UNIVERSITY OF BELGRADE

FACULTY OF ECONOMICS

MASTER THESIS

PREDICTING SUCCESS
OF FINTECH STARTUPS
USING MACHING
LEARNING ALGORITHMS

Kristina Bradic (22/2801) **Phd Dragana Radojicic**

BELGRADE
AUGUST 2024

Statement of Academic Integrity

Student: Kristina Bradić

Student ID number: 2801/22

The author of the work entitled:

Predicting Success of FinTech Startups using Machine Learning Algorithms

By signing, I declare:

- that the work is solely the result of my own research work;
- that I indicated or cited the work and opinions of other authors that I used in this paper in accordance with the Instructions;
- that all works and opinions of other authors are listed in the list of literature/references that are an integral part of this work and written in accordance with the Instructions; that I have obtained all permissions for the use of the author's work that are fully included in the submitted work and that I have clearly stated this;
- that I am aware that plagiarism is the use of other people's works in any form (such as quotations, paraphrases, images, tables, diagrams, designs, plans, photographs, films, music, formulas, websites, computer programs, etc.) without stating the author or presenting other people's works as mine, punishable by law (Act on Copyright and Related Rights, Official Gazette of the Republic of Serbia, No. 104/2009, 99/2011, 119/2012), as well as other laws and relevant acts of the University of Belgrade;
- that I am aware that plagiarism includes presenting, using and distributing the work of lecturers or other students as one's own;
- that I am aware of the consequences that proven plagiarism can have on the submitted master's thesis and my status;
- that the electronic version of the master's thesis is identical to the printed copy and I agree to its publication under the conditions prescribed by the University's acts.

Belgrade, 28.8.2024.

Signature Брадић Кристина

Izjava o korišćenju

Dozvoljavam da se objave moji lični podaci vezani za dobijanje akademskog naziva master ekonomiste, kao što su ime i prezime, godina i mesto rođenja i datum odbrane rada.

Ovi lični podaci mogu se objaviti na mrežnim stranicama digitalne biblioteke, u elektronskom katalogu i u publikacijama Univerziteta u Beogradu – Ekonomskog fakulteta.

Ovlašćujem biblioteku Univerziteta u Beogradu – Ekonomskog fakulteta da u svoj digitalni repozitorijum unese moj završni (master) rad pod naslovom:

Predicting Success of FinTech Startups using Machine Learning Algorithms

koji je moje autorsko delo.

Završni (master) rad sa svim prilozima predao/la sam u elektronskom formatu pogodnom za trajno arhiviranje.

Moj završni (master) rad, pohranjen u Digitalnom repozitorijumu Univerziteta u Beogradu – Ekonomskog fakulteta i dostupan u otvorenom pristupu, mogu da koriste svi koji poštuju odredbe sadržane u CC BY licenci Kreativne zajednice (Creative Commons), a kojom je dozvoljeno umnožavanje, distribucija i javno saopštavanje dela, i prerade, uz adekvatno navođenje imena autora, čak i u komercijalne svrhe.

Potpis autora

U Beogradu, 28.8.2024.

Apstrakt

Predviđanje uspeha FinTech startupa je izazovno kako za investitore, tako i za istraživače. Ipak, zahvaljujući kompanijama poput Crunchbase koje prikupljaju podatke o startupovima, postalo je moguće da se kreira i evaluira model mašinskog učenja na bazi primera postojećih startup kompanija na tržištu.

Cilj master rada je kreiranje modela mašinskog učenja koji će uputiti na binarnu klasifikaciju FinTech startupova, tj. omogućiti predviđanje da li će startup doživeti uspeh ili ne. Iskorak u odnosu na prethodna istraživanja u kojima su primenjivani algoritmi mašinskog učenja u predviđanju uspeha startupova podrazumeva filtriranje isključivo kompanija koje pružaju FinTech usluge.

Čitalac se kroz rad bolje upoznaje sa fenomenom FinTech industrije i njemu pripadajuće InsurTech industrije, trendovima razvoja, kao i elementarnim informacijama vezanim za algoritme mašinskog učenja i njihovu primenu u literaturi. U radu su kombinovane istraživačke metode poput pregleda literature i kvantitativne analize sekundarnih podataka.

Uporedili smo tri algoritma mašinskog učenja – Random Forest, Support Vector Machine i Extreme Gradient Boosting. Oslanjajući se na rezultate metrika evaluacije poput f1-score, accuracy, recall, true positive rate, i true negative rate, Extreme Gradient Boosting algoritam se pokazao kao superioran u odnosu na preostale algoritme. Takođe, utvrdili smo da karakteristike poput geografske lokacije, usluga koje FinTech startupovi pružaju, iznosa investicije po rundama i tehnologije, imaju najveći uticaj na uspeh FinTech startupova.

Ključne reči: FinTech, startup, algoritmi mašinskog učenja, analiza podataka

Abstract

Predicting the success of FinTech startup has been challenging for both investors and researchers. However, thanks to companies like Crunchbase that collect data about startups, it became possible to create and evaluate machine learning model based on existing startup companies on the market.

The objective of the master thesis is to create machine learning model that would lead to binary classification of FinTech startups, videlicet whether they will achieve success or not. A step forward in comparison with previous literature that have applied machine learning algorithms in predicting startup success is filtering exclusively FinTech companies.

The reader is better informed with FinTech phenomenon and its belonging InsurTech industry, trends, and impacts, so as elementary information about machine learning algorithms and their application in literature. We combined two research methods: literature review and quantitative analysis of secondary data.

We compared three machine learning algorithms - Random Forest, Support Vector Machine and Extreme Gradient Boosting. Based on the results of evaluation metrics such as f1-score, accuracy, recall, true positive rate, and true negative rate, the Extreme Gradient Boosting algorithm is shown to be superior in comparison with remaining algorithms. We also found that characteristics such as geographic location, services provided by FinTech startups, amount of investment per round, and technology used, have the greatest impact on the success of FinTech startups.

Key words: FinTech, startup, machine learning algorithms, data analysis

Table of Contents

INTRODUCTION

In the rapidly evolving financial technology landscape, grasping the determinants of startup success has become increasingly crucial. The FinTech sector, characterized by its innovative approach to traditional financial services, has recorded a significant surge in activity over recent years. This surge is fueled by technological advancements, shifting consumer behaviors, and regulatory changes. Within this context, this master thesis aims to explore the complex dynamics of fintech success, focusing on merger and acquisitions (M&As) and Initial Public Offerings and the role of machine learning algorithms in predicting startup success.

The idea of the thesis is to approach the FinTech phenomenon and machine learning application in uncertain and evolving industries such as financial technology. Moreover, the thesis investigates the predictive power of machine learning algorithms in forecasting the success of fintech startups. By leveraging data from the Crunchbase, we undertook a comprehensive analysis to identify patterns and insights that contribute to the predictive model.

The aim of the study is to create and select the machine learning model that performs binary classification (successful or not) of startups the most accurately. It differs from the previous studies due to its industry criterion and implementation of three algorithms (Random Forest, Support Vector Machine and XGBoost) in that context. During the research, we expected to achieve several technical objectives, including data analysis experiment setup, and the presentation of results.

Central to this research are hypotheses that specific factors such as geographical location and investor count are the most relevant contributors to a startup’s success besides the timing of market entry relative to the financial crises. The data analysis of these variables, complemented with machine learning techniques, strives to offer an understanding of what drives fintech success. Additionally, we aim to discover whether the results of different machine learning algorithms differ.

The thesis encompasses two research methods: literature review and quantitive analysis of secondary data. A literature review involves researching, reading, analyzing and comparing scholarly literature. On the other hand, through a meticulous process of data collection, preprocessing, and analysis, we develop and evaluate the machine learning model. Employing algorithms such as Random Forest, Support Vector Machine (SVM), and XGBoost, the research engages in a comparative analysis to determine the optimal predictive approach. Additionally, it discovers the importance of features within the models, aiming to confirm the initial hypotheses regarding critical factors for startups' success.

The first chapter is about FinTech startups. It contains several sections that concern the definition, trends, the perspective of the FinTech industry, the InsurTech industry as a perspective subcategory of FinTech, the determinants of success and the definition of success in the research. The reader is informed about the basic concepts of the FinTech industry, merger and acquisition, initial public offering and the features that possibly affect the success of startups.

The second chapter regards the machine learning phenomenon. We aim to explain basic concepts, trends and perspectives of the machine learning field and chosen algorithms. Additionally, we look back on previous machine learning applications on startup success prediction.

Finally, the last chapter is the presentation of data collection, preprocessing, analysis and prediction model. This segment establishes a practical framework for understanding prediction models and critical success factors. We succeed in validating our hypotheses regarding the result of machine learning algorithms and the most relevant startup features.

This thesis advances the scholarly discourse by furnishing empirical insights into the key determinants of FinTech startup success, while also illustrating the role of machine learning in refining predictive models. By integrating theoretical frameworks with empirical analysis, this work seeks to elucidate the intricate dynamics that underpin the prosperity and growth of FinTech ventures.

1. Fintech startups

1.1 FinTech Startup: Definition, impact, and trends

In the past decade, we have witnessed grinding technological evolution manifesting through increased use of abundant data, artificial intelligence and less in-person communication. The Great Recession and regulatory reform followed by the COVID-19 crisis caused a deterioration of public trust and disruption to financial services. The novel finance start-up companies redefined the existing paradigms in many aspects of the financial landscape. (Anagnostopoulos, 2018).

As more technology entrepreneurs enter the industry, the more they modify it to the social needs. Frequent rapid changes in the FinTech industry cause experts in the field and the external stakeholders find FinTech term ambiguous. (Zavolokina et al., 2016) Therefore, it is unsurprising that an emergent body of literature strives to define the FinTech phenomenon. Harasim (2021) points out that the FinTech ecosystem mainly consists of small, young technology companies whose core business is delivering innovative products and financial services, aiming to improve user experience. Not only does FinTech implement new digital technologies to financial services, but also applies digital processes in all aspects of business model and products. (*OECD Digital Economy Outlook 2015 | READ Online*, n.d.) The authors describe FinTech as a service sector leveraging mobile-centric IT technology to enhance financial system efficiency (Kim et al., 2016). Varga (2017, p. 201) further asserts that the primary objective of Fintech to offer innovative, technology-driven financial services with added value. Finally, Zavolokina et al. (2016) characterize FinTech as a financial service designed to achieve cost reduction, streamline business process, and foster rapidity, flexibility, and innovation through the application of cutting-edge technologies.

Specialized FinTech service fields continuously multiply, offering solutions for both B2B and B2C customers (Bethlendi & Szőcs, 2022). While the digitalization of payments and customer service remains central to the revolution of FinTech, the broader digital transformation has ushered in new capabilities and domains, such as Machine Learning, Internet of Things, Artificial Intelligence, Big Data Analytics. (Bethlendi & Szőcs, 2022). As noted by the World Economic Forum, the array of FinTech services encompasses deposits and lending, payments, insurance, market provisioning, capital raising and investment management (*The Future of FinTech*, n.d.). By analyzing Martin company, Gakman identified seven similar FinTech business models: Payments, Lending, International money transfer, Equity financing, Insurance, and Personal finance (Gakman, 2022.).

Innovation and resource-based theories suggest that the rise of high-growth fintech startups is largerly fueled by entrepreneurial expertise and the capacity to recognize lucrative opportunities (Acs et al., 2009; Alvarez & Busenitz, 2007). Cojoianu et al. (2021) argue that, in the early stages of startup development, expertise in the IT sector plays a more crucial role than that of the financial services sector. However, as these startups expand and seek financing, the significance of financial sector knowledge becomes more pronounced. Additionally, regions characterized by higher levels of trust in financial services tend to attract greater FinTech investment. (Cojoianu et al., 2021)

Social norms also impact potential entrepreneurs to think entering a new field is legitimate and likely to bring success (York & Lenox, 2014). Haddad and Hornuf (2019) explore the economic and technological catalysts that compel entrepreneurs to found a FinTech enterprise. The positive impact on the development of startup formations is proved in well-developed countries where an encouraging economic climate is cultivated. Increasingly qualified labour force, available venture capital, Internet and sophisticated mobile network infrastructure are the key drivers of the emergence of FinTech startups. (Haddad & Hornuf, 2019) Laidroo et Avarmaa (2020) discover that the fruitful ground for the FinTech community are small countries, countries with developed ICT technologies and clusters and countries that have undergone a crisis during the recent decade. (Laidroo & Avarmaa, 2020) Zavolokina et al. (2016) conclude that an amalgam of regulatory, financial, and technological factors provokes the financial innovation and not one factor exclusively.

Eventually, FinTech companies act as a new financial intermediary between clients and banks, chipping off the most lucrative horizontally and vertically integrating services from traditional banks (A. Boot et al., 2021). The traditional bank business model relies on customer accounting records and payment flows, but with the digital surge, non-financial data has become available for financial decision-making. “Digital footprint” or non-financial customer data is obtained through consumer platforms in e-commerce, online search, and social media. (A. Boot et al., 2021) Berg et al. (2020) prove the incorporation of “digital footprint” into credit score analysis performs as well as traditional borrower risk assessment. While FinTech may easily obtain non-financial data, banks do not have access to this data type due to rigid regulatory policies. Therefore, the ability of FinTech startups to collect and process non-financial data leverages their lack of access to borrowers' financial information. (A. W. A. Boot, 2000; Botsch & Vanasco, 2019).

Nevertheless, the relationship between FinTech entrants and banks is not always described as rivalry. FinTech innovations enhance the cost efficiency of the banks and reform the technology used by banks. (Kou et al., 2021) Li et al. (2022) prove the clustering of financial data aids banks in fraud detection, default rate prediction and credit analysis. Similarly, FinTech improves SME's efficiency and competitiveness (Abbasi et al., 2021). Small and Medium Enterprises (SMEs) play a crucial role in the economy, given their contribution towards GDP, employment and tax revenue (Rosavina et al., 2019). Therefore, it is essential to identify distinct apparatus that may affect SMEs' efficiency. FinTech startups incorporate big data in their borrowers' default rate prediction, enabling SMEs to qualify for loans at lower interest rates (Jagtiani & Lemieux, 2019). FinTech lending causes search cost-reduction for SMEs as FinTechs are able to process loan applications much faster than traditional banking lending (Gomber et al., 2018; Rosavina et al., 2019; Sangwan et al., 2019). Further**,** the implementation of innovative technologies (such as Fintech) increases SME's survival rate (Hassan et al., 2018). FinTech provides robo-advisors as a cost-efficient way to tailor SMEs' portfolios. Robo-advisors acquire information, predefine parameters of investment goals, aversion to risk, financial background and process data to develop portfolio allocation and investment recommendations with little or no human intervention (Gomber et al., 2018). SMEs are likely to improve revenue and exploit lucrative business opportunities, as FinTech provides quick access to funds (Gomber et al., 2018; Sangwan et al., 2019). Generally, FinTech startups have been proven to ease innovation in the financial sector as a whole (M. A. Chen et al., 2019).

There is no doubt the days of brick-and-mortar banking are over. New entrants challenge incumbents by offering user-friendly interfaces and cost-efficient communication channels via mobile apps. To prevent market roles replacement, banks will strive to upgrade their information systems and processes, shift to cloud computing, and offer digital platforms as their products and services. However, the massive IT transition is likely to be hindered due to the banks' organizational complexity and regularity policies. The FinTech startups may successfully compete with banks, but their success still has evident limits. Banks are true examples of big bureaucratic organizations, committed to their existing product mix, and subjected to rigid regulation, but they have the advantage of big consumer base, reputation, economies of scale and economies of scope. (Stulz, 2019)

The future trend analyses indicate that the BigTech companies are the ones to pose a more serious threat to traditional banks rather than FinTech (Stulz, 2019). The FinTech startups aim to compete with banks for a specific product line, while BigTechs have the capacity to challenge banks by attacking numerous market niches such as lending SMEs and consumer finance (A. Boot et al., 2021).

The BigTechs develop a sheer ecosystem around their cloud services, including business-to-business (B2B) marketplaces, as their non-financial core businesses allow them to gather large amounts of data via web-scraping. Prominent financial industry influencers vividly described the difference between these two phenomena, stating FinTech are making faster horses while BigTech are working with airplanes. (*Jim Marous*, n.d.)

According to S&P Global Market Intelligence data, FinTech funding dropped globally by 36% year over year to $6 billion in the third quarter of 2023. While venture capital pressures eased for mature start-ups, as deal count and funding values increased by 30% year over year, seed-stage fintech investments plunged significantly. With declining personal savings rates and rising interest rates, the consumer market has become less attractive. Some of fintech entrepreneurs are responding to these trends by transforming their business models to business-to-business (B2B) oriented. The B2B business model represents a stable and more lucrative strategy for its higher revenue potential and subscription-based revenue. (*Fintech Funding*, n.d.)

