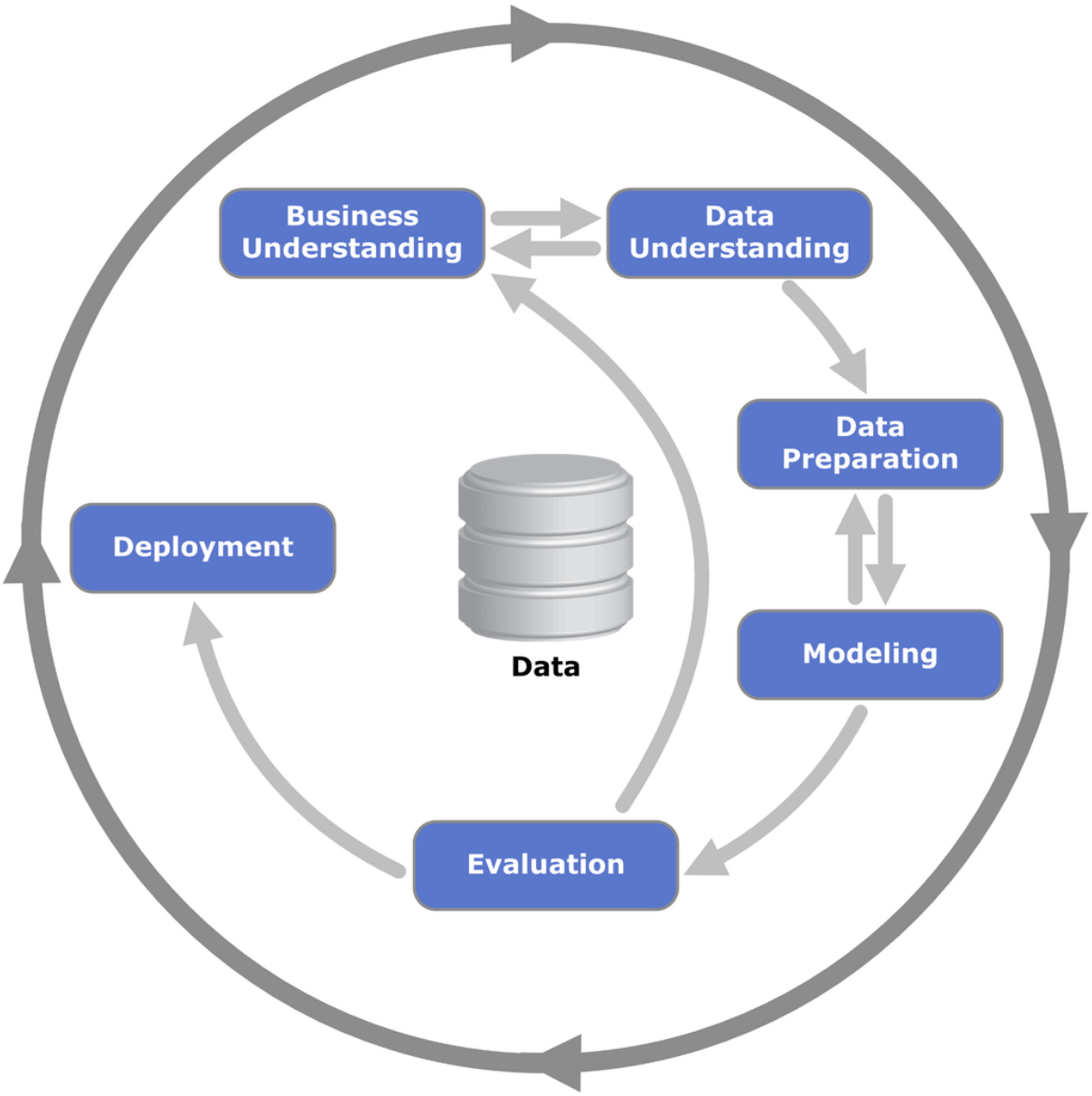
Metodologie CRISP-DM. Exemple de proiecte DS

Proiectele de DS sunt diverse si greu de prins intr-o schema unitara. Totusi, exista un consens larg asupra pasilor care trebuie sa se execute intr-un astfel de proiect. Mai mult, exista o metodologie care da secventa de pasi de urmat. Metodologia se numeste CRISP-DM (Cross-industry standard process for data mining).Este un standard deschis, convenind pasii pe care expertii in Data Mining ii urmeaza. Conceput in 1996, a fost extins in 2015 de catre IBM prin Analytics Solutions Unified Method for Data Mining/Predictive Analytics (ASUM-DM).



