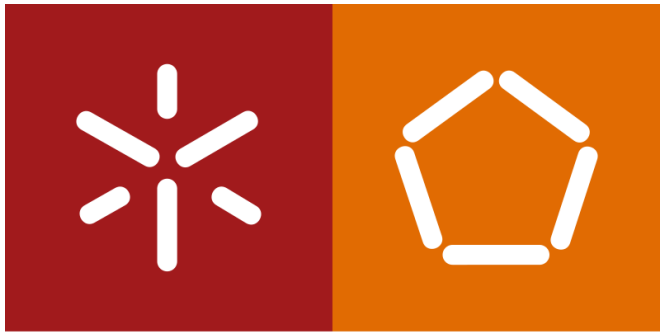


Universidade do Minho
Escola de Engenharia

S istemas de **A** prendizagem e **E** xtração de **C** onhecimento

José Machado

Diana Ferreira

REGRESSÃO LINEAR COM O RAPIDMINER

CONTEXTO E PRESPECTIVA

Recordam-se da Sara, a gerente de vendas regional do exemplo da aula de correlações? O seu negócio está em expansão, existindo cada vez mais novos clientes, e ela quer ter a certeza que a empresa será capaz de responder a este nível de procura.

A Sara sabe que há alguma correlação entre os atributos no seu conjunto de dados e agora questiona-se se poderá usar o mesmo conjunto de dados para prever o uso de óleo de aquecimento para novos clientes.

Os novos clientes ainda não começaram a consumir óleo de aquecimento. A Sara quer saber quanto óleo é necessário manter em *stock* para atender à demanda destes novos clientes.

O Data Mining pode ajudá-la a examinar os vários atributos e as quantidades de consumo de óleo de casos anteriores para antecipar e responder às necessidades dos novos clientes.

BUSINESS UNDERSTANDING

O novo objetivo de Sara é bastante claro: ela quer antecipar a demanda por óleo de aquecimento.

A Sara tem um *dataset* com 1.218 observações, usado na aula de correlações, que fornece um perfil de atributos para cada casa, juntamente com o consumo anual de óleo de aquecimento dessas casas. Ela pretende usar os dados desse *dataset* como dados de treino para construir um modelo capaz de prever o consumo dos novos clientes.

Para atender o objetivo da Sara, vamos usar um modelo de **regressão linear**, uma abordagem de modelação estatística que calcula uma relação entre uma resposta escalar (ou variável dependente) e uma ou mais variáveis explicativas (ou variáveis independentes) e que depois usa essa relação para efetuar a previsão.

DATA UNDERSTANDING

Sendo assim, o *dataset* usado na aula de correlações será usado para **treinar** o modelo. Recorde-se que este *dataset é* composto pelos seguintes atributos:

- **Insulation**: classificação de densidade que varia de 1 a 10 e indica a espessura do isolamento de cada casa. Uma casa com uma classificação de densidade de um é mal isolada, enquanto uma casa com uma densidade de dez possui um excelente isolamento.

- **Temperature**: temperatura ambiente média externa de cada casa no ano mais recente, medida em graus Fahrenheit.

- **Heating_Oil**: número total de unidades de óleo de aquecimento adquiridas pelo proprietário de cada casa no ano mais recente.

DATA UNDERSTANDING

- **Num_Occupants**: número total de ocupantes que vivem em cada casa.
- **Avg_Age**: idade média dos ocupantes que vivem em cada casa.
- **Home_Size**: classificação, numa escala de 1 a 8, do tamanho geral da casa. Quanto maior o número, maior a casa.

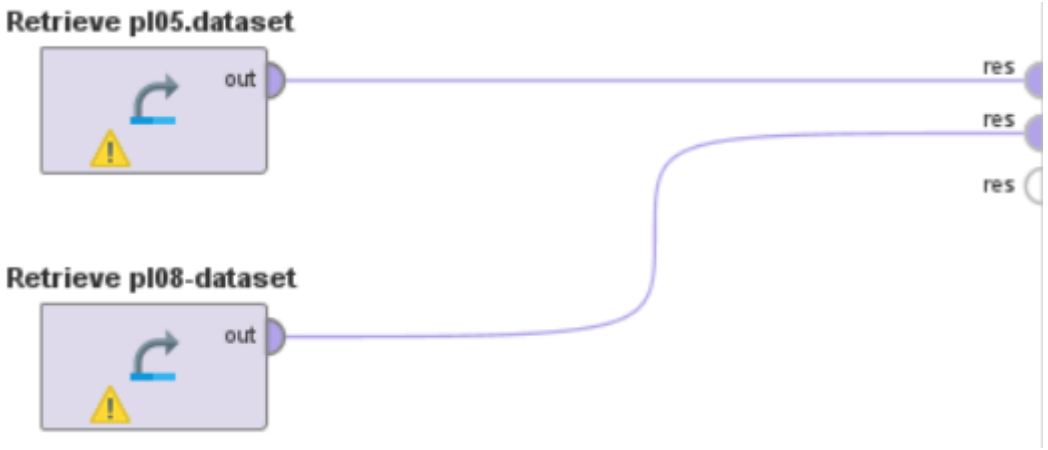
A Sara reuniu num ficheiro CSV os dados dos novos clientes contendo todos estes atributos, exceto, é claro, o **Heating_Oil**. Este conjunto de dados será o *dataset* usado para **testar** o modelo de regressão linear.

