Federal State Autonomous Educational Institution for Higher Education

NATIONAL RESEARCH UNIVERSITY
«HIGHER SCHOOL OF ECONOMICS»

Faculty «Saint Petersburg School of Economic and Management»
Department Economics

Garmanova Vera Ivanovna, Mironov Mikhail Vladimirovich

PREREQUISITES FOR IMPLEMENTATION OF BLOCKCHAIN TECHNOLOGY IN ENTERPRISE: AN EMPIRICAL STUDY OF USE CASES ACROSS INDUSTRIES

BACHELOR THESIS

EDUCATIONAL PROGRAMME 38.03.01 «Economics»

Students' group number № 191
Programme «Economics»

Supervisor:
Associate Professor, Department of Economics in HSE
Tatiana A. Lezina

Academic adviser:
Ph.D., Full Professor, Department of Economics and Business in UPF
Helena Ramalhinho

Saint-Petersburg
2023

ABSTRACT

Blockchain is becoming an important technology of this decade pushing humanity towards the Industrial Revolution 4.0. The last few years have shown that companies have seen significant advantages in the new technology not only for the finance industry, but also for use in various areas of business. This study aims to fill a gap in the research of the types of companies that need to introduce new technology in order to identify an idea of those companies that benefit from changing current business processes in the future. The authors of the current paper studied the TOP-2000 companies from the USA and China as the main competing countries seeking to implement blockchain. The study examined the industries and use cases of blockchain in order to answer the research question: what are the key characteristics of companies that need to change their business model and implement blockchain technology in order to be stable and profitable in the near future? While the types of companies were studied using clustering models, the assumption that only a part of the industry needs to implement a certain type of use cases has been tested within the framework of classification models. The results showed that companies from the financial, entertainment and IT industries need to take a closer look at their competitors and make a possible decision to implement blockchain in their companies in order not to become lagging behind among competitors in the nearest future. Thesis format: academic (empirical).

CONTENTS

INTRODUCTION

Background

Blockchain technology has started gaining momentum among different industries only a short while. For different businesses this technology gives a breakthrough concept of how to rebuild logistics chains, process and store information, conduct monetary transactions in a company. The capitalization of the blockchain market is expected to grow in the near future from 3 billion in 2020 to 39,7 billion by 2025 (Ashraf, 2021). For the financial system, monetary transactions based on blockchain are a serious alternative to the established financial system, for businesses - it is a stimulus to change business models to achieve higher results than their competitors on the market.

Despite the fact that the topic of blockchain is highly discussed, it is too young to form an unambiguous opinion about its positive and negative aspects. The business strives to introduce a disruptive technology as early as possible in order to bypass a competitor and be the first to occupy a niche, taking on a significant risk in case of failure. Full-fledged research studies usually appear later and evaluate these successes and failures of companies. Thus, the pioneers in such cases are always companies that make a decision based on intuition.

At this moment it can be observed that such a race is surrounded by uncertainty among companies and may last the nearest five-ten years. The first companies that successfully implement blockchain will gain a strategic advantage. Nevertheless, it is highly important to understand for each decision maker that not everyone will be a winner. Since the unexplored return on investment in such technology imposes a huge risk on the entrepreneur, these investments with a high probability will not pay off.

The study of key players who have now rushed into battle can help other players in making decisions. Large market sharks have huge development departments and several dozen analysts assess the prospects and directions of the company's future development. In this case, the rest of the companies can rely on the signals given by such pioneers. After weighing all the pros and cons, large companies are the first to decide that changing partly current business models has more advantages for them in the future. In this case, researchers can pay attention to those key characteristics of companies that have decided to implement or not blockchain technology in their business.

Despite the fact that all technical tools are available to researchers to measure the impact of the introduction of blockchain technology, only some researchers strive to do this. Thus, there have been almost no studies presented before 2018, which can be considered the year of the beginning of publications on the topic of blockchain implementation in companies (Happy et al., 2023). During

these few years, mainly works with a focus in the field of management were written. The economic approach and quantitative measurements of the market are poorly represented on this topic for several reasons. Firstly, the field is new and it is logical to argue that papers based on discussions and analysis appear first, and early hypotheses are formed in those works. Secondly, only by today it can be assumed that some database is accumulating on companies that have decided to introduce new technology over the years and have publicly announced that decision.

The vast majority of papers list the advantages of the technology and the scope of potential implementation of blockchain. For most companies such analytics are primarily suitable since company management is often based on familiar management schemes that do not require numerical measurement. With the development of a new field, researchers are gradually getting involved, who seek to measure numerically the impact of decisions made. However, quantitative measurement is not cost-effective in every case. It requires data, which are not always available and measurable, time and labour resources.

In this case, numerical measurement of blockchain technology is more relevant today than ever, since the return on such investments is not fully measured. Numerical evaluation is important for new growth points where the cost of error is high for small companies investing large resources in rebuilding business models. For such cases, companies look at the pioneers who are financially more stable in the market, and at their early successes, finally implementing when it is not too late a successful technology, but with less risk and lower costs.

However, there may be either the advantage of the first or second move. There are many examples in history when the pioneers went bankrupt or made less profit than the players who, following them in the second roles, got their place in the sun at a lower cost. At the same time, the theory of the advantage of the first move suggests that it is important to have time to grab a piece of the pie before competitors. This is a dilemma for companies.

The largest competing countries studying and implementing blockchain are the US and China. Both countries are striving for blockchain dominance. China nowadays extensively submit blockchain patent applications - over 84% of all applications worldwide in 2022[1] and according to various indicators of activity China has been signaling in recent years that it is going to take its all advantages from this technology[2]. It is estimated that China will receive $440 billion in 2030, or GDP growth of 1,7% followed by the US, which expects growth of $407,2 billion (Lin et al., 2022).

[1]Website "Cointelegraph" [Online resource]: `https://cointelegraph.com/news/china-accounts-for-84-of-all-blockchain-patent-applications-but-there-s-a-catch` (date of assessing: 02.01.2023)

[2]Website "FinancesOnline" [Online resource]: `https://financesonline.com/blockchain-statistics/` (date of assessing: 02.01.2023)

The current study seeks to reduce the ambiguity in the benefits of implementing blockchain technology by numerically measuring the characteristics of the TOP-2000 companies in the US and China. Studying the specifics of industries and the method of using blockchain companies that have decided to implement blockchain in their business model will give other companies a certain vector showing where the introduction of new technology is cost-effective.

Problem statement and research question

The current research is aimed at answering the questions "what" type of blockchain, "when" and "for whom" it is important to implement blockchain technology. This study is focusing on industries, types of blockchain use cases and financial characteristics of companies. The authors conduct research on various companies and identify how they are utilizing blockchain technology. Main goals include 1) revealing the main characteristics of the company that influence the decision on the introduction of blockchain technology, 2) studying the shares of blockchain use cases across industries, 3) identifying specific types of companies for which the implementation of blockchain is essential to increase their competitiveness in the market. The research question is what are the key characteristics of companies that need to change their business model and implement blockchain technology in order to be stable and profitable in the near future?

Relevance of the study

Despite potential benefits that blockchain technology offers, such as increased transparency, security and efficiency, automating data storage and eliminating intermediaries, many businesses are unsure about the payoff of actually implementing the new technology. It is valuable to study types of businesses and use cases for those companies that can benefit from new types of business models. The current study aims to partially fill the research gap and give a boost to new researchers for other quantitative economic research.

Contribution of the authors

Each of the authors sought to make the maximum and equivalent contribution to the research work. To evaluate the contribution of each of the participants, it is possible to list those sections to which a greater contribution was made by one of the authors. A greater contribution to the formulation of the research question, the study of relevant literature, primary data analytics and the construction of clustering models in belongs to Vera Garmanova, while the formation of the dataset and obtaining the main statistically significant result in the process of constructing various

classification models belongs to Mikhail Mironov. The final work is a single work in which the contribution of each of the participants is significant. Joint cooperation has allowed authors to achieve interesting results in a poorly empirically studied research area.

The remainder of this paper is organized as follows. “Literature review” describes the main concepts and finding of relevant studies, “Collection and primary data processing” provides the whole process of collecting required data, “Methodology” focuses on technical tools of quantitative measurements, “Results and discussions” provides the final results of the study, “Conclusion” sum up the paper.

1. LITERATURE REVIEW

1.1. Blockchain application in supply chain

Over the past few years, the number of research papers devoted to blockchain technology has increased significantly. Despite this, Happy et al. (2023) points out that until 2018 there was practically no work on the implementation of blockchain into the supply chain. Only a third of the 211 recent articles reviewed by the authors are quantitative, which is a distinctive feature of the young field of research. The number of articles varies by country (Lim et al., 2021). China (25 articles with 101 citations in a whole between 2017 and 2020 inclusive) and the United States (16 with 252 respectively) are the two countries most interested in blockchain technology.

The poor availability of preliminary data can be used to explain the lack of attention in empirical blockchain data and the limitations of our knowledge about its capabilities. In recent years, the volume of data has grown significantly, which has complicated data storage. As a result, a bright spot for the future is the scalable commercial use of blockchain technology in modern intelligent installations (Deepa et al., 2022). Blockchain is a peer-to-peer (P2P) technology that generates new types of business models with characteristics potentially capable of improving the organizational structure.

Cole et al. (2019) is one of the earliest publications focusing on the advantages of blockchain in terms of operations and supply chain management (OSCM). The authors emphasize that when considering the totality of previous studies, there is not enough evidence of when the introduction of blockchain is really necessary. To determine whether it is appropriate to introduce a new technology, which is not an easy task, managers should carefully study the characteristics of the product, supply chains and services. This gives the business a chance to soon surpass competitors and change its corporate structure. The new technology can be implemented in three stages. They are 1) adoption evaluation, 2) technology implementation, and 3) integration. The majority of businesses are in the evaluation or beginning implementation phases. It is now only feasible to quantify the incentives and intention of the company to incorporate blockchain in its business model, making it difficult to divide situations into full and partial adoption.

Product security or safety (Pfizer, AbbVie, Genentech, AmerisourceBergen, and McKesson), management quality improvement (Maersk and Renault), the elimination of intermediaries (Fairtrade), the reduction of illegal counterfeiting (Provenance), the establishment of reliable supply chain relationships (Kouvola Innovation), and inventory management (The New York Shipping Exchange, GE) are potential applications where implementing blockchain may be essential from

company to company. Looking at the experiences of large corporations will help small firms determine whether adopting new technologies is crucial to their company strategy. While second-place competitors like medium-sized and small firms might turn to the majors and learn from them in this race, big corporations, as leaders with financial sustainability, make a decision and explore multiple approaches to stay afloat.

The vast majority of recent quantitative research (Chowdhury et al., 2022), (Agi and Jha, 2022), (Deng et al., 2022), (Al-Zaqeba et al., 2022), (Jum'a, 2023), (Kurdi et al., 2023), (Pan et al., 2020) were questionnaire- and management-based. Although their results may be biased as a result of subjective data, it may still be interesting to compare them to the findings of the current study. The important thing that has to be noted is that these studies aim to provide answers to the issues of what impact incorporating blockchain technology into a business will have and which kinds of businesses would benefit from it.

However, in terms of approach and data, there is one significant study that is comparable to the present one (Farnoush et al., 2022). According to the authors, there is minimal proof that businesses are already utilizing blockchain and their insight characteristics. The authors of this report used the TOP-1000 companies, of which 8% were using blockchain technology. In the study, various financial indicators are used to cluster businesses based on how stable their finances are. This is a data-driven research paper basically similar to the subject of our study. The volume of results that the authors were able to gather is where this study's limitations lie. They were able to make the fundamental prediction that companies with strong financial standing will have more success implementing blockchain technology.

1.2. Main findings from previous studies

It is beneficial to take a closer look at the findings from earlier studies using quantitative questionnaire methodologies that were based on managers' perceptions of the input and output data. Thus, one of the studies based on questionnaires of respondents show that supply chains based on blockchain technology in evidence of Jordanian manufacturing sector achieve higher levels of productivity, lead times, customer service and relationships with supply chain members, while reducing transaction costs (Jum'a, 2023). Another linear regression model proves that the efficiency of the supply chain is positively and significantly impacted by different characteristics of blockchain technology (Al-Zaqeba et al., 2022). Different enterprises from China were measured by conducting structural equation modeling (SEM) demonstrating that relative advantage and cost savings with competitive pressure are significant determinants, while technological readiness, and

financial readiness did not show significant impact on blockchain adoption (Deng et al., 2022).

