IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 7.105 ⁜ Vol. 9, Issue 4, April 2022

DOI: 10.17148/IARJSET.2022.94122

Risk Analytics in Banking and Financial services

Nisarg Chakravarty, Dr. Sunil Maggu

Information Technology, Maharaja Agrasen Institute of Technology Delhi, India

Abstract: India has been ranked 5th in the list of countries with the highest non-Performing Assets. It implies India possesses more defaulters each year regarding the subject's ability to repay their loans. Noticeably this is an alarming figure and with the rise of Bank frauds, figured at high as 100 Crore per day. With this as our motivation, we've decided to work on a problem that suitably allows us to identify whether or not a bank can engage in business with a client based on their previous history. Using ML techniques taught, our project aims at creating an efficient model where we can appropriately find the right set of metrics from a diverse group of features that help us in identifying whether a subject can repay their loan once taken.

Keywords: Banks, loans, repay, Non-performing assets, frauds, defaulters

INTRODUCTION

The loan-providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it to their advantage by becoming a defaulter. Suppose you work for a consumer finance company that specializes in lending various types of loans to urban customers. You have to use EDA to analyze the patterns present in the data. This will ensure that the applicants who are capable of repaying the loan are not rejected.

When the company receives a loan application, the company has to decide on loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:

- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business for the company
- If the applicant is not likely to repay the loan, i.e. they are likely to default, then approving the loan may lead to a financial loss for the company.

This system aims to identify patterns that indicate if a client has difficulty paying their installments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc and ensure that probable loss to the bank is minimized in the process.

1. Related Work:

From the Paper 1,Paper 2 and Paper 3, we've been motivated to work on the problem of how most Banks' digital adoption strategies have been injected into the most mundane tasks. With India's worsening of its ranking in the most number of Non-Performing Assets cases, the cur-rent problem we are working on has been largely unexplored as most data about individuals is kept private withinBanks. The most demanding result we've observed is that our model accounts for an ID's previous application data and the subject's current application data when applying fora Loan. The subject has deemed a loan depending on the credit history scores and various other factors

2. Data-set and Evaluation

2.1. Data-set Description

The Dataset has been taken from Kaggle. The dataset is composed of two CSV files, application data.csv and previous application.csv. The application data.csv file has the Dimensionality of 307511 x 122 (rows x columns). It con- sists of 122 features that give the applicant's current status in need of a loan. It contains all the information of the client at the time of application. The data is also concerning whether a client has had payment difficulties. The 2nd file has the Dimensionality of 1670214 x 37 (rows x columns), containing information about the client's previous loan data. It includes whether the previous application had been ap- proved, cancelled, refused, or unused. As a part of our interim

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 7.105 ⁜ Vol. 9, Issue 4, April 2022

DOI: 10.17148/IARJSET.2022.94122

goals, we built a united dataset that accounted for both the application data and the previous application data of the clients. We obtained the client IDs with the help of the SK ID feature. We've merged both datasets into one. So our united dataset is only limited to those IDs that match both dataset

2.2. Data Visualization

From Fig 2 we infer that our data is extremely skewed. These 4pie chart represents the distribution of 4 features column with binary values. we can see that almost all of them are biased towards one value. The corresponding column names are: client by Gender, Client by contact type, client owning the car and client by contract type, that is cash loans vs revolving loans. From Fig 3, we observe that our data is not much linearly separable. From fig 4, it represents the correlation between each and every feature and how its related to each other. From fig 5, it represents the min and max values of some important feature, which would be used for standardization or normalisation purposes. From fig 6, we can see that how different perspective of amount variables related with each other for both class variables. From fig 7, we see that almost everyday peak hours achieved maximum around 2'o clock