DATA PREPARATION

Download do dataset: pl05-dataset.csv + pl08-dataset.csv

1. Importe os *datasets* para o repositório rapidminer (Import Data -> My Computer).

2. Mude para a perspetiva de *design* e arraste os dois *datasets* para a janela do processo. Ligue ambas as portas *out* às portas *res*, como mostrado na figura abaixo, e depois execute o modelo.



DATA PREPARATION

Os intervalos para todos os atributos nos dados de teste devem estar dentro dos intervalos para os atributos correspondentes nos dados de treino.

PORQUÊ?

Um conjunto de dados de treino não pode ser utilizado para prever um atributo nos dados de teste com observações cujos valores estejam fora dos valores do conjunto de dados de treino.

Insulation	Integer	0	Min 2	Max 10	
Temperature	Integer	0	Min 38	Max 90	
Heating_Oil	Integer	0	Min 114	Max 301	Average 197.394
Num_Occupants	Integer	0	Min 1	Max 10	Average 3.113
Avg_Age	Real	0	Min 15.100	Max 72.200	Average 42.706
Home_Size	Integer	0	Min 1	Max 8	Average 4.649

dados de treino – pl05.dataset

DATA PREPARATION

Os intervalos são os mesmos para todos os atributos, exceto para o atributo Avg_Age. Os dados de teste possuem observações onde a Avg_Age está ligeiramente abaixo do limite inferior do conjunto de dados de treino de 15,1, e algumas observações onde a Avg_Age está ligeiramente acima do limite superior do conjunto de treino de 72,2.

			Min	Max	Average
Insulation	Integer	0	2	10	5.989
Temperature	Integer	0	38	90	63.962
Num_Occupants	Integer	0	1	10	5.489
Avg_Age	Real	0	15	73	44.040
Home_Size	Integer	0	1	8	4.495

