IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Analysis of Bank Statement Using Secured & Optimized Machine Learning Algorithms

Ishan Padalkar, Mahesh Kulkarni, Pranav Kulkarni, Sukrut Pendharkar, Prof. Anita Shinde

[1] Marathwada Mitra Mandal's College of Engineering, Savitribai Phule Pune University, Pune

***Abstract*: In today's world large amount of data is generated in every field and banking industry is one of them. This data contains valuable information. Hence, it is very important to store, process, manage and analyze this data to extract knowledge from it. It helps to increase business profit. Banking industry plays very important role in economy of country. Customers are the main asset of the bank. Hence it is necessary to focus on problems faced by the banks. Frauds occurring now a days in banking sector are increasing a lot due to integration mobile banking and digital wallets. Fraud detection is not the only solution to tackle such problem, but it will surely help banks to keep track of fraudulent transactions and take necessary actions on them. Banks do have tremendous amount of sensitive data which cannot be shared with unauthorized person hence data security in this case plays a big role while extracting a specific knowledge from it. Here, we are working on customers and companies involved in transactions. Matching (checksum) the balance of transactions made and flag them as fraudulent if any. In this work, fuzzy logic algorithm is implemented for search purpose.**

***Keywords*: Fuzzy Logic, Transaction Type, Fraud Detection, Data Analysis, Report Generation.**

I. INTRODUCTION

In recent years, financial fraud, including credit card fraud, corporate fraud and money laundering, has attracted a great deal of concern and attention. The Oxford English Dictionary defines fraud as "wrongful or criminal deception intended to result in financial or personal gain." Phua describes fraud as leading to the abuse of a profit organization's system without necessarily leading to direct legal consequences. Although there is no universally accepted definition of financial fraud, Wang define it as "a deliberate act that is contrary to law, rule, or policy with intent to obtain unauthorized financial benefit." The Banking industry generates a massive volume of data every day. It contains customer account information, transaction information, all financial data etc. Data analytics can be used to analyze large volume data to extract meaningful information from it. It helps to uncover hidden information, hidden patterns and to discover knowledge from the large volume data. Banks statement analysis facing various challenges like customer's and companies' involvement, fraud detection, categorization of transaction type and report generation. It needs to focus on these challenges to increase business profit. Customer identification is effective method for the growth of the banks. In the banking sector churn and fraud becomes major problem today so it is important to identify customer's behavior and retain them. To retain customers first it is necessary to identify which customers are active and inactive. Fraud detection is difficult because it usually involves class imbalanced data, since there are significantly fewer fraudulent transactions. We will be applying different oversampling and under sampling methods to deal with the imbalanced nature of the data. Subsequently, different machine learning algorithms are tested on both imbalanced dataset and different versions of the balanced data.

II. LITERATURE REVIEW

This paper presents different technologies used to build this project. Fraud is one of the fast-growing issues in today's world especially when it is considered in the context of money. It has become a billion-dollar business and will keep on increasing every year.

Column Heading Matcher and Instance Matcher are employed to enhance the matching accuracy. A set of experiments are applied to real-world web tables and the results demonstrate that method has higher precision and accuracy [1].

There are many other applications other than fraud detection that have to deal with imbalanced nature of the data. This is one of the major issues faced in the field of machine learning that how to deal with imbalanced data [2]

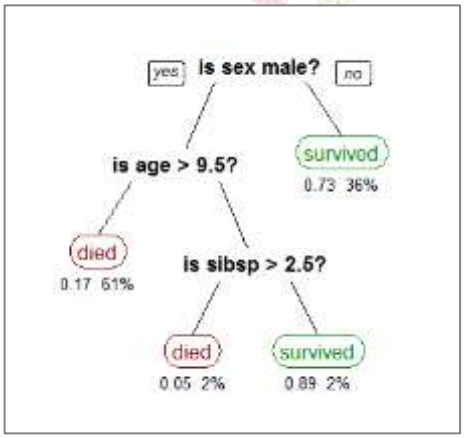
Handling data at backend is necessary to keep track of changes made to the data. Database is used to store a huge amount of data. Only data storage will not work in this case. Database connectivity is important to handle this kind of data. Different languages which support database connectivity are given in this paper [3].

