I.J. Modern Education and Computer Science, **2022, 1, 65-77**
Published Online February 2022 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2022.01.06

Detection of False Income Level Claims Using Machine Learning

Anil Kumar K.M
Department of Computer Science and Engineering, Sri Jayachamarajendra College of Engineering, JSS Science and Technology University, Mysuru, Karnataka, India
Email: anilkm@jssstuniv.in

Bhargava S
Department of Computer Science and Engineering, Sri Jayachamarajendra College of Engineering, JSS Science and Technology University, Mysuru, Karnataka, India
Email: bhargava0311@gmail.com

Apoorva R
Department of Computer Science and Engineering, Sri Jayachamarajendra College of Engineering, JSS Science and Technology University, Mysuru, Karnataka, India
Email: apoorvabharadwaj02@gmail.com

Jemal Abawajy
Deakin University, Geelong, Australia
Email: jemal.abawajy@deakin.edu.au

Received: 02 December 2020; Revised: 12 January 2021; Accepted: 25 February 2021; Published: 08 February 2022

Abstract: Data driven social security fraud detection has been given limited attention in research. Recently, social schemes have seen significant expansion across many developing countries including India. The fundamental aims of social schemes are to alleviate poverty, enhance the quality of life of the most vulnerable and offer greater chances to those relegated to the fringe of society to engage more enthusiastically in the society. Although governments channel billions of dollars every year in support of these social schemes, quite significant number of the eligible people are excluded from the program mainly through fraud and dishonesty. Although fraud is considered an illegal offence and morally reprehensible, it is unfortunate that the prevalence of fraud in social benefit schemes is rampant and a significant challenge to address. In this paper, we studied the viability of machine learning techniques in identifying fraudulent transactions in the context of social schemes. We focus on the detection of the false income level claims made by the fake beneficiaries to get the privileges of government scheme. We used the standard classifiers like Logistic Regression, Decision Trees, Random Forests, Support Vector Machine (SVM), Multi-Layer Perceptron and Naïve Bayes to identify fake beneficiaries of the government scheme from those deserving people. The results show that the Random Forest Classifier perform best providing an accuracy of 99.3% with F1 score of 0.99. The outcome of this research can be used by the government agencies entrusted with the management of the schemes to wade out the abusers and provide the required benefits to the right and deserving recipients.

Index Terms: Banking, Classification, Machine learning, Fraud Detection, Social Scheme

1. Introduction

United Nations and Governments across the world are committed to eradication of poverty by 2030 [36]. For instance, Government of India has come up with a number of social welfare schemes like Deen Dayal Antyodaya Yojana(DAY), Atal pension Yojana, Mid-day Meal Scheme etc. The fundamental aims of social schemes are to alleviate poverty, enhance the quality of life of the most vulnerable and offer greater chances to those relegated to the fringe of society to engage more enthusiastically in the society. One of the world's largest food securities schemes introduced in India is called Targeted Public Distribution System (TPDS) [17]. The Government of India classifies all the families based on their income levels into Above Poverty Line (APL) and BPL (Below Poverty Line) classes [1] to avail the schemes. Currently the Government of India has provided the classification responsibility to income tax department. This body issues income certificates to determine whether a family belongs to APL or BPL category.

Unfortunately, the methods followed by the government to issue the income certificates have a lot of drawbacks. There is a lot of inconsistency in records maintained by the departments. Quite a number of families are excluded and misclassified due to various reasons like fraud of government officials, inadequate information, lack of awareness etc. It is possible that the families belonging to APL can submit false income certificate and avail benefits of social schemes. It is unfortunate that the prevalence of fraud in social benefit schemes is rampant and a significant challenge to address.

There is a need to develop frame works and methods to address this problem and make sure these schemes reach to the actual beneficiaries. Government of India has an aim to provide a unique ID to all citizens of the country called Aadhar Card issued by the Unique Identification Authority of India (UIDAI) [3]. The Government is making it mandatory to link this unique ID to all the administrative functionalities such as our Mobile numbers, Bank Accounts, Ration Cards, and Permanent Account Number (PAN) Cards etc. [4]. Hence, it won't be possible to hide bank details from the government in the future. We make use of bank data to detect the fraud in false income claim.

In this paper, we studied the viability of machine learning techniques in identifying genuine and fake beneficiaries of social schemes. The contribution of this paper is to effectively classify the beneficiaries as BPL or APL using bank transaction dataset. This study provides different perspective of fraud plaguing government and other institutes in developing and underdeveloped countries. It also shows how machine learning can be employed to detect these kinds of frauds. The outcome of this paper can be used by Government agencies to identify genuine beneficiaries and abusers of social welfare schemes and for effective management of schemes. The rest of the paper is organized as follows: Section 2 discusses related works. Methodology followed for the experimentation is described in Section 3. Experiments and results are discussed in Section 4, finally Conclusion is discussed in Section 5.

2. Methodology

2.1 Dataset

It is always a challenge to obtain financial data set for research purposes due to privacy and security concerns. It is even more challenging to obtain bank transaction data. Thanks to the Principles and Practices Knowledge Discovery Challenge [15] in 1999, when a Czech Republic bank released1million plus bank transactions of its customers to the public as anonymous data in order to understand their customers and provide better services to them. The challenge of using this dataset is that it is not a labelled dataset and required labelling of the dataset.

2.2 Description of Dataset

The dataset consists of the fields like Client, Account, Card, Disposition, Permanent order, Transaction, Loan and Demographic data. The detailed description of the fields are provided in Annexure 1. Each record in client, Account, Card, Permanent Order and Demographic data describes characteristics of the irrespective fields. Each record in Loan field describes a loan granted for a given account. Each record in Disposition field relates together a client with an account i.e. this relation describes the rights of clients to operate accounts. Each record in transaction field describes one transaction of an account.

2.3 Data Preprocessing

We have to clean the data in all the individual data sets available. We use Python libraries Pandas and Numpy [19] for pre-processing. We deduce the age of the client from client data set. We convert all transaction types which are represented in Czech Republic Language to English for our analysis. In the URL of the dataset, there is description about the terms and their translations [15]. We did skip the demographic data as a part of our analysis as it contains data about the districts which doesn't provide any individual insight about any client. As the dataset consists of individual transactions, we cannot use the dataset directly for our analysis. Hence, we obtained the data at an individual level and obtained as much information about each individual as possible to make an effective analysis.

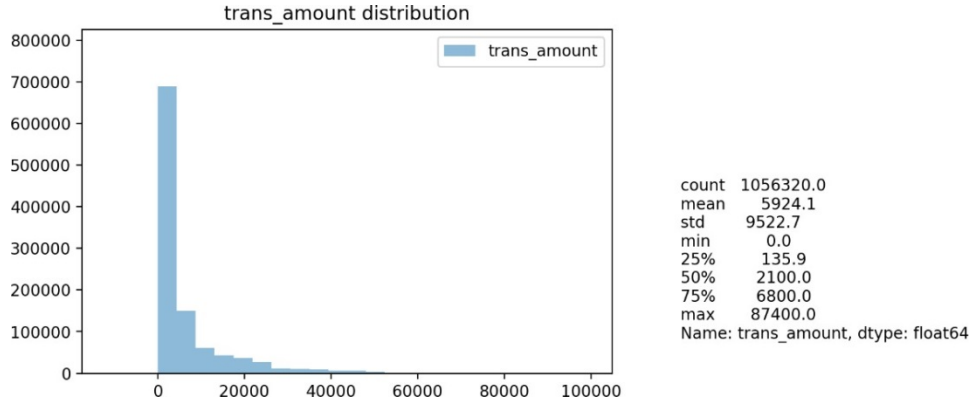

Fig.1 Distribution of account balance after transaction

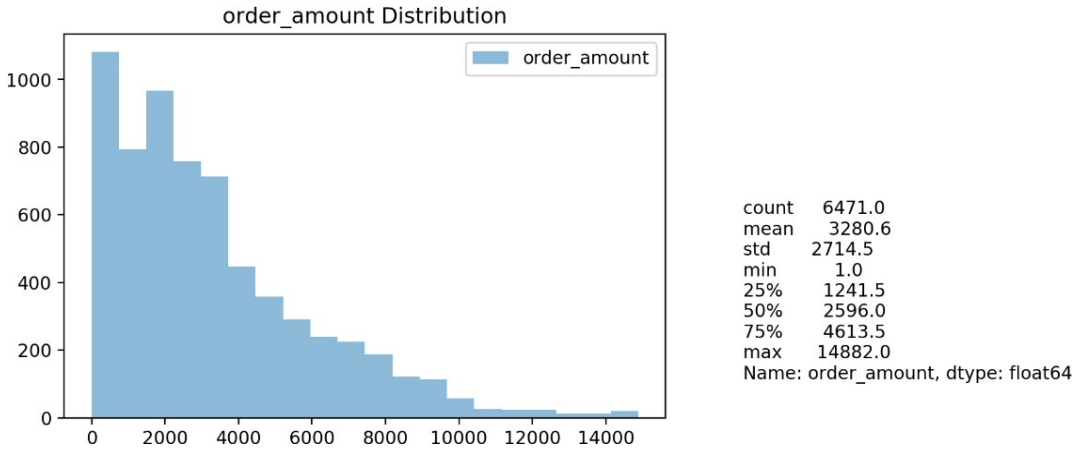

Fig. 2 Distribution of Order Amount

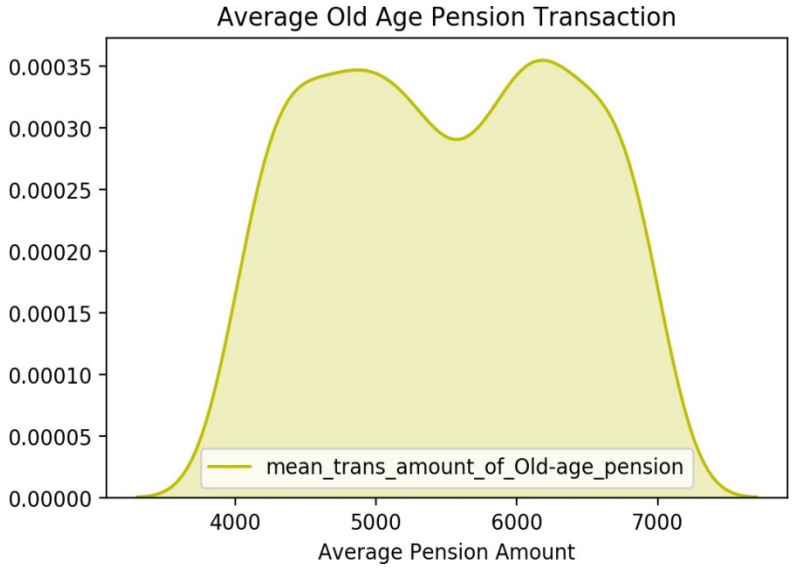

Fig. 3. Distribution of Average Pension Amount

To give a glimpse of the way the data is distributed Fig. 1 depicts the transaction amount distribution and Fig. 2 depicts the order amount distribution. It is evident from these figures that the distribution is more concentrated on lower amount. Hence, we cannot use a single attribute like transaction amount or order amount to identify the potential beneficiaries. A number of visualizations were obtained by using the Python's Matplotlib and pyplot [21] libraries to understand which at tributes to consider for classification process. As an example, consider Fig.3 which is the distribution of Average Old Age Pension Distribution. It is clear that this distribution is not skewed like the other two mentioned earlier. A number of aggregations and merges were performed to obtain individual level information as mentioned below.

- Aggregating Order on account ID by sum, mean and size to obtain total sum of order amount. Aggregating Order on order amount by mean and size to obtain average order amount.
- Aggregating Transaction on transaction amount by sum, mean and size to obtain total sum of Transaction Amount.
- Aggregating Transaction on account ID and transaction symbol by sum, mean and size to obtain average Transaction Amount.
- Aggregating Transaction on transaction type by mean and size to obtain average transaction amount in each transaction type.
- The transaction table is merged on account ID.
- The transaction table is merged on transaction type.
- The transaction table is merged on date intervals.
- The transaction table is merged on value of the sum of transaction amount per year.
- The above obtained transaction table and the order table are merged based on account ID.
- The resulting table above is merged with client table based on client ID.

- The resulting table above is merged with account based on account id.
- The resulting table above is merged with disponent based on disponent ID.
- The resulting table above is merged with card based on disponent ID.

After performing all the above operations, whole data are grouped by account ID to obtain dataset consisting of 5369 number of account IDs.

2.4 Labelling

The dataset is not labelled and needed a procedure to label the data. Hence, we took the help of human evaluators to label this data. A total of 560 human evaluators were employed for this purpose. Instructions were provided on how to label the dataset by providing the necessary information as well as the quartile ranges of the attributes in order to achieve relative classification rather than absolute as the demographics of the data and the timeframe of the data is totally different. Labels were chosen as either APL or BPL, when more than 60% of evaluators voted for the same. We obtained a labelled dataset with 2,200 beneficiaries being labelled as BPL and 3,169 beneficiaries labelled as APL and remaining were discarded because of noise and lack of agreement by the evaluators.

2.5 Data Preparation

After all the aggregations, merges and labelling, we finally obtain a dataset with 76 attributes. We used this dataset for classification purpose. The attributes include all the aggregated measures of each table that were carried out in the pre-processing stage like the average transaction amount per year, average transaction amount by type of orders, amount of loan etc.

We have to prepare this data before presenting them to classification. All the null values and "nan" values were labelled with '0's and some of the columns that had important information like account id and client id missing were removed from the data set. Finally, we obtain 5369 rows describing detailed analysis of each client in the data set.

2.6 Classification

After the labelling and data preparation exercises, we have a dataset that is suitable for classification. We use the following classifiers in our experiment:

2.6.1 Logistic Regression: It is used to describe data and relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio- level independent variables [12]. Logistic regression is an efficient and general method for predict 0/1 response. It is one of the popular methods used in Fraud detection.

2.6.2 Decision Trees: It is a classic and natural model of learning. It derives its name from its nature in which we write our set of questions and guesses in a tree format. Features are the questions which we can ask in general and feature values are the responses to these questions [13]. The output is called the label. Training data will consist of a set of feature values, paired with labels.

2.6.3 Support Vector Machines: It is similar to logistic regression in that it is driven by a linear function [12].

$$wT + b \tag{1}$$

The support vector machine does not provide probabilities unlike logistic regression. It provides outputs as a class identity. The SVM predicts that the positive class is present when wT +b is positive. Likewise, it predicts that the negative class is present when wT + b is negative [14].

2.6.4 Random Forests: Training is one of the most computationally expensive parts of decision trees. The expensive part is choosing the tree structure. The usage of trees with fixed structures and random features reduces this computational overhead. Collections of trees are called forests, and hence the classifiers built like this are called random forests.

The Random Forest algorithm takes three arguments: the data, a desired depth of the decision trees, and a number K of total decision trees to build [13]. The algorithm generates each of the K trees independently. A complete binary tree of required depth is generated for each tree. The features used at the branches of this tree are selected randomly. The resulting classifier is then just a voting of the K-many random trees. The Random Forest algorithm has been found to provide a good estimate of the generalization error and to be resistant to overfitting.

2.6.5 Neural Networks: Learning here is called training. There are two types of training methods supervised and unsupervised. In supervised training, samples of both fraudulent and non-fraudulent records are used to create models [14]. However, unsupervised training simply seeks those transactions that are most dissimilar. Also, unsupervised techniques do not require earlier knowledge of fraudulent and non-fraudulent transactions in database.

2.6.6. Naïve Bayes: This approach makes the simple assumption that all the attributes are independent. This leads to a simpler and effective classifier. The independence assumption implies that the likelihood can be decomposed into a product of dimension-wise probabilities [37].

$$\mathrm{P}(x|c_i = P(x_1, x_2, \ldots, x_d|c_i) = \prod_{j=1}^{d} P(x_j|c_i) \quad (2)$$

The classifier uses the sample mean diagonal sample co-variance matrix for each class c_i.Thus, in total 2d parameters have to be estimated, corresponding to the sample mean and sample variance for each dimension.

By applying classification machine learning technique on the data that was cleaned and pre-processed as mentioned above we are ready with our system to predict if a given beneficiary's income category which is our goal. The pen and paper based system is replaced with a machine learning system.

3. Experimental Setup

After the labelling, cleaning and preparation, our dataset is ready for use with the classification models.

3.1 Experimental Environment

We used python 3.6 scikit-learn [20] library version 0.20.2 to build our model with the following Experimental Settings: In Logistic Regression model, L1 regularization is chosen. Coordinate descent optimization is chosen. Maximum iterations for convergence is initialized to 100. In SVM Classifier Penalty parameter C of the error term is selected as 1. The kernel type is linear. The gamma value is equal to the number of features * Variance of the variable. In Multi-Layer Perceptron model, we used Relu as the activation function. Alpha value was set to 0.05. Beta value was set to 0.9. Number of hidden layers was 40. Learning rate was initialized to 0.001. Optimization algorithm used was Limited Memory Broyden Fletcher Goldfarb Shanno (BFGS). In decision Tree Classifier the criteria used for splitting is Gini Index. Nodes are expanded until all leaves are pure. The minimum number of samples required to split an internal node is selected to be equal to 2. The number of features to consider when looking for the best split is selected to be equal to the number of features. In Random forest model the number of trees in a forest is initialized to 10. The splitting criteria is chosen to be Gini and other parameters relating to the tree are same as the decision trees.

3.2 Performance Metrics

True Positives (TP): They are the cases, when the actual class of the data point was true and the predicted is also true.

True Negatives (TN): They are the cases, when the actual class of the data point was False and the predicted is also false.

False positives (FP): They are the cases, when the actual class of the data point was False and the predicted is true. False is because the model has predicted incorrectly and positive because the class predicted was a positive one.

False negatives (FN): They are the cases, when the actual class of the data point was 1(True) and the predicted is 0(False). False is because the model has predicted incorrectly and negative because the class predicted was a negative one.

Accuracy: It is defined as the number of correct predictions made by the model overall predictions made. It is a standard measure to compare our model.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

Precision: It is defined as the number of true positives over the number of true positives plus the number of false positives. Precision talks about how precise our model is. It is important in our case as costs of False Positive i.e. a non-fraudulent case predicted as fraudulent is high.

$$Precision = \frac{TP}{TP+TN} \quad (4)$$

Recall or Sensitivity: It is defined as the number of true positives over the number of true positives plus the number of false negatives. In our case it gives the accuracy of fraud cases which is important as we should have a measure of how many fraud cases can we detect accurately.

$$Sensitivity = \frac{TP}{TP+FP} \quad (5)$$

Specificity: It measures the proportion of actual negatives that are correctly identified as such. In our case it gives the accuracy of non-fraudulent cases which is important as we should have a measure of how many non-fraud cases can we detect accurately.

$$Specificity = \frac{TN}{TN+FP} \quad (6)$$

F1 score: It is defined as the harmonic mean of precision and recall. F1 Score provides a better measure of balance between Precision and Recall. Accuracy does not focus on False Negative and False Positive much which are a useful measure in our case as it is important for us to find the misclassification rate.

$$F1\ Score = 2 \times \frac{P \times R}{P + R} \tag{7}$$

Another reason for choosing these particular performance metrics is that these are the metrics used in almost all the papers we referred to in our Literature Survey which makes it easier for us to evaluate our model.

3. Result Analysis and Discussion

Table 1 provides the results for 50:50 training and testing split, and we observe that Random Forest provides the best result with an accuracy of 99.3% against the other classifiers. Table 2 provides the results for 60:40 training and testing split, we found Random Forest provides the best result with an accuracy of 99.39% against other classifiers. Also, Table 3 provides the results for 70:30 training and testing split, we observe that Random Forest provides the best result with an accuracy of 99.4% against other classifiers. From the above, we observe that the random forests and decision tree models are well suited for classification of data into BPL or APL. That is because the way in which these trees work and the way our intuition works to label is almost the same. Both use the rule-based procedures for each attribute and takes a decision at each stage. Hence, we conclude that Decision trees and Random Forest are the best models for identification of fake beneficiaries.

Table 1. Results for 50:50 split

Sl No.	Classifier	Accuracy	Sensitivity	Specificity	Precision	F1 Score
1	Logistic Regression	0.955	0.974	0.900	0.966	0.965
2	Naïve Bayes	0.895	0.870	0.969	0.988	0.897
3	Linear SVM	0.951	0.949	0.960	0.985	0.964
4	Multi-Layer Perceptron	0.770	0.994	0.092	0.764	0.863
5	Decision Tree	0.989	0.991	0.983	0.994	0.988
6	Random Forest	0.993	0.995	0.991	0.997	0.995

Table 2. Results for 60:40 split

Sl No.	Classifier	Accuracy	Sensitivity	Specificity	Precision	F1 Score
1	Logistic Regression	0.951	0.969	0.900	0.967	0.965
2	Naïve Bayes	0.853	0.808	0.980	0.893	0.890
3	Linear SVM	0.912	0.890	0.973	0.937	0.948
4	Multi-Layer Perceptron	0.770	0.994	0.090	0.760	0.860
5	Decision Tree	0.986	0.988	0.980	0.990	0.991
6	Random Forest	0.993	0.996	0.987	0.995	0.995

Table 3. Resultsfor70:30split

Sl No.	Classifier	Accuracy	Sensitivity	Specificity	Precision	F1 Score
1	Logistic Regression	0.954	0.969	0.912	0.969	0.970
2	Naïve Bayes	0.855	0.811	0.983	0.892	0.892
3	Linear SVM	0.946	0.978	0.856	0.964	0.952
4	Multi-Layer Perceptron	0.773	0.991	0.090	0.773	0.871
5	Decision Tree	0.986	0.986	0.985	0.990	0.991
6	Random Forest	0.994	0.994	0.990	0.995	0.995

Figure 4 shows the accuracy of six different classifiers namely Random Forest, Decision Tree, Linear SVM, Logistic Regression, Multi-Layer Perceptron & Naïve Bayes used for the experiment with different train and test cases.

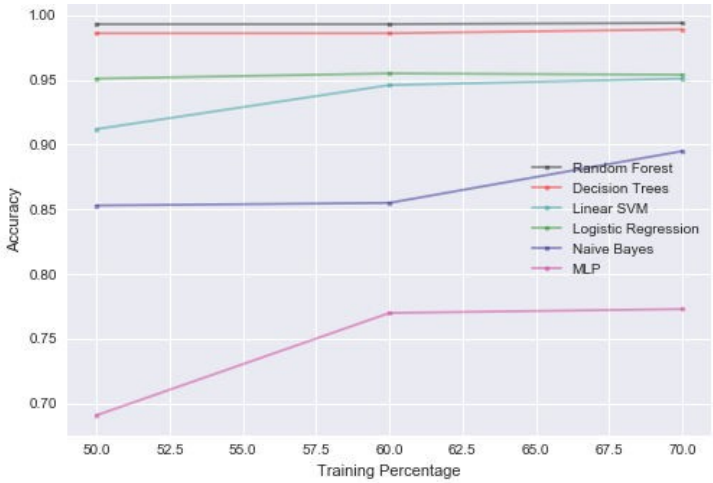

Fig. 4 Accuracy of all classifiers

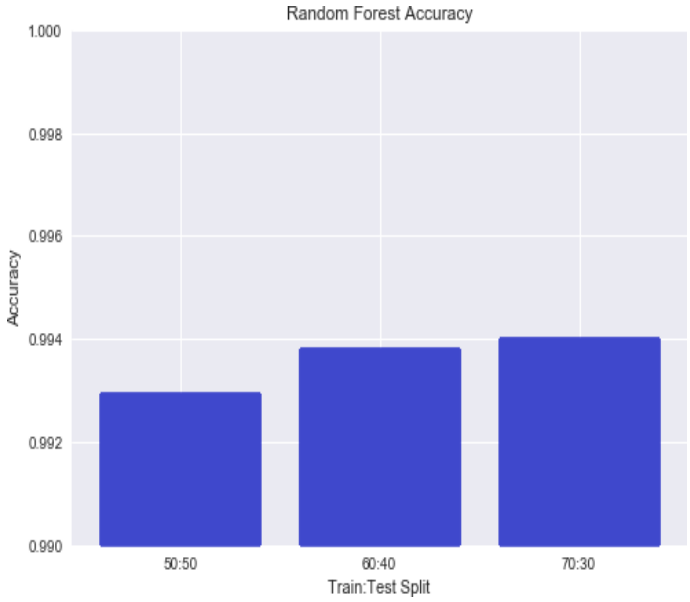

Fig. 5 Accuracy of Random Forest for Different splits

Figure 5 shows the accuracy of Random Forests for different train and test cases for the given dataset. Also, from Figure 5, we observe that there is no significant difference in terms of accuracy with different train and test cases hence we conform there is no overfitting.

Table 4. Comparison of Results from Literature

	Paper	Classifier	Accuracy	Sensitivity	Specificity	Precision	F1 Score
1	John et al [16]	Logistic Regression	0.980	0.080	0.970	0.980	0.160
2	-do-	Naïve Bayes	0.973	0.807	0.974	0.050	0.095
3	-do-	KNN	0.969	0.883	0.971	0.410	0.560
4	Sahil et al[11]	Logistic Regression	0.939	0.940	0.939	0.940	0.940
5	-do-	Naïve Bayes	0.905	0.910	0.953	0.910	0.910
6	-do-	Linear SVM	0.932	0.930	0.946	0.930	0.930

7	-do-	Decision Tree	0.908	0.910	0.980	0.994	0.910
8	-do-	Random Forest	0.945	0.950	0.966	0.950	0.950
9	Siddhartha et al[10]	Logistic Regression	0.966	0.740	0.967	0.100	0.177
10	-do-	SVM	0.955	0.687	0.957	0.072	0.131
11	-do-	Random Forest	0.978	0.812	0.979	0.157	0.264
12	Our results	Logistic Regression	0.954	0.969	0.912	0.969	0.970
13	-do-	Naïve Bayes	0.855	0.811	0.983	0.892	0.892
14	-do-	Linear SVM	0.946	0.978	0.856	0.964	0.952
15	-do-	Multi-Layer Perceptron	0.986	0.986	0.985	0.990	0.991
16	-do-	Decision Tree	0.994	0.994	0.990	0.995	0.995
17	-do-	Random Forest	0.773	0.990	0.090	0.770	0.870

Table 4 provides results obtained by different authors in detection of fraud as reported in literature. Even though our task is different from them, it shows effectiveness of algorithms in dealing with different fraud.

We observe that Logistic Regression provides a better result with an accuracy of 95.5%. Also, we observe that Random Forest provides a better result with an accuracy of 99.4%. Decision Tree provides a better result with an accuracy of 98.6%. These algorithms provide afore mentioned result for train and test split of 70:30. On an average, Random Forest provides an accuracy of 99.3%, Decision Tree provides an accuracy of 98.7% and Logistic Regression provides an accuracy of 3%. These results are found to be better than results reported in literature. We found Random Forest algorithm to be better compared other algorithms and can use the same for effectively identifying fake beneficiaries availing social scheme.

4. Related Work

A Study [18] has shown that Public Distribution System suffers from nearly 61% of error in exclusion and 25% of error in inclusion of beneficiaries resulting in the misclassification of the poor as non-poor and vice-versa. According to TIICMS India corruption study 2007 report [2] nearly 3.5 million BPL households paid Indian Rupees 1,224 million as bribe.

S. Bhattacharyya et al [23] discuss serious and growing problem related to credit card problem. There are many predictive models employed for credit card fraud detection. Application of data mining for research on credit card fraud detection is less due to non-availability of data. They have experimented with logistic regression, support vector machine and random forest on real data set and find random forest providing good result compared to others. E.W.T.Ngai et al [28] have presented a review of data mining techniques that can be useful for detection of financial fraud detection. They have listed four categories of fraud in banking sector, insurance sector, commodities and others. They have also listed different techniques like classification, regression etc., are useful for fraud detection. Sandeepkumar et al [42] presents a model using Enhanced Deep Feed Forward Neural Network Model to predict the customer whittling down in the Banking Domain. They compare various classes of machine learning algorithms Logistic regression, Decision tree, Gaussian Naïve Bayes Algorithm, and Artificial Neural Network

Kuldeep et al [24] address the financial fraud using machine learning algorithms. They have used ensemble methods Ada boost and Majority voting to detect credit card frauds. Also, they have used publicly available data set for evaluating their model and reports that majority voting method provides better result in fraud detection. M. Seera et al [25] have employed Genetic Algorithm with fuzzy c-means for detection of automobile insurance fraud. They have divided their test set into genuine class and suspicious class. They have further analyzed suspicious cases with a set of machine learning algorithms. They report that the support vector machine provides a better result compared to others.

Gosh and Reily [5] train the neural network to detect fraud accounts. They trained their network on large sample of data of labelled data set and tested on data that was collected for two-month period. They report that trained network performed better in fraud detection and that substantially reduce fraud and human review. Minegishi et al [6] presented a new type of decision tree learning called Very Fast Decision Tree (VFDT). It considers data as data stream and analyzes the same for fraud detection. They also propose set of criteria for implementation of VFDT for its node construction and evaluate the feasibility of such method on imbalanced distribution data streams.

Sherly K.K et al [8] proposed three classification algorithms that showed to be very effective in fraud detection. They also suggested advanced algorithms can be leveraged with basic ones resulting in hybrid approach that has better fraud coverage and less false positives. Varre Perantalu and Bhargav Kiran [9] proposed fraud detection using predictive modelling, logistic regression, and decision tree. The data set used for the experiment contains credit card transactions in September 2013 by European cardholders. The data set is unbalanced and made up of transactions that happened in 2 days with 492 frauds out of 284,807 transactions, approximately accounting to 0.172 % of fraud on all transactions. Doaa Hassan [40] investigates the impact of false negative cost on cost sensitive learning developed using Bayes Minimum risk as an applied mechanism for making a classifier cost sensitive. She applied her approach on Credit card fraud detection dataset and her results show that classifiers behave differently at different costs of FN including the real and average amount of transaction, and a range of random constant costs that are greater or less than the average amount.

O. S. Yee et al [10] have used data mining techniques for understanding the patterns of suspicious and non-suspicious transactions. They also have used machine learning and data mining to identify those patterns and show that they achieve a better result with set of classifiers. Clifton Phua et al [22] provide a survey paper that categorizes, compares, and summarizes articles in automated fraud detection from a particular time frame. They have discussed about fraudsters, different types of frauds and provide an outline of industries affected by frauds. The survey covers much more technical articles and proposes alternative data and solutions for different industries. Sunil Kappal [44] application used various classification methods and analytical techniques to identify a potential fraud. Sunil presented a hybrid fraud detection using the Bayesian Classification technique followed by Benford's Law to detect a fraudulent transactions. He used the Bayesian model on Conversational Analytics Output to calculate the probabilities of fraud into high medium or low classes.

Sahil Dhankhad, et al [11] use supervised machine learning algorithms to create ensemble classifier to detect credit card fraud. They use it against a real-world data set. They have identified important parameters that provide better result against other classifiers mentioned in literature. Adrian and Banarescu [26] discusses how technology can be leveraged for efficient detection and prevention of frauds in both public and private economic bodies. Lynnette Purda and David Skillicorn [27] developed a tool used to distinguish between fraudulent and non-fraudulent reports based on the language used in reports. They assigned probability score to each report for distinguishing the same. Also, their work assigns probability of truth with both quantitative and language-based approaches. They infer that the language-based approach is very effective in identifying fraudulent reports. Similarly other studies by Maja Puh ; Ljiljana Brkić [38] and Vaishnavi Nath Dornadula and S Geetha [39] have discussed fraud related to credit cards and employed machine learning to detect fraud.

Van Vlasselaer et al [31] propose a novel approach to detect frauds in online transactions. They use features Recency, Frequency and Monetary (RFM) along with network-based features for providing a suspicious score for new entity. The results show both are effective in detecting online frauds. Zhou, H et al [29] present several machine learning algorithms before finalizing on gradient boosting algorithm. They use enrolment behaviour for enhancing the feature set. They plan to use transaction history in their future work. They also inform that their work have been used in new design for mobile payment fraud detection system.

Johannes Jurgovsky et al [30] detects fraud using a sequence of classification task and use long short-term memory (LSTM) as part of the sequence. They report good accuracy achieved with usage of LSTM. Shantanu Rajora et al [33] provide comparative study of machine learning algorithms for detecting frauds in credit cards. They have experimented with ensemble learning and classification algorithm with and without time as a feature. They have found two ensemble learning algorithms perform better leaving out the time feature. They found majority of classifiers employed for the experiment provides a good score for all the features. Hamza O. Salami et al [41] conducted a study on building decision tree models for automatically detecting anomalies in student's examination results. They divided anomalies into two types: Course based anomalies and Student based anomalies. After applying Decision Tree classifiers were applied on these anomalies to classify the examination results.

Y. Bouzembrak et al [34] have developed a tool called MedISys-FF that processes food reports published in media to identify food fraud. Their results indicate that the tool is very effective and collects publications on food fraud with highest relevance. It is also used to create data set on food fraud. Hans J. Pet al [35] present the significance of big data in the area of food safety. They have discussed several new means in dealing with food safety. One of them is the contribution of mobile phones and social media in identifying issues related to food safety. Suresh et al [43] presents a hybrid approach for detecting money laundering activities in suspicious accounts using Data Mining Techniques. They use hashing technique detect high frequency activities in banking accounts and apply graph theoretic approach to detect suspicious activities in a bank account.

We have gone through a number of literatures related to fraud detection and have mentioned only a few above. Majority of these literatures discuss fraud related to credit card, insurances, network intrusion etc. Existing literature use rules to detect fraud based on historical data, data mining, machine learning and combination of data mining and machine learning to detect fraud in credit card, insurances, network intrusion etc. A few of them have used text mining techniques to obtain data from social media and use it to detect fraud. However, our work is different, we focus on

detecting fraud in social schemes by the nature of transactions carried out by the beneficiaries. We use machine learning techniques to identify fake beneficiaries availing the social welfare schemes.

5. Conclusion

In this paper, we have addressed the identification of fake beneficiaries of social scheme through their fraudulent transactions. We have taken an Indian context and classified the beneficiaries as BPL or APL for identification of abusers. We have used different machine learning algorithms for the experimentation and found Random Forest algorithm with an accuracy of 99.3% to be effective in detection of fake beneficiaries or abusers of social welfare scheme. This system is much more rationale and efficient when compared to the existing pen and paper based model, where all the decision making power lies in the hands of middle men and corrupted officials. It is an improved and systematic approach with minimum human intervention which helps to prevent the existing problems. This study is significant to wade out abusers of social schemes, as governments in developing and underdeveloped countries are implementing various schemes for the welfare of its citizens. Although the use case taken here is TPDS, this system can be extended to any social welfare schemes that the government decides to exercise.

Description of Dataset

- **Client (5369objects):**The client record has the following fields:
  - client_id: client identifier
  - birthnumber: the number is in the form YYMMDD for men; the number is in the form YYMM+50DD for women, where YYMMDD is the date of birth
  - district_id: address of the client

- **Account (4500objects):**The Account record has the following fields:
  - account_id: Identification of the account
  - district_id: Location of the branch
  - date: Date of creating of the account in the form YYMMDD
  - date_frequency: Frequency of issuance of statements

- **Card (892objects):**The Card record has the following fields:
  - card_id: Record identifier
  - disp_id: Disposition to an account
  - type: Type of card of which possible values are "junior", "classic", "gold"
  - issued: Issue date

- **Disposition (5369objects):**The Disposition record has the following fields:
  - disp_id: Record identifier
  - client_id: Identification of a client
  - account_id: Identification of an account
  - type: Type of disposition (owner/user)

- **Permanent Order (6471objects):**The Permanent Order record has the following fields:
  - order _id: Record identifier
  - account_id: Account, the order is issued for
  - bank_to: Bank of the recipient
  - account_to: Account of the recipient
  - amount: Debited amount
  - K_symbol: Characterization of the payment

- **Transaction (1056320objects):** The Transaction record has the following fields:
  - trans_id: Record identifier
  - account_id: Account, the transaction deals with
  - date: Date of transaction in the form YYMMDD
  - type: +/-Transaction
  - operation: Mode of transaction
  - amount: Amount of money

- balance: Balance after transaction
- k_symbol: Characterization of the transaction
- bank: Bank of the partner each bank has unique two-letter code
- account: Account of the partner

- **Loan (682objects):** The Loan record has the following fields:
  - loan_id: Record identifier
  - account_id: Identification of the account
  - date: Date when the loan was granted in the form YYMMDD
  - amount: Amount of money
  - duration: Duration of the loan
  - payments: Monthly payments
  - status: Status of paying off the loan

- **DemographicData (77object):**The Demographic record has the following fields:
  - A1=district id district code
  - A2=district name
  - A3=region
  - A4=no. of inhabitants
  - A5=no. of municipalities within habitants<499
  - A6=no. of municipalities with habitants between 500-1999
  - A7=no. of municipalities with habitants between 2000-9999
  - A8=no. of municipalities with habitants >10000
  - A9=no. of cities
  - A10=ratio of urban inhabitants
  - A11=average salary
  - A12=unemployment rate '95
  - A13=unemployment rate'96
  - A14=no. of entrepreneurs per 1000inhabitants
  - A15=no. of committed crimes '95
  - A16=no. of committed crimes '96

References

[1] http://planningcommission.nic.in/reports/genrep/pov/rep0707.pdf
[2] https://www.transparency.org/_les/content/pressrelease/IndiacorruptionStudy2007KeyHighlights.pdf
[3] https://uidai.gov.in/about-uidai/unique-identi_cation-authority-of-india/vision-mission.html
[4] https://rbidocs.rbi.org.in/rdocs/noti_cation/PDFs/85454.pdf
[5] Ghosh, S. and Reilly, D. L. 1994. Credit card fraud detection with a neural network in Proceedings of the27th Annual Hawaii International Conference on System Science vol.3.
[6] Minegishi, Tatsuya & Niimi, Ayahiko. (2013). Proposal of Credit Card Fraudulent Use Detection by Online-type Decision Tree Construction and Verification of Generality. International Journal for Information Security Research.
[7] G. R. Faulhaber,O Design of service systems with priority reservation O in Conf. Rec. 1995 IEEE Int. Conf. Communications, pp. 38.
[8] Sherly K.K," A Comparative Assessment of Supervised Data Mining Techniques for Fraud Prevention", TIST. Int. J. Sci. Tech. Res., Vol.1 (2012), 1-6.
[9] Varre Perantalu, Bhargav Kiran-"Credit card Fraud Detection using Predictive Modeling: a Review".
[10] O. S. Yee, S. Sagadevan, N. Hashimah, A. Hassain, Credit Card Fraud Detection Using Machine Learning As Data Mining Technique, vol.10,no. 1, pp.23-27.
[11] Sahil Dhankhad, et al., \Supervised Machine Learning Algorithms for Credit Card Fraudulent Transaction Detection: A Comparative Study", IEEE International information Reuse and Integration or Data Science, pp122-125,2018
[12] AllenB.Downey,ThinkStats,O'ReillyMedia,Inc.,2011
[13] Mohammed J. Zaki , Wagner Meira Jr, Data Mining and Analysis: Fundamental Concepts and Algorithms, Cambridge University Press, New York, NY, 2014
[14] Ian Goodfellow, Yoshua Bengio, Aaron Courville, Deep Learning ,The MIT Press,2016
[15] https://web.archive.org/web/20161019192412/http://lisp.vse.cz/pkdd99/berka.htm
[16] J.O. Awoyemi, A. O. Adetunmbi, S. A. Oluwadare, "Credit card fraud detection using machine learning techniques: A comparative analysis", 2017 International Conference on Computer Networking and Informatics (ICCNI), pp.1-9, 2017.
[17] Planning Commission. TPDS Definition. url: https://www.gktoday.in/gk/targeted- public-distribution networks

