Projet 8

Compétition Kaggle

Ubiquant Market Prediction

Marie-France LAROCHE-BARTHET

09/04/2022

INTRODUCTION

Contexte

- **Ubiquant Investment Co.**
  - Fond spéculatif d'investissements quantitatifs
  - Basé à Pékin
- **But compétition**
  - Prédire valeur de la target
    - ➡ Taux de retour d'un investissement
    - ➡ Modèle de régression
- **Evaluation**
  - Coefficient de corrélation de Pearson
  - Utilisation d'une API spécifique pour soumission
- **Timing**
  - 18 janvier au 18 avril 2022

Données

- **Accès**

  https://www.kaggle.com/competitions/ubiquant-market-prediction/data

- **Dataset lourd (17. 2Go)**
  - Version allégée utilisée (3.6 Go, format parquet)
- **Description**
  - 3,141,410 lignes
  - 304 colonnes
    - ➡ row_id
    - ➡ time_id
    - ➡ investment_id
    - ➡ target
    - ➡ features f_0 à f_299

Objectif

- **Influence preprocessing**
  - Réduction dimension
  - Scaling

➡ **Features**

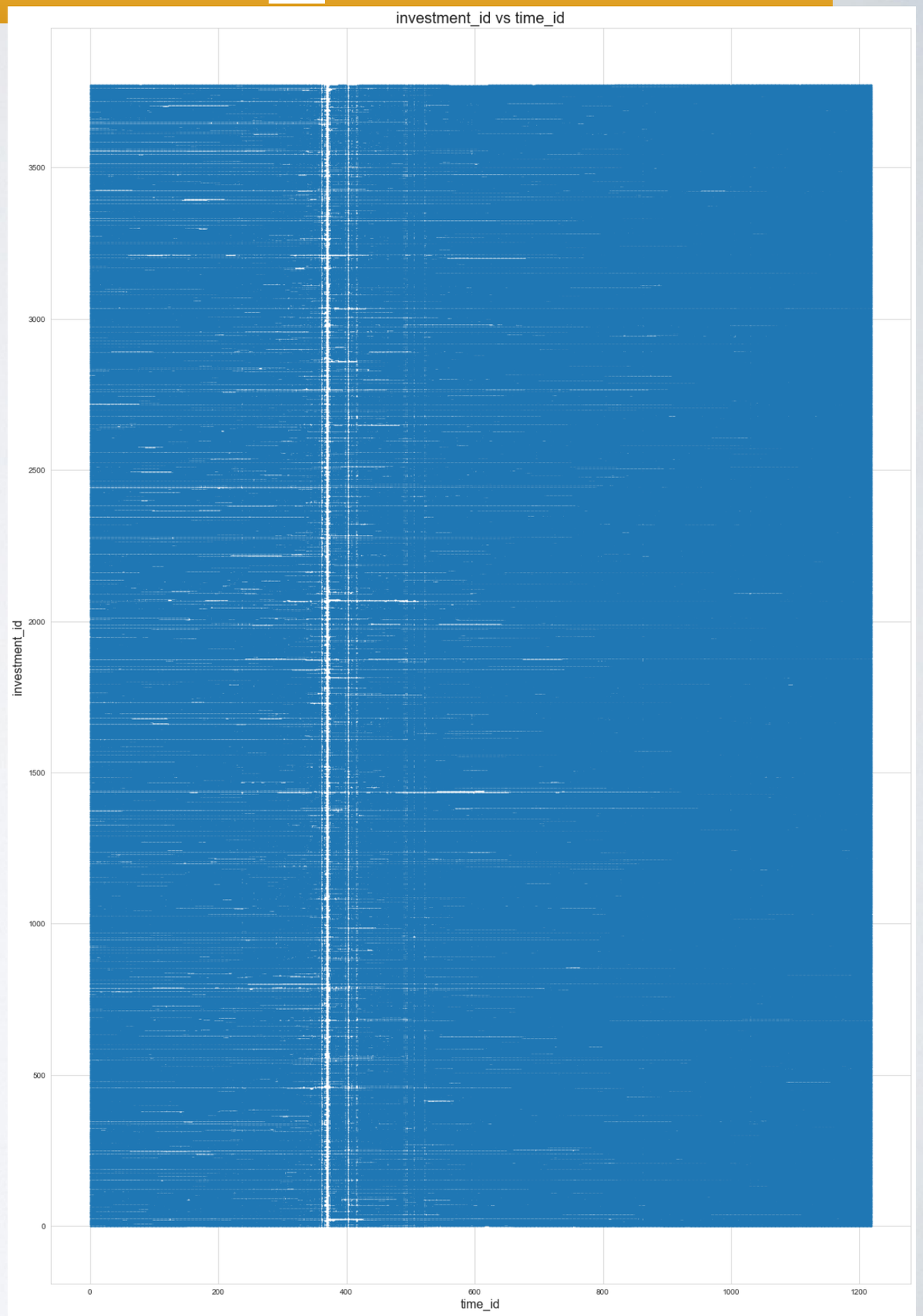
- **Evaluation modèles Machine Learning**
  - LinearRegression
  - ElasticNet
  - XGBRegressor
  - Réseaux de neurones convolutionnels 1D (Conv1D)

