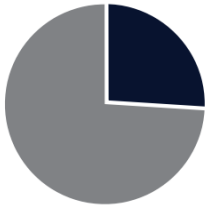
FRAUD DETECTION MODEL FOR MOBILE BANKING

By: Lien Pham

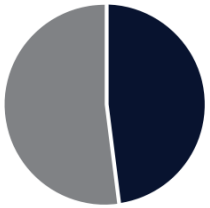
WHAT IS FRAUD?

SOME STATISTICS ABOUT FRAUD IN 2020



26%

of financial institutions experienced at least one spear-phishing or business email compromise attack in 2019 where user credentials were compromised and/or fraud was committed.



48%

of financial services organizations have limited or no visibility when it comes to identifying the impact of a phishing attack.

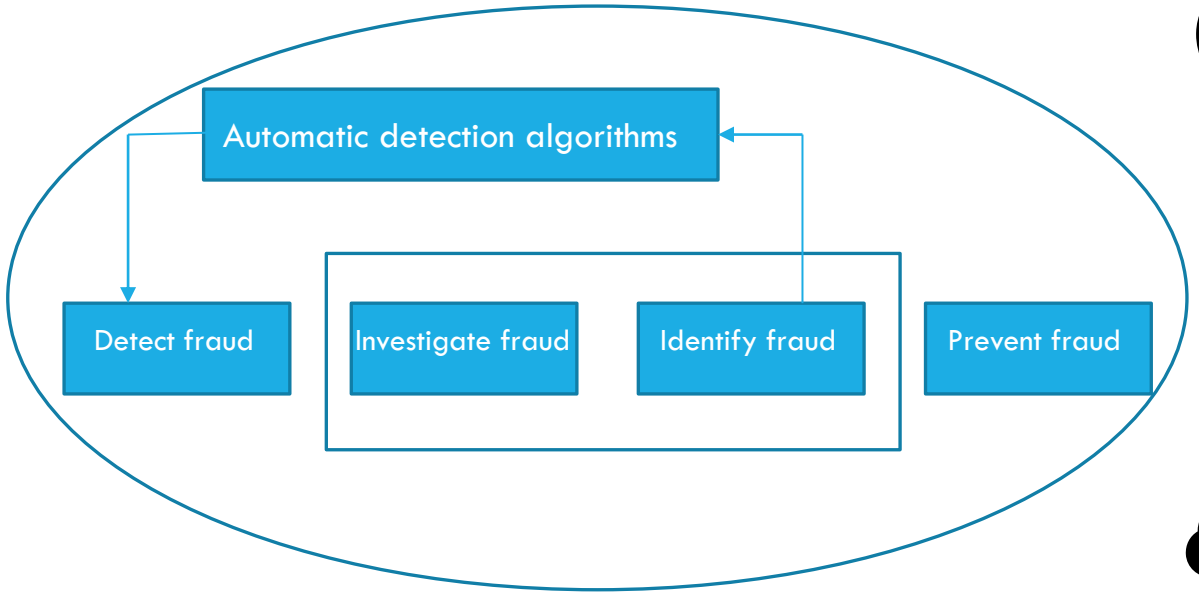
57%

say the lack of awareness of socially engineered fraud schemes among customers and partners is a serious concern.

CHARACTERISTICS OF FRAUD

- Done secretly and on purpose
- Behaviors and methods change over time and on purpose
- Directly or indirectly for personal economic services
- Negatively affect the assets, revenue, and surplus value of the business organizations
- Breach of duty of loyalty to the company

CYCLE OF FRAUD DETECTION



- **Fraud detection:** Apply the detection model on the new, never-observations and label each observation with a fraud risk flag

- **Fraud investigation:** operation audit team examines fraud depending on the level of severity and complexity of fraud

- **Fraud confirmation:** final fraud labeling determination and field study required

- **Fraud prevention:** helps detect fraud before fraudsters implement fraudulent behaviors

The scope of machine learning fraud detection modeling will focus on fraud detection across the fraud lifecycle, to provide warnings about suspicious behavior, transactions, or fraudulent scores high cheat. Then these suspicious acts will be evaluated by fraud experts, checking to see if they are really a fraud

