Bankruptcy Prediction: A Comparative Study on various Missing Value Imputation and Sampling Techniques

Chirag Bhatt
Department of Mathematics
Indian Institute of Technology, Delhi
Delhi 110016, India
mt1180750@maths.iitd.ac.in

Subhalingam D
Department of Mathematics
Indian Institute of Technology, Delhi
Delhi 110016, India
mt1180770@maths.iitd.ac.in

Abstract

This project aims towards experimenting with different methods for improving the existing models in the literature used for predicting corporate bankruptcy for Polish companies. More specifically, we aim to get better recall values as the existing models are suffering from poor recall. To tackle class imbalance, we use sampling techniques like SMOTE, RUS, SMOTE-ENN, and to impute missing values, we use Simple Imputer, MissForest, KNN. The performance of the models improved with SMOTE and simple imputation.

Keywords—Bankruptcy, Classification, SMOTE, Feature Selection, Imputation, Sampling

1 Introduction

Bankruptcy is a legal proceeding carried out to allow businesses unable to make payment to its creditors get freedom from their debts, while providing creditors an opportunity for repayment based on the business's assets. Only in USA more than 21 thousand businesses filed for bankruptcy in year 2020 [Investopedia]. When a company becomes bankrupt and goes for into liquidation, stock shares becomes practically worthless. The common shareholders may, at best, get a portion of their value back but most often they do not get anything at all. Some well-known examples of Indian companies that went bankrupt include *Yes Bank*, *Jet Airways* and *Videocon*.

Thus, knowing the financial health of a firm and predicting whether it has possibility of getting bankrupt in near future is of prime importance for the investors and the creditors.

1.1 Problem Statement

The Corporate Bankruptcy prediction project aims to predict the possibility of a firm getting bankrupt in thr next 1 to 5 years, given the financial measures of the firm, like *net profit/sales*, *total assets/total liabilities for a financial year*.

1.2 Past Works

Zikeba et al. (2016) prepared the Polish companies bankruptcy dataset (discussed further in Section 2) and proposed an ensemble Boosted Decision Tree model with synthetic feature generation for the prediction task. Quynh and Thi Lan Phuong (2020) proposed a new approach to take best ensemble models as base models and to make predictions based on "*fair voting*", and a procedure to handle missing values using Random Forest algorithm. Ren (2020) searched for common financial indicators of bankruptcy by comparing diverse financial markets in China and Poland.

1.3 Motivation

A preliminary data analysis showed a very high class imbalance and presence of missing values in the data. The class imbalance is natural as the number of companies that actually get bankrupt is very less. Zikeba et al. (2016) used accuracy as the metric to evaluate their models – however, given such a high class imbalance, a model that is completely biased towards '*not bankrupt*' would also have an accuracy of about 95%. Quynh and Thi Lan Phuong (2020) proposed an approach to impute missing values using Random Forests for the same task, which motivated us to explore more techniques in this area. Ren (2020) showed that the best achieved recall value (without sampling) was 56% for the fifth year, which showed the performance of the model towards '*bankrupt*' label.

1.4 Our Contributions

We explored and surveyed various sampling techniques and missing values imputation techniques in the literature for the bankruptcy prediction task.

In particular, for sampling, we experimented on: 1. Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002); 2. Random Under-Sampling of majority class (RUS); and 3. combined over-sampling and under-sampling using SMOTE and Edited Nearest Neighbours (SMOTE-ENN) (Batista et al., 2004). For missing values imputation, we experimented on: 1. Simple Imputation (replacing the missing values by mean); 2. MissForest (algorithm that operates on Random Forest) (Stekhoven and Buhlmann, 2011); and 3. KNN-based imputation (Troyanskaya et al., 2001), We performed a comparative study on the performance of the models by considering different combinations of the techniques listed above, aiming at improving the performance of prediction on '*bankrupt*' class.

2 Dataset

We use the "*Polish companies bankruptcy data Data Set*" available on UCI Machine Learning Repository[1] for this project.

2.1 Description

The bankrupt companies were analyzed in the period 2000-2012, while the still operating companies were evaluated from 2007 to 2013. Based on the collected data, five classification cases were distinguished depending on the forecasting period[2]:

- **1st Year**: the data contains financial rates from 1st year of the forecasting period and corresponding class label that indicates bankruptcy status after 5 years
- **2nd Year**: the data contains financial rates from 2nd year of the forecasting period and corresponding class label that indicates bankruptcy status after 4 years
- **3rd Year**: the data contains financial rates from 3rd year of the forecasting period and corresponding class label that indicates bankruptcy status after 3 years
- **4th Year**: the data contains financial rates from 4th year of the forecasting period and corresponding class label that indicates bankruptcy status after 2 years.
- **5th Year**: the data contains financial rates from 5th year of the forecasting period and corresponding class label that indicates bankruptcy status after 1 year.

The dataset contains numerical data for 64 economic measures and a class label $\{0, 1\}$ signifying whether the firm gets bankrupt in the forecasting period for each corporate company. The list of economic measures available is given in Table 1

3 Exploratory Data Analysis

3.1 Imbalanced Data

We found that the distribution of firms belonging to different classes in the given data set is highly skewed, with 95.1% companies belonging to class 0 (*not getting bankrupt*) and 4.9% belonging to class 1 (*getting bankrupt*). The distribution between classes for each year is listed in Table 2.

Year	#0	#1	Total
1	6756	271	7027
2	9773	400	10173
3	10008	495	10503
4	9277	515	9792
5	5500	410	5910

Table 2: Distribution of different classes in the dataset

3.2 Missing Values

We observed that 53% examples in the data set contains at least one attribute missing and overall 1.48% of total values are missing. Notably, *Attr37* had the highest percent of missing values (43%) for all 5 years. Following that, *Attr21* has the second highest percentage of missing values for year 1 and 2 (27.8%). Other attributes have missing values less than 7%. The top 3 attributes with highest percentage of missing values for each year is tabulated in Table 3.

3.3 Feature Selection

Feature Selection is performed to find out the attributes which are more significant in predicting the class. We used `sklearn.feature_selection` library which outputs a score (between 0-1) for each attribute signifying the relative importance of that attribute. The score sums up to 1. We performed

[1] https://archive.ics.uci.edu/ml/datasets/Polish+companies+bankruptcy+data

[2] It can be observed that the n^{th} year refers to "*bankruptcy status after* $(5 - n) + 1$ *years*", and a similar terminology is used throughout in the report

Attr	Description	Attr	Description
1	net profit / total assets	34	operating expenses / total liabilities
2	total liabilities / total assets	35	profit on sales / total assets
3	working capital / total assets	36	total sales / total assets
4	current assets / short-term liabilities	37	(current assets - inventories) / long-term liabilities
5	[(cash + short-term securities + receivables - short-term liabilities) / (operating expenses - depreciation)] * 365	38	constant capital / total assets
		39	profit on sales / sales
6	retained earnings / total assets	40	(current assets - inventory - receivables) / short-term liabilities
7	EBIT / total assets		
8	book value of equity / total liabilities	41	total liabilities / ((profit on operating activities + depreciation) * (12/365))
9	sales / total assets		
10	equity / total assets	42	profit on operating activities / sales
11	(gross profit + extraordinary items + financial expenses) / total assets	43	rotation receivables + inventory turnover in days
12	gross profit / short-term liabilities	44	(receivables * 365) / sales
13	(gross profit + depreciation) / sales	45	net profit / inventory
14	(gross profit + interest) / total assets	46	(current assets - inventory) / short-term liabilities
15	(total liabilities * 365) / (gross profit + depreciation)	47	(inventory * 365) / cost of products sold
16	(gross profit + depreciation) / total liabilities	48	EBITDA (profit on operating activities - depreciation) / total assets
17	total assets / total liabilities	49	EBITDA (profit on operating activities - depreciation) / sales
18	gross profit / total assets		
19	gross profit / sales	50	current assets / total liabilities
20	(inventory * 365) / sales	51	short-term liabilities / total assets
21	sales (n) / sales (n-1)	52	(short-term liabilities * 365) / cost of products sold)
22	profit on operating activities / total assets		
		53	equity / fixed assets
23	net profit / sales	54	constant capital / fixed assets
24	gross profit (in 3 years) / total assets	55	working capital
25	(equity - share capital) / total assets	56	(sales - cost of products sold) / sales
26	(net profit + depreciation) / total liabilities	57	(current assets - inventory - short-term liabilities) / (sales - gross profit - depreciation)
27	profit on operating activities / financial expenses		
		58	total costs /total sales
28	working capital / fixed assets	59	long-term liabilities / equity
29	logarithm of total assets	60	sales / inventory
30	(total liabilities - cash) / sales	61	sales / receivables
31	(gross profit + interest) / sales	62	(short-term liabilities *365) / sales
32	(current liabilities * 365) / cost of products sold	63	sales / short-term liabilities
		64	sales / fixed assets
33	operating expenses / short-term liabilities	*label*	1 for companies that got bankrupt and 0 for those which did not