ANALYSE DES DONNÉES

Time_id et investment_id

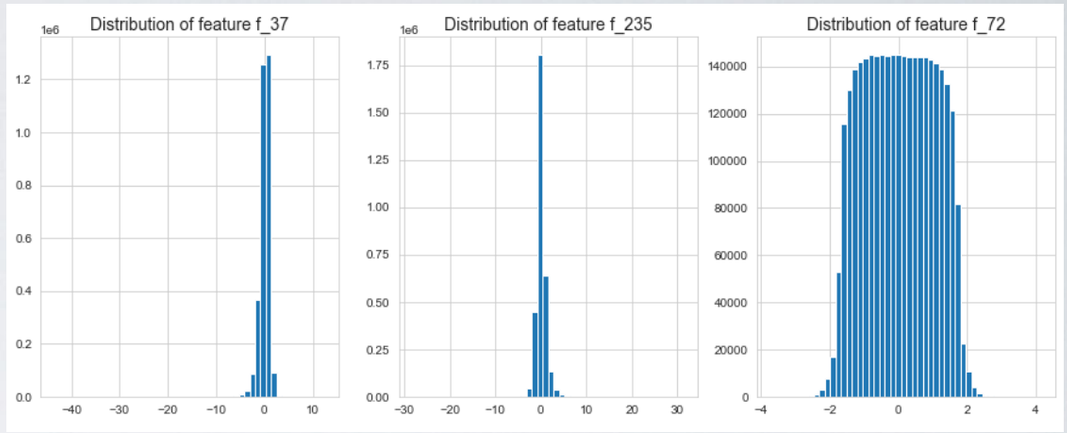
- **Time_id**
  - 1211 valeurs différentes
- **Investment_id**
  - 3579 valeurs différentes

