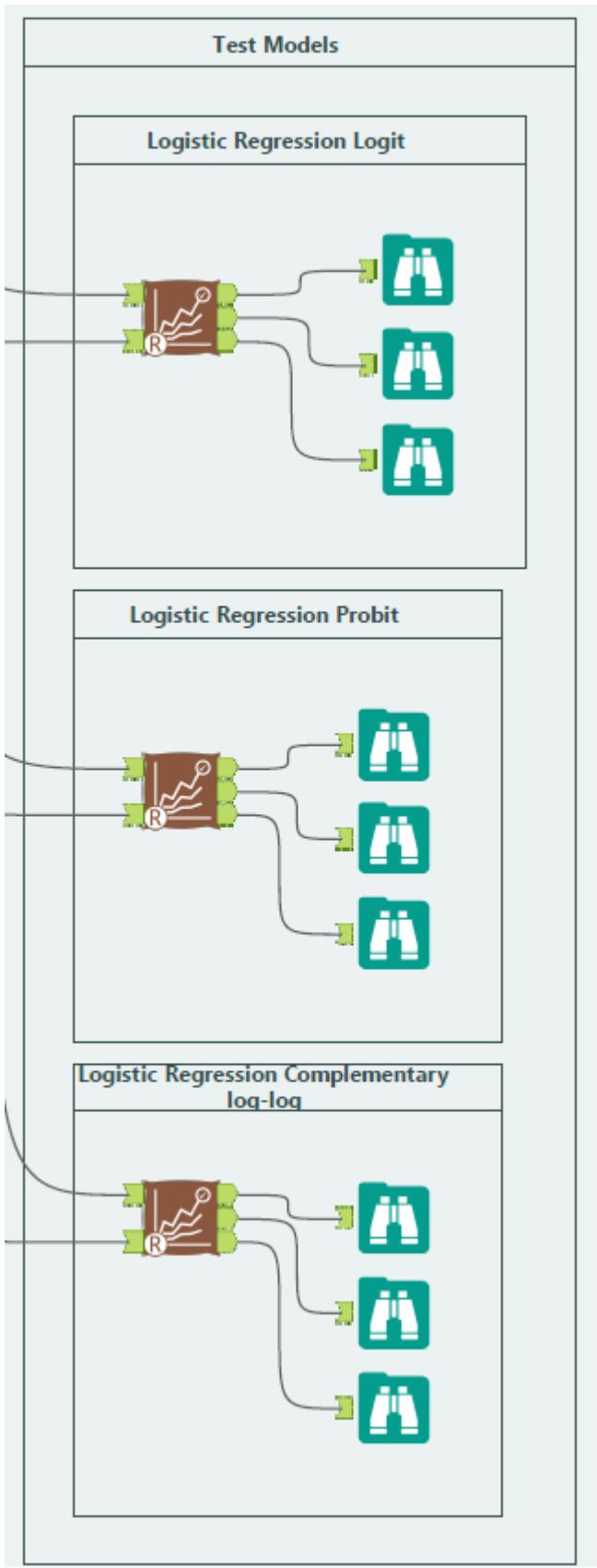
Setup Window:

- Type model name: Logistic_Regression_Complementary log-log
- Select target variable: Survived
- Select predictor variables:
 - Selected: 22 Fields: 22
 - Show: All Selected
 - Checklist of variables:
 - Pclass
 - Age
 - SibSp
 - Parch
 - Fare
 - Sex_female
 - Sex_male
 - Cabin_A
 - Cabin_B
 - Cabin_C

Customize Window:

- Model tab selected.
- Checkboxes:
 - Use sampling weights in model estimation (optional)
 - Use regularized regression
- Enter positive class for target variable (optional): [empty field]
- Select model type: complementary log-log

Test Models:



Results:

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_0	Accuracy_1
Logistic_Regression_Logit	0.8202	0.8596	0.8539	0.9074	0.6857

Model	Accuracy	F1	AUC	Accuracy_0	Accuracy_1
Logistic_Regression_Probit	0.8165	0.8563	0.8543	0.9012	0.6857

Model	Accuracy	F1	AUC	Accuracy_0	Accuracy_1
Logistic_Regression_Complementary log-log	0.8315	0.8703	0.8592	0.9321	0.6762

Confusion matrix of Logistic_Regression_Logit

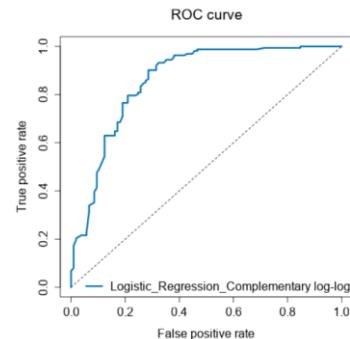
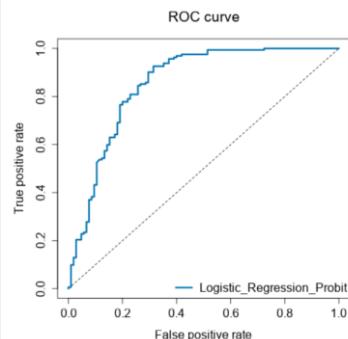
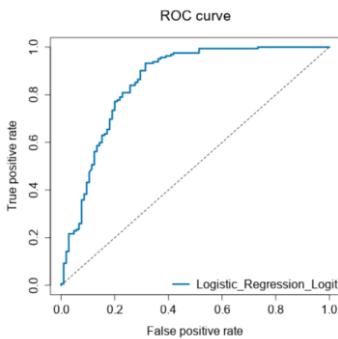
	Actual_0	Actual_1
Predicted_0	147	33
Predicted_1	15	72

Confusion matrix of Logistic_Regression_Probit

	Actual_0	Actual_1
Predicted_0	146	33
Predicted_1	16	72

Confusion matrix of Logistic_Regression_Complementary log-log

	Actual_0	Actual_1
Predicted_0	151	34
Predicted_1	11	71



Best Model

As we can see the **Complementary Log-Log** model perform the best with an AUC value of 0.8592, follow closely by the Probit model with AUC of 0.8543, and lastly the Logit model with an AUC of 0.80539. We can also see that the Complementary Log-Log is the best model from the Confusion Matrix.