Due to the lack of quantitative valuation methods that correctly calculate the fundamental value of early-stage companies, venture capitalists often bank on heuristics or gut to reach investing decisions. Pattern recognition based on previous experience, gut, is inevitably included in the evaluation process of early-stage companies, but the quantitative approach has proven to be more useful. (Corea et al., 2021) Human heuristics tend to have issues in processing abundant data, and as well reflect all sorts of biases ranging from sample selection bias from confirmatory and hindsight to overconfidence (Åstebro & Elhedhli, 2006). On the contrary, a more data-driven evaluation process (supported with data mining and machine learning) lowers the risk associated with investing in early-stage companies and eventually brings higher returns on investments (ROI) (Cao et al., 2023).

1.2 InsurTech – The Evolving Frontier in the FinTech Ecosystem

The FinTech revolution has deeply transformed the landscape of the financial industry, disrupting traditional banking, investment management and payment processing with innovative technologies and business models. However, one segment of the financial sector that has been relatively slow to experience this wave of innovation is the insurance industry. InsurTech, a specific branch of the FinTech industry consisting of traditional and non-traditional market players, aims to deliver specific solutions to the insurance industry by exploiting information technology. Despite its slower start rate in comparison with other FinTech subsectors, InsurTech has gained significant momentum in recent years. (Stoeckli et al., 2018)

At its core, insurance facilitates the transfer of risk from the customer to the insurer. InsurTech, like other FinTech sectors, is distinguished by the integration of BigTech companies that introduce innovative platforms and leverage advanced technologies. We can categorize InsurTech firms into three distinct types: distributors, technology solution providers, and full-stack carriers. Distributors collaborate with established insurers to market policies via their platforms, whereas full-stack carriers are licensed insurance entities employing state-of-the-art technologies to underwrite policies and manage claims. Technology solutions providers, meanwhile, target specific segments of the insurance value chain, enabling traditional insurers to enhance operational efficiency. (Bian et al., 2023) It is noteworthy that over the half of InsurTech transactions are concentrated in distribution, while the premiums garnered by full-stack carriers represent less than 1% of the total premiums within the insurance industry. This constrast underscores the differing competitive dynamics between InsurTech and the traditional insurance sector, as opposed to the competition between FinTech lending and conventional banking. (Watson, Re, & Insights, 2020) Table 1 represents an overview of the categories that fall into the three main InsurTech business model types (Braun & Schreiber, 2017).

Table 1 – *Overview of InsurTech categories*

Description	**What They Offer**
Comparison Portals	Provide online comparisons between different insurance products and provider classes
Digital Brokers	Brokerage of insurance policies through mobile apps or portals
Internet of Things	Collect vast data via smart devices
Big Data Analytics & Insurance Software	Deliver software solutions
Peer-to-Peer Insurance	Connect private parties for mutual insurance coverage
On-Demand Insurance	Deliver coverage for defined periods of time
Insurance Cross Sellers	Offer insurance as complements to the product mix
Digital Insurers	Provide digitalized insurance solutions available via online channels
Blockchain & Smart Contracts	Deliver solutions for temper resistant database system for transaction

Source: Braun, A., & Schreiber, F. (2017). *The Current InsurTech Landscape: Business Models and Disruptive Potential* (Research Report 62). I.VW HSG Schriftenreihe. https://www.econstor.eu/handle/10419/226646

Technological innovations have led to changes in the behavior of the customers and the specifications of the objects. Vehicles, houses, factories, and watches are digitally equipped with sensors and connectivity, allowing insurance companies to systematically monitor each operation with the Internet of Things (IoT). (*Unlocking the Potential of the Internet of Things* | *McKinsey*, n.d.) On the other hand, insurance companies can exploit lucrative opportunities in the field of health and life insurance. For example, more than 40% of people in the US possess wearable technology products. A significant part of daily communication is now digital. Consumers are more prone to shop online than in the physical store. Thus, more than half of decision-relevant shopping information derives from digital sources. (vor dem Esche & Hennig-Thurau, 2014)

The UK-based InsurTech firm proves how competitive advantage is based on innovative business models rather than superior technology or products. Insurethebox company integrated available technology into its business model, offerring a car insurance policy with several additional features such as a bonus for safe driving behavior. The information about each client is gathered by a telematics box installed into his or her car. Hence, we can conclude that ongoing digitalization nurtures much of the potential and unexploited opportunities for all InsurTech providers. (Braun & Schreiber, 2017)

The InsurTech market is increasingly shaped by the advent of digital distributors, such as the Sure platform in the United States, Ant in China, and Insurethebox in the United Kingdom. These innovators are revolutionizing the insurance landscape by automating premium payments via smart contracts, collaborating with traditional insurers to deliver innovative, scenario-based products, and pioneering digital marketing approaches. A case study reveals that InsurTech is instrumental in reducing market concentration within the non-life insurance sector, particularly where products are largely commoditized. Through digital distribution channels, InsurTech firms enable small and medium-sized non-life insurers to expand their reach, offering platforms that lower search and commission costs while transcending geographic limitations. However, when it comes to life insurance products, consumers prefer to purchase them in person from large opulent insurance companies. The reputation of the insurance company represents a pivotal role in penetrating life insurance market due to the high premiums and long duration. (Bian et al., 2023)

InsurTech shares real-time information among numerous stakeholders. The phenomenon helps in optimizing insurance processes, but also strives to protect readily available data. Due to the nature of the insurance sector, regulations vary across regions, hindering startups’ ability to scale. The essential aspects of the InsurTech paradigm are a strong internet network platform, intelligent systems, flexible organization and competencies, and automated control. (Nicoletti, 2020)

1.3 Start-up Success: IPOs and M&A

The evolution of every startup initially starts with the pre-seed funding stage, when entrepreneurs invest their own financial capital or seek external financial aid from friends, family, or angel investors. Angel investors are opulent market participants who deploy their capital in emerging businesses in exchange for ownership equity. (Morrissette, 2007) The pre-seed funding stage is followed by the seed funding stage when the company creates a minimum valuable product (MVP). In this phase, typical investors are angel investors, incubators, and venture capitalists. Venture capitalists secure funding from institutional investors to invest in entrepreneurial ventures and portfolio companies, with the objective of achieving substantial capital returns. (Da Rin & Hellmann, 2020). The next stage of startup funding is called series funding. Series funding provides venture capital in one up to five funding rounds. Finally, the financial lifecycle of a company terminates when the company faces success. (*Finding The Most Significant Predictors of Startup Success with Machine Learning*, n.d.)

The route to the success of a start-up is determined by two main exit strategies. The start-up can either receive additional financing as its stakeholders sell shares to the public through an IPO (Initial Public Offering) or it can be acquired or merged (M&A) with an existing company. (Guo et al., 2015)

The lifecycle of every startup commences with an entrepreneurial idea supported by private equity capital. As a startup evolves, it seeks to raise additional capital through IPO. (Jain & Kini, 1999) An IPO refers to the process of the first stock sale by a private company (Liu & Li, 2014). Nevertheless, the post-IPO period may shape a startup into a failure, subject of acquisition or a thriving independent company. There are several reasons why firms go public. One of the explanations is because a company reaches the growth stage and needs to finance existing and future investments. (Jain & Kini, 1999) Other evidence shows that the IPOs are initiated to restore companies' accounts after a period of high growth and investments (Brau & Fawcett, 2006; Farinos & Sanchis, 2009). The short-run underpricing of an IPO occurs in every stock market, but the difference is in the amount of underpricing. The average initial return depends on selling mechanism, institutional constraints, and differences in features companies going to the public. The higher institutional constraints bind, and the younger companies go to the public, the higher initial returns tend to be. (Loughran et al., 1994)

Merger and Acquisition is the process of complete or partial consolidation of companies’ property rights under predefined conditions to further develop their competitive advantage. M&A plays a crucial role in a highly competitive market for its synergy effects. As soon as complementarity between companies from different features such as brand, channel or technology is established, economies of scale, tax advantage and increasing market power arise. (Gaughan, 2010; Weber & Dholakia, 2000; Wei et al., 2009)

The universally optimal exit strategy for a company does not exist, for it is contingent on multiple factors. The financial market conditions at the time of the exit, profitability of the startup, the asymmetry of information between potential buyers and new investors, and the characteristics of venture capitalists among many other factors determine which exit strategy is the most lucrative to be executed. (Guo et al., 2015) The paper additionally proves that start-ups with a higher expected value are more prone to exit through IPO. In contrast, those with a lower expected value tend to search for an acquirer company. Moreover, there is a positive correlation between the investment value and the likelihood of a successful exit, especially an IPO exit. (Guo et al., 2015)

To obtain a clear vision of whether gainful synergy will be created after M&A or IPO, acquirers execute meticulous due diligence analyses (Sirower & O’Byrne, 1998). Research has shown that a thorough due diligence process should include analysis of the financial aspects of business, such as debt capital, balance of equity, and sale of assets but also organizational culture and human capital (*Strategic Management: Competitiveness and Globalization (Concepts and Cases) -ORCA*, n.d.).

1.4 Determinants of FinTech Startup Success

Early-stage technology start-ups encounter fierce competition while operating in a rapidly changing business landscape (Bhave, 1994; Gentry et al., 2013). Most startups fail within two years from their foundation as potential investors perceive them as a source of risk and uncertainty (Crowne, 2002). Given that startups lack tangible assets used as collateral and face information issues, what attracts early-stage investors to invest in these companies (Bernstein et al., 2017; Hall & Lerner, 2010)? According to the competing theories of the company, there are three crucial features of start-ups:

the identity of current lead investors, the founding team and the company's traction (i.e. sales base). Bernstein et al. (2017) provide evidence that the key factor that strongly affects potential investors and, hence, entrepreneurial success is the human capital of the company (Bernstein et al., 2017).

Due to the current trends in the fintech industry, entrepreneurs cannot solely rely on their technology-related knowledge but also market and leadership competencies. The essential value of human capital lies in its inter-disciplinarity (Kopera et al., 2018). Additionally, Saura et al. (2019) use word mining techniques to identify that the key factors for the startup success are the attitude of startup managers, artificial intelligence, and machine learning processes. On the contrary, if they are poorly managed, the relationship with business angels, the programming language used and the quality of job offers may hinder startup success (Saura et al., 2019).

The authors indicate that FinTech startups founded by a single entrepreneur tend to reach a break-even point faster (Carbó-Valverde et al., 2022). In addition, the company has a higher probability to obtain positive profits earlier if it originates from a FinTech accelerator program or incubator. Contrary to the previous studies (Haddad & Hornuf, 2019; Hornuf et al., 2020; Klus et al., 2019), the study proves there is no distinctive benefit of establishing alliances between FinTech and banks as their investors. Unlike (Gazel & Schwienbacher, 2021), Carbo-Valverde et al. (2022) also do not find evidence suggesting companies located near FinTech technological clusters are likely to become profitable (Carbó-Valverde et al., 2022).

Even though FinTech has emerged in both advanced and developing economies, the adoption rates differ significantly (Frost, 2020). Haddad et al. (2018) discovered that the increase in GDP per capita and in the labour market led to an increase in fintech startup formation. Moreover, the large demand for fintech services is discovered in countries where the rights of borrowers and lenders are strongly protected. Finally, the more easily it is for market participants to access loans, the lower is the number of fintech startups in a country. (Haddad & Hornuf, 2019)

Historically, FinTech ventures are more prone to receive larger financing after the global financial crises, especially in regions without major financial center. At the same time, these startups are more likely to result in liquidation. The venture capital is moving toward the direction of the media-hyped industry and not toward the industry developing efficient technologies and innovation. To increase the quality of fintech venture capital deals, policymakers should enforce unique regulatory standards and not solely focus on large financial institutions. (Cumming & Schwienbacher, 2018)

2. Machine Learning

2.1 Machine Learning: Core methods, Trends and Perspectives

Over the past few decades, mobile devices and embedded computing generated vast amounts of data, a phenomenon known as “Big Data” (Jordan & Mitchell, 2015). Big Data comes in different forms, such as structured, semi-structured, unstructured and “metadata” (data about data). (Han J, Pei J, Kamber M., 2011; McCallum, 2005) To obtain valuable insights, forecasts, and decisions in a timely and intelligent way from abundant data, scientists and engineers adopt machine-learning methods. Conceptually, machine learning represents a paradigm that elevates system performance by employing advanced computational techniques to derive insights from experiential data. (Jordan & Mitchell, 2015). Moreover, machine learning algorithms construct a model built from experienced data, making predictions on new observations. It is one of the fastest-growing fields of study, intersecting with computer science and statistics. (Zhou, 2021)

Machine Learning techniques range greatly from ones aiming to solve function approximation problems (e.g. credit-card transaction is given as an input, a “fraud” or “not fraud” label is an output) to ones aiming to find parameters’ values that optimize the performance metric of an implicit function (Rw et al., 2014). Depending on the nature of the data and target output, learning algorithms are divided into four major methods: supervised, unsupervised, semi-supervised and reinforcement learning (Mohammed et al., 2016).

Supervised machine learning algorithms strive to predict an output Y in response to an input X based on sample input-output pairs (Han J, Pei J, Kamber M., 2011). The outputs may take one or two values (simple binary classification problem), one or more of K labels (multiclass or multilabel classification), partial order on a specific set and combinatorial object (general structured prediction). A wide variety of algorithms, such as decision trees, logistic regression, support vector machines, Bayesian classifiers, and kernel machines, reflect different types of mathematical structures and applications, trading off between amount of data, performance, and computational complexity (Jordan & Mitchell, 2015). The most typical supervised tasks are “regression” that predicts continuous dependent variables and “classification” that predicts the label of a given input data (Zhou, 2021).

Contrary to supervised learning, which is defined as a task-driven process, unsupervised learning is a data-driven approach. In unsupervised learning, data does not have a label due to the expensive service of manual labeling or to the innate nature of the data itself. Consequently, unsupervised machine algorithms aim to identify relevant patterns and structures, obtain generative characteristics and group in results without any human guidance. (Mohammed et al., 2016) The most common tasks are finding association rules and clustering (Zhou, 2021).

A clustering problem is identifying the groups with heterogeneous traits among them and homogeneous traits among the observations of each group. On the other hand, the association rule problem discovers patterns that portray large portions of data, such as that people who acquire product X are inclined to buy product Y. (Aggarwal, C. C., 2015; Han J, Pei J, Kamber M., 2011; Kantardzic, 2003; Mitchell, 2006, 2006)

The second paradigm in machine learning research involves formulating assumptions about the structural characteristics of datasets—such as probabilistic or combinatorial properties. It seeks to explicitly identify the underlying manifold from the data. Techniques for dimensionality reduction within this framework encompass factor analysis, manifold learning, random projections, principal component analysis, and autoencoders. These methods render different assumptions about the underlying manifold (e.g. that it is a linear subspace). Furthermore, the defined criterion function includes these assumptions through statistical principles (e.g. maximum likelihood, Bayesian integration, etc.) and allows optimization. Nevertheless, the computational complexity exhibits great concern in both cluster and dimension reduction as these techniques involve processing large amounts of data, and the efficiency of machine learning algorithms becomes vital. (Jordan & Mitchell, 2015)

A hybrid version of supervised and supervised methods, the semi-supervised learning, analyzes both labeled and unlabeled data. (Han J, Pei J, Kamber M., 2011) Many data scientists have found that the amalgamation of labeled and unlabeled data in a model brings significant improvement in learning accuracy. Compared to the acquisition of labeled data that entails physical experiments or human agents, the acquisition of unlabeled data is relatively inexpensive and therefore, the practical value of semi-supervised learning exceeds all incurred costs. Fraud detection, text classification and machine translation among other application fields find solutions in semi-supervised machine learning algorithms. (Alloghani et al., 2020)

Reinforcement learning is an environment-driven approach that aspires to increase the operational efficiency of supply and manufacturing chain logistics, robotics and driving tasks (Zhou, 2021). It helps software agents and machines to automatically estimate the optimal behavior by maximizing the reward and minimizing the risk but without specifying which action would bring the best long-term effect. There are two main sets of techniques for solving reinforcement-learning problems. The first approach represents statistical techniques and programming methods used to evaluate the utility of certain actions in a dynamic environment. The second strives to look for the best behavior in the space of behaviors using genetic algorithms and genetic programming. (Mohammed et al., 2016; Schmidhuber, 1997)

Once we collect, clean, select and transform the data, the need for validation and evaluation rises. The data set is split into training and test data, used to build and validate the test model. The smaller portion of the data is used for model testing, and the test metrics are tested on holdout data. In the case of a small training dataset, the technique of cross validation is implemented. By separating the datasets into subsets (folds), the cross-validation technique trains the model on one portion and validates it on another. This process repeats multiple times, so a more robust evaluation of the model performance is done. The overall model performance is always judged on the holdout folds. (Mohammed et al., 2016)

The trend of data mining applications is present across many fields of business, science, and government. The ongoing Fourth Industrial Revolution brings an abundance of low-cost internet data, allowing machine learning to tailor the products and services based on the current preferences and needs of people. One of the leading pillars of Industry 4.0., the Internet of Things (IoT) improves all aspects of people's lives by predicting traffic in smart cities, parking availability and energy utilization. (Group et al., 2015; Sarker, 2021) Sensors and actuators embedded in physical devices transmit abundant data to computers for analysis. The infrastructure of networked physical objects aids the connection between thing to things, humans, humans and things and thus reduces errors, increases efficiency and incorporates flexible organizational systems. (Group et al., 2015; Sarker, 2021)

The most common machine learning application is found in intelligent decision-making processes driven by data predictive analysis. Mining large crime datasets causes a decrease in crime rate, navigating local police to specific locations and time periods and identifying suspects. Historical traffic records are used to minimize air pollution, accidents and fuel prices. At the same time, abundant medical data sets are seized to solve diagnostic problems, improve patient management, and uncover which patients will respond best to which therapy. (Jordan & Mitchell, 2015) The customers purchase and browsing histories feed machine learning algorithms developing customized shopping experience. E-commerce companies optimize logistics and inventory management while expanding their existing customer base with new ones. (Sarker, 2021).

