Paper: Onderzoeksvoorstel

Data-analyse van historische verkoop data voor het verminderen van voedselverspilling in supermarkten.

Research Methods, 2024-2025

Renaud Vermeiren en Kenzo Cherroudi

E-mail: renaud.vermeiren@student.hogent.be en Kenzo.Cherroudi@student.hogent.be
Project repo: https://github.com/HoGentTIN/paper-research-methods-nl-24-25-rmvermeirencherroudi.git

Samenvatting

Supermarkten verspillen jaarlijks grote hoeveelheden verse producten zoals groenten en fruit. Een nauwkeurige voorspelling van de vraag naar deze producten kan helpen om overschotten en dus verspilling te beperken. Om te achterhalen of machine learning-modellen geschikt zijn voor vraagvoorspelling, werd een vergelijking gemaakt van verschillende machine learning-technieken. De analyse bestaat uit een literatuurstudie waarin ML-modellen beoordeeld werden op accuraatheid op basis van de absolute mean error en snelheid. Dit resulteert in een overzicht van de algoritmen die ingezet kunnen worden voor dit doel. Uit de resultaten blijkt dat boosting, stacking, algoritme-ensemble en modeloptimalisatie worden gezien als de meest effectieve strategieën voor de verbetering van ML-methoden. Deze bevindingen tonen aan dat supermarkten met behulp van goed gekozen ML-technieken hun voorraadbeheer aanzienlijk kunnen verbeteren. Zo leveren de voorspellingen van deze modellen niet alleen economische voordelen op voor de supermarkten, maar dragen ze ook bij aan het verminderen van voedselverspilling.

Keuzerichting: AI & Data Engineering
Sleutelwoorden: Data-analyse, Machine learning

Inhoudsopgave

1. Inleiding

In 2022 verspilde de wereld naar schatting 1,05 miljard ton voedsel in de voedseldiensten- en huishoudelijke sectoren gecombineerd. Dit komt neer op 132 kilogram per persoon over een volledig jaar (Programme, 2024). En supermarkten spelen hier een grote rol in. Volgens Programme (2024) verspillen winkelketens jaarlijks ongeveer 131 miljoen ton voedsel. Want zij hebben het probleem dat er dagelijks grote hoeveelheden onverkochte groenten en fruit worden weggegooid.

Ondanks inspanningen op het vlak van voorraadbeheer en promoties, blijft voedselverspilling een hardnekkige uitdaging. Het gaat hierbij niet alleen om economische verliezen, maar ook om een negatieve impact op duurzaamheid en bedrijfsimago. En het onderzoek van Teller e.a. (2018) toont aan dat een van de grootste oorzaken van voedselverspilling in supermarkten de beperkte voorspelbaarheid van de klantvraag over het hele assortiment is en onvoorspelbaar klantgedrag bij het selecteren of hanteren van producten.

Dit onderzoek wil gebruikmaken van historische verkoopdata van de winkelketens van Lidl in Vlaanderen, alsook externe factoren zoals weer en feestdagen. Om te proberen deze klantvraag naar verse producten zoals groenten en fruit zo goed mogelijk te gaan modelleren. Op deze manier wordt er een antwoord gezocht op de onderzoeksvraag: “Hoe kan machine learning gebruikt worden om op een geschikte manier de vraag naar groenten en fruit te voorspellen bij de supermarkten van Lidl, en hoe kunnen deze bijdragen aan het verminderen van voedselverspilling?”.

Dit wordt gedaan door de volgende deelvragen te beantwoorden:

- Welke factoren hebben de meeste invloed op de vraag naar groenten en fruit?
- Hoe presteren verschillende machine learning-modellen op het vlak van nauwkeurigheid

door de mean absolute error te gebruiken als meetstaaf voor de accuraatheid en snelheid in deze opdracht?

- Welke gegevens zijn minimaal nodig om een model te trainen dat even goed doet als de traditionele manier om de vraag naar groenten en fruit te voorspellen?

