Norges miljø- og biovitenskapelige universitet

Masteroppgave 2021 30 stp

Fakultet for Realfag og Teknologi
Tor Kristian Stevik
Jesper Frausig

Predikering av kursutviklingen for Bitcoin, ved bruk av maskinlærings modeller trent på markedsinformasjon.

Magnus Lillehaug
Mastergrad i Industriell økonomi

Forord

Innleveringen av denne masteroppgaven markerer avslutningen på et spennende og alternativt masterløp. Studiet er gjennomført ved Norges Miljø- og Biovitenskapelige Universitet, NMBU.

I denne oppgaven har jeg sett på hvilke variabler som påvirker en maskinlæringsmodell evne til å korrekt predikere kursutvikling til Bitcoin. Jeg ønsker å takke mine veileder, Tor Kristian Stevik og Jesper Frausig for gode, interessante samtaler og konstruktive tilbakemelding. Stor takk går til min samboer som har gjort det mulig å gjennomføre både denne oppgaven og resten av studiet, og min lille sønn som sørge for svært kjerkommene pauser. Til slutt en, stor takk til venner og familie som har hjulpet med å sitte barnevakt, støtte, positive ord og spesielt korrekturlesning.

Magnus Vaa Lillehaug

Bergen 01.06.2021

Sammendrag

Denne oppgaven har tatt utgangspunkt i kryptovalutta markedet, mer spesifikt Bitcoin. Dette er et marked som hatt en eksponentiell vekst det siste tiåret.
Målet har vært å se hvordan ulike variabler kan anvendes for å predikere kursutvikling ved hjelp av maskinlæringsmodeller. Maskinlæringsmodellen som anvendes er en toppmoderne sekvensiell læringsarkitektur, *long short term memory*.
Det å utføre prisprediksjon med denne typen maskinlæring er et felt som gjennom nyere litteratur viser har et potensiale. Andre forsøk på denne type arbeid, med long short term memory modeller, har benyttet primært tekniske indikatorer og tradisjonell pris data som grunnlag for disse modellene.
I denne oppgaven utvides datasettet ved å utnytte det brede spektre med variabler som finnes i markedsinformasjonen til kryptovalutta. Dette tilsvarer opp mot 100 ulike variabler som gir informasjon om ulike felt innenfor BTC. Av resultatene viser denne typen data potensial til å bidra til å øke nøyaktigheten for kurspredikering av Bitcoin. Modellen som blir presentert, blir analyser ved å andvende SHAP analyse, dette bidrar til å gi innsikt i hvordan modellen andvender de ulike variablene.

Abstract

This theses has been based on the cryptocurrency market, more specifically Bitcoin. The aim has been to see how different features can be used to predict future price, using machine learning models. The specific machine learning model used is a state-of-the-art sequential learning architecture, *long short term memory*.
Performing price prediction with this type of machine learning has been done in recent literature, and has shown to have a potential. Other attempts at this type of work, have been primarily done through technical indicators, news, and Bitcoin network data, as datasets for the the models. This thesis has expanded by utilizing the broad spectrum of features, called on-chain data. This is done by utilizing the wide range of variables contained in on-chain for cryptocurrencies. By looking at up to 100 different variables this thesis will explore their impact on a deep learning model. The results shows the potential to help increase the accuracy of Bitcoin's price prediction. The model that is presented is then analyzed by using SHAP analysing tools. This helps to provide insight into how extract value from the various variables. By doing this, the model becomes transparent and we get a insight into why it predicts as it does.

Innhold

1 Innledning

Teknologisk utvikling har revolusjonert måten finansielle aktivum handles. De største finansielle markedene er preget av automatiserte systemer. Algoritmisk trading står for mellom 60-73 % av all aksjehandel i det amerikanske markedet (Intelligence, 2021)[1]. Med en stadig økning av tilgjengelig data, kommer det til et punkt der det blir for mange variabler for et menneske å holde oversikt over. For å utnytte store mengder data på en effektiv måte, skapes det behov for verktøy som kan bidra i en beslutningsprosess.

Long short-term memory (LSTM) nettverk er en av de mest avanserte læringsarkitekturene for sekvensiell læring. Dette innbefatter oppgaver som bildegjenkjenning, håndskriftgjenkjenning eller tidsserieprediksjoner (Graves et al., 2013). Det å anvende LSTM til prisprediksjon innen finansielle markeder har vært gjort med positive resulter, der modellene har gitt en risikojustert avkastning som er høyere en markedet [2]. Som vist av Fischer and Krauss (2017) presterer LTSM best av flere deep learning modeller, samt enkle regresjonsmodeller [3]. Videre er dette testet innenfor kryptovalutamarkedet av Alonso-Monsalvea et al. (2019). Resultatene viste at ved å anvende LTSM trent på tradisjonelle finansielle data med tekniske analyseindikatorer, kunne de predikere prisendringer.

Dietvorst et al. (2015) har undersøkt fenomenet algoritmeaversjon. Mennesker har problemer med å stole på algoritmer som er imperfekte. Dette er et viktig aspekt dersom maskinlæring skal brukes som beslutningsstøtte innenfor prisprediksjon. En modell vil ha en nøyaktighetet som ikke kan forventes å være høyere 55-60 % [4], derfor er det essensielt at det er transparent og forståelig for et menneske hva som vektlegges av modellen, og dersom mulig, enkelte av dens prediksjonsmønster.

Markedet jeg ønsker å benytte som finansielt instrument er kryptovalutamarkedet, mer spesifikt Bitcoin. Dette er et ungt marked som har hatt en eksponentiell prisvekst, siden Bitcoin ble oppfunnet i 2009, har prisen steget fra 0.0008$ til over 60 000$ på sitt høyeste.

[1] I dette ligger det hvordan handler er utført, hvor stor andel av marked som er hel automatisert er ikke klart

[2] Det er verdt å merke seg at det er vanskelig å få et komplett bilde fordi det er mange variabler innenfor maskinlæring som kan medføre at man får gode resultater, men som ikke er overførbare til et reelt marked.

[3] Dette forsøket ble utført på det amerikanske aksjer markedet, mer spesifikt på SP 500, over en tidsperiode fra 1992 til 2015 med 1 minutts tidsintervall

[4] Dette er et estimat, basert på annen litteratur

De siste årene har det oppstått et nytt kommersielt felt innenfor dataanalyse av det digitale valutamarkedet. Dette er on-chain data eller markedsinformasjon som det blir referer til i oppgaven. Markedsinformasjon er en datavariablel som bidrar med informasjon om mange ulike aspekter ved kryptovalutamarkedet, med blant annet informasjon om hvordan digitale eiendeler beveger seg mellom ulike adresser på offentlige blokkjedenettverk. Denne informasjonen anvendes som grunnlag for en rekke signifikante statistiske beregninger som kan anvendes som indikatorer(CryptoQuant, 2021). Denne typen data kan potensielt bidra med informasjon som fortelle om årsaken til prisbevegelser for Bitcoin (Coinbase Institutional, 2020). Dersom dette stemmer, kan dette anvendes for å underbygge analyser med prisprediksjoner for fremtidig bevegelser.

Jeg ønsker å bruke denne oppgaven til å bygge en LTSM modell som trenes på datasett bestående av markedsinformasjon, data direkte fra en kryptovaluta børs og tekniske indikatorer. Hensikten er å teste om variablene kan brukes til prisprediksjon for Bitcoin. Videre ønsker jeg å se hvordan en slik modell kan ha potensiale som et verktøy for et menneske som er involvert i handel av Bitcoin.

1.1 Problemstilling

For å forstå hvordan deep learning kan brukes for å finne mønster i komplekse og kaotiske data for å predikere prisbevegelser innenfor et finansielt aktiva, tar oppgaven utgangspunkt i Bitcoin. Tilgangen på informasjon gjør at denne oppgaven kan teste et stort antall variabler og hvordan disse påvirker prediksjonene til en maskinlæringsmodell. Modellarkitekturen som vil bli benyttet er long short term memory. Dette er på bakgrunn av annen litteratur som har vist at denne arkitekturen har potensiale til å løse slike problemer. Videre vil det bli undersøkt hvilken informasjon en trader bør ha dersom en slik modell skal brukes som et verktøy til beslutningstagning. Med utgangspunkt i dette har jeg formulert følgende problemstilling:

- **Finnes det variabler som en LSTM-modell kan andvende for å predikere kortsiktig kursutvikling for Bitcoin?**

Med grunnlag i arbeidet som blir gjort i forbindelse med problemstillingen, er det å se på hvordan dette kan overføres til en anvendelse i et reelt markeds miljø. Vinklingen som vil bli belyst i denne oppgaven er hvordan dette kan gjøres og hvilke utfordringer det er å ta i bruk et maskinlærings verktøy for et mennekse. Utfra dette har jeg definert følgende forskningsproblem:

- **Under en implementering av en maskinlæringsmodell som et støtteverktøy til en trader i investeringsøyemed. Hvilken teknisk informasjon om modellen kan traderen ha nytte av?**

1.2 Begrepsliste

- Modell: Modell blir mye brukt i denne oppgaven. Med ulike modeller menes resultatet av en LSTM som er trent på ulike datasett, dvs. datasett 1 = modell 1, datasett 2 = modell 2.

- Trader: Er en person som kjøper og selger finansielle aktiva, i denne oppgaven er dette Bitcoin. Målet for en trader er å maksimere profitt og minimere risiko. Personen antas å være godt informert om markedet. Jeg begrenser hans trading til å basere seg rundt høyfrekvent trading.

- Long: er en markedsposisjon som profitterer på prisoppgang

- Short: er en markedsposisjon som profittere på prisnedgang

- Nøyaktighet: med dette refereres det til maskinlæringsmodellens nøyaktighet, formelen er $nyaktighet = \frac{Rettpositiv+rettnegativ}{Rettpositiv+rettnegativ+feilpositiv+feilnegativ}$

- Treningssett: Dette er det datasettet modellen lærer fra, ved at den får fasiten.

- Tests-sett: Med dette begrepet menes den delen av datasettet som blir adskilt fra treningssettet. Det er på denne delen modellens nøyaktighet blir målt.

- Overfitted: når en maskinlæringsmodell er trent på en slik måte at den er spesialisert på treningssettet. Dette vil medføre dårlige generalisering der modellene ikke fungerer ved å deployere den på et reelt marked.

- Med 1 run, menes en fullstendig trening av modellen.

- Closing price: er prisen ved slutten av et tidsintervall. Dataformatet i oppgaven er 1 minutt, så dvs. verdien ved slutten av det minuttet.

1.3 Program- og maskinvare

Programvare og maskinvare som er benyttet i oppgaven:

python - 3.8
tensorflow-gpu - 2.5.0
scikit-learn - 0.24.2
shap - 0.39.0
optuna- .7.0
seaborn - 0.11.1
Boruta 0.3
numpy - 1.19.5
neptune-client - 0.9.13
pandas - 1.2.4
TA-Lib 0.4.20
Windows 10
AMD 5600
Nvidi RTX 3080
HyperX 32 GB ram

1.4 Avgrensninger

Tekniske begrensninger

For å ha nok datapunkter for bruk av LSTM, settes tidsintervallet til datapunktene til ett minutt. Modellen som blir konstruert vil ha en tidsbegrenset periode der den kan ses som gyldig. Etter dette vil det forekomme en degradering av nøyaktighet knyttet til prediksjonene. Det er et stort antall ulike maskinlæringsmodeller, og flere har vist potensialet til prisprediksjon. I denne oppgaven andendes kun LSTM. Det å utføre prediksjoner, uansett marked, er en svært kompleks oppgave, og for å bruke maskinlæring til det stilles det store krav til datasettet som skal benyttes. For å ikke gjøre oppgaven for generell, vil det kun bli brukt bitcoin som aktiva.

Besvarelse av forskningsspørsmål to begrenses til å undersøke hvordan LSTM-modellen kan analyseres for hente ut relevant informasjon som en person som skal bruke modellen potensielt vil ha behov for.

Analyse av viktige variabler som ikke blir tatt med i oppgaven, men som har potensialet til å påvirke kursutvikling

For å hente data til denne oppgaven benyttes tre kilder: (1) markedsinformasjon hentet fra en datatilbyder spesialisert på kryptovalutta, (2) data direkte fra kryptovalutta børsen Binance og (3) tekniske indikatorer som konstrueres fra stengningspris og volumindikatorer. Dette resulterer i ca 140 ulike variabler. Det er ellers mange viktige variabler, spesielt makroøkonomiske, som vil ha potensielt stor påvirkning på de som brukes som ikke blir dekket i oppgaven.

Til tross for at de ikke tas med så er det viktig å forsøke å avdekke de og ha et forhold til disse variablene. For å strukturere dette gjennomføres en PESTEL-analyse for å forsøke å avdekke de viktigste variablene som ikke blir brukt i oppgaven, men potensielt kan ha stor påvirkning på prisutvikling. Det å avdekke disse er utfordrende. Mange av hendelsene er ikke godt dokumenter, og fallet sammenfaller ofte med nyhetene om endringer i de makroøkonomiske variablene[5].

1. **Politisk**: Skatteregler. I følge Sharma (2020) kan endringer i skatteregler påvirke prisen til BTC.

2. **Økonomisk**: Lave styringsrenter, spesielt i USA, har potensielt en positiv innvirkning på prisen til BTC DiCamillo (2021)

3. Kjøpsstyrke som konsekvens av stimmulus check [6] i USA, kunne potensielt øke etterspørselen etter BTC (Yahoo finance, 2021)

4. **Sosiale**: Twitter har vist seg å ha en påvirkning til BTC sitt handelsvolum som vist av Shen et al. (2019). Oppdukkende sosiale trender endre investeringsmønster. Eksempel på dette er overføringseffekten fra situasjonen med GameStop [7] som ledet oppmerksomhet til kryptovalutta som investeringsobjekt Sigalos (2021)

[5]Viktig å merke seg at kildene som er brukt i denne seksjonen er nyheter. Dette er for å trekke frem nylige hendelsene. Denne avveining tas for å få frem hendelser som kan ha påvirket oppgavens datasett

[6]Stimmulus check var støttemidler alle amerikanske statsborgere over 18 år mottok som COVID-19 støtte, dette ble gjort to ganger og var på 1400$

[7]GameStop aksjen ble et internettfenomen våren 2021, GameStop var på dette tidspunktet en av de mest aksjene med mest spekulasjon på short siden, dette var utført av hedge fond. Småinvestorer begynte å kjøpe opp aksjen og dette ble til en sosial fenomen, aksjeprisen steg fra 20 % til 300 % på kort tid. Dette medførte enorme tap for hedgfonende.

5. **Teknologiske**: Konsekvensen av utviklingen av en Quantumkomputer kan komprimere BTC blokkjeden Barmes (2019)

6. **Miljø**: Et stort strømbrudd i et av de største områdene for BTC mining i Kna, medførte et signifikant fall i BTC hashrate. Et kraftig prisfall for BTC sammenfalt med hasrate falletwea.

7. **Juridiske**: Ulike regelverk, som totalforbud, kan bidra til plutselig tilførsel av BTC til markedet som kan lede til prisnedgang. Eksempelvis India sitt totalforbud som sammenfalt med et fall i BTC pris Makowski (2021).

1.5 Beskrivelse av lignende litteratur

I dette kapittelet blir det gjennomgått lignende litteratur, dette er et teknisk kapitel som bruker mange begreper som blir gjennomgått i teoridelen av denne oppgaven. Det trekkes frem hvilke resultater andre oppnår i sitt arbeid, det er viktig å bemerke at slike resultater ikke nødvendigvis ville vært oppnåelige i et reellt marked.
McNally et al. (2018) testet LSTM RNN og ARIMA på et datasett med BTC priser [8]. De oppnådde en prediksjon nøyaktighet på fremtidig pris på 52.79% og en RMSE på 6.87%. Chen et al. (2020) og forbedret dette resultatet ved å utvide treningsdatasettet. Ved å legge til flere nettverksstatistikk, Google trends[9] og gullprisen. Ved å trene modellen på høydimensjonal data, oppnådde de en nøyaktighet på 67.2 %.

Bao et al. (2017) testet LSTM og RNN opp mot Kjøp og hold strategier på 6 ulike indekser [10]. Ved å bygge modellen til å predikere neste dags stengepris, oppnådde LSTM en årlig profitabilitet som gjorde det bedre enn kjøp og hold[11] i gjennomsnitt over en 6 års periode. De innførte også en ny variant av LSTM, ved å kombinere wavelet transform og stacked autoencoder med LSTM til en modell de kalte *WSAEs-LSTM*. Denne modellen oppnådde bedre resultater en standard LSTM.

Alonso-Monsalvea et al. (2019) demonstrerer hvordan en kombinasjon av konvolusjon neural nettverk sammen med LSTM. Måten Alonso-Monsalvea et al. (2019) behandler data på er fundamentalt annerledes en i denne oppgaven. De omgjorde prisen til de ulike digitale valuta om til bilder. Så å kjørte de disse gjennom en hybriden C-LSTM. Dette gir en signifikant bedre prisprediksjon på flere kryptovaluta [12] en multilayer perceptron, enkel CNN og radial basis function neural network. Datasettet de anvendte innbefatter både tekniske indikatorer og OHLC-V forulike kryptovaluta. Ved å gjøre dette på flere alternative digitale valuta kunne de vise at prispredikering også fungerer på andre kryptovaluta enn BTC.

[8] Datasettet inneholdt OHLC-V fra 2013 til midten av 2016. Den innehold også BTC nettverksdata og to simple moving avrages

[9] Google trends er aggregert søkedata hentet fra google på spesifikke søkeord. I denne oppgaven var dette BTC

[10] CSI 300, Nifty 50, Hang Seng, Nikkei, S&P 500 og DJIA

[11] Kjøp og hold er en enkel handels strategi der man kjøper og holder et aktivium over en lengre tidsperiode.

[12] BTC, Dash, Ether, Litecoin, Monero og ripple

Basert på mine egne undersøkelser, er det ingen artikler som anvender LSTM eller andre maskinlæringsalgoritmer på et datasett som inkluderer et vidt spekter med markedsinformasjon variabler. Som demonstrert av Alonso-Monsalvea et al. (2019); Chen et al. (2020); Bao et al. (2017); McNally et al. (2018) har LSTM som modell potensialet til å predikere fremtidig pris for ulike kryptovaluta. Resultatene har vist seg å bli mer nøyaktige med mer sammensatt og komplette datasett. Dette har gitt inspirasjon til å prøve ut LSTM på et komplekst datasett med data direkte fra børs i kombinasjon med et stort utvalg tekniske indikatorer og markedsinformasjon variabler. I den hensikt å undersøke hvordan et slik verktøy potensielt vil kunne støtte beslutningstageren, eventuelt potensialet for helautomatisk tradingsystem bygget med LSTM modell.

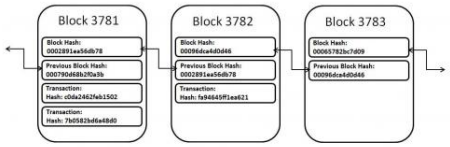
Del 1

Teori og metode

2 Teori

2.1 Kryptovaluta

Kryptovaluta startet med oppfinnelsen av BTC i 2008 av Satoshi Nakamoto. I 2009 ble teknologien utgitt som "open-source", fritt tilgjengelig for alle. BTC er et individ til individ elektronisk valuta system (Nakamoto, 2009). Systemet bygger på kryptografisk bevis, istedenfor tillit. Dette gjør at en transaksjon kan utføres direkte mellom to parter uten at det er behov for en tredjepart for å gjennomføre transaksjonen(Nakamoto, 2009). Blockchain (blokkjede) til BTC er data. Samlet i blokker og sammensatt i linkede kjeder, danner dette en distribuert database. Databasen inneholder alle transaksjonene som er gjennomført på nettverket(Nakamoto, 2009).
BTC løser to problemer som er sentrale for å lage en fungerende digital valutta. Det er: (1) å kontrollere hvordan valuta skapes. (2) den er uforanderlig, det er ikke mulig å duplisere, endre, manipulere, falsifisere eller duplisere (Velde, 2013).
Måten BTC løser dette problemet på er ved at det kreves et "proof of work" fra minere[13]. For at en transaksjon skal tilføres blokkjeden må den inneholde en løsning til et avansert matematisk problem. Dette er kostbart og krever datakraft, elektrisitet og tid. Det kryptografiske problemet er vanskelig å løse, men enkelt å verifisere løsningen. Problemet som må løses er knyttet opp mot verifiseringen av transaksjonen (Velde, 2013). Det å skape en virtuell valuta har vært forsøkt før BTC ble oppfunnet, men ingen av de forsøkene evnet å integrere dette i det tradisjonelle finansielle systemet (Glaser et al., 2014).



Figur 1: En illustrasjon av hvilken informasjon som er i blokkene i en blokkjede, illustrasjon hentet fra link

[13]Mining betyr at man ved å anvende datakraft kan løse kryptografiske ligninger.

Transaksjoner skjer mellom to "wallets" (digitale lommebøker), og alle disse er offentlige tilgjengelig. Dette gjør at man kan hente ut informasjon om alle bevegelsene i nettverket. *Dette er grunnlaget for markedsinformasjons dataen. Siden alle wallets er offentlig, kan man ekstraktere data om ulike felt for valutta, som ikke er tilgjengelig i andre finansielle markeder.* Det er ingen informasjon knyttet opp til hvem som eier de ulike adressene. For å sikre databasen er det et nettverk med fullestendige-node[14] som kjører software som ivaretar databasen. De som bidrar med datakraft inn i nettverket mottar insentiver. Dette gjøres ved å utstedt nye BTC som belønning til minerene (skjer ved et fast intervall). Fra starten av ble det tildelt 50 nye BTC pr. blokk. Dette tallet halveres pr 210 000 blokk (seks blokker per time, så hvert 4 år), og totalt vil det ende opp med $2 \times 50 \times 210,000 = 21 millioner BTC$ (Velde, 2013).

Glaser et al. (2014) hevder at kryptovaluta representerer en ny aktivaklasse som har flere likhetstrekk som kan sammenlignes med et spekulative aktiva, enn andre tradisjonell valuta. Dette er basert på mangelen på en håndgripelig verdi Alonso-Monsalvea et al. (2019).
Som et resultat av blokkjedeteknologien til BTC, har det blitt utviklet mange andre digital valuta med et vidt spekter av bruksområder. Det totale antallet kryptovaluta er over 4500 ifølge Statista (2021b). Totalt har kryptovalutamarkedet passert en samlet verdi på 1.6 billioner dollar. BTC har en pris på 36 000 dollar[15].

[14]En fullstendig-node er et program som har samtlige validerte transaksjoner på BTC-nettverket. Disse bidrar i arbeidet med å akseptere nye transaksjoner fra andre fulle noder, ved å validere transaksjoner og bloker og sender dem videre til andre fulle noder

[15]Data hentet fra tradingview den 30.05.2021

Figur 2: Illustrasjon av BTCs fulle prishistorikk, månedlig candlestick fra børsen BITSTAMP, for å få bildet i fullskala trykk på link

En av karakteristikkene til BTC og kryptovaluta er volatile perioder, med prisendringer som er svært hurtige og eksplosive, illustrert i figur 3

Dette er ikke et anomali, men hendelser som forekommer i kryptovalutamarkedet. Årsaken til disse hendelsene er vanskelig å forklare. I etterkant av kraftige prisbevegelser er det ofte ulike hendelser eller nyheter som knyttes til prisutviklingen som en potensiell årsak. Dette er ofte påstander som er svært utfordrende å bevise, og gjør at markedsdeltagere har behov for andre verktøy til å støtte dem i beslutningstagning.

Figur 3: Illustrasjon av BTC volatilitet, prisen falt 20.49 % på 50 minutter, for fullt bilde klikk på link

2.2 Markedsinformasjons

I dette delkapittelet blir det introdusert hva som menes med markedsinformasjon og forklart hvordan ulike variabler fungerer. ”On-chain” er den engelske terminologien for denne type data, i denne oppgaven blir den definert som markedsinformasjon.

Data-analyse av offentlige blokkjeder er et felt som analyserer hvordan digitale eiendeler beveger seg mellom ulike adresser. Dette er i en tidlig fase, men opplever en økende grad av utvikling og tilgjengelighet (Coinbase Institutional, 2020). Kryptovaluta har åpnet opp for helt nye typer datasett. Innen tradisjonell trading, er det meget vanskelig å innhente spesifikke tradingdata på grunn av personvern og sikkerhetslovverk og generelt hemmelighold. Dette er helt motsatt i kryptovaluta markedet. Blokkjedene er offentlig tilgjengelig, derfor kan det det hentes ut data som ikke er tilgjengelig i andre typer finansielle markeder (Zheng et al., 2020).

Markedsinformasjon er signifikante statistiske beregninger av aktiviteter på ulike blokkjede-nettverk. Denne type data har potensialet til å bidra til å forstå pris-bevegelser til ulike digitale valuta. Det finnes flere firma som driver denne typen analyse og selger disse dataene. I denne oppgaven har jeg valgt å bruke CryptoQuant. Dette valget er tatt med bakgrunn i type data de tilbyr, prising og hvor oppløselig datapunktene er.

Cryptoquant er et firma som selger data som innbefatter markedsdata, markedsinformasjon, både lang og kortsiktige indikatorer for BTC, Ethereum, stablecoins og ERC20 tokens (CryptoQuant, 2021).

For å forstå konseptet bak markedsinformasjon, sett fra et investeringsståsted, må man forstå de viktigste markedsdeltagerne. Disse kan kategoriseres som **børsene** - dette er hvor trading forekommer. Dette kan sees på som den primære etterspørselssiden i markedet. De siste årene har det vokst frem desentraliserte børser [16], handel på disse, blir ikke regnet med i markedsinformasjon. **Minerene** er aktøren som kontrollere tilførselen til markedet, ved at de utvinner nye coins som øker det totale antallet av en digital valuta. Siste er en **Whale** [17]. Denne aktøren utnytter markedets ineffektivitet ved å kontrollere begge sider av tilbud og etterspørsel. Dette er en viktig aktør fordi de har kapital til å bevege markedet i en ønsket retning (CryptoQuant, 2021). Det er et stor antall indikatorer tilgjengelig. Disse er delt inn i flere kategorier og dekker BTC, ETH [18], ERC-20 [19] og Stablecoin [20] blokkjeder.

Markedsinformasjon data, innebefatter flere nye variabler, det foreligger det hverken data eller annen forsking på innvirkningene de potensielt kan ha for å froklare eller predikere kursendringer ved bryk av maskinlæring. Ved å kombinere egen erfaring fra markedet og CryptoQuant beskrivelser, knyttes variablene til økonomiske mekanismene som tilbud og etterspørsel. I tabell 9 på side 87 er alle variabler brukt i oppgaven listet[21].

[16] En desentralisert børs er et markedssted der kryptovalutta kan veksles til andre direkte fra walleten.

[17] Utrykket beskriver en aktør med signifikante verdier innen kryptovaluta

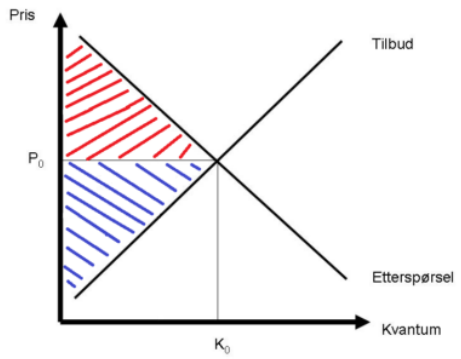
[18] Ethereum (ETH) er den nest største kryptovalutaen med en markedsandel på 196 milliarder dollar data hetnet fra tradingview den 28.03.2021

[19] Alle kryptovaluta som bygges på ETH nettverket betegnes som ERC-20

[20] Stablecoin er en kryptovaluta som er bygget i den hensikt å speile en tradisjonell valuta. Det er ulike tekniske løsninger for dette, bla. USD Tether (USDT), en kryptovalutta som holder verdien til 1 tradisjonell dollar. USDT har en påstått backing, så for hver USDT så holder selskapet bak 1 tradisjonell 1, altså en rate på 1:1.