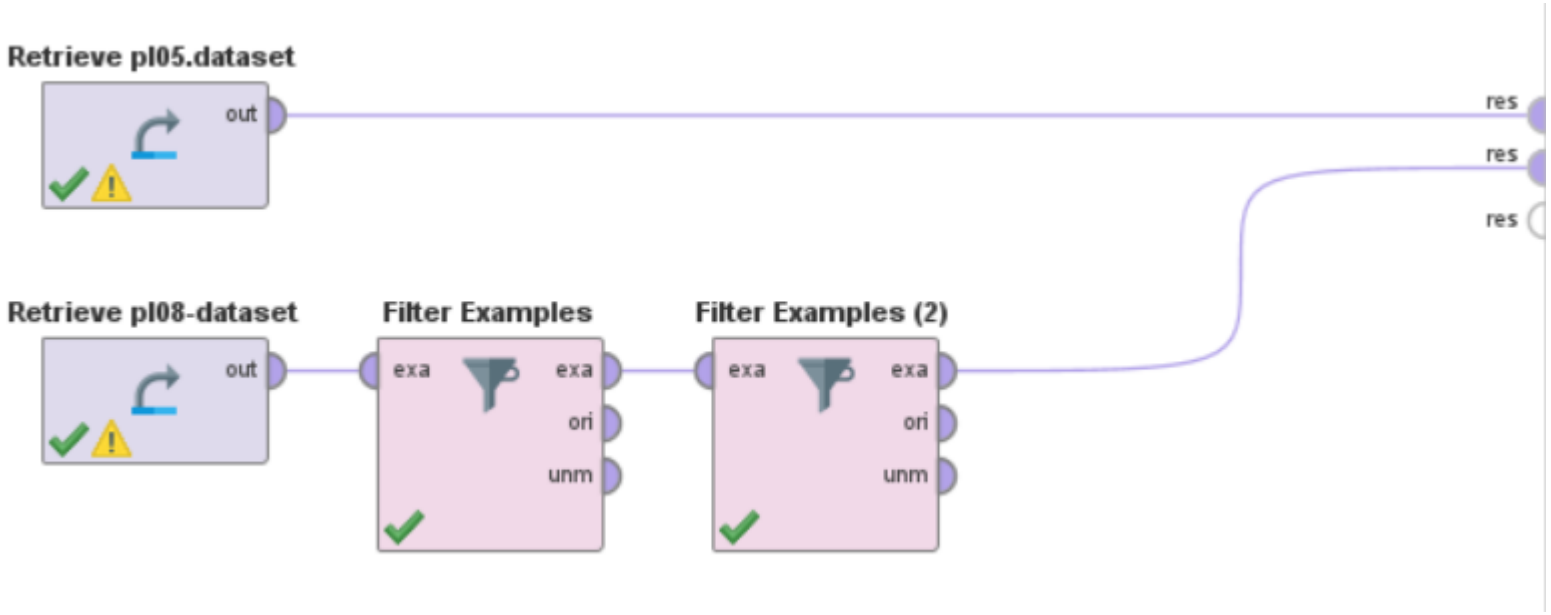
dados de teste – pl08.dataset

DATA PREPARATION

É necessário remover estas observações do conjunto de dados de teste.

3. Volte para a perspetiva de *design.* No separador Operadores, no canto inferior esquerdo, use a caixa de pesquisa para encontrar o operador 'Filter Examples'. Arraste dois operadores deste tipo para a janela do processo. Configure o parâmetro 'condition class' para 'attribute_value_filter' e o parâmetro 'parameter string' para:

Avg_Age >= 15.1
Avg_Age <= 72.2



DATA PREPARATION

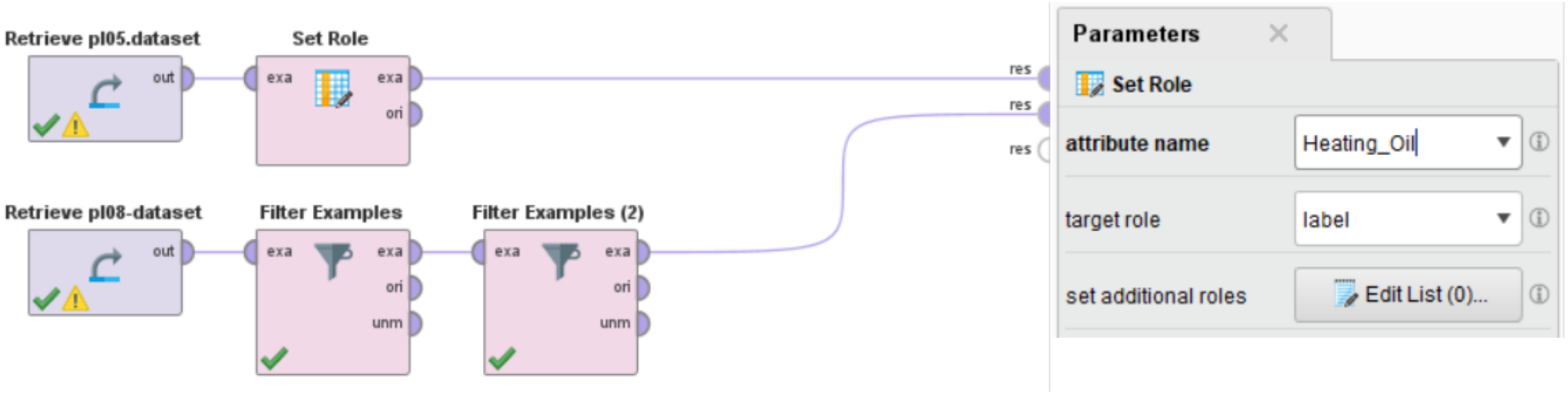
4. Execute o modelo. O conjunto de dados de teste possui agora 42.042 observações. Verifique novamente os intervalos dos atributos para se certificar de que nenhum dos atributos de teste possui intervalos fora dos valores dos atributos de treino.

Insulation	Integer	0	Min 2	Max 10	Average 5.988
Temperature	Integer	0	Min 38	Max 90	Average 63.949
Num_Occupants	Integer	0	Min 1	Max 10	Average 5.489
Avg_Age	Real	0	Min 15.100	Max 72.200	Average 43.674
Home_Size	Integer	0	Min 1	Max 8	Average 4.497