Nevertheless, another study that measures determinants for implementation of blockchain shows that the quantity of the company's total assets among Chinese blockchain technology enterprises significantly influences the implementation whereas size of the staff and sales revenue are insignificant determinants in the model (Pan et al., 2020). The findings indicate that one of the key driving forces behind the adoption of blockchain technology is the growth of the enterprise asset scale as well as blockchain implementation significantly positively influences improving asset turnover rate and reducing sales expense rate.

Another important concept is that only large companies are financially stable enough to accept the risk of changing their business models, and they are also prepared for the possibility that the change may not be successful. For nowadays the opportunity costs of integrating blockchain in OSCM have not been thoroughly studied. According to a quantitative study (Xu and Choi, 2021) a negative cross-channel effect between online and offline channels (when online demand influences offline demand) should motivate businesses to adopt blockchain technology as a Pareto improvement strategy for both the manufacturer and the online platform. The simplest way to determine whether the implementation may have a beneficial influence is to focus on examples of organizations that have already determined that new technology will be efficient for them in the future and have begun using it in their operations.

The same conclusions are reached by Morkunas et al. (2019) regarding the necessity of a thorough analysis of the different business models utilizing blockchain. Existing business models in various supply chain business blocks can be transformed by blockchain. "These nine blocks cover the four main areas of a business: its customers, the offer, the infrastructure, and the financial viability" (Morkunas et al., 2019). Due to escalating competition, it is crucial for firms to understand their most critical choice on whether or not to embrace blockchain technology as soon as possible.

In comparison with previous studies the current study is focusing on various blockchain use cases that might be used in a variety of industries. The main factors that should influence the decision to implement blockchain in our study are the industry and financial performance of the companies. Compared to Farnoush et al. (2022), the authors of the current study apply more sophisticated tools, such as clustering analysis and classification model with bootstrap for stable results. Our research is based on quantitative measurements of data on the TOP-2000 companies in China and the US which are now major players of blockchain adoption.

2. COLLECTION AND PRIMARY DATA PROCESSING

2.1. Sample and data

The current research required specific data that was not available in any out-of-the-box dataset. Due to that the dataset was manually created from several open sources. Python and its powerful modules like Scrapy (quick and scalable web crawlers), Requests (basic HTTP operations), and Selectolax (high-speed html parsing) were used to collect extensive data on companies. The data contained both financial characteristics and industry affiliation as well as the type of blockchain adoption. Since the US and China are the two countries driving blockchain implementation, the current analysis is based on the study of these two countries focusing on TOP-2000 companies for 2019-2022. The list of that companies with their market and structural was obtained from the website "Value.Today"[3]. The primary dataset was merged with financial data collected from well-known online financial platform "Yahoo Finance"[4], publishing various data for investors.

In order to analyze the type of blockchain adoption, it was gathered how firms and their subsidiaries use blockchain in their businesses. Website "Blockdata"[5] occurred to be the main information source of information about the fact of blockchain implementation in the company. This website specializes in creating a database of companies that implement blockchain on their webpage, so the majority of the businesses were featured on the website along with blockchain use cases. The missing data for use cases was gathered manually analyzing current news of the companies for which it was already known from that webpage a status "True". Finally, use cases and industries were combined into larger groups to simplify the analysis of numerous categorical data. A small proportion of the missing values were later filled in with the median value for each of the variables when building models.

Final dataset consisted of 2000 companies from the US and China with 8 numeric financial and 4 categorical variables.

2.2. Variable selection and description

Financial indicators data for 2019-2022 were weighted averaging with weights from 1 to 4 to increase the significance of the company's performance in recent years. Ultimately the final dataset with which further work was carried out consisted of the following variables:

[3]Website "Value.Today" [online resource]: `https://www.value.today/` (date of assessing: 05.02.2023)
[4]Website "Yahoo Finance" [online resource]: `https://finance.yahoo.com/` (date of assessing: 20.02.2023)
[5]Website "Blockdata" [online resource]: `https://www.blockdata.tech/` (date of assessing: 20.03.2023)

Table 1: Variable description

Variable	Description	Data Type
Country	1 - for US, 0 - for China	Binary
Status	Describing whether a company has integrated or integrating blockchain technology in its business. 1- if company implements blockchain, 0 - if not	Categorical: binary
Market capitalizaion	Provides an estimate of the strength on the market. Annual weighted average for 2019-2022	Numeric: float
Industry	Industries that describe companies' business areas: • Communication&Entertainment • Estate • Finance&Investments • Food&Restaurants • Healthcare • IT&Software • Logistics&Transportation • Machinery&Manufature • Resources&Materials • Retail&ECommerce • Services&Travel • Technology&Electronics • Transport&Aerospace	Categorical: str
Basic EPS	Provides an estimate of profitability on an absolute basis on the financial market widely used by investors. Annual weighted average for 2019-2022	Numeric: float

Table 1: Variable description

Variable	Description	Data Type
Use cases	Types of blockchain for companies that implement it. List of use cases: • Cybersecurity • Data Management&Security • Finance&Tokens • Food&Restaurants • Logistics&Supply Chain • IT&Software • NFT&Gaming • Payments • Research&Consulting • Smart Contracts	Categorical: str
Company age	The difference between the year of establishment and the year 2023 of writing the current work	Numeric: int
Revenue	Provides an estimate of sales volumes. Annual weighted average for 2019-2022	Numeric: int
Net income	Provides an estimate of profitability. Annual weighted average for 2019-2022	Numeric: float
Number employees	Provides an estimate of labour force volume. Data for 2022	Numeric: int
Total equity	Represents the value of an investor's share in a company, i.e. provides an estimate of what role does an investor play in decision-making. Annual weighted average for 2019-2022	Numeric: float
Investments cashflow	Provides an estimate of contribution to innovation. Negative values mean investments, positive - sale of capital. Annual weighted average for 2019-2022	Numeric: float

Correlation matrix can be seen below - Figure (1). Most correlated but not less important for the analysis variables are investments with capitalization (-0.84), income with capitalization (0.73), number of employees with revenue (0.72), income with revenue (0.62), income with investments (-0.64), capitalization with equity (0.61). Some variables are almost uncorrelated with others. They are country, company age, basic EPS and status.

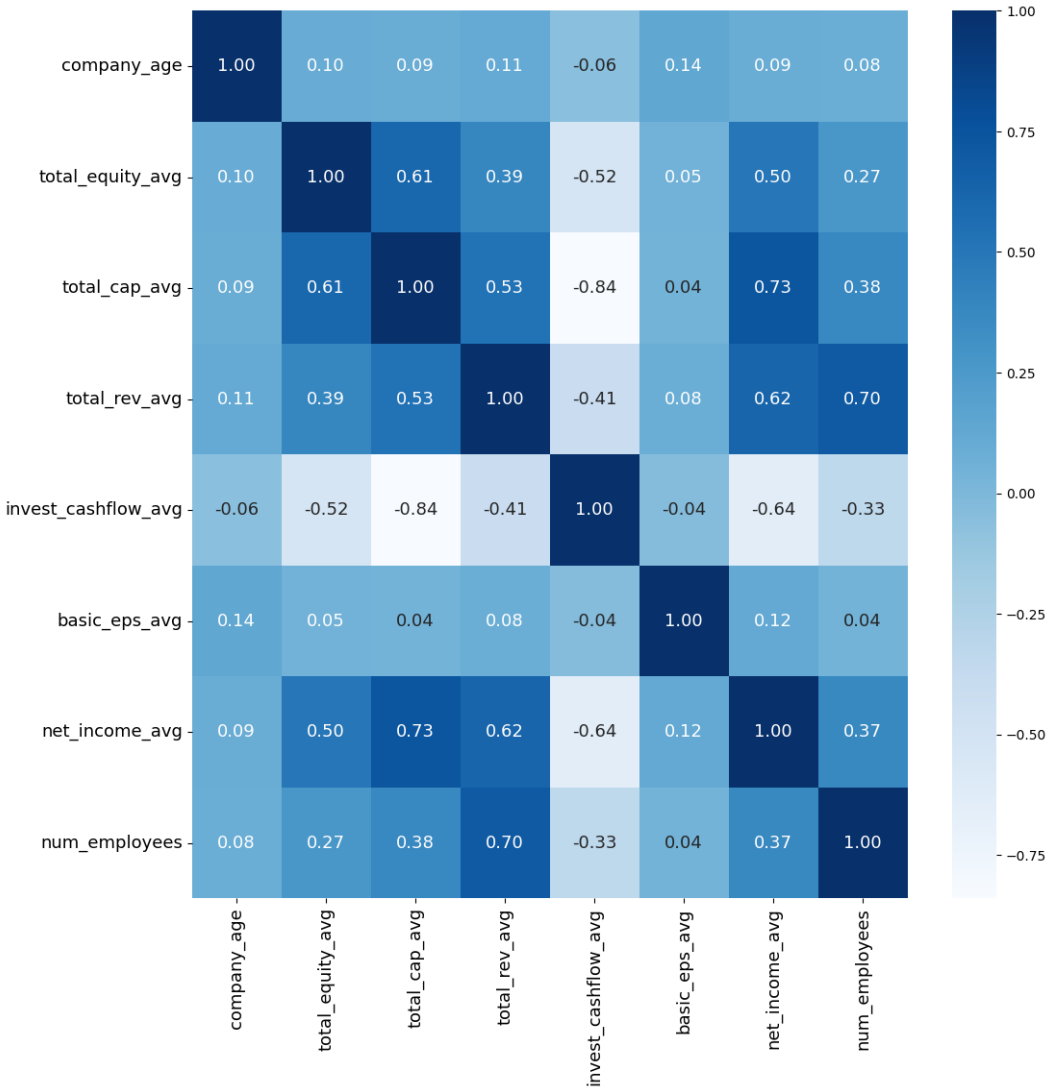

Figure 1: Correlation matrix (authors' picture)

3. METHODOLOGY

3.1. Descriptive analysis

Final dataset consisted of TOP-1000 countries from the US and TOP-1000 companies from China. A total of 180 businesses, or 9% of them, have started implementing blockchain. Over $\frac{2}{3}$ of those companies belong to the US. Comparing numbers of companies across industries in the US and China, it is interesting to point out that companies in Resources & Materials (230 vs 147 companies from China and US respectively), Technology & Electronics (147 vs 62), Transport & Aerospace (83 vs 43) industries are twice as many represented from China than from the US. Meanwhile, Estate (35 vs 74) and Retail & Commerce (39 vs 71) industries are twice as many provided from the US and three times more from the US for IT & Software industry (46 vs 145). The Finance & Investments industry (95 vs 132) is well represented in both countries, but slightly more in the US. All statistics can be seen on the tables in *Appendix 1*.

Interesting characteristics of industries can be seen if you look at their main financial indicators. Large sharks that have accumulated a large capitalization (41 %) and collected large profits (29%) are not surprisingly represented in the Finance & Investments industry. Most revenue was collected by Resource & Materials (19%), Retail & E-Commerce industry (14%) and Healthcare (13%) industries. Proportions of employees working across industries are almost uniformly distributed except Estate (2.8%) and Machinery & Manufacture (2.3%) industries. Most investments are accumulated in Finance & Investments (50%), Resource & Materials (13%) and Communication & Entertainment (8%) industries. All statistics are presented in fractions of the total values and can be observed in *Appendix 1*. For average indicators, the trend is similar.

Statistically on average companies that implement blockchain have higher values for all financial indicators, but it is the topic for current research whether all the indicators really influence the company's decision on the implementation of blockchain. Blockchain is being implemented three times more actively in the US and in such industries as Finance & Investments (16 cases for China vs 44 for the US), IT & Software (5 vs 37) and Communication & Entertainment (2 vs 19). In Picture 2 there is a bar chart illustrating the number of all companies and the share of companies implementing blockchain across industries.