[21] De viktigste variablene blir beskrevet i teksten, dersom det er ønskelig å lese mer om disse er det forklare på følgende link link

Et marked består av faktiske og potensielle kjøpere og selgere for en bestemt vare eller aktiva. For alle priser (høyere en nåværende) er det en kvantitet som tilbydere vil være villige til å selge for. Høyere priser fører til at selgeren blir mer villige til å selge. Kjøpere er på den andre siden, blir mindre villige til å kjøpe ved en høyere pris (Frank and Bernanke, 2007). Markedsinformasjondata har flere variabler som har potensialet til å kvantifisere tilbudssiden og forandringer knyttet til om tilbudet for kryptovaluta er stigende eller synkende.



Figur 4: Illustrasjon av tilbud - etterspørselskurve, hentet fra link

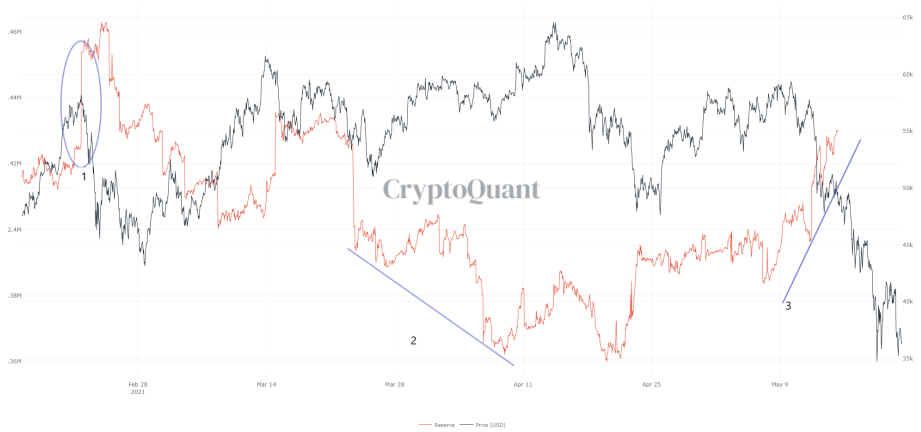
For å gi leseren en generell forståelse av de ulike kategoriene som faller innen markedsinformasjon, er påfølgende seksjon delt inn etter kategori. Enkelte beskriver hele kategorien generelt, mens andre trekker det ut i signifikante variabler som blir beskrevet mer detaljert.

Børs bevegelser

Børs bevegelser (på engelsk: Exchange Flows forkortet: XF) , er en kategori med variabler som gir antallet kryptovaluta som holdes på adresser tilknyttet ulike børser. Disse variablene gir et bilde av det kollektive potensialet av tilbudssiden. En signifikant økning vil kunne indikere økt sannsynlighet for kursnedgang. Dersom antallet tilgjengelig kryptovaluta synker vil dette medføre lavere tilbudsside som kan gi et signal på at prisen skal stige.

Det er av relevans for denne oppgaven, å se effekt av denne variabelen, som kan endringerer i pris skjer etter endringer i variabelen. I figur 5 er det markert inn 3 tydelige endringer i perioden Februar til Mai. Det fremgår av disse at markante endringer i BTC reserven får en innvirkning på pris. Store bevegelser til børs og ut av børs ikke nødvendigvis fører til prisutvikling som er invers korrelert. Dette er basert på egne observasjoner. Når tilbudssiden

økes fort, fører dette til mer volatil nedgang i pris, enn hvis et stort antall BTC trekkes ut av børsene. Dette kan være knyttet til at når BTC sendes til børs, er det i den hensikt å selge den til enten en annen kryptovalutta eller til dollar. Dersom BTC trekkes av børs, kan dette ha implikasjoner for pris, men ikke nødvendigvis like hurtig.



Figur 5: BTC reserver på børs, her er gul linje antallet BTC mens sort linje illustrer prisutvikling på samme tid. 1,2,3 illustrer områder av interesse. Her er det tydelig endring av tilbudet til BTC på børs, endringene skjer før endringen i pris.

Bevegelses indikator

Miner position index (MPI) er en variabel i kategorien bevegelses indikator (på engelsk: Flow Indicator, forkortelse: FI). MPI er en ratio for alle transaksjoner som gjøres fra BTC minere, delt på det årlige gjennomsnittet. Dette er en indikatorer som gir en indikasjon på om fortjenesten fra mining blir solgt på børs, eller om minerene holder. Dette er en indikator som gir et bilde av miner oppførsel som igjen vil ha potensiale for å påvirke tilbudssiden. **Stablecoin ratio**: en variabel som tar totalt antall BTC og deler det på antall tilgjengelig stabelcoin. Et lavt forholdstall vil indikere lite kjøpekraft, et høyt forholdstall vil indikere potensiell kjøpekraft.

Markeds indikator

Estimert girings rate (på engelsk: Estimated leverage ratio) er en variabel innenfor markeds indikatorer (på engelsk: Market Indicator, forkortelse: MI). Den tar de totale leverage verdiene delt på verdier holdt samlet på børsene. Dette gir et bilde av hvor mye midler som er giret i markedet. Dersom markedet har en signifikant bevegelse, vil høy verdi for girede midler kunne medføre enda kraftigere markedsbevegelser på grunn av potensialet for likvideringer av girede posisjoner. Denne effekten beskrives av Rathgeber et al. (2020), de fant en sammenheng mellom giring i markedet og markante markedsbevegelser [22].

Nettverks data

Nettverks data (på engelsk: Network Data, forkortelse: ND), gir data på en rekke ulike datapunkter innenfor aktiviteten på blokkjeden til BTC. De viktigste variablene i denne subkategorien er: antall transaksjoner, antall aktive adresser, fortjenesten til minere, mengde datakraft i nettverket, antall nye BTC som utstedes hver blokk. Disse variablene er med på å danne et bilde av hvor stor aktivitet nettverket har.

Nettverks indikatorer

Nettverks indikatorer (på engelsk: Network Indicator , forkortelse: NI), er et sett med indikatorer som er bygget på aktiviteten til BTC nettverket. De er utviklet for å forsøke å finne BTC´s faktiske verdi. Den mest kjente av disse er *Stock To Flow Ratio*, som er forholdet mellom nåværende eksisterende BTC delt på antall nylige utvinnende BTC. Denne modellen er ment å kvantifisere knapphetsfaktoren til BTC, modellen ble introdusert av pseudonymet PlanB (2020). De ulike nettverksindikatorene er omdiskutert, og det er usikkerhet knyttet til om de gir et godt bilde på BTC faktiske verdi.

[22] Denne undersøkelsen ble gjort på det tyske aksjemarkedet i perioden 1999 til 2014. Dette er en effekt som kan anses å ha lik effekt på tvers av ulike markeder

Miner bevegelser

Miner bevegelser (på engelsk: Miner Flows , forkortelse: MF) er en subkategori der det tallfestes hvordan store adresser tilknyttet signifikante miner operasjoner handler. Variablene i denne kategorien overvåker bevegelsene fra adressene til ulike børser. Data som hentes ut fra dette er hvor stort antall BTC beveger seg, totale verdien og hvor mange adresser er involvert.

Markeds data

Markeds data (på engelsk: Market Data, forkortelse: MD)er sammenfattes av data som er hentet ut fra aktivitet på børsene.
Taker Buy Sell Volume/ Ratio: Enhver transaksjon krever både en kjøper og en selger. Ved å se på om ordretakeren, er en kjøper eller selger, (om en transaksjon skjer på etterspurt pris eller på tilbudtpris) kan volumet skilles mellom kjøpers volum og selgers volum. "Taker" referer til ordre som er utført før de når ordreboken. Dette er kjøp av typen markedsordre. Maker ordre, er av typen som går inn i ordreboken før de blir utført. Dette er typisk limit ordre. Denne typen ordre gir et volum til ordreboken.
Maker en som er villig til å kjøpe et produkt til en lavere pris en budprisen[23]. En maker på salgssiden ønsker å plassere ordrene på en høyere pris en budprisen. Dersom en taker kjøper på budprisen satt av makerordre, vil antallet verdier handlet mellom de to legges til under taker sell volum. Dette er volum som har potensiale til å drive prisen nedover. Når en taker selger på etterspurt pris, vil verdiene som handles legges til takers kjøpsvolum. Dette er volumet som har potensialet til å presse prisen oppover.

Dersom markedet har mer taker kjøpsvolum enn takers salgsvolum, vil det tendere til å øke prisen. Dersom takers salgsvolum er høyere en kjøpsvolumet vil dette ha motsatt effekt, og prisen vil potensielt falle.

Open interest: er antallet girede posisjoner. Dette er det totale antallet shorts og longs som er åpne på derivatmarkedet for BTC/USD. Dette kan ha ulike effekter, dette ble beskrevet under estimated leverage ratio.

[23]Budpris er prisen på den siste gjennomført transaksjonen, og representerer den faktiske verdien

Funding rate: er en periodisk utbetaling som skjer mellom de girede short- og longposisjonene. Funksjonen til denne mekanismen er slik: siden av markedet som har høyest verdi (summen av verdien på antallet posisjoner) betaler den med lavest verdi. Det er en flytende prosent som varierer med forholdet mellom de to sidene. Hensikten med denne mekanismen er å skape insentiver for å opprette posisjoner på motsatt side av den med mest verdi. Dette er for å skape en balanse i markedet. Dersom fundratene er positive, gir dette en indikasjon på et positivt markedssentiment, majoriteten tror på prisoppgang. Da betaler de som har long posisjoner de som holder short posisjoner. Dersom den er negativ, vil markedet ha et negativt sentiment og short betaler long posisjoner.

2.3 Teknisk analyse

Kara et al. (2011) brukte tekniske indikatorer som en del av treningssett for en LSTM modell, de viste med dette en bedre prediksjons-nøyaktighet. I denne oppgaven har jeg valgt å anvende et stort antall tekniske indikatorer[24], som ble vist å ha en effekt innenfor prisprediкering med bruk av kunstig intelligens av Kara et al. (2011). Dette settet omfatte populære momentumindikatorer, ulike moving avrages og relativ styrkeindeks med stokastisk versjon. Indikatorene programmeres slik at de baserer seg på rådata hentet inn i oppgaven. Utregningene for disse indikatorene gjøres i preprosesserings arbeidet.
Ved å anvende LSTM på tekniske indikatorer ønsker man å trene modellen til å prosessere og tilegne seg informasjon om hvordan tekniske indikatorer påvirker pris. Ved å trene modellen på dette vil modellen kunne automatisere og etterligne hvordan tradere som baserer avgjørelsene sin på tekniske indikatorer vil handle (Alonso-Monsalvea et al., 2019). Tradere som handler ut i fra tekniske indikatorer vil bruke et sett med trendlinjer, prismønster i kombinasjon med tekniske indikatorer, for å gi kjøp- og salgssignaler.

Teknisk analyse har en lang historie innenfor investeringsdomenet, men akademikere har lenge vært skeptiske til feltet. Det er blitt vist at tidligere prisutvikling kan ha en predikasjonsverdi (Dempster et al., 2001). Kara et al. (2011) trente neural nettverk på ulike tekniske indikatorer for å predikere bevegelsene på Istanbul Stock Exchange. Indikatorene som ble brukt var momentum, relativ Strength Index, moving avrages, sotchastics k/d%, moving avrage convergence-divergence (MACD). Jeg har valgt å også inkludere on-blance volume og bollinger band. Dette med bakgrunn i deres popularitet.

Som vist av Patel et al. (2015), er det effektivt å anvende tekniske indikatorer som diskrevet variabler for å representere pristrender, ved anvendelse av tekniske indikatorer innenfor deep learning. Hver enkelt teknisk indikator har generelt en verdi som indikerer om prisen er inne i en stigende eller fallende trend ved å omforme de slik at de gir verdier som enten er +1, 0, -1. +1 indikerer at prisen går opp, 0 er nøytral og -1 indikerer at prisen går ned.

[24]Tabell 11 på side 102 viser alle de ulike typene som er benyttet

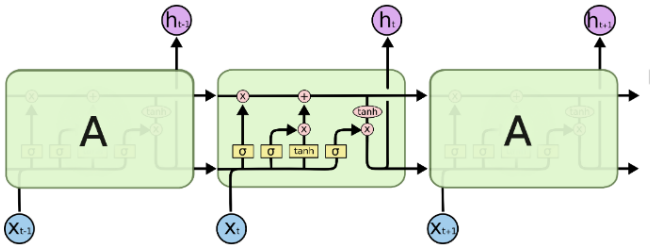
2.4 Long Short Term Memory

Tidsseriedata er en sekvens av vektorer som er avhenge av tid. Definert som: (1)

$$\vec{x}(t), t = 0, 1, 2... \tag{1}$$

Tid er generelt sett på som en diskret variabel. Dette gjør at man får en verdi av $\vec{x}$ ved hvert fastsatte tidsintervall (Dorffner, 1996). I denne oppgaven vil dette være tilnærmingen til tidsseriedata.

Prediksjonen av tidsseriedata er både kompleks og utfordrende. Innenfor tidsserieproblemer får man i motsetning til linearregresjon og klassifikasjon, element med tidsavhengighet mellom ulike observasjoner. Utfordringen er knyttet opp mot behovet for spesialisert behandling av data som kreves når man trener og evaluerer tidsseriemodeller. Maskinlæringsmetoder har vist seg å effektivt kunne håndtere komplekse tidsserieproblemer, med datasett som inneholder et stort antall variabler med komplekse ikke-lineare forhold (Brownlee, 2018).



Figur 6: Illustrajon av en LTSM celle, hentet fra link, den 25.02.2021

En sentral utfordring og begrensning for standard RNN arkitektur, er at de ikke har evnen til "langtidshukommelse". Det vil si å kunne ta informasjon tilbake i tid og anvende det på en nåværende utfording (Olah, 2015). Dette skjer fordi gradienten er multiplisert bakover gjennom nettverket. Dette kan medføre to problemer: gradienten får en ekstremt høy verdi, eller en ekstremt lav verdi. Dersom verdien på gradienten blir svært høy, vil vektparameteren bli oscillerende og ikke i stand til å lære. Dersom verdiene blir for små, vil dette gjøre at nettverket overser verdier og ikke lærer av de. I 1997 introduserte (Hochreiter and schmidhu-

ber, 1997) en løsning på dette problemet: Long Short Term Memory. LTSM kan ses på i to deler: Long-term memory fordi modellen er i stand til å tilegne seg kunnskap fra parametere som endres sakte. Short-term som er evnen til å lære av celler som endrer seg for hvert tidssteg. Hochreiter and schmidhuber (1997). Den matematiske formuleringen til LSTM, basert på Hochreiter and schmidhuber (1997) blir angitt fra ligning 2 til 8 og illustrert i figur 6: Inputverdien som gis inn i nettverket.

$$X = \begin{bmatrix} x_t \\ h_{t-1} \end{bmatrix} \tag{2}$$

Glemmeporten, aktiverignsfunksjonen sigmoid legges til ved elementmetoden.

$$f_t = \sigma \left(W_f \times U_f h_{t-1} + b f_f \right) \tag{3}$$

inputporten, aktiverignsfunksjonen sigmoid legges til ved elementmetoden.

$$i_t = \sigma \left(W_i x_t \times U_i h_{t-1} + b_i \right) \tag{4}$$

outputporten, aktiverignsfunksjonen sigmoid legges til ved elementmetoden.

$$o_t = \sigma \left(W_o x_t \times U_o h_{t-1} + b_o \right) \tag{5}$$

Cellesatus, aktiverignsfunksjonen sigmoid legges til ved elementmetoden.

$$c_t = f_t \odot c_{t-1} + i_t \odot \widetilde{c}_t \tag{6}$$

ny cellestatus, aktiverignsfunksjonen tanh legges til ved elementmetoden.

$$\widetilde{c}_t = tanh \left(W_c x_t \times U_c h_{t-1} + b_c \right) \tag{7}$$

Skjulte tilstand

$$h_t = \sigma_t \odot tanh(c_t) \tag{8}$$

2.5 Menneskelig beslutningstager med støtte fra kunstig intelligens

Innenfor litteratur som omhandler kunstig intelligens blir maskiner enkelte ganger avbildet som direkte konkurrent til mennesket. I en rekke fagfelt vil ikke dette være hensiktsmessig eller gjennomførbart (Grgić-Hlača et al., 2019). I delkapittel rundt avgrensning trekkes det frem flere makroøkonomiske faktorer som vil være vanskelige å kvantifisere og implementere i en maskinlærings modell. Ved å ha en menneskelig trader som har ansvaret for avgjørelser, vil han kunne ha kjennskap til disse faktorene samtidig som han kan anvende en maskin som kommer med informasjon hentet fra utvalgte variabler.

Ved å ha en slik tilnærming har maskinlæring potensialet til å omgjøre store mengder informasjon til enkel og tolkbar informasjon til et menneske. Dette kan være i form av sannsynligheter for utfall eller ved binære prediksjoner som opp / ned.

Dersom et menneske skal bruke en maskinlæringsmodell som en del av en beslutningsprosess, er det viktig for personen å forstå **hvorfor** modellen kommer med prediksjonene. Dette er grunnlaget for å skape tillit til prediksjoner (Ribeiro et al., 2016). Komplekse modeller som er bygget på store datasett med intrikate variabler fører til at selv eksperter på feltet kan slite med å forstå hvorfor modellen kommer frem til ulike prediksjoner. Dette skaper problemstilling mellom nøyaktighet og tolkbarhet (Lundberg and Lee, 2017a).

Ribeiro et al. (2016) definerer tillit i to deler: "(1) *tillit til en prediksjon, om en bruker har tillit til en individuell prediksjon i den grad at man tar en avgjørelse basert på den*", og (2) *"om brukeren stoler på at modellen handler rasjonelt hvis den blir distribuert"* (Ribeiro et al., 2016) . Begge to er direkte knyttet opp mot hvor mye personen som bruker verktøyet forstår modellens adferd.

Det er viktig for prosjekter å ha en kombinasjon av tillit og forståelse for en maskinlæringsmodell hvis den skal brukes. Forståelse av modellen, gjennom gjennomsiktighet, gjør det mulig for et mennesket å vurdere prediksjonene modellen kommer med. Dette vil kunne skape tillit, som kan være med på å effektivisere hvordan modellen brukes (Hall and Gill, 2019).

Ved å forstå modellen, kan man også avdekke styrker og svakheter til modellen. Ved å gjennomføre grundige variable analyse kan man danne et grunnlag til å forbedre både modellen og variablene som den bruker.

De to primære verktøyene som blir andvendt innenfor analyse av maskinlæringsmodeller er Local Surragate (LIME) og Shapley Additive exPlanations (SHAP). For å forstå hvordan de ulike variablene i oppgaven, benyttes SHAP som analyse metode.Dette gjøres ved å bruke Kernel SHAP kan man hente ut svært detaljerte analyser av hver enkelt variabel og hvordan samspillet er mellom ulike variabler.

2.6 Shapley verdien

Shapley verdien (SV) ble utviklet av Lloyd Shapley i 1951, for å vurdere og rangere bidraget til spillere i et samarbeidsspill. Lipovetsky and Conklin (2001) viste i sitt arbeid hvordan SV kan anvendes innenfor regresjon som metode for å forklare viktigheten av de ulike inngangsverdiene. SHAP beregner hvor stor påvirkning hver enkelt variabel har på prediksjonen. I SHAP er en inngangsverdi $x = [x_1, ...x_p]$ for en trent modell f. SHAP forenkler modellen, g, slik at den kan forklare bidraget til hver enkelt verdi. Antallet variabler som modellen analyserer er beskrevet med p. Formelen er beskrevet under.

$$g(z) = \phi_0 + \sum_{n=1}^{p} \phi_n z_n \tag{9}$$

En forenkling av inngangsverdien x, hvor den korresponderende verdien til z er klassifisering 1 og verdien z korresponderer til variabler som ikke blir brukt er 0. Variabelen er brukt for å predikere $z = [z_1, , , Z_p]^T$

$$\phi_n = \sum_{z \subseteq p}^{p} \frac{|x|!(p - |x| - 1)!}{p!} [f(z) - f(z \setminus n)] \tag{10}$$

Lundberg and Lee (2017b) overførte denne metoden til et fungerende program som for analyse av modeller. Målet med deres arbeidet var å utviklet et verktøy som gjør det mulig for mennesker å enkelt analysere hva som er viktig for en maskinlæringsmodell og forstå hvorfor den handler som den gjør. De gjorde dette ved å utnytte SV ved å omgjøre prosessene i maskinlæring slik at spillet er oppdraget til maskinlæringsmodellen. Spillerne representerer variablene og utbyttet vil være prediksjonen ved en gitt observasjon trukket fra den gjennomsnittlige verdien til hele datasettet.
Lundberg and Lee (2017b) demonstrerer i sitt arbeid hvordan ulike inngangsverdier påvirker modellens prediksjoner og at dette kan visualiseres på en god måte som gjør det enkelt å hen-

te ut informasjon om variablene. SHAP verktøyet har muligheten til å forklare både lokale og globale prediksjoner. I denne oppgaven velger jeg å fokusere på det globale bidraget. En svakhet ved denne metoden er hvor mye datakraft som kreves for å regne ut SHAP verdien for å effektivisere ved at man ikke regner verdien for hele datasettet, men heller et utdrag.

2.7 ROC-kurve og forvirringsmatrise

For å bidra til å gi et mer utfyllende svar på modellens nøyaktighet brukes ROC kurve og forvirringsmatrise. ROC-kurve (Reciever Operating Characteristics) gir en oversikt over rett positiv og rett negativ, anvist som en kurve. Dette gjøres for alle modellens klassifiserings terskler. Ved å regne ut areal under kurven, AUROC, ved å bruke AUROC får man ut modellens absolutte klassifisering, dersom modellen har klassifiseringsterskler (I denne oppgaven er det ikke brukt terskel). ROC kurven gir også en anvisning på hvordan potensiell prediksjons nøyaktighet hadde vært dersom en terskel ble implementert.
For å analysere om modellen fungerer bedre til å predikere opp eller nedgang best, anvendes forvirringsmatrise. Dette er et verktøy som presenterer modellens: rett positiv, falsk positiv, rett negativ og falsk negativ på matriseformat.

$$nyaktighet - short = \frac{2 \cdot RN}{RP + RN + FP + FN} \tag{11}$$

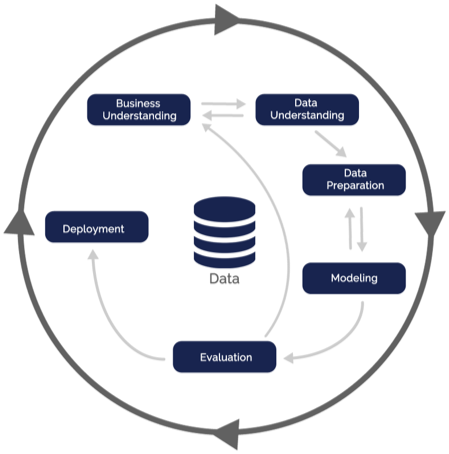
$$nyaktighet - long = \frac{2 \cdot RP}{RP + RN + FP + FN} \tag{12}$$

Problemet med å kun bruke nøyaktighet som eneste målevariabel har vært grundig utforsket. Det viser seg at de som konstruerer modellen har en tendens til å overestimere modellens nøyaktighet, og ikke er i stand til å oppdage mulige feilkilder (Ribeiro et al., 2016). Ved å kombinere SHAP analyse sammen med ROC kurve og forvirringsmatrise er hensikten å gi en dyp analyse av modellens faktiske prediksjonevner.

3 Metode

Innenfor maskinlæringsfeltet er det en rekke ulike oppsett strukturering av arbeide. For å besvare problemstillingen ønsket jeg å strukturere datainnhentingsprosessen, samt se hvordan dette sammenfaller med modelleringsarbeidet. Derfor valgte jeg å bruke Cross-industry standard process for data mining (CRISP-DM). Dette er en prosess som er anvendt for å kvalitetssikre arbeidet med å transformere data til maskinlæringsmodeller Pedro et al. (2020). Prosessen er iterativ frem til man oppnår resultater man anser som gode nok, og den er interaktiv, som betyr at man kan gå frem og tilbake mellom de ulike stegene sett opp mot kvaliteten på løsningen. Modellen er en prosess som bryter ned arbeidet i 6 steg, disse er illustrert i figur 7. Jeg har valgt å trekke inn et element fra Brownlee (2020b) sin metodikk rundt datapreparasjon. Jeg har valgt å endre Business understanding til definering av problemet, og se vekk fra det siste steget, deployment.

.



Figur 7: Illustrasjon av CRISP-DM prosesen, hentet fra smartvision-me

Nedenfor har jeg beskrevet kort de viktigste delene av CRISP-DM metoden, dette er basert på arbeidet av Hapman et al. (2020)

Business understanding (oppdragsforståelse)
Den initielle fasen er å forstå prosjektets mål og krav. I denne oppgaven vil dette være å besvare problemstillingen, deretter å omgjøre denne kunnskapen til hvilke data som trengs for å løse oppgaven og hvilken design som må til for å nå målene.

Data understanding (dataforståelse)
Denne fasen omhandler å starten med å samle inn data og gjøre seg kjent med den, for å indentifisere kvaliteten, mulige problemer og for å få innsikt. På denne måten kan det komme frem om det er interessante hypoteser eller skjult informasjon i datasettet.

Data preparation (Preprosessesering)
Datapreparereingsfasen dekker alle aktiviteter for å konstruere det endelige datasettet som skal brukes til modelleringsverktøy. Dette er en prosess som er sannsynlig at vil bli utført en rekke ganger og ikke i en bestemt rekkefølge. I denne delen blir variabel seleksjon, transformering og rensing av data utført.

Modelling (modellering)
I denne fasen blir ulike modellerignstekniker utført, og parameterene blir tunet for optimale resultater. Det er en rekke ulike teknikker for å gjøre dette. Noen av disse har spesifikke krav til typer av data, derfor vil det ofte være nødvendig å gå tilbake i prosessen og gjenta tidligere steg.

Evaluering
På dette steget av prosjektet har man oppnådd en god kvalitet fra et dataanalytisk ståsted. Før man går videre er det viktig med en grundig evaluering av de stegene som er utført for å komme til dette punktet.

Oversikt over prosessen

3.1 Oppdragsforståelse

Oppdragsforståelsen for prosjektet er forankret i problemstillingen.
Finnes det variabler som en LSTM-modell kan andvende for å predikere kortsiktig kursutvikling for Bitcoin?
For å kunne analysere variablene må det konstrueres en modell som predikerer kursutvikling til det valgte aktivum. Jeg har valgt å definere problemet modellen skal løse som et binært klassifikasjonsproblem.