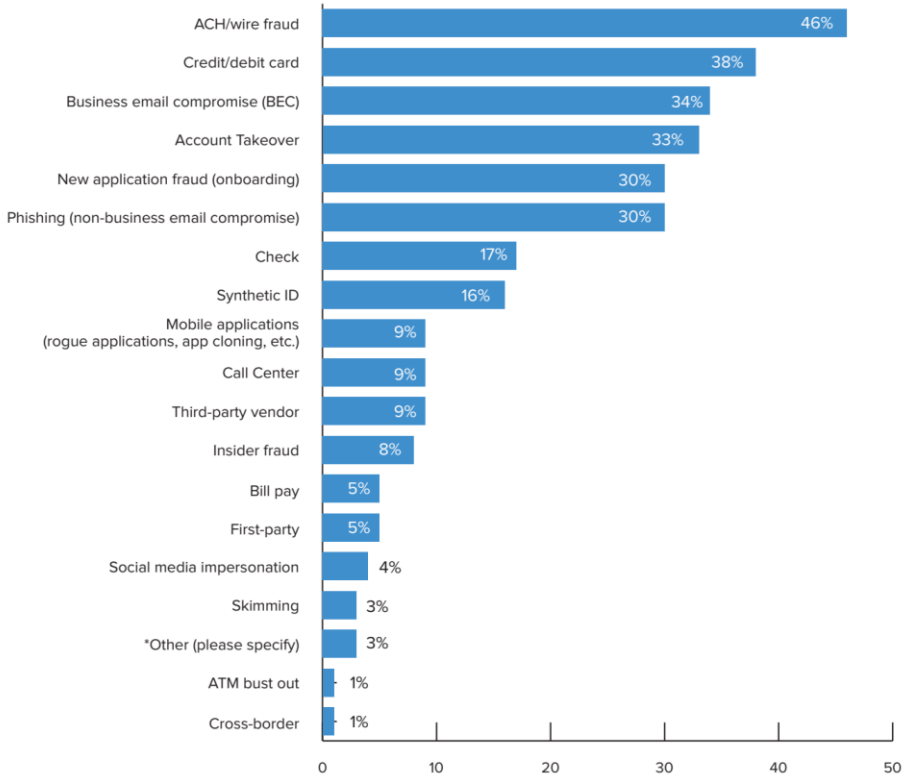
THE MOST POPULAR BANKING FRAUD

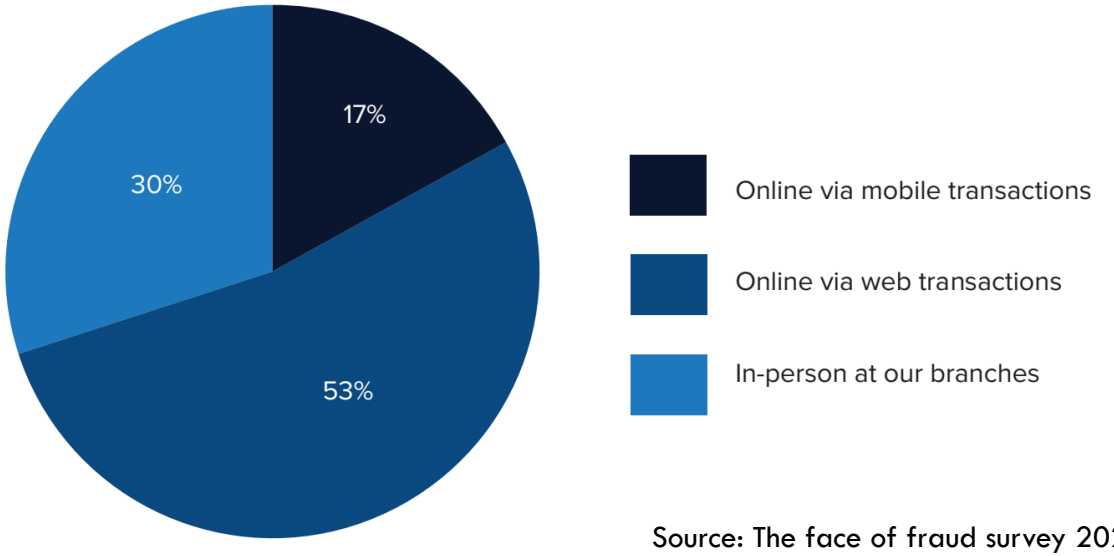
The Most popular mobile banking fraud:

1. ACH/wire fraud is the greatest area of concern, presumably because of the potential transaction size and a high degree of difficulty in orchestrating fund reversal once the fraud has been perpetrated.
2. Credit/debit card fraud (38 percent)
3. Business email compromise attacks are of slightly greater concern than phishing attacks

