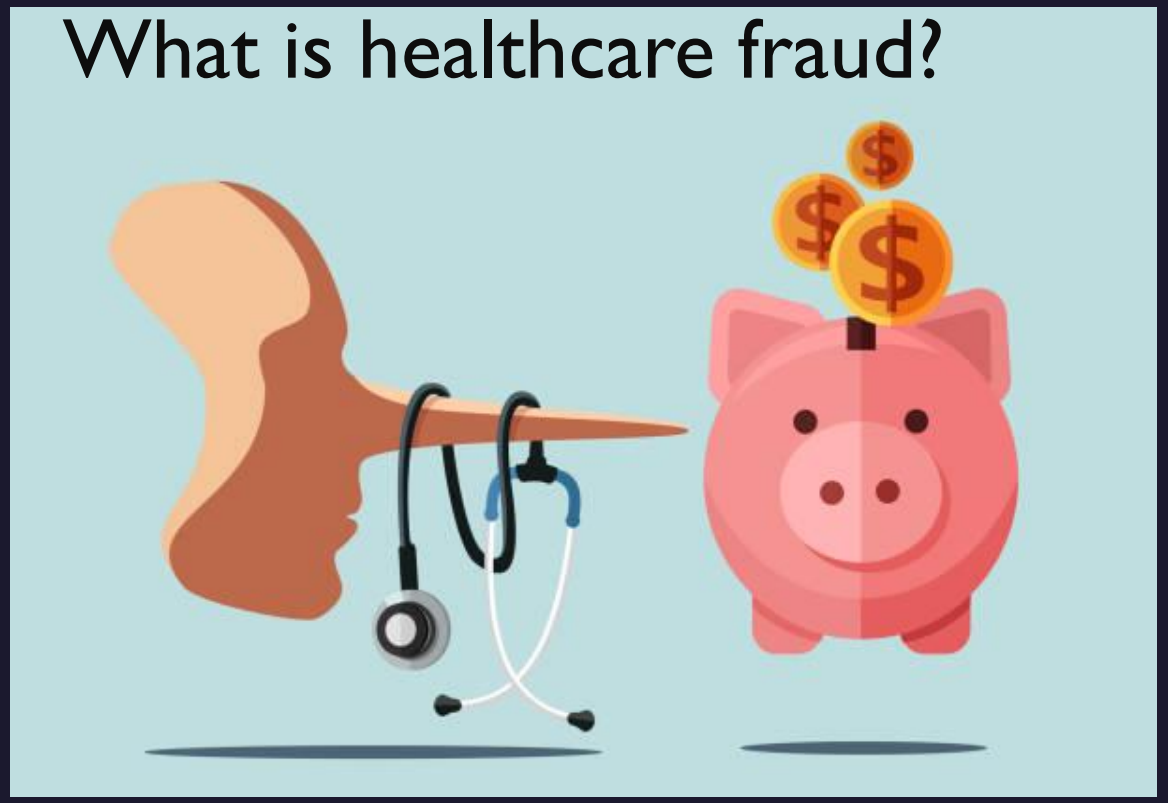
Fraud Detection in Healthcare Insurance Claims

Instructor : Dr Evangelos E.Milios

Group 9:

- Ashpak Rakeeb Mohammad (B00913796)
- Shravya Reddy Gennepally (B00911193)
- Sri Ramya Basam (B00900307)
- Venkata Vijaya Mandapati (B00912916)
- Meagan Sinclair (B00737317)



[1]

