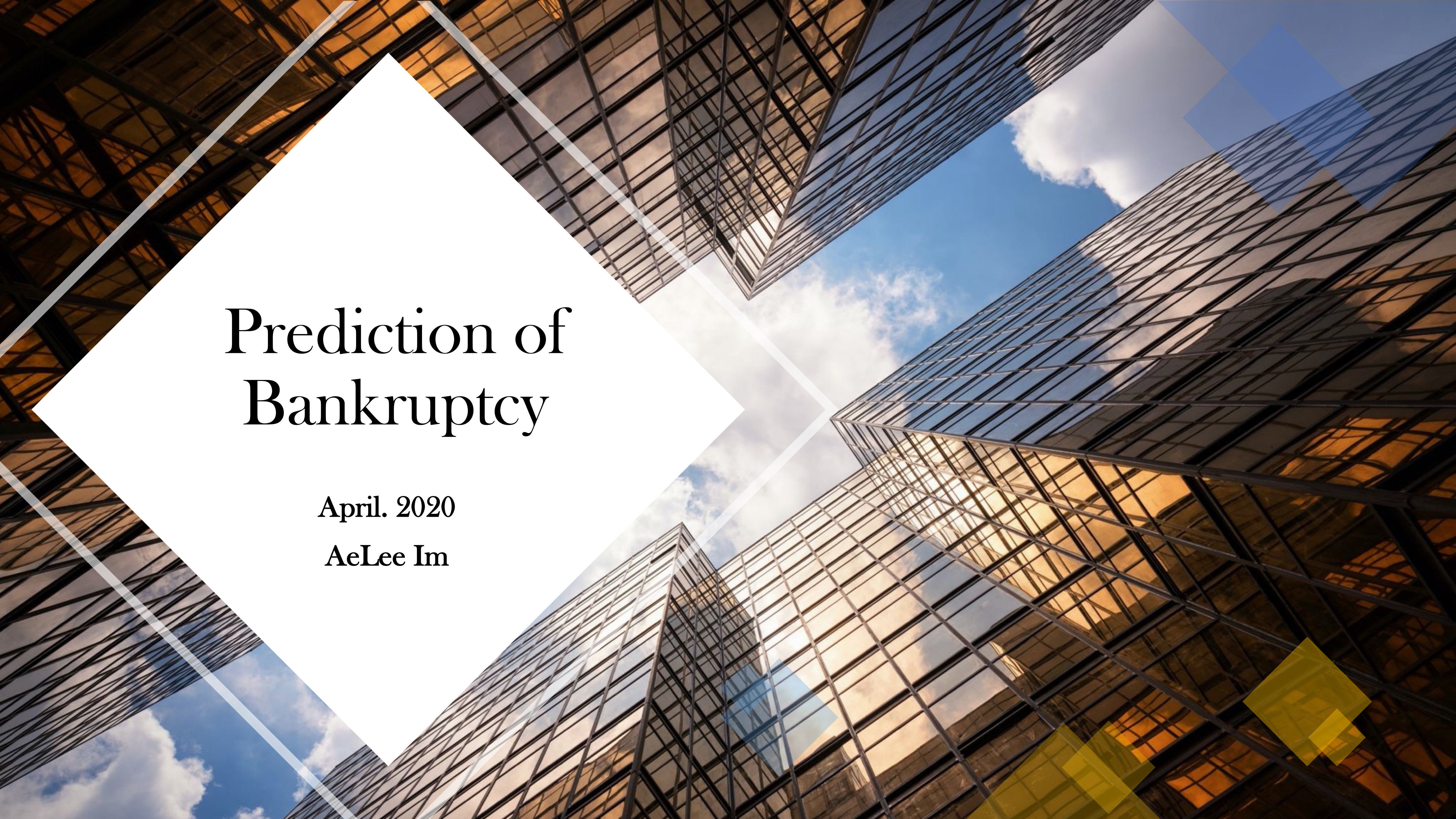


# Prediction of Bankruptcy

April. 2020

AeLee Im

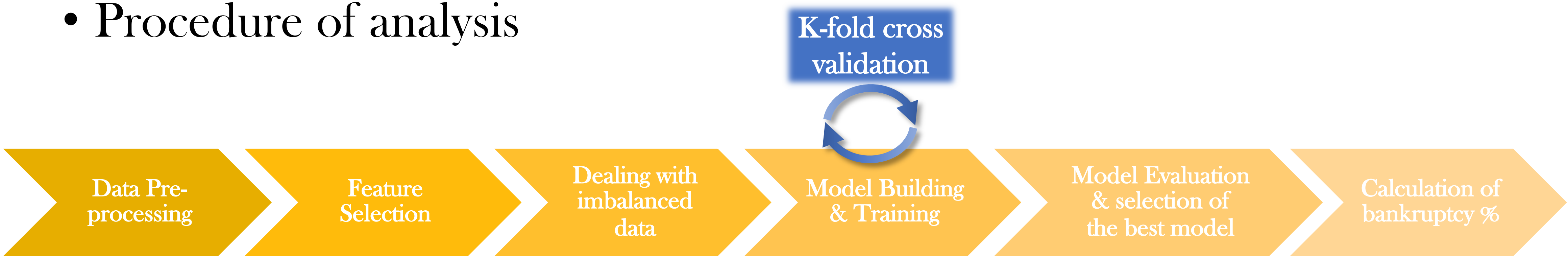
# Contents

- Project overview
    - Objective
    - Procedure
- Analysis methods
- Analysis results
- Next steps
- Appendix

