Classifying Non-Financial Private Corporate (NFPC)Sector* – Issues and Efforts

Brijendra Singh, Meera A. P., Saumya Mishra

From manual profiling of a small number of large companies to construction of a frame for Industrial Activity Codes(IACs), from use of administrative databases (eg. MGT-7) and other surveys (eg. ASI) to exploration of Machine Learning for classification, the efforts to classify the NFPC sector have come a long way. Amidst the balancing concerns of comparability and representativeness, the reclassified estimates have been incorporated in 2017-18. While the impact of reclassification on relatively bigger segments like manufacturing has been minor, smaller segments like storage have been significantly impacted and share of other services (including the residual) has significantly reduced. Even though the classification based on IAC embedded in CIN has now been significantly reduced leading to activities being classified more in line with the actual present business, sources like MGT-7 also have to be treated with caution on multiple counts: lack of proper understanding of codes, skewed classification of vertically integrated businesses, erroneous classification as Financial company based on property receipts in case of presently inactive business etc. Probably, Machine Learning algorithms using industry specific structural ratios coupled with administrative databases would help in improving the classifications further.

**NFPC in the present context excludes Quasi Corporations and only includes entities registered under Companies Act*

1. Background and Evolution of Classification

The Non-Financial Private Corporate (NFPC) Sector is one of the important institutional sectors of the Indian economy in terms of contribution to Gross Domestic Product (GDP), Gross Fixed Capital Formation (GFCF) etc. and hence compilation of National Accounts Statistics for this sector assumes a crucial importance, both for research and policy-making. Since the last base revision, the estimates of macro- economic aggregates of NFPC Sector (used in the present context as a referent to only the non-financial entities registered under Companies Act) are compiled by analyzing the MCA-21 database as provided by Ministry of Corporate Affairs (MCA). Though the MCA-21 database inter-alia includes data of Financial Corporations and PSUs as well, the same is excluded from the compilation process of NFPC Sector. The PSUs are identified using Corporate Identification Number (CIN) (SCG and GOI as constituent) as well as the list received from concerned NDE unit. As regards Financial Corporations, identification is based on the codes embedded CIN, list received from concerned unit, MGT-7 etc. and the same is finalized in consultation with the Unit dealing with Financial Corporations to obviate duplication and exclusions.

Earlier, in general, along with manual profiling of a small segment of large companies, the industrial activity codes embedded in the CIN of the companies were used to categorize the companies into different industry groups. However, the limitations of using CIN to reflect industrial activity of the company were evident right from the beginning. The same had led to manual profiling of large sized companies. The reasons for divergence of the activity from the one indicated in the CIN is largely the diversification of the company from the initial intent, over time. Even though there is a provision of getting the CIN changed as and when industrial activity of a company changes, the table below indicates that the same is resorted to less often by the companies. About two third changes are on count of change in listing status, conversion from one type of company to another and the company changing its registered address from one state to another, even though businesses are increasingly resorting to acquisitions, mergers, product diversification etc. leading to rapid change in the primary business activity.

Table 1: Summary of Reasons for CIN Change

Type of Change	Percentage w.r.t. total*
Listed/ Unlisted	9.12
Activity	32.92
State	36.66
Type of company	19.55
Other	10.60

** There may be overlapping cases hence the sum is not equal to 100*

Consequently, there is good chance of the company being mis-classified in terms of industrial activity being pursued in case the same is based on CIN. A few examples of such cases are also given in the table below:

Table 2: Examples of companies misclassified as per codes from CIN

Sl. No.	CIN	Name	Code from CIN	Corrected Code
1	U72900GJ2007PLC105869	Reliance Jio Infocomm Limited	K3	I5
2	L32102KA1945PLC020800	Wipro Limited	D1	K3
3	L74999MH1994PLC077041	JSW Energy Limited	K5	E1
4	L32100GJ1996PLC030976	Vodafone Idea Limited	D1	I5

2. Efforts towards improving classification

With a view to give more accurate picture of the performance of different sectors of the economy, multiple efforts have been made towards refining the classification. These include manual profiling of a bigger set of companies, construction of a frame for Industrial Activity Codes (IACs), use of parallel data sources etc. These are detailed in the following sections.

2.1 Manual Profiling of Bigger Companies, Frame of IACs & use of CIN Change history

As a small corrective step, initially for some of the bigger companies the industry codes were verified manually from their annual reports or from the websites of the companies. Even though count wise such effort was quite limited, the intent was to ensure that at least half of the value addition in the economy was correctly tabulated. The table below indicates summary statistics of the same for the year 2015-16.

Table 3: Activity Classification based on Manual profiling and CIN based information

Sl. No.	Attribute	IAC based on Manual Profiling		IAC based on CIN	
		Total*	Share(%)	Total*	Share(%)
1	Count	31603	5.38	555852	94.62
2	Share Capital (in Rs. Crore)	501241	34.35	958060	65.65
3	GVA (in Rs. Crore)	1766254	59.66	1194167	40.34

*Based on unadjusted values

Initially it was thought that IAC for current year estimation may be copied from the previous year and only in remaining cases CIN based information would be used. Further, after such classification was done, top companies in each compilation category (CC)* would again be scrutinised for new large entrant to

CC: most diasaggregated Industrial activity group used for supplying of data for National Accounts Statistics ensure that bigger companies were correctly classified in each CC through manual profiling. However, it was observed that in some cases the IACs assigned earlier on the basis of manual profiling were missed subsequently as the company did not file in intermittent years or changed its CIN after initial assignment. The following table illustrates the loss of information after manual profiling. Accordingly, frame of the IACs ever assigned to a company was first constructed and the same was considered together with the CIN change history to ensure that the gains made earlier were carried forward.

Table 4: Examples to illustrate loss of information after manual profiling

Sl No	Name of Company	Last year when correctly profiled	Correct classification step	Year when the mistake was detected	Reasons for information loss	Step that led to Identification	Corrective step
1	Jk Tyre & Industries Limited	2012-13	CIN based classification from Financial Activities changed to Manufacturing activity through manual profiling	2015-16	Change of CIN from L67120WB1951PLC019430 to L67120RJ1951PLC045966	Mistake identified during the scrutiny of new large companies, CC wise.	Both Frame and CIN Change history is now being used while classifying
2	Transport Corporation Of India Limited	2013-14	CIN based classification from Real Estate changed to Transport activity in 2012-13 through manual profiling was carried forward	2015-16	Change of CIN from L70109AP1995PLC019116 to L70109TG1995PLC019116		

Since 2016-17, MCA has also started sharing the master frame of companies maintained by them enabling a possibility of applying industry-wise multiplier in the future. Presently, overall shortfall between the reporting companies and the active companies is uniformly distributed across all compilation categories using a single scaling up/blow up factor. Initially, assessment of industry wise gap in representation was not possible for want of frame of all active companies and the broad classification industry wise classification of the universe of active companies could not be used due to large scale inaccuracies in classification as was evident from the manual profiling.

2.2 Use of Annual Survey of Industries (ASI) data

The Annual Survey of Industries (ASI) is the principal source of Industrial Statistics in India. As regards Manufacturing Sector which is having a major share within the NFPC Sector, the use of information from ASI for appropriately classifying companies, was also explored as the ASI data also contains information on CIN since 2015-16. The MCA data is based on enterprise approach whereas ASI follows establishment approach. At company level comparison was done between Net Sale Value from ASI data & Revenue from Operation from MCA data and if the Net Sale Value lied within certain range (-30% to +30%), then it was construed that manufacturing was the major activity of the company and the company was classified accordingly. However ASI being a sample survey it has got its limitations and it can be considered only for those companies which are covered in the survey. For multi establishment company in ASI, values in respect of all units were added for comparison with enterprise level MCA data. Table 5 demonstrates the assignment of code using ASI (The names of company and DSL No.s are fictitious).

Table 5: Assignment of code using ASI

S. No.	Name of the Company	From ASI 2015-16		From MCA 2015-16	Code Assigned
		DSL No	Net Sale Value (Rs. Billion)	Revenue from Operations (Rs. Billion)	
1	ABC LIMITED	123456	75.79		D1
		121321	26.83		
		188601	19.36		
		167549	14.47		
		177865	0.06		
		Total	136.52	136.54	
2	DEF LIMITED	110953	4.79		D1
		122534	2.46		
		198765	1.53		
		120678	1.27		
		140892	1.09		
		176211	0.54		
		Total	11.67	11.66	

2.3 Use of MGT-7 data

MGT-7 is an electronic form (Annexure I) provided by the MCA to all the corporates in order to fill their annual return details. This form inter-alia, collects information on "principal business activities" (Main Activity group code (based on NIC), Business Activity Code, % of turnover) of a company (Item II of Annexure I). Using this information the activity corresponding to maximum % of turnover was identified and accordingly classified. Table 6 demonstrates the assignment of code using MGT-7.

