# **Credit card default Predicting using a classification model**

# **(pradyumna vikram singh)**

**Data science trainees,**

**AlmaBetter, Bangalore**

**Abstract:**

### This project is aimed at predicting the case of customers’ default payments in Taiwan. From the perspective of risk management, the result of predictive accuracy of the estimated probability of default will be more valuable than the binary result of classification - credible or not credible clients**.** Credit cards are now the most preferred way for customers to transact either offline or online. All our digital transactions through credit card statements are far more easily compared with cash transactions or bank statements. One downside that has been witnessed over the past few years of this increasing digital phenomenon is the rise of fraud on credit cards. Global fraud has increased by almost three times, from $9.84 billion to $32.39 billion in less than a decade (2011 to 2020).

***Keywords: Exploratory Data Analysis, Train -Test split Classification, Machine learning model, (LR, DT, RF, KNN, SVM, XGB)***

1. **Problem Statement**

After understanding the gravity of the fraud situation worldwide, particularly in the United States and some of the major European countries, the next automatic question that comes to mind is how we prevent this fraud and the damage it causes to the overall economy and especially to the customer sentiments and trust in financial institutions. let us discuss some of the challenges that we face while dealing with credit card fraud as below.

**Data imbalance:** The fraud and non-fraud data are generally much skewed. To give an example in the sample open-source dataset that we will be dealing with here, we have **6636** frauds out of a total of 30000 transactions. This is roughly only **21%** of all the transactions. So, it is easy to achieve almost **80 %** accuracy with a naive model which just predicts all the transactions as non-fraud.

**Customer friction:**The most likely outcome if a model predicts a current transaction as fraud is to decline the transaction outright to prevent any financial loss. However, we will soon see that it sometimes proves to be a bone of contention with genuine customers, who might get declined if the model has too many false positives or Type 1 errors.

**Real-Time Detection:**For most of the fraud detection models in practice they have to work under very stringent timing conditions. We can take an example of a transaction-level fraud detection model. Irritate the customer who is waiting to do the transaction, and if we process too fast, we may improve on customer experience, but it might lose out on accuracy.

.

**2. Introduction**

## Recently, the state vigorously promotes the economic construction of large- and medium-sized cities, which not only improves people’s living standards but also changes people’s consumption concept and consumption mode. People are more and more inclined to spend ahead of time and mortgage their “credit” to the bank to enjoy certain things in advance. However, when consuming, people often lack rational thinking and overestimate their ability to repay loans to banks in time. On the one hand, it increases the loan risk of banks; on the other hand, it increases the credit crisis of consumers themselves. With a large number of banks selling credit cards, the phenomenon of credit card default emerges one after another.

## **3. Related Work**

In the prediction of “two classifications”, a few categories are called positive examples (default), and most categories are called counterexamples (no default). However, most of the credit card loan data are unbalanced. Let us now discuss the dataset that we will be working with to build models and decide their effectiveness in fraud detection.  In this model SMOTE algorithm is used to change the data distribution, and then the importance of data features is calculated by using different models All subsections are structured as follows; 3.1 provides information on the data transformation techniques utilized in related works, 3.2 illustrates the details of widely used machine learning models for bike-share prediction.

## **3.1 Data information& Data Transformation**

We had to perform a few imputations and transformations on our dataset for us to create the desired visualizations. There were no major inconsistencies or mismatches in the data. We rename some columns and extract useful information from the date column.

The sample size of this data is 30,000, of which 6,636 are in the positive category (default) and 23,364 are in the negative category (no default). The sample has a total of 25 variables. In this experiment, considering that the variable ID has no relationship with the target variable, the deletion process was performed. 23 characteristic variables and 1 target variable were selected. Our data set have the value: -

