Service Data

DÉTECTION DE FAUX BILLETS

ONCFM

D. GESSER
Service Data

DÉTECTION DE FAUX BILLETS

Introduction

Contexte

- **Situation**
  - Contrat pour l'ONCFM, ONG de lute contre la ***fausse monnaie***
  - Mise en place de ***moyens de detection de contrefaçons***
- **Mission**
  - Mise en place d'un ***modèle*** de detection de faux billets
- **Moyens**
  - Cahier des charges
  - Données d'entrées

Objectifs

- **Créer un programme de détection**
  - Définir si un billet est vrai ou faux
    - En fonction des ***caractéristiques géométriques*** du billet
- **Explorer les données**
  - Comprendre les données fournies
- **Préparer les données**
  - Nettoyer et preparer le jeu de données
- **Comparer et choisir le Meilleur modèle**
  - Générer ***différents*** modeles
  - ***Comparer*** ces modèles
  - ***Sélection*** du meilleur modèle
  - Mise en situation

ONCFM

D. GESSER
Service Data

Service Data

EXPLORATION / PRÉPARATION DES DONNÉES

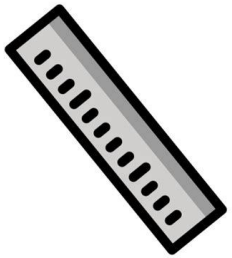
ONCFM

D. GESSER
Service Data

DÉTECTION DE FAUX BILLETS

Exploration des données

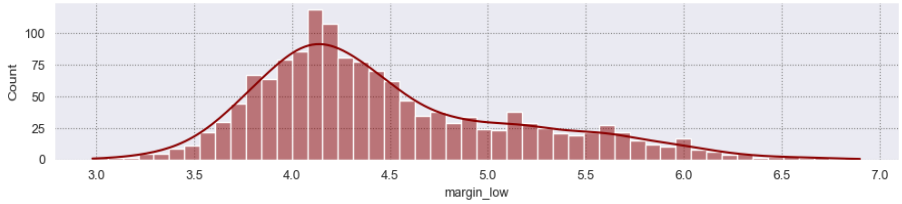
Données fournies



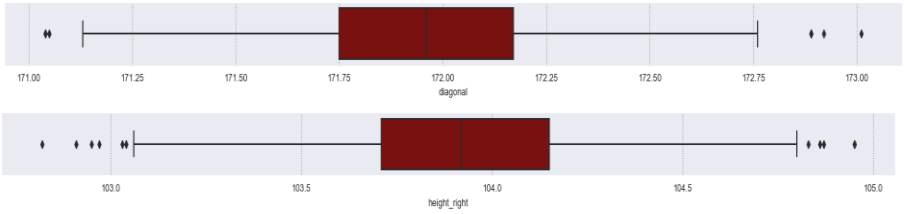
- **Déscription des données**
  - Taille du dataset
    - ***1500*** entrées
    - ***7*** variables
    - ***1000*** Billets vrais
    - ***500*** Billets faux
    - ***37*** données manquantes
  - Type de données
    - Dimensions des billets en mm (float)
    - Type : Vrai / Faux (bool)

Premières exploration

- **Distribution des données par variables**



- **Outliers**
  - Quelques outliers visibles (boxplot)
  - Causes possibles
    - *Etat des billets*
    - *Impression*
    - ...