Table 1: List of attributes in the dataset

	# NA	% NA
Attr37	2740	38.99
Attr21	1622	23.08
Attr27	311	04.42

(a) Year 1

	# NA	% NA
Attr37	451	44.41
Attr21	316	31.10
Attr27	70	06.93

(b) Year 2

	# NA	% NA
Attr37	4736	45.09
Attr21	807	07.68
Attr27	715	06.80

(c) Year 3

	# NA	% NA
Attr37	4442	45.36
Attr27	641	06.54
Attr60	614	06.27

(d) Year 4

	# NA	% NA
Attr37	2548	43.11
Attr27	391	06.61
Attr45	268	04.53

(e) Year 5

	% NA
Total missing values:	1.48%
Datapoints with at least one value missing:	53%

(f) Overall

Table 3: Distribution of missing values

feature selection for each year and averaged over them to find the top 10 most important attributes (Table 4). Notably, we observed that *Attr27* has very high score as compared to its successors, which corresponds to ”*profit on operating activities / financial expenses*”.

[H]	Attribute	Relative Score
1	Attr27	0.0918
2	Attr21	0.0257
3	Attr46	0.0223
4	Attr11	0.0202
5	Attr34	0.0194
6	Attr24	0.0189
7	Attr35	0.0188
8	Attr58	0.0181
9	Attr22	0.0180
10	Attr51	0.0176

Table 4: Top 10 attributes with highest significance

3.4 Correlation

Pairwise correlation between attributes were computed (Figure 2). For attribute pairs with very high correlation (> 0.9), attributes having higher percent of missing values are dropped (Section 4.1). Table 6 and Table 5 show top 10 attribute pairs with highest and least correlation respectively.

Attr #1	Attr #2	Correlation
Attr3	Attr51	-0.981
Attr2	Attr6	-0.836
Attr38	Attr2	-0.819
Attr2	Attr3	-0.814
Attr2	Attr10	-0.804
Attr2	Attr25	-0.735
Attr49	Attr43	-0.730
Attr49	Attr44	-0.724
Attr51	Attr6	-0.688
Attr51	Attr25	-0.686

Table 5: Attribute pairs with least correlation

Attr #1	Attr #2	Correlation
Attr7	Attr14	0.999
Attr17	Attr8	0.998
Attr23	Attr19	0.998
Attr46	Attr4	0.997
Attr19	Attr31	0.992
Attr31	Attr23	0.990
Attr44	Attr43	0.985
Attr62	Attr30	0.963
Attr11	Attr22	0.961
Attr22	Attr48	0.958

Table 6: Attribute pairs with highest correlation

4 Methodology

A schematic of the model pipeline is given in Figure 1 and is discussed in detail in the following subsections.

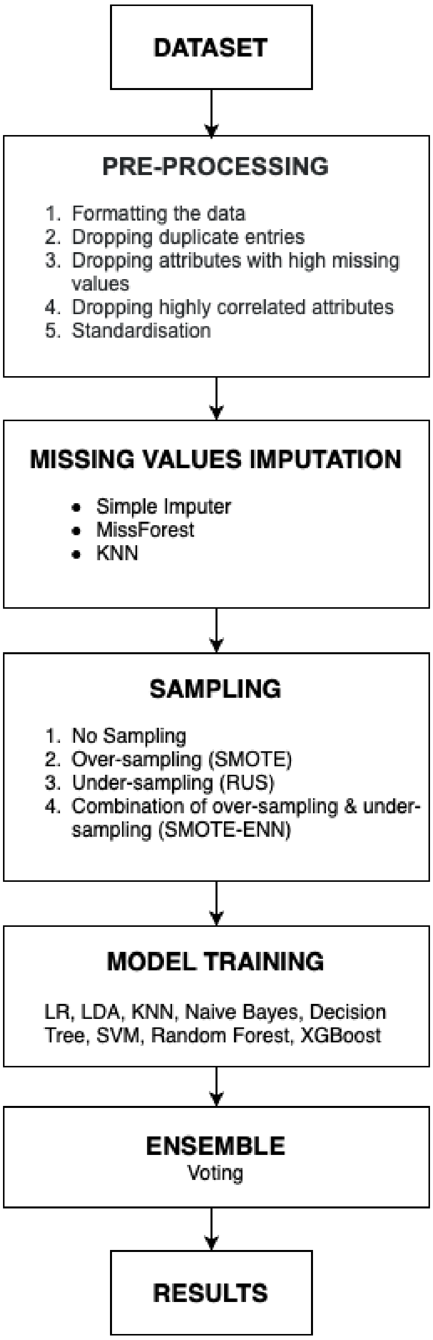

Figure 1: Model Pipeline

4.1 Pre-processing

The following tasks were performed as part of pre-processing:

- The data was converted into a `.csv` format so that data analysis and training could be performed with ease.
- The dataset contained duplicate entries. Such entries were dropped.
- `Attr37` had over 40% of the values missing and hence, was dropped. (Section 3.2.
- In some cases, a pair of attributes in the dataset were highly correlated with each other (Section 3.4). In such cases, the attributes with higher number of missing values were dropped. The list of attributes dropped are:

  {Attr3, Attr4, Attr6, Attr7, Attr15, Attr19, Attr32, Attr38, Attr43, Attr44, Attr49, Attr51, Attr60, Attr62}

- Some models required the data to be standardised (removing the mean and scaling to unit variance).

4.2 Missing values imputation

To tackle the missing values problem (Section 3.2), the following techniques were studied:

- **Simple Imputer**: Missing values are imputed with mean of each column in which the missing values are located.
- **MissForest Imputer**: Missing values are imputed using Random Forests in an iterative fashion (Stekhoven and Buhlmann, 2011).
- **KNN Imputer**: Missing values are imputed using k-Nearest Neighbors approach (Troyanskaya et al., 2001).

4.3 Sampling

To tackle the imbalance in class distributions (Section 3.1), over-sampling, under-sampling or a combination of both were performed and a comparative study was performed on the performance, along with the case where no sampling was performed.

- **Over-sampling**: Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002)
- **Under-sampling**: Random Under-Sampling of majority class (RUS)
- **Over-sampling and Under-sampling**: combined over-sampling and under-sampling using SMOTE and Edited Nearest Neighbours (SMOTE-ENN) (Stekhoven and Buhlmann, 2011).

4.4 Modelling

The goal of the experiment was to identify the best classification model in terms of F1-score. We took under consideration the following classification methods:

- **LR**: Logistic Regression
- **LDA**: Linear Discriminant Analysis (Altman, 1968)
- **KNN-5**: k-Nearest Neighbours (with 5 neighbours) (Guo et al., 2004)
- **KNN-10**: k-Nearest Neighbours (with 10 neighbours)
- **GNB**: Gaussian Naive Bayes
- **DT**: Decision Tree
- **SVC**: Support Vector machine Classifier (Cortes and Vapnik, 1995)
- **RFC**: Random Forest Classifier (Tin Kam Ho, 1995)
- **XGB**: eXtreme Gradient Boosting (Chen and Guestrin, 2016)

4.4.1 Voting

The top performing models were combined to build an ensemble model, in which a majority vote (*hard* voting) would be used to predict the class labels. The aim is to balance out the individual weaknesses of each model.

Based on the results, we chose **LR**, **DT**, **RFC**, **XGB** with individual weights as 1 and **KNN-10**, **SVC** with weights 0.5 each in the voting to build a "Voting" Classifier.