Het doel is om te beoordelen in welke mate machine learning-technieken geschikt zijn voor het voorspellen van de vraag naar verse producten zoals groenten en fruit in supermarkten, met als uiteindelijk doel het verminderen van voedselverspilling door efficiënter voorraadbeheer. Om de kwaliteit van de modellen te meten, zal de Mean Absolute Percentage Error gebruikt worden.

2. Literatuurstudie

2.1. Impact van voedselverspilling

In supermarkten wordt nog steeds veel voedsel weggegooid. Jaarlijks gaat wereldwijd 13 procent van al het voedsel verloren voordat het ooit bij de consument terechtkomt. Dat blijkt uit een rapport van het United Nations Environment Programme uit 2024 (Programme, 2024).

Het gaat hier om producten die snel bederven en dus snel verkocht moeten worden. Denk aan fruit, groenten en vlees. Maar deze producten snel verkopen lukt niet altijd. Volgens Teller e.a. (2018) komt dit vooral door een slechte inschatting van de vraag, beperkte houdbaarheid of een slechte logistiek.

Voedselverspilling heeft ook economische gevolgen, ze leiden tot financiële verliezen bij supermarkten. Volgens Teller e.a. (2018) zijn inefficiënte voorraadbeheerstrategieën en slecht ingeschatte klantvraag belangrijke oorzaken voor deze verliezen. Hun studie geeft ook aan dat supermarkten die hier onvoldoende rekening mee houden, hogere afvalpercentages genereren

Wanneer een product niet verkocht wordt in de supermarkt, zorgt dit voor een kettingreactie. Volgens Pilarski e.a. (2024) zorgt dit ook voor een verspilling van arbeid, transportkosten en voorraadkosten bij bedrijven vóór het product in de supermarkt terechtkomt. De studie combineert simulatie en reinforcement learning en toont aan dat met geoptimaliseerde voorraadbeheerstrategieën de kosten met meer dan 20% kunnen worden verlaagd.

In een studie van Visschers e.a. (2017) wordt een andere interessante stelling besproken, namelijk dat geld motiveert. Hun onderzoek toont dat wanneer mensen beseffen hoeveel voedselverspilling eigenlijk kost, dit motiveert om verspilling tegen te gaan. Ze laten zien dat kleine aanpassingen zoals aangepaste portiegroottes en betere promoties verspilling kunnen minimaliseren, en minder verspilling betekent meteen minder kosten.

Naast de financiële impact heeft voedselverspilling een nog grotere impact op het milieu. Voedsel dat niet wordt opgegeten, werd nog altijd geproduceerd en draagt dus alsnog bij aan de broeikasgasemissies. Denk aan de productie en het transport. Volgens het UNEP Food Waste Index Report is voedselverspilling wereldwijd verantwoordelijk voor ongeveer 8 à 10 procent van de totale uitstoot van broeikasgassen (Programme, 2024).

Teller e.a. (2018) voegen daaraan toe dat supermarkten vaak indirect bijdragen aan milieuschade door onrealistische kwaliteitsnormen, die ertoe leiden dat perfect eetbaar voedsel niet wordt verkocht. Ook zorgen inefficiënte winkelindelingen en koelstrategieën voor extra energieverbruik.

Om voedselverspilling te verminderen, moeten we inzicht krijgen in de factoren die de klantvraag beïnvloeden. Want een groot deel van de verspilling is het gevolg van een onevenwicht tussen aanbod en vraag.

2.2. Factoren die de vraag beïnvloeden

Er zijn veel factoren die invloed hebben op de vraag naar een product; denk aan het effect van het weer op het koopgedrag van consumenten. Of welke promoties er gegeven worden en op welke producten. Ook feestdagen kunnen invloed hebben (Liu e.a., 2021).

Er wordt eerst gekeken naar de invloed die het weer heeft op het koopgedrag. Uit onderzoek van Liu e.a. (2021) blijkt dat op warme dagen meer producten zoals ijs en koude dranken worden verkocht dan op koude dagen. Op regenachtige dagen zien we dat er minder winkelbezoekers zijn, maar de gemiddelde besteding per bezoek is wel hoger. Ook het seizoen speelt een rol, in de zomer zijn sommige producten populairder dan in de winter en omgekeerd. Het weer heeft dus inderdaad een impact op het koopgedrag.

