GreenBox AI by Decerno

Emelie Chandni Jutvik
Mattias Festin

Artificiell Intelligens inom Decerno

2024-06-14

Contents

1 Pitch

Föreställ dig en AI-lösning som är enkel att integrera men kraftfull nog att lösa komplexa affärsproblem. Vår AI-toolchain, tillgänglig som ett NuGet- eller npm-paket, automatiserar databehandling, modellval, träning och verifiering, vilket minskar utvecklingstiden och säkerställer att kundens data förblir skyddad. Distribuerbar både on-premise och i molnet, stödjer den avancerade tekniker som transfer learning och federated learning. Få snabba prediktioner och insikter direkt från dina affärssystem med vår plug-and-play AI-lösning – allt utan att behöva bli en AI-expert.

1.1 Syfte

- Utveckla ett verktyg (toolchain) som förenklar inlärningsprocessen och arbetsflödet för att implementera AI-modeller i Decernos projekt.
- Erbjuda en specialistlösning som skiljer Decerno från andra aktörer på marknaden.
- Ge anställda på Decerno möjlighet att vara delaktiga i utvecklingen av en teknikmodul i framkant.
- Fungera som kvalitetssäkring för framtagna AI-modeller genom att verktyget ställer krav på arbetsflödet och datahanteringen.

1.2 Nytta

- Fler utvecklare inom Decerno får kunskap om AI.
- Fler utvecklare inom Decerno kan implementera AI-modeller i sina projekt.
- Utvecklingstiden för att implementera AI minskas.
- Möjligheten att bygga kundspecifika lösningar.
- Möjligheten att kapsla in lösningen och säkerställa att kundens data inte lämnar deras system.

1.3 Värde

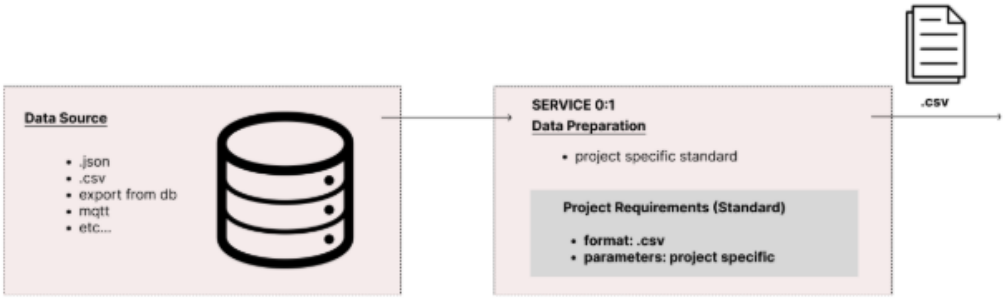
- En modul/toolchain som kan tillämpas inom flera projekt möjliggör att kostnader för vidareutveckling och förädling kan fördelas mellan projekten.

- Produkten kan användas som säljargument för att visa upp Decerno som ett företag som utvecklat en unik toolchain som förenklar vägen från idé till realisering av AI-modeller.
- Produkten kan utökas enkelt med fler modeller för att lösa fler problem i framtiden och per behov.
- Data pipeline-processen kan enkelt förändras och guardrails kan läggas till.