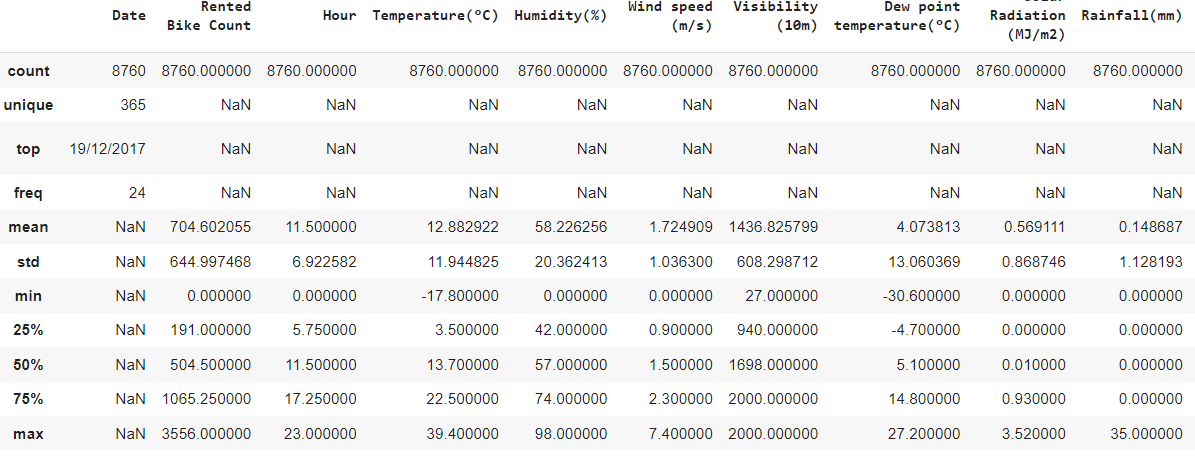
**Date**

# **3.2 Machine learning Models:**

A bike-share system data majorly constitutes time-dependent features. These features fluctuate randomly making it impossible to build a predictive model using static stochastic time series techniques. We start fitting our feature or data from Linear Regression Model and then step-wise move forward to Lasso and regression to more improvement of the linear model. we also try to fit data on the decision tree and visualize the tree. Random Forest also gives a better result then move forward for the Gradient boosting and we find that model performance get increases but score still below 80% so we used next Model that is XGBoost and fit the data to this model and achieve the performance of more than 82% on the training data

**4. Dealing with Outliners:**

We see no outlier in the data set so no worry about dealing with an outlier. we just make our focus on data extraction and correlation

****

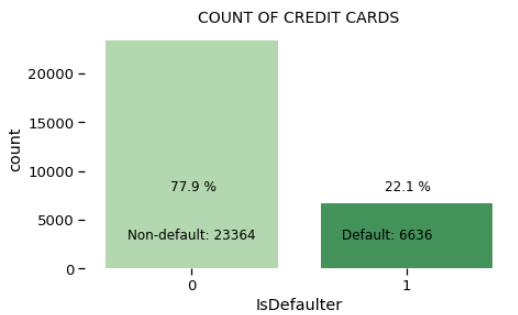
**5. Methodology:**

The existing methodologies for predictions are Logistic regression, decision trees, random forests, KNN, SVM, XGBoost, etc. This research work allows having insight into the performance of various prediction algorithms and walks through the whole process of prediction.

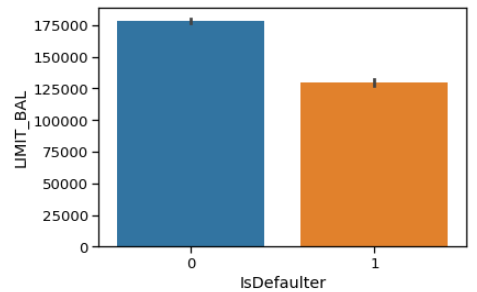
* **Data pre-processing and transformation**
* **Developing and optimizing Logistic Regression model**
* **Developing and optimizing KNN model**
* **Developing and optimizing DECISION TREE**
* **Developing and optimizing RANDOM FOREST**
* **Developing and optimizing SVM (support vector machine)**
* **Developing and optimizing Xtream Gradient Boosting**

**5.1Data pre-processing and transformation**

We can see that the problem of category imbalance is mainly solved from the following two perspectives: the first perspective is to balance the data by changing the number of samples. This method can also be divided into three aspects. On the one hand, it is to improve the oversampling method. On the other hand, it is based on the principle of under-sampling to change the data distribution. On the third hand, it is the method of combining oversampling and undersampling. The second perspective is to improve the classifier algorithm to improve the prediction performance of the model and at the same time use relevant evaluation indicators to evaluate the prediction results

