I. J. Computer Network and Information Security, **2026, 2, 94-111**
Published Online on April 8, 2026 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijcnis.2026.02.06

Assessment of Social Capital from Mobile Communication Data: A Cascade Model Based on Random Forest and Logistic Regression

Irada Y. Alakbarova
Institute of Information Technology, Ministry of Science and Education of Azerbaijan, Baku, Az1141, Azerbaijan
E-mail: airada.09@gmail.com
ORCID iD: https://orcid.org/0000-0002-9876-3035

Received: 16 November 2025; Revised: 18 January 2026; Accepted: 26 February 2026; Published: 08 April 2026

Abstract: With the rapid development of mobile technologies, analyzing data generated by mobile devices is becoming increasingly important. A wide range of applications, from marketing to healthcare, require the development of effective methods for extracting valuable information from this data. This study is devoted to developing a methodology for assessing an individual's social capital based on the analysis of mobile communication data. To assess social capital, we propose a two-stage Cascade Model that combines the advantages of the Random Forest (RF) and Logistic Regression (LR) algorithms. In the first stage, RF is used to select the most significant features reflecting various aspects of social capital. In the second stage, LR is used to assess of social capital, taking into account nonlinear relationships between features. The results of the study open up new opportunities for studying social phenomena and can be used in as sociology, marketing, and urban planning.

Index Terms: Social Capital, Mobile Communication Data, Random Forest, Logistic Regression, Cascade Model.

1. Introduction

In today's world, mobile devices generate vast amounts of data every day, providing valuable insights into people's behavior, interests, and communication habits. Mobile devices have created convenient interfaces for accessing applications, fostering strong social relationships, and maintaining social capital [1]. Mobile devices are portable technologies that can be carried anywhere with network access. Mobile devices include cell phones, smart phones, tablets, smart watches, and even some cars.

These devices have become an integral part of modern life, collecting vast amounts of data on people's behavior [2]. Every interaction, whether it is a phone call, SMS, app use, or location update, provides an opportunity to look into people's digital lives. Such data can be used for a variety of purposes, from understanding a person's mobility and social behavior to enhancing user experience. However, there is little research on how mobile data is used to analyze social capital, as most studies have considered mobile services only in the context of telecommunication. With the growing popularity of mobile communication services (MCS), the relevance of developing methods for studying the impact of MCS on the formation of social capital has increased [2]. It is important to analyze not only the content of messages but also user demographics, interaction patterns, and network connections to determine each person's social capital.

Social capital, a term popularized by sociologists Pierre Bourdieu, James Samuel Coleman, and political scientist Robert David Putnam, refers to the resources and advantages that people can access through their social networks. The concept of social capital encompasses trust, mutual support, and the ability to mobilize resources through connections [3]. The nature and characteristics of social capital, as well as its role in society, were first studied in the 1980s by the French sociologist, philosopher, ethnologist and political journalist P. Bourdieu [4]. He viewed social capital as a network-based resource essential for improving socio-economic status and achieving upward mobility. In his concept of "Forms of Capital", P. Bourdieu argues that capital exists in three main forms: financial capital, social capital and cultural capital. In this concept, social capital is presented as a real and potential resource that can only be obtained if there are stable networks. The amount of social capital that an individual possesses depends on the density of relationships and the exchange of capital in a group [5].

Professor of the University of Chicago and President of the American Sociological Association J. S. Coleman pays special attention to information resources in his research and notes that information is an important resource of social

This work is open access and licensed under the Creative Commons CC BY 4.0 License.

capital and is activated through social relations [6]. Harvard University professor R. D. Putnam notes that social capital depends on the interactions of people in society with each other and that social capital is one of the main factors of economic development and well-being and is based on trust between people. [7].

Research has shown that social capital refers to the networks, relationships, and mobile connections that provide people with access to resources, information, and opportunities. Mobile communications are technologies that allow various devices to connect to the Internet and communicate with each other wirelessly. With mobile communications, people can call and message anyone anywhere, receive notifications of new messages, calls, and events, use the full capabilities of the Internet, and use Internet applications for various purposes. Mobile communications are used to communicate with others, creating, changing, increasing, and sometimes even decreasing social capital [8]:

- Mobile connections make social capital accessible. Mobile connections increase informativeness, increase collective trust, and develop cooperation.
- Mobile connections change social capital. They provide a means for people with similar interests and ideologies to get to know each other and crea te new groups around their interests.
- Mobile connections increase social capital. The number of social connections increases and becomes denser. People not only create new online relationships, they also import these relationships. As a result, virtual social capital is created.
- Assessing social capital involves measuring the extent and quality of an individual's relationships and their ability to access resources through these relationships.

The evolution of mobile communications offers an unprecedented window into individuals' social interactions, providing data that can be analyzed to determine levels of social capital. This paper aims to explore how mobile contact data can be used to determine social capital. The key objectives of this study are:

- Determine the role of social capital in the modern world.
- Identify the most relevant features reflecting various aspects of social capital.
- Assess social capital based on mobile contact data using a Cascade Model with Random Forest (RF) and Logistic Regression (LR).
- Comparative analysis of the Cascade Model (RF and LR) with some advanced machine learning models.

2. Related Works

Early studies on assessing social capital focused on the level of authority in society, social relationships, and trust. These indicators were regarded as measures of social capital [9]. In the modern world, social capital depends on the influence of information and communication technologies (ICT). Yang et al. [10] argue that ICT services are used to form and maintain social capital. Researching what factors influence the decision to choose specific ICT services helps to understand the motives of users and why they use these services.

In [11], social capital is studied mainly based on short messages (SMS) and voice messages. The authors argue that both SMS and voice messages are used to bond social capital, however, SMS is used more for bonding purposes than voice messages. The authors also present a new method for analyzing social capital in the context of different MCS. This paper presents a new method for studying social capital in the context of MCS and demonstrates its results in a pilot study. The proposed approach combines principles of quantitative analysis of mobile communication data and social network analysis.

Even though ICT has a strong impact on increasing social capital, it is believed that the main approach to solving problems of quantitative assessment of social capital remains the methods of survey analysis [12]. However, these methods require a lot of attention, are expensive, and are not effective in analyzing big data. Also, these data in most cases do not cover all the information about a person, raise doubts about the correctness, and limit the scalability of the proposed approaches. The authors recommend preferring the "social capital generators" method, which more accurately reveals network resources [13]. Each person has unique characteristics that affect their behavior and social capital. In [14], the authors argue that people with approximately similar personal profiles have the same behavior and habits. It is proposed to implement two machine learning models for behavior assessment: Decision Tree and RF. According to the applied assessment metrics, the models demonstrated good forecasts.

In [15] the authors argue that ML (machine learning) methods such as Logistic Regression (LR), Support Vector Machines (SVM), Artificial Neural Networks (ANN), and Decision Trees (DT) have demonstrated remarkable capabilities in solving complex tasks by analyzing high-dimensional and big data. One such task is analyzing social capital and Credit Scoring classification based on demographic data, where the authors propose a comprehensive mathematical model using algorithms such as particle swarm optimization (PSO) and others. In [16], the authors propose machine learning methods for classifying social media data based on mobile device data. They classified the data into 10 classes to demonstrate the effectiveness of using machine learning for data analysis. Testing and evaluating these models showed that the proposed models are able to accurately classify the data. The article [17] states that the sharp increase in the number of social media users leads to big data that can be used to analyze people's behavior and

opinions. Using a new classification model, election results are predicted based on tweets.

In [18] the authors propose a new approach called "Mobile Data Science". This approach involves collecting data from various sources and building models that make dynamic decisions based on this data using machine learning techniques. The goal explained is to create intelligent mobile applications, that is, an assistant system with artificial intelligence that better understands the needs of users and helps them. The study explains the main elements of mobile data science and the various modules involved in the process (for example, modules for data collection, processing, and application).

In [19], the authors used statistical analyses (e.g., correlation or regression analyses) to examine the links between mobile phone usage (voice calls and text messages), geographic proximity patterns, and social capital. The analysis used indicators such as age, gender, education, race, household income, voice calls, text messages. Studies have shown that mobile phone call data is a source of information that provides valuable information not only about technical parameters, but also about the social and personal characteristics of users [20].

In [21], a multilevel cyber-physical-social system (CPSS) is proposed for the analysis, modeling, and management of mobile network data (in particular, Call Detail Records). The authors claim that through the proposed CPSS system (in the case of the city of Milan) it is possible to accurately identify the behavior of network users and the areas where the Internet is most frequently used.