Classification, feature selection, clustering or sequence labeling machine methods allow automated recognition of patterns, images, and speech in data. Moreover, natural language processing (NLP) aids computers to read and understand spoken or written language. (Otter et al., 2019; Sarker et al., 2021) Opinion AI (Sentiment Analysis), NLP sub-field, strives to excerpt public judgment and mood through social media, news, forums, etc. (Ravi & Ravi, 2015; Sarker, 2021)

Even though machine learning produces imposing advances in various fields of practical and research areas, there are still many unexplored opportunities. The researchers have realized machine learning models are still lagging from the natural learning methodologies in many aspects. Humans gain supervised and unsupervised knowledge from many different data sources and training experiences compared to the machine learning algorithms that can learn one data model or a specific function. (Nicholson & Smyth, 2013; Taylor & Stone, 2009; Thrun & Pratt, 1998) The desired progress in this field is the construction of a lifelong computer that operates and learns thousands of complementary skills and knowledge. On the other hand, people tend to work in teams throughout the whole process of data collection and analysis. Therefore, experts from diverse backgrounds will complement machine learning algorithms to gather, analyze and draw conclusions about complex data sets. (Jordan & Mitchell, 2015).

The efficiency and effectiveness of machine learning solutions depend on both the characteristics of the data and machine learning algorithms (Sarker, 2021). However, the collection of personal data raises ethical questions about who will have access to online data and what the social benefits are. The most common data owners, corporations, have little or no motive to share data thus neglecting potential social prosperity. As a result, collecting all kinds of personal data hinders people’s privacy and eventually brings in doubt long-term consequences to society. To illustrate the challenging trade-off between personal privacy and social benefits, contemplate using online data (location data, credit-card transactions, security cameras, emergency room admissions) to decrease the risk of global pandemic spread. Every person would be alerted to the potential infection if the person he or she was in contact with was admitted to the hospital with infectious disease. In this simple example, society does mitigate the pandemic threat, but public privacy is compromised. Machine learning tends to be the leading transformative technology in the 21st century, so it appears crucial that society adapt laws and culture to take the most out of this phenomenon called Big Data. (Jordan & Mitchell, 2015)

2.2 Chosen Machine Learning Algorithms

2.2.1 Random Forest

Aiming to predict categorical class labels of new instances, classification algorithms create a model based on previous observations (Sadiq 2020). A single decision tree represents a model in the form of a tree that aids individuals to assess future choices, or courses of action and its probability (Zebari 2020a). The decision trees' algorithms strive to deliver the most homogeneous sub-nodes of a tree by splitting the nodes on all available variables (Kumar 2016, Li et al 2019). Nevertheless, utilizing a single classifier raises the inaccuracy of the estimation data set. Thus, the results are precarious (Zebari et al 2019a).

One of the most representative Machine Learning algorithms is Random Forest for its flexibility, diversity, and especially its accuracy when it comes to classifying a huge volume of data. As a matter of fact, Random forest is a Decision Tree-Based classifier that aggregates many decision trees to limit error and elects the best classification tree via voting. (Abdulkareem & Abdulazeez, 2021) The random forest model is based on a bootstrap method that allows Decision trees to obtain multiple subsets of sample and finally integrates several randomly built Decision Trees into a Random Forest (Denisko & Hoffman, 2018). To reduce the possible correlation between decision trees, Random Forest selects different subsamples of the feature space and computes probability distributions of the classes. The probability distributions of different classes are estimated by counting the percentage of different classes of the observations at the leaf node where the concerned observation belongs to. The calculated probability distribution for each decision tree is indeed useful for classification but only assumed to be precise. The precision of the classification result is questioned if the testing data diverges from the training data. Additionally, the smaller the amount of the training data is, the more reduced accuracy is expected. The case of abundant data in which only a few examples fall into the leaf node is another reason for incorrect classification results. (Breiman, 2001; Parmar et al., 2019)

The possible solution to obtain a more accurate prediction for the Random Forest is to assign a weighted average of trees or subsets of trees according to the tree prediction accuracy. The defined weights can be considered as additional training parameters used to get a maximin or robust decision about predicted values. The algorithm's key parameters include: (a) the number of data points sampled per tree, (b) the number of potential splitting directions at each node (c) the total number of trees and (d) the maximum number of examples allowed per node, constrained by the number of sampled data points within the tree node size (Biau & Scornet, 2016) According to the R package randomForest, the default value for the node size parameter is 1 for classification tasks and 5 for regression tasks (Díaz-Uriarte & Alvarez de Andrés, 2006). On the other hand, an increase in M value causes a decrease in forest's variance and computational complexity. In this respect, conflicting consequences of tuning the M parameter pose a challenge to determine a priori minimum number of trees. The parameters prove to affect the performance of the model, but the theory does not confirm that exact default values for these parameters exist. (Biau & Scornet, 2016).

Random Forests are an easy-to-use, quick, and effective machine learning algorithm that can deal with missing data details without losing accuracy. A random forest does not require any cross verification, and it is not prone to over-fitting. Additionally, it provides techniques to evaluate relevant variables, variable relationships, incomplete data sets, and metadata. (Bhattacharyya et al 2019) On the other hand, Random Forests seldom do effectively deal with multi-valued and multi-dimensional attributes as they prefer multi-level categorical variables. Another limitation of the model is manifested in the regression tasks as over-fitting particular data sample. (Shaik & Srinivasan, 2019)

2.2.2 Support Vector Machine

Alongside supervised learning algorithms such as decision trees, deep learning networks and naïve Bayes, Support Vector Machine (SVM) learns by example to assign class labels to new observations. During the training phase, existing input-output pairs feed SVM algorithm allowing it to identify relevant data classification patterns with balanced reliability and accuracy. In essence, the SVM decision function develops a 'hyperplane' that aims to allocate observations in specific class labels based on existing patterns in those observations, thus defining the most probable label for unseen, new data. The patterns of information about the observations defined as features are most often data derived from interpolation in the feature selection phase. Furthermore, the support vectors represent a result of features referenced by coordinates and their relationship to each other. SVM function ultimately strives to optimize its accuracy while ensuring the classifier is applied to new data. These two complementary goals are bound by the informativeness of the used features and several examples used to train the model. (Pisner & Schnyer, 2020).

The Support Vector Machine (SVM) analysis involves three pivotal stages: (1) feature selection, (2) model training and evaluation, and (3) performance assessment. It is crucial to recognize that these stages are not exclusive to SVM but are prevalent in most machine learning methodologies. (Pisner & Schnyer, 2020)

The core input for the SVM model comprises a set of features derived from transforming the original raw training data. Feature selection techniques utilize specific criteria to prioritize features according to their hierarchical significance. The principal methods of feature selection include embedded, filter, and wrapper approaches. (Pisner & Schnyer, 2020)

Embedded methods perform feature selection as part of the SVM training process, utilizing kernel techniques. These techniques generate a kernel matrix, capturing pairwise similarities between observation patterns in an N x N matrix (where N denotes the number of observations). By mapping raw data into a higher-dimensional feature space, the kernel matrix helps prevent overfitting, which is particularly useful when the number of features exceeds the number of training samples available. (Pisner & Schnyer, 2020)

As features with near-zero variance and highly correlated features do not boost predictive power but only add complexity to the SVM model, filter methods commence preparation for training a classifier with the feature reduction. The model training process thus reduces redundancy in the raw data, which consists of a greater proportion of sample training data relative to the dimensionality of the features. Furthermore, the feature reduction decreases the computational load and aids in distinguishing what data have the most predictive power to the final classifier. (Pisner & Schnyer, 2020)

Wrapper methods succeed in discarding the data points that have the least relevancy in discriminating between class labels, relying on the feedback from every training iteration. One of the most common types of wrapper method is recursive feature elimination (RFE), which arrays features into smaller and smaller subsets of features through cross-validation. Cross-validation is a multipermutation method that iteratively splits the original training data into new training and testing data and re-estimates model performance during each iteration. (Pisner & Schnyer, 2020)

The training of the SVM classifier is based on the example observations whose labels are known in advance. In the linear decision function f(x) = a + k*x training an SVM strives to define the parameters a and k so the hyperplane maximally distinct the members of two classes. Additionally, the choice of the hyperparameter values (e.g. soft margin, number of k-best features) strongly affects the accuracy of the classifier. Moreover, the absolute value of weight mirrors the importance of features in differentiating the two classes. (Ben-Hur & Weston, 2010; Pisner & Schnyer, 2020)

Finally, the performance of SVM is evaluated on the unseen data though a prism of accuracy and reproducibility. In addition to cross-validation, data is divided into train and test data, so the latter data type is kept for the final evaluation of the model performance. Permutation testing iteratively evaluates hyperplanes with randomly permuted class labels, preventing biases to one class over another. (Pisner & Schnyer, 2020)

2.2.3 Extreme Gradient Boosting (XGBoost)

Among various machine learning and data-driven approaches, gradient tree boosting delivers state-of-the-art results on many real-world predictions. The wide range of challenges vary from store sales prediction over web text classification to motion detection. From a statistics point of view, the concept of boosting is based on modifying a “poor hypothesis” into a very “accurate hypothesis”. In that way, the model aims to transform a “weak learner” into a “better learner” by combining rough and insufficiently accurate rules-of-thumb. (Chen & Guestrin, 2016; Chen & He, n.d.; Ramraj et al., 2016)

Extreme Gradient Boosting (XGBoost) is a cutting-edge algorithm that has refined its core concept of gradient boosting. In a nutshell, gradient boosting incorporates differentiable loss functions into the framework, so the novel algorithms are not required to be derived for every loss function. Moreover, a “weak learner” represents a decision tree in gradient boosting. In the case of the regression trees, trees produce their real value outputs for splits. Values created in a preexisting sequence of trees are iteratively added up to the output produced in new trees, improving the final output of the model. The additive model of the trees’ prediction thus ensures optimization of the loss function. (Ramraj et al., 2016)

However, XGBoost surpasses the simple gradient boosting algorithm in its additive model. As a matter of fact, the procedure of addition of decision trees is not performed one after another but much faster and with improved performance. Its addition model represents a multi-layered approach in which adequate utilization of the CPU core of the machine is used. Furthermore, the XGBoost algorithm automatically deals with missing and sparse data, continuously trains already fitted models on a new data and enables quicker machine learning. (T. Chen & Guestrin, 2016; Ramraj et al., 2016)

The scalable tree boosting system, XGBoost, develops various objective functions such as regression, classification, and ranking. The XGBoost package can automatically perform parallel computation on Linux and Windows, deal with various data types of input data (dense matrix, sparse matrix, local data files and xgboost own class), support customized objective functions and exert better performance on several different datasets. (T. Chen & Guestrin, 2016)

2.3 Previous Machine Learning Implementation in Predicting Startup Success

Early-phase startups operate in a chaotic, grinding, evolving ecosystem encountering intense competition (Bhave, 1994; Gentry et al., 2013). The technology sector continuously grows and offers novel solutions, posing serious threats for all market participants (Lanza & Passarelli, 2014; Rose, n.d.). Many failures of technology startups derive from entrepreneurs' inability to deal with uncertainties (Butler et al., 2010). The uncertainties increase not only the probability of failure but also barriers to entry. Therefore, comprehending business threats and opportunities is crucial for both entrepreneurs and investors. The data analytic approach applied to startup evaluation creates machine learning models that provide valuable insights about the probability of success and failure.

Tomy and Pardede (2023) develop a model, which employs internal capabilities and resources and external factors as variables, aiming to predict technology startups' success. The study shows that the Naïve Bayes classifier achieves higher accuracy and lower error rate than k-Nearest Number and Support Vector Machine algorithms. Nevertheless, it would be beneficial to overcome the limitations of the model and enrich it with data from different sectors and geographical locations. (Tomy & Pardede, 2023)

Machines have not yet surpassed human intelligence. The challenge lies in their inability to effectively interpret "soft" information, data that is difficult to quantify. Additionally, scenarios characterized by extreme uncertainty demand intuitive decision-making, a capability that remains inherently human and not yet replicable by machines. (Attenberg et al., 2015; Baer & McKool, 2014; Dellermann et al., 2021; Liberti & Petersen, 2019). Hence, the authors propose a Hybrid Intelligence method that combines machine and collective intelligence to predict success of tech startups under extreme uncertainty (Dellermann et al., 2021; Gregor & Hevner, 2013; Hevner, 2007). The approach offers a formal analysis of “hard” information using different machine learning algorithms such as Logistic Regression, Naïve Bayes, Support Vector Machine (SVM), Artificial Neural Network (ANN) and Random Forest. It is further developed by calculating the unweighted average of the ordinal evaluation non-expert and expert participants provide into a single score. A single score represents the probabilistic classification of series A funding success, the definition of ultimate startup success. (Keuschnigg & Ganser, 2017) Finally, two-way analysis of variance

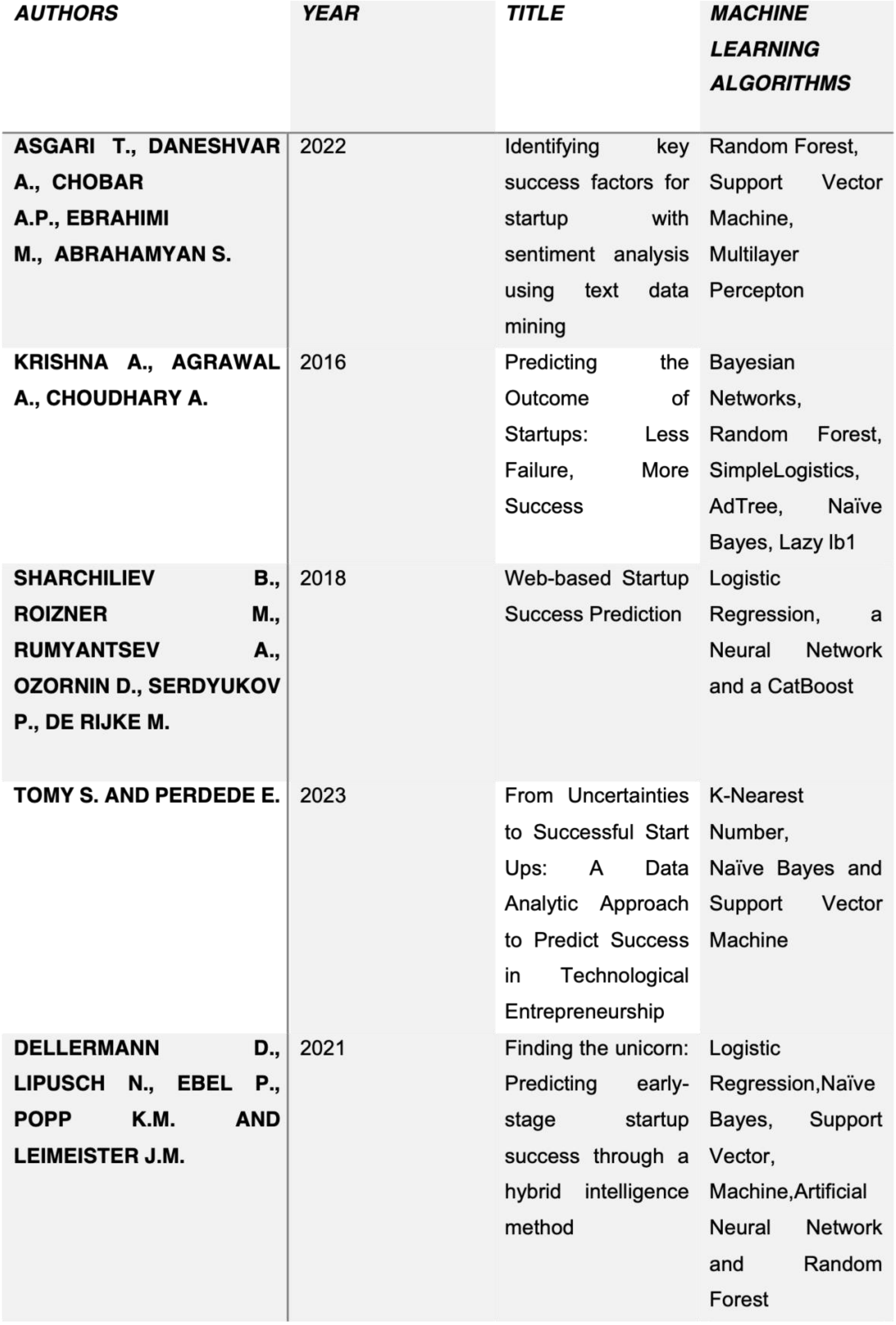
(ANOVA) aids us in identifying the most efficient machine learning algorithm which is complemented by collective intelligence (Dellermann et al., 2021).

