Inhalt

[1 Einführung 1](#_Toc520800629)

[1.1 Was ist Kaggle 1](#_Toc520800630)

[1.2 Was ist die Winton Stock Market Challenge 1](#_Toc520800631)

[1.3 Was ist die Forschungslücke 1](#_Toc520800632)

[2 Daten 1](#_Toc520800633)

[2.1 Daten erklären 1](#_Toc520800634)

[2.2 EDA/Descriptive Statistik 1](#_Toc520800635)

[2.3 Einfluss Variablen und und Core Features 5 und 7 1](#_Toc520800636)

[2.4 Zielvariablen 1](#_Toc520800637)

[3 Methodik 1](#_Toc520800638)

[3.1 Aufbereiten der Daten 1](#_Toc520800639)

[3.2 Benchmarks/ZeroBenchmark 1](#_Toc520800640)

[3.3 Base Models 1](#_Toc520800641)

[3.4 Advanced Feature Engineering 1](#_Toc520800642)

[3.5 Time sensitive CV 1](#_Toc520800643)

[3.6 Model Optimization 1](#_Toc520800644)

[3.6.1 Hyper Parameter Tuning 1](#_Toc520800645)

[3.6.2 Feature Selection 1](#_Toc520800646)

[4 Results 1](#_Toc520800647)

[4.1 Welche Ansätze wie und warum haben welche Ansätze funktioniert 1](#_Toc520800648)

[5 Conclusion 1](#_Toc520800649)

[5.1 Zusammenfassung der Aussagen und wo der Wertbeitrag liegt 1](#_Toc520800650)

# Einführung

## Was ist Kaggle

Kaggle ist eine Plattform auf der Unternehmen Machine Learing Wettbewerbe veranstalten können, indem sie Datensätze zur Verfügung stellt und ein Problem definiert, dass dann von Teams und Mitgliedern der Community durch prädiktive Modelle und Methoden der Datenanalyse gelöst werden soll. Die besten Algorithmen werden anschließend dann mit Geldprämien ausgezeichnet. Gegründet wurde die Plattform 2010 von Anthony Goldbloom und konnte seitdem viele namhafte Unternehmen wie 2Sigma, Facebook und NASA als Kunden für sich gewinnen. Im Jahr 2017 wurde Kaggle schließlich von Google aquiriert und in Google Cloud eingegliedert. Allerdings blieb das Team weitgehend unverändert und Kaggle operiert nach wie vor als eigentständige Marke. (Goldbloom 2017)

Aufbau von Kaggle Challenges

Grundsätzlich laufen die Wettbewerbe so ab, dass der Veranstalter beraten durch ein Team von Kaggle-Mitarbeitern simulierte oder reale und in irgendeiner Art und Weise anonymisierte Trainings- und Testdaten hochläd und das Problem in einem kurzen Text beschreibt.

Der Trainingsdatensatz enthält neben den Einflussvariablen auch die Zielvariable und dient dazu Algorithmen zu trainieren, wohingegen bei dem Testdatensatz die Zielvariable geschätzt werden muss. Die Genauigkeit der Schätzung wird durch eine durch den Veranstalter festgelegte Metrik ermittelt, wie beispielsweise die mittlere quadrierte Abweichung vom wahren Wert.

Zusätzlich gibt es für jede Challenge ein Diskussionsforum, bei dem Mitglieder untereinander ihre Ergebnisse oder Fragen zur Aufgabenstellung diskutieren können oder auch direkt mit dem Veranstlater in Konakt treten können und eine Echtzeit-Rangliste, dass die Zwischenergebnisse der Teilnehmer durch einen Teil der Testdaten evaluiert und ihnen so Feedback über die Qualität ihrer Schätzungen gibt. Die finalen und für die Endwertung entscheidendten Ergebnisse werden dann nach Ablauf des Wettbewerbs am restlichen, noch unberührten Teil der Testdaten bemessen. (Chatterjee et al. 2018, S. 220–222)

Ein zentraler Kerngedanke von Kaggle Competitions ist Crowdsourcing und das Teilen von Informationen, in diesem Zusammenhang sind insbesondere auch die sog. Kernels zu nenne, die Teilnehmer innerhalb von den Wettbewerben für alle einsehbar veröffentlichen können, damit andere Wettbewerber auf den eigenen Ergebnissen aufbauen können und sich zu neuen Ansätzen zu Problemlösung inspirieren lassen und vice versa. (Marr 2016, S. 281–286)

Bei den Kernels handelt es sich um R oder Python Scripte, die über den Browser in der Cloud erstellt und ausgeführt werden können ohne dass der Nutzer eine sich eine eigene Entwicklungsumgebung einrichten oder installieren muss und da der Code so auf den Servern von Kaggle ausgeführt wird, wird die Rechenleistung der lokalen Rechner nicht beansprucht. Die Kernels können außerdem mit Competitions oder Datensätzen verknüpft werden und so direkt auf diese zurückgreifen, dadurch entfällt das hochladen von u.U. großen Datensätzen und fördert das einfache Teilen von Informationen und Ergebnissen zwischen den Mitgliedern. (Guo 2017)

Zukünftig ist außerdem geplant, dass nicht nur direkt von den Kernels auf die Daten einer Challenge zugegriffen werden kann, sondern auch Vorhersagen bzw. Lösungen direkt darüber abzugeben und nicht wie bisher üblich ein extra Dokument dafür zu exportieren, es herunterzuladen und anschließend als Lösungsdatei in einer Challenge wieder hochzuladen, was nicht nur nutzerfreundlicher ist, sondern auch neue Möglichkeiten eröffnet wie Reinforcement Learning Wettbewerbe oder Wettbewerbe bei denen neben der eigentlichen Lösung auch deren Sparsamkeit und Effizienz bewertet wird. (Goldbloom 2018)

