CENSUS DISPLAY MODELLING

PRÉPARÉ PAR:

EZZOUINE AMINA

DAOUAJI SOUKAINA

SIDIBE MOUSSA

ENCADRÉ PAR:

PR MOUSSANIF AHMED

Plan de la présentation :

1. Exploration de la base de données.
2. Description des modèles.
3. Comparaison entre les modèles sur les différentes datasets.
4. Conclusion.

1. Exploration de la base de données

cor_sales_in_vol : ventes corrigées en volume

cor_sales_in_val : ventes corrigées en valeur

turnover : chiffre d'affaires du magasin

value : les valeurs

enseigne

VenteConv : les ventes converties

Feature : Enseigne ayant reçu un prospectus

- Notre base de données est constituée de 200.737 observations et 8 variables.
- Parmi les 8 variables on a 5 variables quantitatives (float et int) et 3 variables qualitatives.
- Notre base de données ne contient aucune valeur manquante.
- Notre variable dépendante est Display, elle prend la valeur 0 ou 1.

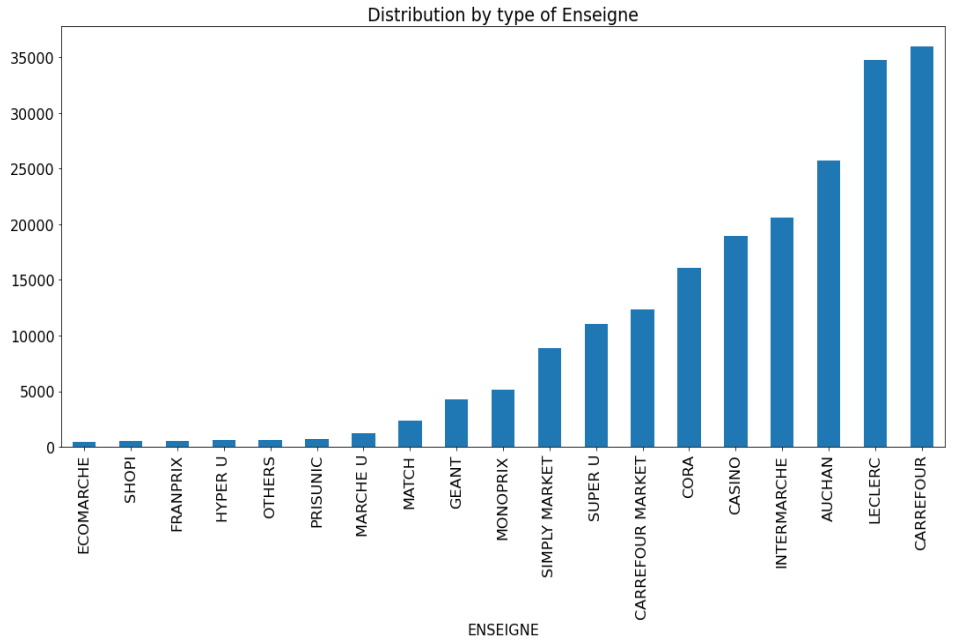
	Display	cor_sales_in_vol	cor_sales_in_val	turnover	value	ENSEIGNE	VenteConv	Feature
0	No_Displ	11.0	35.21	19622	8	MONOPRIX	88.0	No_Feat
1	No_Displ	3.0	13.32	19622	12	MONOPRIX	36.0	No_Feat
2	No_Displ	6.0	8.10	19622	5	MONOPRIX	30.0	No_Feat
3	No_Displ	13.0	19.35	19622	8	MONOPRIX	104.0	No_Feat
4	No_Displ	13.0	90.09	19622	28	MONOPRIX	364.0	Feat

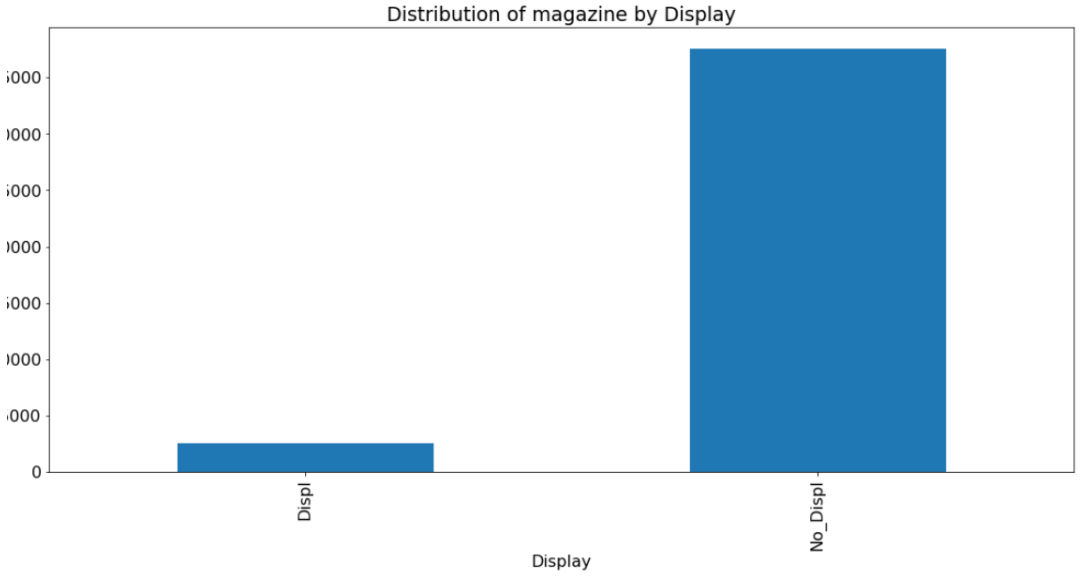
```
CDM_data.shape
```

```
(200737, 8)
```