Ook de impact van promoties is groot, en heeft invloed op de korte, maar ook lange termijn. Uit een onderzoek van Dai e.a. (2017) blijkt dat op korte termijn promoties de aankoopkans van het gepromote product vergroten en zo meer gekocht wordt. Op lange termijn moeten we rekening houden met het feit dat wanneer het herhaaldelijk dezelfde promoties zijn, klanten dit opmerken en leren wachten in plaats van de volle prijs te kopen. Ook moeten we de promoties van de concurrentie bijhouden, want promoties bij de

ene verkoper leiden tot minder aankopen bij de verkoper zonder promotie (Dai e.a., 2017).

Volgens Sacks en Zafar (2022) heeft het seizoen ook een impact. Veel productcategorieën tonen duidelijke seizoensgebonden schommelingen in de vraag. Dit komt doordat meer consumenten het product willen kopen, niet omdat bestaande klanten opeens meer kopen.

Gezien het grote aantal factoren dat de klantvraag beïnvloedt, zoals weer, promoties en seizoenen. Wordt het duidelijk hoe complex het voorspellen van deze vraag is. Daarom zijn traditionele voorspellingsmethoden vaak niet goed genoeg en zijn geavanceerdere technieken nodig.

2.3. Vraagvoorspelling: methoden en uitdagingen

Het is van groot belang voor een efficiënt voorraadbeheersysteem dat de voorspelling van de klantvraag zo nauwkeurig mogelijk is. Een foute inschatting kan leiden tot te veel voorraad en dus voedselverspilling, of tot te weinig voorraad, wat kan leiden tot financieel verlies (Teller e.a., 2018).

Chopra en Meindl (2016) hield een studie naar de traditionele vraagvoorspellingsmethoden, zoals de *moving average* en *exponential smoothing*. Deze zijn redelijk makkelijk te implementeren en worden vaak gebruikt bij supermarkten. Het nadeel van deze traditionele methoden is dat ze geen rekening houden met de externe factoren die we hierboven besproken hebben. Denk aan de impact van warme of koude dagen uit de studie van Liu e.a. (2021), dit is moeilijk accuraat te modelleren met klassieke methoden.

Er zijn dus veel uitdagingen waarmee rekening gehouden moet worden. Een van de grootste is het onvoorspelbare gedrag van consumenten; daarnaast is er het probleem van *demand cannibalization*. Dit probleem stelt dat wanneer er promotie wordt gegeven op een product, dit kan leiden tot een lagere vraag voor gelijkaardige producten (Dai e.a., 2017).

Ook geeft Pilarski e.a. (2024) aan dat een van de grootste problemen de kwaliteit van de historische verkoopdata is. Zaken zoals onvolledige data of foutieve registraties zorgen ervoor dat de modellen getraind worden op foutieve input.

Traditionele methoden zijn vaak niet in staat om zaken zoals het seizoen te betrekken doordat ze vaak uitgaan van lineaire trends en hebben het moeilijk om herhalende schommelingen te verwerken (Sacks & Zafar, 2022).

Aangezien klassieke methoden tekortschieten, onderzoeken we in de volgende sectie welke ML-technieken beter geschikt zijn.

2.4. Machine learning-technieken voor vraagvoorspelling

Een veelbelovend alternatief voor deze traditionele methoden is machine learning (ML). ML is in staat om patronen te herkennen in grote hoeveelheden gestructureerde en ongestructureerde data en op basis daarvan voorspellingen te doen zonder expliciet geprogrammeerde regels (Chopra & Meindl, 2016). Een nauwkeurige voorspelling van klantvraag maakt het mogelijk efficiënt in te kopen en dus verspilling tegen te gaan.

