

Credit Scoring Active Telegram Channels Offering Stock Signals

Babak Sohrabi*

*Corresponding Author, Professor, Department of Information Technology Management, Faculty of Management, University of Tehran, Tehran, Iran. (Email: bsohrabi@ut.ac.ir)

Ahmad Khalili Jafarabad

Ph.D., Department of Information Technology Management, Faculty of Management, University of Tehran, Tehran, Iran. (Email: akhalili@ut.ac.ir)

Saba Orfi

MSc. In Management Information Systems, Department of Information Technology Management, Faculty of Management, University of Tehran, Tehran, Iran. (Email: orfi.saba@ut.ac.ir)

Iranian Journal of Finance, 2022, Vol. 6, No.3, pp. 142-164

Publisher: Iran Finance Association

doi: <https://doi.org/10.30699/IJF.2022.309877.1279>

Article Type: Original Article

© Copyright: Author(s)

Type of License: Creative Commons License (CC-BY 4.0)

Received: January 26, 2022

Received in revised form: March 07, 2022

Accepted: May 02, 2022

Published online: May 18, 2022



Abstract

The impact of personal judgment on the assessment of an individual's financial situation has been drastically reduced through the development of credit scoring. The systems are capable of deciding based on an applicant's total score which is a combination of several factors and indicators. Over the past few decades, credit scoring has been considered an essential tool for evaluation in various institutions and has also been able to transform the industry as a

whole.

Most of the research conducted in the field has taken into account traditional credit scoring, but considering the ever-evolving technological world that we live in and the increasing emergence of new social media networks, such research has now become obsolete. Such technological advancements have not only paved the way for far more sophisticated credit scoring systems but also essentially rendered the previous generations useless. It should be noted that credit scoring and its features have widely been discussed across the globe but, considering the various aspects and models that have to be taken into account, no one best method has been designed or suggested for it so far. This study shows that social media channels tend to perform relatively well in predicting stock market trends when the overall index is growing positively. The research also illustrates that a higher number of days of activity and a large number of signals released do not necessarily mean that the channels can or have credited their offered stock return on a one-month time frame.

The methodology used is "CRISP-DM," which consists of six steps. The main variables include social and financial variables that are examined for six months. In the research, we seek to identify, analyze and categorize active telegram channels in stock signals using the data mining model and the RFM method. The k-means algorithm is selected for this category. Then, in each cluster, the importance of social variables and the performance of the channels are extracted by the EXTRATREECLASSIFIER algorithm, and channel performance is measured by considering the changes in the total index.

Keywords: Credit scoring, Social credit scoring, RFM, K Means, CRISP-DM Methodology.

Introduction

Credit scoring and credit, in general, are pivotal instruments used by both private and public financial institutions. Credit offers the cash flow required for the development of all types of economic activity (Doumpos et al., 2019). The system of credit scoring allows analysis according to numerical values and decisions based on the applicant's overall score (Abdou & Pointon, 2011).

The enhanced credit scoring model considers a person's cognitive capacity and financial data, whereas the traditional credit scoring model is only founded on financial data (Kulkarni & Dhage, 2019). Lenders should expect to lower

their risk in the short term by integrating network-based solutions, according to financial sector managers (Wei et al., 2016). To properly determine a person's financial situation, reliability and overall personality traits are also required (Kulkarni & Dhage, 2019). Credit and behavioral scoring have emerged as essential techniques for predicting financial risk in lending activities and assisting institutions in dealing with the risk of default in consumer loans (Thomas, 2000).

When consumers purposefully adjust their networks to enhance their score and deceitfully reap the rewards that result from it, the effectiveness of social scoring does not necessarily have to decline (Wei et al., 2016). There is no one ideal statistical methodology for designing credit scoring models and no technique exists for all data sets (Abdou & Pointon, 2011). Social scoring is more likely to be effective in online or offline communities and societies where customers have strong relationships (Wei et al., 2016).

Adding qualitative variables such as the depth and breadth of a person's social media relationships and alternative payment data modifies the algorithms used to calculate credit scores. As a result, the credit score will be determined by the users' prestige and social connections. Users will have access to their credit scores using the application, encouraging fiscal discipline in their lives (Lohokare et al., 2017).

Credit scoring has become a critical responsibility since the credit corporate has doubled over the last two decades. Because of its linked memory characteristic and generalization capabilities, the artificial neural network is becoming a preferred solution in credit scoring models (Lee et al., 2002). The main issue of credit scoring, which is a critical study topic in the banking sector, is anticipating bankrupts to acquire potentially profitable clients. The suggested technique divides consumers into five categories: best, good, satisfactory, bad, and worst (Marikkannu & Shanmugapriya, 2011).

In recent years, social media used for suggesting trades on the stock market has gained noticeable traction in Iran. Therefore, the introduction of new validation models to predict the validity of signals published by active telegram channels in the stock is one of the basic needs of the present age, which is the result of this compelling and practical research in this field. By examining various articles and methods used in them, it can be said that selecting the appropriate technique for the validation of active telegram channels offering stock signals is still an open research issue. The selection of the proper technique based on data and the use of advanced data mining techniques can lead to an efficient model for the sample. Selecting appropriate features based

on the data collected will also be important in validating the review due to the lack of criteria to rely on the opinions expressed in the Iranian Telegram channels about the offers to buy and sell shares. So, the need to conduct this research seeks to implement new validation methods using the Telegram channels as one of the most popular offering platforms. Buying and selling stocks are considered in this analysis. What is worth mentioning is that we have implemented a novel approach for identifying reliable active stock traders through Iranian Telegram channels. By reviewing previous papers, we recognised that to identify attributes affecting the validation of stock traders that are active in the field of offering stock signals, we should utilize the concept of credit scoring to meet our scientific goal. So, we set our first question as below:

1. What are the factors affecting the validation of telegram channels active in the field of offering stock signals?

As social media platforms in Iran that recommend stock market trading have gained a lot of attention, we should do the related credit scoring method by considering financial and non-financial features existing in previous studies. Studies have also emphasized that a person's social media involvement can be used to assess their financial situation (Kulkarni & Dhage, 2019).

