PROJECT “BIG DATA AND BUSINESS INTELLIGENCE” - SPACESHIP TITANIC

Introductie:

We zijn in het jaar 2912 en hebben een boodschap ontvangen van ‘Spaceship Titanic’. Tijdens een onoplettendheid is het schip een aanvaring gehad met een wormgat waarbij ± de helft van de bemanning verloren is geraakt.

Het was onze taak om op basis van de bestaande data te voorspellen wie er een grotere kans heeft om geteleporteerd te zijn.

Stappenplan:

We maken eerst een beschrijvend data-onderzoek met behulp van PowerBI, dit om de dataset beter te begrijpen en betere keuses te kunnen maken in de volgende stappen.

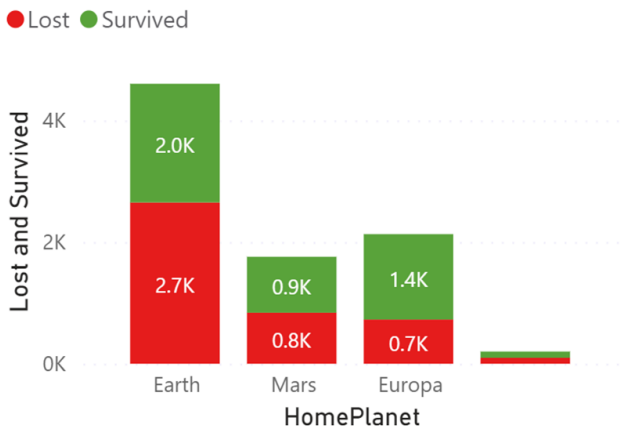
Vervolgens gaan we de data pre-processen met behulp van Python, waarbij we op zoek gaan naar ontbrekende data, deze behandelen en overbodige data verwijderen.

We zullen ook een correlatiematrix opstellen om de features te vinden die het meest correleren. Met behulp van scikit-learn zullen we regressie toepassen op de dataset en een model bouwen / testen met de aangeleverde datasets.

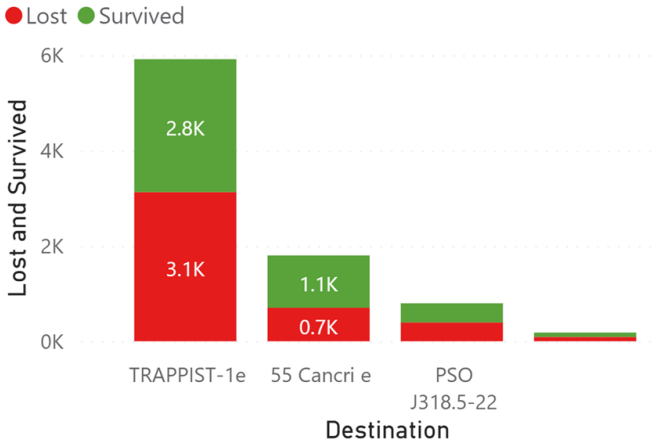
Tot slot laten we onze submission testen door Kaggle om onze precisie te bepalen.

Data onderzoek m.b.v PowerBi

We kunnen een aantal dingen concluderen uit het PowerBi dashboard:

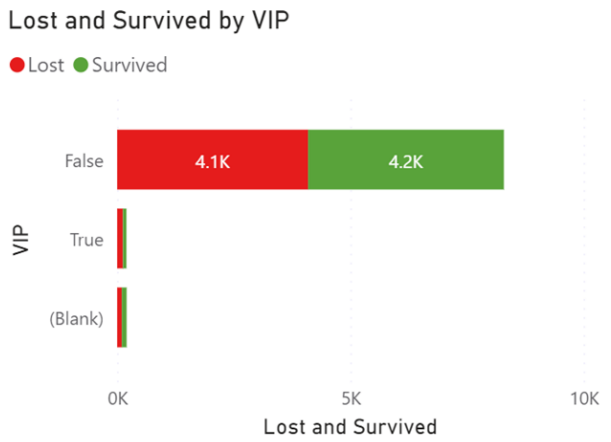


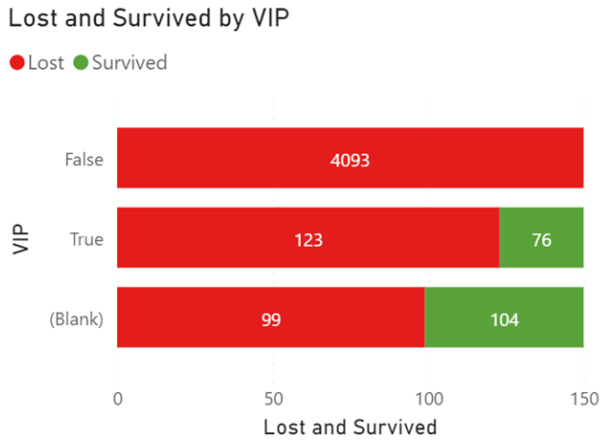
De meeste passagiers vertrokken van Aarde, daarnaast hangt de overlevingskans niet af van waar de passagier vertrok *(overal ongeveer 50%)* en er zijn 201 passagiers waarvan de vertrek locatie **niet** gekend is

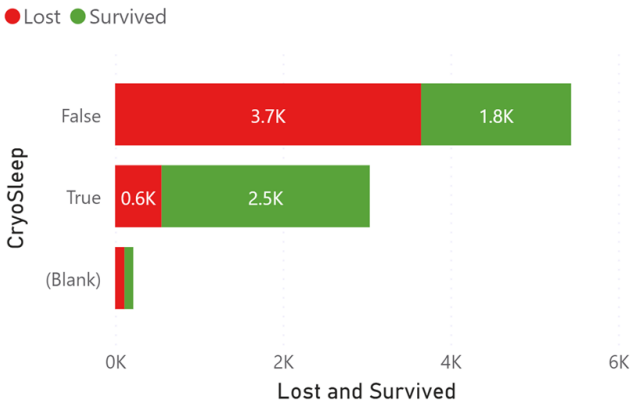


De meeste passagiers gingen naar Trappist. Ook hier hangt de overlevingskans niet af van waar de passagier naartoe ging en er zijn 182 passagiers waarvan de destinatie **niet** gekend is.

Tot onze verbazing is er een iets **lagere overlevingskans** met **VIP**, hiernaast zijn er ook 203 passagiers waarvan de keuze **niet** gekend is







Hier zagen we een interessant resultaat. We kunnen afleiden dat er een **aanzienlijk grotere overlevingskans** is als men koos voor **cryosleep**.

Data Cleaning:

We hebben afgeleid dat er een groot aantal waarden ontbreekt in rijen zoals: homeplanet, destination, roomservice, VrDeck, age, cryosleep…

Blanke waarden in een dataset zijn niet interessant omdat ze geen informatie bevatten die het model kan gebruiken om voorspellingen te doen. Als er te veel ontbrekende waarden zijn, kan dit een negatieve invloed hebben op de nauwkeurigheid van het model.

Homeplanet & Destination

In de kolommen "homeplanet" en "destination" kozen we er voor om elke planeet om te zetten in een uniek nummer. Dit is handig omdat veel machine learning algoritmes alleen met numerieke data kunnen werken. Vervolgens berekenen we het gemiddelde van de gecodeerde waarden om de ontbrekende planeten in te vullen. Met andere woorden, we vervangen de ontbrekende waarden door het gemiddelde van de nummers die de planeten in die kolom vertegenwoordigen.

Age

Voor age hebben we besloten om de blanke waarden simpelweg te vervangen door de gemiddelde leeftijd van alle passagiers.