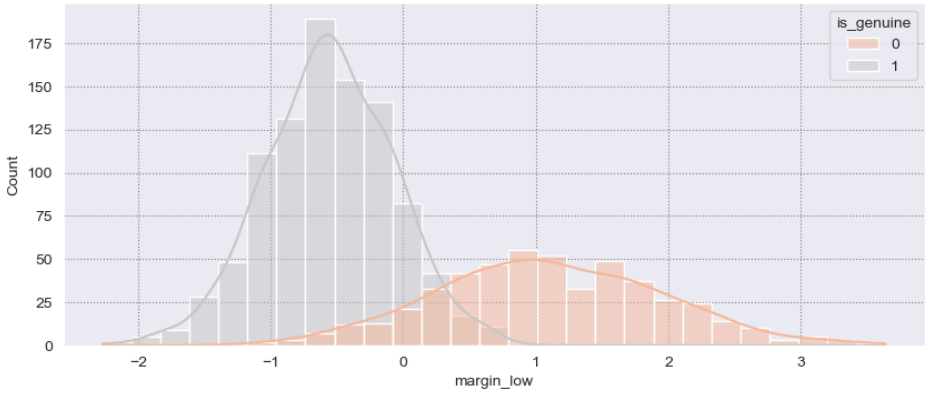
ONCFM
D. GESSER
Service Data

DÉTECTION DE FAUX BILLETS

Exploration des données

Données manquantes

- **Informations**
  - Concerne ***37*** billets
    - ***29 vrais***
    - ***8 faux***
- **Importance de la variable**
  - Pairplot avec difference entre les billets
    - margin_low pourrait aider à différencier les billets



Conclusions

- **Pourcentage concerné**
  - Concerne ***2,46%*** du dataset (nb. de lignes)
- **Pistes possibles**
  - Suppression de ces lignes
  - Imputation par la moyenne ou la médiane
  - ***Régression linéaire***
  - …
- **Solution retenue**
  - ***Régression linéaire***

ONCFM

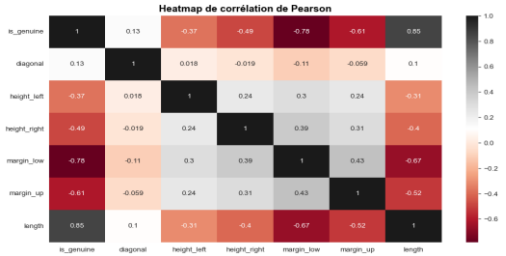
D. GESSER
Service Data

DÉTECTION DE FAUX BILLETS

Préparation des données

Régression Linéaire Multiple

- **Préparation**
  - Changement de True/False en 0/1
  - Standardisation des données
- **Données explicatives**
  - Heatmap pour une première idée
  - Fonction backward selection
    - Prise en compte de toutes les variables explicatives
    - Recherche des variables minimisant l'écart entre la droite de régression et la target
    - Elimination des variables p_value > 0,05
  - Variables retenues
    - ***is_genuine***
    - ***margin_up***

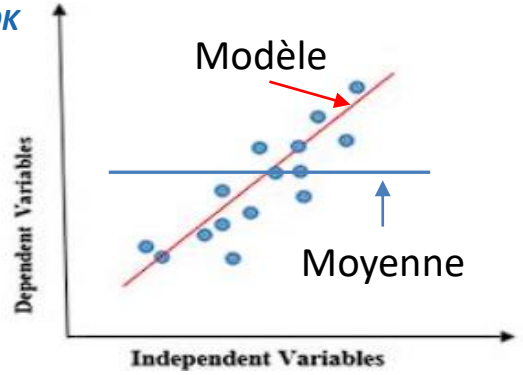


Evaluation du modèle

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

test set — predicted vaue — actual value

- **Evaluation**
  - MSE : ***0,375***
  - R^2 (variance modèle / variance données) : ***61,8 %***
  - Normalité des résidus : ***NOK***
    - QQ-plot
    - Shapiro : ***p_value < 0,05***
  - Homscédasticité des résidus (variance constante) : ***NOK***
    - Test de Levène : ***p_value < 0,05***
  - Multicolinéarité : ***OK***
    - VIF