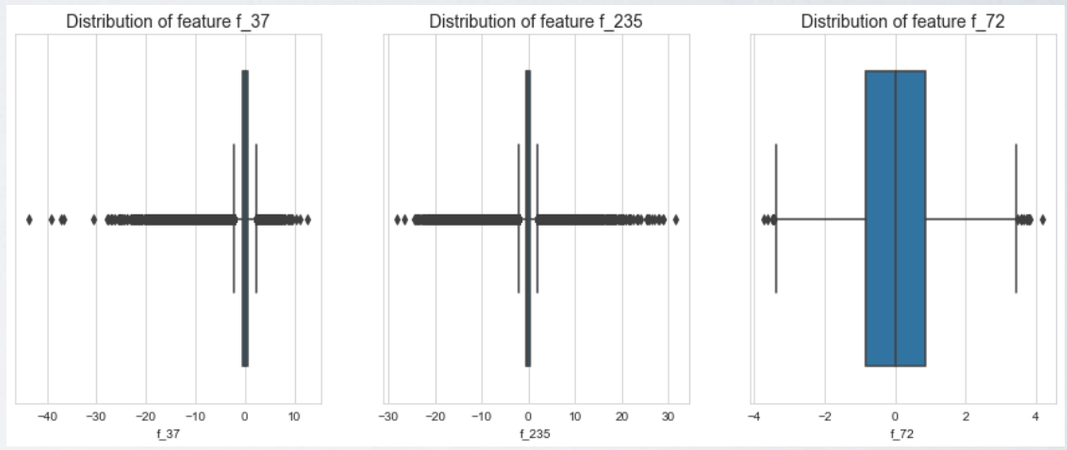
➡ + d'investment_id pour time_id élevés



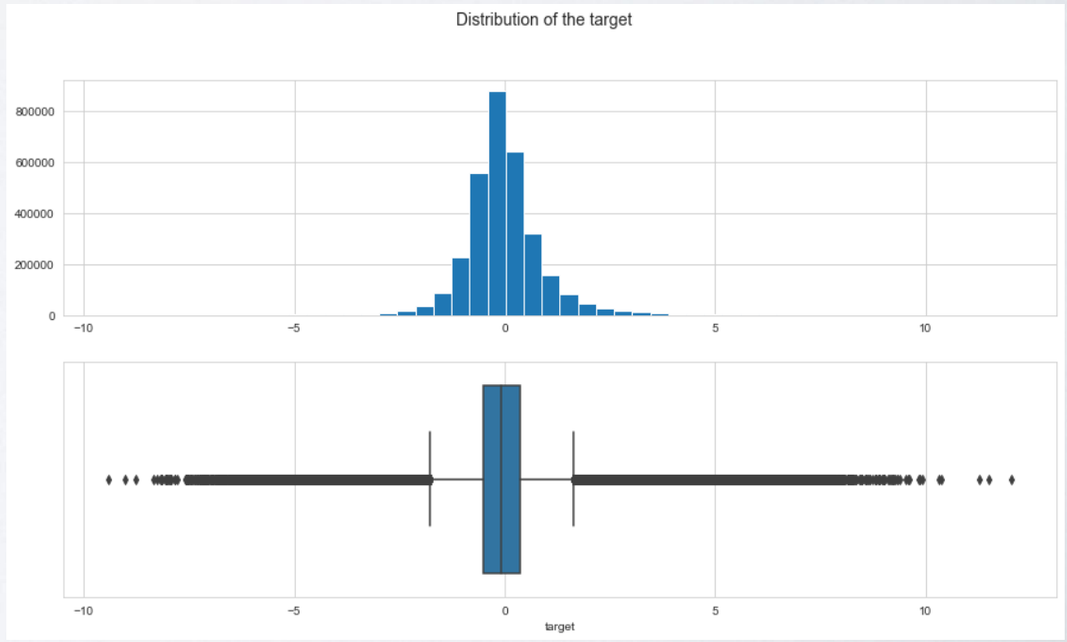
Features et target

- **Features**
  - Moyennes globalement proches de zéro
  - Présence d'outliers



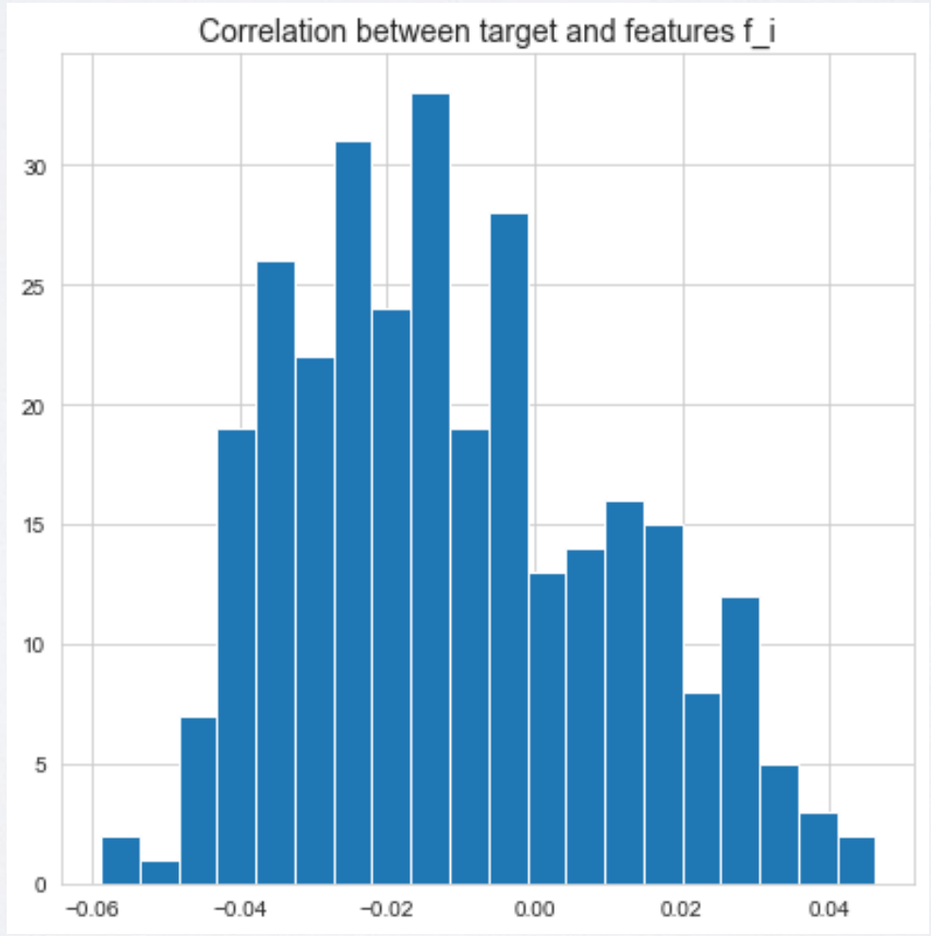


- **Target**
