LAB University of Applied Sciences

# Yrityksen myynnin ennustaminen

LAB-ammattikorkeakoulu
Insinööri (AMK)
2025
Kaisa Kovanen

**Tiivistelmä**


| Tekijä(t) | Julkaisun laji | Valmistumisaika |
|---|---|---|
| Kaisa Kovanen | Opinnäytetyö, AMK | 2025 |
| | Sivumäärä | |
| | 29 | |


Työn nimi

**Yrityksen myynnin ennustaminen**

Tutkinto ja koulutusala

Insinööri (AMK), tieto- ja viestintätekniikan koulutus

Toimeksiantajaorganisaatio (jos opinnäytetyöllä on toimeksiantaja)

Yritys X


Tiivistelmä

Opinnäytetyön tavoitteena oli toteuttaa myyntiennuste Yritys X:n seuraavan puolen vuoden myynneistä. Nykyään liiketoiminnan ennustaminen numeerisilla tiedoilla on yleistä, ja tärkeä osa liiketoiminnan suunnittelua. Koneoppiminen mahdollistaa tarkat ja ajantasaiset ennusteet.

Yritys X on pieni liikunta-alan yritys pk-seudulta. Ennusteiden tarkoituksena on antaa yrittäjälle lisää tietoa yrityksen jatkoa koskevien päätösten tekemisen tueksi. Ennuste toteutettiin koneoppimismenetelmiä hyödyntäen sekä kokonaismyynnille, että kahdelle tärkeimmälle tuotekategorialle erikseen. Malleja koulutettiin Holt-Wintersin metodilla, SARIMA-mallilla, Prophetilla ja XGBoostilla. Malllien suorituskykyä arvioitiin eri menetelmin. Myyntiennusteet toteutettiin jokaisen mallin parhaalla versiolla.

Ensimmäisen kuukauden osalta Prophet suoriutui parhaiten malleista. Vaikka ensimmäisen kuukauden ennuste oli suhteellisen lähellä oikeaa myyntiä, tulee muistaa, että mallin tarkkuus todennäköisesti heikkenee ajan kuluessa. Myyntiennusteet olivat yrittäjälle mielenkiintoisia ja antoivat tukea päätöksien tekemiseen.



Asiasanat

Koneoppiminen, aikasarja, ennustaminen

**Abstract**


| Author(s) | Type of Publication | Published |
|---|---|---|
| Kaisa Kovanen | Thesis, UAS | 2025 |
| | Number of Pages | |
| | 29 | |


Title of Publication

**Sales forecasting**

Degree, Field of Study

Engineer (UAS), Software Engineering

Organisation of the client (if the thesis work is commissioned by another party)

Company X


Abstract

The purpose of the thesis was to implement a sales forecast of the company X's sales for the next six months. Business forecasting with numerical data is common, and an important part of business palnning. Machine learning enables accurate and up-to-date predictions.

The purpose of sales forecasts is to help the entrepreneur to make more informed decisions about the company's future. The forecast was implemented for both total sales and two most important product categories separately. Sales forecasts were implemented using machine learning methods. Models were trained using the Holt-Winters method, SARIMA model, Prophet and XGBoost. The performance of these models was evaluated. Forecasts were implemented with the best version of each model.

For the first month, Prophet performed best on models. Although the forecast for the first month was relatively close to actual sales, it's supposable that the model's accuracy is likely to decline over time. The entrepreneur found sales forecasts interesting and helpful for making decisions.



Keywords

Machine learning, time series, forecasting

Sisällys

# 1 Johdanto

Koneoppimista hyödynnetään liike-elämän tarpeisiin kasvavissa määrin. Yksi osa-alue on menneiden havaintojen perusteella tulevaisuuden ennustaminen. Myyntiennusteiden avulla voidaan optimoida resursseja, hallita varastoja ja tehdä liiketoiminnan kasvuun ja kannattavuuteen vaikuttavia strategisia päätöksiä. Koneoppiminen tarjoaa mahdollisuuksia tarkkojen ja dynaamisten ennustemallien kehittämiseen. Koneoppiminen mahdollistaa suurten tietomäärien analysoinnin ja monimutkaisten riippuvuuksien tunnistamisen.

Toimeksiantaja on anonyyminä pysyttelevä pieni liikunta-alan yritys, myöhemmin yritys X, pääkaupunkiseudulta. Yritys X:n toimintaan kuuluu erilaiset tanssi- ja akrobatiatunnit, tarvikemyynti, workshopit vierailevien ohjaajien pitäminä ja kenkämyynti. Tanssitunteja järjestetään lyhyempinä kursseina, syys-, kevät- ja kesäkausina sekä online-toteutuksina. Kenkämyyntiin kuuluu kenkien maahantuominen Yhdysvalloista.

Opinnäytetyön tavoitteena on ennustaa yritys X:n tulevan puolen vuoden myynti. Yrityksen liiketoiminnan kannalta merkittävimmät kategoriat ovat liikuntatunneille osallistumiset, eli kategoria käynnit, ja kenkämyynti, joista suurin osa liikevaihdosta koostuu. Myyntiennusteet toteutetaan sekä kokonaismyynnille että erikseen kenkämyynnille ja kategorialle käynnit. Ennusteen toteuttamiseksi tutkitaan, millä tilastollisilla ja koneoppimismenetelmillä saadaan aikaan paras ennuste. Myyntiennusteen avulla yrittäjä suunnittelee yrityksensä seuraavia askeleita. Työ toteutetaan analysoimalla eri menetelmin muodostettuja koneoppimismalleja ja vertailemalla niiden suorituskykyä.

# 2 Koneoppiminen ja aikasarjat

## 2.1 Koneoppiminen

Koneoppiminen on yksi tehokkaimmista data-analytiikan työkaluista (Nelli 2018, 5). Koneoppimista hyödynnetään, kun oletetaan, että havaintojen välillä on jokin yhteys, mutta ei tiedetä, millainen (Alpaydin 2021, 47). Shalev-Shwartz ja Ben-David (2014, 3) määrittelevät koneoppimisen käytön tarpeen tehtäviin, jotka ovat liian monimutkaisia ohjelmoitaviksi; esimerkiksi erittäin suurien ja monimutkaisten tietojoukkojen käsittely.

Koneoppimisessa dataa voi ajatella laskentataulukkona, jossa eri sarakkeissa on eri ominaisuuksia. Koneoppimisessa otetaan yksi näistä sarakkeista ennustamisen kohdepiirteeksi, jota koitetaan selittää muiden sarakkeiden avulla. Algoritmin tavoitteena on saavuttaa mahdollisimman pieni virhe ennustuksien ja oikeiden arvojen välille. Kun koneoppimismallin virhe on pienin mahdollinen, sitä käytetään ennustamiseen uudesta datasta. (Amr 2020, 11.)

### 2.1.1 Ohjattu ja ohjaamaton oppiminen

Ohjatussa oppimisessa jokaiselle syötteelle on haluttu tulos. Tarkoituksena on muodostaa malli, jonka avulla voidaan ennustaa uusien syötteiden tulos oikein. Opetusaineiston tulee olla niin suuri ja monipuolinen, että malli toimii myös yleistettäessä. Ohjatussa oppimisessa ei ole tarkoitus toistaa opetustapauksia, vaan opetustapaukset ovat vain pieni osajoukko kaikista mahdollisista tapauksista. Tavoitteena on tehdä oikeita ennustuksia uusista tapauksista. (Alpaydin 2021, 56–57.) Yleensä kaikki ongelmat, joissa pyritään automatisoimaan tai jäljentämään jotakin olemassa olevaa prosessia, ratkeavat parhaiten ohjatun oppimisen keinoin. Ohjatun oppimisen menetelmät ovat hyödyllisiä ja tehokkaita. (Johnston & Mathur 2019, 11–14.)

Ohjaamattomassa oppimisessa ei ole saatavilla tai tiedossa tunnettuja lopputuloksia. Ohjaamattoman oppimisen menetelmät mallintavat dataa koulutusprosessiin suunniteltujen rajoitusten tai sääntöjen avulla. Yksi yleinen ohjaamattoman oppimisen muoto on klusterointi, jossa datajoukko jaetaan tiettyyn määrään eri ryhmiä. KNN-klusteroinnin tapauksessa jokainen datasjoukon näyte merkitään tai luokitellaan näytteen k-lähimpien pisteiden enemmistön mukaan. Koska ennalta määriteltyjä luokkia ei ole, ohjaamattoman oppimisen algoritmien suorituskyky voi vaihdella suuresti käytettävän datan ja mallin parametrien mukaan. Tunnettujen ja tavoiteltujen tulosten puute koulutuksen aikana johtaa siihen, että ohjaamattoman oppimisen menetelmiä käytetään yleisesti esimerkiksi tutkivassa analyysissä. (Johnston & Mathur 2019, 13–14.)

### 2.1.2 Koneoppimismallin luominen

Koneoppimismallin luominen lähtee ongelman määrittelystä. Erilaisia koneoppimistekniikoita on useita, ja ne voidaan ajatella yksinkertaisesti matemaattisiksi prosesseiksi, tai algoritmeiksi, kuten neuroverkoiksi, syväneuroverkoiksi tai satunnaismetsäalgoritmiksi. Ongelman määrittelyn jälkeen määritellään, minkälaisella datajoukolla ongelmaa voi lähteä ratkomaan. Kun datajoukko on hankittu ja puhdistettu, voidaan määritellä ja suunnitella mitä koneoppimismallia käytetään. (Johnston & Mathur 2019, 11.)

Kun käytettävä malli on valittu, voidaan määritellä mallin tarkat arvot. Arvojen määrittelyä toistetaan arvioiden mallin tuotosta suhteessa olemassa olevaan tietoon. Tätä optimointiprosessia kutsutaan mallin kouluttamiseksi. Kun koulutus on suoritettu, mallia tulee arvioida joidenkin viitetietojen perusteella kokonaissuorituskyvyn vertailuarvon saamiseksi. Kriittisimmät vaiheet koneoppimismallin luomisessa on usein ongelman määrittely sekä datan keruu, jotka määrittelevät monia myöhempiä päätöksiä tai koneoppimismallin suunnittelun valintoja. (Johnston & Mathur 2019, 11.)

## 2.2 Aikasarjojen perusteet

Tilastollisen asetelman luomiseksi aikasarja voidaan määritellä kokoelmaksi satunnaismuuttujia, jotka on indeksoitu sen mukaan, missä järjestyksessä ne on kerätty. Aikasarjoissa peräkkäisten havaintojen välinen korrelaatio rajoittaa perinteisten tilastollisten menetelmien käyttöä. (Shumway & Stoffer 2011, 1,11.)

Aikasarja voi olla jatkuva tai diskreetti. Jatkuvissa aikasarjoissa havainnot mitataan tietyin väliajoin, kuten vaikka lämpötila. Diskreetissä aikasarjassa havainnot on mitattu erillisissä pisteissä, kuten esimerkiksi valuuttakurssi. Jatkuvan aikasarjan voi muuntaa diskreetiksi yhdistämällä dataa tietyltä aikaväliltä. (Adhikari & Agrawal 2013, 12.)