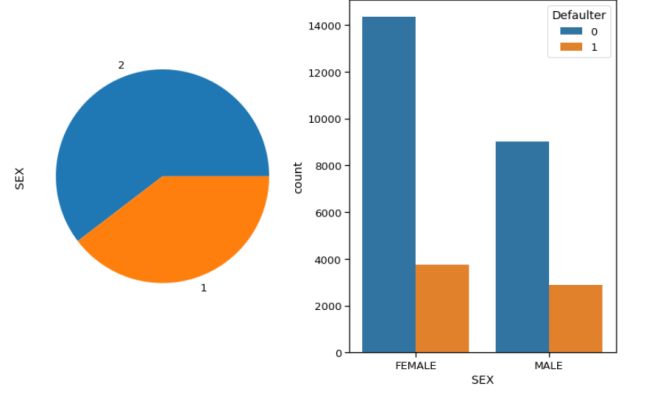


This is a basic graph shows that imbalance in data.



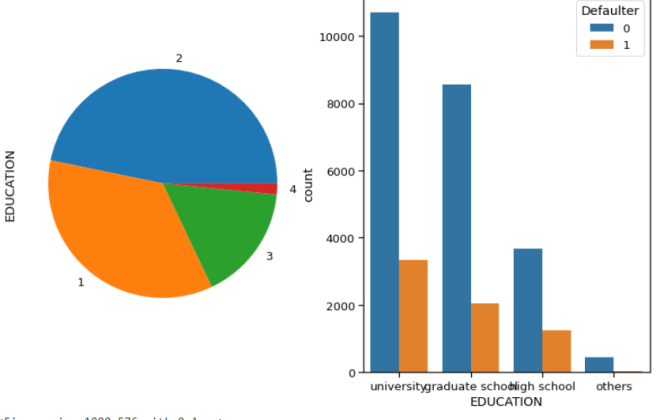
After different categorical feature SMOTE algorithms and feature engineering, we transform our data for reading purposes for a better understanding of the data set.

**Which gender is most defaulter?**

****

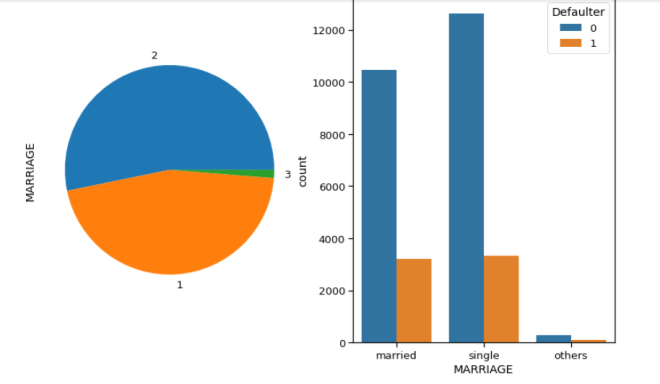
In This graph, we can observe that the female gender is the most no of time defaulter.

**Which candidate from the different educational backgrounds is the most defaulter?**

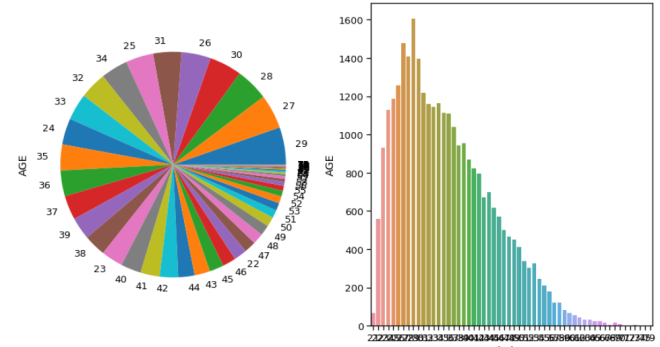
****

For the given plot we can see that University graduates are the most no of time facing defaulter and then school.