  - Globalement centrée en zéro
  - Présence d'outliers



Corrélation target-features

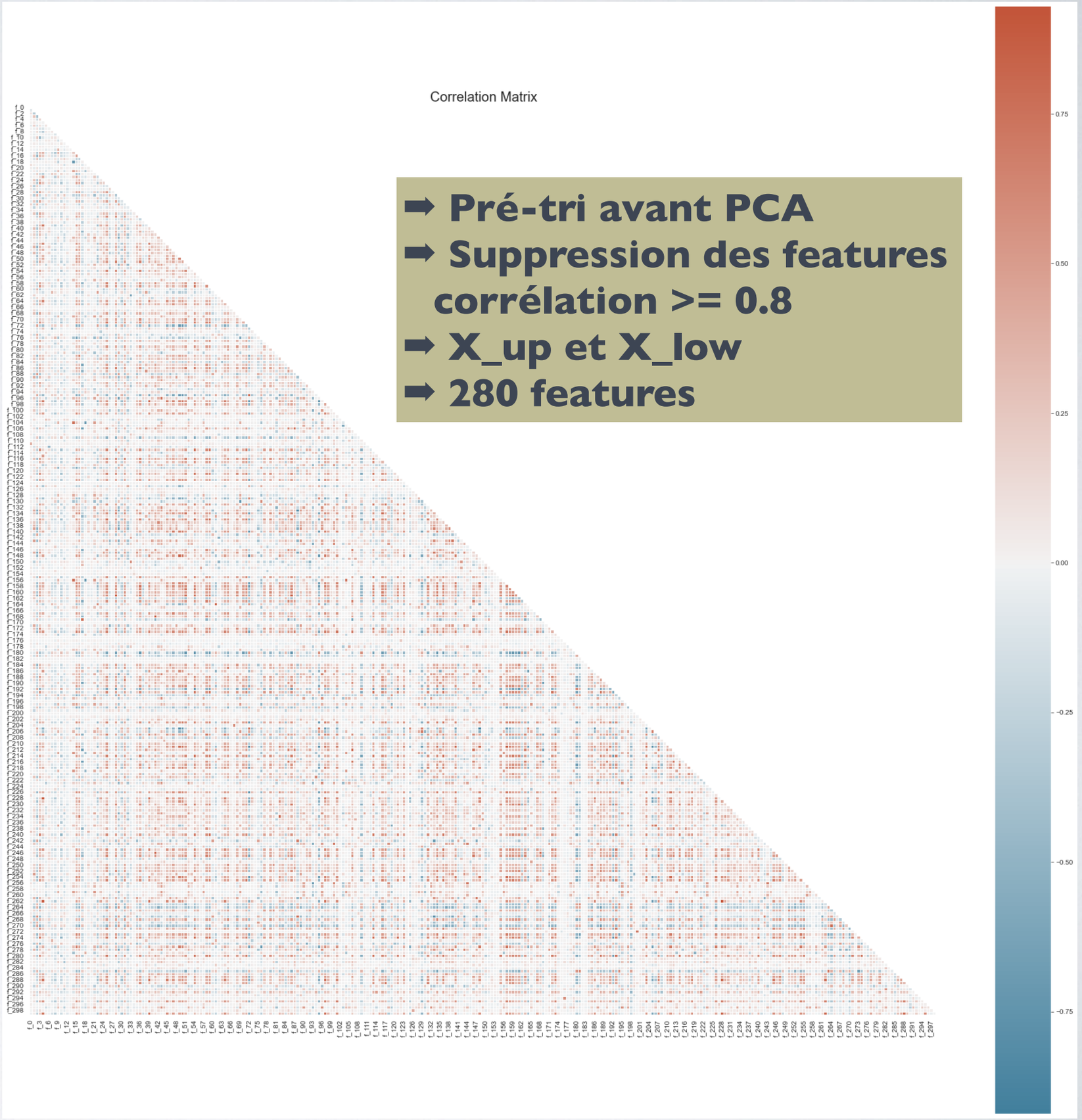
- **Echantillon du dataset** 157070 x 304
  - 5% des données gardées
  - Choix aléatoire
  - Trié par time_id puis investment_id
- **Histogramme corrélations target-features**