In [22], another study focuses on mobile data science and intelligent applications. Researchers presented a concept for developing intelligent mobile applications using artificial intelligence (AI) technologies on smartphones. In [23], a review is conducted examining various methods and approaches available for estimating and predicting human behavior using mobile phone data. The study is a comprehensive literature review covering 148 scientific papers published between 2013 and 2021 and selected from eight major scientific databases. The aim of the study is to provide an overview by analyzing the strengths and weaknesses of existing scientific methods in the field of human behavior modeling based on mobile phone data.

Another study on social capital and mobile communication data examines how different smartphone usage patterns affect the level of social capital among Chinese college students, and identifies the important moderating role of communication skills in this relationship [24]. The results show that students with high frequency and balanced usage patterns have higher social capital, especially if they have high communication skills. The study used Cross-sectional Survey, Latent Class Analysis, and Multiple Regression Analysis methods.

This study examines how mobile social networking sites affect the accumulation of social capital. The primary aim is to determine how the use of WeChat, China’s most popular mobile social networking sites, influences individuals’ social trust [25]. The research also investigates whether the benefits of WeChat use are distributed equally across different demographic groups (age, education level, and place of residence). The analysis utilizes statistical analysis methods (regression analysis and mediation models). The study's data is sourced from the Chinese General Social Survey.

The article [26] examines how the use of digital technologies (e.g., the Internet, smartphones, social media) among older adults affects their levels of social capital. Given the existing evidence that social capital declines with age, the article examines how digital technologies can halt or replace this decline. The research was a systematic review, which involved collecting, evaluating, and summarizing all relevant studies according to established guidelines. The study found a positive association between digital technology use and social capital among older adults.

Research has shown that there is no universal metric for measuring social capital. Different studies use different indicators, which makes it difficult to create a single scale. Research that has been conducted primarily using machine learning methods in recent years typically aims to understand the dynamics of social relationships and their implications for marketing, public health, and network optimization. We conducted studies also revealed that:

- Most studies measuring social capital are based on surveys (self-reports). Such an approach is susceptible to the memory bias and subjective assessment of the respondent.
- Surveys are usually limited to small or medium sample sizes. This makes it difficult to generalize the results.
- Traditional statistical methods (regression) are sometimes weak in capturing complex nonlinear relationships. The best ML models are very slow and are frequently characterized by high computational costs.
- Cross-sectional surveys measure social capital only at a certain point in time (static).

To mitigate memory bias and subjectivity in the assessment of social capital, it is methodologically sounder to utilize real and automatically recorded behavioral data (such as mobile communication parameters, e.g., call frequency and duration) rather than relying on individuals' subjective statements (self-reports). Given that mobile network data inherently encompasses millions of subscribers, the efficient analysis of these Big Data sets necessitates meeting the methodological requirements of achieving simultaneously high predictive accuracy and high computational speed.

3. Methodology

The main dataset analyzed in this paper was collected using mobile phone-based measurements. These measurements were implemented using data collection software installed on the participants’ mobile phones with their

consent – informed consent for data collection was also required from all participants. During this study, data was collected from 40 participants over one month. To protect the privacy of the participants, the phone numbers were encrypted.

3.1. Data Privacy and Ethics

In the assessment of social capital using mobile communication data, the use of sensitive data such as call frequency, duration, and personal identifiers (such as name, age, position, and workplace) requires extensive ethical deliberation. The following principles were followed to ensure the methodological and ethical integrity of the study:

- Despite the anonymization of mobile communication data, unique behavioral parameters such as call frequency, duration, and geographic location pose a risk of subject re-identification. Appropriate technical and organizational measures were implemented with the informed consent of all employees and management to minimize this risk.
- Considering that analytical results such as the social capital score may lead to discrimination or dissatisfaction among employees, the research results were designated for strictly academic purposes only, and a specialized coding system was employed in the database in lieu of employee names.
- Responsible parties for the collection, processing, and protection of mobile communication data were formally designated. A clear and concise data protection report was prepared, detailing the type of data collected, the purpose of processing, the storage period, and retention protocols.
- The collection, processing, and protection of mobile communication data were conducted in accordance with the Law of the Republic of Azerbaijan "On Personal Data" [27]. Data processing was mandated to occur only based on the subject's explicit and freely given consent, and was only performed when serving legitimate public purposes clearly justified by relevant legislation.
- The analysis of mobile communication data was strictly limited to predetermined purposes, and the sale of this data for commercial purposes or any usage incompatible with the original purpose is strictly prohibited.
- To protect the security and integrity of mobile communication data, appropriate organizational and technical security measures (including encryption, access monitoring, and data breach notification procedures) were implemented across all stages of the data collection, transmission, and storage processes.

Adherence to these principles renders the use of mobile communication data for measuring employees' social capital ethically and legally justified.

3.2. The Analytical Part of the Study

The analytical part of the study is structured in five sequential stages to ensure data quality and model accuracy:

- Identification of key metrics.
- Data preprocessing.
- Text feature encoding.
- Model building.
- Model evaluation.

A. Identification of Key Metrics

The mobile user dataset is categorized into two areas:
Mobile contact data. The dataset includes:

- Call logs: records of outgoing and incoming calls with timestamps, including call duration and frequency.
- Message logs: the number of text messages sent and received, including the frequency and content of conversations (if available).
- Contact lists: the number of contacts stored on the device, including the frequency of communication with each contact.

Demographics can provide valuable context for understanding mobile user behavior. The following metrics are considered: education, age, residence, gender, location, annual income, marital status, and employment.

B. Data Preprocessing

Data preprocessing was performed to ensure the quality and reliability of the analytical results. In the process of data processing, one of the main tasks is data cleaning. Messages from mobile devices were translated into English using the Google Translate API. The collected text was then subjected to standard preprocessing techniques, including tokenization, removal of duplicates, and stop word removal [28].

Tokenization means breaking the text into individual words or subwords (tokens). Algorithms are used to detect and remove records containing the same information. The most frequently occurring words that do not carry a

significant semantic load are excluded from the text (for example, “the”, “and”, “of”, “in”) [29]. In machine learning, scaling is a method used to transform the features of a dataset to a common scale [30].

To prevent features with larger ranges from dominating the training process, the data was scaled. The Min-Max Scaler was applied to transform features to a common range, typically from 0 to 1. It is suitable for algorithms that require input features to be in a certain range:

$$X_{scaled} = \frac{\mathrm{X} - \mathrm{X}_{\min}}{X_{\max} - X_{\min}} \tag{1}$$

where $X_{\min}$ – is the minimum value of the feature in the data set, and $X_{\max}$ – is the maximum value of the feature in the data set. If the data has a normal distribution, you can use the z-score. If the distribution is strongly skewed or contains outliers, it is better to use more robust methods.

C. Text Feature Encoding

Since the majority of machine learning algorithms operate exclusively with numerical data, categorical variables (e.g., gender, education level) or text data (e.g., text messages, survey comments) must be converted into a numerical format. In our research, we primarily utilize the TF-IDF (Term Frequency – Inverse Document Frequency) method for processing text data. TF-IDF is a powerful statistical measure that reflects how important a word is to a document in a collection or corpus. In this approach, each document is represented as a vector, where the value of each dimension corresponds to the calculated TF-IDF score of a specific term within that document. TF-IDF is particularly valuable for tasks involving document classification and information retrieval [31].

4. Model Building: A Cascade Model with RF and LR

The relationships between different aspects of social capital can be nonlinear and complex. Given these shortcomings, we propose a two-stage Cascade Model based on RF and LR methods: Cascade models are a sequential application of multiple machine learning models, where the output of one model becomes the input of the next [32].

Cascade learning is a powerful tool for deep learning in big data. The final output of such an approach is the result of the optimal computation of each stage [32, 33]. In the context of social capital assessment, a Cascade Model combining RF and LR combines the positive aspects of both models and can look like this:

- RF is a classification (or regression) algorithm that evaluates the importance of features based on how they affect the accuracy of predictions. RF's feature importance metrics, derived from its decision tree structure, are used to evaluate the contribution of each feature to the prediction accuracy of social capital.
- LR uses the output of RF as additional features and refines the predictions, taking into account more subtle relationships between the data. We get a more accurate assessment in determining social capital.