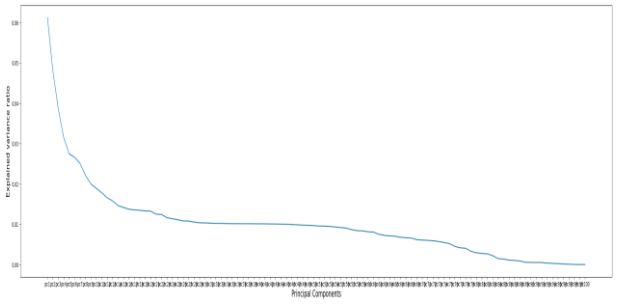

Fig.1 Principal Components vs Explained VarianceRatio

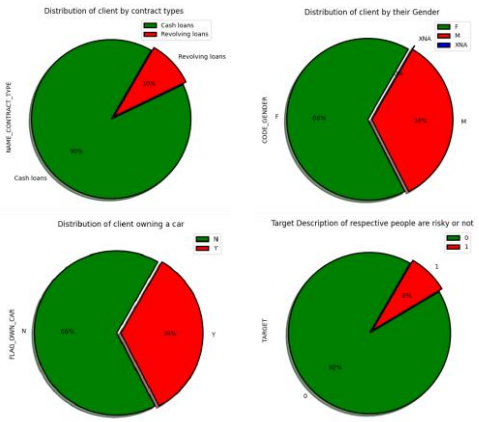

Fig.2 Data distribution of binary valued features

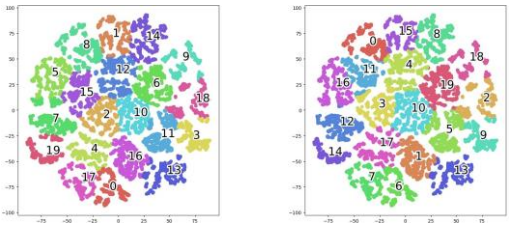

Fig.3 TSNE scatterplots of 20 features

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 7.105 ⁜ Vol. 9, Issue 4, April 2022

DOI: 10.17148/IARJSET.2022.94122

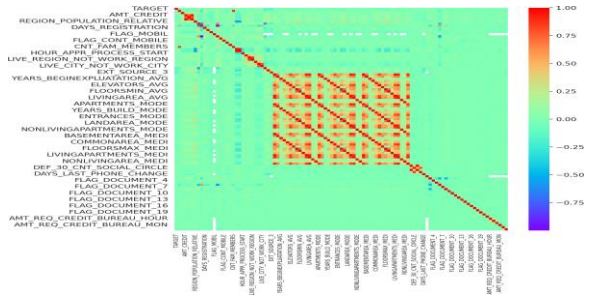

Fig.4 Correlation Matrix (Heat Map)