➡ **Peu de corrélations**

Corrélation entre features

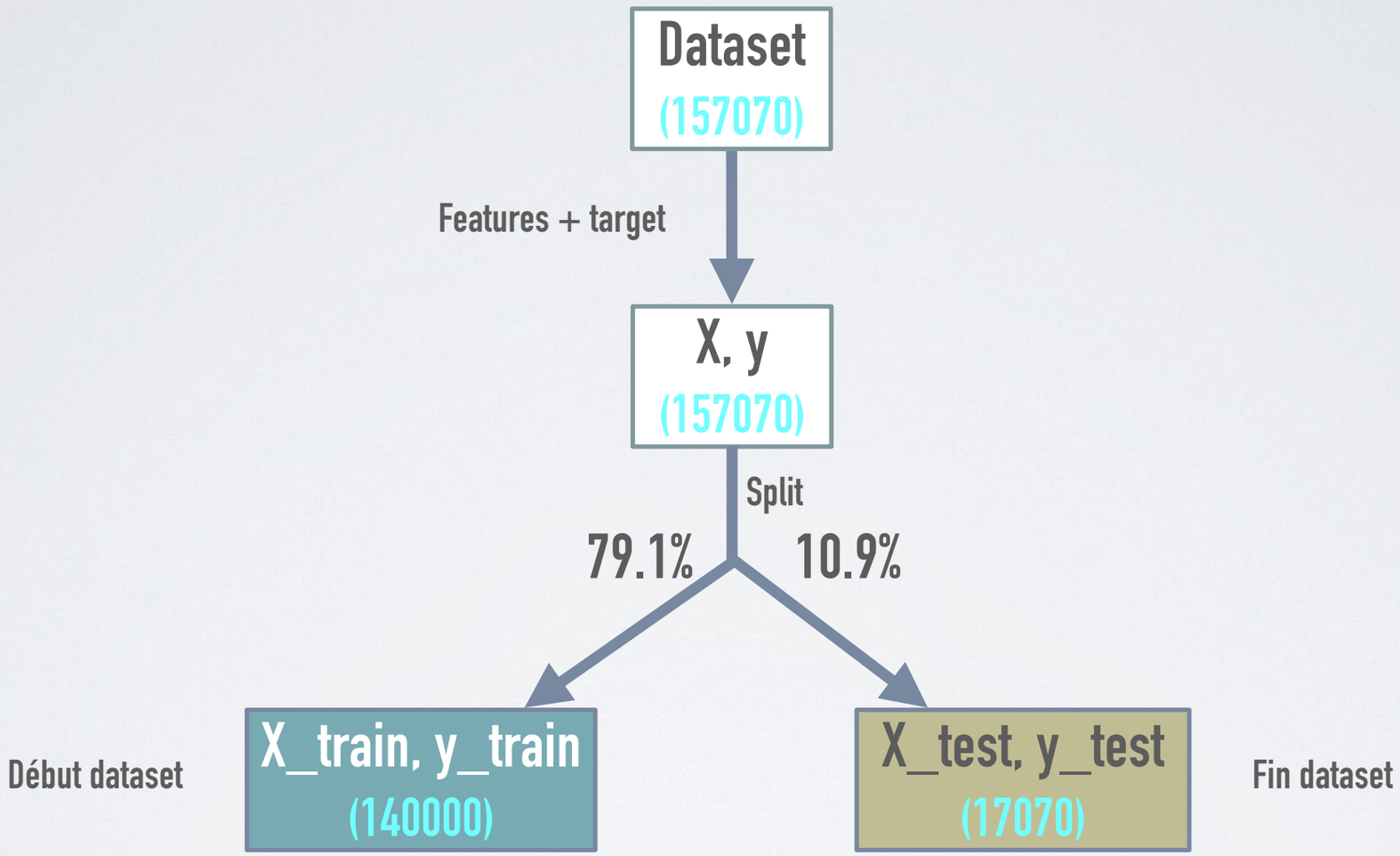
- **Matrice corrélations**
  - Majoritairement faibles corrélations
  - Quelques fortes corrélations



➡ **Pré-tri avant PCA**
➡ **Suppression des features corrélation >= 0.8**
➡ **X_up et X_low**
➡ **280 features**

PREPROCESSING

Split des données



Scaling et PCA

- **Scaling**
  - RobustScaler
    - ➡ Amoindrit influence outliers
- **PCA (Principal Analysis Component)**
  - 85% variance expliquée
    - ➡ Réduction entre 55 et 97% du nombre de features

