Projectmateriaal

Uitgangssituatie: De grote fusie en uitbreiding

De retailwereld van outdoorartikelen is recent behoorlijk opgeschud. Een voorheen kleine speler, Great Outdoors, heeft met de bouw van een Data Warehouse (DWH), verschillende dashboards en Machine-Learningmodellen een zeer sterke marktpositie weten te veroveren. Zij dreigt nu niet alleen een van de grootste spelers te worden op de kampeer-, maar als gevolg van een organisatie-uitbreiding ook op de sportmarkt.

Het bedrijf AenC, een kledingdetailhandel in juist deze sportmarkt met meerdere filialen, besluit op advies van hun financieel adviseurs en aandeelhouders te gaan fuseren met 2 organisaties:

- AdventureWorks: een grote fietsenhandelaar. Ook dit bedrijf heeft namelijk als missie om (potentiële) klanten te stimuleren om meer tijd "outdoor" door te brengen.
- Northwind: een grote voedselhandelaar. AenC hoopt op deze manier een alles-in-éénorganisatie te worden voor mensen die (een aantal dagen lang) willen sporten.

Het fusiebedrijf zal *United Outdoors* gaan heten. Men is zich er zeer van bewust dat om de grootste concurrent, Great Outdoors, de baas te zijn, er snelle, maar tegelijkertijd goede beslissingen genomen moeten worden. Het uitvoeren van hedendaagse én voorspellende data-analyses wordt erkend als hét middel zowel de snel- als de juistheid van deze te nemen besluiten te waarborgen. Om dit mogelijk te maken willen de drie bedrijven zo snel mogelijk hun bestaande data-infrastructuur bundelen in een DWH. “Schaalbaarheid” is hier een belangrijke eigenschap: het moet zo makkelijk mogelijk zijn om data uit nieuwe bronnen aan het DWH toe te voegen.

Daarnaast bestaat binnen *United Outdoors* een driestrijd over de verwachtingen omtrent de uitvoerbaarheid van Machine-Learningtechnieken bij de bovengenoemde data-analyses.

- Een groep mensen (de “ML-pessimisten”) is stellig van mening dat het optimaliseren van de voorspelbaarheid voor geen enkele verzamelde datasoort van meerwaarde is. De link tussen voorspellende analyses (bijv. over verkoopverwachtingen van de volgende maand) en de bovengenoemde te nemen besluiten zien zij (nog) niet.
- Een groep (de “ML-twijfelaars”) ziet wél het nut in van het uitvoeren van voorspellende analyses, maar twijfelt of deze op accurate resultaten komen, gegeven de hoeveelheden en soorten verzamelde data.
- De derde groep (de “ML-optimisten”) wil liever vandaag dan morgen aan de slag met het weerleggen van de zorgen van de bovengenoemde twee groepen. Volgens hen verzamelt United Outdoors voldoende data om voorspellingen met voldoende businesswaarde te kunnen doen. Bovendien zijn zij van mening dat er genoeg externe databronnen op internet te vinden zijn die in voorspellende analyses geïntegreerd kunnen worden om tot nóg accuratere voorspellingen te komen.

Projectopdracht

Aan jouw projectteam wordt door United Outdoors het volgende gevraagd:

1. Een schaal- en overdraagbaar Data Warehouse (DWH) in SQL Server Management Studio (SSMS). Hierin zijn in ieder geval alle feit- en dimensietabellen uit de brondatabases van de drie gefuseerde organisaties opgenomen. Alle data die heel erg sterk aan elkaar gelinkt is staat hierin in één tabel, hiervoor vinden dus de nodige samenvoegingen plaats. Denk bijvoorbeeld aan één tabel die álle klantinformatie bevat, één tabel die álle productinformatie bevat, enzovoort.
2. Een shortlist met mogelijk relevante afhankelijke variabelen met daarbij per variabele twee kenmerken…
    a. De relevantie van de afhankelijke variabele: een nadere uitleg over welke beslissing(en) je kan nemen als je nu al weet welke waarde(n) deze komende maand/kwartaal/jaar gaat hebben.
    b. Een eerste inschatting van bijpassende onafhankelijke variabelen: “op basis van welke onafhankelijke variabelen kan naar jouw verwachting de afhankelijke variabele het beste voorspeld worden?”.
3. Meerdere Machine-Learningmodellen die de daadwerkelijke voorspellingen maken m.b.t. de bovengenoemde shortlist. Op basis van diens resultaten (mean absolute error, accuracy, enzovoort) wordt verwacht dat je iteratief optimaliseert. Je past de juiste hyperparameters aan (bijvoorbeeld max_depth, k, enzovoort), kijkt of het tot de juiste resultaten leidt, past vervolgens weer hyperparameters aan, kijkt of het tot nóg betere resultaten leidt, enzovoort.
4. Een dashboard met visualisaties van zowel hedendaagse sturingsinformatie als voorspellende analyses. Met dergelijke dashboards moeten de in-house data-analisten de resultaten van beide soorten analyses inzichtelijk kunnen maken én, wat Machine-Learningresultaten betreft, de kwaliteit en betrouwbaarheid. Een hoog interactief karakter van dit dashboard is zeer gewenst: denk aan zinvolle filters, drill-downstructuren, enzovoort. In eerste instantie kan dit dashboard gemaakt en gepresenteerd worden in Power BI, maar op de langere termijn willen de data-analisten van United Outdoors dat het dashboard ook via een webomgeving te raadplegen is.

<u>Extra aandachtspunten voor het laatste FDP</u>

- Tijdens het FDP wordt kritisch gekeken in hoeverre de 4,5 projectweken **<u>ten volle</u>** zijn benut om het resultaat af te krijgen. Met andere woorden: men staat uitvoerig stil bij de vraag: “is het projectresultaat in lijn met wat we verwachten qua hoeveelheid verricht werk in 4,5 week?”.
- Men is bereid om degenen die een effectieve webomgeving demonstreren een nóg hogere beloning uit te keren, dan degenen die zich beperken tot Power BI.

Context

De context van United Outdoors is als volgt:

- Het bedrijf is nu primair bezig met het regelen van de fusie en geeft jullie volledige autonomie over het opzetten van de data-analyse binnen de hierboven geschetste opdracht
- Omdat de medewerkers van United Outdoors niet beschikbaar zijn zal jouw team de benodigde kennis voor de uitvoering van de opdracht zelfstandig elders moeten inwinnen.
- De (toekomstige) data-analisten van United Outdoors zijn sterk in het bekijken van data, maar niet technisch begaafd. Het gebruik van jargon uit de outdoorbranche wordt door de analisten zeer op prijs gesteld.
- De brondata wordt aangeleverd as-is. United Outdoors kan niet garanderen dat alle data correct is en dat de verschillende bronnen (optimaal) op elkaar aansluiten. Het kan zijn dat op een later moment, eventueel na afloop van het project, geupdate versies van de bronbestanden worden aangeleverd. Uiteraard moeten de dashboards dan automatisch updaten.
- United Outdoors vraagt zowel kwaliteit (zinnige en goed functionerende overzichten in het dashboard) als kwantiteit (zoveel mogelijk relevante overzichten).

Databestanden

De volgende databestanden zijn nodig tijdens het uitvoeren van het project:

- De database van **AenC** , dit is een downloadbaar Microsoft-Accessbestand
- De database van **AdventureWorks** (bron: Microsoft), dit is een downloadbaar .bak-bestand. Gebruik de instructies op deze **link** om deze te importeren in SQL Server Management Studio.
- De database van **Northwind** (bron: Microsoft), dit is een downloadbaar .txt-bestand. Maak in SQL Server Management Studio een database genaamd "Northwind" aan, open de SQL-editor en kopieer & plak ten slotte de inhoud van dit bestand daarin.