Introduction

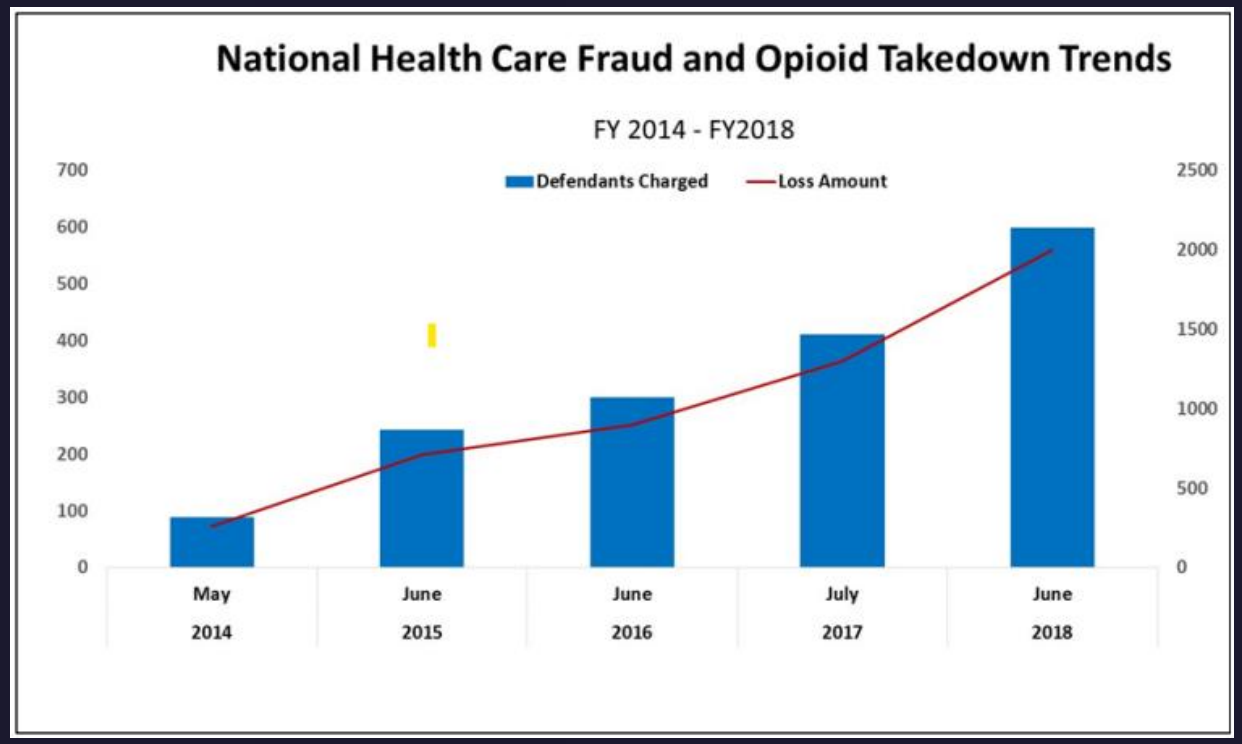
Fraudsters find the healthcare industry to be an attractive target, resulting in significant monetary losses. Here, we examined and identified fraud in healthcare insurance claims where the healthcare practitioner submits the claim on the beneficiary's behalf. One of the main issues that Medicare is now facing is this type of deception. Due to this medical expenditure has rapidly grown.

Literature Review

Clustering methods [8, 9]:
- Bayesian co-clustering
- Cannot conclusively detect fraud