The financial situation in our case is the rate of return on stocks which was suggested by Iranian Telegram channels at one-month intervals. In other words, we had determined the ability of Iranian Telegram channels to propagate the right signal about buying or selling Iranian stocks. The importance of credit scoring Iranian Telegram channels propagating the right signals is that it can guide stockholders to gain economic benefits. This result has the potential to be categorized as a credit scoring concept. Due to the meaning of credit scoring, utilizing scoring in credit evaluation eliminates the need for personal judgment (Abdou & Pointon, 2011). We have also implemented the mentioned concept to identify the credit qualities of Iranian Telegram channels offering stock signals for reducing personal judgment.

As mentioned earlier, by doing a literature review, social and financial variables were extracted to identify factors affecting the validation of Iranian Telegram channels active in offering stock signals. Furthermore, to enrich our results, we realised the relation between Iranian market conditions and the rate of return on stocks recommended by Iranian Telegram channels offering stock signals. This analysis helps stockholders figure out appropriate features that can broaden their minds to find the creditworthiness of Iranian Telegram channels that propagate the right signals about buying or selling Iranian stocks.

As a result, we set our first question as below:

2. Due to what market conditions can the credibility of telegram channels active in the field of stock offering be trusted?

This research is divided into five main chapters: In the first chapter, a brief description of the main subject of the study and its importance is given with a short reference to the method and steps of the study. In the second chapter, the literature is the subject, which includes an overview of the validation concepts, models of validation implementation, data mining, and its application in the issues related to validation. The third chapter focuses on the research methodology and provides a complete description of how the data have been collected and the variables have been defined. Also, in this chapter, the research structure based on Crisp methodology is explained step by step to explain all the steps. Then, the results of data processing and implementation of different models with the help of calculations, comparisons, and analysis of results based on the Crisp method are examined in detail. Finally, the research achievements and their innovative aspects are discussed, and the suggested tips for future research are stated.

Literature Review

The findings suggest that combining financial and non-financial elements leads to a more accurate prediction of future default occurrences than separately using each of these components (Grunert et al., 2005). The conventional technique of credit scoring relied on statistical analysis and the common judgment method. However, the novel techniques are centered on utilizing social media data, mobile data, and psychometrics (Hendricks & Budree, 2019). Classical credit reporting, which is primarily focused on financial data, may only provide such good assessment of SMEs that are restricted in financial data but flourishing in non-financial data, like big data from business, government, social media, and networks (Yadi et al., 2019). The significance of identifying financial and non-financial indicators to evaluate the credit risk associated with microfinance loan applications is inevitable (Abdullah et al., 2020).

Since some consumers' credit information is inadequate or unavailable, credit managers cannot assess their actual credit position. This difficulty, nevertheless, may be efficiently handled with the use of social data, particularly behavioral data, and a credit rating system (Yu et al., 2019). Social media has become more significant in spreading individual perspectives on a

wide range of financial topics, constituting credit risk in investment decisions (Fei et al., 2015). With the collection of data on personal behavior and the advancement of machine learning techniques, social media data may now be used for personal credit rating (X. Yu et al., 2020). When financial or non-financial evaluations are absent or inaccurate, and credit analysts' subjective opinions affect the decision, social media data may be precious in evaluating organizations' credibility (Gül et al., 2018).

Credit scoring is becoming a reality due to the advancement of big data and machine learning models and algorithms (X. Yu et al., 2020). Artificial neural networks will allow the final credibility score to be calculated based on the different obtained data characteristics. Through SMS, a smartphone application will gather bank transaction data and data connected to internet transactions. Adding qualitative variables such as the depth and breadth of a person's social media networks and alternative payment data refines the algorithms used to calculate credit scores (Lohokare et al., 2017).

The proposed study effort intends to contribute to examining the behavioral features of online influencers is accomplished by calculating credit scores and examining the time development of these scores and time-series trends (Daou, 2019).

Related to the basis of the credit scoring model based on the decision tree algorithm, the borrowed credit score, the number of successes, prestige, the number of failures, repayment duration, and forum currency is the most important characteristics for forecasting default based on the credit scoring model and classification criteria (Zhang et al., 2016).

The Long Short-Term Memory model dynamically includes daily news from social media to account for market players' and public opinion views. Leveraging unstructured news crawled from social media to mitigate the influence of financial fraud on default probability prediction. Also, a neural network technique is presented for default probability prediction that integrates both structured financial parameters and unstructured social media data with proper time alignments (Zhao et al., 2019).

To build a credible personal credit evaluation system using social media data, to identify abnormal users, the logistic regression approach was utilized to score the individual credits of users before and after cleaning the data and discovered that the rank order of personal credit scoring has changed substantially (X. Yu et al., 2020).

Through the utilization of Logistic Regression, while combining a variety of information sources such as consumer purchasing habits, social media activity, and geography details with conventional credit scoring methodologies yields fresh insights for appropriate and more accurate credit scoring (Hindistan et al., 2019).

To verify borrowers employing classification and regression trees, digital lenders usually gather vast quantities of data from their clients, such as communication patterns, social media activity data, and extensive mobile phone usage (Shema, 2019). By improving an ensemble learning model with a random forest, a novel computerized approach for extracting public emotions encoded in social postings is created to enhance standard financial indicators, like return-on-assets, for forecasting corporate credit ratings (Yuan et al., 2018). Regardless of whether a person has incredible, good, or bad credit, it is critical to review their credit situation. Money-related reliability, as well as a person's overall identity characteristic, are needed. This may be determined by a person's social media interaction with the means of the Naive Bayes classifier. Combining these two scores will provide the weighted advanced credit score, which will be more precise than the current method (Kulkarni & Dhage, 2019).

Research questions

1. What are the factors affecting the validation of telegram channels active in the field of offering stock signals?
2. Due to what market conditions can the credibility of telegram channels active in the field of stock offering be trusted?