# Project Overview

- Objective of project : Prediction of bankruptcy of company
  - Forecast the likelyhood of bankruptcy of companies
  - What is the probability of bankruptcy after 1 year and  2, 3 years?

- Procedure of analysis



K-fold cross validation

Data Pre-processing

Feature Selection

Dealing with imbalanced data

Model Building & Training

Model Evaluation & selection of the best model

Calculation of bankruptcy %

Data Pre-processing:
- Missing value analysis
- Missing value imputation
  - K-NN
  - Random Forest
- Correlation analysis

Feature Selection:
- by Boruta algorithm
- Selection of the most important features

Dealing with imbalanced data:
- by SMOTE sampling

Model Building & Training:
- 5 Models
  - G. Naive Bayes
  - Logistic Regression
  - Decision Tree
  - Random Forest
  - Balanced Bagging

Model Evaluation & selection of the best model:
- by 5 metrics
  - Accuracy
  - AUC
  - Precision
  - Recall
  - Confusion Metrix with TP,TN,FP,FN

Calculation of bankruptcy %:
- After 1year, 2,3 years

# Analysis Methods

## How to deal with missing values?

✓Impute missing value with k-NN, Random Forest imputation technique

**Benefit of solution**

Reduce the loss of data & beneficial imputation for high dimensional data

## How to deal with highly correlated & numerous features?

- Without removing correlated features, choose the Boruta selection algorithm

Considering of correlation between features during selection of important features

## How to deal with imbalanced dataset?

- Make each class balanced by SMOTE oversampling technique

No loss of data via oversampling & preparing well balanced dataset for high dimensional data

## How to deal with not enough training data?

- Training with every data and iteration with K-fold cross validation

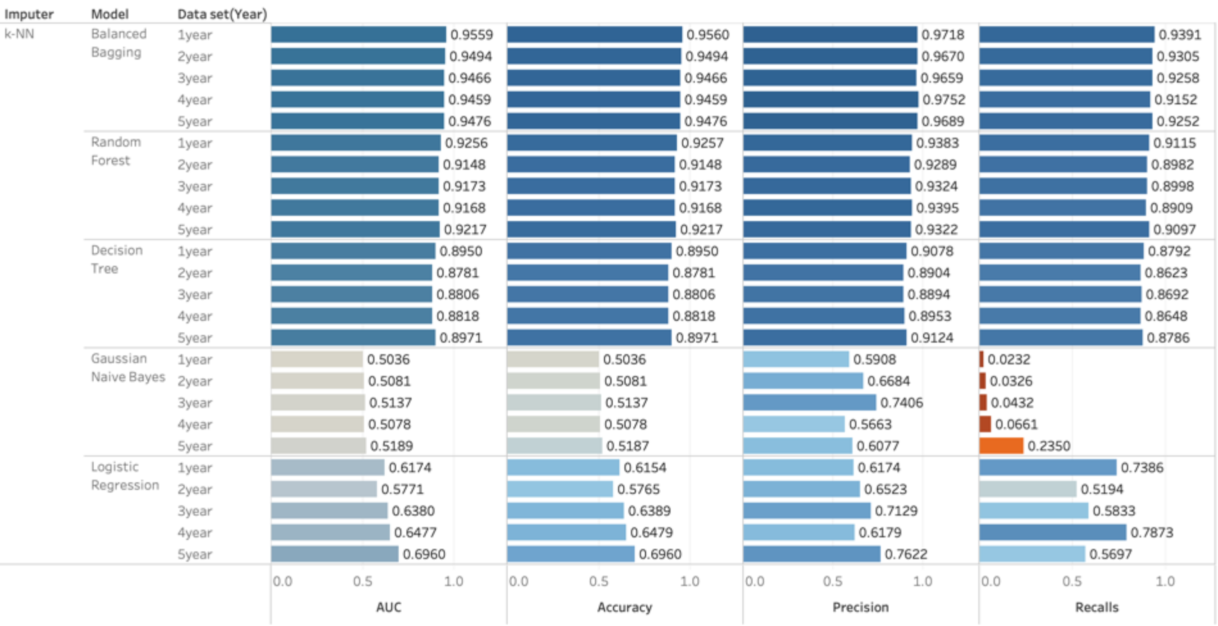
No bias with certain part of dataset & reliable model training results (No over-/underfitting problem)