5 Experiment

5.1 Settings

- The models were implemented[3] in Python using Scikit-learn (Pedregosa et al., 2011). MissingPy and Imbalanced-learn (Lemaître et al., 2017) were used for imputing missing values and balancing classes in the dataset. xgboost package (Chen and Guestrin, 2016) was used to train XGBoost model (to leverage concurrency). Matplotlib (Hunter, 2007) was used to plot graphs and Pandas (Reback et al., 2020; Wes McKinney, 2010) was used for data analysis.
- Pre-processing as mentioned in Section 4.1 were performed.
- The `random_state` was set to 2021 wherever possible.
- The dataset was split into train-test sets using `sklearn.model_selection.train_test_split()` with `test_split` set to 0.25.
- The experiment was conducted for each of the following 12 combinations of missing value imputation and sampling techniques (as mentioned in Section 4) for each year:
  1. **Simple** imputation with **No** sampling
  2. **Simple** imputation with **SMOTE** sampling
  3. **Simple** imputation with **RUS** sampling
  4. **Simple** imputation with **SMOTE-ENN** sampling
  5. **MissForest** imputation with **No** sampling
  6. **MissForest** imputation with **SMOTE** sampling
  7. **MissForest** imputation with **RUS** sampling
  8. **MissForest** imputation with **SMOTE-ENN** sampling
  9. **KNN** imputation with **No** sampling
  10. **KNN** imputation with **SMOTE** sampling
  11. **KNN** imputation with **RUS** sampling
  12. **KNN** imputation with **SMOTE-ENN** sampling
- The sampling was performed only on the train set. However, imputation had to be carried out in both train and test sets.
- The desired ratio of the number of samples in the minority class over the number of samples in the majority class after resampling is set to 0.6.
- Each model was tuned by performing exhaustive search over list of parameter values and fitted with 2-fold cross validation with F1-score as validation metric to obtained the best

[3] Code available at https://github.com/cb3101/bankruptcy_prediction

model. The list of parameters were chosen based on heuristics, considering the computational costs.

5.2 Results

We use the following notations to define the performance metrics used in this report:

- **TP** (True Positive): Model predicts Positive and it is actually Positive.
- **FP** (False Positive): Model predicts Positive but it is actually Negative.
- **FN** (False Negative): Model predicts Negative but it is actually Positive.
- **TN** (True Negative): Model predicts Negative and it is actually Negative.

Now, we define the performance metrics of our interest.

$$\text{Accuracy } (Acc) = \frac{TP+TN}{TP+FP+FN+TN}$$

$$\text{Precision } (Prec) = \frac{TP}{TP+FP}$$

$$\text{Recall } (Rec) = \frac{TP}{TP+FN}$$

$$\text{F1-Score } (F1) = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

We list the model and the setting which gave the best results for each year in Table 7 and enclose the complete results obtained for each settings for each year in Table 8.

5.3 Analysis

From the results in Table 7 and Table 8, the following observations can be drawn:

- XGBoost gave the best results, in terms of F1-score.
- The performance of the models had improved when over-sampling (specifically SMOTE) was used, in general.
- A surprising result was that Simple Imputation outperformed MissForest and KNN Imputation techniques, even though the latter methods are relatively more sophisticated and computationally expensive.
- Both over-sampling and under-sampling (including combination of both) improved the recall, while the precision dropped. This can be reasoned by the fact that the sampling techniques increase the percentage of examples positive labels (*minority* class) in the dataset.

6 Other Approaches

Since many attributes had high correlation values with each other, we were motivated to use Dimensionality Reduction for the data-set, Principal Component Analysis (PCA) (Espirito Santo, 2012) in particular. We reduced the dimension of the dataset to 20 and experimented with different sampling techniques and Simple Imputation. However, we observed that the results did not improve significantly with PCA (Table 8).

7 Conclusion

In this project, we surveyed various missing value imputation and sampling techniques and conducted a comparative study for bankruptcy prediction task. The performance of the models had improved significantly with the application of SMOTE (Table 7 and Table 8). Ren (2020) conducted the experiment for Year 5 without sampling and obtained a recall of 56%. In our case, the best model for Year 5 had a recall of 66% (and a F1-score of 64%). Moreover, the accuracy values of the best models obtained were over 95% which were similar to accuracy values obtained by Zikeba et al. (2016). However, in our case, the model had learnt to predict companies that actually go "*bankrupt*" (as the F1-scores were above 60% on average); however no such conclusion can be drawn from the models proposed by Zikeba et al. (2016) as the F1-scores were not reported and owing to the high class imbalance, a model with high bias towards the class "*not bankrupt*" could also have similar accuracy values (as discussed in Section 1.3).

References

Edward I. Altman. 1968. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, 23(4):589–609.

Gustavo E. A. P. A. Batista, R. Prati, and Maria Carolina Monard. 2004. A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explor.*, 6:20–29.

Year	Imputer	Sampling	Model	Acc	Prec	Rec	F1
1	Simple	SMOTE	XGBoost	0.98	0.69	0.66	0.68
2	Simple	None	XGBoost	0.97	0.97	0.38	0.54
3	Simple	SMOTE	XGBoost	0.96	0.67	0.53	0.60
4	Simple	SMOTE	XGBoost	0.96	0.63	0.59	0.61
5	Simple	SMOTE	XGBoost	0.95	0.63	0.66	0.64

Table 7: Best results in each year

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. ACM.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Mach. Learn.*, 20(3):273–297.

Rafael Espirito Santo. 2012. Principal component analysis applied to digital image compression. *Einstein (São Paulo, Brazil)*, 10:135–9.

Gongde Guo, Hui Wang, David Bell, and Yaxin Bi. 2004. Knn model-based approach in classification.

J. D. Hunter. 2007. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95.

Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. 2017. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5.

Wes McKinney. 2010. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference*, pages 56 – 61.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

T. D. Quynh and T. Thi Lan Phuong. 2020. Improving the bankruptcy prediction by combining some classification models. In *2020 12th International Conference on Knowledge and Systems Engineering (KSE)*, pages 263–268.

Jeff Reback, Wes McKinney, jbrockmendel, Joris Van den Bossche, Tom Augspurger, Phillip Cloud, gfyoung, Sinhrks, Adam Klein, Matthew Roeschke, Simon Hawkins, Jeff Tratner, Chang She, William Ayd, Terji Petersen, Marc Garcia, Jeremy Schendel, Andy Hayden, MomIsBestFriend, Vytautas Jancauskas, Pietro Battiston, Skipper Seabold, chris b1, h vetinari, Stephan Hoyer, Wouter Overmeire, alimcmaster1, Kaiqi Dong, Christopher Whelan, and Mortada Mehyar. 2020. pandas-dev/pandas: Pandas 1.0.3.

Yifan Ren. 2020. A comparison of important features for predicting polish and chinese corporate bankruptcies.

D. J. Stekhoven and P. Buhlmann. 2011. Missforest–non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118.

Tin Kam Ho. 1995. Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 1, pages 278–282 vol.1.

Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B. Altman. 2001. Missing value estimation methods for DNA microarrays . *Bioinformatics*, 17(6):520–525.

Maciej Zikeba, Sebastian K Tomczak, and Jakub M Tomczak. 2016. Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction. *Expert Systems with Applications*.