Table 6: Assignment of code using MGT-7

CIN	Name	Information from MGT-7					Code as per MGT-7	Corrected code
		Buss. Act. Code	Main Act. Code	Main Act Gp. Desc.	Buss. Act. Description	% of Turnover		
U51900KA2010PTC053234	AMAZON SELLER SERVICES PRIVATE LIMITED	J7	J	Information and communication	Data processing, hosting and related activities; web portal	100	K3	K3
U34100TN2005FTC078835	RENAULT INDIA PRIVATE LIMITED	G1	G	Trade	Wholesale Trading	100	G1	G1
U11100GJ1989PLC032116	NAYARA ENERGY LIMITED	C5	C	Manufacturing	Coke and refined petroleum products	100	D1	D1

3. Reclassification and its impact on 2017-18 estimates

The recent availability of MGT-7 data on principal business activity for a large number of companies (about 6.8 lakh) in 2017-18, from MCA, has enabled large scale reclassification of companies instead of earlier general practice of using industrial activity code embedded in CIN for most of the medium and small sized companies. During the reclassification exercise, MGT-9 data (accessed from Annual reports of Listed Companies) and activity information from Annual Survey of Industries have also been used.

3.1 Different sources considered for Reclassification

As indicated above, for refining the classification multiple sources were considered. Table 7 represents the % share of different sources used for classifying companies of 2017-18 Frame (excluding PSU).

Table 7: Share of different sources used for classifying companies of 2017-18 Frame (excluding PSU).

Code assigned using	All cases		No change in IAC due to use of alternative data source		Alternative data source leading to reclassification (Change in IAC)	
	% Share		% Share w.r.t. all cases		% Share w.r.t. all cases	
	Count	PUC	Count	PUC	Count	PUC
ASI	1.14	9.00	0.88	7.36	0.25	1.64
Manually	0.50	6.94	0.43	3.39	0.07	3.55
MGT-7	52.96	69.55	25.31	42.72	27.65	26.83
SYNTAX	45.41	14.51	45.41	14.51	0.00	0.00
Total	**100**	**100**	**72.03**	**67.98**	**27.97**	**32.02**

3.2 Making Sense of the Sectoral Shift

While working at reclassification, much movement was observed between different compilation categories. To make sense of dynamics at a more aggregate level, two way tables using both new and old classification were constructed. The inter sectoral shift (both in count and in GVA) on account of reclassification in case of Private Corporations for 2017-18 are presented in Annexure II & III respectively. However a sample illustration of the same is given in the table below. The diagonal elements cases wherein there is no change in classification whereas off diagonal elements represent instances of reclassification. Moving across a row would indicate reclassification of older IAC into various new IACs or reduction on count of newer classification (except for the diagonal element) whereas movement down the column indicates addition to newer IAC from different older IACs(except for the diagonal element). Cases of bulk movements were further scrutinised to see if the movement was making sense. The row totals represents values as per old classification and column totals represents values as per new classification.

Table 8: Illustration (shift in terms of count):

Reclassified Code > / Code in use V	A1	.....	D1	...	F1	G1	....	I1	I2	I3	I4	K1	....	K5	...	O4	Grand Total
A1	..	..	1408		:												13913
A2			101		:												799
:			:		:												
F1	301	..	960	..	47690	1469		84	5	8	66	25	..	1692	..	145	56036
:			:		:												
I1			37		:			1553									2439
I2			7		:				268								653
:			:		:												:
I4	..	..	96		..		..	1255	250	179	5768	..	..			..	10113
:			:		:												
K1	..	..	337		8138							28115	..			..	42095
:			:		:												
O4	..	..	4975	..	1047	2626	..	..	..	..	..	..	..	4098	..	6400	29613
Grand Total	..	..	115960	..	64781	72753	..	4250	748	..	8936	41691	..	92088	..	..	623686

Prior to reclassification the number of companies in "Supporting and Auxiliary Transport" Sector was 10113 which was reduced to 8936 after reclassification. This is mainly because many companies got correctly classified into Land Transport (1255), Water Transport (250) and Air Transport (179) from this sector during reclassification. Similarly among other shifts 8138 companies which were earlier classified under "Real Estate" Sector were reclassified to "Construction" Sector as they are involved in construction activities as well. This has resulted in increase in the count of companies under "Construction" Sector from 56036 to 64781.

3.3 Impact of Reclassification

Though there is noticeable impact of reclassification at sectoral level in 2017-18, the same is likely to become more consistent in subsequent years as large scale changes in major activity, on annual basis, is unlikely. The reclassification using new MGT-7 data will be attempted once for each National Accounts Statistics. Some general shifts observed while reclassifying companies are listed below:

- Some companies which were earlier classified under "Supporting and auxiliary transport" are shifted to land transport and water transport.
- Companies engaged in food processing, manufacturing of food products, production of animal feeds, milk products etc. got reclassified in into "manufacturing" sector.
- Some companies which were earlier misclassified under real estate sector are involved in construction activities as well and hence are now reclassified into "Construction" sector.
- With the availability of information on Principal Business Activity, many companies which were earlier misclassified under "Other Services" sector could be classified into appropriate Sectors.

Table 9 depicts the Industrial Activity wise share in count and GVA estimates both prior to and post reclassification in 2017-18 along with the percentage change. Even though the reclassification has impacted all the sectors, it is particularly pronounced in Mining, Trade, Real Estate, Storage and Other Services.

Table 9: Industry Wise share in respect of count and GVA of companies prior to reclassification and post reclassification along with percentage change*

Sl. No.	Economic Activity	NIC Classification	Prior to reclassification Share (%)		After reclassification Share(%)		% Change	
			Count	GVA	Count	GVA	Count	GVA
1	Agriculture, forestry & fishing	A1, A2, A3, B1	2.75	0.73	2.62	0.49	-6.25	-33.06
2	Mining & quarrying	C1,C2,C3	1.02	2.02	0.74	1.39	-28.41	-31.23
3	Manufacturing	D1	23.47	46.23	20.14	47.76	-15.62	3.23
4	Electricity, gas, water supply and other utility services	O1, E1, E2, E3,E4	1.26	3.30	1.40	3.09	9.13	-6.20
5	Construction	F1	9.57	5.18	11.25	5.87	15.61	13.35
6	Trade, repair, hotels and restaurants	G1,G2,G3,H1	14.67	5.12	21.80	6.36	46.12	23.95
7	Transport, storage, communication & services related to broadcasting	IR,I1,I2,I3,I4,I5,I6,I7,IP	3.66	6.15	5.23	7.69	40.62	24.94
8	Real estate, ownership of dwelling and professional services	K1,K2,K3,K4,K5	34.79	27.52	30.62	24.94	-13.47	-9.47
9	Other services	M1,N1,O2,O3,O4	8.81	3.76	6.19	2.41	-30.88	-35.89
	Total		**100.00**	**100.00**	**100.00**	**100.00**	**-1.67**	**-0.08**

***Considering unadjusted values.**

4. Issues with reclassification and concerns with the parallel data sources

Sources like MGT-7 also have to be treated with caution on multiple counts: lack of proper understanding of codes, skewed classification of vertically integrated businesses, erroneous classification of non- financial companies as finance company etc. For example a company engaged in both mining and manufacturing of products from the ore will normally be classified as a manufacturing company as the company is likely to sell the end product (unless the company also sells the mined resource without processing and revenue from such sale is greater than the revenue from the sale of the manufactured product). This may lead to skewed classification. Also, sometimes due to certain reasons, a non- financial company may not be able to perform its primary business activity (say, manufacturing) and generates its revenue only from "interest income". In such case, even if the company is earning from interest and has no revenue from sale of products/services, it is not to be classified as a finance company even if MGT-7 data indicates things on the contrary. Following Table gives few examples of MGT-7 information leading to misclassification.