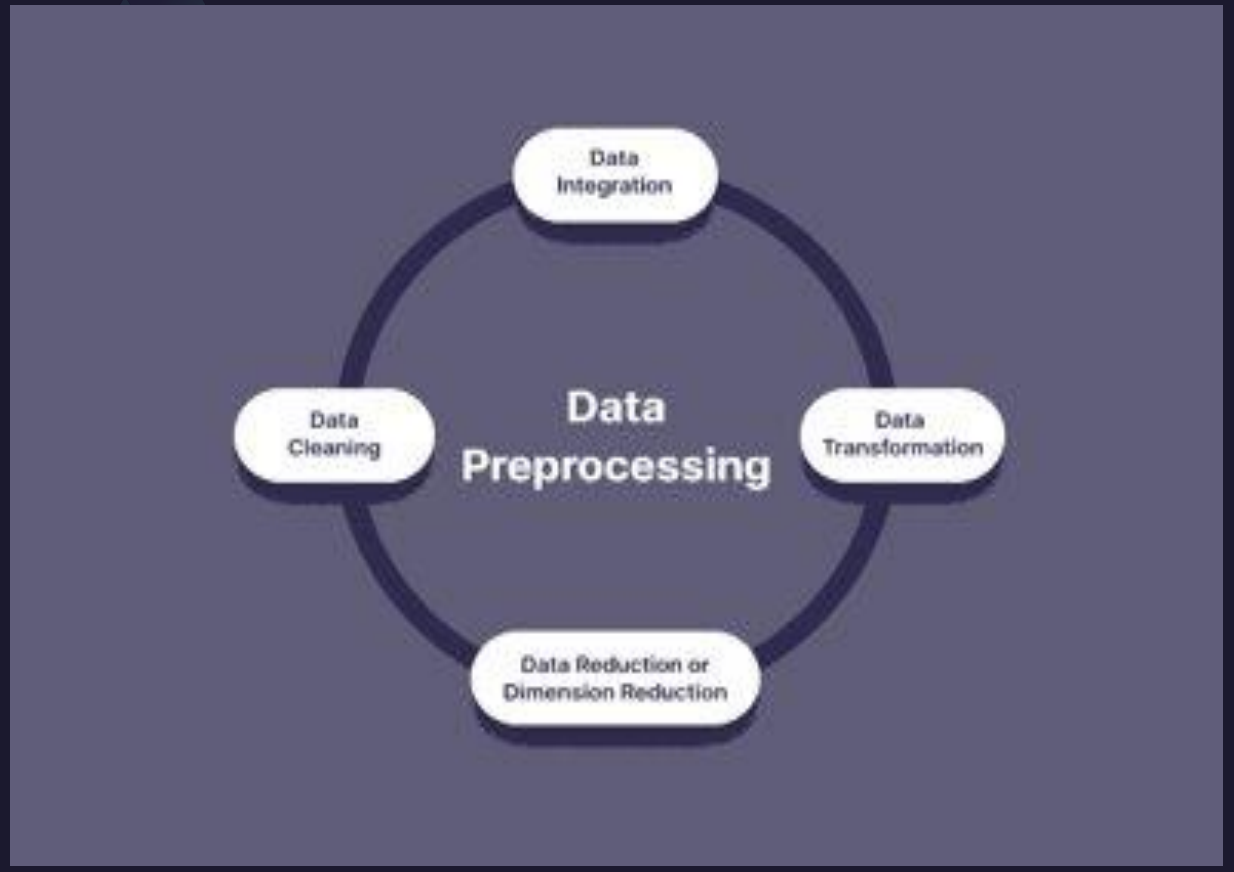
Kaggle data and notebook [7]:
- Logistic regression, random forest, auto-encoders
- Good accuracy
- No parameter tuning

Limited use of ensembled models in literature [10]



Ref : https://oig.hhs.gov/publications/docs/hcfac/FY2018-hcfac.pdf

Methodology

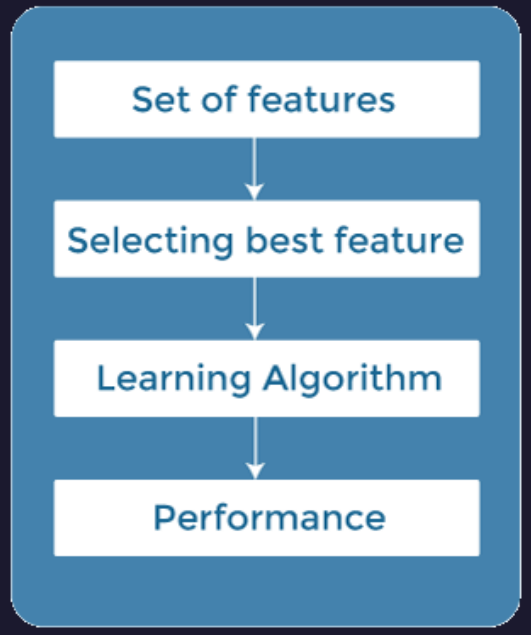


Ref : https://www.v7labs.com/blog/data-preprocessing-guide

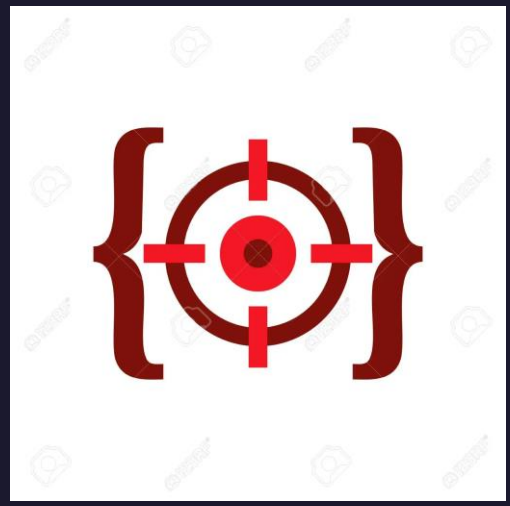
- Data Source
  - Medical Provider Fraud Detection from Kaggle
- Data exploration
  - Beneficiary Data, In-patient and Out-patient datasets (693k records 56 columns collectively)
- Data preprocessing
  - Combined all the three different datasets, eliminate data redundancy
  - Derived features
- Data quality plan
  - Continuous and Categorical features reports to handle missing values.