# Analysis Results : Model comparision & selection

- The best performed model : Balanced Bagging Model
  - ✓ The most accurate prediction performance with highest metrics score (e.g. AUC 0.94 ~ 0.95 etc.)
  - ✓ Excellent performance for every dataset (1year ~ 5year)
  - ✓ No difference between k-NN and Random Forest imputation methods

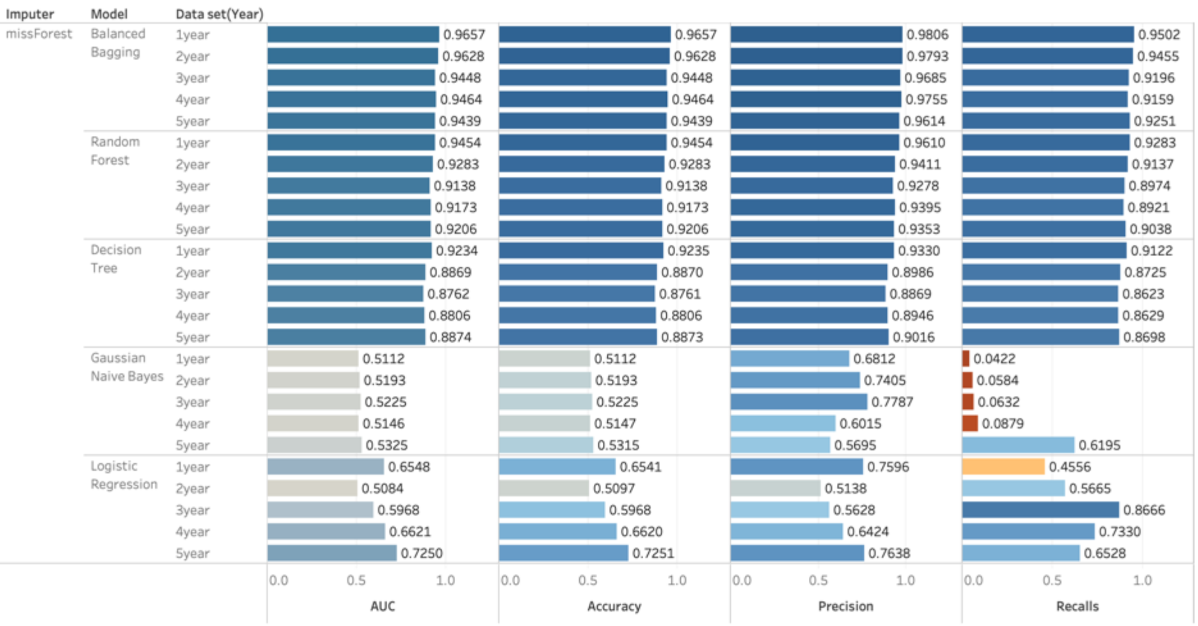
→ **Final Model Decision : Balanced Bagging Model (trained with Random Forest imputed dataset)**