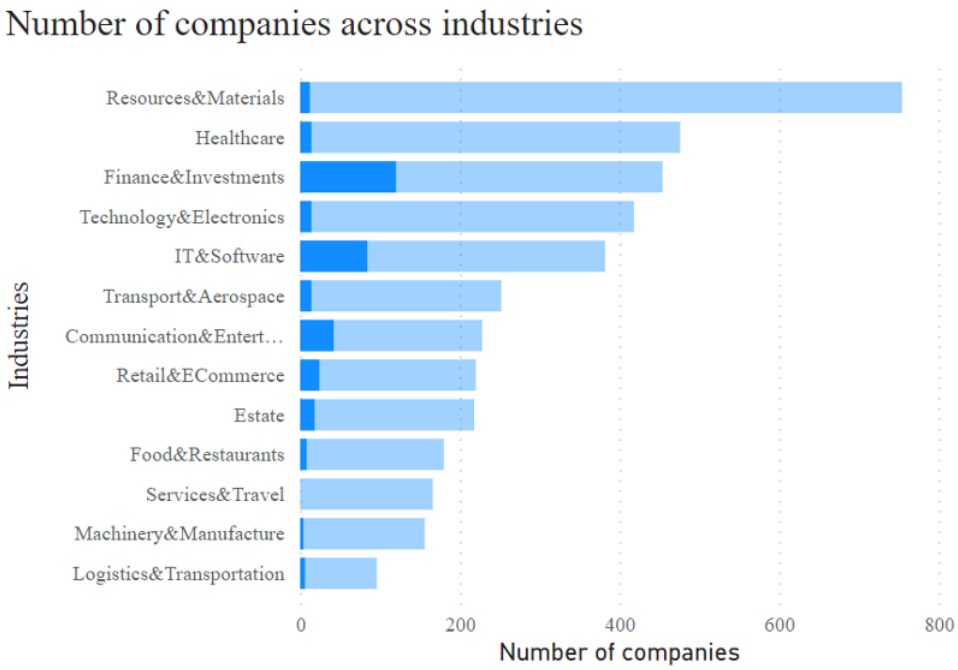

Figure 2: Number of all companies and the share of companies implementing blockchain across industries (authors' picture)

Another important analysis of companies is the study of the types of blockchain so-called use cases. Another bar chart in Figure (3) demonstrates the distribution of the companies implementing blockchain across use cases and the share of those companies from the US. According to this bar chart there is a significant share of companies from the US. Most popular use cases are Finance & Tokens (60 cases), Data management & Security (46), Logistics & Supply chain (25) and NFT & Gaming (22). These use cases mostly implement companies respectively in Finance & Investments (17 cases out of 60), IT & Software (38 out of 46), Retail & E-Commerce with Resources & Materials (10 out of 25), Communication & Entertainment (13 out of 22). The cross-table for use cases and industries is presented in Table (2). The most active industries in implementing blockchain are Finance & Investments (60 cases), IT & Software (42), Communication & Entertainment (21) and Retail & E-Commerce (12). All these industries accumulate 75% of all cases of blockchain implementations.

While the US focuses mostly on Finance & Tokens (32% of all cases) and only then on Data management & Security (18%) use cases, for China it is more important to firstly focus on Data management & Security (38% of all cases) and secondly on Finance & Tokens (23%) Noticeably large investments are presented for Finance & Tokens and Payments use cases.

Additional pie charts and tables of these statistics can be seen in *Appendix 1*.

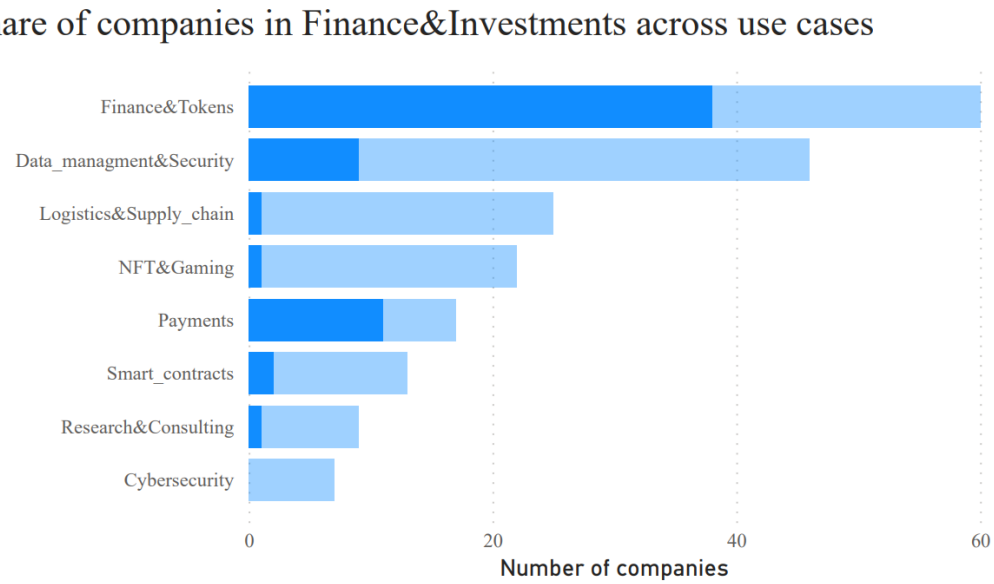

Figure 3: The distribution of the companies implementing blockchain across use cases and the share of those companies from the US (authors' picture)

Table 2: Number of companies implementing blockchain across use cases and industries (authors' table)

industry / use case	Cyber security	Data Management Security	Finance Tokens	Logistics Supply Chain	NFT Gaming	Payments	Research Consulting	Smart-contracts	Total
Finance&Investments	0	**9**	**38**	1	1	**11**	1	2	**60**
IT&Software	4	**17**	**8**	4	1	3	**6**	2	**42**
Communication&Entertainment	1	3	4	0	**13**	0	1	0	**21**
Retail&ECommerce	0	3	1	5	4	1	0	2	**12**
Estate	1	2	6	0	0	1	0	2	**9**
Healthcare	0	6	0	0	1	0	0	0	**7**
Technology&Electronics	0	3	1	2	1	0	0	1	**7**
Transport&Aerospace	1	2	0	3	0	0	1	0	**7**
Resources&Materials	0	0	1	5	0	0	0	2	**6**
Food&Restaurants	0	0	1	1	1	1	0	1	**4**
Logistics& Transportation	0	1	0	2	0	0	0	1	**3**
Machinery&Manufacture	0	0	0	2	0	0	0	0	**2**
Total	**7**	**46**	**60**	**25**	**22**	**17**	**9**	**13**	**180**

3.2. Clustering models

In order to understand more clearly what types of companies we are dealing with in the study, data clustering was carried out with the data described in the section *Variable Selection and Description 2.2*. The dataset was standardized to eliminate the effect of different dimensions of variables. To choose the number of clusters, the Elbow method, Silhouette coefficient and Dendrogram were applied. K-means, DBSCAN and Agglomerative clustering were used as the main models. To check the stability of the results, bootstraping was conducted. Additionally, Principal Component Analysis (PCA) was applied which reduced the dimensionality of the data

and showed the main components that best describe companies in the dataset.

Finally, all clusters were plotted and described, which made it possible to identify the types of companies and link these types with the fact of blockchain implementation. The results of cluster analysis are presented in section *Results of clustering analysis 4.1*.

3.3. Classification models

We decided to go for classification as a primary tool to determine what affects the decision of the company to implement blockchain in its business. Our dataset contains binary variable status which indicates if a given company has been known to have implemented blockchain in its business operations in any way. Therefore, this problem could be easily set up as a binary classification problem where we aim to predict the target variable whether a company has worked with blockchain given other regressor variables describing it in terms of finances and structural data. We will train most common classification models such Logistic Regression, Random Forest and various boosting models such XGBoost and others. As a result, we aim to find the best model that is able to correctly label companies whether they should implement blockchain technology or not. We will use repeated Stratified K-Fold cross validation to compare the models between each other as well as tune their hyperparameters to achieve the highest performance. Results of classification analysis are presented in section *Results of classification analysis 4.2*

3.4. Expected results

To determine the kinds of businesses the current study was done with, the methodology mentioned above should be helpful. The main assumptions were that the decision on the introduction of blockchain should be influenced by the financial characteristics of the company and its affiliation to a particular industry. Thus the expected results from cluster and classification models: 1) combining companies into clusters that most accurately reflect the type of business, 2) identification of industries where the use of blockchain technology has a statistically significant impact, 3) identification of other statistically important indicators, such as investments and earnings, that are taken into consideration when considering whether to implement blockchain.

4. RESULTS AND DISCUSSIONS

4.1. Results of clustering analysis

Bases on the Elbow method, Silhouette coefficient and Dendrogram 5 clusters were chosen for conduction of three different models. K-means showed the best result among all models (silhouette coefficient equals 0,45). Bootstrap showed that these clusters are quite stable, but only for companies from the US. Obviously the average company from the US with its financial indicators dominates the average company from China. Because of that 5 clusters can be stable only for US companies. Nevertheless clusters turned out to be well interpreted and transparent for analysis. The following Table (3) shows the results:

Table 3: Clustering analysis results in USD billions (authors' table)

Cluster	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
num_companies	27	164	675	**711**	423
total_cap_avg	**318.34**	28.48	13.98	9.36	65.52
net_income_avg	**25.27**	2.2	1.15	0.54	5.3
total_rev_avg	**186.16**	20.4	13.44	7.58	16.73
invest_cashflow_avg	**-43.27**	-3.04	-1.54	-0.94	3.13
basic_eps_avg	5.13	3.7	4.02	-0.14	**76.47**
total_equity_avg	**243.07**	16.28	7.76	6.23	26.51
company_age	**64**	57.55	63.49	26.87	38.02
US (share)	0.7	0.77	**1**	0	0.43

First cluster includes the biggest 27 companies with all high indicators, 70% of them are located in the US and 59% are implementing blockchain. Second cluster is another type of companies that with 100% sure implement blockchain. Their average investments are three times higher than for other companies, but noticeable less than for sharks from the first cluster. The average age for these companies are also high as for the first cluster, that is, all companies that implement blockchain on average have been on the market for a long time.

Other three clusters include companies which are not implementing blockchain. The third cluster consists of US companies with a high average age. The fourth cluster includes companies from China and that companies are very young. The fifth cluster consists of companies from both the US and China that are investment attractive for investors according to high average basic EPS.

PCA showed that 80% of all information about companies can be described in 5 components. Figure (4) shows the correlation matrix between principal components and initial variables describing main characteristics of companies:

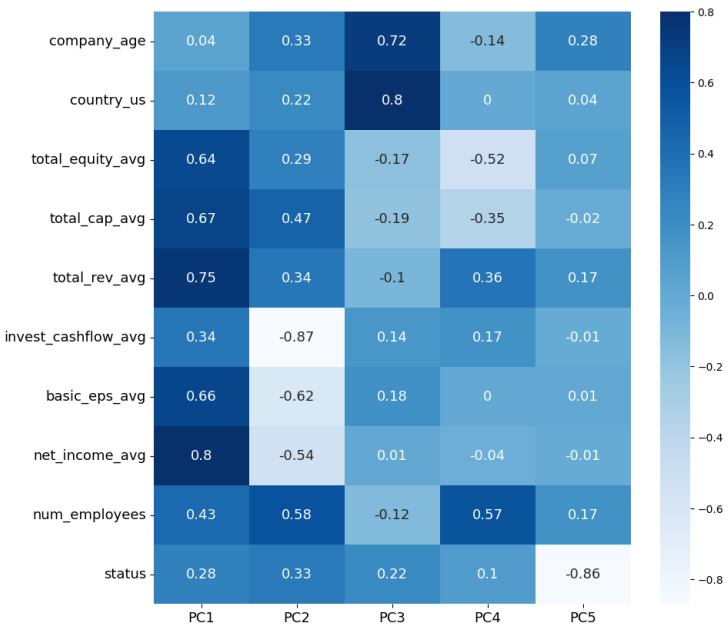

Figure 4: Correlation matrix between Principal Components (PC) and initial variables

PC 1 explains general financial success, while PC 2 catches investment activity of the company. PC 3 includes the information about country affiliation whereas PC 4 explains scale of equity and number of employees. Finally, the last PC 5 catches the status of implementing blockchain.