4.1. Building Random Forest

Using randomly selected features as input, we construct a random forest model and select the most relevant features for the social credit assessment task. RF provides measures of feature importance and can efficiently handle multidimensional data [34]. To construct RF, we first need to define the following hyperparameters:

- The number of decision trees (n_estimators) is determined. The default number of trees is 150.
- The maximum depth of trees is determined (trees are expanded until all leaves are pure (max_depth)).
- The minimum number of samples for splitting a node is determined. By default, the number of samples is set to 2 (min_sample_split). When the algorithm builds a tree, it looks for the best way to split the data at each node to maximize the purity of child nodes. The goal is to split classes (tree nodes) into two child nodes.
- The minimum number of samples for the leaf node is determined: set to 1 (min_samples_leaf).
- The maximum number of random features is determined (max_features), for the best split in classification problems, we take (the square root of the total number of features): $max_features = \sqrt{n}$, where n is the number of features. Each new division has its own random features.
- The best feature for splitting a node is determined (for example, the Gini index or information gain). We will use the Gini index method [35]. The purpose of calculating the index is to select an attribute that can maximally and quickly divide the data into homogeneous classes:

$$Gini = 1 - \sum_{i=1}^{J} (P_i)^2 \tag{2}$$

where P_i – probability of selecting objects of class i in a node $i \in \{1,2,\ldots,J\}$. If, $Gini = 0$ – this means that all objects belong to the same class.

The Random Forest Algorithm:

Step 1. Let the number of trees be: $n = 1,2,...,N$.(model = Random Forest Classifier (n_estimators=150)

Step 2. Train the model on a sample, where X is the input and y is the output. Using the bootstrap, we generate $X = \{x_1, x_2, ..., x_i\}$

Step 3. It is necessary to build a decision tree $b_n(x)$ by sample x_i . The decision tree is built until there are no more than n objects left in each leaf.

Step 4. We will train our classifier on each sample $b_n(x)$. A prediction result will be obtained from each decision tree. The feature weights in a random forest are determined during the construction of each tree.

Step 5. Defining the class to which object x belongs. We apply the model, which describes the principle of operation of the voting algorithm (Majority vote) [36]. The result with the largest number of votes is selected:

$$a(x) = \arg\max_{y \in Y} \sum_{n=1}^{N} [b_n(x) = y] \tag{3}$$

where $b_n(x)$ – decision tree: *n*-th tree prediction for feature value *x*, argmax – operator that selects the maximum value*y*, $\sum_{i=1}^{N} b_n(x)$ – the sum of the predictions of all trees for*x*. *N*– number of trees.

Step 6. Select the most important features: (importances = model.feature_importances_)

Step 7. Create a separate dataset for "visualization:final_df = pd.DataFrame{"Features":pd.DataFrame(X). columns,"Importances"}"

The tree-based strategies used by random forests are naturally ranked by how well they improve the purity of the model in terms of the data. Thus, by "prune" trees below a certain factor (using the Pre-pruning method), we can select the most important features. The importance of a feature is determined by how often that feature is used in the trees of the forest to split the data.

4.2. Training and Evaluation of the Quality of an RF Model

To evaluate the quality of the random forest model, the Confusion matrices table [37] is used. In the table, the parameters true positive (TP), false positive (FP), false negative (FN), and true negative (TN) determine the accuracy, recall, reliability, and F1-mertics. These metrics are used to evaluate the quality of the machine learning algorithms:

$$\mathrm{Re}\,call = \frac{TP}{TP + FN} \tag{4}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$\mathrm{Pr}\,ecision = \frac{TP}{TP + FP} \tag{6}$$

$$F1 - score = \frac{\mathrm{Re}\,call \times \mathrm{Pr}\,ecision}{\mathrm{Re}\,call + \mathrm{Pr}\,ecision} \tag{7}$$

F1-score is a harmonic mean between precision and recall. F1-score shows the balance between the number of correct positive predictions and the recall of all positive examples. If fewer people have high social capital, then high Accuracy can be misleading since the model can predict "low social capital" for everyone. For this reason, we will not take into account the Accuracy metric.

To assess the quality of the model, we calculate the area under the ROC curve. The AUC-ROC (Area Under the Curve – Receiver Operating Characteristic) metric is used to assess the quality of binary classification models. In other words, AUC-ROC (ROC Curve) is a metric that evaluates the ability of a classifier to distinguish between positive and negative classes [38].

The ROC curve is a graph that illustrates the performance of a classification model at all possible classification thresholds [39]. Often in their studies, the ROC curve is defined by authors as the TPR versus FPR curve [40]. The true positive rate (TPR) and false positive rate (FPR) are needed to calculate the model performance metric (AUC-ROC). TPR is known as Recall, and is defined as the proportion of correctly classified positives among all positives in the data

[41]. FPR defines the proportion of misclassified negatives relative to all negatives:

$$FPR = \frac{FP}{TN + FP} \tag{8}$$

To evaluate the quality of the RF algorithm, we calculate the AUC-ROC for the training and test parts of the model and implement the task using the colab.research.google in Python:

```
model = RandomForestClassifier(n_estimators=100, random_state=23)
model.fit(X, y) # train
a = model.predict(X2) # prediction
print ("AUC-ROC (oob) = ", roc_auc_score(y, model.oob_prediction_))
print ("AUC-ROC (test) = ", roc_auc_score(y2, a))
```

The “random_state” operator determines which features will be selected at each step. Different values of “random_state” result in different bootstrap samples and random feature selections, which increases the diversity of trees in the forest and improves the generalization ability of the model.

Each time we build a new tree in the random forest; we bootstrap the original data. This means that we take a random sample with replacement from the original dataset. In this process, some features may be sampled multiple times, while others may not be sampled at all. About 37% of the features from the original sample will not be sampled in any of the bootstrap samples.

Those features that are not sampled to build a given tree are called out-of-bag (oob) samples. Since each tree is built on about 63% of the original data, we have enough out-of-bag samples to estimate accuracy. For each “oob” sample, we run it through all the trees that did not train it. Then, using voting (for classification), we get a prediction for that sample. TPR is also called specificity (Fig. 1).

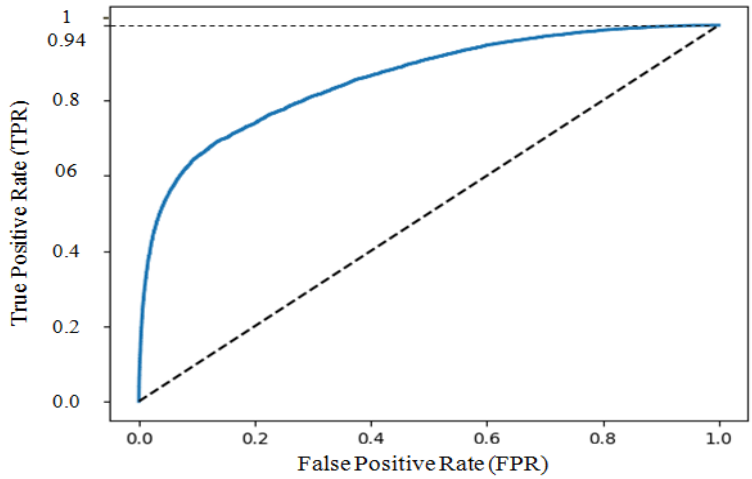


Fig.1. ROC curve for the performance of a classification model built on mobile user data. The area under the curve (AUC) reflects the overall performance of the model

An ideal classification model would tend to the point in the upper left corner of the graph where TPR is 1, and FPR is 0. To evaluate the quality of a feature, we will use the expression: $2 \times \left| AUC_ROC - 0.5 \right|$. An ideal classification model will tend to a point where TPR is 1, and FPR is 0. If the 2 classes are strictly separated by this feature, then the quality scores of the feature are maximal; if they are “mixed”, then the quality scores of the feature are minimal. Choosing the optimal number of trees in a random forest is one of the key parameters that influences the quality of the model. Too few trees can lead to under-fitting, and too many trees can lead to over-fitting.

Figure 2 is a graph showing the dependence of the ROC AUC metric on the n_estimators parameter. The graph shows how the model quality (ROC AUC) changes depending on the number of trees in the random forest. It can be seen that as n_estimators increase, the value of ROC AUC increases. This means that the model becomes better as the number of trees increases.

The growth of ROC AUC is most noticeable at the beginning when n_estimators increases from 0 to about 100. After that, the growth slows down, and the graph becomes flatter. Choosing the optimal value of n_estimators is a trade-off between model quality and computational resources. The more trees, the longer the model takes to train.

The graph shows that for a given data set; the optimal value of n_estimators is between 100 and 200. Further increase of n_estimators will not significantly improve the model quality.