Volgens Pilarski e.a. (2024) presteren ML-modellen uitstekend bij het voorspellen van productvraag, mede dankzij hun vermogen om goed onregelmatigheden in data te verwerken. Denk aan onverwachte pieken tijdens feestdagen en de invloed van het weer op de verkoop. Uit het onderzoek bleek dat het toepassen van reinforcement learning de kosten met 20% verlaagde en ook de verspilling werd verminderd. Traditionele modellen, zoals lineaire regressie, verwachten lineaire trends, terwijl verkoopgegevens in supermarkten vaak niet-lineair zijn. Machine learning kan dit wel en zoekt non-lineaire relaties in de data. Liu e.a. (2021) laat zien dat ML-modellen zoals Random Forest duidelijk beter presteren dan klassieke voorspellers.

Overzicht van ML-technieken

Random Forest (RF):

Modellen zoals Random Forest zijn bijzonder geschikt voor het analyseren van tabulaire gegevens, zoals de dagelijkse productverkoop met kenmerken als promoties, seizoenen en weekdagen. Ook kunnen ze goed omgaan met ontbrekende gegevens in de data. Ze leveren hoge nauwkeurigheid en kunnen via feature importance aangeven welke kenmerken het meest bijdragen aan de voorspelling, al blijft het soms moeilijk te bepalen waarom het model sommige beslissingen neemt. Ook zijn ze minder geschikt voor tijdreeksen (Dai e.a., 2017).

LSTM-netwerken:

Recurrente neurale netwerken zijn hier beter in, vooral de Long Short-Term Memory-modellen. Deze zijn ideaal voor het herkennen van seizoensgebonden patronen en het voorspellen van promotie-effecten over tijd. Dai e.a. (2017) toonde aan dat LSTM-netwerken beter presteren dan ARIMA-modellen bij het voorspellen van wekelijkse verkoopcijfers over langere perioden. Deze netwerken onderscheiden zich doordat ze over een intern geheugen beschikken, waarmee ze informatie over op-

eenvolgende tijdstappen kunnen vasthouden en verwerken.

Reinforcement learning:

Pilarski e.a. (2024) beschrijft hoe reinforcement learning wordt gebruikt voor voorraadoptimalisatie, waarbij het model continu leert van feedback en zijn strategie aanpast. Het model evalueert continu factoren zoals verspilling, tekorten en voorraadkosten. Een belangrijke beperking is de aanzienlijke rekenkracht die nodig is voor training en uitvoering van deze modellen.

Pilarski e.a. (2024) benadrukt dat er ook uitdagingen horen bij machine learning. De eerste uitdaging is de kwaliteit van de data, het is moeilijk goede en volledige data te vinden. Ook de beschikbaarheid van gegevens vormt een probleem. Verschillende supermarkten hebben geen volledige historie van verkoopdata. Daarnaast vormt overfitting een veelvoorkomend probleem bij ML, dit gebeurt wanneer het model zich te sterk aanpast aan de trainingsdata, waardoor het niet goed reageert op nieuwe situaties. Denk aan onverwachte storingen zoals een pandemie of economische crisis (Liu e.a., 2021).

De besproken ML-technieken tonen aan dat het mogelijk is om niet-lineaire en weersafhankelijke vraagschommelingen correct te modelleren. Hierdoor kunnen supermarkten verspilling doelgericht beperken en hun voorraadbeheer optimaliseren, op voorwaarde dat er wordt geïnvesteerd in infrastructuur en datakwaliteit.

3. Methodologie

Fase 1 De eerste fase is het verzamelen van de data en deze gestructureerd opslaan. Deze fase duurt ongeveer 20 uur. Dankzij Lidl hebben we al de historische verkoopdata van alle winkels in Vlaanderen. Deze data is opgeslagen in een csv-file met de volgende kolommen:

(i) datum, (ii) product-ID, (iii) verkochte hoeveelheid, (iv) prijs, (v) of ze in promotie stonden (ja/nee).

Deze data wordt opgeslagen in een MongoDB databank. De reden hiervoor is dat de systematisch literatuurstudie van Khan e.a. (2022) aan toont dat de grote variatie van gegevens tijdens de analyse ervoor zorgt dat efficiënte methoden voor data-analyse en kennisextractie van groot belang zijn tijdens het kiezen van een geschikt databasemanagementsysteem.

Verder zegt deze studie dat traditionele SQL-databases hierdoor niet het meest geschikt zijn om grote diverse datasets efficiënt te verwerken.