The top three most concerning fraud schemes in mobile banking:

1. **Identity thief**
2. **Property Fraud**
3. **Account opening fraud**





Source: The face of fraud survey 2020

CONTEXT AND OBJECTIVES OF MOBILE BANKING FRAUD

1

PROPERTY FRAUD

- Property fraud is when crooks gain access to property that doesn't belong to them, changing access and personal information to commit theft

2

ACCOUNT OPENING FRAUD

- Account opening fraud is when a fraudster opens a new account with information that has stolen the account holder's identity

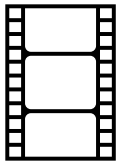
3

IDENTITY THIEF

- Identity theft occurs when someone uses another person's personal identifying information, like their name, identifying number, or credit card number, without their permission, to commit fraud or other crimes

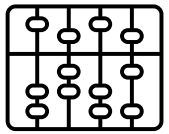
MACHINE LEARNING STRUCTURE

SUPERVISED MACHINE LEARNING



Text, documents, images, and audio used for training

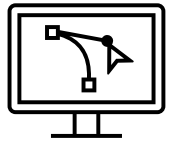
Labels



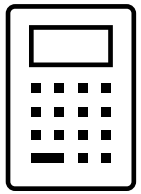
New - Text, documents, images, and audio for testing

Attribute vector

Attribute vector



Machine learning



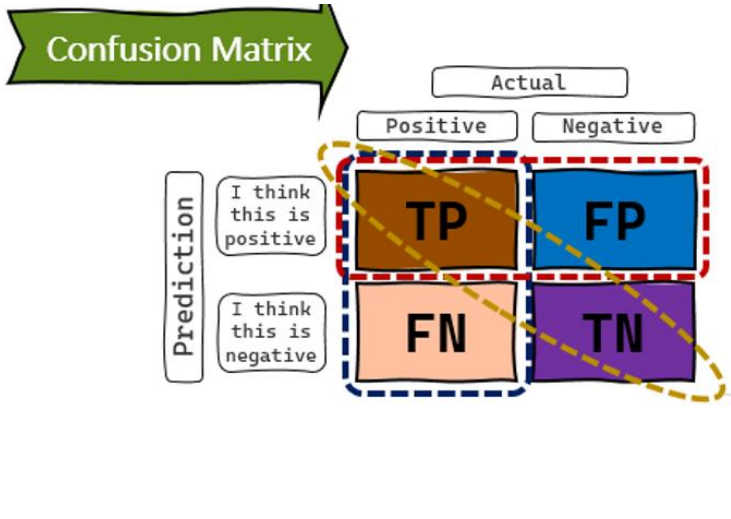
Models

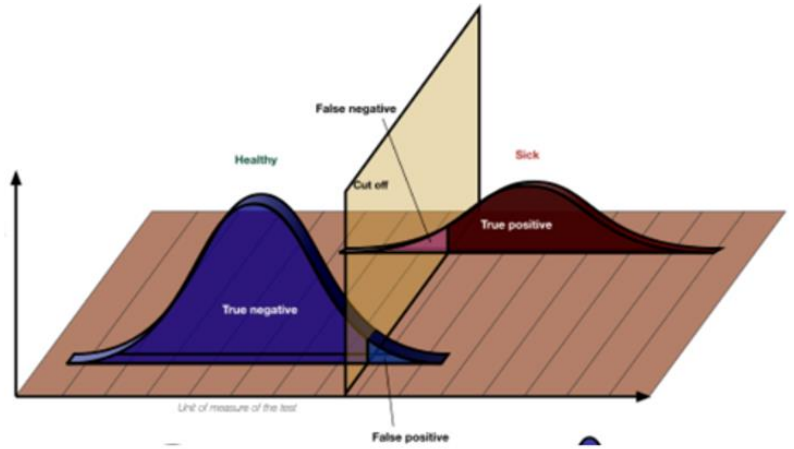
FRAUD ANALYSIS AND DETECTION SCENARIOS

1. Transaction outside trading hours (mean of time follows the von-mises distribution)
2. Customers successfully and unsuccessfully logged in 12 hours
3. Speed using a mobile banking app
4. Strange/new browsers that customers do not often use
5. Transaction locations in xxx minutes