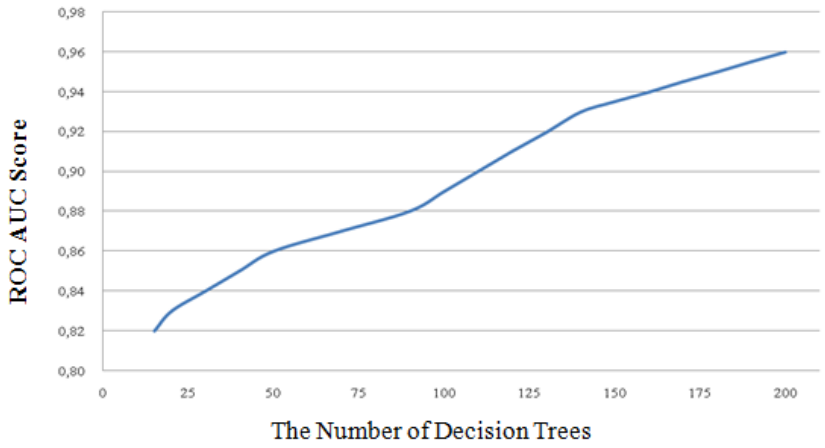


Fig.2. Dependency diagram between the RF model quality metric (AUC-ROC) and the number of trees

The evaluation was carried out using the Weka program based on a custom data base created using mobile contact data of 40 people. The result is shown in Table 1.

Table 1. Performance indicators of RF classification model

Class	Precision	Recall	F1-score	ROC AUC
High social capital	0.921	0.933	0.927	0.944
Low social capital	0.838	0.813	0.826	0.944
Weighted Avg.	0.896	0.897	0.897	0.944

Weighted Avg. is a weighted average of the scores for both classes (high social capital and low social capital). This metric shows generally good results for both classes, with all scores above 0.89.

A high Kappa value (0.75) indicates good agreement between the predicted and true classes; the model effectively classifies objects. This means that the RF model makes very accurate predictions (Table 2).

MAE (Mean absolute error), or mean absolute error, is a metric used to evaluate the accuracy of forecasting models. It shows the average absolute deviation of the predicted values from the actual ones. MAE = 0.1605 indicates that the model we propose is accurate enough.

Table 2. Classification metrics in RF

Metrics	Evaluations
Correctly Classified Instances	897
Incorrectly Classified Instances	103
Kappa statistic	0.7526
MAE	0.1605
Total Number of Instances	1000

4.3. Evaluation of Feature Importance Using the Permutation Importance Method

There are different methods of feature selection in RF: Permutation Importance (random permutation of feature values and evaluation), Threshold Selection (features whose importance exceeds a given threshold are selected), top-k feature selection (k features with the highest importance are selected), Filtering methods (statistical tests to evaluate the significance of each feature), Recursive Feature Elimination (removes the least important features at each step), etc.

For evaluation, we use the Permutation Importance method. The idea of the method is to estimate how much the quality of the model will deteriorate with a random permutation of feature values [42]. A feature is considered important if permuting its values reduces the quality of the machine learning algorithm.

4.4. Algorithm for Assessing the Importance of Features

Step 1. We select the AUC-ROC metric as the evaluation criterion. AUC-ROC will show how well the model copes with predictions after we "spoil" one of the features.

Step 2. For each feature, we randomly shuffle its values in the test sample. This breaks the connection between this feature and the target variable, making it essentially "random noise".

Step 3. We use the training set statistics from the scikit-learn library. We define the minimum number of samples for the terminal node to 10. Thus, when repeating the experiment, we limit the ability of trees to overfit:rf.set_params(classifier__min_samples_leaf=10). fit(X_train, y_train).

Step 4. Calculate the AUC-ROC model quality metric on the modified sample.

Step 5. Calculate the difference between the quality of the original and modified samples. This difference will be the measure of the importance of the feature. The importance of the *i*-th feature: $\omega(i) = metric(f(X), y) - metric(f(\hat{X}), y)$

Where X – a matrix of features where each column corresponds to one feature, y –vector of target variables, $f(X)$–model trained, $metric(f(X), y)$ – model quality metric, based on the original data, $metric(f(\hat{X}), y)$ – model quality metric on data with permuted feature. X_i – *i*-th column of matrix X (*i*-th features), $\hat{X}_i$ – copy of column X_i with valuesswapped. $\hat{X}$ – matrix X in which the *i*-th column is replaced by $\hat{X}_i$.

The difference between these values ($\omega(i)$) shows how much the quality of the model's quality deteriorated after the feature was permuted. If this difference is large, it means that the rearrangement of the i-th feature has significantly deteriorated the quality of the model, therefore, this feature was important for decision making [43]. If $\omega(i) = 0$, it means that the rearrangement of the feature had virtually no effect on the quality of the model, therefore, this feature is not important for decision making.

The figure below shows the importance of permutation features for RF (Fig. 3). Features are ranked in descending order of their weight (regression coefficient). Fewer features reduce the training time of the model and help to better interpret the results. The script selects 9 features from our dataset that provide the best performance for a random forest classifier. As shown in the graph, in the assessment of social capital as a result of ranking the number of calls per day (Coll), place of work (W), activity (Act), the number of messages sent and received per day (Com), the size of text messages per day (Text), education (Ed), age (Age), place of residence (Res), frequency of places visited per month based on location data (Loc) is included in a set of more important signs. If we change the values of these features, this will lead to a decrease in the assessment of the accuracy of the model.

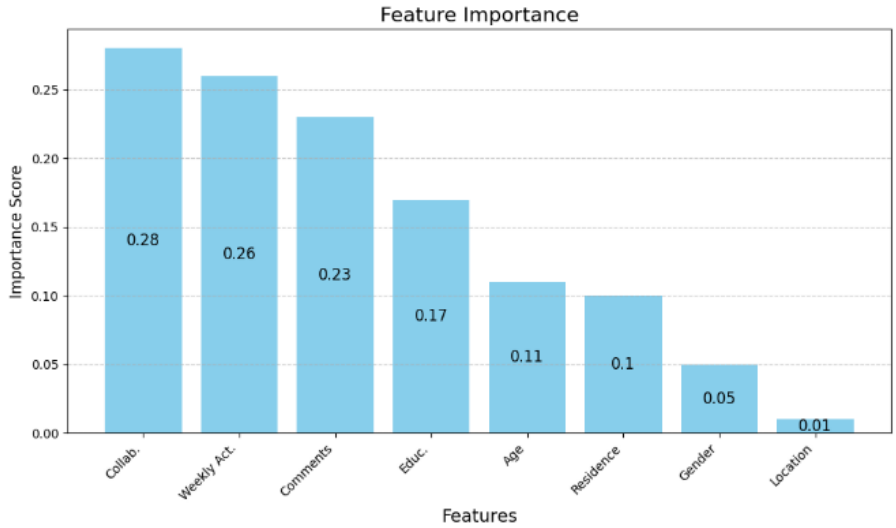


Fig.3.-Graph of parameter importance in a Random Forest model

After identifying more important features, we exclude unimportant features. For example, the tone of messages, gender and income of users may not always accurately reflect the quality and depth of social ties. If data on co-mobility are limited, some attributes of geographical mobility may be less meaningful. Also, the regularity of visits to certain places (work, home, vacation places). may not always reflect social capital. In some cases, message metadata (time, text size) may be more informative than the text itself. If data on joint mobility is limited, some attributes of geographic mobility may be less meaningful. We build a LR model that models the probability of an object belonging to a certain class (for example, a high level of social capital) depending on the values of the features.

4.5. Building an LR Model

LR assumes a linear relationship between the features and the target variable [44]. In LR, we calculate the logit (the logarithm of the ratio of the probability of an event occurring to the probability of its non-occurrence) and then apply the sigmoid function to obtain the probability. The general formula for the logit is:

$$Logit(P_{sc}) = \beta_0 + \beta_1 \times Coll + \beta_2 \times Loc + \beta_3 \times Com + \beta_4 \times Text + \beta_5 \times Age + \beta_6 \times W + \beta_7 \times Ed + \beta_8 \times \mathrm{Re}\, s + \beta_9 \times Act + \varepsilon \quad (9)$$

$Logit(P_{SC})$ – social capital index. β_0 – free regression coefficient. Shows the base probability of high social capital when all independent variables are equal to zero. β_n – coefficient orweight of *n*-th feature. $\beta_1, \beta_2, \ldots, \beta_n$ – regression coefficient прифor the corresponding independent variables. $\beta_n \times X_n$ – multiplying the input feature by its coefficient (weight). ε – model error (noise).

Logit plays a key role in LR, providing a way to model the probabilities of binary events (social capital being above -1 or below -0) within a linear model. For example, the number of calls per day may have more weight than the frequency of attending sporting events. They show how each feature affects the probability of high social capital.

A positive coefficient means that an increase in the value of the feature increases the probability, a negative coefficient means that it decreases it. In logistic regression, the coefficients of the corresponding independent variables are trained on the data to best separate the classes [44].

The result of the linear combination is passed to the sigmoid function, which transforms it into a probability [45]. The standard sigmoid formula is:

$$g(z) = \frac{1}{1 + e^{-z}} \tag{10}$$

To determine the probability that an object belongs to a positive class (high social capital) for a given set of

$$1 + \exp(-\log it(P_{sc})) \geq 1 \tag{11}$$

where $exp(-log\ it(P_{SC}))$ – exponent of a negative P_{SC} value, $1 + exp(-log\ it(P_{SC}))$ – the denominator of a fraction, which is always greater than 1.