Sharchilev et al. (2018) expand the previous work on the problem of predicting the success of startup companies including open web-based sources into the database. The authors define the startup success as its potential to attract further or larger funding rounds, indicating high current or future intrinsic value (Davila et al., 2003). Finally, they train the diversified machine learning pipeline, obtaining information from both structured data about the startup ecosystem and openly available web data. The methods rely on gradient boosting algorithms (Logistic Regression, a Neural Network and a CatBoost) that outperform neural network models used by previous authors. (Davila et al., 2003)

The work done in paper (Krishna et al., 2016) aims to increase the success rate of early-stage companies developing the prediction model based on key features. The traits that have the most profound impact on the company's growth are foundation date, volume of seed funding, series A and B funding, time between funding rounds, so as low burn rate, management system, etc. The authors classify startups into 9 models based on more than 20 factors. As the classification group gradually moves from Model 1 to Model 9, the additional funding round adds to the key factors of the next model. Finally, the last classification group, Model 9, includes all the key factors. The classification model trains and tests the dataset imported from CrunchBase, allowing authors to compare results from six different machine learning algorithms (Bayesian Networks, Random Forest, SimpleLogistics, AdTree, Naïve Bayes, Lazy lb1). The authors present the results for each model in the form of Area Under Curve (AUC), Recall and Precision Values. AUC shows that the best classification schemes are SimpleLogistics and Random Forest, while Recall Vaues range (from Model 0 to Model 9) from 73.3% to 96.3%. (Krishna et al., 2016)

The research of Cojoianu et al. (2021) can be described as exploratory- survey research, as text mining and machine learning techniques yield valuable insights from comments with #Startup. The Twitter data is unstructured, but highly useful for identifying gap between the company and the market. Thus, sentiment analysis aims to divide sentiments into three groups: positive, negative, and neutral, and predict the consequences of the opinions expressed on Twitter. The prediction models, random forest, support vector machine and multilayer perceptron, test the classification of unstructured data and provide accuracy of 0.78, 0.8 and 0.81 respectively. The authors conclude that the most positive factor affecting startup success is product marketing management, while the most negative factor is climate change. Business aspects of a company has showed to be neutral for the company's success. However, it is advisable to enrich the database with more than one hashtag (startup #) such as cohesion of the startup team and investing in a startup. In that way, the quality of data collection would increase. Moreover, deep learning methods can be added to the classification model collection to create a more powerful model. Table 2 represents a summary of previous studies, including the names of the authors, publication years, articles, and the machine learning algorithms they implemented. It offers comprehensive overview of the evolution of machine learning approaches in the context of FinTech startups. (Cojoianu et al., 2021)

Table 2- Previous Machine Learning Implementation in Predicting Startup Success



AUTHORS	***YEAR***	***TITLE***	***MACHINE LEARNING ALGORITHMS***
ASGARI T., DANESHVAR A., CHOBAR A.P., EBRAHIMI M., ABRAHAMYAN S.	2022	Identifying key success factors for startup with sentiment analysis using text data mining	Random Forest, Support Vector Machine, Multilayer Percepton
KRISHNA A., AGRAWAL A., CHOUDHARY A.	2016	Predicting the Outcome of Startups: Less Failure, More Success	Bayesian Networks, Random Forest, SimpleLogistics, AdTree, Naïve Bayes, Lazy lb1
SHARCHILIEV B., ROIZNER M., RUMYANTSEV A., OZORNIN D., SERDYUKOV P., DE RIJKE M.	2018	Web-based Startup Success Prediction	Logistic Regression, a Neural Network and a CatBoost
TOMY S. AND PERDEDE E.	2023	From Uncertainties to Successful Start Ups: A Data Analytic Approach to Predict Success in Technological Entrepreneurship	K-Nearest Number, Naïve Bayes and Support Vector Machine
DELLERMANN D., LIPUSCH N., EBEL P., POPP K.M. AND LEIMEISTER J.M.	2021	Finding the unicorn: Predicting early-stage startup success through a hybrid intelligence method	Logistic Regression,Naïve Bayes, Support Vector, Machine,Artificial Neural Network and Random Forest

3. Methodology

The research is based on the Knowledge Discovery in Databases (KDD) approach or data mining technique. It strives to extract knowledge from relational databases, for existing database volume overpowers human capabilities to analyze such data. (Han et al., 1992) The applied machine learning models for prediction consist of the following experiment setup steps:

(1) **Data Collection**: Import the relevant dataset from CrunchBase for the research question.

(2) **Data Preprocessing**:

(a) Clean any inconsistencies found in data, i.e. missing values and duplicates.

(b) Filter data based on criteria like category group or year.

(c) Create new variables and transform categorical variables into dummy variables.

(3) Model Preparation: Separate the data into train and test sets and standardize features.

(4) **Model Building:** Implement machine learning algorithms and use pipelines to streamline the preprocessing and modeling steps.

(5) **Model Evaluation:** Use GridSearchCV for hyperparameter tuning to find the best model

(6) **Model Comparison**: Compare the performance of different models (XGBoost, RandomForest, SVM).

3.1 Data Collection

The research was conducted using the database from the website CrunchBase (www.crunchbase.com). Crunchbase is a platform that provides insights about public and private startups and corporations, their employees, leaders or founders, investors, funding stages. (*Crunchbase: Discover Innovative Companies and the People behind Them*, n.d.; Fischer, 2017)

The data was approved for the author’s academic research on December 18, 2023. Table 3 show the descriptive overview of the obtained dataset consisting of 17 tables in CSV (comma-separated-values) files (*Legacy CSV Export*, n.d.):

Table 3 – CrunchBase Dataset

NAME	***DESCRIPTION***
ORGANIZATIONS	Organization profiles
ORGANIZATION_DESCRIPTIONS	Detailed organization description
ACQUISITIONS	List of all acquisitions available on Crunshbase
ORG_PARENTS	Mapping between parent organization and subsidiers
IPOS	Detail for each IPO
CATEGORY_GROUPS	Mapping between category groups and organization categories
PEOPLE	People profiles
PEOPLE_DESCRIPTIONS	Detailed description about people profiles
DEGREES	Detail for people’s education background
JOBS	List of all jobs
INVESTORS	List of active investors (organization or/and people)
INVESTMENTS	List of all investments
INVESTMENT_PARTNERS	Partners responsible for their companies’ investments
FUNDS	Investors’ investments funds
FUNDING_ROUNDS	Details about investors’ investments round

EVENTS	Event details
EVENT_APPEARANCES	Details about event participation

Source: (*Legacy CSV Export*, n.d.)

The *organizations* table holds financial information about companies such as total funding, number of exits, number of funding rounds, and dates of first and last funding. Additionally, the table provides basic information about the organization including name, the role of the company (investor or/and company), address, social media links, website, email, phone, and number of employees. Each company is further described by the industry it operates within by both category list and category group list features. (*Legacy CSV Export*, n.d.)

The *acquisitions* and *ipos* tables include information about the dates of such events, details about participants in the acquisition events, and money raised. The *funding rounds* table contains information about the investment type (seed, angel, round A,B, etc.), the number of investors, the raised money, and the universal unique identifier of both investors and companies. (*Legacy CSV Export*, n.d.)

The *jobs* table describes individuals who are investors, founders, employees and their positions and types of jobs within the organization. It also gives insights into the date the individual commenced and finished their employment within the company. (*Legacy CSV Export*, n.d.)

The *organization* table serves as the central or core table within our database schema, acting as the primary source of foundational data about entities (companies). The other three tables (*acquisition, ipos, funding rounds* and *jobs*) are designed as supporting tables, each containing additional details that enhance, elaborate, or expand upon the information found in the organization table. These supporting tables will be integrated into the core organization table through specific common identifiers, allowing for a comprehensive and multidimensional view of the data. (*Legacy CSV Export*, n.d.)

Other tables that were not used in the research are out of scope and do not contain relevant information. This setup provides a centralized approach to data management, where the organization table anchors the dataset, while the supporting tables contribute additional layers of detail and context. The simplified entity-relationship diagram (ERD) of the Crunchbase table is shown in Fig. 1. (Żbikowski & Antosiuk, 2021)

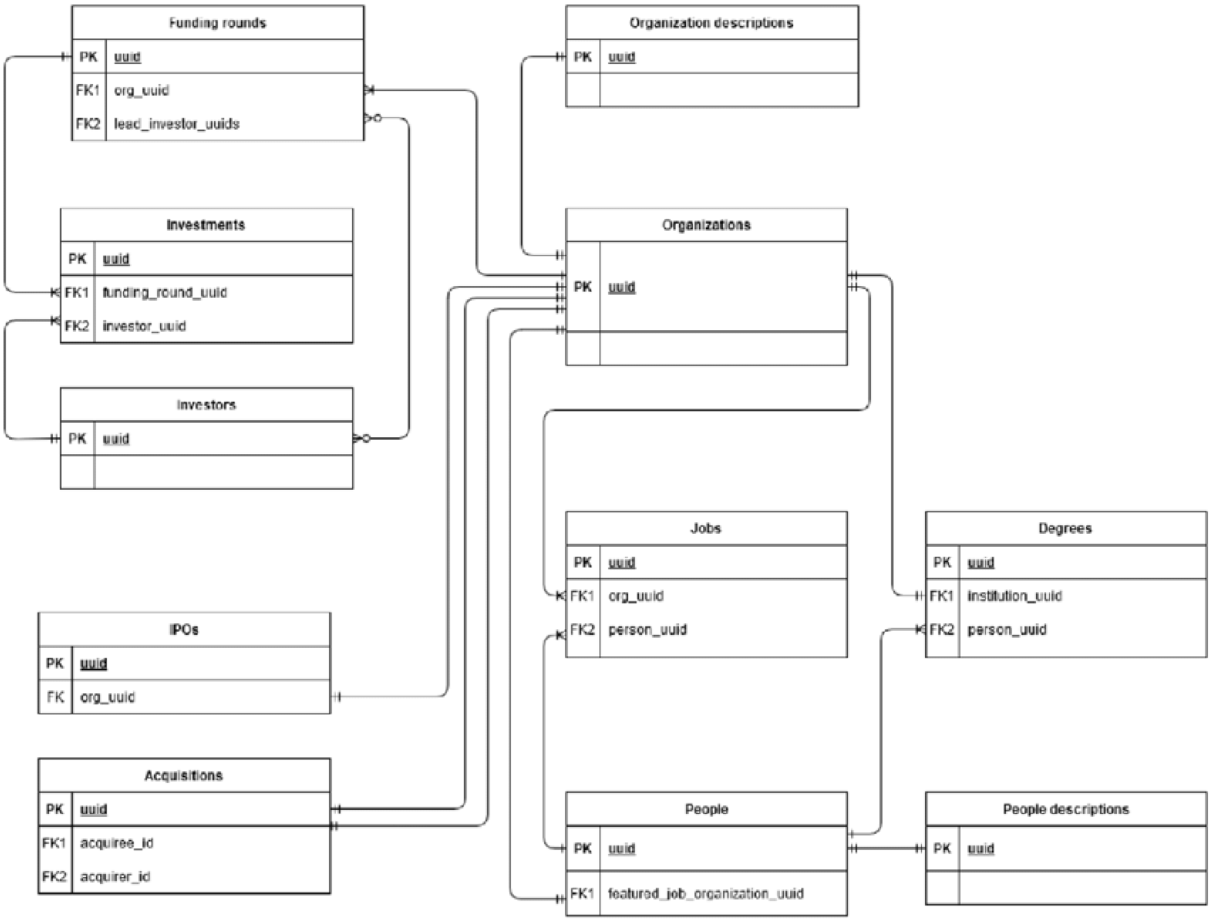


Figure 1. – Simplified ERD diagram of CrunchBase dataset

Source: (Żbikowski & Antosiuk, 2021)

3.2 Data Preprocessing

Data preprocessing is a crucial step in the data analysis and machine learning pipeline. The process involves preparing and cleaning raw data to make it more suitable for analysis and modeling. Finally, improving the data quality leads to the success of the supervised machine learning model.

It is a 3-step process:

(1) **Data Cleaning**: This step involves handling missing data, noise from data, and removing outliers. The technique included imputation (filling missing values with mean) and filtering out outliers.

(2) **Data Selection**: As the context of the research is the success prediction of FinTech startups, a dataset is filtered to include only companies operating within the FinTech industry.

(3) **Data Transformation**: The process consists of converting data into a suitable format for analysis and creating new variables from existing ones to capture the underlying pattern in the data better.

3.2.1 Data Cleaning

Data cleaning is a fundamental step of data preprocessing, ensuring the quality and reliability of data before it undergoes analysis or is fed into machine learning algorithms. It is an iterative and subjective process that requires a deep understanding of data and the context in which it was collected.

The data cleaning process commenced with a strategic step to streamline the dataset by removing columns that were not aligned with the primary objective of predicting the success of fintech startups. This initial phase of data cleaning is crucial as it focuses the analysis on relevant variables, enhancing the efficiency and accuracy of subsequent predictive modeling.

The rationale behind deleting specific columns is rooted in their lack of direct relevance to the factors influencing a fintech startup's success. By eliminating these variables at the outset, we ensure the dataset is optimized for uncovering insights related to the success metrics of fintech companies.

The columns removed from the *organizations* table are as follows:

- **Basic information**: *name, short description, postal_code, primary_role, type, rank, created_at.*
- **Contact information and media links**: *cb_url, legal_name, address, email, phone, facebook_url, linkedin_url, twitter_url, logo_url, permalink, domain*.

- **Additional financial information**: *total_funding, total_funding_currency_code*. For the purpose of the model, only the column (total_funding_usd) total funding denominated in USD currency is used.
- **Dates of dataset creation and update**: *created_at, updated_at.*
- **Other***: allias1, allias2, allias3.*

The next fundamental step in the data cleaning process involved tackling the issue of the missing (NaN) values. Missing data can significantly impair the quality of predictive modeling leading to weak predictive performance. Therefore, addressing these missing values is paramount to ensure the integrity of the dataset for our analysis.

The *investor_count* and *num_funding_rounds* columns are indicative of the startup's viability and success for they address the level of interest and confidence investors have in fintech startups. Given the nature of these features, missing values are interpreted as the absence of investors or funding rounds, rather than a lack of data. Consequently, NaN values are replaced with zeros in these columns. The rationale behind this approach is that missing value in the column presenting the number of investors can be reasonably interpreted as the startup not having attracted any investors yet. Replacing NaN with zero indicates startups are in the early stage or not appealing to investors. On the other hand, replacing NaN with zero in the column *num_funding_rounds* suggests that startups did not initiate or complete any funding rounds. With this careful treatment of missing data, we retain valuable information about startup's investor appeal and funding history.

Finalizing the data cleaning process involves eliminating outliers. Outliers substantially affect the process of estimating statistics, leading to potentially misleading results (Kwak & Kim, 2017). The *total_funding_usd* column, which represents the total amount of funding a startup has received in USD, is essential for evaluating investors' perceptions about a startup's potential for success.