DATA PREPARATION

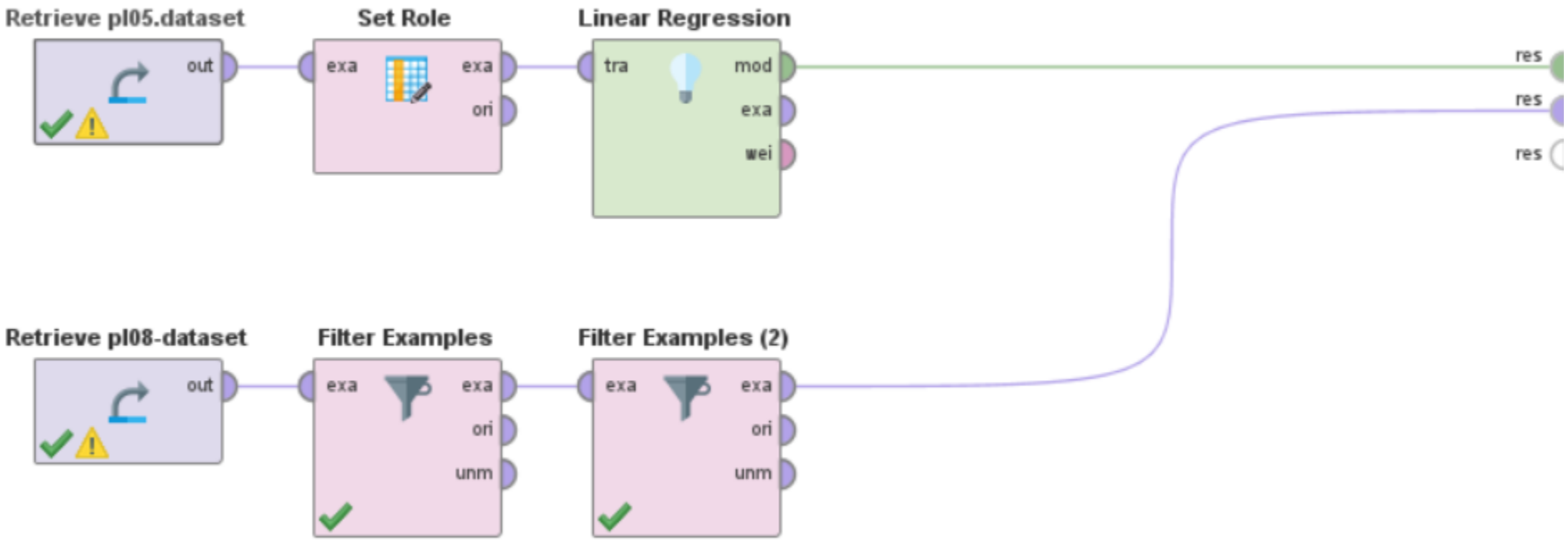
A regressão linear é um modelo preditivo e, portanto, precisa de um atributo para ser designado como *label* - este é o atributo alvo, aquilo que se pretende prever.

5. Volte à perspectiva de Design. Procure o operador “Set Role” e arraste-o para a janela de processo. Associe este operador ao fluxo de treino. Altere os parâmetros para designar Heating_Oil como o atributo alvo para este modelo.