**Defaulter Family background**

from the given plot we can see that unmarried or single person have more defaulter chance than the married

**Defaulter AGE limits background**

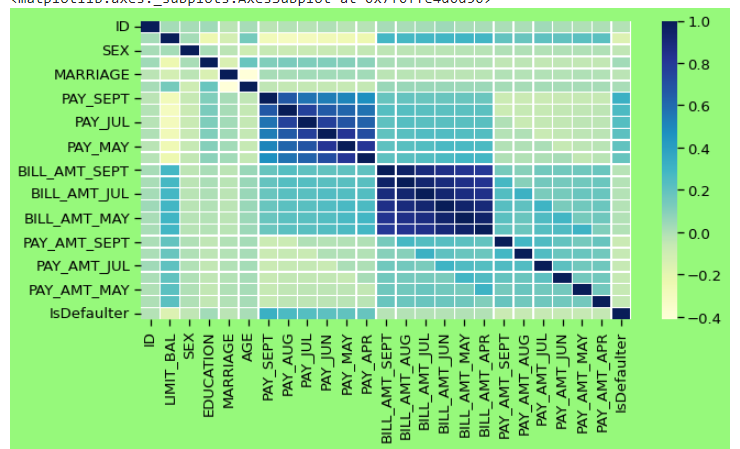
****

This plot does not give lot much clarity about the age limit but after careful observation of the plot, we can say that 29 age limit people get defaulted more than the other. we know that these age limits have lots of ups and downs as well as financial responsibility

.

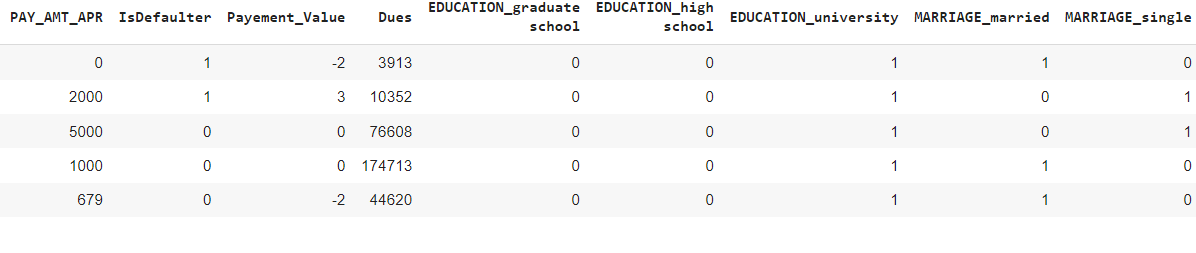
**Correlation between weather parameters:**

We can see that no correlation between the features so there is no need to remove or drop some features.

****

**Data set after the feature Engineering and dummy variable:**

It is a process in which analysts use domain knowledge about the data and create new features in the data set in a way such that the new features help in improving the model accuracy. There is no definite path for feature engineering, but it depends on the skills of the analyst and the type of data. Feature engineering needs to be done on both training and testing data and is a very important part of building a good prediction model. We used One Hot Encoding to produce binary integers of 0 and 1 to encode our categorical features because categorical features that are in string format cannot be understood by the machine and needs to be converted to the numerical format. You can see the sample part of the data set here after feature engineering



**5.2Developing and optimizing Logistic Regression model:**

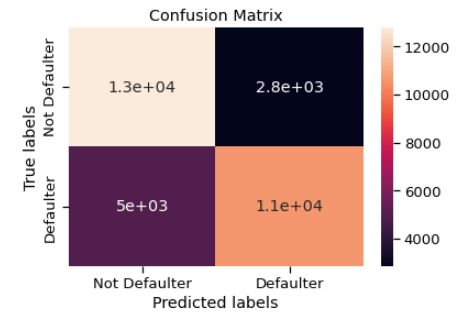
The logistic regression model gives follows the sigmoid curve which has limits 0 to 1. this classification technique is used a threshold value for categorizing the result into 0 or 1 .in this project after applying the logistic regression we get some of the useful results as given below the **accuracy on train data is 0.75 and accuracy on test data is 0.75.**

