Abnormal Distribution Handling with k-Nearest Neighbour (KNN) for Financial Fraud Detection

Dr. Sikander Singh Cheema*, Harsimrat Deo**
*Assistant Professor, Department Of Computer Science & Engineering, Punjabi University, Patiala,
Punjab, India.
**Assistant Professor, Computer Science Department, Mata Gujri College, Fategarh Sahib, Punjab, India.
*sikander@pbi.ac.in
**harsimratdeo@gmail.com

ABSTRACT—The financial frauds are rising with the increasing number of customers as well as increasing financial scheme offerings to its customers. The financial frauds are known to decrease the organizational profits by variable margins, which may be lower or extremely higher. The financial frauds can be detected by analysing spending or purchasing behaviour of financial service user, which is technically based upon the historical data. The historical data based observational analysis can be also used to determine frauds in advance by applying the predictive analysis methods. The predictive analysis methods are designed as the supervised models, which learns the calculative factors based upon the data point values in the given variables. This learning is known as training, and used to apply the calculated factors on the testing data in order to evaluate its observation, which determines the category of the target data in the form of fraud or clean customer. The proposed model is designed with different combinations of the data processing techniques and deploys the k-nearest neighbour (KNN) classification model to predict the possible financial outliers. The proposed model has outperformed the existing models based upon different error based parameters.

KEYWORDS—Financial fraud detection, Predictive analysis, Supervised Classification, Abnormal distribution handling.

INTRODUCTION

In this paper, the work has been carried upon the credit card fraud detection technique using the multiple feature processing methodology. The proposed feature processing methodology is based upon the combination of variety of techniques under the feature engineering paradigm of data processing. The data processing methods are used to validate, transform, scale or correct the variations or magnitude of the data points. Also, the data processing methods are used to correct the distribution of the data, which signifies the spread of the data points across the standard deviation, and used to evaluate the distribution density of the observed data. The following figure signifies the normal distribution, where the next figure to following shows the data with abnormal distributions.

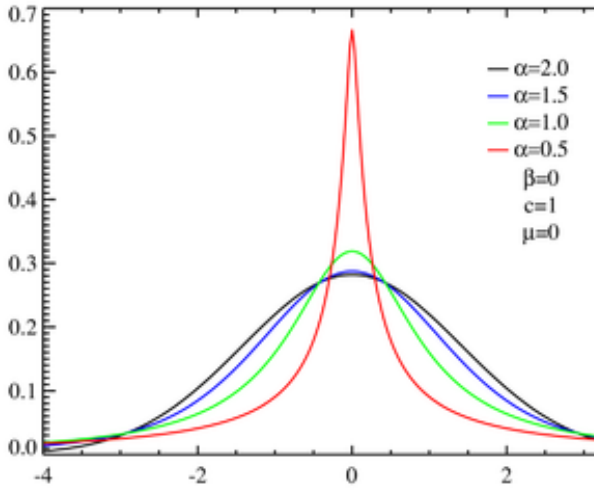


Figure 1: The chart representing the normal distribution

The above figure is showing the normal distributions of different variables or observation with different density around 0 mean values. However, the different distribution curves are showing the different magnitudes for the variables, and wit different kurtosis values.

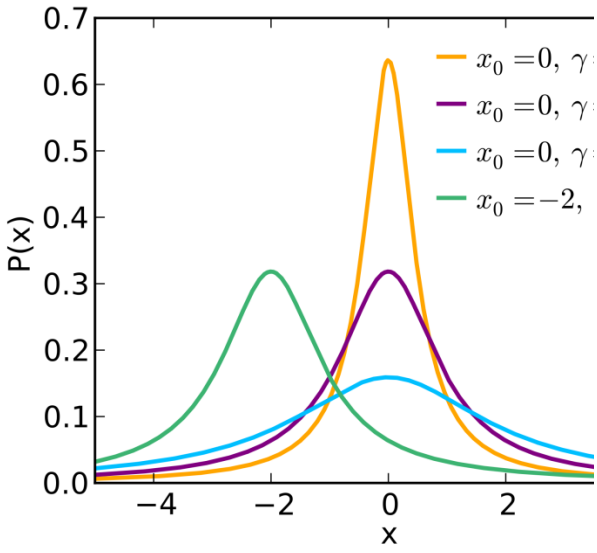


Figure 2: The chart representing the normal and abnormal distributions

The chart representing the mixture of normal and abnormal distributions, where the green curve is representing the left skewness, and slightly higher kurtosis on left side, which increases the area under the curve beyond standard deviation of -4. This means the weight of the values deviated towards the left edge is higher than usual in this case, which makes it different from all other curves.