6. Login fails in recent 30 days
7. Login fails in xxx months
8. Flags notified the new browsers used by the customers
9. Internet Service Provider
10. Mean of transaction time that follows Von-mises distribution
11. Standard deviation of transaction time that follows Von-mises distribution
12. Others

PREDICTIVE MODEL EVALUATION CRITERIA – CONFUSION MATRIX





RECALL
A high recall means a high rate of cases missed by the detection model

PRECISION
A high precision rate means a high prediction rate by the detection model

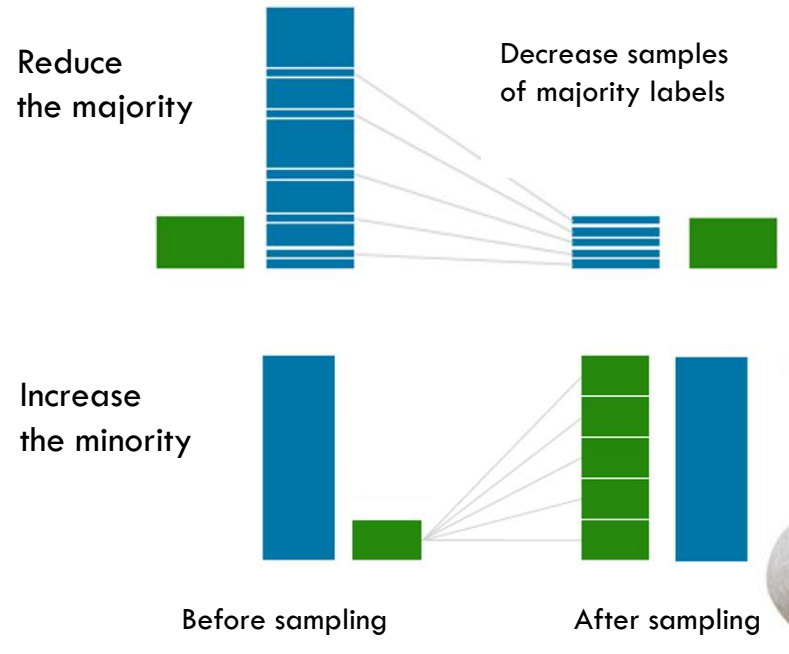
ACCURACY
The accuracy rate of correct prediction for both cases (fraud and non fraud)

THE SOLUTION TO THE IMBALANCE DATA

Class imbalance promotes a huge challenge in detecting the characteristics of fraudulent activities and extracting fraud patterns

Random Oversampling is used to balance classes by simply replicating observations as needed until the balance between classes is reached. Our aim is to modify the behavior of the classification model to concentrate on both the minority class (fraud) and majority class (legitimate) equally

Under sampling: The straightforward way is on data-level like over- or undersampling is used to balance the classes before applying any classification algorithm



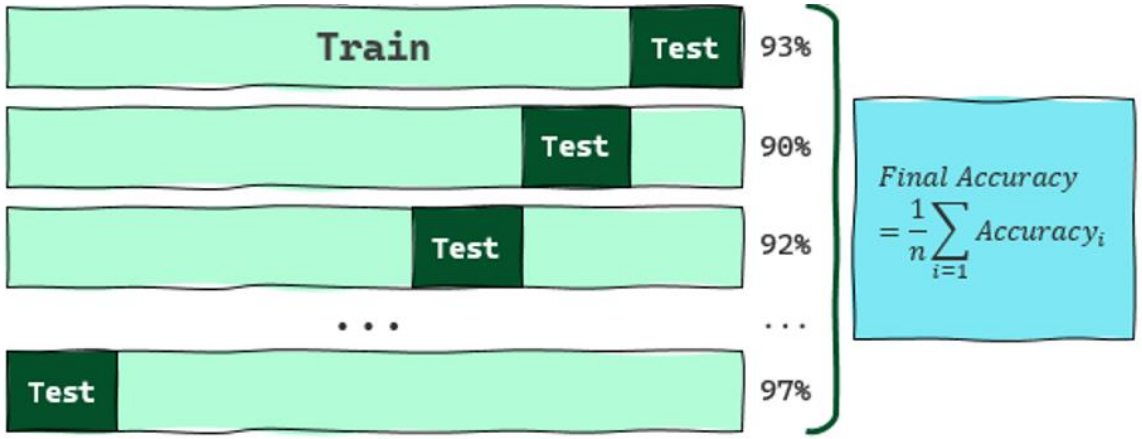
TRAINING MODEL AND CROSS VALIDATION

Cross-validation is a technique for evaluating ML models by training several ML models on subsets of the available input data and evaluating them on the complementary subset of the data