Modellen skal: predikere om prisen til BTC er høyere eller lavere på et gitt tids steg frem i tid.
Målet med prosessen er å sørge for et komplett datasett som maskinlæringsmodellen kan trenes på for å predikere fremtidig prisutvikling. Matematisk fremstilling av problemet modellen skal løse: X_{ti} er læringsmatrisen ved gitt tidsintervall betegnet som ti, f er antallet ulike feature på inputverdiene. Y_{ti} anviser outputverdien 1,0, $y_{ti} = 1$ representer at prisen vil være høyere og $y_{ti} = -1$ at prisen vil være lavere enn på nåværende tidspunkt. Nåværende tidspunkt er gitt som n, fremtidig tidspunkt er gitt som fr.

$$X_{ti} = \left[x^f\right]_{ti} \tag{13}$$

$$Y_{ti} = \{y\}_{ti} \tag{14}$$

Målet til modellen er å minimere prediksjonsfeilen til L_{ti}. Dette gjøres ved å måle distansen mellom modellens predikerte verdi opp mot den reelle verdien. For binære klassifikasjonsproblemer er den mest anvendte taps-funksjonen kategorisk kryss entropi Brownlee (2017). Denne er definert i ligning 16,

$$L = -\sum\nolimits_{ti=1} Y_{ti} log(X_{ti}) \tag{15}$$

Målvariabelen modellen skal predikere er prisen ved slutten av hvert 1 minutts intervall, price usd close. Dette er prisen til BTC målt i dollar ved slutten av tidsintervallet. Tidsintervallet modellen skal predikere settes til 8 minutter. Valget tas på bakgrunn av testing. Ved for kort tidsintervall falt resultatene, samme gjaldt ved for lange tidsintervall.
For å unngå at den generelle retningen på markedet skal påvirke modellen, så legges det inn en begrensning slik at modellen tar med et likt antall short og long prediksjoner i utregningen av modellens nøyaktighet. Dette er en avgjørelse som vil påvirke modellens nøyaktighet negativt, selv om den optimalt f.eks ville hatt en skjevhet mot det generelle markedets retning.

Suksesskriteriene for prosjektet settes til: utarbeide en fungerende maskinlæringsmodell med LSTM arkitektur som kan trenes på høyoppløselige datasett.

3.2 Data forståelse

For å sørge for god dataforståelse blir følgende prosesser utført: valg av datakilde, datainnhenting, beskrivele og gjennomgang av av data og verifisering av datakvalitet (Hapman et al., 2020).

3.2.1 Valg av data

Kryptovaluta kan handles på 440 ulike børser [25]. Av disse, er Binance en av de ledende børsene. Binance har en aktiv brukerbase på 13.5 millioner brukere (Statista, 2021a), et handelsvolum pr. 24 timer på henholdsvis 24 og 44 milliarder dollar for spot [26] og derivatmarkedet [27]. I denne oppgaven vil Binance bli benyttet som hovedbørsen til å hente ut prisdata. Dette gjøres ved å anvende Binance sin API. Den tillater 1200 forespørsler pr. minutt som er en forespørsel pr 50 millisekund (Binance, 2021), tillater 10 ordre pr. sekund. Dette gjør det mulig å hente ut data med høy oppløsning samt at den er egnet for høyfrekvent trading.
Den andre kilden til data er CryptoQuant. Som beskrevet tidligere er dette en profesjonell datatilbyder som distribuerer markedsinformasjon data til kryptovaluta. Det finnes flere ulike firma som tilbyr denne tjenesten. En av de primære årsakene til at CryptoQuant anvendes, er at de tilbyr en av de laveste oppdateringsfrekvensene. Dette betyr at ny informasjon knyttet til transaksjoner skjer nå det produseres en ny blokk. Dette skjer med en varierende tidsintervall som er knyttet opp mot nettverksaktivitet, det ligger på ca 14 min. Dette ligger i at ny informasjon knyttet til transaksjoner skjer når det produserer en ny blokk, dette skjer med et varierende tidsintervall, men dette ligger på ca. 14 minutter.

[25] Data hentet fra Coingecko link den 28.03.2021

[26] Spot trading er når et aktivum handles som skal leveres umiddelbart. Dette innebærer at kjøperen eier aktivumet fysisk

[27] Derivater er et finansielt tradinginstrument som muliggjør spekulasjon på kurspriser uten at man fysisk eier aktivumet. Dette muliggjør flere mer komplekse instrumenter og gjør at man kan spekulere på både oppside og nedside

3.2.2 Datainnhenting

Datainnhentingen til oppgaven gjennomføres ved å programmere automatiserte sanntids datainnhenting og logging fra Binance og CryptoQuant. Fra Binance logges 1-minutts trading data, variablene er: open, high, low, close, volume og antall handler. Dette må gjøres kontinuerlig. For å oppnå dette kjøres programmet på en en skydatamaskinløsning. Denne fungerer som en pc som kjører kontinuerlig uten nedetid. Fra cryptoQuant logges det ca. 100 unike variabler.

3.2.3 Data beskrivelse

Datasettet deles opp i treningssett og testsett sett. Treningssettet består av 90 % av hele datasettet, og testsettet settet består av 10 %. Testsettet jeg har valgt å sette av 10% av datasettet til tests data. Det blir ikke kjørt tester på et separat testsett.
Fordelingen kan virke lav, men dette er fordi det innen finansmarkedet vil forekomme nedbrytning. Med dette mener jeg at dersom en modell trenes, er den kun valid en kort periode fordi mønstrene den potensielt oppdager ikke vil vare over lange tidsperioder. I et ungt marked, som BTC er, vil markedet endre seg hurtig. Dette gjør også at maskinlæringsmodeller ikke vil generalisere godt over lengre tidsperioder.

Treningssettet er fra 2020-12-10 00:00:00 til 21-04-13 05:15:00
Testsettet er fra 21-04-13 05:16:00 til 2021-04-27 00:00:00
Totalt antall tidsintervall: 197 907
Antall tidsintervall i treningsdata: 178 114
Antall tidsintervall i testdata: 19 793 Kapabiliteter

For å undersøke om det er ubalanse i datasettet, signifikant flere datapunkter for oppgang eller nedgang gjennomføres en enkel test på datasettet: BTC prisutvikling i testperioden: 18 537.8$ ved start, 53 941.5 $ ved slutt (en økning på ca 190 %).

Statistikk for balansering i antall perioder som er høyere enn gitt antall tidssteg frem i tid deles inn i testdata og testdata:

Treningssett:

$\text{price usd close}_7 > \text{price usd close}_0 = 88654$

$\text{price usd close}_7 < \text{price usd close}_0 = 89477$

Test sett:

$\text{price usd close}_7 > \text{price usd close}_0 = 10060$

$\text{price usd close}_7 < \text{price usd close}_0 = 9732$

Dette gir følgende: For testsettet er det 823 flere tidsintervall med prisnedgang en oppgang, dette tilsvarer 0.9 % For testsettet er det 328 flere tidsintervall med prisoppgang en nedgang, dette tilsvarer 3.3 % testsettet er litt bias mot prisoppgang. Dette må tas med i betraktning dersom modellen har en betydelig bias mot prediksjoner med prisoppgang. En viktig ting å bemerke, det ble ikke utført "out of sample testing" i denne oppgaven. Dette er å teste modellen på et nytt datasett modellen ikke har hatt tilgang til før. Dette er en viktig del av testingen før modellen kan bli andvendt i et reelt miljø, og burde vært utført.

3.2.4 Formatering

Jeg har valgt å omgjøre hele datasettet til 1 minutt. Variabler som har tregere oppdateringsfrekvens enn 1 minutt, holdes konstante til det forekommer en endring. Dette gjøres ved å anvende en forwardfill metode. For å gjøre dette aggregeres datapunktene. Dersom det ikke er ny data ved neste tidsintervall tas forrige verdi og brukes på nytt.

Tabell 1: Oppdateringsfrekvens for markedsinformasjons variabler

Kategori	Exchange flyt	Flyt indikator	Netverks indikatorer	Miner flyt	Markeds indikator (MI)	Markeds data (MD)	Nettverks data - (ND)
Oppdateringsfrekvens	Blokk	Blokk / 24 timer	24 timer	Blokk	24 timer	pr. minutt	blokk

3.2.5 Kvalitetskontroll

For å kontrollere datasettet underveis formateres det som CSV filer. Dette er et effektivt filformat, det er enkelt å lese datasettet for programmet og det er enkelt å gjennomføre manuelle kontroller ved å bruke Excel. Underveis i arbeidet med datasettet lagres det kontrollfiler der det manuelt blir kontrollert at variabler er korrekt plassert med timestamps [28]. Programmet legger inn korrekt målverdier og andre mulige feilkilder som kan oppstå. For å sikre datasettet validitet gjennomføres det stikkprøver der data som er samlet gjennom API sammenlignes opp mot verdier på CryptoQuantCryptoQuant sine sider, samt opp mot lignende datatilbydere som GlassNode, for å kryssjekke verdiene.

[28] Timestamp er et tidstempel som angir dato og tid for data som representeres

3.3 Preprossesering

3.3.1 Konstruksjon og integrering av data

På grunn av hvordan en datamaskin fungerer, er det signifikant mer oppløsning i spennet mellom 0 til 1 i forhold til tall som er innenfor et større spekter. Derfor er det viktig å normalisere datasettet (Brownlee, 2020b). Datasettet fra Binance og CryptoQuant settes sammen, og det utføres teknisk analyse som også legges til datasettet.Variablene fra Binance inneholder enkelte hull på closing pris som konsekvens av nedetid børsen har hatt i tidsperioden. Problemet løses ved å hente closingprisen fra CryptoQuant sin aggregerte prisdata [29]. Dette vil gi et lite avvik, men dette er ikke signifikant.

Datasettet er skalert ved å anvende Keras sin innebygde funksjon minmaxscaler. Den omgjøre samtlige variabler til en skalering fra 1 til -1. Dette gjøres for å forbedre prosesseringstid. Flere av variablene har svært høye tall som gjør det helt nødvendig å gjennomføre en nedskalering. Dette kan medføre en risiko for at modellen ikke ser samtlige mønstre, men i et omfattende datasett med svært stor variasjon i størrelsen på datasettet anses det som et nødvendig kompromiss. Skaleringen gjøres etter at datasettet er splittet mellom trening og testsett for å unngå datalekkasje, som kan forekommer dersom dette gjøres i en enkelt operasjon (Brownlee, 2020a).

Før data kan anvendes til trening av modellen går den gjennom en pre-prosessering. Her blir datasettet fra Binance og CryptoQuant satt sammen, og det utføres teknisk analyse som også legges til datasettet. Variablene fra Binance inneholder enkelte hull på closing pris som konsekvens av nedetid børsen har hatt i tidsperioden. Problemet løses ved å hente closingprisen fra CryptoQuant sin aggregerte prisdata . Dette vil gi et lite avvik, men dette er ikke signifikant.

[29] Denne prisdata tar gjennomsnittet, vektet etter volum fra de viktigste kryptobørsene

3.3.2 Konstruksjon av datasett - Utvelgelse av variabler

Variabel utvelgelse er prosessen der antallet variabler reduseres i den hensikt å redusere modellens behov for datakraft, samt i enkelte tilfeller å forbedre modellens prediksjoner (Brownlee, 2020b). Dette er en viktig og utfordrende prosess, uten en generell optimal løsning. Prøving og feiling er en del av prosessen.

Initialt forsøk: Den initiale planen for testing av markedsinformasjon variablene var å gruppere de i sett etter kategorien de tilhører. Ved å kjøre analyse på variablene gruppert på denne måten fikk jeg resultat som vist i tabell 2. Nøyaktigheten var lav, med kun markeds data som ga resultater som var verdt å ta med videre.Av resultatene ble det tydelig at dette måtte løses på en annen måte.

Neste forsøk var å sette sammen variabler i tilfeldig konstruerte sett på 15 stk i hvert sett, dette ga 10 ulike datasett. For deretter å kjøre analyse på de oppdelte settene. Deretter gikk jeg gjennom og gjennomførte en SHAP analyse av samtlige sett. Problemet som oppstod med denne metoden var at nøyaktigheten var relativt lav på samtlige sett, nøyaktighet mellom 50-51 %, og variablenes viktighet for modellen varierte mye ved nye simulering på samme sett [30].

Tabell 2: Resultater av markedsinformasjon data grupper etter kategori

Exchange flyt	Nettverks indikatorer	Markedsflyt	Markeds data	Flyt indikator
50.5 %	50.9 %	50.2 %	51.9 %	50.5 %

Som resultat av de forsøkene nevnt over, valgte jeg å se på metoder for å automatisere prosessen. Boruta er en algoritme som anvendes for variabel utvelgelse. Algoritmen ble introdusert av Kursa and Rudnicki (2010). Dette er en metode som automatisk velger ut de viktigste inngangsverdiene til en modell ved å forsøke å finne alle inngangsverdier som bidrar til modellens prediksjoner.

Det er svakheter med å utføre variabel utvelgelse med denne metoden. Dette er knyttet opp til at Boruta algoritmen ikke er konstruert for å se høydimensjonale sammenhenger, slik en LSTM har potensialet til. På denne måten kan man miste variabler som har en nytte for en LSTM modell. Selv om metoden har svakheter, har den vist seg å ha potensialet i kombinasjon med LSTM (Ahmed et al., 2021), der Boruta ble anvendt for å velge ut variabler til en LSTM modell. Jeg velger derfor å bruke dette som en av metodene til å konstruere datasettet med. For å forminske de potensielle svakhetene til Boruta utvelgelsen, gjennomføres også

[30]Det virker som sett med lav nøyaktighet så ser modellen ulike svake mønstre, men disse resultatene er av liten verdi for mønstrene er svak

Modell navn	Modell I	Modell II	Modell III	Modell IV	Modell V	Modell VI	Modell VII
Datasett	Samtlige markedsinformasjon	Samtlige markedsinformasjon og TI	Boruta:markedsinformasjon	Boruta:markedsinformasjon og TI	Samtlige TI	Shap filtrerte TI	Kvalitativ utvelgelse
Fra Binance	0	5	5	5	5	2	0
Markedsinformasjon	94	94	42	24	0	0	14
Tekniske indikatorer	0	47	0	26	47	10	5
Totalt antall variabler	94	146	47	55	52	12	19
Hyperparameterisering	Generell	Generell	Gegnerrel	Generell	Generell	Genrell	Spesifikk

Tabell 3: Oversikt over de ulike modellene og hvilke variabler de er trent på.

analysen på datasett bestående av samtlige variabler. For de tekniske indikatorene ble det utført en utvelgelse der de variablene som hadde størst bidrag av SHAP verdier ble tatt med videre til å danne modell VI.

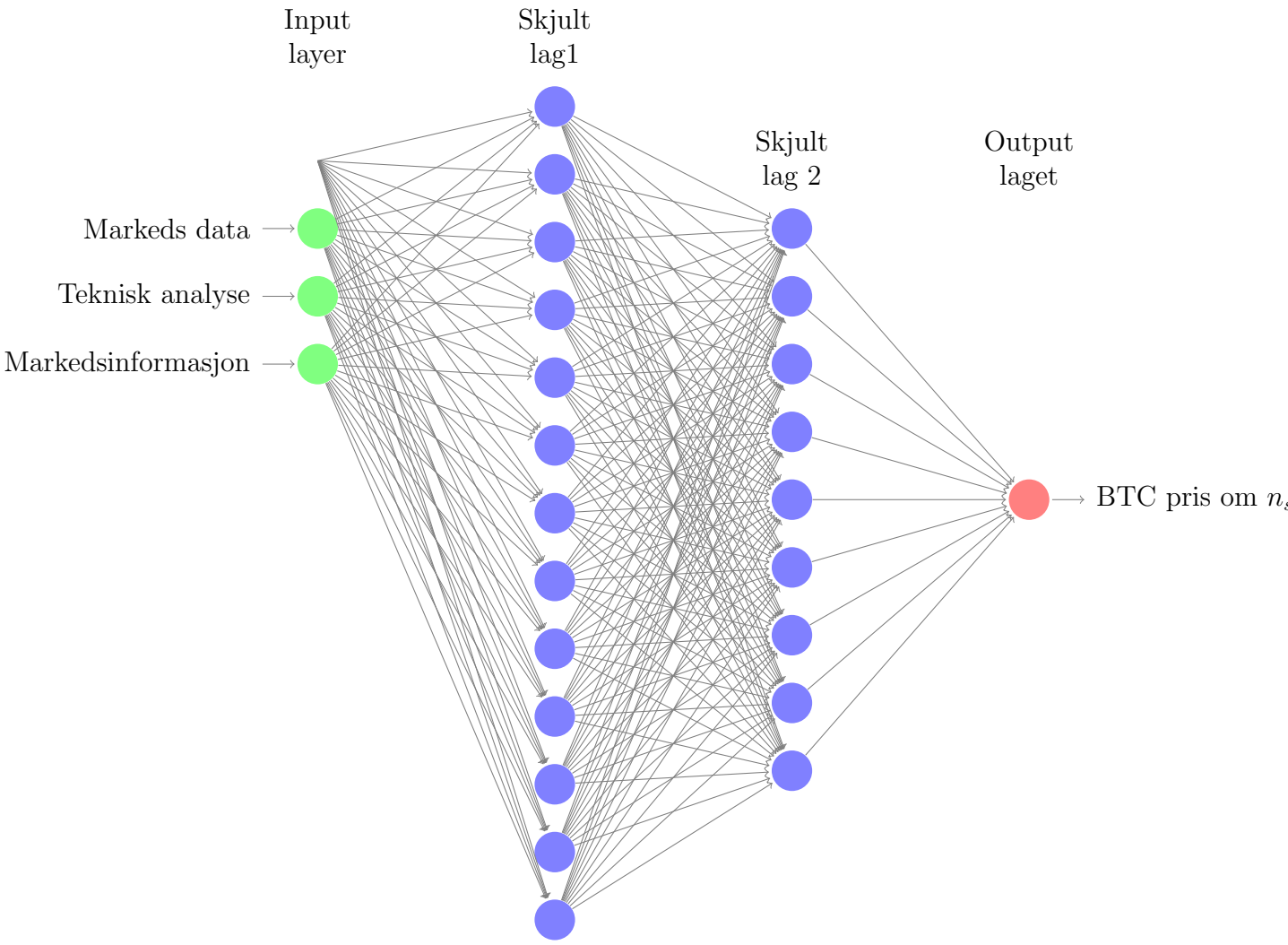
Modell VII blir også referert til som den endelige modellen i oppgaven.

Tabell 4: Variabler brukt i modell VII

Kvalitativ utvalgte variabler	
price usd close	
CCI	
BOP	
OBV	
MFI sell pressure	
MFI buy pressure	
Taker sell volume	
Taker buy volume	
Taker sell ratio	
Taker buy ratio	
Taker sell ratio binance	
Taker buy ratio binance	
Taker sell ratio Binance	
Transaction count inflow	
Stablecoins ratio	
Taker buy volume Binance	
Coinbase premium	
Funding rates binance	
Stablecoin supply ratio	
markedsinformasjon	14
Teknisk analyse	5
Sum	19

3.4 Modellering

I dette delkapitlet blir det gjennomgått hvilke type modell som vil bli valgt, hvordan den blir sammensatt og prameter innstillinger.



Figur 8: Egenprodusert illustrasjon av et nevral nettverk, produsert i LaTex med tikzpicture modul

LSTM modellen konstures ved å anvende Keras sitt rammeverk i python. Keras er et open-source bibliotek som fungerer som interface for TensorFlow. Den har de nødvendige komponentene for å bygge en rekke ulike typer nevrale nettverk. Modellen er sekvensiell , og dette valget tas på grunn av dens enkelhet i kombinasjon med at den vil kunne løse problemstillingen (Chollet, 2017). Modellen gir prediksjonene 1 og 0.

1 = modellen predikerer prisoppgang
0 = modellen predikerer prisnedgang

3.4.1 Hyperparametere

Den eneste forskjellen på modell I til modell VI er at de er trent på ulike datasett, ellers har de identisk struktur. Modell VII gjennomgåes under hyperparametre.

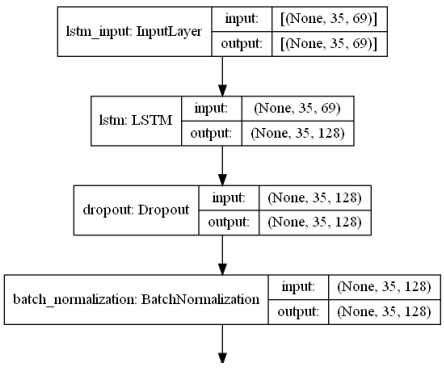
Maskinlæringsmodeller har hyperparametre. Dette er konfigurerbare parametre som gjør at man kan stille inn modellen for å løse ulike oppgaver. Disse innstillingene blir justert av utvikleren bak modellen. For optimal utnyttelse av modellens potensiale bør dette gjøres for hvert spesifikke datasett. Maskinlæringsmodeller har en rekke ulike hyperparametre, og disse interagerer på en ikke-lineær måte, som gjør optimalisering av disse utfordrende (Brownlee, 2020c).

For å løse hyperparametres optimalisering benyttes to ulike metoder. Datasett I til VI er hyperparametrene identiske og ikke tunet til hvert enkelt datasett [31]. Innstillingene er valgt på bakgrunn av erfaring med at de fungerer på flere ulike datasett med tidsserieprediksjon. For mest nøyaktige og best resultater, bør man fintune hyperparametren. Dette er imidlertid svært tidkrevende, derfor er det kun utført på modell VII.

For modell VII, benyttes Optuna, et hyperparameter optimaliseringsrammeverk. Optuna er et modulært rammeverk som anvender sampling og pruning [32] mekanismer som kosteffektivt søker etter optimale hyperparametre. Dette rammeverket har vist seg å være effektivt som optimaliseringsverktøy (Akiba et al., 2019). Ved å integrerer dette rammeverket sammen med koden til LSTM, kjøres det over 500 fullstendige runs på datasettet til modell VII. Den simuleringen som oppnår høyest nøyaktighet blir lagret og de hyperparametrene den ble

[31] Fullstendig oversikt over hyperparametre som blir brukt er vedlagt i Appendix, på side **??**

[32] Pruning er en metodikk, der forsøk som viser dårlig potensiale blir stoppet tidlig, dette gjør prosessen mer effektiv



Figur 9: Utdrag av modell itil visin nettverks-arkitektur, fullstendig nettverk er fremvist på side: 81

Hyperparamatere	Modell I til 6	Modell VII
Antall skjulte lag	3	7
Aktiveringstype	relu	relu
Klassifikator	softmax	softmax
Optimeringstype	Adam	Adam
Learning rate	0.0001	0.0002
Decay	1e-6	1.4e-5
Tapsfunksjon	Spare categorical crossentropy	Spare categorical crossentropy
Tidsitnervall predikert	7	7
Antall epoker	8* (tidlig stopping)	7

Tabell 5: Oppsummering av hyperparameter innstillingene for modellene

gjennomført med lagres og disse er de som blir andvendt. Hyperparameteren med arkitektur er vedlagt i appendix på side 81. Hyperparameterene for modell VII er i appendix på side **??**

3.4.2 Kvalitetssikringsprosesser

For å sikre at resultatene som oppnås er plausible og gir korrekte svar gjennomføres det kvalitetssikringsarbeid. En utfordring er balansen mellom optimalisering av modellen og hvordan den generaliserer. Generalisering er hvordan modellen deretter fungerer på data den ikke har sett før (Chollet, 2017). Som vist av McNally et al. (2018), er en av hovedutfordringene med prisprediksjon ved anvendelse av deep learning modeller, overfitting. Problemstillingen oppstår mellom det å optimalisere og samtidig oppnå en god generalisering som fungerer på data som modellen ikke er trent på. For å forhindre dette, implementeres dropout funksjon i alle modellens lag. Dropout fjerner tilfeldig valgte input features. Dette gjør at modellen ikke blir for tett knyttet opp mot enkelte av disse. Et annet grep som gjøres for å motvirke overfitting er å bruke tidlig stopping. Denne funksjonen stopper treningen av modellen når den ser at nøyaktigheten stabiliserer. Dette gir en variabel i antall gjennomførte treningsepoker.

For å sikre dokumentasjon av data til modellen integreres Neptun.Ai, et eksperimentstyring- og samarbeidsverktøy som effektivt organiserer og lagrer, hyperparameter, resultater, informasjon om tap, nøyaktighet pr. epoke, samtlige figurer som generes, med mer(neptune.ai, 2020). Resultatene av forsøkene som tas med i oppgaven har en digital logg med innstillinger og tilknyttende data lagret og organisert.

3.5 Evalueringskriterier

For å evaluere modellene og velge ut hvilke som blir tatt med videre, er det andvendt ulike kriterier. Den første evalueringskriteriet som tas er den totale nøyaktigheten.

(Rett positiv: RP, rett negativ: RN, feil positv: FP, feil negativ: FN)

$$nyaktighet = \frac{RP + RN}{RP + RN + FP + FN} \tag{16}$$

$$nyaktighet - short = \frac{2 \cdot RN}{RP + RN + FP + FN} \tag{17}$$

$$nyaktighet - long = \frac{2 \cdot RP}{RP + RN + FP + FN} \tag{18}$$

Deretter er det forholdet mellom testtapet og nøyaktighetsutviklingen som vurderes. Dersom testtapet er stigende mens nøyaktigheten er stabil eller fallend underveis i trening, kan dette være en indikasjon på at modellen overfitter. Etter dette er undersøkt, blir modellens ROC-kurve og forvirringsmatrisen vurdert. Forvirringsmatrisen angir nøyaktighet short og long, men for å illustrere har jeg valgt å multiplisere tallet med 2 slik at man får frem nøyaktigheten isolert sett for prediksjonsrettningen.

3.6 Arbeidsprosessen

Undervies i arbeidet, har det vært ulike utfordringer. Det å finne en god metode for variabelutvelgelse viste seg å være krevende. Dette er en viktig del av oppgaven fordi problemstillingen omfatter å vurdere et stort antall ulike variabler. Det er viktig at ikke variabler som potensielt kan ha en verdi blir valgt bort. Samtidig så fungerer modellen dårlig når den får for mange variabler inn. Måten variablene settes sammen på er også viktig, dette ble tydelig etter de mislykkede forsøkene på å organisere datasettene for markedsinformasjon kun etter kategori.

Under evalueringen kom det frem at datasettene som var konstruert med Boruta metoden, ikke ga gode resultater. Modellene fikk en relativt lav total nøyaktighet, men primært lå problemet i at de tenderte til å få et for stort retnings bias. Derfor valgte jeg å gjennomføre en variabelutvelgelse basert på resultatene fra samtlige datasett og sette sammen et sett som tok for seg det jeg anså som de mest lovende variablene. Deretter ble det gjennomført en initiell analyse hvor variablene med liten til ingen SHAP verdi ble fjernet. Resultat av denne prosessen er modell VII. Det er denne modellen som danner grunnlaget for resultatkapittelet. I appendix ligger resultater og data fra de seks andre modellene.

Del 2

Resultater, diskusjon og konklusjon

4 Resultater

I dette kapittelet blir det gjennomgått modell VII. Det blir først presentert resultatene med nøyaktighet, ROC kurve og forvirringsmatrise. Deretter blir det gjennomført en analyse med SHAP verktøyet. I første del av denne analysen blir hele datasettet analysert før det gåes i detalj på de fem viktigste variablene rangert etter shap-verdien deres. Tilslutt blir det presentert en oppsummering av kapitelet og resultatene fra de andre modellene.

For å vurdere en maskinlæringsmodell er det ikke tilstrekkelig å kun måle den utifra klassifiseringsnøyaktighet (Doshi-Velez and Kim, 2017). Det er behov for flere verktøy for å forklare hvordan modellen kommer til valgene. For enkelte problemstillinger vil det være nok å kun få ut en prediksjon. Behovet for å forstå modellen kommer med ikke optimale problemformuleringer (Doshi-Velez and Kim, 2017), som betyr at for enkelte problemer er det ikke nok å kun få en prediksjon. Modellen må også forklare hvordan den kom fram til prediksjonen.

Nøyaktigheten til modellen måles ved to parametere, nøyaktighet og arealet under ROC kurven AUROC. Ved å se på ROC-kurven kan man sammenligne ulike modeller direkte, selv om enkelte andvender terskler og andre ikke. Det fremkommer også visuelt tydelig om modellen har predikative ferdigheter. Er AUROC over 0.5, vil det si at modellen har predikative ferdigheter. Nøyaktigheten blir presentert mer detaljert ved å bruke en forvirringsmatrise.

