Détecter des faux billets

CME

Contexte

Vous êtes consultant Data Analyst dans une entreprise spécialisée dans la data. Votre entreprise a décroché une prestation en régie au sein de l'**Organisation nationale de lutte contre le faux-monnayage (ONCFM)**.

Cette institution a pour objectif de mettre en place des méthodes d'identification des contrefaçons des billets en euros. Ils font donc appel à vous, spécialiste de la data, pour mettre en place une modélisation qui serait capable d'identifier automatiquement les vrais des faux billets. Et ce à partir simplement de certaines dimensions du billet et des éléments qui le composent.

Voici le cahier des charges de l'ONCFM ainsi que le jeu de données

Le client souhaite que vous travailliez directement depuis ses locaux sous la responsabilité de Marie, responsable du projet d'analyse de données à l'ONCFM. Elle vous laissera une grande autonomie pendant votre mission, et vous demande simplement que vous lui présentiez vos résultats une fois la mission terminée. Elle souhaite voir quels sont les traitements et analyses que vous avez réalisés en amont, les différentes pistes explorées pour la construction de l'algorithme, ainsi que le modèle final retenu.

Après avoir lu en détail le cahier des charges, vous vous préparez à vous rendre à l'ONCFM pour prendre vos nouvelles fonctions. Vous notez tout de même un post-it qui se trouve sur le coin de votre bureau, laissé par un de vos collègues :

Importation des fichiers

```
data <- read.csv("data_raw/billets.csv", sep =";")
```

Résumé des datas

```
summary(data)
```

```
  is_genuine           diagonal      height_left     height_right
 Length:1500        Min.   :171.0   Min.   :103.1   Min.   :102.8
 Class :character   1st Qu.:171.8   1st Qu.:103.8   1st Qu.:103.7
 Mode  :character   Median :172.0   Median :104.0   Median :103.9
                    Mean   :172.0   Mean   :104.0   Mean   :103.9
                    3rd Qu.:172.2   3rd Qu.:104.2   3rd Qu.:104.2
                    Max.   :173.0   Max.   :104.9   Max.   :105.0

   margin_low      margin_up         length
 Min.   :2.980   Min.   :2.270   Min.   :109.5
 1st Qu.:4.015   1st Qu.:2.990   1st Qu.:112.0
 Median :4.310   Median :3.140   Median :113.0
 Mean   :4.486   Mean   :3.151   Mean   :112.7
 3rd Qu.:4.870   3rd Qu.:3.310   3rd Qu.:113.3
 Max.   :6.900   Max.   :3.910   Max.   :114.4
 NA's   :37
```

Nous avons un dataframe de 7 colonnes et 1 500 lignes 1 colonne de type character 6 colonnes numériques

Description des variables

```
if (!require(skimr)) install.packages("skimr")
```

```
Le chargement a nécessité le package : skimr
```

```
library(skimr)

skim(data)
```

Table 1: Data summary

Name	data
Number of rows	1500
Number of columns	7
_____	
Column type frequency:	
character	1
numeric	6
_____	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
is_genuine	0	1	4	5	0	2	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
diagonal	0	1.00	171.96	0.31	171.04	171.75	171.96	172.17	173.01	
height_left	0	1.00	104.03	0.30	103.14	103.82	104.04	104.23	104.88	
height_right	0	1.00	103.92	0.33	102.82	103.71	103.92	104.15	104.95	
margin_low	37	0.98	4.49	0.66	2.98	4.02	4.31	4.87	6.90	
margin_up	0	1.00	3.15	0.23	2.27	2.99	3.14	3.31	3.91	
length	0	1.00	112.68	0.87	109.49	112.03	112.96	113.34	114.44	

Nous avons 37 valeurs manquantes dans la colonne margin_low

```
valeur_unique <- unique(data$is_genuine)
print(valeur_unique)
```

```
[1] "True"  "False"
```

Nous avons 2 valeurs uniques dans la colonne is_genuine => True ou False

```
valeur_compte <- table(data$is_genuine)
print(valeur_compte)
```

```
False  True
  500  1000
```

Il y a **500 valeurs False** et **1 000 valeurs True**

En résumé

Nous avons un tableau regroupant les données de 1 500 billets

1 colonne décrivant s'il s'agit de vrais ou faux billets :
- il y a 1 000 vrais billets et 500 faux billets

6 colonne décrivants le format de ces billets :
- diagonale
- hauteur gauche
- hauteur droite
- marge basse
- marge haute
- longueur

37 billets n'ont pas l'information de la marge basse dans le tableau.