Table 10: MGT-7 information leading to misclassification

CIN	Name	Information from MGT-7					Code as per MGT-7	Corrected code
		Buss. Act. Code	Main Act. Code	Main Act Gp. Desc.	Buss. Act. Description	% of Turnover		
L27204RJ1966PLC001208	Hindustan Zinc Limited	C7	C	Manufacturing	Metal and metal products	100	D1	C1
U51395HR2006PTC064080	Panasonic India Private Limited	G1	G	Trade	Wholesale Trading	93.24	G1	D1
L01111DL1985PLC021329	Focus Agro Products Limited	K8	K	Financial and insurance Service	Other financial activities	100	J1	D1
L45203MH1996PLC281138	Gmr Infrastructure Limited	K8	K	Financial and insurance Service	Other financial activities	67	J1	F1

Many a times the information on activity of a company available in MGT-9 Form (attached along with Directors Report) differs from the activity indicated in MGT-7. Even though the information contained in MGT-9 Form is more authentic, it is difficult to be used because the form is not available in digitized format. Table 11 gives an example of mismatch between MGT-7 and MGT-9 information.

Table 11: Mismatch between MGT-7 and MGT-9 information

CIN	Name	Information from MGT-7					IAC as per MGT -7	Industry: MGT-9	IAC MGT -9
		Buss. Act. Code	Main Act. Code	Act. Group	Buss. Act. Description	% of Turn over			
L24222 HR1902 PLC065 611	Shalimar Paints Ltd	R1	R	Arts, entertainment and recreation	Creative, arts and entertainment activities	99.9	O3	Manufacturing	D1
L24110 GJ1993 PLC019 094	Diamond Infosystems Ltd	C6	C	Manufacturing	Chemical & chemical products, pharmaceutical, medicinal chemical & botanical products	100	D1	Data processing, software development & computer consultancy services	K3
L45201 DL1983 PLC016 821	Ansal Housing Ltd	L1	L	Real Estate	Real estate activities with own or leased property	98.8	K1	Construction of Building	F1

5. Machine Learning for Classification

As discussed above there arise several issues while trying to classify the companies through manual interventions using different methods detailed above which hinder the efforts to correctly reflect the true economic picture. One of the possible solutions to correctly classify a company is by using supervised classification techniques in Machine Learning. In this work, attempt is made to address the problem of misclassification by introducing some machine learning algorithms which combines several parameters and meta-data (financial variables in this case) of a firm. In particular, the classifiers that have used, exploit the training set to correlate financial variables such as Property Plant & Equipment, Inventories, cost of material consumed etc. to two labels or classes i.e., the industry group "Construction" and "Real Estate". In the sequel, it applies this information to classify the rest of the firms. To implement these classification algorithms, high level language "Python" is used.

5.1 Supervised Classification Methods used

The study of classification in statistics is vast, and there are several types of classification algorithms that can be used depending on the dataset one is working with. Below are six of the most common algorithms in machine learning that have been used in this paper:

1. Decision Tree(DT)
2. Random Forest(RF)
3. K- Nearest Neighbour (KNN)
4. Logistic Regression(LR)
5. Multilayer Perceptron(MLP)
6. Support Vector Machine (SVM)

```
DT_model = DecisionTreeClassifier()
RF_model = RandomForestClassifier()
knn_model = KNeighborsClassifier()
lreg=LogisticRegression()
mlp=MLPClassifier()
svm=LinearSVC()
```

The definitions of the above methods are given in Annexure IV. Besides the supervised classification, several unsupervised algorithm were also available for classification but due to paucity of time we had to restrict ourselves to the above mentioned classificatory models.

5.2 Data Analysis

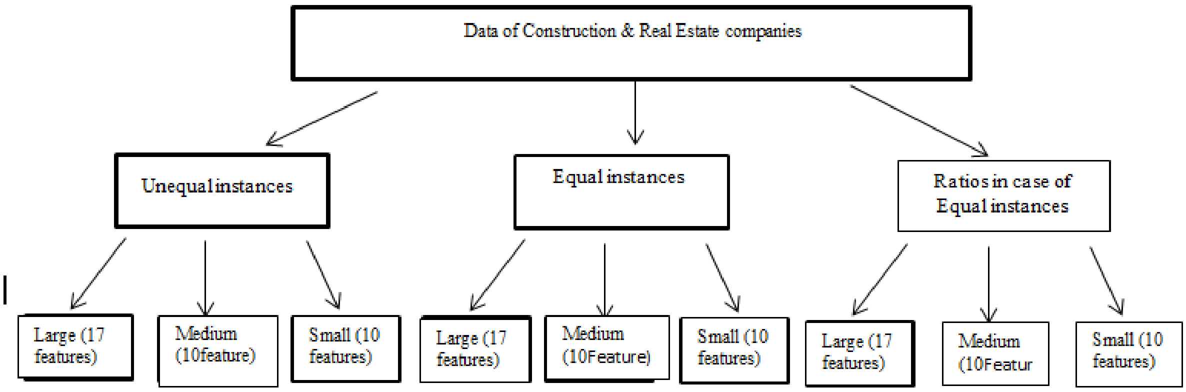
In this section, the classification performance of the above mentioned six classifiers for large and small datasets with different set of features and parameters was analysed. The objective of this comparison was to get the idea as how the model behaves when the input data and model parameters are changed and finally select an appropriate model for the specific problem.

5.2.1 Data sets

The data used for classification is data of companies registered under companies Act. For this paper, the data of the companies engaged in the Construction or Real Estate was taken into consideration. It might be the case that the values of features are affected by the size of the data point i.e. a firm. A bigger firm could have large values compared to smaller firm and this might affect the decision making of classification algorithm, particularly the KNN which classify based on the distance measure. To nullify the effect of the size, the **ratios** of features values were also taken into account.

It was known from the theory that some of the supervised classification algorithm, do not work well if the classes have unequal number of instances, therefore, a set of data having **equal number of instances** for both the classes is also created .

Thus, the input data set is divided into three categories; entire data with unequal number of instances & different set of features, data with equal number of instances & different set of features and ratios of equal instances with different set of features. The flow chart shows the data set used for classification:



5.2.2 Selection of Features

In machine learning, a feature is described as the characteristic of the instances being observed. Selecting the features which are informative and discriminatory is one of the crucial steps in any classification algorithm. Initially, in this exercise, **small set of features** were selected on a priori basis assuming that companies engaged in Construction activities have high Cost of material consumed and low inventories than those performing Real estate activities. On the similar lines, the Property plant & equipment and Purchase of stock in trade could be one of the classifying factors.

It was also felt that the inclusion of other relevant features could lead to improved results in terms of accuracy of classification. For this purpose, the information of top 5 companies in both the classes was examined and the features which showed discriminatory behaviour for two classes are taken into consideration. This resulted in **large set of 17 features**. Among these, some features which had blank or zero values for several instances and did not appear to be behaving as classificatory variables were dropped. This led to third **medium sized set** containing only 10 features. List of features is given in Annexure V.

5.2.3 Model Design

The data used to train the algorithm comprised of 70% of the entire dataset and the remaining 30% was used for testing purposes. A general syntax in python for train-test-split is as follows:

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.3)
```

where test_size=0.3 indicates the proportion of the entire data to be used for testing the algorithm.

The train and test data split is made **randomly** by each algorithm and so every time the algorithm was performed the accuracy of classification, defined as the correct prediction of the input data into labelled classes, came out be different and kept on fluctuating. To overcome this issue of varying accuracy in every run, the process was **repeated 100 times** and **average of all the values of accuracy** obtained in iterative process was taken into account for comparative purposes. This is illustrated below:

```
for i in range(1,100):

  X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=i) #splittng the data set

   model_i = Classifier()

  model_i.fit(X_train, y_train)

  y_predict_i =   RF_model_i.predict(X_test)

  accuracy_i = metrics.accuracy_score(y_test, y_predict_i)

  result.append(accuracy_i)

  i=i+1                                    # put the result on a list within the for-loop

  avg_accuracy=mean(result)                              #computing average accuracy

  min(result)                                   #minimum value of accuracy in iterative process

  max(result)

  standard_dev=stdev(result, average _accuracy)
```

But even in the iterative process, due to random split of train_test data, the accuracy would vary and the range of all the 100 accuracies could be large. To understand how **consistent** the classifier was in repeated run, the **minimum and maximum** value along with **the standard deviation** were also noted.

One of the ways to improve the accuracy of the classification is to select the optimal hyperparameters. Parameters which define the architecture of a model are called hyperparameters and choosing the optimal values of these parameters to train the algorithm is called **hyperparameter tuning**. In python, using the Randomised Search Cross Validation, machine itself is able to do this tuning by sampling the best set of parameters from the parameter grid. The same is illustrated for Logistic Regression Classification method:

```
model= LogisticRegression()

cv=RepeatedStratifiedKFold(n_splits=10,n_repeats=3,random_state=1)

space=dict()

space['solver']=['newton-cg','lbfgs','liblinear']

space['penalty']=['none','l1','l2','elasticnet']

space['C']=[1e-5,1e-4,1e-3,1e-2,1e-1,1,10,100]

search=GridSearchCV(model, space, scoring="accuracy",n_jobs=-1,cv=cv)

result=search.fit(x,y)

result.best_score_
```

5.3 Results

Table 12 compares the result of all the six classifiers on different datasets with different features.

Table 12: Comparison of classifiers

Classification Algorithm	Data set	Feature Set	Avg_accuracy	std dev	min accuracy	max accuracy
Decision Tree	Unequal instances	Large	0.61	0.01	0.47	0.73
		Medium	0.64	0.01	0.62	0.66
		Small	0.65	0.01	0.61	0.67
	Equal instances	Large	0.68	0.05	0.55	0.78
		Medium	0.67	0.06	0.52	0.82
		Small	0.66	0.06	0.51	0.80
	Equal instances Ratio	Large	0.65	0.01	0.63	0.67
		Medium	0.64	0.01	0.62	0.67
		Small	0.66	0.01	0.62	0.68
Random Forest	Unequal instances	Large	0.69	0.05	0.57	0.82
		Medium	0.69	0.01	0.66	0.71
		Small	0.68	0.01	0.66	0.70
	Equal instances	Large	0.76	0.04	0.66	0.85
		Medium	0.75	0.05	0.63	0.85
		Small	0.72	0.05	0.58	0.84
	Equal instances Ratio	Large	0.72	0.01	0.70	0.74
		Medium	0.70	0.02	0.66	0.75
		Small	0.73	0.05	0.60	0.84
K-Nearest	Unequal	Large	0.61	0.05	0.48	0.73

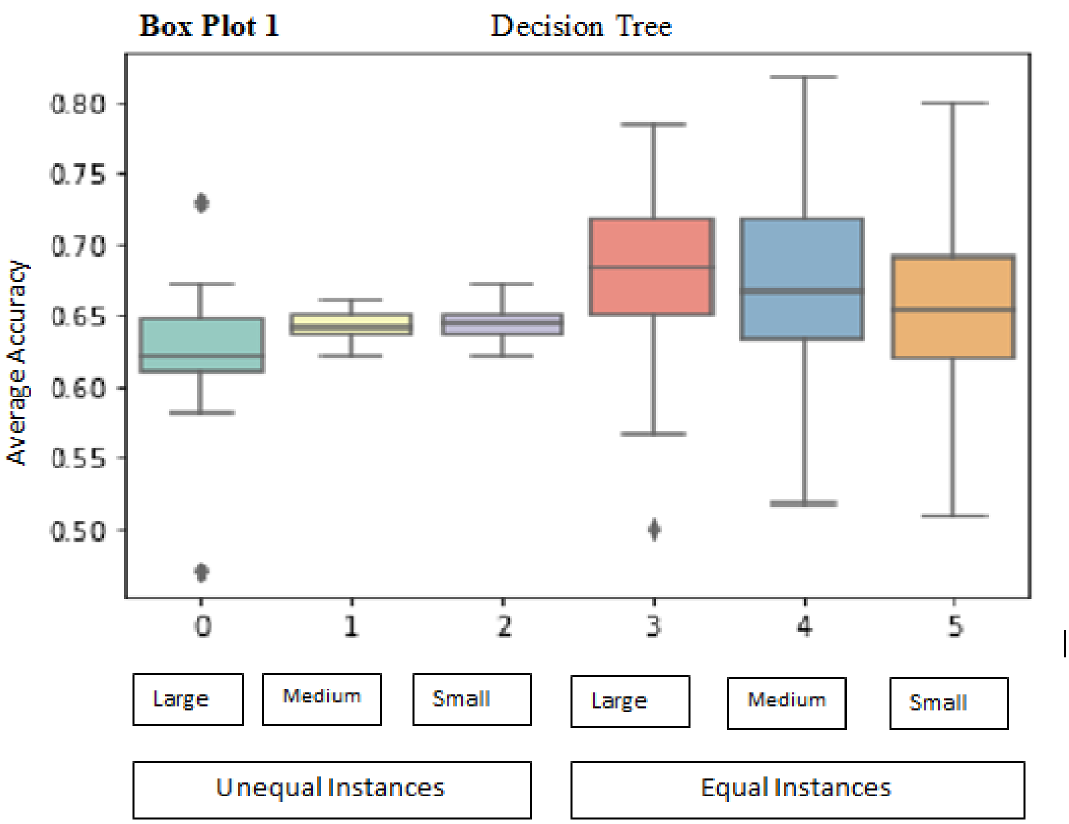
Neighbour	instances	Medium	0.63	0.05	0.52	0.78
		Small	0.60	0.05	0.51	0.78
	Equal instances	Large	0.77	0.04	0.65	0.88
		Medium	0.78	0.05	0.65	0.90
		Small	0.70	0.05	0.58	0.82
	Equal instances Ratio	Large	0.69	0.01	0.67	0.72
		Medium	0.69	0.01	0.67	0.72
		Small	0.68	0.01	0.66	0.71
Logistic Regression	Unequal instances	Large	0.63	0.02	0.61	0.69
		Medium	0.63	0.02	0.60	0.7
		Small	0.71	0.02	0.59	0.74
	Equal instances	Large	0.69	0.06	0.48	0.82
		Medium	0.72	0.06	0.58	0.84
		Small	0.67	0.07	0.41	0.83
	Equal instances Ratio	Large	0.62	0.06	0.44	0.76
		Medium	0.64	0.06	0.46	0.78
		Small	0.49	0.10	0.33	0.72
Multilayer Perceptron	Unequal instances	Large	0.64	0.03	0.51	0.69
		Medium	0.63	0.04	0.51	0.7
		Small	0.63	0.09	0.45	0.73
	Equal instances	Large	0.67	0.06	0.48	0.82
		Medium	0.67	0.07	0.50	0.82
		Small	0.69	0.07	0.50	0.83
	Equal instances Ratio	Large	0.65	0.07	0.52	0.78
		Medium	0.63	0.06	0.50	0.78
		Small	0.68	0.09	0.30	0.85
Support Vector Machine	Unequal instances	Large	0.53	0.07	0.34	0.68
		Medium	0.51	0.08	0.33	0.69
		Small	0.55	0.12	0.28	0.74
	Equal instances	Large	0.59	0.10	0.30	0.84
		Medium	0.59	0.09	0.38	0.78
		Small	0.60	0.11	0.35	0.8
	Equal instances Ratio	Large	0.62	0.07	0.46	0.78
		Medium	0.63	0.07	0.30	0.8
		Small	0.52	0.11	0.26	0.83

It was observed that the accuracy of the classifiers was affected by the following factors:

1. **Input data:** It is evident from the table that for a given algorithm, the average accuracy for the dataset with **unequal number of instances** with either of the feature set is **lower** than that obtained from the dataset with equal instances. For example, in Random forest, the average accuracy

for dataset with unequal instances and large features set is 69% and has improved to 76% for equal instances dataset with same set of features.

2. **Feature pruning:** Pruning is a data compression technique in which the size of the classifier is reduced by eliminating the sections which are non-critical for classifying the data points. In case of equal instances, where the classifiers generally performed better, **feature pruning beyond an extent reduced the performance** for DT, RF, KNN, & LR as all the classifiers showed lower accuracy with small feature set. However, MLP and SVM showed no impact (large to medium feature set) or improvement (medium to small feature set) due to pruning . Opposite performance of pruning in case of DT (improvement from 61% to 65% & deterioration from 68% to 66%) for equal & unequal instances respectively is shown below:



Box Plot 1 shows that not only feature pruning led to improvement in average accuracy of DT with unequal instances, it also led to some moderation in variability, which initially decreased upon pruning and then increased slightly in case of small feature set. Whereas in unequal instances the performance on both counts i.e. average accuracy and the variability decreased. Even though equal instances classifier cases performed better in case of DT as assessed by average accuracy over 100 iterations, the variability in the performance was much more (average standard deviation 0.06) compared to unequal instances cases (average standard deviation 0.01), across all feature set.

Efforts were made at dimension reduction (for both large and medium feature set) using **Principle Component Analysis (PCA)** and thereafter using the classifier on the reduced set of dimensions. However, no improvement in the classification was observed. This happens at times because PCA is based on extracting the axes on which data shows highest variability and can be of much use in unsupervised learning algorithms, though there is no guarantee that the new axes are consistent with the discriminatory feature of supervised classification problem as the PCA is agnostic to target variable (class label).

