



Proef ingediend met het oog op het behalen van de graad van
Master of Science in Toegepaste Economische Wetenschappen

Machine-learningapplicaties voor overnames en fusies: startupwaardering met XGBoost en neural networks

Seppe HOUSEN

0525444

Academiejaar 2021-2022

Promotor: Sam VERBOVEN

Jury: Dirk DE SLEGTE

Sociale Wetenschappen & Solvay Business School

VERKLARING VAN AUTHENTICITEIT

De ondertekende verklaring van authenticiteit is een integrale component van het geschreven werk (Bachelorproef of Masterproef) dat wordt ingediend door de student.

Met mijn handtekening verklaar ik dat:

- ik de enige auteur ben van het ingesloten geschreven werk¹;
- ik dit werk in eigen woorden heb geschreven;
- ik geen plagiaat heb gepleegd zoals gedefinieerd in artikel 118 van het Onderwijs- en Examenreglement van de VUB; waarbij de meest voorkomende vormen van plagiaat zijn (niet-limitatieve lijst):
 - aard 1: tekst overnemen van andere auteurs, weliswaar met bronvermelding maar zonder gebruik van aanhalingstekens waar het om een letterlijke overname gaat;
 - aard 2: tekstfragmenten overnemen van andere auteurs, al dan niet letterlijk, zonder bronvermelding;
 - aard 3: verwijzen naar primair bronmateriaal waar de tekst en bronvermelding al dan niet letterlijk wordt overgenomen uit niet-vermelde secundaire bronnen;
 - aard 4: tekstfragmenten overnemen van andere auteurs, al dan niet met bronvermelding, met geringe en/of misleidende tekstaanpassingen.
- ik in de tekst en in de referentielijst volledig heb gerefereerd naar alle internetbronnen, gepubliceerde of ongepubliceerde teksten die ik heb gebruikt of waaruit ik heb geciteerd;
- ik duidelijk alle tekst heb aangeduid die letterlijk is geciteerd;
- ik alle methoden, data en procedures waarheidsgetrouw heb gedocumenteerd;
- ik geen data heb gemanipuleerd;
- ik alle personen en organisaties heb vermeld die dit werk hebben gefaciliteerd, dus alle ingediende werk ter evaluatie is mijn eigen werk dat zonder hulp werd uitgevoerd tenzij uitdrukkelijk anders vermeld;
- dit werk noch een deel van dit werk werd ingediend aan een andere instelling, universiteit of programma;
- ik op de hoogte ben dat dit werk zal gescreend worden op plagiaat;
- ik alle origineel onderzoeksmateriaal onmiddellijk zal indienen op het Decanaat wanneer hierom wordt gevraagd;
- ik op de hoogte ben dat het mijn verantwoordelijkheid is om na te gaan dat ik word opgeroepen voor een hoorzitting en tijdens de periode van hoorzittingen beschikbaar te zijn;
- ik kennis genomen heb van artikel 118 van het Onderwijs- en Examenreglement van de VUB omtrent onregelmatigheden en dat ik op de hoogte ben van de disciplinaire sancties;
- de afgedrukte kopie die ik indienende identiek is aan de digitale kopie die ik oplaadde op Turnitin.

Student familienaam, voornaam: Moussen Seppe

Datum: 08/08/2022

Handtekening:

¹ Voor groepswerken zijn de namen van alle auteurs verplicht. Hun handtekeningen staan collectief borg voor de volledige inhoud van het geschreven werk.

Abstract

Recent waren er verschillende overnames binnen de game-industrie. Deze kaderden veelal in de strijd om de meta-verse. Ook in vele andere sectoren zijn *mergers and acquisitions* (M&A) belangrijk voor de groei van de ondernemingen. De literatuurstudie in dit onderzoek onderscheidt drie domeinen waarin *machine learning* kan bijdragen in het M&A-proces. Zo kan machine learning ingezet worden om het voltrekken van M&A te voorspellen, te voorspellen wat de voordelen zijn van M&A of om de waardering van een onderneming te voorspellen. In dit onderzoek worden verschillende machine-learningmodellen vergeleken voor het waarderen van startups binnen de M&A-context. XGBoost en *neural networks* bewijzen hierbij een verbetering te zijn ten opzichte van eenvoudigere modellen zoals lineaire regressie.

Kernwoorden: machine learning (ML), mergers and acquisitions (M&A), valuation, XGBoost, neural networks

Dankwoord

In dit dankwoord wens ik graag de tijd te nemen om enkele mensen te bedanken, wiens steun onmisbaar was tijdens het schrijven van deze thesis.

Eerst en vooral wil ik mijn promotor prof. dr. Verboven bedanken. Hij heeft het onderzoek over dit boeiend onderwerp mogelijk gemaakt. Daarnaast zorgde zijn advies voor een completer werkstuk van hogere kwaliteit.

Verder wens ik ook mijn familie en vrienden te bedanken. Zij ondersteunde mij tijdens dit schrijfproces, toonde interesse en wisten op tijd en stond voor de nodige afleiding te zorgen.

Seppe

Inhoudsopgaven

| | |
|---|----|
| Abstract | i |
| Dankwoord | ii |
| 1. Inleiding | 1 |
| 2. Literatuurstudie | 3 |
| 2.1. Mergers and acquisitions | 3 |
| 2.1.1. Neo-klassieke verklaring..... | 3 |
| 2.1.2. Behavioral-verklaring | 3 |
| 2.1.3. Theory of the firm..... | 4 |
| 2.2. Machine learning..... | 4 |
| 2.2.1. Wat is machine learning? | 5 |
| 2.2.2. Types van machine learning | 6 |
| 2.2.3. Waarom machine learning gebruiken..... | 6 |
| 2.2.4. Uitdagingen voor machine learning | 7 |
| 2.2.5. Machine-learningalgoritmes | 8 |
| 2.2.6. Hyperparameteroptimalisatie | 15 |
| 2.2.7. Modelinterpretatie..... | 15 |
| 2.3. Machine learning en mergers and acquisitions | 16 |
| 2.3.1. M&A-voltooiing | 17 |
| 2.3.2. M&A-prestatie..... | 17 |
| 2.3.3. M&A-waardering..... | 18 |
| 2.3.4. Variatoepassingen | 19 |
| 2.4. Onderzoekskloof..... | 19 |
| 3. Methoden..... | 21 |
| 3.1. Dataset | 21 |
| 3.2. Feature engineering | 22 |
| 3.2.1. Aanvullen van ontbrekende data | 22 |
| 3.2.2. Variabelen aanpassen | 22 |

| | | |
|--------|--|----|
| 3.2.3. | Nieuwe variabelen..... | 23 |
| 3.2.4. | Rijen verwijderen..... | 24 |
| 3.3. | Verkennde analyse | 24 |
| 3.3.1. | Dealspecifieke informatie..... | 25 |
| 3.3.2. | Acquirer | 25 |
| 3.3.3. | Acquiree..... | 27 |
| 3.4. | Hyperparameter tuning..... | 30 |
| 4. | Resultaten..... | 32 |
| 4.1. | Voorspellingsfoutverdelingen | 33 |
| 4.2. | Verklaring van de voorspellingen | 35 |
| 5. | Discussie | 42 |
| 5.1. | Betekenis van de voorspelling..... | 42 |
| 5.2. | Link met de literatuur..... | 43 |
| 5.3. | Vergelijking met eerder onderzoek..... | 44 |
| 5.4. | Limitaties van het onderzoek | 45 |
| 5.5. | Toekomstig onderzoek | 46 |
| 6. | Conclusie | 47 |
| 7. | Bronnenlijst | 49 |
| 8. | Appendix..... | 56 |
| 8.1. | Literatuur studie..... | 56 |
| 8.2. | Dataset | 60 |
| 8.3. | Feature engineering | 62 |
| 8.4. | Verkennde analyse | 63 |
| 8.5. | Hyperparameter tuning..... | 65 |
| 8.5.1. | SVR..... | 65 |
| 8.5.2. | XGBoost | 66 |
| 8.5.3. | Neural network..... | 66 |
| 8.6. | Resultaten..... | 67 |

1. Inleiding

Recentelijk waren er enkele grote overnames binnen de videogamesindustrie. Zo kocht Microsoft de gameontwikkelaar Activision Blizzard voor zestig miljard euro. Enkele weken later maakte ook Sony bekend dat het zijn gamestudio's verder uitbreidt, door het opkopen van Bungie. Volgens de pers zouden deze overnames onderdeel zijn van de strategische strijd rond *the metaverse*. Zo hopen grote technologiebedrijven dat deze gamestudio's de nodige content kunnen genereren die nodig is om de nieuwe virtuele wereld interessant te maken (Schillewaert, 2022).

Deze opvatting van overnames als strategische groeistrategie wordt ook erkend door de wetenschappelijke literatuur. Zo geven Lageranna en Crawford (2018) aan dat overnames en fusies steeds vaker ingezet worden om de activa van een bedrijf uit te breiden, om nieuwe markten te betreden of om complementaire competenties te bekomen met als doel sterker te staan in de concurrentie met andere ondernemingen. Echter, de drijfveren blijken sterk te verschillen per sector. Men moet dus eveneens alert blijven voor de sectorspecifieke redenen voor fusies en overnames (Özgür & Wirl, 2020).

Eveneens is het niet verbazingwekkend dat twee overnames (in dezelfde sector) elkaar snel opvolgen. Cretin et al. (2015) geven aan dat de overname-activiteiten een sterke cyclische aard kennen. Er zijn periodes waarin vele en grote fusies en overnames plaatsvinden, en andere periodes waarin dit volume sterk afneemt. Historisch gezien kunnen zes verschillende golven worden waargenomen (Cretin et al., 2015). Meer recent, tijdens de coronacrisis, was het aantal fusies en overnames lager in aantal en in waarde dan in de jaren 2014 tot 2019 (IMAA, 2021).

Fusies en overnames, ook wel *mergers and acquisitions* (M&A) genoemd, zijn complexe processen (Trichterborn et al., 2016) waarin een bedrijf gekocht, verkocht, verdeeld of gecombineerd wordt met als doel het stimuleren van de groei van de onderneming (Xiang et al., 2012). De termen *merger* en *acquisition* worden steeds vaker als synoniemen gebruikt. Toch zijn er enkele verschillen. Bij een *merger* spreken twee bedrijven af om samen één onderneming te vormen. Bijgevolg verliezen beide bedrijven hun onafhankelijkheid. Bij een *acquisition* koopt één bedrijf een ander bedrijf op. Het kopende bedrijf wordt hierbij de eigenaar van het gekochte bedrijf (Hassan et al., 2018).

K. Lee et al. (2020) geven aan dat er verschillende actoren betrokken zijn bij een M&A-proces. Ten eerste is er de betrokkenheid van de verwervende onderneming, ook wel de *acquirer* genoemd. Dit is de onderneming die een ander bedrijf opkoopt. Daarnaast is ook het doelwit of de *acquiree* van belang. Een ander bedrijf tracht deze onderneming op te kopen. Ten slotte is er ook het belang van investeerders, in het bijzonder *merger arbitrageurs*. Deze personen proberen winst te maken door het

aankopen van aandelen van *target firms* om deze vervolgens tegen een hogere prijs te verkopen aan de acquirer. K. Lee et al. (2020) merken op dat deze drie groepen voordeel kunnen halen uit het correct voorspellen of een M&A al dan niet succesvol wordt afgerond. De mogelijkheden van *predictive analytics* in M&A worden eveneens onderstreept door Fanning en Drogts (2014).

Machine learning en *deep learning* worden heden ten dage al snel vooruitgeschoven als het goudhaantje van de predictive analytics. Zo worden deze technieken in Big Tech-ondernemingen, startups en zelfs in academisch onderzoek toegepast (J. H. Lee et al., 2018). J. H. Lee et al. (2018) gaan zelfs zo ver dat ze de impact van machine learning vergelijken met de impact die internet gehad heeft op de samenleving.

De ideeën van machine learning zijn niet per se nieuw. Onderzoek naar machine learning startte in de jaren vijftig en zestig van de vorige eeuw (Alzubi et al., 2018). Toch is de populariteit van machine learning de afgelopen jaren gestegen. Enkele memorabele momenten zijn de Deep Blue-computer van IBM die in 1997 de wereldkampioen schaken, Garry Kasparov, versloeg en het DeepFace-algoritme van Facebook dat erin slaagt om menselijke gezichten te herkennen in foto's (Alzubi et al., 2018). Deze recente ervaring van machine learning heeft te maken met de technologische vooruitgang. Zo heeft een verbetering in hardwaremogelijkheden ertoe geleid dat er meer data en meer rekenkracht beschikbaar zijn om geavanceerde modellen te trainen (J. H. Lee et al., 2018).

Toch zijn er ook bedenkingen bij de populariteit van machine learning. J. H. Lee et al. (2018) geven aan dat de aandacht voor machine learning momenteel mogelijk overdreven is. Ze wijzen hierbij op het ontbreken van recente conceptuele doorbraken op het vlak van machine learning en het feit dat het domein vooral ideeën verfijnt die al bestonden in de jaren zeventig en tachtig. Daarnaast geven ze ook aan dat de vaak gebruikte analogie met het menselijk brein een te grote vereenvoudiging is. Ten slotte waarschuwen ze ook voor het toepassen van machine learning zonder theoretische achtergrond. Greene et al. (2019) voegen daaraan toe dat er een sterke bezorgdheid bestaat over hoe machine learning kan worden gebruikt op een ethische manier. Naïef gebruik van machine learning moet dus vermeden worden. Desalniettemin zijn er verschillende aanwijzingen dat machine learning een grote impact kan hebben op het domein van M&A.

In deze thesis wordt onderzocht hoe machine learning (ML) ingezet kan worden in het domein van mergers and acquisitions (M&A). In het volgende deel van de thesis wordt de literatuur rond M&A en machine learning besproken. In het derde hoofdstuk wordt de methodologie besproken voor het trainen van een machine-learningmodel dat de dealprijs voorspelt. Hoofdstuk vier bevat de resultaten. Ten slotte wordt afgerond met een discussie en een conclusie.

2. Literatuurstudie

In de literatuurstudie wordt dieper ingegaan op het onderzoek dat reeds uitgevoerd is rond M&A en machine learning. In het eerste deel wordt het fenomeen van mergers and acquisitions toegelicht. Concreet wordt hierbij de vraag gesteld waarom M&A plaatsvindt. Daarnaast wordt ingegaan op het concept van machine learning. Ten slotte wordt verder uitgeweid over het gebruik van predictieve modellen in M&A. Hierbij wordt nagegaan hoe machine learning gebruikt kan worden om M&A te voorspellen.

2.1. Mergers and acquisitions

In deze sectie van de literatuurstudie staat het waarom van M&A centraal. Harford (2005) deelt de verklaringen voor M&A op in twee groepen: 'neo-klassiek' en *behavioral*. De neo-klassieke theorie verklaart M&A aan de hand van industrieschokken die de markt herorganiseren. De behavioral-theorie gaat ervan uit dat managers de overgewaardeerde aandelen van hun onderneming gebruiken om andere ondernemingen op te kopen. Desalniettemin maken Rhodes-Kropf en Robinson (2008) gebruik van de *theory of the firm* om mergers and acquisitions te verklaren. Deze theorie veronderstelt dat M&A het gevolg is van bedrijven die op zoek gaan naar synergiën. Deze drie theorieën worden hieronder kort toegelicht.

2.1.1. Neo-klassieke verklaring

De neo-klassieke verklaring van M&A gaat ervan uit dat schokken in de markt zorgen voor het herorganiseren van de activa in die markt (Harford, 2005). Hierbij kopen grote bedrijven met efficiënt management kleinere bedrijven met minder efficiënt management op. Dit zal ervoor zorgen dat deze activa efficiënter zullen worden ingezet en dus tot gevolg hebben dat de productiviteit verhoogd zal worden (Maksimovic & Phillips, 2001). Voor het opkopen van deze minder efficiënte bedrijven zijn voldoende liquide middelen nodig. De beschikbaarheid van liquide kapitaal heeft een cyclisch karakter en draagt op die manier bij aan het verklaren van het cyclisch karakter van mergers and acquisitions (Harford, 2005). Volgens deze theorie kan de markt van mergers and acquisitions gezien worden als een competitie tussen managementteams die strijden voor het recht om een bedrijf te mogen leiden (Li et al., 2018).

2.1.2. Behavioral-verklaring

Volgens de behavioral-verklaring kunnen mergers and acquisitions gezien worden als het gedrag van managers die wensen een 'imperium' uit te bouwen (Maksimovic & Phillips, 2001). Ze maken daarbij gebruik van momenten waarop de aandelen van hun bedrijf overgewaardeerd zijn om ondergewaardeerde spelers op te kopen. Deze ondergewaardeerde bedrijven gaan vervolgens in op

het voorstel, omdat ze enkel oog hebben voor de kortetermijnwinsten of een overschatting maken van de synergie tussen de twee bedrijven (Harford, 2005).

2.1.3. Theory of the firm

Rhodes-Kropf en Robinson (2008) maken gebruik van de theory of the firm om M&A te verklaren. Zij postuleren dat bedrijven op zoek gaan naar activa die een synergie opleveren met hun eigen bedrijf. Deze synergie kan enkel behaald worden als de activa toebehoren aan dezelfde onderneming. Hoe beter de activa samengaan, hoe groter de gecreëerde meerwaarde van de M&A-deal. Hoe deze meerwaarde verdeeld wordt tussen de acquirer en de *target* ligt aan de onderhandelingspositie van beide bedrijven. Deze onderhandelingspositie wordt bepaald door de schaarsheid van de activa. Hierdoor ontstaat een wisselwerking tussen de gecreëerde meerwaarde en de onderhandelingspositie. Het uiteindelijke deel van de gecreëerde meerwaarde bepaalt de waarde van de M&A-deal voor elke partner. Dit leidt ertoe dat bedrijven van ongeveer dezelfde grootte met elkaar fuseren, in plaats van dat grote bedrijven kleine bedrijven opkopen. Rhodes-Kropf en Robinson (2008) geven aan dat *like buys like* een meer toepasselijke omschrijving is van M&A dan *high buys low*.

De theory of the firm is nauw verwant aan de neo-klassieke verklaring van M&A, maar verschilt in het feit dat de theory of the firm een onderscheid maakt tussen supplementaire M&A en complementaire M&A. Wanneer men uit is op het aankopen van supplementaire activa om deze efficiënter in te zetten, zullen de verschillen tussen de acquirer en target groter zijn. Wanneer men uit is op de aankoop van complementaire activa, waarbij het doel is om een synergie te creëren met de huidige activa, dan zullen de verschillen tussen acquirer en target kleiner zijn (Rhodes-Kropf & Robinson, 2008). Bij de neo-klassieke theorie gaat men ervan uit dat de M&A altijd supplementair is. Een groot bedrijf met voldoende liquide middelen koopt minder efficiënte activa van andere bedrijven op.

2.2. Machine learning

Emerson et al. (2019) geven aan dat het gebruik van machine-learningtechnieken en toepassingen in het financieel onderzoeksdomein de laatste jaren sterk is toegenomen. Deze sterke groei die machine learning binnen finance meemaakt, schrijven zij toe aan twee evoluties. Ten eerste geven ze aan dat de kracht van computers is toegenomen. Dit zorgt ervoor dat meer complexe taken uitgevoerd kunnen worden. Daarnaast heeft ook de ontwikkeling van eenvoudiger programmeertalen, zoals Python en R, machine learning meer toegankelijk gemaakt.

Hieronder wordt dieper ingegaan op machine learning. Het doel is om een beter begrip te krijgen van wat machine learning precies is, welke typen ML-systemen er bestaan, waarom machine learning wordt gebruikt, en wat de grootste uitdagingen zijn bij het gebruik van machine-learningtechnieken. Daarnaast worden ook enkele machine-learningalgoritmen besproken en wordt dieper ingegaan op

het optimaliseren van deze algoritmen. Ten slotte wordt ook de verklaarbaarheid van deze modellen besproken.

2.2.1. Wat is machine learning?

Machine learning kan gezien worden als de intersectie van computerwetenschappen en statistieken (Mitchell, 2006). Het onderzoeksdomein van computerwetenschappen probeert een antwoord te vinden op de vraag hoe computers gebruikt kunnen worden om problemen op te lossen. Statistiek focust veelal op de manieren waarop conclusies kunnen worden getrokken op basis van data. Machine learning voegt deze twee aspecten samen. Hierbij gaat men na hoe computersystemen ontwikkeld kunnen worden die automatisch verbeteren naarmate ze meer ervaring opdoen (Mitchell, 2006). Emerson et al. (2019) verduidelijken dat deze ervaring veelal bestaat uit data. Binnen machine learning worden statistische methoden ingezet die computermodellen de mogelijkheid geven om te leren van een dataset. Dit moet ervoor zorgen dat deze computermodellen zelfstandig specifieke taken kunnen aanleren zonder dat ze hiervoor expliciet geprogrammeerd zijn (Géron, 2019).

Het gebruik van machine-learningtechnieken kan beschreven worden aan de hand van een iteratief proces (Provost & Fawcett, 2013). Ten eerste is het belangrijk dat duidelijk wordt geformuleerd welk probleem men probeert op te lossen. Ten tweede moeten de nodige data worden verzameld, die aan de basis liggen van de oplossing. Het is hierbij belangrijk dat de betrouwbaarheid van de data wordt nagegaan. Zo is vaak niet alle data even geschikt om een specifiek probleem op te lossen. Ten derde moet de data worden gepresenteerd in een vorm die de beste resultaten oplevert. Deze vorm verschilt vaak aanzienlijk van de vorm waarin de data verzameld werden. Hiervoor is datamanipulatie van belang. In een vierde stap wordt de data gepresenteerd aan het machine-learningalgoritme. Dit algoritme leert vervolgens de structuur van de data. Deze structuur wordt opgeslagen in een model. Het model kan vervolgens gebruikt worden om voorspellingen te maken op basis van nieuwe data. In een vijfde stap moet worden nagegaan of de voorspellingen van het model voldoen aan de voorwaarden om het probleem op te lossen. Als dit het geval is, kan men overgaan tot het in gebruik nemen van het model. Voldoet het model niet aan de verwachtingen, dan moet het probleem opnieuw worden geanalyseerd en begint het iteratief proces opnieuw.

Dit iteratief proces toont aan dat het gebruik van machine-learningtechnieken niet enkel gestuurd wordt door de computer, maar dat ook menselijke beslissingen van belang zijn. Het is daarom interessant om machine-learningtechnieken te visualiseren op een continuüm (Beam & Kohane, 2018). Aan de ene kant van dit spectrum zijn de technieken te zien die volledig machinegestuurd zijn, aan de andere kant staan de technieken die volledig mensgestuurd zijn. Beam en Kohane (2018) beschrijven dit continuüm als het machine-learning spectrum.

2.2.2. Types van machine learning

Volgens Dixon en Halperin (2019) kunnen de machine-learningtechnieken voor het trainen van een algoritme opgedeeld worden in *supervised machine learning*, *unsupervised machine learning* en *reinforcement learning*. Géron (2019) vermeldt ook nog de categorie *semisupervised learning*.

Bij supervised machine learning wordt een bepaalde output-variabele voorspeld aan de hand van een reeks input-variabelen (Emerson et al., 2019). Hierbij kan men proberen om een specifiek nummer te voorspellen. Dit wordt regressie genoemd. Daarnaast wordt gesproken van classificatie; dit is het indelen van observaties in vooraf bepaalde groepen.

Voor unsupervised learning is er geen gelabelde output-variabele. Het machine-learningalgoritme moet deze output zelf ontdekken vanuit de data (Géron, 2019). In deze groep zijn bijvoorbeeld clustering-technieken terug te vinden. Hierbij worden gelijkaardige observaties gegroepeerd (zonder dat vooraf een groep bepaald werd). Anderzijds kan men ook trachten om de informatie uit een dataset te capteren in een kleinere set van variabelen onder de noemer *dimensionality reduction* (Dixon & Halperin, 2019).

Semisupervised learning bestaat meestal uit een combinatie van supervised en unsupervised algoritmes. Deze algoritmes kunnen omgaan met data waarbij slechts een aantal observaties gelabelde output-variabelen hebben (Géron, 2019).

Reinforcement learning verschilt sterk van de hierboven beschreven technieken. Bij deze techniek worden algoritmes getraind door gebruik te maken van *trial and error* met een feedbackcomponent (Emerson et al., 2019). Zo kan het algoritme zijn omgeving waarnemen en is het in staat om acties te nemen binnen deze omgeving. Deze acties worden vervolgens beloond of afgestraft. Op deze manier ontwikkelt dit algoritme een strategie om zo veel mogelijk beloningen te krijgen. Dit uit zich in het aanleren van een actie in een bepaalde situatie (Géron, 2019).

2.2.3. Waarom machine learning gebruiken

Machine learning reikt nieuwe hulpmiddelen en technieken aan voor economische analyse die onderzoekers in staat moeten stellen om traditionele, maar ook nieuwe vragenstukken te beantwoorden (Mullainathan & Spiess, 2017). Hierbij is machine learning veelal geschikt voor predictieve taken, waarbij wordt getracht een bepaalde variabele te voorspellen. Machine learning is minder geschikt voor verklarende vraagstukken, waarbij het doel is om de invloed van een variabele te begrijpen (Mullainathan & Spiess, 2017). Desalniettemin zijn er de laatste jaren verschillende inspanningen gedaan om de verklaarbaarheid van machine-learningmodellen te vergroten (Lundberg et al., 2020; Lundberg & Lee, 2017). Mullainathan en Spiess (2017) onderscheiden vier groepen van

mogelijke toepassingen voor machine learning. Ten eerste kan machine learning ingezet worden om voorspellingen te doen op basis van nieuwe typen data. Zo zijn machine-learningalgoritmen in staat om grote hoeveelheden data en alternatieve datatypes zoals tekst en foto's te verwerken. Door gebruik te maken van machine-learningtechnieken kunnen ook deze alternatieve datatypes onderzocht worden, samen met de invloed die zij hebben binnen economische vraagstukken (Hansen & Borch, 2021). Ten tweede kan machine learning gebruikt worden als onderdeel van onderzoek naar verklarende variabelen. Zo hebben deze onderzoeken vaak eerst een voorspellende taak. Daarna kunnen de variabelen worden geïdentificeerd die van invloed zijn. Ten derde kan machine learning ook ondersteunen in beleidsbeslissingen. Voor het nemen van een beslissing is het van belang om te weten wat de impact zal zijn van een bepaalde beslissing. Machine learning kan gebruikt worden om de impact van deze beslissing te voorspellen. Ten vierde kan machine learning ingezet worden om theorieën te toetsen die inherent over voorspelbaarheid gaan. Zo kan machine learning worden gebruikt om na te gaan of bepaalde gebeurtenissen voorspelbaar zijn. Anderzijds kan een machine-learningmodel dienen als vergelijkingspunt om te zien hoe goed een theorie een gebeurtenis verklaart.