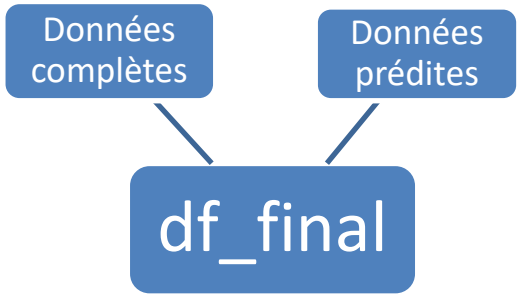
ONCFM
D. GESSER
Service Data

DÉTECTION DE FAUX BILLETS

Imputation des valeurs manquantes

Régression Linéaire Multiple

- **Application**
  - ***Validation*** du modèle
  - ***Utilisation*** du modèle sur le dataset incomplet
- **Regroupement des données**
  - Concatenation du dataset avec nouvelles données et celui des billets complets
  - Obtention du ***dataset final*** pour la suite du projet



Dataset final

- **Infos sur le dataset**
  - Comparaison entre le dataset de base et final
  - Pas de difference notable

```
]: # df complet avec données prédites
   df_final_clean['margin_low'].describe().round(4)

]: count    1500.0000
   mean       -0.0048
   std         0.9943
   min        -2.2694
   25%        -0.6871
   50%        -0.2652
   75%         0.5787
   max         3.6379
   Name: margin_low, dtype: float64

]: # df sans les données manquantes
   df_scaled_not_nan['margin_low'].describe().round(4)

]: count    1463.0000
   mean        0.0000
   std         1.0003
   min        -2.2694
   25%        -0.7097
   50%        -0.2652
   75%         0.5787
   max         3.6379
   Name: margin_low, dtype: float64
```