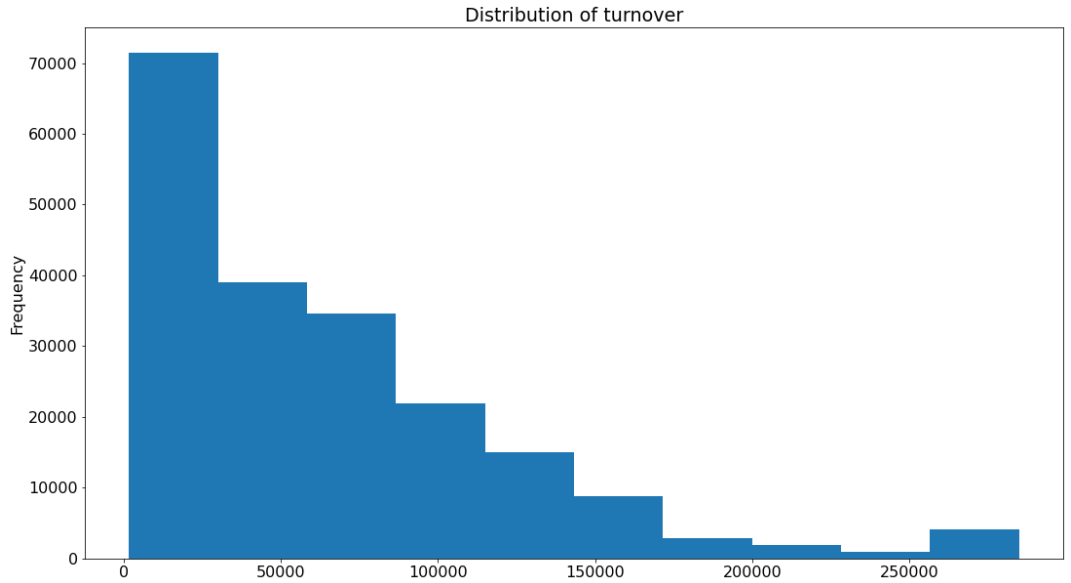
```
CDM_data.info()
```

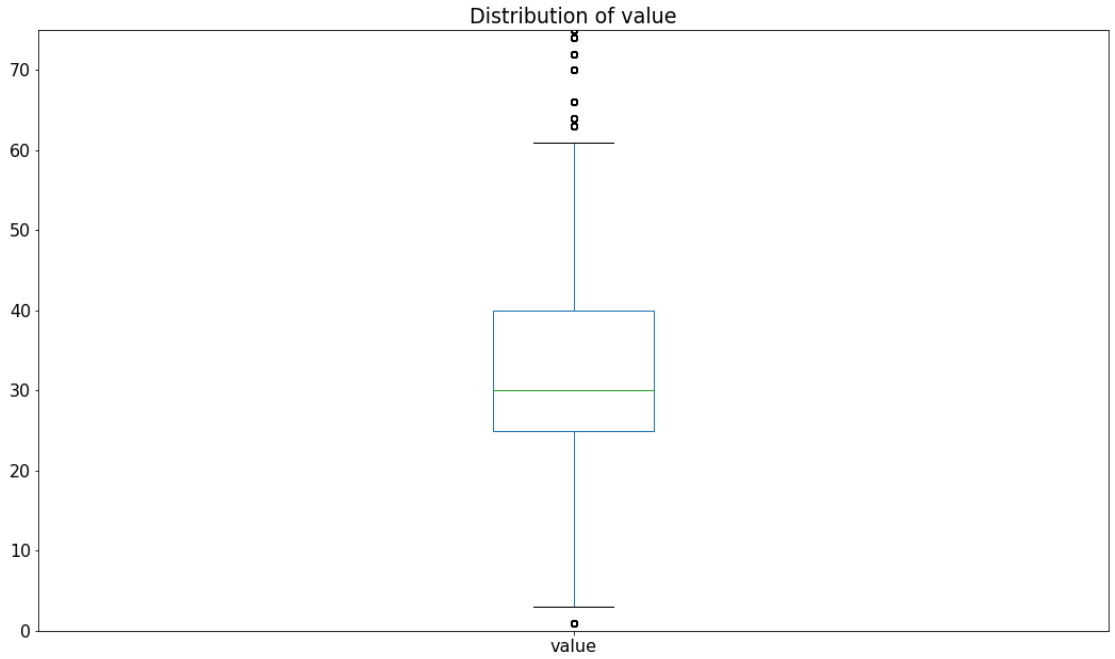
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200737 entries, 0 to 200736
Data columns (total 8 columns):
 #   Column            Non-Null Count   Dtype
---  ------            --------------   -----
 0   Display           200737 non-null  object
 1   cor_sales_in_vol  200737 non-null  float64
 2   cor_sales_in_val  200737 non-null  float64
 3   turnover          200737 non-null  int64
 4   value             200737 non-null  int64
 5   ENSEIGNE          200737 non-null  object
 6   VenteConv         200737 non-null  float64
 7   Feature           200737 non-null  object
dtypes: float64(3), int64(2), object(3)
memory usage: 12.3+ MB
```





- Nous voyons une disparité ou sous representation des enseigne ayant un display et d'autre qui n'ont pas. Cette irrégularité peut influencer le modèle vers la classe surrepresenté en apprenant davantage de caractéristique pour cette classe contrairement au sous représenté.
- On observe la même phénomène pour les modalités de la variable Enseigne.





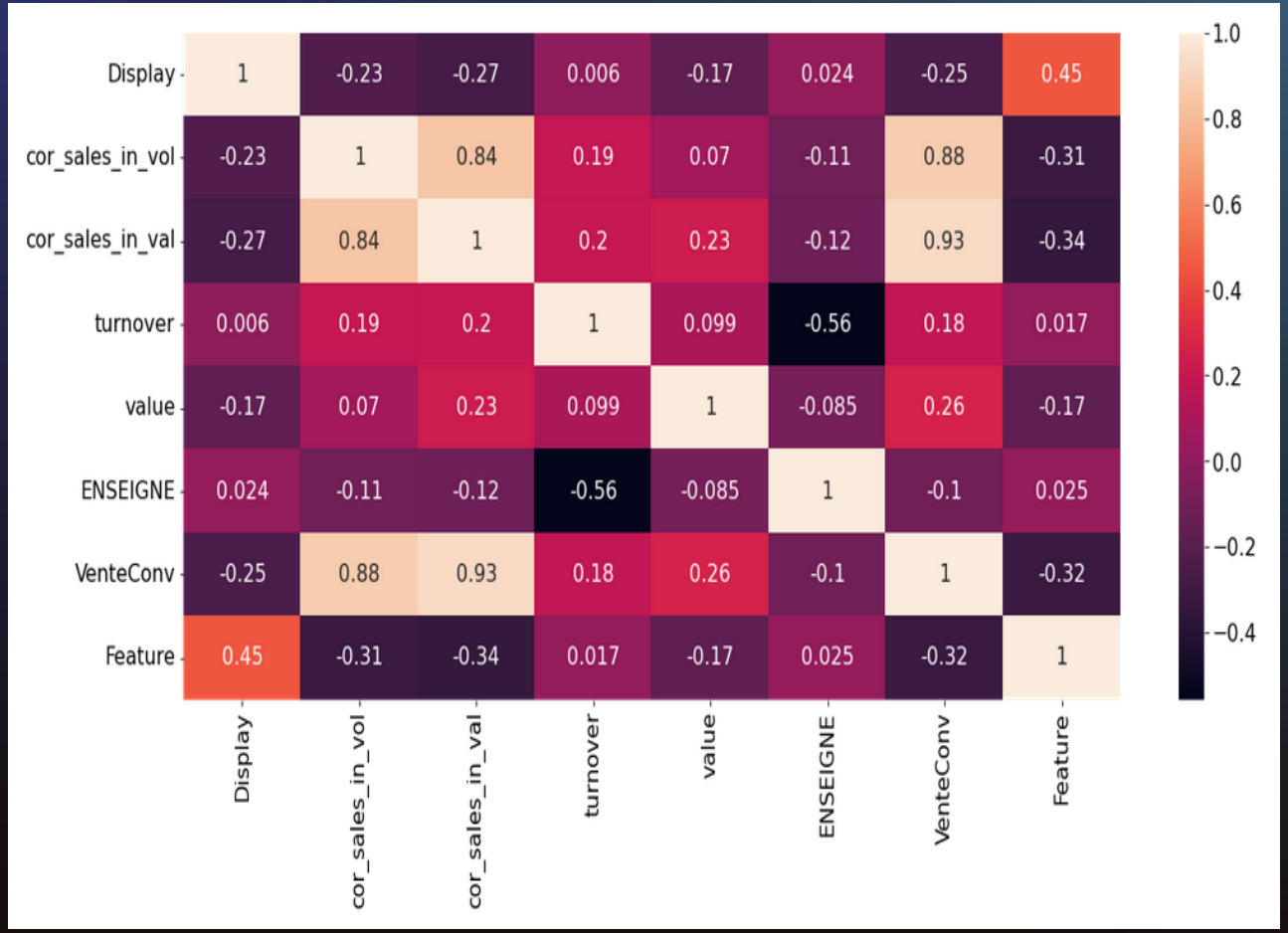
- Avec la distribution du chiffre d'affaires des enseigne, nous pouvons remarquons que l'écrasante majorité de cette liste sont des enseigne avec faible revenue.
- Pour ce qui concerne la variable value, sa Moyenne tourney autour de 30 mais présente beaucoup de valeur aberrante.

On remarque qu'il y a :

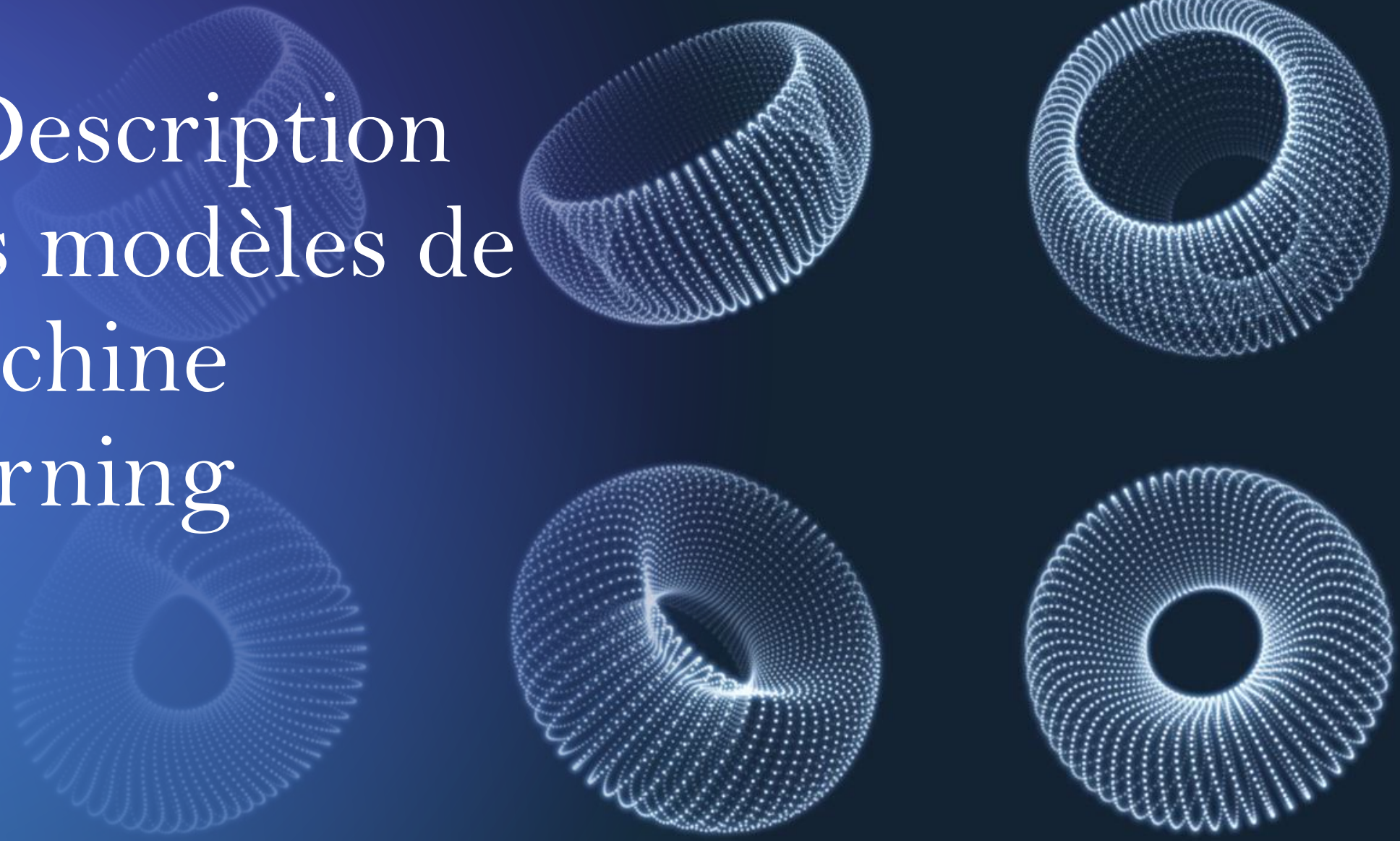
- Une forte corrélation entre les ventes corrigées en volume et les ventes converties ainsi que les ventes corrigées en valeur .

- Une forte corrélation entre les ventes corrigées en valeur et les ventes converties ainsi que les ventes corrigées en volume .

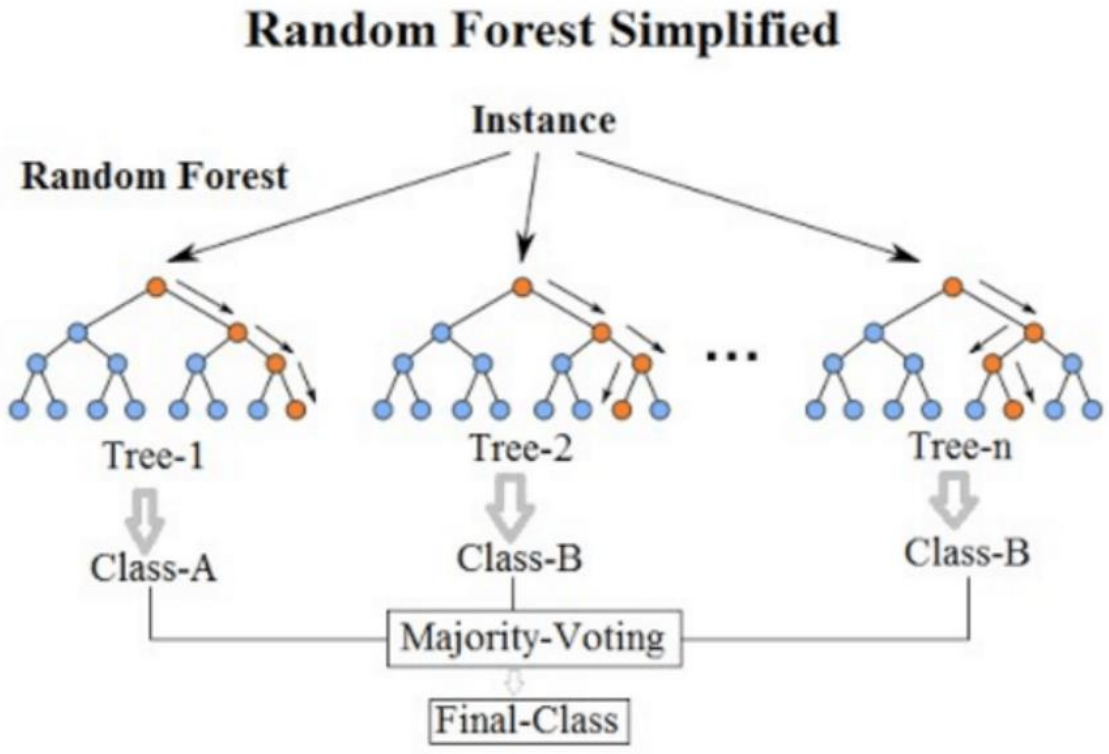
- Une forte corrélation entre les ventes converties et les ventes corrigées en volume ainsi que les ventes corrigées en valeur .



2.Description des modèles de machine learning



Random forest

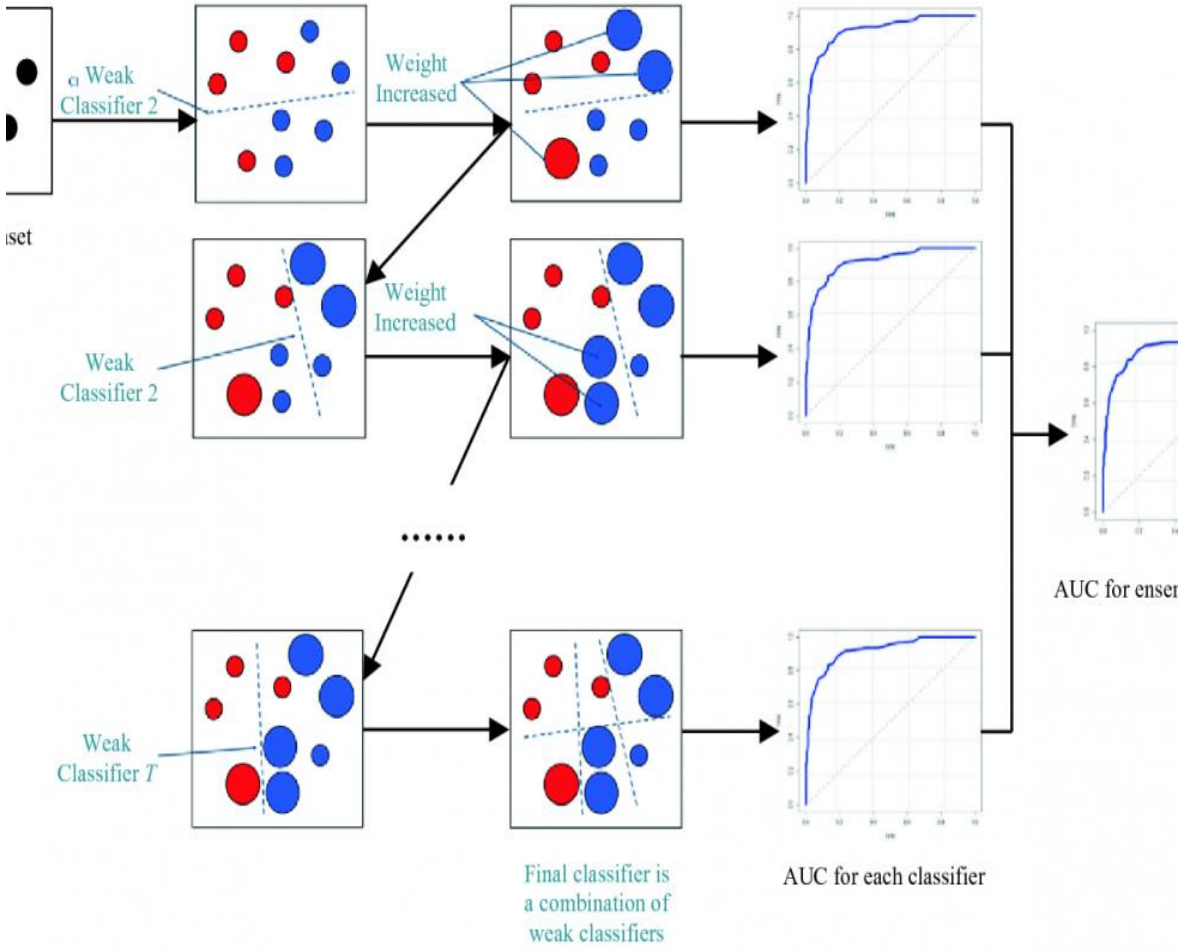


Il s'agit d'une méthode d'ensemble arbres de décision générés sur un ensemble de données réparties de façon aléatoire. Cet ensemble d'arbres de décision classificateurs est également connu sous le nom de forêt. Les arbres de décision individuels sont générés à l'aide d'un indicateur de sélection d'attribut tel que le gain d'information, le ratio de gain et l'indice de Gini pour chaque attribut. Chaque arbre dépend d'un échantillon aléatoire indépendant.