Dividing one by the denominator yields a value between 0 and 1. LR for a binary classification (high/low social capital) can be represented as follows:

$$P(y = 1|x) = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 \times Coll + \beta_2 \times Loc + \ldots + \beta_9 \times Act + \varepsilon)} \tag{12}$$

If, using the sigmoid function, we get a probability close to 1, this indicates a high probability of high social capital for this user. The resulting probability is an estimate based on the model and data. For example, if the model predicts a probability of 0.8, this means that it believes with 80% certainty that the object belongs to the positive class.

4.6. Optimization in LR

Regression coefficients are calculated during the optimization step of the LR algorithm. Optimization is aimed at finding such weights (coefficients) of the model, at which the loss function is minimal. There are different loss functions (Cross Entropy, Mean Square Error, Hinge Loss, Exponential Loss), which can be used in logistic regression. The most commonly used function is the cross-entropy function (Cross Entropy) as a loss function. This function is also called log loss:

$$L(\theta) = \frac{-1}{N \times \sum (y_i \times \log(h_\theta(x_i)) + (1 - y_i) \times \log(1 - h_\theta(x_i))}, \tag{13}$$

where $L(\theta)$ – loss function, θ – vector of model parameters. During model training, these parameters must be optimized to minimize the value of the loss function. N – number of training examples. y_i – true class (0 or 1) for the *i*-th example, $h_\theta(x_i)$ – predicted probability of positive class ($y_i=1$) for the *i*-th example, with parameters θ (sigmoid output). In LR $h_\theta(x_i)$ is calculated using the sigmoid function as follows:

$$h_\theta(x_i) = P\ (y_i = 1|x_i; \theta) = p_i \tag{14}$$

$$L(\theta) = \frac{-1}{N \times \sum (y_i \times \log(p_i) + (1 - y_i) \times \log(1 - p_i)}, \quad 0 \leq p_i \leq 1 \tag{15}$$

p_i – is the predicted probability of the positive class, y_i – true value (0 or 1). According to probability theory, the probability that an individual belongs to class 0 ($y_i=0$) will be:

$$P\ (y_i = 0|x_i;\theta) = 1 - P\ (y_i = 1|x_i;\theta) \tag{16}$$

As we can see, model p_i produces a probability $y_i=1$ or $y_i=0$.

$$-\log(P\ (y_i = 1|x_i;\theta)) = -\log(p_i), y_i = 1 \tag{17}$$

$$-\log(1 - P\ (y_i = 1|x_i;\theta)) = -\log(1 - p_i)), y_i = 0 \tag{18}$$

From formula (13) it is clear that the closer the probability output by the model is to one, the smaller the error. Let's assume that an individual's social credit is rated 0. Then, with values close to zero, the model error will be minimal. Using formulas (13) and (14), sum of probabilities belonging to class 1 can be written as:

$$L(\theta) = \begin{cases} -\log(p_i), y = 1 \\ -\log(1 - p_i), y = 0 \end{cases} \tag{19}$$

The presented loss function is a key component of the LR algorithm. It allows us to evaluate the quality of the model and find the optimal values of its parameters. By minimizing this function, we obtain a model that most accurately predicts the probability of an object belonging to a certain class. For problems with more than two classes, generalizations of binary cross-entropy are used, such as cross-entropy with SoftMax.

5. Evaluation of a Cascade Model

Training a model means fitting a chosen machine learning algorithm to pre-processed data. The goal of training a model is to find the parameter values θ that minimize the loss function $L(\theta)$. This means that the model will best fit the training data and make the most accurate predictions. The scikit-learn libraries in Python were used for the experiment. As can be seen from Table 3, LR is better suited to the analysis and determination of human social capital. Its AUC-ROC indicator is 0.92, the F1-score indicator is 0.81, which is inferior only to RF and K-NN, but Training time (1.2), Prediction time (0.1) and Interpre-tability are higher than those of RF and K-NN.

Table 3. Comparative analysis of machine learning models

Algorithm Name	Accuracy	Precision	Recall	F1-score	AUC-ROC	Training time (sec.)	Prediction time (ms)	Interpretability
Support Vector Machines (SVM)	0.82	0.85	0.78	0.81	0.92	2.5	0.2	High
Logistic Regression (LR)	0.82	0.85	0.78	0.81	0.92	1.2	0.1	High
Naive Bayes (NB)	0.78	0.80	0.75	0.77	0.90	0.1	0.01	Very High
Decision Tree (DT)	0.80	0.82	0.77	0.79	0.91	1.5	0.2	High
Random Forest (RF)	0.85	0.88	0.82	0.85	0.94	10.5	0.5	Low
K-Nearest Neighbors (K-NN)	0.83	0.86	0.80	0.83	0.93	2.0	0.3	Low

Combining the strengths of the RF and LR algorithms, the model allows us to achieve high accuracy and flexibility in data analysis. Thus, we obtain a more accurate assessment in determining social capital. The proposed a Cascade Model can be successfully applied in various fields, such as sociology, marketing and urban planning. The results of the study can be used to develop programs and policies aimed at strengthening social capital and improving the quality of life of people.

After a comparative analysis of simple machine learning algorithms, we compare the proposed approach with more complex methods. Table 4 presents a comparative analysis of the Cascade Model with advanced machine learning models such as MLPClassifier (Multi-Layer Perceptron Classifier), BernoulliNB, XGBoost (gradient boosting) and ANN. The MLPClassifier is actually the simplest and most common neural network model, and it is implemented in the scikit-learn libraries. When we compare the MLPClassifier model with the ANN, we are actually comparing a more specific ANN model with a general category.

The study utilized the Python programming language and the Scikit-learn, XGBoost, and PyTorch libraries. Each algorithm was trained on a pre-prepared dataset and evaluated using the Accuracy, F1-score, and ROC AUC metrics. Performance, interpretability, and other metrics were also calculated during the evaluation. The results obtained from the comparative analysis of the classifiers are presented in Table 4.

Table 4. Comparison of the performance of the Cascade Model (RF and LR) with modern algorithms (MLPClassifier , BernoulliNB, XGBoost, ANN and SVM)

Algorithm Name	Accuracy	F1 Score	ROC AUC	Training time (sec.)	Prediction Time (ms)	Interpretability
Cascade Model	Very High	0.94	0.96	Very High	15.0	Medium
MLPClassifier	Very High	0.95	0.97	Medium	20.5	Low
BernoulliNB	Low	0.65	0.70	Very High	10.2	High
XGBoost	Very High	0.92	0.95	Medium	18.7	Low
ANN	Very High	0.96	0.98	Low	30.1	Very Low

The presented results reflect an experiment conducted on a specific dataset (Table 4). ANN (0.96 F1 and 0.98 ROC AUC) and MLPClassifier (0.95 F1 and 0.97 ROC AUC) demonstrate the highest performance. This suggests that deep learning models are the most effective for this task. Cascade Model (0.94 F1, 0.96 ROC AUC) shows very high results compared to XGBoost (0.92 F1, 0.95 ROC AUC) and BernoulliNB (0.65 F1, 0.70 ROC AUC), but lags behind neural networks. BernoulliNB is significantly weaker in terms of performance. However, it provides the fastest training, just like Cascade Model.

ANN has the slowest speed (30.1 sec) because it is a complex model. Cascade Model (15.0 ms) and BernoulliNB (10 ms) have the highest predictive ability. ANN is the slowest model, which can lead to potential delays in real-time applications. Although the Cascade Model's performance is lower than that of complex models such as ANN and MLPClassifier, it is characterized by a longer training time and a higher inference rate than these models. If the sole goal of data analysis is to maximize F1 Score and ROC AUC, MLPClassifier or ANN may provide better results, but if speed and transparency of data processing are required, the Cascade Model is the most successful option. Therefore, using the Cascade Model for social process prediction, social capital assessment, or demographic studies may be the most successful option.

Table 5 compares the performance of machine learning models using K-fold Cross-validation. Cross-validation is used to measure how reliable and stable a model is on different pieces of data. The table shows the Mean and the 95% Confidence Interval (CI) of this result for each model. The advantage of the Cascade Model here is that, although it lags behind ANN/MLP in terms of performance and is slightly unstable, it compensates for these minor shortcomings with fast prediction and average inference, providing the best overall balance between speed-accuracy-transparency.