For å bruke en maskinlæringsmodell som et verktøy til prisprediksjon, vil det for en trader være essensielt at han forstår hvordan den fungerer og på hvilke grunnlag den kommer med prediksjoner. Dette behovet er forankret i hvor komplekst et finansielt markedet er. Gode avgjørelser krever en oversikt over ulike faktorer, det kan være faktorer som spiller inn som modellen ikke har med i sitt variabelgrunnlag. Dette gjør at modellen f.eks. vil gi gode prediksjoner i gitte situasjoner, mens i andre vil en trader måtte ta hensyn til andre faktorer som vurderes som viktigere. For å løse dette, har jeg valgt å bruke SHAP, som beskrevet i delkapittel 2.5 på side 26.

4.1 Baseline

En baseline blir ofte brukt for å ha et sammenligningsgrunnlag forankret i egen data. Dette er viktig fordi det er svært mange variabler som gjør det lite hensiktsmessig å sammenligne egne resultater direkte med andre modeller. Derfor brukes baseline for å ha et referansepunkt til å vurdere om det er en effekt av å gjøre ulike endringer til maskinlæringsarbeidet. I denne oppgaven er disse endringene primært knyttet opp mot ulike variabler.
Innen maskinlæring er er det utfordrende å forutsi om data man anvender vil ha nytte for modellen (Brownlee, 2017). Ved å etablere en baseline finner man ut hvordan modellen fungerer på den enkleste formen av data. Dette gjør at man er i stand til å vurdere effektivt om nye parametere man tilfører et datasett bidrar med verdi til modellen. Til sammenligning oppnådde McNally et al. (2018) LSTM modell som var trent på basis data en nøyaktighet på 52 % [33]. I baselinetesten oppnådde modellen en gjennomsnittlig nøyaktighet på 51 %, med en maksimalverdi på 51.7 %. Datasettet de er trent på har en svak bias, 1.5 % av radene er høyere ved prediksjonsintervallet som kan være en forklaring på at resultatet er svakt over 50 %. Utfra disse tallene dannes et grunnlag for å se om det som tilføres modellen har et positivt bidrag. 6
Modell 6 trenet på 1 minutt antall, tidsserie datapunkt er 300 000. Datasettet er fra 17.

Tabell 6: Deskriptiv statistikk av prediksjonsnøyaktighet til modellen

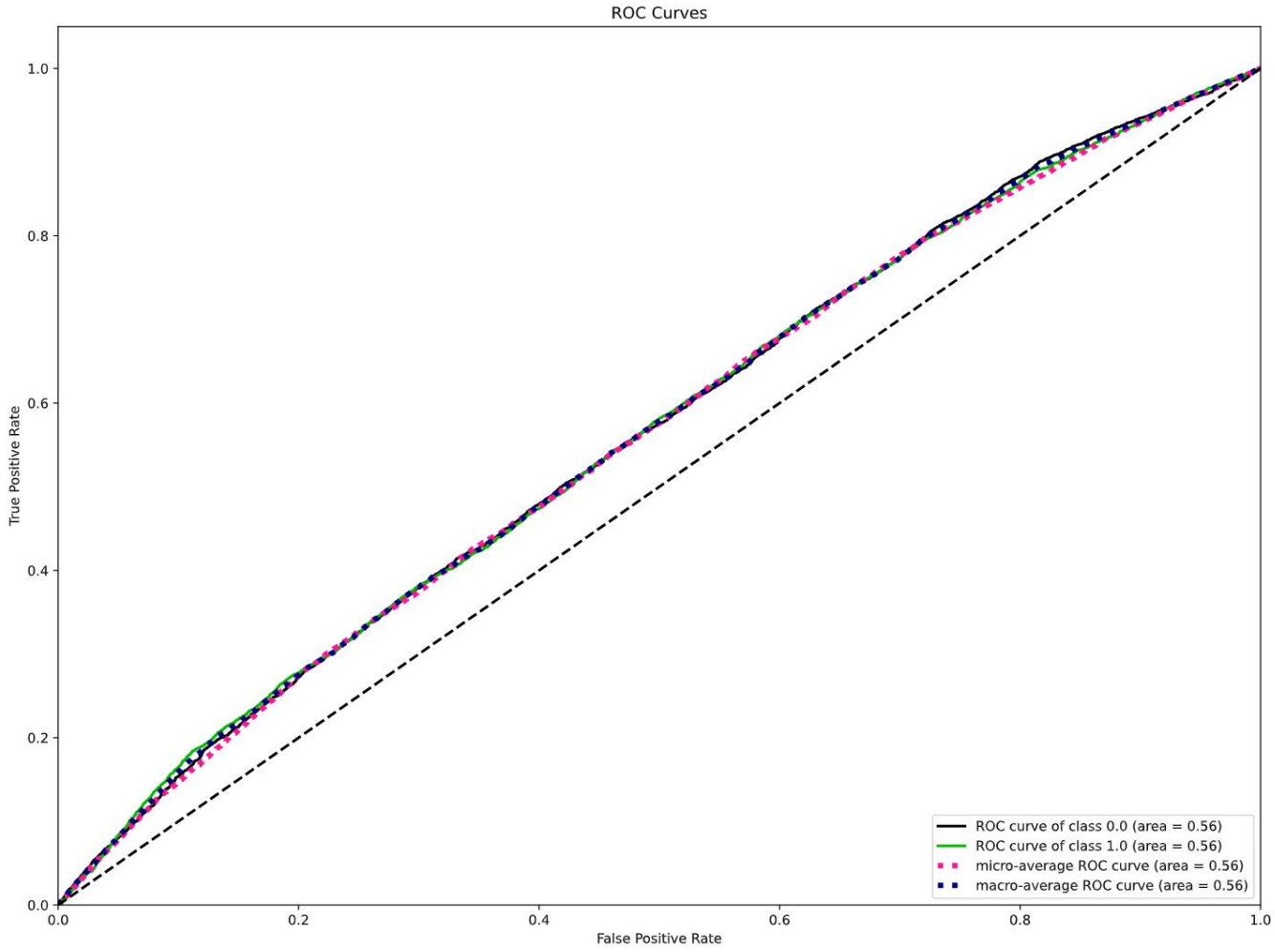
	Gjennomsnitt	Median	Standard avvik	Minimum	Maksimum
BTC	0.510	0.512	0.007	0.495	0.517

februar 2020 til 31. mars 2021. Datasettet inneholder pris ved slutten av hvert tidsintervall, volumet og antall handler utført. Statistikken er basert på 5 runs.

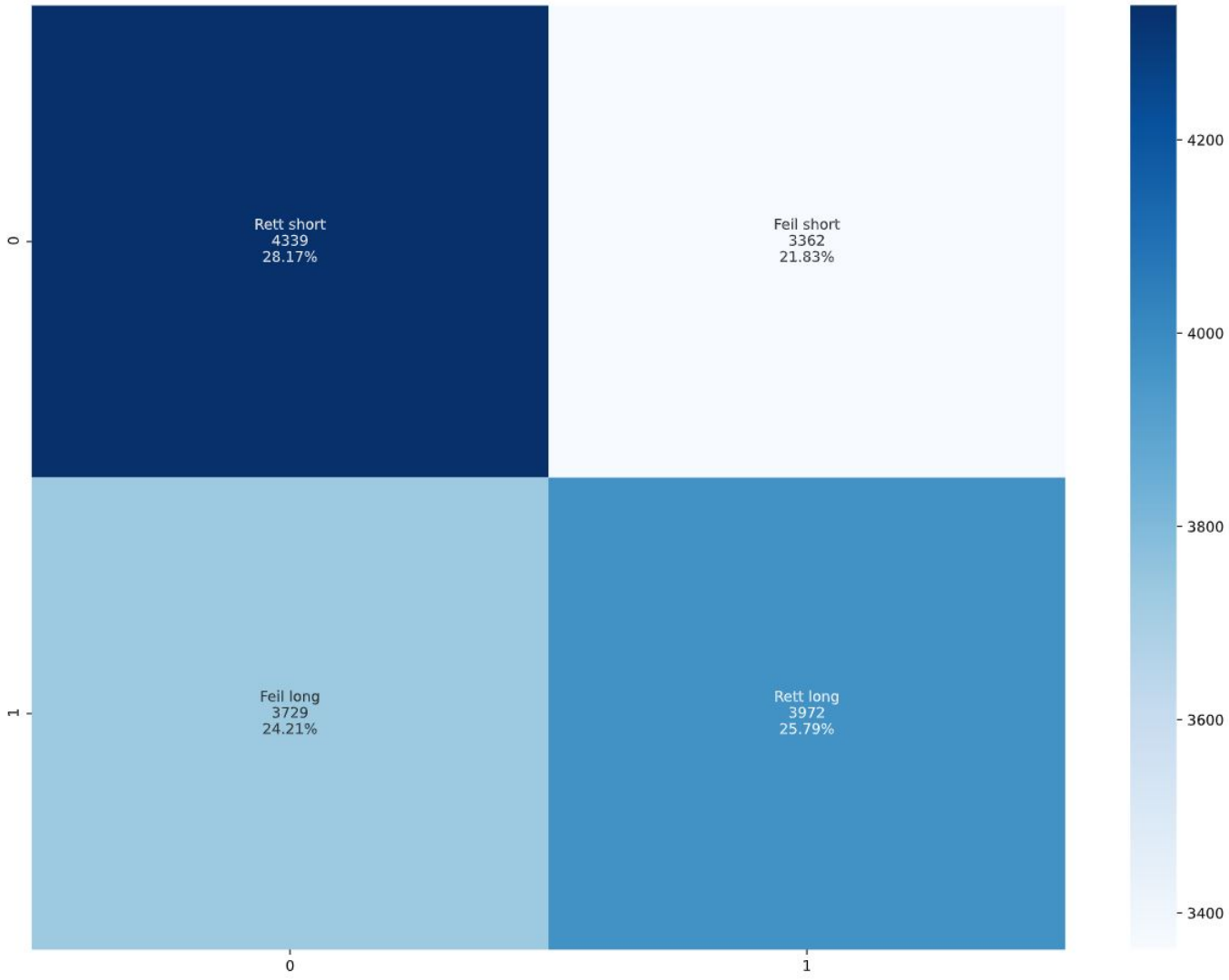
[33]Dette er en sammenligning som gjøres med forbehold. Datasettet denne dataen er hentet fra er anderledes en det som brukes i mitt tilfelle, men det gir et grunnlag for å si at resultatene på baslineresultatene her er plausible.

4.2 Analyse av resultater fra endelig modell

Den endelige modellen oppnår en total nøyaktighet på 54.0 %. For å forstå bedre hva som ligger bak starter jeg å se på ROC-kurven som vist i figur 10. Denne gir raten modellen klassifiserer korrekt. Både kurven for short og for long prediksjonen er over den stiplede midtre linjen (går lineært fra punktet 0.0 til 1.1) som representerer en modell med ingen predikative ferdigheter. Arealet under kurven er likt for både long og short posisjoner med 0.56 for begge.



Figur 10: ROC kurve for endelig modell



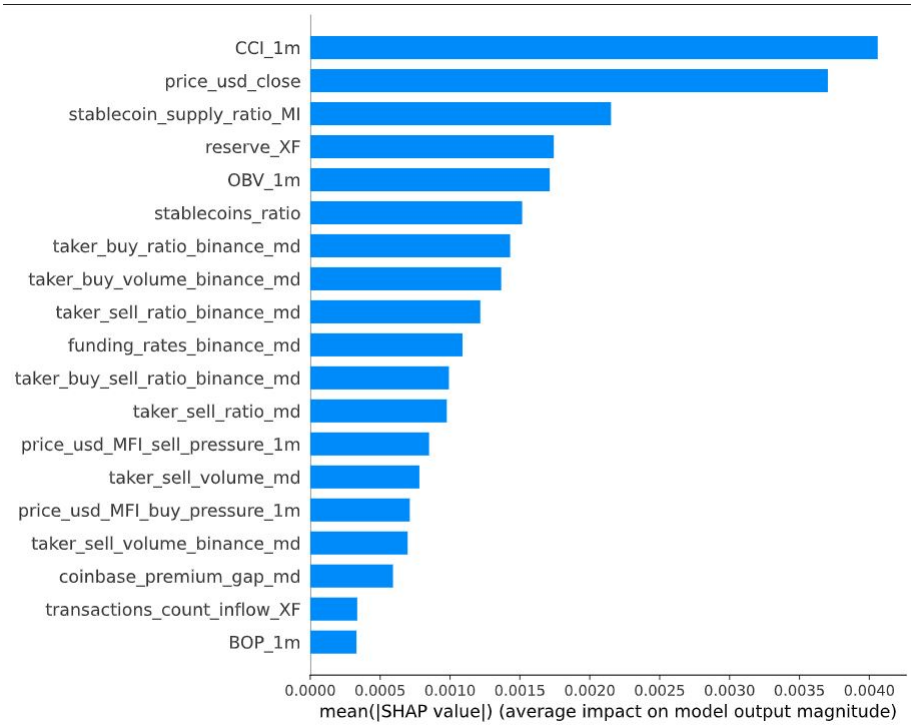
Figur 11: Forvirringsmatrise for endelig modell

Forvirringsmatrisen gir et bildet på modellens klassifiseringsevne. Den har har rett på 56.34 % av prisnedgangs prediksjonene og 51.58 % på prisoppgangs prediksjonene. [34].

[34]Tallet på figuren er multiplisert med 2 for å få ut den isolerte prosenten av de to retningene

4.2.1 SHAP analyse av endelig modell

Nå ønsker jeg å gå mer i dybden i modellen for å få svar på spørsmålet knyttet opp mot variablenes viktighet for prediksjonene. For å løse dette benyttes analyseverktøyet SHAP. Først presenteres det et barplot som gir den absolutte (positivt og negativt bidrag - altså bidrag til prisnedgangs prediksjoner eller prisoppgangs prediksjoner) SHAP verdien for alle variablene som brukes i datasettet. Deretter kommer et oppsummeringsplot der SHAP verdien for de to ulike klassifiseringene blir tydeligjort, og her blir de fem viktigste variablene gjennomgått med hva figuren gir av informasjon.

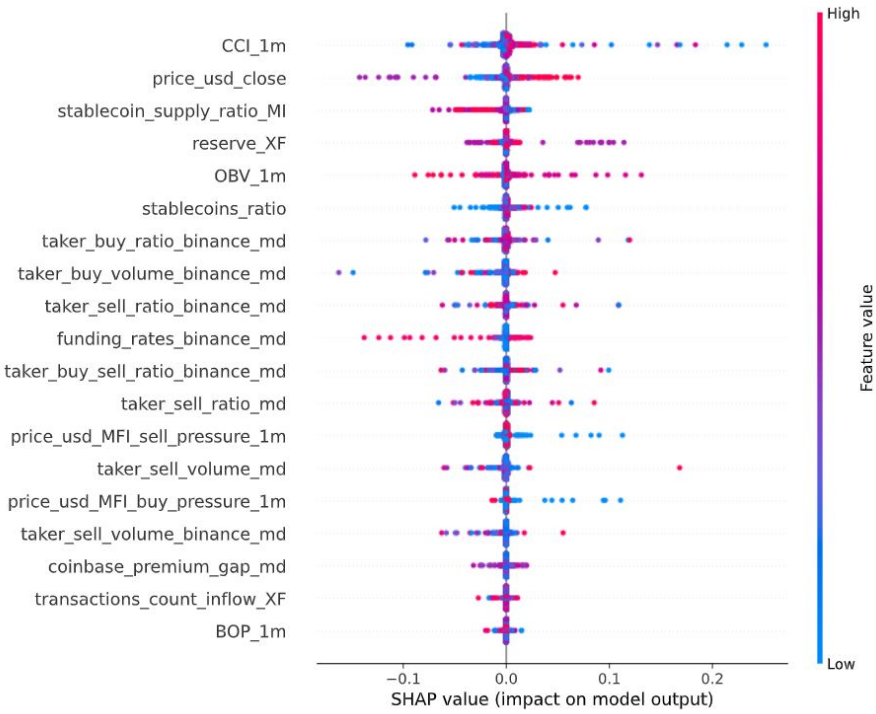


Figur 12: Bar plot for endelig modell

Det fremkommer av figur 12, at både commodity channel index (CCI) og [35]. De skiller seg ut fra resten av variablene. Av markedsinformasjonsvariablene så ser jeg at de som gir informasjon om markedstilbudet for BTC, reserve, og de som angir tilgjengelig USD i markedet: stabalecoin supply ratio, stbalecoins ratio vektlegges. Videre er det flere variabler som har en ganske lik SHAP verdi. Disse har alle med lokal tilbud og etterspørsel som er knyttet opp mot ordreboken til kryptobørsen Binance.

[35]CCI - er en av de tekniske indikatorene. Det er en momentumbasert oscillator som indikerer pristrend og styrke. Verdier over 0 er en indikasjon på videre prisstigning, verdier under, er indikasjon på nedadgående prisMitcham (2005)

SHAP-oppsumeringsplot



Figur 13: SHAP oppsumeringsplot for de 5 viktigste variablene

Forklaring av Figuren: Variablenes viktighet sorteres i synkende rekkefølge. Verdiene er plottet mot målvariabelen som er prisen n steg med fram i tid. Hver enkelt dott representerer en prediksjon og hvordan variablene påvirket modellen til prediksjonen. Fargene representerer verdien innad i variabelen, blå lav verdi, rød høy verdi, mens lilla er nøytral Lundberg and Lee (2017c). Posisjonen langs x-aksen avgjør hvilken retning prediksjonen er, short til venstre og long til høyre.

I påfølgende del vil det bli det presentert en forklaring på de fem viktigste variablene fra figur 13.

Commodity channel Index

Det oppstår en klynge på høyresiden av middelverdien, variabel verdi for CCI er høy (dette fremkommer tydelig av de røde dottene). Dette er en indikasjon på at variabelen bidrar til å predikere at prisen skal øke. Det er det samme på venstre side men med lav CCI verdi. Samtidig er det også tydelige ytterliggende prediksjoner der lav CCI har bidratt til å predikere at prisen skal øke (leses av de blå dottene som ligger ute på høyre side langs x-aksen). Dette er av interesse og blir diskuterte i figur 14.

Price usd close

Høy variabel verdi for price usd close (PUC) indikerer prisoppgang, lav verdi er med på å predikere prisnedgang. Dette er en indikasjon på at markedet trender i begge retningene. Det er en oppsamling av punkter som har en negativ SHAP verdi ved nøytral PUC verdi. Dette kan forklares ved at modellen ser et mønster som indikerer prisen-reversering. Dette kan f.eks være når prisutviklingen går fra oppgang til å flate ut, som kan indikere at prisoppgaven er over og nå vil prisen reversere.

Stablecioin supply ratio

Det forekommer et negativt bidrag når variabelens indre verdi er høy. Når den er lav har den har et moderat bidrag til å predikere pirsøkning. Dette er informasjon som gir en karakteristikk av en variabel som er viktigst for predikeringer av prisnedgang.

Exchange-reserve

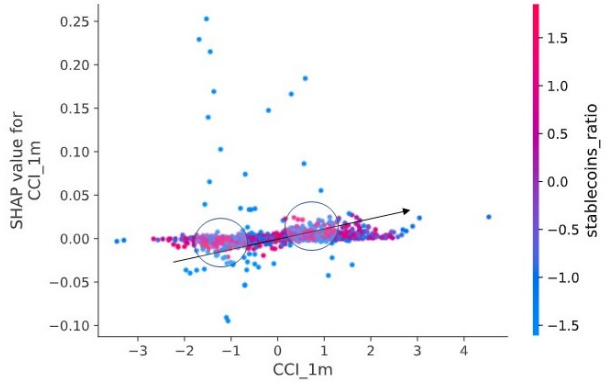
Nøytral verdi for variabelen har en tendens til å gi positive prediksjoner. Dette kan være knyttet opp mot at tilbudssiden i markedet holdes konstant. Det er ellers vanskelig å hente ut mer fra oppsummeringsplottet.

On balance volume

Med høye OBV verdier, har modellen både positive og negative SHAP verdier, mens lave verdier for variabelen resulterer i at modellen gir lave negative SHAP verdier.

SHAP-avhengighetsplot

I denne seksjonen vil avhengighetsplottet til de fem viktigste variablene bli gjennomgått. Hensikten med å gjennomføre en slik analyse er å avdekke mulige forhold mellom de ulike variablene i datasettet.

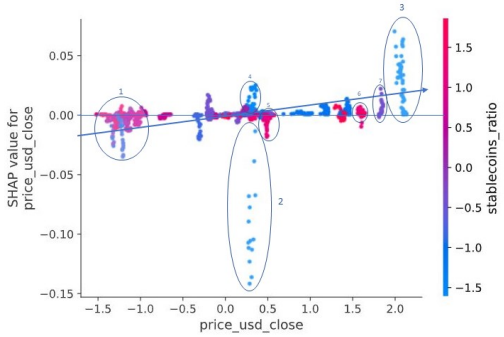


Figur 14: Avhengighetsplot av CCI mot stablecoin ratio

Forklaring av Figuren: Avhengighetsplot er en detaljert illustrasjon av hvordan en variabel samspiller med en annen og hvordan dette påvirker modellens prediksjoner. Variabel 1 (x-aksen og venstre y akse) velges, mens variabel 2 (høyre y-akse) blir valgt automatisk ved at SHAP kernel finner den som har mest samspill med valgt variabel. Prikkene sin posisjon i y-aksen, angir SHAP verdien for variabelen, X-aksen avgjøre den interne verdien for variabelen. Fargen på dotten illustrer indre verdien til den automatisk valgte variabelen Lundberg and Lee (2017c)

Av avhengighetsplottet til CCI, ser jeg en trend (som anvist med pilen), fra negativ SHAP verdier for lav CCI verdi og positv SHAP verdi for høy CCI.
Det er to klynger der jeg ser sammenhengen mellom de to variablene, anvist av sirkelene. Lav CCI og Høy stablecoin ratio (SR) har en svakt negativ SHAP verdi mens høy CCI og lav / blandet SR bidrar til positiv SHAP verdi.
Det er tydelige sterke prediksjoner som jeg ser fremvist med de blå dottene over hovedansamlingen, disse er lave CCI og SR verdier. Dette kan være indikasjon på at modellen har identifisert en reversering i pris der indikatorene er lave, men inne i en skiftende trend.

Price usd close

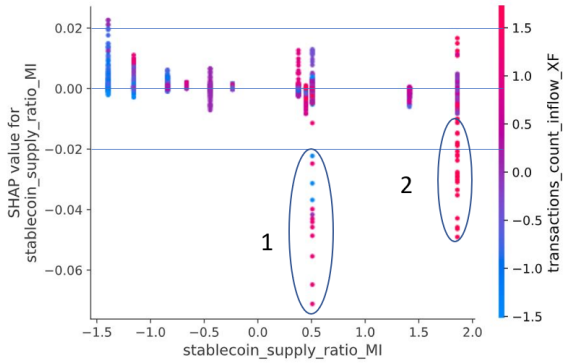


Figur 15: Avhengighets plot for variablene: PUC mot SR

Figur 15 viser en økning fra negativ SHAP verdi fra venstre til en økning mot høyre. Dette indikerer at prediksjonene tenderer mot å trende, dvs økning i pris tenderer til videre økning. Det samme gjelder med nedgang i pris, samme som ble sett i figur 13.
Områdene markert med 1:3 er de mest høyest SHAP bidrag. Området 1 er SR blandet, mens i 2 og 3 er den tydelig lav. Området 2 ser jeg en modell som har funnet en sammenheng mellom svak økning i pris kombinert med lav stablecoins ratio. Dette gir en signifikant negativ SHAP verdi. I området 3 ser jeg en høy PUC, og lav stablecoin ratio gir en positiv prediksjon. Områdene 4:7 viser hvordan modellen finner mindre signifikante sammenhenger men de er ganske tydelige. Det fremkommer at de høyeste SHAP verdiene er negative der maksverdiene er over -0.15, de positive verdiene er på maks 0.6.

Figuren illustrerer også hvordan en maskinlæringsmodell er i stand til å identifisere sammenhenger som er avhengige av variabelens interne verdi, mens det ikke er et lineært forhold fra lav til høy, men at ulike grader har en betydning og at den må ses i sammenheng med andre variabler.

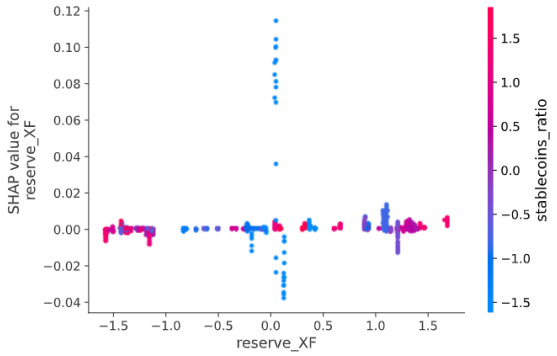
Stablecioin supply ratio



Figur 16: Avhengighets plot for variablene: Stablecoin supply ratio mot transactions count inflow

Det fremkommer en trend der lav stablecoin supply ratio (SSR) og lav transaksjonens count inflow (TCI) gir en svak positiv SHAP bidrag. Når de to variablene blir høyere ser jeg en økende grad av negativt SHAP bidrag, og dette gir også en indikasjon på at de har et lineært forhold, noe som er logisk mtp hva variablene er basert på.
Det er en tydelig samling av punkter som skjer med SSR rundt 0.5 (markert med 1), der den får et bidrag til negativ SHAP verdi. Området 2 er også signifkant
Denne analysen bekrefter det som ble funnet i oppsummeringsplottet, SSR er viktigst for modellens negative SHAP verdi, og den er dermed viktigst for å predikere negativ prisutvikling.

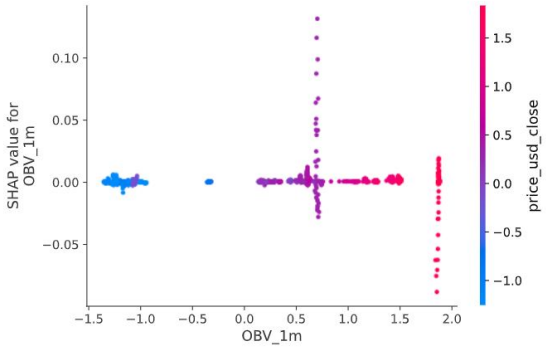
Exchange-reserve



Figur 17: Avhengighets plot for variablene: Exchange reserve

Fra avhengighetsplottet ser jeg at prediksjonene primært ligger innenfor svært lave SHAP verdier. Rundt midtpunktet (nøytral) ser jeg et større SHAP bidrag, både positivt og negativt (det positive bidraget har en betydelig høyere SHAP verdi). For samtlige av disse prediksjonene er SR verdien lav. Det er tydelig at modellen finner et mønster mellom de to variablene der den har et betydelig positivet SHAP bidrag eller et moderat negativt bidrag.

On balance volume



Figur 18: Avhengighets plot for variablene: OBV mot price usd close

Figur 18 viser hvordan en lav OBV [36] verdi er av liten betydning for modellen, variabelen har et signifikant positivt bidrag når OBV variabelen er svakt positiv, i området 0.7 til 0.8 og price usd er nøytral. Når OBV er høy, over 1.6 til 2 og POC er høy ser jeg et delt positivt/negativt SHAP bidrag, men med mer signifikante negative verdier. Variablene har både et bidrag til short og long prediksjoner.

