IES601: MSc-PhD Seminar Report

AI Driven Credit Scoring Model For Farmers

Aakash Roy | 21i190007

Supervisor: Prof. Usha Anantakumar

Industrial Engineering and Operations Research
Indian Institute of Technology, Bombay

Spring, 2021-2022

Contents

Chapter 1

Introduction

Credit scoring for farmers is a lot more different and demanding when compared to credit scoring of financial institutions or businesses. Challenges are many due to low level of financial inclusion of farmers and their limited access to credit. This is compounded by lack of structured data and related information with respect to farmers. This project aims to address this problem by first identifying relevant data that can help in coming up with a road-map of quantifying their credit risk. Subsequently, appropriate machine learning techniques will be utilized to come up with a credit scoring model for farmers. This study is also expected to provide insights on promoting financial inclusion to farmers. We will first introduce the problems and then we will go through the literature survey and after that my project works till now.

1.1 Major Challenges

- **Data Verification:** The data provided by them is valid or not is one of the biggest concerns and this needs to be verified by other available relevant sources.
- **Natural Disaster:** This is tough to overcome but can be mitigated by using meteorological prediction.
- **Misuse:** One could misuse by taking advantage of various schemes. This needs check.
- **Digitization:** Though digitization is a boon, adapting to new technologies may be tough for farmers in the initial phase.
- **Lack of education:** Lack of education may be a major impediment. This requires continuous training.

1.2 Broad Data Heads

To come up with a credit score, we need to understand various aspects about the farmers and related details. For that, we need information about the following:

- **Demographics**
- **Farming details**
- **Capital Strength**

Chapter 2

Literature Survey

Traditional financial institutions are evaluating the creditworthiness of the borrowers based on subjective methods focusing mainly on the 5Cs: character, capacity, collateral, capital, and conditions. This method largely is unable to assess the borrowers who have no loan history and have limited banking transactions, particularly the customers residing in rural areas. In addition, this method fails to provide a comprehensive profile of a potential borrower as there may be a chance of missing the vital and relevant information required for the credit assessment. Hence, banking institutions gradually started following the digital method of credit assessment to fulfill the requirement of potential borrowers with their exact loan eligibility and also to minimize their non-performing loans. Under the digital methods for credit assessment, two distinct lines are bifurcating between the standard econometric model-like models based on logistic regression and ML-based models. ML-based models are mainly categorized into the five broad areas: generalized line models (most basic including ordinary least square method and logistic regression), Bayesian models, ensemble models, support vector machines (SVM), and nearest-neighbor models.

2.1 Solution Approaches

- **Logistic Regression**
- **KNN: K Nearest Neighbours**
- **Ensemble Methods**

2.2 Obtained Results

Credit scoring is a classification problem between the identification of defaulters and non-defaulters. Both SVM and RF are black-box models and are sensitive to hyperparameters. Researchers proposed a modified harmony search random factor that is more robust in terms of performance, explain ability, and computational time. However, it is not completely known that AI-ML algorithms will not cause bias especially against minorities like small and marginal holdings, women communities, etc., Hence, to reduce the impact of human biasness it is necessary to train the data appropriately so that more robust AI-ML algorithms can be formed. Referring to the below table of the comparative analysis of credit scoring techniques, it can be stated that the hybrid model (whether it is AI-ML with AI-ML-based or AI with any other method of credit scoring) could be the best fit for credit score assessment, whereas logistic regression

yields a lesser impact on credit scoring. This analysis has been performed by assigning the weightage (sum of weightage is one) to the below given four parameters: accuracy (0.3), performance (0.3), robustness (0.2), and volume of data (0.2). This weightage has been assigned on the basis of the importance of these parameters. Accuracy and performance are the two main features of any ML-based model, then comes the almost equal weightage of robustness and the size of the data handling. The rating has been assigned from 1 (very low performance) to 5 (very high performance), depending on how these models performed and what the results have been interpreted as after having the empirical studies of the existing literature.