**Model performance with k-NN imputed dataset**

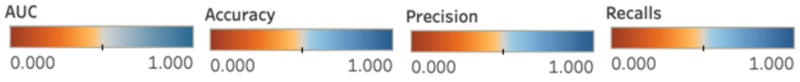


Evaluation metrics

**Model performance with Random Forest(missForest) imputed dataset**



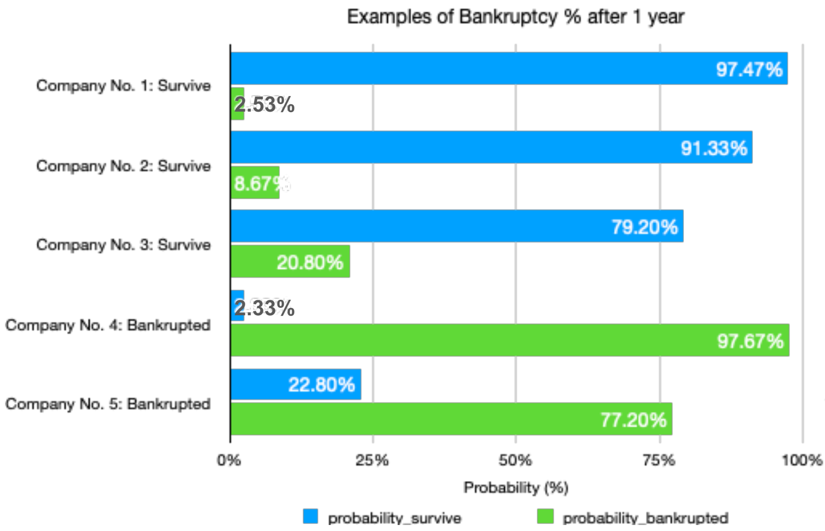
Evaluation metrics

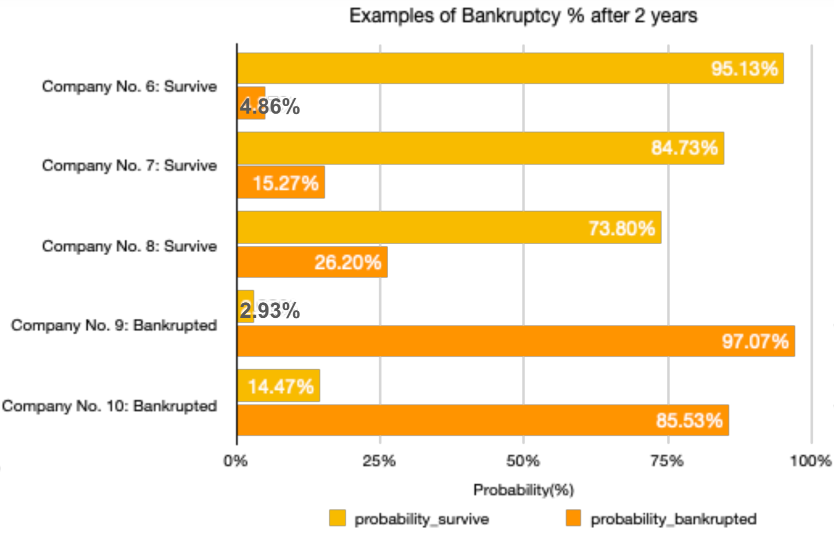


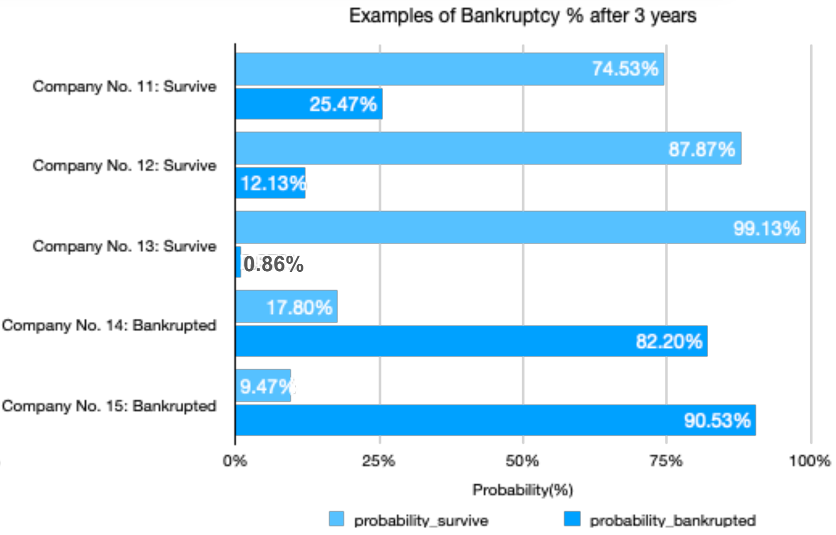
# Analysis Results : Computation of Bankruptcy %

- Test computation of chance of bankruptcy (%) after 1year, 2, 3 years
  - ✓ By applying selected prediction model : Pre-trained model with Balanced Bagging algorithm
  - ✓ Shows the probability(%) of survive and bankruptcy as an evidence of bankruptcy prediction