Research Methodology

The Cross-industry standard process expands on earlier efforts to describe knowledge discovery techniques, commonly known as CRISP-DM, which are applied for data mining. The six critical steps in the process are provided below (Wirth & Hipp, 2000):

1. Business understanding
2. Data understanding
3. Data preparation
4. Modeling
5. Evaluation
6. Deployment

1. **Business understanding:** This is the first stage in which we define the business's goal. In this paper, the goal is to figure out how stockholders can rely on telegram channels' social behavior to gain one-month returns. Among the purposes pursued in this area is to identify financial factors affecting the identification of high-credit channels, criteria related to their social activity, and determine the impact of each of the variables affecting the credit of active Telegram channels to understand shareholders. Improving the current situation and providing solutions to properly evaluate the active telegram channels in stocks have been among the objectives of this research. The RFM model is used to understand the channel conditions in different groups. Sahmeto.com is an intelligent stock social analysis system that helps the audience with everything they need from stock exchange information and social media to choose good stocks, trades and appropriate times. With the help of this system, we crawled Telegram channels data in a time of six months.
2. **Data understanding:** Data gathering is the first stage in this phase. We begin analyzing the data when it has been collected. We investigate data insights and detect quality issues, as well as discover intriguing data subsets. This type of data analysis is a crucial stage in the study of the data. The period that was selected is approximately from July 2020 to January 2021. Moreover, 270 telegram channels were selected, based on existing data, to identify their social behavior.
3. **Data preparation:** It includes all actions necessary to create the final dataset from raw data. This step is used to choose the table, and attributes, and to transform/separate junk data/from clean and useful data for modeling techniques. In this study, the available data are actual figures and statistics obtained from active telegram channels in stock offers by real users. In this study, analysis was performed on all data, and no sampling was performed.

Based on the three criteria in the RFM algorithm, a database of the performance of telegram channels in these three sections was created (Birant, 2011). Part of this database can be seen in the following tables:

1. **The recency of the last broadcast signal (R):** This factor represents the last time the signal was released, which means the time interval of the last broadcast signal of the current channels. The shorter the time interval, the larger the value.
2. **Signal Frequency (F):** This factor indicates repetition. It represents the number of signals emitted by the channels in the time interval.

3. The monetary value of broadcast signals (M): This factor expresses monetary value. In this study, the difference between the performance of the right and wrong signs in each of the broadcasted signals of the channels has been calculated. In this variable, if they function correctly, the buy and sell signals must have led to the growth and decline of the stock, respectively. Neutral signs have also been removed to reduce complexity.

Table 1. RFM dataset before Normalization

publisher_id	Recency	Frequency	Monetary
24	0	87	0.777777778
25	0	73	-0.098445596
26	0	86	-0.269565217
27	0	122	0.185185185
28	0	85	-0.53271028
29	0	99	-0.257831325
30	3	100	-0.676767677
31	0	93	-0.196428571
33	0	114	-0.481927711
34	14	77	-0.401273885
35	0	88	0.300492611
37	4	66	-0.348148148
38	0	68	0.335877863
39	4	26	0
40	13	18	-0.866666667

Table 2. RFM dataset before Normalization

publisher_id	Monetary	Frequency	Recency
24	0.8888889	0.6640625	1
25	0.4507772	0.5546875	1
26	0.3652174	0.65625	1
27	0.5925926	0.9375	1
28	0.2336449	0.6484375	1
29	0.3710843	0.7578125	1
30	0.1616162	0.765625	0.9795918
31	0.4017857	0.7109375	1
33	0.2590361	0.875	1
34	0.2993631	0.5859375	0.9047619
35	0.6502463	0.671875	1
37	0.3259259	0.5	0.9727891
38	0.6679389	0.515625	1
39	0.5	0.1875	0.9727891
40	0.0666667	0.125	0.9115646
41	0.125	0.7421875	1

In the next step, a dataset related to understanding the state and importance of both financial and social criteria, which are imported from the literature review, is created:

Financial criteria:

1. Stock return rate: The following formula is used to compare the returns of shares. This return is compared to time stability. In such a way, the considered periods are compared with each other seasonally, annually, etc. The efficiency achieved in fewer time intervals is the best (Kumar et al., 2021).

$$\text{Return} = \frac{\text{Actual End Price}}{\text{Start Price} * \text{Stock Holding Period}}$$

Due to the limited time frame of the present study, this return has been considered only for thirty days. In this study, the denominator of the deduction is the final price of the share on the first day in thirty days, and the numerator also includes the final price thirty days later. Finally, the average of these returns for the channels is calculated (to reduce complexity, records for which there was no end price were separated from the others).

2. Performance score: The following formula has been used to calculate the performance score for the rate of positive and negative emotions published on Twitter, which includes the amount of difference between negative and positive emotions of each user divided by the total difference between negative and positive emotions of users (Leitch & Sherif, 2017).

$$\text{Score} = \frac{\sum POS - \sum NEG}{\text{Number of POS} - \text{Number of NEG}}$$

In this study, according to the previous variable formula, the signals emitted by each channel, the difference between the performance of the right and wrong signals in each channel divided by all the signals with a positive and negative performance of the channels has been calculated. In this variable, if they function correctly, the buy and sell signals must have led to the growth and decline of the stock, respectively. In this case, TRUE and otherwise FALSE are considered. Neutral signals have also been removed to reduce complexity.

Social criteria:

1. Normal activity rates: The behavior of some users on social networks is not normal in such a way that they publish a lot of information in a limited

period and then remain silent for a long time. Scores greater than 0.05 and less than 10 are considered good performance in this criterion. The denominator of the deduction includes the number of user activities (in this study, the number of published signals), and the denominator of the deduction consists of the difference between the first and last days of signal propagation in the interval (X. Yu et al., 2020).

$$A_i = \frac{n_i}{T_i}$$

In the present study, this criterion is calculated weekly and in total.

2. Monthly average of published signals: To calculate the past behavior in relation to the loan recharge, the frequency of this recharge every month has been calculated as validation variables (Shema, 2019).

This study includes the total number of signals published in the period in question and the average weekly and monthly number of channel signals to offer shares.

3. Monthly average of active days to spread the signal: To calculate the past behavior concerning loan characteristics, the amount of active days after activating the account on a monthly basis has been calculated as validation variables (Shema, 2019).