Methodology (Cont.)

- Feature selection and Techniques
  - SelectKBest with mutual information for feature selection
  - Shuffle split (80% train, 20% test)
- Model selection
  - Supervised model, Learning curves, Scalability and Performance of model



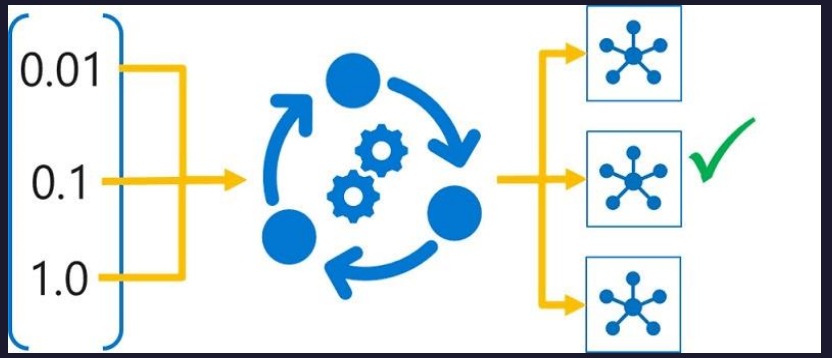
[4]



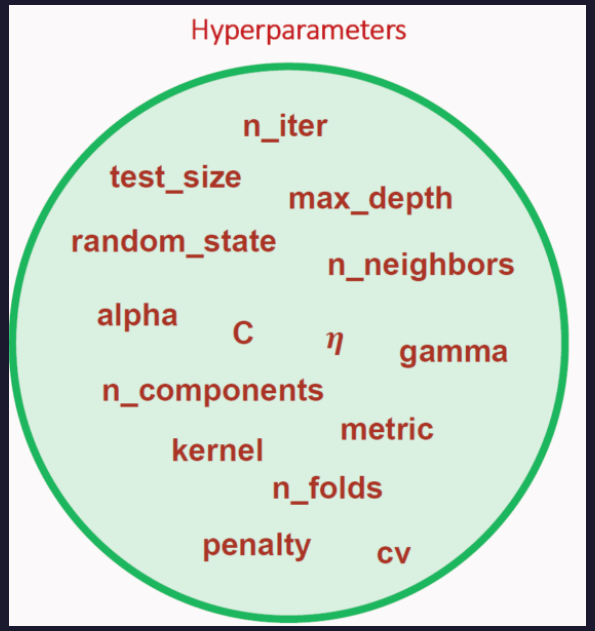
[5]

Methodology (Cont.)

- Parameter Tuning
  - Adjustable parameters that control the model training process
  - Model performance heavily depends on hyperparameters. Values of hyperparameters might improve or worsen the model's accuracy.
  - Methods of Hyperparameter tuning: GridSearch, RandomSearch, InformedSearch
- Evaluation
  - Accuracy
  - F1 Score



[6]