2.2.4. Uitdagingen voor machine learning

Ondanks dat machine learning vele mogelijkheden biedt binnen economisch en financieel onderzoek, en dat het steeds eenvoudiger wordt om deze technieken toe te passen, is het van belang om alert te blijven dat deze technieken niet naïef toegepast worden (Mullainathan & Spiess, 2017). Zo moeten onderzoekers ervoor waken dat ze voldoende kwaliteitsvolle data koppelen aan een geschikt algoritme om een resultaat te behalen dat het probleem kan oplossen. Een machine-learningmodel zal niet goed werken als het getraind is op een kleine, niet-representatieve dataset met veel irrelevante variabelen (Géron, 2019). Daarnaast moet het machine-learningmodel ook de juiste complexiteit hebben om de structuren in de data correct te beschrijven (Géron, 2019). Dixon en Halperin (2019) identificeren vier vaak voorkomende problemen in verband met de adoptie van machine-learningtechnieken in finance. Ten eerste kan het co-integreren van machine learning met bestaande financiële analyse ontbreken. Dit leidt tot modellen waarvan de uitkomsten niet verklaard of verdedigd kunnen worden. Ten tweede kan men er blind op vertrouwen dat de toekomst zal verlopen zoals het verleden. Dit leidt tot het ontbreken van statistische tests en het gebruik van de verkeerde modellen. Bijgevolg kunnen enkel suboptimale resultaten behaald worden. Ten derde kan een te grote focus ontstaan op het model, waardoor de bredere context van een concurrerende markt vergeten wordt. Een model met uitsluitend een hoge accuraatheid is niet voldoende. Zo moeten modellen ook snel en efficiënt werken en gemakkelijk te onderhouden zijn. Ten vierde kan er een terughoudendheid zijn voor het accepteren van nieuwe ontdekkingen gemaakt door machine-learningalgoritmen. Hierbij worden de uitkomsten

van het model enkel aanvaard wanneer ze voldoen aan de verwachtingen en opvattingen van de gebruiker.

2.2.5. Machine-learningalgoritmes

Aan de basis van machine learning liggen de algoritmes die leren op basis van data. Voor supervised machine learning, en dus ook voor de toepassing die in dit onderzoek centraal staat, leren deze algoritmes een functie die de inputvariabelen afbeeldt op de outputvariabele. Dit moet ertoe leiden dat van de inputgegevens de output kan worden voorspeld. Vier algoritmes die veel gebruikt worden in economisch en financieel onderzoek zijn: lineaire regressie, support-vectorregressie, XGBoost en *neural networks*. Deze algoritmen worden hieronder besproken.

2.2.5.1. Lineaire regressie

Lineaire regressie is een zeer bekende statistische techniek die reeds toegepast werd in een brede waaier aan domeinen. Het basisidee van lineaire regressie is het zoeken naar een lijn die zo goed mogelijk de data beschrijft. Deze redenering kan echter ook uitgebreid worden naar het beschrijven van de data aan de hand van een hypervlak. Op deze manier kan een lineaire regressie een voorspelling maken van de outputvariabele op basis van een gewogen som van de inputvariabelen plus een constante term (Géron, 2019). Het model kan als volgt beschreven worden:

$$y = b_0 + b_1x_1 + b_2x_2 + \dots + b_nx_n$$

Waarbij:

- y de voorspelde waarde aangeeft;
- n het aantal variabelen is;
- x_i de waarde is van de variabele met volgnummer i ;
- b_j de factor is met volgnummer j ; deze wordt door het model bepaald.

Een lineaire regressie kan getraind worden door het minimaliseren van de gemiddelde kwadratische afstand (MSE) (Géron, 2019). Dit is de gemiddelde gekwadrateerde afstand tussen de voorspelde waarden en de geobserveerde waarden.

Het lineaire-regressiemodel heeft enkele sterktes en zwaktes. De voornaamste sterkte is dat het model eenvoudig is. Hierdoor kan worden ingeschat hoe sterk de invloed van de verschillende inputvariabelen is op de voorspelling. Zo zullen variabelen die een groter/kleiner gewicht toegekend krijgen een grotere/kleinere invloed hebben op de voorspelling. Daarnaast is het model ook relatief eenvoudig te berekenen en schaalte het efficiënt met het aantal observaties in de dataset (Géron,

2019). Daartegenover staat dat lineaire regressie minder efficiënt kan omgaan met datasets die veel variabelen bevatten (Géron, 2019). Een tweede minpunt is het feit dat het model eenvoudig is. Dit zorgt ervoor dat het minder goed complexe patronen kan beschrijven.

Desalniettemin is een lineaire regressiemodel ook gebruikt worden om meer complexe data te modelleren. Zo kunnen ook interactie-effecten tussen variabelen en niet-lineaire data gemodelleerd worden. Dit wordt gedaan door aan de variabelen respectievelijk producten en machten van de oorspronkelijke variabelen toe te voegen (Géron, 2019).

2.2.5.2. Support-vectorregressie

Een *support vector machine* (SVM) is een sterk en veelzijdig machine-learningalgoritme dat zowel in staat is om regressie- als classificatieproblemen op te lossen (Géron, 2019). In zijn regressievariant, ook wel support-vectorregressie (SVR) genoemd, tracht het algoritme om een lijn met marge (ook wel een straat) te trekken over de datapunten, zodat zoveel mogelijk datapunten binnen de marge liggen en er zo min mogelijk punten buiten de marge liggen. De marge is een belangrijke (hyper-)parameter van het model. Deze parameter moet de onderzoeker zelf bepalen en wordt niet getraind door het model.

Het trainen van een support-vectorregressie kan gezien worden als het minimaliseren van zowel de marge en de *predictive error* (Awad & Khanna, 2015). Zo kan het optimalisatieprobleem geschreven worden als:

$$\text{minimaliseer } \frac{1}{2} \|w\|^2$$

Waarbij w de factoren (ook wel gewichten) van de variabelen aangeeft en alle voorspellingen binnen de marge liggen.

Het is echter niet altijd mogelijk om deze lijn met marge te bepalen zodat alle datapunten binnen deze marge liggen. In sommige gevallen is het wenselijk om toe te laten dat enkele punten buiten de marge vallen. Hiervoor kunnen zogenaamde slack-variabelen worden gedefinieerd (Smola & Schölkopf, 2004). Dit leidt tot het trainen van het SVR-model aan de hand van volgende functie:

$$\text{minimaliseer } \frac{1}{2} \|w\|^2 + C \sum (\xi + \xi^*)$$

Hier valt op te merken dat het eerste deel van de functie hetzelfde is. Het doel is nog steeds om de factoren bij de variabelen te minimaliseren. Dit kan ook worden geïnterpreteerd als het zoeken van de meest vlakke lijn (met marge). Daarnaast wordt toegelaten dat bepaalde datapunten buiten de marge liggen aan de hand van de variabelen ξ en ξ^* . Deze variabelen bepalen de strafscore voor datapunten

buiten de marge. De variabele C moet zelf gekozen worden en bepaalt de verhouding tussen de 'vlakheid' van de lijn en het aantal datapunten buiten de marge.

Daarnaast kan de SVR uitgebreid worden om ook niet-lineaire data te modelleren. Hiervoor kan een *kernel trick* worden gebruikt. Hierbij worden de data getransformeerd naar een hogere dimensionale voorstelling (Awad & Khanna, 2015; Smola & Schölkopf, 2004). Op deze manier kan ook niet-lineaire data gemodelleerd worden met SVR.

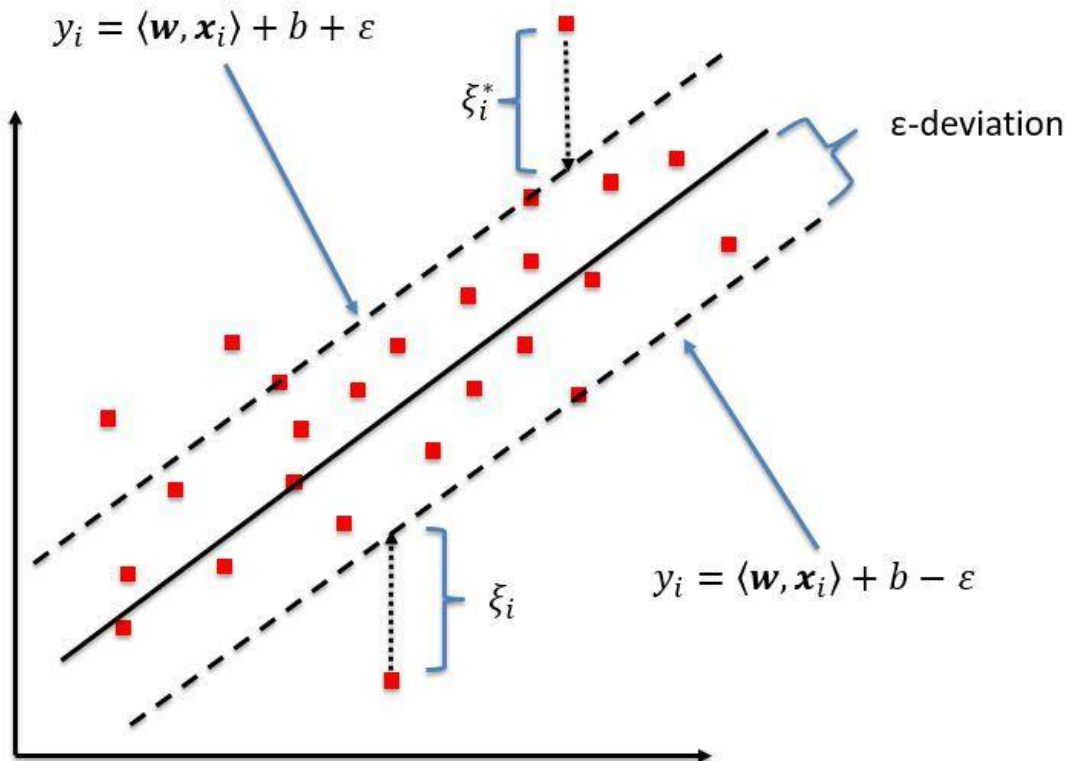


Figure 1: Visual representation of SVR (Kleynhans et al., 2017)

2.2.5.3. XGBoost

XGBoost of *extreme gradient boosting* is een uitbreiding op het decision-tree-algoritme (DT). *Decision trees* proberen de dataset te verdelen in homogene groepen. Om dit te bereiken, zoekt het algoritme naar de splitsing van de dataset die de MSE minimaliseert (Géron, 2019). De MSE wordt berekend aan de hand van de gewogen som van de MSE van elk deel van de splitsing. De gewichten van elk deel worden bepaald op basis van de proportie van de observaties in elk deel (Géron, 2019). Zo kan een splitsing bepaald worden aan de hand van volgende formule:

$$\text{minimaliseer } \frac{m_a}{m} \text{MSE}_a + \frac{m_b}{m} \text{MSE}_b$$

Waarbij:

- a en b de delen van de splitsing aangeven;
- m staat voor het aantal observaties in de dataset (of per deel, indien aangegeven met een subscript).

Het decision-tree-algoritme blijft splitsingen maken totdat het een stop-criteria bereikt of totdat alle observaties eindigen in een unieke groep (Géron, 2019). Deze splitsingen kunnen vervolgens in een boomstructuur gegoten worden, zoals voorgesteld in figuur 2. Een nieuwe waarde kan voorspeld worden door deze splitsingsstructuur te volgen. De voorspelde waarde is het gemiddelde van de geobserveerde groep waarvan de nieuwe waarde deel uit maakt (Géron, 2019).

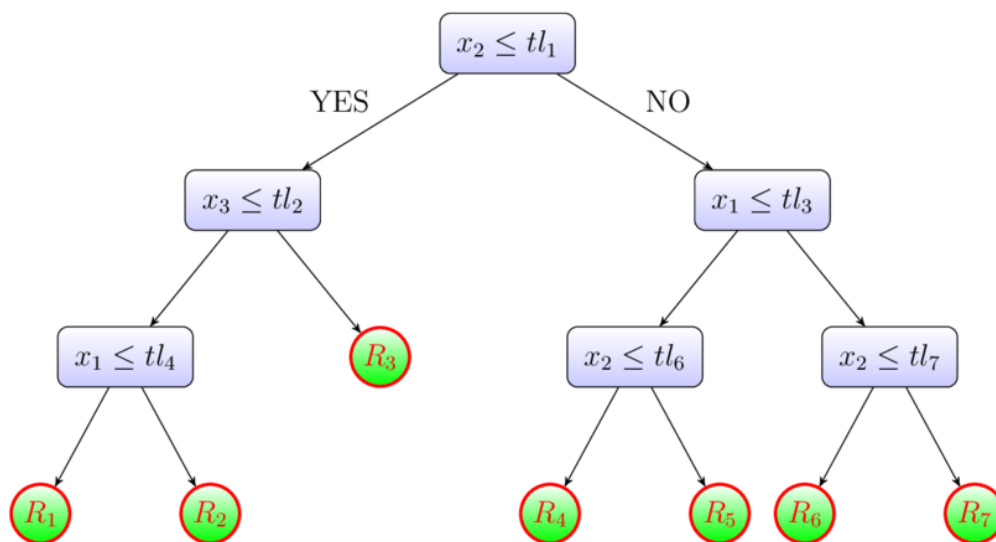


Figure 2: Visual representation of a decision tree model (Akbari et al., 2021)

Het grote voordeel van decision-treemodellen is dat ze eenvoudig te begrijpen zijn, grafisch weergegeven kunnen worden en gemakkelijk geïnterpreteerd kunnen worden (James et al., 2013). Daar staat echter wel tegenover dat decision trees meestal niet dezelfde predictieve accuraatheid behalen als andere machine-learningalgoritmes (James et al., 2013). Eveneens worden decision trees sterk beïnvloed door kleine veranderingen in de dataset. Hierbij kan worden gesproken van een hoge variantie van het algoritme (Olaru & Wehenkel, 2003). Evenwel kunnen deze negatieve aspecten verminderd worden door bagging- en boostingtechnieken toe te passen (Painsky & Rosset, 2017).

Bagging en *boosting* zijn voorbeelden van ensemble-methoden. Bij ensemble-methoden worden verschillende predictieve modellen afzonderlijk getraind. Vervolgens worden hun voorspellingen gegroepeerd om tot een gezamenlijke voorspelling te komen (Géron, 2019).

XGBoost is een algoritme dat gebruikmaakt van boostingtechniek. Meer specifiek maakt het gebruik van *gradient boosting*. Dit is een techniek waarbij decision-treemodellen sequentieel getraind worden. Het eerste DT-model probeert de output-variabelen te voorspellen. Het zal hierbij geen perfecte voorspelling maken en dus een voorspellingsfout maken. Vervolgens wordt een tweede DT-model getraind om deze voorspellingsfout in te schatten. De voorspelling van dit tweede DT-model wordt vervolgens gebruikt om de voorspelling van het eerste DT-model bij te sturen. Op deze manier worden verschillende decision-treemodellen getraind die telkens de voorspelling van de vorige reeks bijsturen (Géron, 2019). Al deze decision-treemodellen, die elkaars voorspelling verbeteren, vormen samen het XGBoost-model.

2.2.5.4. *Neural networks*

Een *artificial neural network* is een netwerk van neuronen, ook wel nodes genoemd. Een neuron ontvangt als input verschillende getallen en geeft als output één getal. Om deze output te genereren, berekent de neuron een gewogen gemiddelde van de inputs. Vervolgens wordt deze gewogen som getransformeerd door een activatiefunctie (Géron, 2019). Een voorbeeld van een dergelijke activatiefunctie is de *rectified linear unit* (ReLU). Deze functie geeft de oorspronkelijke waarde van de input terug als de input positief is, en nul in de andere gevallen.

Een neural network is opgebouwd uit verschillende *layers* die verschillende van deze nodes bevatten. Drie types van layers worden onderscheiden (Géron, 2019). De eerste layer is de *input layer*. Deze laag bevat de gegevens van de inputvariabelen. Vervolgens zijn er verschillende *hidden layers*. Deze lagen transformeren de inputgegevens. Ten slotte is er de *output layer*. Deze laag geeft een voorspelling op basis van de getransformeerde informatie uit de hidden layers.

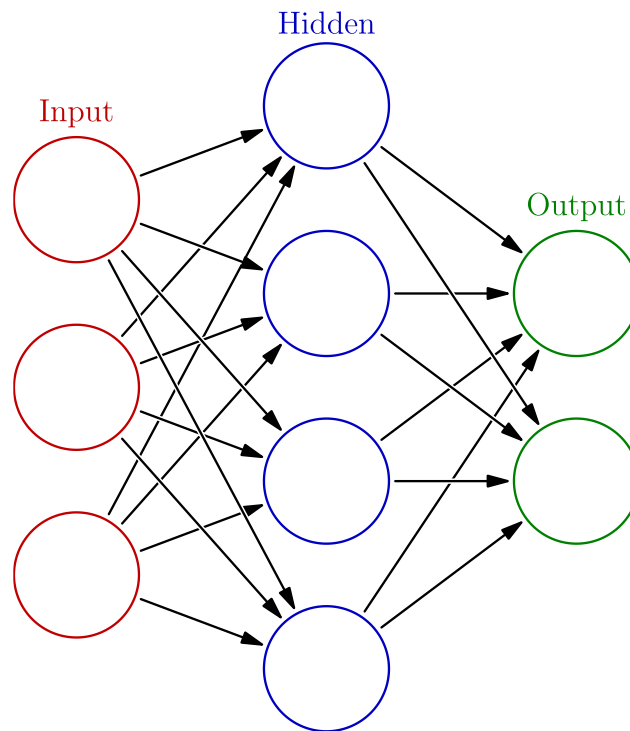


Figure 3: Visual representation of a neural network (Cburnett, 2006)

Een *neural network model* kan getraind worden door middel van *backpropagation* (Géron, 2019). Dit algoritme is gebaseerd op *gradient descent* en kan doorlopen worden in vier stappen (Géron, 2019; Ng et al., 2022). Voordat het model getraind kan worden, moeten de gewichten van de connecties gedigitaliseerd worden met gerandomiseerde waarden. Vervolgens kunnen volgende vier stappen herhaald worden tot het model getraind is.

- 1) **Forward propagation.** In de eerste stap wordt het neural network gebruikt om een voorspelling te maken. Op basis van de inputs wordt een output gegenereerd. Dit gebeurt door de inputs door te geven aan de hidden layer. De resultaten van deze hidden layer worden vervolgens doorgegeven aan de volgende hidden layer. Dit gaat door tot de output layer bereikt is. Dit resulteert in één voorspelling per case.
- 2) **Loss-functie berekenen.** Ten tweede moet de loss- of kostenfunctie berekend worden. Deze functie geeft aan hoe sterk de voorspelling afwijkt van de geobserveerde waarden.
- 3) **Backpropagation.** In de derde stap wordt de (partiële) afgeleide van de kostenfunctie berekend. Dit geeft inzicht in hoe sterk de kostenfunctie beïnvloed wordt door de verschillende connecties in het netwerk en in welke richting deze connectie een effect heeft. Deze berekening vertrekt van de voorspelling en werkt zo terug naar de input layer. Dit leidt tot de naam van het algoritme.

- 4) **Training step.** Ten vierde moet het model worden geüpdatet. Dit gebeurt door de connecties bij te stellen aan de hand van de hellingen die berekend werden in de vorige stap.

Voor een regressieprobleem kan ervoor gekozen worden om in het gehele netwerk de ReLU-activatiefuncties te gebruiken. Dit zorgt ervoor dat de voorspelde output altijd positief is. Daarnaast kan voor de kostenfunctie worden geopteerd om de gemiddelde kwadratische afstand (MSE) of de gemiddelde absolute afstand (MAE) te gebruiken (Géron, 2019).

2.2.5.4.1. Hyperparameters

Neurale netwerken hebben een grote flexibiliteit. Hierdoor moeten verschillende keuzes gemaakt worden tijdens het ontwerp van deze netwerken (Géron, 2019). Deze keuzes worden hyperparameters genoemd. De hyperparameters van een neural network kunnen in twee grote groepen verdeeld worden. Enerzijds zijn er de hyperparameters die de training van het model beïnvloeden, anderzijds zijn er de hyperparameters die de architectuur van het model bepalen (Yu & Zhu, 2020). Deze laatste categorie kan volgens Liao et al. (2022) verder opgedeeld worden in hyperparameters omtrent de globale architectuur van het netwerk en hyperparameters die layer-specifiek zijn.

De belangrijkste hyperparameters die de architectuur van het neural network bepalen, zijn het aantal layers, het aantal neuronen per layer en de regularisatie van het model (Müller & Guido, 2016). Ten eerste moet men een keuze maken ten aanzien van het aantal hidden layers in het netwerk. Een hoger aantal layers zorgt ervoor dat het netwerk efficiënt complexere structuren kan aanleren. Dit kan leiden tot betere voorspellingen. Anderzijds verhoogt het de kans op overfitting (Benardos & Vosniakos, 2007). Naast het bepalen van het aantal hidden layers, moet men ook bepalen hoeveel neuronen elke laag zal bevatten. Voor de input- en outputlayer ligt dit aantal vast en is dit gelijk aan respectievelijk het aantal inputvariabelen en het aantal gewenste outputvariabelen. Voor de hidden layers kan het aantal neuronen vrij gekozen worden. Het aantal neuronen van de hidden layers ligt hierbij meestal tussen het aantal van de input layer en het aantal van de output layer (Yu & Zhu, 2020). Ten slotte kan men ook opteren om regularisatie toe te voegen aan het netwerk. Dit is een extra strafscore voor de complexiteit van het model. Het toevoegen van regularisatie beperkt de kans op overfitting (Géron, 2019).

Voor het trainingsproces van het neural network zijn vooral de *learning rate* en de *optimizer* van belang (Yu & Zhu, 2020). De learning rate bepaalt de grootte van elke trainingsstap in het backpropagation-algoritme (Liao et al., 2022). Een hogere learning rate betekent dat het model sneller leert. Evenwel is het mogelijk dat deze grote stappen de optimale waarde voorbijschieten. Hierdoor opteert men in sommige gevallen om de learning rate te verlagen naarmate de training vordert (Yu & Zhu, 2020). Ten tweede zijn er verschillende methoden van *optimization*. Dit is het aanpassen van de gewichten van

de connecties tussen de neuronen in de *training step* (Liao et al., 2022). Zo is het mogelijk om rekening te houden met deze informatie uit voorgaande iteraties. Dit kan leiden tot snellere optimalisatie (Géron, 2019).

2.2.6. Hyperparameteroptimalisatie

Een eenduidige, optimale combinatie van hyperparameters die voor alle toepassingen de beste resultaten oplevert, bestaat niet (Victoria & Maragatham, 2021). Hierdoor moet op zoek worden gegaan naar de hyperparameters die voor het huidige vraagstuk het beste model opleveren. Dit wordt hyperparameteroptimalisatie genoemd. Deze optimalisatie is niet voor de hand liggend, omdat er bijzonder complexe interactie-effecten bestaan tussen deze hyperparameters met betrekking tot hun invloed op de uiteindelijke prestatie van het model (Breuel, 2015).

Een eenvoudige manier om een optimale combinatie van hyperparameters te vinden, is door middel van *grid search*. Bij deze techniek worden een aantal mogelijkheden voor verschillende hyperparameters gedefinieerd. Vervolgens worden de combinaties van deze hyperparameters gebruikt om verschillende modellen te trainen en te scoren (Müller & Guido, 2016). Deze methoden vragen vaak zeer veel rekentijd. Daardoor wordt vaak geopteerd om een *random search* toe te passen. Hierbij worden een aantal willekeurige combinaties geselecteerd en vergeleken. Dit verlaagt de nodige rekentijd en resulteert in een oplossing die dicht bij de optimale oplossing ligt (Feurer & Hutter, 2019; Liao et al., 2022).

Daarnaast kan er ook voor gekozen worden om *Bayesian optimization* toe te passen. Deze methode beoogt het gericht zoeken naar veelbelovende hyperparametercombinaties. Hierbij slaagt de methode erin om de benodigde rekentijd verder te verlagen (Dernoncourt & Lee, 2017). Frazier (2018) geeft aan dat het algoritme uit twee componenten bestaat. Ten eerste wordt een kansverdeling ingeschat die voor elke hyperparametercombinatie een verwachte prestatie van het model genereert. Ten tweede kan worden bepaald welke combinatie van hyperparameters getest dient te worden door een afweging te maken tussen het verkennen van de kansverdeling en het optimaliseren van de prestaties van het model. De geselecteerde hyperparameters worden gebruikt om een model te trainen. De uitkomst van deze test kan vervolgens gebruikt worden om de kansverdeling te updaten en een nieuwe testcase te bepalen. Dit proces wordt herhaald tot een geschikte hyperparametercombinatie gevonden is.