This study includes the total number of active days to broadcast the signal in the period and the average weekly and monthly number of active days of the channels to post.

Table 3. Validation variables from review studies

A	B	C	D	E	F	G	H	I	J	K
publisher_id	rate_sig_act_t	rate_sig_act_w	weekly_day_mean	monthly_day_mean	total_day_active	weekly_sig_mean	monthly_sig_mean	total_sig	return_30d_later	sent_score_finalchoice
168	19.65546218	8.735191638	11.03846154	35.875	119	96.42307692	313.375	2339	0.032451237	TRUE
265	18.04854369	5.77173913	14.15384615	46	103	81.69230769	265.5	1859	0.030811681	FALSE
105	21.65384615	9.803278689	7.038461538	22.875	78	69	224.25	1689	0.031528724	FALSE
283	12.77777778	4.628361858	15.73076923	51.125	126	72.80769231	236.625	1610	0.030813582	TRUE
27	12.21311475	10.18120805	5.730769231	18.625	122	58.34615385	189.625	1490	0.036708572	FALSE
100	11.70588235	6.817351598	8.423076923	27.375	119	57.42307692	186.625	1393	0.033081061	FALSE
305	15.98837209	4.296675192	15.03846154	48.875	86	64.61538462	210	1375	0.036264538	FALSE
302	14.10989011	4.03562341	15.11538462	49.125	91	61	198.25	1284	0.035106952	FALSE
159	9.746153846	4.934256055	11.11538462	36.125	130	54.84615385	178.25	1267	0.03104859	FALSE
222	10.53333333	4.34502924	13.15384615	42.75	120	57.15384615	185.75	1264	0.033784086	FALSE
116	10.30833333	5.733050847	9.076923077	29.5	120	52.03846154	169.125	1237	0.029737585	TRUE
338	10.22807018	3.327433628	17.38461538	56.5	114	57.84615385	188	1166	0.032864204	FALSE
121	10.69724771	5.595652174	8.846153846	28.75	109	49.5	160.875	1166	0.031440153	TRUE
90	9.888888889	6.024154589	7.961538462	25.875	117	47.96153846	155.875	1157	0.030603866	FALSE
235	10.18584071	3.982758621	13.38461538	43.5	113	53.30769231	173.25	1151	0.032421765	FALSE
257	13.43902439	4.08849558	13.03846154	42.375	82	52.26923077	169.875	1102	0.031715149	FALSE
268	14.09090909	3.92173913	13.26923077	43.125	77	52.03846154	169.125	1085	0.03022966	FALSE
83	10.27835052	6	6.923076923	22.5	97	41.53846154	135	997	0.032839482	FALSE
258	11.30232558	3.575581395	13.23076923	43	86	47.30769231	153.75	972	0.034078051	FALSE
254	9.979166667	3.462857143	13.46153846	43.75	96	46.61538462	151.5	958	0.02718941	FALSE
233	9.316831683	3.51497006	12.84615385	41.75	101	45.15384615	146.75	941	0.030201498	FALSE
145	10.32183908	4.495689655	8.923076923	29	87	40.11538462	130.375	898	0.027069243	FALSE
161	8.613861386	3.935114504	10.07692308	32.75	101	39.65384615	128.875	870	0.030713636	TRUE
276	10.97435897	3.197740113	13.61538462	44.25	78	43.53846154	141.5	856	0.029252615	FALSE
155	10.1547619	4.217573222	9.192307692	29.875	84	38.76923077	126	853	0.030754561	FALSE
107	9.161290323	4.795	7.692307692	25	93	36.88461538	119.875	852	0.030651987	FALSE
38	12.33823529	8.273584906	4.076923077	13.25	68	33.73076923	109.625	839	0.034748785	TRUE
346	19.25581395	3.017994859	14.96153846	48.625	43	45.15384615	146.75	828	0.027941517	FALSE

4,5. Modeling and Evaluation: For modeling, we use data mining techniques in this step. Then, for evaluation, one of the primary goals is to determine whether any critical business issues have been overlooked. A choice on how to use the data mining results should be made near the end of this stage.

Our example study utilized a basic K-means method to define multiple clusters based on the RFM criteria. We used the R tool for the clustering process, a free software environment for analyzing data and implementing statistical computing and graph representation (Hornik, 2012). The K-means clustering method attempts to arrange related objects into groups. K represents the number of groupings. K-means clustering is a well-known and very effective unsupervised machine learning technique. It is used to address a wide range of complex unsupervised machine learning issues (Malik, 2011). The elbow technique is the most used strategy for addressing this issue when using the k-means clustering algorithm. We compute WCSS, the within-cluster sum of a square, for each value of K. WCSS is the sum of the squared distances between each point in a cluster and its centroid. When we display the WCSS with the K value, we get an Elbow. The WCSS value decreases as the number of clusters rises. (Syakur et al., 2018).

By implementing the elbow method, four clusters are the best choice for this dataset (as it is shown in the following figure).

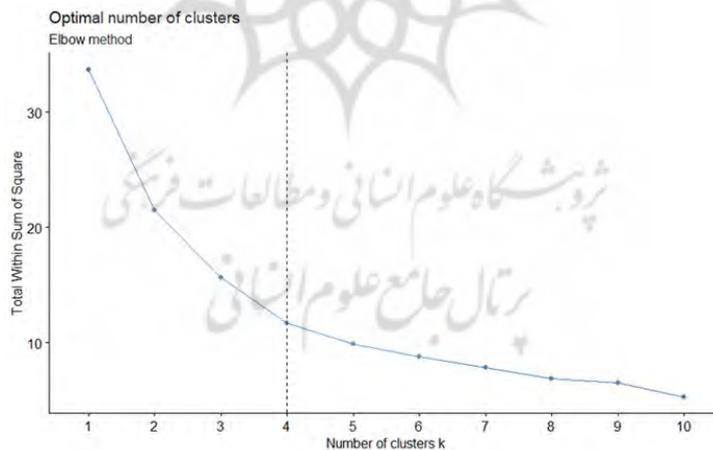


Figure 1. The Elbow Method

The center of each cluster and other information is detailed in table 4 :

