CASE: VIBORG FODSPORTS FORENING

FORUDSIGELSE AF UDNYTTELSESGRAD OG DATAANALYSE:

EN STRATEGISK DATAMODEL FOR VFF

UDARBEJDET AF:

Gruppe 1

André Sittrup Olesen
Anja Leth Jensen
Camilla Agnes Sparre Andersen
Patrick Schumann

1. semester interne eksamensprojekt

Vejleder: Lotte Soelberg Kronbæk
Bjarne Taulo Sørensen

Anslag: 26.793
6. januar 2025

ERHVERVSAKADEMI DANIA

1. Resumé

Formålet med dette projekt er at undersøge Viborg Fodsports Forening's (VFF) udfordringer med at forudsige fremmødet af VIP-gæster med guldmenu-billetter. Uoverensstemmelser mellem bestilte og benyttede billetter medfører ineffektiv ressourceanvendelse, madspild og forringede gæste/netværks-oplevelser, hvilket strider imod organisationens mission om at skabe unikke oplevelser og udnytte deres grønne potentiale.

Dataindsamlingen omfattede en Excel-fil med interne data fra VFF, eksternt indhentede vejrdata fra DMI og kampinformationer fra Superstats. Disse blev suppleret med semistrukturerede interviews med fire medarbejdere og tre praktikanter fra organisationen, som repræsenterede flere afdelinger. Interviewene gav indsigt i interne organisatoriske datarelaterede udfordringer såsom silo-dannelser og manglende dataintegration.

I projektet blev CRISP-DM anvendt som ramme for analyseprocessen. R-studio blev brugt til modellering og visualisering i undersøgelsen for forudsigelse af udnyttelsesgraden, hvor Random Forest-modellen viste højeste præcision blandt flere testede prædiktionsmodeller. Vejrforhold og kampkontekst blev identificeret som afgørende faktorer for fremmødet, men begrænsede datamængder reducerede modellernes anvendelighed og præcision. For at optimere modellerne og fremmødeanalysen anbefales VFF at udvide dataindsamlingen med detaljerede kundedata, herunder adfærds- og demografiske oplysninger. Dette vil styrke VFF's grundlag for at træffe mere informerede og datadrevne beslutninger.
Derudover foreslås organisatoriske forbedringer, såsom implementering af Kotter's 8-trins model for at reducere silo-dannelser samt fremme en fælles datakultur og derved opnå et højere datamodenhedsniveau i organisationen. På længere sigt bør VFF realiserer visionen om at etablere et Datawarehouse for at sikre bedre integration og anvendelse af data.

Projektets resultater fremhæver vigtigheden af at kombinere organisatoriske forbedringer med tekniske løsninger for at realisere VFF's mål om effektiv ressourceudnyttelse og optimerede gæsteoplevelser. Når de organisatoriske ændringer er implementeret, vil VFF være bedre rustet til at integrere tekniske løsninger og maksimere værdien af deres data.

Indholdsfortegnelse

2. Problemstilling

VFF har samarbejdsaftaler, hvor partnere (VIP-gæster) tildeles guldmenu-billetter til Vesttribunen ved hjemmebanekampe (Interview 1, 2024). VFF oplever i den forbindelse uoverensstemmelse mellem antal bestilte guldmenu-billetter og antal fremmødte VIP-gæster med disse billetter (1. semesterprøven, 2024).

Dette skaber problemer for flere interessenter (Interview 1, 2024), f.eks. madleverandører, køkkenpersonale og Eventii.

For disse interessenter medfører de uforudsigelige fremmøde-mønstre madspild, forkert allokering af personale og uudnyttede VIP-faciliteter på Vesttribunen.

Baggrunden for problematikken vedr. data-anvendelse og udnyttelse af tilgængelige data, kan sandsynligvis tilskrives organisationens infrastruktur og datamodenhedsniveau (Respondenter VFF, 2024). Der er på nuværende tidspunkt manglende integration af data på tværs af systemer i de forskellige afdelinger i organisationen samtidig med at parternes billet-brug varierer, hvilket medfører en reduceret stadionoplevelse for fremmødte, hvor den samlede stemning og atmosfære forringes (Respondenter VFF, 2024). Dette er problematisk, da VFF's mission er at skabe fællesskaber og oplevelser blandt de bedste i Danmark, og visionen er at forene mennesker for at realisere det grønne potentiale (Tea Nørgaard Marketing- og Kommunikationschef, 2024).

Derudover medfører det varierende billet-brug, økonomisk spild samt ineffektiv ressourceallokering på kampdage (Interview 1, 2024).