Given the importance of this financial feature, identifying, and removing outliers from the *total_funding_usd* column was crucial to ensure the accuracy and reliability of the subsequent analysis and model development. The interquartile range (IQR) method was used, so that only observations that fall into the middle 50% of the data are kept. Observations that fall below the first quartile minus 1.5 times the IQR or above third quartile plus 1.5 times the IQR are defined as outliers.

3.2.2 Data Selection

The process of data selection was meticulously carried out by applying two key filters to the startup data frame, ensuring the analysis would align with the specific objectives of the research, and predicting the success of FinTech startups.

The first filter applied to the dataset was based on the *category_list* column, with the criteria set to include only startups that operate in the FinTech industry. By focusing exclusively on the startups that operate within the financial technology sector, the study aims to discover patterns specific to this innovative sector. Diving into the subcategories of the FinTech category, we were able to uncover a diversified palette of goods and services that fintech startups offer.

The second filter aimed to exclude startups established before 1980, effectively removing the early pioneers from the dataset. This temporal boundary was deliberately chosen to concentrate the analysis on more contemporary entities, reflecting the context of market conditions, technological advancements (starting with the development of Artificial Intelligence, telephone, and online banking), market conditions and consumer behaviors that is more representative of the current FinTech ecosystem. (*Fintech*, 2023)

These strategic data selection criteria effectively narrowed the dataset to include only those entities that are most relevant to the research's objective. The category and foundation period criteria ensure that each step in subsequent analysis, from feature engineering to machine learning model, is precisely targeted towards understanding the factors that contribute to the success of contemporary FinTech startups. Consequently, the targeted approach enables the creation of predictive models that accurately reflect and predict the pathway of modern FinTech industry participants.

3.2.3 Data Transformation

The process of data transformation is pivotal in preparing the dataset for predictive modeling. It ensures that the data is in the appropriate format for analysis and aligns with the specific requirements of the chosen machine learning algorithms. The transformation of the data in the research was methodically executed in two successive phases, aiming to refine the dataset to enhance its analytical value and predictive potential. The phases of the data transformation process include changes in the original data and the creation of new variables.

3.2.3.1 Changes in Original Data

The first step in the data transformation process included converting categorical features into a format that could be effectively utilized by the predictive models. The *employee_count* feature was initially categorized with labels indicating the range of employee numbers (e.g. '1-10', '11-50', '51-100', etc.). To allow the machine learning algorithms to understand and leverage the inherent order within the features, this feature was transformed into ordinal values. The label '1-10' became value 1, '11-50' transformed to value 2 and so on. Such conversion allows algorithms to recognize that startups with '11-50' are larger than those with '1-10', hence increasing the impact of the company size on the success of FinTech startups.

The second transformation addressed the format of year-related columns within the dataset (columns *founded_at, acquired_on, went_public_on, started_on, finished_on*). These columns, initially formatted as objects, were standardized into float values. This uniformity is crucial for performing the calculation of period, the time elapsed between key events (e.g., foundation and IPO or acquisition). By ensuring these year-related columns are in a consistent and calculable format, the dataset enables more sophisticated temporal analyses, providing insight of how the age of a startup or the timing of its funding rounds may influence its chances of success.

These transformations significantly enhance the dataset's utility for predictive modeling. The approach ensures that machine learning algorithms can effectively interpret and analyze the features, thereby improving the accuracy and relevance of the predictions regarding the success of FinTech startups.

3.2.3.2 Creating of New Variables

In the second phase of the data transformation process, the focus shifted to augmenting the dataset with new variables that capture additional aspects of information relevant to predicting the success of FinTech startups. The new variables were meticulously engineered to provide deeper insights into various aspects of startup financial and operative performance and market conditions. Table 4 shows thorough representation of these newly created variables, illustrating their comprehensive inclusion and the breadth of information they offer.

Table 4 – New Variables in the FinTech startups dataset

NEW VARIABLE NAME	***DEFINITION***	***SIGNIFICANCE***
WORK_EXPERIENCE	Calculated as difference between the ended_on and started_on features (people.csv) for individuals associated with the startup.	Provides insights about collective work experience and know-how.
INVESTMENT_TYPE	Type of investment received by the startup, with unique values angel, series A, series B, etc.	Reflects investor confidence in the startup success potential.
INVESTOR_COUNT	Represents the number of investors, funding the startup.	Offers insights about the level of investor support and interest for startup business model.
AVG_AMOUNT_RAISED	Calculated as the total funding amount divided by the number of funding rounds (*total_funding_usd / num_funding_rounds*).	Reflects the average amount of capital raised per funding round, serving as a proxy for the startup's fundraising efficiency, investor confidence and financial health.
ACQUISITION_YEAR	Calculated as the difference between the acquisition year date (*acquired_on*) and the founding year (*founded_on*).	Indicates the startup's strategic value within the industry and attractiveness to potential acquirers.
IPO_YEAR	Calculated as the difference between the year startup went public (*went_public_on*) and the year it was founded (*founded_on*).	Offers insights about startup maturity, potential for sustained growth and valuation.

SUCCESS_AGE	Converges the IPO year and acquisition year to determine the startup's success age, duration between foundation, and IPO/acquisition event.	Captures duration and nature of startup journey to achieve success (IPO/ acquisition).

3.3 Dataset Breakdown

We delve into the distribution of FinTech startups across different countries, regions, and cities, shedding light on the geographical landscape of the FinTech industry. Understanding the spatial distribution of FinTech startups is pivotal for identifying trends and underlying factors that influence the formation of these innovative ventures.

According to the data shown in Figure 2, FinTech formation is the greatest in the United States of America, Great Britain, Canada, Brazil, Germany, and Finland respectively (Figure 2). The literature meticulously explored myriad factors that influence FinTech distribution, proving the consistency of the stated results. Countries that have been more financially developed or experienced crises during the last decade tend to have a larger FinTech ecosystem. Additionally, Figure 3 proves that the greater tertiary education enrollment, fixed-line availability and cooperation between industry and university, the greater FinTech formation is. Ultimately, the intensity of FinTech formation is more pronounced in nations with well-established ICT services clusters—evaluated by the proportion of ICT services exports relative to total exports—elevated per capita income, and advanced quality of internet and mobile communication infrastructure. (Laidroo & Avarmaa, 2020; Wójcik, 2021)

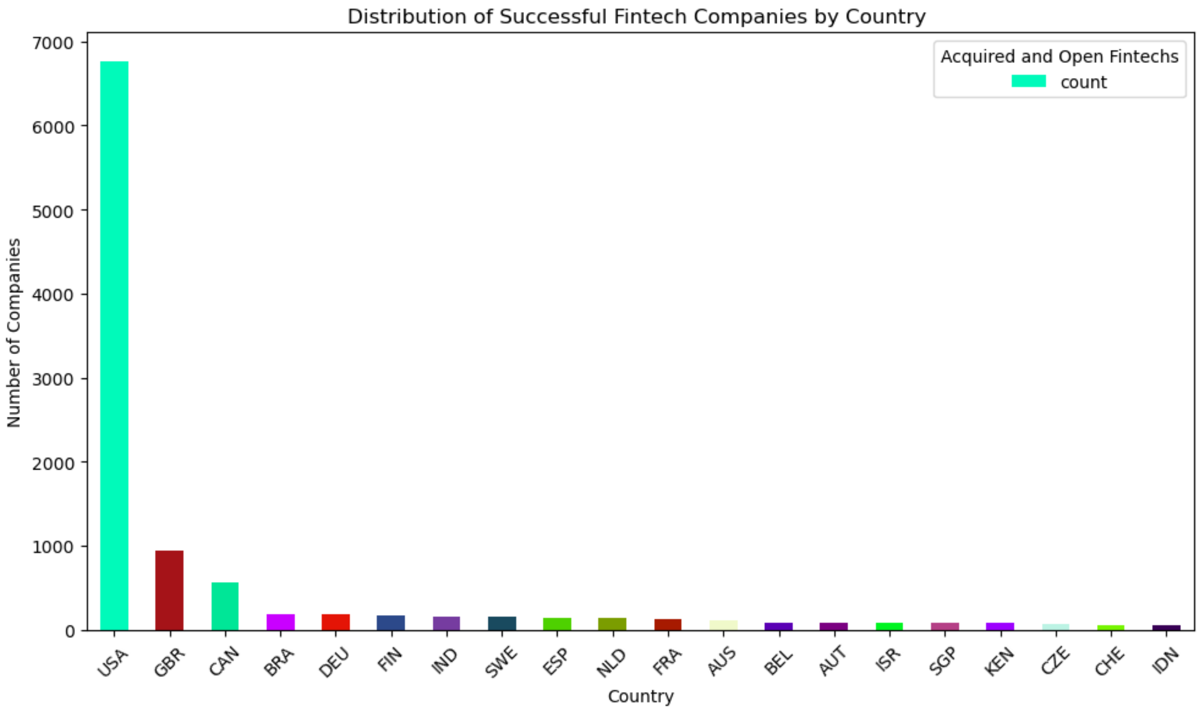


Figure 2 – Distribution of FinTech startups based on country criterion

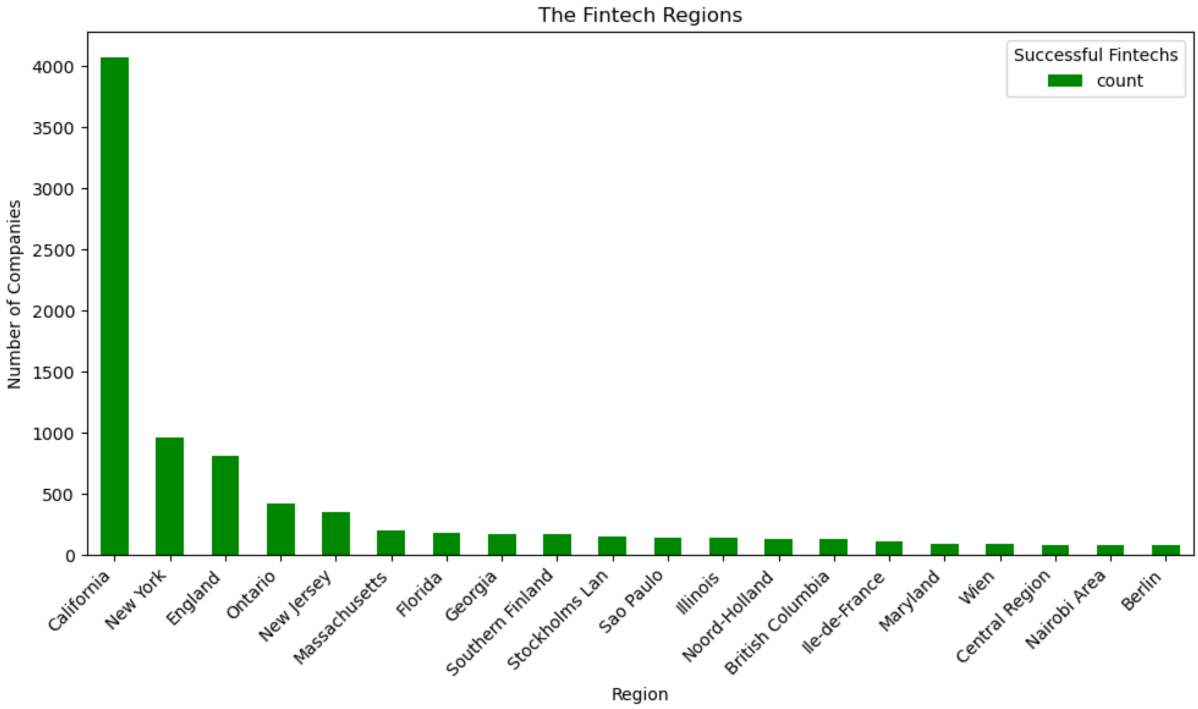


Figure 3 – Distribution of FinTech startups based on region criterion

The proliferation of FinTech startups in urban centers worldwide marks a crucial trend in the revolution of financial services. The dense networks of tech companies, academic hubs, and financial institutions contribute to the grinding FinTech innovations. Figure 4 represents visual distribution of FinTech startups across key cities, providing a snapshot of the global urban landscape of FinTech innovation. The top five successful FinTech centers are: San Jose, San Francisco, New York, London and Redwood City. In this ranking report, there were no cities in China. According to *Global FinTech Hub Report* (2018) Claessens et al. (2018), cities like New York, London, San Francisco have developed their FinTech sector on demand from institutional investors, the tech sector and regulatory innovation, while the dominant source of funding in China are individuals. Furthermore, the uneven access to the internet represents additional constraint for the FinTech formation in China. (Hasan et al. 2020).

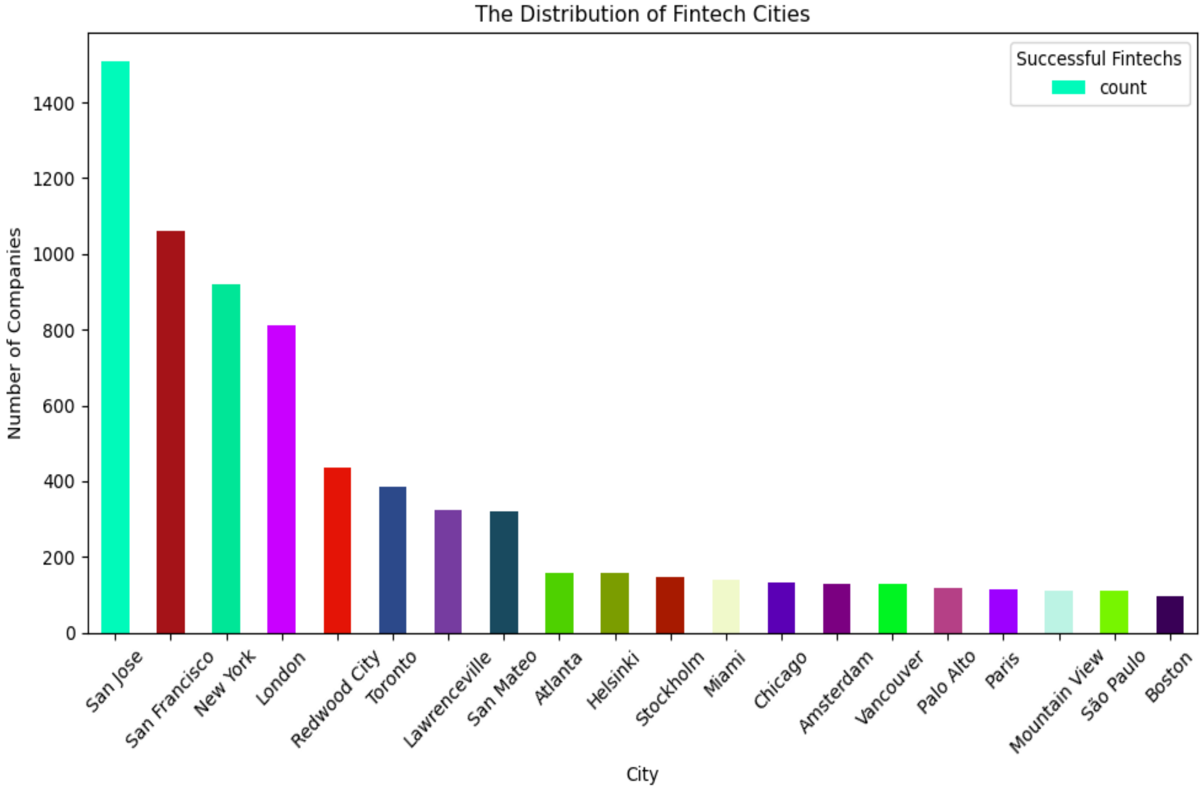


Figure 4 – Distribution of FinTech startups based on city criterion

Figure 5 illustrates the most attractive FinTech categories based on average founding amount per round. High funding levels in specific categories demonstrate strong beliefs of investors in the market potential. On the contrary, categories with lower funding amounts may represent niches with higher barriers to widespread adoption or unrecognized potential. Sectors such as InsurTech, Artificial Intelligence in FinTech and Wealth Management Tech showcase how implementation of technological advancements aims to tailor financial services to the individual's needs and facilitate access to wealth management and insurance services. Additionally, categories such as Payment, Cryptocurrency, Banking prove that investors comprehend the transformative potential of these services in facilitating efficient, secure and accessible financial transactions worldwide. By comparing average funding amounts, we can identify which sectors are perceived as having the highest growth potential in the FinTech landscape. Startups whose core businesses are risk and wealth management, e-commerce, cryptocurrency and banking services proved to be the pivotal forces that will navigate the future of financial services.