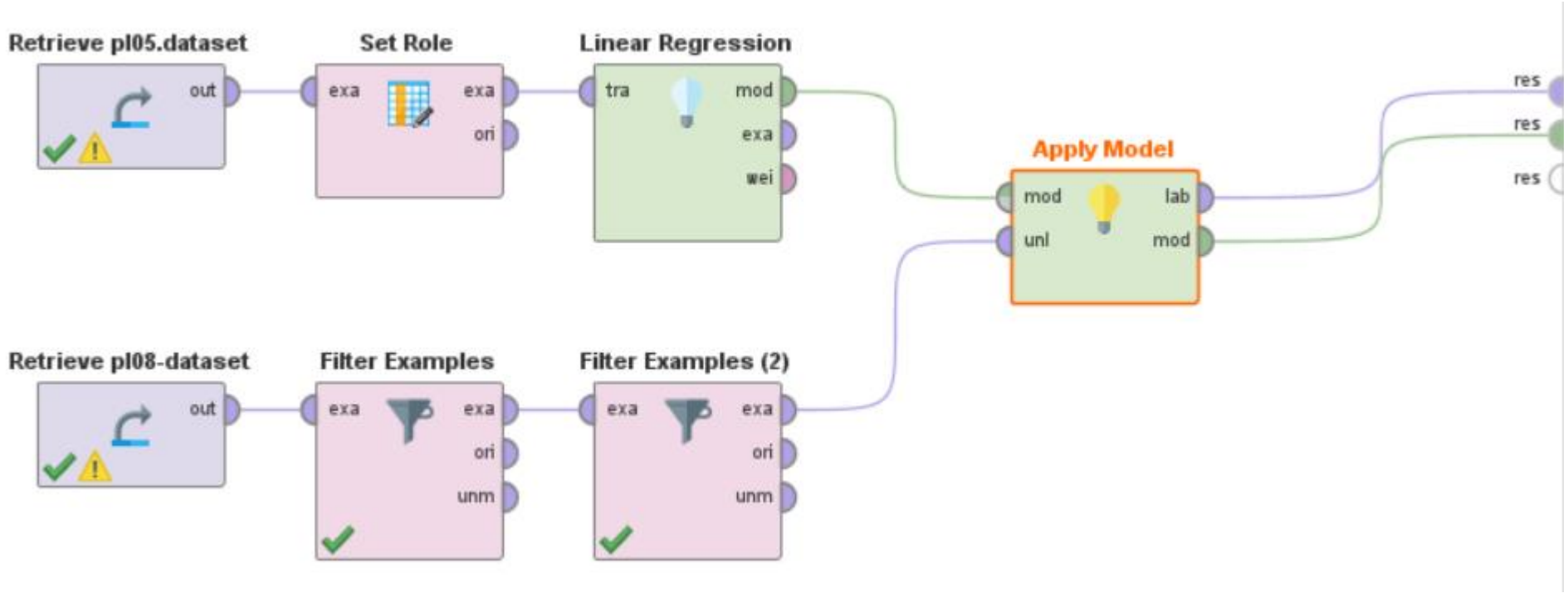
MODELING

1. Encontre o operador ‘Linear Regression’ e arraste-o para a janela do processo. Associe este operador ao fluxo de treino, como mostrado na figura abaixo.



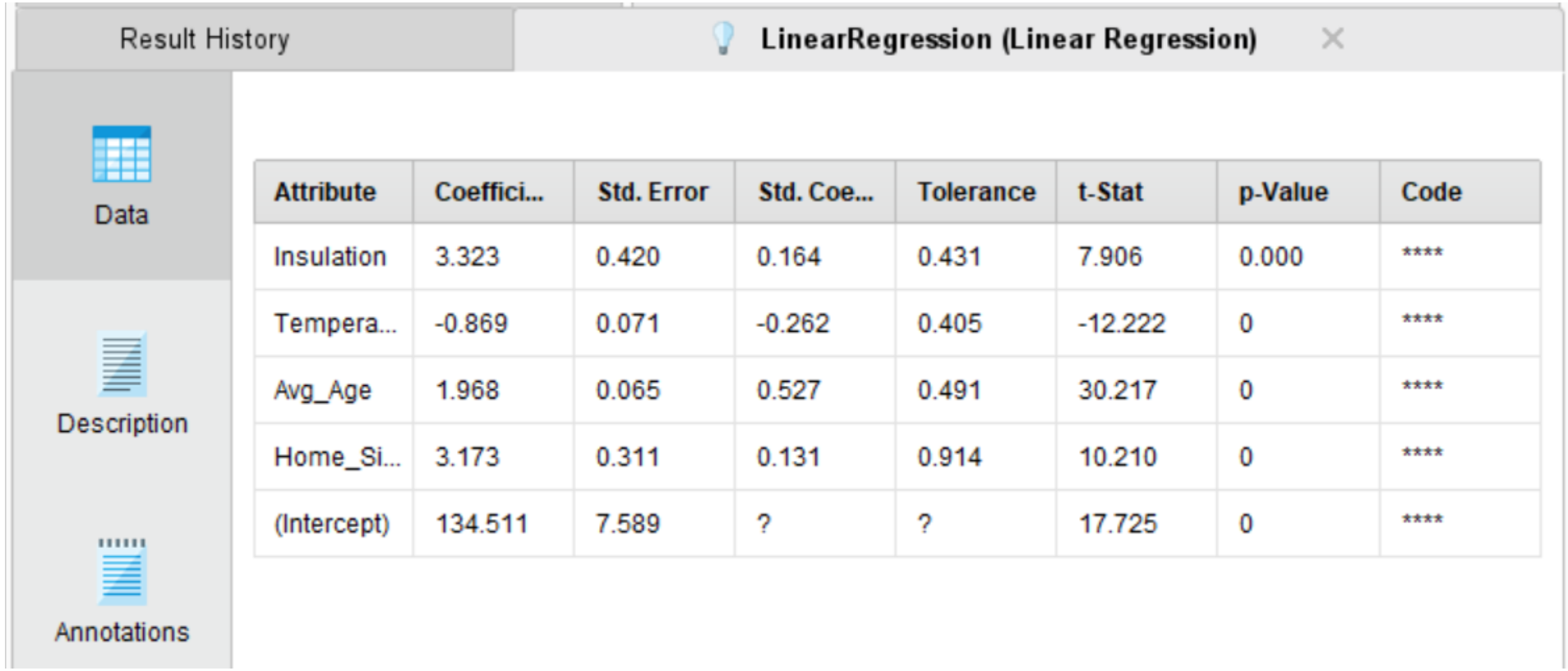
MODELING

2. O passo final para completar o modelo é usar um operador do tipo ‘Apply Model’ para conectar o fluxo de treino ao fluxo de teste. Procure este operador e arraste-o para a janela do processo. Certifique-se de conectar as portas *lab* e *mod* às portas res como ilustrado na figura.



EVALUATION

1. Corra o modelo. O facto de existirem duas saídas do operador 'Apply Model' conectadas às portas *res,* resultará em dois separadores na perspectiva de resultados. Vamos examinar primeiro o separador LinearRegression.



Attribute	Coeffici...	Std. Error	Std. Coe...	Tolerance	t-Stat	p-Value	Code
Insulation	3.323	0.420	0.164	0.431	7.906	0.000	****
Tempera...	-0.869	0.071	-0.262	0.405	-12.222	0	****
Avg_Age	1.968	0.065	0.527	0.491	30.217	0	****
Home_Si...	3.173	0.311	0.131	0.914	10.210	0	****
(Intercept)	134.511	7.589	?	?	17.725	0	****

EVALUATION

A modelação de regressão linear tem como objetivo determinar a proximidade de uma determinada observação com uma linha imaginária que representa a média ou o centro de todos os pontos no conjunto de dados.

Variável Independente
Insulation, p.e.

O que queremos prever (label)
Heating_Oil

$$y = mx + b$$

Constante
Valor do Intercept na tabela
134.511

Coeficiente do atributo
Segunda coluna da tabela
Para a Insulation = 3.323
(maior coeficiente de todos)

Se tivéssemos uma casa com densidade de isolamento de 5, a nossa fórmula usando esses valores de isolamento seria $y = (5 \times 3.323) + 134.511$

EVALUATION

Como podemos configurar esta fórmula linear quando temos várias variáveis independentes?

O resultado do operador LinearRegression possui apenas quatro atributos. O que aconteceu com o atributo Num_Occupants?

EVALUATION

O resultado do operador LinearRegression possui apenas quatro variáveis. O que aconteceu com Num_Occupants?

O Num_Occupants não era uma variável estatisticamente significativa para prever o uso de óleo de aquecimento neste *dataset* e, portanto, foi removido pelo RapidMiner.

Quando o RapidMiner avaliou a influência que cada atributo do *dataset* exercia sobre o uso de óleo de aquecimento para cada residência representada no *dataset* de treino, o número de ocupantes era tão pouco influente que o seu peso na fórmula foi definido como zero.

EVALUATION

Como podemos configurar a fórmula linear quando temos várias variáveis independentes?

$$y = mx + mx + mx \ldots + b$$

Por exemplo:

- Insulation: 6
- Temperature: 67
- Avg_Age: 35.4
- Home_Size: 5

$$y = (6 * 3.323) + (67 * -0.869) + (35.4 * 1.968) + (5 * 3.173) + 134.511 = 181.758$$

A previsão para o número anual desta casa de unidades de óleo de aquecimento encomendadas (y) é 181.758, ou seja, basicamente 182 unidades.

DEPLOYMENT

Ainda na *view* dos resultados, mude para o separador ExampleSet. Podemos observar que o modelo desenvolvido no RapidMiner fez uma previsão rápida e eficaz do número de unidades de óleo para aquecimento que cada um dos novos clientes da empresa da Sara provavelmente usará no seu primeiro ano.