**The precision on test data is 0.68**

**and recall on test data is 0.78**

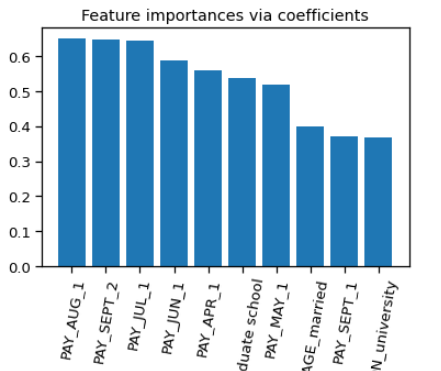
**also, f1 on test data is 0.73**

**and roc\_score on test data is 0.75**

****

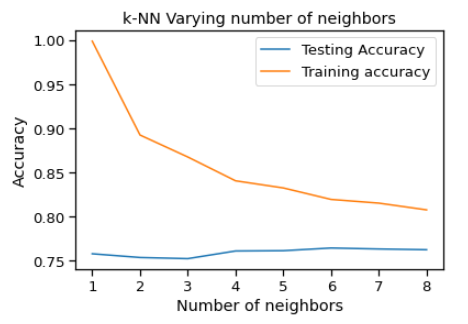
A confusion matrix is very informative in classification problem decision-making.

**Features importance according to logistic regression:**

****

as per the logistic regression model payment month is an important feature with respect to different month defaulter rates and rest of least.

**5.3 Developing and optimizing KNN**

****

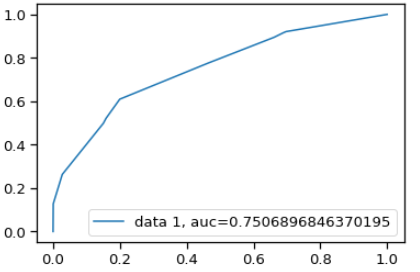
From the given plot we can take the best possible no of neighbors for our dataset. we can see here 7 is the best k value for further analysis. Roc score is 0.83% for KNN after grid search cv.

**5.4 Developing and optimizing Decision Tree model**

Decision Tree is another very popular algorithm for classification problems because it is easy to interpret and understand. An internal node represents a feature, the branch represents a decision rule, and each leaf node represents the outcome. Some advantages of decision trees are that they require fewer data preprocessing, i.e., no need to normalize features. However, noisy data can be easily overfitted and results in biased results when the data set is imbalanced.

**The accuracy on train data is 0.80**

**The accuracy on test data is 0.78 and AUC –ROC score is 0.75%**

****

**5.5 Developing and optimizing the SVM model**

the running time of the SVM model is too long, close to 6 minutes; compared to other models, the running efficiency of SVM is very low. If the amount of data is very large, it is not a wise choice for us to use SVM for prediction.

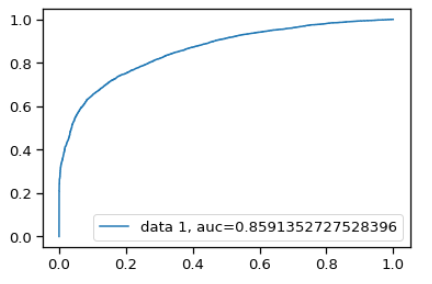
**The accuracy of test data is 0.78**

**The precision on test data is 0.71**

**The recall on test data is 0.81**

**The f1 on test data is 0.76**

**The auc\_score on test data after GridSearchCv is 0.85**

****

**5.6 Developing and optimizing Random Forrest Tree**

Random Forest is a supervised learning algorithm; it creates a forest and makes it somehow random. The "forest “it builds, is an ensemble of Decision Trees. the score of the model are given below:

**The accuracy of test data is 0.83**

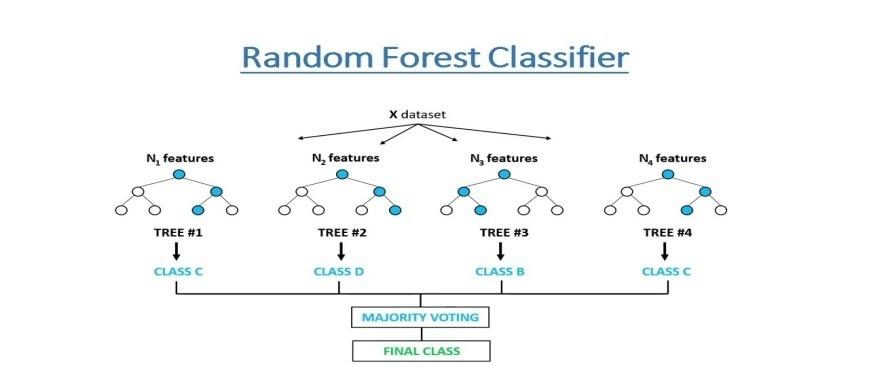
**The precision on test data is 0.80**

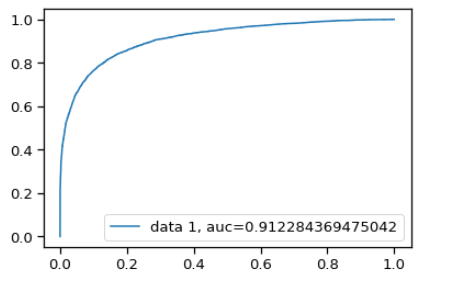
**The recall on test data is 0.85**

**The f1 on test data is 0.82**

**The auc\_score on test after Grid search CV on test data is 0.91**

.

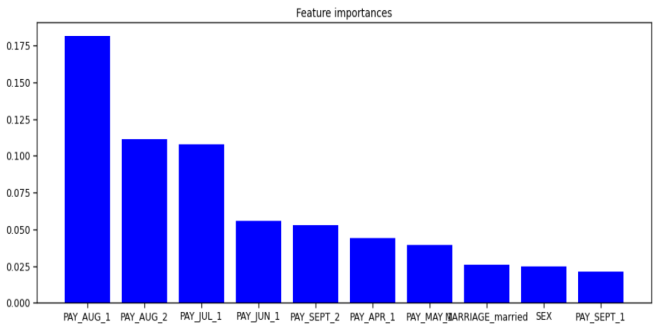


****

We can achieve the best of auc area under of curve.

**5.6 Developing and optimizing Xtream Gradient Boosting (XGB)**

XGBoost is a gradient boosting algorithm. Which is faster and usually converts the weak learner into strong. We use mostly the default parameters and fit the model. We then plot the feature importance as per the model. and we found that payment in aug is one of the most important features of all available.Xgboost is a sequential model which make a weak learner into a strong one



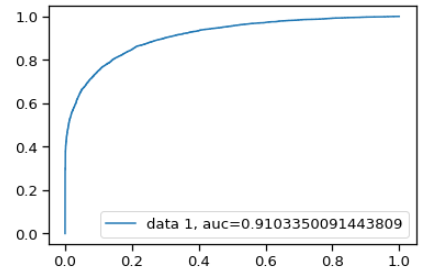
**The accuracy of test data is 0.82**

**The precision on test data is 0.78**

**The recall on test data is 0.85**

**The f1 on test data is 0.82**

**The auc\_score on test data after GridSearchCV is 0.91**

****

**6. Conclusion**

We investigated the data, checking for data unbalancing, visualizing the features, and understanding the relationship between different features. We then investigated two predictive models. The data was split into three parts, a train set, a validation set, and a test set. For the first three models, we only used the train and test set. We started with Logistic Classifier and then with RandomForrestClassifier, for which we obtained an AUC code of 0.81 and 0.91, respectively, when predicting the target for the test set. We followed with a knn model, with a lower AUC score (0.91) for the prediction of the test set target value. We then followed with an SVM, with the AUC score after training iterations 0.85.

We then experimented with an XGBoost model. In this case, we used the validation set for validation of the training model. The best validation score obtained was 0.91. Then we used the model with the best training step to predict the target value from the test data the AUC score obtained was 0.9253.

**7. REFERENCES**

[1] Z. Feng and M. Feng, “Research on credit card scoring model based on AHP,” *Finance Theory and Practice*, vol. 1, pp. 74–77, 2016.View at: [Google Scholar](https://scholar.google.com/scholar_lookup?title=Research%20on%20credit%20card%20scoring%20model%20based%20on%20AHP&author=Z.%20Feng%20&author=M.%20Feng&publication_year=2016)