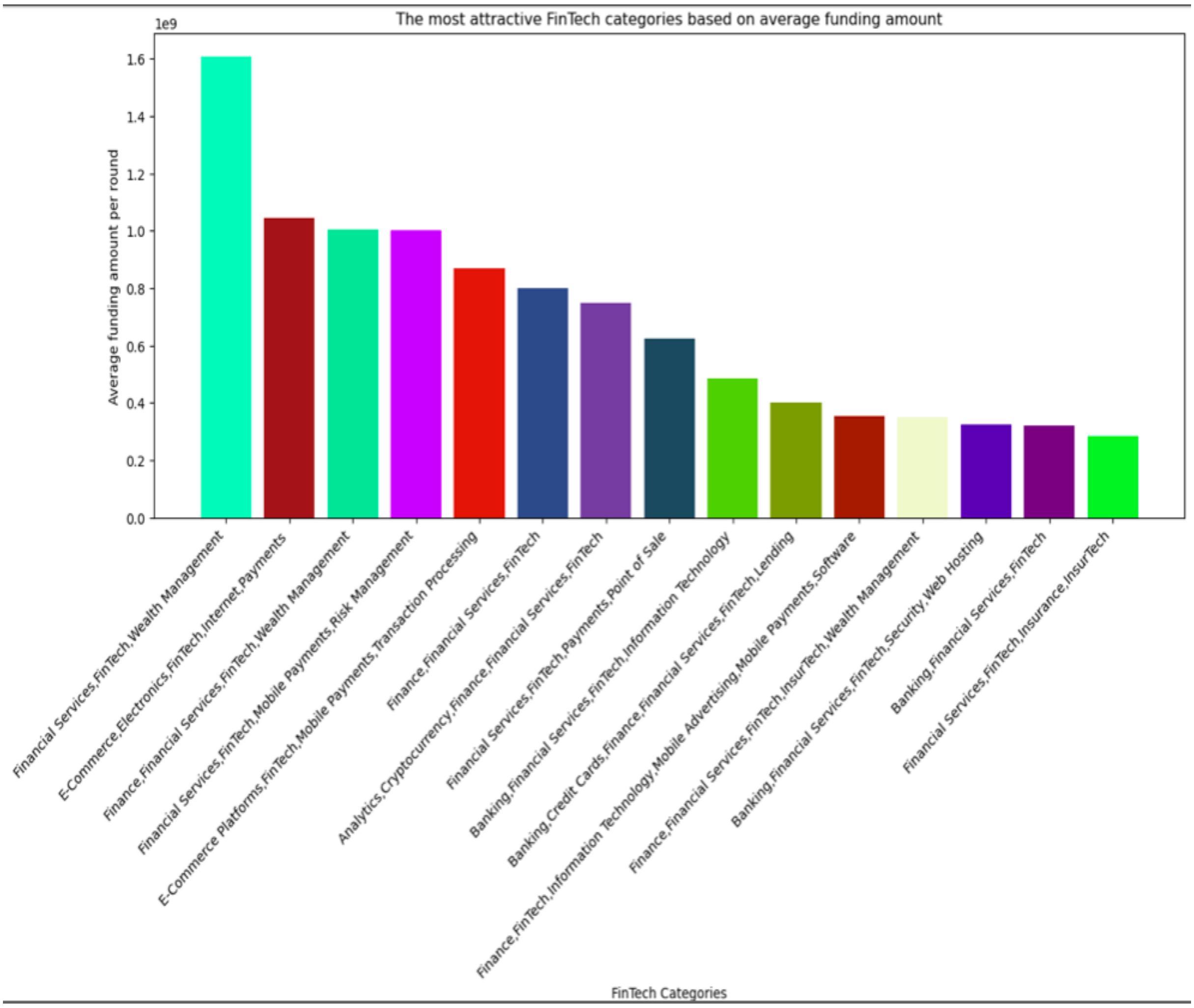


Figure 5 – The most attractive FinTech categories based on avarage funding amount per round criterion

Understanding the duration from a startup’s founding to a significant liquidity event, such as an IPO (Initial Public Offering) or acquisition, provides valuable insights into the evolution of the industry across different time periods and, consequently, different market conditions. Strategic decisions, technological innovation, and regulatory and market forces determine the dynamics of the industry. Finally, Figure 6 describes the average success age among early pioneers, FinTech companies founded between 2008 and 2018 and those founded after 2018.

The division of the dataset into categories based on evolutionary context, specifically into FinTech 2.0 and FinTech 3.0, reflects an analysis of the FinTech industry's development stages and their distinct characteristics. Timeframe FinTech 2.0 refers to the period between 1967 and the global financial crisis of 2008. By the beginning of the 21st century, banks 'internal processes and interactions with stakeholders were fully digitized. On the other hand, regulators perceived that e-banking solutions provide more credit risks, as competition among lenders rises. As geographical limits were removed, the borrowers had a greater pool of lenders. The transition from FinTech 2.0 to Fintech 3.0 means a shift from the digitalization of traditional financial services to innovative financial paradigms, resulting in faster success achievement.

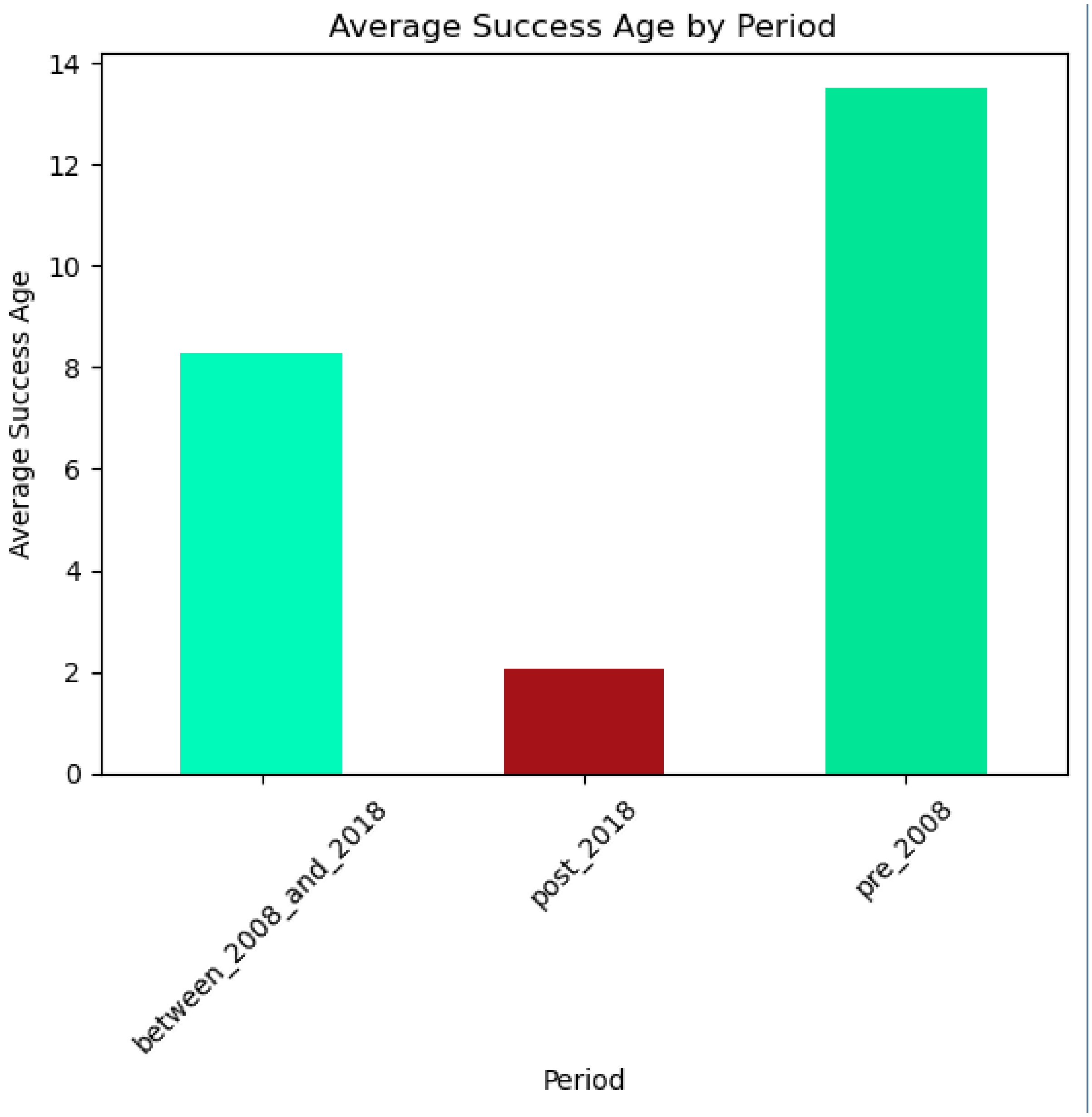


Figure 6 – Average Success Age by Period

The most notable impact of the economic crises was the twisted brand image of traditional banks. According to the 2015 survey, American public trust levels in technology firms providing financial services exceeded their confidence in banks. The 2008 economic crisis proved that providing banking services is necessary but not from banks. (Arner et al., 2015)

The FinTech 3.0, the current and ongoing phase, is characterized by technological advancements such as Artificial intelligence, blockchain and Big Data analytics in financial services. The post-crise regulatory obligations turned in favour of technology companies as increased capital requirements led to a decreased number of competing banks. We divided startups founded in the FinTech 3.0 era into those established before and after 2018 to highlight the period of surge in FinTech innovation and the maturation phase. We realize that startups founded during FinTech 3.0 experienced a shorter success age compared to earlier cohorts, as they benefited from increased investor trust, supportive regulations and a more mature FinTech ecosystem. The post-crises regulatory obligations turned in favor of technology companies. Public perception of banks coincided with increased regulatory pressures that limited banks ' ability to innovate. Consequently, the described market context provided fruitful ground for new FinTech startups to emerge and establish a new paradigm in the finance industry. (Arner et al., 2015).

3.4 Experiment Setup and Results

The first issue we encountered was the sparsity of Crunchbase database. The dataset contained missing values ('NaN's), so the machine learning algorithms could not handle these 'NaN 'values directly. The fact that notable entities and features are frequently reviewed while new incomplete profiles of relatively young companies exponentially grow, a high sparsity level of Crunchbase dataset is inevitable. To fill in missing values, we implemented a zero-imputation strategy. Single imputation of missing data is based on the idea that any value in a study sample can be replaced with only one estimate from the same source population. (Ar et al., 2006) In this approach, the estimate was zero, with the assumption that the instances in features *('num_exits', 'investor_count 'and 'num_funding_rounds')* do not contain value, it is because the value is zero.

While this strategy is straightforward and easy to implement, it can lead to biased estimates and inaccurate predictions. In contrast, a more sophisticated imputation techniques, such as multiple imputation, generate multiple plausible values for each missing data point, based on observed data and the relationship between features. Standard softwares such as SAS and S-Plus provide access to these techniques.

Additionally, the idea behind creating the new variable *'work_experience'* was that it shows the number of years an employee, CEO, founder, etc., spent in one or more companies since the graduation. Unfortunately, we do not possess data about their previous work experience, and the feature itself contains a large proportion of NaN values. Thus, we excluded this variable from the prediction model.

The second issue found in the data analysis was a large class imbalance between successful and non-successful companies. After pre-processing, only 15.3% of the dataset was classified as non-successful companies. The class imbalance issue causes that model trained on such data tend to be biased towards the majority class. Machine learning algorithms perform best when the number of samples in each class is approximately equal. To mitigate class imbalance within our dataset, we explored the application of the Synthetic Minority Oversampling Technique (SMOTE) across several predictive models, including Random Forest, Support Vector Machine (SVM), and XGBoost. SMOTE is a sophisticated oversampling approach that generates synthetic instances of the minority class by interpolating between existing minority class examples. These synthetic examples are crafted as linear combinations of two proximate samples from the minority class, which are assumed to have identical expected values and variances. For high-dimensional datasets, the efficacy of SMOTE is enhanced through careful variable selection, which helps to reduce Euclidean distance and improve performance. Although SMOTE can offer advantages over simple oversampling methods, its impact may be limited for classifiers that depend on mean values, and it may lead to a loss of sample independence. The technique has proven effective in a range of applications, including network intrusion detection, species distribution, and breast cancer prediction. (Biau & Scornet, 2016; Blagus & Lusa, 2013; Breiman, 2001; Chawla et al., 2002; Denisko & Hoffman, 2018)

However, the application of SMOTE and other oversampling techniques introduces additional computational complexity by generating new synthetic data points and eventually increasing the size of the training dataset. Additionally, the computational cost of training predictive models on this augmented dataset can be considerably higher, especially for inherently more computationally demanding algorithms. Given this consideration, we opted to implement the SMOTE technique exclusively in the context of the Random Forest model. Random Forest is particularly well-suited to handle larger datasets and is relatively efficient in terms of computational resources compared to algorithms such as SVM and XGBoost. Furthermore, Random Forest models are known for their robustness and ability to handle imbalanced data, which makes integration of SMOTE a strategic choice to maximize performance gains while managing computational cost. This decision allowed us to rationalize computational expenses that may have been incurred if we applied the SMOTE technique in all considered algorithms. (Biau & Scornet, 2016; Blagus & Lusa, 2013; Breiman, 2001; Chawla et al., 2002; Denisko & Hoffman, 2018)

The experiment results were evaluated in the context of different useful metrics, such as accuracy, precision, recall, support and F1-score. We first introduce the accuracy metric that answers the question, “Out of all the predictions made, how many were true?” It is calculated as the ratio of the sum of true positives (TP) and true negatives (TN) to the total sum of true positives, true negatives, false negatives (FN), and false positives (FP). In simpler terms, it represents the proportion of correct predictions out of all predictions. (Yacouby & Axman, 2020)

Table 5 – *Accuracy Metric of Machine Learning Models*

Machine Learning Model	**Accuracy**
Random Forest	87.01%
Support Vector Machines	84.26%
XGBoost	98.61%

Source: Author’s Calculation

When comparing the performance of machine learning models, relying on the accuracy metric shown in the Table 5, XGBoost significantly outperforms other models like Random Forest and SVM with notable differences (XGBoost at 98.6% vs Random Forest at 87% and SVM at 84%). The explanation for such disparity may be the sequential manner of the XGBoost model construction. The approach builds models such that each new model corrects errors made by the previous model, leading to better nuance capture in the data. The boosting mechanism is especially preferable in cases where relationships between features and the target variable are non-linear and complex. Additionally, the superior performance of XGBoost may be attributed to its intrinsic ability to handle imbalanced classes more effectively. The XGBoost parameters like ‘scale_pos_weight ’help in adjusting the algorithm’s focus towards minority classes without careful tuning or additional techniques like SMOTE. Nevertheless, to investigate and validate these results further, it is pivotal to evaluate model performance across a range of other metrics. (Chen & Guestrin, 2016; Chen & He, n.d.; Ramraj et al., 2016)

The classification is a performance evaluation metric used in machine learning to assess the accuracy of a classification model. It provides a comprehensive overview of **the precision, recall, f1-score and support** for each class in a classification problem. This report is particularly useful for understanding the performance of a model where some classes are more important than others or when dealing with an imbalanced dataset.

a) Precision is the ratio of correctly predicted positive observations to the total predictive positives. It is an especially useful metric when the cost of a false positive is high. The formula is TP/ (TP+FP).

b) Recall (Sensitivity) is the ratio of correctly predicted positive observations to all observations in actual class. Recall is crucial when the cost of a false negative is high. Formula: TP / (TP+FN).

c) F1-Score is the weighted average of precision and recall. Thus, it considers both false positives and false negatives. It is a better measure than examining precision and recall independently, especially in cases of imbalanced datasets.

d) Support is the number of actual occurences of the class in the specified dataset. For each class, it shows how many instances of the class were present in the dataset, providing insight into the dataset's balance across classes.

Table 6 – *Classification Report of Machine Learning Models*

Random Forest		**precision**	**recall**	**f1-score**	**support**
	0	0.58	0.58	0.58	418
	1	0.92	0.92	0.92	2314
	Macro avg	0.75	0.75	0.75	2732
	weighted avg	0.87	0.87	0.87	2732
		precision	recall	f1-score	support

SVM	0	0.00	0.00	0.00	429
	1	0.84	1.0	0.91	2303
	macro avg	0.42	0.5	0.46	2732
	Weighted avg	0.71	0.84	0.77	2732
XGBoost		precision	recall	f1-score	support
	0	0.98	0.93	0.95	429
	1	0.99	1.0	0.99	2302
	macro avg	0.98	0.96	0.97	2732
	Weighted avg	0.99	0.99	0.99	2732

Source: Author’s Calculation

Table 7- *Confusion Matrix*

Models	**TPR**	**TNR**	**FPR**	**FNR**
XGBoost	0.997	0.9277	0.0723	0.003
RF	0.8493	0.1678	0.8322	0.1507
SVM	0.9996	0.0	1.0	0.0004

Source: Author’s Calculation

Due to class imbalance, we rely on the macro average precision, recall and F1-score. The macro average gives equal weight to every class, no matter how many instances of that class are in the dataset. On the other hand, the confusion matrix shown in Table 6 provides insights into the true positive rate (TPR), true negative rate (TNR), and false positive rate (FPR), false negative rate (FNR). The ideal model would have high TPR values and low FPR and FNR values, indicating that it accurately classifies positive and negative cases while minimizing errors. Ultimately, the XGBoost algorithm is the chosen one, performing the best score in each evaluation metric.