Imputer	Sampling	Metric	LR	LDA	KNN-5	KNN-10	GNB	DT	SVC	RFC	XGB	Voting
Simple	None	Acc	0.96	0.96	0.96	0.96	0.07	0.98	0.96	0.98	0.98	0.98
		Prec	0.65	0.00	0.33	0.00	0.04	0.81	0.00	0.87	0.95	0.97
		Rec	0.19	0.00	0.01	0.00	0.99	0.56	0.00	0.50	0.54	0.50
		F1	0.30	0.00	0.03	0.00	0.08	0.66	0.00	0.64	0.69	0.66
	SMOTE	Acc	0.87	0.94	0.84	0.85	0.06	0.85	0.90	0.98	**0.98**	0.98
		Prec	0.15	0.15	0.10	0.13	0.04	0.18	0.23	0.77	**0.69**	0.81
		Rec	0.47	0.09	0.40	0.53	1.00	0.76	0.60	0.53	**0.66**	0.62
		F1	0.23	0.11	0.16	0.21	0.08	0.29	0.33	0.63	**0.68**	0.70
	RUS	Acc	0.87	0.87	0.81	0.85	0.66	0.85	0.92	0.91	0.92	0.94
		Prec	0.15	0.12	0.10	0.11	0.06	0.19	0.10	0.27	0.29	0.35
		Rec	0.51	0.34	0.47	0.38	0.49	0.84	0.12	0.74	0.81	0.76
		F1	0.23	0.17	0.16	0.17	0.10	0.31	0.11	0.40	0.43	0.48
	SMOTE-ENN	Acc	0.81	0.88	0.80	0.78	0.09	0.95	0.79	0.97	0.97	0.97
		Prec	0.13	0.11	0.11	0.10	0.04	0.39	0.10	0.67	0.63	0.61
		Rec	0.69	0.31	0.56	0.54	0.99	0.65	0.54	0.54	0.71	0.68
		F1	0.22	0.16	0.18	0.16	0.08	0.48	0.17	0.60	0.67	0.64
MissForest	None	Acc	0.96	0.96	0.96	0.96	0.95	0.94	0.96	0.96	0.96	0.96
		Prec	0.17	0.00	0.00	0.00	0.09	0.22	0.00	0.00	0.80	0.00
		Rec	0.01	0.00	0.00	0.00	0.03	0.21	0.00	0.00	0.06	0.00
		F1	0.03	0.00	0.00	0.00	0.04	0.21	0.00	0.00	0.11	0.00
	SMOTE	Acc	0.87	0.94	0.82	0.82	0.07	0.84	0.93	0.95	0.94	0.95
		Prec	0.13	0.10	0.11	0.10	0.04	0.11	0.18	0.32	0.31	0.33
		Rec	0.40	0.06	0.50	0.46	0.99	0.44	0.21	0.28	0.40	0.37
		F1	0.20	0.07	0.18	0.17	0.08	0.17	0.19	0.30	0.35	0.35
	RUS	Acc	0.92	0.87	0.79	0.87	0.91	0.80	0.95	0.86	0.85	0.91
		Prec	0.16	0.12	0.09	0.07	0.04	0.08	0.09	0.16	0.16	0.17
		Rec	0.22	0.37	0.47	0.21	0.06	0.41	0.04	0.57	0.69	0.34
		F1	0.18	0.18	0.15	0.11	0.05	0.14	0.06	0.25	0.26	0.23
	SMOTE-ENN	Acc	0.76	0.84	0.77	0.76	0.10	0.84	0.85	0.92	0.89	0.89
		Prec	0.12	0.08	0.09	0.09	0.04	0.12	0.12	0.22	0.18	0.18
		Rec	0.78	0.29	0.54	0.56	0.93	0.51	0.43	0.37	0.51	0.53
		F1	0.20	0.13	0.16	0.15	0.07	0.20	0.18	0.27	0.27	0.27
KNN	None	Acc	0.96	0.96	0.96	0.96	0.07	0.94	0.96	0.96	0.96	0.96
		Prec	0.00	0.00	0.17	0.00	0.04	0.25	0.00	0.17	0.77	0.00
		Rec	0.00	0.00	0.01	0.00	0.99	0.25	0.00	0.01	0.15	0.00
		F1	0.00	0.00	0.03	0.00	0.08	0.25	0.00	0.03	0.25	0.00
	SMOTE	Acc	0.87	0.94	0.83	0.83	0.09	0.89	0.93	0.94	0.95	0.95
		Prec	0.13	0.11	0.10	0.11	0.04	0.15	0.18	0.26	0.39	0.34
		Rec	0.41	0.07	0.40	0.50	0.96	0.40	0.25	0.32	0.46	0.24
		F1	0.20	0.09	0.15	0.18	0.08	0.22	0.21	0.29	0.42	0.28
	RUS	Acc	0.91	0.86	0.79	0.82	0.12	0.84	0.92	0.87	0.86	0.91
		Prec	0.15	0.12	0.09	0.09	0.04	0.12	0.10	0.18	0.17	0.18
		Rec	0.28	0.38	0.47	0.38	1.00	0.47	0.12	0.60	0.68	0.37
		F1	0.20	0.18	0.15	0.14	0.08	0.19	0.11	0.27	0.27	0.24
	SMOTE-ENN	Acc	0.82	0.84	0.78	0.76	0.09	0.77	0.78	0.90	0.94	0.92
		Prec	0.10	0.11	0.10	0.09	0.04	0.10	0.10	0.17	0.33	0.21
		Rec	0.46	0.44	0.56	0.57	0.96	0.63	0.57	0.44	0.57	0.34
		F1	0.16	0.18	0.17	0.16	0.08	0.18	0.17	0.25	0.42	0.26

(a) Year 1

Imputer	Sampling	Metric	LR	LDA	KNN-5	KNN-10	GNB	DT	SVC	RFC	XGB	Voting
Simple	None	Acc	0.96	0.96	0.96	0.96	0.06	0.97	0.96	0.97	**0.97**	0.97
		Prec	0.33	0.40	0.00	0.00	0.04	0.71	0.00	0.85	**0.97**	0.95
		Rec	0.02	0.02	0.00	0.00	0.93	0.25	0.00	0.22	**0.38**	0.20
		F1	0.04	0.04	0.00	0.00	0.07	0.37	0.00	0.35	**0.54**	0.33
	SMOTE	Acc	0.86	0.93	0.81	0.80	0.08	0.92	0.86	0.97	0.97	0.97
		Prec	0.12	0.07	0.08	0.07	0.04	0.20	0.13	0.74	0.67	0.67
		Rec	0.40	0.06	0.36	0.34	0.90	0.36	0.41	0.28	0.39	0.34
		F1	0.19	0.06	0.13	0.12	0.07	0.26	0.19	0.41	0.49	0.45
	RUS	Acc	0.82	0.84	0.73	0.81	0.88	0.86	0.87	0.87	0.88	0.91
		Prec	0.10	0.08	0.08	0.08	0.06	0.16	0.07	0.18	0.21	0.24
		Rec	0.46	0.28	0.53	0.36	0.14	0.60	0.20	0.66	0.69	0.62
		F1	0.17	0.12	0.13	0.13	0.09	0.25	0.11	0.28	0.32	0.35
	SMOTE-ENN	Acc	0.72	0.83	0.74	0.73	0.05	0.88	0.78	0.96	0.96	0.95
		Prec	0.08	0.07	0.07	0.07	0.04	0.15	0.09	0.52	0.50	0.42
		Rec	0.55	0.28	0.43	0.48	0.99	0.41	0.51	0.26	0.39	0.45
		F1	0.13	0.11	0.12	0.12	0.08	0.21	0.15	0.35	0.44	0.43
MissForest	None	Acc	0.96	0.96	0.96	0.96	0.08	0.95	0.96	0.96	0.97	0.96
		Prec	0.33	0.25	0.00	0.00	0.04	0.37	0.00	0.12	0.80	1.00
		Rec	0.02	0.04	0.00	0.00	0.89	0.21	0.00	0.01	0.16	0.01
		F1	0.04	0.07	0.00	0.00	0.07	0.27	0.00	0.02	0.27	0.02
	SMOTE	Acc	0.83	0.93	0.80	0.79	0.09	0.83	0.83	0.92	0.96	0.94
		Prec	0.10	0.09	0.08	0.07	0.04	0.11	0.10	0.20	0.45	0.25
		Rec	0.40	0.08	0.38	0.35	0.89	0.46	0.43	0.31	0.34	0.28
		F1	0.16	0.09	0.13	0.11	0.07	0.18	0.16	0.24	0.39	0.26
	RUS	Acc	0.82	0.86	0.72	0.80	0.92	0.77	0.86	0.83	0.84	0.87
		Prec	0.08	0.08	0.07	0.06	0.06	0.09	0.06	0.14	0.15	0.14
		Rec	0.33	0.24	0.50	0.27	0.07	0.52	0.18	0.62	0.64	0.45
		F1	0.13	0.12	0.12	0.10	0.07	0.15	0.09	0.22	0.24	0.22
	SMOTE-ENN	Acc	0.71	0.83	0.73	0.72	0.09	0.82	0.76	0.88	0.93	0.88
		Prec	0.07	0.07	0.07	0.07	0.04	0.09	0.08	0.14	0.27	0.15
		Rec	0.55	0.24	0.47	0.46	0.89	0.40	0.50	0.37	0.37	0.45
		F1	0.13	0.10	0.12	0.12	0.07	0.15	0.14	0.20	0.31	0.23
KNN	None	Acc	0.96	0.96	0.96	0.96	0.07	0.94	0.96	0.96	0.96	0.96
		Prec	0.33	0.40	0.00	0.00	0.04	0.19	0.00	0.12	0.73	0.33
		Rec	0.02	0.02	0.00	0.00	0.91	0.18	0.00	0.01	0.16	0.01
		F1	0.04	0.04	0.00	0.00	0.07	0.18	0.00	0.02	0.26	0.02
	SMOTE	Acc	0.85	0.93	0.80	0.78	0.08	0.86	0.83	0.93	0.92	0.93
		Prec	0.12	0.07	0.08	0.06	0.04	0.12	0.10	0.19	0.21	0.23
		Rec	0.42	0.06	0.36	0.32	0.91	0.39	0.40	0.27	0.35	0.31
		F1	0.18	0.06	0.12	0.10	0.07	0.19	0.16	0.22	0.27	0.26
	RUS	Acc	0.81	0.85	0.71	0.79	0.89	0.82	0.86	0.82	0.82	0.88
		Prec	0.10	0.09	0.07	0.06	0.06	0.11	0.07	0.12	0.12	0.15
		Rec	0.44	0.30	0.52	0.31	0.13	0.49	0.19	0.59	0.58	0.45
		F1	0.16	0.14	0.12	0.11	0.08	0.18	0.10	0.20	0.20	0.23
	SMOTE-ENN	Acc	0.70	0.82	0.73	0.72	0.05	0.67	0.75	0.86	0.93	0.88
		Prec	0.08	0.07	0.06	0.07	0.04	0.07	0.08	0.12	0.24	0.14
		Rec	0.59	0.28	0.43	0.46	0.99	0.56	0.51	0.42	0.37	0.40
		F1	0.14	0.11	0.11	0.11	0.08	0.12	0.14	0.19	0.29	0.20