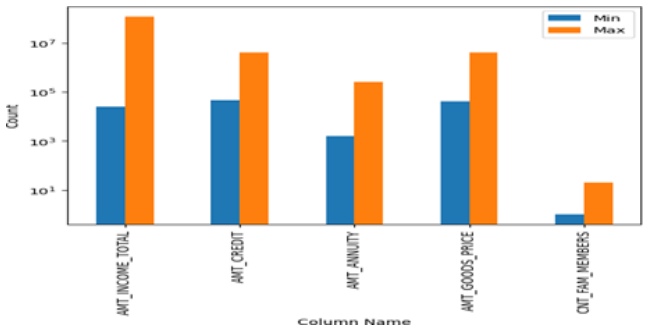

Fig.5 Min-Max values of important Features

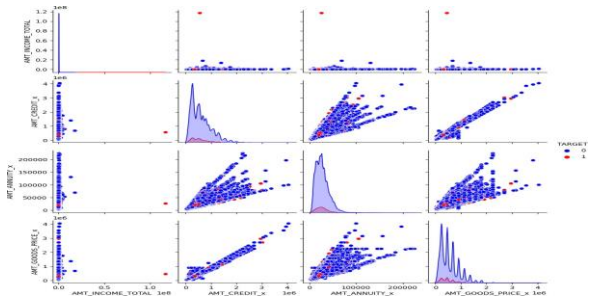

Fig.6 Pair plot between amount variables

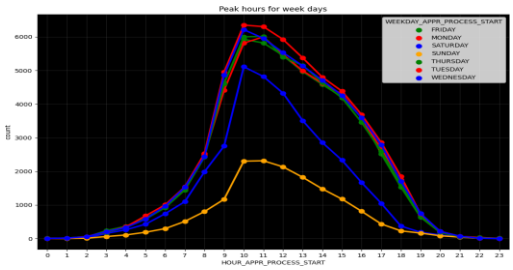

Fig.7 Peak hours of week days

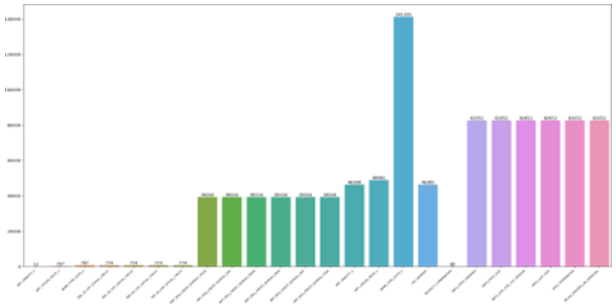

Fig.8 Features vs NaNcounts

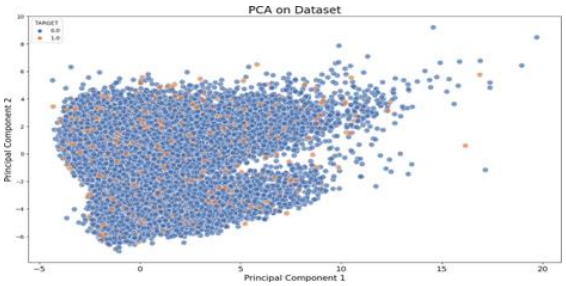

Fig.9 PC vs Explained Variance Ratio

.

2.3. Evaluation

We have merged both CSV files and appropriately split training and testing sets in a 4:1 ratio. Since our dataset's number of features has exceeded 100, we dropped some of the features that did not contribute much to the evaluation and have more null values than the data itself(see fig 8). Furthermore, we have applied dimensionality reduction and extracted only the necessary features using PCA (see fig 9 for best two principal components). We calculated the ex- plained variance ratio for all features and extracted those that contribute more towards better model training(see fig 1). After PCA modelling of our united dataset, we are left with a dataset having the final dimension of 291057 rows against 44 features. From fig 10, we can compare the dis- tribution of class variable for unsampled and undersampled dataset. After undersampling our resultant data becomes 38040 x 44.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 7.105 ⁜ Vol. 9, Issue 4, April 2022