Use cross-validation to detect overfitting, ie, failing to generalize a pattern

There are some common methods that are used for cross-validation

1. Validation Set Approach
2. Leave-P-out cross-validation
3. Leave one out of cross-validation
4. K-fold cross-validation
5. Stratified k-fold cross-validation



The steps for k-fold cross-validation are:

- Split the input dataset into K groups
- For each group:
  - Take one group as the reserve or test data set
  - Use the remaining groups as the training dataset
  - Fit the model on the training set and evaluate the performance of the model using the test set

Note:

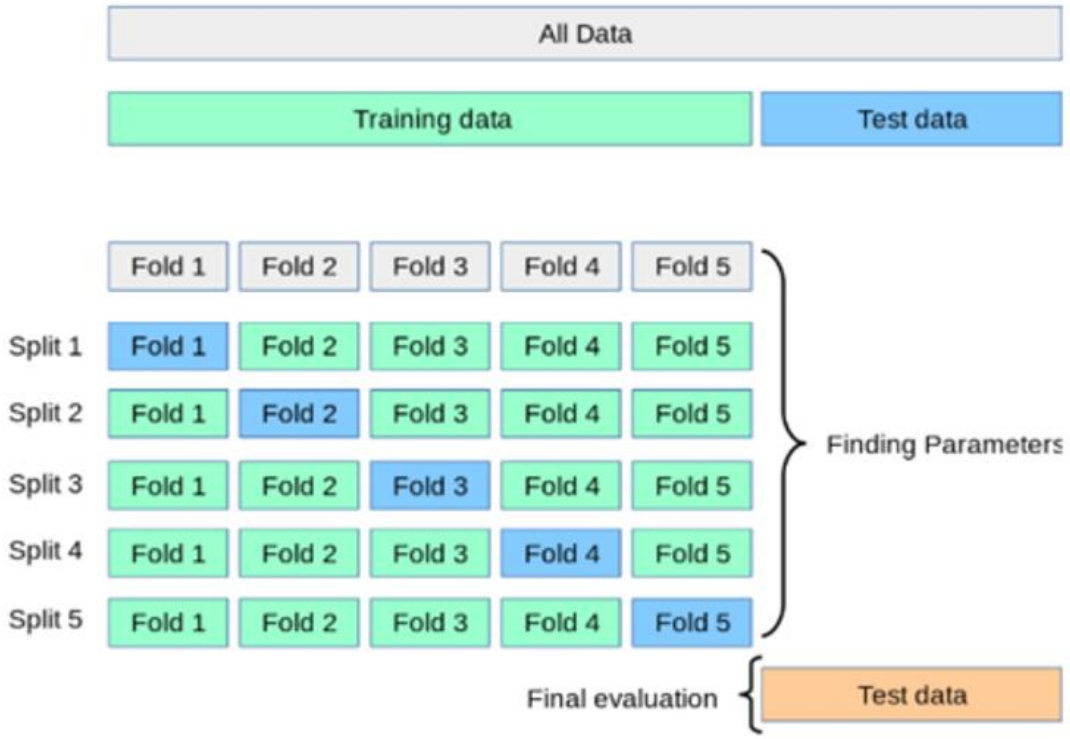
- The approach is suited for small- to modestly-sized datasets and/or models that are not too computationally costly to fit and evaluate
- Each time the procedure is run, a different split of the dataset into k-folds can be implemented, and in turn, the distribution of performance scores can be different, resulting in a different mean estimate of model performance

TRAIN THE MODEL WITH MULTIPLE TIMES

A single run of the k-fold cross-validation procedure may result in a noisy estimate of model performance. Different splits of the data may result in very different results.

Repeated k-fold cross-validation provides a way to improve the estimated performance of a machine-learning model. This involves simply repeating the cross-validation procedure multiple times and reporting the mean result across all folds from all runs.