(b) Year 2

Imputer	Sampling	Metric	LR	LDA	KNN-5	KNN-10	GNB	DT	SVC	RFC	XGB	Voting
Simple	None	Acc	0.95	0.95	0.94	0.95	0.07	0.96	0.95	0.96	0.97	0.97
		Prec	0.06	0.00	0.00	0.00	0.05	0.66	0.00	0.79	0.87	1.00
		Rec	0.01	0.00	0.00	0.00	0.99	0.42	0.00	0.34	0.45	0.31
		F1	0.01	0.00	0.00	0.00	0.09	0.51	0.00	0.48	0.59	0.47
	SMOTE	Acc	0.85	0.92	0.81	0.80	0.06	0.93	0.86	0.96	**0.96**	0.96
		Prec	0.16	0.15	0.11	0.10	0.05	0.35	0.18	0.54	**0.67**	0.62
		Rec	0.49	0.15	0.40	0.43	0.99	0.51	0.52	0.45	**0.53**	0.45
		F1	0.24	0.15	0.17	0.17	0.09	0.42	0.26	0.49	**0.60**	0.52
	RUS	Acc	0.88	0.87	0.77	0.80	0.12	0.92	0.85	0.89	0.89	0.93
		Prec	0.23	0.14	0.10	0.11	0.05	0.32	0.11	0.26	0.29	0.37
		Rec	0.64	0.33	0.49	0.46	0.96	0.57	0.33	0.75	0.82	0.68
		F1	0.34	0.19	0.17	0.18	0.09	0.41	0.17	0.39	0.42	0.48
	SMOTE-ENN	Acc	0.74	0.83	0.74	0.72	0.11	0.89	0.74	0.95	0.96	0.94
		Prec	0.12	0.12	0.09	0.09	0.05	0.23	0.12	0.44	0.54	0.39
		Rec	0.71	0.41	0.49	0.57	0.94	0.61	0.68	0.47	0.59	0.62
		F1	0.20	0.18	0.15	0.16	0.09	0.33	0.20	0.45	0.57	0.48
MissForest	None	Acc	0.95	0.95	0.95	0.95	0.08	0.93	0.95	0.95	0.96	0.95
		Prec	0.05	0.05	0.00	0.00	0.05	0.20	0.00	0.08	0.82	0.50
		Rec	0.01	0.01	0.00	0.00	1.00	0.14	0.00	0.01	0.15	0.01
		F1	0.01	0.01	0.00	0.00	0.09	0.17	0.00	0.01	0.25	0.02
	SMOTE	Acc	0.86	0.93	0.81	0.79	0.06	0.79	0.86	0.92	0.95	0.93
		Prec	0.15	0.11	0.10	0.10	0.05	0.11	0.11	0.28	0.46	0.26
		Rec	0.45	0.08	0.37	0.44	0.99	0.50	0.29	0.40	0.30	0.33
		F1	0.23	0.09	0.15	0.17	0.09	0.19	0.16	0.33	0.36	0.29
	RUS	Acc	0.87	0.89	0.75	0.79	0.40	0.70	0.85	0.83	0.84	0.86
		Prec	0.14	0.10	0.09	0.10	0.06	0.10	0.11	0.16	0.18	0.20
		Rec	0.36	0.19	0.48	0.44	0.77	0.69	0.30	0.65	0.67	0.63
		F1	0.21	0.13	0.16	0.16	0.11	0.18	0.16	0.26	0.29	0.30
	SMOTE-ENN	Acc	0.76	0.86	0.74	0.72	0.12	0.82	0.77	0.88	0.94	0.88
		Prec	0.13	0.12	0.08	0.09	0.05	0.12	0.12	0.17	0.38	0.21
		Rec	0.69	0.33	0.46	0.50	0.98	0.44	0.63	0.37	0.42	0.54
		F1	0.21	0.18	0.14	0.15	0.10	0.19	0.20	0.23	0.40	0.31
KNN	None	Acc	0.95	0.95	0.94	0.95	0.07	0.94	0.95	0.95	0.96	0.95
		Prec	0.05	0.00	0.00	0.00	0.05	0.21	0.00	0.07	0.83	0.00
		Rec	0.01	0.00	0.00	0.00	0.99	0.14	0.00	0.01	0.12	0.00
		F1	0.01	0.00	0.00	0.00	0.09	0.17	0.00	0.01	0.21	0.00
	SMOTE	Acc	0.85	0.92	0.80	0.78	0.06	0.85	0.85	0.91	0.95	0.92
		Prec	0.14	0.14	0.10	0.10	0.05	0.13	0.12	0.22	0.45	0.23
		Rec	0.44	0.14	0.40	0.44	0.99	0.35	0.37	0.33	0.37	0.26
		F1	0.22	0.14	0.16	0.16	0.09	0.19	0.19	0.26	0.40	0.24
	RUS	Acc	0.85	0.88	0.76	0.79	0.10	0.82	0.85	0.80	0.83	0.84
		Prec	0.15	0.11	0.10	0.11	0.05	0.10	0.11	0.15	0.17	0.17
		Rec	0.46	0.22	0.50	0.46	0.97	0.34	0.33	0.66	0.70	0.58
		F1	0.22	0.15	0.16	0.17	0.09	0.15	0.17	0.24	0.28	0.26
	SMOTE-ENN	Acc	0.74	0.83	0.73	0.71	0.11	0.84	0.75	0.91	0.91	0.91
		Prec	0.12	0.12	0.09	0.09	0.05	0.13	0.12	0.21	0.28	0.21
		Rec	0.70	0.40	0.49	0.55	0.94	0.41	0.67	0.35	0.53	0.36
		F1	0.20	0.18	0.14	0.15	0.09	0.20	0.20	0.27	0.36	0.27