[36]On-Balance Volume (OBV) er en pris og volum trend kvantifisering. Indikatoren er populært brukt ved at man går long når OBV verdien er over gjennomsnittet for et satt antall tidligere perioder, og gå short når prisen er under gjennomsnittet av gitte tidligere tidsperiode (Mitcham, 2005)

4.3 Resultater fra alle datasettene

Tabell 7: Oppsummering av resultater

Modell	Nøyaktighet	Nøyaktighet short	Nøyaktighet long	A-ROC
Modell I	52.4 %	60.0 %	44.7 %	0.53
Modell II	52.3 %	64.1 %	40.5 %	0.53
Modell III	52.1 %	60.8 %	41.3 %	0.53
Modell IV	52.6 %	45.8 %	59.5 %	0.53
Modell V	53.2%	50.0 %	56.4 %	0.54
Modell VI	54.6 %	55.5 %	53.76 %	0.55
Modell VII	54.0 %	56.3 %	51 .6 %	0.56

Tabell 7 gir de overordnede resultatene avdekket i oppgaven. Alle datasettene har en høyere nøyaktighet enn både den gjennomsnittlige og maksverdien etablert i baseline resultatene, med hhv. 51.0 % og 51.7 %.
A-ROC verdien for alle modellene er over 0.5, og det kan også leses av på ROC kurvene. Dette er en indikasjon på at modellen har predikative ferdigheter.

I delkapittelet lignende litteratur ble det trukket frem nøyaktighetene som ble presenter for andre datasett som predikerte BTC sin prisutvikling. Disse lå innenfor 52- 60 % prediksjons nøyaktighet. Som drøftet tidligere kan ikke dette sammenlignes direkte, men gir en indikasjon på at resultatene her er plausible.

Ved å gå litt dypere i resultatene, fremkommer det at flere av modellene får et bias mot enten prisoppgang eller -nedgang. Modellene I til IV har en signifikant høyere nøyaktighet i en spesifikk retning, og årsaken er kan være overfitting, eller en ubalanse i datasettet modellene er trent på. Dette kan ha med å gjøre at flere av markedsinformasjons variablene gir informasjon som kun påvirker en spesifikk retning. En annen forklaring er at modellene har blitt overfitted. Det kan være ved mer arbeid med disse modellene kan man optimalisere de for kun å predikere en bestemt retning. Implikasjonene av å andvende slik de er nå, ville trolig gitt dårlig avkastning. Modell V,VI og VII ligger innenfor det som kan anses som et akseptabel avvik mellom de to retningene og anses som å gi plausible resultater.

5 Diskusjon

I dette kapitellet blir de viktigste funnene fra resultatkapittelet drøftet for å belyse ulike sider knyttet til problemstillingen.

Resultatene viser modellens prediksjoner blir mer nøyaktig ved å trene på tekniske indikatorene. Dette samsvarer med resultater fra Chen et al. (2020), Li and Dai (2020) og Alonso-Monsalvea et al. (2019).I denne oppgaven bennytes de samme tekniske indikatorene, men datasettet blir utvidet med markedsinformasjons variabler. Fra resultatene oppnåes det positive SHAP verdier i kombinasjon med positiv nøyaktighet for samtlige av modellene. Dette danner en basis for å trekke slutningen at variablene kan ha et bidrag til maskinlæringsmodellen som styrker dens predikativ effekt. Det viktigste variablene er de som ble utvalgt for modell VII.

Det er av relevans for oppgaven å få frem bidraget variabler som gir informasjon om tilbudssiden til både BTC og USD har for modellen. Fra modell VII får *stablecoin supply ratio* og *reserve*, begge signifikante SHAP verdier. Det fremstår som plausibelt, ved å få informasjon om antallet tilgjengelige BTC og USD på børsene, med frekvent informasjonsoppdatering, er med på å øke prediksjonsnøyaktigheten.
For det amerikanske aksjemarkedet ble det utført en studie som viser en sammenheng mellom prisendring og ubalanse mellom kjøp og salgsordre i ordreflyten (Hopman, 2007). Gjennom markedsinformasjon har modellen variabler som gir denne informasjonen[37]. Ved å få informasjon om ordrer i ordeboken[38], kan modellen overvåke ordreflyten.

CCI er en teknisk indikator som gir informasjon om potensiell fremtidig volatilitet. Denne indikatoren har vist å kunne brukes som basis til profitable tradingstrategier. (Shah, 2019) testet den ut på options-kontrakter og positive resultater profitable resultater. Basert på SHAP verdiene, er dette den viktigste variabelen for modellen. SHAP har funnet at den har mest signifikant interaksjon med stablecoin ratio variabelen. Ved å analysere plottet var det en svak trend, men vanskelig å si noe entydig. Den informasjonen som kan hentes fra dette, er hvordan modellen har en effekt av å kombinere både tekniske indikatorer og markedsinformasjon.

[37] taker buy ratio binance, taker sell ratio binance, taker sell ratio binance og taker buy ratio
[38] En ordrebok, er oversikten over alle ventende kjøp- og salgs ordre

Ved å analysere avhengighetsplottet til den tekniske indikatoren OBV, er det et tydelig mønster mellom høy OBV og høy price usd close. Dette medfører en negativ SHAP verdi. Dette er konterintuitivt. Basert på hvordan OBV brukes tradisjonelt, så er en høy verdi en indikasjon på økende kjøpsvolum. Dette kombinert med høy price USD ville indikert en positiv kursendring. Det kan være andre sammenhenger som ikke fremkommer av denne figuren som fører til dette. Dette er med på å understreke viktigheten av å undersøke mønstrene som fremkommer av en SHAP analyse.

Resultatene som er oppnådd i denne oppgaven for alle modellene er over den etablerte baseline nøyaktigheten på 51.0 %. Modellen som oppnådde best nøyaktighet består av kun tekniske indikatorer, den endelige modellen fikk en total nøyaktighet på 54 %, og en AUROC 0.56. Videre oppnådde alle modellen en positiv AUROC og nøyaktighet over 50 %.
Det å direkte sammenligne resultater med annen litteratur, er ikke hensiktsmessig. Dette på bakgrunn av alle de ulike variablene som vil spille inn, dette er spesielt tydelig for prisprekskjonsarbeid der det ikke finnes en standard. Likefremt, er det av relevans til oppgaven å danne et bilde av hva som kan anses som plausible resultater. Chen et al. (2020) oppnådd 67 % prediksjons nøyaktighet på BTC. [39] Alonso-Monsalvea et al. (2019) presenterer en nøyaktighet på over 60 %.
En mulig årsak til differansen innen nøyaktighet mellom resultatene, kan ligge i en utvikling i markedet. Datasettet disse modellene er trent på er eldre og representerer ulike markedskondisjoner. Med prisutviklingen til BTC skjer det endringer i markedsaktørene, der det er naturlig å forvente en økt konkurranse, flere sofistiskere markedsdeltagere. Dette medfører større utfordringer for modellen og vil føre til lavere nøyaktighet. Videre kan det også være knyttet til mindre restriksjoner på modellene eller potensielt overfitting. Det er viktig å bemerke, høy nøyaktighet i testfasen vil ikke nødvendigvis bety gode resultater i et reelt scenario (Babyak, 2004).

I denne analysen, legges det mye vekt på resultatene fra verktøyet SHAP, og at dette gir et korrekt bilde av hva som skjer på innsiden av maskinlæringsmodellen. I en studie utført av Slack et al. (2020) ble det konstruert spesifikke datasett i den hensikt å utfordre SHAP og LIME [40], de utførte dette ved å konstruere datasett der både SHAP og LIME sine resultater

[39] Dette resultatet er svært høy, så dette vurderes som et lite plausibelt resultat.

[40] LIME er en metode som ligner på SHAP som brukes som verktøy for å forklare maskinlæringsmodeller og hvordan de fungerer

viste en tydelig bias som ikke reflekterte de faktiske underliggende biasene. Med andre ord så klarte de å fremprovosere feil representasjon av variablenes viktighet for modellen. Dette ble gjort ved å konstruere et spesifikt datasett som utnyttet en svakhet ved kategoriske variabler, slik at de manipulerte resultatene (Slack et al., 2020).

Dette er viktig å ta med, for selv om man oppnår gode resultater ved å bruke ulike analyseverktøy, så må det undersøkes om det er plausible forklaringer og resultater man finner. Videre er det ikke mange av variablene som modellen benytter seg av som er kategoriske, så for denne oppgaven anses det som lite sannsynlig at dette vil være et problem.
Til tross for svakhetene,er metoden andvendt i flere ulike fagfelt. Eksempelvis er den andvendt for å kartlegge hva som er de viktigist faktorene for et vellykket bedriftsoppkjøp Futagami et al. (2021). Vurdering av seismiske skader på ulike strukturer (Mangalathu et al., 2020) og ved avdekking av årasaken til trafikkulykker (Parsa et al., 2020). Dette er ulike felt, men hensikten bak er de samme, å forsøke å forstå hvilke variabler som er viktigst for maskinlæringsmodellene, og svaret hentes fra en SHAP analyse.

Selv om det fremkommer at flere av variablene har en predikativ verdi, er modellens nøyaktighet på et nivå der en implementering på et live marked med handelskostnader [41] entropi, usikkerheten og andre variabler som spiller inn, gjør at det ikke oppnås profitable prediksjoner. Til tross for dette, er det flere funn som jeg mener har en reel verdi. Både tekniske indikatorer og markedsiformasjonsvariablene ser ut til å ha et potensiale. De tekniske indikatorene har blitt grundig utprøvd og presentert i annen litteratur, men markedsinformasjons variablene er nye. Disse har vist å kunne gi informasjon om deler av markedet som er av relevans med tanke på kursutvikling. Jeg ønsker å spesielt trekke frem muligheten til å få en kvantifisering av tilbuds siden for både BTC og USD, og ordreflyten knyttet opp til børsene sine ordrebøker. Dette underbygges av arbeid fra andre finansielle markeder som beskrevet av (Hopman, 2007) for aksjemarkedet.

[41] Handelskostnader for BTC på Binance er på 0.1 % pr handel. Dersom man skal innta en short posisjon ville kostnadene vært flytende.

I lignende litteratur som McNally et al. (2018),Alonso-Monsalvea et al. (2019), Chen et al. (2020), har fokuset ligget på helautomatiserte handelsstrategier. Jeg har lyst til å se på hvilken teknisk informasjon som kan være relevant for en trader dersom en maskinlæringsmodell skulle blitt brukt som et beslutningstakningsverktøy, eller en form for teknisk indikator. Som vist tidligere er prediksjonsnøyaktigheten som kan forventes innenfor prisprediksjon, i sjiktet mellom 54 - 60 %. Dette vil være en barriere.
Som vist av Grgić-Hlača et al. (2019) er mennesker tvilsomme til algoritmer som er imperfekte. Så for at jeg skal bruke dem må det etableres tillit Ribeiro et al. (2016). I denne oppgaven har modellen gitt binære prediksjoner, opp eller ned. Som beslutningsverktøy tror jeg ikke dette er hensiktsmessig. Da tror jeg det vil være bedre med en modelloutput som gir en prosentsannsynlighet for prisendringer i gitte retninger. Dette gjør at dersom modellen er usikker, kan traderen bruke den informasjonen, og dersom modellen er sikker vil dette kunne lede til en annen tilnærming.

Den tekniske informasjonen, som i dag er mulig å hente ut, vil kunne bidra til, om ikke å skape tillit, potensielt gi en bakgrunnsinformasjon om hvorfor modellen predikerer slik som den gjør. Ved å utarbeide analyse av mønstre som leder til sterke prediksjoner, som f.eks. 15 på side 55. Videre bør det foreligge forvirringsmatrise, ROC-kurve og god dokumentasjon på datasettet modellen er trent på, slik at traderen har forståelse for hvilket type marked modellen er tilpasset. En interessant mekanisme er at for å analysere en slik modell, så bør dette gjøres av en med markedsforståelse og god innsikt i hvordan de ulike tekniske indikatorene og markedsmekanismene tradisjonelt blir brukt. Den som har best forståelse for dette er traderen, så en tanke er at han/hun bør være involvert i prosessen med å utvikle et slikt støtteverktøy, der traderen påvirker hvilke variabler som blir brukt, hvordan problemstilling modellen løser og viktigst, være med å forsøke å forstå om modellen finner faktiske mønstre. På det siste punktet, kan løses på samme måte som resultatene har blitt presentert i denne oppgaven. Ved å gå igjennom en SHAP oppsummering, for deretter å analysere alle variablene individuelt. Dette kan også gjøres i enda større detalj ved å bruke enda flere av funksjonene til SHAP.

Dersom det viser seg at modellen faktisk ser et gitt mønster, og traderen er kjent med dette på forhånd så han verifiserer at modellen handler på en plausibel måte, er nok det en måte et slik verktøy kan ha en form for verdi.
Til tross for dette, tror jeg, med de analyseverktøyene som er tilgjengelige i dag, at det fortsatt

vil være behov for bedre og mer forståelig metoder før det blir et verktøy som kan brukes effektivt i et finansmarked.

Resultatene jeg har kommet frem til er en konsekvens av en prosess med en rekke valg. Det er ingen standardisert metode for utvikling og evaluering av maskinlæringsmodeller. For å sørge for en systematisk prosess er metoden bygget rundt CRISP-DM med enkelte elementer fra teorien til Brownlee (2017). Selve evalueringsprosessen rundt det å si om en variabel har eller ikke har informasjon som kan bidra til prisprediksjon, er utfordrende. I likhet med prosessen er det på dette feltet heller ingen standard for å løse dette. Jeg har valgt å vektlegge AUROC og SHAP verdier. Dette er samme metoden som Seki et al. (2021) har andvendt i sitt arbeid, men på et helt annet felt. De analyserte sannsynligheten for dødsfall, i det en person innlegges på sykehus ved å bruke maskinlæring. Fokuset deres var å kartlegge de viktigste variablene, og dette ble gjort ved å evaluere modellen ut fra AUROC og SHAP. Dette er en helt annen problemstilling enn det som gjennomgås i denne oppgaven, men jeg mener essensen i å se om dataene kan ha predikative egenskaper.

5.1 Metodiske valg

I arbeidet har en sentral utfordring i prosessen vært knyttet opp mot det å forutsi på forhånd hvordan ulike valg vil påvirke resultatet, og hva det er som vil fungere eller ikke.
I etterkant av arbeidet, er det flere tydelig svakheter som har blitt oppdaget. Disse kommer av valg som er tatt underveis, som med fordel kunne vært gjort anderledes, som potensielt ville gitt bedre resultater. Modellen er pålagt flere restriksjoner, og av disse er det viktig å trekke frem at modellen må ta med likt antall opp og ned prediksjoner i resultatberegningene. Det var et valg som ble tatt for å motvirke ubalanse i datasettet og gi et bedre bilde av prediksjons-nøyaktigheten.

Den viktigste designavgjørelsen som i etterkant kan ha hatt en negativ påvirkning på resultatet, er at modellen gjennomfører en prediksjon hvert eneste minutt. Dette gjør at modellen tar en avgjørelse, selv om marginene for prediksjonen er svært små. Med dette menes at modellen finner en sannsynlighet for om den predikerer pris oppgang eller nedgang, estimatene skiller ikke på om modellen er 50.1 % sikker på en retning eller 60 %. Dersom modellen hadde vært bygget som multiklasse klassifiserings, der den kunne ha en nøytral prediksjon, ville man kunne hente ut færre prediksjoner, mens de som ble gitt ville potensielt hatt høyere treff rate.

Den siste konstruksjonsendringen er å gå fra en enkelt modell som er konstruert for å predikere to retninger, til to modeller som er spesifisert for en enkelt retning. Det er enkelte variabler som har en tydelig SHAP verdi i en enkelt retning, mens lite bidrag i en annen. Dersom man konstruerer to modeller hvor den ene er tiltenkt å predikere prisøkning og den andre prisnedgang. På denne måten kan man spesialisere en modell for en spesifikk retning, med ulike variabler for de to.
For å kunne si mer om hvordan modellene ville fungert i et reelt miljø, burde det vært gjennomført testing på et adskilt testsett i tilegg.

5.2 Verdien av arbeidet i oppgaven

Selv om modellene ikke ga resultater som gjør at de vil kunne anvendes, er arbeidet med å se på et stort antall markedsinformsjonsvariabler av verdi. Det fremkommer av resultatene der modellen henter ut verdi fra disse som bidrar til å øke korrekte prediksjoner. Dersom disse anvendes i en mer sofistikert markedsoptimalisert modell, vil denne type datasett ha potensialet for å øke graden av nøyaktighet. Dette kan lede til gode prediksjoner for kursutviklings prediksjoner av Bitcoin.
Kryptovaluta er en ung aktivumklasse, som er preget av spekulasjon og vollatile hurtige prisendringer. Investorer og markedsdeltagere har vist at det ikke alltid er rasjonelle avgjørelser som blir fattet. Ved å utvikle bedre og mer presise verktøy, vil dette kunne bidra til å gjøre markedet mer effektivt. Dette vil også være med på å styrke markedet som helhet, da mange vil føle at kryptovalutamarkedet, slik det er i dag, er for usikkert med for store vollatile bevegelser.

6 Konklusjon

I denne oppgaven har det blitt gjennomgått et vidt spekter av variabler som kvantifiserer mange av variablene som er innenfor BTC markedet. Gjennom arbeidsprosessen har det vært flere utfordringer som var vanskelig å forutse, men som er med på å understreke utfordringene knyttet til maskinlæring, spesielt innenfor finansmarkedet.

Problemstillingen som oppgaven forsøker å gi svar på er: **Finnes det variabler som kan bidra til å predikere bitcoin prissetting på korte tidsintervall?**
Gjennom resultatene, kan det konkluderes med at ved å kombinere både tekniske indikatorer og markedsinformasjon, klarte modellen å vise prediktive kapabiliteter.
Spesielt tekniske indikatorer vist potensiale, modell VI oppnådde høyest nøyaktighet, og denne ble ikke optimalisert med egen hyperparameter. Det er oppnådd både nøyaktighet og AUROC resultater som er bedre en det en utrent modell ville vært i stand til å leverer.
Til tross for dette, er ikke resultatene gode nok til at dette kan bli anvendte et live miljø, hverken som beslutningsstøtte eller en helautomatisert strategi. I et live miljø er det svært sannsynlig at modellen ville produsert dårligere resultater.

Det er relevant å trekke frem hvordan SHAP analysen ned på variabel nivå, bidrar med å gi et visuelt bilde av hvilken mønster modellen ser og vektlegger. Dette er en metode som blir enda mer tydelig dersom modellen oppnår bedre nøyaktighet. Det å kunne gå igjennom variabelgrunnlaget og arbeide metodisk med å forstå modellen kan man anvende videre for å optimalisere og spisse modellene mot spesfikke retninger. Spesielt innen finansmarkedet, med høy konkurranse, tror jeg dette kan gi resultater som kan føre til profitt eller et godt beslutningsstøtte verktøy.

Besvarelse av forskningsspørsmålet: **Under en implementering av en maskinlæringsmodell som et støtteverktøy til en trader i investeringsøyemed. Hvilken teknisk informasjon om modellen kan traderen ha nytte av?**
Det å arbeide med å forstå maskinlærings modeller er viktig, spesielt dersom disse skal implementeres der de skal fungere sammen med mennesker. Ved å gjennomføre SHAP analyser så bidrar man til å forstå hva som skjer inne i selve modellen.
Dersom det utvikles maskinlæringsmodeller med nøyaktighet som er gode nok til at det kan være aktuelt å implementere de som et verktøy, vil en slik analyse potensielt bidra til å skape

tillit og forståelse mellom en traderen og modellen. Det å få en grundig analyse av et verktøy som skal anvendesvil gjøre at man som trader på en langt mer effektiv måte forstår muligheter og begrensninger til verktøyet. Dette blir på mange måter som en ny type indikator, som er bygget på en rekke andre. Som vist i denne oppgaven er det nå verktøy for å forklare modellen, og disse kan brukes for å gi gode illustrasjoner som kan andvendes i en implementering.

6.1 Viderearbeid

For videre arbeid, ville det vært av interesse å konstruere en modell som er optimalisert rundt å maksimere profitt. Ved å konstruere to modeller, en for å predikere økning i pris og en for å predikere nedgang, for så å finne de viktigste variablene for disse to prissettingene. Deretter kan det være av interesse å se hvilken funksjon modellen optimeres rundt. Ved å optimere for profitt, istedenfor å minimere tap, vil man potensielt kunne få til en modell som ikke nødvendigvis har flere korrekte prediksjoner, men den vil potensielt generere mer profitt, og det er dette man som oftest ønsker. Videre kan dette tilpasses ved at man f.eks optimaliserer etter risikojustert avkastning, eller andre parametere der risiko blir en del av dette.
Markedssen timent er en annen interessant datakilde, som kan være interessant å inkludere i et datasett. En metode for å samle denne dataen er foreslått av Huang et al. (2021) som gjør dette for BTC.

Del 3

Referanseliste, figurlist, tabelliste og appendix

7 Referanseliste

Referanser

A. A. M. Ahmed, R. C. Deo, A. Ghahramani, N. Raj, Q. Feng, Z. Yin, and L. Yang. Lstm integrated with boruta-random forest optimiser for soil moisture estimation under rcp4.5 and rcp8.5 global warming scenarios. *Stochastic Environmental Research and Risk Assessment*, 2021. doi: 10.1007/s00477-021-01969-3.

T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama. Optuna: A next-generation hyperparameter optimization framework. KDD '19, page 2623–2631, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362016. doi: 10.1145/3292500.3330701. URL https://doi.org/10.1145/3292500.3330701.

S. Alonso-Monsalvea, A. L. Suárez-Cetrulo, A. Cervantes, and D. Quintanac. Convolution on neural networks for high-frequency trend prediction of cryptocurrency exchange rates using technical indicators. pages 1, 8 – 11, 2019. doi: https://doi.org/10.1016/j.eswa.2020.113250.

M. A. Babyak. What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine*, 66(3):411–421, 2004. doi: 10.1097/01.psy.0000127692.23278.a9.

W. Bao, J. Yue, and Y. Rao. A deep learning framework for financial time series using stacked autoencoders and long-short term memory. *PLOS ONE*, 12(7), 2017. doi: 10.1371/journal.pone.0180944.

I. Barmes. Quantum computers and the bitcoin blockchain, 2019. URL https://www2.deloitte.com/nl/nl/pages/innovatie/artikelen/quantum-computers-and-the-bitcoin-blockchain.html.

Binance. Python binance, 2021. URL https://python-binance.readthedocs.io/en/latest/overview.html.

J. Brownlee. How to get baseline results and why they matter, Jun 2017. URL https://machinelearningmastery.com/how-to-get-baseline-results-and-why-they-matter/.

J. Brownlee. *Deep Learning for Time Series Forecasting: Predict the Future with MLPs, CNNs and LSTMs in Python*. Machine Learning Mastery, 2018. URL https://books.google.no/books?id=o5qnDwAAQBAJ.

J. Brownlee. How to avoid data leakage when performing data preparation, Aug 2020a. URL https://machinelearningmastery.com/data-preparation-without-data-leakage/.

J. Brownlee. How to choose a feature selection method for machine learning. https://machinelearningmastery.com/what-is-data-preparation-in-machine-learning//, 2020b.

J. Brownlee. How to choose a feature selection method for machine learning. https://machinelearningmastery.com/hyperparameter-optimization-with-random-search-and-grid-search/, 2020c.

Z. Chen, C. Li, and W. Sun. Bitcoin price prediction using machine learning: An approach to sample dimension engineering. *Journal of Computational and Applied Mathematics*, 365: 342, 2020. doi: 10.1016/j.cam.2019.112395.

F. Chollet. page 60:65. Manning Publications, 2017.

Coinbase Institutional. Crypto h1 2020 in review. https://static-assets.coinbase.com/custody/Coinbase-Institutional-Crypto-H1-2020-Review.pdf, 2020.

CryptoQuant. Cryptoquant data guide. https://dataguide.cryptoquant.com/, 2021.

M. A. H. Dempster, T. W. Payne, Y. Romahi, and G. W. P. Thompson. Computational learning techniques for intraday fx trading using popular technical indicators. *IEEE Transactions on Neural Networks*, 12(4):744–754, 2001. doi: 10.1109/72.935088.

N. DiCamillo. The relationship between us government debt and bitcoin, explained, Jan 2021. URL https://www.coindesk.com/the-relationship-between-us-government-debt-and-bitcoin-explained.

B. J. Dietvorst, J. P. Simmons, and C. Massey. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1): 114–126, 2015. doi: 10.1037/xge0000033.

G. Dorffner. *Neural networks for time series processing*. Osterr. Forschungsinst. fur Artificial Intelligence, 1996.

F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable machine learning, 2017. arXiv:1702.08608.

T. Fischer and C. Krauss. Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research 270(2), 654–669.*, pages 655, 667, 2017. doi: 10.1016/j.ejor.2017.11.054.

R. Frank and B. Bernanke. McGraw-Hill/Irwin, 2007.

K. Futagami, Y. Fukazawa, N. Kapoor, and T. Kito. Pairwise acquisition prediction with shap value interpretation. *The Journal of Finance and Data Science*, 7:22–44, 2021. ISSN 2405-9188. doi: https://doi.org/10.1016/j.jfds.2021.02.001. URL https://www.sciencedirect.com/science/article/pii/S2405918821000015.

F. Glaser, K. Zimmermann, M. Haferkorn, M. C. Weber, and M. Siering. Bitcoin - asset or currency? revealing users' hidden intentions. Twenty Second European Conference on Information Systems, Tel Aviv 2014, 2014. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2425247.

A. Graves, A.-r. mohamed, and G. Hinton. speech recognition with deep recurrent neural networks". Department of Computer Science, University of Toronto https://arxiv.org/pdf/1303.5778.pdf, 2013.

N. Grgić-Hlača, C. Engel, and K. P. Gummadi. Human decision making with machine assistance. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–25, 2019. doi: 10.1145/3359280.

P. Hall and N. Gill. *An Introduction to Machine Learning Interpretability.* O'Reilly Media, Inc., 2019.

P. Hapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth. Crisp-dm, Mai 2020. URL https://the-modeling-agency.com/crisp-dm.pdf.

S. Hochreiter and J. schmidhuber. Long short-term memory. pages 1–40, 1997.

C. Hopman. Do supply and demand drive stock prices? *Quantitative Finance*, 7(1):37–53, 2007. doi: 10.1080/14697680600987216.

X. Huang, W. Zhang, Y. Huang, X. Tang, M. Zhang, J. Surbiryala, V. Iosifidis, Z. Liu, and J. Zhang. Lstm based sentiment analysis for cryptocurrency prediction, 2021.

M. Intelligence. Algorithmic trading market: Growth, trends, and forecasts (2020 - 2025), 2021. URL https://www.mordorintelligence.com/industry-reports/algorithmic-trading-market.

Y. Kara, M. A. Boyacioglu, and Ömer Kaan Baykan. predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the istanbul stock exchange". pages 1, 8 – 11, 2011. doi: 10.1016/j.eswa.2010.10.027.

M. B. Kursa and W. R. Rudnicki. Feature selection with the boruta package. *Journal of Statistical Software, Articles*, 36(11):1–13, 2010. ISSN 1548-7660. doi: 10.18637/jss.v036.i11. URL https://www.jstatsoft.org/v036/i11.

Y. Li and W. Dai. Bitcoin price forecasting method based on cnn-lstm hybrid neural network model. *The Journal of Engineering*, 2020(13):344–347, 2020. doi: 10.1049/joe.2019.1203.

S. Lipovetsky and M. Conklin. Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry*, 17(4):319–330, 2001. doi: https://doi.org/10.1002/asmb.446. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/asmb.446.

S. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions, May 2017a. URL https://arxiv.org/abs/1705.07874v1.

S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017b. URL https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017c. URL https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

M. Makowski. How the india crypto ban could impact investors, Mar 2021. URL https://investmentu.com/india-crypto-ban/.

S. Mangalathu, S.-H. Hwang, and J.-S. Jeon. Failure mode and effects analysis of rc members based on machine-learning-based shapley additive explanations (shap) approach. *Engineering Structures*, 219:110927, 2020. doi: 10.1016/j.engstruct.2020.110927.

