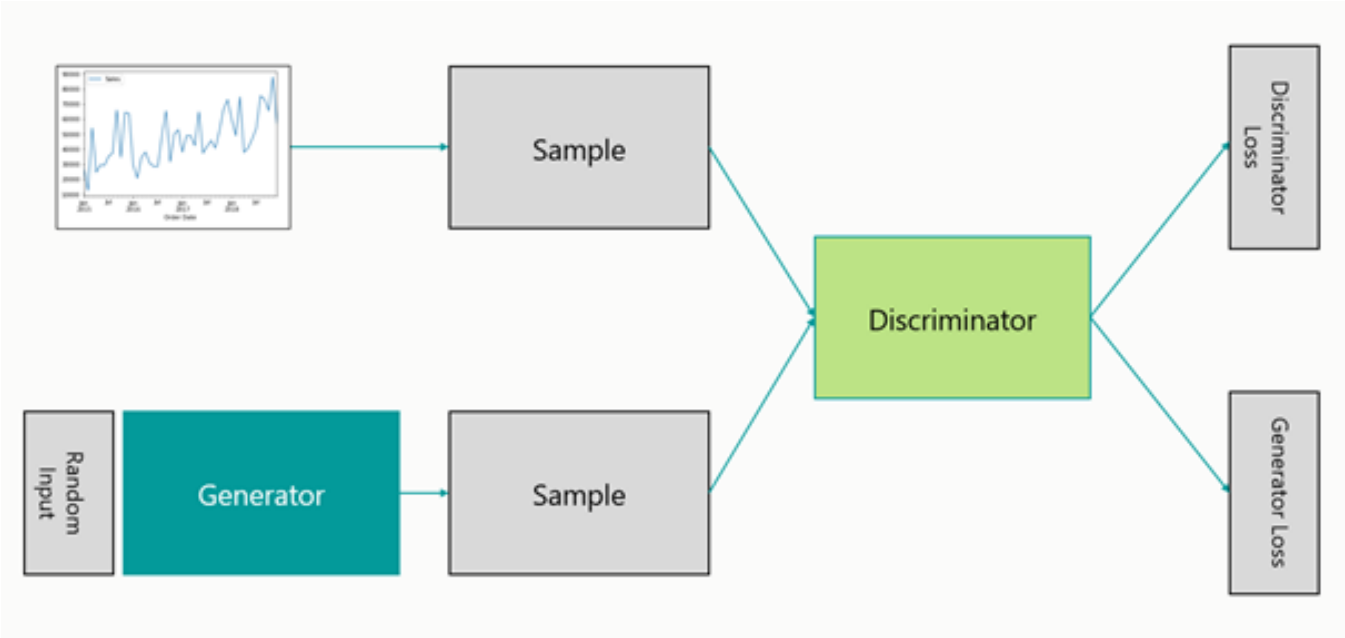
3. Hackathon der Franconian Data Science League

Am 18.02.2022 fand die 3.Runde der Franconian Data-Science-League (virtuell) statt. Dieses Mal waren wir, die DATEV, Gastgeber. Unter den Teilnehmern befand sich jeweils ein Team der GfK, des IAB und von Adidas. Die FAU konnte leider nicht genügend Teilnehmer finden.