Aikasarjoista voidaan erottaa osia. Aikasarjoista on yleensä havaittavissa trendi, kausivaihtelu ja epäsäännölliset tai sykliset komponentit. Trendi on yleinen suunta, jota kohti jokin asia kehittyy tai muuttuu, kuten esimerkiksi planeetan lämpötilojen nousu. Kausivaihtelu puolestaan ilmenee tiettyinä, spesifeinä ajankohtina, kuten viikoittain tai vuosittain, kuten jäätelönmyynnin kasvaminen kesäisin. (Auffarth 2021, 14–16.) Syklinen vaihtelu puolestaan tulee esiin pidemmällä aikavälillä, yleensä vähintään kahden vuoden jaksoissa. Aikasarjojen neljäs komponentti, virhe (residual) muodostuu epäsäännöllisistä ja ennustamattomista tekijöistä, jotka eivät toista tiettyä kaavaa. Tällaista vaihtelua aiheuttavat esimerkiksi lakot, maanjäristykset tai muut vastaavat. (Adhikari & Agrawal 2013, 13.)

Aikasarjojen neljää komponenttia voi lähestyä multiplikatiivisella mallilla, joka esitetään kaavassa 1. Toinen tapa on additiivinen malli, joka on nähtävissä kaavassa 2.

$$Y(t) = T(t) \times S(t) \times C(t) \times I(t) \qquad (1)$$

$$Y(t) = T(t) + S(t) + C(t) + I(t) \qquad (2)$$

joissa $Y(t)$ on havainto, ja $T(t)$, $S(t)$, $C(t)$ ja $I(t)$ vastaavasti trendi, kausittainen komponentti, syklinen komponentti ja virhe ajassa $t$. (Adhikari & Argawal 2013, 13.)

Multiplikatiivisessa mallissa oletetaan, että aikasarjan komponentit eivät ole toisistaan riippumattomia ja saattavat vaikuttaa toisiinsa, kun taas additiivisessa mallissa oletetaan, että komponenteilla ei ole vaikutusta toisiinsa. (Adhikari & Argawal 2013, 13.) Aikasarjoja käsiteltäessä on otettava huomioon, onko data yksi- vai monimuuttujaista, sisältääkö se kausivaihtelua, onko se stationaarista ja onko se lineaarista (Atwan 2022, 332).

### 2.2.1 Stationaarisuus

Aikasarjoja analysoitaessa mitataan sarjan arvojen välistä riippuvuutta. Jos riippuvuusrakenne muuttuu jatkuvasti, riippuvuuden mittaaminen on vaikeaa. Merkitsevän tilastollisen analyysin kannalta on kriittistä, että ainakin keskiarvo- ja autokovarianssifunktiot täyttävät stationaarisuuden ehdot, ainakin jollakin kohtuullisella aikavälillä. (Shumway & Stoffer 2011, 58.) Auffarthin (2021, 50–52) mukaan stationaarisuus on aikasarjan ominaisuus, jossa sen jakauma ei muutu ajan myötä. Jos aikasarja on stationaarinen, sillä ei ole trendiä eikä determinististä kausivaihtelua, eli ei-stationaarisen aikasarjan voi muuttaa stationaariseksi poistamalla trendin ja kausivaihtelun. Monet koneoppimismallit olettavat stationaarisuutta, eivätkä välttämättä toimi kunnolla, jos data on ei-stationaarista. Stationaarisuutta voi testata esimerkiksi täydennetyllä Dickey-Fuller-testillä.

Trendin voi poistaa datasta esimerkiksi derivoimalla, usein yksi tai kaksi derivointia riittää. Aikasarjadatassa on havaittavissa usein sekä trendi että kausivaihtelua. Derivointi sopii myös kausivaihtelun poistamiseen. Ensin poistetaan kausivaihtelu ja sen jälkeen trendi. (Montgomery ym. 2016, 50–52.)

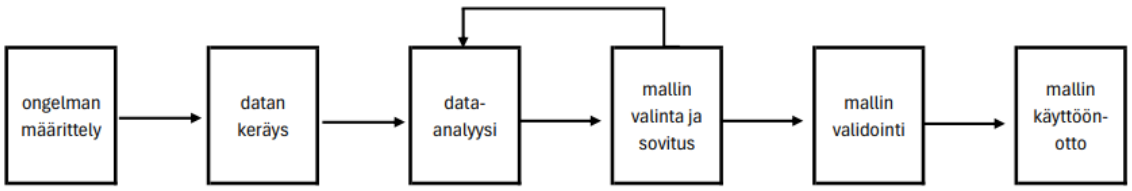
### 2.2.2 Aikasarjaennustaminen

Ennustaminen jaetaan usein lyhyen, keskipitkän ja pitkän aikavälin ennusteisiin. Lyhyt aikaväli on päivistä kuukausiin, keskipitkät yhdestä kahteen vuoteen ja pitkän aikavälin ennusteet menevät vielä pidemmälle. Tilastolliset menetelmät ovat usein hyödyllisiä lyhyen ja keskipitkän aikavälin ennusteissa. (Montgomery ym. 2016, 2.)

Aikasarjoilla ennustettaessa menneitä havaintoja kerätään ja analysoidaan sopivan matemaattisen mallin luomiseksi. Mallin on tarkoitus koota taustalla oleva datan luontiprosessi aikasarjalle. Tätä mallia käytetään tulevien tapahtumien ennustamiseen. (Adhikari & Argawal 2013, 15.) Tekniikat ennustamisen toteuttamiseen vaihtelevat käyrän sovittamisesta ekstrapolaatioon, nykyisten trendien analysoinnista monimutkaisiin koneoppimistekniikoihin (Auffarth 2021, 95).

Yleensä ennustetta ajatellaan yhtenä numerona, joka edustaa parasta arviota kiinnostavan muuttujan tulevasta arvosta. Tätä kutsutaan pistearvoksi tai piste-ennusteeksi. Usein nämä ennusteet ovat kuitenkin vääriä, eli tapahtuu ennustevirhe. Tämän takia on hyvä liittää ennusteeseen arvio siitä, kuinka suuri ennustevirhe saattaa olla. Yksi tapa tämän toteuttamiseen on tarjota ennusteväli (PI, prediction interval) piste-ennusteen lisäksi. Ennusteväli tarkoittaa haarukkaa ennusteelle, joka on todennäköisesti paljon hyödyllisempää päätöksen teossa kuin yksittäinen luku. (Montgomery ym. 2016, 5.)

Aikasarjaennusteen luomiseen kuuluu monia vaiheita, jotka näkyvät kuvassa 1. Ongelman määrittelyn ja datajoukon keräämisen jälkeen datalle suoritetaan data-analyysi. Datajoukosta tehdyistä kuvioista etsitään trendiä ja kausivaihtelua. Mallin valinnassa ja sovituksessa on kyse yhden tai useamman ennustemallin valitsemisesta ja mallin sovittamisesta datajoukkoon. Mallin validointivaiheessa arvioidaan ennustemallin suorituskykyä. Validoinnissa tutkitaan, minkä suuruisia ennustevirheitä tulee, kun mallia käytetään uusien tietojen ennustamiseen. Sovitusvirheet ovat aina pienempiä kuin ennustevirheet. Näiden vaiheiden jälkeen mallin voi ottaa käyttöön. (Montgomery ym. 2016, 13–15.)




Kuva 1. Aikasarjaennustemallin luomisen vaiheet (mukailtu Montgomery ym. 2016, 14)

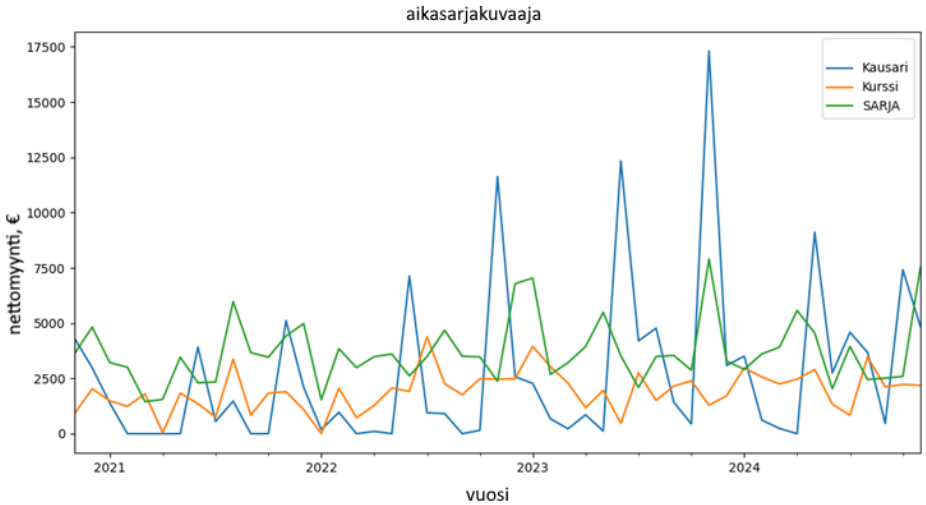
Kun malli on otettu käyttöön, sen suorituskykyä tulisi tarkkailla jatkuvasti hyvien ennusteiden takaamiseksi. Ennusteille on tyypillistä, että ehdot muuttuvat ajan mittaan, eli hyvin toimivan mallin suorituskyky todennäköisesti heikkenee ajan kuluessa. Suorituskyvyn

heikkeneminen näkyy suurempina tai systemaattisempina ennustevirheinä. (Montgomery ym. 2016, 16.)

### 2.2.3 Aikasarja-analyysi

Ensimmäinen vaihe ennen datajoukon ominaisuuksien muuntamista ja koneoppimismallin luomista on tutkiva aikasarja-analyysi. Aikasarjojen analysointiin kuuluu datajoukon ominaisuuksien tunnistaminen, määrittely ja jakaminen osiin (decomposition). Yksiulotteisessa analyysissa jokaiselle muuttujalle tehdään samat toimenpiteet erikseen, esimerkiksi yhteenvetotilastoja, histogrammeja tai puuttuvien ja poikkeavien arvojen etsimistä sekä stationaarisuuden testaamista. (Auffarth 2021, 42.) Jos mahdollista, havaintoja tulisi olla vähintään 50, ja mieluiten vähintään 100 tai enemmän (Box ym. 2016, 49).

Aikasarjan analysoiminen on hyvä aloittaa datan graafisella esittämisellä, monet aikasarjan yleiset piirteet voidaan havaita visuaalisesti. Aikasarjakuvaajassa on $y_t$ piirretty ajanhetkeä $t$ vastaan, missä $t = 1, 2, \ldots, T$. Aikasarjakuvaajasta näkee yleensä trendin ja kausittaisuuden, esimerkiksi kuviossa 1 on havaittavissa kausittaisuutta. Joskus on hyödyllistä asettaa alkuperäisen aikasarjakuvaajan päälle sen tasoitettu versio, jotta alkuperäisen datan rakenteet ja kuviot hahmottuvat paremmin. Tähän voidaan käyttää useita erilaisia datan tasoitusmenetelmiä, yksi yksinkertaisimmista ja yleisimmin käytetyistä on liukuva keskiarvo. (Montgomery ym. 2016, 26–27.)




Kuvio 1. Esimerkki aikasarjakuvaajasta, jossa on nähtävillä kausittaisuutta

Aikasarjan graafisen esittämisen jälkeen seuraava vaihe on trendin ja kausivaihtelun poistaminen, joko derivoimalla tai sovittamalla sopiva malli dataan. Jos vaihtelu on suhteellista aikasarjan keskiarvoon, datatransformaatioiden käyttäminen voi olla perusteltua. Näiden vaiheiden tarkoituksena on luoda stationaarinen aikasarjadata. (Montgomery ym. 2016, 62.)