S. McNally, J. Roche, and S. Caton. Predicting the price of bitcoin using machine learning. In *2018 26th Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP)*, pages 339–343, 2018. doi: 10.1109/PDP2018.2018.00060.

C. Mitcham. Encyclopedia of science, technology, and ethics, 2005. URL https://www.amazon.com/Encyclopedia-Technical-Market-Indicators-Second/dp/0070120579.

J. J. Murphy. *Technical analysis of the financial markets: a comprehensive guide to trading methods and applications*. New York Institute of Finance, 1999.

S. Nakamoto. Bitcoin: A peer-to-peer electronic cash system. https://bitcoin.org/bitcoin.pdf, 2009. URL https://bitcoin.org/bitcoin.pdf.

neptune.ai. Neptune: experiment management and collaboration tool, 2020. URL https://neptune.ai.

C. Olah. Understanding lstm networks, 2015. URL https://colah.github.io/posts/2015-08-Understanding-LSTMs/.

A. B. Parsa, A. Movahedi, H. Taghipour, S. Derrible, and A. K. Mohammadian. Toward safer highways, application of xgboost and shap for real-time accident detection and feature analysis. *Accident Analysis amp; Prevention*, 136:105405, 2020. doi: 10.1016/j.aap.2019.105405.

J. Patel, S. Shah, P. Thakkar, and K. Kotecha. Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. *Expert Systems with Applications*, 42(1):259–268, 2015. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2014.07.040. URL https://www.sciencedirect.com/science/article/pii/S0957417414004473.

L. Pedro, A. Ogbechie, J. Diaz-Rozo, D. A. Alonso, C. Bielza, and C. Puerto-Santana. *Industrial applications of machine learning*. CRC Press, 2020.

U. PlanB. Modeling bitcoin value with scarcity, Jul 2020. URL https://medium.com/@100trillionUSD/modeling-bitcoins-value-with-scarcity-91fa0fc03e25.

A. W. Rathgeber, J. Stadler, and S. Stöckl. The impact of the leverage effect on the implied volatility smile: evidence for the german option market. *Review of Derivatives Research*, Sept. 2020. doi: 10.1007/s11147-020-09171-3. URL https://doi.org/10.1007/s11147-020-09171-3.

M. T. Ribeiro, S. Singh, and C. Guestrin. "why should i trust you?". In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages =1135*. ACM, Aug. 2016. doi: 10.1145/2939672.2939778. URL https://doi.org/10.1145/2939672.2939778.

T. Seki, Y. Kawazoe, and K. Ohe. Machine learning-based prediction of in-hospital mortality using admission laboratory data: A retrospective, single-site study using electronic health record data. *PLOS ONE*, 16(2), 2021. doi: 10.1371/journal.pone.0246640.

P. K. Shah. An empirical study on options trading strategy using 'commodity channel index' for nse's nifty options in india. *SSRN Electronic Journal*, 2019. doi: 10.2139/ssrn.3323746.

R. Sharma. How cryptocurrency taxes affect bitcoin price, Sep 2020. URL https://www.investopedia.com/news/how-cryptocurrency-taxes-affect-bitcoin-price/.

D. Shen, A. Urquhart, and P. Wang. Does twitter predict bitcoin? *Economics Letters*, 174: 118–122, 2019. ISSN 0165-1765. doi: https://doi.org/10.1016/j.econlet.2018.11.007. URL https://www.sciencedirect.com/science/article/pii/S0165176518304634.

M. Sigalos. Yes, bitcoin could be the new gamestop, Feb 2021. URL https://www.cnbc.com/2021/02/01/how-bitcoin-could-be-the-new-gamestop.html.

D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 180–186, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450371100. doi: 10.1145/3375627.3375830. URL https://doi.org/10.1145/3375627.3375830.

Statista. Coinbase vs binance - active users, 2021a. URL https://www.bitdegree.org/crypto/coinbase-vs-binance.

Statista. Number of cryptocurrencies worldwide from 2013 to 2021, 2021b. URL https://www.statista.com/statistics/863917/number-crypto-coins-tokens/.

F. R. Velde. Bitcoin: a primer. *Chicago Fed Letter*, (Dec):2, 2013. URL https://ideas.repec.org/a/fip/fedhle/y2013idecn317.html.

Yahoo finance. Nearly 10% of the $380 billion in stimulus checks may be used to buy bitcoin and stocks: survey, 2021. URL https://finance.yahoo.com/news/nearly-10-of-the-380-billion-in-stimulus-checks-may-be-used-to-buy-bitcoin-and-stocks.html.

P. Zheng, Z. Zheng, J. Wu, and H.-N. Dai. Xblock-eth: Extracting and exploring blockchain data from ethereum. *IEEE Open Journal of the Computer Society*, 1:95–106, 2020. doi: 10.1109/OJCS.2020.2990458.

8 Figurliste

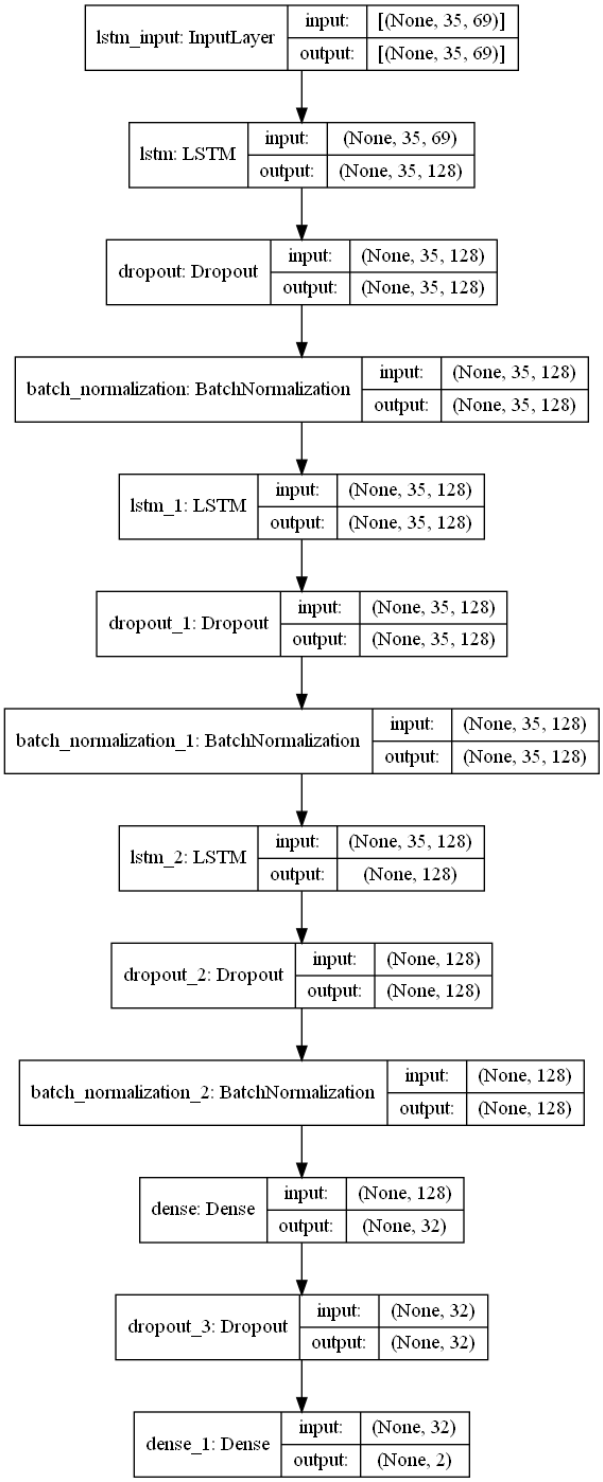
Figurer

9 Tabelliste

Tabeller

10 appendix

Figur 19: Arkitektur til modell I til VI



10.1 Tekniske indikatorer

Verdisetting her er basert på Murphy (1999). Omgjørelse av tekniske indikatorer til diskrete variabler. (I denne oppgaven blir stengepris beskrevet som C_t.)

- SMA: når stengeprisen er over SMA gir den verdien +1, når stengeprisen er under -1.
- EMA er gis det output 1, ved følgende viktige krysninger: ema_4, ema_9 og ema_18, EMA produserer også signal ved kryssing av hurtig EMA_{20} over treg EMA_{200} gir +1, -1 ved hurtig under treg.
- STCK%, STCD% og Williams R% er stochastiske oscillatorer, MACD er trendfølgende men følger samme metodikk. Dette gjør at dersom indikatorene er økende: verdien er høyere ved tiden t en $t-1$ representere det en stigende trend og output "+1", dersom den er lavere generes output 1".
- RSI er en indikator som brukes for å finne om et aktiva er overkjøpt: verdi over 70 gir output 1", dersom verdien er under 30 indikerer det at den er over-solgt gir verdien "+1", mellom 30-70 gir den "0".
- CCI fungerer likt som RSI, for denne har jeg valgt verdien +200 for overkjøpt og -200 for oversolgt.
- BB, dersom C_t er under lower band "+1", mellom lower band og upper band "0", over upper band 1".

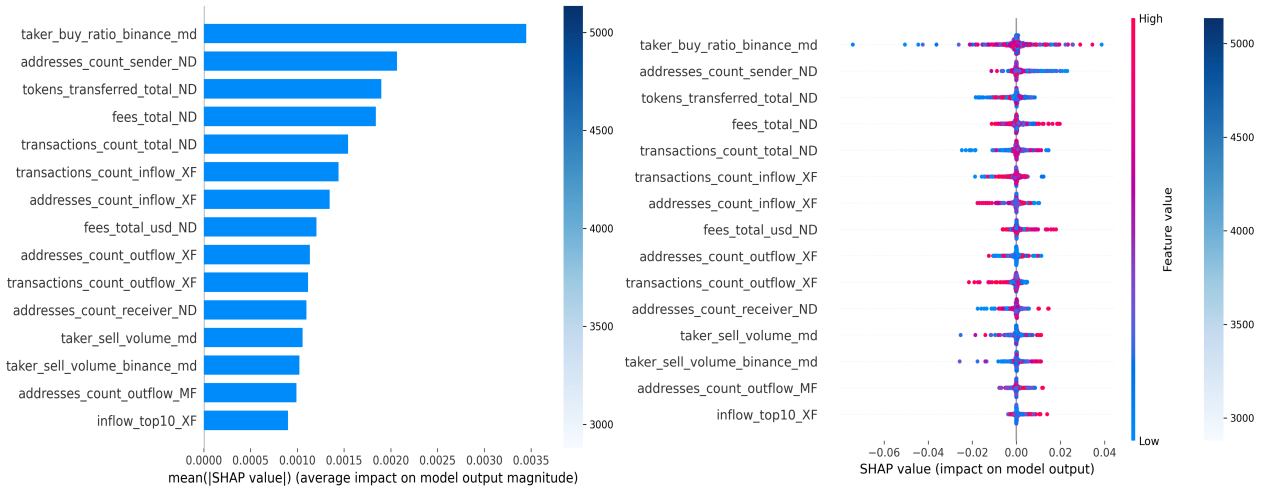
Teknisk indikator	Formel
Simple moving avarage	$C_t + C_{t-1} + ... + C_{-9}$
Weighted 10 day moving avarge	$\frac{(n)C_t+(n-1)C_{t-1}+...C_{10}}{(n+(n-1)+...+1)}$
Momentum	$C_t - C_{t-n}$
Stochastic K%	$\frac{C_t - LL_{t-n}}{HH_{t-n} - LL_{t-n}} \times 100$
Stochastic D%	$\frac{\sum_{i=0}^{n-1} K_{t-i}\%}{n}$
Relativ Strength Index (RSI)	$100 - \frac{100}{1+\sum_{i=0}^{n-1} \frac{Up_{t-i}}{n} / \sum_{i=0}^{n-1} \frac{Dw_{t-i}}{n}}$
MACD	$MACD(n)_{t-1} + \frac{2}{n} + 1 \times (DIFF_t - MACD(n)_{t-1}$
Larry Williams R%	$\frac{H_n - C_t}{H_n - L_n} \times 100$
Accumulation / Distribution	$\frac{H_t - C_{t-1}}{H_t - L_t}$
Commodity channel Index	$\frac{M_t - SM_t}{0.015 D_t}$
On-balance volume	$OBV_{prev} + \begin{cases} volume & if close > close_{prev} \\ 0 & if close = close_{prev} \\ -volume & if close < close_{prev} \end{cases}$
Bollinger Band	$\begin{cases} Middleband = & sma_{20} \\ Upperband & Middleband + 2StdDev \\ Lowerband & Middleband - 2StdDev \end{cases}$
BOP	SMA of [(Close - Open) / (High - Low)]

Tabell 8: Oversikt over de viktigste tekniske indikatorene som er brukt i oppgaven. Disse er konstruert ved å bruke modulen TA-lib: link. Det er flere som blir andvendt i oppgaven, for informasjon om disse se TA-lib.

10.2 Modell I, trent på samtlige markedsinformasjon variabler

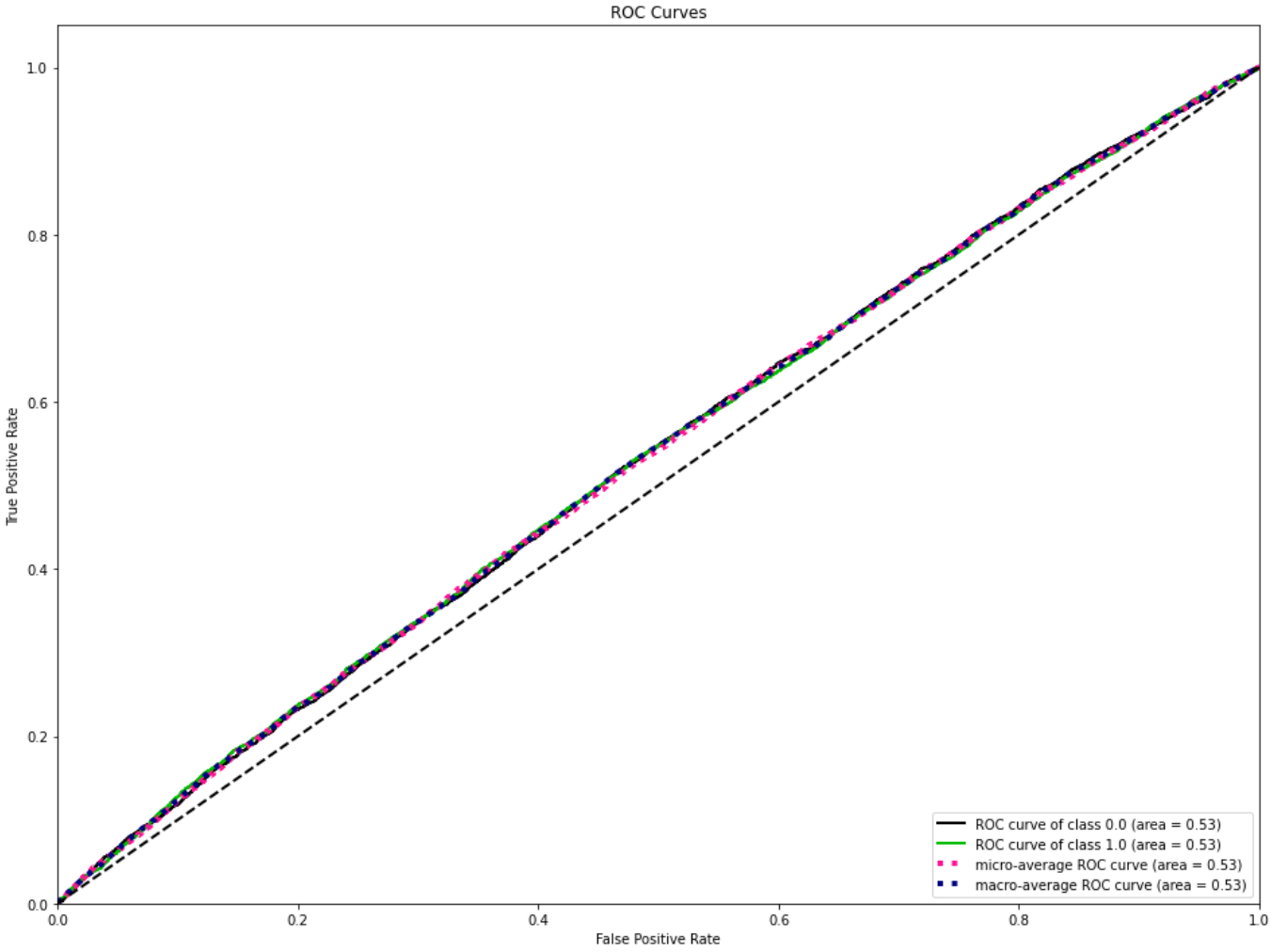
Test utført på inngangsverdier som anvist i tabell 9 på side 87 (med untakk av kolonnen med tekniske indikatorer)

Nøyaktighet: 52,26 %
Tap 0.725
Korrekt prediksjon av kursnedgang: 64.1 %
Korrekt prediksjon av kursoppgang: 40.5 %

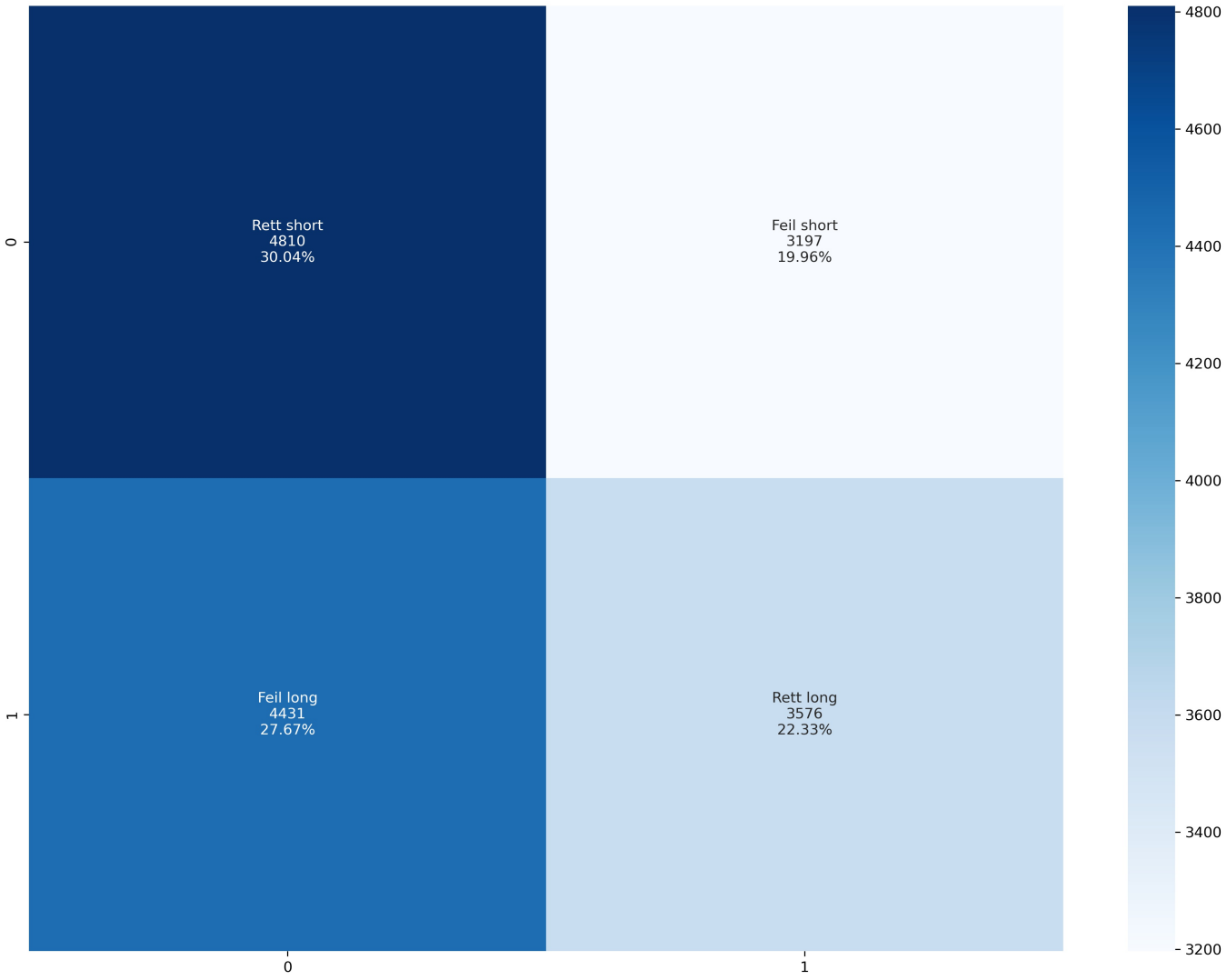


Figur 20: SHAP analyse av modell I, illustrerer de 15 variablene med høyet absolutt SHAP verdi.

Figur 21: ROC kurve av modell I



Figur 22: Confusion matrise av modell I



Direkte fra binance	Miner flyt	Markeds data	Markeds indikator	Flow indikator	Network i indikatorer	Bors flyt	Nettverks data	Tekniske indikatorer
BTCUSDT open	reserve	price usd close	estimated leverage ratio	mpi	stock to flow	reserve	supply total	EMA (5,12,20,50,100,200)
BTCUSDT high	reserve usd	taker buy volume md	estimated leverage ratio binance	exchange whale ratio	stock to flow reversion	reserve usd	supply new	SMA(5, 12, 20,50, 100,200)
BTCUSDT low	inflow total	taker sell volume md	mvrv	fund flow ratio	nvt	netflow total	transactions count total	Bollinger Bands (up, mid, low)
BTCUSDT close	netflow total	taker sell ratio binance md		stablecoins ratio	nvt golden cross	transactions count inflow	transactions count mean	MACD
BTCUSDT volume	inflow top10	taker buy sell ratio binance md		stablecoins ratio usd	nvts	transactions count outflow	addresses count active	MACD signal
	inflow mean	funding rates md			puell multiple	inflow total	addresses count new	MACD histogram
	outflow total	taker buy volume binance md				inflow top10	addresses count receiver	RSI
	outflow top10	taker sell volume binance md				inflow mean	tokens transferred total	CCI
	outflow mean	funding rates binance md				outflow total	tokens transferred mean	MFI
	addresses count inflow	long liquidations md				outflow top10	tokens transferred median	OBV
	addresses count outflow	taker buy ratio md				outflow mean	block bytes	WILLR
		taker sell ratio md				addresses count inflow	block interval	RSI fast
		taker buy sell ratio md				addresses count outflow	utxo count	RSI slow
		market cap md					fees block mean	ULTOSC
		realized cap md					fees block mean usd	CMO
		average cap md					fees total	BOP
		delta cap md					fees total usd	ADOSC
		thermo cap md					fees reward percent	ATR
		coinbase pre um gap md					fees transaction mean	TRIME
		coinbase pre um 1 es md					fees transaction mean usd	KAMA
		price usd close md					fees transaction median	BTCUSDT MACD sign long
		price usd high md					fees transaction median usd	BTCUSDT MACD sign short
		price usd low md					blockreward	BTCUSDT CCI signal long
		open interest md					blockreward usd	BTCUSDT CCI signal short
		open interest binance md					difficulty	BTCUSDT MFI signal long
		short liquidations md					hashrate	BTCUSDT MFI signal short
		long liquidations usd md					block count	BTCUSDT MFI sell pressure
		short liquidations usd md					velocity supply total	BTCUSDT MFI buy pressure
								BTCUSDT Signal medium
								BTCUSDT Signal long
								BTCUSDT Signal ema price over EMA50
								BTCUSDT Signal RSI bear
								BTCUSDT Signal RSI bull
5	11	28	3	5	6	13	28	47

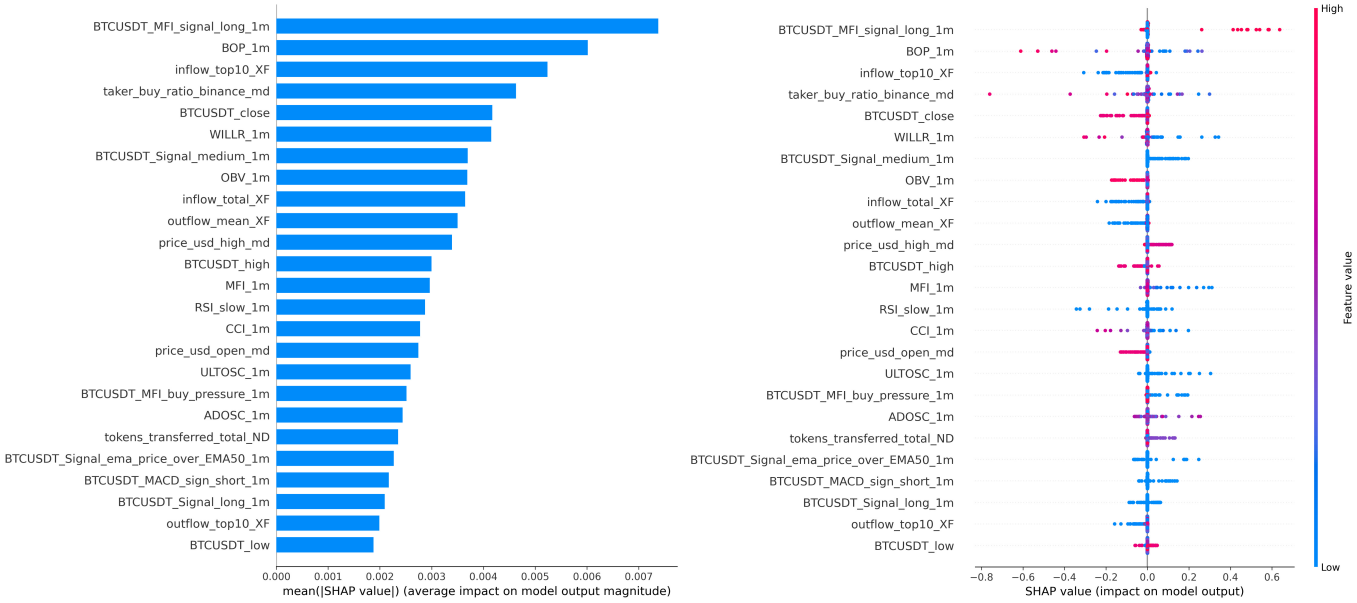
Tabell 9: Samtlige variabler benyttet i oppgaven