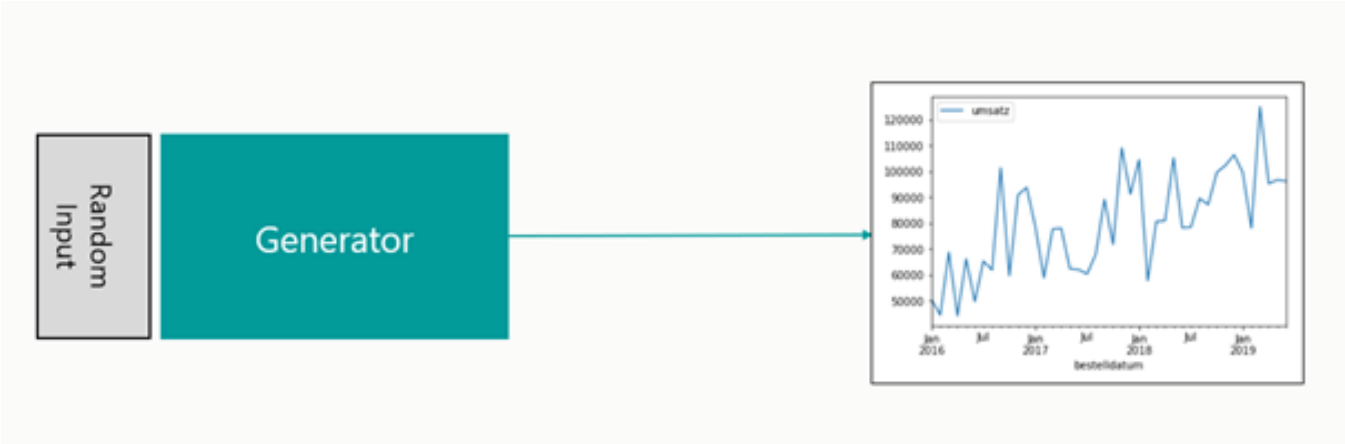
Datensatz

Da Datenschutz bei der DATEV eine zentrale Rolle spielt, war es nicht einfach einen geeigneten Datensatz zu finden, welcher sowohl eine gute Hackathon Aufgabe abgibt, als auch mit unserer Domäne verbunden ist. Deshalb kam folgende Idee auf: Ein künstlicher Datensatz, welcher die latente Struktur von echten Daten enthält. Latent bezieht sich dabei auf die vorhandene, aber nicht offensichtlichen Eigenschaften wie saisonale Schwankungen oder auch ein möglicherweise jährlich steigender Umsatz.

Mittels eines Generative Adversarial Networks (GAN) kann ein solcher Datensatz erzeugt werden. Dabei handelt es sich um die Architektur eines neuronalen Netzes, welches aus einem Generator und einem Discriminator besteht. Der Discriminator versucht dabei zu erkennen, ob Daten echt sind, oder aus dem Generator entstanden sind. Weil der Generator zu Beginn jedoch nur ‘zufälligen‘ Output erzeugt, ist der Discriminator gut bei der Unterscheidung. Im Laufe des Trainings jedoch, wird der Generator immer besser, da die erzeugten Outputs immer mehr den Trainingsdaten ähneln. Eine Unterscheidung wird immer schwerer, gar unmöglich.

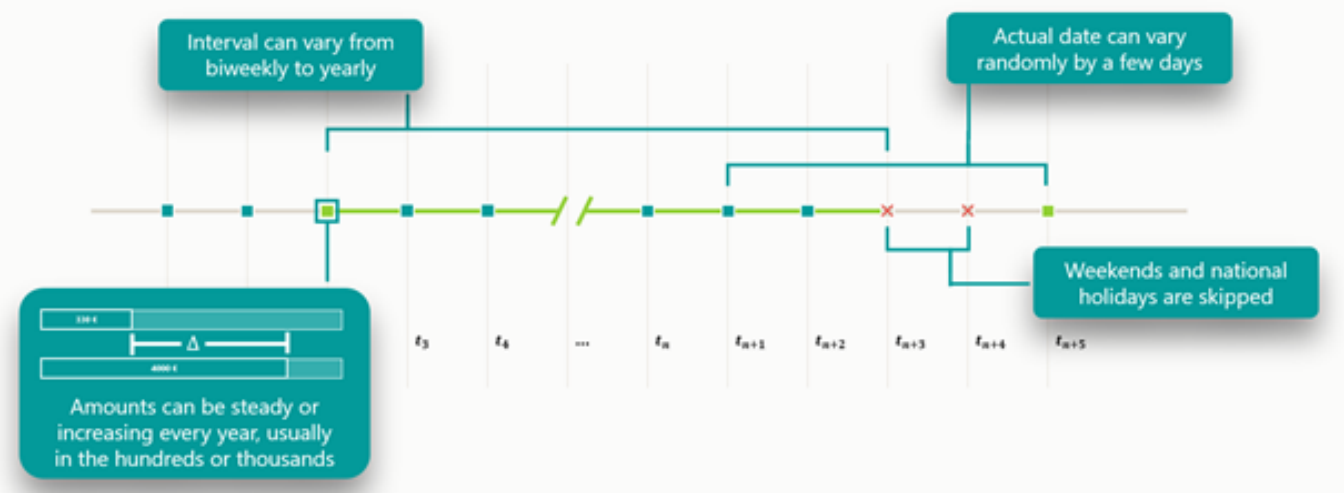


Dieses Verfahren haben wir genutzt, um einen Basisdatensatz zu erzeugen. Dabei haben wir Kaggle Daten zum Training genutzt, welche die Umsätze eines Onlineshops beinhalten.



	date	weekofyear	ship mode	sales	customer	product	city	zipcode
0	2016-01-02	2015-W53	Standard Shipping	8	KND-0160	PRK-0055	Mainz, Stadt	55116
1	2016-01-03	2015-W53	Premium Shipping	2669	KND-0250	PRK-0010	Kassel, documenta-Stadt	34117
2	2016-01-03	2015-W53	Standard Shipping	1	KND-0097	PRK-0058	Hamburg, Freie und Hansestadt	20038
3	2016-01-03	2015-W53	Standard Shipping	3	KND-0177	PRK-0118	Rostock, Hansestadt	18055
4	2016-01-03	2015-W53	Standard Shipping	87	KND-0190	PRK-0411	Magdeburg, Landeshauptstadt	39104