Organisationen ser ud til at mangle en model eller et system, der kan analysere data og forudsige udnyttelsesgraden på Vesttribunen under VIP-afsnittet. Uden en sådan løsning spildes ressourcer, og det bliver sværere at træffe informerede beslutninger, der kan sikre en sammenhængende og optimeret kampdags-oplevelse.

3. Problemformulering

"Hvordan kan VFF anvende eksisterende data om bestilte, benyttede og maksimale guld menu-billetter til at forudsige udnyttelsesgraden blandt VIP-gæster? Og hvordan kan organisationen samtidig fremme sin data-modenhed og infrastruktur, og hvilke konkrete initiativer kan understøtte integrationen af modellen i praksis?"

4. Afgrænsning

Data om fremmøde på Energi Viborg Arena er afgrænset til VIP-afsnittet på Vesttribunen og gælder kun guldmenu-billetter udleveret gennem samarbejdsaftaler. Datasættet udelukker perioden under corona-krisen og fokuserer på kampe i de sæsoner, hvor VFF har spillet i Superligaen.
Vi antager desuden, at VFF placerer sig jævnt i Superligaen.

4.1. Industri

I dette projekt betragtes VFF som en del af oplevelsesindustrien, hvor organisationen konkurrerer om publikums tid og opmærksomhed med lokale aktiviteter, der finder sted samtidig med hjemmebanekampene. Dette på baggrund af VFF's mission om at skabe fællesskaber og oplevelser, der er blandt de bedste i DK.

4.2. ChatGPT

ChatGPT er i denne rapport blevet anvendt som en sparringspartner og inspirationskilde. Værktøjet har bidraget til at forbedre formuleringer og skabe en mere ensartet struktur i teksten. Derudover er det blevet brugt til idéudvikling og til at sikre klarhed i rapportens indhold.

5. Definitioner og forkortelser

VIP-gæster = partnere med guldmenuer inkluderet i deres samarbejdsaftale
Viborg Fodsports Forening = VFF
Energi Viborg Arena = Viborg Stadion
Desuden anvendes almindelige sproglige forkortelser, som typisk forekommer i skriftligt dansk, hvor det er relevant.

6. Videnskabsteori og metode

6.1. Ontologi

6.1.1. Problem

Projektet tager udgangspunkt i kritisk realisme, da denne tilgang gør det muligt at analysere sammenhænge mellem observerbare data, som antallet af bestilte og anvendte billetter, samtidig med at undersøge de underliggende mekanismer, som kan påvirke fremmødet, hvilket er relevant for VFF.

Kritisk realisme giver mulighed for at undersøge på Bhaskars tre niveauer (Egholm, 2014):

1. På det virkelige niveau undersøger vi de strukturer og faktorer, der påvirker udnyttelsesgraden af guld menu-billetter blandt VIP-gæster, samt analyserer VFF's data-modenhed og infrastruktur.
2. På det faktiske niveau udvikler vi en model til forudsigelse af udnyttelsesgraden samt analysere VFF's data modenhedsniveau med fokus på områder til forbedring.
3. På det empiriske niveau evaluerer vi modellens præcision og kommer med et løsningsforslag til, hvordan organisationen kan styrke sin datamodenhed og infrastruktur.

6.1.2. Paradigmevalg

Vi anvender paradigmet kritisk realisme (Egholm, 2014), da det gør det muligt at analysere både observerbare sociale fænomener (fremmødemønstre) og de underliggende mekanismer, der driver dem (virksomhedskultur og eksterne faktorer som vejr og Superliga-placering). Ved at antage dybdeniveauers eksistens opstilles hypoteser om potentielle sammenhænge, som testes for at validere de bagvedliggende strukturer. Kritisk realisme, der kombinerer natur- og samfundsvidenskabelige metoder, gør det derved muligt at identificere faktorer, der påvirker VIP-gæsters fremmøde, forbedre forudsigelsen af guldmenu-billetudnyttelse og udvikle løsninger, der fremmer organisationens datamodenhed.

6.1.3. Retroduktion

Ved at anvende en retroduktiv tilgang (inden for kritisk realisme) (Egholm, 2014) kan vi i undersøgelsen kombinere kvantitative data (guldmenu-billetter og x-variabler) med kvalitative indsigter, der afdækker underliggende mekanismer i organisationen, som indirekte kan påvirke dataudnyttelse og dermed forudsigelsens nøjagtighed i fremmødemønstre. Retroduktionen

muliggør en analyse af både det observerede og de dybere strukturer, der påvirker fremmødet, ved at bevæge os mellem induktion (identifikation af datamønstre) og deduktion (hypotesetest). Dette sikrer et solidt grundlag for udviklingen af realistiske og tilpassede løsninger til VFF.