3. **Model Optimization (Hyperparameter tuning):**

Usually, the models are finalised keeping in mind **variance-bias trade** off which results from under fitting/over fitting of models. Models performing poorly over train and test set are called **under fitted** whereas those performing too well on train data with significant drop in performance on test data are called **over fitted**. Over fitted models have high variance whereas very simple models have high bias. To get an optimal model with maximum prediction accuracy (minimum total error on account of bias and variance), a workable way could be to choose the parameters of the classifiers such that:

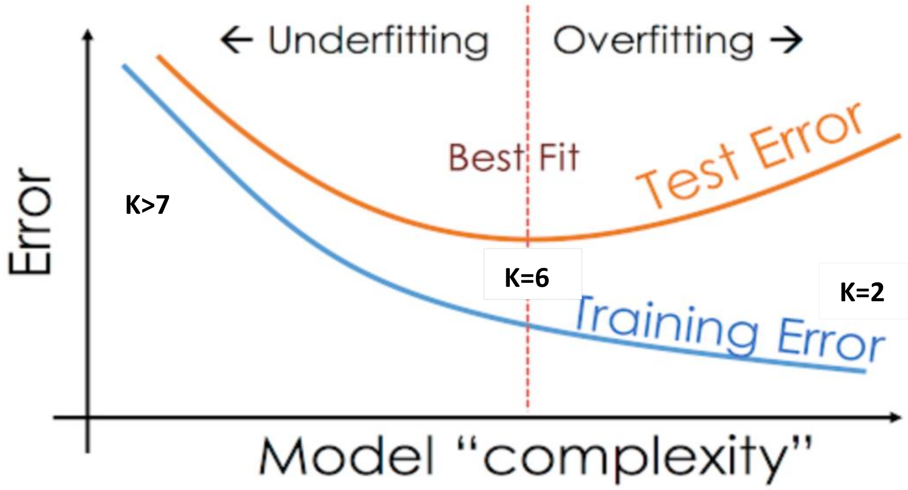
- the testing score is the highest, and
- both the test score and the training score are close to each other

For example, in KNN, using small number (K=1) of neighbours over specified the model to fit each data point in the training set resulting in perfect prediction(100 per cent accuracy) in training set and less accuracy(75%) in test data. The accuracy in case of test data improved with increasing neighbours to an extent (K=5/6) and thereafter the accuracy decreased due to oversimplification/generalisation and the optimal model as a trade-off was found at 6 neighbours.

Table 13: Finding value of neighbour for optimal KNN

	K=1	K=2	K=3	K=4	K=5	K=6	K=8
Test accuracy	0.75	0.73	0.77	0.78	0.80	0.81	0.76
Train accuracy	1.0	0.88	0.86	0.86	0.81	0.82	0.79

K denotes the number of neighbours



Apart from manually trying to locate best parameters, **Randomised Search** (RandomizedSearchCV) and **Grid Search** (GridSearchCV) options available in sklearn library can be used. Improvement in average accuracy using the best parameters parameters thrown up by Gridsearch for the logistics regression is shown below:

Logistics Regression Classifier : Hyper Parameter Tuning Results

	Avg Accuracy	Std Dev	Min. Accuracy	Max. Accuracy
Default Parameters	0.72	0.06	0.58	0.84
Tuned Parameters	0.77	0.06	0.62	0.92

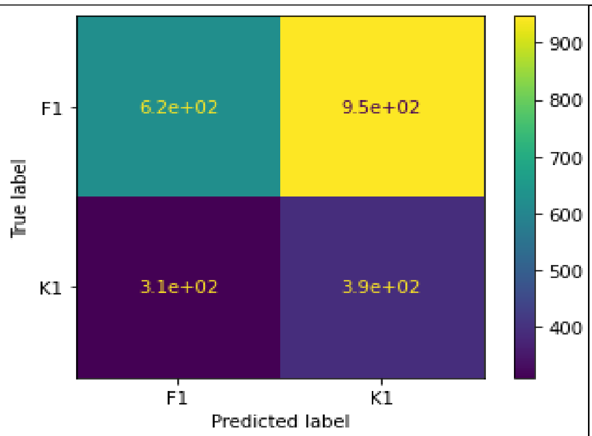
Above results are based on 100 iterations

4. **Size Effect :** It was expected that differentials in the size of the cases and different features (for same cases) might affect the performance of classifiers. Hence converting features' values into **ratios** was expected to yield better results. However, no improvement in accuracy was observed except for SVM with equal instances and large/medium set of features where it has improved from 59% to 62%/63%. Elimination of size differential in features was also undertaken . It was expected that **feature scaling** might lead to improvement in the accuracy, specially in case of distance based classification algorithms such as SVM and KNN even though classifiers like decision tree and random forest are usually scaling invariant. However, no significant improvement was observed probably because the features were not drawn from very different scales to start with. Standard scaler available in *sklearn* library was used for the exercise:

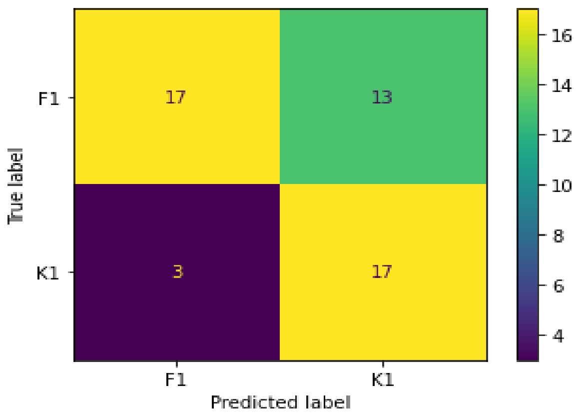
```
x=StandardScaler().fit_transform(x)    #x denoted the set of features
```

5. **Skewness in Misclassifications:** Behaviour of equal and unequal instances in different classifiers, in terms of skewed misclassification was assessed. It was observed that distribution of misclassified labels was also dependent on the values of features in training data set besides the equality of the cases in the same. Even after checking for equal representation of instances post train-test split, skewed misclassification was thrown up in many iterations. Confusion matrix indicating same is given below:

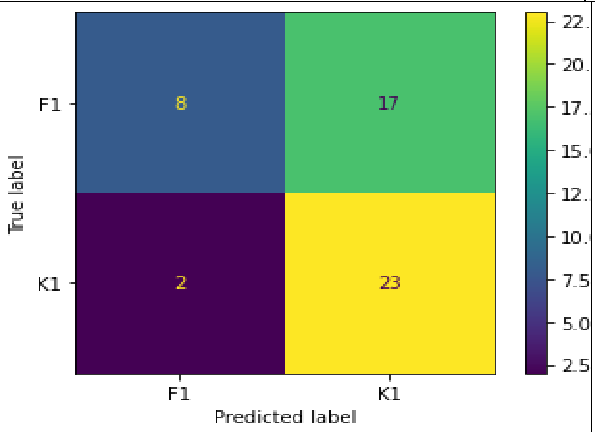
Table 13: Confusion matrix for SVM



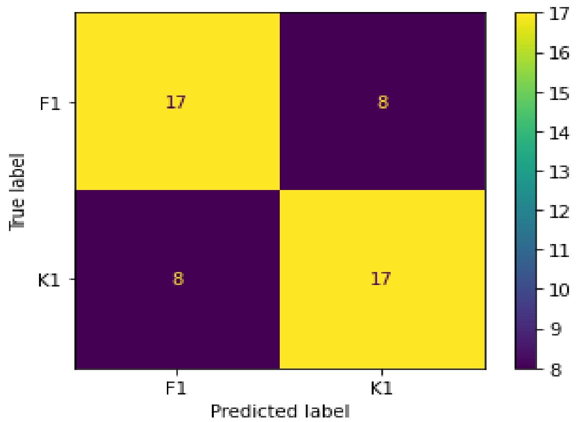
Unequal Instances



Equal instances, Unequal after train-test split, Skewed misclassification



Equal instances, Equal after train- test split, Skewed misclassification

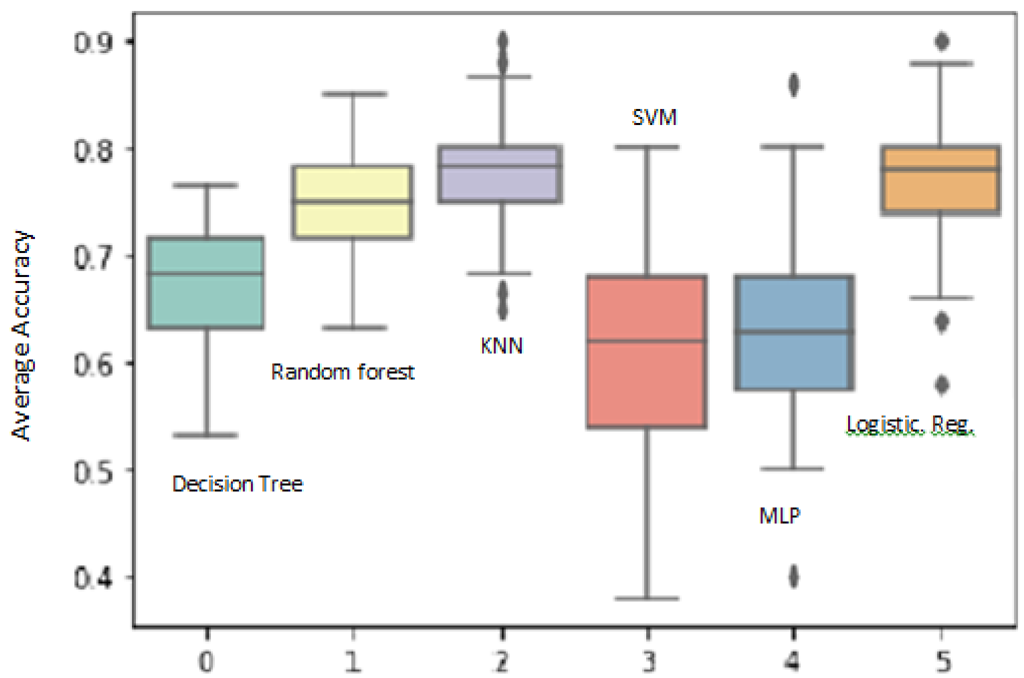


Equal instances, Equal misclassification

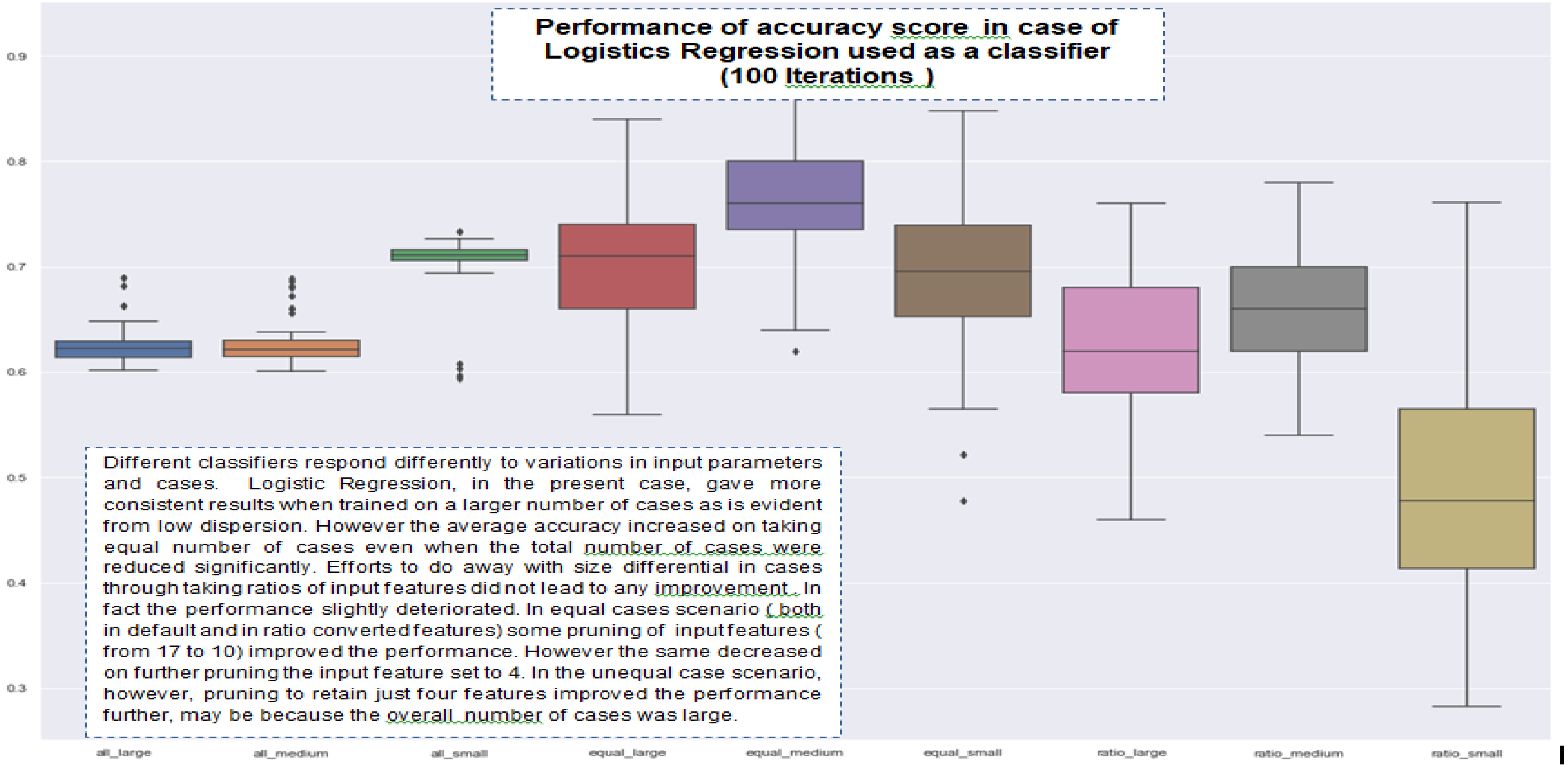
6. **Selection of classification model:** Model selection can play an important role in achieving high accuracy of classification. Behaviour of different

classification algorithm on the same input data and same set of features may be seen from table 12. In general, **random forest, KNN and logistic regression performed better** with average accuracies whereas **SVM showed worst performance** in terms of lower average accuracy as well as **high standard deviation** with accuracies ranging from 26% to 84%. Performance of different classifiers with equal instance of the categories in input data and medium feature set is given in the box plot below. Random forest, KNN and Logistic regression had average accuracies more than 75% (averaging for 100 iterations) and less dispersion .