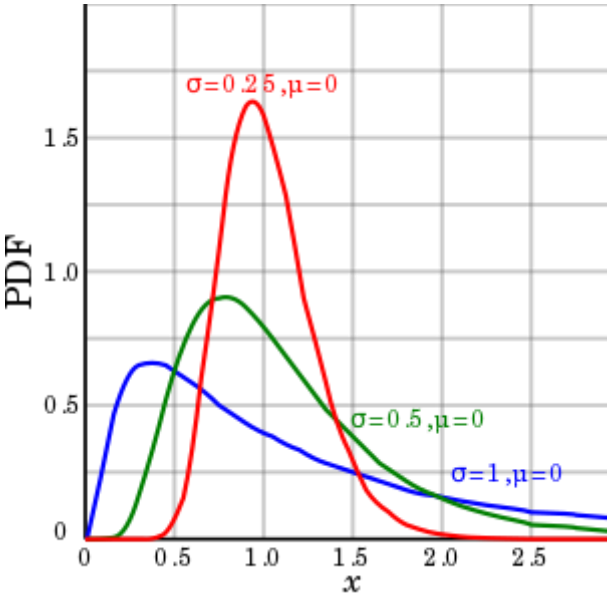


Figure 3: The chart displaying the abnormal (skewed) distributions

The above figure (figure 3) shows the left skewed data for all of the variables, where all green, blue and red curves shows the significant left skewness with significantly higher kurtosis. This type of data must be corrected on the distribution scale and produces a significant error during the classification otherwise. In this paper, the focus on the distribution correction is kept, and certain methods for the abnormal distribution correction is applied over the data using the log based conversion of the target data.

LITERATURE REVIEW

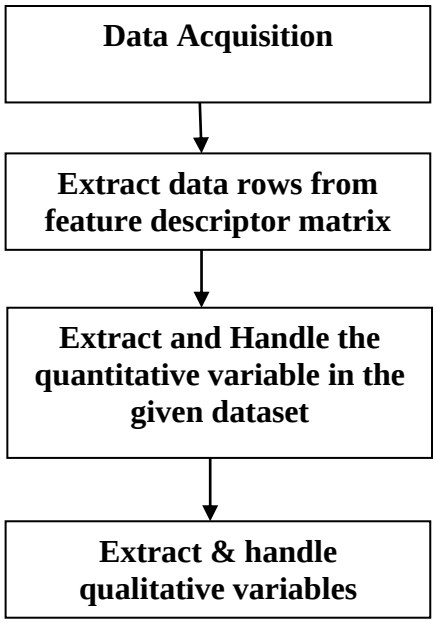
Kulkarni, Pallavi et. al. [13] has worked on the unbalanced financial data for the credit card fraud detection using the regression model. Traditionally, machine learning area has been developing algorithms that have certain assumptions on underlying

distribution of data, such as data should have predetermined and fixed distribution. Bahnsen, Alejandro Correa et. al. [14] has worked towards the feature engineering in order to improve the feature descriptors for the purpose of credit card fraud investigations. In this paper the authors have expanded the transaction aggregations strategy, and propose to create a new set of features based on analysing the periodic behaviour of the time of a transaction using the von Mises distribution. Dal Pozzolo, Andrea et. al. [15] has developed the financial risk prediction model based upon cloud systems to minimize the online threats on financial systems. Based on the threat model, they have proposed Secure-Logging-as-a-Service (SecLaaS), which preserves various logs generated for the activity of virtual machines running in clouds and ensures the confidentiality and integrity of such logs.