6.2. Epistemologi

Kritisk realisme balancerer objektive data og subjektive indsigter, hvilket muliggør en kombination af kvantitative analyser og kvalitative interviews (Egholm, 2014).

6.2.1. Kvalitative data og kvantitative data

Den metodiske pluralisme i kritisk realisme (Egholm, 2014) understøtter brugen af kvantitative data fra guld menu-registreringer til at identificere faktorer, der kan påvirke fremmødet, og kvalitative data fra interviews til at analysere organisationens datamodenhed.

6.2.2. Sekundære og primære data

Vores primære data indsamler vi gennem vores egne metoder, såsom interviews.
Vores sekundære data består af de eksisterende data, vi har fået fra VFF (guld-ark) samt eksterne data fra kilder som Superstats (Superstats, u.d.) og DMI (DMI, u.d.), som understøtter vores undersøgelse.

6.3. Metodologi

6.3.1. Undersøgelsesdesign

Blandede metoder (Egholm, 2014), fordi de muliggør kombination af kvantitative analyser af fremmødemønstre og kvalitative interviews for at afdække organisatoriske udfordringer.

6.3.2. Metoder

Projektet er et casestudie, der analyserer udfordringer med datamodenhed og billetudnyttelse. Gennem semistrukturerede interviews og dataanalyse (som CRISP-DM sætter rammen for) undersøges faktorer, der hhv. påvirker organisationens strategiske brug af data og faktorer der kan have betydning for udnyttelsesgraden, herunder udvælgelse af machine learnings model (Gareth James, 2023) til forudsigelse.
R-studio anvendes som det primære værktøj til dataanalyse og modellering. Med brug af fleksible pakker som tidyverse, caret og ggplot2 blev datarensning, visualisering og udvikling af prædiktive modeller struktureret i sektioner for dataimport, EDA (Exploratory Data Analysis) og modellering.

Dette sikrede gennemsigtighed og reproducerbarhed i analyseprocessen og understøttede en systematisk anvendelse af CRISP-DM-metoden i projektet.
Kotter's 8-trins model (Bang, 2024) bruges til at koble teori og praksis i udviklingen af løsninger.

6.3.2.1. Datagenereringsprocessen

Processen for indsamling af empirisk data er kort beskrevet nedenfor. For uddybelse af processen se **bilag 1.**
Empirisk data blev indsamlet via semistrukturerede interviews med fire VFF-medarbejdere og tre praktikanter fra forskellige afdelinger. Fokusområder inkluderede infrastruktur, VIP-fremmøde og netværksoplevelser, hvilket danner grundlag for projektets analyser og løsninger.

6.3.2.2. Præsentationer og præsentationsanalyse

Transskribering af interview 1 vedlagt som:	**Bilag 2**
Transskribering af interview 2 vedlagt som:	**Bilag 3**
Interviewanalyse vedlagt som:	**Bilag 4**

6.3.2.3. Præsentationer og præsentationsanalyse

Præsentationsanalysen er baseret på præsentationerne fra Dataafdelingen (Dataafdelingen, 2024) og Marketingsafdelingen (Marketingafdelingen, 2024) og er vedlagt som **bilag 5.**

7. Analyse

CRISP-DM (Bang, 2024) bruges som nævnt til at sætte rammen for projektet, da denne model sikrer en systematisk tilgang og værdi fra data. En kort beskrivelse og modellens relevans for projektet findes i **bilag 6**.

7.1. Forretningsforståelse

VFF udfordres med at forudsige fremmødet af VIP-gæster med guldmenu til hjemmekampe. Projektet sigter mod at udvikle en machine learning-model, der optimerer ressourcer og forbedrer gæsteoplevelsen.
Præsentations- og interviewanalyser afslørede, at VFF's udfordringer skyldes både forudsigelsesproblemer og silo-tendenser, der hæmmer dataintegration og kommunikation.

VFF har implementeret en ordning, hvor uaktiverede billetter frigives før kampstart, hvilket reducerer fremmødeusikkerhed.