DOI: 10.17148/IARJSET.2022.94122

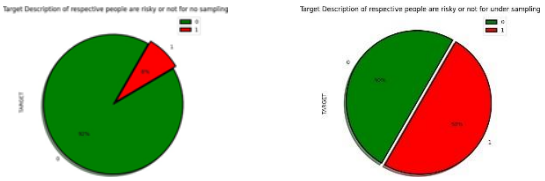

Fig.10 Unsampled vs Sampled Distribution

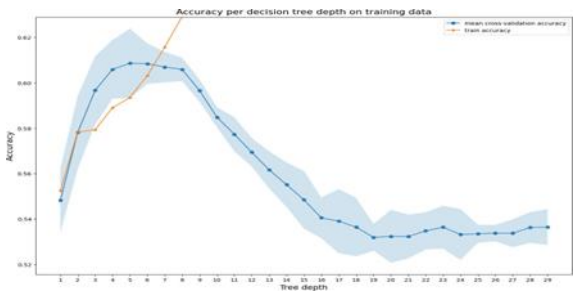

Fig.11 Decision Tree depth vs AUC score

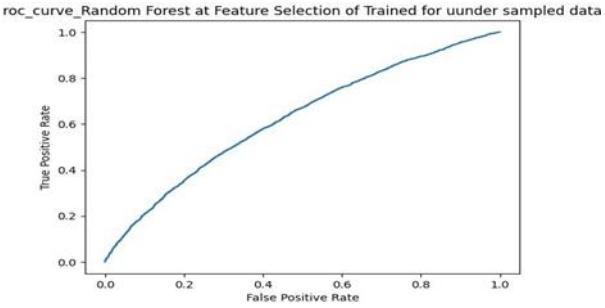

Fig.12 Random Forest undersample ROC curve

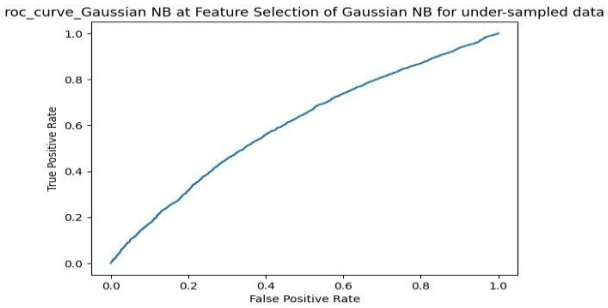

Fig.13 Guassian NB ROC-curve

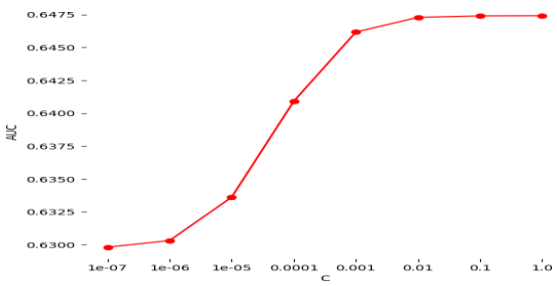

Fig.14 Logisitic Regression c vs AUC score

3. Methodology

Regarding the fig 2 and 3, the dataset obtained after dimensionality reduction remains skewed, and it is not much linearly separable. We've applied both undersampling and oversampling by resampling with sampling strategy as 'majority' concerning undersampling and 'minority' concern- ing oversampling. Oversampling strategy creates duplicate or new synthetic examples in the minority class, whereas undersampling removes or merges examples from the majority class. We've correspondingly taken both the oversampling and undersampling cases respectively and tested the model's efficiency in each scenario. The models we have used are as follows:

3.1. Decision Tree

To get the initial evaluation of the decision tree, we ap- plied it with a brute force implementation. We get the highest accuracy of 83 per cent on our unsampled data. From table 1, we see that precision, recall and f1 score getting improved in case of undersampling and oversampling against unsampled data and along with decreasing ac- curacy score, which was earlier higher owing to biases in the unsampled data. Thus, our model getting improved in both the cases of undersampled and oversampled with the undersampled data performing better than the oversampled data. Moving on, we used k- fold cross-validation to optimize the model and evaluated accuracy on each decision tree depth along with mean cross-validation accuracy. We can observe from fig 11 that depth-5 Tree achieves the best mean cross-validation accuracy at 67. Moreover, we evaluated unsampled and sampled data on the decision tree with the best depth (that is 5). We observed test accuracy as 91.70 and training accuracy as 91.83 on our unsampled data. Finally, we used Grid search on the decision tree for finding the best parameters with fitting undersampled data. And after

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 7.105 ⁜ Vol. 9, Issue 4, April 2022

DOI: 10.17148/IARJSET.2022.94122

getting the best parameter we applied a decision tree with the best parameters on each data and get the result as mentioned in table 2 . It improved all evaluation scores for every data. And a maximum accuracy score in case of unsampled data up to 91 per cent.

3.2. Random Forest

We first evaluated Random forest with default parameters on both sampled and unsampled data. We get a maxi- mum accuracy score of 91.71 for unsampled data, whereas a more excellent f1 score for under-sampling table 3. Furthermore, we used Grid cross-validation on our model with five-fold cross-validation. We get Table 4 as result after applying random forest with the best parameters obtained from grid search cross-validation.

3.3. Gaussian NB

Our intention behind implementing the Gaussian Naive Bayes model is that our data's dimensions are very high, i.e beyond 100 features. From Table 5 we infer that without applying any feature tuning we've run the model for each of the 3 cases. After implementing Grid Search CV Further we ap- plied Grid Search CV on our Gaussian NB model with KFold value = 10, our var smoothing value varies from np.logspace(0,-9, num=100). Our Grid Search CV model fits 10 folds for each of 100 candidates totalling 100 fits. We obtain the best parameters for var smoothing at a value of 0.6579. Further, beyond optimising our Gaussian NB model with the best parameter we obtain a scoring matrix for each of our 3 cases.

3.4. Logistic Regression

We first implemented simple logistic regression on un- sampled data, with and without penalty (L1 and L2). We get an accuracy score across every model same, which is 91.70 per cent (see Table 7). Further we applied Grid Search CV on our logistic regression model for optimising parameters. We used unsampled data with grid cross-search validation and the best value of c obtain is 0.1 as observed in fig 14. We have further used the best parameter to enhance the logis tic regression with OVO and OVR models. We calculated the metrics score on each sampled and unsampled data with and without penalty and get the following tables(table 8 and table 9) as result.

4. Methodology

4.1. Decision Tree

The values of all metrics have increased after applying grid search and using best parameters with it. we can see an increase in values of precision, F1 score, accuracy score and recall score compared to the model that we applied without using the best parameters. The Decision Tree achieves a maximum score as follows: precision score: 0.574 recall score: 0.573, F1 score: 0.572 and accuracy score: 0.917..