2 Befintlig Kodbas

2.1 Modul 1

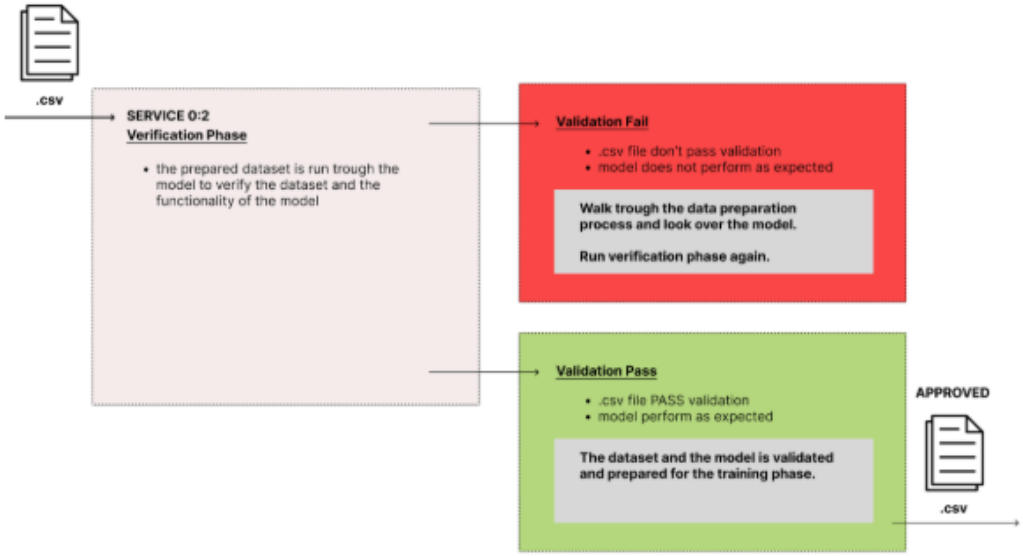
Hanterar olika typer av dataformat och levererar relevant data för beräkningen.



2.2 Modul 2

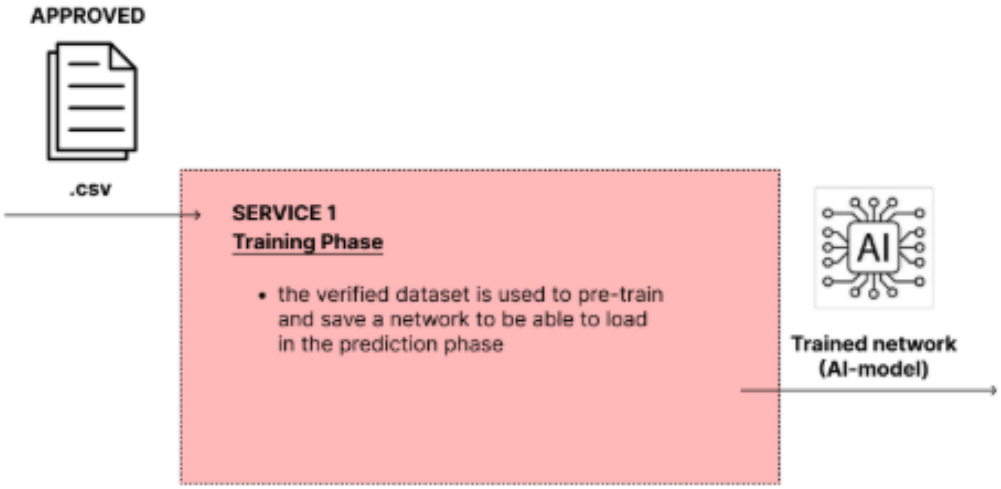
Läser data från Modul 1 och förbereder denna för att kunna läsas av ett neuralt nätverk. Den förberedda datan och nätverket genomgår ett valideringstest:

- Godkänd validering → Data och modell är redo för att skickas till Modul 3.
- Icke godkänd validering → Modellen behöver justeras och körs genom valideringstest tills dess att önskat resultat uppnås.