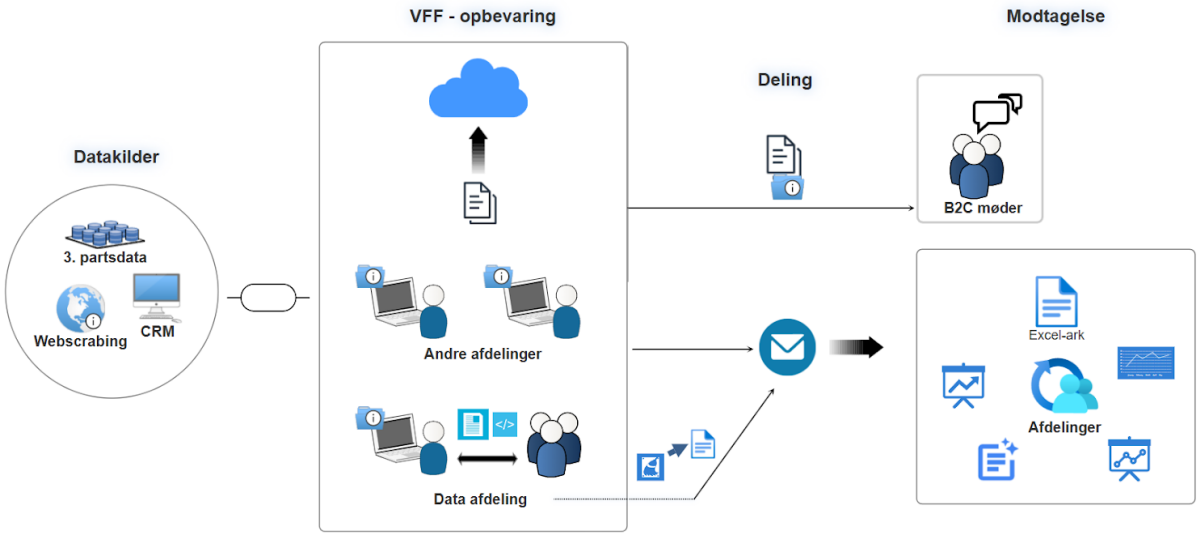
Dette initiativ er et skridt mod at reducere uforudsigeligheden omkring fremmødet. Samtidig fremhæver det behovet for en mere struktureret tilgang til datahåndtering, hvor en forudsigelsesmodel kan anvendes som et redskab til at udnytte eksisterende data mere effektivt og understøtte organisationens overordnede mål om at styrke datamodenheden

7.1.1. Organisationsstruktur og kultur

VFF driver professionel fodbold og arbejder målrettet på at skabe fællesskaber og unikke oplevelser (Bierholm, 2023-2024). Virksomheden har en vision om at være blandt de bedste i Danmark og stræber efter at gøre deres arrangementer så oplevelsesrige, at fodboldkampene bliver en mindre del af en større helhedsoplevelse (Interview 1, 2024).
Virksomheden er opdelt i to hovedafdelinger: en sportsafdeling og en administrativ afdeling. Den administrative afdeling består af cirka 20-25 ansatte og opererer med en flad struktur.
Den administrative afdeling er yderligere opdelt i funktioner med egne specialiseringer. Nogle af funktionerne har egen leder. Den horisontale arbejdsdeling gør det muligt for medarbejderne at dele viden og ekspertise på tværs af organisationen.
Den funktionsopdelte struktur hos VFF har ført til silo-dannelser, hvor afdelingerne arbejder med egne systemer og data. I organisationens decentraliserede tilgang træffes beslutninger tæt på medarbejderne. Dette giver en medarbejder mulighed for at dele en interessant opdagelse i sine egne data med en kollega i en relevant afdeling, som kan anvende den (Interview 1, 2024).
Siloerne har dog haft en negativ indflydelse på den interne kommunikation, idet forskellige afdelinger ofte anvender forskellige definitioner og fortolkninger af centrale områder, f.eks. forklaring af VIP-billetter. Organisationen oplever det som en ressourcekrævende udfordring at integrere data mere effektivt på tværs af afdelingerne (Interview 1, 2024).
For at forbedre analyserne anbefales det, at VFF indsamler flere kundedata om VIP-gæster med guldmenu-billetter. Dette vil ikke kun understøtte organisationens datamodenhed, men også give mulighed for at analysere, hvordan netværksoplevelsen påvirker fremmødemønstre. En dybere forståelse af denne dimension kan bidrage til mere målrettede tiltag og forbedret gæsteoplevelse.

Udviklingen i sportens verden har ikke kun haft indflydelse på VFF sportsafdelingen, men også i den administrative afdeling. I de seneste år har VFF besluttet at bruge data i deres beslutningstagning, i stedet for at gå med mavefornemmelser og oplevelser (Interview 1, 2024).

7.1.2. Business Proces Mapping



Figur 1, Egen fremstilling udarbejdet vha. Smartdraw.com

VFF har de seneste år arbejdet på at blive en datadrevet organisation, hvor beslutninger i højere grad træffes på baggrund af data. Dette har medført en øget brug af data på tværs af afdelinger, men der er stadig udfordringer med varierende IT-kompetencer blandt medarbejderne. Data indsamles fra forskellige kilder og deles på forskellige måder, især som Excel-filer. Opbevaring og organisering af data sker ofte i et ad hoc-setup, hvor data gemmes på medarbejdernes individuelle computere. VFF's kommunikation er fleksibel, men denne fleksibilitet fører til ineffektivitet, da den skaber silo-dannelse, manglende overblik og tidsspild. Samlet set har VFF gjort fremskridt mod at blive en mere datadrevet organisation, men der er behov for at standardisere processer for opbevaring, organisering og deling af data for bedre at udnytte data (se **bilag 7**).

