Hello,

I am new to RapidMiner and data science in general. I study Business Administration and have decided to write my bachelor thesis on predictive analytics. For this purpose, I would like to develop a model from the sales data of a retail company in RapidMiner, with which I can predict stock levels for future smartphone models.

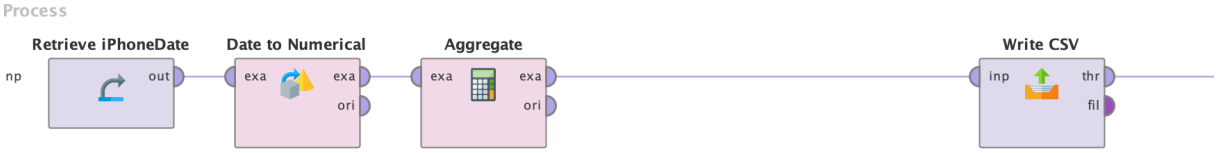
I have already been able to acquire some very helpful information in this forum and in the learning environment of rapidminer. But now I am at a point where I am stuck, because I also lack some basic knowledge.

I have received sales data from certain product categories of the last three years from a retailer. The data includes lines for each sale and some information about the product sold. I then edited the data in excel and divided some technical specifications as individual attributes. I have subdivided the specification "capacity" into "low", "medium" and high rather than the existing gigabytes (32GB, 64GB, ...) in order to eliminate discrepancies resulting from technical progress. As a further attribute, I have added the age of the product.The aim of me was to classify the individual attributes of a product as generally as possible. I have added some examples of the edited Data as screenshot. For legal reasons, I am unable to upload the entire datasets.

A	B	C	D	E	F	G	H	I	J
LFNR_Artikel	**Model Text**	**Date**	**Amount**	**Age**	**Modellreihe**	**Modell 1**	**Modell 2**	**Capacity**	**Color**
MG4H2ZD/A	iPhone 6; 64GB; Silver	15.02.16	1	1 year	6	N	normal	medium	Silver
MKQN2ZD/A	iPhone 6s; 64GB; Space Gray	15.02.16	1	current	6s	S	normal	medium	SpaceGray
MKQN2ZD/A	iPhone 6s; 64GB; Space Gray	15.02.16	1	current	6s	S	normal	medium	SpaceGray
MKQN2ZD/A	iPhone 6s; 64GB; Space Gray	15.02.16	1	current	6s	S	normal	medium	SpaceGray
MKQN2ZD/A	iPhone 6s; 64GB; Space Gray	15.02.16	1	current	6s	S	normal	medium	SpaceGray
MKQP2ZD/A	iPhone 6s; 64GB; Silver	15.02.16	1	current	6s	S	normal	medium	Silver
MKQR2ZD/A	iPhone 6s; 64GB; RoSE; Gold	15.02.16	1	current	6s	S	normal	medium	Gold
MKQK2ZD/A	iPhone 6s; 16GB; Silver	16.02.16	1	current	6s	S	normal	low	Silver
MKQL2ZD/A	iPhone 6s; 16GB; Gold	16.02.16	1	current	6s	S	normal	low	Gold
MKQN2ZD/A	iPhone 6s; 64GB; Space Gray	16.02.16	1	current	6s	S	normal	medium	SpaceGray
MKQN2ZD/A	iPhone 6s; 64GB; Space Gray	16.02.16	1	current	6s	S	normal	medium	SpaceGray
MKQP2ZD/A	iPhone 6s; 64GB; Silver	16.02.16	1	current	6s	S	normal	medium	Silver
MKQR2ZD/A	iPhone 6s; 64GB; RoSE; Gold	16.02.16	1	current	6s	S	normal	medium	Gold
ME432DN/A	iPhone 5S 16GB; Spacegrau	17.02.16	1	2 years	5S	S	normal	low	SpaceGray

My idea was then to create a sum of the sales per week for the respective products, since the sales figures of one week are very close to the basic stock level. Furthermore, I removed attributes, which in my opinion are no longer relevant for the further processing.

Here a screenshot of the preparation process and a example of the output data:



Fehler! Es wurde kein Dateiname angegeben.