Sursa: https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining#/media/File:CRISP-DM_Process_Diagram.png (https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining#/media/File:CRISP-DM_Process_Diagram.png)

Sunt 6 faze majore intr-un proces de DS/DM:

1. Business understanding
2. Data understanding
3. Data preparation
4. Modeling
5. Evaluation
6. Deployment

Pasul 1. Determinarea obiectivelor de business (Determine business objectives)

1. Ce se doreste de la proiect?
    - setarea obiectivelor - traducerea din perspectiva de business in obiective data science. De exemplu, se doreste determinarea tenditei clientilor de a se muta la un competitor; deciderea asupra caror clienti li se va adresa o anumita campanie (cui trimiti pliante/solicitari de donatii etc.)
    - perspectiva DS: specificarea tipului de problema: clasificare, regresie, descriere, sumarizare, grupare, detectare de outliers etc.
    - definirea metricilor de succes - cum se decide daca proiectul se incheie cu succes sau nu, metode masurabile de cuantificare a profitului, pe scurt masurarea impactului procesului (parte din Key Performance Indicators, KPIs)
    - producerea unui plan de actiune - pasii care urmeaza sa se execute in restul proiectului, inclusiv alegerea uneltelor si a tehnicilor folosite
2. Estimarea situatiei curente
    - inventar de resurse: personal (experti in domeniu, pentru domain knowledge; suport tehnic; experti ML/DS, ingineri DS); date (fisiere CSV, date relationale, data warehouse, data lake, REST endpoints etc.); resurse hardware (platforme HW, CPU/GPU, stocare, resurse cloud, backup, modalitate de comunicare); software (unelte gratuite/platite - deja disponibile sau achizitionabile, alt software relevant);
    - cerinte, presupuneri, constrangeri:
        - calendar al activitatilor, securitatea datelor/privacy, deliverables (rapoarte, proof of concept etc.)
        - presupuneri care trebuie validate - coerenta datelor
        - constrangeri: resurse necesare, timp de rulare, modalitate de deploy, constrangeri tehnologice
    - managementul riscului - care sunt portiunile riscante din proiect? exista alternative? care e costul lor?
    - terminologie - termeni de business, dar si termeni de ML (accuracy, precision, recall etc.)
    - costuri si beneficii - pentru fiecare rezultat potential, care sunt costurile de proiect pentru atingerea lui?
3. Definirea scopurilor
    - scopuri de business - descrierea rezultatelor intentionate, legate de obiectivele de business (cresterea numarului de clienti, limitarea pierderilor dintr-un proces etc.)
    - metrici de DS - gradul de acuratete atins, scor de tip mean squared error/mean absolute error etc.
4. Producerea unui plan de proiect
    - pasii de executat, durata, resurse cerute, intrari, iesiri, dependinte; iteratiile trebuie explicit date; detectarea zonelor de risc + alternative
    - evaluare initiala a tehnicilor si uneltelor folosite; necesar SW; deciziile pot avea impact de durata mare

Pasul 2. Intelegerea datelor

1. Lista surselor de date
2. Descrierea datelor
    - care sunt datele disponibile, in ce format, efort necesar pentru convertirera lor; verificarea faptului ca satisfac niste cerinte de calitate minimale
3. Explorarea datelor - interogari particulare, vizualizarea datelor, rapoarte intermediare
    - relatii intre perechi sau multimi mici de atribute
    - rezultatele unor agregari simple
    - subpopulatii de date (grupare pe genul persoanelor, provenienta geografica, localizare, nivel de educatie, cunostinte de ordin financiar etc)
    - analiza statistica simpla
    - finalizare cu raport de explorare a datelor - primele chestiuni descoperite, impactul lor preconizat asupra proiectului si metodelor utilizate (de ex: ocurenta valorilor lipsa, clase debalansate etc.)
4. Verificarea calitatii datelor
    - asigura toate cazurile cerute?
    - contin erori? chestiune in care domain knowledge e esential; detectare de valori eronate, outliers
    - missing values? daca da, cum se trateaza (stergere de date/atribute, missing value imputation)
5. Raport asupra calitatii datelor
    - lista rezultatelor obtinute in urma investigatiilor din acest pas
    - sugerarea rezolvarilor - dependenta de domain knowledge, tipuri de probleme, unelte DS disponibile