(c) Year 3

Imputer	Sampling	Metric	LR	LDA	KNN-5	KNN-10	GNB	DT	SVC	RFC	XGB	Voting
Simple	None	Acc	0.94	0.94	0.95	0.94	0.10	0.95	0.95	0.96	0.97	0.96
		Prec	0.29	0.22	0.00	0.00	0.05	0.58	0.00	0.75	0.89	0.94
		Rec	0.03	0.03	0.00	0.00	0.94	0.33	0.00	0.23	0.45	0.23
		F1	0.06	0.05	0.00	0.00	0.10	0.42	0.00	0.36	0.59	0.36
	SMOTE	Acc	0.88	0.91	0.81	0.81	0.08	0.88	0.89	0.93	**0.96**	0.95
		Prec	0.21	0.18	0.12	0.13	0.05	0.24	0.22	0.41	**0.63**	0.51
		Rec	0.45	0.20	0.41	0.43	0.95	0.58	0.46	0.52	**0.59**	0.52
		F1	0.28	0.19	0.19	0.20	0.10	0.34	0.30	0.46	**0.61**	0.51
	RUS	Acc	0.86	0.88	0.80	0.83	0.90	0.86	0.90	0.86	0.90	0.90
		Prec	0.19	0.17	0.13	0.15	0.12	0.21	0.18	0.24	0.33	0.30
		Rec	0.53	0.33	0.48	0.45	0.14	0.59	0.25	0.73	0.82	0.66
		F1	0.28	0.22	0.20	0.22	0.13	0.31	0.21	0.36	0.47	0.42
	SMOTE-ENN	Acc	0.79	0.84	0.75	0.73	0.09	0.88	0.84	0.94	0.93	0.92
		Prec	0.14	0.16	0.12	0.11	0.05	0.25	0.16	0.44	0.42	0.36
		Rec	0.59	0.46	0.56	0.57	0.96	0.61	0.48	0.42	0.63	0.58
		F1	0.23	0.24	0.19	0.18	0.10	0.35	0.25	0.43	0.51	0.45
MissForest	None	Acc	0.95	0.94	0.94	0.95	0.13	0.92	0.95	0.95	0.95	0.95
		Prec	0.27	0.22	0.00	0.14	0.05	0.26	0.00	0.47	0.79	0.78
		Rec	0.02	0.03	0.00	0.01	0.90	0.23	0.00	0.06	0.17	0.05
		F1	0.04	0.05	0.00	0.01	0.10	0.24	0.00	0.11	0.28	0.10
	SMOTE	Acc	0.84	0.88	0.81	0.82	0.12	0.86	0.89	0.92	0.94	0.93
		Prec	0.17	0.15	0.12	0.12	0.05	0.20	0.19	0.31	0.43	0.33
		Rec	0.50	0.27	0.41	0.41	0.87	0.57	0.33	0.45	0.45	0.34
		F1	0.25	0.19	0.19	0.19	0.09	0.30	0.24	0.36	0.44	0.34
	RUS	Acc	0.82	0.84	0.80	0.88	0.87	0.81	0.86	0.84	0.85	0.88
		Prec	0.16	0.14	0.12	0.18	0.11	0.15	0.14	0.18	0.19	0.22
		Rec	0.54	0.38	0.46	0.37	0.22	0.56	0.34	0.59	0.61	0.54
		F1	0.24	0.20	0.20	0.24	0.15	0.24	0.20	0.28	0.29	0.32
	SMOTE-ENN	Acc	0.75	0.82	0.76	0.74	0.12	0.81	0.83	0.89	0.90	0.88
		Prec	0.13	0.14	0.11	0.11	0.05	0.16	0.15	0.23	0.29	0.21
		Rec	0.64	0.48	0.52	0.57	0.88	0.59	0.51	0.51	0.55	0.50
		F1	0.22	0.22	0.19	0.19	0.09	0.25	0.23	0.32	0.38	0.30
KNN	None	Acc	0.94	0.94	0.94	0.94	0.10	0.91	0.95	0.95	0.96	0.95
		Prec	0.31	0.22	0.00	0.00	0.05	0.19	0.00	0.47	0.92	0.80
		Rec	0.04	0.03	0.00	0.00	0.94	0.22	0.00	0.06	0.17	0.03
		F1	0.07	0.05	0.00	0.00	0.10	0.20	0.00	0.11	0.29	0.06
	SMOTE	Acc	0.88	0.91	0.79	0.80	0.09	0.82	0.90	0.92	0.93	0.92
		Prec	0.20	0.18	0.11	0.11	0.05	0.14	0.20	0.31	0.38	0.30
		Rec	0.45	0.20	0.39	0.41	0.95	0.46	0.27	0.41	0.49	0.38
		F1	0.28	0.19	0.17	0.18	0.10	0.22	0.23	0.35	0.43	0.33
	RUS	Acc	0.86	0.88	0.77	0.86	0.90	0.83	0.90	0.84	0.85	0.88
		Prec	0.18	0.16	0.11	0.15	0.12	0.16	0.19	0.18	0.20	0.22
		Rec	0.48	0.32	0.46	0.37	0.14	0.51	0.27	0.60	0.64	0.49
		F1	0.26	0.21	0.17	0.22	0.13	0.24	0.22	0.28	0.30	0.30
	SMOTE-ENN	Acc	0.80	0.84	0.73	0.71	0.09	0.79	0.81	0.86	0.88	0.89
		Prec	0.14	0.16	0.11	0.10	0.05	0.13	0.15	0.19	0.24	0.22
		Rec	0.54	0.46	0.55	0.54	0.96	0.52	0.56	0.52	0.57	0.46
		F1	0.22	0.23	0.18	0.16	0.10	0.20	0.24	0.28	0.34	0.30