7.1.3. Datamodenhed

VFF's data-modenhedsniveau analyseres i **bilag 8** for at identificere, hvor organisationen befinder sig på nuværende tidspunkt. Dette gøres vha. CMMI Institutes Data Management Maturity Model (Datadrevet bog kap. 2, s. 47), da vi med denne model kan identificere de områder, hvor

organisationens nuværende praksis understøtter succes samt de områder, hvor de kunne bruge forbedring.

Modenhedsvurdering af datahåndtering

Analysen viser, at organisationen samlet set befinder sig på niveau 2: Reactive, med variationer på tværs af de seks hovedkategorier. Selvom der er etableret grundlæggende processer, hæmmer manglen på ensartede retningslinjer, kvalitetssikring og systemintegration organisationens effektivitet, hvilket fremhæver behovet for standardisering af processer og et styrket samarbejde på tværs af afdelingerne.

Denne vurdering danner et solidt grundlag for vores videre udvikling af et realistisk løsningsforslag til VFF.

7.2. Dataforståelse

I denne fase af projektet undersøges de tilgængelige data for at opnå en dybere indsigt i deres struktur og relevans. Dette omfatter en analyse af data om guldmenuer, fremmødemønstre og billetbrug for at identificere mønstre og potentielle udfordringer.

I **bilag 9** undersøges en mulig sammenhæng mellem VFF's hjemmekampe og lokale arrangementer. Der blev ikke fundet en betydelig sammenhæng, hvilket tilskrives begrænset dataadgang. Med historiske data fra f.eks. Viborg Kommunes 'Det sker' (Kommune, u.d.) kunne analysen have været mere omfattende.

7.2.1. Datakilder

De tre primære datakilder for programmeringsanalysen er:

1. Data fra Excel-filen: Guld.xlsx. Filen inkluderer data trukket internt fra VFF med detaljer om antal bestilte-, anvendte- og et estimeret antal max for VIP-gæster med guldmenu billetter.
2. DMI's API-data: Vejrdata som inkluderer temperatur, vindhastighed og nedbør m.m., blev hentet og koblet på kampdatoer.
3. Webscrapede kampdata: Kampe, modstandere, tilskuertal og andre relevante oplysninger fra Superstats.

7.2.2. Undersøgelse

Udnyttelsesgraden i de interne data i Guld.xlsx, blev fundet ved at dividere guld_menu_stk med antal_bestilte af VIP-gæster.

API-data fra DMI bidrog med vejrforhold, såsom temperatur, vindhastighed og nedbør m.m., som potentielle prædiktorer for fremmødet (x-variabler).

Superstats-data blev udvidet med yderligere information, herunder oplysninger om modstandere og publikum (x-variabler).

Nye funktioner fra den opdaterede kode

- Variablen *rolling_goals_3* blev lavet for at beregne gennemsnittet af mål scoret over de seneste tre kampe, hvilket bruges som en indikator for holdets præstationshistorik.
- *Weather_cat* blev tilføjet som en kategorisk variabel, der grupperer vejrforhold i tre kategorier (Good, Moderate, Bad) baseret på temperatur og nedbør. Denne laves for at gøre modellen enklere.
- Analyse af korrelationer viste stærke forbindelser mellem vejrforhold, modstander og udnyttelsesgrad.

Interaktioner mellem vejr variabler blev også inkluderet, som fx *cloud_cover* × *humidity*, for at identificere skjulte mønstre.

7.2.3. Visualisering

Data blev visualiseret for at identificere unormale værdier og finde mulige mønstre. For eksempel viste korrelationerne, at der var en stærk sammenhæng mellem temperatur og udnyttelsesgraden. Residualplots viste, at præcisionen var blevet forbedret i de opdaterede modeller.

Analyser af interaktionerne afslørede, at både vejret og kampens kontekst spiller en vigtig rolle i variationen af udnyttelsesgraden.

7.3. Dataforberedelse

Data klargøres til analyse ved at gennemgå processer som rensning af fejlbehæftede værdier, håndtering af manglende data og transformation til et format, der er egnet til modellering.