### 2.2.4 Datan esikäsittely

Oikeasta elämästä kerätty datajoukko on usein epätäydellistä ja epäjohdonmukaista. Esikäsittelyn tarkoituksena on parantaa datan laatua koneoppimismallin laadun parantamiseksi. Datan esikäsittely jakautuu ominaisuusmuunnoksiin (feature transforms) ja ominaisuuksien suunnitteluun. Ominaisuusmuunnoksiin kuuluu skaalaaminen, potenssi- ja logaritmimuunnokset sekä imputointi. (Auffarth 2021, 68.)

Esikäsittelyn tavoitteena on luoda koneoppimismallin kouluttamiselle sopiva syöte, jonka avulla malli on helpompi kouluttaa ja arvioida. Ominaisuuksien tulee ennustaa kohdetta, ja jos malli ei sovi tarkoitukseensa, datan kerääminen, ominaisuuksien suunnittelu ja mallin rakentaminen uudelleen tai paremmin on perusteltua. Tietojen analysointi, esikäsittely ja koneoppimismallin luominen on iteratiivinen prosessi. (Auffarth 2021, 68–69.)

**Datan puhdistaminen**

Aikasarjamallien kehittäminen ja niillä ennustaminen vaatii datajoukon, jonka voi ajatella syötteenä aikasarjamallin tuottamalle tulosteelle. Kun datajoukko on kerätty, sille tehdään muutoksia esimerkiksi päällekkäisten tietueiden poistamiseksi ja ongelmien, kuten puuttuvien tietojen ratkaisemiseksi. Tätä vaihetta kutsutaan datan puhdistamiseksi. (Montgomery ym. 2016, 16.)

Datan puhdistamisessa tutkitaan dataa mahdollisten virheiden, puuttuvan datan, poikkeamien tai epätavallisten arvojen, tai jonkun muun epäjohdonmukaisuuden löytämiseksi ja korjaamiseksi. Puuttuvan datan lisäämistä tai virheiden korjaamista jollakin arviolla kutsutaan datan imputoinniksi. Imputoinnissa puuttuvat tai virheelliset arvot korvataan todennäköisellä arvolla, joka perustuu muuhun saatavilla olevaan dataan. Puuttuva arvo voidaan korvata esimerkiksi saatavilla olevien arvojen keskiarvolla. Jos datassa on trendi tai syklistä vaihtelua, stokastinen keskiarvoimputointi lisää keskiarvoon satunnaisvaihtelua, jotta imputoitu data heijastaisi paremmin alkuperäisen datan hajontaa. (Montgomery ym. 2016, 16–19.)

**Datan muunnokset**

Datajoukon havainnot eivät usein ole normaalijakauman mukaisia, minkä seurauksena perinteiset mallit voivat tuottaa virheellisiä tuloksia. Dataan on mahdollista tehdä muunnoksia,

joilla siitä tehdään niin normaalijakauman mukaista kuin mahdollista. Kuitenkin on olemassa epälineaarisia menetelmiä, joiden tulokset eivät riipu datajoukon jakaumasta. (Affarth 2021, 69–70.)

Skaalaamisen tarkoitus on tasoittaa ominaisuuksien välistä vaihteluväliä, ja se toteutetaan usein minmax-skaalaimella tai normalisoimalla esimerkiksi standard normal variate (SNV) -menetelmällä. SNV muuntaa datan siten, että sen keskiarvo on 0 ja keskihajonta 1. Logaritmimuunnoksia käytetään vähentämään jakauman vinoutta. Myös potenssimuunnoksia käytetään muuntamaan data lähemmäs normaalijakaumaa, esimerkiksi Box-Cox-muunnoksella. Datamuunnoksia tehdessä tulee varmistaa, että kohdepiirteen varianssi pysyy riittävällä tasolla eikä tarkkuutta menetetä. (Auffarth 2021, 70.)

**Datan jakaminen**

Kun aikasarja jaetaan koulutus- ja testausdataan, on tärkeää säilyttää kronologinen järjestys ja välttää datan vuotaminen. Testausdatan tulee olla tuoreempaa kuin koulutusdatan. Data jaetaan erottamalla tietty prosenttiosuus koulutus- ja testausdataksi. (Stan 2021.)

Testausdata voidaan jakaa edelleen testaus- ja validointidataan. Jos on saatavilla esimerkiksi yhden vuoden tiedot, ne voidaan jakaa koulutusdataan 9 kuukauden ajalta, validointidataan kahden kuukauden ajalta ja viimeinen kuukausi testausdataksi. Validointi- ja testausdatan käyttö on päällekkäistä, testausdatan käyttö tarkistaa myös validointidatan. Joskus käytetään ainoastaan testausdataa. (Auffarth 2021, 106.)

### 2.2.5 Mallin valitseminen, sovittaminen ja validointi

Käytettävissä oleva raakadata jaetaan kahteen ryhmään, koulutusdataan ja testausdataan. Koulutusdatan havaintoja käytetään mallin luomiseen. Usein osaa testausdatasta käytetään validointidatana. Kun malli on luotu, sitä käytetään ennustuksien luomiseen. Testausdatan havaintoja käytetään mallin tarkkuuden arviointiin. Tarvittaessa datalle suoritetaan käänteinen muunnos takaisin alkuperäiseen asteikkoon. Malleja vertaillaan ja niiden ennustetarkkuutta arvioidaan tarkastelemalla mallien suhteellista suorituskykyä testidatassa. (Adhikari & Argawal 2013, 42.)

Säästäväisyyden periaatteen (principle of parsimony) mukaan mallit tulisi luoda pienimällä mahdollisella parametrimäärällä. Useista sopivista malleista kannattaa harkita yksinkertaisinta, joka kuitenkin säilyttää aikasarjan ominaisuudet ja riittävän kuvan taustalla olevasta datasta. Mallin monimutkaistuessa riski poiketa mallin oletuksista kasvaa. Mallin parametrien määrän kasvaessa riski ylisovittamiselle kasvaa. Ylisovitettu malli toimii testausdatalle

hyvin, mutta ei välttämättä sovellu tulevaisuuden ennustamiseen. (Adhikari & Argawal 2013, 15–16.)

On yleistä, että on useita eri malleja, joita voisi käyttää dataan. Yksi tapa valita sopivin, on sovittaa malleja menneeseen dataan, sillä huonosti sopiva malli ei todennäköisesti tuota hyviä ennusteita. Toisaalta parhaiten menneeseen dataan sopiva malli ei yleensä tuota parhaita ennusteita uudella datalla. On tärkeää välttää ylisovittamista. (Montgomery ym. 2016, 63, 74.)

Mallin validoinnissa arvioidaan, kuinka todennäköisesti malli toimii tarkoitetussa tehtävässä. Mallin sopivuutta ei tule arvioida ainoastaan menneeseen datan sovittamalla, vaan on tutkittava minkä suuruisia ennustevirheitä tulee, kun mallia käytetään uuteen dataan. Sovitusvirheet ovat aina pienempiä kuin ennustevirheet. (Montgomery ym. 2016, 15.)

### 2.2.6 Mallin suorituskyvyn arviointi

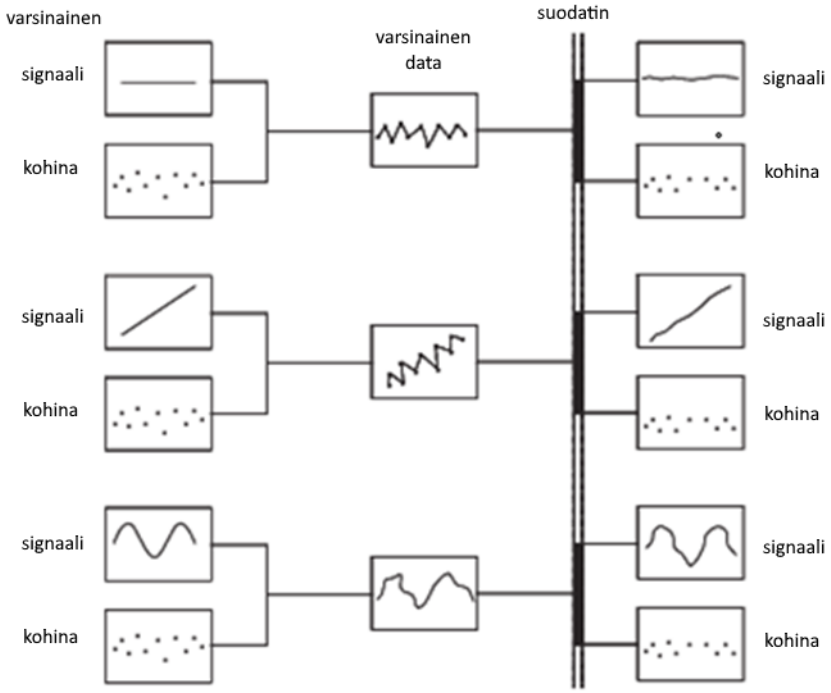
Usein käytettyjä suorituskyvyn mittareita on esimerkiksi keskimääräinen ennustevirhe (MFE, Mean Forecast Error), keskimääräinen absoluuttinen virhe (MAE, Mean Absolute Error), keskineliövirhe (MSE, Mean Squared Error) ja keskineliövirheen neliöjuuri (RMSE, Root Mean Squared Error). Näillä mittareilla on erilaisia ominaisuuksia ja niiden tulokset eroavat toisistaan, eli mallia arvioitaessa on hyvä ottaa huomioon useampi kuin yksi suorituskyvyn mittari. Useampaa mittaria käyttämällä on helpompi ymmärtää ennustevirheen määrää, suuruusluokkaa ja suuntaa. (Adhikari & Argawal 2013, 43–45.)

Keskimääräisen ennustevirheen tulisi olla mahdollisimman lähellä nollaa. Jos ennuste eroaa tuntuvasti nollasta, ennusteessa on vinouma. Jos keskimääräinen ennustevirhe ajautuu mallia käytettäessä kauemmas nollasta, aikasarja saattaa olla muuttunut tavalla, jota ennustemalli ei ole pystynyt ottamaan huomioon. Keskineliövirhe mittaa ennustevirheiden vaihtelevuutta, ja tämänkin luvun olisi hyvä olla mahdollisimman pieni. Yleistäen on hyvä valita malli, jonka keskihajonta jäännösvirhe on yhden askeleen ennusteissa pienin, kun mallia käytetään dataan, jota ei ole käytetty sovittamiseen. Suorituskyvyn mittaamiseen olisi hyvä varata vähintään 20 tai 25 havaintoa datajoukosta. (Montgomery ym. 2016, 65–66, 74.)

# 3 Aikasarjaennustamisen malleja

## 3.1 Eksponenttinen suodatus

Datajoukon voidaan ajatella koostuvan kahdesta erillisestä osasta, signaalista ja kohinasta. Signaali edustaa mitä tahansa kuviota, jonka aiheuttaa sen prosessin sisäinen dynamiikka, josta data kerätään. Suodatus on tekniikka, jolla pyritään erottamaan signaali kohinasta niin hyvin kuin mahdollista. Kuvassa 2 esitetään, miten tasoittamalla signaaleja niitä voidaan rekonstruoida, ja signaalin taustakuvio saadaan jossain määrin palautettua. (Montgomery ym. 2016, 233–234.)