10.3 Modell II : Samtlige inngangsverdier

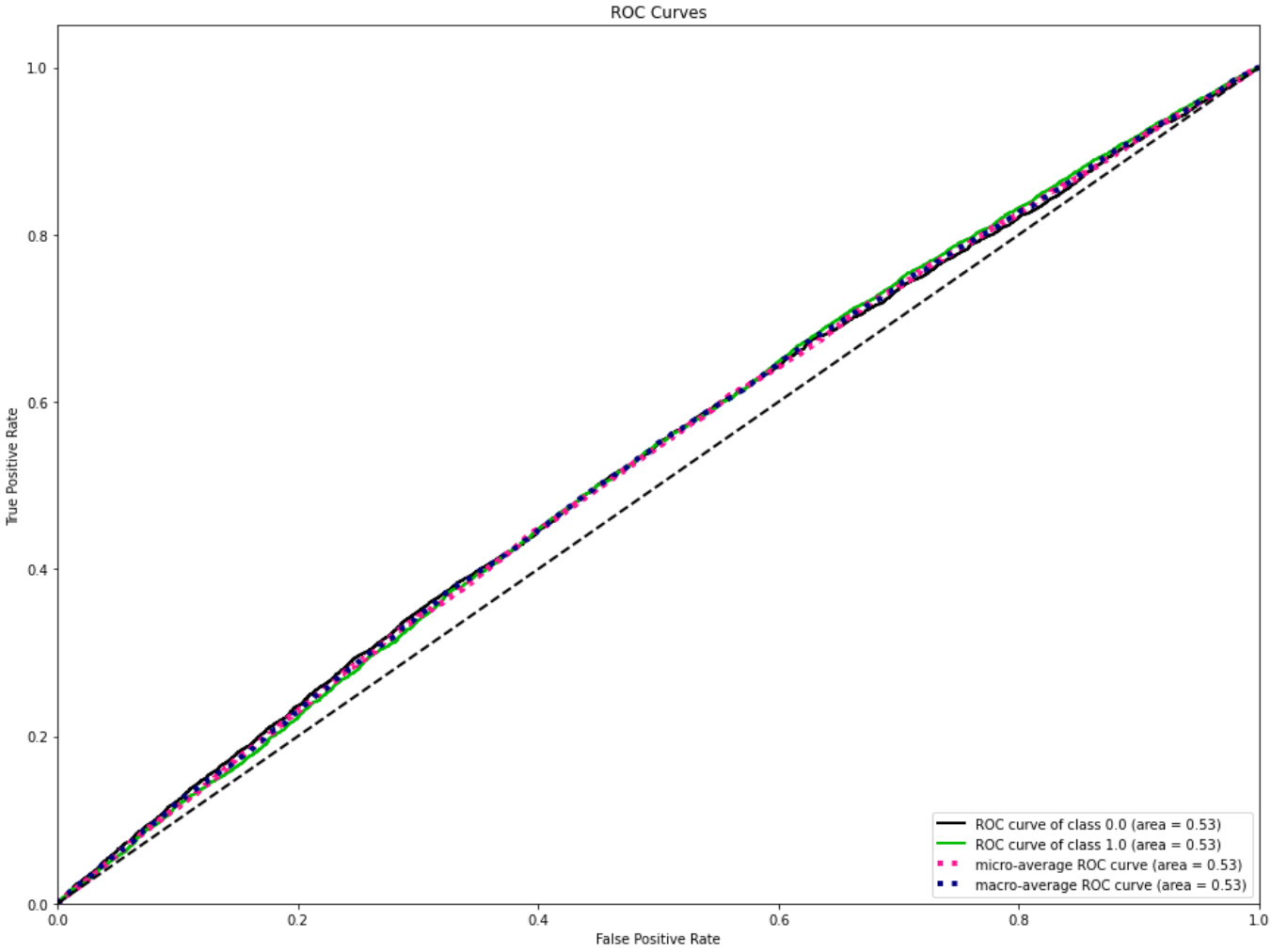
Test utført på inngangsverdier som anvist i tabell 9 på side 87

Nøyaktighet: 52,36 %
Tap: 0.725
Korrekt prediksjon av kursnedgang: 60.0 %
Korrekt prediksjon av kursoppgang: 44.66 %

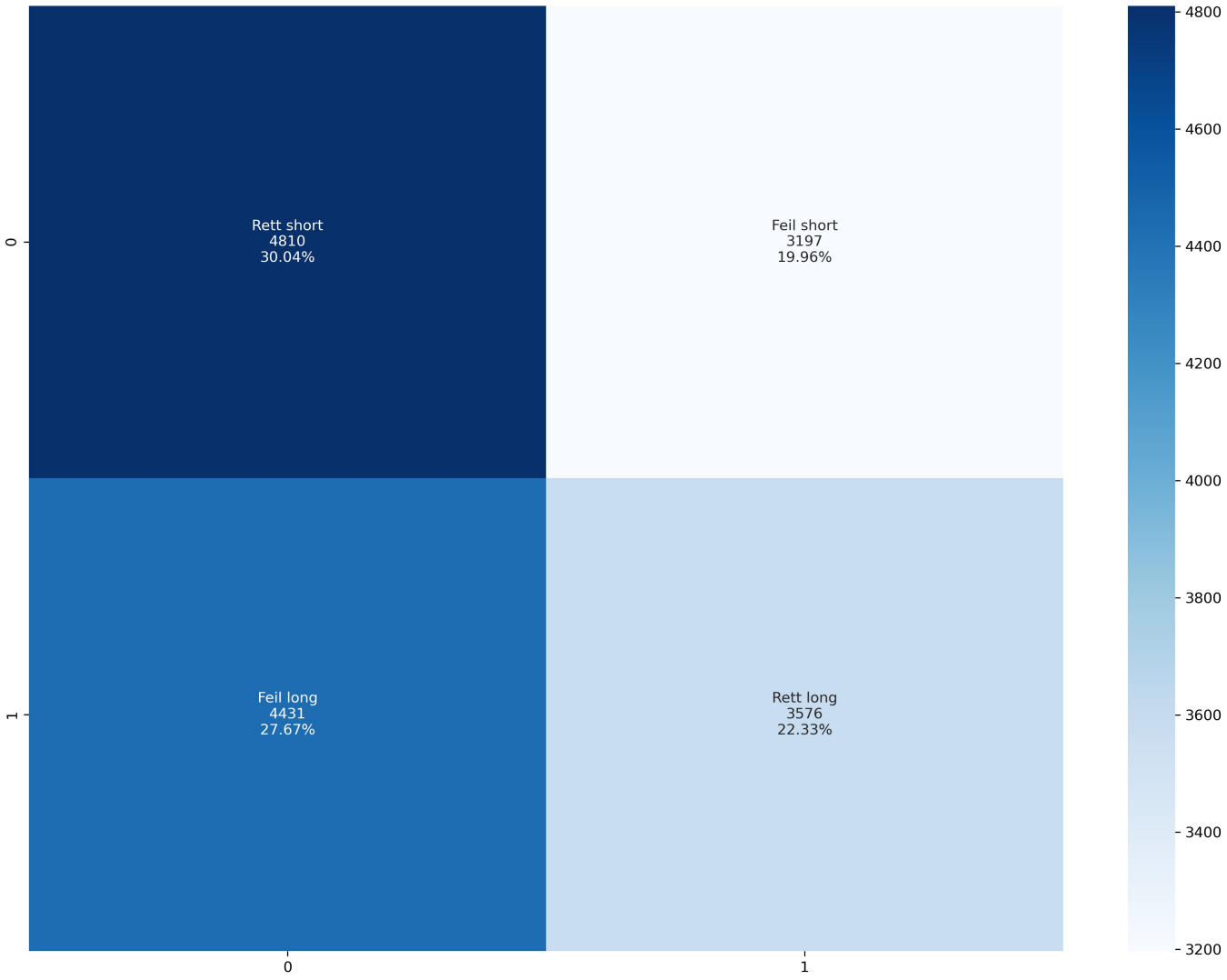
Figur 23: SHAP analyse av modell II



Figur 24: ROC kurve for modell II



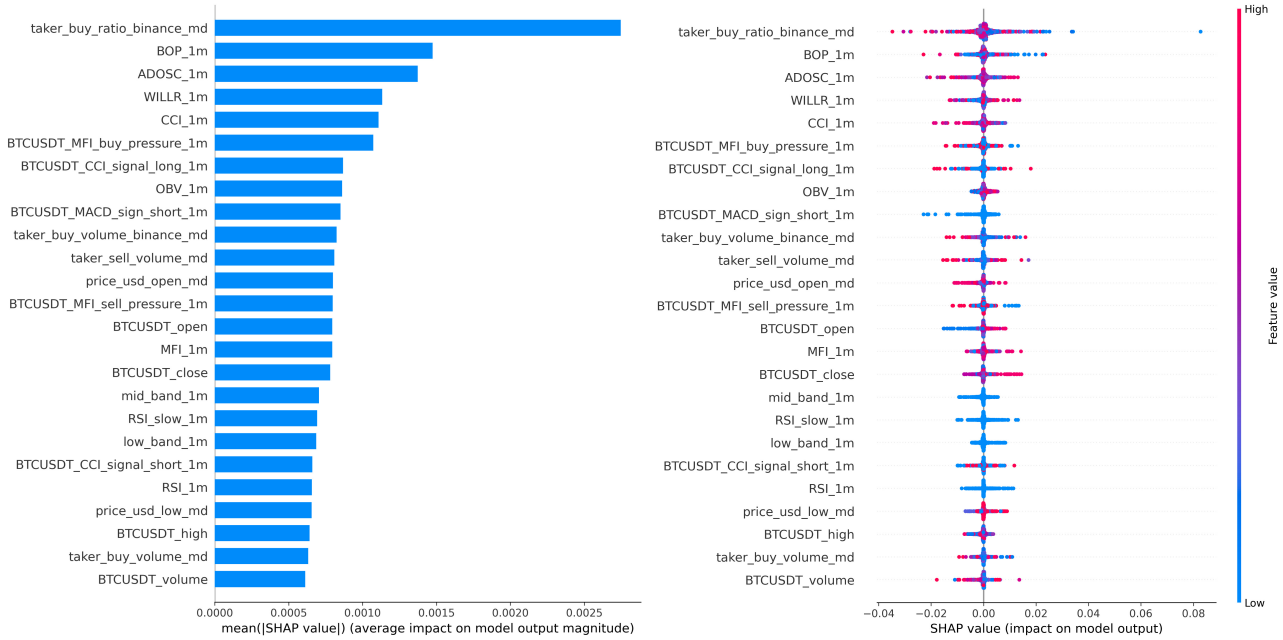
Figur 25: Forvirringsmatrise for modell II



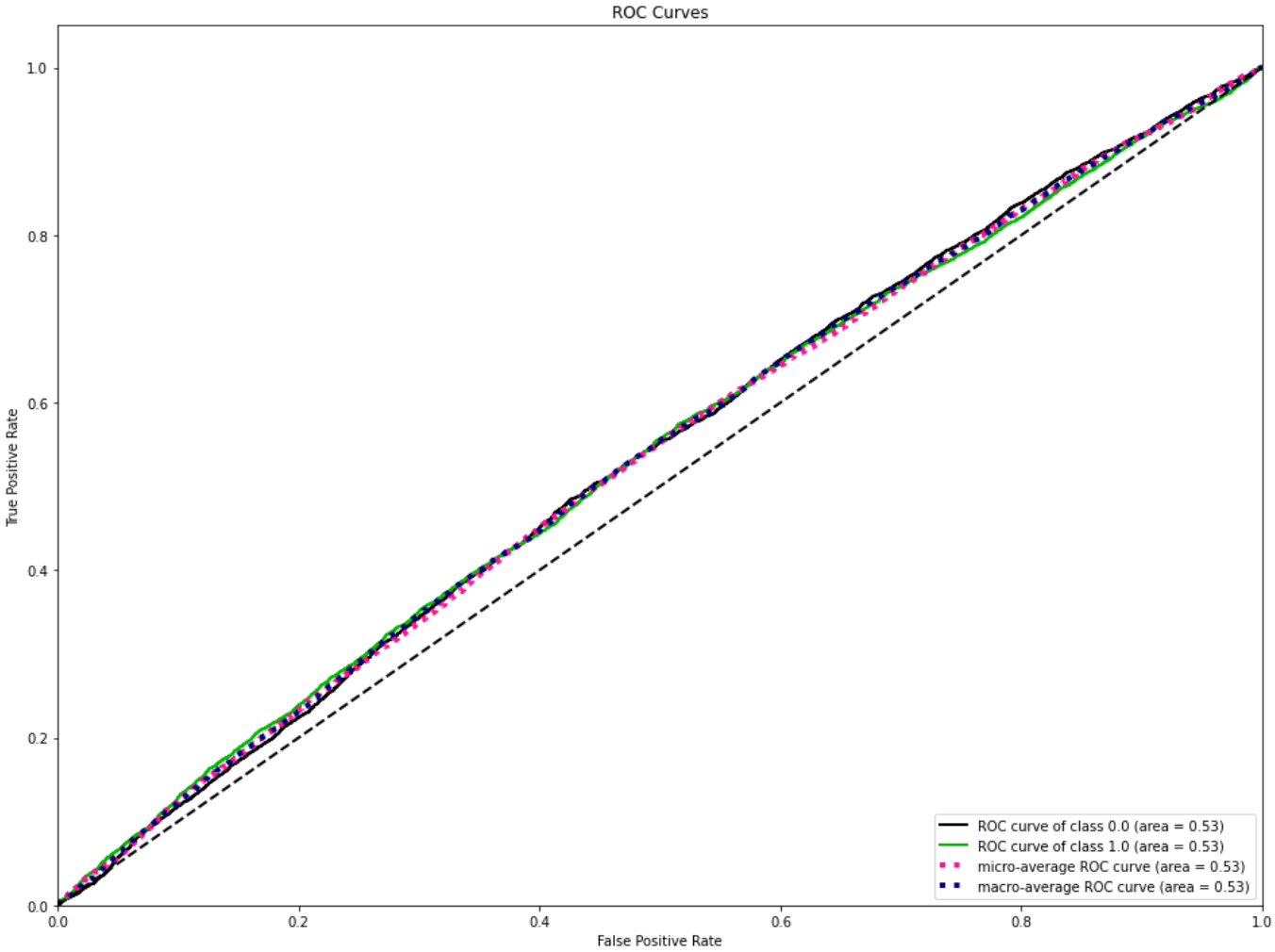
10.4 Modell III: Boruta utvalgte inngangsverdier, markedsinformasjon og tekniske indikatorer

Test utført på inngangsverdier som anvist i tabell på side 94

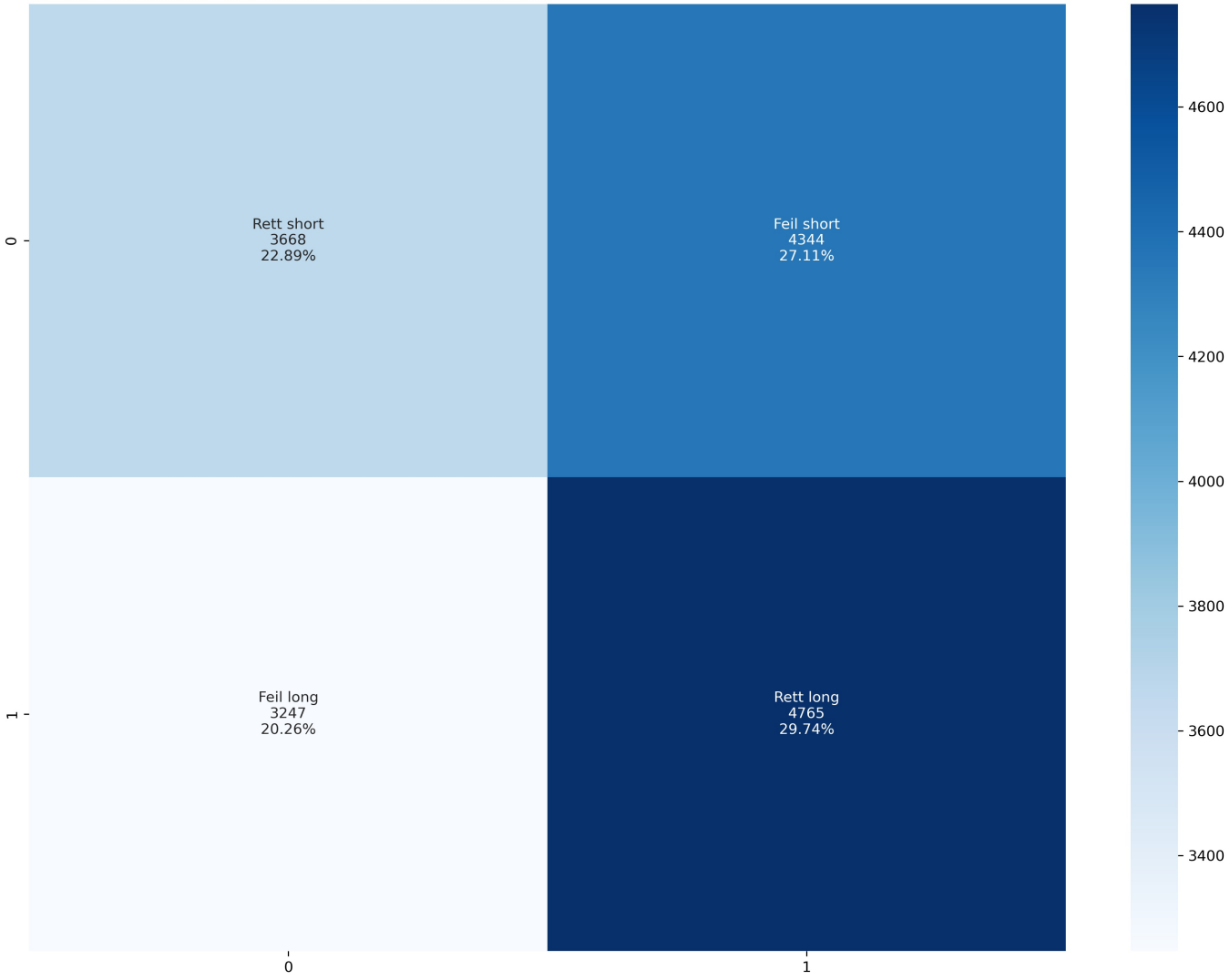
Nøyaktighet: 52,6 %
Tap: 0.705
Korrekt prediksjon av kursnedgang: 45.8 %
Korrekt prediksjon av kursoppgang: 59.5 %



Figur 26: SHAP analyse av modell III



Figur 27: ROC kurve av modell III



Figur 28: Forvirringsmatrise av av modell III

Variabler for modell III	
BTCUSDT high	RSI
BTCUSDT low	CCI
BTCUSDT close	MFI
BTCUSDT volume	OBV
Taker buy volume	WILLR
Taker sell volume	RSI slow
Taker buy ratio binance	ULTOSC
Taker sell ratio binance	CMO
Taker buy sell ratio binance	BOP
Funding rates	ADOSC
Taker buy volume binance	ATR
Taker sell volume binance	TRIME
Tunding rates binance	KAMA
Long liquidations	BTCUSDT MACD sign short
Taker buy ratio	BTCUSDT CCI signal long
Taker sell ratio	BTCUSDT CCI signal short
Taker buy sell ratio	BTCUSDT MFI signal long
Coinbase premium gap	BTCUSDT MFI signal short
Coinbase premium index	BTCUSDT MFI sell pressure
Price usd close	BTCUSDT MFI buy pressure
Price usd open	Open interest binance
Price usd high	Short liquidations
Price usd low	Long liquidations usd
Open interest	Short liquidations usd
MACD	Boolinger band - up band
MACD signal	Bollinger band - mid band
MACD histogram	Bollinger band - low band

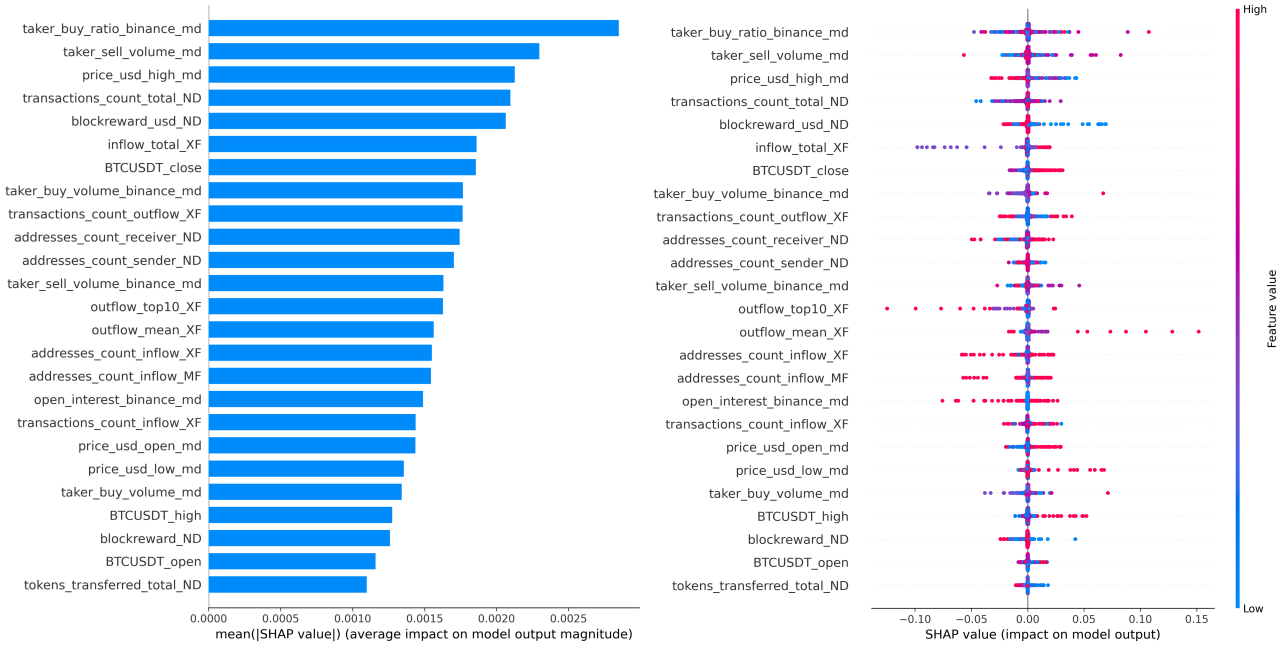
Direkte fra binance	5
markedsinformasjon	24
Teknisk analyse	26
Sum	55

10.5 Modell IV: Boruta utvalgte inngangsverdier kun markedsinformasjon

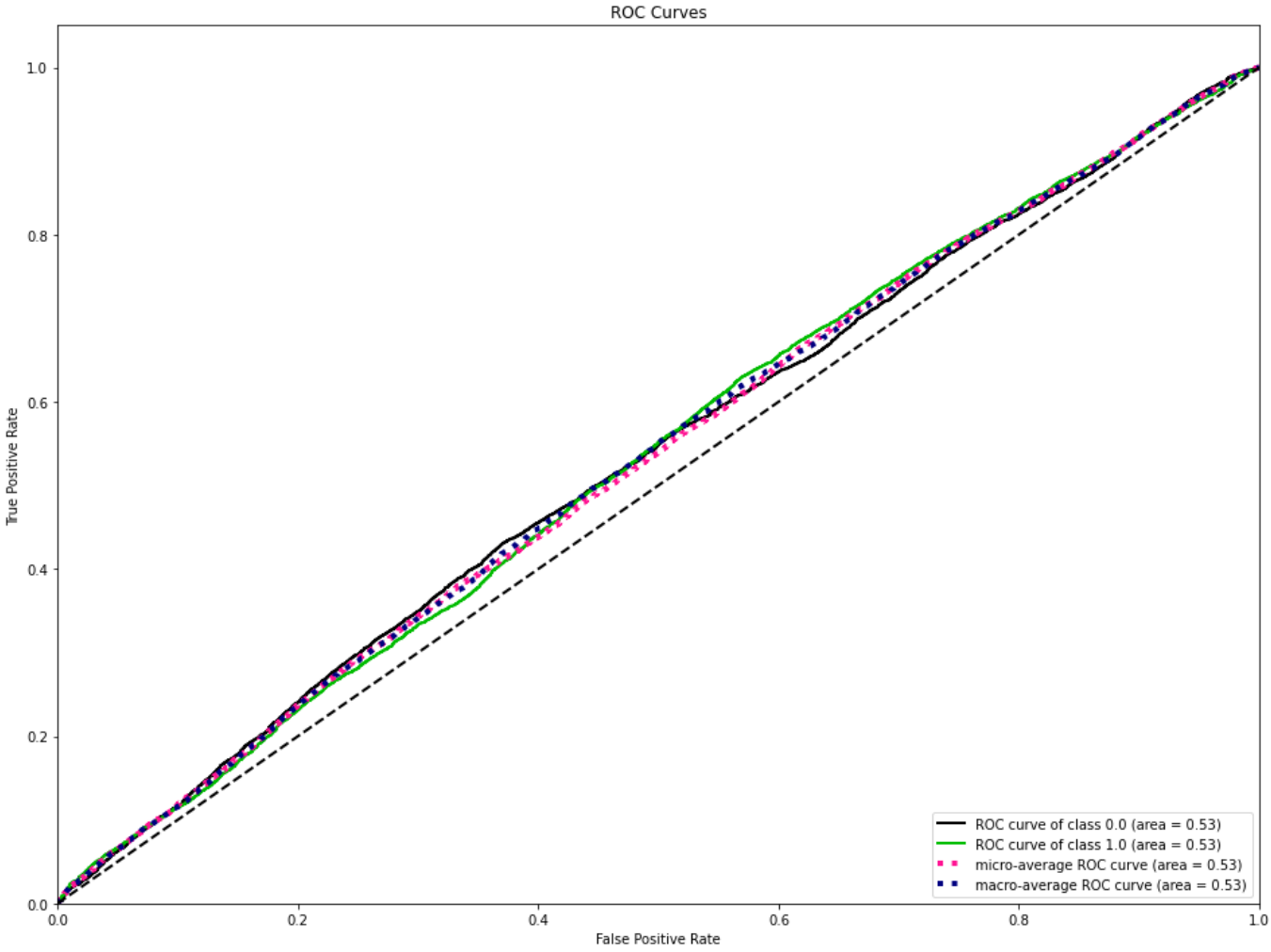
Test utført på inngangsverdier som anvist i tabell på side 98

Nøyaktighet: 52,1 %
Tap: 0.705
Korrekt prediksjon av kursnedgang: 60.8 %
Korrekt prediksjon av kursoppgang: 41.3 %

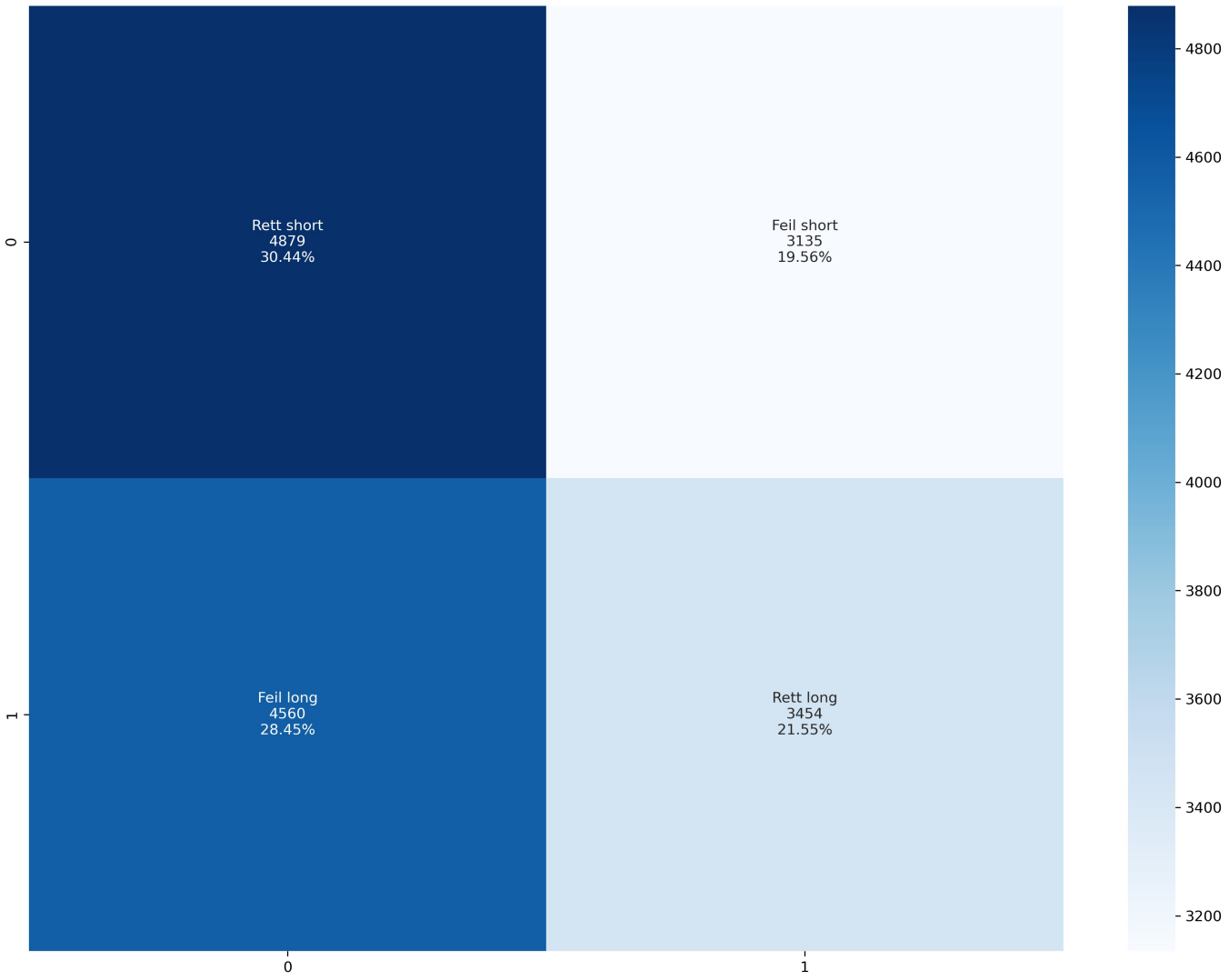
Figur 29: SHAP analyse for Boruta utvalgte inngangsverdier, markedsinformasjon