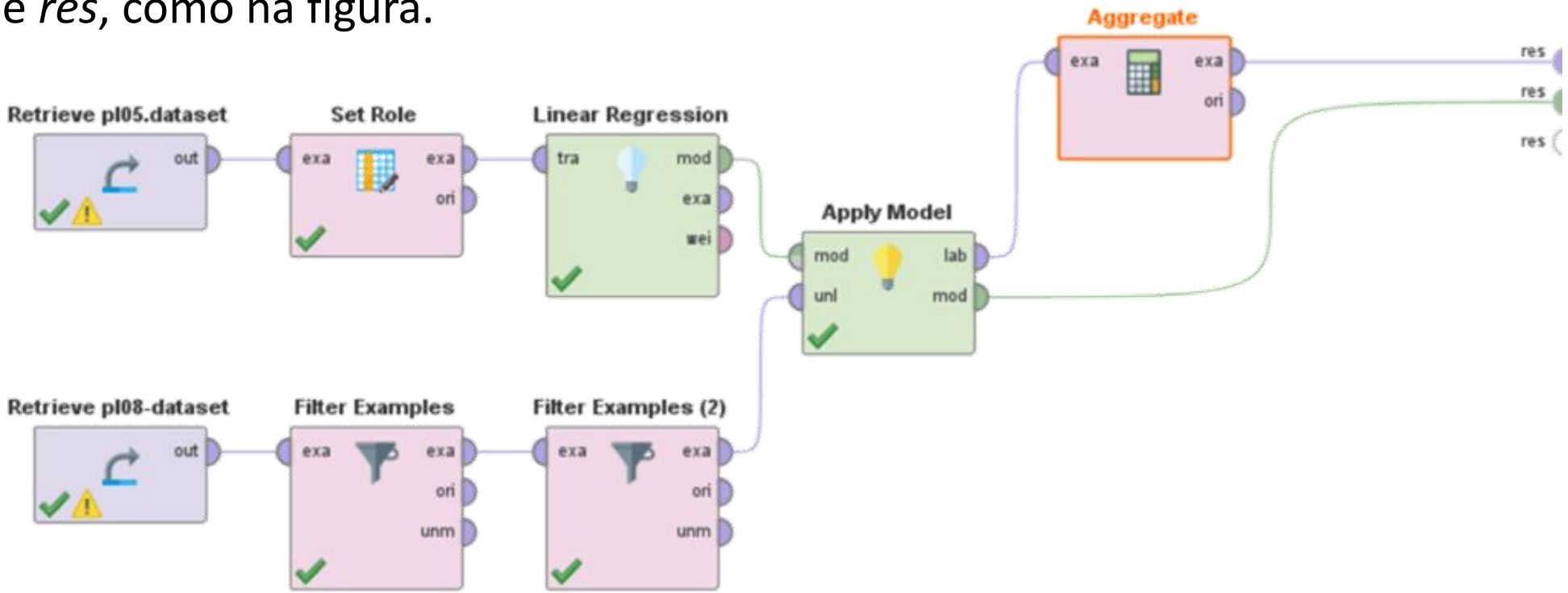
$$(5 * 3.323) + (69 * -0.869) + (70.1 * 1.968) + (7 * 3.173) + 134.511 = 251.321$$

Row No.	prediction(H...	Insulation	Temperature	Num_Occup...	Avg_Age	Home_Size
1	251.321	5	69	10	70.100	7
2	216.028	5	80	1	66.700	1
3	226.087	4	89	9	67.800	7
4	209.529	7	81	9	52.400	6
5	164.669	4	58	8	22.900	7
6	180.512	4	58	6	37.400	3
7	221.188	6	51	2	51.600	3
8	164.001	2	73	5	37.400	4

DEPLOYMENT

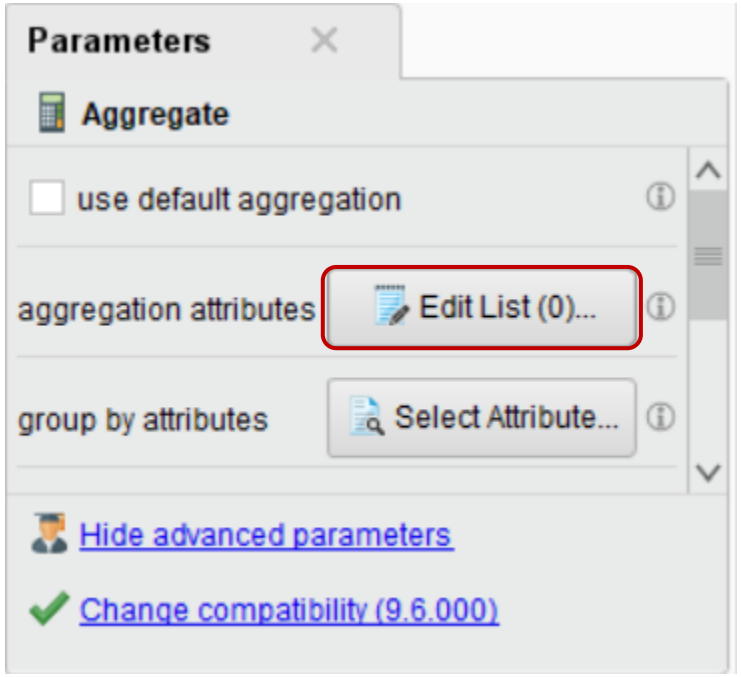
A Sara tem agora uma previsão do consumo de óleo para as casas de cada um dos novos clientes, à exceção daqueles com valores de Avg_Age fora do intervalo. Isso irá dizer-lhe o total de novas unidades de óleo de aquecimento que a sua empresa precisará de fornecer no próximo ano.