[18] CH Shah. Programme Evaluation Organisation, Planning Commission" Evaluation Report on First Year'sWorking of Community Projects"(Book Review)". In Indian Journal of Agricultural Economics 9.2 (1954), p.54
[19] Travis Oliphant, NumPy: A guide to Numpy, USA: Trelgol Publishing.
[20] Scikit-learn: Machine Learning in Python, Pedregosaetal, JMLR12, pp.2825-2830, 2011.
[21] J.D.Hunter,"Matplotlib: A2D Graphics Environment", Computing in Science & Engineering, vol.9, no.3, pp.90-95, 2007.
[22] Clifton Phua, Vincent C.S.Lee, Kate Smith Miles and Ross W. Gayler, "A Comprehensive Survey of Data Mining-based Fraud Detection Research"', CoRR, abs/1009.6119, 2010.
[23] S.Bhattacharyya, S.Jha, K.Tharakunnel and J.C.Westland, Data mining for credit card fraud: A comparative study,"Decis. Support Syst., vol.50,no.3, pp.602-613,2011.
[24] K.Randhawa, C.K.Loo, M.Seera, C.P.Lim, and A.K.Nandi, Credit Card Fraud Detection Using Ada Boost and Majority Voting, "IEEE Access,vol.6,pp.14277- 14284,2018
[25] M.Seera, C.P.Lim, K.S.Tan, and W.S.Liew, Classification of transcranial Doppler signals using individual and ensemble recurrent neural networks, "Neurocomputing, vol.249, pp.337-344
[26] Adrian and Banarescu, "Detecting and Preventing Fraud with Data Analytics", Emerging Markets Queries in Finance and Business 2014, EMQFB2014, 24-25October 2014, Bucharest, Romania.
[27] Lynnette Purda and David Skillicorn, "Accounting Variables, Deception, and a Bag of Words: Assessing the Tools of Fraud Detection", contemporary research, volume32, issue3, Pages: 815-1318, 2015.
[28] EWT Ngai, Y Hu, YH Wong, Y Chen, X Sun - Decision support systems, "The application of data mining techniques in financial fraud detection: A classification frame- work and an academic review of literature", Decision support systems, 2011-Elsevier Volume50, Issue3, February 2011, Pages 559-569.
[29] Zhou H, Chai H. & Qiu M. Frontiers Information Technologies Electronic Engg. (2018)19:1537
[30] Jurgovsky J, Granitzer M, Ziegler K, et al., 2018. Sequence classification for credit-card fraud detection. Expert Syst Appl, 100:234-245.
[31] Veronique Van Vlasselaer, Cristian 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 SupportSystemsVolume75,July2015,Pages 38
[32] Johannes Jurgovsky, Michael Granitzer, Konstantin Ziegler, Sylvie Calabretto, Pierre-Edouard Portier, Liyun He-Guelton, and Olivier Caelen, "Sequence classification for credit-card fraud detection", Expert Systems with Applications Volume 100, 15 June 2018, Pages 234-245.
[33] Shantanu Rajora, Dong-Lin Li, , Chandan Jha, Neha Bharill, Om Prakash Patel, Sudhanshu Joshi, DeepakPuthal and Mukesh Prasad, "A Comparative Study of Machine Learning Techniques for Credit Card FraudDetection Based on Time Variance", IEEE Symposium Series on Computational Intelligence (SSCI), 2018,Pages1958-1963.
[34] Y. Bouzembrak, B. Steen, R. Neslo, J. Linge, V. Mojtahed and H.J.P. Marvin, \Development of food fraud media monitoring system based on textmining", Food Control, Volume 93, November 2018, Pages 283-296
[35] Hans J.P. Marvin, Esmee M. Janssen, Yamine Bouzembrak, Peter J.M. Hendriksen, and Martijn Staats, Big data in food safety: An overview", Critical Reviews In Food Science And Nutrition 2017, Vol.57, No.11, 2286-2295.
[36] https://www.un.org/sustainabledevelopment/poverty/
[37] https://doi.org/10.1016/B978-0-12-381479-1.00009-5
[38] Maja Puh and Ljiljana Brkić, “Detecting Credit Card Fraud Using Selected Machine Learning Algorithms “International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 20-24 May 2019, DOI:10.23919/MIPRO.2019.8757212, Opatija, Croatia.
[39] Vaishnavi Nath Dornadula and S Geetha, “Credit Card Fraud Detection using Machine Learning Algorithms”, International Conference on Recent Trends in Advanced Computing 2019, India, 2019
[40] Doaa Hassan, "The Impact of False Negative Cost on the Performance of Cost Sensitive Learning: A Case Study in Detecting Fraudulent Transactions", International Journal of Intelligent Systems and Applications, Vol.9, No.2, pp.18-24, 2017.
[41] Hamza O. Salami, Ruqayyah S. Ibrahim, Mohammed O. Yahaya,"Detecting Anomalies in Students' Results Using Decision Trees", International Journal of Modern Education and Computer Science, Vol.8, No.7, pp.31-40, 2016.
[42] Sandeepkumar hegde, Monica R Mundada, "Enhanced Deep Feed Forward Neural Network Model for the Customer Attrition Analysis in Banking Sector", International Journal of Intelligent Systems and Applications, Vol.11, No.7, pp.10-19, 2019.
[43] Ch.Suresh, K.Thammi Reddy, N. Sweta,"A Hybrid Approach for Detecting Suspicious Accounts in Money Laundering Using Data Mining Techniques", International Journal of Information Technology and Computer Science, Vol.8, No.5, pp.37-43, 2016.
[44] Sunil Kappal, "Deplyoing Advance Data Analytics Techniques with Conversational Analytics Outputs for Fraud Detection", International Journal of Mathematical Sciences and Computing, Vol.5, No.1, Pp.42-52, 2019.

Authors' Profiles

Dr. Anil Kumar K.M is currently working as Associate Professor, Department of Computer Science & Engineering, JSS Science and Technology University, Mysuru, Karnataka, India. He did his post doc from Deakin University under Professor Jemal Abawajy and Ph.D. from University of Mysore under the supervision of Prof. Suresha, Chairman, DOS in Computer Science. He has teaching experience of 20 years and research experience of 12years. His research interest includes Text mining, Sentiment Analysis, Datamining, Opinion mining, Web Mining, Data Analytics, Computer Networks, Cyber Security. He has received 5 grants from different Government and Private funding agencies for Research & Development. He has published nearly 39 Research papers in National and International proceedings.

Bhargava S is currently working in Hewlett Packard Enterprise, Bengaluru as a Software Development Engineer. He graduated from SJCE in 2019. He worked with Dr. Anil Kumar for the conducting experiments and documented results for this research paper as a part of his interest in research in Machine Learning and Data analytics.

Apoorva R is currently working in Applied Materials, Bengaluru as a Software Development Engineer. She graduated from SJCE in 2019. She worked with Dr. Anil Kumar for the conducting experiments and documented results for this research paper.

Dr. Jemal H. Abawajy is a full professor at School of Information Technology, Faculty of Science, Engineering and Built Environment, Deakin University, Australia. He is currently the Director of the Parallel and Distributing Computing Tutorial. He is a Senior Member of IEEE Computer Society; IEEE Technical Committee on Scalable Computing (TCSC); IEEE Technical Committee on Dependable Computing and Fault Tolerance and IEEE Communication Society.

He is actively involved in funded research supervising large number of PhD students, postdoctoral, research assistants and visiting scholar in the area of Cloud Computing, Big Data, Network and System Security, Decision Support System, and E-healthcare. He is the author/co–author of five books, more than 250 papers in conferences, book chapters and journals such as IEEE Transactions on Computers and IEEE Transactions on Fuzzy Systems. He also edited 10 conference volumes. More info at http://www.deakin.edu.au/~jemal

How to cite this paper: Anil Kumar K.M, Bhargava S, Apoorva R, Jemal Abawajy, "Detection of False Income Level Claims Using Machine Learning", International Journal of Modern Education and Computer Science(IJMECS), Vol.14, No.1, pp. 65-77, 2022.DOI: 10.5815/ijmecs.2022.01.06