**Q. Will Company be bankrupted after 1 year? Or 2 years? Or 3 years?**


Examples of Bankruptcy % after 1 year


Examples of Bankruptcy % after 2 years

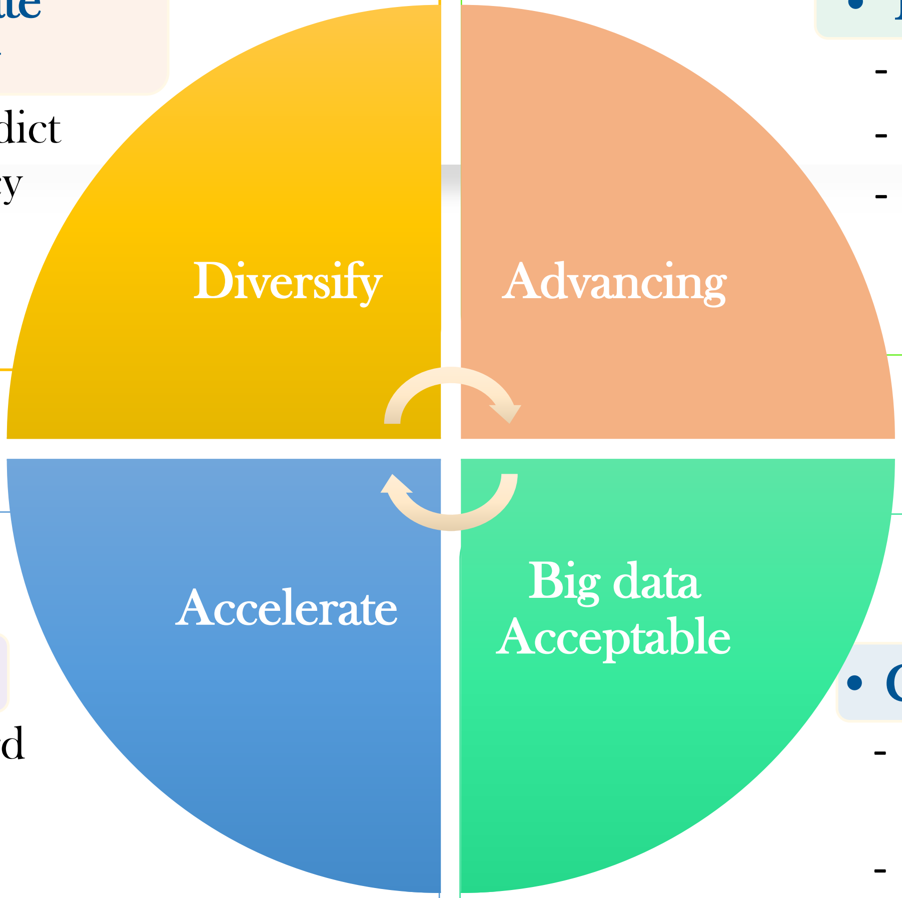

Examples of Bankruptcy % after 3 years

According to the prediction analysis result,
→ Company No. 1 will be survived with 97.47% probability after 1 year.
→ However, Company No. 10 will be bankrupted with 85.53% probability after 2 years!

# Next steps

- Additional prediction of accurate date/month/year of bankruptcy
  - Applying Time series data to predict the precise moment of bankruptcy
  - Q. When this company will be bankrupted?

- Try another Ensemble model
  - Boosting method (e.g. Xgboost)
  - Beneficial for high-bias data
  - Compare with our current best model (i.e. by Bagging method) for further model enhancement



- Real time monitoring available
  - Generate auto-updating dashboard
  - Deployment of developed model with implementation of database
  - No waiting time for reporting

- Convert to Cloud based analysis
  - With big data analysis systems (e.g. Hadoop, HIVE, Apache)
  - Deep learning approach available due to increase of data volume

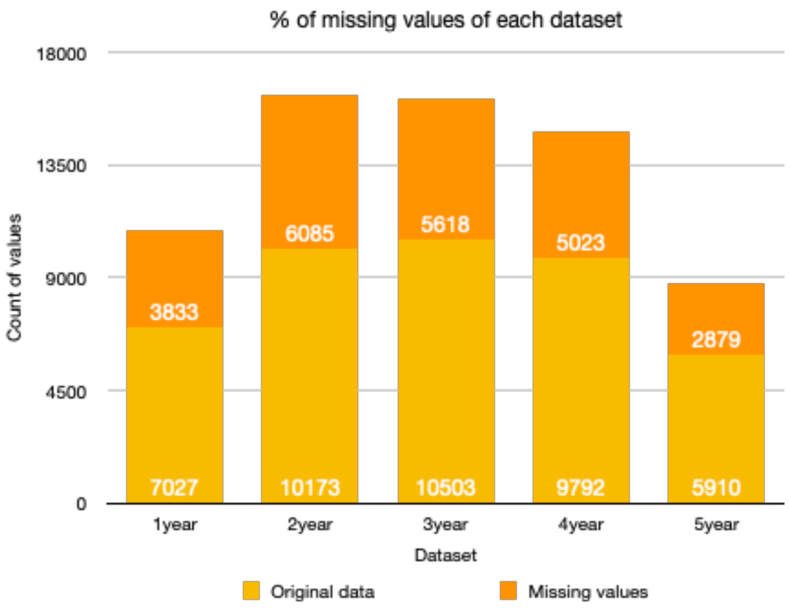
# Appendix.