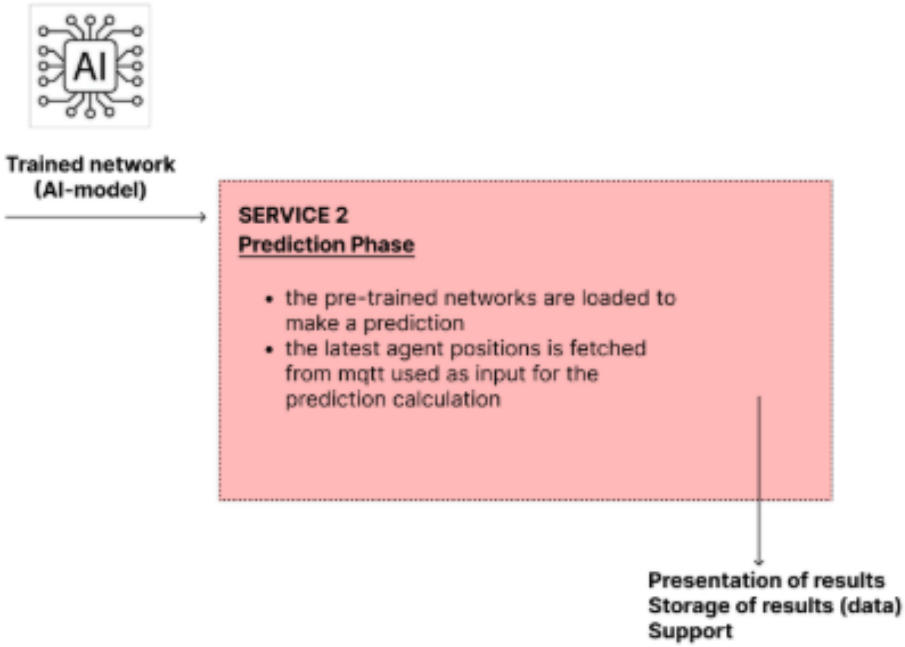
2.3 Modul 3

Det neurala nätverket tränas och sparas.



2.4 Modul 4

Det sparade/tränade neurala nätverket används för att göra en prediktion. Resultatet skickas i form av rådata. Presentationen av resultatet och jämförelsen med verklig data görs lokalt.



3 Paketering

Planen är att använda en existerande AI-lösning som redan är beprövad och effektiv. Denna AI-lösning kommer att paketeras med ett lättanvänt API runt sig, vilket gör det enkelt för användare att integrera och använda AI-funktionaliteten i sina applikationer. API:et kommer att innehålla tydliga datastrukturer och funktioner, vilket gör det möjligt att snabbt komma igång med AI-lösningen.

De definierade datastrukturerna är viktiga för nästa steg, där vi strävar efter en förbättrad AI-process. Detta inkluderar att tillhandahålla verktyg för automatisk modellval, hyperparameter-justering och prediktiva analyser i realtid. Genom denna paketering kan vi erbjuda en robust och skalbar AI-lösning som enkelt kan anpassas och utökas med tiden, men vi kan skapa värde så snabbt som möjligt.

3.1 MVP (Minimum Viable Product)

- Hantering av olika typer av indata-filer (.json, .csv till att börja med).
- Användargränssnitt som hanterar:
  - Uppladdning av indata-fil.
  - Val av neuralt nätverk.
  - Specifikation av in-parametrar som ska användas för beräkningen (string).
  - Specifikation av ut-parameter (string).
  - Inställning av hyper-parametrar.
  - Presentation av verifieringstest (plot).
  - Möjligheten att starta träning av ett nätverk.
  - Output $\rightarrow$ Tränad modell.

4 Förbättrad AI-process

4.1 Översikt

Vi utvecklar ett AI-system som använder olika specialiserade AI-modeller för att lösa specifika problem. Systemet består av flera moduler som samarbetar för att behandla indata, välja rätt AI-modell, generera en output och säkerställa att resultatet är godtagbart.