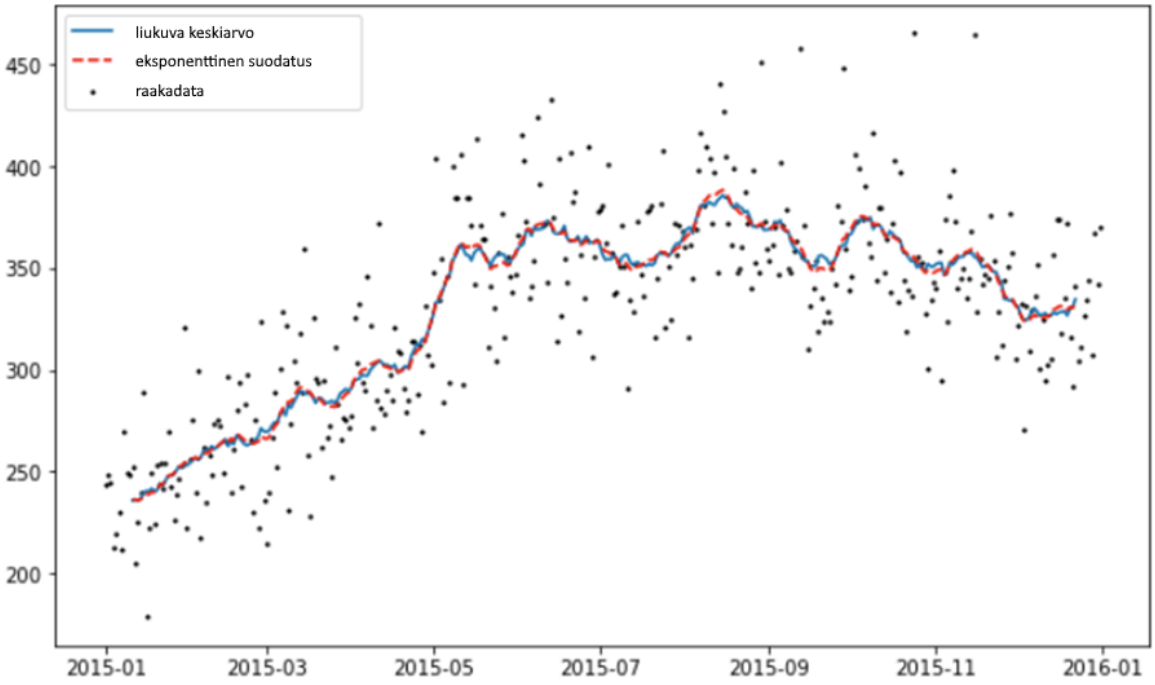



Kuva 2. Signaalin ja kohinan taustakuvion palauttaminen tasoittamalla (mukailtu Montgomery ym. 2016, 233)

Liukuva keskiarvo on mahdollisesti yksinkertaisin ennustamisen muoto, jonka tavoitteena on löytää suorempi linja paljon vaihtelua sisältävän datan läpi. Jokainen datapiste säädetään *n* ympäröivän datapisteen keskiarvoon, jossa *n* on ikkunan koko. Isompi ikkunan koko poistaa kohinaa ja saa trendin esiin datajoukosta, mutta ennusteet ovat trendistä jäljessä. (Rafferty 2021, 19.)

Ennusteet reagoivat nopeammin muutoksiin, jos käytetään eksponenttista suodatusta (Rafferty 2021, 19). Eksponenttisella suodatuksella menneiden havaintojen painotettujen

keskiarvojen painotus heikkenee eksponentiaalisesti havaintojen vanhetessa. Tämä tuottaa luotettavia ennusteita nopeasti monenlaisille aikasarjoille. (Hyndman & Athanasopoulos 2021.) Kuviosta 2 on nähtävillä, kuinka liukuvan keskiarvon linja on paljon rosoisempi kuin eksponentiaalisen suodatuksen linja.




Kuvio 2. Liukuva keskiarvo verrattuna eksponenttiseen tasoitukseen (mukailtu Rafferty 2021, 19)

Jos datassa ei ole selkeää trendiä eikä kausivaihtelua, siihen sopiva eksponenttinen tasoituksen metodi on nimeltään simple exponential smoothing (SES). Jos datassa on trendi, tulee käyttää Holtin lineaarisen trendin metodia. Jos datassa on trendin lisäksi kausivaihtelua, tulee käyttää Holt-Wintersin metodia. (Hyndman & Athanasopoulos 2021.)

Yleistäen eksponenttinen suodatus on nopea ja tehokas tekniikka aikasarjojen analyysin parantamiseksi, poikkeamien havaitsemiseen ja ennustamiseen. (Atwan 2022, 355). Nämä metodit reagoivat kuitenkin hitaasti trendin muutokseen, eli ennusteet jäävät jälkeen todellisuudesta eivätkä sovi pitkän ajan ennusteiden luomiseen (Rafferty 2021, 20).

## 3.2 SARIMA

Liukuvasta keskiarvosta (MA, moving average) on jatkokehitetty ARIMA-malli. Autoregressiomalli (AR, autoregressive) käyttää edellisten aikavaiheiden havaintoja syötteinä

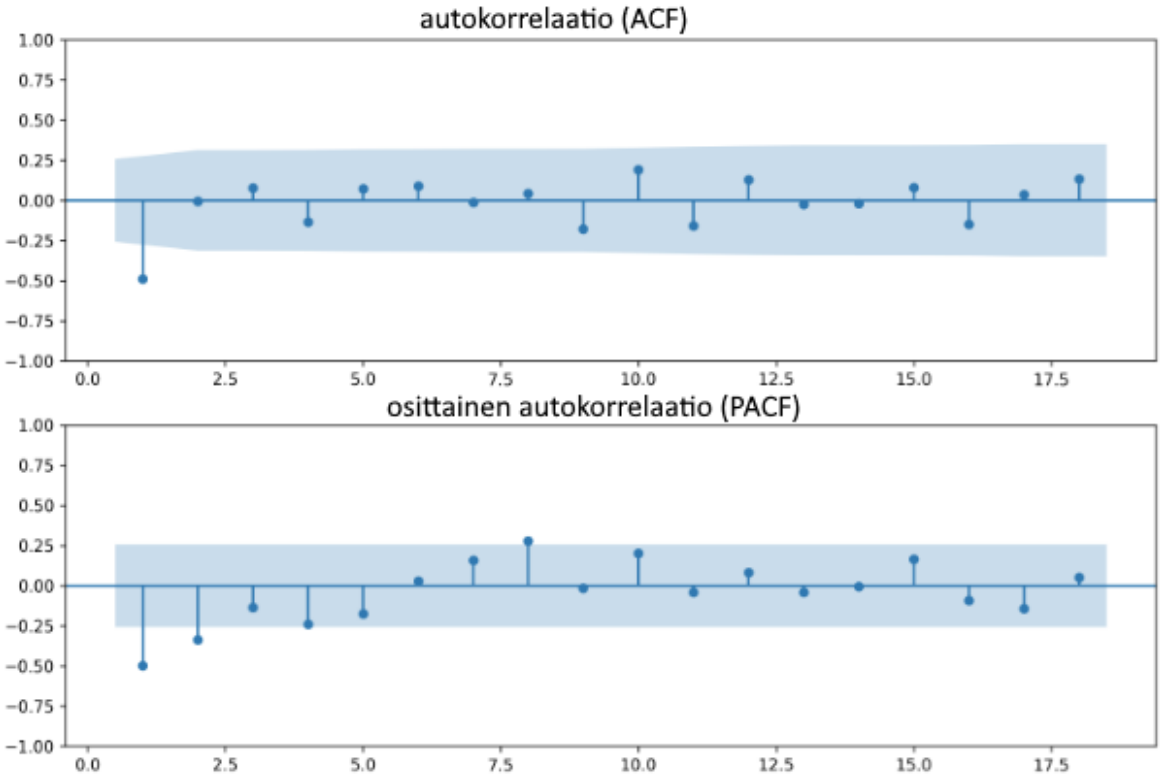
regressioyhtälöön seuraavan vaiheen ennustetun arvon määrittämiseksi. Sitä voidaan kuvata muuttujan regressioksi itsensä aikaisempaan versioon. Autoregressiomallista on johdettu ARIMA, jossa derivoinnille (I) on lisätty oma muuttuja stationaarisuuden saavuttamiseksi. Kausittaiselle datalle käytetään ARIMAn variaatiota, SARIMAa (Seasonal Autoregressive Integrated Moving Average). (Adhikari & Argawal 2013, 18; Atwan 2022, 364–366.)

ARIMA-mallia, joka on nähtävissä kaavassa 3, rakentaessa tulee valita parametrit *p*, *q* ja *d.*

$$ARIMA(p, d, q) \qquad (3)$$

jossa *p* on autoregressiivisen mallin viiveiden lukumäärä, *d* on derivointiaste ja *q* on liukuvan keskiarvon mallin viiveiden lukumäärä. (Atwan 2022, 357.)

Parametreja *p* ja *q* kutsutaan viiveeksi, koska ne kuvaavat viivettä eli sitä, kuinka monta ajanjaksoa siirrytään taaksepäin. Parametri *d* tarkoittaa integrointiastetta, eli *d*:n arvo valitaan sen mukaan, kuinka monta derivointia tarvitaan stationaarisuuden saavuttamiseksi. Parametrin *p* määrittämiseen voidaan käyttää osittaisen autokorrelaation (PACF) kuvaajaa ja *q*:n määrittämiseen autokorrelaation (ACF) kuvaajaa. Kuviossa 3 on esimerkki ACF- ja PACF-kuvaajista. Näissä kuvaajissa esitetään arvot, jotka vaihtelevat -1:stä 1:een pystyakselilla, ja vaaka-akselilla esitetään viive. Merkittävä viive on mikä tahansa viive, joka ylittää varjostetun luottamusvälin. Kuvion 3 ACF-kuvaajassa merkittävä viive on 1. Aina ei kuitenkaan ole selvää, mitkä optimaaliset arvot *p*:lle ja *q*:lle ovat, jolloin on hyvä testata ARIMA-mallia useilla eri arvoilla. (Atwan 2022, 339, 357–360.)




Kuvio 3. ACF- ja PACF-kuvaaja (mukailtu Atwan 2022, 340)

ARIMA-mallin rakentaminen voi olla iteratiivinen prosessi, johon kuuluu erilaisten mallien luominen ja testaaminen (Atwan 2022, 363). ARIMA-mallilla saadaan tyypillisesti hyviä tuloksia, mutta sen virittäminen ja optimointi on usein resursseja vievää ja tulosten hyvyys voi riippua ennusteen tekijän taitotasosta ja kokemuksesta (Rafferty 2021, 21).

SARIMA-mallissa ei-kausittaisille komponenteille on parametrit (*p, d, q*), jonka lisäksi kausittaiselle osuudelle (*P, D, Q, s*), missä *s* on kausivaihtelun pituus. Kirjaimet tarkoittavat edelleen samaa kuin ARIMA-mallissa, ja kirjaisinkoolla ilmoitetaan mihin komponenttiin viitataan. Parametri *s* ilmoittaa syklin askelten määrän, esimerkiksi kuukausittaisessa datassa *s*=12. (Atwan 2022, 369.)

## 3.3 Prophet

Prophet on Facebookille sisäisesti kehitetty aikasarjaennustamisen menetelmä. Käytettävissä olevat automattisemmat ennustetyökalut olivat yleensä liian joustamattomia ja kykenemättömiä käsittelemään lisäoletuksia, kun taas vankemmat ennustetyökalut tarvitsivat kokeneen analyytikon, jolla on datatieteen taitoja. Prophetin käyttö ei vaadi syvää matemaattista tai tilastollista osaamista. Jokainen ongelma ratkaistaan samalla kaavalla, jolloin aikaa ennusteen optimointiin kuluu vähemmän aikaa. (Rafferty 2021, 15–24.)

Prophet on suunniteltu varsinkin liiketoiminnan ennustamiseen, jolle on omat erityispiirteensä. Liiketoiminnassa on usein kausivaihtelua. Lomat ja pyhät sekä muut yksittäiset tapahtumat saattavat aiheuttaa epäsäännöllisyyttä, dataa voi puuttua ja trendi voi muuttua merkittävästikin jonkin toimenpiteen seurauksena. Prophet tuottaa oletusasetuksillakin tyypillisesti laadukkaita ennustuksia, mutta myös kustomointi on tarvittaessa mahdollista. Pohjimmiltaan Prophet on additiivinen regressiomalli. (Rafferty 2021, 24.)

Prophetin toimintalogiikka on samanlainen kuin additiivisella mallilla. Kaavassa 4 on eritelty Prophetin komponentit, jotka ottavat huomioon trendin, kausivaihtelun ja pyhäpäivät.

$$y(t) = g(t) + s(t) + h(t) + e(t) \qquad (4)$$

jossa

*g(t)* = trendi

*s(t)* = kausivaihtelu

*h(t)* = pyhäpäivien vaikutus ennusteeseen

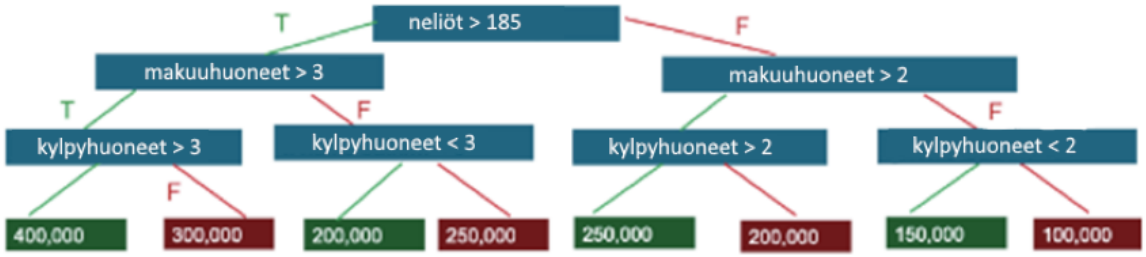
*e(t)* = virhe

*y(t)* = ennuste. (Geeksforgeeks, 2024.)

Prophet toimii oletusarvoisesti paloittain lineaarisella (piece-wice linear) mallilla, joka toimii parhaiten datalle, jolla on lineaarisia ominaisuuksia. Ei-lineaariselle datalle tulee käyttää logistisen kasvun mallia. (Geeksforgeeks, 2024.) Kriegerin (2021) mukaan Prophet tunnistaa automaattisesti pisteet, jossa trendin *g* suunta muuttuu. Muutospisteet voi asettaa tarvittaessa manuaalisesti. Kausivaihtelu *s* noudattaa Fourier'n sarjaa, jota voi ajatella useiden peräkkäisten sinien ja kosinien summana. Prophet käsittelee automaattisesti puuttuvan ja virheellisen datan.

## 3.4 XGBoost

Päätöspuu on diagrammi, jossa on sarja valintoja ja niiden seurauksia. Kuviossa 4 on nähtävillä esimerkki yksinkertaisesta päätöspuusta, jossa tehdään asunnon hinta-arvio kokoon, makuuhuoneiden ja kylpyhuoneiden lukumäärään perustuen. Päätöspuu-algoritmit ovat suosittuja, sillä ne ovat helposti koulutettavia, toimivat puuttuvista arvoista huolimatta, valitsevat automaattisesti olennaiset ominaisuudet mallille ja käsittelevät sekä numeerista että kategorista dataa. Päätöspuu-algoritmit käsittelevät kohinan ja poikkeavat havainnot hyvin. (Wang ym. 2023, 97.)




Kuvio 4. Esimerkki päätöspuusta (mukailtu Nvidia)

Gradienttia tehostavat (gradient boosting) mallit oppivat yksittäisten päätöspuiden virheistä säätämällä nykyistä puuta aiempien puiden virheiden perusteella. Päätöspuut eivät ole eristettyjä, vaan ne rakentuvat toistensa päälle. Gradienttia tehostavien mallien toiminta perustuu heikkojen mallien avulla paremman mallin luomiseen. Ensimmäinen päätöspuu, niin sanottu perusoppija, ei saa olla liian tarkka. Mallin oppimisen periaatteena on analysoida virhettä, eli eroa mallin ennusteiden ja todellisten arvojen välillä. (Wade 2020, 84–88.)

XGBoost (eXtreme Gradient Boosting) on skaalautuva ja tarkka toteutus gradientin tehostamiseen perustuvasta koneoppimisesta. XGBoostissa päätöspuita rakennetaan rinnakkain eikä peräkkäin, kuten perinteisissä boostausmenetelmissä. XGBoost noudattaa tasokohtaista strategiaa, jossa gradienttiarvot skannataan ja näiden osittaisia summia hyödynnetään arvioitaessa mahdollisten jakojen laatua harjoitusdatan eri kohdissa. Useissa vertailututkimuksissa on todettu XGBoost-mallien tarjoavan parhaan yhdistelmän ennustustarkkuutta ja laskennallista tehokkuutta. (Nvidia.)

Boschettin ja Massaron (2018, 240–241) mukaan laskennallisen tehokkuuden taustalla on muun muassa algoritmi, joka voi hyödyntää harvoja matriiseja (sparse matrix) säästäen muistia. Nolla-arvot käsitellään laskenta-aikaa säästävällä tavalla. Data tallennetaan Column Block -nimisellä ratkaisulla, jossa data on levyllä sarakkeittain, joka optimoi algoritmin toimintaa. Puuttuvien arvojen imputointi on toteutettu tehokkaasti saatavilla olevan datan perusteella. XGBoostin keskeiset parametrit ovat:

- *eta:* Kuinka nopeasti algoritmi oppii, eli kuinka monta puuta tarvitaan. Korkeammilla arvoilla oppimisprosessilla on parempi konvergenssi, mutta koulutusaika ja tarvittavien puiden määrä on suurempi.
- *gamma:* Pienin häviön vähentäminen, joka tarvitaan uuden leaf noden luomiseen. Toimii pysäyttävänä kriteerinä puun kehityksessä. Korkeammat arvot tekevät oppimisesta konservatiivisempaa.

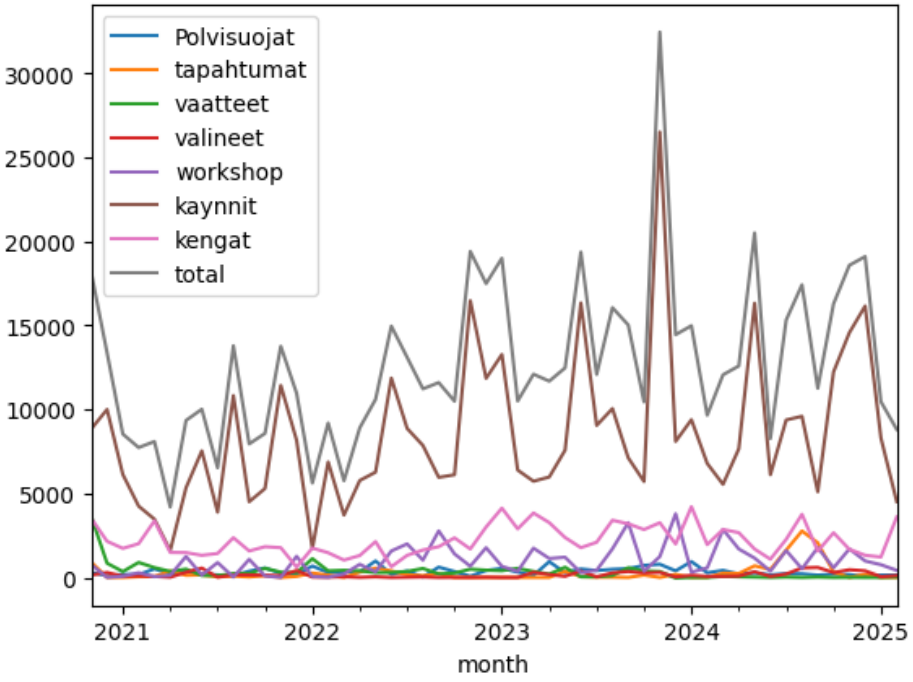
- *min_child_weight:* Minimipaino puun lehtisolmussa (leaf node). Korkeammat arvot estävät ylisovituksen.
- *max_depth:* Vuorovaikutusten määrä puissa.
- *subsample:* Kussakin iteraatiossa käytettävä koulutusdatan esimerkkien määrä.

# 4 Myynnin ennustaminen

## 4.1 Työn tavoite ja datajoukko

Työn tavoitteena on ennustaa yritys X:n tulevaa myyntiä menneen myynnin perusteella. Käytettävissä oleva data on yrityksen myyntidata kuukausittain, marraskuulta 2021 helmikuulle 2025. Mallien ennustusteiden hyvyyden arviointiin käytettiin maalikuun 2025 myyntidataa.

Datajoukon esikäsittelyyn kuului puuttuvien tuoteryhmien lisääminen tuotteille. Negatiiviset arvot poistettiin vähentämällä kyseiset arvot edellisen kuukauden myynneistä. Tuoteryhmiä yhdisteltiin selkeämmän kuvan muodostamiseksi. Kuviosta 5 on havaittavissa, että eniten myyntiä tuottavat kategoriat ovat käynnit, eli liikuntatunneille osallistuminen, ja kengät. Kokonaismyynnin lisäksi myyntiennusteet toteutettiin kategorioille kengät ja käynnit. Nämä sarjat käyttäytyvät hieman eri tavalla.



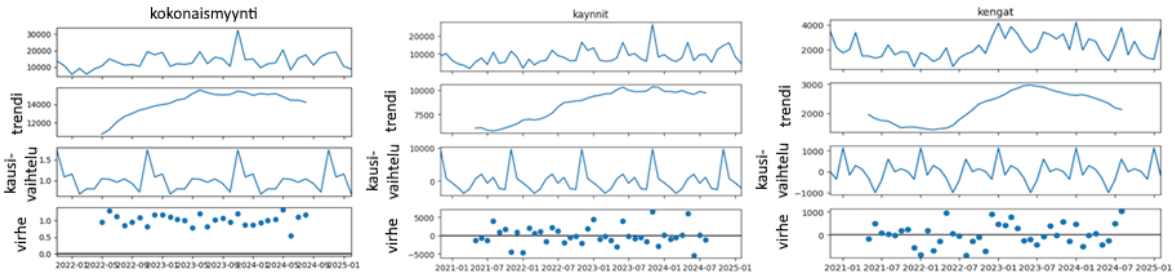

Kuvio 5. Tuotteiden nettomyyntimäärät kategorioittain

Myynnin ennustamiseen käytetään nettomyyntejä euroissa. Käytettävissä oleva datajoukko on melko pieni ja tuoteryhmiä on yhdistelty ennusteiden toteuttamista varten. Esimerkiksi kategoriassa käynnit tuotteiden hinta vaihtelee 8 euron kertaostoksesta yli 300 euron

kausikortteihin ja kuukausittain laskutettaviin jatkuviin tilauksiin, joten nettomyynti tarjoaa enemmän informaatiota kuin esimerkiksi kappalemäärän ennustaminen. Yrityksen toimintahistorian aikana esimerkiksi liikuntapalveluiden verokanta on muuttunut, mutta nettomyynnissä verot eivät ole mukana.

## 4.2 Data-analyysi

Datan analysointi aloitettiin jakamalla aikasarja komponentteihin. Kuvan 3 aikasarjan erittelykuvaajasta voidaan nähdä, että kaikkien kategorioiden trendi on ensin nouseva, kunnes tasaisen jakson jälkeen kääntyy laskuun. Kausivaihtelussa on näkyvissä varsinkin käynneissä selkeät piikit kausipaikkojen tullessa myyntiin, mikä on liikuntapaikoille tyypillistä.




Kuva 3. Kokonaismyynnin, käyntien ja kenkämyynnin aikasarjan erittelykuvaaja

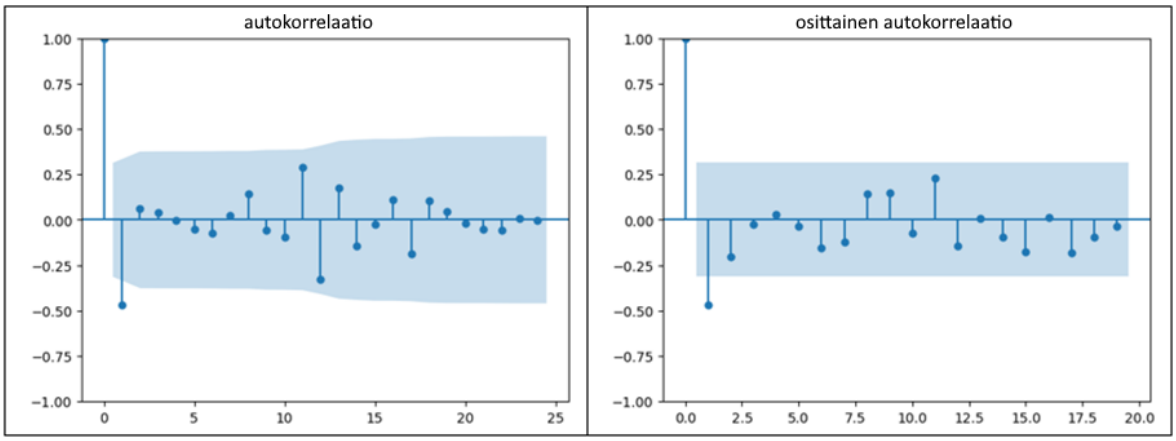
Aikasarjojen erittelykuvaajat toteutettiin sekä multiplikatiivisella että additiivisella mallilla, jotka silmämääräisesti näyttivät samanlaisilta. Paremman mallin arviointiin otettiin avuksi jäännösvarianssin (residual variance) laskeminen, jonka perusteella multiplikatiivinen malli näyttäisi sopivan paremmin jokaiselle kategorialle.

Stationaarisuutta tutkittiin sekä KPSS-testillä että täydennetyllä Dickeyn-Fullerin testillä, joiden tulokset ovat nähtävillä taulukosta 1. Jos aikasarjalla on stationaarinen keskiarvo ja trendi, täydennetty Dickeyn-Fullerin testi luokittelee aikasarjan stationaariseksi, mutta KPSS-testi tunnistaa sen ei-stationaariseksi. Ensimmäisen asteen derivointi riitti saavuttamaan stationaarisuuden molemmilla testeillä. Liian derivoidulla datalla koulutettu malli johtaa vähemmän tarkkoihin tuloksiin.

| kategoria | testi | stationaarisuus | 1. derivointi |
|---|---|---|---|
| kokonaismyynti | ADF | stationaarinen | stationaarinen |
| | KPSS | ei-stationaarinen | stationaarinen |
| kengät | ADF | stationaarinen | |
| | KPSS | stationaarinen | |
| käynnit | ADF | stationaarinen | stationaarinen |
| | KPSS | ei-stationaarinen | stationaarinen |

Taulukko 1. Stationaarisuuden testitulokset

Kuviossa 6 on esitetty kenkämyynnin autokorrelaation ja osittaisen autokorrelaation kuvaajat. Molemmissa kuvaajissa näkyy selvä piikki lagissa 1, mikä viittaa siihen, että sarjassa on tilastollisesti merkitsevä yhteys edelliseen havaintoon.

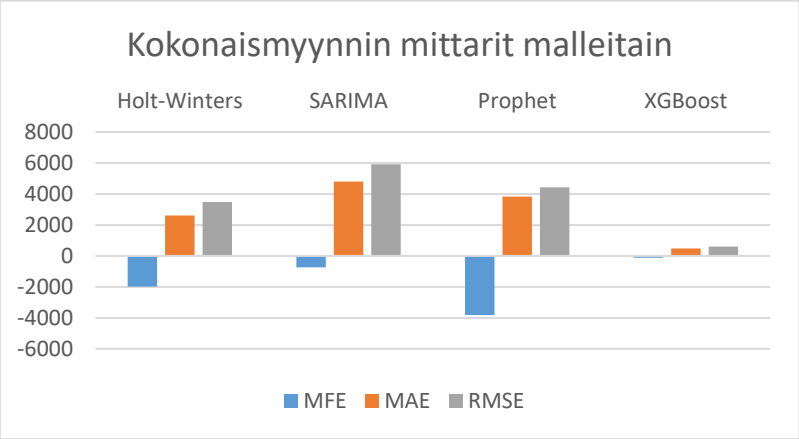



Kuvio 6. Kenkämyynnin autokorrelaatiokuvaajat

Datajoukko on melko pieni koneoppimismallien koulutukseen. Koulutusdatana käytettiin 44 havaintoa ja testausdatana 8 havaintoa. Datajoukon pieni koko ohjasi käytettyjen mallien valintaa. Eksponenttien tasoitus on melko yksinkertainen tilastollinen aikasarjaennustusmetodi, joka toimii pienilläkin datajoukoilla. SARIMA-mallissa varsinkin derivointi voi vaikuttaa näin pienen datajoukon ennustettavuuteen. Myös Prophet ja XGBoost on suunniteltu suuremmille datajoukoille. Kaikilla malleilla trendin ja kausivaihtelun arvioiminen pienellä datajoukolla on haastavaa ja ylisovittamisen riski kohonnut. Malleja koulutettiin useita ja parametreja säädettiin varsinkin Holt-Wintersilla ja SARIMAlla. Koulutetuista malleista parhaiten suoriutuneet otettiin mukaan mallien arviointiin ja ennustuksen luomiseen.

## 4.3 Mallien ennustetarkkuuden arviointi

Mallien arviointiin käytettiin keskimääräistä ennustevirhettä, keskimääräistä absoluuttista virhettä ja keskineliövirheen neliöjuurta. Jokaisella mittarilla tavoitellaan arvoa mahdollisimman lähelle nollaa. Varsinkin keskineliövirheen neliöjuuri oli suuri muissa malleissa kuin XGBoostissa. Käytettävillä mittareilla arvioituna XGBoost näyttää selkeästi parhaimmalta mallilta. Kuviosta 7 on nähtävissä, että toiseksi paras on Holt-Winters, sitten Prophet ja viimeisenä SARIMA. Datajoukon pieni koko voi vaikuttaa mallien suoriutumiseen. Näin pienessä datajoukossa laskenta-ajoissa ei ollut merkittäviä eroja, XGBoost oli kuitenkin selkeästi hitain.




Kuvio 7. Kokonaismyynnin mallien arviointi

Kuvassa 4 on nähtävissä eri mallien ennusteet. Kuvioissa on visualisoitu koulutusdata sinisellä, ennusteet punaisella ja testidata vihreällä. Holt-Wintersin mallille ennusteet on visualisoitu sekä punaisella että keltaisella. Näiden visualisointien perusteella XGBoost ja Prophet vaikuttavat luotettavimmilta. Varsinkin XGBoostilla on kohonnut ylisovittamisen riski pienen datajoukon vuoksi.


Kuva 4. Mallien ennusteet testausdatalle

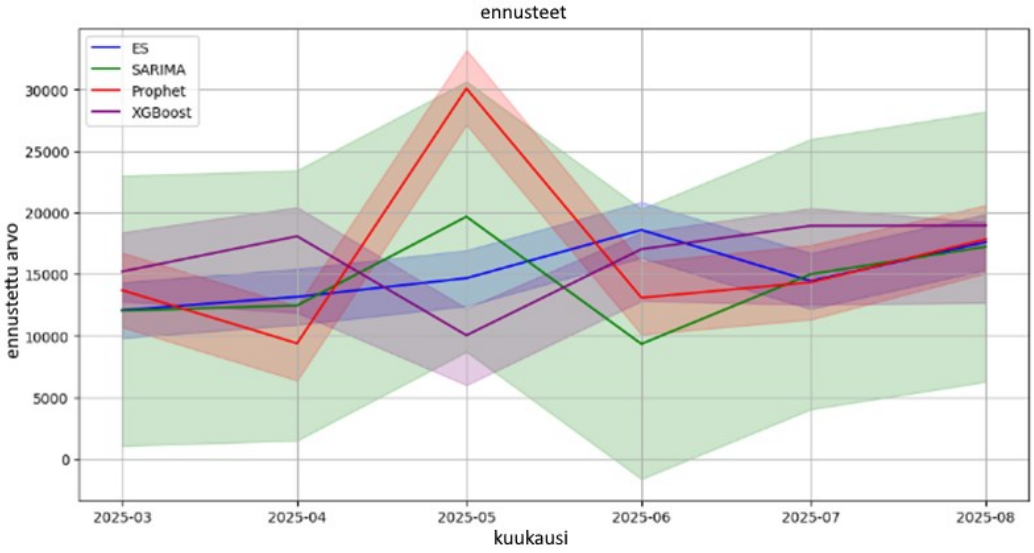
## 4.4 Mallien puolen vuoden ennusteet

Ensimmäinen ennuste on maaliskuulle 2025. Maaliskuun toteutunut kokonaisnettomyynti on 13677,97 euroa. Taulukosta 2 on luettavissa, että Prophetilla muodostettu ennuste osui lähimmäksi todellista nettomyyntiä poiketen vain +15,21 euroa. XGBoostin ennuste yliarvioi myyntiä merkittävästi, kun taas Holt-Winters ja SARIMA ennustavat huomattavasti alhaisemmat myynnit.