Ein bisschen ausgegliedert aus der eigentlichen Plattform betreibt Kaggle auch noch den hauseignenen Blog „No Free Hunch“ auf dem neben allgemeinen Informationen über die Entwicklung der Plattformen oder aktuelle Themen in Bereich Data Science insbesondere auch Interviews mit den Gewinnern verschiedener Challenges veröffentlicht, in dem diese erklären wie sie vorgegangen sind und was ihre wesentlichen Erkenntnisse waren um diese auch den weniger erfolgreichen Teilnehmern zugänglich zu machen und deren Lernprozess zu unterstützen. [Quelle fehlt]

## Was ist die Winton Stock Market Challenge

The Winton Stock Market Challenge war die zweite Recruitment Competition die von der Investment Management Firma Winton auf Kaggle veranstaltet wurde und neben den Preisgeldern gab es außerdem Bewerbungsgespräche mit dem Analytics Team von Winton zu gewinnen. Ziel des Wettbewerbs war es zukünftige Wertpapierrenditen vorherzusagen auf Basis deren Vergangenheit, sowie 25 maskierten Merkmalen oder Kennzahlen. Die Maskierung ist insbesondere gewählt worden um Teilnehmern mit Branchenwissen oder Kenntnissen im Bereich Trading keinen Vorteil gegenüber anderen zu verschaffen und somit für Chancengleichheit zu Sorgen.

Gestartet wurde der Wettbewerb im Oktober 2015 und endete im Januar 2016, anders als die Meisten Kaggle-Competitions lag die Laufzeit nicht bei zwei sondern drei Monaten und auch das Teilen von Code oder veröffentlichen von Kernels wurde nicht gestattet, wodurch auch nach Abschluss nur begrenzte Informationen verfügbar waren.

Designt wurde die Challenge von zwei Mitarbeitern der Forschungsabteilung von Winton, die nach eigenen Angaben versucht haben einen guten Überblick ihres Tagesgeschäfts abzubilden und den Problemen mit denen sie dort konfrontiert sind. Explizit genannt werden das Umgehen mit großen verrauschten Datenmengen und daraus resultierend einem starken Hang zum overfitten.

Bereitgestellt wurde ein Trainingsdatensatz der abgesehen von den 25 Features außerdem zwei Vergangene Tagesrenditen

Kurzüberblick über die Daten

Messgröße

Wie üblich gab es ein für alle sichtbares live Leaderboard, dass den Teilnehmern nach dem Hochladen ihrer Vorhersagen den WMAE von 25% der Testdaten als Punktestand errechnete. Für die entscheidende Endwertung des Wettbewerbs wurde ein nach Abschluss ein privates Leaderboard veröffentlicht, das den WMAE der restlichen 75% der Testdaten abbildete und für die Platzierung ausschlaggebend war.

Ein Punkt der außerdem noch erwähnt werden sollte ist, dass während des Wettbewerbs der Testdatensatz geändert werden musste. Einigen Teilnehmern gelang es durch das wiederholte Abgeben von Vorhersagen für verschiedenen Datenpunkte, die Art wie die Testdaten in öffentliches Leaderboard und privates Leaderboard aufgeteilt wurde rück zu rechnen und somit auch in der Lage die wahren Werte des privaten Leaderboards zu entschlüsseln. Winton entschied sich nach bekanntwerden des Fehlers dazu die Testdaten anzupassen, wohlwissen das dies ein markanter Eingriff in den Wettbewerb war, mit der Begründung, dass Entschlüsseln der Werte keine statistischen Modelle darstellt und somit nicht zugelassen werden können und Teilnehmer mit robusten statistischen Modellen auch in der Lage sein werden mit diesen den neuen Datensatz ähnlich gut vorherzusagen.

## Was ist die Forschungslücke

# Daten

## Daten erklären

## EDA/Descriptive Statistik

## Einfluss Variablen und und Core Features 5 und 7

## Zielvariablen

# Methodik

## Aufbereiten der Daten

## Benchmarks/ZeroBenchmark

## Base Models

## Advanced Feature Engineering

## Time sensitive CV

## Model Optimization

### Hyper Parameter Tuning

### Feature Selection

# Results

## Welche Ansätze wie und warum haben welche Ansätze funktioniert

# Conclusion

## Zusammenfassung der Aussagen und wo der Wertbeitrag liegt

Literaturverzeichnis

Chatterjee, Samir; Dutta, Kaushik; Sundarraj, Rangaraja P. (Hg.) (2018): Designing for a digital and globalized world. 13th international conference, DESRIST 2018, Chennai, India, June 3-6, 2018 : proceedings. DESRIST; International Conference on Design Science Research in Information Systems and Technology. Cham: Springer (Lecture notes in computer science, 10844).

Goldbloom, Anthony (2017): Kaggle Joins Google Cloud. Kaggle.

Goldbloom, Anthony (2018): Reviewing 2017 and Previewing 2018. Kaggle. Online verfügbar unter http://blog.kaggle.com/2018/01/22/reviewing-2017-and-previewing-2018/, zuletzt geprüft am 31.07.2018.

Guo, Yufeng (2017): Introduction to Kaggle Kernels. towardsdatascience.com. Online verfügbar unter https://towardsdatascience.com/introduction-to-kaggle-kernels-2ad754ebf77.

Marr, Bernard (2016): Big data in practice. How 45 successful companies used big data analytics to deliver extraordinary results. Chichester, West Sussex: Wiley. Online verfügbar unter http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&AN=1205831.