7.3.1. Rensning

Overflødige kolonner såsom 'Gule_poletter_stk' blev fjernet for at fokusere på relevante variabler.
NA-værdier i numeriske kolonner blev importeret med medianer for at sikre et mere realistisk billede afi modellerne.
Dato formater blev standardiseret for at sikre konsistens på tværs af alle datasæt. Dette inkluderede konvertering af alle datoer til et ensartet 'ÅÅÅÅ-MM-DD' format, hvilket letter sammenkoblingen af data fra forskellige kilder og sikrer præcis datojustering ved sammenfletning af datasæt.
Ikke-numeriske data som 'goals' og 'time' blev konverteret til numeriske variabler ved hjælp af string parsing, og forskellige dato formater blev tilpasset til et standardformat.
Der blev anvendt strengere korrelationsfiltre for at fjerne unødvendige gentagelser mellem variablerne og dermed gøre modellen stærkere.

7.3.2. Integration

Data fra kilder som DMI's API, Guld.xlsx og webscraping blev sammenkoblet på datoerne for at skabe en integreret dataramme.
Vejrdata og kampdata blev kombineret med udnyttelsesgraden ved at justere de præcise datoer, så eksterne faktorer som temperatur, luftfugtighed og nedbør blev taget med.

7.3.3. Feature Engineering

Ny variabel 'weather_cat' blev lavet for at gruppere vejret i kategorierne 'Good', 'Moderate', 'Bad', hvilket forenkler modelleringen.
Tidsmæssige variabler som 'minutes_since_midnight' og 'weekday' blev tilføjet for at fange tidsmæssige mønstre.
Historiske tendenser blev indarbejdet gennem 'rolling_goals_3' og 'rolling_audience_3', samt nye variabler som 'goal_diff'.
Interaktionsvariabler som 'cloud_cover * humidity' blev lavet for at undersøge de kombinerede effekter.

7.3.4. Normalisering og kodning

Kategoriske variabler som kampnavne, modstandere og 'weather_cat' blev omdannet til dummy-variabler ved hjælp af dummy-encoding for at se forholdene mellem "good, moderate og bad" i forhold til kampe og modstandere.
Numeriske variabler blev skaleret ved hjælp af centerring og standardisering for at undgå bias i modellerne, og forbedre modellens evne til at arbejde med kategoriske data.

7.4. Modellering

Her anvendes data mining-teknikker og algoritmer til at udvikle modeller, der kan forudsige udnyttelsesgraden for VIP-gæster.

7.4.1. Valg af modeller og træning

For at udforske og forbedre udnyttelsesgraden, testede vi fire forskellige statistiske modeller:

1. **Lineær regression:** Den har vi brugt som en simpel baseline for at skabe et udgangspunkt for vores projekt og for at kunne se tendenserne rimelig hurtigt i forløbet.
2. **Lasso regression (L1-regularisering):** Den er effektiv til at fjerne unødvendige variabler ved kun at vælge de mest relevante, hvilket forbedrede modellens nøjagtighed og reducerede MSE med cirka 15%. Det giver også god indsigt i, hvilke faktorer der virkelig betyder noget, ved at sætte mindre relevante variabler til 0. For eksempel kan vi se, at 'Antal_bestilte', 'weather_cat', 'rolling_audience_3' og 'weekday/time' har en tydelig indvirkning.
3. **Ridge regression (L2-regularisering):** Brugt til at håndtere multikollinearitet og overfitting, men viste sig ikke at være mere effektiv end Lasso.
4. **Random Forest:** Fokuserede på ikke-lineære sammenhænge og vigtigheden af specifikke variabler. Denne model blev optimeret med justerede hyperparametre for at forbedre dens præstation, hvilket resulterede i en MSE på 140, en forbedring fra tidligere 150, og et R^2 på 0.92.

7.4.2. Træning og test

Vi delte datasættet op, så 80% blev brugt til træning, og de resterende 20% blev brugt til test. For at vurdere modellernes anvendelighed og stabilitet benyttede vi 5-fold cross-validation. Vi brugte også en "CutDown" hvor vi reducerede variablerne til de vigtigste variabler som havde størst

indflydelse på vores resultater. På den måde hjalp det os med at mindske risikoen for overfitting og sikre, at modellerne kunne give stabile resultater.