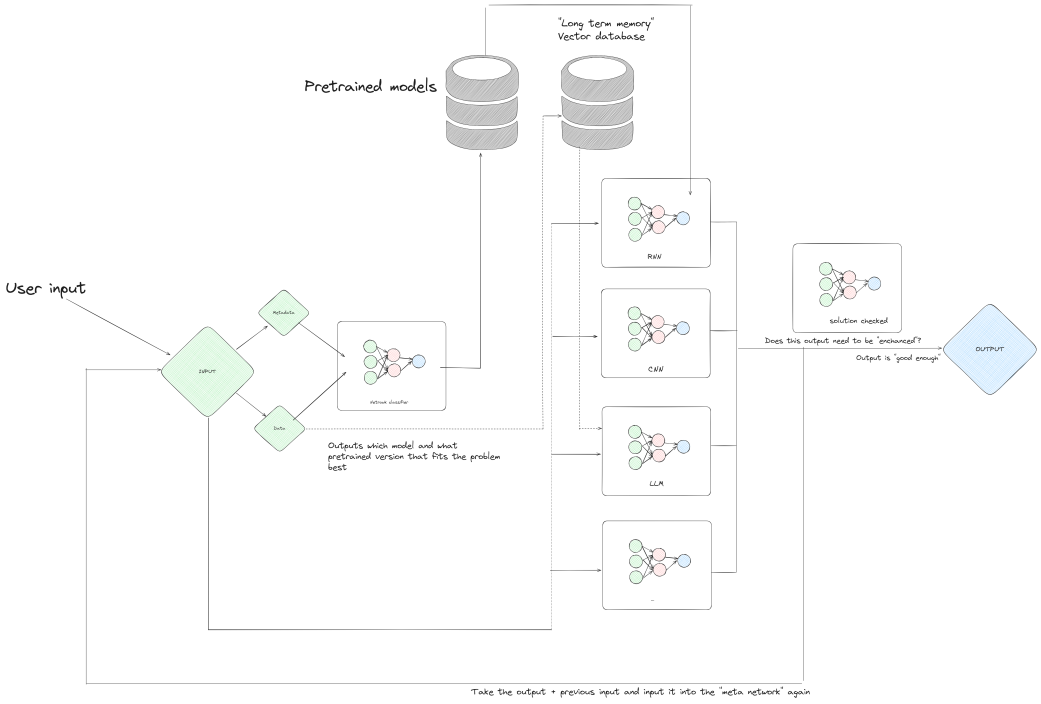
4.2 Steg-för-steg Process

1. **Indatamodul:**
    - **Funktion:** Behandlar och förbereder indata så att det passar den valda AI-modellen.
    - **Detaljer:** Hanterar olika typer av indata-format (t.ex. CSV, JSON) och transformerar dessa till ett format som är kompatibelt med AI-modellen. Om indata behöver en specifik representation, som text som behöver transformeras till vektorrepresentationer, utförs detta i detta steg. Vi kan använda tekniker som embeddings för detta.
2. **Konduktör-AI:**
    - **Funktion:** Klassificerar indatat och väljer rätt förtränad AI-modell för uppgiften.
    - **Detaljer:** En klassificeringsalgoritm avgör vilken AI-modell som är mest lämplig baserat på indatans natur. Exempelvis kan en RNN användas för tidsserieprognoser, en DNN för klassificeringsproblem, och en LLM för textbaserade uppgifter.
3. **AI-Exekvering:**
    - **Funktion:** Den valda AI-modellen körs och genererar en output.
    - **Detaljer:** Den valda AI-modellen tar den förberedda indata och bearbetar den för att producera ett resultat.
4. **Review-AI:**
    - **Funktion:** Utvärderar resultatet från AI-modellen och avgör om det är godtagbart.

- **Detaljer:** En sekundär AI-modell granskar resultatet för att säkerställa att det uppfyller förväntningarna. Om resultatet inte är tillfredsställande, omkörs processen med en annan AI-modell eller justeringar görs för att förbättra utfallet.

5. **Output-AI:**

   - **Funktion:** Formaterar resultatet i rätt format för presentation.
   - **Detaljer:** Beroende på behovet kan resultatet presenteras som text, rådata, bild, eller annan önskad form. Output-AI ser till att presentationen är korrekt och användbar.



4.3 Exempel på hjälpfunktioner

- **Vektordatabas:** För långtidsminne och effektiv lagring av embeddings, vilket är användbart för LLM-modeller som behöver hålla reda på historiska data.

5 Framtida Utveckling

- **Recursion och Augmentation:** Introducera mekanismer för att iterativt förbättra resultat genom att använda olika AI-modeller och justeringar i processen.
- **Transfer Learning:** En teknik där en modell tränad på en stor datamängd återanvänds och finjusteras för en specifik uppgift med mindre datamängd. Detta minskar träningskostnader och tid genom att dra nytta av befintlig kunskap.
- **Federated Learning:** En metod där modeller tränas över flera decentraliserade enheter eller servrar som innehåller lokala dataexempel, utan att dela data mellan enheterna. Detta förbättrar datasekretessen och möjliggör träning på känsliga data utan att de behöver överföras.