This mean result is expected to be a more accurate estimate of the true unknown underlying mean performance of the model on the dataset, as calculated using the standard error



MODEL SELECTION

Visualization of recall and precision

- X-axis: Precision, Y-axis: Recall on the validation dataset of every sampling

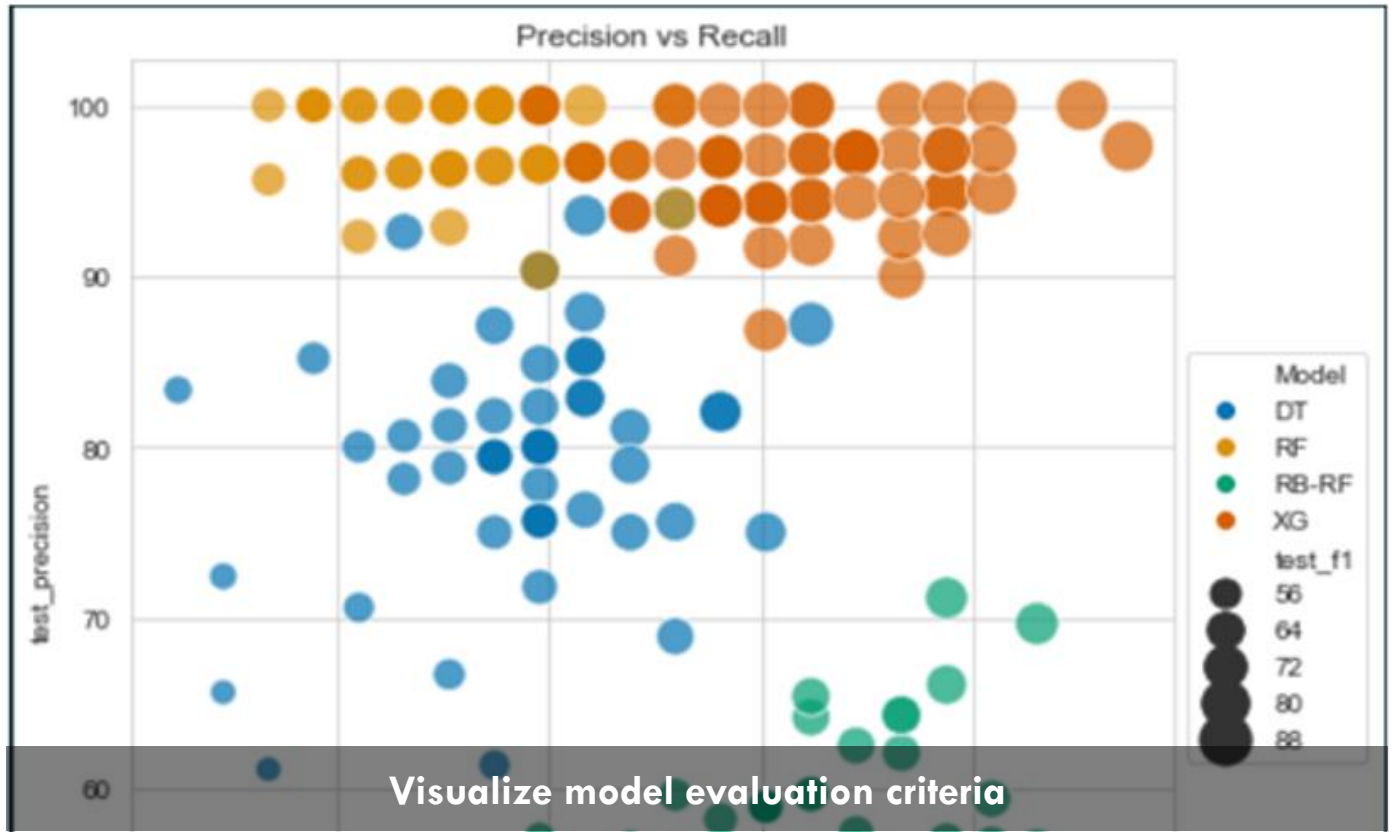
- The colored balls are the results of the pairs: recall and precision

- Each color of the balls is a different ML algorithm

- The location of the balls (1st quarter – on the left) shows the prediction power of the ML model

- The bigger the ball, the higher F1-score

Perform continuous sampling on the dataset representing the evaluation criteria for each training time



Visualize model evaluation criteria

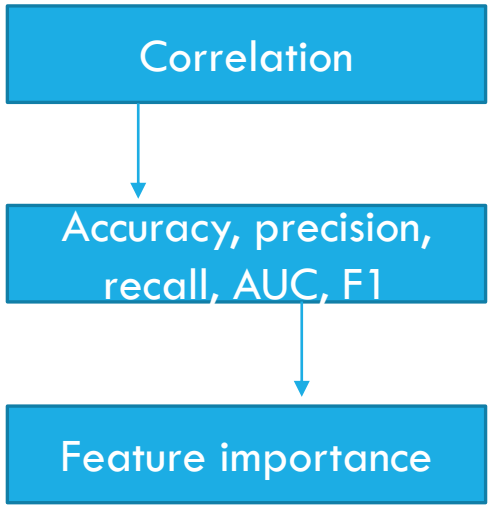
XGboost is the best model of all, according to the above criteria