Table. 5. Comparison of K-fold Cross-validation results and 95% confidence intervals

Algorithm Name	Mean F1 Score	F1 CI (95%)	Mean ROC AUC	ROC CI (95%)
Cascade Model	0.940	[0.927, 0.955]	0.960	[0.955, 0.963]
MLPClassifier	0.950	[0.941, 0.957]	0.970	[0.964, 0.977]
BernoulliNB	0.650	[0.588, 0.669]	0.700	[0.670, 0.724]
XGBoost	0.920	[0.914, 0.937]	0.950	[0.945, 0.960]
ANN	0.960	[0.950, 0.964]	0.980	[0.974, 0.984]

The chart in Figure 4 presents a comprehensive comparison of the performance of various machine learning models (Cascade Model, MLPClassifier, BernoulliNB, XGBoost and ANN).

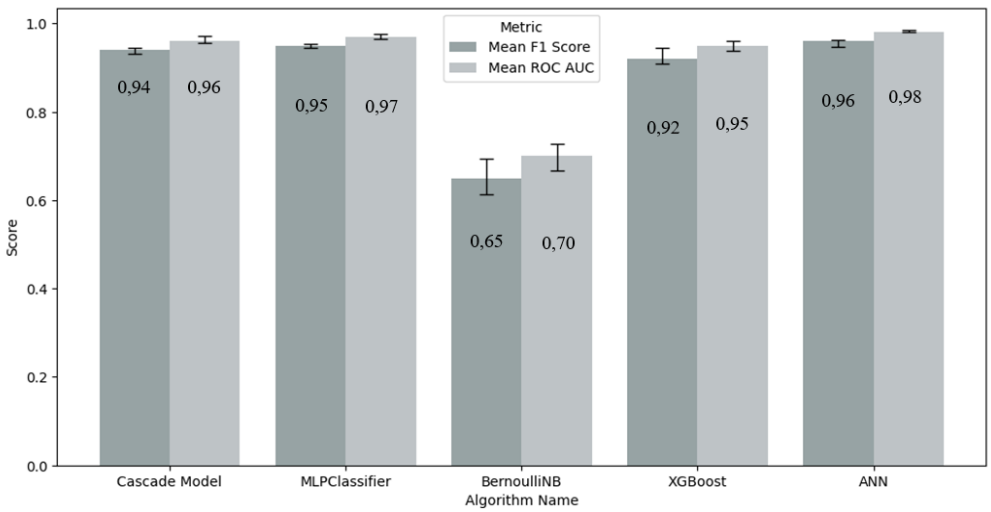


Fig.4. Comparison of K-fold Cross-validation results and confidence intervals of machine learning models

This comparison covers not only the metric values but also their statistical reliability. The chart shows that the Cascade Model (0.96), MLPClassifier (0,97) and ANN (0.98) have the highest ROC AUC values. The most important feature of the chart is the black lines located above each column. These lines represent 95% confidence intervals. This demonstrates the statistical reliability of the average results obtained by the model as a result of the k-fold cross-validation process.

As we can see from the chart, the Cascade Model demonstrates stable and consistent results throughout the k-fold cross-validation process, meaning it produces the same results on different test sets. Here, we see that the BernoulliNB model's results vary more across test sets, indicating they are less reliable.

6. Calculating Social Capital on Specific Data by Applying the Cascade Model

Using call and message data collected from mobile devices (with employee consent), a social capital analysis was conducted for 10 individuals within the organization. Communication data on the number of daily calls and messages, as well as features such as workplace, education, and age, were collected and anonymized. For each employee, an expanded set of metrics influencing social capital was calculated:

- Calls (Coll): Average number of daily calls.
- Messages (Com): Number of messages sent and received.
- Text Size (Text): Average size of messages sent.
- Workplace (W): Employee's affiliation with a department or structural unit.
- Education (Ed): Level of education.
- Age (Age): Employee's age.
- Activity (Act): Type of role performed and level of activity.
- Residence (Res): Proximity of residence to the workplace.
- Visit Frequency (Loc): Frequency of places visited per month based on location data.

The calculations were performed using Python, Colab.research and the Scikit-learn libraries. Based on this set of features, each employee was assigned a social capital score from 1 to 10. The use of a Cascade Model determined that certain features (Coll, Com, W, Act, Text, and Ed) have a higher significance in social capital assessment.

F1 and F2 achieved the high scores across all key metrics (9.5 - 8.0). Coll and Com indicators make them central hubs for information dissemination within the organization (Table 6). The significance of their workplaces and education levels reinforces this trend. F3, F4, F5, F6, F7, and F8 demonstrate balanced profiles (7.9 - 5.0).

Table. 6. Results of social capital analysis for 10 people in an organization

Employee	Coll (Avg/day)	Com (Avg/day)	Text (Avg size)	W (Significance)	Ed (Level)	Social Capital Score
F1	15+	40+	High	High	High	9.5
F2	10-15	30-40	High	High	High	8.8
F3	8-10	20-30	Medium	Medium	High	8.1
F4	5-8	10-20	Low	Medium	Medium	7.5
F5	10-15	20-30	Medium	Medium	Medium	7.0
F6	5-8	10-20	Low	Low	High	6.4
F7	8-10	10-20	Low	Low	Medium	5.9
F8	5-8	5-10	Low	Medium	Medium	5.2
F9	5-8	5-10	Low	Low	Low	4.5
F10	5-8	5-10	Low	Low	Low	4.0

Their social capital is sufficient for performing daily tasks, but certain features prevent them from achieving a higher score. For instance, F4 and F7 have low Text and Com indicators. F9 and F10 have low overall scores (4.9 - 1.0). This may be due to less active communication or less significant work roles. According to the Cascade Model, all their key indicators (Coll, Com, Text, W, Ed) are at a low level, suggesting they are less integrated into the organization's information flow. The solution to the problem can be divided into the following steps:

Step 1. Random Forest Model. The first step uses a random forest model. The output shows the "Permutation Feature Importance," which shows the contribution of each feature to the model's predictions. The algorithm identified features such as "Com" (number of messages), "W_IT" (workplace: IT), and "Text" (average message size) as the most important in this dataset.

Step 2. Logistic Regression Model. In the second step, the random forest model predictions are used as an additional feature in the logistic regression model.

Step 3. The Cascade Model is evaluated. The Cascade Model evaluation metrics (MSE and R2) evaluate how well the cascade model predicts a continuous numerical value (in our case, Social Capital Score). They show a mean square error (MSE) of 0.08 and a correlation coefficient (R2) of 0.98 (Fig.5).

```
Stage 1: Training the Random Forest Model...

Permutation Feature Importance (from RF Model):
                    Feature  Importance
15         remainder__Text    0.192878
14          remainder__Com    0.181541
1             onehot__W_IT    0.163232
4          onehot__Act_High    0.140159
7       onehot__Ed_Bachelor    0.080278
9         onehot__Ed_Master    0.067235
13          remainder__Coll    0.056603
16           remainder__Age    0.033440
17           remainder__Loc    0.029371
2       onehot__W_Marketing    0.011114
3           onehot__W_Sales    0.009169
8   onehot__Ed_High School    0.008674
10           onehot__Ed_PhD    0.008028
5           onehot__Act_Low    0.004280
0             onehot__W_HR    0.003839
6        onehot__Act_Medium    0.003746
12          onehot__Res_Far    0.003746
11        onehot__Res_Close    0.002668

Stage 2: Training the Linear Regression Model...

Cascaded Model Evaluation:
Mean Squared Error (MSE): 0.08
R-squared (R2) Score: 0.98

Predictions for the first 5 test samples:
   Actual Score  Predicted Score
0          6.44             6.62
1          3.25             3.56
2          1.63             1.99
3          5.36             5.58
4          4.52             4.60
```

Fig.5. Cascade Model measurement metrics

The lower the MSE, the better. An MSE of 0.08 is a very low value, indicating high accuracy. R2 indicates the proportion of the target variable's variance explained by the model. An R2 of 0.98 means that the model explains 98% of the variability in the data, which is a very good result.

While high-performance ML models often offer higher accuracy, they suffer from the 'Black Box Problem,' making it nearly impossible to explain why the prediction was made. Our cascaded approach achieves high predictive accuracy while maintaining interpretability. This is done by using the RF's output, which allows us to analyze the relative influence of the initial input features (as shown in the Feature Importance table), grounding the final, highly accurate prediction in an explainable framework. Unlike standard ensembles (like simple bagging or boosting), which deliver one aggregate result, our two-stage architecture separates the modeling tasks. The RF focuses on information maximization and handling non-linearity, and the LR focuses on minimizing residual error. This leads to a more stable, efficient, and methodologically sound solution to the problem of modeling complex social processes.

7. Conclusions

Mobile devices have become an integral part of modern life, generating huge amounts of data that offer valuable insights into human behavior and social capital. Mobile data provides a unique opportunity to study people's social relationships as well as the demographic state of a country.