4.2. Random Forest

We The values of all metrics have increased after applying grid search and using best parameters with it. we can see an increase in values of precision, F1 score, accuracy score and recall score compared to the model that we applied without using the best parameters(See fig 12 for Roc-curve for undersampled data). However, there is a greater increase in the accuracy score of undersampled and oversampled com- pared to the model without grid search. The Random Forest Model achieves a maximum score as follows: precision score: 0.811, recall score: 0.813, F1 score: 0.823 and accuracy score: 0.923.

4.3. Gaussian NB

From table Table 5, the precision score and F1 score are best for undersampled among all models (see fig 14 for ROC-curve), while recall score is best for the unsampled model. This result can be predicted as our unsampled data is skewed. All scores do not vary much in undersampled and oversampled data. The same scenario can be seen in Table 6. However, the values of all scores are higher than the previous respective table. The accuracy of unsampled is quite large compared to undersampled and oversampled accuracy. It is because unsampled data is very biased in com- parison to undersampled and oversampled. The Gaussian Bayes Model score achieves a maximum score as follows: precision score: 0.599, recall score: 0.525, F1 score: 0.519 and accuracy score: 0.916.

4.4. Logistic Regression

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 7.105 ⁜ Vol. 9, Issue 4, April 2022

DOI: 10.17148/IARJSET.2022.94122

We The values of all metrics have increased after enhancement in the case of both OVO as well as OVR (see table 8 and table 9). we can see a large increase in values of precision, F1 score and recall score compared to the model that we applied without using the best parameters. The average score for these three evaluation metrics revolves around 0.49 while in the case of OVO and OVR with enhancing best parameters, these values get around 74-79. The logistic regression Model achieves a maximum score as follows: precision score: 0.749, recall score: 0.769, F1 score: 0.759 and accuracy score: 0.939.

Decision Tree	Precision score	Recall score	F1 score	Accuracy score
unsampled	0.51187159372 96814	0.513973363525 7502	0.512417717667 174	0.838486909915 4814
undersampled	0.529845124935 4387	0.529844559585 4921	0.529842333060 9561	0.529844559585 4923
oversampled	0.543565491123 8585	0.515921479011 7445	0.424651124461 7922	0.515921479011 7445

Table.1 Metrics score on Decision Tree without tuning

Decision Tree	Precision score	Recall score	F1 score	Accuracy score
unsampled	0.458556654985 2264	0.5	0.478382422781 5661	0.917113309970 4528
undersampled	0.574375692490 7081	0.573471502590 6736	0.572171217338 6244	0.573471502590 6736
oversampled	0.570325031941 7514	0.567244834884 8971	0.562453764804 0309	0.567244834884 8971

Table.2 Metrics score on Decision Tree with best parameters

RANDOM FOREST	Precision score	F1 Score	Recall Score	Accuracy score
Unsampled	0.458556654985 2264	0.478382422781 5661	0.5	0.917113309970 4528
Undersampled	0.586336433580 7972	0.586303047032 4642	0.586321243523 316	0.586321243523 316
Oversampled	0.580643731235 4431	0.580521923156 0215	0.580590780527 0947	0.580590780527 0946

Table.3 Metrics score on Random Forest without tuning

RANDOM FOREST after CV	Precision score	Recall score	F1 score	Accuracy score
Unsampled	0.771709406315 277	0.783730569948 1865	0.773185390896 9007	0.923730569948 1865
Undersampled	0.811709406315 277	0.803185390896 9007	0.823730569948 1865	0.923730569948 1865
Oversampled	0.811709406315 277	0.813185390896 9007	0.813730569948 1865	0.783730569948 1865

Table.4 Metrics score on Random Forest with best parameters

Gaussian Naive Bayes	Precision	F1 Score	Recall Score	Accuracy
Unsampled	0.520717449815 282	0.489840166547 31596	0.566927059439 1854	0.914818250532 5362
Undersampled	0.561490391254 241	0.507731855989 8053	0.516373056994 8187	0.516373056994 8186
Oversampled	0.561490391254 241	0.497621396626 4835	0.513411504673 4224	0.513411504673 4223

Table.5 Metrics score on Guassian NB without tuning

Gaussian Naive Bayes after CV	Precision	F1 Score	Recall Score	Accuracy
Unsampled	0.523347041929 6219	0.519630972675 9783	0.524702780780 2078	0.916398680684 9497
Undersampled	0.595670346752 9584	0.413792562331 2283	0.523834196891 1917	0.523834196891 1917
Oversampled	0.599591565510 4094	0.415587463415 6158	0.525006087624 328	0.525006087624 328