Cyrosleep

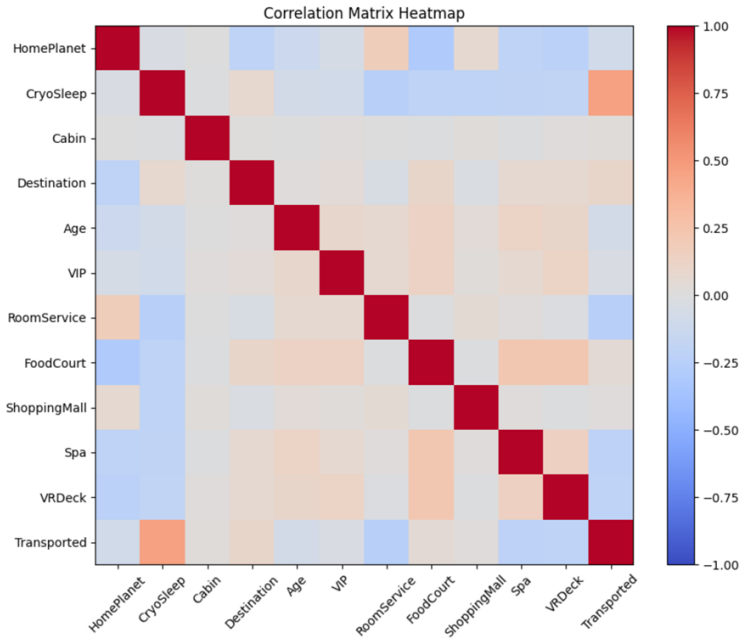
Bij cryosleep hebben we belist om alle ontbrekende waarden te vervangen door ‘False’ omdat deze het meest voorkomt en dus het minst zal doorwegen als we die er bij tellen.

Diensten (VRDeck, Spa, Foodcourt…)

Hier hadden we 2 scenario’s. Als de passagier koos voor cryosleep werdt de onbrekende waarde ingevulde door 0. Dit omdat de passagier tijdens die toestand onmogelijk gebruik kon maken van deze diensten.

Als de passagier niet in cryosleep toestand was, namen we het gemmidelde gebruik van de andere passagiers.

Correlatiematrix



Uit de corrolatie matrix kunnen we afleiden dat cryosleep positief gecorolleerd is met de overlevingskans. Dit betekend dat passagiers in cryosleep een grotere overlevingskans hadden dan de passagiers die wakker waren.

We kunnen ook afleiden dat het gebruik van diensten negatief gecorolleerd is met de overlevingskans. Dit betekend dat hoe minder er gebruik werd gemaak van een dienst, hoe groter de overlevingskans was. Dit viel ook te verwachten aangezien mensen in cryosleep geen diensten kunnen gebruiken.

Supervised Learning

In het laatste deel van ons project hebben we data gebruikt om een model te maken dat voorspellingen kan doen over testgegevens.

We hebben hiervoor verschillende classifiers geprobeerd en uiteindelijk gebruik gemaakt van de Random Forest-classifier omdat deze ons de hoogste nauwkeurigheid gaf.

Na het trainen van de gegevens hebben we het model gebruikt om te voorspellen of passagiers geteleporteerd werden of op het schip bleven en deze resultaten ingediend op Kaggle. Onze inzending scoorde **0.79728**.

Conclusie

Uit ons onderzoek blijkt dat er een aantal belangrijke factoren zijn die invloed hebben op de overlevingskansen van passagiers op Spaceship Titanic. Zo blijkt dat het gebruik van diensten aan boord negatief gecorreleerd is met de overlevingskans, terwijl de keuze voor cryosleep juist positief gecorreleerd is. Ook hebben we ontdekt dat het ontbreken van data een negatieve invloed kan hebben op de nauwkeurigheid van modellen die worden gebruikt om voorspellingen te doen over de overlevingskansen van passagiers.

We hebben verschillende methoden gebruikt om de dataset schoon te maken en te analyseren, en hebben uiteindelijk een Random Forest-classifier gebruikt om voorspellingen te doen over testgegevens. Onze inzending op Kaggle leverde een score van **0.79728** op.

Al met al denken we dat ons onderzoek waardevolle inzichten heeft opgeleverd voor toekomstige ruimtevluchten.

Reflectie

Tijdens dit project hebben we als team veel geleerd over het belang van data-analyse en machine learning met Python.

We hebben geleerd dat het ontbreken van data een grote invloed kan hebben op de nauwkeurigheid van voorspellingen en dat het belangrijk is om de dataset grondig te analyseren en te reinigen voordat we deze gebruiken om modellen te trainen.

We zijn trots op het resultaat en zijn blij dat we nieuwe vaardigheden hebben geleerd die we kunnen toepassen in toekomstige projecten.