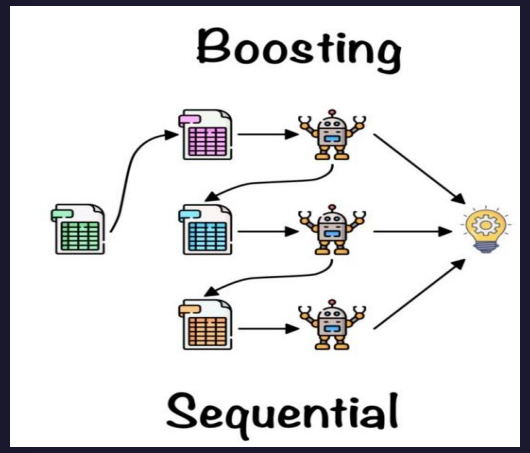


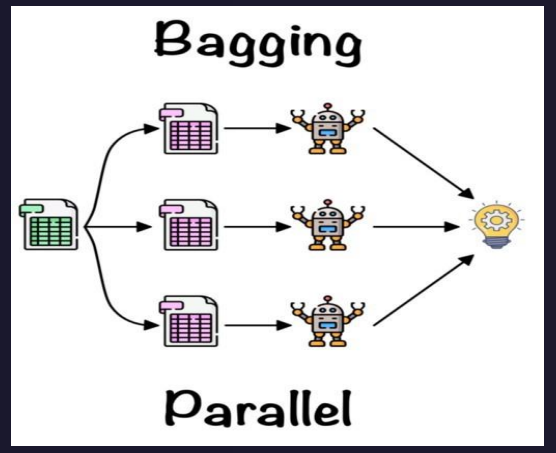
[6]

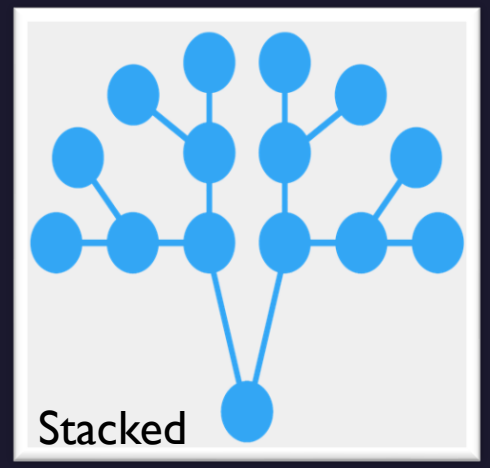
Experiments

Performed 4 different models

Ensemble Technique







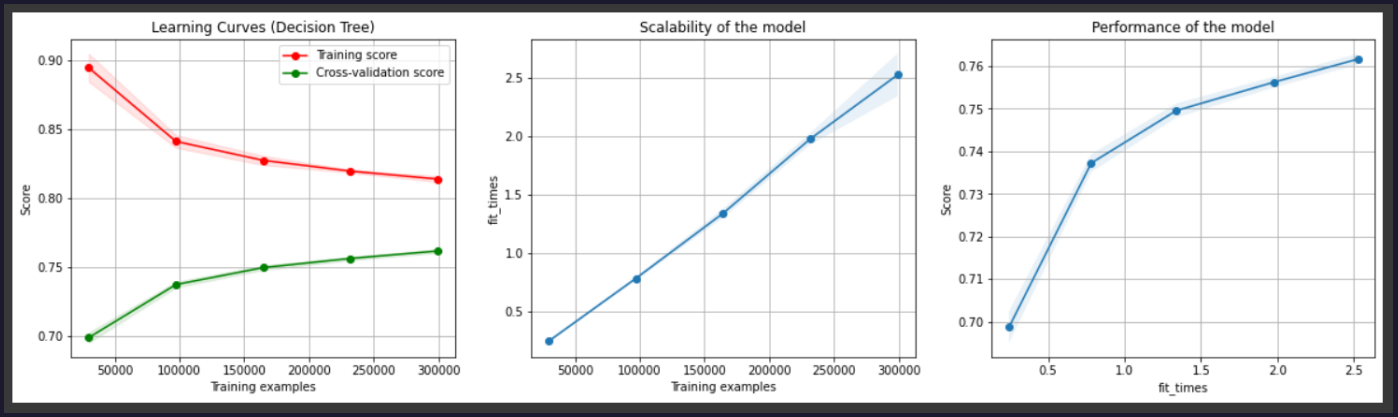
- Decision tree classifier
- Hyper-tuned with depth and features parameters of the tree

- XGBoost with weak classifier as a Decision tree
- Hyper-tuned tree depth, for optimizing model complexity

- Bagging classifier based on decision tree
- Hyper-tuned the number of base classifiers

- Stacked base estimators: Logistic regression, k-neighbors classifier, Decision tree, and Gaussian naïve Bayes
- Final estimators: Logistic regression