Service Data

MODÈLE DE PRÉDICTION

ONCFM

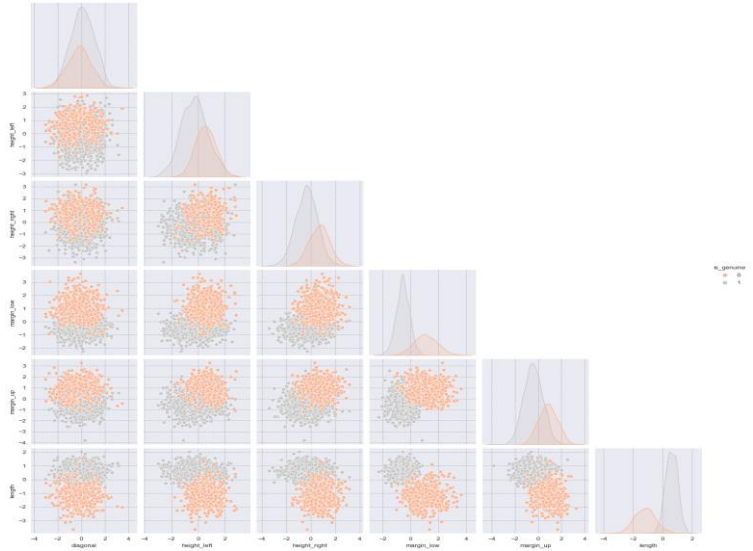
D. GESSER
Service Data

DÉTECTION DE FAUX BILLETS

Exploration des données

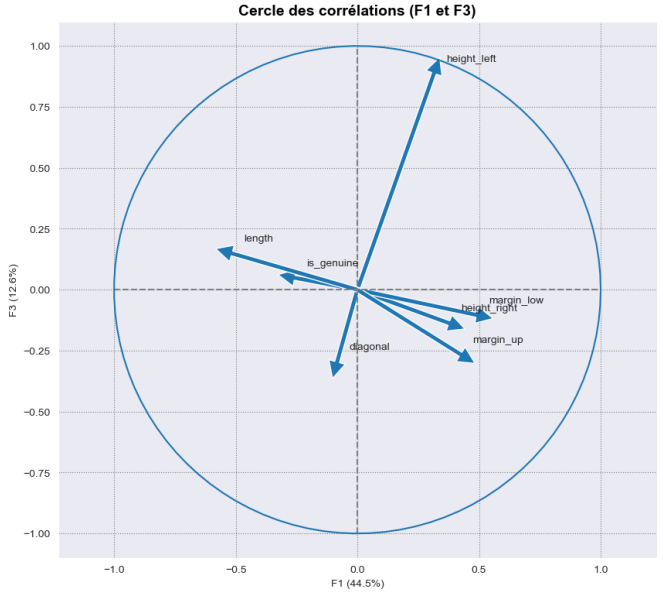
Recherche de pistes

- **Corrélations**
  - Pairplot avec separation du type
  - Heatmap
  - ACP



Analyse Composantes Principales

- **Cercle de corrélation**
  - Le type de billet est fortement ***corrélé*** à length
  - Le type est fortement ***anticorrélé*** à margin_up, low et height_right



ONCFM

D. GESSER
Service Data

DÉTECTION DE FAUX BILLETS

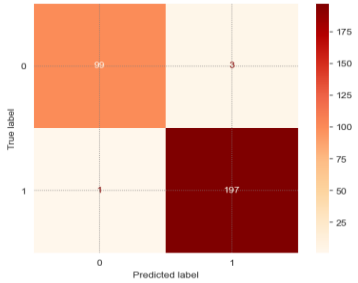
Modèle de prédiction – Régression logistique

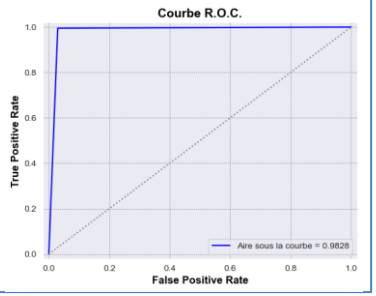
Création du modèle

- **Données explicatives**
  - Fonction backward
  - Prise en compte de toutes les variables explicatives
    - Itération
    - Elimination des variables p_value > 0,05
  - Variables retenues
    - *length*
    - *margin_up*
    - *margin_low*
    - *height_right*
- **Entrainement du modèle**
- **Sauvegarde du modèle**
  - Utilisation de JobLib

Evaluation du modèle

- **Scores**
  - Train/Test set
    - Entrainement du modèle sur Trainset
    - Evaluation du modèle sur Testset
- **Scores**
  - Score global : ***98,667 %***
  - Precision : ***98,5 %***
  - Recall : ***99,495 %***
  - F1 Score : ***98,995 %***
- **Matrice de confusion**
  - Réel vs Prédict
- **Courbe ROC**
  - Aire sous la courbe : ***0,9828***





ONCFM
D. GESSER
Service Data

DÉTECTION DE FAUX BILLETS

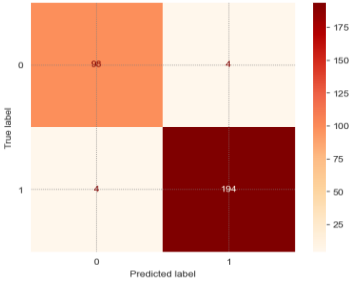
Modèle de prediction – KMeans

Création du modèle

- **Définition des centroïdes**
  - GroupBy sur is_genuine
  - Calcul des centroïdes
  - Paramètrage de Kmeans
    - Nombre de cluster
    - Centroïdes
- **Remarque**
  - Kmeans est un algorithme de clustering
- **Entrainement du modèle**
- **Sauvegarde du modèle**
  - Utilisation de JobLib

Evaluation du modèle

- **Scores**
  - Train/Test set
    - Entrainement du modèle sur Trainset
    - Evaluation du modèle sur Testset
- **Scores**
  - Score global : ***Pas disponible***
  - Precision : ***97,98 %***
  - Recall : ***97,98 %***
  - F1 Score : ***97,98 %***
- **Matrice de confusion**
  - Réel vs Prédict
- **Courbe ROC**
  - Aire sous la courbe : ***0,9703***