VARIABLE SELECTION

A long list of variables consumes various resources to build a model and causes misleading predictions

Therefore, after selecting Xgboost with resampling (increase minority), select the most important variables to increase the model calculation rate but still keep good prediction power of the model

Selection criteria

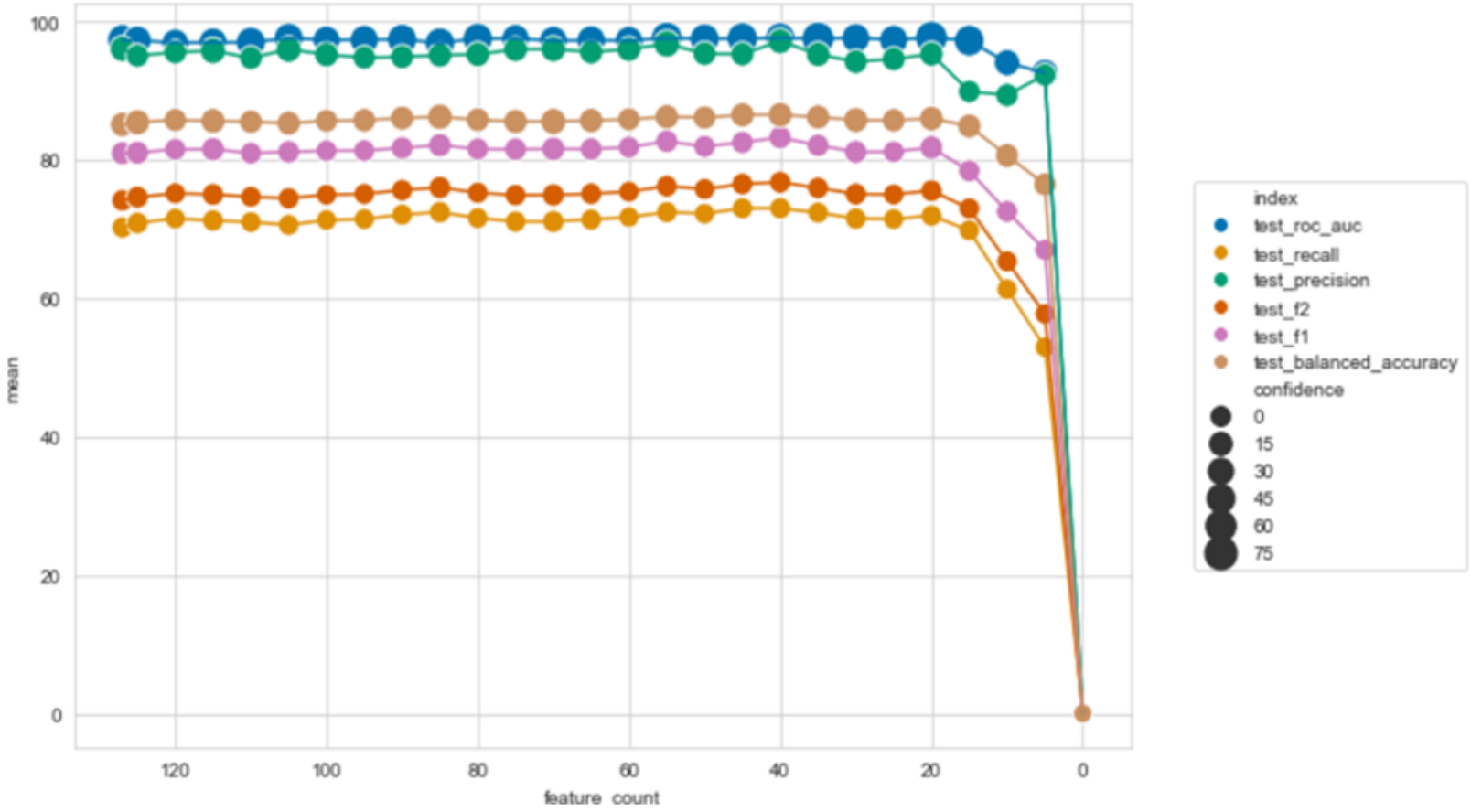


FILTER AND REMOVE VARIABLES

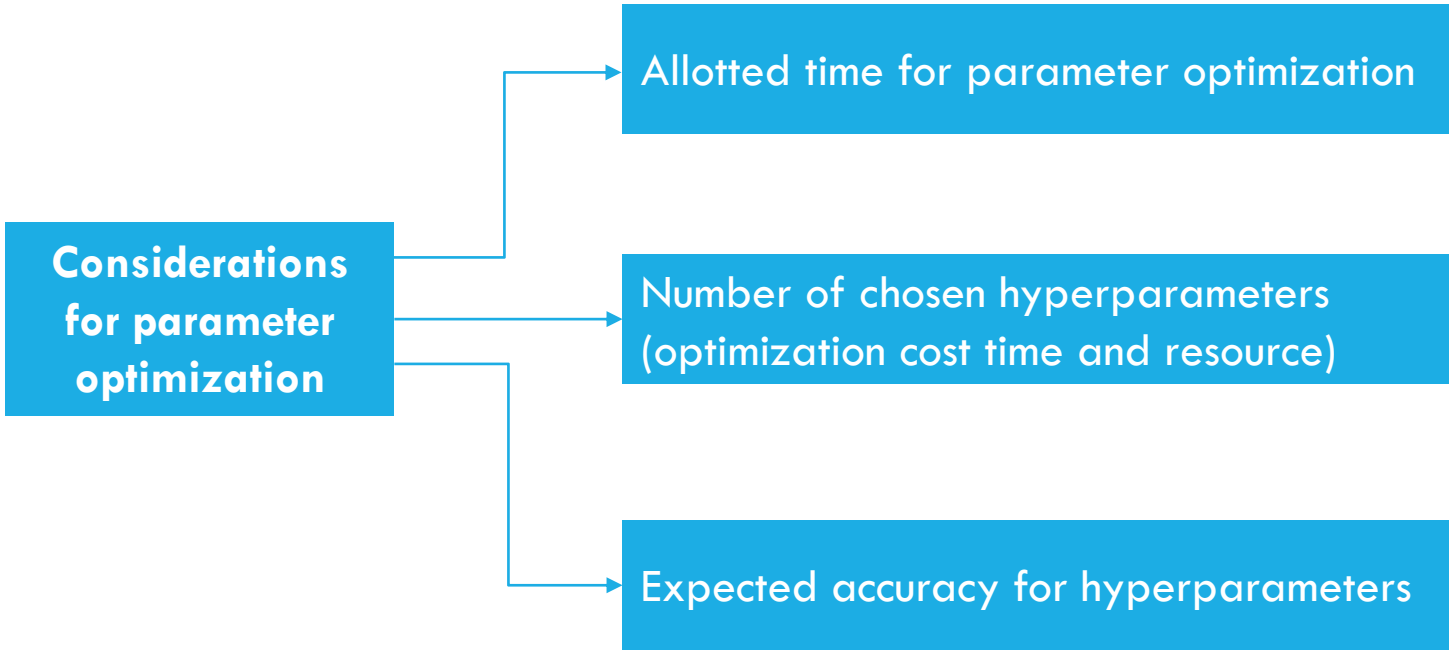
Filter and remove variables to select the best variables that have good predictive power for the model

Using the set of criteria (Accuracy, Precision, Recall, UAC, F1 score) for every filter then compare with the model that consists of all the variables

Each colored line presents the selection criteria after each filter



HYPERPARAMETER OPTIMIZATION



Hyperparameter optimization, also called hyperparameter tuning, is the process of searching for a set of hyperparameters that gives the best model results on a given dataset

- In machine learning, a hyperparameter refers to a parameter that is not learned from the training data but is set by the practitioner before the training process

•The "hyper-" prefix implies that they are higher-level parameters that control the learning process.

•Some examples of hyperparameters include:

- Number of hidden layers in a neural network
- Number of leaves of a decision tree
- Learning rate of a gradient descent
- The ratio between the training set and test set

- Tuning hyperparameters helps machine learning models generalize well. Generalization refers to the ability of the model to perform well on training data as well as on new data. A model fails to generalize due to:
  - Overfitting
  - Underfitting

THE END
THANK YOU!