Table 4. Cluster Centers

Cluster Number	Recency	Frequency	Monetary	Total number of telegram channels in each cluster
۱	۰,۹۹۱۲۶۲۳	۰,۶۹۱۴۰۶۲۵	۰,۳۱۷۲۴۷۳	۱۰۲
۲	۰,۸۷۹۵۹۱۸	۰,۲۲۶۱۷۱۸۷	۰,۲۱۷۵۵۴۰	۶۰
۳	۰,۵۰۱۰۲۸۴	۰,۰۹۴۶۶۹۱۲	۰,۸۱۹۳۵۱۷	۱۷
۴	۰,۹۶۲۰۹۹۱	۰,۴۳۱۳۱۸۶۸	۰,۵۶۲۸۹۸۳	۹۱

The view of the scattering of clusters is illustrated in figure 2:

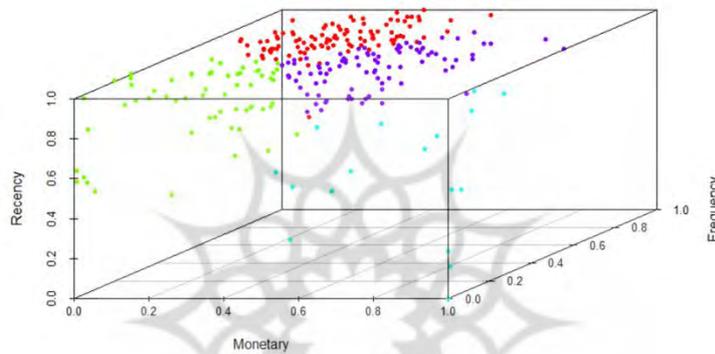


Figure 2. Clusters Scatter Plot

Also, we used the Python tool, which has proven to be a complete programming solution to advance data science goals. Python's ease of learning and flexibility make it one of the fastest-growing languages. Python's evolving libraries also make it a convenient option for data analysis (Nagpal & Gabrani, 2019). We have used a Tree-based Pipeline Optimization Tool (TPOT) to create machine learning pipelines with competitive classification accuracy and identified innovative pipeline operators (Olson et al., 2016). The results of running this algorithm on the validation variables are as follows: ExtraTreesClassifier is proposed as the algorithm that performs best. ExtraTreesClassifier is an ensemble learning approach, also known as Extremely Randomized Trees (Bhati & Rai, 2020). It mainly consists of heavily randomizing both attribute and cut-point selection while splitting a tree node. In the most extreme example, it generates completely randomized trees with models independent of the learning sample's output values (Geurts et al., 2006). These algorithms were implemented in the Python tool. Then, cross-

validation was implemented, which is a resampling procedure used to evaluate machine learning models. For each cluster, the accuracy of implementing this algorithm is: cluster1=0.705, cluster2=0.717, cluster3=0.764, cluster4=0.714.

It should be noted that the performance score variable is considered the target variable because it means the channel's overall performance, in terms of rising stock prices following the signal, compared to other channels. As a result, the importance of other criteria is considered concerning this variable.

```

Generation 1 - Current best internal CV score: 0.7148148148147
Generation 2 - Current best internal CV score: 0.7148148148147
Generation 3 - Current best internal CV score: 0.7148148148147
Generation 4 - Current best internal CV score: 0.7160493827160491
Generation 5 - Current best internal CV score: 0.7160493827160491

Best pipeline: ExtraTreesClassifier(RBFSampler(input_matrix, gamma=0.75), bootstrap=False, criterion=entropy, max_features=0.6000000000000001, min_samples_leaf=7, min_samples_split=12, n_estimators=100)
    
```

Figure 3. TPOT result

Feature importance and summary of variables are shown in the following figures:

Cluster 1:



Figure 4. Factors of importance in cluster 1

	rate_sig_act_t	rate_sig_act_w	weekly_day_mean	monthly_day_mean	total_day_active	weekly_sig_mean	monthly_sig_mean	total_sig	return_30d_lat
count	102.000000	102.000000	102.000000	102.000000	102.000000	102.000000	102.000000	102.000000	102.000000
mean	7.026004	3.393211	10.177802	33.077206	90.500000	32.004148	104.013480	657.990196	0.0317
std	3.724983	1.737210	3.658735	11.890890	15.948401	16.084827	52.275687	405.971601	0.0021
min	2.519481	1.269341	3.961538	12.875000	60.000000	11.961538	38.875000	154.000000	0.0259
25%	4.478448	2.041414	6.951923	22.593750	78.000000	21.326823	69.312500	374.750000	0.0304
50%	5.983696	3.147870	9.846154	32.000000	88.000000	27.115385	88.125000	544.500000	0.0315
75%	9.277946	4.221701	13.451923	43.718750	99.750000	39.432892	128.156250	845.750000	0.0331
max	21.653846	10.181208	17.384615	56.500000	130.000000	96.423077	313.375000	2339.000000	0.0378

Figure 5. Factors statistics in cluster 1

In figures 6,9,12 and 15, line charts named 1 and 2 are the positive and negative performance of signals, respectively.