6 Nyttan med mindre AI-modeller

6.1 Energieffektivitet och miljöfördelar

I takt med att användningen av artificiell intelligens (AI) ökar, blir det allt viktigare att överväga de miljömässiga och energimässiga effekterna av dessa teknologier. Mindre AI-modeller erbjuder en rad fördelar när det gäller energieffektivitet och miljöpåverkan, vilket gör dem till ett attraktivt alternativ för hållbar utveckling.

6.2 Energieffektivitet

- **Minskad Beräkningskraft:** Mindre AI-modeller kräver betydligt mindre beräkningskraft jämfört med större modeller. Detta innebär att de kan köras på enklare hårdvara, vilket minskar energiförbrukningen per operation. Till exempel, en mindre modell kan ofta köras på en vanlig dator eller en mobil enhet, istället för att behöva kraftfulla GPU-kluster.

- **Snabbare Träning och Inferens:** Mindre modeller tränas och körs snabbare än större modeller. Detta reducerar den tid som datorer och servrar behöver vara igång, vilket i sin tur minskar den totala energikonsumtionen. Kortare träningstider innebär även att utvecklare kan iterera snabbare och effektivare, vilket ytterligare optimerar resursanvändningen.

6.3 Miljöfördelar

- **Lägre Koldioxidutsläpp:** Mindre energiförbrukning direkt översätts till lägre koldioxidutsläpp. Eftersom datacenter som kör AI-modeller ofta drivs av elektricitet från fossila bränslen, innebär varje watt som sparas en minskning av koldioxidutsläppen. Genom att använda mindre AI-modeller kan företag och organisationer minska sitt koldioxidavtryck och bidra till en mer hållbar framtid.

- **Minskade Krav på Hårdvara:** Mindre modeller kräver mindre avancerad och mindre energiintensiv hårdvara. Detta inte bara reducerar den omedelbara energiförbrukningen, utan minskar också behovet av att producera och underhålla högpresterande datacenter och hårdvarukomponenter, vilket ytterligare minskar miljöpåverkan.

- **Lokal Bearbetning:** Mindre modeller möjliggör lokal bearbetning av data, vilket minskar behovet av att skicka stora datamängder över nätverk till centrala datacenter. Lokal bearbetning minskar nätverksbelastningen och den energianvändning som är förknippad med datatransporter, vilket bidrar till en mer energieffektiv infrastruktur.

7 Exempelprojekt och Visioner

Det följande avsnitten presenterar två exempelprojekt som illustrerar hur vår AI-toolbox kan implementeras i olika branscher. Dessa projekt är visioner och representerar potentiella användningsområden för att demonstrera flexibiliteten och kraften i vår lösning. Genom att tillämpa vår AI-toolbox kan vi utveckla skräddarsydda lösningar för att lösa specifika problem och förbättra processer inom olika sektorer.

Vår AI-toolbox är utformad för att vara enkel att integrera med befintliga databaser i affärssystem, vilket möjliggör prediktioner och klassificeringar på befintliga data. Detta ”plug-and-play”-koncept innebär att företag snabbt kan komma igång med AI-driven analys och förbättra sina beslut utan omfattande anpassning eller utveckling.

7.1 Prediktion och underhåll av Malmbanan

7.1.1 Bakgrund

Malmbanan är en kritisk infrastruktur för Sverige och Norge, särskilt för transport till Nordnorge som saknar egen tågbana i regionen. Malmbanan har haft flera urspårningar som har diskuterats intensivt lokalt och haft stora ekonomiska konsekvenser. Till exempel, en urspårning kostar LKAB upp till 100 miljoner kronor om dagen, vilket ledde till förluster på cirka 7-10 miljarder kronor under den senaste incidenten. Dessutom påverkar det person- och godstransporter, vilket kan leda till långvariga störningar i flera år.

7.1.2 Problem

De återkommande urspårningarna på Malmbanan har inte bara ekonomiska konsekvenser utan påverkar också försörjningen till Nordnorge, vilket resulterade i att julmat fastnade i Kiruna. NATO har även ställt krav på att banan måste fungera som en del av Sveriges infrastruktur. Det finns ett uppenbart behov av att förutsäga och förebygga sådana incidenter samt optimera underhållsplaneringen för att säkerställa banans kontinuerliga drift.