Pasul 3: pregatirea datelor

1. Selectarea datelor
    - date selectate dupa relevanta
    - motivele pentru care unele date se exclud trebuie sa fie clar documentate
    - constrangeri tehnice (cantitatea de date) considerate
    - caracterul privat al datelor
2. Curatarea datelor
    - modalitate de selectare/filtrare
    - missing value imputation (valori default sau sub-proiecte pentru estimarea valorilor lipsa)
    - raport al etapei de curatare a datelor
3. Constructii auxiliare
    - selectarea sau extragerea trasaturilor (chestiuni diferite)
    - generarea de inregistrari suplimentare (clienti fara comenzi efectuate, alte situatii preluate din realitate)
4. Integrarea datelor
    - jonctiuni de date, concatenare
    - agregari de date (numar de achizitii, valoarea totala a cumparaturilor, valoare media/mediana) -> posibil sa duca la noi trasaturi informative

Pasul 4. Modelarea

1. Selectarea tehnicii de modelare:
    - documentarea tehnicilor ce urmeaza sa fie folosite
    - presupuneri/cerinte asupra datelor (fara valori lipsa, statistici minimale) - in ce conditii tehnicile de la pct anterior functioneaza
    - adaptarile tehnicilor standard pentru cazurile concrete existente
2. Design-ul pasului de test
    - cum se masoara performanta modelelor (metrici ML/DS)
    - descrierea etapei de validare si testare
    - modalitatea de impartire in train/validation/test susbsets
3. Construirea modelelor
    - antrenarea modelelor, determinarea hiperparametrilor, estimarea erformantelor
    - salvarea modelelor rezultate
    - descrierea modelelor si a modificarilor specifice
    - documentarea dificultatilor intampinate (conversia datelor, durata de antrenare/validare, dependente intre pasi etc.)
4. Estimarea performantei modelelor
    - raportarea rezultatelor, interpretarea lor, explicarea comportamentului modelelor
    - revizuirea hiperparametrilor + a valorilor candidat specifice; reantrenare modele

Pasul 5: evaluarea

1. Evaluarea rezultatelor
    - confruntarea cu obiectivele de business - estimarea castigului realizat
    - daca e posbil: aplicare modele pe piata
    - considerarea rezultatelor colaterale obtinute, impactul asupra proiectului, considerarea lor ca metrici
    - aprobarea modelelor - cele care indeplinesc cerintele de business devin **modele aprobate**
2. Revizuirea (review)
    - reevaluarea pasilor procesului, lessons learned, posibilitate de extindere, verificarea compatibilitatii cu datele actuale (data privacy, prevederi locale sau generale etc.)
3. Determinarea pasilor urmatori
    - finalizare proiectului si mutarea in productie (deployment)? reluarea iteratiilor?
    - evaluarea resurselor ramase poate influenta decizia
    - se produce: lista de posibile actiuni, motivul de alegere al fiecarei optiuni
    - decizia finala, cu motivatie

Deployment

1. Planificarea pasului de deployment
    - se sumarizeaza pasii prin care modelele adoptate ajung in "productie". Se poate decide deployere pe cloud, pregatirea de masini virtuale setate corespunzator, software containers (Docker), rescrierea modelelor in alte limbaje (C++/Java) etc;
    - estimare de efort (timp, oameni, resurse software, calificari - e.g. DevOps)
    - sumar al strategiei de deploy, inclusiv pasii necesari, preconditii
2. Monitorizare si mentenanta
    - se evita utilizarile necorespunzatoare, perioadele de nefunctionare, input necorespunzator, interpretarea eronata a productiei sistemelor DS
    - se consulta jurnalele de activitate (logging)
    - se monitorizeaza performanta modelelor (timpii de reactie, calitatea predictiilor, incarcarea sistemelor); se poate detecta degradarea performantei modelelor, de exemplu din motive de concept drifting;
    - plan de monitorizare: actiuni, etape, frecventa de actiune/interogare a starii sistemelor; KPIs;
3. Producerea raportului final
    - poate fi doar un sumar al prouectului si experientele dobandite
    - poate fi un raport extensiv al rezultatelor obtinute
    - artefacte: raport final + prezentare finala
4. Revizuirea proiectului
    - documentarea experientei; lectii invatate; greseli in pasii urmati;
    - poate utiliza rapoartele realizate de-a lungul pasilor

Exemplul 1

Predictia supravietuirii pe Titanic (https://www.districtdatalabs.com/how-to-start-your-first-data-science-project)

Exemplul 2

Data science worklow example
(https://github.com/aakashtandel/misc_projects/blob/master/Data%20Science%20Workflow%20Project/Data%20S