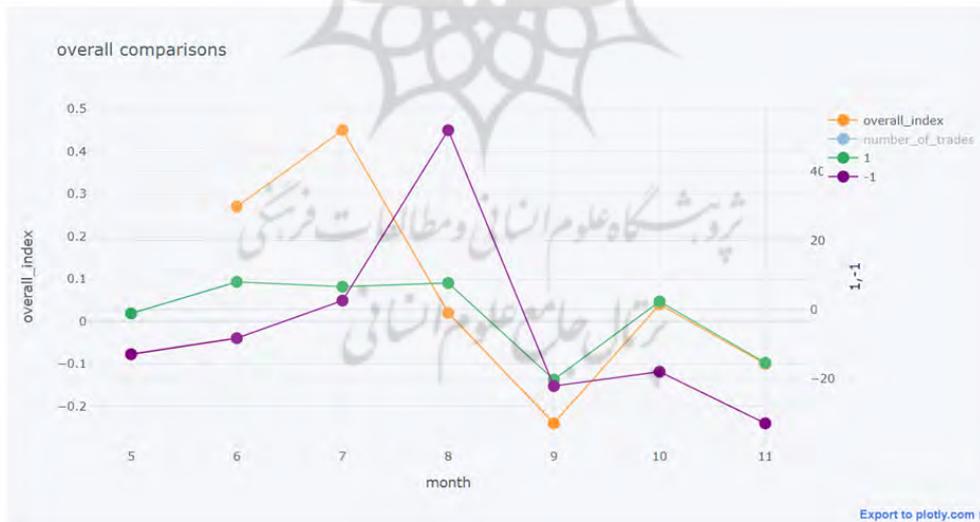


Figure 6. signal trends in cluster 1

Table 5. Channels Analysis

clusters	The number of channels included	Analysis results
1	102	<p>Examining the behavior of these channels regarding the natural activity rate around stock issuance, we conclude that they are more inclined to propagate large volumes of signal in a limited period. The performance ratio of the false signals of this channel is increasing in times of unfavorable market situations when the overall index is declining. In other words, the state of propagation of the correct signals in this category is not very desirable.</p> <p>Therefore, to measure the performance score of these channels in stock signal propagation, we can pay attention to the total number of active days around signal propagation and their activity rate, which, compared to other clusters, tends to create pressure around signal propagation during limited weeks or days.</p>
2	60	<p>Examining the behavior of these channels regarding the natural activity rate around the stock release, we conclude that unlike the behavior of the first cluster channels, they do not have much tendency to propagate a large volume of the signal in a limited period. The performance ratio of the incorrect signals of this channel increases at the beginning when the total index is rising and then falling, but gradually this ratio decreases. Similar to the first cluster, the state of propagation of correct signals in this category is not very desirable.</p> <p>Also, to measure the performance score of these channels in stock signal propagation, we can pay attention to the total number of active days around signal propagation and their activity rate in weekly intervals. However, compared to the first cluster, less tendency to create pressure around Signal propagation is limited to weeks or days. However, some channels in the cluster still propagate large volumes of signal over a limited period.</p>
3	17	<p>Examining the behavior of these channels regarding the activity rate around the issuance of stocks, we conclude that they have a more natural behavior than the previous two clusters. This means that they do not emit a large signal volume in a limited period. The performance ratio of the correct signals of this classification is observed when the condition of the overall index is improving. The performance ratio of the incorrect signals of this channel increases at the beginning when the total index rises and then falls. Still, like the first two clusters, this ratio gradually decreases. These 17 channels have the most favorable state of propagation of correct signals.</p> <p>Also, to measure the performance score of these channels in the release of the stock offer signal, we can pay attention to the one-month returns and their average monthly activity. The channels in this cluster seem to have more credibility, but only if the overall</p>

		index is experiencing positive growth.
4	91	<p>Examining the behavior of these channels concerning the natural activity rate around the release of stocks, we conclude that there is less tendency to exert pressure on the release of signals during the week or limited days. However, there are still channels in this cluster that release large volumes of signal in specific periods. The performance ratio of the correct signals in this category is better than the first and second clusters. This correct release situation is mostly seen at the time of the rise of the total index and to some extent at the time of the decline of this index. The performance ratio of the incorrect signals of this channel increases at the beginning when the total index rises and then falls. Still, as in the previous three clusters, this ratio gradually decreases.</p> <p>Compared to the first and second clusters, the channels of this cluster have a better situation in transmitting the correct signals but still have a lower figure than the accurate financial performance of the third cluster. Also, to measure the performance score of these channels in the release of the stock offer signal, we can pay attention to their average weekly and monthly activity.</p>

4. Deployment: During the deployment process, we have created a report that includes the results. Through clustering, the telegram channels, hidden patterns in their behavior can be traced to identify the validity of their signal and improve shareholder decisions about relying on these signals to buy or sell shares.

Discussion and Conclusion

Now, using the data and results presented in this research, channels active in the field of stock signal propagation have been categorized, and for each category, an appropriate validation strategy should be adopted so that firstly, shareholders choose reputable channels to rely on for their analysis and secondly, channels are encouraged to improve the quality of their transmitter signals. With the help of this research, it is possible to validate the accuracy of the financial status of channels active in the field of stock signal propagation according to their different activities in the Telegram social network.

Now we should answer the research questions:

1. What are the factors affecting the validation of telegram channels active in the field of offering stock signals?

According to the literature review and the results of the algorithms implemented in this research are: Social variables include the rate of normal activity around the weekly and total stock signal propagation (abnormal behavior of some users on social networks is such that they publish a large amount of information in a limited period, and then do not send signals for a relatively long time) / weekly, monthly average, and the total number of signals and the weekly, monthly average and the total number of activity days around the signal propagation. Financial variables also include the rate of return on stocks at one-month intervals and the performance score (the difference between the right and wrong signals per channel over the total signals with positive and negative channel performance). Among the mentioned criteria, to identify the channels whose stock financial return status is less consistent with the type of their published signals, we can refer to the total number of active days when the stock signal is released (in total desired return) and calculate the normal performance rate. These channels paid more attention to this (as weekly reviews). The fact that channels whose number of activities peaks each week on certain days and then do not work cannot be trusted in the one-month horizon to the returns of stocks that have recommended buying or selling. Also, identifying channels with more credibility for one-month returns does not necessarily mean that they have the highest daily activities or the highest proportion of signals.

2. Due to what market conditions can the credibility of telegram channels active in the field of stock offering be trusted?

Examining the clusters; the primary and most reliable part when viewing the published signals around stock trading is that the channels usually perform more correctly around the forecast for stock buying or selling when the overall index grows positively. It would not be wise to rely on telegram channel signals when the growth of the overall index is negative, and the stock market is not in a favorable position. Also, the high number of activity days and the high amount of channel signals do not necessarily mean that they can give high credits to the returns of their proposed stocks in the one-month horizons.

In any research, some limitations reduce the accuracy of doing so. There are limitations in this research as well. The most important of these are:

1. Limited access to data in a short period, six months
2. Only 270 channels that could be shared between variables were extracted. Their number can be increased in future research.

Research suggestions: Other researchers are suggested to take action in the following cases:

1. If it is possible to access the data for a more extended period, in addition to using the RFM model, the LRFM model can also be used. This model, along with the previous three variables, has the variable L, which indicates the duration of the channels' activity in propagating the stock signal in the telegram.
2. The weight of RFM variables in this study is the same. But depending on the industry of research and the opinion of experts, these weights can have different degrees of importance.
3. A combination of other clustering algorithms can be used in future research.
4. Using neutral signals (neither buy nor sell) in credit scoring
5. Validation variables have been extracted according to the literature review. In future research, the opinion of stock industry experts can also be used in this regard.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest concerning the research, authorship and, or publication of this article.

Funding

The authors received no financial support for the research, authorship and, or publication of this article.

References

- Abdou, H. A., & Pointon, J. (2011). Credit Scoring, Statistical Techniques And Evaluation Criteria: A Review Of The Literature: Credit Scoring, Techniques & Evaluation Criteria: A Literature Review. *Intelligent Systems In Accounting Finance & Management*, 18(2-3), 59-88.
- Abdullah, M. A., Ahmad, A., Nayan, N. A. M., Azhar, Z., & Ahmad, A.-R. (2020). Credit Risk Assessment Models Of Retail Microfinancing: The Case Of A Malaysian National Savings Bank's Branch. *International Journal Of Financial Research*, 11(3), 73.
- Bhati, B. S., & Rai, C. S. (2020). Ensemble-Based Approach For Intrusion Detection Using Extra Tree Classifier. In *Intelligent Computing In Engineering* (Pp. 213-

- 220). Springer Singapore.
- Birant, D. (2011). Data Mining Using RFM Analysis. In Knowledge-Oriented Applications In Data Mining. Intech.
- Brownlee, J. (2016, June 2). *Ensemble Machine Learning Algorithms In Python With Scikit-Learn*. Machinelearningmastery.Com. <https://Machinelearningmastery.Com/Ensemble-Machine-Learning-Algorithms-Python-Scikit-Learn/>
- Daou, H. (2019). Identifying Influencers Using Time Series Analysis. *2019 Sixth International Conference On Social Networks Analysis, Management And Security (SNAMS)*.
- Doumpos, M., Lemonakis, C., Niklis, D., & Zopounidis, C. (2019). *Introduction To Credit Risk Modeling And Assessment*. Springer.
- Fei, W., Gu, J., Yang, Y., & Zhou, Z. (2015). Credit Risk Evaluation Based On Social Media. *Procedia Computer Science*, 55, 725–731.
- Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely Randomized Trees. *Machine Learning*, 63(1), 3–42.
- Grunert, J., Norden, L., & Weber, M. (2005). The Role Of Non-Financial Factors In Internal Credit Ratings. *Journal Of Banking & Finance*, 29(2), 509–531.
- Gül, S., Kabak, Ö., & Topcu, I. (2018). A Multiple Criteria Credit Rating Approach Utilizing Social Media Data. *Data & Knowledge Engineering*, 116, 80–99.
- Hendricks, M. K., & Budree, A. (2019). Can A Mobile Credit-Scoring Model Provide Better Accessibility To South African Citizens Requiring Micro-Lending? *International Journal Of Electronic Finance*, 9(3), 157.
- Hindistan, Y. S., Kiyakoglu, B. Y., Rezaeinazhad, A. M., Korkmaz, H. E., & Dag, H. (2019). Alternative Credit Scoring And Classification Employing Machine Learning Techniques On A Big Data Platform. *2019 4th International Conference On Computer Science And Engineering (UBMK)*.
- Hornik, K. (2012). The Comprehensive R Archive Network: The Comprehensive R Archive Network. *Wiley Interdisciplinary Reviews. Computational Statistics*, 4(4), 394–398.
- Kulkarni, S. V., & Dhage, S. N. (2019). Advanced Credit Score Calculation Using Social Media And Machine Learning. *Journal Of Intelligent & Fuzzy Systems*, 36(3), 2373–2380.
- Kumar, R., Mukherjee, S., Kumar, B., & Bala, P. K. (2021). Deriving Wisdom From Virtual Investing Communities: An Alternative Strategy To Stock Recommendations. *Journal Of Modelling In Management*, 16(2), 668–688.
- Lee, T.-S., Chiu, C.-C., Lu, C.-J., & Chen, I.-F. (2002). Credit Scoring Using The

- Hybrid Neural Discriminant Technique. *Expert Systems With Applications*, 23(3), 245–254.
- Leitch, D., & Sherif, M. (2017). Twitter Mood, CEO Succession Announcements And Stock Returns. *Journal Of Computational Science*, 21, 1–10.
- Lohokare, J., Dani, R., & Sontakke, S. (2017). Automated Data Collection For Credit Score Calculation Based On Financial Transactions And Social Media. *2017 International Conference On Emerging Trends & Innovation In ICT (ICEI)*.
- Malik, P. P. L. G. (2011). Generating Customer Profiles For Retail Stores Using Clustering Techniques. " *International Journal On Computer Science And Engineering (IGCSE)*, 3(6).
- Marikkannu, P., & Shanmugapriya, K. (2011). Classification Of Customer Credit Data For Intelligent Credit Scoring System Using Fuzzy Set And MC2 — Domain-Driven Approach. *2011 3rd International Conference On Electronics Computer Technology*.
- Nagpal, A., & Gabrani, G. (2019). Python For Data Analytics, Scientific And Technical Applications. *2019 Amity International Conference On Artificial Intelligence (AICAI)*.
- Olson, R. S., Urbanowicz, R. J., Andrews, P. C., Lavender, N. A., Kidd, L. C., & Moore, J. H. (2016). Automating Biomedical Data Science Through Tree-Based Pipeline Optimization. In *Applications Of Evolutionary Computation* (Pp. 123–137). Springer International Publishing.
- Shema, A. (2019). Effective Credit Scoring Using Limited Mobile Phone Data. *Proceedings Of The Tenth International Conference On Information And Communication Technologies And Development - ICTDX '19*.
- Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., & Satoto, B. D. (2018). Integration K-Means Clustering Method And Elbow Method For Identification Of The Best Customer Profile Cluster. *IOP Conference Series. Materials Science And Engineering*, 336, 012017.
- Thomas, L. C. (2000). A Survey Of Credit And Behavioural Scoring: Forecasting Financial Risk Of Lending To Consumers. *International Journal Of Forecasting*, 16(2), 149–172.
- Wei, Y., Yildirim, P., Van Den Bulte, C., & Dellarocas, C. (2016). Credit Scoring With Social Network Data. *Marketing Science*, 35(2), 234–258.
- Wirth, R., & Hipp, J. (2000). *CRISP-DM: Towards A Standard Process Model For Data Mining. 1*.
- Yadi, L. I. U., Yuning, S., Jiayue, Y. U., Yingfa, X. I. E., Yiyuan, W., & Xiaoping, Z. (2019). Big-Data-Driven Model Construction And Empirical Analysis Of Smes Credit Assessment In China. *Procedia Computer Science*, 147, 613–619.