(d) Year 4

Imputer	Sampling	Metric	LR	LDA	KNN-5	KNN-10	GNB	DT	SVC	RFC	XGB	Voting
Simple	None	Acc	0.93	0.93	0.93	0.93	0.92	0.96	0.93	0.95	0.96	0.96
		Prec	0.56	0.39	0.62	0.56	0.32	0.85	0.00	0.81	0.94	0.97
		Rec	0.19	0.07	0.15	0.15	0.07	0.55	0.00	0.41	0.49	0.38
		F1	0.28	0.12	0.24	0.23	0.11	0.67	0.00	0.55	0.65	0.55
	SMOTE	Acc	0.89	0.89	0.84	0.85	0.93	0.91	0.90	0.94	**0.95**	0.94
		Prec	0.31	0.31	0.21	0.23	0.40	0.41	0.34	0.55	**0.63**	0.58
		Rec	0.54	0.44	0.47	0.52	0.06	0.74	0.53	0.60	**0.66**	0.60
		F1	0.40	0.36	0.29	0.32	0.10	0.53	0.42	0.57	**0.64**	0.59
	RUS	Acc	0.86	0.88	0.86	0.90	0.92	0.89	0.89	0.87	0.91	0.91
		Prec	0.27	0.28	0.28	0.34	0.29	0.38	0.31	0.32	0.42	0.43
		Rec	0.55	0.49	0.59	0.51	0.12	0.77	0.49	0.74	0.76	0.74
		F1	0.36	0.36	0.38	0.41	0.17	0.51	0.38	0.45	0.54	0.54
	SMOTE-ENN	Acc	0.84	0.85	0.81	0.82	0.92	0.92	0.84	0.92	0.93	0.91
		Prec	0.26	0.24	0.20	0.21	0.34	0.44	0.25	0.46	0.49	0.41
		Rec	0.68	0.52	0.57	0.59	0.20	0.69	0.67	0.55	0.69	0.67
		F1	0.38	0.33	0.29	0.31	0.25	0.53	0.36	0.50	0.57	0.51
MissForest	None	Acc	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.94	0.95	0.94
		Prec	0.54	0.40	0.56	0.56	0.33	0.46	0.25	0.64	0.80	0.72
		Rec	0.19	0.08	0.14	0.15	0.07	0.42	0.01	0.23	0.35	0.23
		F1	0.28	0.13	0.22	0.23	0.11	0.44	0.02	0.33	0.49	0.34
	SMOTE	Acc	0.89	0.90	0.84	0.84	0.93	0.87	0.90	0.93	0.94	0.93
		Prec	0.32	0.34	0.20	0.22	0.33	0.28	0.35	0.47	0.57	0.51
		Rec	0.56	0.43	0.44	0.52	0.06	0.60	0.52	0.55	0.49	0.58
		F1	0.41	0.38	0.27	0.31	0.10	0.38	0.42	0.51	0.53	0.54
	RUS	Acc	0.88	0.88	0.86	0.90	0.92	0.88	0.89	0.87	0.87	0.90
		Prec	0.30	0.28	0.27	0.35	0.27	0.32	0.33	0.31	0.30	0.37
		Rec	0.56	0.49	0.58	0.50	0.12	0.64	0.52	0.69	0.69	0.66
		F1	0.39	0.36	0.37	0.41	0.16	0.43	0.40	0.43	0.42	0.48
	SMOTE-ENN	Acc	0.84	0.87	0.80	0.81	0.92	0.85	0.83	0.90	0.90	0.89
		Prec	0.26	0.29	0.18	0.20	0.39	0.27	0.24	0.36	0.36	0.34
		Rec	0.67	0.58	0.54	0.58	0.23	0.66	0.69	0.57	0.63	0.64
		F1	0.37	0.38	0.27	0.29	0.29	0.38	0.36	0.44	0.46	0.44
KNN	None	Acc	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.94	0.94
		Prec	0.57	0.39	0.54	0.58	0.33	0.46	0.25	0.56	0.67	0.67
		Rec	0.20	0.07	0.14	0.15	0.07	0.31	0.01	0.23	0.29	0.16
		F1	0.29	0.12	0.22	0.23	0.11	0.37	0.02	0.32	0.41	0.25
	SMOTE	Acc	0.88	0.89	0.83	0.84	0.93	0.89	0.90	0.92	0.94	0.93
		Prec	0.30	0.31	0.19	0.21	0.40	0.35	0.34	0.46	0.56	0.49
		Rec	0.54	0.45	0.43	0.49	0.06	0.62	0.52	0.57	0.49	0.54
		F1	0.38	0.37	0.26	0.29	0.10	0.45	0.41	0.51	0.52	0.51
	RUS	Acc	0.86	0.87	0.86	0.90	0.92	0.84	0.89	0.87	0.87	0.88
		Prec	0.27	0.28	0.27	0.35	0.32	0.25	0.33	0.31	0.31	0.33
		Rec	0.55	0.53	0.60	0.52	0.13	0.69	0.49	0.69	0.71	0.66
		F1	0.36	0.36	0.37	0.42	0.18	0.37	0.39	0.42	0.43	0.44
	SMOTE-ENN	Acc	0.84	0.85	0.79	0.80	0.92	0.84	0.84	0.91	0.91	0.91
		Prec	0.27	0.24	0.18	0.19	0.36	0.24	0.25	0.40	0.41	0.39
		Rec	0.70	0.51	0.53	0.59	0.20	0.61	0.68	0.65	0.63	0.65
		F1	0.38	0.32	0.26	0.29	0.25	0.34	0.37	0.49	0.49	0.49

(e) Year 5

Table 8: Model performance for different settings for each year

Year	Sampling	Metric	LR	LDA	KNN-5	KNN-10	GNB	DT	SVC	RFC	XGB	Voting
1	None	Acc	0.96	0.96	0.96	0.96	0.08	0.96	0.96	0.97	0.97	0.97
		Prec	0.00	0.00	0.00	0.00	0.04	0.55	0.00	0.88	0.89	0.87
		Rec	0.00	0.00	0.00	0.00	0.99	0.38	0.00	0.32	0.37	0.29
		F1	0.00	0.00	0.00	0.00	0.08	0.45	0.00	0.47	0.52	0.44
	SMOTE	Acc	0.87	0.96	0.85	0.84	0.10	0.87	0.94	0.96	0.95	0.96
		Prec	0.10	0.09	0.09	0.09	0.04	0.17	0.33	0.51	0.36	0.53
		Rec	0.29	0.01	0.31	0.32	0.97	0.62	0.41	0.38	0.47	0.40
		F1	0.15	0.03	0.14	0.14	0.08	0.27	0.37	0.44	0.41	0.45
	RUS	Acc	0.88	0.94	0.79	0.87	0.09	0.76	0.94	0.89	0.90	0.92
		Prec	0.09	0.08	0.06	0.09	0.04	0.11	0.10	0.20	0.21	0.27
		Rec	0.22	0.06	0.26	0.26	0.99	0.74	0.06	0.62	0.59	0.54
		F1	0.12	0.07	0.09	0.14	0.08	0.19	0.07	0.30	0.31	0.36
	SMOTE-ENN	Acc	0.78	0.95	0.80	0.78	0.06	0.90	0.86	0.95	0.93	0.94
		Prec	0.09	0.07	0.07	0.07	0.04	0.20	0.15	0.37	0.27	0.32
		Rec	0.51	0.01	0.34	0.40	1.00	0.47	0.56	0.41	0.47	0.40
		F1	0.15	0.02	0.12	0.13	0.08	0.28	0.23	0.39	0.34	0.35
2	None	Acc	0.96	0.96	0.96	0.96	0.06	0.96	0.96	0.96	0.96	0.96
		Prec	0.00	0.00	0.43	0.00	0.04	0.38	0.00	0.61	0.74	0.80
		Rec	0.00	0.00	0.03	0.00	0.96	0.16	0.00	0.14	0.17	0.12
		F1	0.00	0.00	0.06	0.00	0.08	0.23	0.00	0.23	0.28	0.21
	SMOTE	Acc	0.95	0.96	0.83	0.83	0.07	0.92	0.94	0.95	0.93	0.95
		Prec	0.00	0.00	0.06	0.07	0.04	0.18	0.23	0.31	0.22	0.32
		Rec	0.00	0.00	0.23	0.25	0.96	0.30	0.26	0.23	0.29	0.16
		F1	0.00	0.00	0.10	0.11	0.08	0.22	0.24	0.26	0.25	0.21
	RUS	Acc	0.92	0.93	0.78	0.88	0.48	0.77	0.95	0.85	0.88	0.92
		Prec	0.03	0.06	0.07	0.11	0.04	0.07	0.03	0.13	0.14	0.18
		Rec	0.03	0.05	0.37	0.28	0.50	0.39	0.01	0.51	0.42	0.31
		F1	0.03	0.06	0.12	0.15	0.07	0.12	0.01	0.21	0.22	0.23
	SMOTE-ENN	Acc	0.90	0.96	0.79	0.78	0.08	0.87	0.88	0.94	0.91	0.95
		Prec	0.06	0.00	0.07	0.06	0.04	0.13	0.13	0.26	0.18	0.30
		Rec	0.09	0.00	0.32	0.31	0.93	0.37	0.33	0.26	0.31	0.25
		F1	0.07	0.00	0.11	0.10	0.07	0.19	0.18	0.26	0.22	0.27
3	None	Acc	0.95	0.95	0.95	0.95	0.10	0.95	0.95	0.96	0.96	0.96
		Prec	0.00	0.00	0.27	0.22	0.05	0.38	0.00	0.65	0.88	0.94
		Rec	0.00	0.00	0.05	0.02	0.95	0.26	0.00	0.16	0.23	0.12
		F1	0.00	0.00	0.08	0.03	0.09	0.31	0.00	0.26	0.36	0.22
	SMOTE	Acc	0.95	0.95	0.83	0.83	0.13	0.90	0.90	0.95	0.94	0.95
		Prec	0.08	0.00	0.11	0.11	0.05	0.23	0.21	0.47	0.36	0.48
		Rec	0.02	0.00	0.35	0.38	0.94	0.45	0.40	0.34	0.49	0.34
		F1	0.03	0.00	0.16	0.17	0.09	0.31	0.27	0.39	0.42	0.40
	RUS	Acc	0.88	0.89	0.78	0.82	0.12	0.88	0.92	0.86	0.87	0.91
		Prec	0.15	0.15	0.09	0.10	0.05	0.22	0.04	0.20	0.22	0.29
		Rec	0.33	0.28	0.40	0.37	0.94	0.59	0.03	0.66	0.65	0.56
		F1	0.20	0.19	0.15	0.16	0.09	0.32	0.04	0.30	0.32	0.38
	SMOTE-ENN	Acc	0.88	0.95	0.78	0.77	0.14	0.84	0.86	0.94	0.92	0.93
		Prec	0.10	0.07	0.10	0.09	0.05	0.16	0.18	0.39	0.32	0.31
		Rec	0.20	0.01	0.46	0.46	0.94	0.57	0.54	0.40	0.56	0.43
		F1	0.13	0.01	0.16	0.16	0.09	0.25	0.27	0.40	0.40	0.36