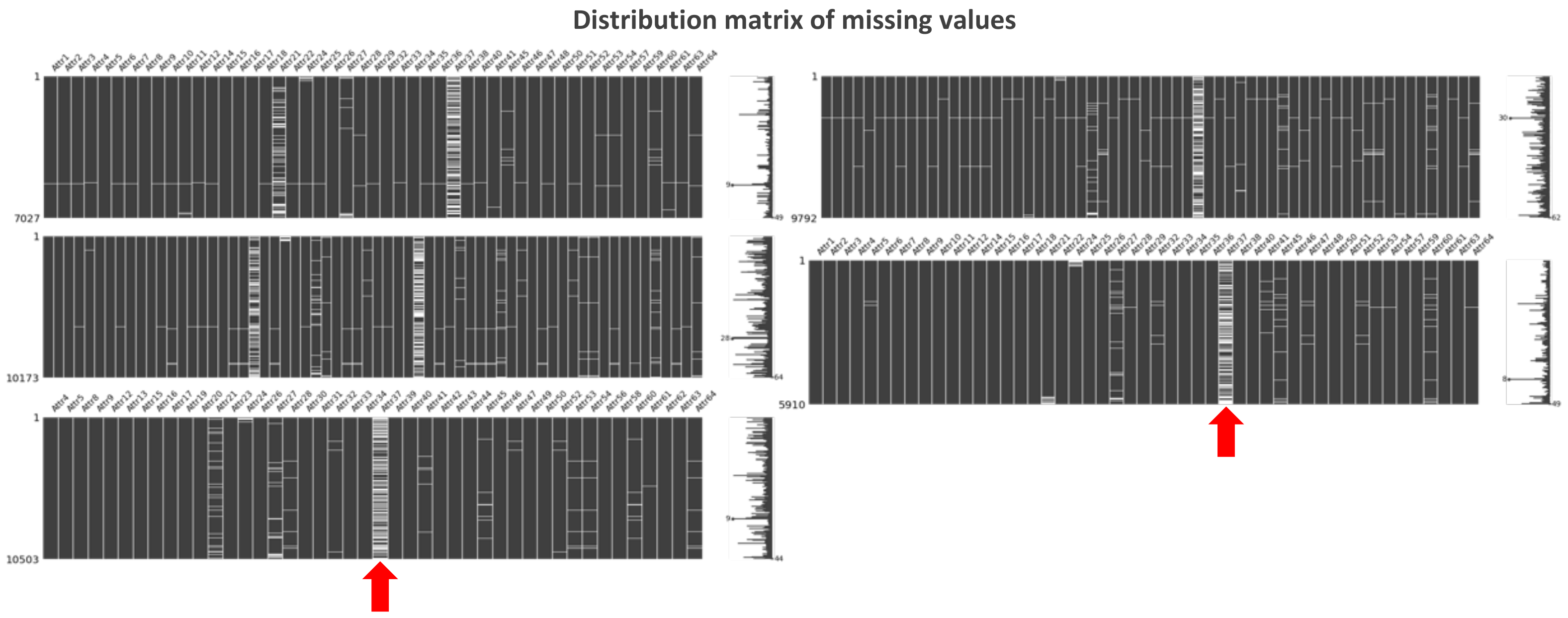
## Results of data processing

1. Missing value analysis
2. Correlation analysis
3. Data balancing analysis
4. Feature selection

# Appendix. 1. Missing value analysis

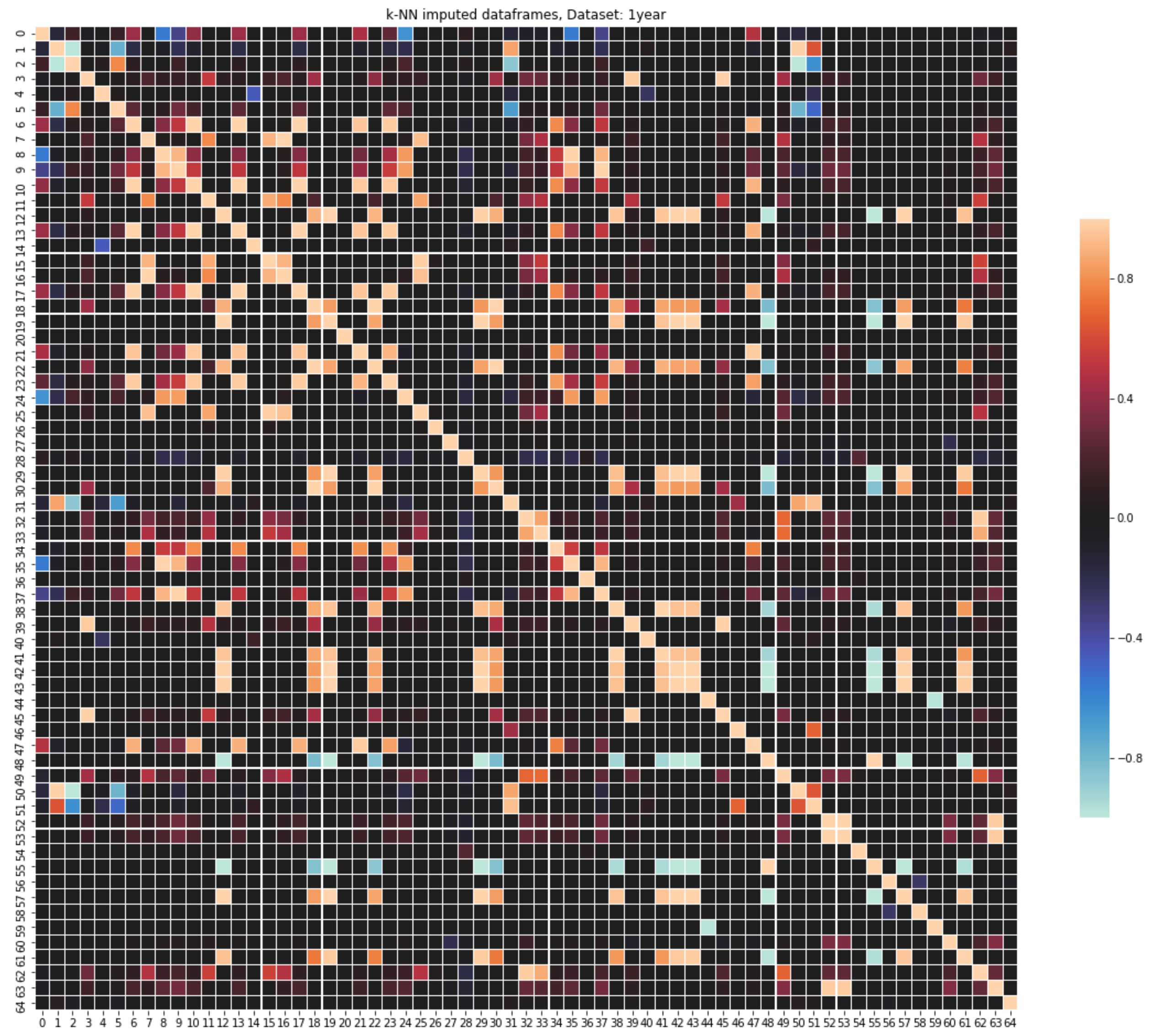
- Original dataset has high volume of missing values (53.99%)
  - ✓ Each dataset has around 50% of missing values (48.71 ~ 59.82%)
  - ✓ Most of missing values are concentrated on certain attributes
    (e.g. Attr 37 has 80.99% of missing value among the total missing value of whole dataset)


% of missing values of each dataset


Distribution matrix of missing values

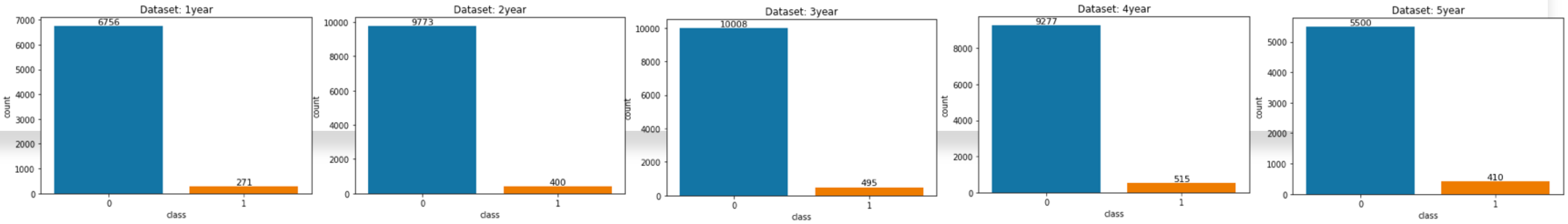
# Appendix. 2. Correlation analysis

- How to know correlation from this heat map?
  - Dark area : No correlated cells
  - Red/Light Red : Positive correlated cells
  - Blue/Light Blue : Negative correlated cells
- How's the correlation of our dataset?
  - 39 out of 64 features(60.93% of data) shows high correlation over +/- 0.95
  - Highly correlated data(high bias)



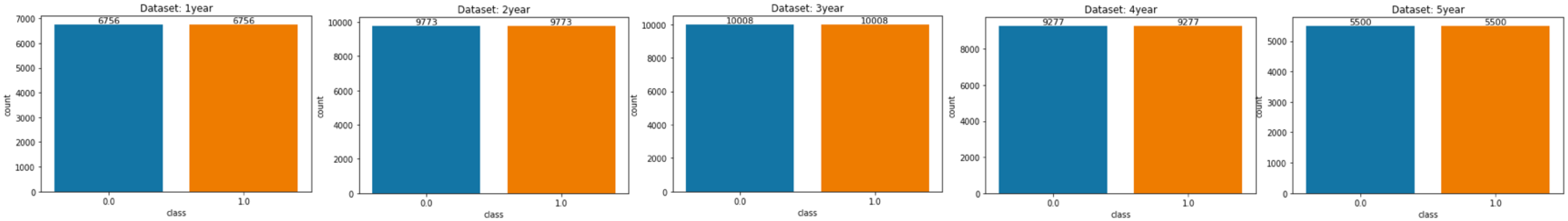
# Appendix. 3. Data balancing analysis

## <Before : Highly imbalanced dataset>



| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Shape | 6756 | 271 | 9773 | 400 | 10008 | 495 | 9277 | 515 | 5500 | 410 |

## <After : Balanced dataset after SMOTE sampling>



| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Shape | 6756 | 6756 | 9773 | 9773 | 10008 | 10008 | 9277 | 9277 | 5500 | 5500 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| FALSE | FALSE | FALSE | FALSE | TRUE | TRUE | FALSE | FALSE | TRUE | FALSE | FALSE | TRUE |
| 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 |
| TRUE | FALSE | FALSE | TRUE | FALSE | FALSE | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 |
| FALSE | TRUE | TRUE | FALSE | TRUE | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE |
| 37 | 38 | 39 | 40 | 41 | 42 | 43 | 44 | 45 | 46 | 47 | 48 |
| FALSE | TRUE | FALSE | FALSE | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE | FALSE | FALSE |
| 49 | 50 | 51 | 52 | 53 | 54 | 55 | 56 | 57 | 58 | 59 | 60 |
| FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE |
| 61 | 62 | 63 | 64 | | | | | | | | |
| FALSE | FALSE | FALSE | FALSE | | | | | | | | |