A Cascade Model in machine learning is a sequential application of several ML models, where the output of one model serves as the input to the next. The first stage of the Cascade Model uses an RF Classifier. Its "predict_proba" method obtains a probability score for each prediction. Examples with this probability score less than 0.95 are considered "unreliable" and are passed to the LR model in the second stage. This approach allows a powerful model like RF to quickly provide predictions, while LR can better analyze examples that are considered more difficult.

This separation of tasks results in higher model accuracy, since each model is optimized for its part of the task. It also reduces the risk of over-fitting, since the model becomes more robust to noise in the data. Cascading allows each model to focus on its strengths, resulting in higher overall data classification accuracy and more accurate and

interpretable results when assessing social capital from mobile data. Deep learning models process data in a very complex way, and although they produce high results, it's difficult to understand why the model produces these results. This is the main drawback of the ANN model. Our proposed Cascade Model easily solves this problem. The Cascade Model, while ensuring high accuracy, runs faster and makes predictions easier to interpret. This feature makes it a good choice, especially in the field of social research, where it is important to understand why the results are obtained the way they are.

Because the study presented in this article was conducted on a small sample (40 participants over one month), the generalizability of its results to a wider audience is limited. To overcome this limitation and enhance the robustness of the proposed model, future research will focus on the following strategic areas:

- To expand the scope of the study, it is planned to collect data from larger and more diverse groups of participants in terms of demographic characteristics (age, gender, employment, income level, etc.). This will allow for a more accurate assessment of the model's performance across various social groups.
- Short-term (one-month) data collection may not fully capture the dynamics of social capital. Therefore, future studies are expected to collect data from participants over a longer period (e.g., six months). This approach will allow for an examination of the impact of changes in social relationships and behavior over time on the predictive power of the model.
- It is planned to recruit participants from different regions of the country so that the study's results are not limited to a single geographic region. This will help determine how social capital is formed in different societies and more accurately assess the applicability of the proposed model.

Implementation of these strategic steps will improve the reliability of the Cascade Model and significantly expand its scope of application in real life.

Author Contributions Statement

The author is the sole contributor to this work, including conceptualization, methodology, software implementation, data curation, model training, validation, formal analysis, and writing (original draft, review, and editing).

All authors have read and agreed to the published version of the manuscript.

Conflict of Interest Statement

The author declares no conflicts of interest.

Funding Declaration

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data Availability Statement

The datasets analyzed in this study are not publicly available due to privacy constraints and the terms of the participants' informed consent. Data may be shared only upon reasonable request and with explicit permission from the subjects.

Ethical Declarations

This article does not contain any studies with human participants or animals performed by any of the authors.

Acknowledgments

We sincerely thank the experts for their professional evaluation and valuable recommendations, which have contributed to improving the quality of the experiment and the reliability of its results.

Declaration of Generative AI in Scholarly Writing

During the preparation of this article, the author used Google Gemini only at the end of the "Conclusion" section to improve the manuscript's English and readability. After using this tool, the author reviewed and edited the content and bears full responsibility for the publication's content.

Abbreviations

The following abbreviations are used in this manuscript:

AI - Artificial Intelligence
ANN – Artificial Neural Networks
AUC-ROC - Area Under the Curve – Receiver Operating Characteristic
CPSS – Cyber-Physical-Social System
DL - Deep Learning

DT - Decision Trees
FN - False Negative
FP - False Positive
FPR - False Positive Rate
ICT - Information and Communication Technologies
LR – Logistic Regression
MAE - Mean Absolute Error
MCS - Mobile Communications Services
ML - Machine Learning
MLPClassifier - Multi-Layer Perceptron Classifier
NLP - Natural Language Processing
PSO – Particle Swarm Optimization
RL - Random Forest
SMS - Short Messages
SVM - Support Vector Machines
TF-IDF - Term Frequency – Inverse Document Frequency
TN - True Negative
TP - True Positive
TPR - True Positive Rate

References

[1] P. Kyung-Gook, S. Han, and L. L. Kaid, "Does Social Networking Service Usage Mediate the Association Between Smartphone Usage and Social Capital?" in *New Media & Society*, Vol.15, No.7, 2013, pp. 1077–1093. doi: 10.1177/1461444812465927

[2] J. C. Wang, C. Y Hsieh, and S. H. Kung, "The Impact of Smartphone Use on Learning Effectiveness: A Case Study Of Primary School Students," in *Education and Information Technologies*, Vol.28, No.6, 2023, pp. 6287–6320, doi: 10.1007/s10639-022-11430-9

[3] W. Zhang, and A. Abitbol. "The Role of Public Relations in Social Capital," in *Journal of Communication and Media Technologies*, Vol.6, 2016, http://ecommons.udayton.edu/cmm_fac_pub/44.

[4] P. Bourdieu, and J. S. *Coleman, Social theory for a changing society*, 389 p., Westview Press. 1991.

[5] P. Bourdieu, G. Sapiro, and B. McHale, "The New Capital: Introduction to a Japanese Reading of State Nobility", in *National Literatures/Social Space*s, Vol.12, No. 4, 1991, pp. 643–653. doi: 10.2307/1772705

[6] J. Coleman, "Foundations of Social Theory", 302 p., Belknap Press, Cambridge, 1994. doi: 10.1007/978-3-658-13213-2_88

[7] R. D. Putnam, "Social Capital: Measurement and Consequences, in *Canadian Journal of Policy Research*, Vol.2, 2001, pp. 41–51. http://www.sietmanagement.fr/wp-content/uploads/2016/04/Putnam_SocialCapital.pdf

[8] R. M Alguliyev, and I. Y. Alakbarova, "Analysis of the Role of Social Capital in the Formation of an Electronic Government", in *Problems of information technology*, Vol.10, No.1, 2019, pp. 35–46. doi: 10.25045/jpit.v10.i1.04

[9] R. D. Putnam, "Bowling Alone: America's Declining Social Capital", in *Journal of Democracy*. Vol.6, No.1, 1995, pp. 65–78. https://link.springer.com/chapter/10.1007/978-1-349-62397-6_12

[10] S. Yang, S. Kurnia, and S. P. Smith, "The Impact of Mobile Phone Use on Individual Social Capital", in *Proceedings of the 44th Hawaii International Conference on System Sciences (*HICSS2011*)*, Kauai, HI, USA, 2011, pp.1–10. doi: 10.1504/IJEF.2013.057277

[11] J. Karikoski, and K. Kilkki, "Building Social Capital with Mobile Communication Services", in *International Journal of Electronic Finance*, Vol.7, No.2, 2013, pp. 115–131, doi:10.1504/IJEF.2013.057277

[12] K. S. Vivek, and I. Ghosh, "Inferring Individual Social Capital Automatically via Phone Logs" in *Proceedings of the ACM on Human-Computer Interaction*, Vol.1, 2017, pp. 1–12. doi: 10.1145/3134730

[13] A. N. Gentry, J. P. Martin, and K. A. Douglas, "Social capital Assessments in Higher Education: a Systematic Literature Review", in *Frontiers in Education*, Vol. 9, 2025, pp. 1498422. doi: 10.3389/feduc.2024.1498422

[14] R. Ketipov, V. Angelova, L. Doukovska, and R. Schnalle, "Predicting User Behavior in e-Commerce Using Machine Learning", in *Cybernetics and Information Technologies*, Vol.22, 2023, No.3, pp. 89 – 101. doi: 10.2478/cait-2023-0026

[15] A. Aljadani, B. Alharthi, M. A. Farsi, H. M. Balaha, M. Badawy, and M. A. Elhosseini, "Mathematical Modeling and Analysis of Credit Scoring Using the LIME Explainer: A Comprehensive Approach", in *Mathematics*, Vol.11, 2023, pp. 1–28. doi: 10.3390/math11194055

[16] N. Ullah Khan, W. Wan, R. Riaz, Sh. Jiang, and X. Wang, "Prediction and Classification of User Activities Using Machine Learning Models from Location-Based Social Network Data", in *Applied Sciences*, Vol.13, No.6, 2023, pp. 3517; doi: 10.3390/app13063517

[17] M. Z. Ansaria, M. B. Aziza, M. O. Siddiquib, H. Mehraa, and K. P. Singha, "Analysis of Political Sentiment Orientations on Twitte", in *Proceedings of the International Conference on Computational Intelligence and Data Science (ICCIDS 2019),* Vol.27, 2020, pp. 1821–1828, doi: 10.1016/j.measen.2023.100768

[18] I. H. Sarker, "Mobile data science: Towards Understanding Data-Driven Intelligent Mobile Applications", in *EAI Endorsed Transactions on Scalable Information Systems*, Vol.5, No.19, 2018, 155866. doi: 10.4108/eai.13-7-2018.155866

[19] S. W. Campbell, and N. Kwak, "Mobile Communication and Social Capital: An Analysis of Geographically Differentiated Usage Patterns", in *New Media & Society*, Vol.12, No.3, 2010, pp. 435–451. doi: 10.1177/1461444809343307

[20] G. M. Jones, P. Santhiya, S. G. Winster, and R. Sundar, "An Intelligent Analysis of Mobile Evidence Using Sentimental Analysis" in *Proceedings of the International Conference on Information Security, Privacy and Digital Forensics*, 2022, pp. 317–330, Singapore: Springer Nature Singapore. doi: 10.1007/978-981-99-5091-1_23

[21] F. Amin, and G. S. Choi, "Analysis and Modeling of Mobile Phone Activity Data Using Interactive Cyber-Physical Social System" in *Computers, Materials & Continua*, Vol.80, No.3, 2024, doi: 10.32604/cmc.2024.053183

[22] I. H. Sarker, M. M. Hoque, M. K. Uddin, and T. Alsanoosy, "Mobile Data Science and Intelligent Apps: Concepts, AI-based Modeling and Research Directions", in *Mobile Networks and Applications*, Vol.26, No.1, 2021, pp. 285–303. doi: 10.1007/s11036-020-01650-z

[23] M. Okmi, L. Y. Por, T. F. Ang, and C. S. Ku, "Mobile Phone Data: A Survey of Techniques, Features, and Applications", in *Sensors*, Vol.23, No.2, 2023, pp. 908. doi: 10.3390/s23020908

[24] S. Sun, X. Wang, and D. Wang, "Smartphone Usage Patterns and Social Capital Among University Students: The Moderating Effect of Sociability", in *Children and Youth Services Review*, Vol.155, 2023, 107276. doi: 10.1016/j.childyouth.2023.107276

[25] J. Liu, and Y. Zhang, "The Impact of Mobile Social Network Sites on Social Trust: Evidence from the China" in *Humanities and Social Sciences Communications*, Vol.12, No.1, 2025, pp. 1–12. doi: 10.1057/s41599-025-05408-4

[26] G. Grotto, and A. Buja, "Association Between Digital Technology Use and Social Capital Among Older Adults: A Systematic Review", in *Digital Health*, Vol.11, 2025, doi: 10.1177/20552076251336973

[27] Law of the Republic of Azerbaijan on Personal Data, 2010, № 998-IIIQ, https://e-qanun.az/framework/19675

[28] A. M. Sarhan, H. Ayman, M. Wagdi, B. Ali, A. Adel, and R. Osama, "Integrating Machine Learning and Sentiment Analysis in Movie Recommendation Systems", in *Journal of Electrical Systems and Information Technology*, Vol.11, No.53, 2024, pp. 1–25, doi: 10.1186/s43067-024-00177-7

[29] Zh. Hu, I., K. Dychka, V. Potapova, and V. Meliukh, "Augmenting Sentiment Analysis Prediction in Binary Text Classification through Advanced Natural Language Processing Models and Classifiers". in *International Journal of Information Technology and Computer Science* (IJITCS), Vol.16, No.2, 2024, pp. 16–31, doi: 10.5815/ijitcs.2024.02.02

[30] H. M. Balaha, A. E. S. Hassan, E. M. El-Gendy, H. Z. Eldin, and M. M. Saafan, "An Aseptic Approach Towards Skin Lesion Localization and Grading Using Deep Learning and Harris Hawks Optimization", in *Multimedia Tools and Applications*; Springer, Vol.83, No.7, 2023, pp. 1–29. doi: 0.1007/s11042-023-16201-3

[31] I. Y. Alakbarova and D. Alizada, "A New Approach to Improving Search Efficiency in Digital Libraries", in *International Journal of Intelligent Systems and Applications*, Vol.16, No.2, 2024, pp. 13–23. doi: 10.5815/ijisa.2024.02.02

[32] A. Bria, C. Marrocco, M. Molinara, and F. Tortorella, "An Effective Learning Strategy for Cascaded Object Detection", in *Information Science*, Vol.340, 2016, pp. 17–26. doi: 10.1016/j.ins.2016.01.021

[33] A. A. Soofi, and A. Awan, "Classification Techniques in Machine Learning: Applications and Issues", in *Journal of Basic and Applied Sciences*.Vol.13, 2017, pp. 459–465. doi: 10.6000/1927-5129.2017.13.76

[34] X. Yuan, S. Liu, W. Feng, and G. Dauphin, "Feature Importance Ranking of Random Forest-Based End-to-End Learning Algorithm" in *Remote Sensing*, Vol.15, No.21, 2023, pp. 5203. doi: 10.3390/rs15215203

[35] L. Breiman, "Random Forests" in *Machine Learning*, Vol.45, No.1, 2001, pp. 5–32. https://link.springer.com/content/pdf/10.1023/A%3A1010933404324.pdf

[36] T. Daniya, M. Geetha, and K. S. Kumar, "Classification and Regression Trees with Gini Index", in *Advances in Mathematics: Scientific Journal*, Vol.9, No.10, 2020, pp. 8237-8247. doi: 10.37418/amsj.9.10.53

[37] S. Dhingra, and P. Bansal, "Employing Divergent Machine Learning Classifiers to Upgrade the Preciseness of Image Retrieval Systems" in *Cybernetics and Information Technologies*, Vol.20, No.3, 2020, pp.75–85. doi: 10.2478/cait-2020-0029

[38] J. W. Abraham, M. Olson, J. Bleich, and D. Mease, "Explaining the Success of AdaBoost and Random Forests as Interpolating Classifiers" in *Journal of Machine Learning Research*, Vol.18, 2017, pp.1–33. https://www.jmlr.org/papers/volume18/15-240/15-240.pdf

[39] D. Chicco., and G. Jurman, "The Matthews Correlation Coefficient (MCC) Should Replace the ROC AUC as the Standard Metric for Assessing Binary Classification", in *BioData Mining*, Vol.16, No.4, 2023, pp.1–23. https://link.springer.com/article/10.1186/s13040-023-00322-4

[40] G. Naidu, T. Zuva, and E. M. Sibanda, "A Review of Evaluation Metrics in Machine Learning Algorithms" in *Artificial Intelligence Application in Networks and Systems*, 2023, pp. 15-25. https://link.springer.com/chapter/10.1007/978-3-031-35314-7_2.

[41] N. Ritushree, P. Samanta, "A Machine Learning Approach for Sentiment Analysis Using Social Media Posts," in *International Journal of Information Technology and Computer Science (IJITCS),* Vol.16, No.5, 2024, pp. 23–35. doi: 10.5815/ijitcs.2024.05.02

[42] A. Altmann, L. Tolosi, O. Sander, and T. Lengauer, "Permutation Importance: a Corrected Feature Importance Measure", in *Bioinformatics*, Vol.26, No.10, 2010, pp. 1340-1347, doi: 10.1093/bioinformatics/btq134

[43] N. Huang, G. Lu, and D. Xu, "Permutation Importance-Based Feature Selection Method for Short-Term Electricity Load Forecasting Using Random Forest", in *Energies*, Vol.9, No.10, 2016, 767, doi: 10.3390/en9100767

[44] H. Schaeben, "Targeting: Logistic Regression, Special Cases and Extensions" in *International Journal of Geo-Information*, Vol.3, No.4, 2014, pp. 1387–1411, doi: 10.3390/ijgi3041387

[45] H. Iqbal, and I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions, SN", in *Computer Science*, Vol.2, 2021, pp. 1–21, doi: 10.1007/s42979-021-00592-x.

Authors' Profiles



Alakbarova Irada Yavar in 1984 she graduated from the Faculty of Automation of production processes, Azerbaijan Institute of Oil and Chemistry named after M. Azizbayov. In the same year, she was accepted for employment at the Institute of Information Technology. In currently holds the post of Sector Chief of the Institute of Information Technology (Ministry of Science and Education of Azerbaijan). In 2018, the defense of the dissertation on the "Development of methods and algorithms for analysis of information war technologies in a wiki environment" and she received his Ph.D. (2018). In currently conducts research in the field of Social Network Analysis, Text Analysis, Clustering, Social Credit Analysis, and Big Data Analytics. She is the author of 60 articles and three books.

How to cite this paper: Irada Y. Alakbarova, "Assessment of Social Capital from Mobile Communication Data: A Cascade Model Based on Random Forest and Logistic Regression", International Journal of Computer Network and Information Security(IJCNIS), Vol.18, No.2, pp.94-111, 2026. DOI:10.5815/ijcnis.2026.02.06