Figure (5) demonstrates the result of clustering analysis in dimensions PC 1 and PC 5, from which all five clusters are clearly visible:

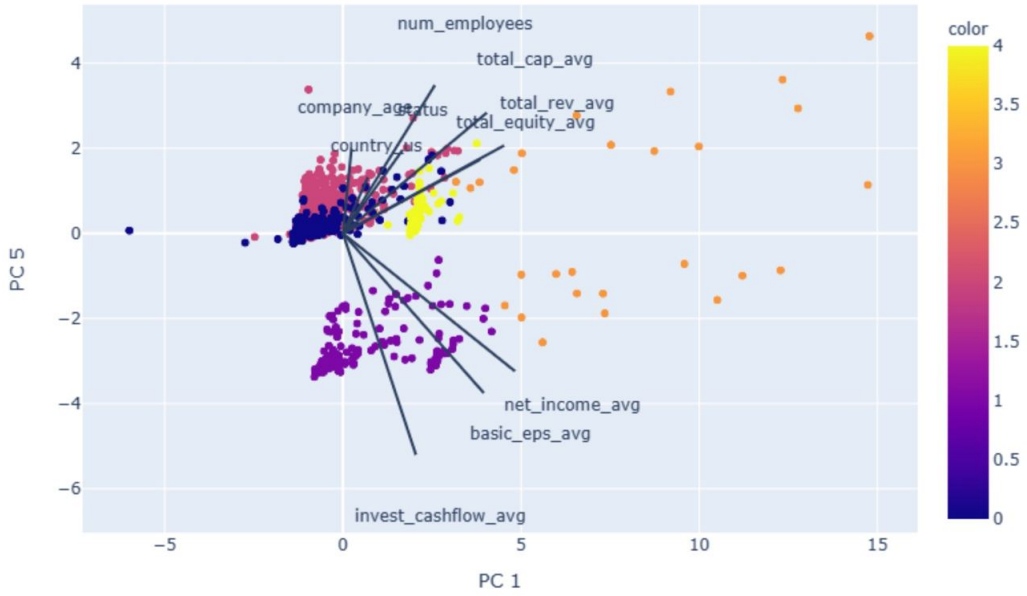

Figure 5: Visualisation of clustering analysis with PCA plot (authors' picture)

Cluster analysis revealed the types of companies we are studying. There are two types of companies that implement blockchain. One of them is the market sharks who rushed into battle because they have a lot of funds for such experiments (the first cluster).

For other companies (the second cluster), high investments are more likely the result of the introduction of blockchain, and not vice versa, as for the first cluster. These companies are more interesting to emulate, since a mistake will cost more, these companies will weigh their decision several times before implementing a new business model. For other organizations that are currently contemplating whether to embrace blockchain, these companies serve as important case studies.

All other graphs complementing cluster analysis can be seen in *Appendix 2*.

4.2. Results of classification analysis

4.2.1. Baseline model. Data imbalance

Due to a high imbalance in the data, meaning that we have companies that have implemented blockchain as a minority class, we should be careful with classification models since they might just predict the majority class all the time maximizing accuracy.

Table 4: Number of companies that have blockchain

	Category
No Blockchain (False)	1529
Blockchain (True)	167

Therefore, as a baseline, we take such a "dummy" predictor that always outputs "False" meaning it always predicts that the company does not have blockchain. Such model will get very high accuracy close to 90% in our case, but obviously such an approach does not help us with interpretation whatsoever. Therefore, we should think of other classification models and most importantly appropriate scoring metrics to evaluate their quality given the imbalance in data. Typically, a confusion matrix - Table (5) is used to analyze what kind of mistakes a classification model makes.

Table 5: Confusion matrix

		Predicted	
		Negative	Positive
Actual		*True Negative (TN)*	*False Positive (FP)*
		False Negative (FN)	*True Positive (TP)*

In order to analyse our imbalanced data we should not use metrics that include TP and TN in the denominator at the same since they also account for predictions of the majority class. Instead we should focus on metrics like Precision and Recall since they allow to evaluate labelling performance of the minority class:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

Precision focuses on the positive predictions made by the model. It measures the proportion of true positive predictions out of all positive predictions (TP + FP). Precision indicates how many of the predicted positive instances are actually correct. High precision implies a low rate of false positives. Recall (also known as sensitivity or true positive rate): Recall focuses on the actual positive instances in the dataset. It measures the proportion of true positive predictions out of all actual positive instances (TP + FN). Recall indicates how well the model identifies the positive instances in the dataset. High recall implies a low rate of false negative.

4.2.2. Logistic regression

As a primary model for classification we will use Logistic Regression. It is a great model for binary classification. In order to address data imbalance we will just adjust the weights for minority class, meaning that model is penalised greatly now if it wrongly classifies the minority class.

Table 6: Classification report for Logistic Regression

	precision	recall	f1-score	support
0	0.968	0.691	0.806	307
1	0.215	0.788	0.338	33
accuracy	0.700	0.700	0.700	0.7
macro avg	0.591	0.739	0.572	340
weighted avg	0.895	0.700	0.761	340

The classification report - Table (6) above is produced for the following confusion matrix - Figure (6). Such model has a high recall meaning that it is able to find almost every company that has implemented blockchain but unfortunately this is achieved mainly at the expense of pre-

cision score. Our model just predicts ”True” all the time that is why it is so successful finding all blockchain companies.

As a result such model shows great ability at finding blockchain companies but this comes at a cost of low precision. All of this corresponds to low f1-score for the minority class. Therefore, our aim is to find a classification model that could either have very high precision or recall. If possible it would be great to balance them, so that both precision and recall are relatively high but this is typically impossible to achieve with imbalanced data.

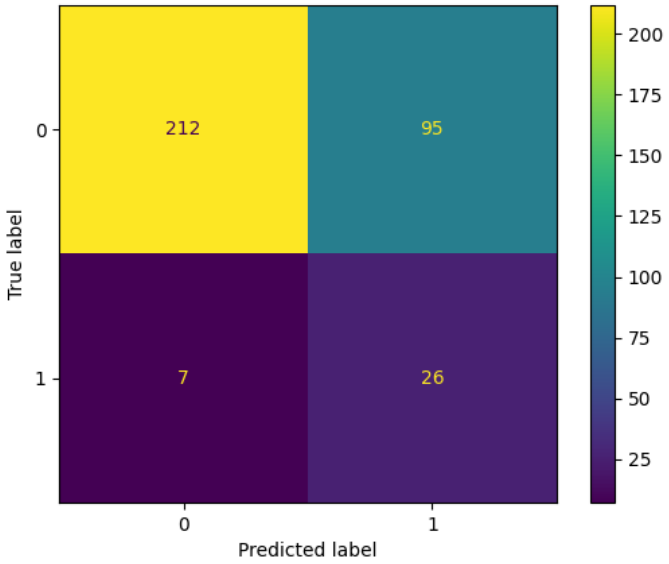

Figure 6: Logistic Regression Confusion Matrix

Therefore, we should find the model that has a great performance in terms of preferably both precision and recall. This could be done with Precision-Recall curves, specifically PR-AUC (Precision Recall Area Under the Curve) which measures overall performance no matter the probability threshold $p \in [0, 1]$. The higher the value of PR-AUC the better the performance of classifier is.

After deciding on the primary quality metric PR-AUC we tried to tune the Logistic Regression model above by adjusting its hyperparameters, namely, *class_weight* which defines misclassfication penalty. After multiple training cycles of the model with GridSearch we were able to find optimal parameters yielding the highest PR-AUC score but still they were not enough to tackle class imbalance.

4.2.3. Oversampling

Random Oversampling (ROS) is a popular approach to class imbalance. This technique simply randomly copies observations of underrepresented minority to balance the number of obser-

vations in each class. This is basically the same as adjusting weights of classes since by copying observations we signal importance of this class to classifier. Another approach is SMOTENC which finds K-Nearest Neighbors and combines them to create new observations. Therefore, such approach fills in minority class with similar observations consisting of linear combinations of others.

4.2.4. Random forest

Random Forest is a great classifier since it allows to train a lot of weak learners (Decision Trees) to then ensemble them into a single prediction. In our case such decision trees are trained on subsamples with both less features and less observations than the original dataset. This way we are trying to prevent overfitting of the model (when the model simply learns the whole training set achieving great target scores but performing very poorly on unseen test data). In order to find optimal depth of Decision Tree, number of such trees as well as other hyperparameters we used Optuna which makes use of Bayesian optimisation. This way we were able to tune the model so it yields the highest PR-AUC (Precision-Recall Area Under the Curve) score on multiple folds of stratified validation sets. We also trained Random Forest Classifier on balanced with SMOTENC dataset. But such approach did not seem to help very much with model performance.

4.2.5. Boosting

As a logical continuation of Random Forests there are boosting techniques when we feed errors of estimators to the next estimator so it is able to make up for the mistakes made by previous estimators. Such approach is also more flexible since it allows to stop training once there is no improvement on validation set in terms of desired metric (PR-AUC in our case). With Random Forest we had to specify number of estimators (Decision Trees) before training process, here since boosting models are trained sequentially we are able to stop training whenever we spot overfitting. This allows us to get the best specification of the model as well as preserve its generality.

We used XGBoost to train classification model. In order to get the best parameters we ran 1000 training rounds with Optuna. Obtained model should not be overfitted since we did early stopping explained above, shown in (7). Alternatively, we could have minimized the difference between train and validation PR-AUC scores but this would result in overall lower scoring metric.

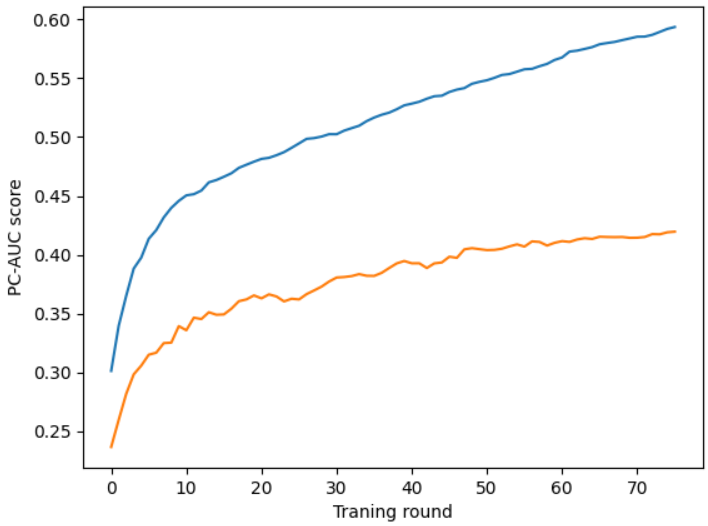

Figure 7: XGBoost Validation and Training scores

4.2.6. Choosing the best model

Trained models are compared based on their Precision-Recall curves on test set. Results are shown in Figure (8) Although, model selection should not be based on test set, since it allows for bias, we have no other way to present visually how models compare to each other since we do not have a separate validation set. But still we can say that models are performing very similarly on test set.

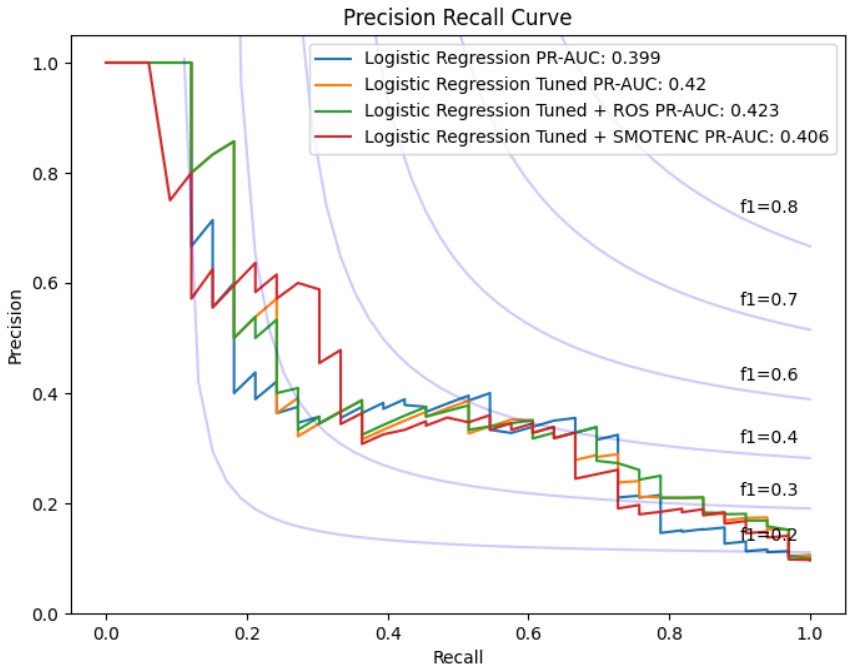

Figure 8: Logistic Regression PR-Curves on test data

In order to make comparison of the models somewhat fair we should use repeated stratified K-Fold crossvalidation same as we used for hyperparameter tuning. The more repetitions we choose the more certain we can be about rankings of the models. We will be able to compare models in terms of their mean PR-AUC score calculated on various crossvalidation sets.