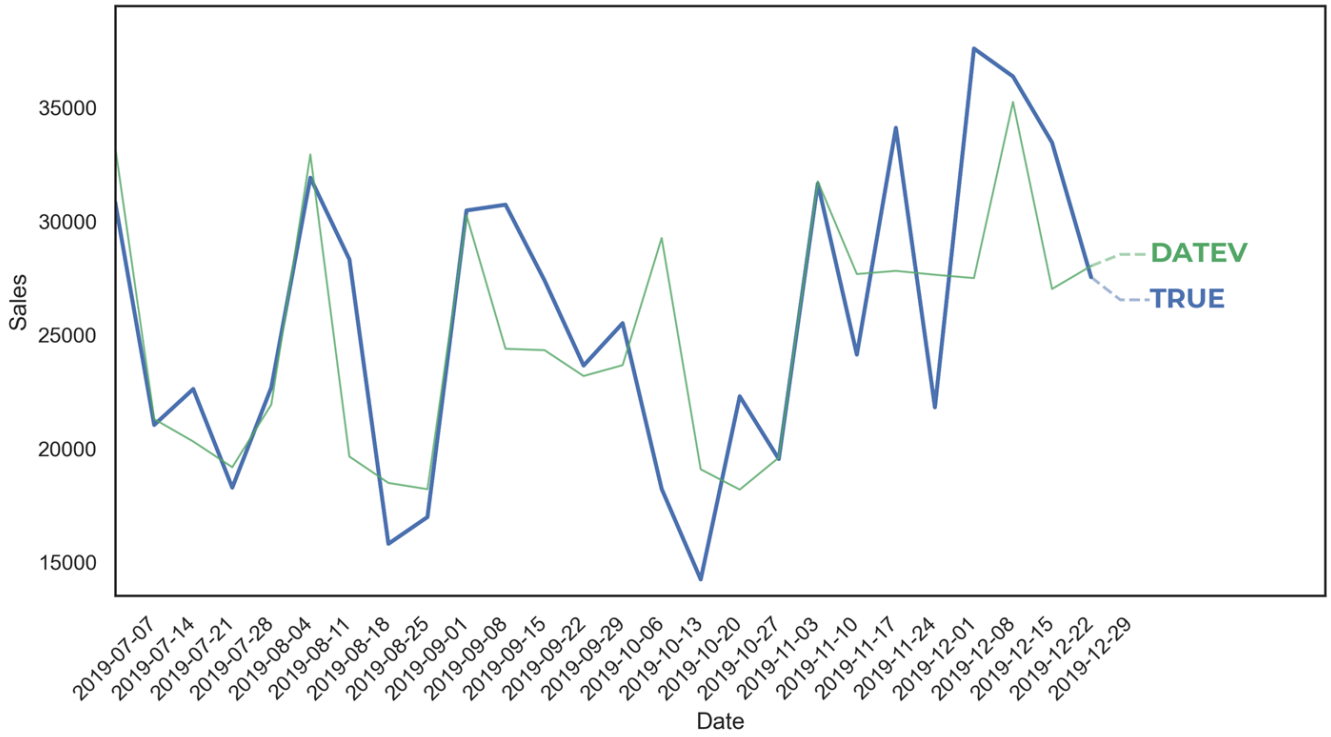
Um die Aufgabe jedoch etwas realistischer und anspruchsvoller zu gestalten, wurden diese Daten noch angereichert. Neben Skalierungen an Feiertagen (Weihnachten), wurden wiederkehrende Buchungen hinzugefügt. Diese fanden einmalig (ad hoc) oder in festen Intervallen (z.B. monatlich oder quartalsweise) mit einer gewissen Schwankung (z.B. jährliche prozentuale Umsatzerhöhung von 10%) statt. Um die Daten möglichst realistisch zu halten, konnten diese Buchungen nur an Werktagen erfolgen.



Aufgabe

Die Teilnehmer erhielten Daten in einem Zeitraum von 3,5 Jahren (Januar 2016 - Juni 2019). Ziel war es, den Umsatz für das folgenden halbe Jahr (Juli 2019 – Dezember 2019) wöchentlich vorherzusagen. Als Metrik wurde MAPE (mean absolute percentage error) genutzt.