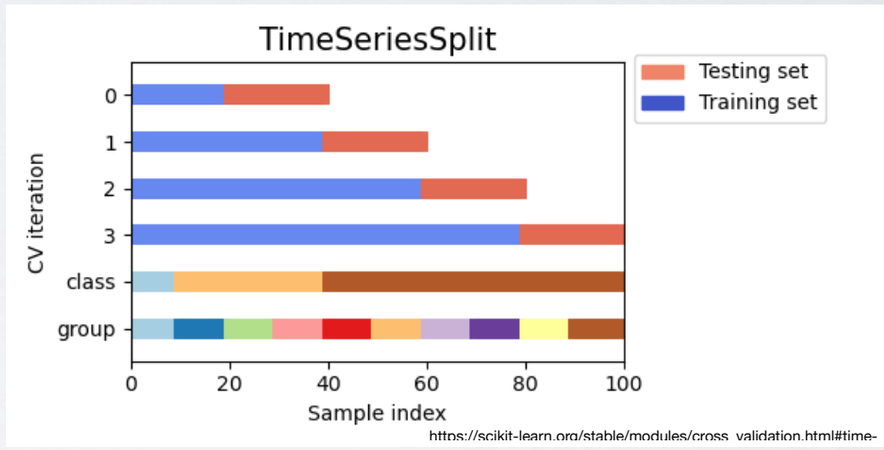
- **X_data utilisées en modélisation**

X_data	Nb features	Scaling	PCA	% Réduction de dimension via PCA
X_train	300	NON	NON	/
X_scaled	300	OUI	NON	/
X_pca85	125	NON	OUI	58 %
X_scal_pca85	10	OUI	OUI	96.7%
X_up_pca85	126	NON	OUI	55 %
X_up_scal_pca85	9	OUI	OUI	96.8%
X_low_pca85	125	NON	OUI	55.4%
X_low_scal_pca85	9	OUI	OUI	96.8%

MODÉLISATION

Entrainement des modèles

- **Modèles testés**
  - LinearRegression
  - ElasticNet
  - XGBRegressor
  - Conv1D
- **Validation croisée**
  - Utilisation de TimeSeriesSplit
    - ➡ 5 folds
    - ➡ Validationset = 20 000 lignes
  - Pas avec Conv1D
    - ➡ Validationset = 20 derniers% trainset initial



- **Métriques utilisées**
  - RMSE (Root Mean Squared Error)
  - Coefficient corrélation Pearson
  - MAE (Mean Absolute Error)
  - MSE (Mean Squared Error)

Optimisation des modèles

- **GridSearchCV - Hyperparamètres**
  - ElasticNet
    - ➡ Scikit-Learn
  - XGBRegressor
    - ➡ Manuellement
- **Conv1D**
  - Entrainement 30 ou 50 epochs maximum
  - Utilisation Earlystopping
    - ➡ Patience 5 ou 10 sur val_rmse
  - Utilisation Checkpoint
    - ➡ Enregistre poids des paramètres correspondant meilleure valeur val_rmse
  - Optimizer Adam
    - ➡ Learning_rate 0.001
  - Batch_Size 4096
  - Différentes architecture testées
    - ➡ Empilement couches
    - ➡ Fonction activation
    - ➡ Kernel_regularizer

Récapitulatif des modèles testés

- **Combinaisons X_data/modèle**

X_data	X_train	X_scaled	X_pca85	X_scal_pca85
Modèles	LinearRegression Conv1D_1 à 7 Conv1D_1_1	LinearRegression ElasticNet Conv1D_1 et 6 Conv1D_6_1	LinearRegression XGBRegressor Conv1D_1	LinearRegression ElasticNet XGBRegressor

X_data	X_up_pca85	X_up_scal_pca85	X_low_pca85	X_low_scal_pca85
Modèles	LinearRegression	LinearRegression ElasticNet	LinearRegression	LinearRegression ElasticNet

RÉSULTATS

Jeu de validation

- **LinearRegression**
  - Même top3 avec les 3 métriques
  - PCA améliore score
  - Scaling dégrade score

X_data	RMSE	MAE	Pearson_coef
X_up_pca85	**0.906208**	**0.618814**	**0.105118**
X_low_pca85	0.906225	0.618842	0.10491
X_pca85	0.906316	0.618908	0.104369
X_low_scal_pca85	0.908127	0.620206	0.07334
X_up_scal_pca85	0.908131	0.620209	0.07328
X_scal_pca85	0.9082	0.620276	0.072287
X_scaled	0.908742	0.621333	0.09949
X_train	0.908742	0.621333	0.09949

- **ElasticNet**
  - Réduction dimension dégrade score

X_data	RMSE	Pearson_coef
X_scaled	**0.90662**	**0.10346**
X_low_scal_pca85	0.90799	0.07408
X_up_scal_pca85	0.90802	0.07402
X_scal_pca85	0.90802	0.07346

- **XGBRegressor**
  - Scaling dégrade score

X_data	RMSE	Pearson_coef
X_pca85	**0.906451**	**0.091055**
X_scal_pca85	0.907621	0.075637

Jeu de validation

- **Conv 1D**
  - Pas même meilleur modèle selon RMSE /MSE et MAE
  - Plus de couches de convolution améliorent score RMSE
- Faible learning rate améliore MAE
- Architecture meilleur modèle
  - ➡ Activation ReLU
  - ➡ Dropout 0.5