Table 7: Crossvalidation PR-AUC scores

	mean validation PR-AUC
Logistic Regression	0.354603
Logistic Regression Tuned	0.379670
Logistic Regression Tuned + ROS	0.356380
Logistic Regression Tuned + SMOTENC	0.372991
RandomForest Tuned	0.345277
RandomForest Tuned + SMOTENC	0.327572
XGBoost Tuned	0.363342
XGBoost Tuned + ROS	0.346332

In this case we ran 6-fold crossvalidation with 50 repetitions meaning that each model has been trained on 300 various subsets of training data and evaluated on another 300 different validation sets. As we can see from the Table (7) above, Logistic Regression tuned with weight balancing is the best classifier in terms of PR-AUC score, but, honestly, all of the classifiers above perform more or less the same. Weight balancing turned out to be slightly ahead of oversampling techniques.

4.2.7. Model interpretation

Now we will move from comparison of the models to their interpretations, we will mostly focus on Logistic Regression. We trained Logit model on the whole dataset, as a result it predicted log probabilities of observations being class 1 (has blockchain) based on financial and structural variables described earlier along with categorical *industry* variable encoded as multiple dummy variables. We reasoned that our data is not perfectly representative of the general population. In order to tackle this, we used Bootstrapping technique, where we resampled dataset multiple times by randomly picking observations into a new bootsrapped dataset which was later used to train the model. By doing this we created 5000 bootstrapped subsamples and trained the model on each of them, then we saved the coefficients and p-values of Wald's significance test. As a result, we were able to create distributions for coefficients - Figure (28) and p-values - Figure (29). Quantiles are shown below in Table (8).

Table 8: Bootstrapped coefficients and p-values for Linear Regression

	coef_5%	coef_median	coef_95%	pval_5%	pval_median	pval_95%
intercept	−5.956	−3.719	−2.679	0.000	0.000***	0.000
company_age	−0.053	0.114	0.270	0.003	0.211	0.901
total_equity_avg	−0.758	−0.113	0.549	0.034	0.482	0.847
total_cap_avg	−0.077	0.266	0.670	0.002	0.146	0.842
total_rev_avg	−0.072	0.179	0.402	0.001	0.158	0.882
invest_cashflow_avg	−0.068	0.099	0.356	0.021	0.395	0.935
basic_eps_avg	−0.160	−0.030	0.133	0.104	0.532	0.951
net_income_avg	−0.156	0.060	0.388	0.033	0.452	0.945
num_employees	0.117	0.348	0.579	0.000	0.009**	0.396
is_us	0.574	0.970	1.423	0.000	0.000***	0.015
Communication_Entertainment	0.609	1.710	4.036	0.002	0.023*	0.537
Healthcare	−1.837	−0.459	2.105	0.016	0.431	0.944
Food_Restaurants	−2.500	−0.516	1.651	0.056	0.512	0.968
Resources_Materials	−2.316	−0.934	1.425	0.003	0.239	0.932
IT_Software	0.813	1.827	3.993	0.001	0.011*	0.248
Retail_ECommerce	−0.861	0.373	2.555	0.078	0.461	0.946
Technology_Electronics	−1.626	−0.183	2.113	0.059	0.512	0.952
Finance_Investments	0.963	1.963	4.131	0.001	0.006**	0.152
Estate	−0.539	0.696	2.920	0.046	0.360	0.932
Services_Travel	−31.083	−10.820	−5.090	0.566	0.907	1.000
Transport_Aerospace	−1.197	0.158	2.579	0.081	0.513	0.949
Machinery_Manufacture	−12.651	−0.546	1.692	0.108	0.594	0.983

Significance levels: p-value "***" ≤ 0.001, "**" ≤ 0.01, "*" ≤ 0.05

Bootstrapping allows us to check how stable values of both coefficients and p-values are when we resample our dataset. Variables such as *num_employees, is_us, Communication&Entertainment, IT&Software, Finance&Investments* are statistically significant meaning that removing them from the model (making their corresponding coefficients equal to 0, therefore rejecting null hypothesis) negatively affects the fitness of the model. We chose to use bootstrapped p-values and coefficients since they allow to have a better picture about real effects of variables on the target variable. This is done mainly to prevent such cases when we train the model on some sample and then remove a few observations to find out that variables that were significant are now insignificant. Therefore, we expect that our interpretations will be more stable and less dependant from the sample size and unaccounted missing observations. Distributions of both p-values and coefficients are presented in *Appendix 3*, Permutation and Feature importances for the variables in *Appendix 3*.

4.3. Limitations and further work

Despite the fact that many different models have been run to obtain current results, there are always limitations in any research. In our case, the results depend on the data presented to test the models. Since business data is not quite available for the vast majority of the companies, our

research focused on the assumption that by studying TOP-2000 companies, we can suppose that the rest of the companies make decisions looking back at the success of large companies in the market. It followed from this assumption that our results could be scaled to a different sample of smaller companies. Taking into account the fact that the hypothesis about the non-impact of the company's scale could not be refuted using our classification model, the results of the study cannot insist that the strategy of implementing blockchain in IT, financial and entertainment will pay off, but he cannot prove the opposite either.

Another important limitation of the current work is the construction of the model itself and the choice of those characteristics that are important for measuring the probability of successful implementation of blockchain. This work is based on the assumption that financial indicators reflect the efficiency and sustainability of the company. Studying across industries suggests that established business models could have been best optimized over the past decades, that changing the business model is unprofitable and expensive. Our model does not take into account the internal structure of the company, the production chain into which the company is embedded. One of the advantages of blockchain is transparency in conducting transactions, signing contracts, logistics, insurance, which is very important for companies that work a lot with third parties.

It is worth mentioning that this study revealed most of the US companies as China was represented to a lesser extent in the sample of companies implementing blockchain. The study does not fully investigate the differences in the approach of each country to the implementation of the blockchain. Thus, China might bear lower costs, which means that for business restructuring, a Chinese company might pay lower costs than an American one. Moreover, most Chinese companies implementing blockchain may be presented among small companies or even startups. Thereby, this is a separate topic for research, why exactly these two economies rushed to implement blockchain, how the approaches of the countries may differ, what costs are paid by each of the countries and who benefits more from the introduction of the new technology.

Future researchers can focus on the data available on the "Blockdata" website, which we used to obtain the true or false status about the implementation of the blockchain. This site presents a wide database of all companies that are currently implementing blockchain. For all these companies, one can try to collect financial data and compare the results with companies that do not implement it. In such a study, the main thing is not to get an unrepresentative dataset that does not reflect the real market, that the results will not be scalable to other samples. It also makes sense to conduct this kind of research in future years, since the number of companies implementing blockchain will certainly grow. Many companies may also not announce now that the transformation process has been launched, so the market may not know all the plans of the companies.

CONCLUSION

The impact of the new blockchain technology on the markets today can no longer be underestimated. The financial sector adopted blockchain technology earlier than other sectors, which explains why there are more businesses using it today. However, other industries are at the stage of evaluating the benefits of introducing a new technology. Databases with information about businesses changing their business models and integrating blockchain in data management, payments, logistics, smart contracts, and in other areas have emerged during the past four years with the growth of big data storage. This enables ongoing and future research to shift to an empirical measurement of the influence of blockchain on the performance of the company, while also considering the characteristics of those companies that need to implement blockchain for future stability and competitiveness.

The current study focused on studying the types of blockchain implementation across different industries. Our models have shown statistical significance for the financial, IT, and entertainment ones. These sectors have been the first to rush into the technology race, pinning hopes on the success of early technology implementation in their industry. The paper traces a pattern in the proportion of companies implementing blockchain and the complexity of business. Thus, the food industry, resource supply, transportation and machinery have been around for a long time and do not require business model optimization in the form of the introduction of a new technology.

In addition, our study suggested the influence of financial indicators on the decision to implement blockchain. The findings support previous research on the insignificance of financial considerations. For other companies, this is an important signal that the new technology is essential to consideration not only for large companies making substantial investments. Researchers that are interested in the introduction of other new technologies may find the current study to be of interest. Our methodology is applicable to the study of the characteristics of companies for which the introduction of any other technology might be efficient.

Finally, our study suggests that small businesses could assess the need to implement blockchain in relation to their industry on the results obtained using the evidence of pioneers and thereby make more informed and lower investment decisions.

APPENDICES

Appendix 1. Descriptive statistics of the data

industry	China	US	All
Communication&Entertainment	45	69	**114**
Estate	35	74	**109**
Finance&Investments	95	132	**227**
Food&Restaurants	50	40	**90**
Healthcare	119	119	**238**
IT&Software	46	145	**191**
Logistics&Transportation	21	27	**48**
Machinery&Manufacture	45	33	**78**
Resources&Materials	230	147	**377**
Retail&ECommerce	39	71	**110**
Services&Travel	45	38	**83**
Technology&Electronics	147	62	**209**
Transport&Aerospace	83	43	**126**
All	**1000**	**1000**	**2000**

Figure 9: Cross-table of all companies across countries and industries (authors' table)

industry	total_cap_sum	net_income_sum	total_rev_sum	num_employees_sum	invest_cashflow_sum	basic_eps_sum	total_equity_sum
Communication&Entertainment	7,41%	9,01%	6,76%	5,77%	7,82%	2,37%	7,90%
Estate	4,62%	3,10%	3,09%	2,77%	2,61%	3,46%	4,64%
Finance&Investments	41,58%	29,29%	14,24%	13,88%	49,81%	23,81%	41,74%
Food&Restaurants	2,75%	4,30%	4,61%	6,41%	1,84%	3,69%	2,38%
Healthcare	6,65%	9,62%	13,25%	8,37%	5,70%	14,62%	6,41%
IT&Software	5,77%	12,18%	9,03%	10,93%	4,26%	6,42%	5,16%
Logistics&Transportation	2,03%	0,34%	3,16%	5,78%	1,95%	1,61%	1,73%
Machinery&Manufacture	1,46%	2,47%	2,17%	2,30%	1,24%	5,47%	1,41%
Resources&Materials	15,27%	14,20%	19,07%	10,22%	13,05%	8,98%	15,73%
Retail&ECommerce	3,14%	6,32%	9,23%	13,68%	3,96%	10,55%	3,24%
Services&Travel	2,29%	0,94%	3,65%	5,14%	1,88%	2,66%	2,12%
Technology&Electronics	3,05%	4,86%	5,11%	7,02%	3,46%	7,13%	3,51%
Transport&Aerospace	3,98%	3,38%	6,63%	7,72%	2,41%	9,24%	4,03%

Figure 10: Sum of financial performance of all companies across industries (authors' table)

status	basic_eps_avg	invest_cashflow_avg	net_income_avg	num_employees_avg	total_cap_avg	total_equity_avg	total_rev_avg	company_age_avg
False	2,17	-1,45	1,01	28070,78	13,68	9,29	11,71	49,04
True	3,85	-8,53	5,17	87085,65	65,80	37,66	40,10	58,84

Figure 11: Sum of financial performance of all companies across industries (authors' table)

industry	China	US	All
Communication&Entertainment	2	19	**21**
Estate	3	6	**9**
Finance&Investments	16	44	**60**
Food&Restaurants		4	**4**
Healthcare	3	4	**7**
IT&Software	5	37	**42**
Logistics&Transportation	2	1	**3**
Machinery&Manufacture		2	**2**
Resources&Materials	1	5	**6**
Retail&ECommerce	3	9	**12**
Technology&Electronics	5	2	**7**
Transport&Aerospace	2	5	**7**
All	**42**	**138**	**180**