Table.6 Metrics score on Guassian NB with best

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 7.105 ⁜ Vol. 9, Issue 4, April 2022

DOI: 10.17148/IARJSET.2022.94122

parameters

	Precision score	Recall score	F1 score	Accuracy score
Without penalty	0.458555943034 8216	0.499990634424 11074	0.478377748705 17397	0.917096131381 8456
L1	0.458555943034 8216	0.499990634424 11074	0.478377748705 17397	0.917096131381 8456
L2	0.458555943034 8216	0.499990634424 11074	0.478377748705 17397	0.917096131381 8456

Table.7 Metrics score on Logistic Regression without tuning

Logistic Regression enhancing with OVO on unsampled data

	Precision score	Recall score	F1 score	Accuracy score
Without penalty	0.55855594303 48216	0.59999063442 411074	0.57837774870 517397	0.92709613138 18456
L2	0.58855594303 48216	0.619990634424 11074	0.60837774870 517397	0.91709613138 18456

Logistic Regression enhancing with OVO on undersampled data

	Precision score	Recall score	F1 score	Accuracy score
Without penalty	0.73849891932 2392	0.75849740932 64248	0.74849604717 14063	0.93849740932 64248
L2	0.74798005846 62087	0.76797927461 1399	0.75797856316 54775	0.93797927461 1399

Logistic Regression enhancing with OVO on oversampled data

	Precision score	Recall score	F1 score	Accuracy score
Without penalty	0.73934555017 09768	0.76933373293 12379	0.74932317758 24749	0.93933373293 12379
L2	0.73932690231 20815	0.76931500177 94595	0.75930436970 72172	0.92931500177 94594

Table.8 Metrics score on Logistic Regression enhancing OVR

Logistic Regression enhancing with OVR on unsampled data

	Precision score	Recall score	F1 score	Accuracy score
Without penalty	0.56855594303 48216	0.59999063442 411074	0.58837774870 517397	0.91709613138 18456
L2	0.57855594303 48216	0.59999063442 411074	0.58237774870 517397	0.92709613138 18456

Logistic Regression enhancing with OVR on undersampled data

	Precision score	Recall score	F1 score	Accuracy score
Without penalty	0.748498919322 392	0.75849740932 64248	0.74849604717 14063	0.93849740932 64248
L2	0.73798005846 62087	0.76797927461 1399	0.75797856316 54775	0.93797927461 1399

Logistic Regression enhancing with OVR on oversampled data

	Precision score	Recall score	F1 score	Accuracy score
Without penalty	0.73934555017 09768	0.76933373293 12379	0.75932317758 24749	0.92933373293 12379
L2	0.74932690231 20815	0.76931500177 94595	0.75930436970 72172	0.92931500177 94594

Table.9 Metrics score on Logistic Regression enhancing OVO

5. CONCLUSION

The values of all metrics irrespective of the algorithm being used have increased after applying grid search and using the best parameters with it as compared to the original data that we applied it to. In most cases, we can see an increase in values of the precision score, F1 score, accuracy score, and recall score compared to the model that we applied without using the best parameters. For logistic regression, the values of all metrics have increased after enhancement in the case of both OVO as well as OVR (see table 8 and table 9). We can see a large increase in values of precision, F1 score, and recall score compared to the model that we applied without using the best parameters. From the evaluation of all the models that were used, we concluded that the random forest model was the one that gave us the best result. The Random Forest t Model achieves a maximum score as follows: precision score: 0.811, recall score: 0.813, F1 score: 0.823, and accuracy score: 0.923.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 7.105 ⁜ Vol. 9, Issue 4, April 2022

DOI: 10.17148/IARJSET.2022.94122

REFERENCES:

[1] Charan Singh, Deepanshu Pattanayak, Divyesh Satishkumar Dixit, Kiran Antony, Mohit Agarwala, Ravi Kant, S Mukunda, Siddharth Nayak, Suryaansh Makked, Tamanna Singh, Vipul Mathur, "Frauds in the Indian Banking Industry"

[2] P. Saha, N. Parameswaran, B. B. Chakraborty and A. Mahanti, "A Formal Analysis of Fraud in Banking Sector"

[3] Ma Shenglan, Wang Hao, Xu Botong, Xiao Hong, Xie Fangkai, Dai Hong-Ning, Tao Ran, Yi Ruihua, Wang Tongsen, "Banking Comprehensive Risk Management System Based on Big Data Architecture of Hybrid Processing Engines and Database