Dans un problème de classification, chaque arbre vote et la classe la plus populaire est choisie comme résultat final. Dans le cas de la régression, la moyenne de toutes les sorties des arbres est considérée comme le résultat final. Il est plus simple et plus puissant que les autres algorithmes de classification non linéaires.

Les paramètres :

- **N_estimators :** Le nombre d'arbres dans la forêt.
- **Criterion** : La fonction permettant de mesurer la qualité d'un fractionnement. Les critères pris en charge sont "gini" pour l'impureté de Gini et "entropie" pour le gain d'information.
- **Max_depth :** La profondeur maximale de l'arbre. Si None, alors les noeuds sont développés jusqu'à ce que toutes les feuilles soient pures ou jusqu'à ce que toutes les feuilles contiennent moins de min_samples_split échantillons.
- **max_features :** Le nombre de caractéristiques à prendre en compte lors de la recherche de la meilleure répartition
- **min_samples_leaf :** Le nombre minimum d'échantillons requis pour être à un noeud feuille. Un point de séparation à n'importe quelle profondeur ne sera pris en compte que s'il laisse au moins min_samples_leaf échantillons de formation dans chacune des branches gauche et droite. Cela peut avoir pour effet de lisser le modèle, en particulier dans la régression.
- **min_samples_split :** Le nombre minimum d'échantillons requis pour diviser un noeud interne



GradientBoostingClassifier

La méthode gradient boosting se base sur un ensemble des méthodes faibles pour construire un modèle avec un fort potentiel de prédiction. Mais avec la méthode de gradient nous il ajuste le poids de chaque observation selon son niveau de difficulté de classification.

Les paramètres :

- Learning rate
- Max depth
- Subsample
- ...

LDA :

- Extension de la régression dans le cas où la variable à expliquer est qualitative.

- Deux aspects :

- **Analyse discriminante à but descriptif :**

l'Analyse Factorielle Discriminante

- Objectif :

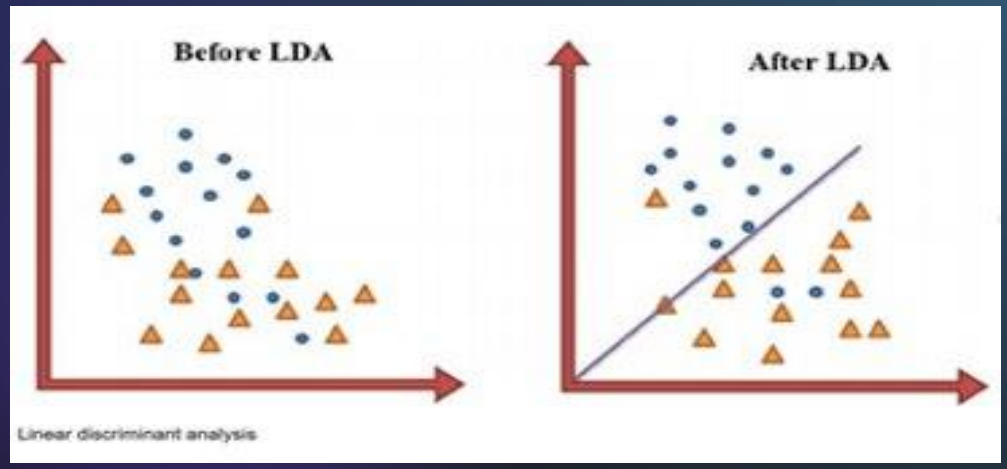
- déterminer les combinaisons linéaires de variables qui permettent de séparer au mieux les différentes classes,

- donner une représentation graphique.

- **Analyse discriminante à but décisionnel**

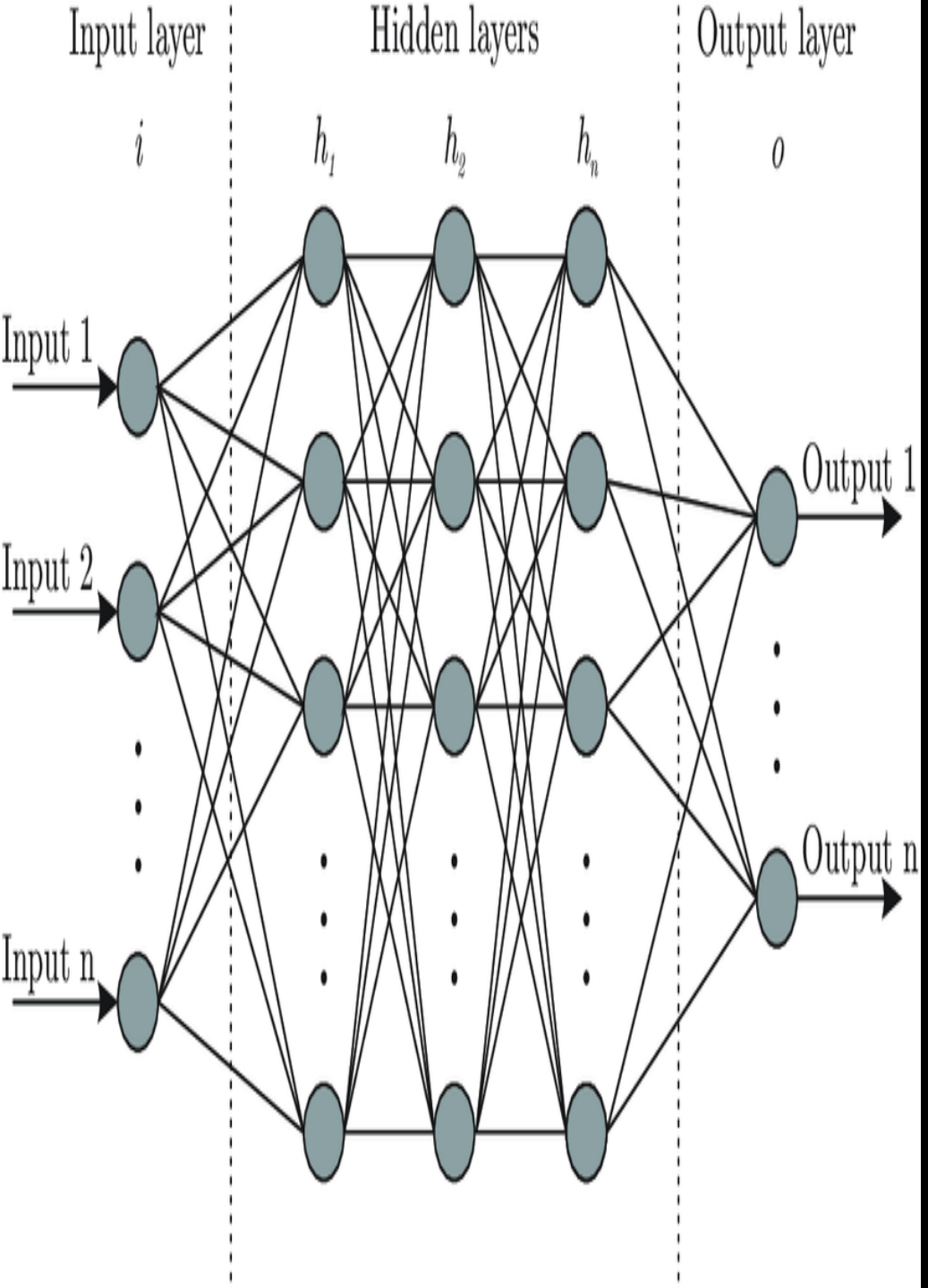
- Objectif :

- prédire la classe d'un nouvel objet décrit par la valeur de ces attributs.



Les paramètres :

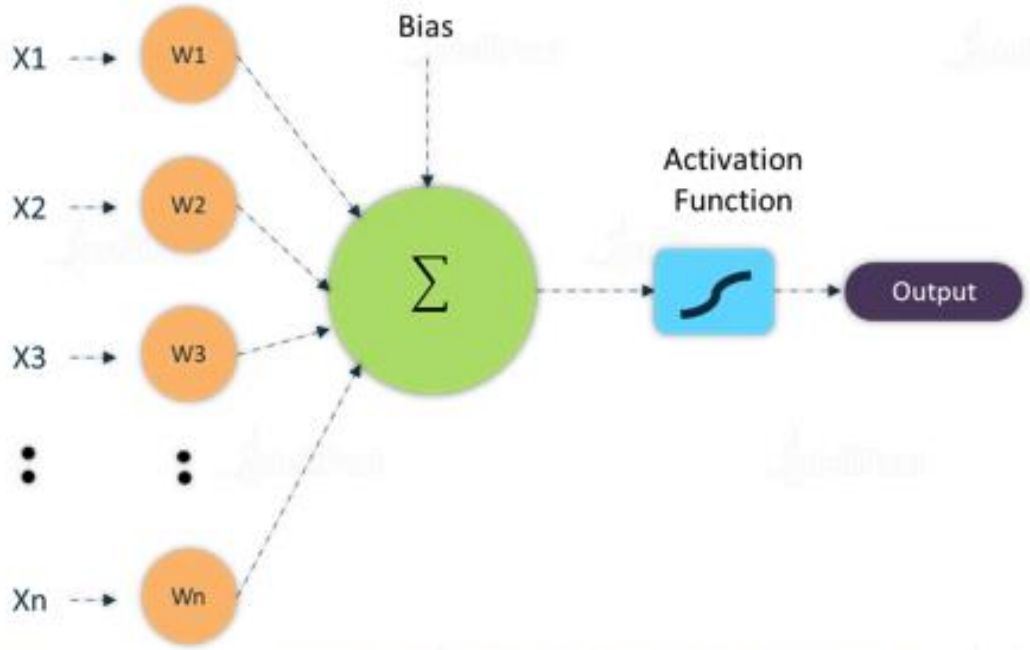
- Solver
- Shrinkage
- priors
- n_components
- store_covariance
- tol



Neural Networks