Sr.no	Paper Name	Publication	Author	Key Findings
1	Analysis of Banking Data Using Machine Learning	IEEE,2018	Priyanka S. Patil, Nagaraj V. Dharwadkar	Artificial Neural Network algorithm for classification
2	Financial Fraud Detection using Machine Learning	University of Waterloo, 2019	Aamenah A. Bashir, Shafaq Z. Iqbal, Sulaiman Olabiyi	Imbalanced Data, SVM, Decision Tree Regression, Cohen-Kappa.
3	Database Connection Technology	International Journal of Advanced Research in Computer Science, 2018	Sonia Kumari, Kumari Seema Rani.	Java, Python, PHP, Hibernate, JDBC, etc
4	The application of data mining techniques in financial fraud detection	The Hong Kong Polytechnic University, 2017	Yong Hu, Y.H. Wong, Xin Sun, Yijun Chen.	Data recovery, Fraud detection, Data mining.
5	Django The Python Web Framework	International Journal of Computer Science and Information Technology Research	Prof. B Nithya Ramesh, Aashay R Amballi, Vivekananda Mahanta	Web Development, Model-View-Template

III. ALGORITHMS:

1. Decision Tree:

A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements. Diagram represents the exact working of



Simple Decision Tree for Survived People

decision tree. In our project decision tree will play a role to segregate fraud transaction detection

- **STEP 1:** Import the .csv file to be analyzed using decision tree.
- **STEP 2:** Once the file is imported, calculate either Information Gain or Entropy of different attributes in table.
- **STEP 3:** Now split the data depending on the attribute with lowest entropy or highest information gain.
- **STEP 4:** Repeat step 2 unless the last attribute is used for splitting dataset.
- **STEP 5:** Once all steps are followed then you will get output of attributes which affect more on the main output.

2. Support Vector Machine (SVM):

A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problems. After giving an SVM model sets of labeled training data for either of two categories, they're able to categorize new examples. Similarly, if we train our model with fraudulent transaction then it can categorize fraud transactions.

- **STEP 1**: Select any classification algorithm that you want to use for classifying the output.
- **STEP 2**: After selecting a precise classifying algorith follow the steps to achieve classification of output depending on the selected algorithm.
- **STEP 3**: Working of initial part of every algorithm is different than other hence keep in mind to follow proper initial steps of algorithm that you select.

3. Naïve Bayes:

Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem with the "naive" assumption of conditional independence between every pair of features given the value of the class variable. In this project Naïve Bayes can help us find probable fraud transactions depending on the previous results of that particular customer.

- **STEP 1:** Import the .csv file to be analyzed using Naïve Bayes algorithm.
- **STEP 2**: Once the file is imported next step comes in picture for calculating the **Conditional Probability** of the given record. This probability tells you about what percent does it belong to output class. This is known as **Maximum A Posteriori (MAP)**.
- **STEP 3**: After calculating MAP for all attributes, the attribute with maximum probability wins and is considered as most likely class for output classification.
- **STEP 4**: Output is then predicted using the most likely class.

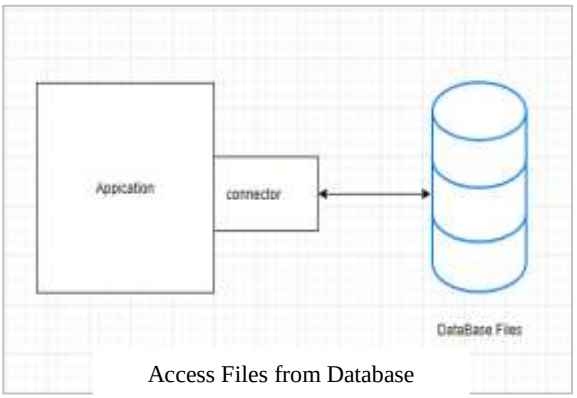
IV. TECHNOLOGY STACK:

1. Django Web Framework:

Django is an extremely popular and fully featured server-side web framework, written in Python. This module shows you why Django is one of the most popular web server frameworks, how to set up a development environment, and how to start using it to create your own web applications. It is a "batteries-included" philosophy. The principle behind batteries-included is that the common functionality for building web applications should come with the framework instead of as separate libraries.

2. SQLite:

SQLite is a relational database management system contained in a C library. In contrast to many other database management systems, SQLite is not a client–server database engine.[13] Rather, it is embedded into the end program. The lite in SQLite means light weight in terms of setup, database administration, and required resource. SQLite is self-contained, serverless, zero-configuration, transactional. SQLite database is integrated with the application that accesses the database. The applications interact with



Access Files from Database

the SQLite database read and write directly from the database files stored on disk.[16]

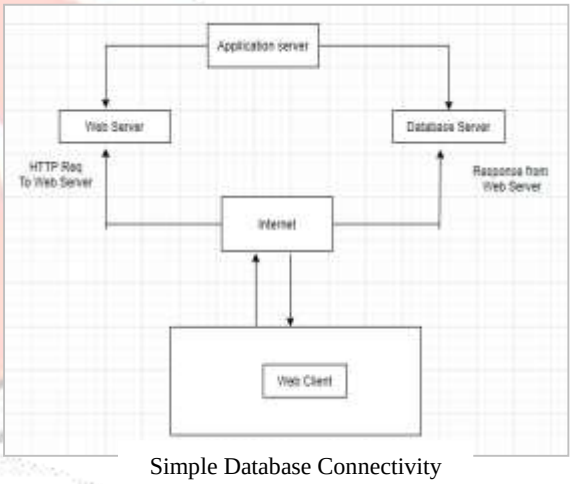
3. Python:

Python is a general-purpose interpreted, interactive, object-oriented, and high-level programming language. Python is designed to be highly readable. It uses English keywords frequently whereas other languages use punctuation, and it has fewer syntactical constructions than other languages. This project is completely programmed in python including database connectivity. Database connectivity in python is easier than any other languages in computer field. Python based Django framework helped in handling front-end of the project and SQLite is used for backend handling.

V. DATABASE CONNECTIVITY:

1. Database Connectivity in Java –

Java is an object-oriented programming language that offers a robust, secure and portable environment. It is unique in the sense that it is platform independent i.e. its programs can run in various platforms such as Linux, Microsoft Windows, Apple Macintosh, etc.[14] Compared to other high level, fully interpreted scripting languages, Java is one of the best in terms of performance. Moreover, it is a dynamic language that fully supports multithreading. Multiple relational databases over the web can be retrieved by its database connectivity interface.[15] Fig. shows the integration of web server and database server.



Simple Database Connectivity

2. Database Connectivity in Python –

a. Download and install MySQL Python connector. It is available in various platforms like Mac OS, Microsoft Windows etc. MySQL Python connector which is a standardized database driver provided by MySQL needed to access MySQL database from Python.[14]

b. Go to the python GUI that comes installed with Python called IDLE.From there type the following command:- **import mysql.connector**

c. Establish connection with some existing MySQL database:

```
import mysql.connector
conn=mysql.connector.connect(user='root', password='123', host='localhost', database='family')
```

Your database will be connected to the project once you complete all the steps mentioned above.

3. Database Connectivity in PHP –

PHP is an important language in the software development market. PHP is at the forefront of Web2.0 and Service Oriented Architectures supports technologies along with other open source projects MYSQL and Apache [12]. For many people, the foremost reason why they acquire knowledge about a scripting language like PHP is of the interaction with database it can offer. In this, I will show you how to use PHP to connect with MYSQL database. PHP is endorsed not only by its large open source community in the IT market such as IBM, Oracle and Microsoft. This paper provides instructions for connectivity to MYSQL database using PHP.[12]

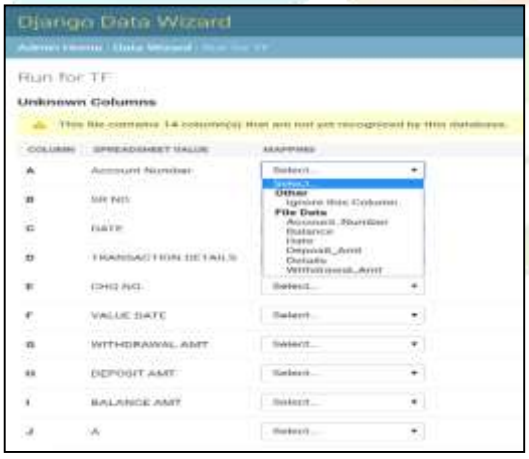
VI. Result –

i. Expected Outcome:

- Develop a machine learning model to filter out fraudulent transactions.
- Use necessary tools and libraries to secure transaction data.
- Get accurate prediction of fraud transactions using trained machine learning model.

ii. Output Screenshots:

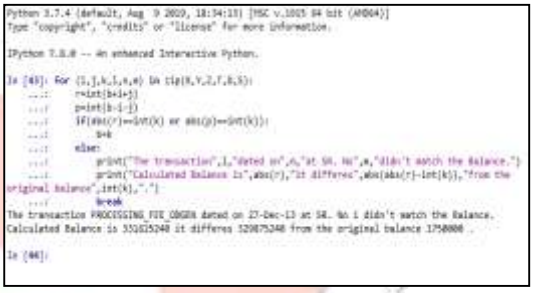
- Selection of columns from excel-



- Mapping of unrecognized columns with columns of system default table-

- Import entire data from external table to system table-

- Checksum the balance of transactions made and inform if fraud occurs-



VII. Discussion –

Outcome of this project is achieved when analysis of transactions is done, and final checksum gives output of fraud transaction. Project is capable of handling and detecting duplicate transactions which further lead to mistakes in calculating balance. Furthermore, project even tells us about a fishy transaction for example if you suddenly withdraw a huge amount when you generally withdraw a less amount in a month.

VIII. Conclusion –

Developing a software application that can be used to analyze the complex bank statement and help the user for an easy retrieval of data with graphical representation and report generation. Moreover, user can even check how much more he spent last month and how much he could save. Fishy transactions can catch user's attention to take necessary actions through this project.

Accepting data through any format which can be in PDF or any other. Fraud Detection can be implemented as future scope and retrieval of bank data using IFSC codes.

IX. References -

[1] IEEE TRANSACTIONS ON FUZZY SYSTEMS, VOL. 4, NO. 2, MAY 2010 "Fuzzy Logic = Computing with Words" Lotfi A. Zadeh, Life Fellow, IEEE.

[2] Priyanka Patil, Nagaraj Dharwadkar "Analysis of Banking data using machine learning " 2017 International Conference on

I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC).

[3] "Interval Type-2 Fuzzy Logic Systems Made Simple" Jerry M. Mendel, Life Fellow, IEEE, Robert I. John, Member, IEEE, and Feilong Liu, Student Member, IEEE.

[4] Abraham Kandel, Naphtali David Rishe (2018). Complex Fuzzy Sets and Complex Fuzzy Logic. 2018 IEEE International Conference on Consumer Electronics (ICCE).

[5] 2. L.A. Zadeh, "A Rationale for Fuzzy Control," J. Dynamic Systems, Measurement and Control, Vol. 94,Series G,1972,pp. 3-4.

[6] Chao Chen, Yue Zhao "Study on column matching" 12th Web Information System and Application Conference (WISA) 2015, pp. 708- 713. doi: 10.1109/CCAA.2015.7148468.

[7] "What is mathematical fuzzy logic "Peter Hájek Institute of Computer Science, Academy of Sciences of the Czech Republic, 182 07 Prague, Czech Republic.

[8] K. Rinartha, W. Suryasa and L. G. S. Kartika, "Comparative Analysis of String Similarity on Dynamic Query Suggestions," 2018 Electrical Power, Electronics, Communications, Controls and Informatics Seminar (EECCIS), Batu, East Java, Indonesia, 2018, pp. 399-404.

[9] Complex Fuzzy Sets and Complex Fuzzy Logic an Overview of Theory and Applications Dan E. Tamir, Naphtali D. Rishe and Abraham Kandel.

[10] Rajesh Ramachandran, "OFAC Name Matching and False-Positive Re-duction Techniques" Cognizant 20-20 insights.

[11] Jingtao Zhou, Mingwei Wang, Han Zhao, Shusheng Zhang, and Chao Zhang,"Concept Capture Based On Column Matching and Clustering "First International Conference on Semantics, Knowledge, and Grid IEEE 2016

[12] Sonia Kumari, Kumari Seema Rani, Manvendra Yadav, "Database Connection Technology" International Journal of Advanced Research in Computer Science Volume 8, No. 5, May-June 2017

[13] Daniel J. Abadi, Adam Marcus, Samuel Madden, and Kate Hollenbach. SW-Store: a vertically partitioned DBMS for semantic web data management. VLDB Journal, 18(2):385–406, 2015.

[14] S. Papastavrou , P. K. Chrysanthis, G. Samaras, E. Pitoura (2000) "An evaluation of the java-based approaches to web database access",International Conference on Cooperative Information Systems. Springer Berlin Heidelberg., 10(4), 102-113

[15] B.Vasavi, Y.V.Sreevani, G.Sindhu Priya(2011) "Hibernate technology for an efficient business application extension", Journal of Global Research in Computer Science, 2(6), 118- 125

[16] J. Keogh, C.J. Date, H. Darwen. (2002). J2EE: The Complete Reference: A Guide to the SQL Standard,New York,NY:Tata McGraw-Hill Education.