Box Plot 2 Models with Equal instances & Medium feature set



The impact of change in input data, set of features, size differential taken altogether on the Logistic Regression model, in the pictorial format is given below:



6. Conclusion and Way Forward

As discussed in this paper the exercise of reclassification of companies, so far, was based on manual profiling of large companies and use of databases like MGT-7, ASI etc. It is likely that though the current interventions have helped in reducing misclassification and reflecting true economic scenario, manual intrusion has got its own limitations while dealing with large scale data sets and use of ML algorithms is a probable alternative for reducing the manual intervention which is expected to yield more objective outcomes. Machine learning can be a useful tool for both diagnostics (when the classification is available from some other data source) and classification (in case no support from another database is available). The inputs like number of features, input data (equal/unequal instances), classifier etc. need to be selected appropriately for task at hand. Such exercise is helpful not only for overlapping cases such as Construction & Real Estate, Manufacture & Trading etc. but also in partitioning Financial and Non-Financial companies in the first place where the algorithm are expected to behave much better.

7. Disclaimer

Though the authors are working in the National Accounts Division, National Statistical Office, MoSPI, Government of India, the views expressed are personal and do not necessarily reflect the position of Government of India.

Data Sources:

1. *Data from MCA*
2. *Data from ASI*

Annexure II

The inter sectoral shift in Count on account of reclassification in case of Private Corporations for 2017-18

Reclassified Code >/ Code in use V	A1	A2	A3	B1	C1	C2	C3	D1	E1	E2	E3	E4	F1	G1	G2	G3	H1	I1	I2	I3	I4	I5	I6	I7	IP	IR	J1	K1	K2	K3	K4	K5	M1	N1	O1	O2	O3	O4	Grand Total
A1	9270	1	86	48	22	1		1408	19	5	4		158	1343	516		73	8			9	7	4	138			161	326	15	19	27	187	15	17	2	2	11	11	**13913**
A2	356	109	7	10		1		101					6	106	29		7							5			11	22	2	3		13	1	6			4	0	**799**
A3	94		109	4	1		1	124	1				6	69	21		7				2	1	2	5			10	12	1	4		12		2	1		1	2	**492**
B1	72	1	2	519	3			91	3				9	88	36		9	3	1		5	1		6			7	15	2	1	1	15		3	1	1	2	2	**899**
C1	43				2618	5	64	1131	12	1	4		94	470	205		10	27	1		8	3	2	8			61	83	8	8	1	101	2	2	1		1	3	**4977**
C2	3		1		22	172	1	89	12	33	1		13	44	20			5	1		8	1		4		1	6	9	2	1		58	3		4		1	0	**515**
C3	2				61		174	46	11				12	85	25		2	5			1	2		2		1	17	8	4			22			1			0	**481**
D1	1584		28	94	371	40	23	89542	449	225	142	1	1639	20044	8862		825	172	50	14	176	588	755	598	8	6	1569	1910	443	686	198	4684	129	894	166	15	187	307	**137424**
E1	15			1	8	6	4	315	2956	37	4	1	137	134	83		4	3	1		2	12	1	2		1	39	38	6	18	3	219	2	2	19		1	4	**4078**
E2	1				1	3		47	24	127		2	4	17	15			1	1		3	3					2	4				16	1		1			1	**274**
E3	7			1	5	2	1	154	484	18	754	10	55	101	76		3	1	1	1	1	8	1	1			20	9	2	5	1	103	2		40		1	6	**1874**
E4	4							84	350	11		146	32	48	22			2				1	1				5	8	2	3	1	44		1	8			2	**775**
F1	301		3	1	107	3	9	960	132	13	9		47690	1469	1125		406	84	5	8	66	97	22	98	1	1	1081	25	168	102	1	1692	56	43	40	4	69	145	**56036**
G1	372		2	32	120	3	24	4692	71	35	19		1642	29696	7569	4	206	159	11	6	152	199	75	121	4	3	4173	3618	242	243	16	2850	42	189	20	12	70	142	**56834**
G2	77		3	5	22		1	1019	20	2	2		211	3139	7299		64	30	2	1	23	109	16	23	3	1	638	438	43	146	4	656	19	68	6	2	42	54	**14188**
G3	2			1		2		363		5			19	603	1333	816	9	31	1		17	2	3	5	1		32	39	18	6	1	120	1	1	1		2	12	**3446**
H1	66		3	3	3		1	192	4	1	1		263	151	160		9326	12		1	143	13	5	7	1		88	500	31	7		252	11	55	1	3	88	47	**11439**
I1	3			3	11		1	37		3	1		30	67	57	1	4	1553	23	13	304	3		74	33	1	28	27	40	7		109	2		2			2	**2439**
I2	2			3				7	1				7	13	11		3	19	268	4	175		1	14	4	2	15	4	6			91	2		1			0	**653**
I3								10					3	10	4		1	12	2	185	69	1	1	4	7		1	5	5	3		39	10	1			1	0	**374**
I4	16			2	22	2	1	96	6	2	2		89	181	154		209	1255	250	179	5768	27	9	313	119	28	179	112	155	35		796	24	17	6	1	38	20	**10113**
I5					1			52	1		1		27	74	46			1		1	1	658	110		12		7	6	1	86	1	107	4	2			6	4	**1209**
I6	2							421	2				24	141	114		26	3	1		4	254	2082	1	1		43	61	18	89		412	39	7		6	645	19	**4415**
I7	96		2	6				61	1		1		18	105	33		7	19	1		52	1		1007	3	4	18	40	9	2		37			2			1	**1526**
IP								10					8	19	13		1	9		3	11	77	12		249		7	10	1	5		58	2				6	0	**501**
IR					2			6					5	9	3			51	2		10				2	87	4	1	1			8						0	**191**
J1	140		3	5	23		2	562	18	6	2	1	396	1455	1065	3	110	47	4	2	39	87	41	27	4		30540	1378	147	132	5	1718	44	37	3	6	34	55	**38141**
K1	335		7	3	43	1	1	337	41	2	6	2	8138	1146	796		339	50		1	23	55	10	46	1	1	1159	28115	129	60	3	1026	33	20	7	15	32	112	**42095**
K2	1				2			22	1				17	20	9		3	10	2		5	5	3	1		1	8	21	117	5		24					3	3	**283**