7.1.3 Projektidé

Vår idé är att implementera en AI-lösning för att prediktera och förebygga urspårningar samt optimera underhållet av Malmbanan. Genom att samla in och analysera data från olika källor kan vi utveckla en modell som förutsäger risker och behov av underhåll. Här är en mer detaljerad beskrivning av projektet:

1. **Datainsamling:**
    - Samla in väderdata, rådata från sensorer på tågen, bilder och annan relevant information.
    - Använda data från Banverkets inspektionståg och sensorer längs banan.
2. **Modellutveckling:**
    - Utveckla en AI-modell som kan analysera de insamlade data för att förutsäga risken för urspårningar.
    - Modellen kan tränas med hjälp av historiska data om urspårningar och andra incidenter för att identifiera mönster och varningstecken.

3. **Prediktivt Underhåll:**

    - Använd AI-modellen för att förutsäga när olika komponenter, såsom växlar, behöver underhåll.
    - Implementera ett system för att planera underhållsaktiviteter baserat på förutsägelser från AI-modellen, vilket minskar risken för oplanerade avbrott.

4. **Planerade Stopp:**

    - Baserat på väderprognoser och andra riskfaktorer kan AI-modellen rekommendera planerade stopp av trafiken för att skydda banan och undvika urspårningar. Dessa rekommendationer kan bidra till att förebygga incidenter under extrema väderförhållanden.
    - Det är viktigt att konsultera sakkunniga inom Banverket och andra relevanta myndigheter för att fastställa detaljer och beslutsprocesser kring planerade stopp och andra säkerhetsåtgärder.

5. **Hantering av Mindre Incidenter:**

    - Modellen kan tränas för att identifiera och förutsäga inte bara urspårningar, utan även mindre incidenter som växelfel, ej fungerande bromsar och andra tekniska problem.
    - Genom att analysera data om förseningar och mindre tekniska problem kan modellen hjälpa till att upptäcka underliggande infrastrukturella problem och föreslå åtgärder innan de leder till större incidenter.

6. **Integration och Samarbete:**

    - Samarbeta med Addnodes systerbolag som redan arbetar med Banverket för att integrera AI-lösningen i befintliga system.
    - Säkerställa att projektet upphandlas korrekt, med Banverket som huvudansvarig.

7. **Implementering och Drift:**

    - Implementera lösningen som ett skalbart system som kan köras både on-premise och i molnet.
    - Använda tekniker som transfer learning och federated learning för att kontinuerligt förbättra modellens prestanda och minska energiförbrukningen.

7.1.4 Fördelar

- **Ekonomiska Besparingar:** Minska kostnaderna för urspårningar och underhåll genom prediktiva analyser.
- **Effektivitet:** Optimera underhållsplaneringen för att minimera störningar i tågtrafiken.
- **Säkerhet:** Öka säkerheten på Malmbanan genom att förutse och förebygga potentiella risker.
- **Hållbarhet:** Använd mindre och mer energieffektiva AI-modeller för att reducera miljöpåverkan.
- **Proaktiva Åtgärder:** Möjliggör planerade stopp och andra proaktiva åtgärder baserat på väderprognoser och riskanalyser för att skydda banan.
- **Förebyggande av Mindre Incidenter:** Identifiera och hantera mindre tekniska problem som växelfel och ej fungerande bromsar, vilket kan förhindra större incidenter och förseningar.

7.1.5 Sammanfattning

Genom att implementera en AI-baserad lösning kan vi dramatiskt förbättra förutsägelsen och förebyggandet av urspårningar på Malmbanan, samtidigt som vi optimerar underhållet och säkerställer en pålitlig transportväg för både Sverige och Norge. Detta projekt har potential att inte bara spara betydande kostnader utan också förbättra infrastrukturen och säkerheten för alla involverade parter.