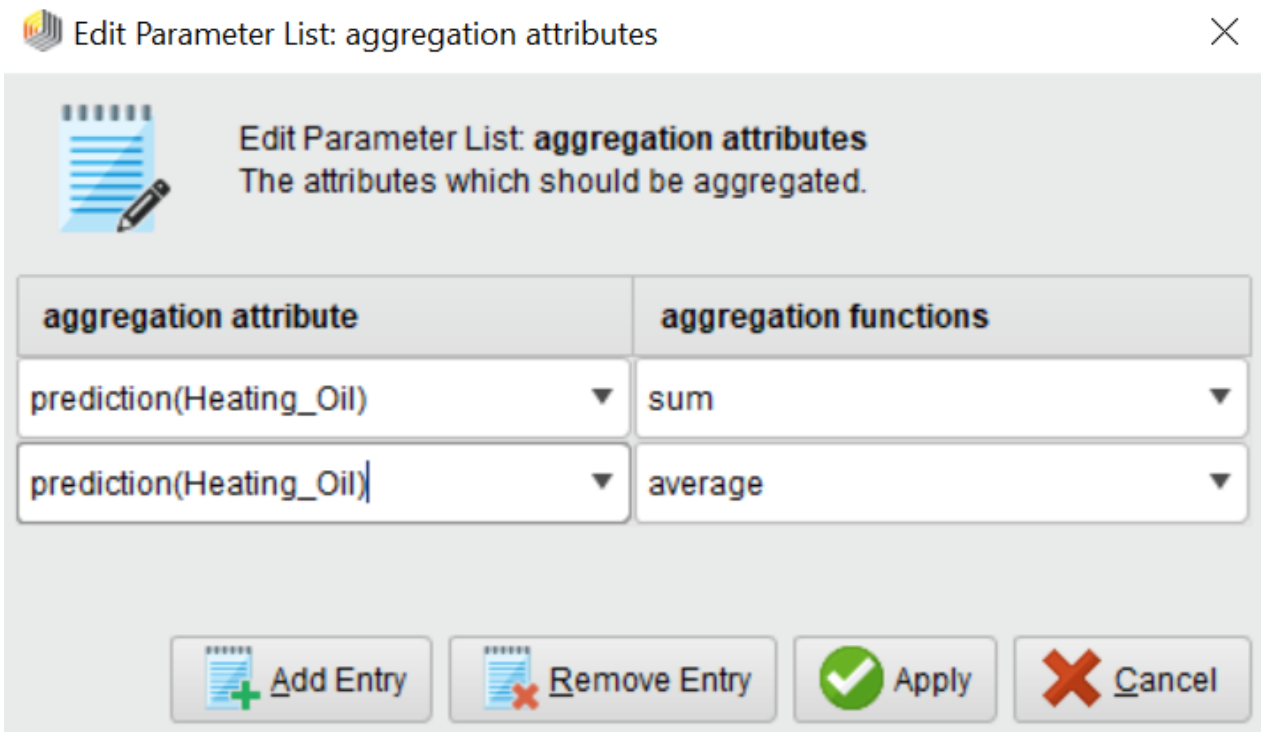
2. Volte para o separador *Design,* procure o operador 'Aggregate' e adicione-o entre as portas *lab* e *res*, como na figura.



DEPLOYMENT

3. Na tab ‘Parameters’, clique no boão *Edit List.* Defina o atributo de previsão (Heating_Oil) como o atributo de agregação e a função de agregação como “sum". Se quiser, pode adicionar outras agregações, como por exemplo a média.





DEPLOYMENT

4. Clique em OK para retornar à janela principal do processo e, em seguida, execute o modelo. Na secção dos resultados, selecione o separador *ExampleSet (Aggregate)* e selecione a opção Data.

Row No.	sum(predict...	average(pre...
1	8368087.536	199.041

A partir destes resultados, podemos ver que a empresa da Sara provavelmente venderá aproximadamente 8.368.088 unidades de óleo de aquecimento para os novos clientes. A empresa pode esperar que, em média, os novos clientes encomendem cerca de 200 unidades cada.

RESUMO

- A **regressão linear** é um modelo preditivo que usa *datasets* de treino e de teste para gerar previsões numéricas. É importante lembrar que a regressão linear usa **dados numéricos** para todos os seus atributos.

- Cada atributo no *dataset* é avaliado estatisticamente pela sua capacidade de prever o atributo do tipo *label*. Os atributos com **fraca capacidade de previsão são removidos** do modelo.

- À medida que mais dados são recolhidos, estes podem ser adicionados ao *dataset* de treino para o tornar mais robusto ou expandir os intervalos de alguns atributos para incluir ainda mais valores. É muito importante lembrar que os **intervalos para os atributos de *scoring*** devem estar **dentro dos intervalos dos atributos de treino** para garantir previsões válidas.