Nous allons aggréger les données sur la colonnne is_genuine
Afficher la valeur moyenne de chaque variable pour les lignes False et True

```
if (!require(dplyr)) install.packages("dplyr")
```

```
Le chargement a nécessité le package : dplyr
```

```
Attachement du package : 'dplyr'
```

```
Les objets suivants sont masqués depuis 'package:stats':

    filter, lag

Les objets suivants sont masqués depuis 'package:base':

    intersect, setdiff, setequal, union
```

```r
library(dplyr)
```

```r
resultats <- data %>%
  group_by(is_genuine) %>%
  summarise(across(where(is.numeric), \(x) mean(x, na.rm = TRUE)))

print(resultats)
```

```
# A tibble: 2 x 7
  is_genuine diagonal height_left height_right margin_low margin_up length
  <chr>         <dbl>       <dbl>        <dbl>      <dbl>     <dbl>  <dbl>
1 False          172.        104.         104.       5.22      3.35   112.
2 True           172.        104.         104.       4.12      3.05   113.
```

```r
# Étape 1: Création du modèle de régression linéaire
# Modèle avec toutes les variables disponibles pour prédire margin_low
model <- lm(margin_low ~ diagonal + height_left + height_right + margin_up + length, data = d
```

```r
# Étape 2: Prédire les valeurs manquantes
# Créer une copie du dataframe avec les NA
data_na <- data[is.na(data$margin_low), ]
```

```r
data_na
```

```
     is_genuine diagonal height_left height_right margin_low margin_up length
73         True   171.94      103.89       103.45         NA      3.25 112.79
100        True   171.93      104.07       104.18         NA      3.14 113.08
152        True   172.07      103.80       104.38         NA      3.02 112.93
198        True   171.45      103.66       103.80         NA      3.62 113.27
242        True   171.83      104.14       104.06         NA      3.02 112.36
252        True   171.80      103.26       102.82         NA      2.95 113.22
```

```
285        True   171.92      103.83       103.76         NA      3.23 113.29
335        True   171.85      103.70       103.96         NA      3.00 113.36
411        True   172.56      103.72       103.51         NA      3.12 112.95
414        True   172.30      103.66       103.50         NA      3.16 112.95
446        True   172.34      104.42       103.22         NA      3.01 112.97
482        True   171.81      103.53       103.96         NA      2.71 113.99
506        True   172.01      103.97       104.05         NA      2.98 113.65
612        True   171.80      103.68       103.49         NA      3.30 112.84
655        True   171.97      103.69       103.54         NA      2.70 112.79
676        True   171.60      103.85       103.91         NA      2.56 113.27
711        True   172.03      103.97       103.86         NA      3.07 112.65
740        True   172.07      103.74       103.76         NA      3.09 112.41
743        True   172.14      104.06       103.96         NA      3.24 113.07
781        True   172.41      103.95       103.79         NA      3.13 113.41
799        True   171.96      103.84       103.62         NA      3.01 114.44
845        True   171.62      104.14       104.49         NA      2.99 113.35
846        True   172.02      104.21       104.05         NA      2.90 113.62
872        True   171.37      104.07       103.75         NA      3.07 113.27
896        True   171.81      103.68       103.80         NA      2.98 113.82
920        True   171.92      103.68       103.45         NA      2.58 113.68
946        True   172.09      103.74       103.52         NA      3.02 112.78
947        True   171.63      103.87       104.66         NA      3.27 112.68
982        True   172.02      104.23       103.72         NA      2.99 113.37
1077      False   171.57      104.27       104.44         NA      3.21 111.87
1122      False   171.40      104.38       104.19         NA      3.17 112.39
1177      False   171.59      104.05       103.94         NA      3.02 111.29
1304      False   172.17      104.49       103.76         NA      2.93 111.21
1316      False   172.08      104.15       104.17         NA      3.40 112.29
1348      False   171.72      104.46       104.12         NA      3.61 110.31
1436      False   172.66      104.33       104.41         NA      3.56 111.47
1439      False   171.90      104.28       104.29         NA      3.24 111.49
```

```r
# Prédire les valeurs manquantes
predicted_values <- predict(model, newdata = data_na)
```

```r
predicted_values
```

```
      73       100       152       198       242       252       285       335 
4.318525 4.393668 4.410457 4.319014 4.650617 3.803308 4.179736 4.127442 
     411       414       446       482       506       612       655       676 
4.135034 4.160539 4.177420 3.768554 4.058764 4.298047 4.160607 4.094065 
     711       740       743       781       799       845       846       872 
```

```
4.439846 4.470650 4.341643 4.080414 3.614306 4.371811 4.093621 4.249629
     896      920      946      947      982     1077     1122     1177
3.893748 3.746333 4.237415 4.710533 4.137780 5.050277 4.802145 5.067584
    1304     1316     1348     1436     1439
5.047570 4.778967 5.726993 5.185862 5.140043
```

```
# Étape 3: Remplacer les NA par les valeurs prédites
data$margin_low[is.na(data$margin_low)] <- predicted_values
```

```
# Voir le résultat
summary(data$margin_low)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.980   4.020   4.310   4.483   4.870   6.900
```

```
skim(data)
```

Table 4: Data summary

Name	data
Number of rows	1500
Number of columns	7
_____	
Column type frequency:	
character	1
numeric	6
_____	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
is_genuine	0	1	4	5	0	2	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
diagonal	0	1	171.96	0.31	171.04	171.75	171.96	172.17	173.01	
height_left	0	1	104.03	0.30	103.14	103.82	104.04	104.23	104.88	
height_right	0	1	103.92	0.33	102.82	103.71	103.92	104.15	104.95	
margin_low	0	1	4.48	0.66	2.98	4.02	4.31	4.87	6.90	
margin_up	0	1	3.15	0.23	2.27	2.99	3.14	3.31	3.91	
length	0	1	112.68	0.87	109.49	112.03	112.96	113.34	114.44	