Results

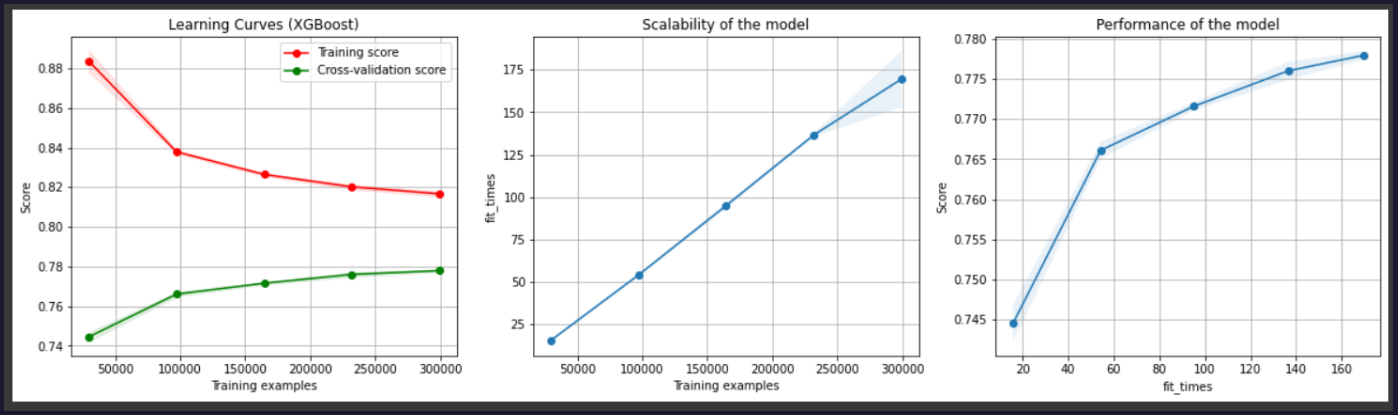



```
Confusion Matrix Val:
 [[96281 18024]
 [25674 44231]]
              precision    recall  f1-score   support

       False       0.79      0.84      0.82    114305
        True       0.71      0.63      0.67     69905

    accuracy                           0.76    184210
   macro avg       0.75      0.74      0.74    184210
weighted avg       0.76      0.76      0.76    184210
```

Decision Tree Evaluation



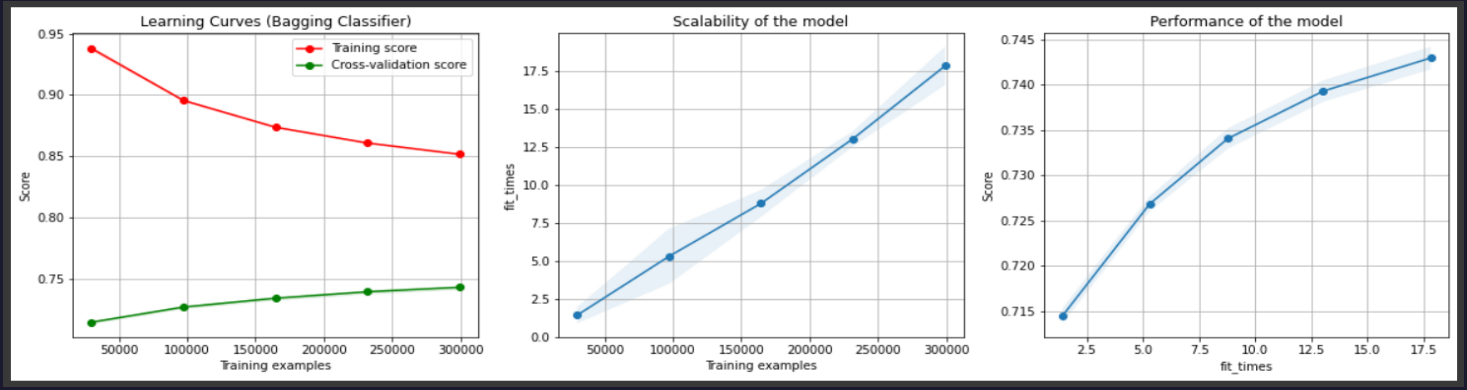

```
Confusion Matrix Val:
 [[97503 16802]
 [24231 45674]]
              precision    recall  f1-score   support

       False       0.80      0.85      0.83    114305
        True       0.73      0.65      0.69     69905

    accuracy                           0.78    184210
   macro avg       0.77      0.75      0.76    184210
weighted avg       0.77      0.78      0.77    184210
```

XG Boost Evaluation

Results (Cont.)