Figure 12: Sum of financial performance of all companies across industries (authors' table)

use_cases	total_cap_ avg	net_income_ avg	total_rev _avg	num_employees_ avg	invest_cashflow_ avg	basic_eps _avg	total_equity_ avg	company_age_ avg
Cybersecurity	10,26	1,01	11,72	22837,14	-0,86	2,88	5,18	20,43
Data_managment&Security	50,30	6,25	47,87	100734,19	-5,71	2,15	29,08	44,00
Finance&Tokens	126,09	10,35	57,46	104811,32	-18,01	5,59	72,76	63,23
Logistics&Supply_chain	33,82	3,19	59,24	158510,28	-1,57	3,43	23,81	84,10
NFT&Gaming	28,19	1,80	34,58	51394,00	-1,55	1,94	17,49	56,11
Payments	76,31	5,70	23,98	44927,00	-14,92	4,91	36,74	68,94
Research&Consulting	15,21	1,66	16,61	91801,88	-1,63	6,35	6,19	62,00
Smart_contracts	31,00	3,56	34,65	87289,08	-2,09	3,79	21,85	49,18
All	**64,58**	**5,86**	**44,43**	**95249,08**	**-8,28**	**3,79**	**37,33**	**58,43**

Figure 13: Cross-table of companies implementing blockchain across countries and use cases (authors' table)

Use cases / country	China	US	All
Cybersecurity	1	6	**7**
Data_managment&Security	18	28	**46**
Finance&Tokens	11	49	**60**
Logistics&Supply_chain	7	18	**25**
NFT&Gaming	2	20	**22**
Payments	2	15	**17**
Research&Consulting	1	8	**9**
Smart_contracts	5	8	**13**
All	**42**	**138**	**180**

Figure 14: Average financial performance of companies implementing blockchain across use cases (authors' table)

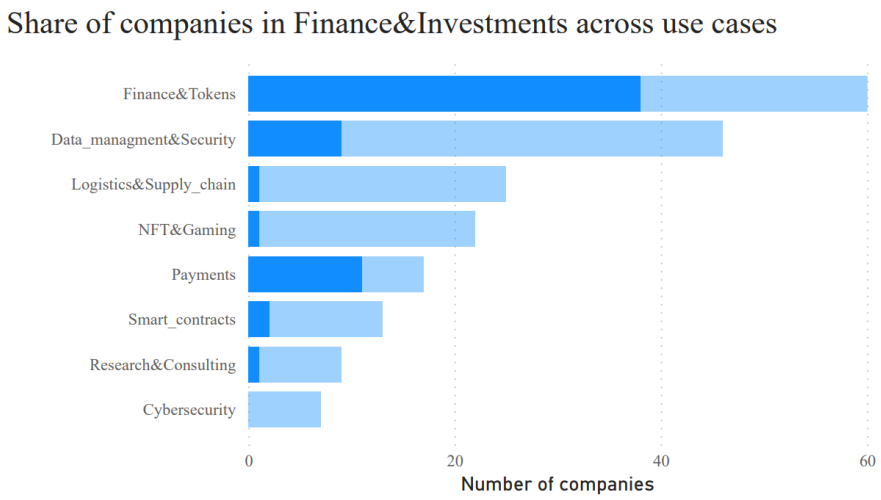

Figure 15: Share of companies in Finance Investments industry implementing blockchain across use cases (authors' picture)

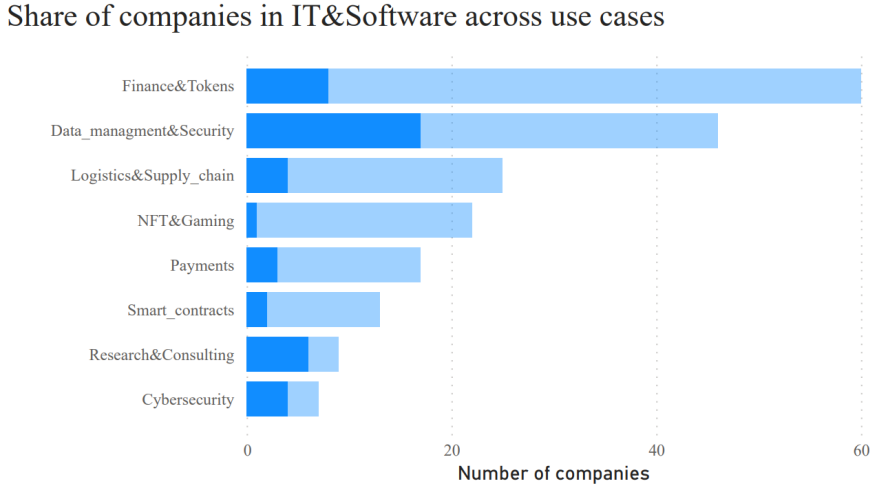

Figure 16: Share of companies in IT Software industry implementing blockchain across use cases (authors' picture)

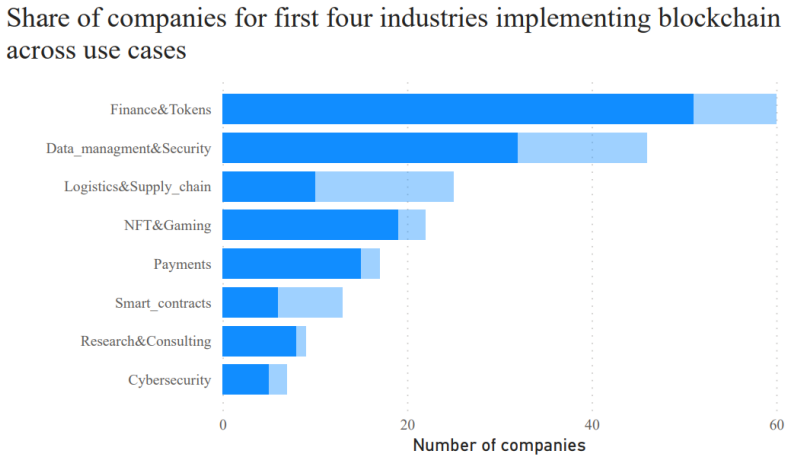

Figure 17: Share of companies in first four industries (Finance, IT, Entertainment, Retail) implementing blockchain across use cases (authors' picture)

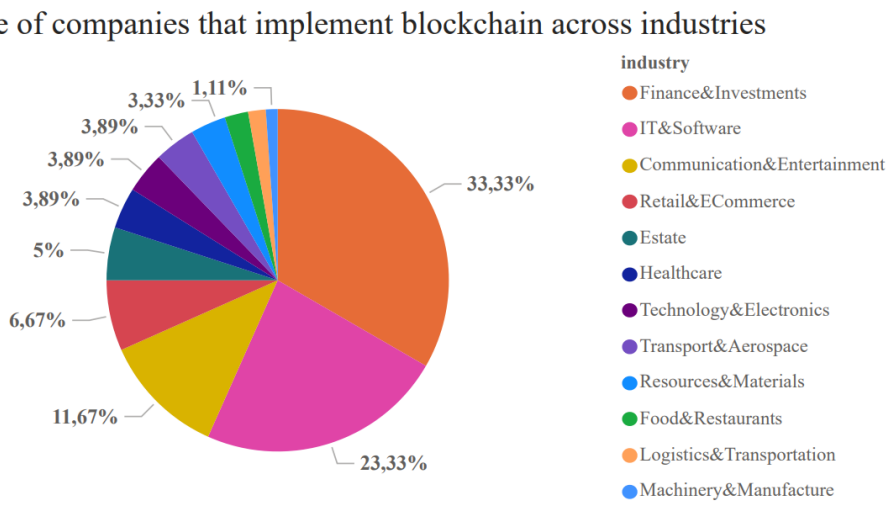

Figure 18: Share of companies implementing blockchain across industries (authors' picture)

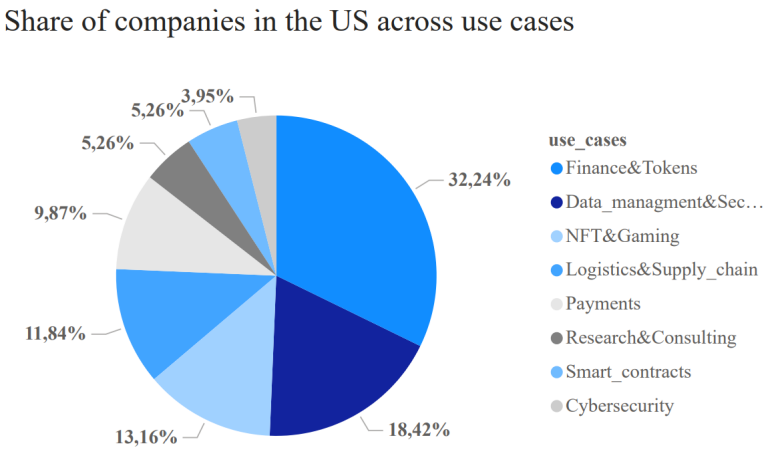

Figure 19: Share of companies implementing blockchain across use cases in the US (authors' picture)

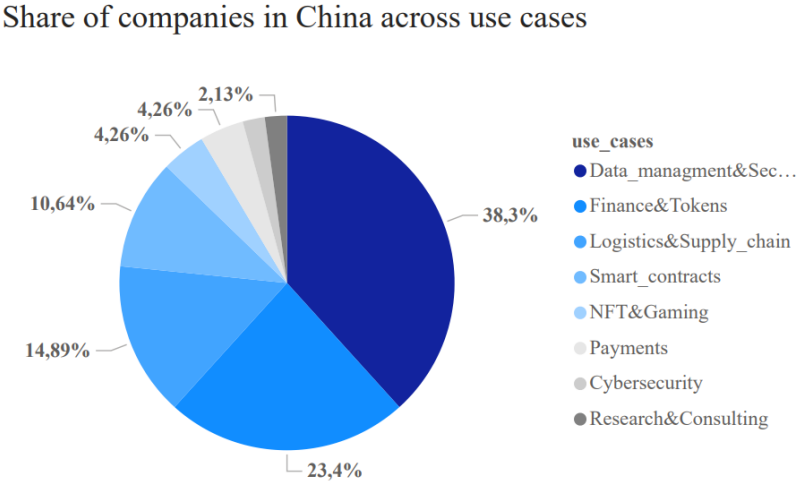

Figure 20: Share of companies implementing blockchain across use cases in China (authors' picture)

Appendix 2. Clustering analysis

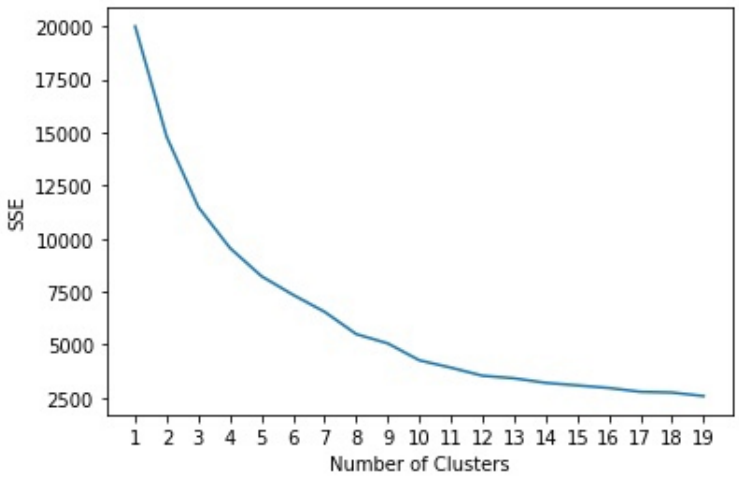

Figure 21: Elbow method for clustering analysis (authors' picture)