[2] R. Mei, Y. Xu, and G. Wang, “Study on analysis and influence factors of credit card default prediction model,” *Statistics and Applications*, vol. 5, no. 3, pp. 263–275, 2016.Viewat: [Publishe site](https://doi.org/10.12677/SA.2016.53026) | [Google Scholar](https://scholar.google.com/scholar_lookup?title=Study%20on%20analysis%20and%20influence%20factors%20of%20credit%20card%20default%20prediction%20model&author=R.%20Mei&author=Y.%20Xu&author=&author=G.%20Wang&publication_year=2016)

[3]M. Zan, G. Yanrong, and F. Guanlong, “Credit card fraud classification based on GAN-AdaBoost-DT imbalance classification algorithm,” *Journal of Computer Applications*, vol. 39, no. 2, pp. 314–318, 2019.View at: [Google Scholar](https://scholar.google.com/scholar_lookup?title=Credit%20card%20fraud%20classification%20based%20on%20GAN-AdaBoost-DT%20imbalance%20classification%20algorithm&author=M.%20Zan&author=G.%20Yanrong&author=&author=F.%20Guanlong&publication_year=2019)

[4]L. Hu, Z. Peng, W. Xiang, and X. Rongze, “A new combination sampling method for imbalanced data,” in *Proceedings of the 2013 Chinese Intelligent Automation Conference: Intelligent Information Processing*, vol. 256, pp. 547–554, Yangzhou, China, 2013.View at: [Google Scholar](https://scholar.google.com/scholar_lookup?title=A%20new%20combination%20sampling%20method%20for%20imbalanced%20data&author=L.%20Hu&author=Z.%20Peng&author=W.%20Xiang&author=&author=X.%20Rongze)

[5]H. Han, W.-Y. Wang, and B.-H. Mao, “Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning,” *Advances in Intelligent Computing*, vol. 3644, no. 1, pp. 878–887, 2005.View at: [Google Scholar](https://scholar.google.com/scholar_lookup?title=Borderline-SMOTE:%20a%20new%20over-sampling%20method%20in%20imbalanced%20data%20sets%20learning&author=H.%20Han&author=W.-Y.%20Wang&author=&author=B.-H.%20Mao&publication_year=2005)

[6]S. Wang, W. Liu, and J. Wu, “Training deep neural networks on imbalanced data sets,” in *Proceedings of the International Joint Conference on Neural Networks*, pp. 4368–4374, IEEE, Vancouver, Canada, July 2016.View at: [Google Scholar](https://scholar.google.com/scholar_lookup?title=Training%20deep%20neural%20networks%20on%20imbalanced%20data%20sets&author=S.%20Wang&author=W.%20Liu&author=&author=J.%20Wu)

[7]J. Jiao, X. Zhang, F. Li, and Z. Niu, “Identically distributed multi-decision tree based on reinforcement learning and its application in imbalanced data sets,” *Journal of Central South University (Science and Technology)*, vol. 50, no. 5, pp. 1112–1118, 2019.View at: [Google Scholar](https://scholar.google.com/scholar_lookup?title=Identically%20distributed%20multi-decision%20tree%20based%20on%20reinforcement%20learning%20and%20its%20application%20in%20imbalanced%20data%20sets&author=J.%20Jiao&author=X.%20Zhang&author=F.%20Li&author=&author=Z.%20Niu&publication_year=2019)

[8]D. Hong, L. Balzano, and J. A. Fessler, “Asymptotic performance of PCA for high-dimensional heteroscedastic data,” *Journal of Multivariate Analysis*, vol. 167, pp. 435–452, 2018.View at: [Google Scholar](https://scholar.google.com/scholar_lookup?title=Asymptotic%20performance%20of%20PCA%20for%20high-dimensional%20heteroscedastic%20data&author=D.%20Hong&author=L.%20Balzano&author=&author=J.%20A.%20Fessler&publication_year=2018)

[9]K. Mens, E. Elzinga, and M. Nielen, “Applying machine learning on health record data from general practitioners to predict suicidality,” *Internet Interventions*, vol. 21, Article ID 100337, 2020.View at: [Google Scholar](https://scholar.google.com/scholar_lookup?title=Applying%20machine%20learning%20on%20health%20record%20data%20from%20general%20practitioners%20to%20predict%20suicidality&author=K.%20Mens&author=E.%20Elzinga&author=&author=M.%20Nielen&publication_year=2020)