K3	39			1	6	1	1	1054	21	5	3		244	1391	1521		57	34		3	78	669	372	16	8	1	496	489	88	32854	81	6149	708	78	7	6	162	60	**46703**
K4	32		2		3			125	3		1		16	44	30		4				4	31	6	2			19	8	2	63	593	355	64	85	9	3	6	5	**1515**
K5	548		13	34	151	22	9	7084	436	60	109	3	2573	7161	5317		891	370	90	58	1073	2758	1052	416	114	11	5554	2225	322	3610	227	64632	1868	1365	224	48	1559	1148	**113135**
M1	5			1	1			45	1				30	71	52		33	6		1	13	83	26				55	73	7	156	5	449	6010	38	1	7	246	26	**7441**
N1	34		1	2	10			604	16	3	5		70	528	288		87	4		2	16	37	14	5	1	1	156	131	15	77	49	469	149	7925	7	19	126	145	**10996**
O1	2							15	1		23		8	15	7		1				1		1	2			4	4		3		32	3	3	238			15	**378**
O2	9		1					18		1	1		10	10	12		12	1	1		3	5	3				29	12			5	149	50	14	4	263	39	27	**679**
O3	9		1		1	1		61	1		2		31	60	69		115	3	1		30	46	311	2			23	80	11	25		286	20	18	1	12	1590	32	**2842**
O4	374	0	11	5	34	7	12	4975	150	24	32	2	1047	2626	1630	0	467	260	28	29	641	541	452	167	47	5	1675	1825	153	574	30	4098	350	399	65	36	442	6400	**29613**
Grand Total	**13916**	**111**	**285**	**784**	**3674**	**272**	**330**	**115960**	**5248**	**619**	**1129**	**168**	**64781**	**72753**	**38627**	**824**	**13321**	**4250**	**748**	**512**	**8936**	**6385**	**5393**	**3120**	**623**	**156**	**47940**	**41691**	**2216**	**39038**	**1253**	**92088**	**9666**	**11292**	**889**	**461**	**5415**	**8812**	**623686**

*1.For illustration purpose only.

2. Though the Financial Corporations as per MCA are also considered while preparing this table for illustration these are excluded to arrive at estimates of NFPC sector

Annexure III

The inter sectoral shift in GVA on account of reclassification in case of Private Corporations for 2017-18

Reclassified Code >/ Code in use V	A1	A2	A3	B1	C1	C2	C3	D1	E1	E2	E3	E4	F1	G1	G2	G3	H1	I1	I2	I3	I4	I5	I6	I7	IP	IR	J1	K1	K2	K3	K4	K5	M1	N1	O1	O2	O3	O4	Grand Total
A1	6209	19	19	33	33	0		7406	42	1	1		140	1641	290		64	21			71	2	1	152			39	104	10	19	54	255	5	34	4	0	27	9	**16704**
A2	722	1872	11	4		1		2017					6	151	14		12							2			1	2	1	3		54	0	3			2	0	**4879**
A3	39		94	0	0		0	580	4				1	62	181		8				0	0	0	2			2	3	0	0		6		0	12		0	0	**997**
B1	300	65	0	1078	6			910	1				7	175	63		9	3	16		270	0		5			1	81	1	3	0	19		1	0	0	0	154	**3169**
C1	9				36339	6	98	3786	10	188	2		298	1158	230		13	131	0		66	2	62	11			2	51	0	49	0	336	1	0	0		2	0	**42850**
C2	1		0		63	2649	0	21758	11	14	0		92	87	14			14	7		106	0		9		0	1	1	4	5		241	0		28		0	0	**25105**
C3	0				846		1454	92	496				304	289	25		1	52			25	0		1		5	2	0	1			14			1			0	**3608**
D1	3135		28	748	3045	760	272	1479591	-525	908	150	0	8756	43311	24301		3092	638	1766	89	3298	4174	9871	999	24	4	3164	2357	818	24123	1499	15865	973	2373	635	19	232	661	**1641154**
E1	9			6	38	-5	68	6522	78026	992	3	16	2885	279	65		252	69	53		2	256	674	1		0	62	132	5	158	0	674	0	1	320		3	15	**91582**
E2	19				21	78		993	29	4678		55	4	18	8			0	6		0	0					0	0				59	0		0			0	**5969**
E3	5			0	12	0	25	1595	5813	31	2085	1063	149	181	36		1	2	31	2	0	9	0	0			-60	2	0	11	0	77	0		90		3	2	**11166**
E4	0							1093	2407	18		2526	82	115	7			0				1	2				0	1	0	8	0	100		66	-1			0	**6425**
F1	64		1	0	1073	4	723	6442	681	47	340		156072	2170	758		988	337	13	42	5737	994	18	641	2	13	1453	61	149	283	1	3445	33	595	232	0	203	233	**183850**
G1	534		0	288	-1680	0	541	47544	214	39	69		1577	22324	13979	436	472	2467	62	278	820	679	613	491	1	11	2134	1303	226	-2647	19	4455	15	473	70	5	36	188	**98034**
G2	138		1	55	60		4	10752	24	2	4		214	2966	20647		160	93	1	9	193	259	58	72	4	0	113	170	312	1354	2	2196	16	90	5	2	51	-109	**39919**
G3	0			0		0		2260		5			36	1203	3829	5308	17	147	1		92	0	14	2	21		-67	12	26	9	172	236	1	2	0		0	10	**13338**
H1	38		0	0	5		0	1828	0	0	2		720	65	190		23983	87		0	209	27	67	5	0		93	655	16	60		907	39	342	2	1	465	817	**30625**
I1	0			1	168		0	697		27	0		604	249	378	359	17	12859	288	145	1760	-2		468	116	201	1	13	269	3		424	50		13			8	**19116**
I2	0			13				30	6				84	9	9		2	76	2196	401	2486		0	70	34	22	55	-2	45			431	0		3			0	**5971**
I3								22					0	8	5		0	122	8	14233	1631	0	0	68	318		0	17	2	0		118	-21	0			0	0	**16532**
I4	43			30	402	675	20	587	21	8	1		2492	385	208		424	9249	3707	1504	34529	45	6	2682	509	286	343	240	1388	417		2369	55	25	78	1	42	196	**62968**
I5					-4			5207	0		0		509	475	1309			1		-1	44	50002	1164		106		-937	129	0	3716	0	1407	20	2			4	106	**63260**
I6	0							893	5				135	189	112		147	2	1		1	16149	22387	0	1		22	208	16	562		725	36	1		29	1509	108	**43239**
I7	83		11	30				259	10		1		12	115	16		9	76	2		275	0		1343	6	9	0	53	6	0		36			3			0	**2356**
IP								73					4	17	43		0	119		279	34	164	5		3067		4	12	0	1		171	0				54	0	**4046**
IR					0			158					2	18	0			251	2		54				1	232	0	0	4			50						0	**775**
J1	98		0	2	53		0	4531	131	27	9	0	1475	1633	557	44	398	140	2	31	52	1352	106	63	4		299282	1152	123	1756	31	4502	65	49	57	3	270	40	**318041**
K1	160		0	1	305	0	0	1227	447	0	16	76	20799	664	382		1188	122		1	20	47	35	65	0	26	599	21855	140	256	6	1610	307	202	10	6	176	217	**50964**