7.2 Lageroptimering för detaljhandel

7.2.1 Bakgrund

Lagerhållning är en av de mest kostsamma aspekterna för detaljhandeln. Att ha för mycket lager innebär onödiga kostnader, medan för lite lager kan leda till att produkter tar slut, vilket missnöjer kunderna och påverkar försäljningen negativt. Att förutse rätt antal produkter att beställa vid rätt tidpunkt är avgörande för att upprätthålla en balanserad och kostnadseffektiv lagerhållning. Speciellt vid hantering av säsongsbaserade varor, som julskum och snöskyfflar, blir denna utmaning ännu större.

7.2.2 Problem

Felaktiga lagerprognoser kan leda till överlager eller lagerbrist, vilket i sin tur medför ekonomiska förluster. Dessutom innebär hanteringen av säsongsbaserade varor ytterligare komplexitet då efterfrågan kan variera kraftigt. Det finns ett tydligt behov av en lösning som kan analysera historisk försäljningsdata och externa faktorer för att optimera lagerhanteringen och minimera kostnaderna.

7.2.3 Projektidé

Vår idé är att implementera en AI-lösning för att optimera lagerhanteringen inom detaljhandeln. Genom att samla in och analysera data från olika källor kan vi utveckla en modell som förutsäger rätt antal produkter att beställa vid rätt tidpunkt, samt hanterar säsongsvariationer effektivt. Här är en mer detaljerad beskrivning av projektet:

1. **Datainsamling:**
    - Samla in historisk försäljningsdata från detaljhandelsföretag.
    - Integrera externa faktorer såsom väderprognoser, marknadstrender och kampanjer som kan påverka efterfrågan.
2. **Modellutveckling:**
    - Utveckla en AI-modell som analyserar de insamlade data för att förutsäga efterfrågan på produkter.
    - Modellen kan tränas med hjälp av historiska försäljningsdata för att identifiera mönster och trender.
3. **Säsongsbaserad Prognostisering:**

- Modellen kan också hantera säsongsvariationer genom att analysera historiska säsongsdata och förutse efterfrågan på säsongsbaserade varor som julskum och snöskyfflar.

4. **Lageroptimering:**
   - Använd AI-modellen för att optimera lagernivåerna och beställningsscheman, vilket minskar kostnaderna för överlager och lagerbrist.
   - Implementera ett system för att planera lagerhållning baserat på förutsägelser från AI-modellen, vilket säkerställer att rätt mängd produkter finns tillgängliga vid rätt tidpunkt.
5. **Integration och Samarbete:**
   - Samarbeta med befintliga ERP-system (Enterprise Resource Planning) och lagerhanteringssystem för att integrera AI-lösningen sömlöst.
   - Säkerställa att projektet upphandlas korrekt, med relevanta intressenter inom detaljhandeln som huvudansvariga.
6. **Implementering och Drift:**
   - Implementera lösningen som ett skalbart system som kan köras både on-premise och i molnet.
   - Använda tekniker som transfer learning för att förbättra modellens prestanda genom att använda förtränade modeller och anpassa dem efter specifika detaljhandelsbehov.

7.2.4 Fördelar

- **Kostnadsbesparingar:** Minska lagerkostnader genom optimerad lagerhantering och minskat behov av överlager.
- **Ökad Försäljning:** Förhindra lagerbrist och säkerställa att populära produkter alltid finns tillgängliga, vilket ökar kundnöjdheten och försäljningen.
- **Effektiv Säsongsplanering:** Hantera säsongsbaserade varor effektivt genom att förutse efterfrågan och planera lagerhållning i förväg.
- **Datadrivna Beslut:** Använd avancerad dataanalys för att fatta informerade beslut om lagerhantering och beställningar.
- **Flexibilitet:** Använda AI-lösningen både on-premise och i molnet, vilket ger flexibilitet i distribution och skalbarhet.

7.2.5 Sammanfattning

Genom att implementera en AI-baserad lösning kan detaljhandeln dramatiskt förbättra sin lagerhantering, minska kostnaderna och öka försäljningen. Denna lösning erbjuder en effektiv metod för att förutsäga efterfrågan, hantera säsongsvariationer och optimera lagerhållningen. Detta projekt har potential att inte bara spara betydande kostnader utan också förbättra kundnöjdheten och företagets konkurrenskraft.