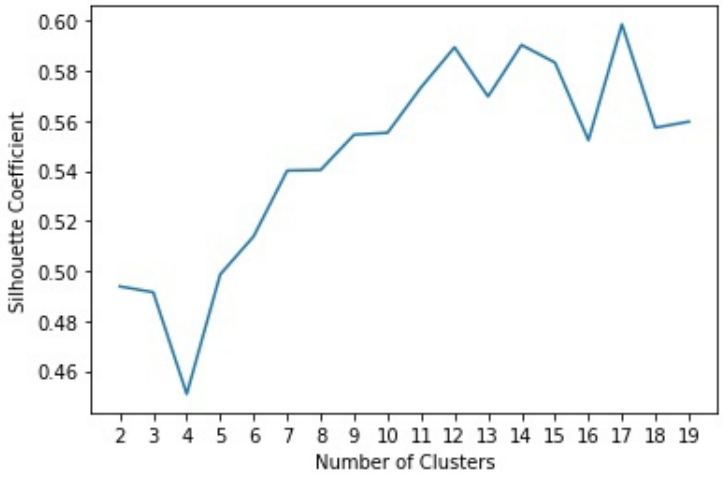

Figure 22: Silhouette coefficient method for clustering analysis (authors' picture)

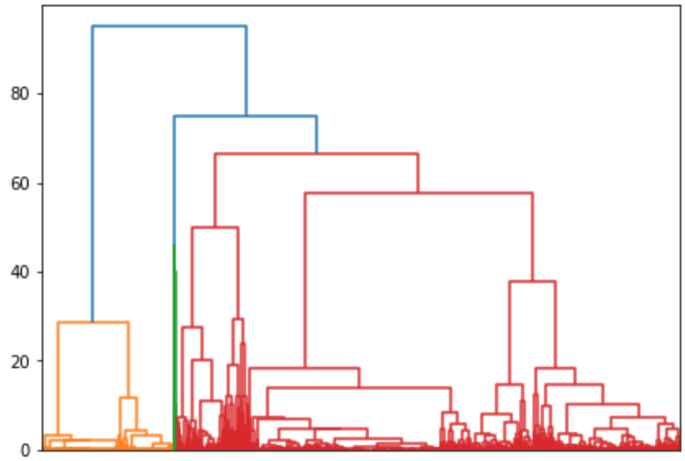

Figure 23: Dendrogram for clustering analysis (authors' picture)

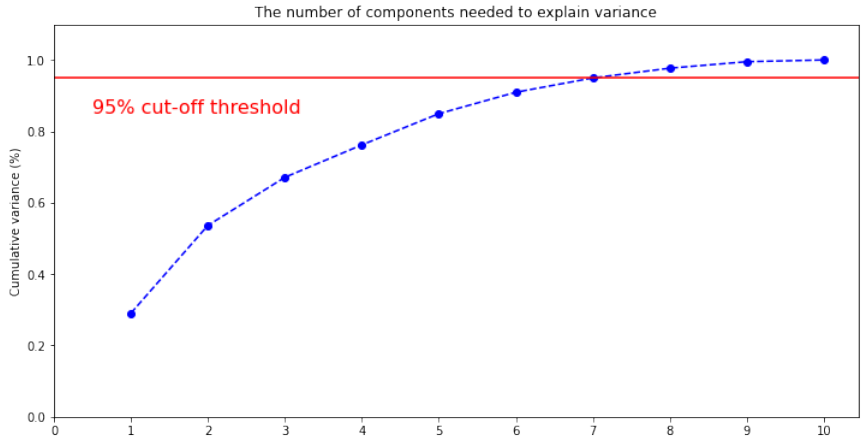

Figure 24: The number of components needed to explain variance (authors' picture)

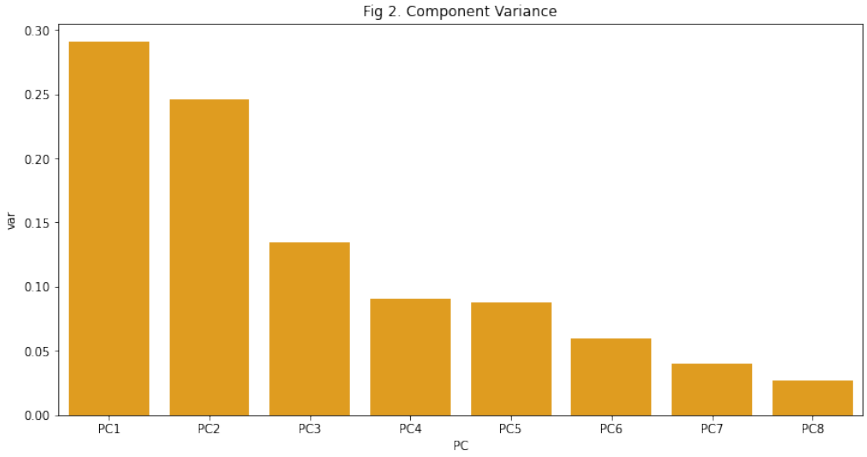

Figure 25: Explained variance by each PC (authors' picture)

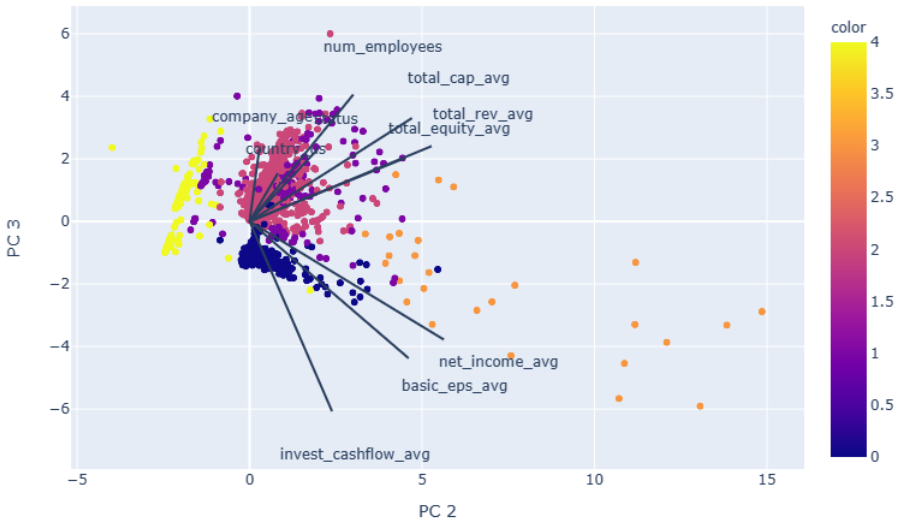

Figure 26: Visualisation of clustering analysis on PCA plot (authors' picture)

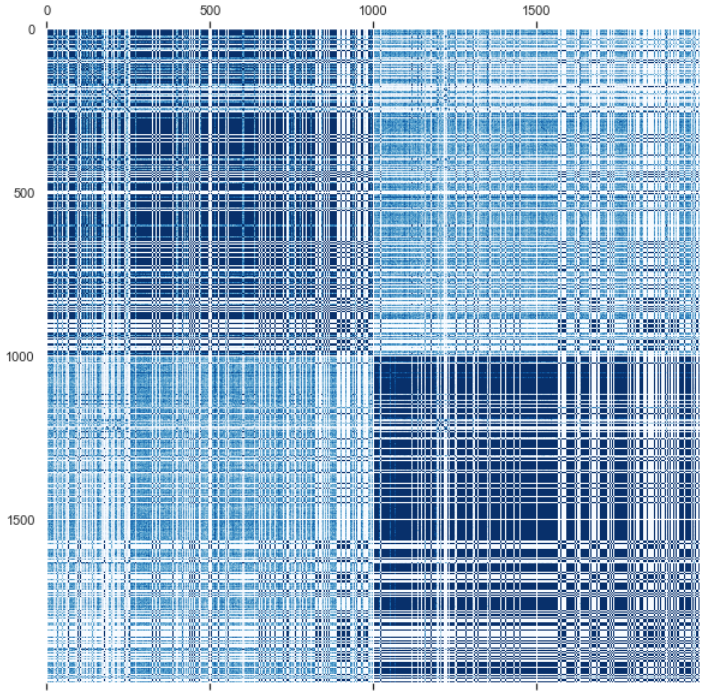

Figure 27: Bootstrapping 5 K-Means clusters to check their stability (authors' picture)

Appendix 3. Classification

Boostrapping logistic regression

We resampled our dataset 5000 times and trained the Logistic Regression model, we ended up with the following distributions of p-values - Figure (28) and coefficients - Figure (29) of predictor variables.

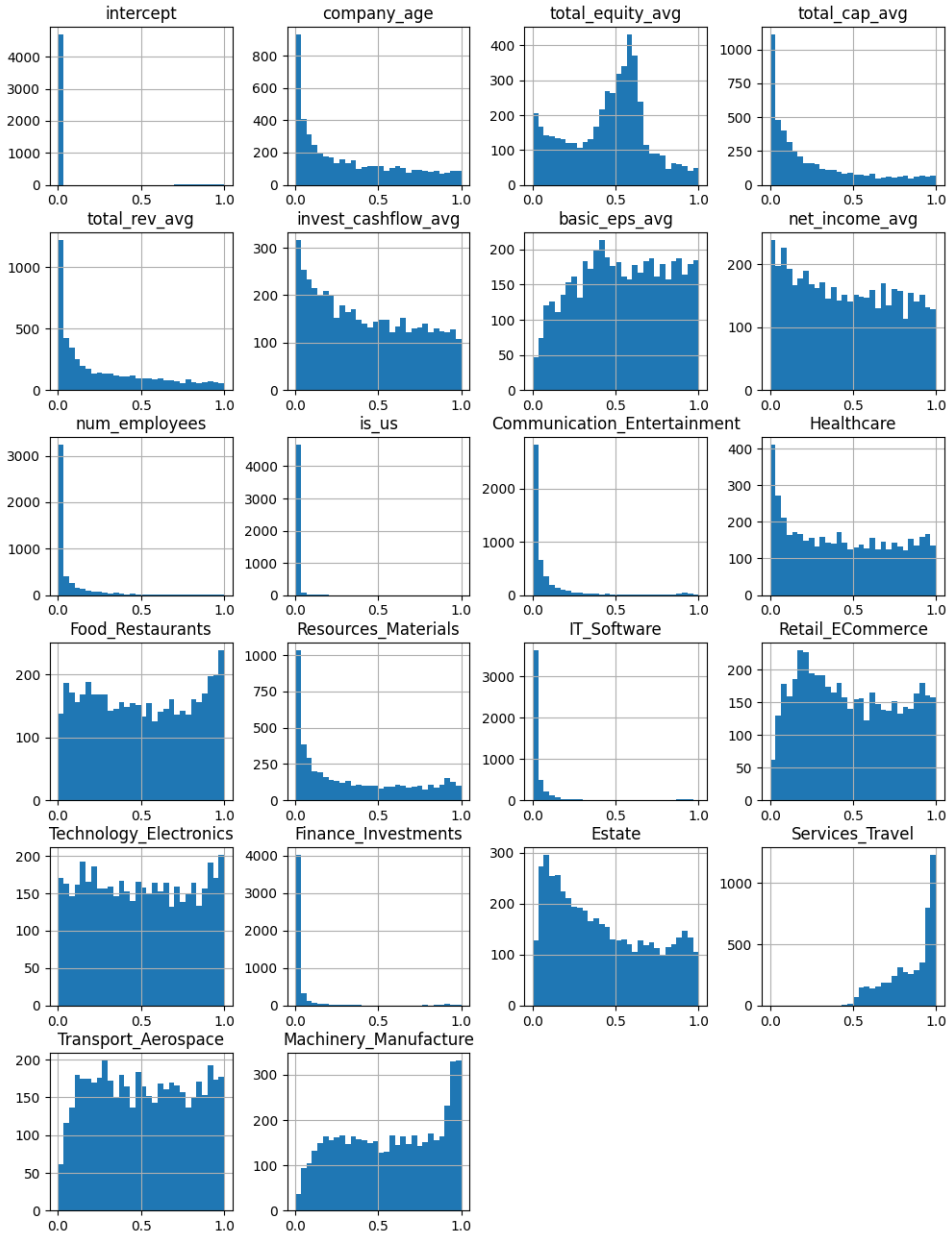

Figure 28: Bootstrapped p-values of Wald's test

Variables which are statistically significant have mostly left-skewed distributions where most of resamples result in the variable being significant at at least 5 % level. Variables which do not seem to be any informative have more or less uniform distribution of p-values.

Below are distributions of regression coefficients themselves. We are able to see that statistically significant variables have their corresponding distributions far away from zero either to the left or to the right. Insignificant variables are mostly centered around 0, meaning that on average they do not seem to have any effect on (*status*) - target variable.