7.4.3. Resultater og fortolkning (MSE, RMSE, R^2)

Resultaterne af modelleringen viste, at Random Forest og Ridge Regression præsterede bedst med hensyn til MSE og RMSE, mens Ridge Regression også havde den højeste R^2-værdi. De detaljerede resultater fremgår af Tabel 1 nedenfor, og en mere udførlig gennemgang findes i **Bilag 10.**

Tabel 1: Egen fremstilling til sammenligning af modellernes præstationer

Model	MSE	RMSE	R^2
Lineær Regression	0.002904	0.053892	0.311
CutDown LM	0.002582	0.050810	0.005
Lasso Regression	0.002226	0.047181	0.023
CutDown Lasso	0.002564	0.050634	0.005
Ridge Regression	0.002130	0.046156	0.242
CutDown Ridge	0.002385	0.048837	0.002
Random Forest	0.001997	0.044683	0.202
CutDown Random Forest	0.001888	0.043457	0.166

Random Forest viste sig at have den laveste MSE og RMSE, hvilket gør den særligt egnet til forudsigelse af VIP-gæsternes fremmøde. På baggrund af datasættets begrænsede størrelse vurderes dog, at alle modeller vil drage fordel af en større datamængde.

De vigtigste variabler

I bilaget fremgår der en visualisering i form af en graf, som viser betydningen af de forskellige variabler for udnyttelsesgraden. Herunder ses en tabel oversigt, over betydningen for variablerne baseret på analysen i R-studio:

Tabel 2: Egen fremstilling

Variabel	Betydning
Antal_bestilte	Angiver den direkte efterspørgsel; grundlæggende for resten af variablerne
weather_cat_Good	Godt vejr øger sandsynligheden for flere fremmødte og en højere udnyttelsesgrad
rolling_audience_3	Tidligere publikumsmønstre hjælper med at forudsige fremtidig efterspørgsel
weekday	Visse ugedage har typisk højere aktivitet og dermed bedre udnyttelsesgrad
minutes_since_midnight	Tidspunktet på dagen påvirker efterspørgslen

7.5. Evaluering

I denne fase vurderes modellernes præstation for at sikre, at de opfylder de forretningsmæssige krav. Det omfatter validering af modellernes nøjagtighed ved hjælp af testdata og vurdering af, om resultaterne er relevante og anvendelige for VFF's beslutningsprocesser.

7.5.1. Præstationsanalyse

Random Forest modellen fortsatte med at levere de bedste resultater i forhold til andre testede modeller, hvilket blev demonstreret ved den laveste mean squared error (MSE) og den højeste R^2. Denne model blev forbedret yderligere gennem optimering med nye funktioner, som bidrog til en mere ensartet fordeling i residualplots, indikerende mindre bias og en stærkere modelgeneraliserbarhed.

7.5.2. Fortolkning

Analyser og fortolkninger af modellen bekræftede, at vejrforhold (kategoriseret som 'weather_cat') og kampkontekst (repræsenteret ved 'rolling_audience_3') er kritiske prædiktive faktorer. 'rolling_audience_3' havde en betydelig indflydelse på modellens præcision, hvilket blev fremhævet gennem både Lasso og Random Forest modeller. Desuden viste temperatur og modstanderes popularitet sig at have en betydelig indflydelse på fremmødet, hvor en højere temperatur og populære modstandere korrelerede med større fremmøde.

7.5.3. Visualisering

Visualiseringerne understøttede disse fund med præcise plots af observeret versus forudsagt fremmøde, som illustrerede Random Forest modellens overlegenhed i præcision. Variable importance-plots afslørede desuden 'Antal_bestilte', 'rolling_audience_3', og 'weekday' som nøglevariable. Disse plots var afgørende for at validere modelpræcisionen og for at vise, hvordan forskellige prædiktorer bidrog til modellens præstationer. Residualplots fra de opdaterede modeller afslørede en reduceret bias og en mere ensartet fejlfordeling sammenlignet med tidligere modeller.

7.6. Implementering

7.6.1. Anvendelse af machine learning-resultater og anbefalinger

De udviklede prædiktive modeller og deres resultater kan bruges i VFF's strategiske beslutningsprocesser for at optimere, hvordan ressourcerne fordeles, og forbedre forberedelserne til hjemmekampe. For at gøre vores machine learnings produkt praktisk anvendeligt, vil vi anbefale at etablere dataarkitektur i organisationen.
Når det overordnede er på plads, kan VFF starte med at anvende vores analyser og modeller til at bygge deres egne modeller eller videreudvikle på dem. Det kræver en grundig indsigt i de data og faktorer, der spiller en rolle for guldmenuerne, så modellerne kan designes og optimeres på bedst mulige måde.

7.6.2. Anbefalinger

På baggrund af de foretaget analyser anbefales det at VFF:
1. Fortsætter den allerede påbegyndte proces med at oprette et Datawarehouse.
2. Anvender Kotter's 8-trins model som en strategisk ramme for at implementere de organisatoriske ændringer, der er nødvendige for at opnå dataintegration og ejerskab.

I **bilag 11** har vi udarbejdet Kotter's 8-trins model for forandring tilpasset VFF, baseret på vores begrænsede kendskab til organisationen infrastruktur og virksomhedskultur.