K2	0				12			143	3				34	62	21		-1	55	0		49	-585	2	0		1	1	144	3891	42		99					6	10	**3990**
K3	-3			0	3	0	0	6847	1100	5	2		680	3855	2051		-84	123		0	-105	14861	1602	108	8	0	3286	1279	119	484747	3680	51068	1623	630	30	0	159	40	**577716**
K4	53		0		79			1597	37		0		30	645	54		24				324	202	70	51			105	6	-1	7601	3181	2588	48	672	35	1	1	0	**17403**
K5	724		18	192	508	-119	73	53044	1155	78	498	226	7368	16865	5049		2824	2382	380	52	3734	5229	4367	1699	570	15	6566	3732	681	40417	1467	153355	2226	2867	641	28	1602	6500	**327009**
M1	0			0	0			85	69				428	228	27		127	5		0	8	61	94				47	-1	12	429	0	640	10585	49	0	3	219	8	**13124**
N1	44		2	1	292			4506	188	4	3		171	871	802		347	35		1	364	235	10	2	21	-82	436	330	102	294	918	1470	227	32001	4	23	64	215	**43898**
O1	0							25	1		86		85	38	6		8				0		386	8			18	13		1		131	0	6	820			231	**1863**
O2	5		1					49		0	0		101	23	13		87	1	6		5	1	20				70	14			3	154	89	5	4	232	28	23	**934**
O3	1		0		-1	0		488	0		0		65	173	102		453	0	0		50	260	3315	0			5	251	43	277		769	4	33	0	5	2853	73	**9220**
O4	178	0	4	0	88	16	102	18465	1377	17	101	162	1975	3525	1392	0	886	1442	84	169	2227	1265	1337	430	297	24	4194	1017	236	2995	29	11466	606	1148	279	137	677	7958	**66299**
Grand Total	**12609**	**1956**	**190**	**2481**	**41766**	**4065**	**3382**	**1694102**	**91783**	**7090**	**3372**	**4123**	**208393**	**106243**	**77171**	**6146**	**35930**	**31122**	**8631**	**17236**	**58432**	**95687**	**46290**	**9451**	**5108**	**768**	**321037**	**35399**	**8649**	**566953**	**11063**	**262521**	**17004**	**41670**	**3376**	**495**	**8690**	**17712**	**3868099**

*1. Based on unadjusted values and for illustration purpose only.

2. Though the values in respect of Financial Corporations as per MCA data are also considered while preparing this table for illustration these values are excluded to arrive at estimates of NFPC sector

Annexure IV

The classification algorithms used in this paper are described below:

Decision Tree:

Decision trees classify the instances by sorting them down the tree from the root to some leaf node, with the leaf node providing the classification to the instance. Each node in the tree acts as a test case for some attribute, and each edge descending from that node corresponds to one of the possible answers to the test case. This process is recursive in nature and is repeated for every subtree rooted at the new nodes.

Random Forest:

Random forest is an ensemble of many decision trees. Random forests are built using a method called **bagging** in which each decision trees are used as parallel estimators. If used for a classification problem, the result is based on majority vote of the results received from each decision tree.

K-nearest neighbours (KNN)

KNN algorithm uses 'feature similarity' to predict the values of new datapoints which further means that the new data point will be assigned a value based on how closely it matches the points in the training set. To determine which of the K instances in the training dataset are most similar to a new datapoint a distance measure, commonly Euclidean distance, is used.

Logistic regression

Logistic regression is a classification algorithm used to assign observations to a discrete set of classes. Logistic regression transforms its output using the logistic sigmoid function to return a probability value.

Multilayer Perceptron (MLP)

MLP is a class of feedforward artificial neural network consisting of at least three layers of nodes: an input layer, a hidden layer and an output layer. The input layer is the initial layer of the network which takes in an input which will be used to produce an output. The hidden layer(s) perform computations and operations on the input data to produce something meaningful. The neurons in the output layer display a meaningful output.

Support vector machine (SVM)

SVM is a representation of the training data as points in space separated into categories by a clear gap that is as wide as possible. New datapoints are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

Annexure V

List of Features

Based on a priori information and data led decision, the features list is generated as follows:

Decision criterion	List of Features
A priori (Small feature set)	Property Plant and Equipment, Cost of material consumed, Purchase of Stock in Trade, Inventory
Data-led after feature pruning (Medium feature set)	Cost of material consumed, Inventory, Non-current investment, Long-term borrowings, Short-term borrowings, Other Income, Other expense, Total Profit, Repairs to machinery, Revenue from Operations
Data-led (Large feature set)	Property Plant and Equipment, Cost of material consumed, Purchase of Stock in Trade, Inventory, Non-current investment, Current investment, Investment Property, Trade receivable, Trade payable, Long-term borrowings, Short-term borrowings, Other Income, Other expense, Total Profit, Rental income, Repairs to machinery, Revenue from Operations