Les réseaux neuronaux reflètent le comportement du cerveau humain, permettant aux programmes informatiques de reconnaître les modèles et de résoudre les problèmes communs dans les domaines de l'IA, de l'apprentissage automatique et de l'apprentissage profond.

Les réseaux neuronaux comptent sur les données d'entraînement pour apprendre et améliorer leur précision au fil du temps. Une fois que ces algorithmes d'apprentissage sont mis au point avec précision, ils sont des outils puissants en informatique et en intelligence artificielle, ce qui nous permet de classer et de regrouper les données à une vitesse élevée.



Schematic Representation of a Neuron in a Neural Network

Les paramètres

- **Le poids :** transforme les données d'entrée dans les couches cachées du réseau. Plus le poids est grand, plus il impactera le réseau.
- **Le biais** : le biais est comme l'interception ajoutée dans une équation linéaire. C'est un paramètre supplémentaire dans le réseau neural qui est utilisé pour ajuster la sortie avec la somme pondérée des entrées au neurone. Par conséquent, le biais est une constante qui aide le modèle d'une manière qu'il peut s'adapter au mieux aux données.

- Ces paramètres vont évoluer durant l'ensemble du processus d'entrainement, lors de la backpropagation.

MPLClassifier

Le modèle MPLClassifier est une classe de Neural Networks permettant de faire un apprentissage sur la base du schéma de perceptron multicouches. Il a les mêmes caractéristiques que ce décrit dans le modèle construit à la main avec neural network ci-dessus.

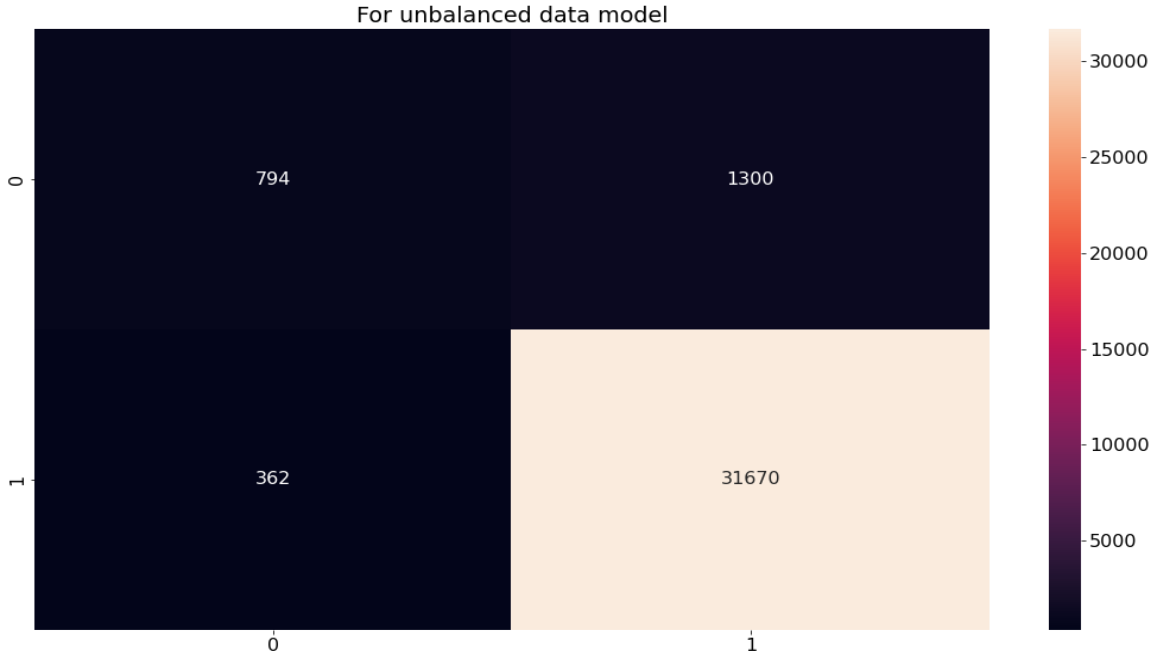
Elle comporte les paramètres suivants permettant de contrôler la performance du modèle:

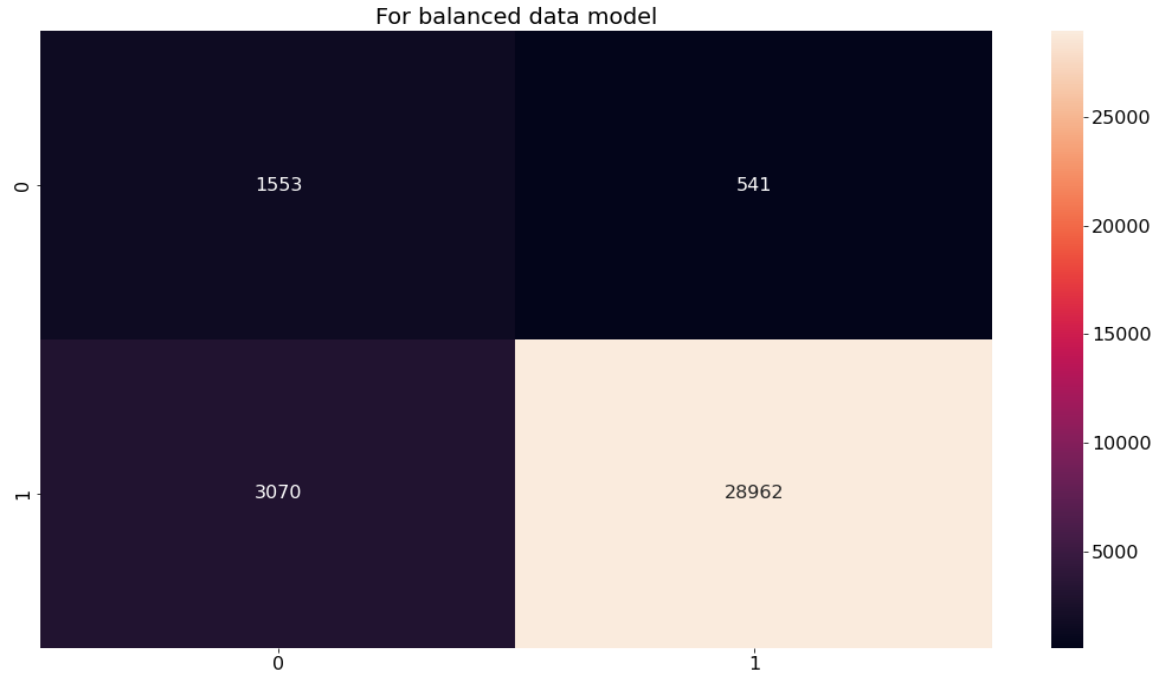
- Hidenn layer sizes
- Learning rate init
- Solver
- Alpha
- Max iteration
- activation

Performances des modèles sur les different datasets

Pour comparer la performance des modèles sur la data unbalanced et balanced par rapport à la variable cible Display, nous nous sommes intéressés aux métriques « matrice de confusion et pour plus détails classification_reports » en procédant comme suit:

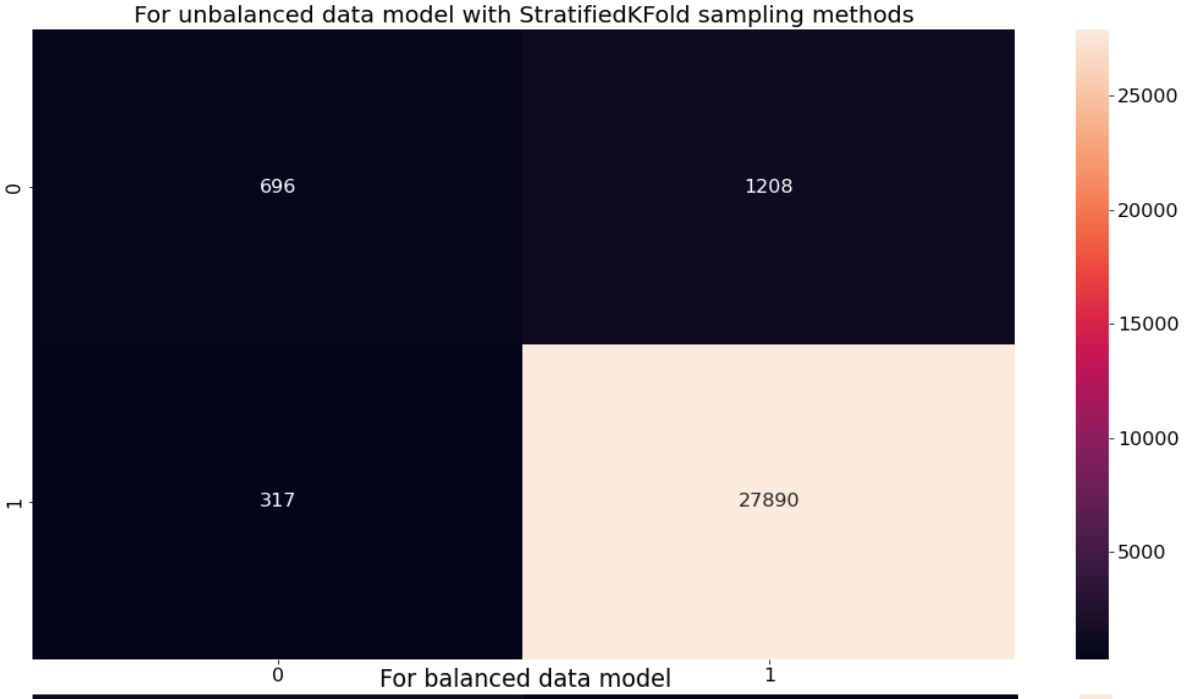
- Unbalanced data-based model vs balanced based model
- Stratified unbalanced data-based model vs balanced based model

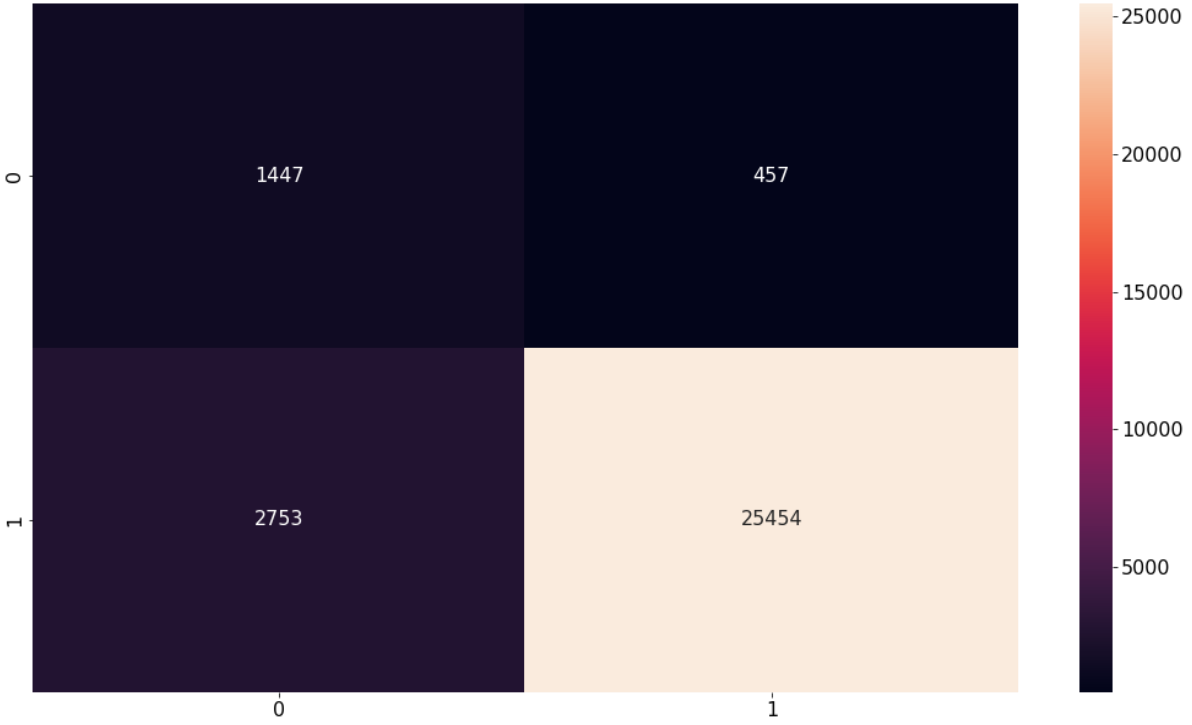




Random Forest

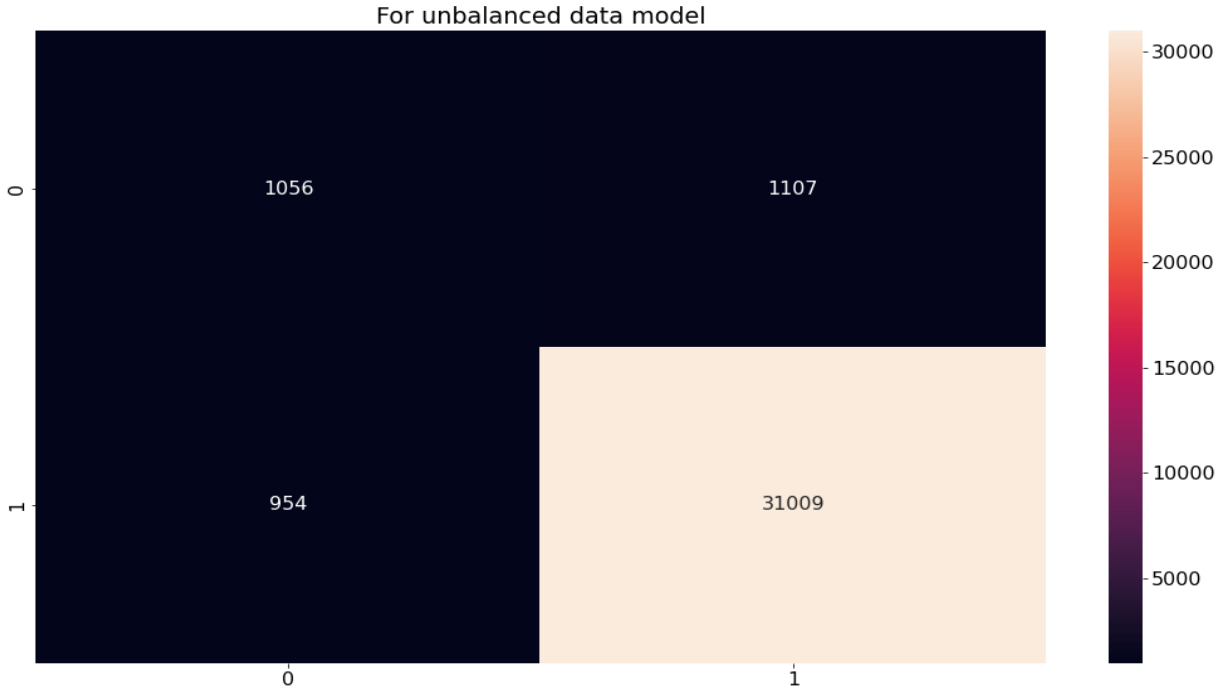
```
====================For unbalanced data====================
              precision    recall  f1-score   support

           0       0.69      0.38      0.49      2094
           1       0.96      0.99      0.97     32032

    accuracy                           0.95     34126
   macro avg       0.82      0.68      0.73     34126
weighted avg       0.94      0.95      0.94     34126

====================For balanced data====================
              precision    recall  f1-score   support

           0       0.34      0.74      0.46      2094
           1       0.98      0.90      0.94     32032

    accuracy                           0.89     34126
   macro avg       0.66      0.82      0.70     34126
weighted avg       0.94      0.89      0.91     34126
```

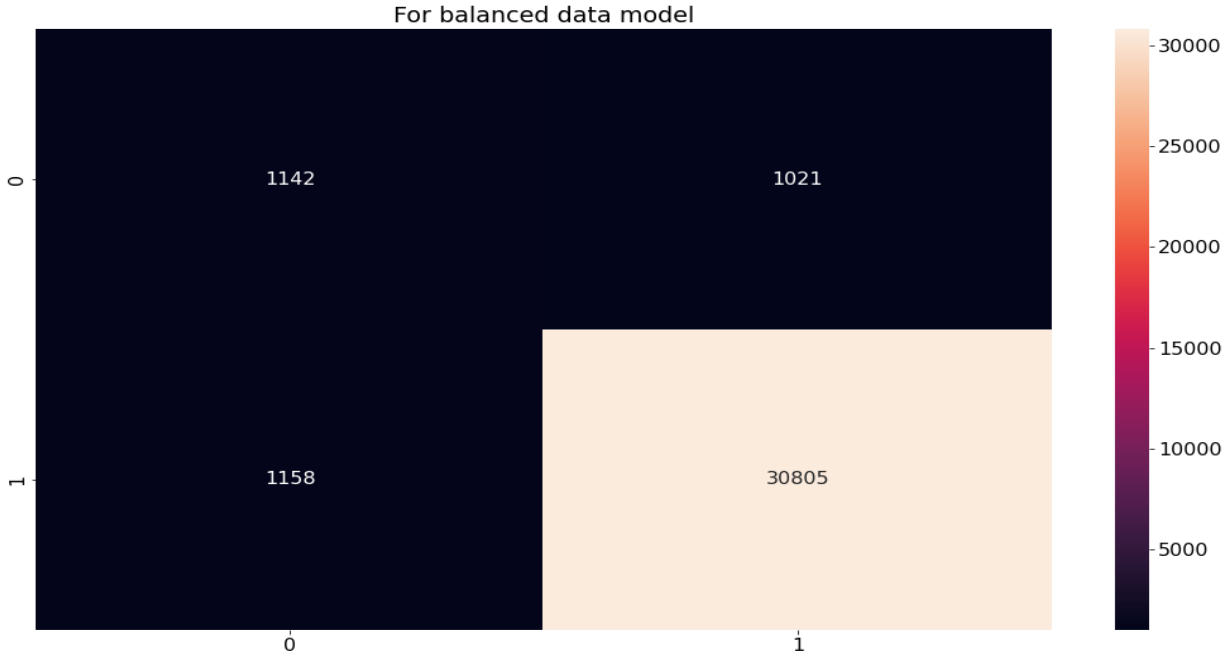




Random Forest

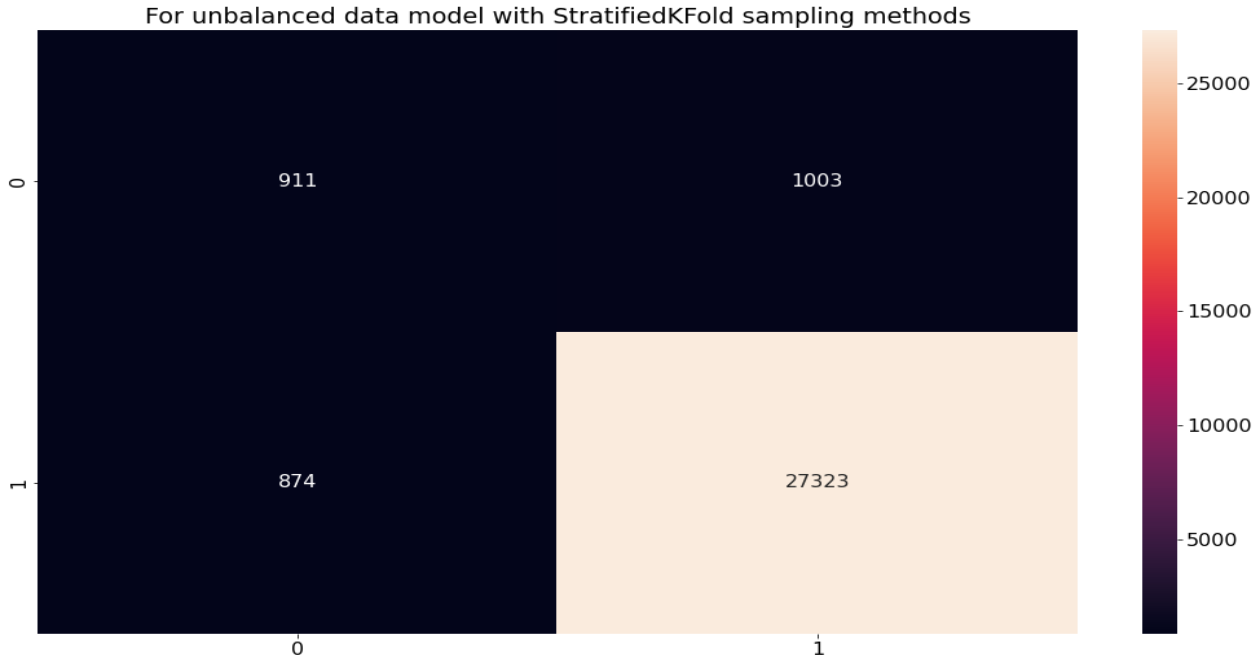
```
==========For unbalanced data model with StratiefKfold=====
              precision    recall  f1-score   support

           0       0.69      0.37      0.48      1904
           1       0.96      0.99      0.97     28207

    accuracy                           0.95     30111
   macro avg       0.82      0.68      0.73     30111
weighted avg       0.94      0.95      0.94     30111

====================For balanced data model================
              precision    recall  f1-score   support

           0       0.34      0.76      0.47      1904
           1       0.98      0.90      0.94     28207

    accuracy                           0.89     30111
   macro avg       0.66      0.83      0.71     30111
weighted avg       0.94      0.89      0.91     30111
```

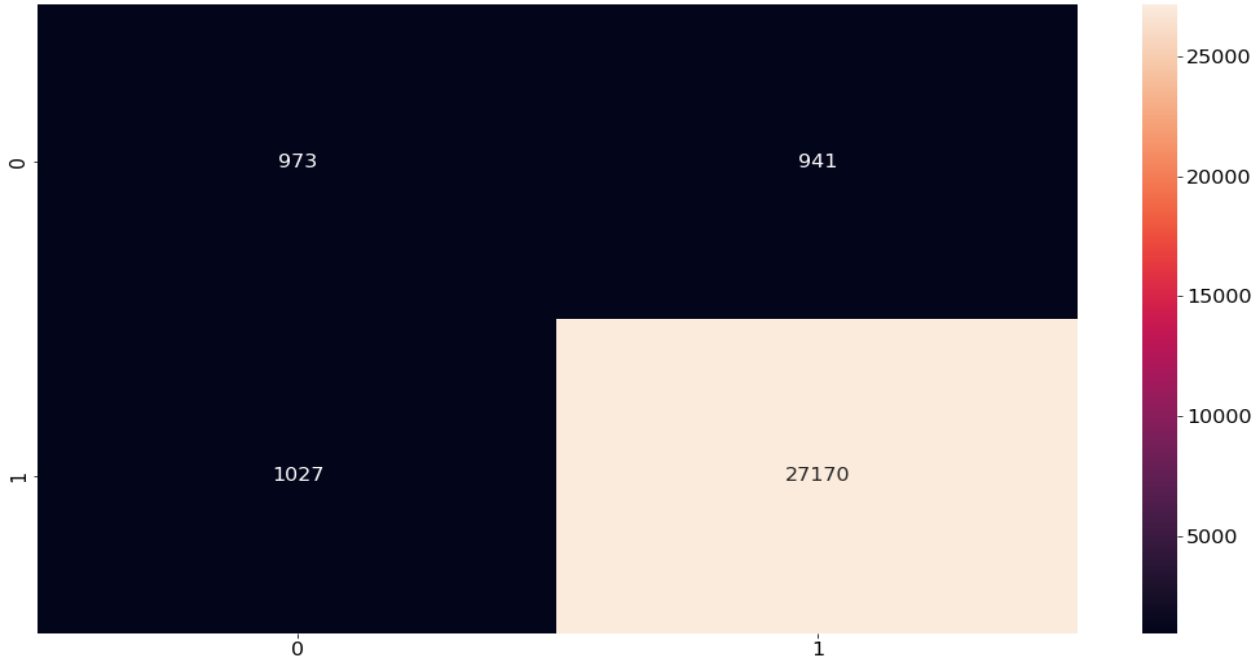




LDA

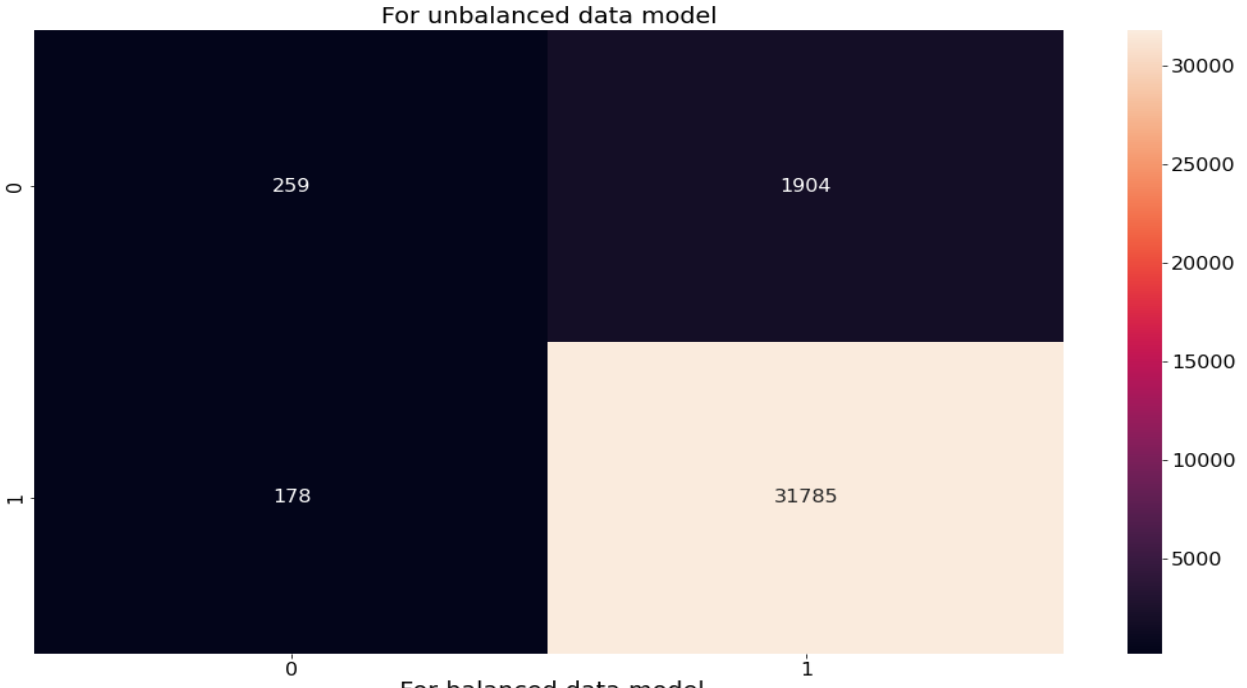
```
====================For unbalanced data====================
              precision    recall  f1-score   support

           0       0.53      0.49      0.51      2163
           1       0.97      0.97      0.97     31963

    accuracy                           0.94     34126
   macro avg       0.75      0.73      0.74     34126
weighted avg       0.94      0.94      0.94     34126

====================For balanced data======================
              precision    recall  f1-score   support

           0       0.50      0.53      0.51      2163
           1       0.97      0.96      0.97     31963

    accuracy                           0.94     34126
   macro avg       0.73      0.75      0.74     34126
weighted avg       0.94      0.94      0.94     34126
```

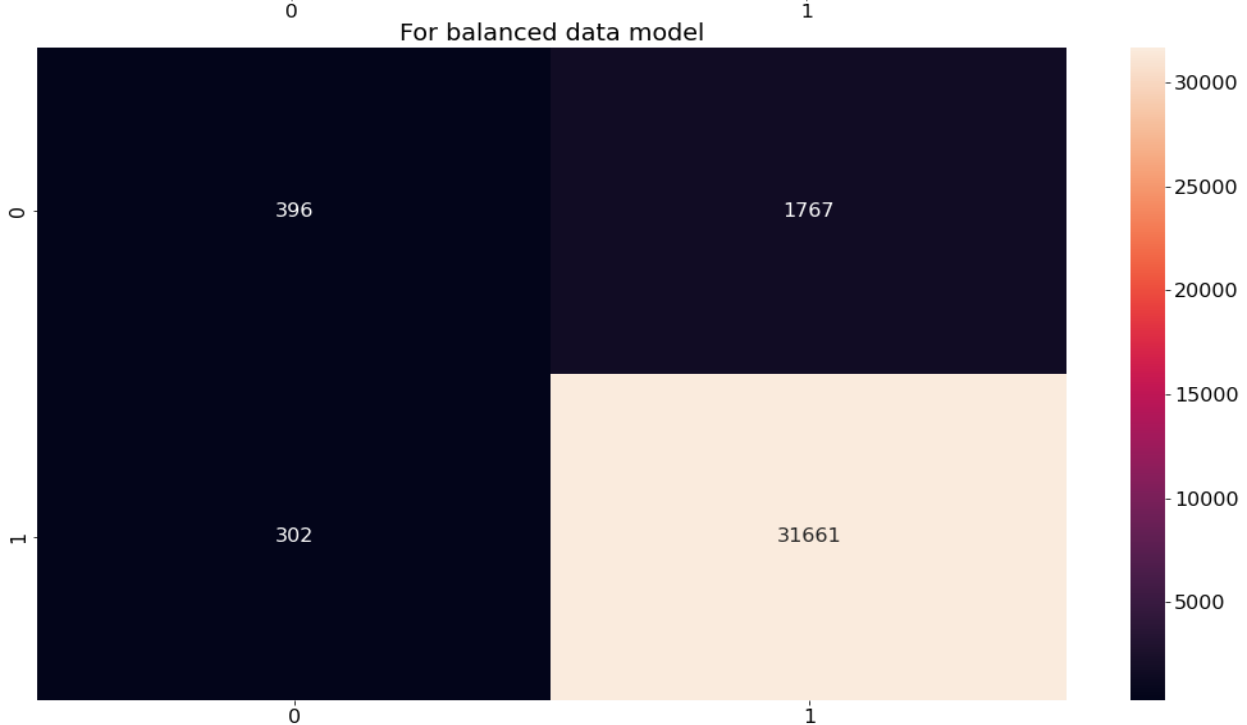




LDA

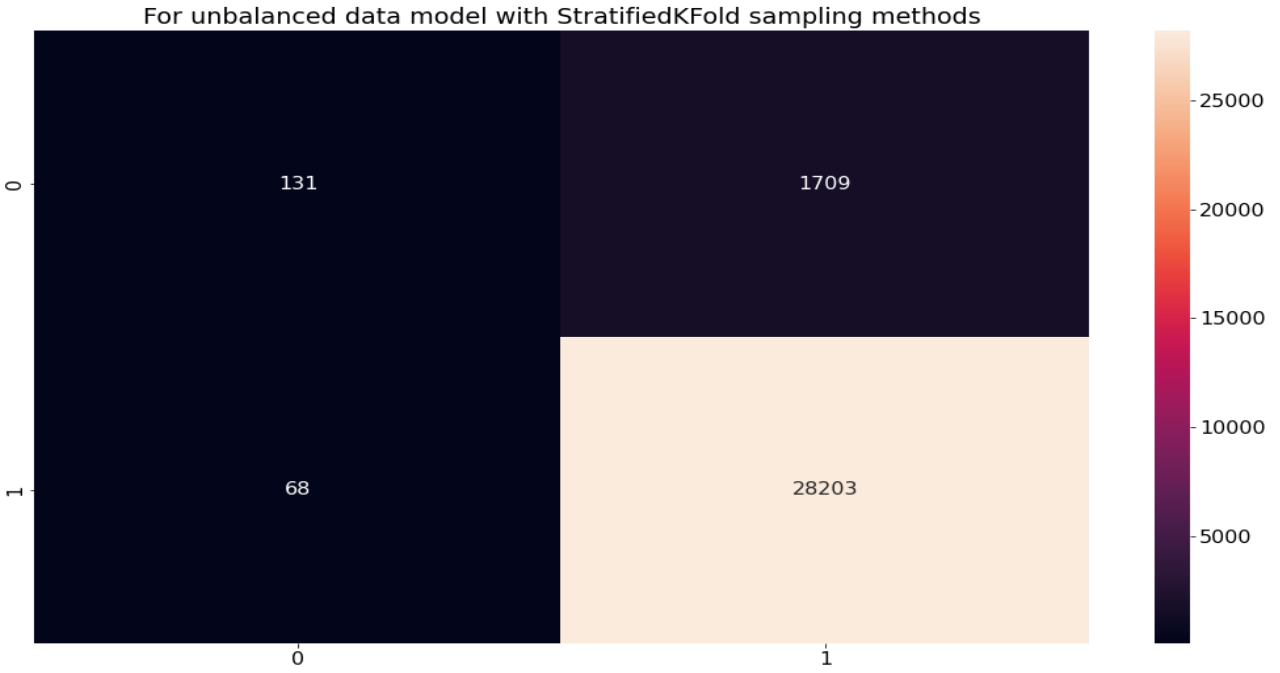
```
==========For unbalanced data model with StratiefKfold======
              precision    recall  f1-score   support

           0       0.51      0.48      0.49      1914
           1       0.96      0.97      0.97     28197

    accuracy                           0.94     30111
   macro avg       0.74      0.72      0.73     30111
weighted avg       0.94      0.94      0.94     30111

====================For balanced data model================
              precision    recall  f1-score   support

           0       0.49      0.51      0.50      1914
           1       0.97      0.96      0.97     28197

    accuracy                           0.93     30111
   macro avg       0.73      0.74      0.73     30111
weighted avg       0.94      0.93      0.94     30111
```

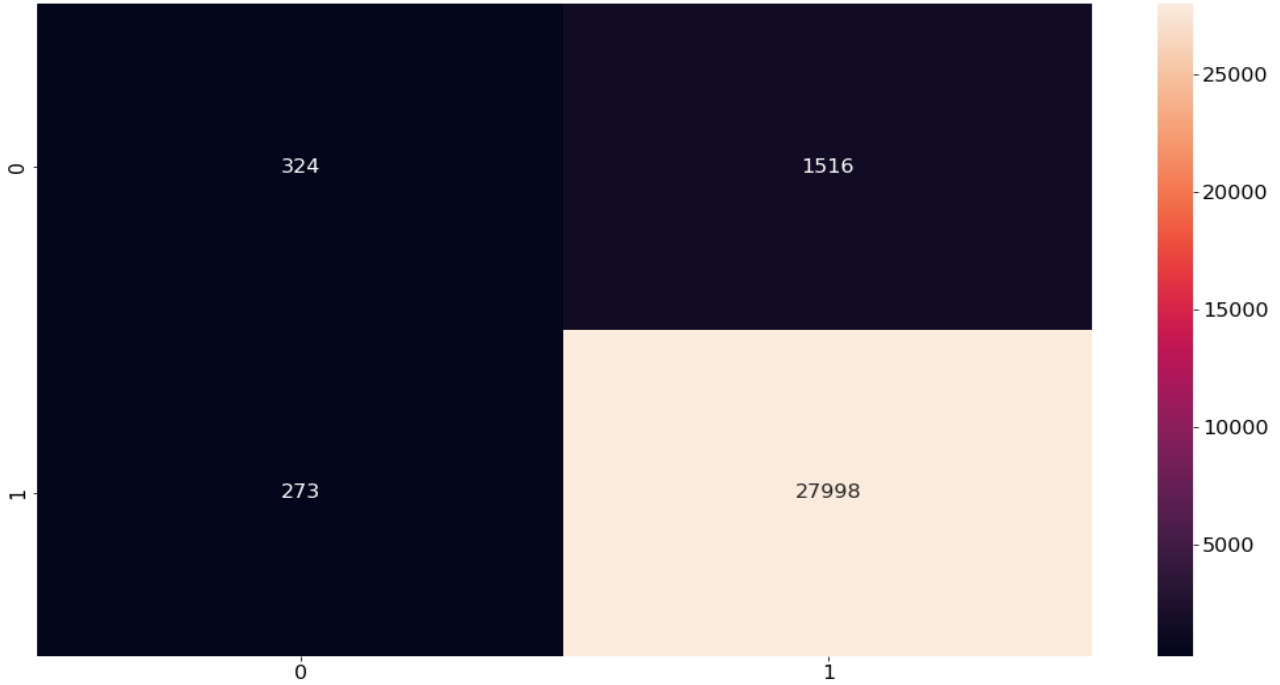




MLPClassifier

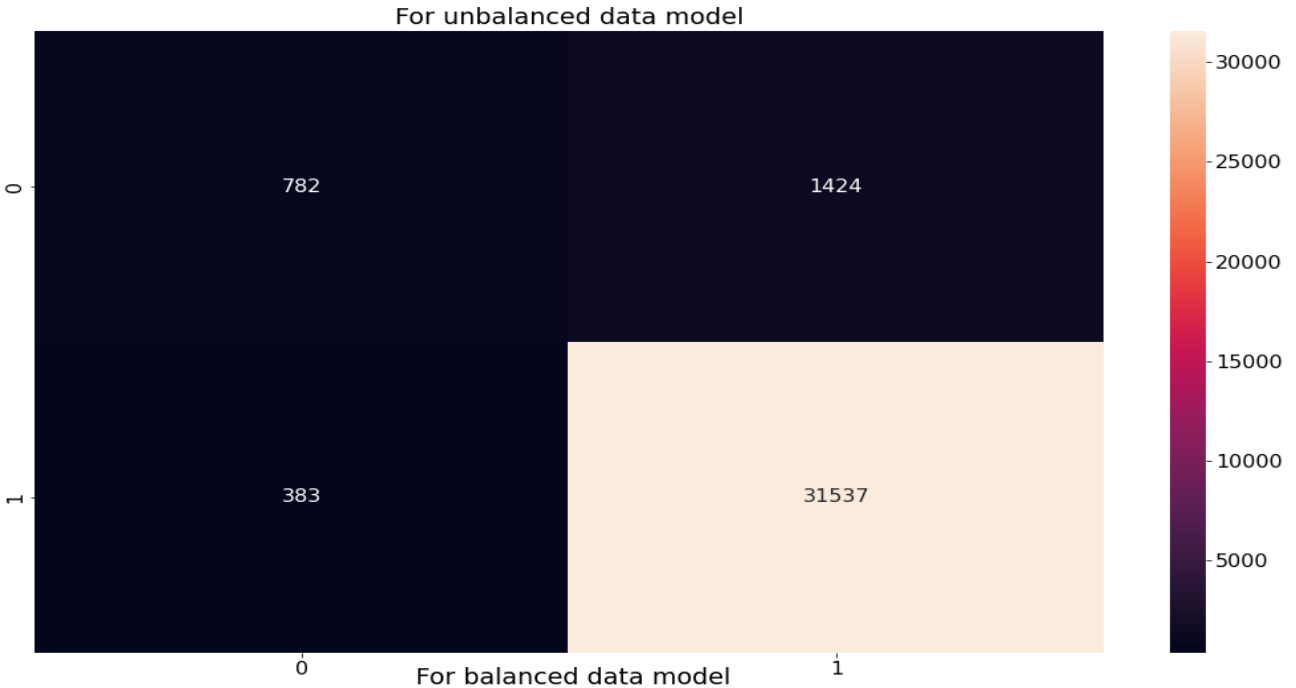
```
====================For unbalanced data==================
              precision    recall  f1-score   support

           0       0.59      0.12      0.20      2163
           1       0.94      0.99      0.97     31963

    accuracy                           0.94     34126
   macro avg       0.77      0.56      0.58     34126
weighted avg       0.92      0.94      0.92     34126

====================For balanced data====================
              precision    recall  f1-score   support

