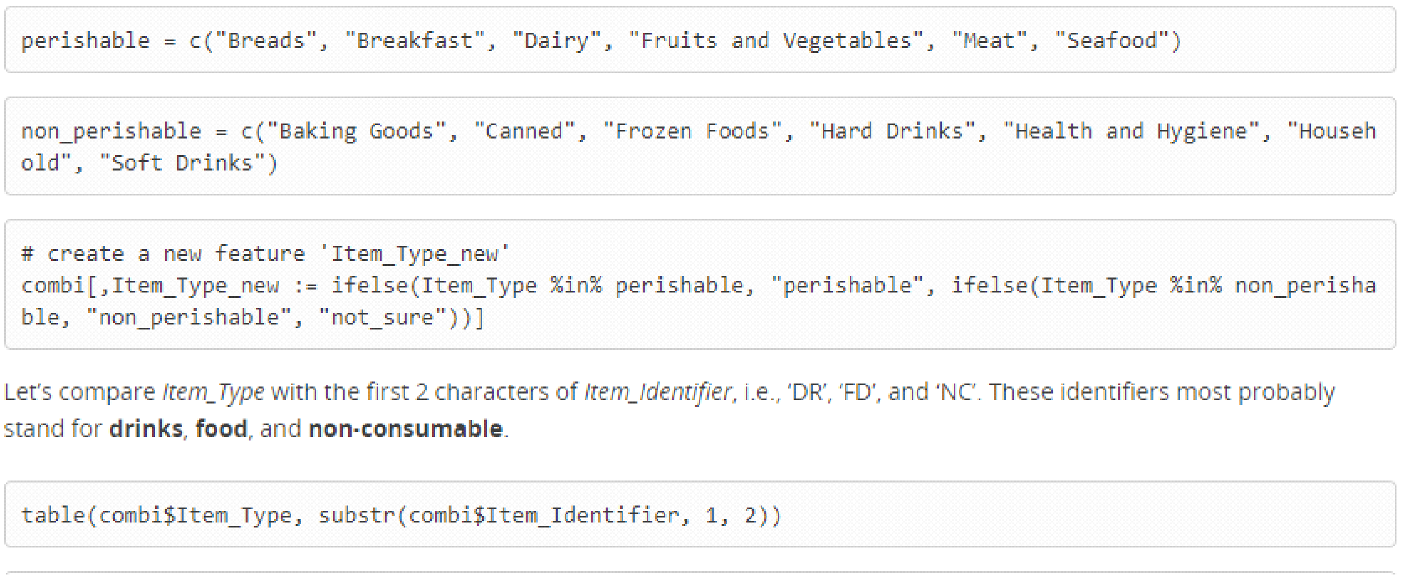
19 November 2018
17:08

Most of the times, the given features in a dataset are not sufficient to give satisfactory predictions. In such cases, we have to create new features which might help in improving the model’s performance. Let’s try to create some new features for our dataset.

In this section we will create the following new features:

- **Item_Type_new**: Broader categories for the variable *Item_Type*.
- **Item_category**: Categorical variable derived from *Item_Identifier*.
- **Outlet_Years**: Years of operation for outlets.
- **price_per_unit_wt**: *Item_MRP/Item_Weight*
- **Item_MRP_clusters**: Binned feature for *Item_MRP*.

We can have a look at the *Item_Type* variable and classify the categories into **perishable** and **non_perishable** as per our understanding and make it into a new feature.