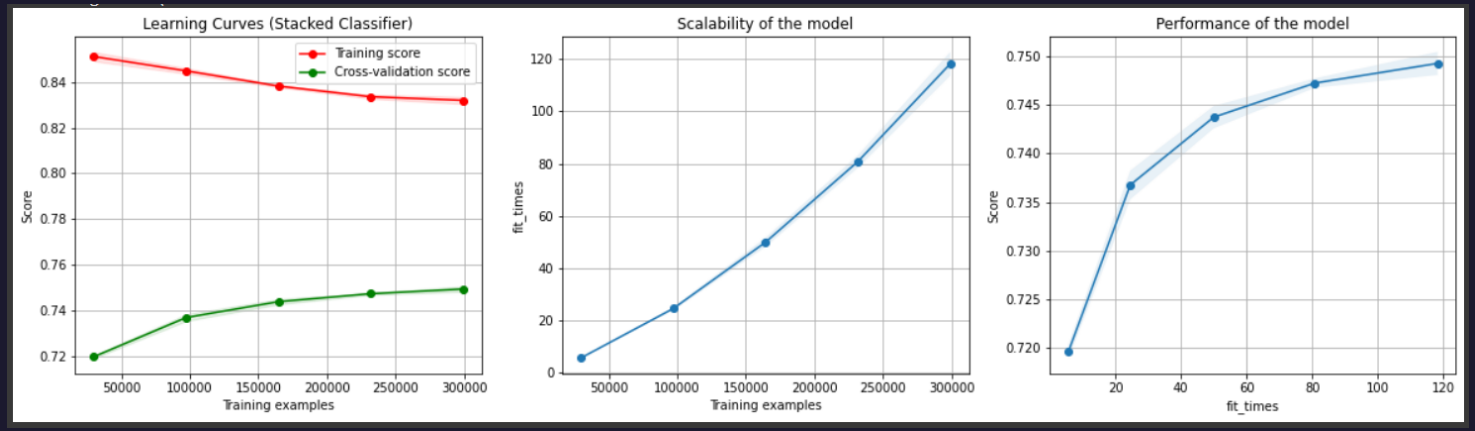


Bagging Classifier Evaluation

```
Confusion Matrix Val:
 [[92384 21921]
 [24771 45134]]
              precision    recall  f1-score   support

       False       0.79      0.81      0.80    114305
        True       0.67      0.65      0.66     69905

    accuracy                           0.75    184210
   macro avg       0.73      0.73      0.73    184210
weighted avg       0.74      0.75      0.75    184210
```



Stacked Classifier Evaluation

```
Confusion Matrix Val:
 [[94483 19822]
 [26326 43579]]
              precision    recall  f1-score   support

       False       0.78      0.83      0.80    114305
        True       0.69      0.62      0.65     69905

    accuracy                           0.75    184210
   macro avg       0.73      0.72      0.73    184210
weighted avg       0.75      0.75      0.75    184210
```

Conclusion

Most accurate current model: XGBoost with 78% accuracy

Further work:

- Assessing bias and variance

References

[1] https://www.istockphoto.com/illustrations/healthcare-fraud
[2] https://www.fbi.gov/how-we-can-help-you/safety-resources/scams-and-safety/common-scams-and-crimes/health-care-fraud
[3] https://oig.hhs.gov/publications/docs/hcfac/FY2018-hcfac.pdf
[4] https://www.javatpoint.com/feature-selection-techniques-in-machine-learning
[5] https://www.123rf.com/photo_101011059_stock-vector-accuracy-code-logo-icon-design.html
[6] https://k21academy.com/microsoft-azure/dp-100/hyperparameter-tuning-in-azure/
[7] https://www.kaggle.com/datasets/rohitrox/healthcare-provider-fraud-detection-analysis
[8] Peng, G. Kou, A. Sabatka, Z. Chen, D. Khazanchi, and Y. Shi, "Application of clustering methods to health insurance fraud detection," in 2006 International Conference on Service Systems and Service Management, vol. 1, pp. 116–120, ISSN: 2161-1904.
[9] T. Ekin, "Application of bayesian methods in detection of healthcare fraud," vol. 33.
[10] R. Bauder, T. M. Khoshgoftaar, and N. Seliya, "A survey on the state of healthcare upcoding fraud analysis and detection," vol. 17, no. 1, pp. 31–55. [Online]. Available: https://doi.org/10.1007/s10742-016-0154-8