7.6.2.1. Implementeringsforslag

Kotter's 8-trins model kan hjælpe VFF med at oprette et Datawarehouse og forbedre datamodenhed gennem standardisering, samarbejde og klare retningslinjer og derved sikre en struktureret og effektiv implementering af nye datainitiativer.

8. Konklusion

Projektet viser potentiale i prædiktionsmodeller til at forudsige udnyttelsesgraden af guldmenu-billetter, men begrænsede datamængder og manglende standardisering reducerer modellernes anvendelighed. Dette understreger, at organisatoriske forbedringer, som centralisering af data og etablering af en datadrevet kultur, skal prioriteres før tekniske løsninger som et Datawarehouse. Derfor blev organisationens datamodenhedsniveau først og fremmest vurderet, vha. CMMI Institutes Data Management Maturity Model-analyse. Resultatet fra modellen viste, at organisationen på nuværende tidspunkt befinder sig samlet set på niveau 2 (reactive), baseret på de tilgængelige data og informationer, som projektet har haft til rådighed til undersøgelsen. Datamodenhedsanalysen dannede grundlag for konkrete initiativer, der kan understøtte integrationen af forudsigelsesmodellen i organisationen ud fra det nuværende datamodenhedsniveau, for at sikre, at løsningsforslaget er realistisk at implementere for VFF. Organisationen er i vækst og bevæger sig mod at blive en endnu mere datadrevet organisation med en vision om at opnå 100% ejerskab over egne data.
Der er på nuværende tidspunkt silo-dannelses tendenser pga. manglende integration mellem de datadrevne systemer. Dette begrænser muligheden for effektivt at udnytte det fulde potentiale, som integration af data kan tilbyde, især ift. at understøtte strategisk beslutningstagning baseret på avancerede datamodeller.
Implementeringen af Kotter's 8-trins model anbefales som et strategisk værktøj til at fremme samarbejde, skabe fælles forståelse og sikre en stærkt organisatorisk grundlag for databenyttelse. Modellen er foreslået som en praktisk ramme, der kan hjælpe VFF med at fremme en datadrevet kultur og styrke forståelsen for dataanvendelse på kort sigt. Samtidig kan modellen, med justeringer, anvendes på længere sigt til at understøtte organisationens evne til at håndtere større forandringer, herunder implementering af Datawarehouse-løsningen som sandsynligvis vil kræve mere omfattende ændringer i datakulturen samt en øget dataforståelse på tværs af alle afdelinger, hvilket gør det nødvendigt at støtte medarbejderne i overgangen til en mere datadrevet kultur.

9. Metodekritik og refleksion

I **bilag 12** præsenteres et afsnit, der omfatter metodekritik og refleksion.

10. Litteraturliste

1. semesterprøven, b. (2024). *Dataanalyse, Viborg*.
Bang, C. G. (2024). *Data-Driven Decision-Making for Business.*
Bierholm. (2023-2024). *Årsrapport for regnskabsåret , s. 15.*
Dataafdelingen. (2024). Præsentation fra dataafdelingen. *Datachef i VFF: Daniel og praktikanter: Frederiks, Olga og Mohammed*.
DMI. (u.d.). Hentet fra https://www.dmi.dk
Egholm, L. (2014). *Videnskabsteori, perspektiver på organisationer og samfund.* Hans Reitzels Forlag.
Gareth James, D. W. (2023). *An Introduction to Statistical Learning, with Applications in R.*
Interview 1, T. N.-o. (2024). (D. o. undervisere, Interviewer)
Kommune, V. (u.d.). *Det sker i Viborg*. Hentet fra https://viborg.dk/oplevelser-og-fritid/det-sker/?ArrKunstner=&Genre=&ArrStartdato=15%2F12%202024&Area=Viborg-postby
Marketingafdelingen. (2024). Præsentation fra Marketing. *Marketing- og Kommunikationschef: Tea Nørgaard og Marketingansvarlig: Daniel Lindemann Jakobsen*.
Palle. (2024). Præsentation Billetsalg.
Respondenter VFF, p. (2024). Samlet indblik i organisationen fra præsentationer.
Superstats. (u.d.). Hentet fra https://superstats.dk
Tea Nørgaard Marketing- og Kommunikationschef, P. (2024).

11. Tabel over figurer

12. Bilag

Bilag er vedlagt i et separat dokument (med sidetal på)

13. Redegørelse af ændringer i forhold til det oprindelige projekt

De eneste ændringer, der er foretaget, omfatter en revision af nogle få steder for at forklare, hvad der sker i koden og rettelser af stavefejl.