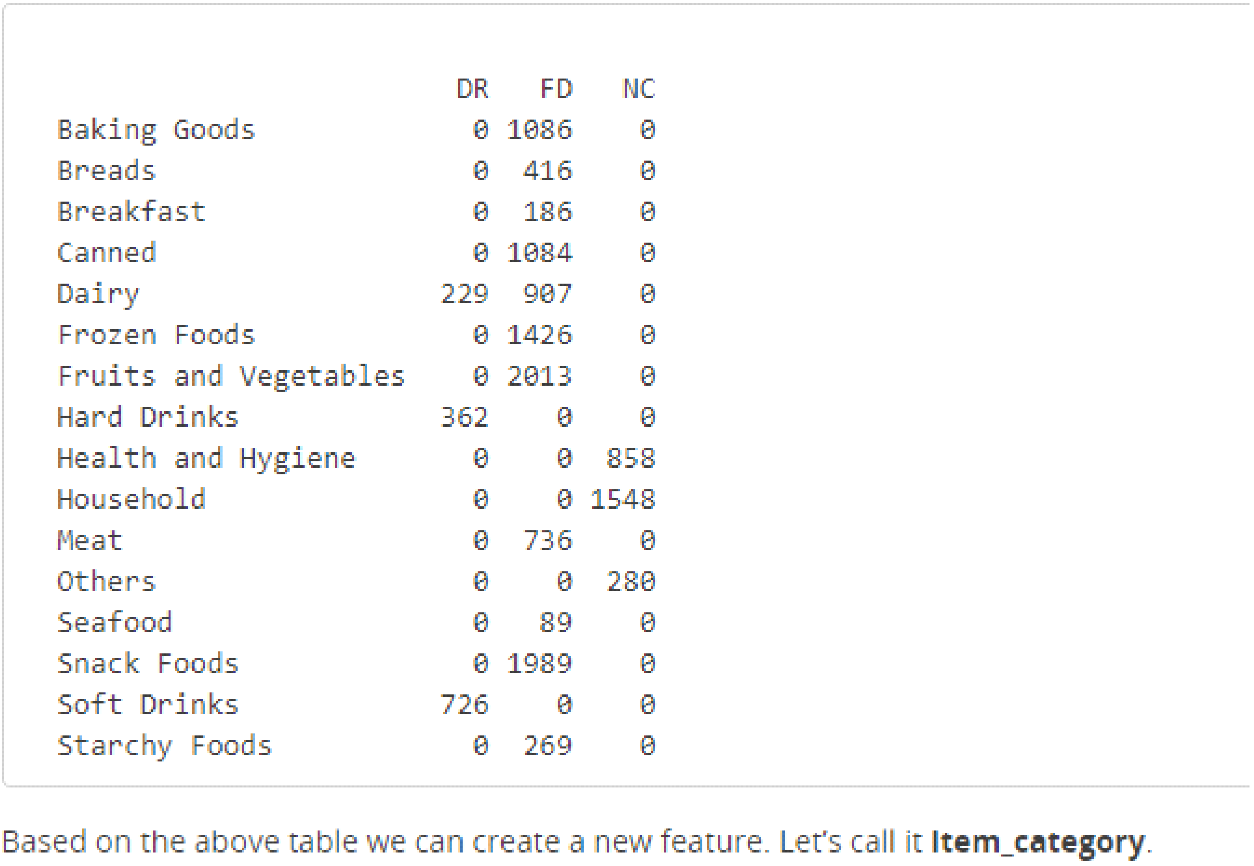
```
perishable = c("Breads", "Breakfast", "Dairy", "Fruits and Vegetables", "Meat", "Seafood")
```

```
non_perishable = c("Baking Goods", "Canned", "Frozen Foods", "Hard Drinks", "Health and Hygiene", "Househ
old", "Soft Drinks")
```

```
# create a new feature 'Item_Type_new'
combi[,Item_Type_new := ifelse(Item_Type %in% perishable, "perishable", ifelse(Item_Type %in% non_perisha
ble, "non_perishable", "not_sure"))]
```