ONCFM

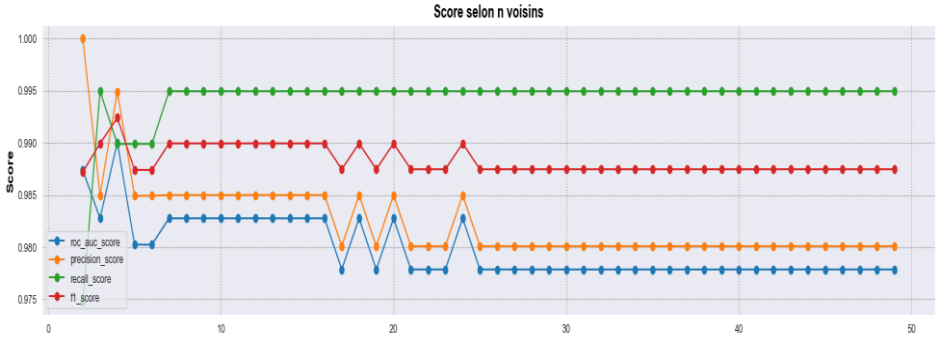
D. GESSER
Service Data

DÉTECTION DE FAUX BILLETS

Modèle de prediction – KNN

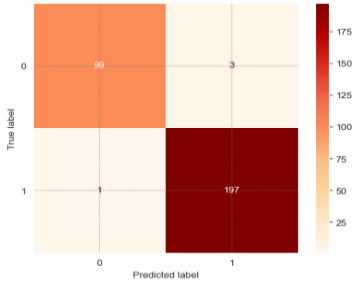
Création du modèle

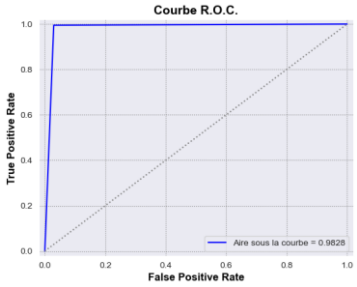
- **GridSearch**
  - Recherche des meilleurs paramètres
    - Possibilité de définir plusieurs paramètres
    - Récupération des meilleurs paramètres selon le score
- **Entrainement du modèle**
- **Sauvegarde du modèle**
  - Utilisation de JobLib



Evaluation du modèle

- **Scores**
  - Train/Test set
    - Entrainement du modèle sur Trainset
    - Evaluation du modèle sur Testset
- **Scores**
  - Score global : ***98,667 %***
  - Precision : ***98,5 %***
  - Recall : ***99,495 %***
  - F1 Score : ***98,995 %***
- **Matrice de confusion**
  - Réel vs Prédict
- **Courbe ROC**
  - Aire sous la courbe : ***0,9828***





ONCFM
D. GESSER
Service Data

DÉTECTION DE FAUX BILLETS

Modèle de détection – Comparaison - Sélection

Comparaison des modèles

- ***Régression logistique***
  - AUC ROC : ***98,277 %***
  - Score global : ***98,667 %***
  - F1 Score : ***98,995 %***
- ***KMeans***
  - AUC ROC : **97,29 %**
  - Score global : ***Pas disponible***
  - Accuracy *(equivalent score global)* : ***97,333 %***
  - F1 Score : ***97,98 %***
- ***KNN***
  - AUC ROC : ***98,277 %***
  - Score global : ***98,667 %***
  - F1 Score : ***98,995 %***

Sélection du modèle

- **Conclusions**
  - Régression logistique et KNN equivalent
  - Kmeans en retrait
  - Choix de la ***regression logistique***
    - Statistiques lors des prediction
    - Ressources système

ONCFM

D. GESSER
Service Data

Service Data

APPLICATION FINALE

ONCFM

D. GESSER
Service Data

DÉTECTION DE FAUX BILLETS

Merci pour votre attention !

ONCFM
D. GESSER
Service Data