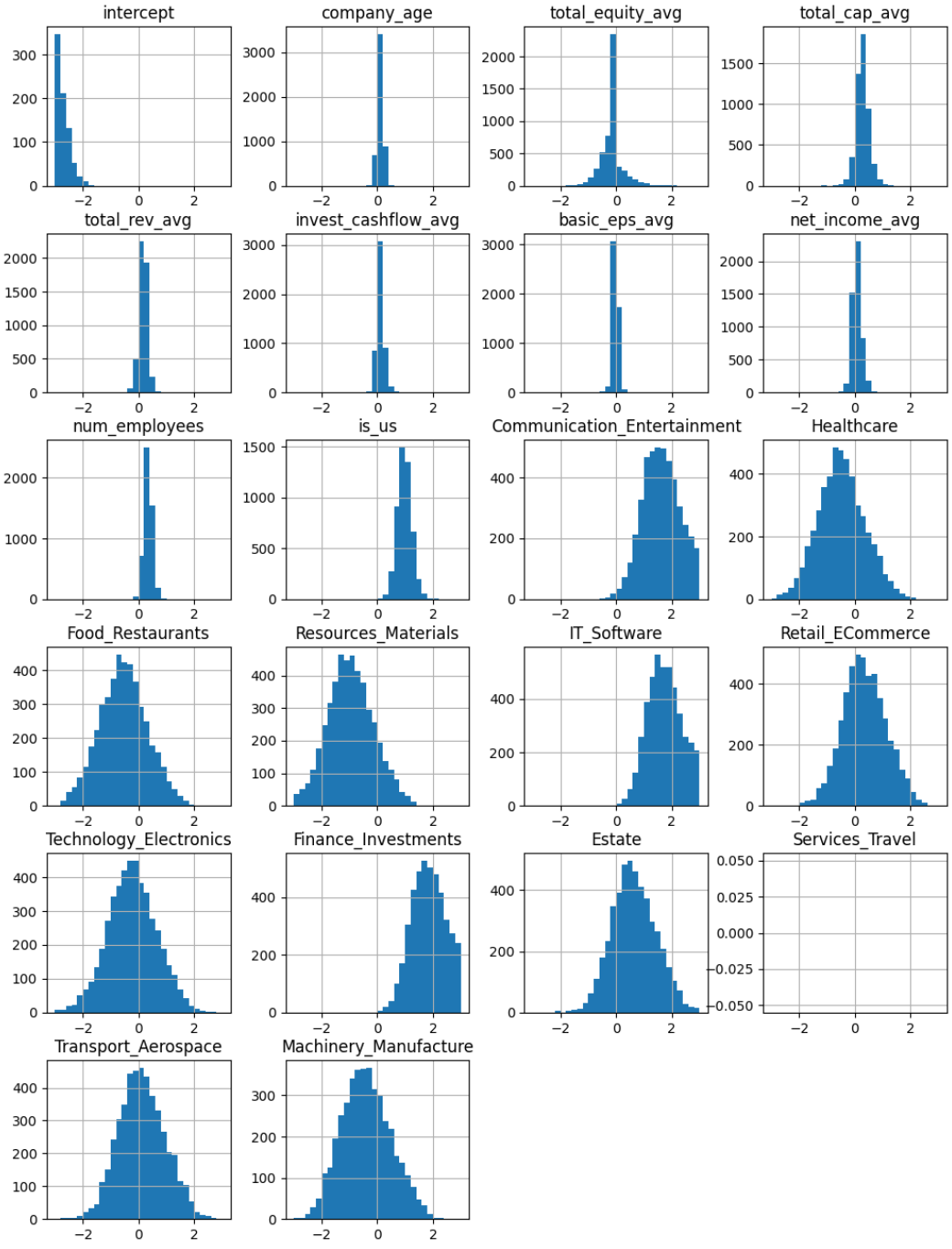

Figure 29: Bootstrapped coefficients of Logistic Regression

Feature importance & Permutation importance

In this section we will have a look at feature and permutation importance of the models, obtained in Classification section.

Permutation importance of a variable refers to the decrease in the scoring metric after randomly shuffling the variable. By this we break down the connection between the variable and target variable and evaluate how much worse the model has become in terms of some scoring metric, in our case PR-AUC. Therefore, variables that do have an impact are those with highest Permutation Importance, meaning that essentially without them model performs worse. Below are permutation importances for RandomForestClassifier - Table (9) and XGBoostClassifier - Table (10) evaluated using test sample. We also made sure to run multiple permutation importance calculations since permutation itself involves randomness with shuffling.

Table 9: Permutation importance for RandomForest Classifier

Variable	Permutation Importance
Finance_Investments**	**0.0441**
total_rev_avg	**0.0416**
Resources_Materials	0.0296
IT_Software*	**0.0269**
net_income_avg	0.0234
basic_eps_avg	0.0192
total_cap_avg	0.0191
Healthcare	0.0104
num_employees**	**0.0091**
total_equity_avg	0.0056
is_us***	**0.0034**
Communication_Entertainment*	**0.0032**
Technology_Electronics	0.0020
Transport_Aerospace	0.0011
Logistics_Transportation	0.0006
Services_Travel	0.0005
Estate	0.0003
Retail_ECommerce	0.0001
Food_Restaurants	-0.0001
Machinery_Manufacture	-0.0014
company_age	-0.0023
invest_cashflow_avg	-0.0049

Table 10: Permutation importance for XGBoostClassifier

variable	Permutation Importance
total_cap_avg	0.0481
IT_Software*	**0.0293**
num_employees**	**0.0245**
is_us***	**0.0242**
Finance_Investments**	**0.0226**
Resources_Materials	0.0202
invest_cashflow_avg	0.0181
net_income_avg	0.0134
total_equity_avg	0.0132
Healthcare	0.0091
company_age	0.0081
Services_Travel	0.0073
Technology_Electronics	0.0035
basic_eps_avg	0.0015
Communication_Entertainment*	**0.0007**
Estate	0.0000
Transport_Aerospace	0.0000
Logistics_Transportation	0.0000
Retail_ECommerce	0.0000
Food_Restaurants	-0.0014
Machinery_Manufacture	-0.0031
total_rev_avg	-0.0044

Feature importance of a variable is another measure of how important a variable is to the model. It is measured as its marginal impact on the evaluation metric. For instance, in order to train RandomForest we used Gini index which decides the optimal way to split data points with current variable. Feature importance measures how much variable inputs to such metric measuring improvement in class separation. Below is the Table (11) with feature importances for XGBoostClassifier:

Table 11: Feature importance for XGBoostClassifier

Variable	Feature Importance
Finance_Investments**	**0.1036**
Resources_Materials	0.0883
IT_Software*	**0.0780**
net_income_avg	0.0715
is_us***	**0.0665**
total_cap_avg	0.0567
total_rev_avg	0.0530
basic_eps_avg	0.0512
Communication_Entertainment*	**0.0470**
num_employees**	**0.0440**
total_equity_avg	0.0427
Technology_Electronics	0.0416
invest_cashflow_avg	0.0410
company_age	0.0383
Services_Travel	0.0369
Machinery_Manufacture	0.0339
Food_Restaurants	0.0332
Healthcare	0.0291
Retail_ECommerce	0.0112
Transport_Aerospace	0.0109
Logistics_Transportation	0.0107
Estate	0.0107

Research notebooks

Research itself has been carried out in Python Notebooks. Below you can find the Github repository will all code used for this project:

Github: `https://github.com/B0R0koko/HSE_Thesis/`

REFERENCES

[1] Agi, M. A. N. and Jha, A. K. (2022). Blockchain technology in the supply chain: An integrated theoretical perspective of organizational adoption. *International Journal of Production Economics*, 247:108458.

[2] Al-Zaqeba, M., Jarah, B., Ineizeh, N., Almatarneh, Z., and Jarrah, M. (2022). The effect of management accounting and blockchain technology characteristics on supply chains efficiency. *Uncertain Supply Chain Management*, 10(3):973–982.

[3] Ashraf, M. U. (2021). A Survey on Data Security in Cloud Computing Using Blockchain: Challenges, Existing-State-Of-The-Art Methods, And Future Directions | Lahore Garrison University Research Journal of Computer Science and Information Technology.

[4] Chowdhury, S., Rodriguez-Espindola, O., Dey, P., and Budhwar, P. (2022). Blockchain technology adoption for managing risks in operations and supply chain management: evidence from the UK. *Annals of Operations Research*.

[5] Cole, R., Stevenson, M., and Aitken, J. (2019). Blockchain technology: implications for operations and supply chain management. *Supply Chain Management: An International Journal*, 24(4):469–483. Publisher: Emerald Publishing Limited.

[6] Deepa, N., Pham, Q.-V., Nguyen, D. C., Bhattacharya, S., Prabadevi, B., Gadekallu, T. R., Maddikunta, P. K. R., Fang, F., and Pathirana, P. N. (2022). A survey on blockchain for big data: Approaches, opportunities, and future directions. *Future Generation Computer Systems*, 131:209–226.

[7] Deng, N., Shi, Y., Wang, J., and Gaur, J. (2022). Testing the adoption of Blockchain Technology in Supply Chain Management among MSMEs in China. *Annals of Operations Research*.

[8] Elbashbishy, T. S., Ali, G. G., and El-adaway, I. H. (2022). Blockchain technology in the construction industry: mapping current research trends using social network analysis and clustering. *Construction Management and Economics*, 40(5):406–427. Publisher: Routledge _eprint: https://doi.org/10.1080/01446193.2022.2056216.

[9] Farnoush, A., Gupta, A., Dolarsara, H. A., Paradice, D., and Rao, S. (2022). Going beyond intent to adopt Blockchain: an analytics approach to understand board member and financial health characteristics. *Annals of Operations Research*, 308(1):93–123.

[10] Happy, A., Chowdhury, M. M. H., Scerri, M., Hossain, M. A., and Barua, Z. (2023). Antecedents and consequences of blockchain adoption in supply chains: a systematic literature review. *Journal of Enterprise Information Management*, 36(2):629–654. Publisher: Emerald Publishing Limited.

[11] Jum'a, L. (2023). The role of blockchain-enabled supply chain applications in improving supply chain performance: the case of Jordanian manufacturing sector. *Management Research Review*, ahead-of-print(ahead-of-print).

[12] Kurdi, B., Alzoubi, H., Alshurideh, M., Alquqa, E., and Hamadneh, S. (2023). Impact of supply chain 4.0 and supply chain risk on organizational performance: An empirical evidence from the UAE food manufacturing industry. *Uncertain Supply Chain Management*, 11(1):111–118.

[13] Lim, M. K., Li, Y., Wang, C., and Tseng, M.-L. (2021). A literature review of blockchain technology applications in supply chains: A comprehensive analysis of themes, methodologies and industries. *Computers & Industrial Engineering*, 154:107133.

[14] Lin, S.-Y., Zhang, L., Li, J., Ji, L.-l., and Sun, Y. (2022). A survey of application research based on blockchain smart contract. *Wireless Networks*, 28(2):635–690.

[15] Morkunas, V. J., Paschen, J., and Boon, E. (2019). How blockchain technologies impact your business model. *Business Horizons*, 62(3):295–306.

[16] Pan, X., Pan, X., Song, M., Ai, B., and Ming, Y. (2020). Blockchain technology and enterprise operational capabilities: An empirical test. *International Journal of Information Management*, 52:101946.

[17] Xu, X. and Choi, T.-M. (2021). Supply chain operations with online platforms under the cap-and-trade regulation: Impacts of using blockchain technology. *Transportation Research Part E: Logistics and Transportation Review*, 155:102491.

Internet resources:

1. Website "Cointelegraph" [Online resource]:
`https://cointelegraph.com/news/china-accounts-for-84-\`
`of-all-blockchain-patent-applications-but-there-s-a-catch` (date of assessing: 02.01.2023)

2. Website “FinancesOnline” [Online resource]:
`https://financesonline.com/blockchain-statistics/` (date of assessing: 02.01.2023)

3. Website “Value.Today” [online resource]:
`https://www.value.today/` (date of assessing: 05.02.2023)

4. Website “Yahoo Finance” [online resource]:
`https://finance.yahoo.com/` (date of assessing: 20.02.2023)

5. Website “Blockdata” [online resource]:
`https://www.blockdata.tech/` (date of assessing: 20.03.2023)