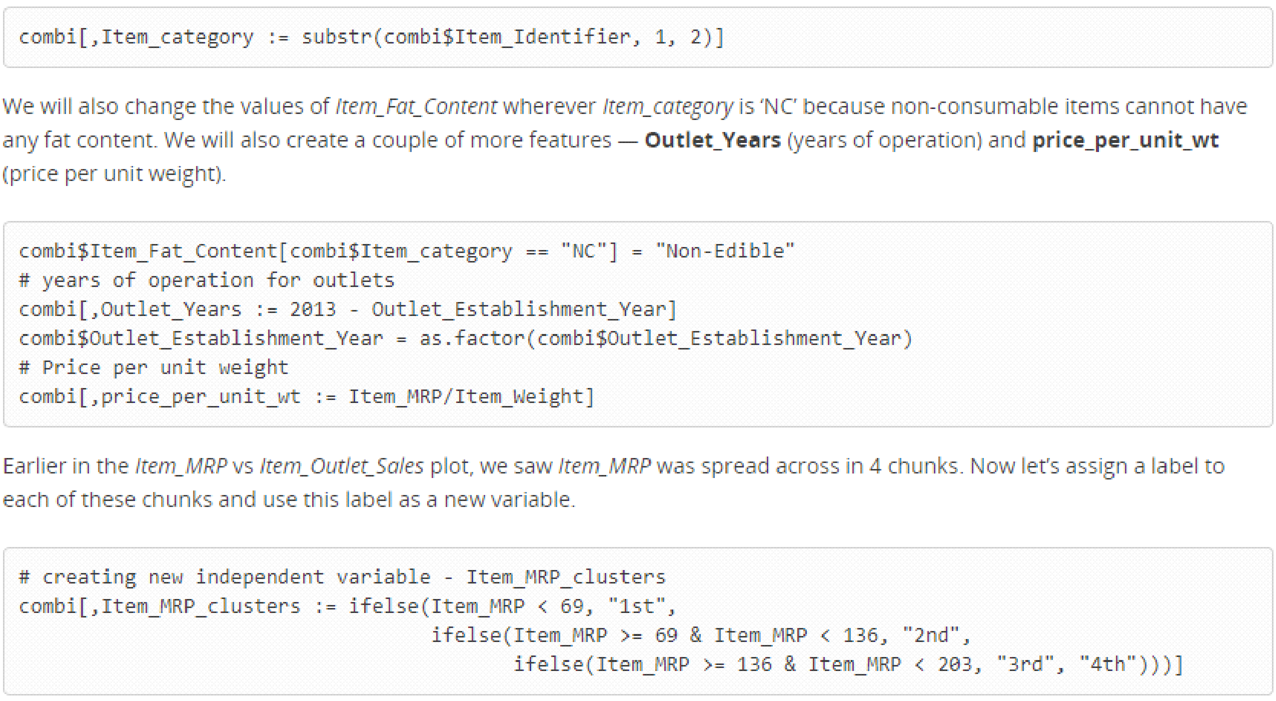
Let’s compare *Item_Type* with the first 2 characters of *Item_Identifier*, i.e., ‘DR’, ‘FD’, and ‘NC’. These identifiers most probably stand for **drinks**, **food**, and **non-consumable**.

```
table(combi$Item_Type, substr(combi$Item_Identifier, 1, 2))
```



```
                        DR   FD   NC
Baking Goods             0 1086    0
Breads                   0  416    0
Breakfast                0  186    0
Canned                   0 1084    0
Dairy                  229  907    0
Frozen Foods             0 1426    0
Fruits and Vegetables    0 2013    0
Hard Drinks            362    0    0
Health and Hygiene       0    0  858
Household                0    0 1548
Meat                     0  736    0
Others                   0    0  280
Seafood                  0   89    0
Snack Foods              0 1989    0
Soft Drinks            726    0    0
Starchy Foods            0  269    0
```

Based on the above table we can create a new feature. Let's call it **Item_category**.



```
combi[,Item_category := substr(combi$Item_Identifier, 1, 2)]
```

We will also change the values of *Item_Fat_Content* wherever *Item_category* is 'NC' because non-consumable items cannot have any fat content. We will also create a couple of more features — **Outlet_Years** (years of operation) and **price_per_unit_wt** (price per unit weight).

```
combi$Item_Fat_Content[combi$Item_category == "NC"] = "Non-Edible"
# years of operation for outlets
combi[,Outlet_Years := 2013 - Outlet_Establishment_Year]
combi$Outlet_Establishment_Year = as.factor(combi$Outlet_Establishment_Year)
# Price per unit weight
combi[,price_per_unit_wt := Item_MRP/Item_Weight]
```

Earlier in the *Item_MRP* vs *Item_Outlet_Sales* plot, we saw *Item_MRP* was spread across in 4 chunks. Now let's assign a label to each of these chunks and use this label as a new variable.

```
# creating new independent variable - Item_MRP_clusters
combi[,Item_MRP_clusters := ifelse(Item_MRP < 69, "1st",
                                   ifelse(Item_MRP >= 69 & Item_MRP < 136, "2nd",
                                          ifelse(Item_MRP >= 136 & Item_MRP < 203, "3rd", "4th")))]
```