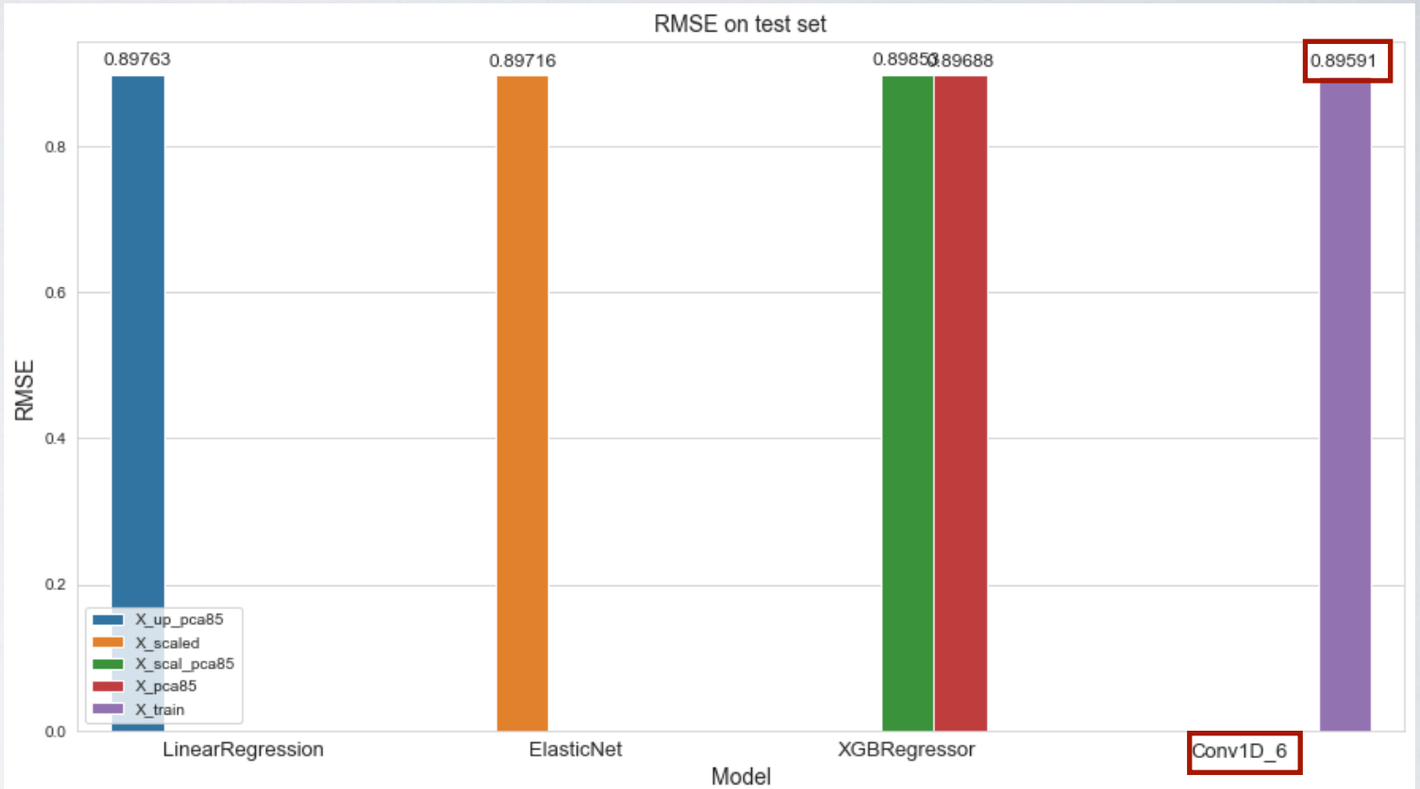
Modèle	X_data	RMSE	Loss = MSE	MAE
Conv1D_6	**X_train**	**0.9041**	**0.81741**	0.61136
Conv1D_6_sc_1	X_scaled	0.90418	0.81755	0.61078
Conv1D_5	X_train	0.90451	0.82206	0.61128
Conv1D_4	X_train	0.90456	0.82366	0.61056
Conv1D_3	X_train	0.9049	0.81885	0.61167
Conv1D_1_1	X_train	0.90516	0.81931	0.61157
Conv1D_2	X_train	0.90526	0.8195	0.6135
Conv1D_1	X_train	0.90528	0.81954	0.61126
Conv1D_scal_1	X_scaled	0.90531	0.8196	**0.60981**
Conv1D_pca_1	X_pca85	0.90548	0.81989	0.61169
Conv1D_7	X_train	0.90569	0.82028	0.61073
Conv1D_6_sc	X_scaled	0.90599	0.82081	0.611

Layer (type)	Output Shape	Param #
Conv1D_1 (Conv1D)	(None, 300, 64)	320
dropout_63 (Dropout)	(None, 300, 64)	0
Conv1D_2 (Conv1D)	(None, 98, 16)	8208
Conv1D_3 (Conv1D)	(None, 96, 16)	784
Conv1D_4 (Conv1D)	(None, 32, 32)	2080
Conv1D_5 (Conv1D)	(None, 8, 64)	8256
flatten_38 (Flatten)	(None, 512)	0
Dense_2 (Dense)	(None, 64)	32832
Dense_3 (Dense)	(None, 32)	2080
Dense_4 (Dense)	(None, 1)	33