- Yu, J., Yao, J., & Chen, Y. (2019). Credit Scoring With AHP And Fuzzy Comprehensive Evaluation Based On Behavioural Data From The Weibo Platform. *Tehnički Vjesnik*, 26(2), 462–470.
- Yu, X., Yang, Q., Wang, R., Fang, R., & Deng, M. (2020). Data Cleaning For Personal Credit Scoring By Utilizing Social Media Data: An Empirical Study. *IEEE Intelligent Systems*, 1–1.
- Yuan, H., Lau, R. Y. K., Wong, M. C. S., & Li, C. (2018). Mining Emotions Of The Public From Social Media For Enhancing Corporate Credit Rating. *2018 IEEE 15th International Conference On E-Business Engineering (ICEBE)*.
- Zhang, Y., Jia, H., Diao, Y., Hai, M., & Li, H. (2016). Research On Credit Scoring By Fusing Social Media Information In Online Peer-To-Peer Lending. *Procedia Computer Science*, 91, 168–174.
- Zhao, Y., Shen, Y., & Huang, Y. (2019). DMDP: A Dynamic Multi-Source Default Probability Prediction Framework. *Data Science And Engineering*, 4(1), 3–13.

Bibliographic information of this paper for citing:

Sohrabi, Babak; Khalili Jafarabad, Ahmad & Orfi, Saba (2022). Credit Scoring Active Telegram Channels Offering Stock Signals. *Iranian Journal of Finance*, 6(3), 142-164.

Appendix

RFM: variables and descriptive statistics

	Monetary	Frequency	Recency
Mean	0.409500704	0.462789352	0.93015873
Standard Error	0.013150358	0.014567467	0.008869728
Median	0.390037205	0.48828125	1
Mode	0	0.6640625	1
Standard Deviation	0.216082438	0.239367913	0.1457445
Sample Variance	0.04669162	0.057296998	0.021241459
Kurtosis	0.463067571	-0.844288437	12.62472802
Skewness	0.53994942	-0.050957292	-3.263520521
Range	1	1	1
Minimum	0	0	0
Maximum	1	1	1
Sum	110.56519	124.953125	251.1428571
Count	270	270	270
Largest(1)	1	1	1
Smallest(1)	0	0	0
Confidence Level(95.0%)	0.025890715	0.02868075	0.017462915

Financial and Non-Financial features: variables and descriptive statistics

	weekly_day_mean	monthly_day_mean	total_day_mean	weekly_sig_mean	monthly_sig_mean	total_sig	rate_activity	rate_signal_activity	%return_30d_later	senti_score_finalchoice
Mean	2.090281124	7.46528973	5.000102516	12.24991727	0.777969019	306.2329317	1.982443221	149.6065879	0.033886557	0.273092369
Standard Error	0.065925315	0.235447554	0.225376299	0.795328837	0.01199719	19.88322093	0.098301898	12.47104143	0.000273029	0.02829232
Median	2.2	7.857142857	3.909090909	8.88	0.857142857	222	1.514229738	92.43396226	0.033120612	0
Mode	2.2	7.857142857	4.454545455	9.8	0.857142857	245	1	117.079646	0.037547907	0
Standard Deviation	1.040283926	3.715299795	3.556378018	12.55007737	0.189312465	313.7519341	1.551177794	196.7897144	0.004308332	0.446445274
Sample Variance	1.082190646	13.80945212	12.64782461	157.5044419	0.035839209	98440.27617	2.406152548	38726.1917	1.85617E-05	0.199313383
Kurtosis	-0.650635552	-0.650635552	4.751712867	6.257895611	0.848531235	6.257895611	15.87595105	10.99440502	5.940651232	-0.957667137
Skewness	-0.052711247	-0.052711247	2.003788003	2.130999977	-1.050638178	2.130999977	3.304172946	2.792745946	1.564540183	1.024738989
Range	4.6	16.42857142	20.7761194	84.88	0.857142857	2122	11.34428259	1456.761943	0.033263351	1
Minimum	0.08	0.285714286	1	0.08	0.142857143	2	1	0.084210526	0.023855991	0
Maximum	4.68	16.71428571	21.7761194	84.96	1	2124	12.34428259	1456.846154	0.057119342	1
Sum	520.48	1858.857143	1245.025527	3050.08	193.7142857	76252	493.6283619	37252.04037	8.437752719	68
Count	249	249	249	249	249	249	249	249	249	249
Largest(1)	4.68	16.71428571	21.7761194	84.96	1	2124	12.34428259	1456.846154	0.057119342	1
Smallest(1)	0.08	0.285714286	1	0.08	0.142857143	2	1	0.084210526	0.023855991	0
Confidence Level(95.0%)	0.129844896	0.46373177	0.443895672	1.566460314	0.023629373	39.16150786	0.193613026	24.56265957	0.000537752	0.055723864

We analysed mutual telegram channels which the related data is available for them.