Year	Sampling	Metric	LR	LDA	KNN-5	KNN-10	GNB	DT	SVC	RFC	XGB	Voting
4	None	Acc	0.95	0.95	0.94	0.95	0.10	0.95	0.95	0.95	0.95	0.95
		Prec	0.00	0.00	0.22	0.22	0.05	0.48	0.00	0.73	0.70	1.00
		Rec	0.00	0.00	0.03	0.02	0.96	0.18	0.00	0.15	0.15	0.12
		F1	0.00	0.00	0.05	0.03	0.10	0.26	0.00	0.25	0.25	0.22
	SMOTE	Acc	0.94	0.94	0.82	0.82	0.08	0.85	0.90	0.91	0.91	0.94
		Prec	0.09	0.08	0.11	0.10	0.05	0.15	0.20	0.27	0.28	0.37
		Rec	0.02	0.01	0.32	0.32	0.98	0.39	0.30	0.40	0.43	0.26
		F1	0.04	0.01	0.16	0.16	0.10	0.22	0.24	0.32	0.34	0.30
	RUS	Acc	0.91	0.93	0.80	0.87	0.90	0.76	0.91	0.83	0.84	0.87
		Prec	0.17	0.09	0.11	0.15	0.06	0.13	0.13	0.18	0.19	0.21
		Rec	0.20	0.04	0.41	0.30	0.06	0.63	0.12	0.60	0.65	0.54
		F1	0.18	0.05	0.17	0.20	0.06	0.22	0.12	0.27	0.29	0.31
	SMOTE-ENN	Acc	0.88	0.94	0.77	0.76	0.08	0.84	0.81	0.92	0.89	0.90
		Prec	0.15	0.09	0.11	0.10	0.05	0.16	0.13	0.30	0.24	0.23
		Rec	0.27	0.02	0.47	0.45	0.98	0.48	0.48	0.40	0.51	0.35
		F1	0.19	0.03	0.18	0.17	0.10	0.24	0.21	0.34	0.33	0.28
5	None	Acc	0.93	0.93	0.92	0.93	0.14	0.95	0.93	0.94	0.95	0.95
		Prec	0.14	0.50	0.21	0.00	0.07	0.74	0.00	0.68	0.77	0.87
		Rec	0.01	0.01	0.03	0.00	0.90	0.41	0.00	0.33	0.32	0.26
		F1	0.02	0.02	0.05	0.00	0.13	0.53	0.00	0.45	0.46	0.41
	SMOTE	Acc	0.90	0.93	0.86	0.85	0.14	0.85	0.86	0.91	0.92	0.92
		Prec	0.28	0.33	0.22	0.21	0.07	0.26	0.24	0.40	0.45	0.46
		Rec	0.31	0.04	0.42	0.42	0.92	0.68	0.45	0.47	0.58	0.51
		F1	0.29	0.07	0.29	0.28	0.13	0.38	0.32	0.43	0.51	0.48
	RUS	Acc	0.89	0.91	0.83	0.84	0.91	0.84	0.82	0.86	0.86	0.88
		Prec	0.30	0.33	0.21	0.22	0.20	0.26	0.18	0.29	0.29	0.31
		Rec	0.40	0.25	0.53	0.51	0.10	0.69	0.46	0.70	0.67	0.65
		F1	0.34	0.28	0.30	0.30	0.13	0.38	0.26	0.41	0.40	0.42
	SMOTE-ENN	Acc	0.82	0.92	0.81	0.80	0.19	0.86	0.82	0.91	0.88	0.89
		Prec	0.22	0.33	0.19	0.18	0.07	0.26	0.21	0.38	0.33	0.33
		Rec	0.66	0.13	0.54	0.53	0.88	0.61	0.60	0.51	0.67	0.62
		F1	0.33	0.18	0.28	0.27	0.13	0.37	0.31	0.44	0.44	0.43

Table 8: Results with PCA and Simple Imputing for each year

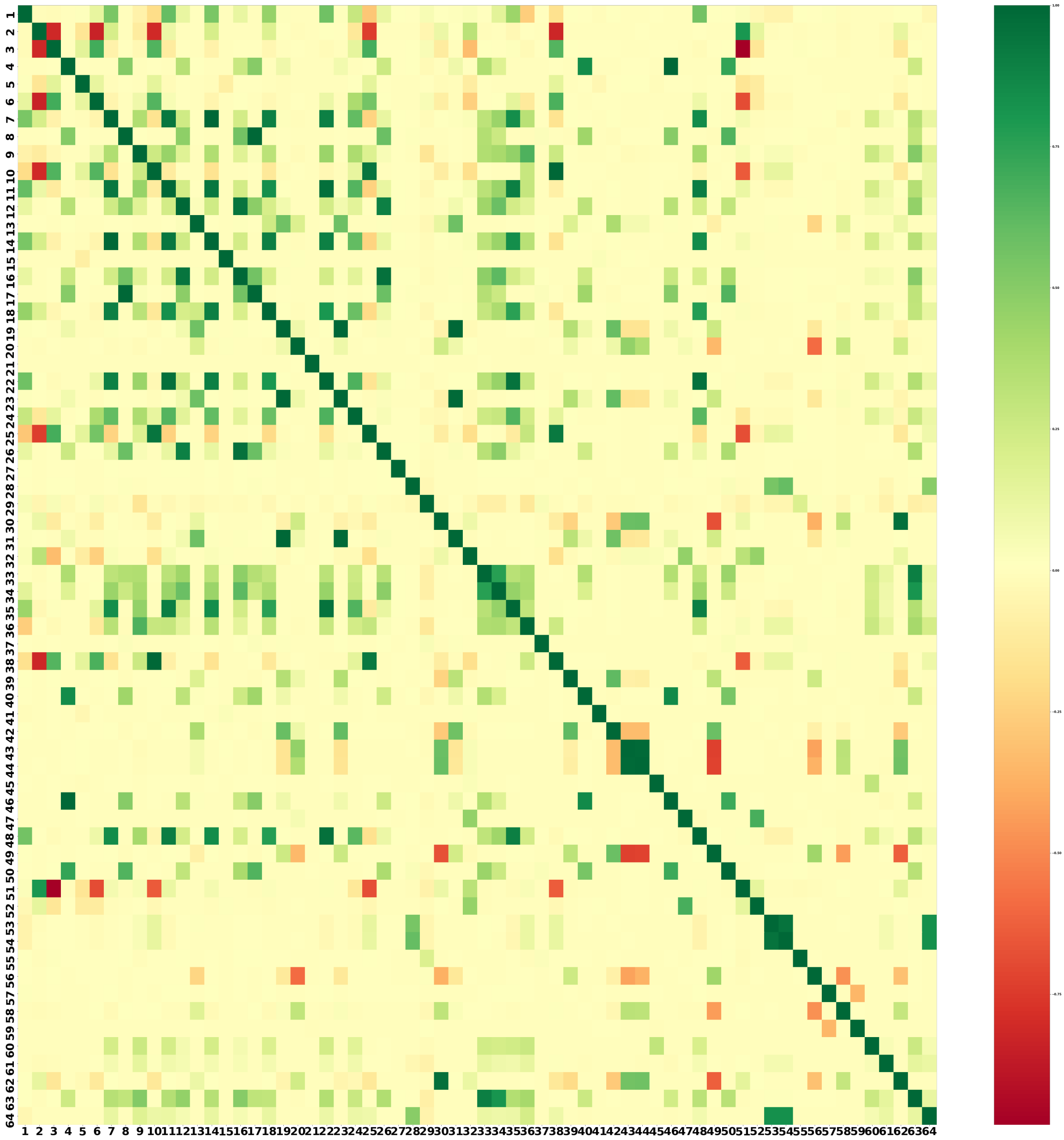

Figure 2: Correlation matrix for the dataset