Figur 30: ROC kurve av modell IV



Figur 31: Forvirringsmatrise av av modell IV



Tabell 10: Variabler brukt i modell IV

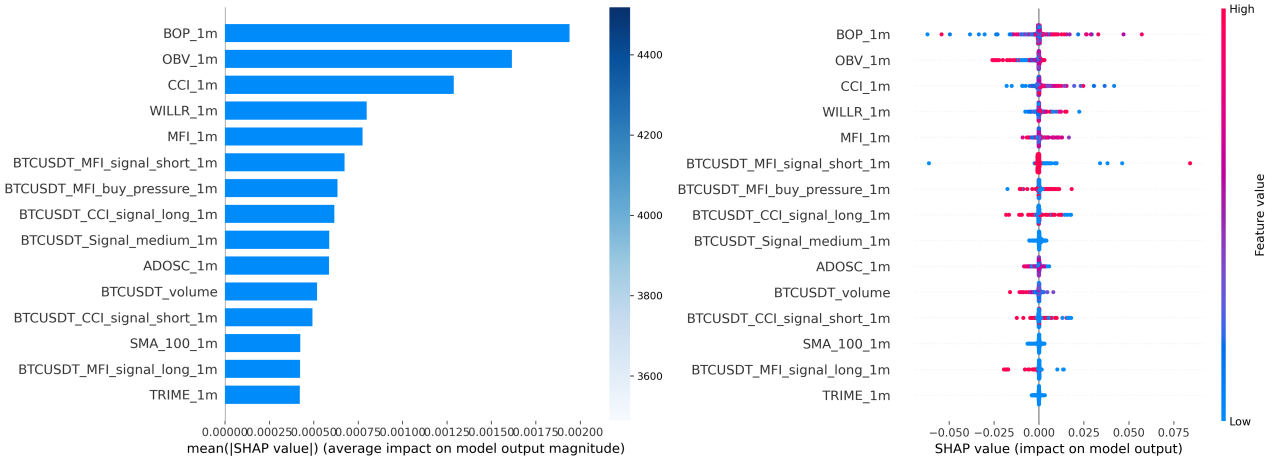
Boruta utvalgte inngangsverdier markedsinformasjon	
BTCUSDT high	Taker buy ratio
BTCUSDT low	Taker sell ratio
BTCUSDT close	Taker buy sell ratio
BTCUSDT volume	Coinbase premium gap
Taker buy volume	Coinbase premium index
Taker sell volume	Exchange whale ratio
Taker buy ratio binance	Price usd close
Inflow total MF	Open interest
Inflow top10 MF	Open interest binance
Inflow mean MF	Short liquidations
Addresses count inflow MF	Long liquidations usd
Taker sell ratio Binance	Short liquidations usd
Taker buy sell ratio Binance	Transactions count inflow
Funding rates	Transactions count outflow
Taker buy volume Binance	Inflow total
Taker sell volume Binance	Inflow top10
Funding rates binance	Outflow top10
Long liquidations	Outflow mean
Tokens transferred total	Addresses count inflow
Fees total	Transactions count total
Fees total usd	Addresses count sender
Fees transaction mean	Addresses count receiver
Blockreward	Blockreward usd

Direkte fra binance	5
markedsinformasjon	42
Teknisk analyse	0
Sum	47

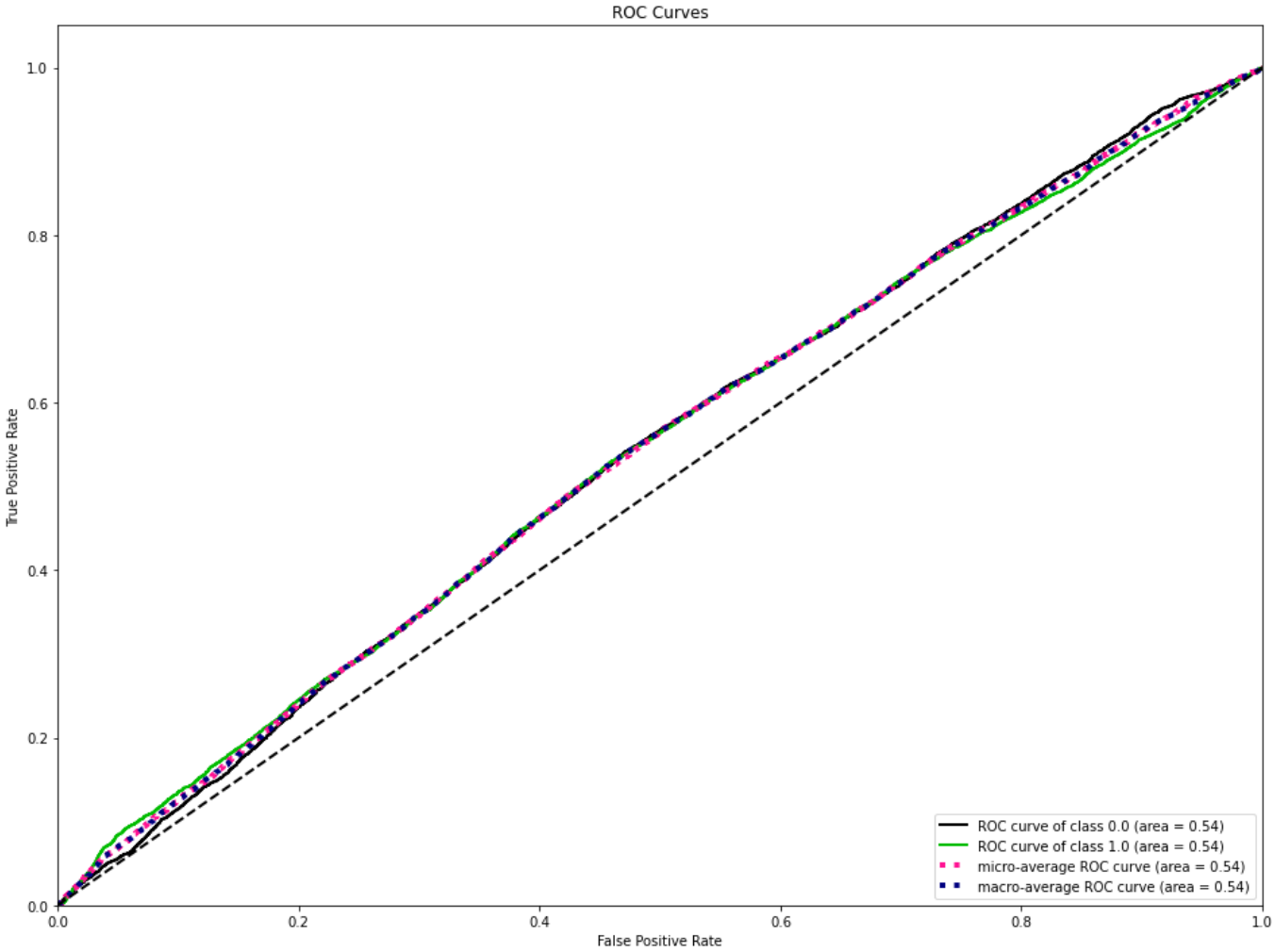
10.6 Modell V : Samtlige tekniske indikatorer

Test utført på inngangsverdier som anvist i tabell 11 på side 102

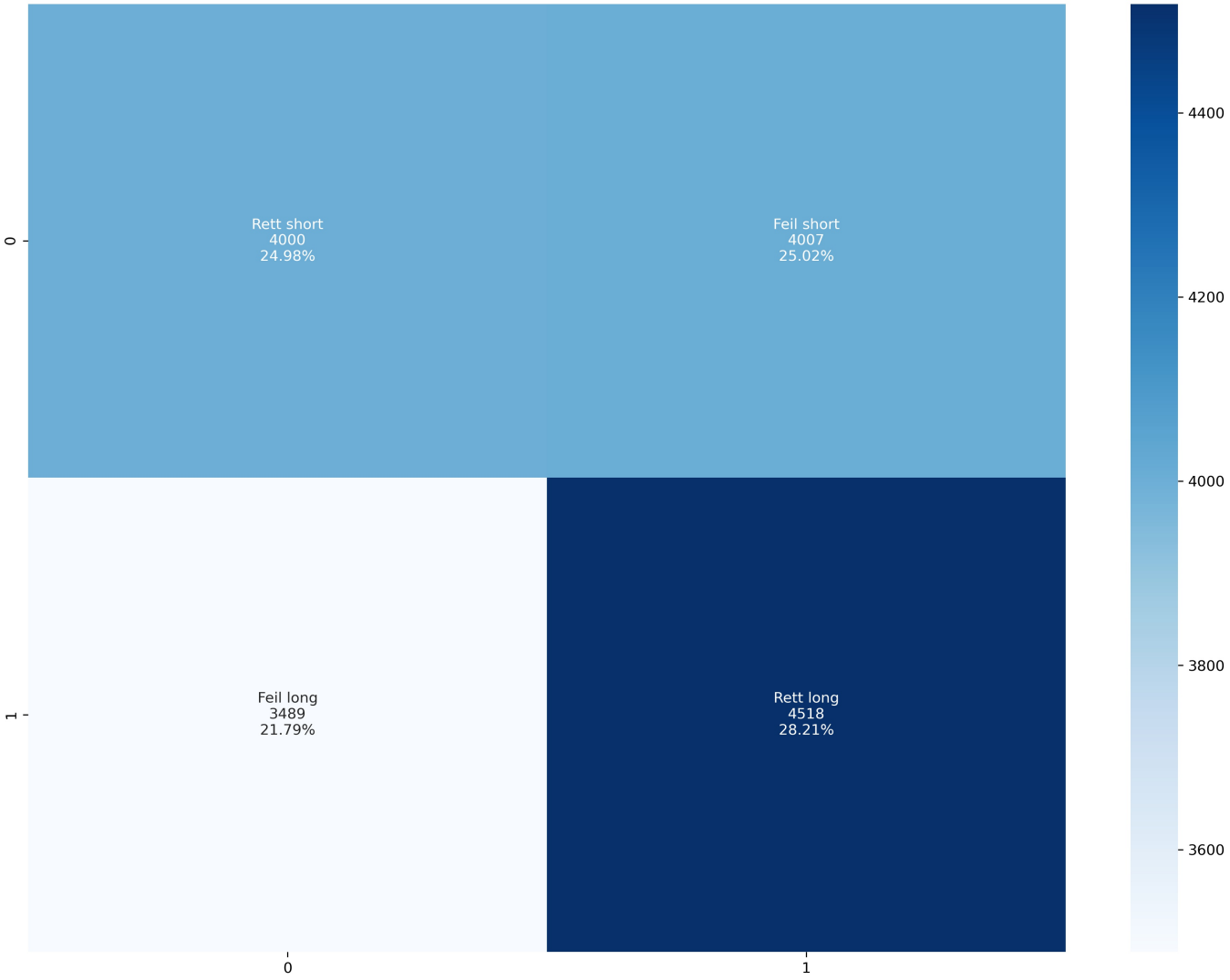
Nøyaktighet: 53,2 %
Tap: 0.694
Korrekt prediksjon av kursnedgang: 50.0 %
Korrekt prediksjon av kursoppgang: 56.4 %



Figur 32: SHAP analyse av modell V



Figur 33: ROC kurve av modell V



Figur 34: Forvirringsmatrise av modell V

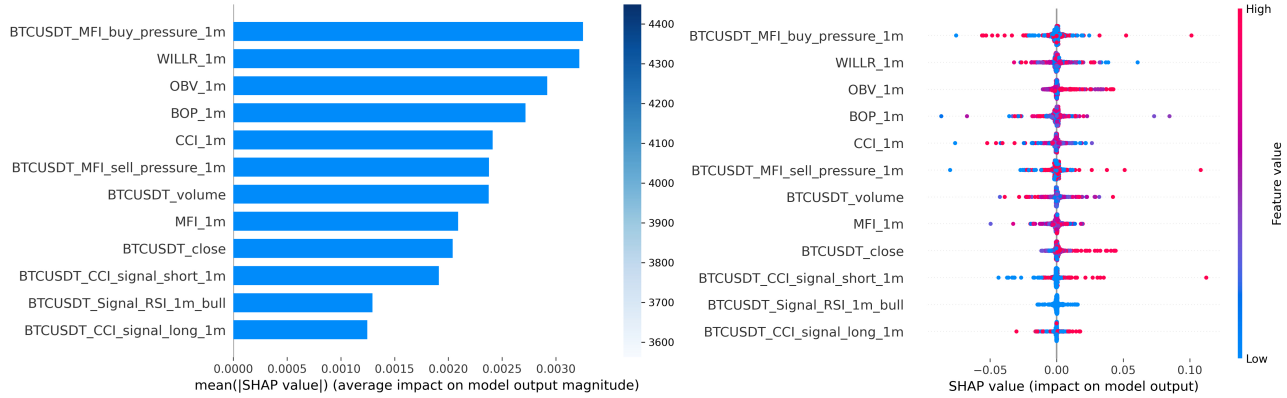
Tabell 11: Variabler til modell V

Samtlige tekniske indikatorer	
EMA (5,12,20,50,100,200)	ULTOSC
SMA(5, 12, 20,50, 100,200)	CMO
Bollinger Bands (up, mid, low)	BOP
MACD	ADOSC
MACD signal	ATR
MACD histogram	TRIME
RSI	KAMA
CCI	BTCUSDT MACD sign long
MFI	BTCUSDT MACD sign short
OBV	BTCUSDT CCI signal long
WILLR	BTCUSDT CCI signal short
RSI fast	BTCUSDT MFI signal long
RSI slow	BTCUSDT MFI signal short
BTCUSDT Signal long	BTCUSDT MFI sell pressure
BTCUSDT Signal ema price over EMA50	BTCUSDT MFI buy pressure
BTCUSDT Signal RSI bear	BTCUSDT Signal medium
BTCUSDT Signal RSI bull	
Antall tekniske indikatorer; 47	

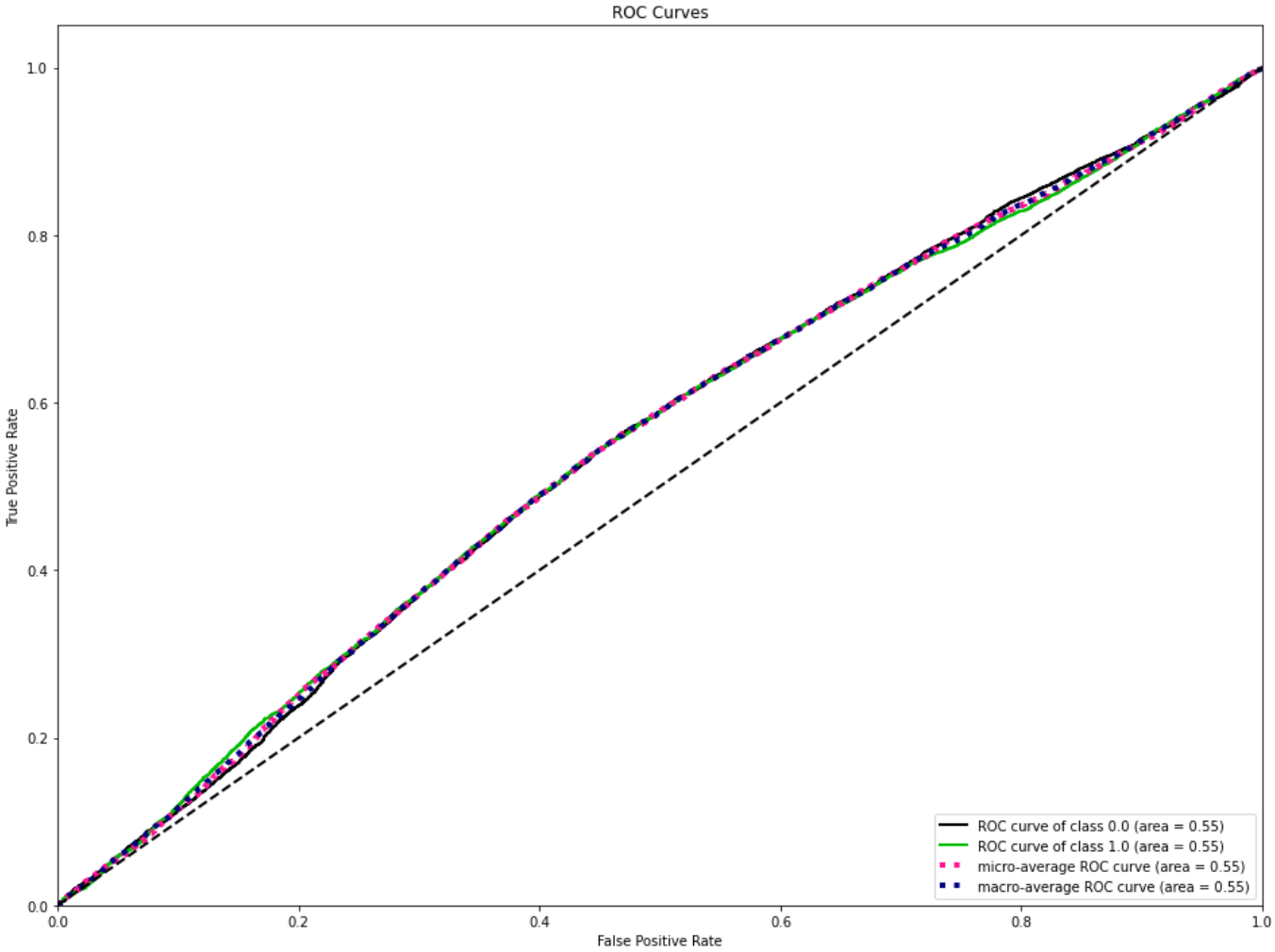
10.7 Modell VI: Shap filtrerte tekniske indikatorer

Test utført på inngangsverdier som anvist i tabell 12 på side 106

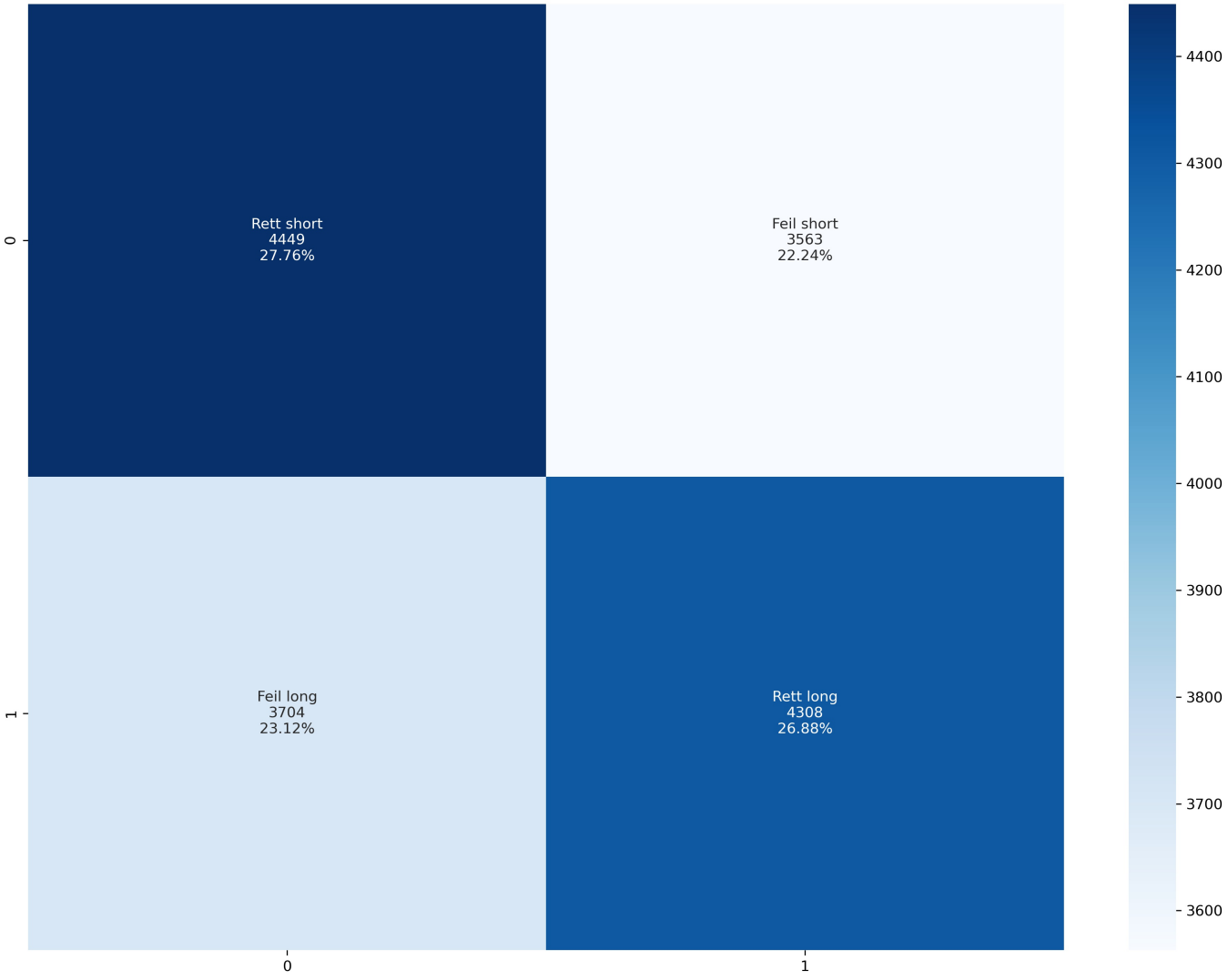
Nøyaktighet: 54,6 %
Tap: 0.705
Korrekt prediksjon av kursnedgang: 55,5 %
Korrekt prediksjon av kursoppgang: 53,76 %



Figur 35: SHAP analyse av modell VI



Figur 36: ROC kurve av modell VI



Figur 37: Forvirringsmatrise av modell VI

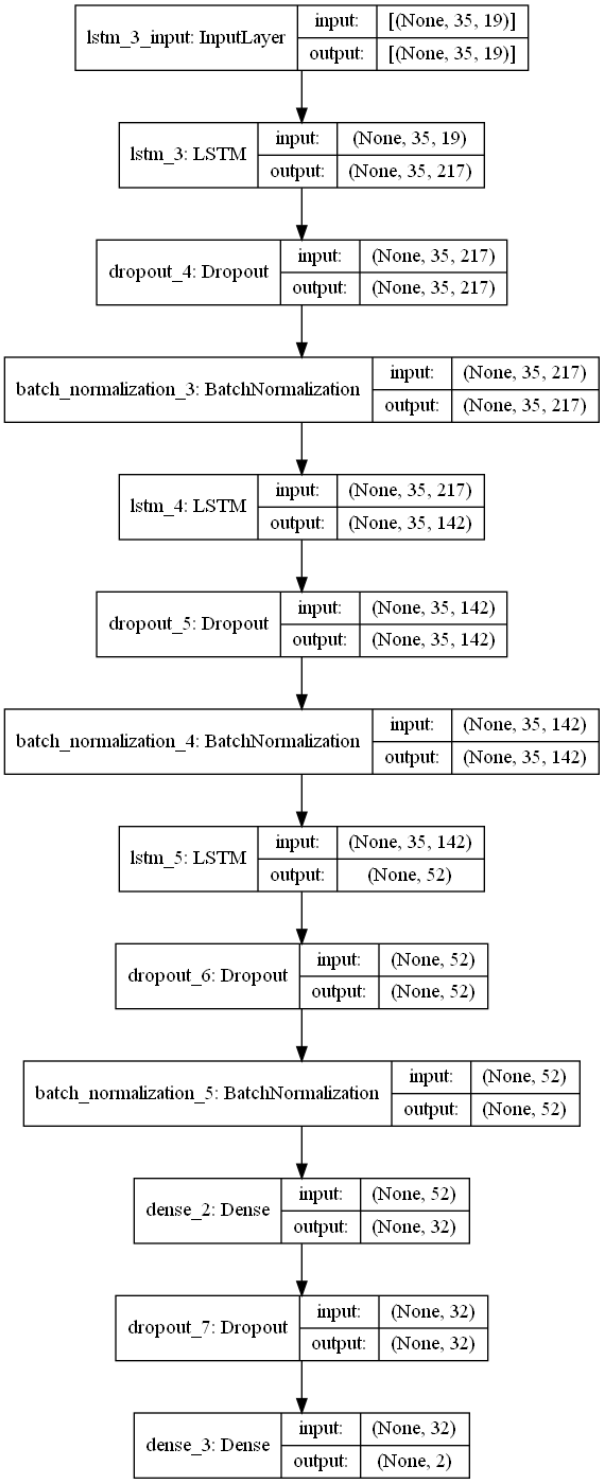
Tabell 12: Variabler brukt i modell VI

Shap filtrerte tekniske indikatorer
BOP
WILLR
BTCUSDT MFI SELL PRESSURE
BTCUSDT MFI BUY PRESSURE
BTCUSDT CCI signal short
BTCUSDT CCI signal long
MFI
BTCUSDT volume
CCI
OBV
BTCUSDT Signal RSI bull
BTCUSDT Close

Tekniske indikatorer	10
Direkte fra binance	2

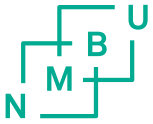
10.8 Modell VII: Kvalitativ utvalgte variabler

Figur 38: LSTM arkitektur og hyperparametere for modell 7



Figur 39: Oversikt over modell VII sin struktur

```
Test loss: 0.6880333602405216
Test accuracy: 0.53960526
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm_3 (LSTM)                (None, 35, 217)           205716
_________________________________________________________________
dropout_4 (Dropout)          (None, 35, 217)           0
_________________________________________________________________
batch_normalization_3 (Batch (None, 35, 217)           868
_________________________________________________________________
lstm_4 (LSTM)                (None, 35, 142)           204480
_________________________________________________________________
dropout_5 (Dropout)          (None, 35, 142)           0
_________________________________________________________________
batch_normalization_4 (Batch (None, 35, 142)           568
_________________________________________________________________
lstm_5 (LSTM)                (None, 52)                40560
_________________________________________________________________
dropout_6 (Dropout)          (None, 52)                0
_________________________________________________________________
batch_normalization_5 (Batch (None, 52)                208
_________________________________________________________________
dense_2 (Dense)              (None, 32)                1696
_________________________________________________________________
dropout_7 (Dropout)          (None, 32)                0
_________________________________________________________________
dense_3 (Dense)              (None, 2)                 66
=================================================================
Total params: 454,162
Trainable params: 453,340
Non-trainable params: 822
_________________________________________________________________
```



Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway