Erfahrungsbericht

Data Analytics Bootcamp Amsterdam

Carolin Vogt

Felix Ley

Robert Meier

Motivation

Wir erzeugen Daten – jeden Tag, jede Minute, jede Sekunde. 2018 wurden 2,5 Trillionen Bytes Daten erzeugt pro Tag, Tendenz exponentiell steigend. 90% der weltweit verfügbaren Daten wurden in den vergangenen zwei Jahren erzeugt.

Die heutige datengetriebene Welt macht auch vor Produktionsprozessen keinen Halt. In unserem Studium haben wir immer wieder die verschiedenen Durchbrüche digitaler Innovationen behandelt, das Potenzial und die Notwendigkeit von Digitalisierung in allen Geschäftsbereichen eines Unternehmens nachvollzogen und die ungenutzten Möglichkeiten von Daten hinterfragt. Uns fehlte jedoch noch ein entscheidender Punkt, um die Zukunft aktiv mitzugestalten können: umfassende Programmierkenntnisse. Vor unserem Berufseinstieg und bestärkt durch einige Schwerpunktfächer in unserem Masterstudium Technologiemanagement am Campus Gummersbach, haben wir uns zum Ziel gesetzt, dass nicht nur Wissen über Lean Exzellenz Teil unserer Vita sein soll, sondern auch Data Exzellenz.

Wir entschieden uns dafür uns für ein Programmierbootcamp zu bewerben, um uns möglichst effizient die nötigen Skills anzueignen, um uns zu interdisziplinär ausgebildeten Fachkräften zu entwickeln. Wir setzten uns intensiv neben unserem Master mit dem Thema auseinander. Vor allem das Thema der Datenanalyse und die Gewinnung von Insights, unterstützt durch verschiedene Algorithmen faszinierte uns. Unweigerlich wird es in Zukunft mehr Bedarf an Fachkräften geben, welche in der Lage sind, komplexe Zusammenhänge zu verstehen und diese auch in Code umzusetzen. Große Datenmengen sind nutzlos, wenn sie nicht verarbeitet werden können. Nach einer erfolgreichen Bewerbungsphase und auch durch die Unterstützung des Vereines zur Förderung des Campus Gummersbach der Technischen Hochschule Köln e.V., konnten wir in den ersten drei Monate des Jahres 2021 das Ironhack Data Analytics Bootcamp in Amsterdam (remote) mit Studierenden aus ganz Europa absolvieren.

Ablauf und Alltag während des Camps

Bei einem Bootcamp handelt es sich um eine intensive Form des Lernens, mit dem Ziel, möglichst viel Inhalt in möglichst wenig Zeit zu vermitteln. Es ist üblich 10+ Stunden am Tag ausgelastet zu sein. Der zu vermittelnde Stoff wird Teilnehmenden bei Ironhack hauptsächlich durch praktische Übungen beigebracht. In den ersten Tagen entwickelte sich schnell eine Art Routine, die einen die Zeit vergessen ließ.

Ein typischer Bootcamp-Tag begann morgens um 9:00 Uhr vor dem heimischen Computer mit einer Tasse Kaffee. Die Vorlesungen fanden über Zoom statt, jede weitere Kommunikation erfolgte über die Plattform Slack (ähnlich MS Teams). In den ersten Minuten eines jeden Tages wurden aufkommende Fragen beantwortet sowie die Schlüsselelemente vom vorherigen Tag wiederholt. Anschließend startete der Unterricht.

Der Unterricht wurde von einem Head Coach geleitet, welcher Unterstützung von zwei Teacher Assistent erhielt. Üblicherweise gab es zu Beginn einer jeden Einheit eine kurze

Einführung in ein neues Thema. Diese kurzen Vorlesungen dauerten in den wenigsten Fällen länger als fünfzehn Minuten. Die vorgestellten Themen wurden im Anschluss in Form eines Beispiels aufgegriffen und wir konnten die neu erlernte Programmierbausteine wiedererkenne. Anschließend konnten wir in Einzelarbeit, das zuvor gezeigte Konzept anhand eines Beispiels eigenständig lösen. Üblicherweise erhielten wir dafür maximal zehn Minuten Zeit. Nach Ablauf der Zeit wurde eine zufällige Person ausgewählt und musste ihre Ergebnisse vorstellen. Sollten Fragen aufkommen, wurden diese direkt besprochen. Diesem Rhythmus gingen wir bis zur Mittagspause nach.

Um 12:00 Uhr wurde uns die tägliche Gruppenaufgabe des Mittags vorgestellt (Lab). Jeden Tag erhielten wir eine neue Gruppenzuordnung, in welcher wir die beiden Labs des Tages erfüllen sollten. Ein Lab bestand aus verschiedenen, meist etwas umfangreicheren Aufgaben, in denen wir die neu erlernten Inhalte anwenden sollten. Offiziell ging der Unterricht jeden Tag bis 16:00 Uhr mit anschließendem Bearbeiten des Nachmittag-Labs. Da in den meisten Fällen die Zeit mittags nicht ausreichte, um das Vormittags-Lab zu erledigen, mussten beide Labs nach Unterrichtsende bearbeitet werden. Bei den in Gruppenarbeit durchgeführten Labs, wechselten die Teacher Assistents regelmäßig durch die Gruppenräume, um bei Fragen zu unterstützen und Hilfestellung zu leisten. Trotz der virtuellen Veranstaltung war es möglich sich mit anderen Teilnehmenden anzufreunden.

Durch das gemeinsame Erarbeiten von Lösungen oder auch das gemeinsame Verzweifeln an Problemen, entstand ein ähnliches Gefühl wie bei Team Building. Unsere Klasse setzte sich aus internationalen Studierenden unterschiedlichen Backgrounds zusammen. Game Design, Marketing, Wirtschaftsingenieurwesen, BWL, Politikwissenschaft oder Biologie war vertreten.

Neben den Unterrichtstagen gab es auch verschiedenen Projektphasen während des Camps. Projektphasen zeichneten sich dadurch aus, dass der reguläre Unterricht durch Einzel oder Gruppenprojekte ersetzt wurde. Es gab ein kurzes Gruppenprojekt über zwei Tage (*Ironhack FIFA Challenge)*, ein einwöchiges Gruppenprojekt (*Credit Card Marketing – A Classification Challenge*), ein Einzelprojekt *(Spotify – Build your own song recommende*r) und ein finales Einzelprojekt, dessen Thema während der ersten Woche des Bootcamps definiert und in der letzten Woche final bearbeitet wurde. Die Projekte werden im weiteren Verlauf dieses Berichtes genauer vorgestellt.

Neben all dem Lernen blieb jedoch auch noch Zeit für ein wenig Ablenkung sowie gemeinsame Aktivitäten. Jeden Freitag wurde der Nachmittag dafür genutzt den Team Feedback zu geben und vor allem, um gemeinsam diverse (online – dank der COVID Situation) Spiele zu spielen. Der absolute Favorit war ‘Scribbl’. (https://skribbl.io). Da wir uns bereits vorher kannten, hatten wir die Motivation, zusammen eine online Party zu organisieren. Als wir die Hälfte der Bootcamp-Tage geschafft hatten, gab es das von uns organisierte Bergfest. Wir organsierten kleine Spielchen mit allen Teilnehmenden. So musste uns jeder ein Bild von sich als Kleinkind zukommen lassen. Dieses Bild wurde dann in einer Präsentation allen gezeigt mit der bitte zu erraten, um wen es sich handelt. Durch solche und ähnliche Spiele konnte man sich noch besser kennen lernen.

Durch Schweiß, Nerven und viel Spaß gelang es uns das Camp souverän zu meistern und uns eine enorme Menge Wissen in kurzer Zeit anzueignen. Dies war nur durch das gut durchdachte Konzept möglich, welches von Anfang verstand, uns praktisch in die Themen einzuarbeiten.

Projekte während des Camps

Quelle: https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.redbull.com%2Fint-en%2Ffifa-17-new-features-red-bull-games&psig=AOvVaw3G9UBX7OKp_zntTxqFwfpv&ust=1629321526603000&source=images&cd=vfe&ved=0CAoQjRxqFwoTCNjW6ur9uPICFQAAAAAdAAAAABAD

Projekt Woche 1: FIFA-Challenge

Die erste Woche des Kurses war extrem anstrengend. Wir haben die Grundlagen der Linearen Regression erlernt, um am Ende der Woche direkt ein eigenes Projekt umzusetzen. Wir erhielten einen Datensatz aus dem Spiel “FIFA 2017”, welchen wir analysieren mussten. Wir sollten in Teams zusammenarbeiten und verschiedene Funktionen erstellen, mit welchen wir die vorkommenden Datentypen des Datensatzes anpassen können. Durch die Teamarbeit merkte man schnell, dass es beim Programmieren auch sehr viel auf zwischenmenschliche Fähigkeiten ankommt, da eine effektive Zusammenarbeit sonst nicht möglich ist. Ziel des Projektes war es ein Modell zu trainieren, welches es einem ermöglicht den “Overall Score” eines jeden Spielers vorherzusagen. Jeder Spieler hat Positionen, welche er am besten beherrscht sowie diverse individuelle Fähigkeiten. Aus diesen Faktoren errechnet das Spiel für den Spieler einen Gesamtscore. Wir haben mit unserem Modell versucht diese Berechnung zu imitieren und den Zusammenhang der Daten zu ergründen.

CREDIT CARD MARKETING

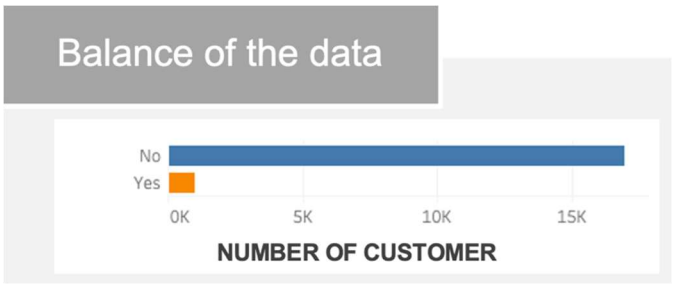
classification challenge

Projekt Woche 5: Credit Card Marketing

Für Woche fünf des Bootcamps war eine Projektwoche angesetzt, in welcher die Teilnehmenden in Gruppen zu je drei Personen eine Lösung für einen bestimmten Use Case erarbeiten sollten. Die Gruppen konnten einen Use Case aus den Kategorien der Regression oder Klassifikation wählen. Wir drei bildeten gemeinsam eine Gruppe und wählten das Projekt aus der Kategorie Klassifikation.

Die Aufgabenstellung des Use Case sah vor, als Berater einer Bank dabei zu helfen, die Ergebnisse einer Kreditkarten Marketingkampagne für Bestandskunden auszuwerten. Dabei wurden Daten von 18.000 Bestandskunden gesammelt, in denen verschiedene Angaben zur Person, als auch zur Reaktion auf die Marketingkampagne festgehalten wurden. Ziel war es zunächst der Bank ein besseres Verständnis für ihre Kunden zu erarbeiten, indem die stärksten Einflussgrößen für eine Annahme des Angebotes analysiert wurden.

Im nächsten Schritt wurde dann ein Machine Learning Modell trainiert, welches Anhand der gepflegten Daten für jeden Kunden individuell ermittelt, ob eine höhere Wahrscheinlichkeit herrscht das der Kunde das Angebot annimmt oder ablehnt. Diese Erkenntnis kann dann bei zukünftigen Kampagnen genutzt werden, um kostenoptimiert und gezielter Angebote an die Kunden heranzutragen und diese nicht mehr pauschal an jeden Kunden zu übermitteln.



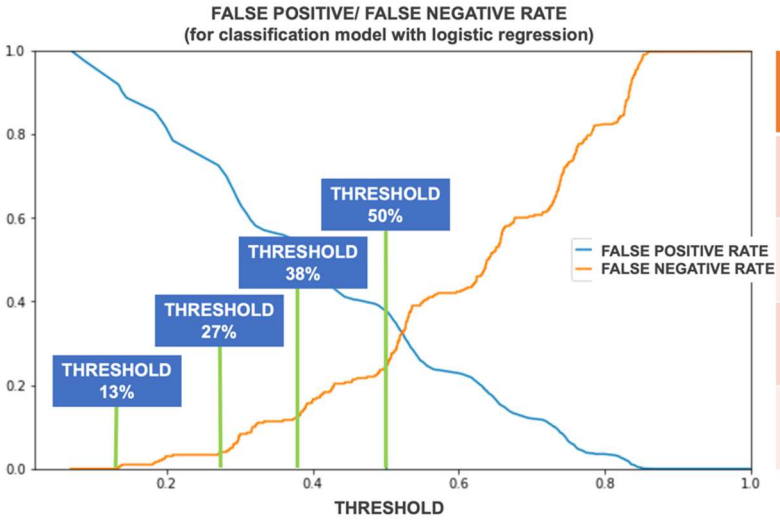
Dadurch, dass in der vorangegangenen Marketingkampagne nur jeder 18. Kunde das Angebot angenommen hatte, wurde das Modell mit wesentlich mehr Daten von Kunden gepflegt, die das Angebot ablehnten. Dies führte dazu, dass das Modell bei Performancetests an den bestehenden Daten zunächst nicht alle Kunden identifizieren konnte, welche das Angebot angenommen hatten. Somit wären durch das Modell bei zukünftigen Marketingkampagnen gegebenenfalls vielversprechende Adressaten für das Angebot von der Marketingkampagne ausgeschlossen worden und Umsätze eingebüßt.

Um diese Quote von zu Unrecht negativ bewerteten Kunden du minimieren, haben wir unseren Algorithmus dahingehend angepasst, dass automatisiert der optimale Schwellenwert

für eine Klassifizierung identifiziert werden kann, um jeden potenziellen Interessenten zu erreichen. Für eine Klassifizierung in positiv oder negativ wäre somit nicht mehr eine Wahrscheinlichkeit von >50% ausschlaggebend, sondern würden Kunden in unserem Fall schon bei einer Wahrscheinlichkeit von >14% als potenzieller Kunde klassifiziert und somit ein Angebot erhalten.

MODEL IMPROVEMENT

Adjusting the threshold is always a trade-off!



THRES-HOLD	FALSE NEGATIVE	FALSE POSITIVE	TRUE NEGATIVE
starting point (50%)	244	6,425	10,574
1st plateau (38%)	134 (- 110)	9,189 (+ 2,764)	7,810 (- 2,764)
2nd plateau (27%)	37 (- 97)	12,296 (+ 3,107)	4,703 (- 3,107)
absolut 0 (13%)	0 (- 37)	15,924 (+ 3,628)	1,075 (- 3,628)

ACHIEVEMENTS & ACTIONS

Clues for improving the marketing campaign for credit cards

ACHIEVEMENTS

Identification of customer features relevant to accept a credit card offer

Changing the mailer type to postcard exclusively.

Development of a machine learning classification model

(1)
Identification of customer most likely to reject the offer
- *Reducing Marketing effort by 6%*

(2)
Tool to adjust the threshold
- *Trade-off between marketing costs and offer acceptance rate*

ACTIONS

Uncertain forecast due to high imbalanced data

Further feature engineering could reduce false positive rate

Diese auf den Use Case abgestimmte Anpassung führte dazu, dass wir von den anderen Gruppen und Lehrenden zur besten Gruppenleistung gewählt wurden.

Die Nutzung solcher Modelle in realen Marketingkampagnen kann dabei helfen, ein Versenden von Angeboten an Kunden, die das Angebot sicher ablehnen werden zu reduzieren und somit Geld und Ressourcen effizienter zu nutzen.

Quelle: https://www.elevationcapital.co.nz/blog/2020/7/5/fund-update-spotify

Projekt Woche 6: Building your own Spotify Song Recommender

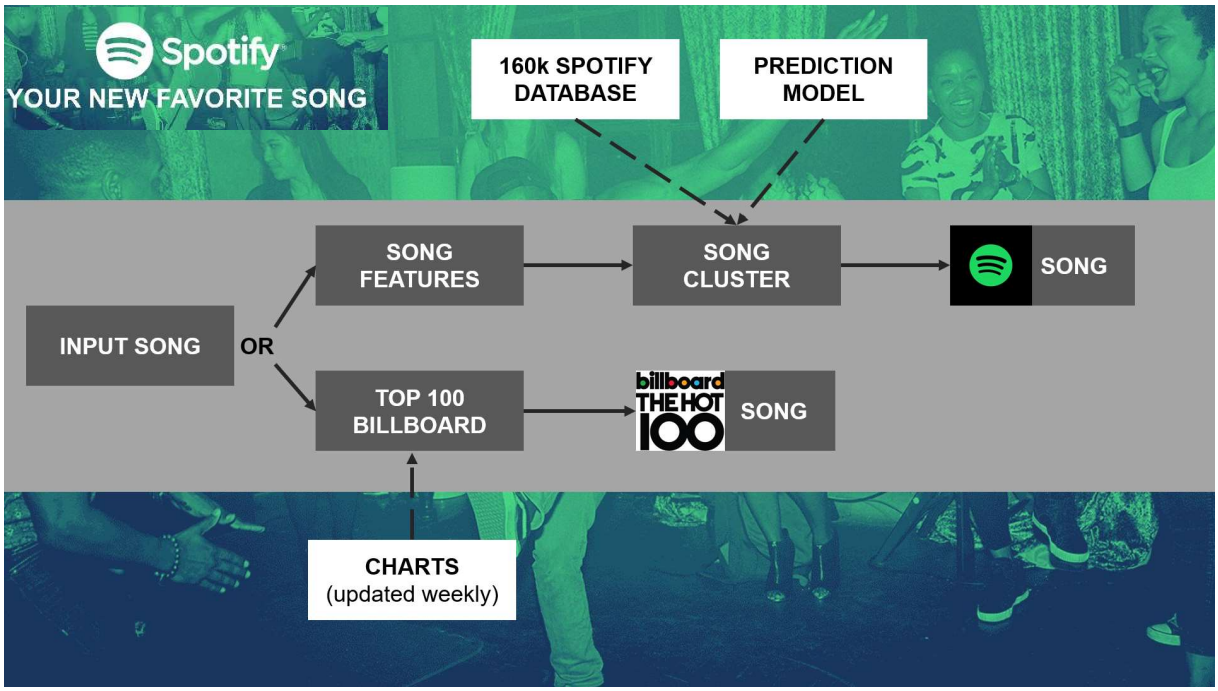
Spotify verzeichnet 365 Millionen monatliche Nutzer:innen, davon 165 Millionen Premium-Nutzer:innen und ist damit die erfolgreichste Musikstreaming Plattform. Warum bekommen wir (scheinbar) immer zur richtigen Zeit den richtigen Song vorgeschlagen?

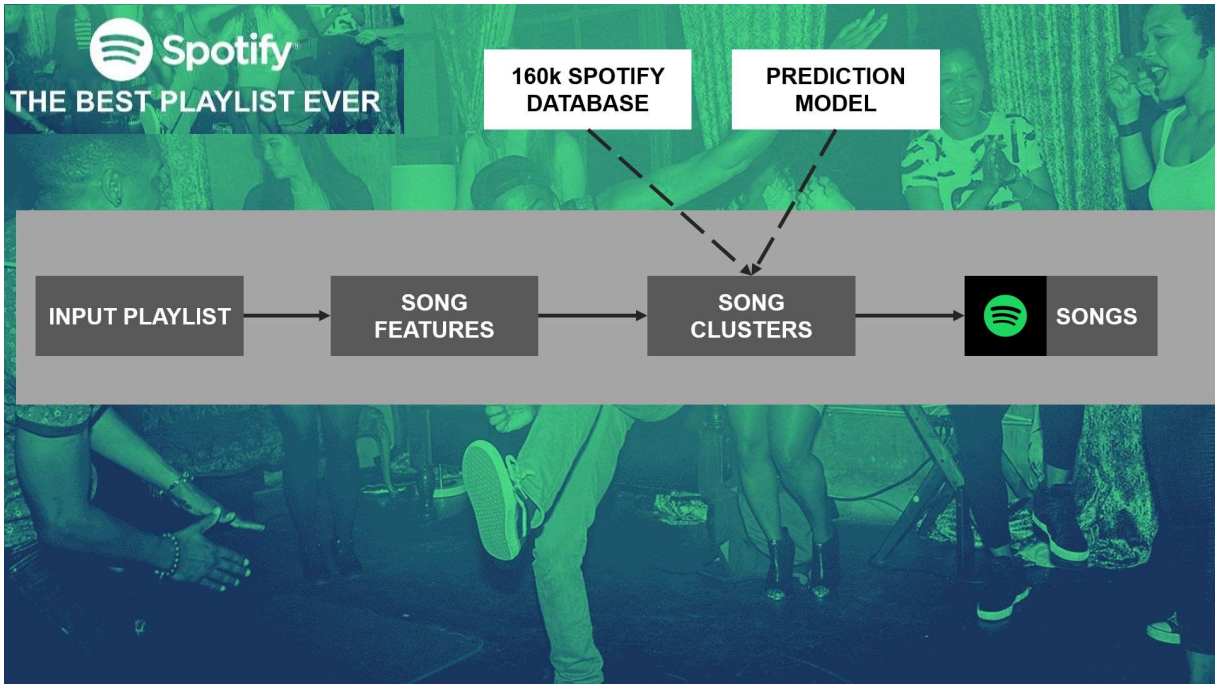
In unserem Projekt in Woche 6 des Ironhack Bootcamps sind wir dieser Frage ein wenig auf den Grund gegangen. Ziel war es eine eigene Song-Datenbank aufzubauen, jeden Song über Features zu beschreiben sowie ein Machine Learning Modell zu erstellen, das dem User neue Songs vorschlägt und dafür eine Maske zu entwickelt, die als Minimum Viable Product (MVP) dienen soll.

Für die Datenbank haben wir unter anderem die Billboard Top 100 Charts mithilfe von Web Scraping ausgelesen, eine 160k Songs große Datenbank von Kaggle und unsere eigenen Spotify Playlists eingelesen. Jeden Song mit Features zu beschreiben, stellte uns vor eine größere Herausforderung. Dazu haben wir die Spotify Web API (Programmierschnittstelle) SpotiPy genutzt. Falls Sie mit APIs nicht vertraut sind: Eine Anwendungsprogrammierschnittstelle ist im Grunde ein Server, auf den Sie zugreifen können, um Daten zu erhalten und zu senden. Im Fall von Spotify bietet das Unternehmen Software- und App-Entwickler:innen über eine Web-API Zugriff auf einige seiner Daten über Nutzer, Wiedergabelisten und Künstler. Und eben auch auf die Features, mit denen Spotify jeden Song beschreibt:

Diese Features sind zur Anwendung eines *Clustering Machine Learning Algorithm* sehr hilfreich. In diesem Projekt haben wir den *K Nearest Neighbors (KNN)* Algorithmus für das Clustering der Songs verwendet. Technisch gesehen klassifiziert der Algorithmus ein unbekanntes Objekt, indem er k seiner bereits klassifizierten, nächsten Nachbarn betrachtet und indem er die nächsten Nachbarn ermittelt, die ähnliche Features wie das unbekannte Objekt haben. Der KNN-Algorithmus ist eine gute Wahl, wenn ein kleiner Datensatz vorliegt und die Daten gelabeled und diese Label ohne große Fehlerrate sind. KNN wird unbrauchbar, sobald der Datensatz groß wird – etwas mehr als 160k Songs sind hierbei im Kontext von Data Science keine großen Datenmengen.

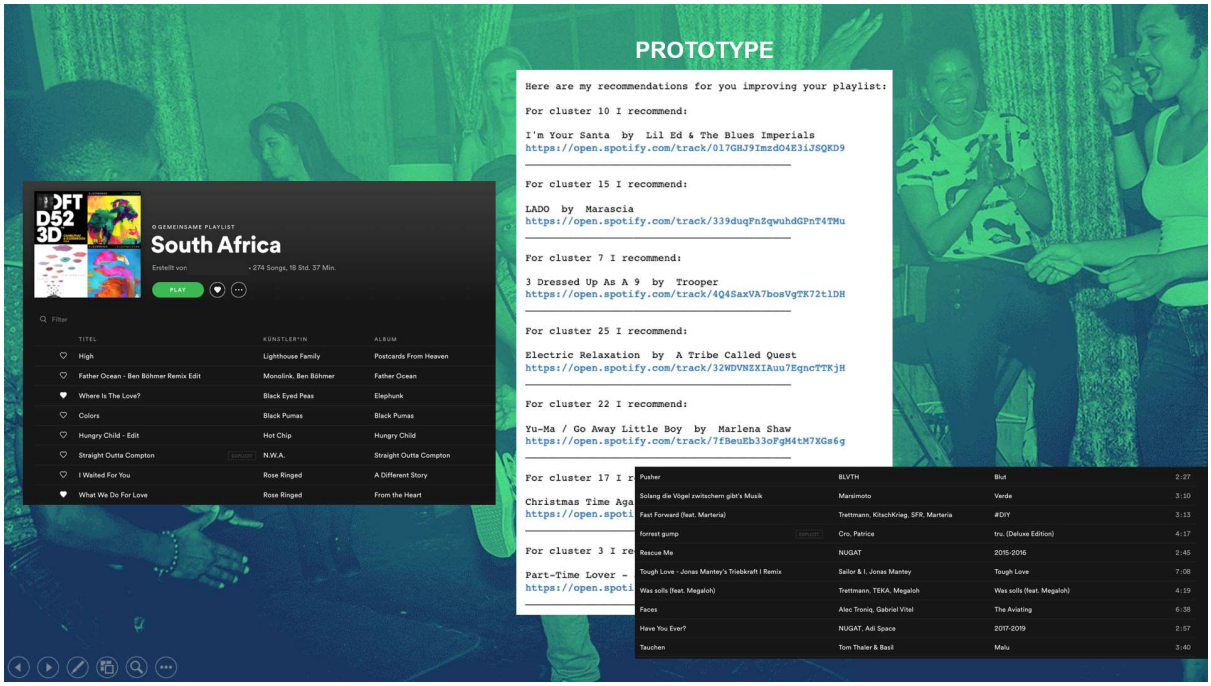
Das trainierte Machine Learning Modell kann anschließend beispielsweise eine persönliche Playlist analysieren und favorisierte Genres (Cluster) identifizieren. Der Abschluss des Projektes bestand in der Präsentation des *Song Recommenders* als MVP. Das Konzept des Song Recommenders auf Basis eines einzelnen Songs oder sogar einer gesamten Playlist ist unten dargestellt:





Zu der Präsentation des MVP Song Recommender gehört außerdem ein Prototyp. Folgende Funktionalitäten waren zu finden:

- Eingabe eines Songs Titels mit oder ohne Artist Name ODER Eingabe eines Playlist-Links
- Abfrage und Bestätigung des richtigen Künstlers durch den User, wenn der Song Titel häufiger bei Spotify gelistet ist
- Korrektur von Rechtschreibfehlern
- Ein neuer Songvorschlag für das identifizierte Cluster des Songs ODER ein Songvorschlag für jedes identifizierte Cluster in der eingegebenen Playlist
- Ausgabe des Song Titels & Artist Namens sowie der Link zu Spotify zum direkten Hören



Unsere Abschlussprojekte

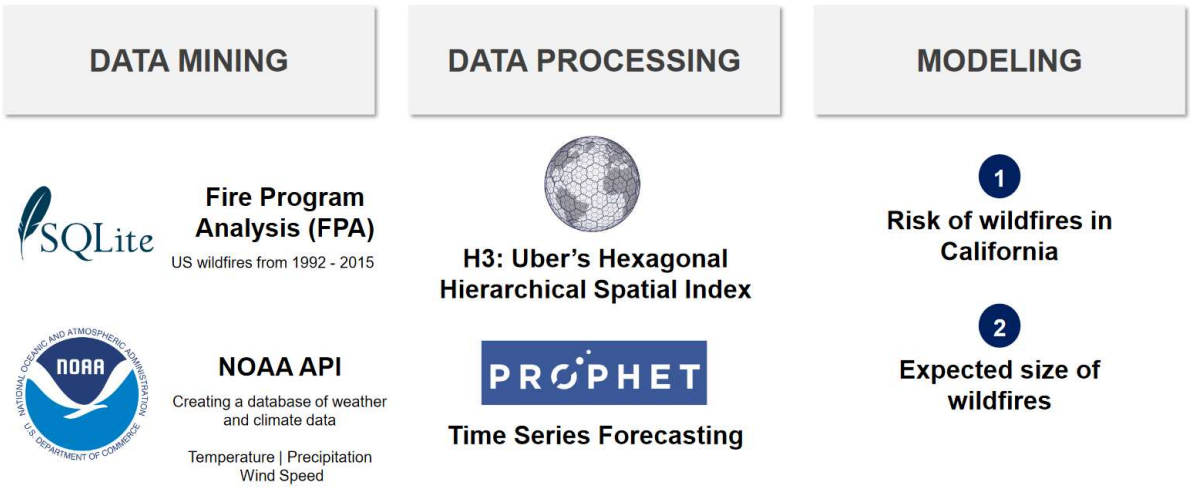
In der letzten Woche des Bootcamps ist für alle Teilnehmenden vorgesehen, dass diese ein FINAL PROJECT Ihrer Wahl ausarbeiten, um gelerntes Wissen und Methoden unter Beweis zu stellen. Neben der reinen Anwendung des vermittelten Stoffs der vorherigen Wochen, lag die Herausforderung in der offenen Aufgabenstellung somit auch in der Auswahl des Projektes und einzuschätzen, was möglich ist und was eher nicht. Dabei wurden teils eigene Interessen verfolgt, Hypothesen aufgestellt und untersucht oder mit Datensätzen aus realen Unternehmen wertvolle Einblicke generiert.

Die Waldbrände an der amerikanischen Westküste nehmen stetig zu, vor allem in Kalifornien. Vielen sind die Bilder des glühenden Himmels über San Francisco oder der Golden Gate Bridge noch präsent. Die Waldbrandsaison 2020 hat die bisher schlimmste Waldbrandsaison 2008 deutlich übertroffen. Bis Ende 2020 hatten 9.639 Brände 4.397.809 Hektar Land verbrannt - mehr als 4 % der kalifornischen Landfläche.

Nicht nur steigt die allgemeine Waldbrandgefahr, die Waldbrandsaison verlängert sich und die Anzahl großer Feuer steigt. Zwischen 1992 und 2015 wurden mehr als 1,8 Millionen Waldbrände in den USA gezählt. Die Zahl der Brände zwischen 4 und 20 km^2 nahm in diesem Zeitraum deutlich zu. Der Einfluss des Klimawandels ist deutlich zu erkennen. Angesichts der vielfältigen Auswirkungen des Klimawandels stellt dieses Projekt die Frage: "Is fire season now a year round reality?".

Konzept

Im Final Project wurde ein Maschine Learning Model zur Vorhersage von Waldbränden in Kalifornien entwickelt. Das Projekt umfasst die Arbeit mit einer Datenbank von etwa 190.000 aufgezeichneten Waldbränden zwischen 1992 - 2015, die Auswertung von Klimadaten aus mehr als dreißig Jahren und die Einordnung des Ganzen in das Gebiet der San Francisco Bay Area.



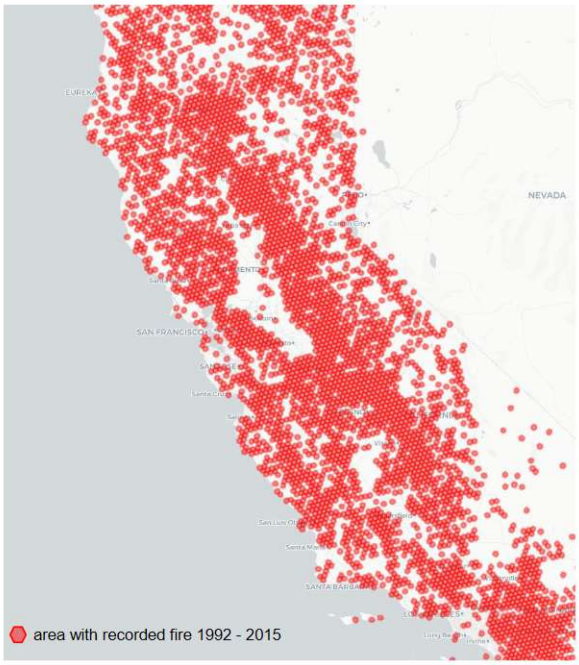
Phase 1: Data Mining

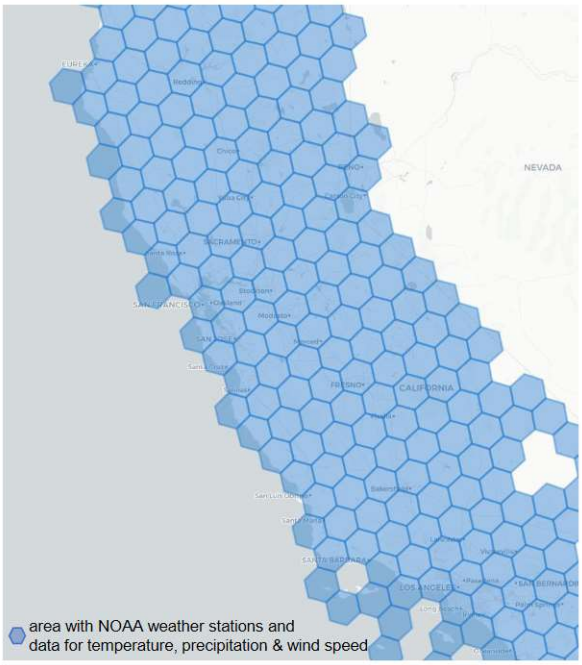
In der ersten Projektphase ging es um die Beschaffung relevanter Daten. Die Erweiterung der Waldbrand Datenband des (FPA) stand hier im Vordergrund. Die National Oceanic & Atmospheric Administration (NOAA) stellt in ihrem National Centers for Environmental Information (NCDC) Portal aktuelle sowie historische Wetter- und Klimadaten zur Verfügung. Für den Zeitraum der aufgezeichneten Waldbrände zwischen 1992 – 2015, sowie für die Jahre 2016 – 2020 wurden für Wetterstationen in Kalifornien Daten zu Temperaturen, Windgeschwindigkeiten, Niederschlag und Verdunstung über diese API (dt.: Programmierschnittstelle) gesammelt. Ziel des Projektes war es zusätzlich für die Jahre 2021 – 2016 einen Trend zu identifizieren, um die Rolle des Klimawandels einzuordnen. Um ein für ein zuverlässigeres Time-Series-Forecasting-Model für die Jahre 2021-2026 zu erstellen, wurde zusätzlich die historischen Daten von 1985 – 1991 gesammelt. Die Dokumentation der NOAA API finden Sie hier.

Die Abfrage der ausgewählten Klimadaten pro Zeitraum ist mit denen im Kurs gelernten Tool des Web Scraping automatisiert und zur weiteren Verwendung pro verfügbare Wetterstation (4427 eindeutige Stations-IDs) gespeichert.

Phase 2: Data Processing

Im nächsten Schritt wurden die Standorte der Wetterstationen mit den lokalisierten Waldbränden zusammengeführt (Data Processing). Orte, an denen Waldbrände zwischen 1992 – 2015 ausgebrochen sind erstrecken sich über den ganzen Bundesstaat. Die Größe der Waldbrände ist in dieser Karte nicht zu erkennen. (Beinahe) für die gesamte Fläche Kaliforniens stehen Klima- und Wetterdaten zur Verfügung.





Wie wurden die Wetterstationen und Waldbrände kartiert?

H3: Uber's Hexagonal Hierarchical Spatial Index ist ein Open Source Project von Uber und ermöglicht es, geografische Informationen zu analysieren. Die grundlegenden Funktionen der H3-Bibliothek ist die Indizierung von Orten, wobei Breiten- und Längengradpaare in einen 64-Bit-H3-Index (Hexagon) umgewandelt werden. Die H3 Bibliothek enthält verschiedene Funktionen. Eine Funktion nimmt einen Breitengrad, einen Längengrad und eine Auflösung (zwischen 0 und 15, wobei 0 für die gröbste und 15 für die feinste Auflösung steht) entgegen und gibt einen Index zurück. Über diesen Index können unterschiedliche Orte in ein Hexagon geclustert werden. Die Dokumentation für *H3: Uber's Hexagonal Hierarchical Spatial Index* Projekt finden Sie hier.

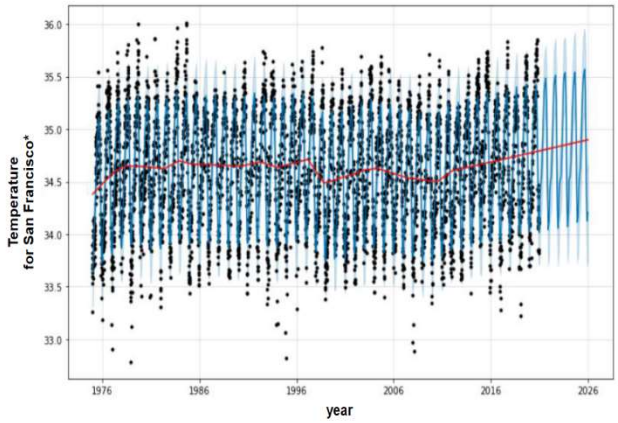
Im Final Project wurde die Uber H3-Bibliothek für die Indizierung der Waldbrände und der Wetterstationen verwendet. Die interaktive Karten der Waldbrände in der Auflösung 5 und der Wetterstationen in der Auflösung 4 steht über GitHub zur Verfügung. Für das weitere Projekt wurden diese Hexagons als Standorte verwendet. Sowohl die unterschiedlichen Wetterdaten (Temperatur, Windgeschwindigkeit, Niederschlag und Verdunstung) also auch die Anzahl und Größe der ausgezeichneten Waldbrände wurden auf dieser Ebene zusammengeführt.

Neben der Lokalisierung der Waldbrände und Wetterstationen musste für ein Time-Series-Forecasting-Modell außerdem eine Zeitreihenanalyse durchgeführt werden. Es gibt viele verschiedene Ansätze zu Zeitreihenanalyse, in diesem Projekt wurde, auch aufgrund der kurzen Zeit, *Prophe*t von Facebook verwendet. Facebook beschreibt dieses Open Source Projekt wie folgt:

> *„Prophet ist ein Verfahren zur Vorhersage von Zeitreihendaten auf der Grundlage eines additiven Modells, bei dem nichtlineare Trends mit jährlicher, wöchentlicher und täglicher Saisonalität (…) angepasst werden. Es funktioniert am besten mit Zeitreihen, die starke saisonale Effekte und mehrere Saisons historischer Daten aufweisen.*

Prophet ist robust gegenüber fehlenden Daten und Trendverschiebungen und kommt in der Regel gut mit Ausreißern zurecht.“

Prophet ermöglicht es Data Scientists und Data Analysts das automatisierte Forecasting manuell an die eigenen Anforderungen anzupassen. Die Dokumentation des *Facebook Prophet* Projektes finden Sie hier.



Für jedes Hexagon und die unterschiedlichen Wetterparameter wurde das Time-Series-Forecasting durchgeführt.

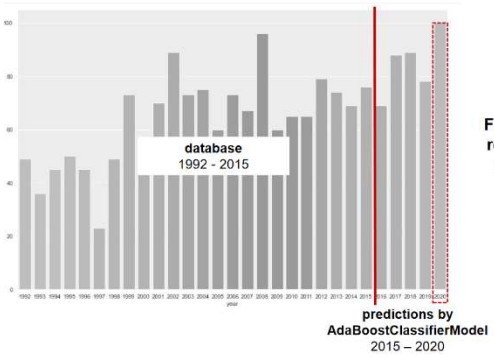
Wie sich die Temperatur in den nächsten fünf Jahren in San Francisco verändern kann, kann der nebenstehenden Abbildung entnommen werden. Die Steigerung der Durchschnittstemperatur ist deutlich zu erkennen.

Phase 3: Modelling

In diesem Projekt wurde ein Two-Step-Approach zur Vorhersage von Waldbränden in Kalifornien verfolgt. Im ersten Schritt des Models wird die Wahrscheinlichkeit berechnet, dass in einem bestimmten Gebiet (Hexagon) in einem bestimmten Monat ein Waldbrand ausbrechen wird. Hier handelt es sich um ein *Classification Problem.* Im Ironhack Bootcamp haben wir einige Machine Learning Modelle kennengelernt, mit denen solche Probleme gelöst werden können.

Ein Schritt, um ein zuverlässiges Machine Learning Modell zu erstellen, ist das sogenannte Feature Engineering. Neben den monatlichen Wetterdaten wurden Klimadaten des letzten Quartals und Jahres berechnet, Gebiete mit hoher Bevölkerungsdichte identifiziert (urban areas) und die Brandgrößenklasse reduziert. Weitere Features wurden bei der Automatisierte Merkmalsauswahl (*Automated Feature Engineering*) aussortiert, Ziel dieses Schrittes ist die Verringerung der Anzahl der Merkmale. Außerdem wurden Wettervariablen als Hauptmerkmale identifiziert.

Der *AdaBoostClassifier* hat für die Berechnung der Wahrscheinlichkeit von Waldbränden in einer bestimmten Region die beste Performance erreicht. Die Genauigkeit des Modells im TEST-Set beträgt: 0.97 und die Kappa des Modells in der TEST-Sets beträgt: 0.894. Hierbei hat insbesondere die Reduzierung der Brandgrößenklassen innerhalb des Feature Engineerings zu einer Erhöhung des Recall-Scores geführt (Reduzierung der falsch-negativen Vorhersagen des Modells). Für den zweiten Schritt des Vorhersagemodells zu der Größe ausgebrochener Waldbrände konnte innerhalb des Final Project kein zufriedenstellendes Modell entwickelt werden. Für die Vorhersage der Größe von Waldbränden war die Unausgewogenheit der Größe der registrierten Waldbrände eine Herausforderung.



For 2020, the model predicts a record number of fires for the area around San Francisco

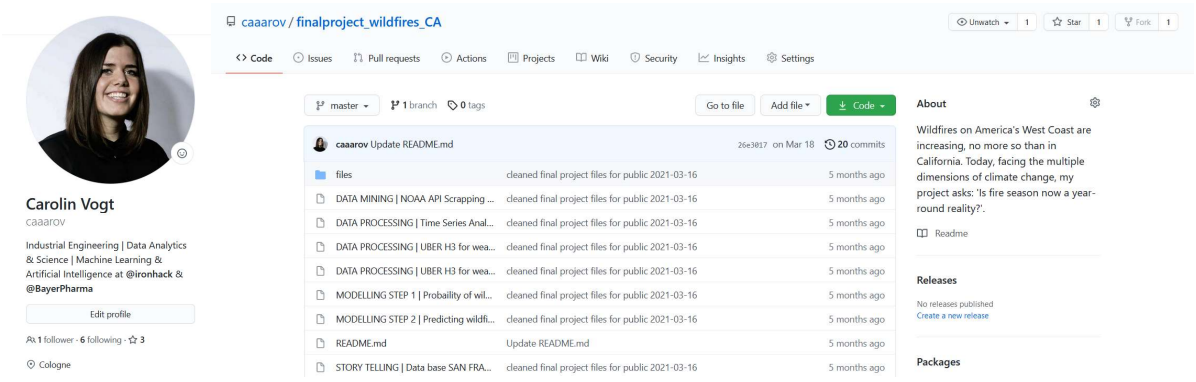
Für die San Francisco Bay Area hat das entwickelte Machine Learning Model eine Rekordzahl von Waldbränden vorhergesagt. Dies deckt sich mit der erschreckenden Waldbrandsaison, die Kalifornien 2020 erlebt hat.

Ausblick

Folgende Ideen wurden entwickelt, um das Projekt weiterzuentwickeln:

- Einrichten einer Live-Verbindung zur NOAA API, um das Modell mit Live-Klima- und Wetterdaten auf dem neuesten Stand zu halten
- Hinzufügen weiterer Funktionen mit Informationen über das Ökosystem und die Artenvielfalt der Hexagone
- Kennzeichnung von Sechsecken mit einem hohen Prozentsatz an bewaldeten Flächen
- Verwendung von Stufe 5 oder 6 für die H3-Indizierung für eine höhere Granularität
- Verwendung von NN, um "versteckte Merkmale" zu finden (z. B. staatliche Vorschriften in bestimmten Zeiträumen zur Verringerung des Waldbrandrisikos)

Den gesamten Code des Projektes und die Visualisierungen der verwenden Daten finden Sie in meinem GitHub Account:



Abwasseraufbereitung von Felix & Robert

Jährlich werden ca. 10 Milliarden Kubikmeter Abwasser in Deutschland aufbereitet. Die Aufreinigung des Abwassers ist mit einem hohen Energie- und Betriebsmitteleinsatz und damit auch hohen Kosten für die Betreiber verbunden. Abwassertechnische Anlagen gehören zu den elektrischen Großverbrauchern. Die von den ca. 10.000 deutschen Kläranlagen verbrauchte Strommenge, ist für die Emission von rund 3 Millionen Tonnen CO^2 verantwortlich. Die Säuberung des Abwassers ist mit einem hohen technischen Aufwand verbunden. Die teils hochkomplexen Reinigungsprozesse produzieren eine große Menge an Prozessdaten, welche von den Mitarbeitern dafür genutzt werden, die Anlagen zu steuern.

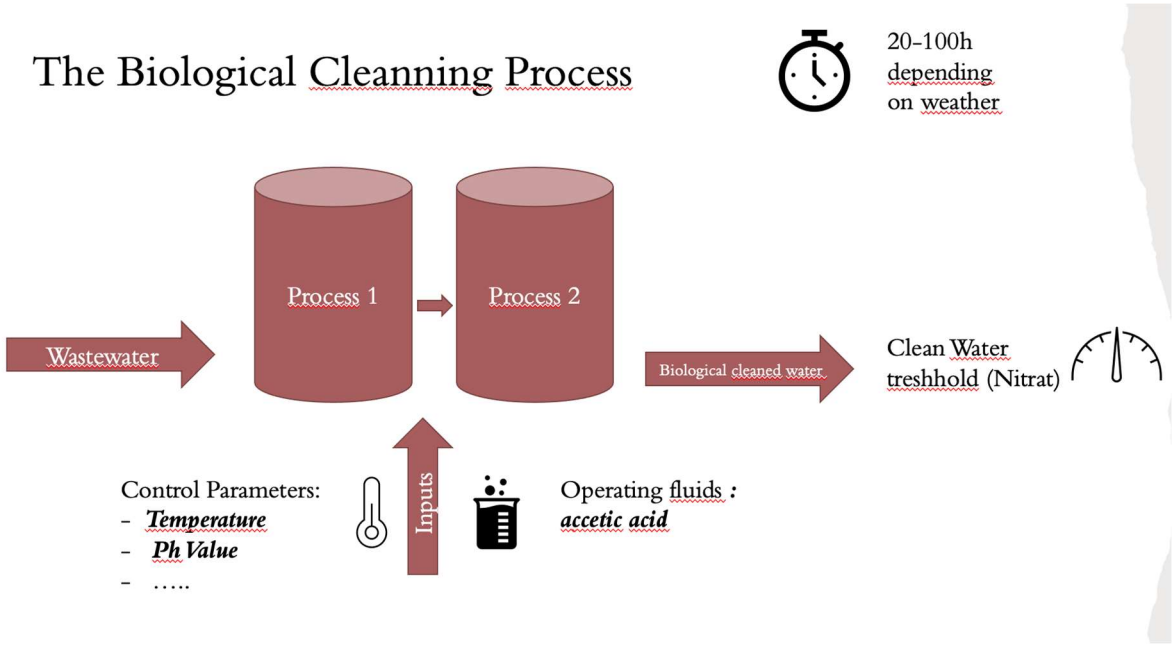
Eine solche Kläranlage ist die Deponie Leppe welche seit 1982 von dem Bergischen Abfallwirtschaftsverband (BAV) betrieben wird. Die Deponie erstreckt sich über eine Fläche von 45 Hektar und hat ein Volumen von insgesamt 10 Millionen Kubikmetern. Zur Deponienachsorge gehört neben der Oberflächenabdichtung, auch die Deponiegasproduktion sowie die Sickerwasseraufbereitung. Die täglich anfallenden Daten werden von den Mitarbeitern nur teilweise ausgewertet und verarbeitet.

Das im Jahr 2021 gegründete Kölner Start-Up nerou, plant eine auf Machine Learning Algorithmen basierende Datenanalyse-Software für Abwasserbetriebe zu entwickeln. Die Anlage des BAV ist Pilotkunde von nerou. Die speziell auf die biologische Aufbereitungsstufe angepasste Software, soll eine wirtschaftliche und nachhaltige Alternative zur optimierten Steuerung der Anlage bieten. Insbesondere die Biologische Abwasserbehandlung stellt die Betreiber durch ihre schwer zu skalierenden komplexen biologischen Prozesse vor Herausforderungen.

Felix und Robert sind für ihr finales Projekt mit Nerou in Kontakt getreten, um eine möglichst industrienahe Aufgabe zu bearbeiten. Da die Abwasseraufbereitung in zwei voneinander unabhängige Probleme geteilt werden kann, konnten beide an individuellen Projekten arbeiten.

Zielsetzung Biologische Wasseraufbereitung

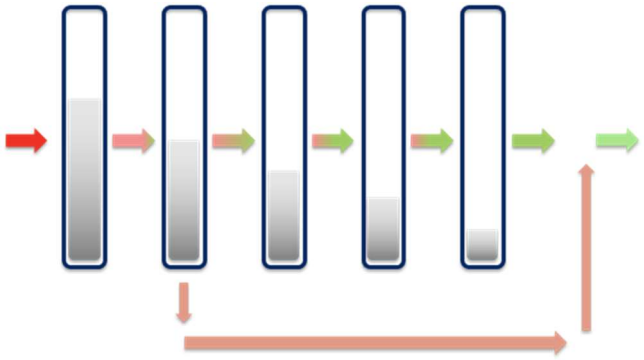
Die Biologische Abwasseraufbereitung ist ein hochkomplexer biologischer Prozess. Die in geschlossenen Behältern stattfindenden chemischen und biologischen Reaktionen unterliegen starken natürlichen Schwankungen. Durch die Farbe und Viskosität der durchfließenden Flüssigkeit gestaltet sich das physische Messen durch Sensoren bei vielen gewünschten Werten als schwierig bis unmöglich. Des Weiteren lassen sich die gewohnten Verhaltensweisen aus Laborversuchen schwierig auf einen hochskalierten Industrieprozess übertragen. Die in die Anlagensteuerung verbauten Regelkreise bilden das Geschehen innerhalb der spezifischen Anlage bei der Deponie Lippe nur unzureichend ab. Da eine jede Abwasseraufbereitungsanlage mit leicht unterschiedlichen Zusammensetzungen des zu reinigenden Abwassers sowie einer anderen Anlagenkonstellation arbeitet ist eine auf festen Grenzen basierende Steuerung nicht ausreichend für eine adäquate Steuerung. Ziel des Projektes ist es, mit Hilfe der historischen Daten die Zusammenhänge in der Anlage durch einen Algorithmus abzubilden. Die Prognosen der zu erstellenden Modelle sollen die Höhe des Grenzwertes vorhersagen, um die Zugabe von Betriebsmitteln dementsprechend anzupassen.



Durch das Auswerten der historischen Daten habe ich versucht die wichtigsten Betriebsparameter zu identifizieren welchen Einfluss auf den Grenzwert haben. Durch dieses vorsortieren reduziert man die benötige Rechenzeit später deutlich und eine Interpretation der Ergebnisse wird vereinfacht. Anschließend nutzte ich das Wissen aus dem Bootcamp und versuchte mit Hilfe von verschiedenen machine-learning Modellen den Zusammenhang in den Daten zu analysieren. Nach einigen Experimenten hatte ich meine ersten Modelle, die mir den zukünftigen Nitrat-Grenzwert vorhersagen konnten. Auf Basis dessen kann eine Betriebsmitteländerung vorgenommen werden.

Zielsetzung Mechanische Wasseraufbereitung:

Die mechanische Wasseraufbereitung folgt auf die biologische Wasseraufbereitung. Dabei wird das Abwasser durch eine Filterstraße mit 5 Aktivkohlefiltern geleitet. Da das Abwasser beim Durchlaufen aller fünf Filter so sauber wäre, dass der gesetzliche Grenzwert bei Weitem unterschritten würde, existiert ein Bypass nach Filter 2. Nach Filter 2 wird somit eine gewisse Menge Abwasser mit stärkerer Kontamination abgeleitet und mit sehr schwach kontaminiertem Abwasser nach Filter 5 vermischt, sodass die gesetzlichen Grenzwerte eingehalten werden. Dies dient dazu die Betriebskosten der Filterstraße möglichst gering zu halten, da die Abnutzung der teuren Filter somit reduziert wird.

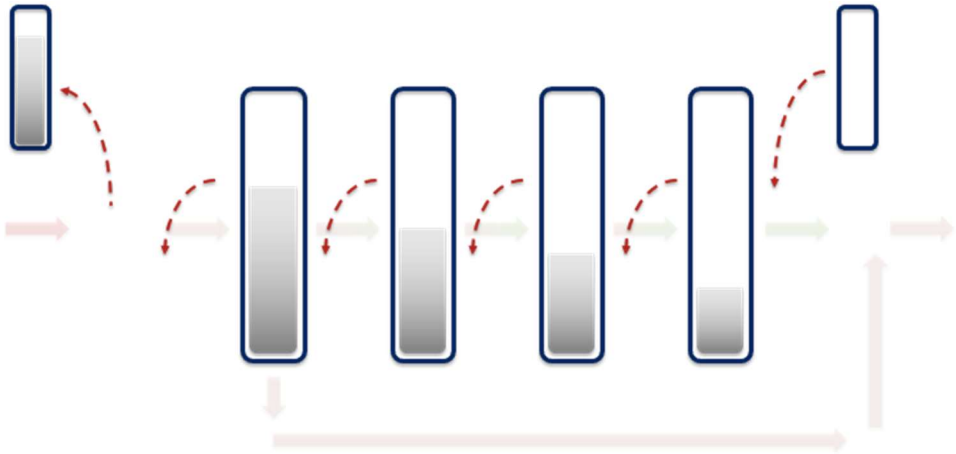


Filterstraße mit 5 Filtern und Bypass

Bisher wurde die Durchflussmenge am Bypass nach Erfahrungswerten geregelt. Um die Effizienz der Anlage zu erhöhen, indem die Abnutzung der Filter durch eine Optimierung des Mischungsverhältnisses reduziert wird, sollen Machine-Learning Ansätze implementiert werden.

Anhand von Vergangenheitsdaten der Deponie wurde während des Projektes ein Modell trainiert, welches es ermöglicht die Kontamination des Abwassers nach jedem einzelnen Filter vorherzusagen. Sind diese Kontaminationen bekannt, kann die Durchflussmenge am Bypass somit optimal geregelt werden, um nach der Vermischung hinter Filter 5 eine Kontamination des Abwassers zu erreichen, welche im Idealfall nur knapp unter dem gesetzlichen Grenzwert liegt.

Die Herausforderung lag darin, die abnehmende Filterleistung der Filter je nach Alter bzw. Abnutzungsgrad und Position in der Filterstraße mit einfließen zu lassen. Des Weiteren kam erschwerend hinzu, dass wenn ein Filter altersbedingt ersetzt wird, der darauffolgende Filter nachrückt und diese Position einnimmt. Somit durchläuft jeder Filter im Laufe seiner Nutzung jede Position innerhalb der Filterstraße. Dabei ist die Belastung eines Filters an jeder Position anders.



Positionswechsel der Filter

Es ist gelungen das im Bootcamp vermittelte Wissen auf diesen Prozess anzuwenden und ein Machine-Learning Modell zu trainieren, welches es ermöglicht die Kontamination des Abwassers an den verschiedenen Stellen im Prozess zu ermitteln und somit die Effizienz dieser Filterstraße zu erhöhen. Bevor ein solches Modell jedoch in Produktion geht, müssen die Vorhersagen noch ausreichend evaluiert werden, um Strafzahlungen wegen Nichteinhalten von Grenzwerten zu vermeiden.

Da Robert sich dazu entschlossen hat sich auch nach Abschluss des Bootcamps weiterhin für Nerou zu engagieren und sogar seine Masterarbeit dort zu schreiben, war das finale Projekt für ihn der Startschuss für eine deutlich genauere Ausarbeitung des Problems und die Möglichkeit mit dem erworbenen Wissen in der realen Welt, mit echten Daten einen Mehrwert zu schaffen.

Werdegang nach dem Camp

Uns war allen schon vorher bewusst, dass ein Bootcamp nur der Anfang für eine Karriere im Bereich Data Science sein kann. Daher haben wir bereits während des Camps angefangen uns für Praktika zu bewerben, um einen möglichst fließenden Übergang zu gewährleisten. Denn - und dies kann man nicht oft genug sagen - „real data is messy“. Wir wollten unsere Fähigkeiten möglichst praktisch weiter ausbauen und mit echten Problemen der Industrie Erfahrung sammeln. Unsere Wahlen fielen recht unterschiedlich aus. Da wir auch im Privaten sehr gut befreundet sind, tauschen wir uns regelmäßig über unsere Tätigkeiten aus und können so von den Erfahrungen der anderen profitieren

Im Folgenden sind unsere Werdegänge nach dem Camp in Kürze dargestellt.

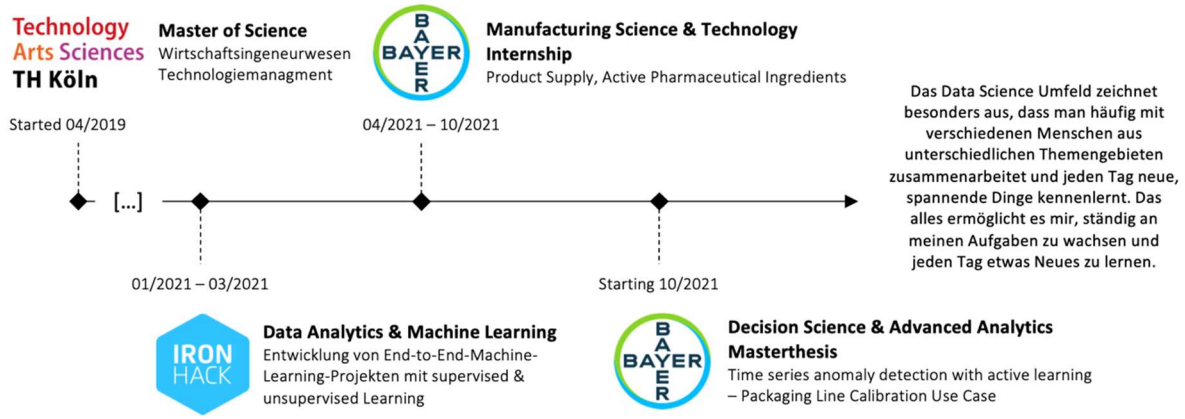
Wer wir sind

CAROLIN VOGT

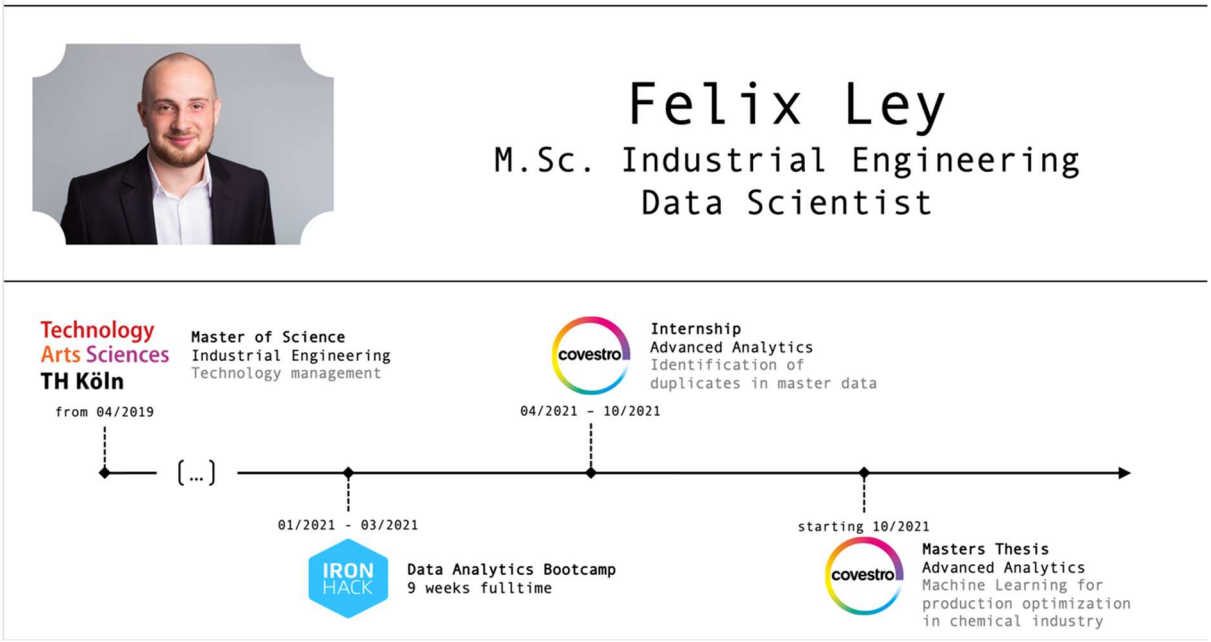
MASTER OF SCIENCE
INDUSTRIAL ENGENEERING
DATA ANALYTICS & SCIENCE

Interdisciplinary, proactive, passionate team player, willing to take over responsibility, motivated to go the extra mile and to continuously learn new skills.

MY JOURNEY TOWARDS DATA SCIENCE

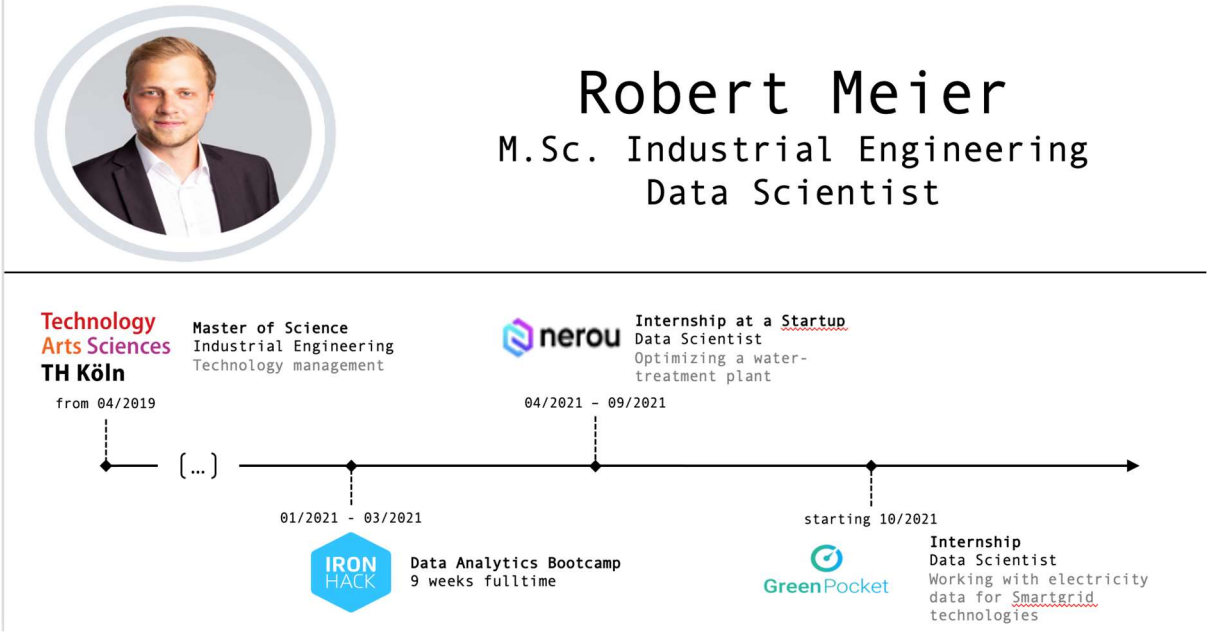


Das Ironhack Bootcamp hat mir die einmalige Möglichkeit geboten, meine gesamte Zeit in dieses für mich weitgehend neue Gebiet zu investieren. Viele Aspekte, die ich an Data Science interessant, spannend und herausfordernd finde, durfte ich hier das erste Mal Kennenlernen. In Kombination mit meinem Studium haben sich für mich neue Türen und Perspektiven eröffne



Während meiner Werkstudententätigkeiten in der chemischen Industrie habe ich erkannt, dass gerade in Bereichen mit einer hohen Automatisierungsgrad großen Datenmenge anfallen und gesammelt werden, welche darauf warten mit den passenden Methoden ausgewertet zu werden. Zwar sind BI Programme und Dashboards weit verbreitet, jedoch stoßen diese bei der tieferen Analyse von nicht direkt ersichtlichen Sachverhalten an ihre Grenzen. Die teils fehlende Expertise Daten programmatisch auszuwerten, tiefere Erkenntnisse zu gewinnen und Prozesse mittels Machine Learning zu optimieren haben meinen Wunsch geweckt, diese Methoden zu erlernen.

Die Teilnahme an dem Bootcamp hat es mir ermöglicht, genau diesen Schritt zu gehen. Aktuell darf ich als Data Scientist in der Advanced Analytics Abteilung der Covestro Deutschland AG meine neu erlernten Fähigkeiten anwenden und mich im Umfeld der chemischen Industrie weiterentwickeln.



Das Bootcamp eröffnete mir neue Möglichkeiten für meine berufliche und private Zukunft. Die Arbeit mit Daten mag für den einen oder anderen recht öde klingeln, ich persönlich empfinde die Arbeit jedoch als durchaus kreativ und sie lässt mich meine Problemlösungsvorstellungen endlich auch in die Praxis umsetzten. Durch meine Praktika möchte ich mir weitere Fähigkeiten aneignen, welche mich dazu befähigen aktiv an Innovationen mitzuarbeiten.

Ich fühle mich gerüstet für die Zukunft und freue mich darauf diese aktiv mitzugestalten

Danksagung

An dieser Stelle möchten wir uns ausdrücklich beim Verein zur Förderung des Campus Gummersbach der Technischen Hochschule Köln sowie allen dahinterstehenden Personen und Institutionen bedanken. Ohne Ihre Förderung hätte sich die Finanzierung dieses Bootcamps für jeden von uns deutlich erschwert. Dank der Teilnahme an diesem Bootcamp und somit auch dank Ihnen, gelang es jedem von uns einen Fuß in die Data Science Welt zu setzen und mit dem vermittelten Wissen in der realen Welt einen Mehrwert zu erzeugen. Nach wie vor ist jeder von uns froh diesen Weg gehen zu dürfen und würde sich erneut dafür entscheiden.
Mit Ihrer Förderung haben Sie es uns ermöglicht noch kurz vor Ende unseres Studiums den Fokus auf einen Interessenbereich zu legen, der uns sonst auf Grund unserer Studienfächer erstmal verwehrt geblieben wäre. Dank Ihnen aber haben es drei junge, motivierte Studierende geschafft, einen wegweisenden Grundstein für eine vielversprechende Karrieremöglichkeit zu legen und datengetriebene Lösungen für reale Probleme liefern zu können.

Wir danken Ihnen vielmals!