Total params: 54,593
Trainable params: 54,593
Non-trainable params: 0

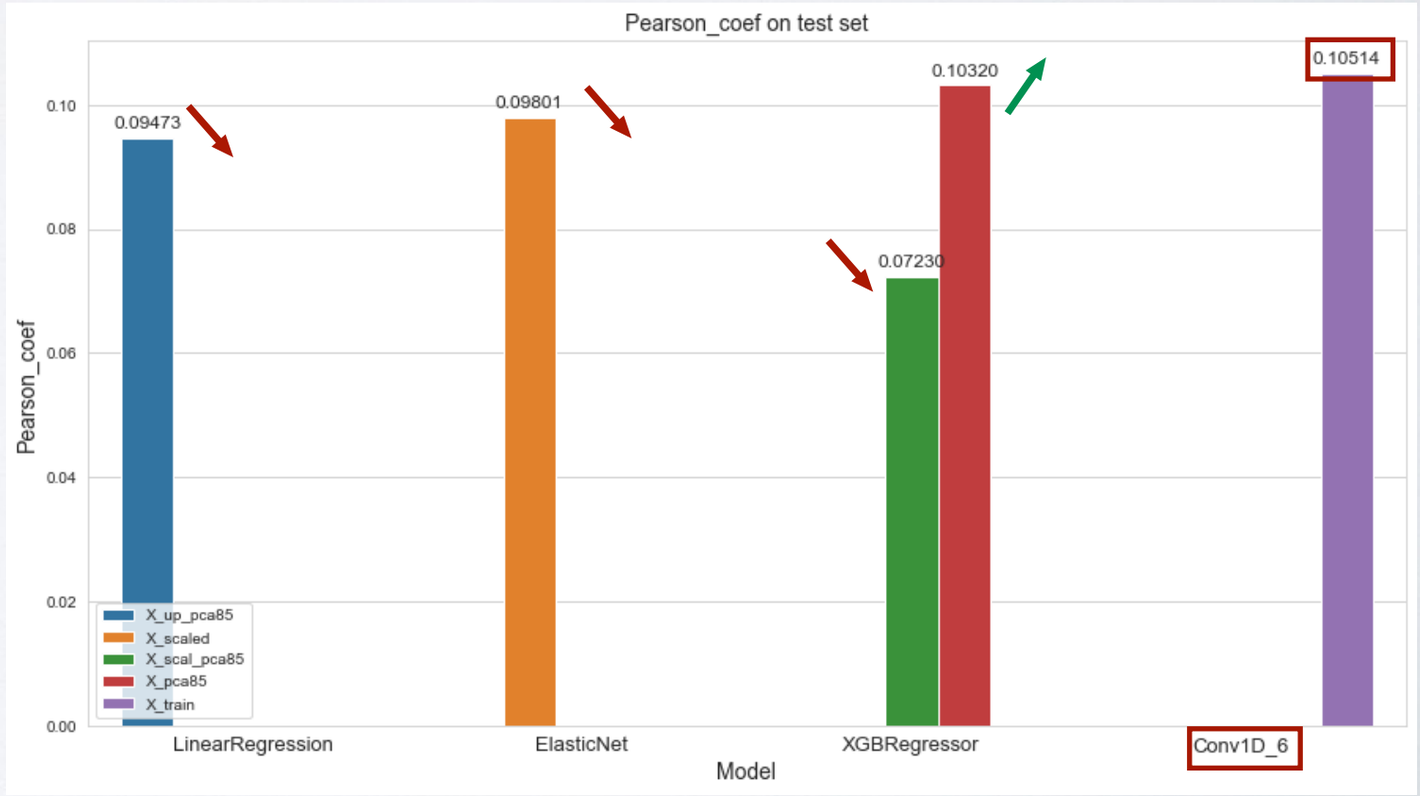
Jeu de test

- **RMSE**
  - Valeurs très proches
  - Scores meilleurs que sur validationset



- **Pearson coefficient**
  - Plus grande variabilité

➡ Conv1D_6 meilleur modèle



CONCLUSION

Conclusion

- **Influence preprocessing**
  - Réduction dimension ➡ **Utile pour LinearRegression**
  - Scaling ➡ **Utile pour ElasticNet**
- **Evaluation modèles Machine Learning**
  - LinearRegression ➡ **Pas si mauvais pour modèle basique**
  - ElasticNet
  - XGBRegressor
  - Conv1D ➡ **Meilleur modèle**

Conclusion

- **Pistes d'amélioration**
  - Autres architectures de réseaux Conv1D
  - Réduction du learning_rate
    - ➡ Via callback
  - Considérer time_id ou investment_id
    - ➡ Introduire nouvelle dimension Conv2D
- **Dataset non trivial**

MERCI

QUESTIONS