           0       0.57      0.18      0.28      2163
           1       0.95      0.99      0.97     31963

    accuracy                           0.94     34126
   macro avg       0.76      0.59      0.62     34126
weighted avg       0.92      0.94      0.92     34126
```

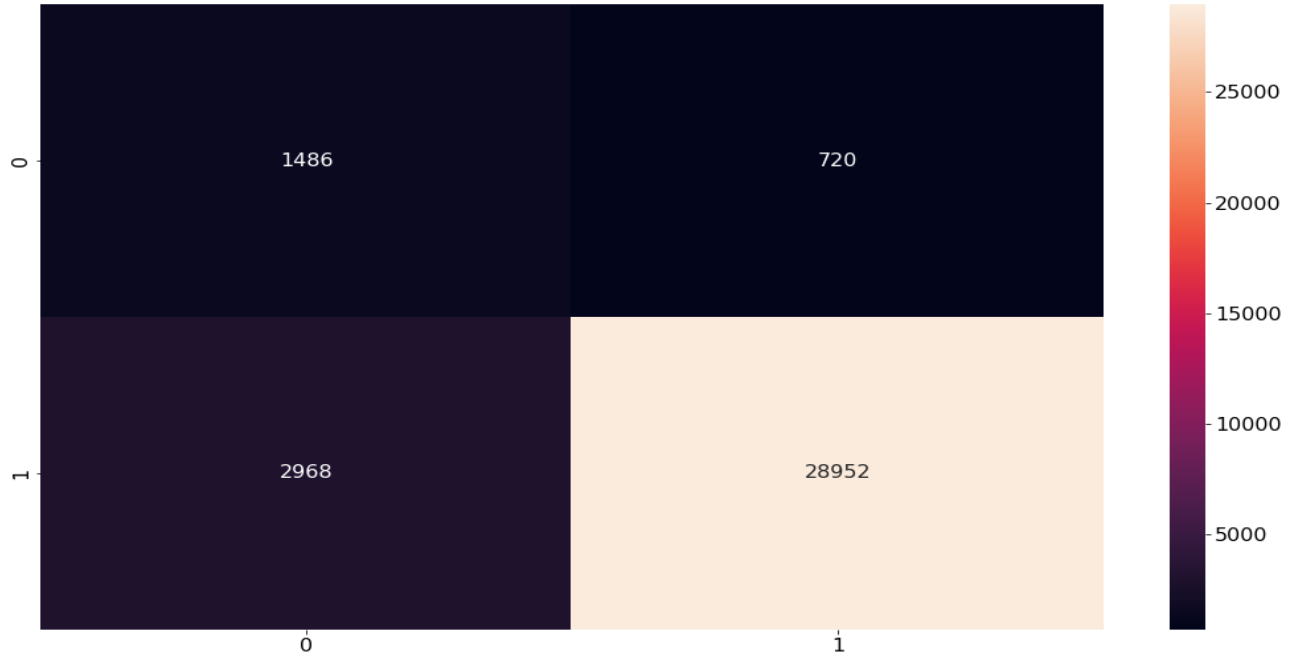




MLPClassifier

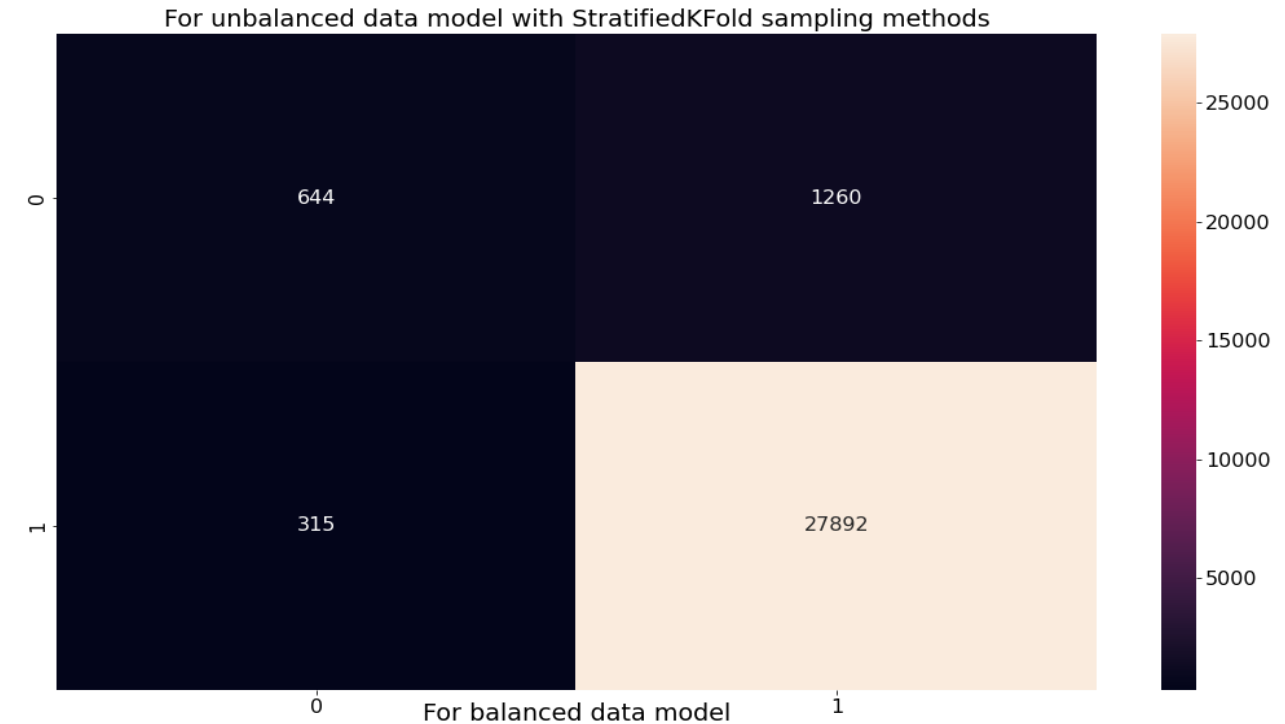
```
==========For unbalanced data model with StratiefKfold======
              precision    recall  f1-score   support

           0       0.66      0.07      0.13      1840
           1       0.94      1.00      0.97     28271

    accuracy                           0.94     30111
   macro avg       0.80      0.53      0.55     30111
weighted avg       0.93      0.94      0.92     30111

====================For balanced data model=================
              precision    recall  f1-score   support

           0       0.54      0.18      0.27      1840
           1       0.95      0.99      0.97     28271

    accuracy                           0.94     30111
   macro avg       0.75      0.58      0.62     30111
weighted avg       0.92      0.94      0.93     30111
```

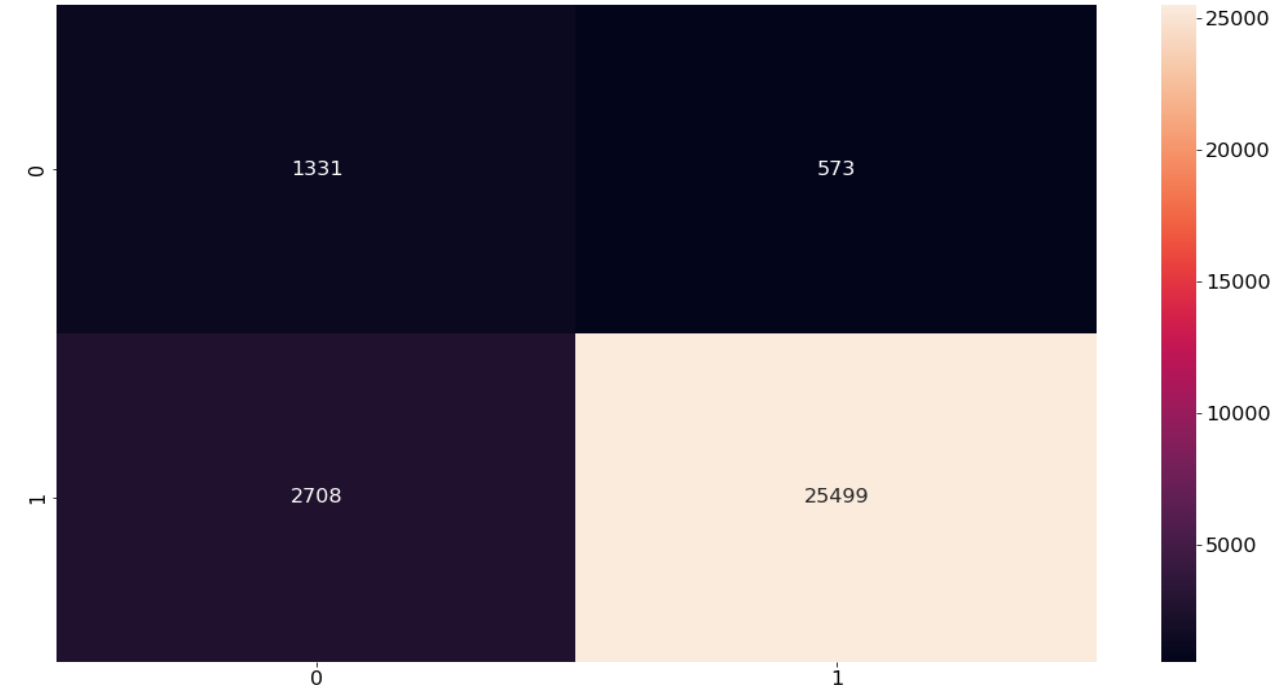




GradientBoost Classifier

```
====================For unbalanced data====================
              precision    recall  f1-score   support

           0       0.67      0.35      0.46      2206
           1       0.96      0.99      0.97     31920

    accuracy                           0.95     34126
   macro avg       0.81      0.67      0.72     34126
weighted avg       0.94      0.95      0.94     34126

====================For balanced data====================
              precision    recall  f1-score   support

           0       0.33      0.67      0.45      2206
           1       0.98      0.91      0.94     31920

    accuracy                           0.89     34126
   macro avg       0.65      0.79      0.69     34126
weighted avg       0.93      0.89      0.91     34126
```





GradientBoostClassifier

```
==========For unbalanced data model with StratiefKfold=====
              precision    recall  f1-score   support

           0       0.67      0.34      0.45      1904
           1       0.96      0.99      0.97     28207

    accuracy                           0.95     30111
   macro avg       0.81      0.66      0.71     30111
weighted avg       0.94      0.95      0.94     30111

====================For balanced data model================
              precision    recall  f1-score   support

           0       0.33      0.70      0.45      1904
           1       0.98      0.90      0.94     28207

    accuracy                           0.89     30111
   macro avg       0.65      0.80      0.69     30111
weighted avg       0.94      0.89      0.91     30111
```

Conclusion

Dans cette étude nous avons éssentiellement, nous avons construit plusieurs modèle avec des données variées notamment celles équilibrée et non équilibré selon la distribution de chaque classe. Ce fut intéressant car nous avons vu que le modèle entrainée avec un deséquilibre présente une capacité plus louable pour la généralisation par au dataset équilibrée.