| kuukausi | malli | alaraja | ennuste | yläraja |
|---|---|---|---|---|
| 01/03/2025 | ES | 9805.94 | 12070.89 | 14335.84 |
| 01/03/2025 | SARIMA | 1053.96 | 12024.40 | 22994.83 |
| 01/03/2025 | Prophet | 10693.36 | 13698.18 | 16737.30 |
| 01/03/2025 | XGBoost | 12790.98 | 15214.22 | 18390.30 |
| 01/04/2025 | ES | 10893.10 | 13158.05 | 15422.99 |
| 01/04/2025 | SARIMA | 1486.56 | 12459.97 | 23433.38 |
| 01/04/2025 | Prophet | 6370.94 | 9378.16 | 12392.11 |
| 01/04/2025 | XGBoost | 11872.42 | 18078.36 | 20434.47 |
| 01/05/2025 | ES | 12409.20 | 14674.15 | 16939.10 |
| 01/05/2025 | SARIMA | 8697.94 | 19671.35 | 30644.76 |
| 01/05/2025 | Prophet | 27141.05 | 30076.60 | 33210.69 |
| 01/05/2025 | XGBoost | 5990.34 | 10035.20 | 12224.80 |
| 01/06/2025 | ES | 16325.60 | 18590.55 | 20855.50 |
| 01/06/2025 | SARIMA | -1641.84 | 9331.57 | 20304.98 |
| 01/06/2025 | Prophet | 10098.42 | 13085.60 | 15959.94 |
| 01/06/2025 | XGBoost | 12832.45 | 17028.24 | 18394.77 |
| 01/07/2025 | ES | 12172.77 | 14437.72 | 16702.67 |
| 01/07/2025 | SARIMA | 4023.85 | 14997.26 | 25970.67 |
| 01/07/2025 | Prophet | 11320.45 | 14339.47 | 17344.92 |
| 01/07/2025 | XGBoost | 12524.90 | 18928.41 | 20348.17 |
| 01/08/2025 | ES | 15340.67 | 17605.62 | 19870.57 |
| 01/08/2025 | SARIMA | 6257.22 | 17230.63 | 28204.03 |
| 01/08/2025 | Prophet | 15078.88 | 17839.20 | 20596.42 |
| 01/08/2025 | XGBoost | 12678.14 | 18954.62 | 19219.60 |

Taulukko 2. Puolen vuoden ennusteet

Kuviossa 8 on ennusteet ja visualisoitu 95 % luottamusväli. Toukokuulle on odotettavissa myynnin kasvua kausipaikkojen tullessa myyntiin, mutta Prophetin ennuste, jopa luottamusvälin alarajalla, vaikuttaa epärealistisen suurelta.




Kuvio 8. Puolen vuoden myyntiennusteet 95 % luottamusvälillä malleittain

Kenkien nettomyynti maaliskuussa oli 3666,98 euroa ja käyntien 7044,02 euroa. Taulukossa 3 on esitetty mallien ennusteet puolelle vuodelle. Kenkämyynnissä lähimmäksi osui SARIMA -279,29 euron erolla ja käynneissä Prophet 490,18 euron ylityksellä. Maaliskuulle osui eräät messut, jotka kasvattivat kenkämyyntiä. Kun kenkämyynnistä vähentää messujen vaikutuksen, Prophetin ennustus on vain 52,41 euroa väärässä. Näiden vertailujen perusteella Prophet suoriutui jokaisessa kategoriassa parhaiten. Messujen vaikutus selittää osaltaan myös Prophetin tarkkuutta kokonaismyynnin ennustamisessa. Testausdatalla Prophetin ennusteet olivat järjestelmällisesti liian suuria, joten yllättävä lisämyynti toimi Prophetin eduksi.

**kengät**

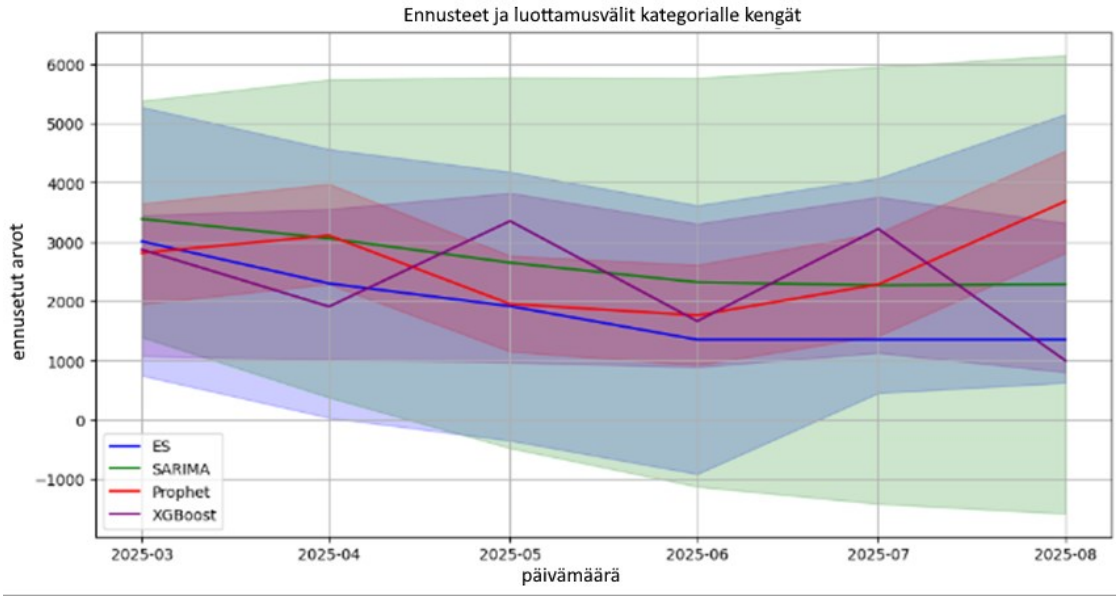
| kuukausi | malli | alaraja | ennuste | yläraja |
|---|---|---|---|---|
| 01/03/2025 | ES | 745.12 | 3010.07 | 5275.02 |
| 01/03/2025 | SARIMA | 1393.65 | 3387.69 | 5381.74 |
| 01/03/2025 | Prophet | 1956.18 | 2815.11 | 3648.62 |
| 01/03/2025 | XGBoost | 1073.22 | 2871.75 | 3449.59 |
| 01/04/2025 | ES | 34.92 | 2299.86 | 4564.81 |
| 01/04/2025 | SARIMA | 378.06 | 3059.29 | 5740.51 |
| 01/04/2025 | Prophet | 2289.44 | 3109.37 | 3978.86 |
| 01/04/2025 | XGBoost | 1016.15 | 1910.17 | 3552.88 |
| 01/05/2025 | ES | -348.81 | 1916.13 | 4181.08 |
| 01/05/2025 | SARIMA | -478.09 | 2649.98 | 5778.05 |
| 01/05/2025 | Prophet | 1152.61 | 1953.13 | 2767.16 |
| 01/05/2025 | XGBoost | 968.50 | 3352.75 | 3827.76 |
| 01/06/2025 | ES | -913.21 | 1351.73 | 3616.68 |
| 01/06/2025 | SARIMA | -1126.59 | 2320.47 | 5767.53 |
| 01/06/2025 | Prophet | 922.88 | 1762.64 | 2613.66 |
| 01/06/2025 | XGBoost | 890.90 | 1664.43 | 3307.04 |
| 01/07/2025 | ES | 453.56 | 1351.73 | 4076.34 |
| 01/07/2025 | SARIMA | -1415.86 | 2268.82 | 5953.51 |
| 01/07/2025 | Prophet | 1417.24 | 2284.45 | 3153.33 |
| 01/07/2025 | XGBoost | 1130.72 | 3220.43 | 3759.13 |
| 01/08/2025 | ES | 626.07 | 1351.73 | 5155.97 |
| 01/08/2025 | SARIMA | -1582.77 | 2283.26 | 6149.29 |
| 01/08/2025 | Prophet | 2816.66 | 3685.59 | 4539.80 |
| 01/08/2025 | XGBoost | 804.16 | 998.52 | 3319.64 |

**käynnit**

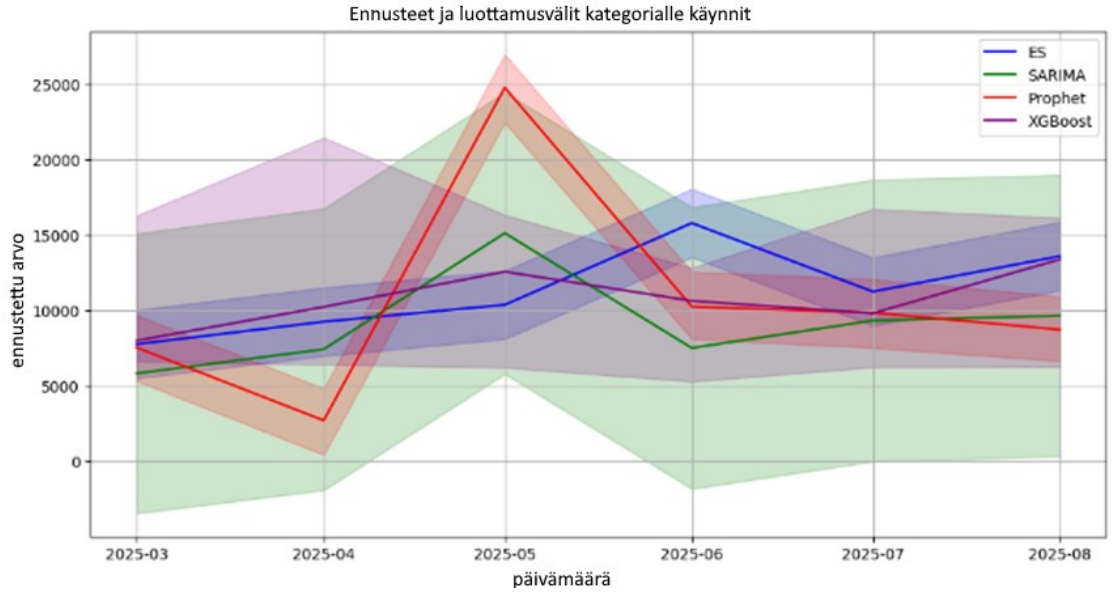
| kuukausi | malli | alaraja | ennuste | yläraja |
|---|---|---|---|---|
| 01/03/2025 | ES | 5500.97 | 7765.91 | 10030.86 |
| 01/03/2025 | SARIMA | -3447.38 | 5825.70 | 15098.77 |
| 01/03/2025 | Prophet | 5322.15 | 7534.40 | 9674.85 |
| 01/03/2025 | XGBoost | 6609.72 | 7996.78 | 16290.18 |
| 01/04/2025 | ES | 6986.68 | 9251.63 | 11516.58 |
| 01/04/2025 | SARIMA | -1906.23 | 7424.80 | 16755.83 |
| 01/04/2025 | Prophet | 431.44 | 2710.61 | 4824.65 |
| 01/04/2025 | XGBoost | 6416.67 | 10232.78 | 21452.77 |
| 01/05/2025 | ES | 8106.86 | 10371.80 | 12636.75 |
| 01/05/2025 | SARIMA | 5792.93 | 15123.96 | 24454.99 |
| 01/05/2025 | Prophet | 22445.01 | 24763.46 | 26968.12 |
| 01/05/2025 | XGBoost | 6217.48 | 12566.42 | 16327.01 |
| 01/06/2025 | ES | 13522.76 | 15787.71 | 18052.66 |
| 01/06/2025 | SARIMA | -1820.68 | 7510.35 | 16841.38 |
| 01/06/2025 | Prophet | 8106.46 | 10248.83 | 12558.86 |
| 01/06/2025 | XGBoost | 5294.03 | 10632.47 | 12820.99 |
| 01/07/2025 | ES | 8981.59 | 11246.54 | 13511.49 |
| 01/07/2025 | SARIMA | -3.17 | 9327.86 | 18658.89 |
| 01/07/2025 | Prophet | 7531.97 | 9847.54 | 12067.04 |
| 01/07/2025 | XGBoost | 6238.06 | 9809.92 | 16718.80 |
| 01/08/2025 | ES | 11329.54 | 13594.49 | 15859.44 |
| 01/08/2025 | SARIMA | 321.25 | 9652.28 | 18983.30 |
| 01/08/2025 | Prophet | 6640.69 | 8729.04 | 10926.38 |
| 01/08/2025 | XGBoost | 6274.56 | 13383.17 | 16131.05 |

Taulukko 3. Kenkien ja käyntien myyntiennusteet

Kenkämyynnin aikasarjan erittelykuvaajassa trendi laski voimakkaimmin alun nousun ja tasaisen jakson jälkeen. Todennäköisesti tämän trendin muutoksen takia SARIMA ja Holt-Winters antavat suuren epävarmuuden ennusteilleen. Kuviossa 9 Prophetin 95 % luottamusvälillä olevalla ennusteella on kaikista kapein euromääräinen vaihtelu, eli Prophetin ennuste on tarkin. Kuviossa 10 käyntien ennusteissa Prophet antaa pienimmän luottamusvälin ennusteille. Kesän lukujärjestyksen tunnit tulevat myyntiin toukokuussa, joten piikki myynnissä on oletettava ja toivottu, mutta Prophetin ennuste vaikuttaa epärealistisen suurelta.