Table 8 – *Classification Report of Random Forest model (after SMOTE application)*

accurancy=0.73	precision	recall	f1-score	support
0	0.33	0.70	0.44	418
1	0.93	0.74	0.82	2314
macro avg	0.63	0.72	0.63	2732
weighted avg	0.84	0.73	0.77	2732

Source: Author’s Calculation

SMOTE technique in the case of the Random Forest model led to lower precision, recall and f1-score (Table 8 compared to Table 7). The possible explanation for this result is that with an increased number of synthetic samples, the model might become too focused on the minority class, potentially at the expense of the majority class performance. The synthetic samples the SMOTE technique generated may also represent unrealistic data points that confuse the model.

3.5 Feature Importance

Identifying influential features can provide valuable insights into the factors that contribute most significantly to a startup’s success, guiding stakeholders in making rational and informed decisions. Even though the existing literature indicates that the fintech services and products are gaining more importance in some areas, we still do not fully distinguish features directly causing a startup’s success.

Table 9 shows our results of the 20 most influential features sorted by importance to the XGBoost model.

Table 9 – *Top 20 features*

Rank	**Feature**	**Importance**
1	USA	0.072
2	Information Technology, Mobile, Mobile, Privacy and Security	0.035

3	avg_amount_raised	0.028
4	Stockholm	0.025
5	Information Technology, Mobile, Payment, Software	0.02
6	Financial Services	0.02
7	Apps, Commerce and Shopping, Internet Services, Media and Entertainment, Mobile, Software	0.0197
8	total_funding_usd	0.0184
9	Consumer Electronics, Hardware, Professional Services	0.0176
10	Massachusetts	0.0174
11	Commerce and Shopping, Financial Services, Lending and Investments, Mobile, Software	0.0168
12	Data and Analytics, Financial Services	0.0157
13	Financial Services, Lending and Investments, Software	0.0149
14	Zurich	0.0146
15	Data and Analytics, Financial Services, Software	0.0144
16	Commerce and Shopping, Financial Services, Internet Services, Payments, Sales and Marketing	0.0138
17	FRA	0.0137
18	Financial Services, Professional Services	0.0135
19	Financial Services, Hardware, Information Technology	0.0121
20	Apps, Commerce and Shopping, Internet Services, Media and Entertainment, Mobile, Software	0.0118

Source: Author’s Calculation

When summarizing the most influential features presented in Table 8, it is pivotal to commence by categorizing them into 4 different groups:

a) Services
b) Geographical Location
c) Funding Details
d) Technological aspects

The nature of the services provided by FinTech companies has proven to play a crucial role in their success. We can see that services such as Lending and Investments, Commerce and Shopping, Software, Hardware, Data and Analytics, and Mobile and Internet services showed strong positive correlations with success. Furthermore, the diversified palette of services offered by FinTech companies, ranging from Financial Services to Commerce and Sales indicates that the startups should strive to expand their expertise but not at the cost of lower product quality. The financial services provided by FinTech startups are especially attractive due to their lenient credit scoring and lower regulatory burdens.

When it comes to geographical location, we identify the United States of America, Massachusetts, Sweden (Stockholm), France and Switzerland (Zurich) as the most positively correlated with FinTech success. These geographical patterns can be explained by the strong ecosystems of investors and partners, high levels of digital adoption among consumers, supportive regulatory frameworks, high GDP per capita and finally, the widespread availability of the Internet and mobile devices. The higher demand for FinTech has proven to be highly correlated with a higher bank regulatory burden and a more concentrated market. Cities such as Stockholm and Zurich are well-known for their proximity to local startup culture, access to talent and thus grinding creation of network opportunities.

The adoption of Artificial intelligence, Big Data, virtual currencies, etc., in a wide variety of startup services brings a more efficient and effective approach to delivering desired products to clients. It also allows more precise and efficient monitoring of the internal startup's operations. On the other hand, both the total funding amount and average funding amount per round are strong indicators investors follow. The higher the funding amount is, the more investors tend to believe in the startup's exit or merger and acquisition.

The multidimensional nature of startup success proves that there is a pivotal urge for conducive ecosystems to support innovation, access to funding, and education while the regulatory framework encourages new business models. Only a framework that promotes these values would make the economy flourish and reap the benefits of this disruptive industry.

CONCLUSION

Predicting the success of FinTech startups is a challenging task, but it is critical to stakeholders who shape the economy by investing. Intuitively, as the company matures and undergoes angel and VC funding rounds, it becomes easier for private and public investors to foresee the future startup outcome.

The main aim of the study was to identify the optimal machine learning algorithm for predicting the success of FinTech startups. We built machine learning models and compared the performance of three algorithms: Random Forest, Support Vector Machine, and XGBoost. The research culminated in the selection of XGBoost algorithm, for an exceptional accuracy rate of 99%, TPR 97%, and TNR of 93%. We witnessed an outstanding model performance of XGBoost precisely intricating patterns underlying FinTech success.

The construction of our predictive model was enriched by creating new variables. Thid process allowed us to capture a more nuanced and comprehensive view of the factors influencing FinTech success. The *investment type, employee count, success age, work experience*, *average amount raised* variables unravelled financial and organizational determinants that propel a FinTech startup towards success.

A crucial aspect of our study addressed the challenge posed by imbalanced classes within our Crunchbase dataset. To counteract this imbalance, we implemented the Synthetic Minority Over-sampling Technique (SMOTE). However, the anticipated enhancement in model performance did not materialize as expected. The inherent complexity of the data and the potential for SMOTE to introduce artificial noise led to the deterioration of Random Forest model results.

Our analysis discovered that features that have the most influence over a startup's success can be broadly classified into four categories: location, funding, technology and service mix. Each category provided valuable insights into the FinTech ecosystem. The model proved the literature's discoveries when it comes to the geographical location. FinTech startups are more prone to achieve success in countries that promote startup culture, supportive regulatory framework and educational centres. The most relevant location features are The United States, France, Sweden and Switzerland. Furthermore, the prediction model proved that service mix elements play a pivotal role, especially in the case of a diversified service mix. These findings not only contribute valuable insights to the academic discourse but also offer practical implications for practitioners seeking to navigate the evolving FinTech landscape.

Future work could increase the recall of the models by enriching the dataset. More detailed data about the founder’s prior education, work experience and company’s service could improve the performance of the models. Text sentiment analysis from relevant websites such as Twitter and Reddit may provide additional relevant sources of features for the dataset, especially when it comes to the company’s reputation. Furthermore, the approach may be improved by gathering snapshots of the CrunchBase database at regular intervals. Thus, modeling using time-series techniques would unravel the dynamics of the company’s growth.

Bibliography

1. *2018 Global FinTech Hub Report*. (n.d.). Cambridge Judge Business School. Retrieved August 14, 2024, from https://www.jbs.cam.ac.uk/faculty-research/centres/alternative-finance/publications/2018-global-fintech-hub-report/
2. Abbasi, K., Alam, A., Du, M. (Anna), & Huynh, T. L. D. (2021). FinTech, SME efficiency and national culture: Evidence from OECD countries. *Technological Forecasting and Social Change*, *163*, 120454. https://doi.org/10.1016/j.techfore.2020.120454
3. Abdulkareem, N. M., & Abdulazeez, A. M. (2021). *Machine Learning Classification Based on Radom Forest Algorithm: A Review*. https://doi.org/10.5281/ZENODO.4471118
4. Acs, Z. J., Braunerhjelm, P., Audretsch, D. B., & Carlsson, B. (2009). The knowledge spillover theory of entrepreneurship. *Small Business Economics*, *32*(1), 15–30. https://doi.org/10.1007/s11187-008-9157-3
5. Aggarwal, C. C. (2015). *Data Mining*. https://link.springer.com/book/10.1007/978-3-319-14142-8
6. Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., & Aljaaf, A. J. (2020). A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science. In *Supervised and Unsupervised Learning for Data Science* (pp. 3–21). Springer, Cham. https://doi.org/10.1007/978-3-030-22475-2_1
7. Alvarez, S. A., & Busenitz, L. W. (2007). The Entrepreneurship of Resource-based Theory*. In *Entrepreneurship* (pp. 207–227). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-48543-8_10
8. Anagnostopoulos, I. (2018). Fintech and regtech: Impact on regulators and banks. *Journal of Economics and Business*, *100*, 7–25. https://doi.org/10.1016/j.jeconbus.2018.07.003
9. Ar, D., Gj, van der H., T, S., & Kg, M. (2006). Review: A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology*, *59*(10). https://doi.org/10.1016/j.jclinepi.2006.01.014
10. Arner, D. W., Barberis, J. N., & Buckley, R. P. (2015). *The Evolution of Fintech: A New Post-Crisis Paradigm?* https://doi.org/10.2139/ssrn.2676553
11. Åstebro, T., & Elhedhli, S. (2006). The Effectiveness of Simple Decision Heuristics: Forecasting Commercial Success for Early-Stage Ventures. *Management Science*, *52*(3), 395–409. https://doi.org/10.1287/mnsc.1050.0468
12. Attenberg, J., Ipeirotis, P., & Provost, F. (2015). Beat the Machine: Challenging Humans to Find a Predictive Model's "Unknown Unknowns." *Journal of Data and Information Quality*, *6*(1), 1:1-1:17. https://doi.org/10.1145/2700832
13. Baer, J., & McKool, S. S. (2014). The Gold Standard for Assessing Creativity. *International Journal of Quality Assurance in Engineering and Technology Education (IJQAETE)*, *3*(1), 81–93. https://doi.org/10.4018/ijqaete.2014010104
14. Ben-Hur, A., & Weston, J. (2010). A User's Guide to Support Vector Machines. In *Data Mining Techniques for the Life Sciences* (pp. 223–239). Humana Press. https://doi.org/10.1007/978-1-60327-241-4_13

15. Berg, T., Burg, V., Gombović, A., & Puri, M. (2020). On the Rise of FinTechs: Credit Scoring Using Digital Footprints. *The Review of Financial Studies*, *33*(7), 2845–2897. https://doi.org/10.1093/rfs/hhz099
16. Bernstein, S., Korteweg, A., & Laws, K. (2017). Attracting Early-Stage Investors: Evidence from a Randomized Field Experiment. *The Journal of Finance*, *72*(2), 509–538. https://doi.org/10.1111/jofi.12470
17. Bethlendi, A., & Szőcs, Á. (2022). How the Fintech ecosystem changes with the entry of Big Tech companies. *Investment Management and Financial Innovations*, *19*(3), 38–48. https://doi.org/10.21511/imfi.19(3).2022.04
18. Bhave, M. P. (1994). A process model of entrepreneurial venture creation. *Journal of Business Venturing*, *9*(3), 223–242. https://doi.org/10.1016/0883-9026(94)90031-0
19. Bian, W., Ge, T., Ji, Y., & Wang, X. (2023). How is Fintech reshaping the traditional financial markets? New evidence from InsurTech and insurance sectors in China. *China Economic Review*, *80*, 102004. https://doi.org/10.1016/j.chieco.2023.102004
20. Biau, G., & Scornet, E. (2016). A random forest guided tour. *TEST*, *25*(2), 197–227. https://doi.org/10.1007/s11749-016-0481-7
21. Blagus, R., & Lusa, L. (2013). SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics*, *14*(1), Article 1. https://doi.org/10.1186/1471-2105-14-106
22. Boot, A., Hoffmann, P., Laeven, L., & Ratnovski, L. (2021). Fintech: What’s old, what’s new? *Journal of Financial Stability*, *53*, 100836. https://doi.org/10.1016/j.jfs.2020.100836
23. Boot, A. W. A. (2000). Relationship Banking: What Do We Know? *Journal of Financial Intermediation*, *9*(1), 7–25. https://doi.org/10.1006/jfin.2000.0282
24. Botsch, M., & Vanasco, V. (2019). Learning by lending. *Journal of Financial Intermediation*, *37*, 1–14. https://doi.org/10.1016/j.jfi.2018.03.002
25. Brau, J. C., & Fawcett, S. E. (2006). Initial Public Offerings: An Analysis of Theory and Practice. *The Journal of Finance*, *61*(1), 399–436. https://doi.org/10.1111/j.1540-6261.2006.00840.x
26. Braun, A., & Schreiber, F. (2017). *The Current InsurTech Landscape: Business Models and Disruptive Potential* (Research Report 62). I.VW HSG Schriftenreihe. https://www.econstor.eu/handle/10419/226646
27. Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324
28. Butler, J. E., Doktor, R., & Lins, F. A. (2010). Linking international entrepreneurship to uncertainty, opportunity discovery, and cognition. *Journal of International Entrepreneurship*, *8*(2), 121–134. https://doi.org/10.1007/s10843-010-0054-x
29. Cao, L., Halvardsson, G., McCornack, A., von Ehrenheim, V., & Herman, P. (2023, September 28). *Sourcing Investment Targets for Venture and Growth Capital Using Multivariate Time Series Transformer*. arXiv.Org. https://arxiv.org/abs/2309.16888v1
30. Carbó-Valverde, S., Cuadros-Solas, P. J., & Rodríguez-Fernández, F. (2022). Entrepreneurial, institutional and financial strategies for FinTech profitability. *Financial Innovation*, *8*(1), 1–36.

31. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953
32. Chen, M. A., Wu, Q., & Yang, B. (2019). How Valuable Is FinTech Innovation? *The Review of Financial Studies*, *32*(5), 2062–2106. https://doi.org/10.1093/rfs/hhy130
33. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785
34. Claessens, S., Frost, J., Turner, G., & Zhu, F. (2018). *Fintech Credit Markets Around the World: Size, Drivers and Policy Issues*. https://papers.ssrn.com/abstract=3288096
35. Cojoianu, T. F., Clark, G. L., Hoepner, A. G. F., Pažitka, V., & Wójcik, D. (2021). Fin vs. tech: Are trust and knowledge creation key ingredients in fintech start-up emergence and financing? *Small Business Economics*, *57*(4), 1715–1731. https://doi.org/10.1007/s11187-020-00367-3
36. Corea, F., Bertinetti, G., & Cervellati, E. M. (2021). Hacking the venture industry: An Early-stage Startups Investment framework for data-driven investors. *Machine Learning with Applications*, *5*, 100062. https://doi.org/10.1016/j.mlwa.2021.100062
37. Crowne, M. (2002). *Why software product startups fail and what to do about it. Evolution of software product development in startup companies*. *1*, 338–343 vol.1. https://doi.org/10.1109/IEMC.2002.1038454
38. Cumming, D., & Schwienbacher, A. (2018). Fintech Venture Capital. *Corporate Governance: An International Review*, *26*. https://doi.org/10.1111/corg.12256
39. Da Rin, M., & Hellmann, T. (2020). *Fundamentals of entrepreneurial finance*. Oxford University Press. https://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=2363737
40. Davila, A., Foster, G., & Gupta, M. (2003). Venture capital financing and the growth of startup firms. *Journal of Business Venturing*, *18*(6), 689–708. https://doi.org/10.1016/S0883-9026(02)00127-1
41. Dellermann, D., Lipusch, N., Ebel, P., Popp, K. M., & Leimeister, J. M. (2021). *Finding the unicorn: Predicting early stage startup success through a hybrid intelligence method* (arXiv:2105.03360). arXiv. http://arxiv.org/abs/2105.03360
42. Denisko, D., & Hoffman, M. M. (2018). Classification and interaction in random forests. *Proceedings of the National Academy of Sciences*, *115*(8), 1690–1692. https://doi.org/10.1073/pnas.1800256115
43. Díaz-Uriarte, R., & Alvarez de Andrés, S. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, *7*(1), 3. https://doi.org/10.1186/1471-2105-7-3
44. Farinos, J., & Sanchis, V. (2009). Factores determinantes de la salida a bolsa en España. *Working Papers. Serie EC*, Article 2009–03. https://ideas.repec.org//p/ivi/wpasec/2009-03.html
45. *Finding The Most Significant Predictors of Startup Success with Machine Learning*. (n.d.). Eindhoven University of Technology Research Portal. Retrieved January 18, 2024, from

https://research.tue.nl/en/studentTheses/finding-the-most-significant-predictors-of-startup-success-with-m