# Appendix 4-1. Feature selection

- Dataset 1year
- TRUE : Feature selected (17)
- FALSE : Not selected(47)

|  |  |  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| FALSE | FALSE | FALSE | FALSE | TRUE | TRUE | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE |
| 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 |
| FALSE | TRUE | FALSE | TRUE | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | TRUE |
| 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 |
| FALSE | TRUE | TRUE | FALSE | TRUE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE |
| 37 | 38 | 39 | 40 | 41 | 42 | 43 | 44 | 45 | 46 | 47 | 48 |
| TRUE | TRUE | TRUE | FALSE | TRUE | TRUE | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE |
| 49 | 50 | 51 | 52 | 53 | 54 | 55 | 56 | 57 | 58 | 59 | 60 |
| FALSE | FALSE | TRUE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE |
| 61 | 62 | 63 | 64 |  |  |  |  |  |  |  |  |
| FALSE | FALSE | FALSE | FALSE |  |  |  |  |  |  |  |  |

# Appendix 4-2. Feature selection

- Dataset 2year
- TRUE : Feature selected (18)
- FALSE : Not selected (46)

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| FALSE | FALSE | FALSE | FALSE | TRUE | TRUE | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE |
| 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 |
| TRUE | FALSE | FALSE | TRUE | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | TRUE | TRUE |
| 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 |
| TRUE | TRUE | TRUE | FALSE | TRUE | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE |
| 37 | 38 | 39 | 40 | 41 | 42 | 43 | 44 | 45 | 46 | 47 | 48 |
| FALSE | TRUE | FALSE | FALSE | TRUE | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE |
| 49 | 50 | 51 | 52 | 53 | 54 | 55 | 56 | 57 | 58 | 59 | 60 |
| FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE |
| 61 | 62 | 63 | 64 | | | | | | | | |
| FALSE | FALSE | FALSE | FALSE | | | | | | | | |

# Appendix 4-3. Feature selection

- Dataset 3year
- TRUE : Feature selected (17)
- FALSE : Not selected (47)

|  |  |  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| FALSE | FALSE | FALSE | FALSE | TRUE | TRUE | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE |
| 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 |
| TRUE | FALSE | TRUE | TRUE | FALSE | FALSE | FALSE | FALSE | TRUE | TRUE | FALSE | TRUE |
| 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 |
| TRUE | TRUE | TRUE | FALSE | TRUE | FALSE | FALSE | FALSE | FALSE | TRUE | TRUE | FALSE |
| 37 | 38 | 39 | 40 | 41 | 42 | 43 | 44 | 45 | 46 | 47 | 48 |
| FALSE | TRUE | TRUE | FALSE | TRUE | TRUE | FALSE | FALSE | TRUE | TRUE | FALSE | TRUE |
| 49 | 50 | 51 | 52 | 53 | 54 | 55 | 56 | 57 | 58 | 59 | 60 |
| FALSE | FALSE | TRUE | FALSE | FALSE | TRUE | TRUE | FALSE | FALSE | FALSE | FALSE | FALSE |
| 61 | 62 | 63 | 64 |  |  |  |  |  |  |  |  |
| FALSE | FALSE | FALSE | FALSE |  |  |  |  |  |  |  |  |

# Appendix 4-4. Feature selection

- Dataset 4year
- TRUE : Feature selected (25)
- FALSE : Not selected (39)

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| TRUE | FALSE | FALSE | FALSE | FALSE | TRUE | TRUE | FALSE | FALSE | FALSE | FALSE | TRUE |
| 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 |
| TRUE | TRUE | TRUE | TRUE | FALSE | TRUE | FALSE | FALSE | TRUE | TRUE | FALSE | TRUE |
| 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 |
| TRUE | TRUE | TRUE | FALSE | TRUE | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE |
| 37 | 38 | 39 | 40 | 41 | 42 | 43 | 44 | 45 | 46 | 47 | 48 |
| FALSE | TRUE | TRUE | FALSE | TRUE | TRUE | TRUE | FALSE | TRUE | TRUE | FALSE | FALSE |
| 49 | 50 | 51 | 52 | 53 | 54 | 55 | 56 | 57 | 58 | 59 | 60 |
| FALSE | FALSE | TRUE | FALSE | FALSE | TRUE | TRUE | FALSE | FALSE | FALSE | FALSE | TRUE |
| 61 | 62 | 63 | 64 | | | | | | | | |
| FALSE | FALSE | FALSE | FALSE | | | | | | | | |

# Appendix 4-5. Feature selection

- Dataset 5year
- TRUE : Feature selected (28)
- FALSE : Not selected (36)