Bahnsen, Alejandro Correa et. al. [19] has worked on the improvement of the credit card fraud detection models with the calibrated probabilities. In this paper two different methods for calibrating probabilities are evaluated and analysed in the context of credit card fraud detection, with the objective of finding the model that minimizes the real losses due to fraud. Aktepe, Adnan et. al. [1] has worked on the customer satisfaction and loyalty analysis with classification algorithms and structural equation modelling. Businesses can maintain their effectiveness as long as they have satisfied and loyal customers. Customer relationship management provides significant advantages for companies especially in gaining competitiveness. In order to reach these objectives primarily companies need to identify and analyse their customers. Gaiardelli, Paolo et. al. [2] has worked towards the classification model for product-service offerings. In this paper, the authors have developed a comprehensive model for classifying traditional and green Product-Service offerings, thus combining business and green offerings in a single model.

EXPERIMENTAL DESIGN

This study begins with the literature study, hence firstly the literature review on the existing techniques of fraud detection was performed, in which we have analysed the problems in the existing models. This gives the perspective of the problems and possible improvements in the existing model to improve the overall accuracy. Afterwards, the feature imbalance correction and imputation methods are designed to create the balanced features to achieve the higher accuracy of classification. The main target of the proposed model is to eliminate or minimize the false negative errors from the classification in order to minimize the financial losses of the credit card organizations.



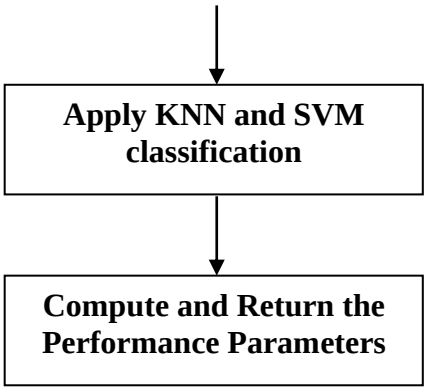


Figure 4: Overall workflow of the proposed model

Further, the implementation of the proposed model is completed using the python environment, which is performed and analysed on the training and testing data, as its being shown in the above figure (figure 4). The performance of the model is evaluated using the various parameters of error and accuracy.

Financial data is particularly collected by examining the customer's in the financial spending, where customers are analysed for their possibility of committing a fraud. Hence, the larger volumes of data can be collected from such financial organizations. In this proposed model, the credit card related observation data has been utilized under the supervised classification. This financial data provides the credit card policy planners with the ability to analyse their characteristics of the unusual spending or chances of committing the bank frauds. The online credit data has been very important in the recent years, as there are many ongoing financial projects in the various parts of the world to integrate the credit card spending pattern data from all citizens. Almost all of these projects are leading towards making this financial data available online, which enables the access of person's credit card and other banking history easier to all of the financial institutions in the certain countries or group of countries. In the proposed model, the combination of the supervised classification algorithm along with various feature engineering models and pre-processing techniques based algorithm has been designed to improve analytical performance of credit card frauds. The following steps describe the basic structure of proposed predictive model:

1. The input data acquisition is done on the real time healthcare data to observe the credit card frauds. The credit card fraud data contains the various attributes to detect and observe the presence or absence of credit card fraud in the certain customers under screening.

2. Each data row is inspected individually to classify the deterministic patterns in the given dataset on the basis of their similarity matrix. The pre-emphasis based class analysis depicts the mode of operation to divide a healthcare entry into certain class on the basis of presence or absence of credit card fraud.

3. The feature important is used to determine the significance of each individual features, which helps to analyse the decisions on pre-labelled data. This data is scaled and corrected for its distribution to observe the highly accurate classification results.

4. Afterwards, the supervised algorithm is executed over the processed data to predict the labels of testing data. The predicted observations are compared to the original

observations to determine the classification accuracy.

RESULT AND EVALUATION

The proposed financial fraud detection model for the credit card organizations is based upon the KNN based classification, which involves the multi-layered feature description. In figure 5, the results of the proposed model are observed against the different classification models, which involve the SVM and logistic regression. The proposed model based upon KNN is observed with 91.6% of recall, approx. 84% f1-error and nearly perfect approx. 100% accuracy with 0.05% standard error. The standard error of 0.05% cancels the possibility of overfitting, which builds a healthy and robust classification model.

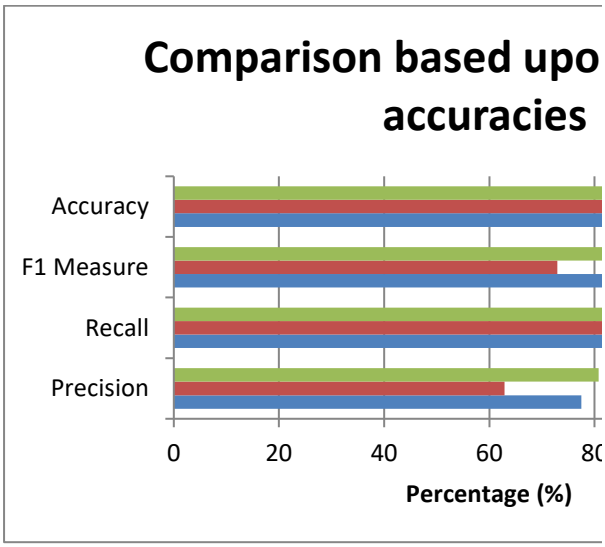


Figure 5: Comparison based upon the different statistical accuracy measures

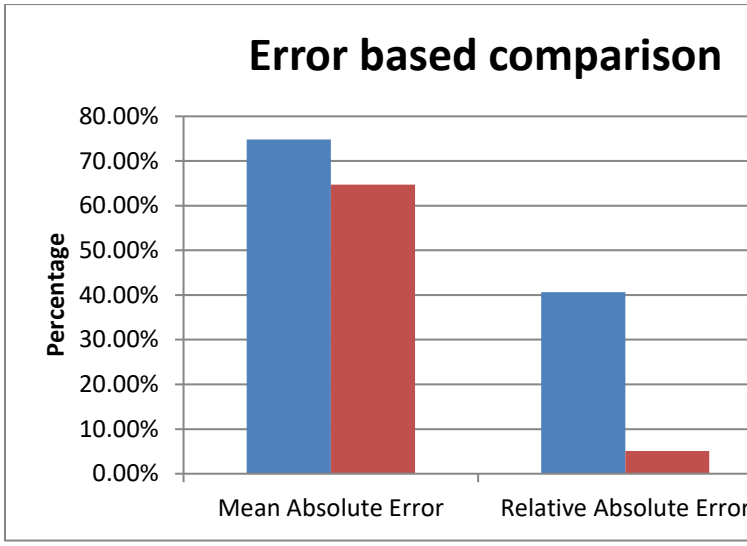


Figure 6: Error based comparison of existing and proposed models

The proposed model is learnt to outperform the existing model on the basis of both mean absolute error (MAE) and relative absolute error (RAE). The proposed model is observed with 0.051% RAE against the existing model's 0.406%, which is considered as the significant improvement of the proposed model. Also, the proposed model is learnt to improve on the basis of MAE with the observed value of approx. 64.50% against approx. 75% of the existing model.

CONCLUSION

The supervised classifier KNN based credit card fraud detection model is evaluated with the existing model on the basis of average readings of errors and accuracy based parameters. The average accuracy of proposed model is significantly improved (99.95%) by using the KNN classifier. The proposed model has achieved the improvement of higher than 3% on the basis of accuracy, which is evident from its observation 99.95% against 96.62% for existing model. The KNN based credit card fraud detection model has outperformed existing model on the basis of mean absolute error (MAE)

with error of 64.71% against 74.83%, which shows a noteworthy improvement of nearly 10%. Similarly, the proposed model produced 0.051% relative absolute error (RAE) against 0.406% of existing model.

REFERENCES

1. Aktepe, Adnan, Süleyman Ersöz, and Bilal Toklu. "Customer satisfaction and loyalty analysis with classification algorithms and structural equation modeling." *Computers & Industrial Engineering* 86 (2015): 95-106.
2. Gaiardelli, Paolo, Barbara Resta, Veronica Martinez, Roberto Pinto, and Pavel Albores. "A classification model for product-service offerings." *Journal of cleaner production* 66 (2014): 507-519.
3. Lu, Ning, Hua Lin, Jie Lu, and Guangquan Zhang. "A customer churn prediction model in telecom industry using boosting." *Industrial Informatics, IEEE Transactions on* 10, no. 2 (2014): 1659-1665.
4. K. Coussement and D. V. Poel, "Churn Prediction in Subscription Services: An Application of Support Vector Machines while Comparing Two Parameter-Selection Techniques," *Expert Systems with Applications*, Vol. 34, No. 1, Jan. 2008, pp. 313-327.
5. W. Verbeke, K. Dejaeger, D. Martens, J. Hur, and B. Baesens, "New Insights into Churn Prediction in the Telecommunication Sector: A Profit Driven Data Mining Approach," *European Journal of Operational Research*, Vol. 218, No. 1, Apr. 2012, pp. 211-229.
6. W. J. Reinartz and V. Kumar, "The Impact of Customer Relationship Characteristics on Profitable Lifetime Duration," *Journal of Marketing*, Vol. 67, No. 1, Jan. 2003, pp. 77-99.
7. P. Datta, B. Masand, D. R. Mani, and B. Li, "Automated Cellular Modeling and Prediction on a Large Scale," Artificial Intelligence Review, Vol. 14, No. 6, Dec. 2000, pp. 485-502.
8. D. Popović and B. D. Bašić, "Churn Prediction Model in Retail Banking Using Fuzzy C-Means Algorithm," *Informatica*, Vol. 33, No. 2, May. 2009, pp. 235-239.
9. C.-P. Wei and I. T. Chiu, "Turning Telecommunications Call Details to Churn Prediction: A Data Mining Approach," *Expert Systems with Applications*, Vol. 23, No. 2, Aug. 2002, pp. 103-112.
10. M. Owczarczuk, "Churn Models for Prepaid Customers in the Cellular Telecommunication Industry Using Large Data Marts," *Expert Systems with Applications*, Vol. 37, No. 6, Jun. 2010, pp. 4710-4712.
11. J. Burez and D. V. Poel, "Handling Class Imbalance in Customer Churn Prediction," *Expert Systems with Applications*, Vol. 36, No. 3, Apr. 2009, pp. 4626-4636.
12. N. Kim, K.-H. Jung, Y. S. Kim, and J. Lee, "Uniformly Subsampled Ensemble (USE) for Churn Management: Theory and Implementation," *Expert Systems with Applications*, Vol. 39, No. 15, Nov. 2012, pp. 11839-11845.
13. Kulkarni, Pallavi, and Roshani Ade. "Logistic Regression Learning Model for Handling Concept Drift with Unbalanced Data in Credit Card Fraud Detection System." In *Proceedings of the Second International Conference on Computer and Communication Technologies*, pp. 681-689. Springer India, 2016.

14. Bahnsen, Alejandro Correa, Djamila Aouada, Aleksandar Stojanovic, and Björn Ottersten. "Feature engineering strategies for credit card fraud detection." *Expert Systems With Applications* 51 (2016): 134-142.
15. Dal Pozzolo, Andrea, Olivier Caelen, Yann-Ael Le Borgne, Serge Waterschoot, and Gianluca Bontempi. "Learned lessons in credit card fraud detection from a practitioner perspective." *Expert systems with applications* 41, no. 10 (2014): 4915-4928.
16. Halvaiee, Neda Soltani, and Mohammad Kazem Akbari. "A novel model for credit card fraud detection using Artificial Immune Systems." *Applied Soft Computing* 24 (2014): 40-49.
17. Van Vlasselaer, Véronique, Cristián Bravo, Olivier Caelen, Tina Eliassi-Rad, Leman Akoglu, Monique Snoeck, and Bart Baesens. "APATE: A novel approach for automated credit card transaction fraud detection using network-based extensions." *Decision Support Systems* 75 (2015): 38-48.
18. Prakash, A., and C. Chandrasekar. "An optimized multiple semi-hidden markov model for credit card fraud detection." *Indian Journal of Science and Technology* 8, no. 2 (2015): 165-171.
19. Bahnsen, Alejandro Correa, Aleksandar Stojanovic, Djamila Aouada, and Björn Ottersten. "Improving credit card fraud detection with calibrated probabilities." In *Proceedings of the 2014 SIAM International Conference on Data Mining*, pp. 677-685. Society for Industrial and Applied Mathematics, 2014.
20. Zareapoor, Masoumeh, and Pourya Shamsolmoali. "Application of credit card fraud detection: Based on bagging ensemble classifier." *Procedia Computer Science* 48 (2015): 679-685.