46. *Fintech: A Brief History | EC1 Partners*. (2023, April 4). https://ec1partners.com/blog/fintech-a-brief-history/
47. *Fintech funding: Why It's dropped & what comes next*. (n.d.). Carta. Retrieved December 7, 2023, from https://carta.com/blog/fintech-funding-falls/
48. Fischer, A. (2017, July 28). *About Crunchbase*. Crunchbase. https://about.crunchbase.com/about-us/
49. Frost, J. (2020). The Economic Forces Driving FinTech Adoption across Countries. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3515326
50. Gaughan, P. A. (2010). *Mergers, Acquisitions, and Corporate Restructurings*. John Wiley & Sons.
51. Gazel, M., & Schwienbacher, A. (2021). Entrepreneurial fintech clusters. *Small Business Economics*, *57*(2), 883–903.
52. Gentry, R. J., Dalziel, T., & Jamison, M. A. (2013). Who Do Start-Up Firms Imitate? A Study of New Market Entries in the CLEC Industry. *Journal of Small Business Management*, *51*(4), 525–538. https://doi.org/10.1111/jsbm.12055
53. Gomber, P., Kauffman, R. J., Parker, C., & Weber, B. W. (2018). On the Fintech Revolution: Interpreting the Forces of Innovation, Disruption, and Transformation in Financial Services. *Journal of Management Information Systems*, *35*(1), 220–265. https://doi.org/10.1080/07421222.2018.1440766
54. Gregor, S., & Hevner, A. R. (2013). Positioning and Presenting Design Science Research for Maximum Impact. *MIS Quarterly*, *37*(2), 337–355.
55. Group, S. M. A., Engineering (NITIE), N. I. of I., Lake, V., Mumbai, Group, I. R. A., Engineering (NITIE), N. I. of I., Lake, V., Mumbai, Group, I. T. A., Engineering (NITIE), N. I. of I., Lake, V., Mumbai, & India. (2015). Internet of Things (IoT): A Literature Review. *Journal of Computer and Communications*, *03*(05), Article 05. https://doi.org/10.4236/jcc.2015.35021
56. Guo, B., Lou, Y., & Pérez-Castrillo, D. (2015). Investment, Duration, and Exit Strategies for Corporate and Independent Venture Capital-Backed Start-Ups. *Journal of Economics & Management Strategy*, *24*(2), 415–455. https://doi.org/10.1111/jems.12097
57. Haddad, C., & Hornuf, L. (2019). The emergence of the global fintech market: Economic and technological determinants. *Small Business Economics*, *53*(1), 81–105.
58. Hall, B. H., & Lerner, J. (2010). Chapter 14—The Financing of R&D and Innovation. In B. H. Hall & N. Rosenberg (Eds.), *Handbook of the Economics of Innovation* (Vol. 1, pp. 609–639). North-Holland. https://doi.org/10.1016/S0169-7218(10)01014-2
59. Han, J., Cai, Y., & Cercone, N. (1992). Knowledge discovery in databases: An attribute-oriented approach. *VLDB*, *18*, 574–559. http://hanj.cs.illinois.edu/pdf/vldb92.pdf
60. Han J, Pei J, Kamber M. (2011). Data mining: Concepts and techniques. *Elsevier*. https://doi.org/10.1007/s42979-021-00592-x
61. Harasim, J. (2021). FinTechs, BigTechs and Banks—When Cooperation and When Competition? *Journal of Risk and Financial Management*, *14*(12), 614. https://doi.org/10.3390/jrfm14120614

62. Hassan, M. U., Iqbal, Z., Malik, M., & Ahmad, M. I. (2018). Exploring the role of technological developments and open innovation in the survival of SMEs: An empirical study of Pakistan. *International Journal of Business Forecasting and Marketing Intelligence*, *4*(1), 64–85.
63. Hevner, A. R. (2007). A three cycle view of design science research. *Scandinavian Journal of Information Systems*, *19*(2), 4.
64. Hornuf, L., Klus, M., Lohwasser, T., & Schwienbacher, A. (2020). *How Do Banks Interact with Fintech Startups?* (SSRN Scholarly Paper 3252318). https://doi.org/10.2139/ssrn.3252318
65. Jagtiani, J., & Lemieux, C. (2019). The roles of alternative data and machine learning in fintech lending: Evidence from the LendingClub consumer platform. *Financial Management*, *48*(4), 1009–1029. https://doi.org/10.1111/fima.12295
66. Jain, B. A., & Kini, O. (1999). The Life Cycle of Initial Public Offering Firms. *Journal of Business Finance & Accounting*, *26*(9–10), 1281–1307. https://doi.org/10.1111/1468-5957.00298
67. *Jim Marous*. (n.d.). Forbes. Retrieved December 7, 2023, from https://www.forbes.com/sites/jimmarous/
68. Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260. https://doi.org/10.1126/science.aaa8415
69. Kantardzic, M. (2003). *Data Mining*. https://books.google.com/books/about/Data_Mining.html?hl=sr&id=ZZ7l6v0CvRMC
70. Keuschnigg, M., & Ganser, C. (2017). Crowd Wisdom Relies on Agents' Ability in Small Groups with a Voting Aggregation Rule. *Management Science*, *63*(3), 818–828. https://doi.org/10.1287/mnsc.2015.2364
71. Kim, Y., Choi, J., Park, Y.-J., & Yeon, J. (2016). The adoption of mobile payment services for "fintech." *International Journal of Applied Engineering Research*, *11*(2), 1058–1061.
72. Klus, M., Lohwasser, T., Holotiuk, F., & Moormann, J. (2019). Strategic Alliances between Banks and Fintechs for Digital Innovation: Motives to Collaborate and Types of Interaction. *The Journal of Entrepreneurial Finance*, *21*(1). https://doi.org/10.57229/2373-1761.1346
73. Kopera, S., Wszendybył-Skulska, E., Cebulak, J., & Grabowski, S. (2018). Interdisciplinarity in tech startups development: Case study of "unistartapp" project. *Foundations of Management*, *10*(1), 23–32. https://doi.org/10.2478/fman-2018-0003
74. Kou, G., Olgu Akdeniz, Ö., Dinçer, H., & Yüksel, S. (2021). Fintech investments in European banks: A hybrid IT2 fuzzy multidimensional decision-making approach. *Financial Innovation*, *7*(1), 39. https://doi.org/10.1186/s40854-021-00256-y
75. Krishna, A., Agrawal, A., & Choudhary, A. (2016). Predicting the Outcome of Startups: Less Failure, More Success. *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, 798–805. https://doi.org/10.1109/ICDMW.2016.0118
76. Kwak, S. K., & Kim, J. H. (2017). Statistical data preparation: Management of missing values and outliers. *Korean Journal of Anesthesiology*, *70*(4), 407–411. https://doi.org/10.4097/kjae.2017.70.4.407

77. Laidroo, L., & Avarmaa, M. (2020). The role of location in FinTech formation. *Entrepreneurship & Regional Development*, *32*(7–8), 555–572. https://doi.org/10.1080/08985626.2019.1675777

78. Lanza, A., & Passarelli, M. (2014). Technology Change and Dynamic Entrepreneurial Capabilities. *Journal of Small Business Management*, *52*(3), 427–450. https://doi.org/10.1111/jsbm.12042

79. Li, T., Kou, G., Peng, Y., & Yu, P. S. (2022). An Integrated Cluster Detection, Optimization, and Interpretation Approach for Financial Data. *IEEE Transactions on Cybernetics*, *52*(12), 13848–13861. IEEE Transactions on Cybernetics. https://doi.org/10.1109/TCYB.2021.3109066

80. Liberti, J. M., & Petersen, M. A. (2019). Information: Hard and Soft. *The Review of Corporate Finance Studies*, *8*(1), 1–41. https://doi.org/10.1093/rcfs/cfy009

81. Liu, J., & Li, D. (2014). The life cycle of initial public offering companies in China. *Journal of Applied Accounting Research*, *15*, 291–307. https://doi.org/10.1108/JAAR-12-2013-0111

82. Loughran, T., Ritter, J. R., & Rydqvist, K. (1994). Initial public offerings: International insights. *Pacific-Basin Finance Journal*, *2*(2–3), 165–199. https://doi.org/10.1016/0927-538X(94)90016-7

83. McCallum, A. (2005). Information Extraction: Distilling structured data from unstructured text. *Queue*, *3*(9), 48–57. https://doi.org/10.1145/1105664.1105679

84. Mitchell, T. (2006). *The Discipline of Machine Learning*.

85. Mohammed, M., Khan, M., & Bashier, E. (2016). Machine Learning: Algorithms and Applications. In *Machine Learning: Algorithms and Applications*. https://doi.org/10.1201/9781315371658

86. Morrissette, S. G. (2007). A Profile of Angel Investors. *The Journal of Private Equity*, *10*(3), 52–66.

87. Nicholson, A., & Smyth, P. (2013, September 30). *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence (2013)*. arXiv.Org. https://arxiv.org/abs/1309.7971v2

88. Nicoletti, B. (2020). *Insurance 4.0: Benefits and Challenges of Digital Transformation*. Springer Nature.

89. *OECD Digital Economy Outlook 2015 | READ online*. (n.d.). Oecd-Ilibrary.Org. Retrieved December 3, 2023, from https://read.oecd-ilibrary.org/science-and-technology/oecd-digital-economy-outlook-2015_9789264232440-en

90. Otter, D. W., Medina, J. R., & Kalita, J. K. (2019). *A Survey of the Usages of Deep Learning in Natural Language Processing* (arXiv:1807.10854). arXiv. https://doi.org/10.48550/arXiv.1807.10854

91. Parmar, A., Katariya, R., & Patel, V. (2019). A Review on Random Forest: An Ensemble Classifier. *International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018*, 758–763. https://doi.org/10.1007/978-3-030-03146-6_86

92. Pisner, D. A., & Schnyer, D. M. (2020). Support vector machine. In *Machine Learning* (pp. 101–121). Elsevier. https://doi.org/10.1016/B978-0-12-815739-8.00006-7

93. Ramraj, S., Uzir, N., Sunil, R., & Banerjee, S. (2016). Experimenting XGBoost algorithm for prediction and classification of different datasets. *International Journal of Control Theory and Applications*, *9*(40), 651–662.

94. Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Systems*, *89*, 14–46. https://doi.org/10.1016/j.knosys.2015.06.015

95. Rosavina, M., Rahadi, R. A., Kitri, M. L., Nuraeni, S., & Mayangsari, L. (2019). P2P lending adoption by SMEs in Indonesia. *Qualitative Research in Financial Markets*, *11*(2), 260–279. https://doi.org/10.1108/QRFM-09-2018-0103

96. Rose, J. (n.d.). *Software Entrepreneurship: Two paradigms for promoting new information technology ventures*.

97. Rw, W., R, H., Nh, S., W, D., & E, H. (2014). Toward enhanced pharmacovigilance using patient-generated data on the internet. *Clinical Pharmacology and Therapeutics*, *96*(2). https://doi.org/10.1038/clpt.2014.77

98. Sangwan, V., Harshita, Prakash, P., & Singh, S. (2019). Financial technology: A review of extant literature. *Studies in Economics and Finance*, *37*(1), 71–88. https://doi.org/10.1108/SEF-07-2019-0270

99. Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, *2*(3), 160. https://doi.org/10.1007/s42979-021-00592-x

100. Sarker, I. H., Hoque, M. M., Uddin, Md. K., & Alsanoosy, T. (2021). Mobile Data Science and Intelligent Apps: Concepts, AI-Based Modeling and Research Directions. *Mobile Networks and Applications*, *26*(1), 285–303. https://doi.org/10.1007/s11036-020-01650-z

101. Sarker, I. H., Kayes, A. S. M., Badsha, S., Alqahtani, H., Watters, P., & Ng, A. (2020). Cybersecurity data science: An overview from machine learning perspective. *Journal of Big Data*, *7*(1), Article 1. https://doi.org/10.1186/s40537-020-00318-5

102. Saura, J., Palos-Sanchez, P., & Grilo, A. (2019). Detecting Indicators for Startup Business Success: Sentiment Analysis Using Text Data Mining. *Sustainability*, *11*, 917. https://doi.org/10.3390/su11030917

103. Schmidhuber, J. (1997). Discovering Neural Nets with Low Kolmogorov Complexity and High Generalization Capability. *Neural Networks*, *10*(5), 857–873. https://doi.org/10.1016/S0893-6080(96)00127-X

104. Shaik, A. B., & Srinivasan, S. (2019). A Brief Survey on Random Forest Ensembles in Classification Model. In *International Conference on Innovative Computing and Communications* (pp. 253–260). Springer, Singapore. https://doi.org/10.1007/978-981-13-2354-6_27

105. Sharchilev, B., Roizner, M., Rumyantsev, A., Ozornin, D., Serdyukov, P., & de Rijke, M. (2018). Web-based Startup Success Prediction. *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 2283–2291. https://doi.org/10.1145/3269206.3272011

106. Sirower, M. L., & O'Byrne, S. F. (1998). The Measurement of Post-Acquisition Performance: Toward a Value-Based Benchmarking Methodology. *Journal of Applied Corporate Finance*, *11*(2), 107–121. https://doi.org/10.1111/j.1745-6622.1998.tb00652.x

107. Stoeckli, E., Dremel, C., & Uebernickel, F. (2018). Exploring characteristics and transformational capabilities of InsurTech innovations to understand insurance value creation in a digital world. *Electronic Markets*, *28*(3), 287–305. https://doi.org/10.1007/s12525-018-0304-7

108. *Strategic management: Competitiveness and globalization (Concepts and cases) -ORCA*. (n.d.). Retrieved December 8, 2023, from https://orca.cardiff.ac.uk/id/eprint/25299/

109. Stulz, R. M. (2019). FinTech, BigTech, and the Future of Banks. *Journal of Applied Corporate Finance*, *31*(4), 86–97. https://doi.org/10.1111/jacf.12378

110. Taylor, M. E., & Stone, P. (2009). Transfer Learning for Reinforcement Learning Domains: A Survey. *The Journal of Machine Learning Research*, *10*, 1633–1685.

111. *The Future of FinTech: A Paradigm Shift in Small Business Finance*. (n.d.). World Economic Forum. Retrieved March 22, 2024, from https://www.weforum.org/publications/future-fintech-paradigm-shift-small-business-finance/

112. Thrun, S., & Pratt, L. (Eds.). (1998). *Learning to Learn*. Springer US. https://doi.org/10.1007/978-1-4615-5529-2

113. Tomy, S., & Pardede, E. (2023). *From Uncertainties to Successful Start Ups: A Data Analytic Approach to Predict Success in Technological Entrepreneurship*. https://doi.org/10.26181/22881692.v1

114. *Unlocking the potential of the Internet of Things | McKinsey*. (n.d.). Retrieved March 20, 2024, from https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-internet-of-things-the-value-of-digitizing-the-physical-world

115. Varga, D. (2017). Fintech, the new era of financial services. *Vezetéstudomány / Budapest Management Review*, *48*(11), 22–32. https://doi.org/10.14267/VEZTUD.2017.11.03

116. vor dem Esche, J., & Hennig-Thurau, T. (2014). *German Digitalization Consumer Report 2014*. https://doi.org/10.13140/2.1.3071.0402

117. Weber, J. A., & Dholakia, U. M. (2000). Including Marketing Synergy in Acquisition Analysis: A Step-Wise Approach. *Industrial Marketing Management*, *29*(2), 157–177. https://doi.org/10.1016/S0019-8501(99)00062-0

118. Wei, C.-P., Jiang, Y.-S., & Yang, C.-S. (2009). Patent Analysis for Supporting Merger and Acquisition (M&A) Prediction: A Data Mining Approach. In C. Weinhardt, S. Luckner, & J. Stößer (Eds.), *Designing E-Business Systems. Markets, Services, and Networks* (pp. 187–200). Springer. https://doi.org/10.1007/978-3-642-01256-3_16

119. Wójcik, D. (2021). Financial Geography I: Exploring FinTech – Maps and concepts. *Progress in Human Geography*, *45*(3), 566–576. https://doi.org/10.1177/0309132520952865

120. York, J. G., & Lenox, M. J. (2014). Exploring the sociocultural determinants of de novo versus de alio entry in emerging industries. *Strategic Management Journal*, *35*(13), 1930–1951. https://doi.org/10.1002/smj.2187

121. Zavolokina, L., Dolata, M., & Schwabe, G. (2016). The FinTech phenomenon: Antecedents of financial innovation perceived by the popular press. *Financial Innovation*, *2*(1), Article 1. https://doi.org/10.1186/s40854-016-0036-7

122. Zhou, Z.-H. (2021). *Machine Learning*. Springer Nature.