Table : Comparative analysis of credit scoring techniques

Parameters	Comparative Analysis—Credit Scoring Techniques												
	Weights	ANN		SVM		RF/XG Boost		Logistic Regression		GA		Hybrid Model	
		Rating	Score	Rating	Score	Rating	Score	Rating	Score	Rating	Score	Rating	Score
Accuracy	0.30	4	1.2	4	1.2	5	1.5	3	0.9	4	1.2	4	1.2
Performance	0.30	4	1.2	3	0.9	5	1.5	3	0.9	4	1.2	4	1.2
Robustness	0.20	3	0.6	3	0.6	3	0.6	3	0.6	4	0.8	5	1
Volume of Data	0.20	3	0.6	3	0.6	3	0.6	2	0.4	3	0.6	5	1
Total	1.00		3.6		3.3		4.2		2.8		3.8		4.4

2.3 Paper Discussion

2.3.1 Applications

In India, small and marginal farmers have the highest proportions among the farming community and being the most heterogeneous group, they have varied, vast, and fragmented farm data. Therefore, AIML-based algorithms for credit scoring show the way for guiding their credit eligibility check in the shortest computational time with perhaps higher accuracy.

2.3.2 Key Takeaways

Majority the methods that are discussed in the paper are applicable only when there is label data available and if so, then we can use the methods to improve the accuracy as well as the overall performance of the model. But in our case this is not the case. Since, credit score to farmers have not been there in India, that's why the data we've collected is not labeled i.e. credit scores are not assigned to the farmers. Hence, we can't directly use the methods discussed in the paper but, once we're done with assigning the credit scores to the respective farmers, we can then work on improving our model with the help of the methods discussed in the paper.

Chapter 3

Project Works

Continuing the 'Key takeaways', the data could be analysed to look for similar behaviour of farmers by applying clustering technique. This will help in getting insights about offering different loan offers for different groups of farmers. This eventually can be used to update their credit score.

3.1 Data Collection

Collecting the data was the biggest challenge for us. To start with modelling, We have collected data of 53 farmers belonging to Maharashtra, Tamil Nadu and Uttar Pradesh.In total we've collected 38 features. We can't show the dataset as it is confidential. All the data collected includes the following:

3.1.1 Demographics:

a. First Name b. Last Name c. Aadhar No. d. Pan No. e. Phone f. Marital status g. Genuineness of the person while interacting on a scale of 1-10[10: Very genuine, 1: Not at all genuine] h. Family Size i. Experience as a farmer j. City/District k. State l. Pin m. Address n. Years of stay

3.1.2 Farming Details

a. Land area b. Soil Type c. Land location d. Major Crops grown e. Quantity produced for each crop f. Selling price for each crop g. Water access on a scale of 1-10(10-Very good availability, 1-Poor availability h. Most used Machineries

3.1.3 Capital Strength

a. Monthly income b. Average Bank balance in last year c. No. of times loans taken d. Amount of loan taken each time e. Repayment duration of loan each time(in months)[Pending: if not repaid] f. Purpose of taking loan choose the appropriate no. [1. Farming, 2.Children's education, 3.Children's marriage, 4.Health issues of self, 5.Health issues of family members, 6.Debt, 7.other(Please specify)] g. Available assets valuation (in lacs)[including their own house, own land own ornaments in stock, cash] h. Sold any asset in the

past 1 year: yes/no i. If yes in the previous column, then which asset? j. If yes in the previous column, then which asset? k. Valuation of the asset sold recently l. Name of partners m. Age of respective partners n. Duration of partnerships o. How healthy is the farmer (1-10) [1: Very unhealthy, 10: Very healthy] p. Other income sources (if any) q. The amount of average monthly support from other sources r. Average Monthly expenditure

3.2 Data Preprocessing

Since the data we have collected was raw data, it needed to preprocessed for making it eligible to run the ML algorithms.