Age	Modell 1	Modell 2	Capacity	Color	Week	sum(Amount)
1 year	N	Plus	low	SpaceGray	5	1
3 years	S	Plus	high	Gold	9	4
current	N	normal	mid	Silver	24	3

In the predicting process, I would like to train a model with the prepared data and apply it to example Data of a new product.
For the data of the new product, i used of course the same attributes as the training data

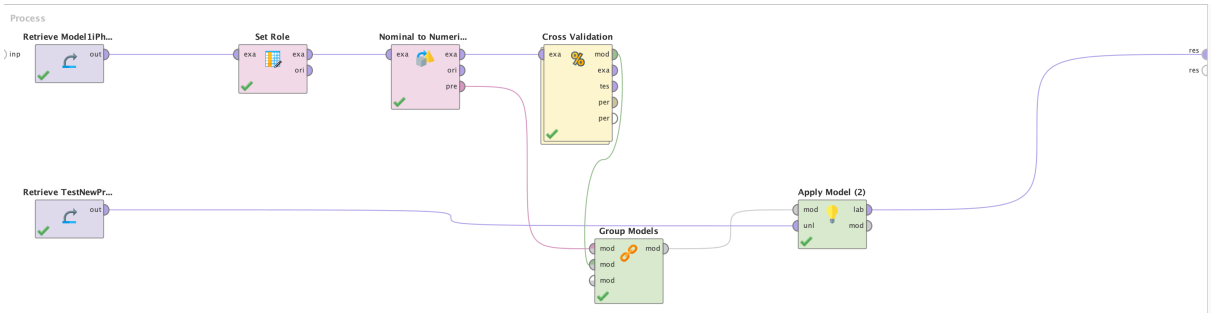
and wantet the prediction for all 52 weeks of a year.

Example New product data:

Age	Modell 1	Modell 2	Capacity	Color	Week
current	S	normal	low	SpaceGray	1
current	S	normal	low	SpaceGray	2
current	S	normal	low	SpaceGray	3
current	S	normal	low	SpaceGray	4
current	S	normal	low	SpaceGray	5
current	S	normal	low	SpaceGray	6
current	S	normal	low	SpaceGray	7
current	S	normal	low	SpaceGray	8
current	S	normal	low	SpaceGray	9
current	S	normal	low	SpaceGray	10
current	S	normal	low	SpaceGray	11
current	S	normal	low	SpaceGray	12
current	S	normal	low	SpaceGray	13
current	S	normal	low	SpaceGray	14
current	S	normal	low	SpaceGray	15
current	S	normal	low	SpaceGray	16
current	S	normal	low	SpaceGray	17
current	S	normal	low	SpaceGray	18
current	S	normal	low	SpaceGray	19
current	S	normal	low	SpaceGray	20
current	S	normal	low	SpaceGray	21
current	S	normal	low	SpaceGray	22
current	S	normal	low	SpaceGray	23
current	S	normal	low	SpaceGray	24

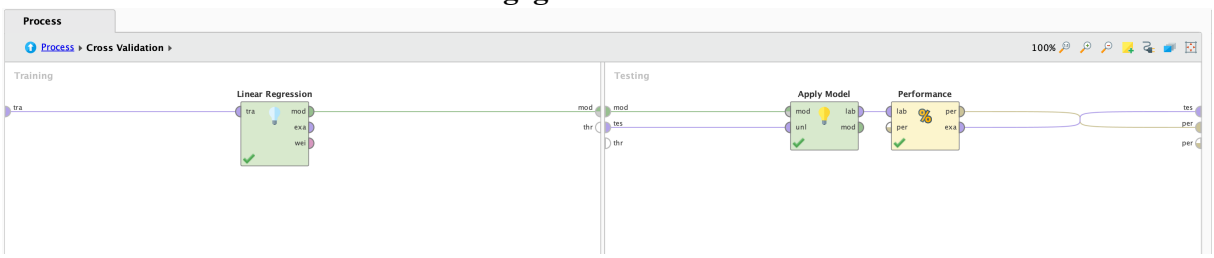
....

I then created with the help of some tutorials this process:



Cross Validation:

Fehler! Es wurde kein Dateiname angegeben.



Now I am faced with the problem that the data on the sales volume per week

"sum(Amount)" generated by this process are not very realistic, or the data differs too little per week (the data output is eg between 9 and 13, while the input data of the Training sets for these specific attributes are between 1 and 72.

My questions are: am I on the right track at all? And is the linear regression the right model for my task?

Many thanks for your help

Andy