2.2.7. Modelinterpretatie

Het kunnen verklaren van de voorspellingen van een machine-learningmodel is uitermate belangrijk. Zo kan namelijk worden begrepen hoe een voorspelling tot stand komt. Dit leidt tot meer vertrouwen in het algoritme, nieuwe inzichten om het model te verbeteren en een beter begrip van het fenomeen

dat voorspeld dient te worden (Lundberg & Lee, 2017). Recent onderzoek heeft gebruikgemaakt van SHAP (SHapley Additive exPlanation) om machine-learningmodellen te verklaren (Parsa et al., 2020). Deze techniek maakt het mogelijk om inzichten te verwerven over het belang van verschillende variabelen bij het maken van voorspellingen (Parsa et al., 2020).

SHAP maakt gebruik van de *Shapley value* uit de *game theory*. Deze waarde drukt de bijdrage uit die een speler had aan het eindresultaat van het spel. Voor het verklaren van machine-learningvoorspellingen kunnen we deze concepten toepassen waarbij de input-variabelen de spelers voorstellen en de voorspelling de uitkomst is van het spel (Rozemberczki et al., 2022).

Om deze Shapley values te berekenen, zijn verschillende methodes ontwikkeld. Voor tree-based machine-learningmodellen is een exacte berekening beschikbaar (Lundberg et al., 2020). Voor neural networks moet men zich baseren op schattingen van de exacte Shapley values (Lundberg & Lee, 2017). Deze SHAP-methode bouwt verder op de DeepLIFT-methode die gebruikt wordt om het belang van variabelen te schatten in een neural network (Shrikumar et al., 2017).

2.3. Machine learning en mergers and acquisitions

Xiang et al. (2012) delen het huidig onderzoek naar *M&A prediction* op in drie groepen. De eerste groep steunt vooral op het gebruik van financiële en managementgegevens om M&A te voorspellen. De tweede groep maakt gebruik van *data mining* en machine-learningtechnieken. De derde groep gaat dieper in op faillissementen en het falen van ondernemingen. Momenteel steunt veel van het onderzoek naar voorspellingen van M&A-succes op het gebruik van logit- en probitmodellen (Karatas & Hirs, 2021). Deze meer traditionele statistische modellen stoten op enkele beperkingen (K. Lee et al., 2020). Daarom is te zien dat recent onderzoek in M&A prediction veelal gebruikmaakt van machine-learningtechnieken zoals neural networks (Karatas & Hirs, 2021; K. Lee et al., 2020).

In dit onderdeel van de literatuurstudie ligt de focus op de onderzoeken waarbij M&A voorspeld wordt aan de hand van machine-learningtechnieken. Deze machine-learningmodellen, die M&A proberen te voorspellen, kunnen ondergebracht worden in vier groepen. Ten eerste zijn er modellen die proberen te voorspellen of de M&A-deal al dan niet succesvol afgerond wordt. De tweede groep van modellen voorspelt of een M&A-deal een positief eindresultaat oplevert. Ten derde zijn er modellen die proberen te voorspellen hoeveel zal worden betaald voor een bepaald bedrijf tijdens een M&A-deal. Ten vierde wordt ook nog een 'varia'-groep besproken. Hierin worden predictieve modellen besproken die niet rechtstreeks in één van de eerdere groepen passen. In appendix zijn een visuele voorstelling en een samenvatting te vinden van deze literatuur.

2.3.1. M&A-voltooiing

Binnen de groep van M&A-voltooiing, of *M&A completion*, kan verder onderscheid worden gemaakt tussen twee categorieën. Ten eerste is er het onderzoek dat zich uitsluitend toespitst op M&A-deals. In deze categorie wordt machine learning toegepast om te voorspellen of een M&A-deal zal worden voltooid. Hierbij wordt gebruikgemaakt van neural networks om een binaire classificatie te verkrijgen met aan de ene zijde een voltooide deal en aan de andere zijde geen deal (Karatas & Hirsra, 2021; K. Lee et al., 2020). In deze toepassingen wordt men geconfronteerd met ongelijk verdeelde data, waarbij meer deals worden gevonden die zullen slagen dan deals die zullen falen. K. Lee et al. (2020) lossen dit probleem op door *parameter tuning* toe te passen om de leersnelheid te verbeteren. Karatas en Hirsra (2021) maken gebruik van een oversamplingtechniek. Beide studies merken op dat het classificeren van een mislukte deal als een geslaagde deal meer kosten met zich meebrengt dan het classificeren van een mislukte deal als een geslaagde deal (Karatas & Hirsra, 2021; K. Lee et al., 2020). Om hiermee om te gaan, is het belangrijk om de juiste evaluatiemaat te selecteren. K. Lee et al. (2020) ontwikkelden een specifieke costfunctie die de ongelijke kosten van de fouten in rekening brengt. Karatas en Hirsra (2021) focussen op de *recall* om het model te evalueren. Deze evaluatiemaat geeft aan welk percentage van de positieve gevallen ook als positief voorspeld worden. In beide onderzoeken scoren de neural networks beter dan de *baseline* logit- of probitfuncties en bewijzen neurale netwerken hun nut in het voorspellen van succesvolle M&A-deals.

Een tweede categorie binnen M&A completion bekijkt het voorspellen van M&A-deals vanuit een investerings- of durfkapitalistperspectief. Hierbij is het doel om te voorspellen welke bedrijven succesvol zullen worden, zodat hier vroegtijdig in kan worden geïnvesteerd. Een bedrijf wordt dan als succesvol gezien indien het een target wordt voor een M&A-deal (Xiang et al., 2012). Ook kan een meer uitgebreide definitie worden toegepast waarbij zowel het worden van een M&A-target, als het behalen van een publieke beursgang (IPO) (Pan et al., 2018) of het behalen van een B-serie-investeringsronde (Żbikowski & Antosiuk, 2021) gezien wordt als succesvol. Hierbij kan worden opgemerkt dat in deze toepassingen zowel binaire classificatie als multiclass-classificatie wordt toegepast. Voor de binaire modellen worden alle situaties gegroepeerd die men als gunstig beschouwd. Toch is het ook mogelijk om elke categorie afzonderlijk te voorspellen (Arroyo et al., 2019). In deze toepassingen worden verschillende algoritmen gebruikt, zoals het *Bayesian network* (Xiang et al., 2012), de decision tree (Tsagkanos et al., 2007), K-NN (Pan et al., 2018) en XGBoost (Żbikowski & Antosiuk, 2021).

2.3.2. M&A-prestatie

Binnen de groep van M&A-prestatie, ook *M&A performance* genoemd, probeert men de effecten van een M&A-transactie te voorspellen. Er wordt getracht een vergelijking te maken tussen de situatie na

de M&A-deal en de situatie indien er geen M&A-deal had plaatsgevonden. De situatie waarin geen M&A plaatsvond, is een hypothetische situatie en moet daarom ingeschat worden. Dit kan worden gedaan op basis van theoretische modellen, zoals het *capital asset pricing model* (CAPM) of op basis van een synthetische controlegroep van gelijkaardige bedrijven (Zohrehvand et al., 2021). Dit soort studies hebben inherent enkele nadelen. Niet elke M&A komt vanuit dezelfde motivatie tot stand. Het is dus belangrijk om het doel van een deal duidelijk te definiëren en de deal vervolgens op dit criteria te evalueren. Dit wordt niet consequent gedaan en is mogelijk een verklaring voor de tegenstrijdige resultaten (Hassan et al., 2018). Daarnaast vergelijkt men vaak de prestatie van een bedrijf post-M&A met een hypothetische situatie waarin de onderneming geen M&A had uitgevoerd. Wanneer men deze vergelijking maakt, test men eveneens of de *proxy* voor de hypothetische situatie een goede proxy is. Dit kan vertekende beelden opleveren (Harford, 2005).

Craninckx en Huyghebaert (2011) maken gebruik van logistische en lineaire regressie om na te gaan of Europese overnames een positief effect hebben op de aandelenkoers en het operationeel resultaat. Daarnaast proberen ze ook te voorspellen of de investeerder, na het voltooien van een M&A-deal, zijn meerderheidspositie zal desinvesteren. De resultaten tonen aan dat slechts vijftig procent van de overnames een positief effect heeft op de aandelenkoers en dat tussen de zestig en de zeventig procent van de overnames leidt tot een verhoging van het operationeel resultaat.

Daarnaast kan de aandelenkoers van bedrijven die betrokken zijn in een M&A ook worden voorspeld op basis van historische gegevens. Wei (2019) maakt gebruik van deep learning om de verandering in de aandelenprijs van grote tech-bedrijven te voorspellen. In dit onderzoek wordt specifiek gekeken naar overnames uitgevoerd door Facebook, Amazon, Google en Apple.

2.3.3. M&A-waardering

Traditioneel worden *discounted cashflows* en *multiples* gebruikt voor de waardering, of *valuation*, van bedrijven (Moro Visconti et al., 2018). Meer recentelijk worden ook bigdata- en machine-learningtechnieken ingezet om de waarde van ondernemingen te bepalen. Zhang et al. (2020) maken gebruik van verschillende machine-learningalgoritmen om de dealprijs voor een overname te voorspellen in de olie- en energie-industrie. Hierbij behaalt het neural network een betere prestatie dan SVR, RF, KNN, Adaboost en DT. Ook Ang et al. (2020) maken gebruik van machine learning om de waarde van een onderneming te voorspellen. Zij maken gebruik van Elastic net en XGBoost met Bayesian optimization om de waarde van startups te voorspellen. Hieruit blijken vooral de investeringen, investeerders, regio en het aantal werknemers belangrijke variabelen om de waarde te bepalen. Vayas-Ortega et al. (2020) voorspellen de waarden van ondernemingen op basis van endogene en exogene variabelen. Hierbij blijken Gaussian Process Regression, SVR en BaggedTree-

modellen geschikt voor verschillende situaties, afhankelijk van de gebruikte variabelen en de sector. Opvallend is dat in deze studie de financiële variabelen juist wel een hoge invloed hadden op de voorspelling. De meest verklarende variabelen waren de kasstromen, de schulden, het BBP en het reële interesttarief. Deze financiële variabelen waren niet aanwezig in het onderzoek van Ang et al. (2020). Deze kunnen mogelijk dus nog een sterke bijdrage leveren in *startup valuation*.

Naast het voorspellen van de waarden kan machine learning ook ingezet worden om nuttige informatie omtrent bedrijfswaarde te verzamelen. Seng en Lai (2010) maken gebruik van *natural language processing* om financiële gegevens te verzamelen. Deze financiële gegevens kunnen vervolgens gebruikt worden om de bedrijfswaarde te bepalen. Ze slagen erin om efficiënt bruikbare gegevens te halen uit de jaarrekening en berichtgeving in het nieuws.

2.3.4. Variatoepassingen

Binnen de variatoepassingen zijn drie groepen van toepassingen te herkennen. Ten eerste zijn er de onderzoeken die gebruikmaken van tekstanalyse binnen M&A. In deze toepassingen worden algoritmen ingezet om automatisch relevante tekstdelen te verzamelen die belangrijk zijn om verder onderzoek te doen voor de M&A-transactie (Kellner, 2012; Roegiest et al., 2018). Veelal gaat het hierbij om contractuele documenten.

Een tweede groep toepassingen maakt gebruik van machine-learningalgoritmes om bepaalde gegevens van ondernemingen te voorspellen. Wang et al. (2020) maken gebruik van deep learning om de vijf belangrijkste (*big five*) persoonlijkheidskenmerken van CEO's en CFO's te voorspellen. Deze voorspellingen werden vervolgens gebruikt in hun verklarend model rond M&A-intensiteit. Li et al. (2018) maken gebruik van natural language processing om de bedrijfscultuur van ondernemingen in te schatten. Deze inschatting wordt vervolgens meegenomen in het model dat de kans op M&A voorspelt.

Ten derde kan machine learning ook ingezet worden om M&A-transacties in groepen te verdelen. Shao et al. (2019) verdelen M&A-deals in vijf categorieën aan de hand van *K-means clustering*.

2.4. Onderzoekskloof

Ondanks dat het waarderen van een onderneming een belangrijke stap is binnen het M&A-proces, is tot op heden nog maar beperkt onderzoek verricht naar de mogelijkheden van machine learning binnen dit domein. Deze thesis draagt op vier manieren bij aan het onderzoek naar machine-learningtoepassingen binnen M&A.

Ten eerste wordt gefocust op het inschatten van de dealprijs van startups. Er lijkt tot op heden geen onderzoek te zijn dat de M&A-dealprijs van startups voorspelt aan de hand van machine-

learningtechnieken. Eerder onderzoek naar dealprijspredictie focust specifiek op overnames in de olie- en energie-industrie (Zhang et al., 2020). Dit onderzoek maakt gebruik van deep-learningtechnieken. Dit leidt tot het vermoeden dat deze techniek ook geschikt is voor het voorspellen van dealprijzen van startups.

Ten tweede maken Zhang et al. (2020) gebruik van financiële gegevens om de aankoopwaarde van ondernemingen te voorspellen. Het huidige onderzoek maakt gebruik van niet-financiële gegevens voor het voorspellen van de dealprijs. Dit kan interessant zijn, omdat financiële gegevens van startups vaak niet beschikbaar zijn (Ang et al., 2020).

Ten derde maakt dit onderzoek op vernieuwende wijze gebruik van een bestaande dataset, namelijk Crunchbase data. Deze data werden eerder gebruikt voor het bepalen van de *post-funding valuation* van startups (Ang et al., 2020). Dit is de waarde van een startup nadat het een significante investering heeft ontvangen. De huidige thesis spitst zich toe op de valuation binnen een M&A-context. Voor dit doeleinde werd de dataset niet eerder gebruikt.

Ten vierde wordt in deze thesis gebruikgemaakt van hedendaagse technieken om de voorspellingen van de machine-learningmodellen te bespreken. Zo kan een beter zicht verkregen worden op de variabelen die van belang zijn voor het voorspellen van de dealprijs. Dit zal verdere inzichten geven in het fenomeen van M&A.

Hierbij is het doel van dit onderzoek om na te gaan hoe machine learning ingezet kan worden om de dealprijs van startups te voorspellen. Eerder onderzoek doet vermoeden dat deep learning de beste resultaten zal halen. Daarom wordt deep learning vergeleken met andere, meer traditionele machine-learningtechnieken, zoals XGBoost, support vector regression en lineaire regressie.

Concreet wordt getracht een antwoord formuleren op de volgende vragen:

- V1.** Welke modellen zijn het meest geschikt om de dealprijs van startups te voorspellen?
- V2.** Hoe accuraat kunnen de M&A-dealprijzen van startups worden voorspeld?
- V3.** Welke variabelen zijn het belangrijkste voor het voorspellen van de dealprijs van startups?

Met dit onderzoek wordt een bijdrage geleverd aan een beter begrip van startup valuation binnen M&A. Daarnaast wordt verder gebouwd aan het gebruik van machine-learning- en deep-learningtoepassingen binnen M&A valuation.

3. Methoden

Het doel van deze thesis is om bij te dragen aan het onderzoeksdomein dat machine-learningtechnieken inzet voor bedrijfswaardering in de context van M&A. Hierbij wordt specifiek toegespitst op de waardering van startups.

Het huidige onderzoek en de modellen die erin worden opgesteld, kunnen gezien worden als een beleidsondersteunende toepassing. Volgens de neo-klassieke theorie en de theory of the firm is M&A er veelal op gestoeld om het operationeel resultaat van de onderneming te verhogen. Volgens beide theorieën is het doel van M&A om efficiëntie-voordelen te behalen. Dit kan door kosten uit te sparen of meer omzet te genereren. Dit zou ten goede moeten komen aan het bedrijfsresultaat. Toch moet men eerst een investering doen voordat men deze vruchten kan plukken. Om te kunnen beslissen of M&A een juiste strategische beslissing is, moeten dus de kosten en de baten afgewogen worden. De kosten die gemaakt worden, betreffen veelal het proces van aankoop van een onderneming en het integreren van deze onderneming met de huidige bedrijfsvoering. De baten uit zich veelal in een stijging van het operationeel resultaat. Een goede inschatting van deze kosten en baten helpen dus bij het nemen van de juiste beslissing. Dit onderzoek tracht bij te dragen aan het maken van betere beslissingen door meer inzichten in de kosten van aankoop te verwerven. Meer specifiek wordt in dit onderzoek toegespitst op het inschatten van de aankoopkosten. Dit kan helpen in het beslissen of M&A een juiste strategische keuze is. Daarnaast zou deze inschatting ook kosten kunnen besparen doordat men geen middelen spendeert aan het onderzoeken van dure M&A-opties. Eveneens zou het kunnen helpen om goedkope M&A-opties te identificeren. Tenslotte wordt gebruikgemaakt van niet-financiële gegevens. Hierdoor kan het verkregen model een uitstekende aanvulling zijn voor het vaak gebruikte discounted-cashflowmodel. Dit is zeker het geval voor startups, waar de cashflows vaak ondergeschikt zijn aan het groeipotentieel van de startup.

Aangezien het doel is om een outputvariabele, in ons geval de dealprijs, te voorspellen op basis van inputvariabelen, is supervised learning de meest geschikte techniek. Meer specifiek wordt gebruikgemaakt van regressieanalyse, onder andere door middel van neural networks. De resultaten die verkregen worden aan de hand van dit algoritme worden vergeleken met algoritmes die al frequenter gebruikt zijn, zoals lineaire regressie, SVM en XGBoost.

3.1. Dataset

De dataset die gebruikt wordt, is de 'Crunchbase 2013 snapshot' (Crunchbase, 2013). Deze dataset bevat informatie over startups en hun betrokkenheid bij M&A-transacties. Na *data cleaning* en *feature engineering* (beschreven in puntje 3.2) blijft een dataset over van 2478 transacties, die beschreven worden aan de hand van 39 variabelen. Deze variabelen bevatten deal-specifieke kenmerken en

informatie over de betrokken ondernemingen, zijnde de acquirer en acquiree. Over de M&A-transacties is de prijs bekend die betaald is en wanneer de deal voltooid was. Voor de betrokken ondernemingen hebben we informatie over de sector waarin ze actief zijn, de plaats en datum van oprichting, investeringen die ze gemaakt en verkregen hebben, en indicatoren voor de grootte van de onderneming. De target variabele is de dealprijs, de prijs die betaald werd tijdens de M&A-transactie. Deze wordt uitgedrukt in US dollar en heeft een bereik tussen tienduizend en 39 miljard. Meer informatie over de dataset is terug te vinden in de appendix.

Hierbij dient opgemerkt te worden dat de definitie van startup wordt overgenomen zoals gebruikt bij Crunchbase in de '2013 snapshot'. De onderzoeker wenst geen verdere restricties op te leggen aan de ondernemingen die opgenomen worden in deze studie om het aantal M&A-deals niet verder te beperken.

3.2. Feature engineering

Om de datakwaliteit te verhogen, worden enkele datacleaning- en feature-engineeringstappen uitgevoerd. Zo werden ontbrekende data aangevuld, variabelen aangepast, nieuwe variabelen aangemaakt en rijen verwijderd waar nodig. Doordat de data veelal afkomstig zijn van bijdragen uit de gemeenschap is het niet altijd duidelijk of een waarde daadwerkelijk nul is, of dat de waarde voor deze variabele onbekend is voor de gemeenschap. In deze gevallen kan geen onderscheid gemaakt worden en worden dus geen aanpassingen gedaan.

3.2.1. Aanvullen van ontbrekende data

Het jaar waarin de acquirer en de acquiree werden opgericht ontbrak bij vele bedrijven. Wanneer dit jaar ontbrak, werd dit aangevuld met de mediaan van de oprichtingsdata van respectievelijk de acquirers en de acquirees. Dit resulteerde enerzijds in het aanvullen van 689 rijen met het oprichtingsjaar 1994 voor de acquirers. Anderzijds werden 1082 acquirees aangevuld met het oprichtingsjaar 2000.

Daarnaast werd ook de industrie van de bedrijven aangevuld wanneer deze data ontbrak. Hier werd gekozen om deze kolom aan te vullen met 'other'. Deze data moesten aangevuld worden voor 216 acquirers en 522 acquirees.

Ten slotte werd ook het aantal *offices* aangevuld. Ondernemingen waarvan het aantal offices niet bekend was, werden aangevuld met nul. Dit gebeurde voor 163 acquirers en 478 acquirees.

3.2.2. Variabelen aanpassen

Vijf aanpassingen werden doorgevoerd om de representatie van de variabelen te verbeteren. Ten eerste zijn niet alle deals uitgedrukt in US dollar. Er is gekozen om alle prijzen om te zetten naar dollar,

aangezien deze valuta de grootste groep vertegenwoordigt. Voor deze omzetting wordt gebruikgemaakt van de gemiddelde wisselkoers tussen 1993 en 2013. Dit is de periode waarin de M&A-transacties plaatsvonden. De concrete wisselkoersen zijn terug te vinden in appendix.

Ten tweede zijn ook de variabelen voor het land en de regio, van zowel de acquirer als de acquiree, omgezet naar dummy-variabelen die respectievelijk aangeven of de onderneming al dan niet in de Verenigde Staten gelokaliseerd is en of de onderneming al dan niet in de San Francisco Bay Area gelokaliseerd is.

Ten derde is voor de acquiree het diploma van de oprichter vertaald naar één dummy-variabele die aangeeft of de oprichter beschikt over een diploma van minstens bachelorniveau.

Ten vierde is de industrievariabele omgezet naar sectorniveau. Deze *mapping* naar sectorniveau reduceert het aantal categorieën. Daarnaast zorgt deze mapping er ook voor dat elke sector genoeg ondernemingen bevat. Voor het groeperen van de industrieën is gebruikgemaakt van de Global Industry Classification Standard (MSCI, 2022). De sectoren Financials en Real Estate zijn echter samengevoegd. Nadat de industrieën vertaald zijn naar een sector is voor elke sector een dummy-variabele aangemaakt. De exacte mapping is terug te vinden in appendix.

Ten vijfde zijn de variabelen ook gestandaardiseerd naar de gemiddelde nul- en standardvariantie. Dit kan de prestatie van de modellen verbeteren en is voornamelijk van belang bij neural networks en support vector machines.

Ten slotte is ook geëxperimenteerd met het omzetten van numerieke variabelen naar dummy-variabelen. Tijdens de verkennende analyse werd duidelijk dat vele variabelen scheef verdeeld zijn. De meeste ondernemingen scoren laag op de variabelen; slechts enkele scoren hoog. Doordat er slechts enkele cases met hoge scores zijn, kan het moeilijk zijn voor het model om deze effecten correct aan te leren. Indien het prijsverschil tussen hoog scoren en zeer hoog scoren beperkt is, kan het nuttig zijn om deze variabelen om te zetten naar dummy-variabelen. Zo werden de variabelen voor *investments*, *funding*, *relationships*, *milestones* en *offices* omgezet naar dummy-variabelen. Dit blijkt niet tot betere resultaten te leiden. De resultaten die verder besproken worden, zijn behaald door deze variabelen op te nemen als numerieke variabelen.

3.2.3. Nieuwe variabelen

Naast het aanpassen van bestaande variabelen werden ook nieuwe variabelen aangemaakt. Ten eerste is het aantal acquisitions dat een acquirer uitgevoerd had, berekend. Dit werd gedaan door het aantal M&A-deals voor een acquirer op te tellen. Deze variabele kan een indicatie geven van de hoeveelheid ervaring die een onderneming heeft met M&A. Anderzijds geeft deze variabele ook een indicatie van

de grootte van de onderneming. Een onderneming met meer budget zal meer overnames kunnen uitvoeren.

Ten tweede zijn ook twee dummy-variabelen aangemaakt om de gelijkheid tussen de acquirer en de acquiree aan te geven. Enerzijds werd nagegaan of de ondernemingen uit hetzelfde land komen en anderzijds werden de sectoren van de beide ondernemingen vergeleken.

Ten slotte is ook de leeftijd van de ondernemingen berekend op het moment van de overnamen. Dit is gedaan door de overname-datum te verminderen met de oprichtingsdatum van respectievelijk de acquirer en de acquiree.

3.2.4. Rijen verwijderen

Naast het aanpassen en creëren van variabelen, moesten ook enkele rijen verwijderd worden. Niet alle cases hebben voldoende (kwaliteitsvolle) data. Er werd in de volgende gevallen voor geopteerd de data te verwijderen:

- Wanneer de datum van de overname voor de datum van oprichting plaatsvond; dit is zowel gecontroleerd voor de acquirer als de acquiree.
- Voor één onderneming was het land niet USA, maar de regio was wel San Francisco Bay. Dit is onmogelijk. Deze case is verwijderd.
- De dealprijzen onder tienduizend USD werden verwijderd. Er waren veel M&A-transacties waarvan de dealprijs nul was. Hier was het onmogelijk om een prijs te voorspellen. Daarnaast waren er ook nog enkele ondernemingen met een zeer lage dealprijs. Vermoedelijk heeft hier een foute dataregistratie plaatsgevonden.

3.3. Verkennende analyse

Alvorens over te gaan tot het trainen van de machine-learningmodellen, is het belangrijk om een beeld te krijgen van de data waarmee wordt gewerkt. Daarom wordt een exploratieve analyse uitgevoerd om een beter begrip te krijgen van de distributie van onze data en de wijze waarop variabelen aan elkaar en aan de dealprijs zijn gelinkt.

Over het algemeen is te zien dat de data scheef verdeeld zijn. Zo is zichtbaar dat de meeste M&A-deals slechts een relatief lage prijs hebben. Een kleine groep uitschieters heeft een veel hogere dealprijs. Deze uitschieters bepalen in grote mate de totale waarde van alle M&A-deals en verhogen zo ook de gemiddelde prijs. Anderzijds zijn ook de kenmerken van de ondernemingen onderhevig aan deze scheve verdeling. Zo scoren de meeste ondernemingen laag op de variabelen. Slechts een beperkt aantal ondernemingen heeft hogere scores. Een voorbeeld hiervan is het aantal kantoren van de

acquirer. De meeste acquirers hebben minder dan tien kantoren. Eén acquirer heeft echter 67 kantoren.

Daarnaast is het duidelijk dat zowel de kenmerken van de acquirer als de acquiree invloed hebben op de dealprijs. Voor de acquirer lijken voornamelijk van belang: het aantal investeringen dat zij reeds deden, het aantal relaties met andere entiteiten en de leeftijd. Voor de acquiree zijn de belangrijkste kenmerken: het aantal relaties, het aantal mijlpalen en het aantal kantoren.

In wat volgt wordt de impact van verschillende variabelen op de dealprijs besproken. Hierbij wordt een indeling gemaakt voor kenmerken gelinkt aan de M&A-deal zelf, kenmerken van de acquirer en kenmerken van de acquiree.

3.3.1. Dealspecifieke informatie

Voor de dealspecifieke kenmerken is het belangrijk om de aandacht te vestigen op het jaar van de overname en de prijs die betaald is voor de M&A-deal. De meeste transacties in de dataset vonden plaats tussen 2005 en 2013. De oudste transactie dateert van 1993.

Tot aan het jaar 2003 blijft het aantal transacties beperkt. De transacties stijgen sterk in de periode van 2003 tot 2008. Vanaf 2008 is er een stagnatie, al blijft een sterke daling uit. Hierdoor lijkt de invloed van de financiële crisis op het aantal transacties beperkt. In het jaar 2010 is echter wel een dip te zien. Deze daling is van korte duur, aangezien het aantal M&A-deals weer stijgt vanaf 2011.

De prijs van de transacties ligt in de meeste gevallen tussen de één en driehonderd miljoen USD. Hierbij is te zien dat lage transactiepreizen vaker voorkomen. Een beperkt aantal outliers loopt op tot een dealprijs van 39 miljard USD.

3.3.2. Acquirer

Voor de bijna 2500 M&A deals zijn er 1265 unieke acquirers. Het is dus mogelijk dat eenzelfde acquirer betrokken is bij meerdere acquisitions. Zo zijn er acquirers die voor 96 acquisitions opgenomen zijn in de dataset. In wat volgt wordt bekeken hoe de kenmerken van de acquirer een mogelijke invloed hebben op de dealprijs.

Ten eerste is de leeftijd van de acquirer belangrijk voor het bepalen van de dealprijs. Zo betalen bedrijven die een vroegere oprichtingsdatum hebben over het algemeen meer voor een M&A dan bedrijven die meer recent opgericht zijn. Dit verband wordt verder versterkt wanneer wordt gekeken naar de leeftijd op het moment van de overname. Hier is te zien dat oudere bedrijven een neiging hebben om een hoger bedrag te betalen voor een M&A.

Daarnaast speelt ook de sector van een onderneming een belangrijke rol. Zo lijkt, in figuur 4, de gemiddelde prijs die bedrijven uit de IT-, luxegoederen- en productiesector betalen voor een M&A-deal, lager te liggen dan andere sectoren zoals finance. Echter, de prijsverschillen zijn in deze laatste sectoren veel groter. Dit komt mogelijk door het verschil in het aantal transacties per sector. Zo heeft de IT-sector de meeste transacties. Dit leidt tot een meer stabiele gemiddelde deal prijs. Anderzijds heeft de sector finance minder transacties. Hierdoor wordt de gemiddelde dealprijs van deze sectoren sterker beïnvloed door outliers.

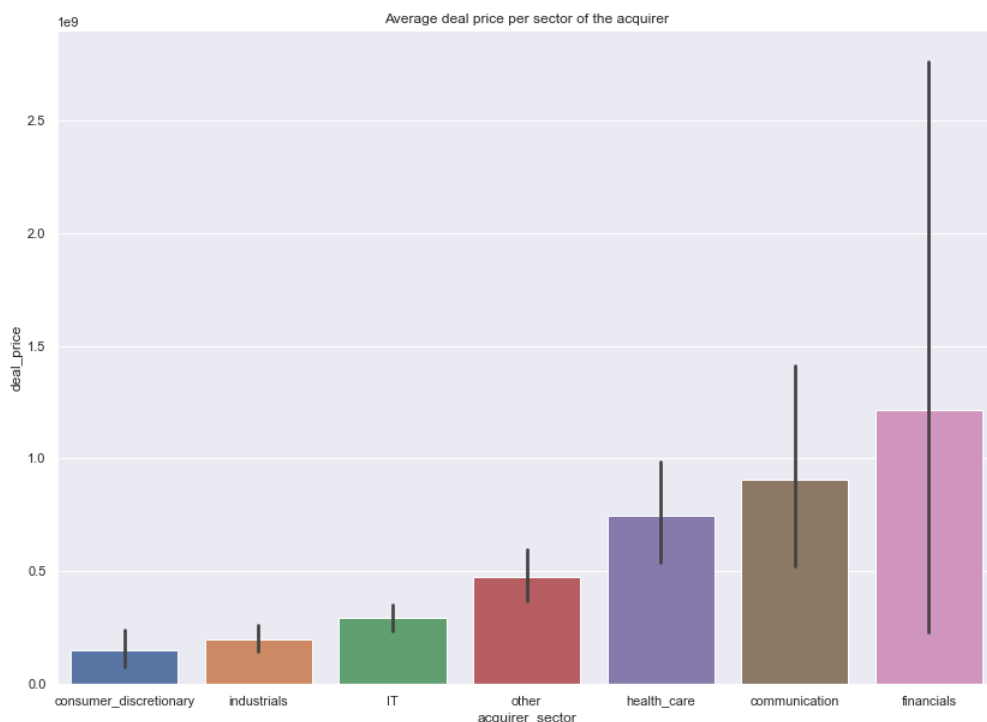


Figure 4: Average deal price per sector of the acquirer

Ook de locatie van de acquirer heeft invloed op de dealprijs. Hierbij lijkt het verschil tussen landen kleiner dan tussen specifieke regio's. Het al dan niet gelokaliseerd zijn in de Verenigde Staten heeft een beperkte invloed op de dealprijs. Is de acquirer echter afkomstig uit de San Francisco Bay Area, dan zal deze gemiddeld gezien een hogere prijs betalen.

Bovendien zijn ook de investeringen en de financiering van de acquirer gelinkt aan de dealprijs. Toch is de relatie van deze variabelen met de prijs voor een overname niet altijd gelijk. Wanneer een acquirer veel investeringen maakt, hangt dat samen met hogere prijzen in M&A-transacties. Het aantal acquisities dat een bedrijf doet, heeft een niet-lineair effect op de dealprijs. Dit kan waargenomen

worden in figuur 5. Zo lijkt de prijs in eerste instantie te stijgen naar mate het aantal acquisities toeneemt. Echter, er is ook één acquirer te zien met zeer veel acquisitions die gemiddeld gezien een lage prijs betaalt.

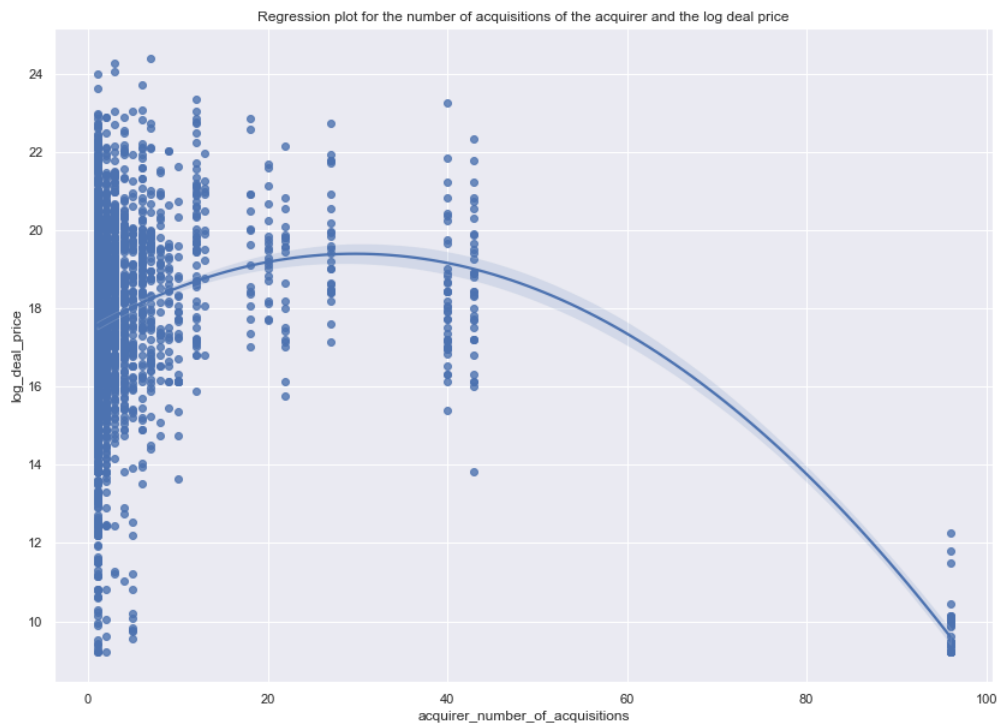


Figure 5: Regression plot for the number of acquisitions of the acquirer and the log deal price

Het tegenovergestelde is te zien voor het aantal financieringsrondes voor de acquirer. Wanneer een acquirer financiering ontvangt, zal de prijs die deze betaalt in een overname lager liggen. Dit effect keert om wanneer er meer dan zes financieringsrondes plaatsvinden.

Ten slotte lijkt ook de grootte van de acquirer de dealprijs van M&A te verhogen. Zo zijn de aantallen mijlpalen en relaties positief gecorreleerd met de dealprijs. Het aantal kantoren lijkt minder van belang.

3.3.3. Acquiree

Net als bij de acquirers, is de leeftijd van de acquiree van belang. Bedrijven die al langer bestaan, genereren een hogere dealprijs. Dit is zowel te zien aan de negatieve correlatie van de oprichtingsdatum als aan de positieve correlatie van de leeftijd van de acquiree.

De sector van de acquiree lijkt een minder grote impact te hebben op de dealprijs dan de sector van de acquirer. Figuur 6 toont dat de gemiddelde dealprijs voor de meeste sectoren van de acquirer eerder gelijk is. Uitschieters zijn finance- en productiebedrijven. Echter, de prijs schommelt binnen deze sectoren sterk. Het sterkste verband met de dealprijs wordt gevonden wanneer de acquiree tot de gezondheidszorgsector behoort. Daarnaast is vastgesteld, in figuur 7, dat het aantal transacties per sector verschilt van moment tot moment. Zo is te zien dat het aantal acquisitions voor de IT- en gezondheidszorgsectoren sterk stijgt in de aanloop naar 2008. Voor andere sectoren ligt de piek van het aantal transacties later en was de stijging meer geleidelijk. Dit kan erop wijzen dat verschillende sectoren anders beïnvloed worden met verloop van tijd.

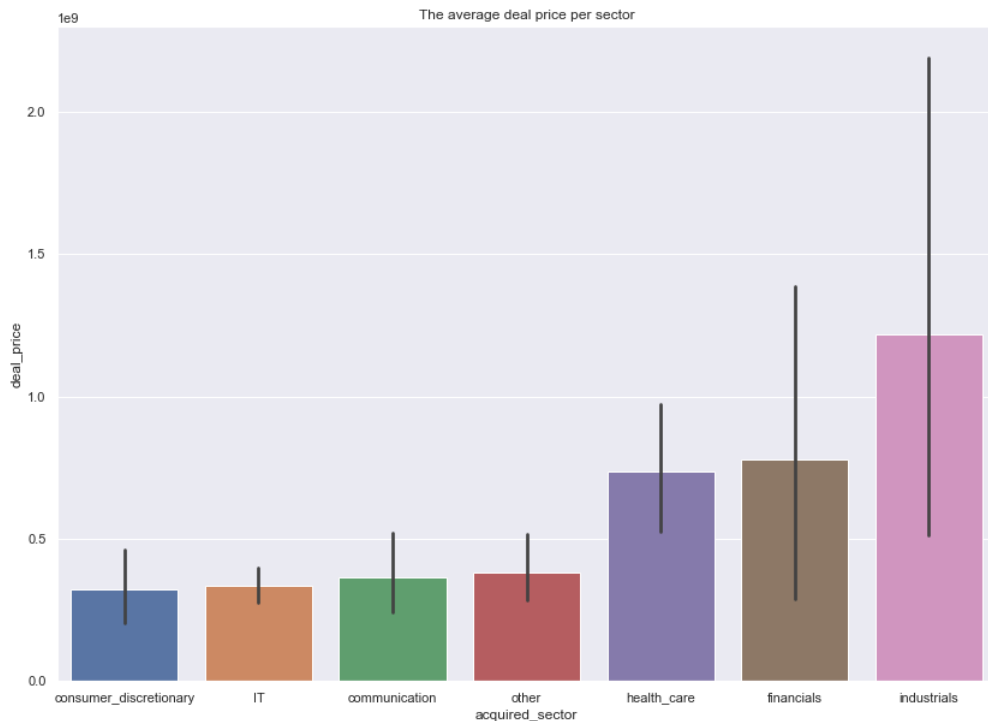


Figure 6: Average deal price per sector of the acquiree

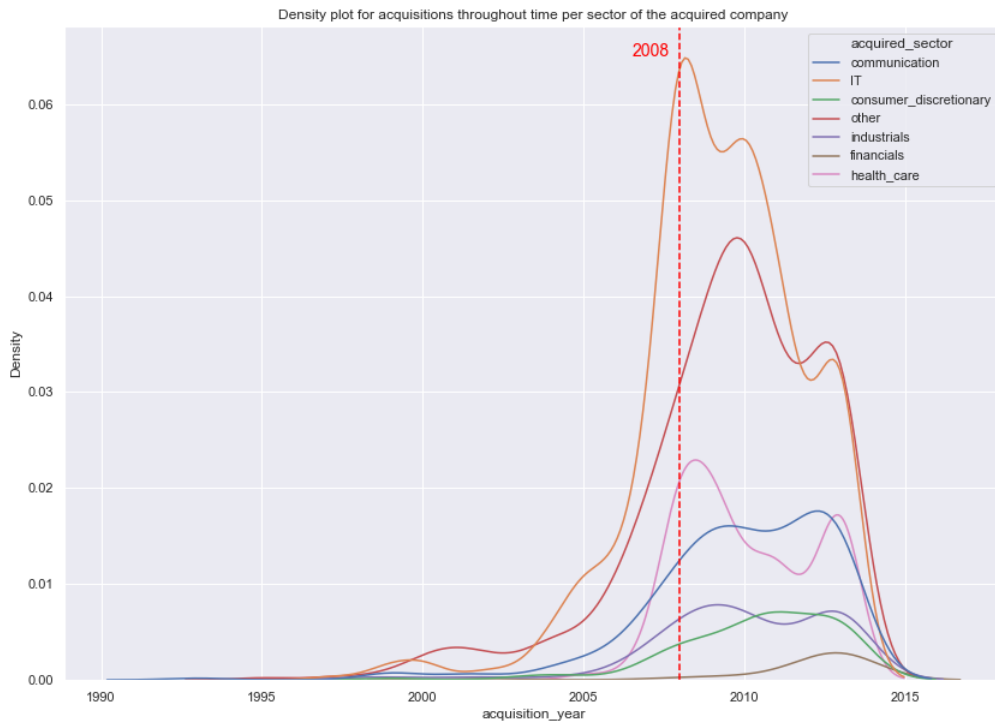


Figure 7: Density plot for the acquisitions throughout time

Ook de locatie van de acquiree lijkt van belang voor het bepalen van de prijs. Voor startups uit de Verenigde Staten wordt een hogere prijs betaald. Desalniettemin is er een lagere prijs voor startups die specifiek uit de San Francisco Bay Area komen. De richting van het effect van de locatie is dus anders voor de acquiree dan voor de acquirer.

Daarnaast is het effect van investeringen en financiering voor de acquiree duidelijk positief. Indien een acquiree meer investeringen doet, zal dit een positief effect hebben op de prijs die voor deze onderneming betaald wordt. Eveneens zijn de aantallen financieringsrondes en fondsen verbonden aan de startup positief gecorreleerd met de dealprijs.

De variabelen die gelinkt zijn aan de grootte van de onderneming hebben een sterke samenhang met de dealprijs. Deze variabelen zijn het aantal relaties, het aantal kantoren en het aantal mijlpalen. De impact van deze variabelen op de dealprijs wordt weergegeven in figuur 8. Het hebben van een score hoger dan nul, op deze variabelen, lijkt gelinkt te zijn aan een hogere dealprijs. Dit komt doordat vele van de startups een gelijkaardige, eerder lage score hebben. Er zijn slechts enkele uitschieters die een hogere waarde hebben. Desalniettemin lijkt de dealprijs voor grotere ondernemingen hoger.



Figure 8: Scatterplot for the acquiree's size and log deal price

Ten slotte heeft ook de opleiding van de oprichter invloed op de verkoopprijs. Hier is te zien dat wanneer een oprichter minstens een bacheloropleiding heeft gehaald, de prijs voor deze onderneming omhoog gaat.

3.4. Hyperparameter tuning

Om een optimale prestatie van de modellen te verkrijgen, dienen de juiste hyperparameters te worden geselecteerd. Voor lineaire regressie is dit niet nodig. Voor SVR en XGBoost werd de hyperparameteroptimalisatie gedaan aan de hand van random search. Het bereik van de hyperparameters die getest zijn, is te vinden in de appendix. De *hyperparameter ranges* voor de SVR zijn gebaseerd op Ito en Nakano (2003); voor XGBoost zijn deze gebaseerd op Bentéjac et al. (2021). Tweeduizend verschillende hyperparametercombinaties werden met elkaar vergeleken. Hierbij werden de continue hyperparameters geselecteerd op basis van een log-uniforme verdeling. Dit werd gedaan omdat de grootteorde van waarde in het bereik verschilt. Voor discrete en categorische hyperparameters werd een even grote kans op selectie toegewezen aan elke waarde. Elke combinatie werd geëvalueerd op basis van de *mean absolute error* (MAE). Hierbij werd gebruikgemaakt van *five-fold cross validation*. Dit houdt in dat de dataset werd verdeeld in vijf delen. Vervolgens zijn vier delen gebruikt om het model te trainen en één deel om het model te evalueren. Dit werd vijf keer herhaald zodat elk deel één keer gebruikt werd als test-set (Provost & Fawcett, 2013). De gemiddelde score van deze vijf tests wordt gebruikt om het model te evalueren. De methode die hierboven besproken is, werd geïmplementeerd met scikit-learn 0.24.2 en XGBoost 1.6.1.

Voor de *hyperparameter tuning* van het neural network werd Bayesian optimization toegepast. Hierbij werden vijftig hyperparametercombinaties gericht geselecteerd. Hierbij werd eveneens een log-uniforme verdeling gebruikt voor de continue hyperparameter en een uniforme verdeling voor de discrete en categorische hyperparameter. Elk netwerk werd getraind voor 600 epochs en gescoord op

basis van de MAE. Hierbij is gebruikgemaakt van een *train-validation split*, waarbij het model getraind wordt op tachtig procent van de cases en geëvalueerd op twintig procent van de cases. Deze methode werd geïmplementeerd met Tensorflow 2.8.2, Keras 2.8.0, en KerasTuner 1.1.2.

4. Resultaten

In dit deel wordt de prestatie van de modellen beschreven en vergeleken. Voor de machine-learningalgoritmen, met uitzondering van het neural network, werd een hyperparameteroptimalisatie uitgevoerd aan de hand van cross-validatie. Vervolgens werden de beste modellen opnieuw getraind op de volledige training set en getest op een test-set. Voor het neural network werd een train-validation-testsplit toegepast. Hierbij werd eerst gezocht naar de optimale hyperparameters. Vervolgens werd het beste model verder getraind voor 2000 epochs¹ en getest op de test-set. Merk op dat de test set voor alle modellen dezelfde cases bevat. Tenslotte werd ook een dummy-regressie geëvalueerd. Dit model voorspelt altijd de gemiddelde waarden en kan gezien worden als een baselinemodel.

Table 1: Results of machine learning models

| Model | Cross validation MAE | Test MAE |
|--------------------|----------------------|--------------------|
| Dummy | 548 340 940 | 730 325 846 |
| Lineaire regressie | 545 626 170 | 589 108 956 |
| SVM | 377 218 224 | 551 766 971 |
| XGBoost | 354 126 647 | 528 058 600 |

Table 2: Results for neural network

| Model | Train MAE | Validation MAE | Test MAE |
|----------------|-------------|----------------|-------------|
| Neural network | 359 574 296 | 291 410 452 | 529 953 181 |

De resultaten geven aan dat zowel XGBoost als neural networks een zekere verbetering zijn ten opzichte van eenvoudigere modellen, zoals de dummy-regressie of de lineaire regressie. Het is iets moeilijker om een duidelijk onderscheid te maken tussen de prestatie van het neural network en XGBoost. Doordat het neural network niet geëvalueerd is met een cross-validation-procedure, kunnen de scores niet rechtstreeks worden vergeleken. In vergelijking met de cross-validation MAE van de XGBoost behaalt het neural network een grotere error op de training-set, maar een kleinere error op de validation-set. Voor de test-set, die voor beide modellen dezelfde is, is te zien dat XGBoost een lagere error behaalt. Deze resultaten kunnen deels verklaard worden door de verschillende dealprijzen van de datasets. Zo is de gemiddelde dealprijs voor de cases in de test-set hoger dan die in de andere datasets. Anderzijds doet XGBoost vermoedelijk een voorspelling die beter te veralgemenen is naar de populatie van M&A-transacties.

¹ De loss-curve voor de train en de validation set kan teruggevonden worden in appendix

4.1. Voorspellingsfoutverdelingen

Naast het vergelijken van de gemiddelde absolute afstand (MAE) is het ook interessant om te kijken hoe deze afwijkingen verdeeld zijn. In dit deel wordt dieper ingegaan op hoe de voorspellingen van XGBoost en het neural network zich verhouden ten opzichte van de reële waarden.

In figuur 9 is te zien dat de voorspellingen van het XGBoost-model op de training-set een systematische onderschatting zijn ten opzichte van de geobserveerde waarde (hierbij dient te worden gelet op de schalen van de assen). Hierbij is echter wel te zien dat de voorspellingen consistent zijn. Hiermee bedoelen we dat het model slechts enkele keren een voorspelling maakt die sterk afwijkt van deze trend. Deze afwijkingen zijn dan veelal grote onderschattingen van de dealprijs.

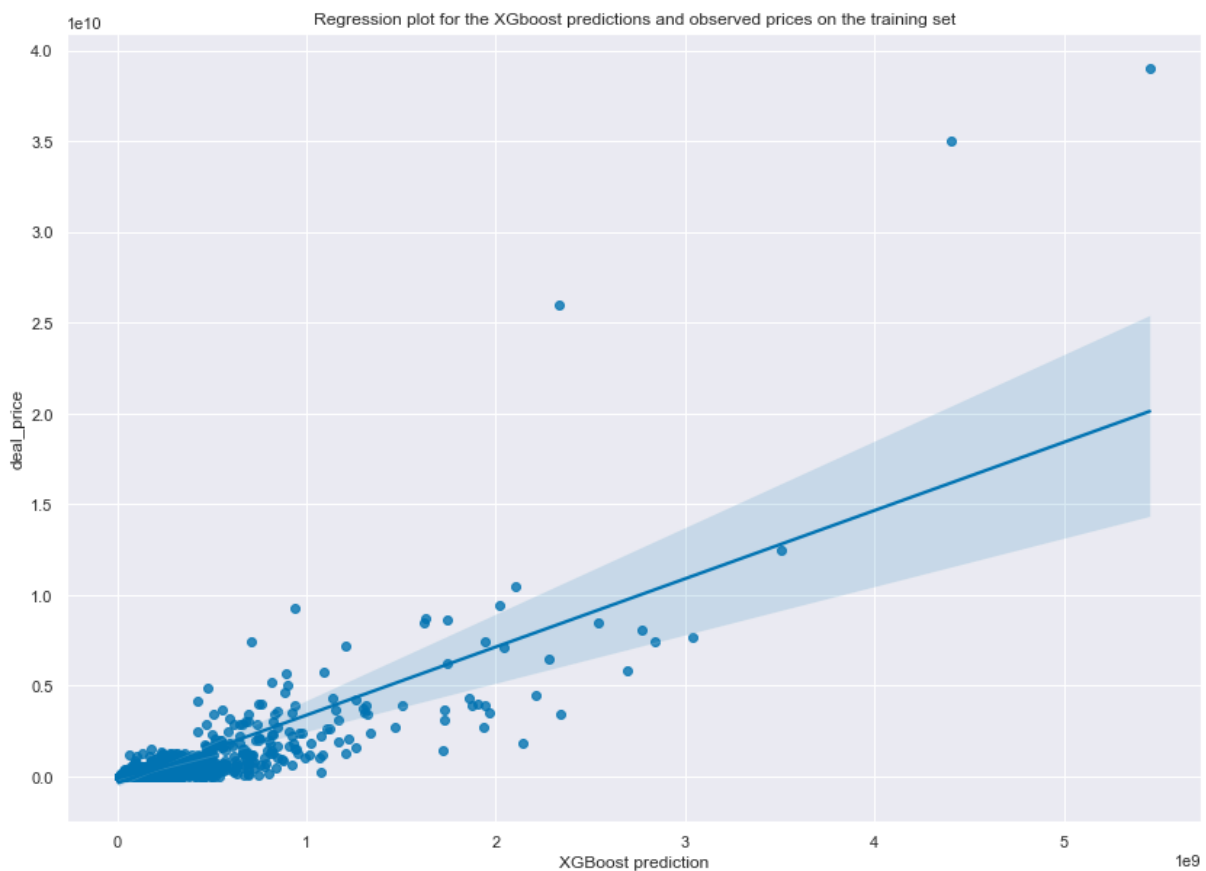


Figure 9: Regression plot for XGBoost predictions and observed deal prices on the training set

Voor de voorspellingen op de test-set, figuur 10, is een gelijkaardige trend waar te nemen. Echter, hier zijn de afwijkingen van de trend meer merkbaar. Zo maakt het model meer voorspellingen waarbij de onderschatting zowel groter als kleiner is dan de algemene trend van het model.

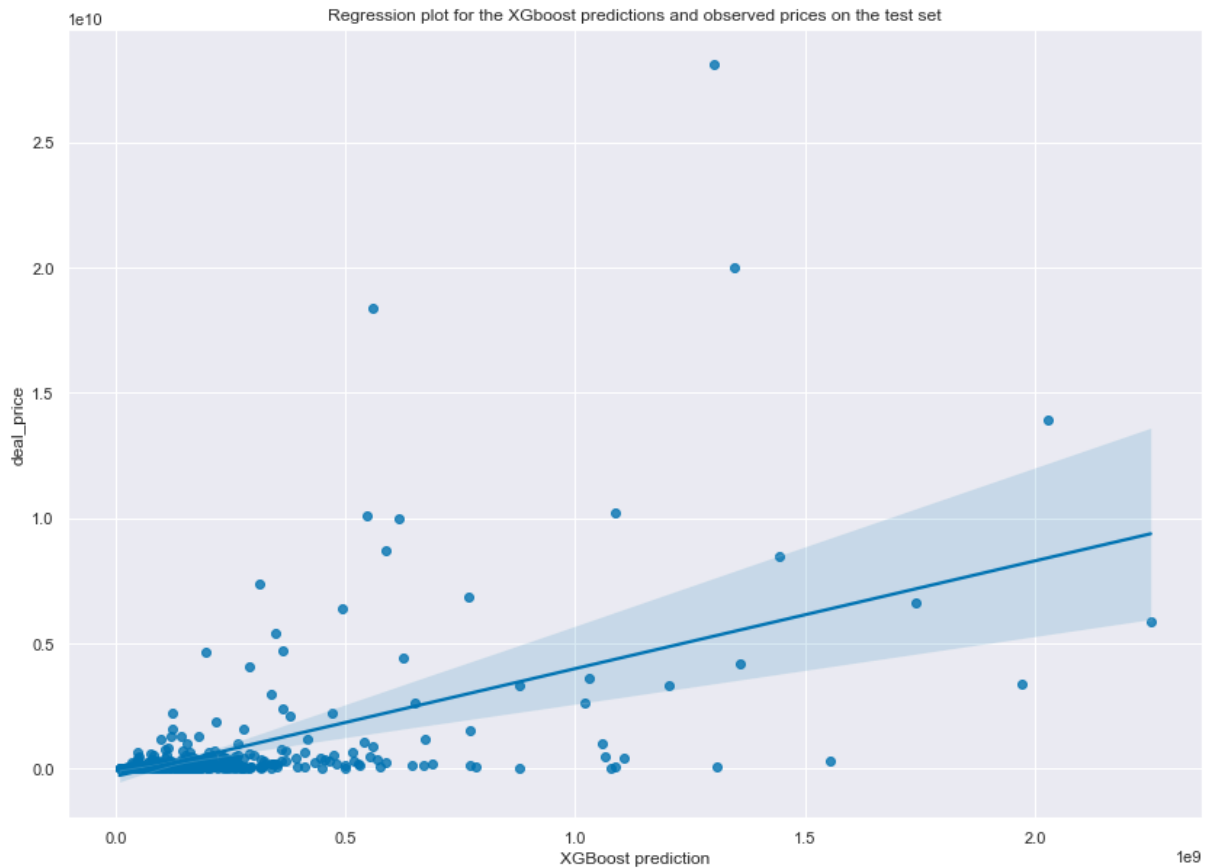


Figure 10: Regression plot for XGBoost predictions and observed deal prices on the test set

Voor het neural network is een ander tendens zichtbaar.² Op de training-set slaagt het netwerk erin om een voorspelling te doen die, gemiddeld gezien, gelijkloopt met de geobserveerde waarden. Echter, er zijn verschillende voorspellingen die sterk afwijken van deze trend. Zo maakt het model zowel over- als onderschattingen van de dealprijs. De onderschattingen zijn meer frequent. Wanneer echter wordt gekeken naar de test-set, is te zien dat het neural network een constante onderschatting maakt. Hier is een gelijkaardige trend te observeren als de voorspellingen van het XGBoost-model.

Het dient opgemerkt te worden dat op het continuüm van de dealprijzen, vele specifieke prijzen ontbreken in de dataset. Zo komen prijzen hoger dan vijfhonderd miljoen maar zelden voor. Daarbij komt dat de meeste dealprijzen laag zijn, maar ook enkele zeer hoge dealprijzen geobserveerd kunnen worden. Dit fenomeen is ook op te merken in de inputgegevens, zoals aangegeven in de verkennende analyse. Deze schaarse data heeft op twee manieren een invloed op de voorspellingen. Ten eerste worden de voorspellingen minder zeker naarmate de dealprijs stijgt. Bij lage dealprijzen kunnen de modellen terugvallen op verschillende vergelijkbare prijzen. Hierbij kunnen ze een consistente

² Figuren voor de voorspellingen van het neural network kan u terugvinden in de appendix

voorspelling maken. Wanneer de dealprijzen stijgen, verlaagt het aantal vergelijkbare cases. Dit maakt de voorspellingen meer onzeker. Ten tweede wordt de gemiddelde absolute afwijking (MAE) sterk beïnvloed door de fouten op deze hogere, minder vaak voorkomende dealprijzen. Op de hogere dealprijzen maakt het model, in absolute waarde, grotere fouten. Deze grotere fouten hebben een sterke invloed op het gemiddelde van de absolute afwijkingen van de voorspellingen ten opzichte van de reële waarden. Hierdoor is het waarschijnlijk dat het model voor lagere dealprijzen nauwkeurigere voorspellingen maakt dan de MAE die hier gerapporteerd werd.

4.2. Verklaring van de voorspellingen

Aan de hand van SHAP kan beter inzichtelijk worden gemaakt welke variabelen belangrijk zijn voor het maken van de voorspellingen. De impact van een variabele wordt gemeten ten opzichte van een baselinevoorspelling van het model. Dit is de voorspelling van het model indien geen informatie beschikbaar was over de input-variabelen (Lundberg & Lee, 2017). Voor het XGBoost-model is deze baselinevoorspelling gelijk aan 217.813.570. Dit is de voorspelling indien geen informatie beschikbaar was over de ondernemingen. Echter, de kennis over de ondernemingen beïnvloedt de voorspelde waarde. Zo zullen verschillende inputvariabelen de voorspelde waarde verhogen of verlagen.

De belangrijkste variabelen voor het XGBoost-model zijn het aantal relaties van de gekochte onderneming, de leeftijd van de gekochte onderneming en het aantal investeringsrondes van de koper. Een volledige lijst van de twintig meest belangrijke variabelen is terug te vinden in de appendix. Het is interessant om te zien dat het model zowel informatie gebruikt van de acquirer als van de acquiree om de prijs van de M&A-transactie te bepalen.

Het effect van de variabelen op de voorspelling van het XGBoost-model is te zien in figuur 11. Voor de meeste variabelen is het effect voor de hand liggend. Zo zal een hoger aantal relaties van de acquiree de prijs verhogen. Andere variabelen zijn minder eenduidig.

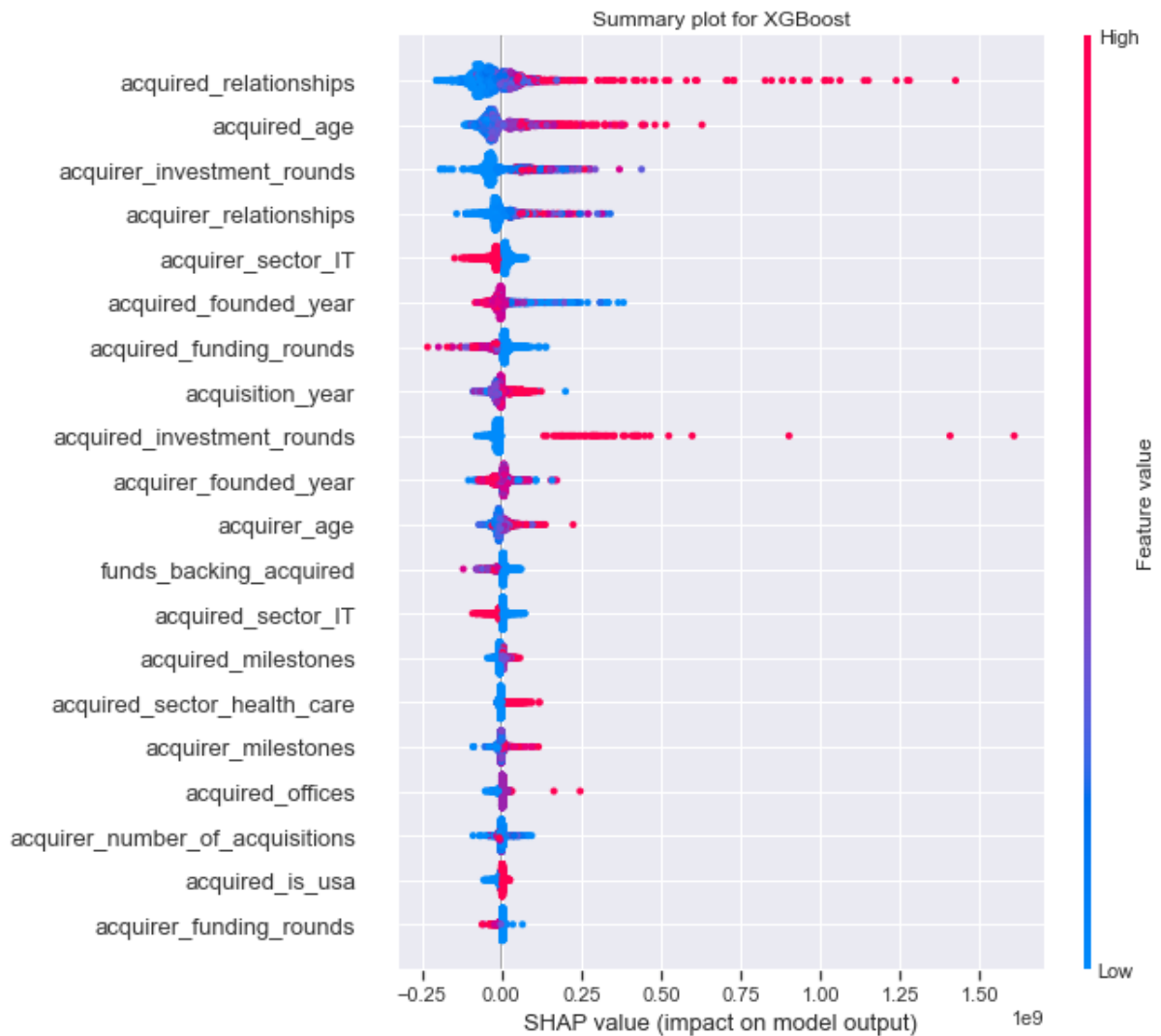


Figure 11: Summary plot of feature importance for XGBoost

Ten eerste lijkt er een optimale leeftijd te zijn om een startup te verkopen. In figuur 12 is zichtbaar dat een jonge leeftijd een negatief effect heeft op de verkoopprijs. Naarmate de startup ouder wordt, verhoogt deze verkoopprijs, totdat deze een maximum bereikt tussen ongeveer twintig en zestig jaar. Daarna lijkt het effect van een hogere leeftijd af te nemen.

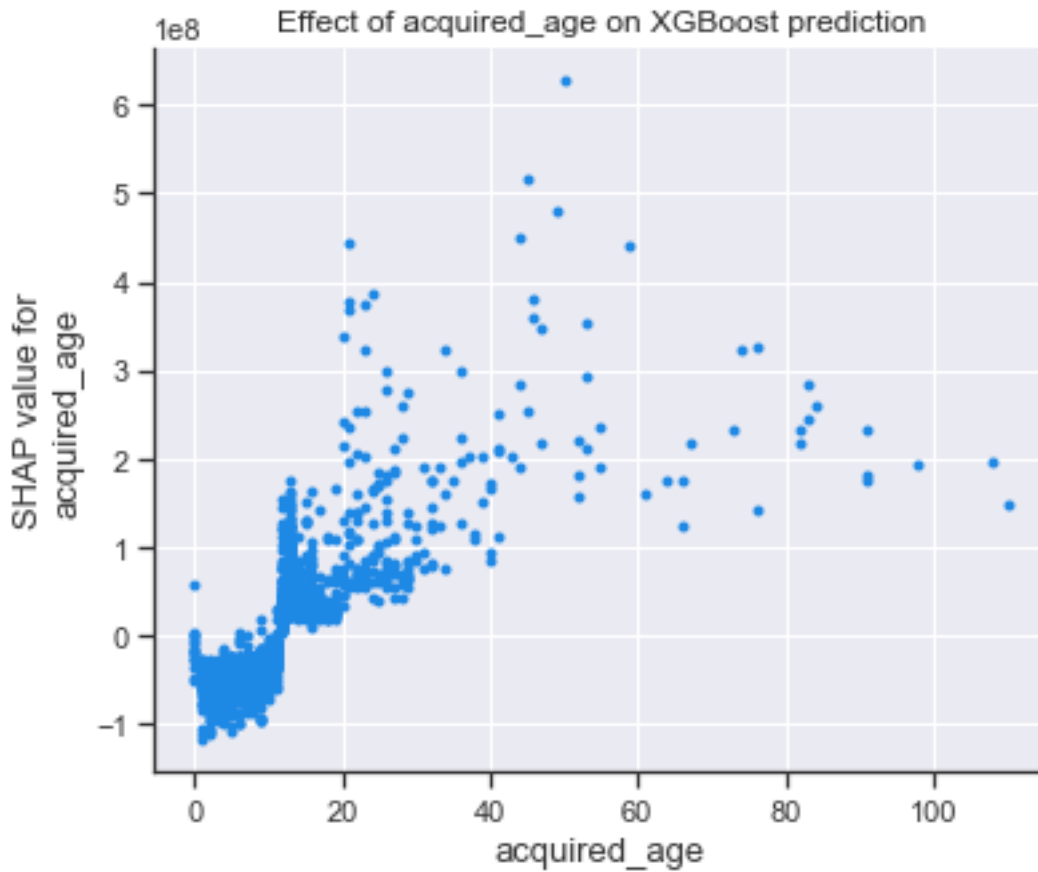


Figure 12: Impact of acquiree age on deal price prediction by XGBoost

Voor de investeringsrondes en het aantal relaties van de acquirer is een gelijkaardig verband in figuren 13 en 14 te observeren. Zo heeft een (bijna) nulscore op deze variabelen een negatief effect op de dealprijs. Wanneer de scores op deze variabelen stijgen, is de impact hiervan echter niet eenduidig. Het lijkt dat er een bepaalde drempel is. Wanneer deze drempelwaarde overschreden wordt, heeft dit een positieve impact op de dealprijs. Echter, een verdere stijging dan deze drempelwaarde heeft geen verdere impact op de dealprijs. Tussen de investeringsrondes en het aantal relaties konden geen duidelijke interactie-effecten waargenomen worden.

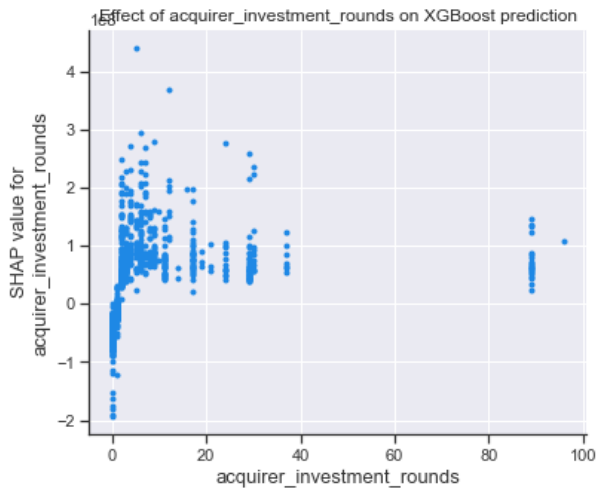


Figure 13: Impact of acquirer investments on deal price prediction of XGBoost

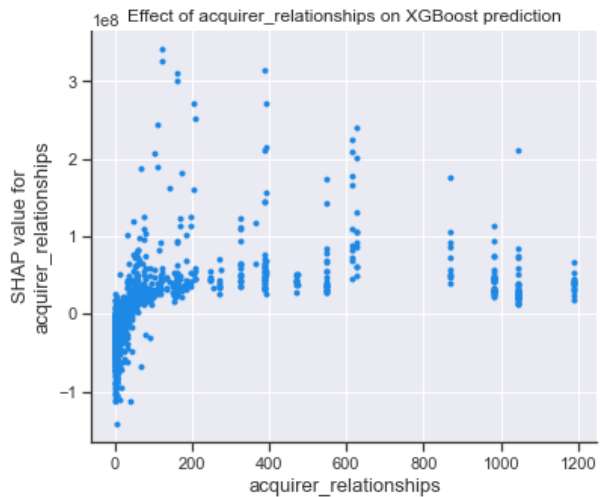


Figure 14: Impact of acquirer relationships on deal price prediction of XGBoost

Wanneer gekeken wordt naar de variabelen relaties, mijlpalen en kantoren voor zowel de acquirer als acquiree worden twee dingen duidelijk. Ten eerste zijn er transacties te zien waarbij zowel hoge en lage waarden voor beide ondernemingen voorkomen als transacties waarbij hoge of lage waarden voor de éne onderneming en lage of hoge waarden voor de andere onderneming voorkomen. Zo toont figuur 15 voorbeelden van ondernemingen met veel relaties die andere ondernemingen met veel relaties opkopen. Daarnaast zijn voorbeelden te zien waarbij de hoeveelheid relaties (in relatieve

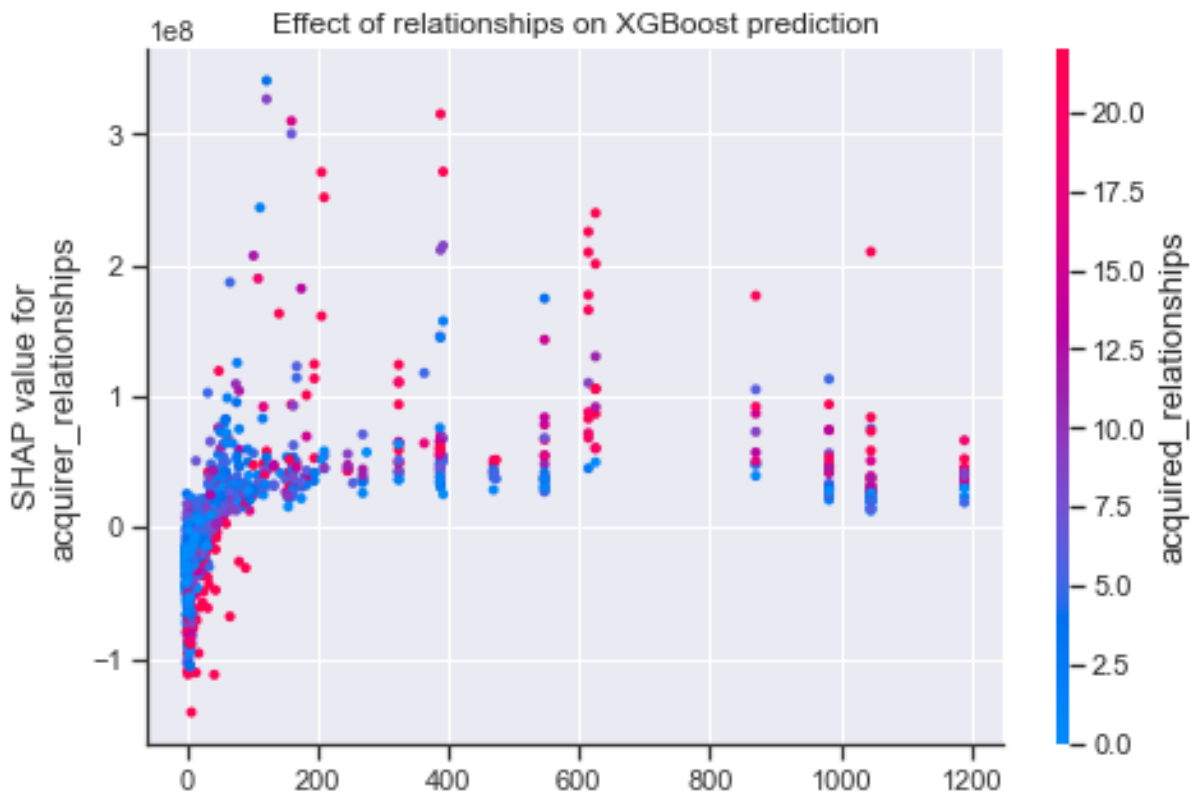


Figure 15: Effect of number of relationships of acquirer and acquiree on XGBoost predictions

termen) sterk verschilt. Zo kunnen ook bedrijven met weinig relaties bedrijven met meer relaties opkopen of vice versa.

Echter, er lijkt een effect te zijn voor de verhouding tussen het aantal relaties, de mijlpalen en de kantoren van betrokken ondernemingen en de voorspelling van het XGBoost-model. In figuur 16 is te zien dat de voorspelde waarde stijgt indien de mijlpalen voor beide ondernemingen rond de vier is. Echter, wanneer het aantal van de mijlpalen dicht bij nul is, of het aantal mijlpalen van de acquirer en acquiree sterk verschilt, dan is het effect op de voorspelling beperkt.

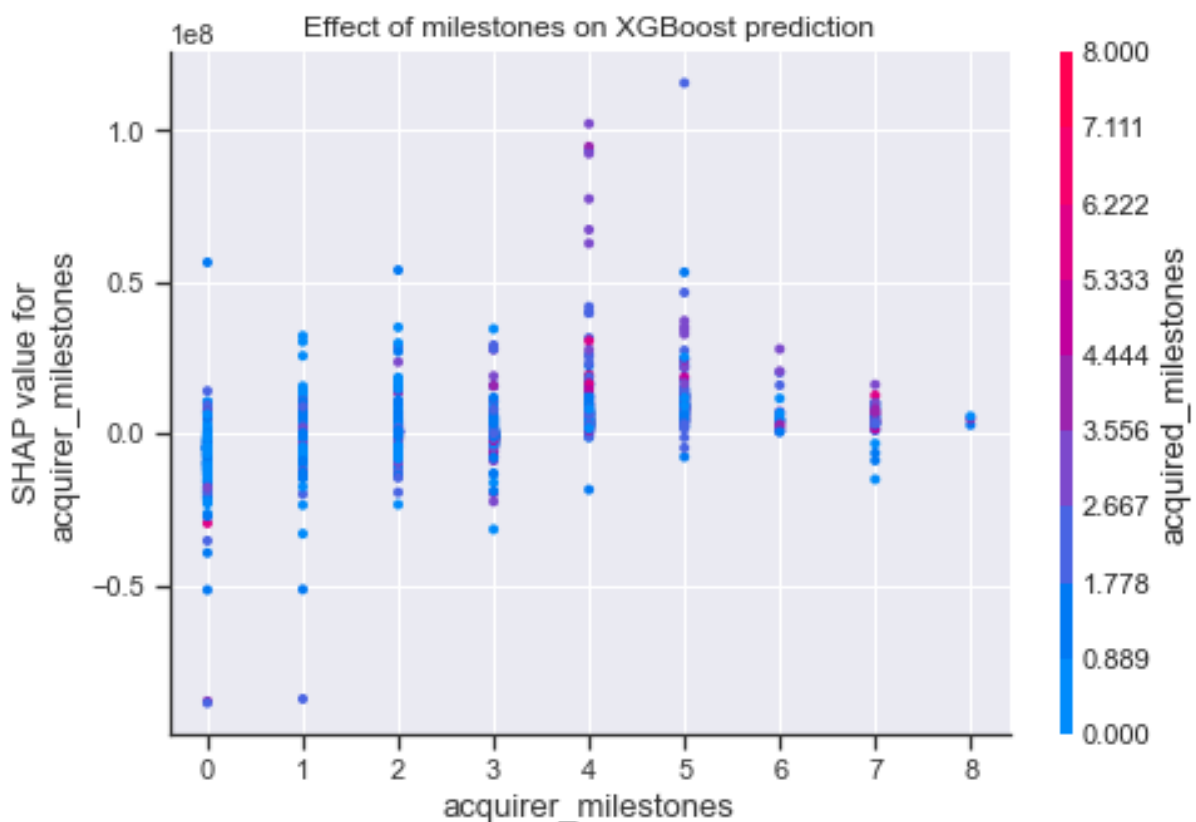


Figure 16: Effect of number of milestones of acquirer and acquiree on XGBoost predictions

Eveneens kan worden geobserveerd in figuur 17 dat wanneer de acquirer een groot aantal kantoren heeft, en dit niet het geval is voor de acquiree, dat dit een neerwaarts effect heeft op de voorspelling van het XGBoost-model. Wanneer beide ondernemingen geen, of een beperkt aantal, offices hebben, dan heeft dit een positief effect op de voorspelling.

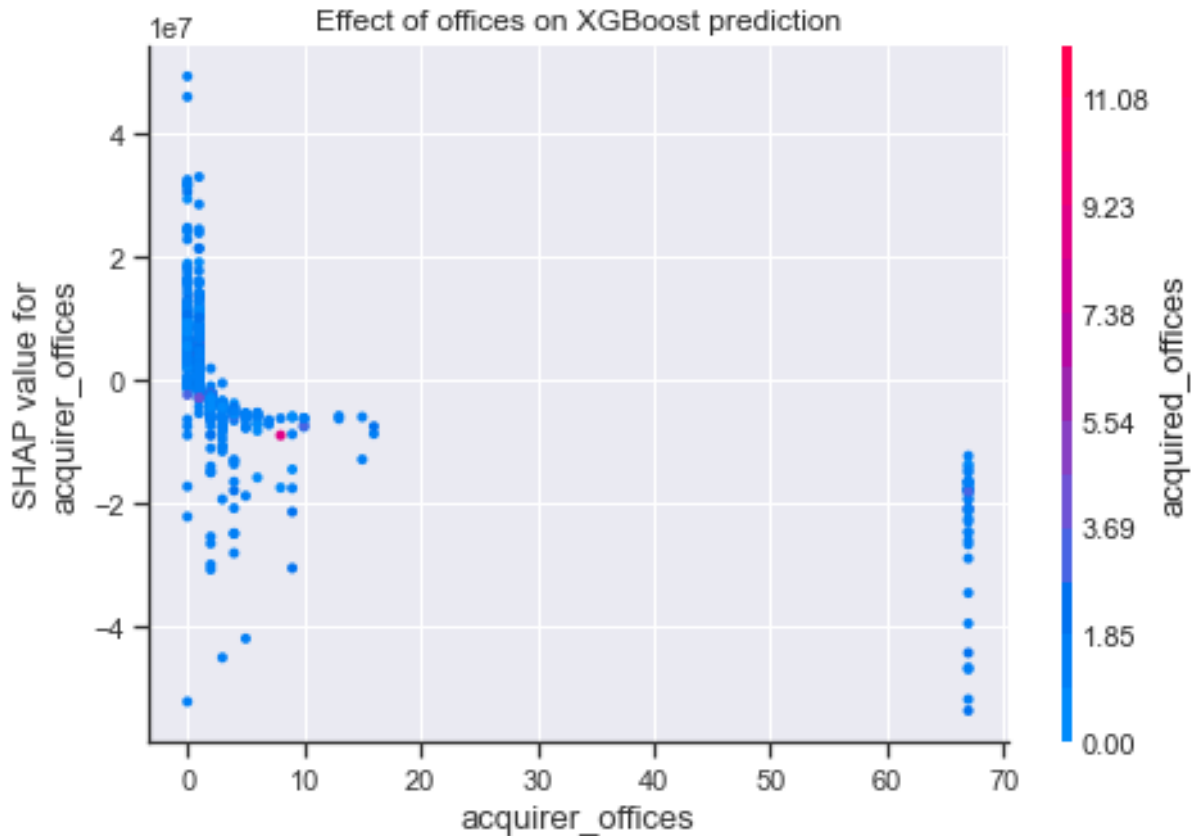


Figure 17: Effect of number of offices of acquirer and acquiree on XGBoost predictions

Het is interessant om op te merken dat de impact van de variabelen in het neural network anders zijn dan voor het XGBoost-model.³ Alhoewel, net als bij XGBoost, het aantal relaties van de acquiree de belangrijkste variabele is, is de impact van deze variabele op de voorspellingen van het neural network groter. Daarnaast hecht het neural network ook veel belang aan het aantal mijlpalen en de relaties van de acquirer. Deze kwamen in het XGBoost-model respectievelijk op de zestiende en vierde plaats. Ten slotte wordt de voorspelling van het neural network ook sterk beïnvloed door de afkomst van de acquiree. Voor het XGBoost-model stond het feit of de acquiree al dan niet uit de Verenigde Staten afkomstig was, slechts op de negentiende plaats.

Daarnaast is de impact van deze variabelen veelal eenduidiger voor het neural network⁴. Zo hebben zowel het aantal relaties van de acquirer en de acquiree, het aantal mijlpalen van de acquirer als het afkomstig zijn uit de Verenigde Staten een positieve impact op de dealprijs. Ten slotte kan worden

³ In appendix kan u een feature importance plot zien voor het neural network.

⁴ Een summary plot voor de impact van de variabelen op de voorspellingen van het neural network kan u terugvinden in appendix.

opgemerkt dat de baselinevoorspelling van het neural network lager ligt dan het XGBoost model en gelijk is aan 153.553.761.

5. Discussie

5.1. Betekenis van de voorspelling

De modellen in dit onderzoek voorspellen de dealprijs van een M&A-transactie. Het is hierbij echter van belang om na te denken wat deze voorspelling precies betekent en hoe deze gebruikt kan worden in de context van een M&A-deal. De modellen nemen gegevens van zowel de acquirer als de acquiree en specifieke dealgerelateerde informatie als input. Hierdoor is de voorspelling een verwachte waarde die één specifieke onderneming zou betalen voor een andere in het kader van een M&A-transactie in een specifiek jaar. De verwachte waarde is gebaseerd op gelijkaardige M&A-deals en duidt dus op een marktwaarde. Hierdoor sluit de voorspelling van de modellen in dit onderzoek nauw aan bij de waarderingstechniek van de multiples. Bij multiple-waardering wordt de waarde van een onderneming geschat door boekhoudgegevens te vergelijken met gelijkaardige bedrijven die eerder op de markt gewaardeerd zijn (Moro Visconti et al., 2018). Het verschil tussen de modellen in dit onderzoek en de multiple-methode zit veelal in de gebruikte gegevens. Voor multiples steunt men veelal op het gebruik van boekhoudgegevens en financiële gegevens. De modellen in dit onderzoek bevatten geen financiële gegevens. Hierdoor kunnen ze een aanvulling vormen op de multiple-methode.

Anderzijds is de voorspelde waarde gebaseerd op de prijzen van andere M&A-deals. De prijzen zijn vaak het gevolg van een biedwedstrijd waarbij verschillende ondernemingen geïnteresseerd zijn om eenzelfde onderneming aan te kopen. In de context van deze biedwedstrijd kan gesproken worden van de *winner's curse* (Bazerman & Samuelson, 1983). Dit is het fenomeen waarbij de winnaar van een biedwedstrijd veelal een te hoge prijs betaalt. Dit komt doordat de winnaar van de wedstrijd vaak degene is die de hoogste prijs betaalt. De bereidheid om een hogere prijs te betalen, is veelal het gevolg van een hogere waardering van het goed dat men wil kopen. Dit leidt ertoe dat de winnaar van een biedwedstrijd meer betaalt dan de waarde van het goed dat men kocht (Bazerman & Samuelson, 1983). Wanneer deze redenering wordt toegepast op de dealprijzen binnen een M&A-context, is te zien dat de prijs geen intrinsieke waarde is van de deal, maar mogelijk een overschatting is van de waardering van de acquiree. Om deze reden kan de voorspelling van het model in dit onderzoek niet gebruikt worden als een benadering van de intrinsieke waarde van een onderneming.

De voorspelling van de modellen is bijgevolg een inschatting van de marktwaarde voor een specifieke deal op basis van niet-financiële gegevens. Hierdoor kan de voorspelling gebruikt worden om een eerste screening te maken van potentiële M&A-deals die een onderneming wenst te verkennen. Op deze manier kunnen deals die boven het budget liggen, geschrapt worden en kosten bespaard worden in het verder onderzoeken van het potentieel van deze deals. Anderzijds geeft de voorspelling de mogelijkheid om zeer goedkope transacties te ontdekken. Hierdoor stuit men mogelijk op M&A-

opportuniteiten die anders niet onderzocht waren. De uiteindelijke prijs die men bereid is te betalen voor een M&A-deal zal moeten worden bepaald aan de hand van de meerwaarde die een acquiree zal leveren na de M&A-transactie.

Daarenboven nemen de modellen informatie over de acquirer in acht. Indien informatie over de potentiële kopers van een specifieke M&A-deal beschikbaar is, kan een inschatting worden gemaakt van het bod dat zij zullen plaatsen.

5.2. [Link met de literatuur](#)

In de literatuur omtrent waardering in de context van M&A aan de hand van machine learning was reeds aangetoond dat het mogelijk was om de dealprijs te voorspellen aan de hand van neural networks die getraind waren op financiële data (Zhang et al., 2020). Dit onderzoek breidt deze inzichten uit op twee manieren. Ten eerste wordt aangetoond dat ook niet-financiële gegevens nuttig kunnen zijn voor de waardering van bedrijven in de M&A-context. Hierbij wordt in dit onderzoek op unieke wijze gebruikgemaakt van technieken die het mogelijk maken de voorspellingen van het model te verklaren. Dit leidt ertoe dat variabelen onderscheiden kunnen worden die een grote invloed hebben op de voorspellingen van de modellen. Ten tweede wordt in het onderzoek aangetoond dat XGBoost minstens dezelfde resultaten behaalt als het neural network. Zo is dit algoritme ook een goede voorspeller in deze context dat vaak minder rekenkracht vereist om het te trainen.

Daarbij is het interessant op te merken dat het XGBoost-model theoretische concepten lijkt aan te leren. In de resultatensectie werden reeds de effecten besproken van het aantal relaties, mijlpalen en kantoren op de dealprijs. Indien deze variabelen worden geïnterpreteerd als indicators van de grootte van de onderneming, kunnen deze verbanden worden gelinkt aan theoretische concepten in M&A-literatuur. Zo postuleerde Harford (2005) dat M&A getypeerd wordt door grote bedrijven die kleine bedrijven opkopen. Rhodes-Kropf en Robinson (2008) weerleggen dit door te stellen dat *like buys like* een betere beschrijving is dan *high buys low*. Hiermee bedoelen ze dat veelal bedrijven van gelijke grootte M&A-deals aangaan. In de dataset komen beide transacties terug. Zo kunnen bedrijven met een hoog aantal relaties, mijlpalen en kantoren zowel ondernemingen met hoog als laag aantal relaties, mijlpalen en offices aankopen. Tevens kunnen 'low buys high'-transacties geobserveerd worden. Het model lijkt hierbij aan te leren dat een gelijke grootte (een relatief gelijk aantal relaties, mijlpalen of offices) een positief effect heeft op de dealprijs. Zo lijkt het model inzichten te verwerven in de belangen van de onderhandelingspositie van de ondernemingen. De onderhandelingspositie van ondernemingen speelt een belangrijke rol in het bepalen van deals volgens Rhodes-Kropf en Robinson (2008). Een even sterke positie leidt tot een meer gelijke verdeling van de meerwaarde van de transactie. Dit leidt tot een hogere prijs. Wanneer de onderhandelingsposities verschillen, heeft de

grote onderneming de mogelijkheid om een lagere prijs te betalen. Het XGBoost-model lijkt dit effect aan te leren.

5.3. Vergelijking met eerder onderzoek

De accuraatheid van de modellen van dit onderzoek vergelijken met andere toepassingen stuit op enkele hindernissen. Zo is er momenteel geen gelijkaardig onderzoek dat gebruikmaakt van de Crunchbase 2013 snapshot voor bedrijfswaardering in de context van M&A. Echter, er kan gekeken worden naar aanverwant onderzoek.

Ten eerste maken Ang et al. (2020) gebruik van meer recente Crunchbase-data om *post-funding valuation* te voorspellen. Hierbij varieert de geobserveerde waardering tussen de tien en de 26 op een logaritmische schaal. Het XGBoost-model in deze studie behaalt een *accuracy* van 96.45% en een MAE van 0.550. Door de schaalverschillen kunnen deze resultaten niet rechtstreeks worden vergeleken met de resultaten uit dit onderzoek. Daarom werd ook een XGBoost-model getraind op een dealprijs met logaritmische schaal. Een *five-fold cross-validation* van dit model behaalde een gemiddelde MAE van 1.242. Op de test-set scoorde dit model 1.222. Dit onderzoek is er dus niet in geslaagd om een nauwkeurigere voorspelling te maken dan Ang et al. (2020). Twee mogelijke verklaringen zijn enerzijds de mindere datahoeveelheid en kwaliteit van de Crunchbase 2013 snapshot ten opzichte van meer recente Crunchbase-data. Anderzijds is het ook mogelijk dat het moeilijker is om de dealprijs te voorspellen ten opzichte van de post-funding valuation.

Ten tweede maken Zhang et al. (2020) gebruik van financiële gegevens om de M&A-prijzen in de energiesector te voorspellen. Hier is de targetvariabele de dealprijs en dus gelijkaardig aan het onderzoek dat hier wordt gepresenteerd. Echter, de dealprijs varieert tussen 0,001 en 69.445 in het onderzoek van Zhang et al. (2020). De MAE voor de modellen in dit onderzoek variëren tussen de vijf en 21. De voorspellingen kunnen worden geschaald zodat ze in een gelijkaardig bereik liggen als de voorspellingen in dit onderzoek. Dit leidt tot de conclusie dat de voorspellingen van Zhang et al. (2020) nauwkeuriger zijn dan de voorspellingen in dit onderzoek. Financiële gegevens lijken dus belangrijk te zijn in het voorspellen van de deal prijs.

Daarnaast kan worden geobserveerd dat de gemiddelde absolute afwijking voor de cross-validation van het XGBoost-model in dit onderzoek gelijk is aan 354.126.647. Dit is hoger dan ruim tachtig procent van de dealprijzen. Op de test-set loopt de afwijking zelfs op tot meer dan 500.000.000. Hierbij merkten we reeds op dat deze gemiddelde afwijkingen sterk beïnvloed worden door afwijkende voorspellingen op M&A-deals met een hoge dealprijs. Voor transacties met een lagere dealprijs is het model nauwkeuriger.

Deze observaties hebben enkele implicaties voor de ingebruikname van het model. Het is belangrijk dat men voorzichtig omgaat met alle voorspellingen van het model. Zo is een substantiële afwijking van de reële waarde mogelijk. Dit is zeker het geval wanneer de dealprijs hoog is, of wanneer de betrokken ondernemingen inputwaardes hebben die weinig of niet voorkomen in de dataset. Daarbij komt dat men altijd de betekenis van de voorspelde waarde in acht moet nemen. In het algemeen lijkt het XGBoost-model dus veelal bruikbaar voor het inschatten van een richtprijs voor M&A's van kleinere ondernemingen.

5.4. Limitaties van het onderzoek

Ondanks dat het onderzoek verschillende bijdragen kan leveren aan de wetenschappelijke literatuur, moeten ook enkele beperkingen worden bekend. Ten eerste kan het onderzoek geen verbetering van de voorspellingsaccuraatheid realiseren ten opzichte van vergelijkbaar onderzoek. Dit is veelal te wijten aan de gebruikte dataset. Zo bevat de Crunchbase 2013 snapshot maar een gering aantal M&A-deals die beschikken over een dealprijs. Dit zorgt ervoor dat de modellen maar een beperkt aantal cases kunnen gebruiken om verbanden tussen de variabelen te leren. Vervolgens is ook de datakwaliteit niet optimaal. Op het moment dat de dataset opgesteld werd, steunde deze data veelal op input vanuit de gemeenschap. Dit leidt tot onzekerheid over de datakwaliteit. Hierbij komt dat de dataset maar een gering aantal cases heeft voor transacties met een zeer hoge dealprijs of waarbij de betrokkenen zeer hoog scoren op de inputvariabelen. Hierdoor kan de accuraatheid van de modellen voor die specifieke cases verder in twijfel worden getrokken.

Een tweede limitatie voor het gebruik van de modellen is de opzet van de voorspelling. In het huidige onderzoek wordt het voorspellen van de dealprijs bestudeerd als een cross-sectioneel probleem. Hierdoor ontstaat een *look-ahead bias*. Zo wordt in dit onderzoek geen rekening gehouden met het jaartal van de M&A-deal voor het trainen en valideren van de modellen. In dit onderzoek is het mogelijk dat deals uit het jaar 2010 gebruikt worden in de training-set en dat het model gevalideerd wordt op deals vanuit het jaar 2008. In werkelijkheid is dit niet mogelijk. Dan zouden de voorspellingen enkel gebaseerd kunnen worden op historische gegevens.

Een derde limitatie is de evaluatie van de modellen. Zo wordt gebruikgemaakt van slechts één vijfvoudige crossvalidatie voor lineaire regressie, SVR en XGBoost. Hierbij worden enkel de gemiddelde waarden van deze modellen vergeleken. Voor het trainen van het neural network wordt een train-validation-set gebruikt. Ten slotte worden ook al deze modellen geëvalueerd op eenzelfde test-set. Deze evaluatiemethode is sterk afhankelijk van de gekozen dataset-splits en de specifieke dataset die wordt gebruikt. Daarnaast worden de modellen geëvalueerd op basis van de gemiddelde absolute afwijking (MAE). Dit is een veel gebruikte evaluatiemaat voor regressieproblemen. Desalniettemin

wordt deze score beïnvloed door outliers. Door de grotere afwijkingen van de voorspellingen voor cases met een hoge dealprijs, kan de MAE een vertekend beeld geven van de algemene voorspellingsaccuraatheid van de modellen.

5.5. Toekomstig onderzoek

Verder onderzoek moet proberen de hierboven vermelde limitaties te verhelpen. Ten eerste moet men de gebruikte data verbeteren. Zowel de datahoeveelheid als de datakwaliteit van Crunchbase-data zijn mettertijd verbeterd. Daarom zou het interessant zijn om meer recente data te gebruiken voor verder onderzoek. Zo zou verder onderzoek de huidige analyse kunnen overdoen met meer data van betere kwaliteit. Evenwel kan men financiële gegevens over de ondernemingen toevoegen. Dit zou kunnen leiden tot een betere voorspelling van de dealprijs. Indien men beschikt over een grotere dataset, kan men eveneens beslissen om toe te spitsen op een subtype van M&A-deals. Zo zou men kunnen opteren om outliers, M&A-deals met zeer hoge prijzen, te verwijderen uit de dataset.

Ten tweede zou men gebruik kunnen maken van een *time-series approach* voor het voorspellen van M&A-prijzen. Zo moet verder onderzoek rekening houden met de jaartallen van de deals voor het opsplitsen van train- en testdata. De modellen kunnen bijvoorbeeld getraind worden op data van 1993 tot en met 2010, en vervolgens gevalideerd worden op data van 2011 tot 2013. Dit geeft een realistischer beeld van de prestaties van de modellen op nieuwe data.

Ten derde kan toekomstig onderzoek meer robuuste methoden inzetten om de modellen te vergelijken. Zo kan men gebruikmaken van *10X10 cross-validation*. In deze methode worden tien crossvalidaties uitgevoerd met telkens tien splits. Dit kan men vervolgens herleiden tot één stabiele score voor elk model. Daarnaast kan men ook de standaardafwijkingen van deze scores interpreteren om inzichtelijk te maken hoe de score van het model verschilt per split. Daarnaast kan men ervoor kiezen om de modellen te evalueren op basis van de mediane absolute afwijking. Deze evaluatiemaat is minder gevoelig voor outliers en kan op die manier een meer genuanceerd beeld geven van de prestatie van het model.

Wanneer men ten slotte, met deze verdere aanpassingen, een model verkrijgt dat voldoende accuraat is, moet dit model ook getest worden in de praktijk. Hierbij is het belangrijk om na te gaan of het model zijn accuraatheid behoudt op nieuwe data. Eveneens moet men nagaan of de voorspellingen van dit model nuttig zijn voor mensen in het werkveld.

6. Conclusie

In deze thesis zijn de mogelijkheden geanalyseerd van *machine learning* binnen de context van *mergers and acquisitions* (M&A). In de literatuurstudie worden drie grote domeinen van machine-learningtoepassingen onderscheiden. Zo bestaan er modellen voor het voorspellen van *M&A-completion*, *M&A-performance* en *M&A-valuation*. Ten slotte wordt hieraan een varia-groep toegevoegd. Hierin worden verscheidene andere modellen in de context van M&A besproken.

In dit onderzoek is een model uitgewerkt binnen het domein van M&A-valuation om zo de verkoopprijs in een M&A-transactie te voorspellen. De bijdragen van dit onderzoek zijn: (1) het waarderen van startups in de context van M&A, (2) het benutten van niet-financiële gegevens in deze context, (3) het gebruikmaken van een dataset die niet eerder werd gebruikt voor dit type onderzoek, en (4) het toepassen van verklaringstechnieken om een beter begrip te krijgen van de manier waarop de modellen voorspellingen maken.

Vier machine-learningmodellen zijn met elkaar vergeleken. Eerst is een verkennende analyse uitgevoerd. Dit gaf aanleiding tot *data cleaning* en *feature engineering* om vervolgens over te gaan tot hyperparameteroptimalisatie en het vergelijken van de modellen.

Hierbij werden drie onderzoeksvragen beantwoord. Ten eerste, welke modellen zijn het meest geschikt om de dealprijs van startups te voorspellen? De resultaten duiden dat XGBoost en het *neural network* vergelijkbare resultaten behalen die een verbetering zijn ten opzichte van lineaire regressie en support-vectorregressie. Ten tweede, hoe accuraat kunnen de M&A-dealprijzen van startups worden voorspeld? XGBoost en het neural network behalen een gemiddelde absolute voorspellingsfout van 350.000.000 tot 530.000.000. Ten derde, welke variabelen zijn het belangrijkste voor het voorspellen van de dealprijs van startups? De variabelen die een sterke impact hebben op de voorspellingen van XGBoost en het neural network zijn verschillend. De variabelen die het meeste invloed hebben op de voorspellingen voor beide modellen zijn: het aantal relaties van beide ondernemingen, de leeftijd van de *acquiree*, het al dan niet afkomstig zijn uit de VS van de *acquiree*, en het aantal investeringen en mijlpalen van de *acquirer*.

Het onderzoek wordt gekenmerkt door enkele limitaties. Zo is de gebruikte dataset enigszins beperkt in omvang en kwaliteit. Daarnaast ondervinden de modellen een *look-ahead bias* door de manier waarop de train- en testdata geselecteerd wordt. Ten slotte is ook de evaluatie van de modellen eenvoudig en onvoldoende robuust.

In verder onderzoek kan de focus worden gelegd op het gebruik van meer kwaliteitsvolle data, het gebruik van een time-series-perspectief en een meer robuuste methode om de modellen te evalueren.

Daarnaast moeten ook de bruikbaarheid en meerwaarde van deze modellen in de praktijk onderzocht worden. Mits deze extra stappen genomen worden, heeft machine learning de mogelijkheid om een waardevolle bijdrage te leveren aan het domein van mergers and acquisitions.

7. Bronnenlijst

- Akbari, A., Ng, L., & Solnik, B. (2021). Drivers of economic and financial integration: A machine learning approach. *Journal of Empirical Finance*, 61, 82–102. <https://doi.org/10.1016/j.jempfin.2020.12.005>
- Alzubi, J., Nayyar, A., & Kumar, A. (2018). Machine Learning from Theory to Algorithms: An Overview. *Journal of Physics: Conference Series*, 1142(1). <https://doi.org/10.1088/1742-6596/1142/1/012012>
- Ang, Y. Q., Chia, A., & Saghafian, S. (2020). Using Machine Learning to Demystify Startups Funding, Post-Money Valuation, and Success. *SSRN Electronic Journal*, 1–28. <https://doi.org/10.2139/ssrn.3681682>
- Arroyo, J., Corea, F., Jimenez-Diaz, G., & Recio-Garcia, J. A. (2019). Assessment of machine learning performance for decision support in venture capital investments. *IEEE Access*, 7, 124233–124243. <https://doi.org/10.1109/ACCESS.2019.2938659>
- Awad, M., & Khanna, R. (2015). *Efficient Learning Machines*. Apress Media.
- Bazerman, M. H., & Samuelson, W. F. (1983). I Won the Auction But Don't Want the Prize. *Journal of Conflict Resolution*, 27(4).
- Beam, A. L., & Kohane, I. S. (2018). Big data and machine learning in health care. *JAMA - Journal of the American Medical Association*, 319(13), 1317–1318. <https://doi.org/10.1001/jama.2017.18391>
- Benardos, P. G., & Vosniakos, G. C. (2007). Optimizing feedforward artificial neural network architecture. *Engineering Applications of Artificial Intelligence*, 20(3), 365–382. <https://doi.org/10.1016/j.engappai.2006.06.005>
- Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. In *Artificial Intelligence Review* (Vol. 54, Issue 3). Springer Netherlands. <https://doi.org/10.1007/s10462-020-09896-5>
- Breuel, T. M. (2015). *The Effects of Hyperparameters on SGD Training of Neural Networks*. <http://arxiv.org/abs/1508.02788>
- Cburnett. (2006). *File:Artificial neural network.svg*. Wikipedia. https://commons.wikimedia.org/wiki/File:Artificial_neural_network.svg

- Craninckx, K., & Huyghebaert, N. (2011). Can stock markets predict M&A failure? A study of European transactions in the fifth takeover wave. *European Financial Management*, 17(1), 9–45. <https://doi.org/10.1111/j.1468-036X.2010.00541.x>
- Cretin, F., Dieudonne, S., & Bouacha, S. (2015). M&A Activity: Where Are We In the Cycle? *Alternative Investment Analysis*, 38–44.
- Crunchbase. (2013). *2013 Snapshot*. <https://data.crunchbase.com/docs/2013-snapshot>
- Dernoncourt, F., & Lee, J. Y. (2017). Optimizing neural network hyperparameters with Gaussian processes for dialog act classification. *2016 IEEE Workshop on Spoken Language Technology, SLT 2016 - Proceedings*, 406–413. <https://doi.org/10.1109/SLT.2016.7846296>
- Dixon, M. F., & Halperin, I. (2019). The Four Horsemen of Machine Learning in Finance. *SSRN Electronic Journal*, 1–24. <https://doi.org/10.2139/ssrn.3453564>
- Emerson, S., Kennedy, R., O’Shea, L., & O’Brien, J. (2019). Trends and Applications of Machine Learning in Quantitative Finance. *8th International Conference on Economics and Finance Research (ICEFR)*.
- Fanning, K., & Drogt, E. (2014). Big Data: New opportunities for M&A. *The Journal of Corporate Accounting & Finance*, 27–34. <https://doi.org/10.1002/jcaf>
- Feurer, M., & Hutter, F. (2019). Hyperparameter Optimization. In F. Hutter, L. Kotthoff, & J. Vanschoren (Eds.), *Automated Machine Learning* (pp. 3–34). Springer. <http://www.springer.com/series/15602>
- Frazier, P. I. (2018). *A Tutorial on Bayesian Optimization*. <http://arxiv.org/abs/1807.02811>
- Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (2nd ed.). O’Reilly Media.
- Greene, D., Hoffmann, A. L., & Stark, L. (2019). Better, nicer, clearer, fairer: A critical assessment of the movement for ethical artificial intelligence and machine learning. *Proceedings of the Annual Hawaii International Conference on System Sciences, 2019-January*, 2122–2131. <https://doi.org/10.24251/hicss.2019.258>
- Hansen, K. B., & Borch, C. (2021). The absorption and multiplication of uncertainty in machine-learning-driven finance. *British Journal of Sociology*, 72(4), 1015–1029. <https://doi.org/10.1111/1468-4446.12880>

- Harford, J. (2005). What drives merger waves? *Journal of Financial Economics*, 77(3), 529–560.
<https://doi.org/10.1016/j.jfineco.2004.05.004>
- Hassan, I., Ghauri, P. N., & Mayrhofer, U. (2018). Merger and acquisition motives and outcome assessment. *Thunderbird International Business Review*, 60(4), 709–718.
<https://doi.org/10.1002/tie.21967>
- IMAA. (2021). *M&A Statistics - Worldwide, Regions, Industries & Countries*. Institute of Mergers, Acquisitions and Alliances. <https://imaa-institute.org/mergers-and-acquisitions-statistics/>
- Ito, K., & Nakano, R. (2003). Optimizing Support Vector Regression Hyperparameters Based on Cross-Validation. *Proceedings of the International Joint Conference on Neural Networks*, 3, 2077–2082.
<https://doi.org/10.1109/ijcnn.2003.1223728>
- James, G., Witten, D., Hastie, T., & Tibishirani, R. (2013). An Introduction to Statistical Learning with Applications in R (older version). *Springer Texts in Statistics*, 426.
- Karatas, T., & Hirsra, A. (2021). *Predicting Status of Pre and Post M&A Deals Using Machine Learning and Deep Learning Techniques*. 1–21. <http://arxiv.org/abs/2110.09315>
- Kellner, C. (2012). The Measure of a MAC: A Quasi-Experimental Protocol for Tokenizing Force Majeure Clauses in M&A Agreements. *SSRN Electronic Journal*, 168(2012), 202–205.
<https://doi.org/10.2139/ssrn.1872568>
- Kleynhans, T., Montanaro, M., Gerace, A., & Kanan, C. (2017). Predicting top-of-atmosphere thermal radiance using MERRA-2 atmospheric data with deep learning. *Remote Sensing*, 9(11).
<https://doi.org/10.3390/rs9111133>
- Lageranna, G., & Crawford, C. (2018). Merger and Acquisition: Conceptual Review. *SSRN Electronic Journal*, 29–35. <https://doi.org/10.2139/ssrn.3208534>
- Lee, J. H., Shin, J., & Realf, M. J. (2018). Machine learning: Overview of the recent progresses and implications for the process systems engineering field. *Computers and Chemical Engineering*, 114, 111–121. <https://doi.org/10.1016/j.compchemeng.2017.10.008>
- Lee, K., Joo, S., Baik, H., Han, S., & In, J. (2020). Unbalanced data, type II error, and nonlinearity in predicting M&A failure. *Journal of Business Research*, 109(December 2019), 271–287.
<https://doi.org/10.1016/j.jbusres.2019.11.083>

- Li, K., Mai, F., Shen, R., & Yan, X. (2018). Corporate Culture and Merger Success: Evidence from Machine Learning. *UBC Sauder Working Paper*. https://editorialexpress.com/cgi-bin/conference/download.cgi?db_name=CICF2018&paper_id=394
- Liao, L., Li, H., Shang, W., & Ma, L. (2022). An Empirical Study of the Impact of Hyperparameter Tuning and Model Optimization on the Performance Properties of Deep Neural Networks. *ACM Transactions on Software Engineering and Methodology*, 31(3), 1–40. <https://doi.org/10.1145/3506695>
- Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1), 56–67. <https://doi.org/10.1038/s42256-019-0138-9>
- Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, 30. <https://github.com/slundberg/shap>
- Maksimovic, V., & Phillips, G. (2001). The market for corporate assets: Who engages in mergers and asset sales and are there efficiency gains? *Journal of Finance*, 56(6), 2019–2065. <https://doi.org/10.1111/0022-1082.00398>
- Mitchell, T. M. (2006). The Discipline of Machine Learning. *Machine Learning*, 17(July), 1–7.
- Moro Visconti, R., Montesi, G., & Papiro, G. (2018). Big data-driven stochastic business planning and corporate valuation. *Corporate Ownership and Control*, 15(3–1), 189–204. <https://doi.org/10.22495/cocv15i3c1p4>
- MSCI. (2022). *Global Industry Classification Standard*. <https://www.msci.com/our-solutions/indexes/gics>
- Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, 31(2), 87–106. <https://doi.org/10.1257/jep.31.2.87>
- Müller, A. C. , & Guido, S. (2016). *Introduction to Machine Learning with Python* (first). O'Reilly Media.
- Ng, A., Katanforoosh, K., & Mourri, Y. B. (2022, February). *Deep Learning Specialization*. Coursera; Coursera.
- Olaru, C., & Wehenkel, L. (2003). A complete fuzzy decision tree technique. *Fuzzy Sets and Systems*, 138(2), 221–254. [https://doi.org/10.1016/S0165-0114\(03\)00089-7](https://doi.org/10.1016/S0165-0114(03)00089-7)
- Özgür, Ş., & Wirl, F. (2020). Cross-border mergers and acquisitions in the oil and gas industry: An overview. *Energies*, 13(21). <https://doi.org/10.3390/en13215580>

- Painsky, A., & Rosset, S. (2017). Cross-Validated Variable Selection in Tree-Based Methods Improves Predictive Performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(11), 2142–2153. <https://doi.org/10.1109/TPAMI.2016.2636831>
- Pan, C., Gao, Y., & Luo, Y. (2018). *Machine Learning Prediction of Companies ' Business Success*.
- Parsa, A. B., Movahedi, A., Taghipour, H., Derrible, S., & Mohammadian, A. (Kouros). (2020). Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis. *Accident Analysis and Prevention*, 136. <https://doi.org/10.1016/j.aap.2019.105405>
- Provost, F., & Fawcett, T. (2013). *Data Science for Business*. O'Reilly Media.
- Rhodes-Kropf, M., & Robinson, D. T. (2008). The Market for Mergers and the Boundaries of the Firm. *Journal of Finance*, 63(3), 1169–1211. <https://doi.org/10.1111/j.1540-6261.2008.01355.x>
- Roegiest, A., Hudek, A. K., & McNulty, A. (2018). A dataset and an examination of identifying passages for due diligence. *41st International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2018*, 465–474. <https://doi.org/10.1145/3209978.3210015>
- Rozemberczki, B., Watson, L., Bayer, P., Yang, H.-T., Kiss, O., Nilsson, S., & Sarkar, R. (2022). *The Shapley Value in Machine Learning*. <http://arxiv.org/abs/2202.05594>
- Schillewaert, N. (2022). Opnieuw grote overname in de gamingwereld: Sony koopt maker van computerspellen “Halo” en “Destiny.” *VRT NWS*.
- Seng, J. L., & Lai, J. T. (2010). An intelligent information segmentation approach to extract financial data for business valuation. *Expert Systems with Applications*, 37(9), 6515–6530. <https://doi.org/10.1016/j.eswa.2010.02.134>
- Shao, B., Asatani, K., & Sakata, I. (2019). Categorization of Mergers and Acquisitions in Japan Using Corporate Databases: A Fundamental Research for Prediction. *IEEE International Conference on Industrial Engineering and Engineering Management, 2019-Decem*, 1523–1527. <https://doi.org/10.1109/IEEM.2018.8607408>
- Shrikumar, A., Greenside, P., & Kundaje, A. (2017). *Learning Important Features Through Propagating Activation Differences*. <http://arxiv.org/abs/1704.02685>
- Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, 14, 199–222.

- Trichterborn, A., Zu Knyphausen-Aufseß, D., & Schweizer, L. (2016). How to improve acquisition performance: The role of a dedicated M&A function, M&A learning process, and M&A capability. *Strategic Management Journal*, 37(4), 763–773. <https://doi.org/10.1002/smj.2364>
- Tsagkanos, A., Georgopoulos, A., & Siriopoulos, C. (2007). Predicting Greek mergers and acquisitions: a new approach. *International Journal of Financial Services Management*, 2(4), 289. <https://doi.org/10.1504/ijfsm.2007.016286>
- Vayas-Ortega, G., Soguero-Ruiz, C., Rodríguez-Ibáñez, M., Rojo-Álvarez, J. L., & Gimeno-Blanes, F. J. (2020). on the differential analysis of enterprise valuation methods as a guideline for unlisted companies assessment (II): Applying machine-learning techniques for unbiased enterprise value assessment. *Applied Sciences (Switzerland)*, 10(15), 1–17. <https://doi.org/10.3390/APP10155334>
- Victoria, A. H., & Maragatham, G. (2021). Automatic tuning of hyperparameters using Bayesian optimization. *Evolving Systems*, 12(1), 217–223. <https://doi.org/10.1007/s12530-020-09345-2>
- Wang, Q., Lau, R. Y. K., & Yang, K. (2020). Does the interplay between the personality traits of CEOs and CFOs influence corporate mergers and acquisitions intensity? An econometric analysis with machine learning-based constructs. *Decision Support Systems*, 139(October), 113424. <https://doi.org/10.1016/j.dss.2020.113424>
- Wei, Z. (2019). Machine learning applications in finance: some case studies. [Doctoral Thesis, University of London]. Imperial College of Science Repository, January.
- Xiang, G., Zheng, Z., Wen, M., Hong, J., Rose, C., & Liu, C. (2012). A supervised approach to predict company acquisition with factual and topic features using profiles and news articles on TechCrunch. *ICWSM 2012 - Proceedings of the 6th International AAAI Conference on Weblogs and Social Media*, 607–610.
- Yu, T., & Zhu, H. (2020). *Hyper-Parameter Optimization: A Review of Algorithms and Applications*. <http://arxiv.org/abs/2003.05689>
- Żbikowski, K., & Antosiuk, P. (2021). A machine learning, bias-free approach for predicting business success using Crunchbase data. *Information Processing and Management*, 58(4), 102555. <https://doi.org/10.1016/j.ipm.2021.102555>
- Zhang, C., Zhang, H., & Liu, D. (2020). A Contrastive Study of Machine Learning on Energy Firm Value Prediction. *IEEE Access*, 8, 11635–11643. <https://doi.org/10.1109/ACCESS.2019.2953807>

Zohrehvand, A., Doshi, A. R., & Vanneste, B. (2021). Generalizing Event Studies Using Synthetic Controls: An Application to the Dollar Tree–Family Dollar Acquisition. *SSRN Electronic Journal*.
<https://doi.org/10.2139/ssrn.3856879>

8. Appendix

8.1. Literatuur studie

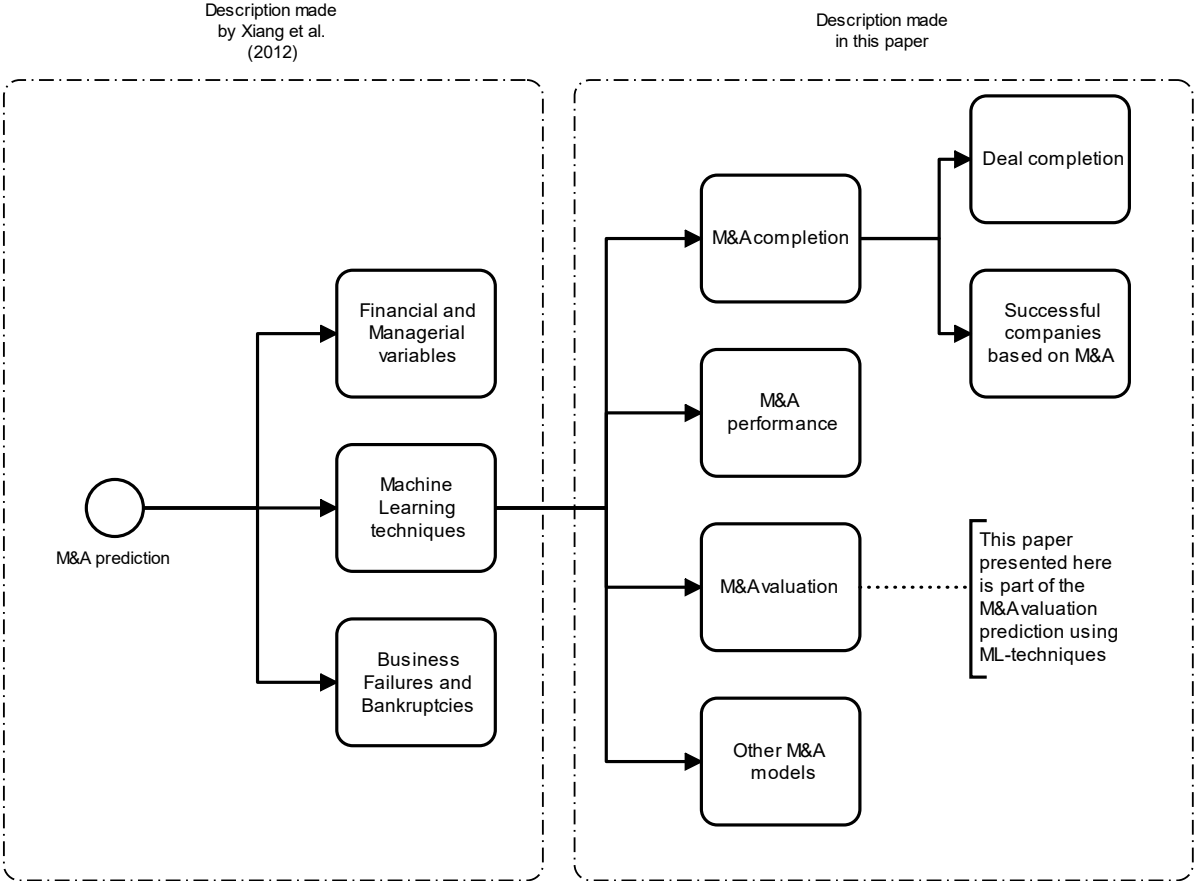


Figure 18: Visual representation of literature review

Table 1: Overview of literature using M&A prediction using machine learning

| Auteurs | Year | Title | Goal | Technique | Algorithm(s) |
|-------------------------|------|--|-------------------------------|-------------------------------|---|
| Ang et al | 2020 | Using Machine Learning to Demystify Startups Funding, Post-Money Valuation, and Success | M&A Valuation | Regression | Elastic nets, XGBoost and neural networks |
| Arroyo et al | 2019 | Assessment of machine learning performance for decision support in venture capital investments | Deal Completion | Classification | Tree based methods and SVM |
| Craninckx & Huyghebaert | 2011 | Can stock markets predict M&A failure? A study of European transactions in the fifth takeover wave | M&A Performance | Regression and Classification | Linear and logistic regression |
| Karatas & Hirs | 2021 | Predicting Status of Pre and Post M&A Deals Using Machine Learning and Deep Learning Techniques | Deal Completion | Classification | Logit, neural networks |
| Kellner | 2012 | The Measure of a MAC: A Quasi-Experimental Protocol for Tokenizing Force Majeure Clauses in M&A Agreements | Legal document identification | NLP | Latent semantic analysis |
| Lee et al | 2020 | Unbalanced data, type II error, and nonlinearity in predicting M&A failure | Deal Completion | Classification | Logit, probit and neural networks |
| Li | 2018 | Corporate Culture and Merger Success: Evidence from Machine Learning | Corporate culture prediction | NLP | Neural network |

| | | | | | |
|--------------------|------|---|-------------------------------|----------------|---|
| Pan et al | 2018 | Machine Learning Prediction of Companies ' Business Success | Deal Completion | Classification | Logistic regression, random forest, KNN |
| Roegiest et al | 2018 | A dataset and an examination of identifying passages for due diligence | Legal document identification | NLP | Conditional random fields, SVM |
| Seng & Lai | 2010 | An intelligent information segmentation approach to extract financial data for business valuation | M&A Valuation | NLP | / |
| Shao | 2019 | Categorization of Mergers and Acquisitions in Japan Using Corporate Databases: A Fundamental Research for Prediction | M&A Categorization | Clustering | K-means |
| Tsagkanos et al | 2007 | Predicting Greek mergers and acquisitions: a new approach | Deal Completion | Classification | Decision Tree |
| Vayas-Ortega et al | 2020 | on the differential analysis of enterprise valuation methods as a guideline for unlisted companies assessment (II): Applying machine-learning techniques for unbiased enterprise value assessment | M&A Valuation | Regression | Linear regression, tree based methods, SVM and gaussian process |
| Wang et al | 2020 | Does the interplay between the personality traits of CEOs and CFOs influence corporate mergers and acquisitions intensity? An econometric analysis with machine learning-based constructs | Personality prediction | NLP | Attention based model |

| | | | | | |
|----------------------|------|---|--------------------|------------------------|--|
| Wei | 2019 | Machine learning applications in finance: some case studies | M&A Performance | Regression | Linear regression, random forest, SVM and neural networks |
| Xiang et al | 2012 | A supervised approach to predict company acquisition with factual and topic features using profiles and news articles on TechCrunch | Deal Completion | NLP, Classification | Bayesian networks, logistic regression and SVM |
| Żbikowski & Antosiuk | 2021 | A machine learning, bias-free approach for predicting business success using Crunchbase data | Deal Completion | Classification | Logistic regression, XGBoost, SVM |
| Zhang et al | 2020 | A Contrastive Study of Machine Learning on Energy Firm Value Prediction | M&A Valuation | Regression | KNN, decision tree, SVR, neural networks, adaboost and random forest |

8.2. Dataset

De dataset die in dit onderzoek gebruikt wordt is een openbare dataset van Crunchbase. Zij publiceerde deze in 2013 onder de naam 'Crunchbase 2013 snapshot'. Deze dataset bevat informatie over startups en hun betrokkenheid bij Mergers and Acquisitions. De startups worden beschreven aan de hand van verschillende tabellen. In dit onderzoek maken we gebruik van de tabellen omtrent: acquisitions, degrees, investments, offices, people, relationships en objects.

Deze tabellen liet ons toe om de acquisitions te beschrijven aan de hand van deal specifieke kenmerken, kenmerken van de acquirer en kenmerken van de acquiree. Concreet waren dit volgende variabelen:

Table 2: Variables of Crunchbase 2013 snapshot

| Variabelen naam | Type | Omschrijving |
|-----------------------------|--------|--|
| Price_amount | Int | Prijs van de overname of fusie |
| Price_currency_code | String | De munteenheid waarin de prijs is uitgedrukt |
| Acquired_at | Date | De datum van de overname of fusie |
| Name_acquirer | String | Naam van de koper |
| Acquirer_industry | String | Industrie van de koper |
| Acquirer_country | String | Land van de koper |
| Acquirer_region | String | Regio van de koper |
| Acquirer_investment_rounds | Int | Het aantal investeringsrondes die een acquirer vervult heeft |
| Acquirer_invested_companies | Int | Het aantal bedrijven waarin de koper geïnvesteerd heeft |
| Acquirer_funding_rounds | Int | Het aantal financieringen rondes die een koper verkregen heeft |
| Acquirer_total_funding | Int | De hoeveelheid financiering die een koper verkregen heeft |
| Acquirer_milestones | Int | Het aantal mijlpalen dat een koper behaald heeft |
| Acquirer_relationships | Int | Het aantal relaties dat een koper heeft |

| | | |
|-----------------------------|--------|---|
| Acquirer_founded_year | Int | Het jaar waarin de koper opgericht werd |
| Acquirer_offices | Int | Het aantal kantoren van de koper |
| Name_acquired | String | Naam van de overgenomen onderneming |
| Acquired_industry | String | Industrie van de overgenomen onderneming |
| Acquired_country | String | Land van de overgenomen onderneming |
| Acquired_region | String | Regio van de overgenomen onderneming |
| Acquired_investment_rounds | Int | Het aantal investeringsrondes die een overgenomen onderneming vervult heeft |
| Acquired_invested_companies | Int | Het aantal bedrijven waarin de overgenomen onderneming geïnvesteerd heeft |
| Acquired_funding_rounds | Int | Het aantal financieringen rondes die de overgenomen onderneming verkregen heeft |
| Acquired_total_funding | Int | De hoeveelheid financiering die de overgenomen onderneming verkregen heeft |
| Acquired_milestones | Int | Het aantal mijlpalen dat de overgenomen onderneming behaald heeft |
| Acquired_relationships | Int | Het aantal relaties dat overgenomen onderneming heeft |
| Acquired_founded_year | Int | Het jaar waarin de overgenomen onderneming opgericht werd |

| | | |
|------------------------|--------|--|
| Acquired_offices | Int | Het aantal kantoren van de overgenomen onderneming |
| Funds_backing_acquired | Int | Het aantal investeringsfondsen die geïnvesteerd hebben in de overgenomen onderneming |
| Highest_degree_founder | String | Het hoogste diploma dat behaald werd door één van de oprichters |

8.3. Feature engineering

Table 3: Exchange rates for the mapping of deal prices

| Currency | Exchange rate |
|----------|---------------|
| EUR | 1.3791 |
| CAD | 0.940018 |
| GBP | 1.654192 |
| JPY | 0.009529 |
| AUD | 0.894184 |
| SEK | 0.15567 |

Table 4: Industry to sector mapping

| Sector | Industry |
|------------------------|---|
| IT | Web, search, software, hardware, social, semiconductor, analytics |
| Communication | News, video games, mobile, network hosting, messaging, music, photo and video |
| Consumer discretionary | Advertising, automotive, travel, fashion, hospitality, local |
| Industrials | Public relations, consulting, transportation, legal, manufacturing, nanotech |
| Financials | Finance, real estate |
| Health care | Biotech, medical, health |
| other | Other, enterprise, security, cleantech, education |

8.4. Verkennde analyse

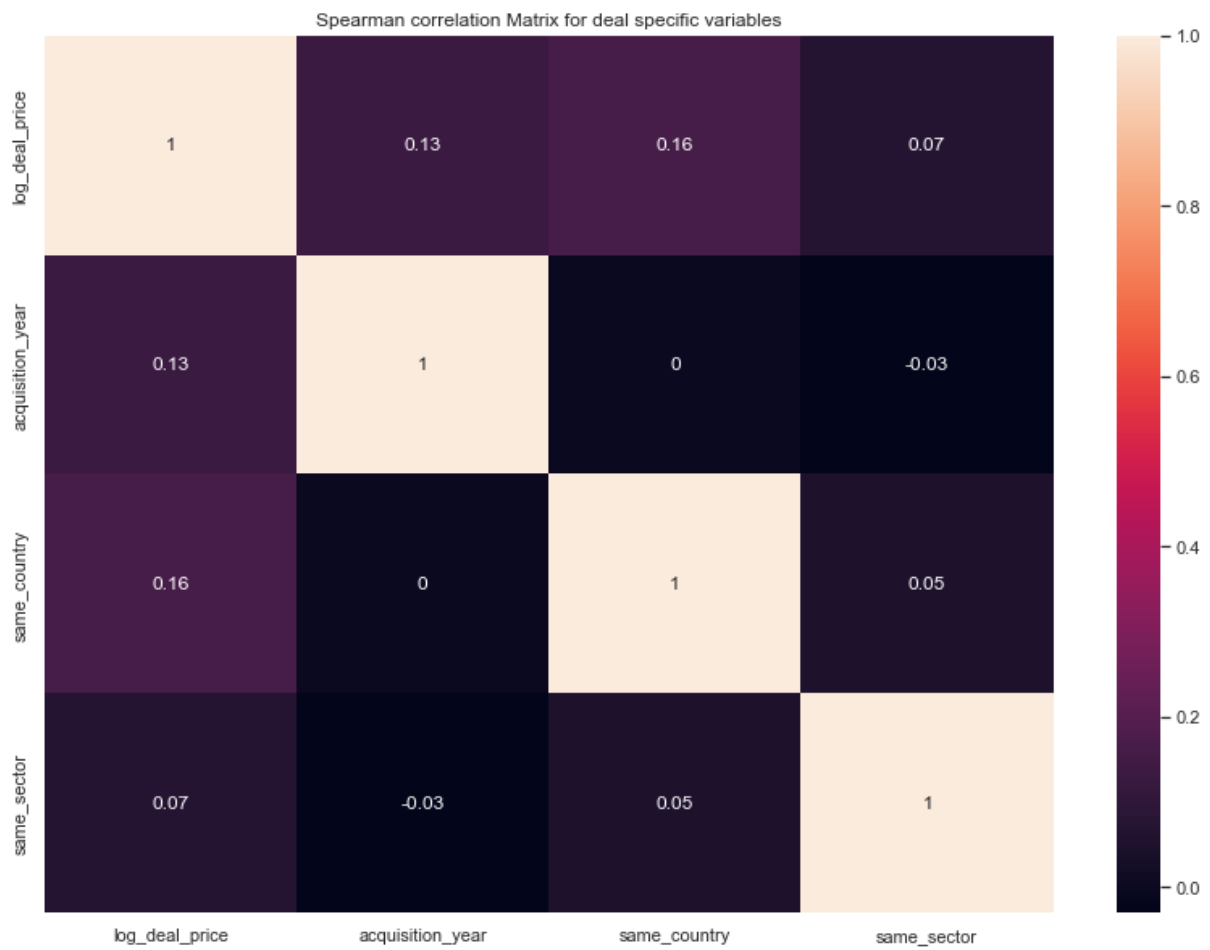


Figure 19: Spearman correlation matrix for deal specific variables

(values equal or above 0.04 are significant at $p < 0.05$)

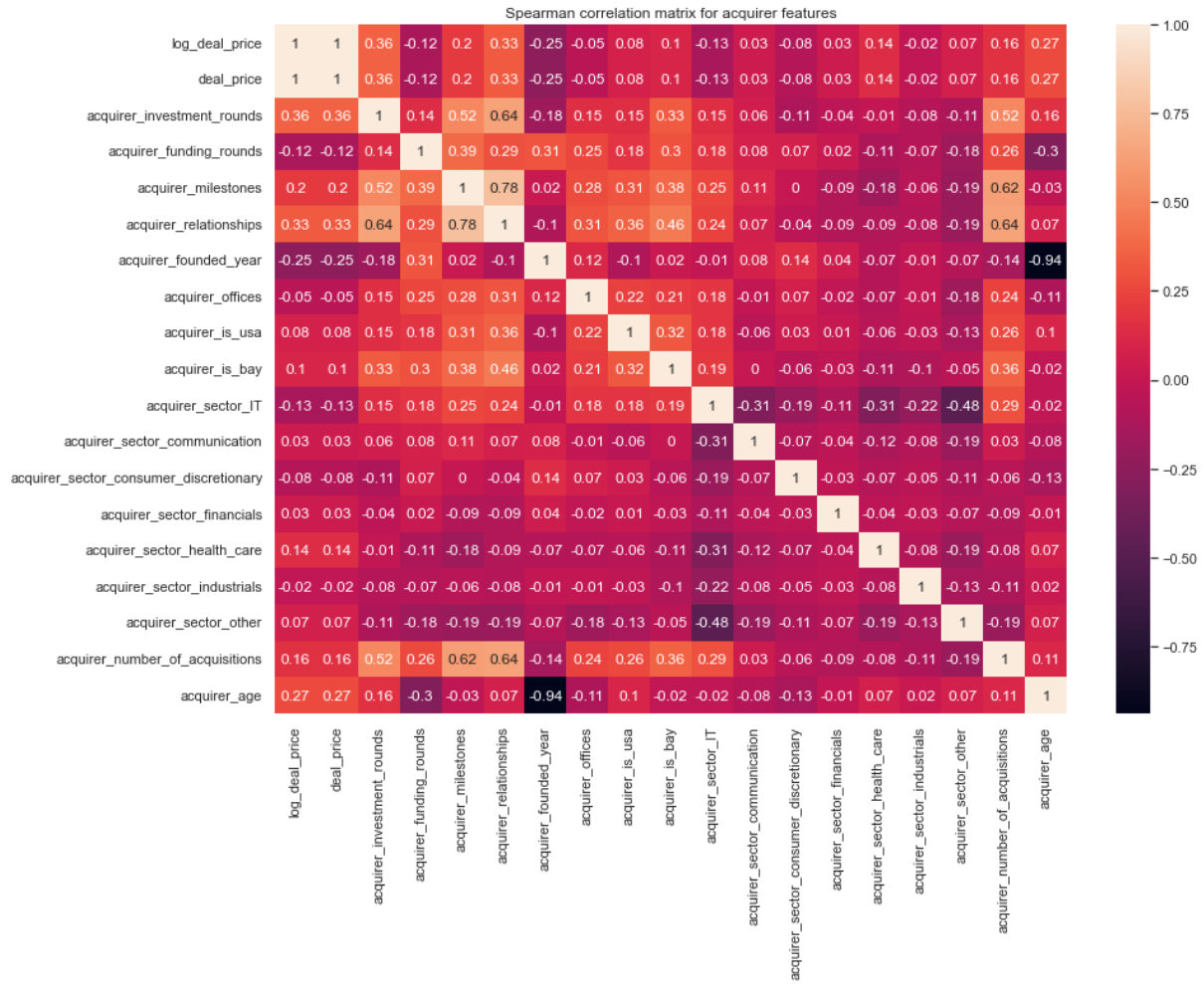


Figure 20: Spearman correlation matrix for acquirer features
 (values equal or above 0.04 are significant at $p < 0.05$)

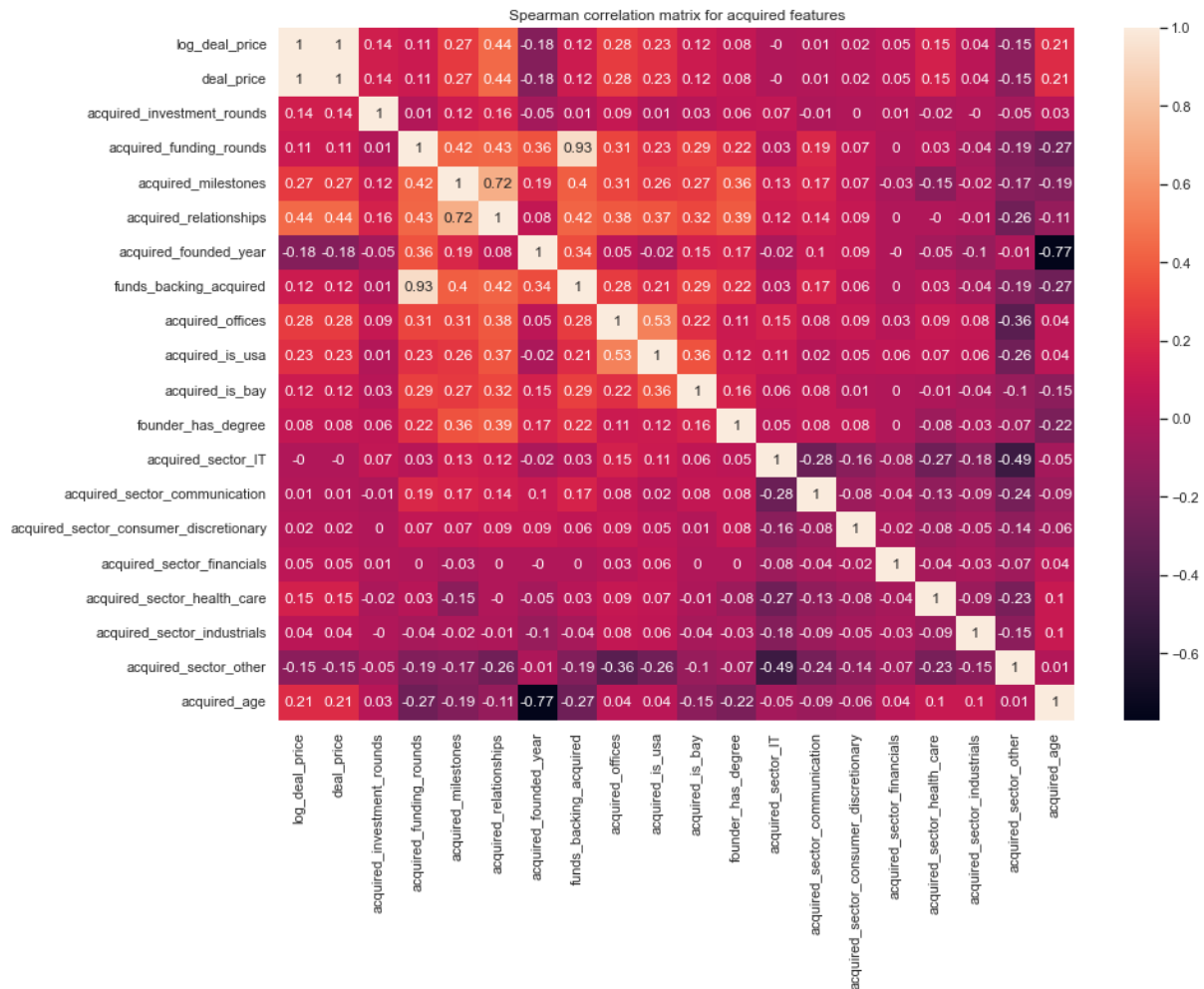


Figure 21: Spearman correlation matrix for acquiree features
(values equal or above 0.04 are significant at $p < 0.05$)

8.5. Hyperparameter tuning

8.5.1. SVR

Table 5: Hyperparameters for SVR

| Hyperparameters | Hyperparameter range | Best model |
|-----------------|----------------------|------------|
| Kernel | Linear, rbf, sigmoid | Linear |
| Gamma | 0.00001 - 1 | 0.0007 |
| C | 0.01 – 100 000 | 99 784 |
| epsilon | 0.001 - 100 | 0.212 |

8.5.2. XGBoost

Table 6: Hyperparameters for XGBoost

| Hyperparameters | Hyperparameter range | Best model |
|-------------------|----------------------|------------|
| Eta | 0.01 – 0.5 | 0.014 |
| Gamma | 0.1 – 1 000 | 32.95 |
| Max_depth | 1 – 10 | 9 |
| min_child_weight | 1 – 10 | 1 |
| Subsample | 0.5, 0.75, 1 | 0.5 |
| Colsample_bylevel | 0.5, 0.75, 1 | 0.5 |
| lambda | 1 - 10 | 7 |

8.5.3. Neural network

Table 7: Hyperparameters for neural network

| Hyperparameters | Hyperparameter range | Best model |
|---------------------|----------------------|------------|
| Num_layers | 1 - 3 | 1 |
| Units | 10 - 50 | 37 |
| Regularization (L2) | 0.00001 – 10 | 0.00001 |
| Optimizer | Adam, sgd, rmsprop | Rmsprop |
| Learning_rate | 0.0001 – 0.1 | 0.1 |

8.6. Resultaten

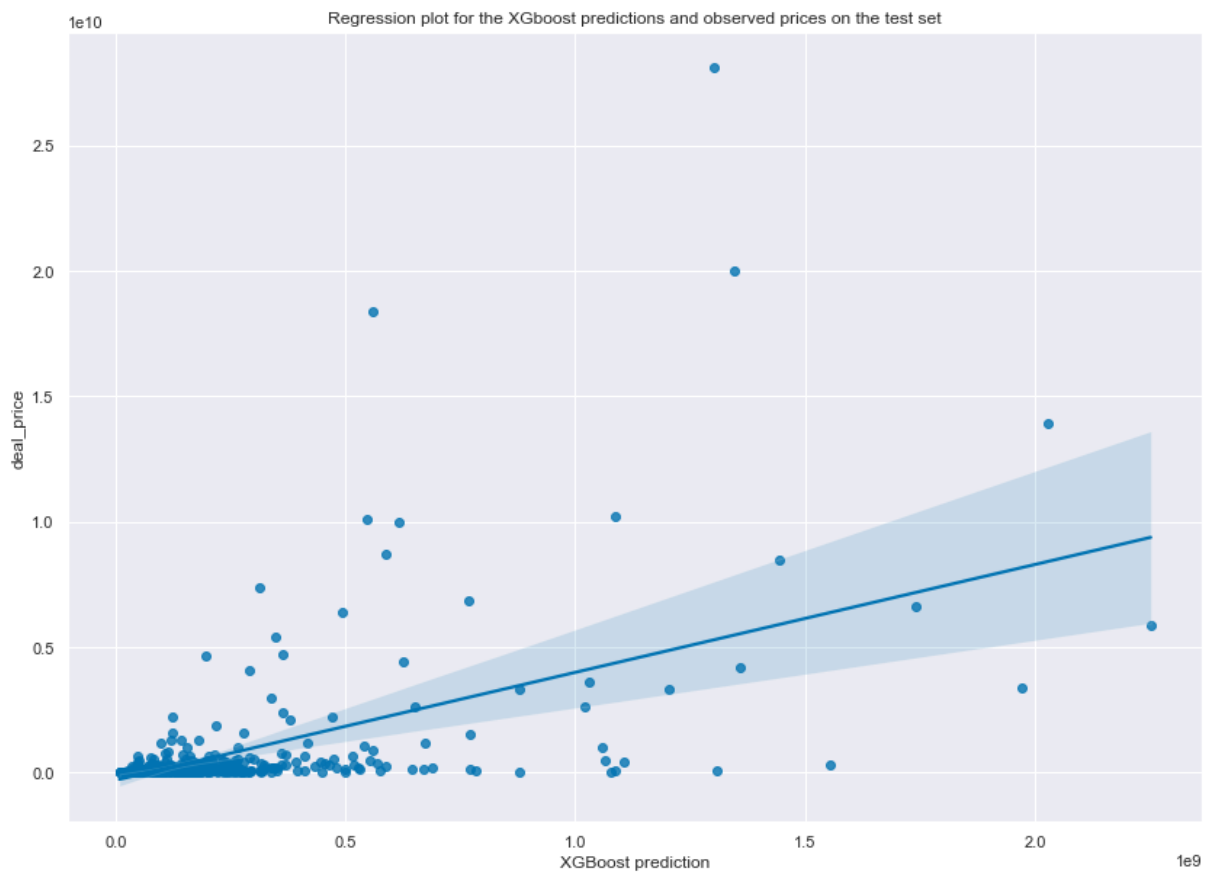


Figure 22: Regression plot for the XGBoost predictions and observed deal price on the test set

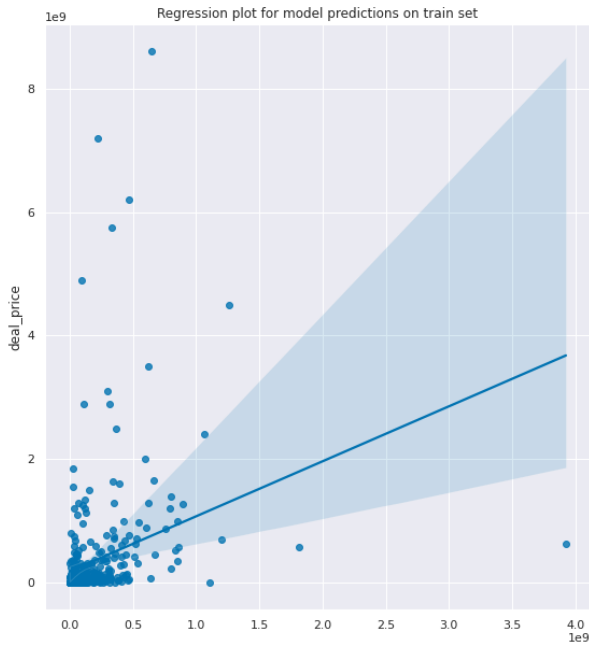


Figure 24: Regression plot for neural network predictions and observed deal price on the train set

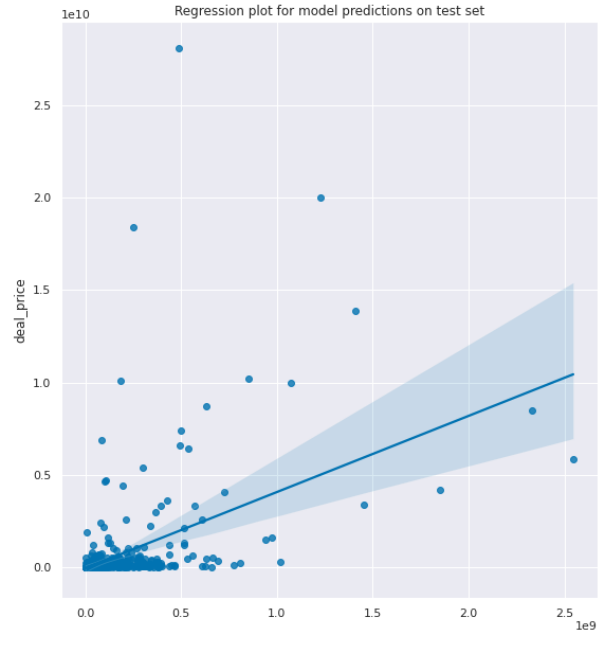


Figure 23: Regression plot for neural network predictions and observed deal price on the test set

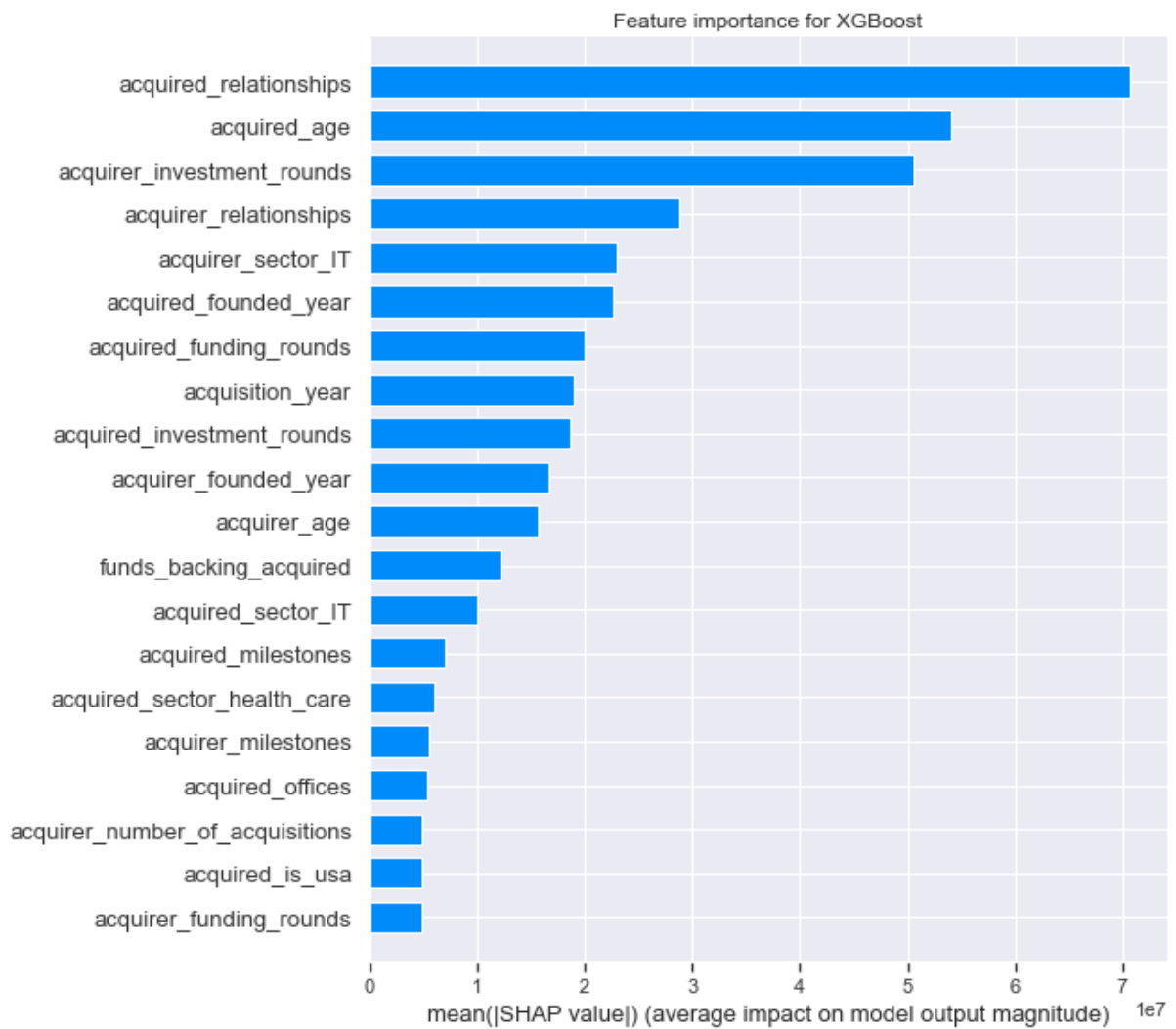


Figure 25: Feature importance of the XGBoost model

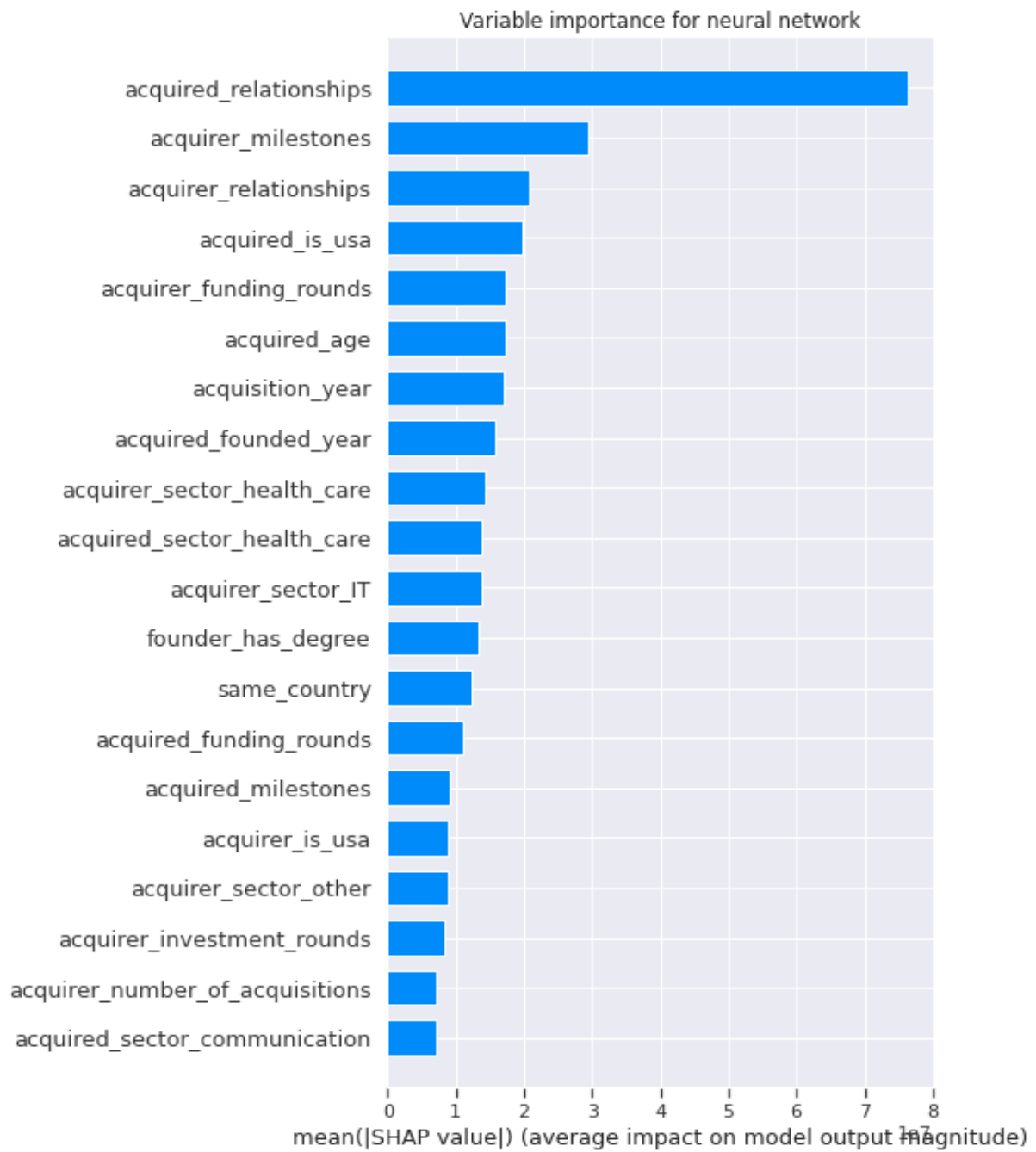


Figure 26: Feature importance for neural network

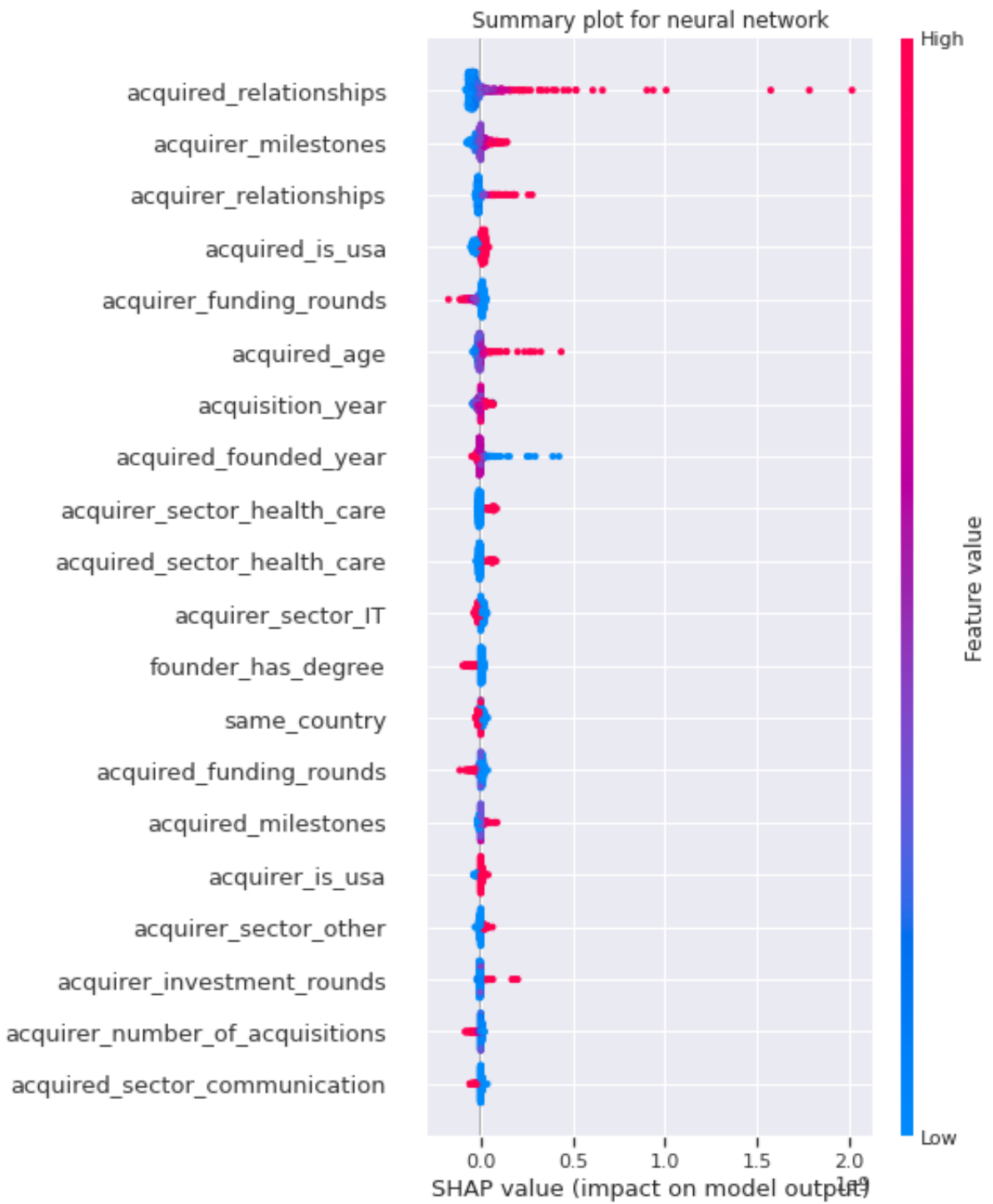


Figure 27: Summary plot for feature impact on predictions of neural network