3.2.1 Taking care of missing data

We removed some rows containing more that 5 empty features and in case the number is less that 5, we replaced them with average values. In some cases we converted Nan to 0. We also removed the following columns 'First Name', 'Aadhar No.', 'Pan No.', 'Phone','Repayment duration of loan each time(in months)[Pending: if not repaid]','Yield','Address','Major Crops grown', 'Purpose of taking loan','Major Crops grown', 'Name of partners','Sold any asset in the past 1 : yes/no' etc. as these have no impact on the model.

3.2.2 Encoding categorical data

Since we can't work with categorical data like 'strings', we need to convert those to sum numerical values. We encoded the following columns 'Last Name', 'Marital Status', 'City/District', 'State', 'Soil type', 'Land Location', 'Major Crops Grown', 'Most used machineries' using 'OneHotEncoder' included in the 'scikitlearn' library.

3.3 Applying Different ML Algorithms for clustering

3.3.1 Finding the optimal number of clusters

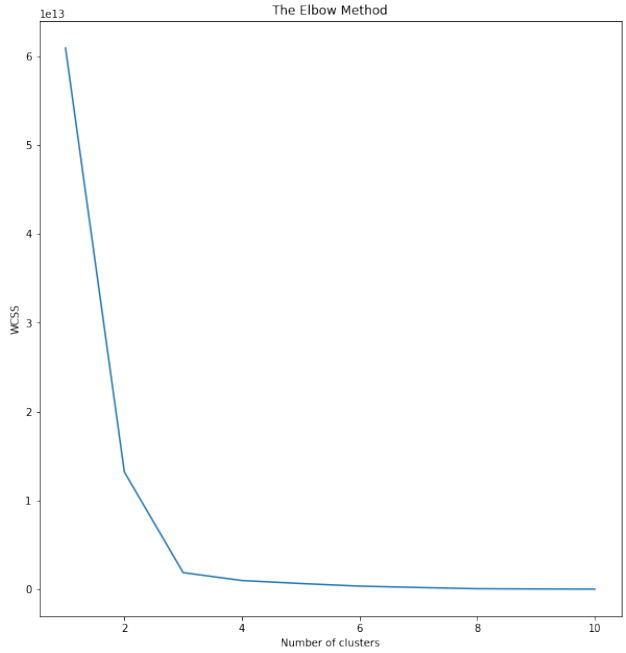
Now, our goal was to determine the optimal number of clusters. For that we used the two methods as follows:

Elbow Method This method calculates the Within-Cluster-Sum of Squared Errors (WSS) for different values of k(no. of clusters), and choose the k for which WSS becomes first starts to diminish. In our case, we used the following code:

```
wcss = []
for i in range(1, 11):
    kmeans = KMeans(n_clusters = i, init = 'k-means++', random_state = 42)
    kmeans.fit(X)
    wcss.append(kmeans.inertia_)
plt.plot(range(1, 11), wcss)
plt.title('The Elbow Method')
plt.xlabel('Number of clusters')
```

```
plt.ylabel('WCSS')
plt.show()
```

This resulted the following graph:

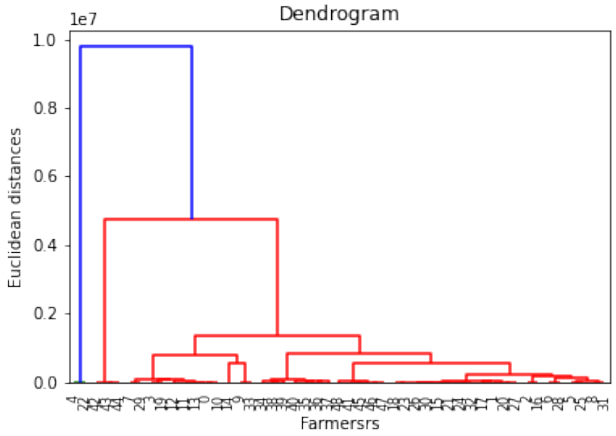


As the change of the wcss value was linear after k=4, we choose the no. of clusters to be 4 in case of K Means Clustering.