Hierdoor geven ze een voorkeur aan NoSQL-databases, omdat die een beter alternatief bieden. Niet enkel omdat ze het mogelijk maken

om grote hoeveelheden data op te slaan maar ook dat ze zeer flexibel zijn dankzij hun dynamische schema's en zo bieden ze een

hoge mate van schaalbaarheid en beschikbaarheid. Wat ideaal is voor deze analyse. Er is specifiek voor MongoDB gekozen, omdat deze open source is en specifiek ontworpen is voor een gedistribueerde omgeving en optimaal is voor JSON Khan e.a. (2022).

Dus omdat we hier met een document-oriented database werken is het belangrijk om de csv data eerst om te zetten naar JSON formaat en pas daarna de data toevoegen aan de databank. De weerdata komt van het Koninklijk Meteorologisch Instituut via https://opendata.meteo.be/.

Aan het einde van de eerste fase is er een werkende databank met alle benodigde data.

Fase 2 In de tweede fase gaan we de data opschonen en al een eerste verkenning van de dataset doen. Dit is een van de grootste fases en zal ongeveer 30 uur in beslag nemen.

Python wordt gekozen voor de data-analyse vanwege het uitgebreide ecosysteem aan gespecialiseerde bibliotheken, zoals Pandas, NumPy en scikit-learn, die efficiënte verwerking en analyse van gegevens mogelijk maken. Bovendien laat Python zich goed integreren met big data-platforms zoals Apache Spark en Hadoop. Ondanks bepaalde schaalbaarheidsuitdagingen ten opzichte van talen zoals Java of Scala, maken de eenvoud, flexibiliteit en brede ondersteuning Python tot een handig hulpmiddel in big data-contexten(Kabir e.a., 2024).

Alle rijen in de data die missende waarden hebben, worden verwijderd en de datums worden gestandaardiseerd naar een datetime-formaat.

Er worden ook wat nieuwe features gemaakt, namelijk de seizoenskenmerken (maand, weeknummer) en een extra lag-feature die de verkoop van de voorafgaande week neemt. Op het einde van de tweede fase is de data klaar om gebruikt te worden om machine learning modellen mee te trainen.

Fase 3 Tijdens de derde fase gaan we de modellen daadwerkelijk beginnen te trainen. Deze fase zal ongeveer 15 uur duren. Maar eerst splitsen we de data op in 80% trainingsdata en 20% testdata volgens de *TimeSeriesSplit* functie in de scikit-learn-module. En voor de reproduceerbaarheid wordt in elke functie waar het kan de *random_state* gelijkgezet aan *42*. Alle modellen worden getraind op een computer met een Ryzen 5 5600x CPU 64bit-architectuur, Nvidia RTX 3070 GPU op Windows 10 versie 22H2. Gebruik maken van Python versie 3.10.11 en Sklearn versie 1.6.1

Vervolgens worden er verschillende regressiemodellen getraind, zoals een Support Vector Regressiemodel en een Multiple Linear Regression-model. We vergelijken deze door gebruik te maken van de absolute mean error als evaluatiecriterium. Hierna worden grafieken gemaakt met behulp van de matplotlib- en seaborn-libraries in Python om de data mee te visualiseren. Op basis van deze resultaten worden de hyperparameters van de modellen aangepast en nieuwe features gemaakt om zo een beter resultaat te bekomen.

Op het einde van fase 3 zijn er verschillende werkende modellen die de vraag naar verse producten kan voorspellen.

4. Verwachte resultaten

Hoewel verwacht werd dat Support Vector Regressie in deze context het beste zal presteren qua nauwkeurigheid en generaliseerbaarheid, worden andere modellen niet uitgesloten. Het zou bijvoorbeeld kunnen dat een eenvoudiger model zoals een lineaire regressie vergelijkbare resultaten oplevert, of dat zelfs de nieuwere LSTM-modellen (Long Short-Term Memory neural networks)(Hochreiter & Schmidhuber, 1997) beter presteren bij sterk seizoensgebonden producten, als er voldoende historische data beschikbaar is.

Naast de voorspelde verkoophoeveelheden geven de resultaten ook inzichten in welke variabelen (features) de meeste impact hebben op de vraag. De feature "*in korting*" heeft de grootste invloed. Dit kan waardevolle informatie opleveren voor het voorraadbeheer en de marketingstrategie van de supermarkt om ervoor te zorgen dat de producten niet slecht worden.

Tot slot leidt het onderzoek tot een onderbouwd adviesrapport over welk ML-model het meest geschikt is voor deze casus, en hoe het kan worden geïntegreerd in een bestaande workflow.

5. Discussie, verwachte conclusie

De resultaten van dit onderzoek tonen aan dat machine learning-technieken een waardevolle aanvulling kunnen zijn voor supermarktketens zoals Lidl om de vraag naar verse producten nauwkeuriger te voorspellen en hun voorraadbeheer te optimaliseren. Dit zorgt voor minder overschotten en dus minder voedselverspilling. Ook krijgen we inzichten rond welke factoren de grootste invloed hebben op de vraag, zoals promoties en het weer. Dit kan supermarkten ook helpen betere marketingstrategieën te ontwikkelen door bijvoorbeeld kortingen op het juiste moment te geven.

Er zijn ook enkele beperkingen in dit onderzoek. De kwaliteit en volledigheid van de data zijn cruciaal voor het trainen van een goed model. Ook onverwachte gebeurtenissen zoals een pandemie kunnen invloed hebben.

Tot slot is een vervolgonderzoek mogelijk door de vraagvoorspelling uit te breiden naar andere productcategorieën.

Referenties

Chopra, S., & Meindl, P. (2016). *Supply Chain Management: Strategy, Planning, and Operation* (6de ed.). Pearson Education.

Dai, W., Zhang, D., & Dong, Y. (2017). How Do Price Promotions Affect Customer Behavior on Retailing Platforms? *SSRN Electronic Journal*, 1–32. https://doi.org/10.2139/ssrn.3029707

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Comput.*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Kabir, M. A., Ahmed, F., Islam, M. M., & Ahmed, M. R. (2024). Python For Data Analytics: A Systematic Literature Review Of Tools, Techniques, And Applications. *ACADEMIC JOURNAL ON SCIENCE, TECHNOLOGY, ENGINEERING MATHEMATICS EDUCATION*, *4*(4), 134–154. https://doi.org/10.69593/ajsteme.v4i04.146

Khan, W., Kumar, T., Cheng, Z., Raj, K., Roy, A. M., & Luo, B. (2022). SQL and NoSQL Databases Software architectures performance analysis and assessments – A Systematic Literature review. https://doi.org/10.48550/ARXIV.2209.06977

Liu, Y., Yu, Y., Chen, Y., & Wang, S. (2021). The impact of weather on consumer behavior and retail performance. *Transportation Research Part E: Logistics and Transportation Review*, *150*, 102327. https://doi.org/10.1016/j.tre.2021.102327

Pilarski, S., Sidhu, A., & Varró, D. (2024). Combining simulation and reinforcement learning to reduce food waste in food retail. *SIMULATION*, *101*(3), 267–285. https://doi.org/10.1177/00375497241299054

Programme, U. N. E. (2024, maart). Food Waste Index Report 2024. Think Eat Save: Tracking Progress to Halve Global Food Waste. Verkregen mei 21, 2025, van https://wedocs.unep.org/20.500.11822/45230

Sacks, D. W., & Zafar, B. (2022). Why Do Retail Prices Fall During Seasonal Demand Peaks? *RAND Journal of Economics*, *53*(3), 567–593. https://doi.org/10.1111/1756-2171.12490

Teller, C., Holweg, C., Reiner, G., & Kotzab, H. (2018). Retail store operations and food waste. *Journal of Cleaner Production*, *185*(5), 981–997. https://doi.org/https://doi.org/10.1016/j.jclepro.2018.02.280

Visschers, V. H. M., Wickli, N., & Siegrist, M. (2017). Sorting out food waste behaviour: A survey on the motivators and barriers of self-reported amounts of food waste in households. *Food Policy*, *63*, 119–129. https://doi.org/10.1016/j.foodpol.2016.12.001