Um die Aufgabe zu validieren, wurde diese Max Schulze Dieckoff, einem Data Scientist der DATEV, gelöst. Mit der Kombination aus einem Interpurchase Time Modell für die wiederkeherenden Buchungen und einem SARIMA Modell für die restliche Zeitreihenanlyse, wurde eine MAPE von 13,56% (niedriger = besser) erreicht.

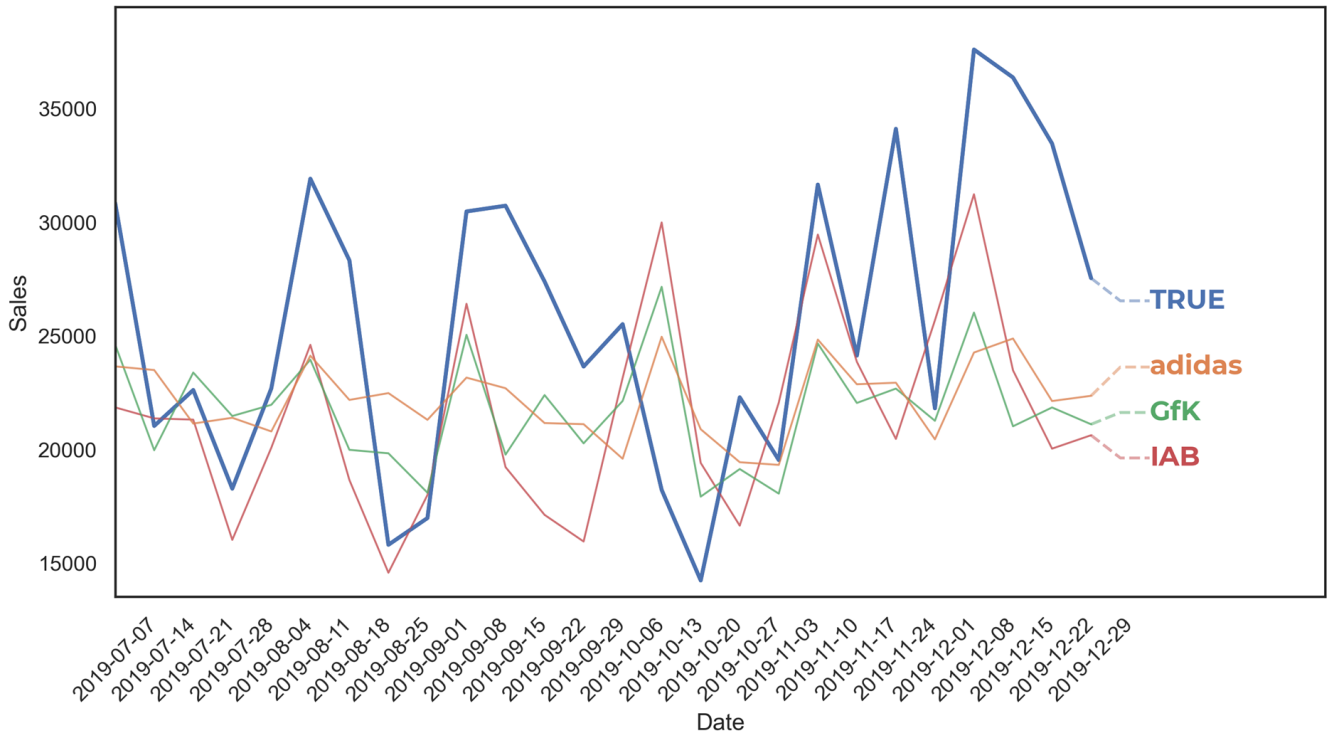


Ergebnis und Learnings

Mit einem MAPE von 22,46% belegte das IAB den dritten Platz. Dabei wurde die Aufgabe mittels eines Neuronales Netzes gelösten.

Adidas hingegen nutzte das bewährte Zeitreihen-Modell SARIMA, reicherten dieses jedoch noch mit zusätzlichen Infos eines Gradient Boosting Modells an. Dies bescherte ihnen den zweiten Platz mit einem MAPE von 21,95%.

Der Gewinner des Hackathons war die GfK mit einem MAPE von 20,37%. Nach Ausprobieren diverser Methoden, wurde schließlich die Prediction mittels einer statistischen Methode durchgeführt. Dabei wurde der Umsatz der gleichen Kalenderwoche im Vorjahr mit dem Median der letzten Wochen angereichert.



Wie man sieht, sind oftmals die simplen bewährten Methoden besser, als einfach alles in ein neuronales Netz zu schmeißen :D.

Ausblick

Zum nächsten Hackathon gibt es noch keine näheren Infos. Jedoch wird die DATEV dann als Teilnehmer agieren und den ersten Sieg einfahren.