Dendrogram Method A dendrogram is a type of tree diagram showing hierarchical clustering — relationships between similar sets of data. They are frequently used in biology to show clustering between genes or samples, but they can represent any type of grouped data. In our case, we used the following code:

```
import scipy.cluster.hierarchy as sch
dendrogram = sch.dendrogram(sch.linkage(X, method = 'ward'))
plt.title('Dendrogram')
plt.xlabel('Farmersrs')
plt.ylabel('Euclidean distances')
plt.show()
```

This resulted the following plot:



From the above plot it is clear that there are mainly 4 groups forming. Hence we choose the no. of clusters to be 4 in case of Hierarchical Clustering.

3.3.2 K Means clustering

K-means is a centroid-based algorithm, or a distance-based algorithm, where we calculate the distances to assign a point to a cluster.each cluster is associated with a centroid. This algorithm tries to minimize the distance of the points in a cluster with their centroid. In our case we imported 'K Means' from the 'Scikitlearn' library. We used the following code:

```
from sklearn.cluster import KMeans
kmeans = KMeans(n_clusters = 4, init = 'k–means++', random_state = 42)
y_kmeans = kmeans.fit_predict(X)
```

3.3.3 Hierarchical Clustering

In this algorithm, we develop the hierarchy of clusters in the form of a tree, and this tree-shaped structure is known as the dendrogram. The hierarchical clustering technique has two approaches:

- **Agglomerative** It is a bottom-up approach, in which the algorithm starts with taking all data points as single clusters and merging them until one cluster is left.

- **Divisive** Divisive algorithm is the reverse of the agglomerative algorithm as it is a top-down approach.

In our case, we used 'AgglomerativeClustering'by importing it from the 'Scikitlearn' library. We used the following code:

```
from sklearn.cluster import AgglomerativeClustering
hc = AgglomerativeClustering(n_clusters = 4, affinity = 'euclidean', linkage = 'ward')
y_hc = hc.fit_predict(X)
```

3.4 Results and Analysis

Since we had run the codes of 'K Means clustering' and 'Hierarchical clustering' by taking number of clusters to be 4, the clusters created for both of them are Cluster 0, Cluster 1, Cluster 2, Cluster 3. Next we wanted to see that if same farmers are assigned to different clusters or not. But, both the algorithms did the exact. This might be because of the small number of data. Farmer's index and their assigned clusters in case of K Means and Hierarchical clustering is given in the next page.
By analysing the results deeply, we saw that 1. 'The quantity produced by each crop', 2.'Surname', 3. 'Selling price for each crop', 4. 'Land Area' are the main factors that are motivating the clusters.

Table 3.1: Clusters assigned to farmers

Farmer index	**K Means**	**Hierarchical**
1	3	3
2	0	0
3	0	0
4	3	3
5	1	1
6	0	0
7	0	0
8	3	3
9	0	0
10	3	3
11	3	3
12	3	3
13	3	3
14	3	3
15	3	3
16	0	0
17	0	0
18	0	0
19	0	0
20	3	3
21	0	0
22	0	0
23	1	1
24	0	0
25	0	0
26	0	0
27	0	0
28	0	0
29	0	0
30	3	3
31	0	0
32	0	0
33	0	0
34	3	3
35	0	0
36	0	0
37	0	0
38	0	0
39	0	0
40	0	0
41	0	0
42	0	0
43	2	2
44	2	2
45	2	2
46	0	0
47	0	0
48	0	0
49	0	0

Chapter 4

Conclusion and Future Work

By analysing the data using appropriate machine learning techniques, we can come up with an efficient credit scoring model that assigns a credit score to farmers or rank the farmers as per their potential to repay the loan thus mitigating the risks involved in lending. Currently, we're working on ranking the clusters and trying to fit a range of credit score to each clusters and later on we will go more deep into the clusters i.e. assigning a credit score range to individual farmers.

END OF REPORT

Bibliography

- **Anil Kumar K, Suneel Sharma, M. Mahdavi. 2021. Machine Learning (ML) Technologies for Digital Credit Scoring in Rural Finance A Literature Review** [https://www.mdpi.com/2227-9091/9/11/192/htm]