Kuvio 9. Ennusteet ja luottamusvälit kategorialle kengät




Kuvio 10. Ennusteet ja luottamusvälit kategorialle käynnit

Vaikka ennusteet on toteutettu puolelle vuodelle, niihin tulee suhtautua varauksella. Data-joukossa on havaittavissa kausittaisuutta, mutta neljä vuotta on erittäin lyhyt aika ennustaa kausivaihtelua pidemmällä aikavälillä. Ennusteiden epävarmuus on suuri, nettomyyntien-nusteet saattavat mennä jopa miinuksen puolelle. Jopa pienimmän luottamusvälin antavalla Prophetilla ennusteiden ylärajan ja alarajan erotus on joka kuussa keskimäärin noin 6000 euroa, joka on todella suuri vaihteluväli verrattuna myynnin mittakaavaan. Vaikka yhden

kuukauden eteenpäin ennustaminen onkin osunut aika tarkasti, kuuden kuukauden päässä on todennäköisesti enemmän epävarmuutta – jopa Prophetilla.

Datassa on trendin muunnos, joka tuo lisää epävarmuutta ennusteeseen. On mahdotonta sanoa, jatkaako trendi laskuaan, pysyykö trendi tasaisena vai kääntyykö se kasvuun. Prophet mallintaa kausivaihtelua joustavasti, myös epälineaarisia trendejä, mutta lyhyen aikasarjan takia sen ennuste voi olla ylikorjaava tai liian herkkä yksittäisille poikkeuksille. Ennustetta kannattaa käyttää korkeintaan suuntaa antavana työkaluna päätöksenteossa tarkkojen lukujen sijaan.

# 5 Yhteenveto ja pohdinta

Opinnäytetyön tavoitteena oli ennustaa yritys X:n seuraavan puolen vuoden myyntiä tilastollisten menetelmien ja koneoppimisen keinoin, ja vertailla eri malleja. Vertailtaviksi malleiksi valikoitui datajoukon perusteella Holt-Winters, SARIMA, Prophet ja XGBoost. Mallien toimivuutta arvioitiin laskemalla keskimääräinen ennustevirhe, keskimääräinen absoluuttinen virhe ja keskineliövirhe. XGBoostilla saavutettiin pienimmät arvot virheille. Malleilla toteutettiin puolen vuoden ennuste maaliskuusta alkaen, ja tästä parhaiten, jopa yllättävän tarkasti, selvisi Prophet. Ennusteiden luomiseen käytettiin ainoastaan myyntidataa, eikä ulkoisia tekijöitä otettu huomioon.

Käytettävä datajoukko oli melko pieni, joten vaikka ensimmäisen kuukauden ennuste Prophetilla oli melko lähellä totuutta, malli todennäköisesti heikkenee ajan kuluessa. Ennusteiden luottamusväli myyntimäärään suhteutettuna oli melko suuri. Ennustuksiin tulee siis suhtautua varauksella. Ensimmäiselle ennustettavalle kuukaudelle osui myös eräät messut, jotka nostivat kenkämyyntiä. Tämä kenkämyynnin nousu vaikutti kokonaismyyntiin vajaalla tuhannella eurolla, ja selittää osaltaan Prophetin tarkkuutta kokonaismyynnin ennustamisessa. Yrittäjälle voisi olla hyödyllisempää tarkastella erillisiä kenkämyynnin ja käyntien ennusteita kokonaismyynnin ennusteiden sijaan.

Toimeksiannosta haastavinta oli työn toteuttaminen itsenäisesti. Toimeksiantajalla ei ole osaamista koneoppimisesta tai aikasarjoista. Työtä tehtäessä tutustuttiin erilaisiin tilastollisiin metodeihin ja koneoppimismalleihin, joilla voi käsitellä aikasarjoja. Tästä tutustumisesta oli hyötyä sopivimpien mallien löytämiselle, Prophet ei ollut mukana alkuperäisessä ajatuksessa käytettävistä malleista.

Malleissa ei ole käytetty ulkoisia tekijöitä selittävinä tekijöinä. Maailmantilanne on epävakaa, Suomen työttömyysluvut ovat nousussa ja liikuntapaikkojen alv-korotuksesta on lyhyt aika. Kirjoitushetkellä ei ole tiedossa, että EU olisi asettanut Yhdysvalloille vastatulleja jalkineiden osalta, mutta todennäköisesti maailmantilanteella tulee olemaan vaikutusta kenkien hintoihin. Varsinkin, kun yritys maahantuo kiinalaisvalmisteisia kenkiä Yhdysvalloista. Vaikka ennustukset on toteutettu nettohinnoilla, eli verokannan muutos ei näy myynnissä euroissa, voi kenkien nouseva hinta ja kallistuneet tuntimaksut vaikuttaa myyntimääriin.

Jatkokehityksenä yrityksen myyntidatan kategorioiden rakennetta voisi yhtenäistää, jos on tarkoitus jatkaa koneoppimismallien hyödyntämistä, esimerkiksi maaliskuun vertailuun käytetty data tuli puhdistaa ja kategorioita yhdistää vastaamaan mallien kouluttamiseen käytettyä dataa. Ennusteiden tarkkuutta voisi yrittää parantaa lisäämällä ulkoisia tekijöitä mallien koulutusdataan. Näitä ulkoisia tekijöitä voisi olla esimerkiksi markkinointikampanjat,

inflaatio- ja työttömyysluvut tai jopa keli – ehkä sateisena kesänä on enemmän kävijöitä kuin aurinkoisena? Toisaalta, vaikka dataa keräisi vielä toiset neljä vuotta, datajoukko on silti melko pieni koneoppimismallien kouluttamiseen. Kenkämyynnille voisi toteuttaa kappalemäärään perustuvan myyntiennusteen, sillä kyseisen kategorian kohdalla sillä voisi saada parempia tuloksia.

## Lähteet

Adhikari, R. & Agrawal, R. K. 2013. An Introductory Study on Time Series Modeling and Forecasting. Saarbrücken, Germany. LAP LAMBERT Academic Publishing.

Alpaydin, E. 2021. Koneoppiminen. Suomentanut Kimmo Pietiläinen. Helsinki: Terra Cognita Oy.

Amr, T. 2020. Hands-On Machine Learning with Scikit-learn and Scientific Python Toolkits: A Practical Guide to Implementing Supervised and Unsupervised Machine Learning Algorithms in Python. Birmingham, UK: Packt Publishing.

Atwan, T. A. 2022. Time Series Analysis with Python Cookbook : Practical Recipes for Exploratory Data Analysis, Data Preparation, Forecasting, and Model Evaluation. Birmingham, UK: Packt Publishing.

Auffarth, B. 2021. Machine Learning for Time-Series with Python : Forecast, Predict, and Detect Anomalies with State-of-the-art Machine Learning Methods. Birmingham, UK: Packt Publishing.

Boschetti, A. & Massaron, L. 2018. Python Data Science Essentials : A Practitioner's Guide Covering Essential Data Science Principles, Tools, and Techniques. Birmingham, UK: Packt Publishing.

Box, G. E. P., Jenkins, G. M., Reinsel, G. C. & Ljung, G. M. 2016. Time Series Analysis : Forecasting and Control. Fifth Edition. Hoboken, New Jersey: Jonh Wiley & Sons, Inc.

Geeksforgeeks. 2024. Time Series Analysis using Facebook Prophet. Viitattu 17.4.2025. Saatavissa https://www.geeksforgeeks.org/time-series-analysis-using-facebook-prophet/

Hyndman, R. J. & Athanasopoulos, G. 2021. Forecasting: principles and practice, 3rd edition, OTexts: Melbourne, Australia. Viitattu 28.3.2025. Saatavissa OTexts.com/fpp3

Krieger, M. 2021. Time Series Analysis with Facebook Prophet: How it works and How to use it. towards data science. Viitattu 17.4.2025. Saatavissa https://towardsdatascience.com/time-series-analysis-with-facebook-prophet-how-it-works-and-how-to-use-it-f15ecf2c0e3a/

Johnston, B. & Mathur, I. 2019. Applied Supervised Learning with Python : Use Scikit-Learn to Build Predictive Models from Real-world Datasets and Prepare Yourself for the Future of Machine Learning. Birmingham: Packt Publishing Ltd.

Montgomery D. C., Jennings C. L. & Kulahci, M. 2016. Introduction to Time Series Analysis and Forecasting. Hoboken, New Jersey: John Wiley & Sons, Inc.

Nelli, F. 2018. Python Data Analytics. With Pandas, NumPy, and Matplotlib. Second Edition. Rome, Italy: Apress.

Nvidia. XGBoost. Viitattu 5.4.2025. Saatavissa https://www.nvidia.com/en-us/glossary/xgboost/

Rafferty, G. 2021. Forecasting Time Series Data with Facebook Prophet : Build, Improve, and Optimize Time Series Forecasting Models Using the Advanced Forecasting Tool. Birmingham: Packt Publishing.

Wade, C. 2020. Hands-On Gradient Boosting with XGBoost and Scikit-learn : Perform Accessible Machine Learning and Extreme Gradient Boostin with Python. Packt Publishing.

Wang, Z., Irfan S. A, Teoh, C. & Bhoyar, P.H. 2023. Numerical machine learning. Sharjah: Bentham Science Publishers.

Shalev-Shwartz, S. & Ben-David, S. 2014. Understanding Machine Learning. From Theory to Algorithms. New York, USA: Cambridge University Press.

Shumway, R. & Stoffer, D. 2011. Time Series Analysis and Its Applications. With R Examples. Third edition. London: Springer Science+Business Media.

Stan. 2021. How to do Time Series Split using Sklearn. Medium. Viitattu 17.4.2025. Saatavissa https://medium.com/@Stan_DS/timeseries-split-with-sklearn-tips-8162c83612b9