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

Einleitender Allgemeiner Satz der Zum thema hinleitet sowas wie aus den Zeotungsartikeln blablabla Viele unternehmen haben die bedeutngng von Big Data für ihr Geschäft erkannt, häufig fehlt es aber an personal die großen Datenmengen zu verarbeiten

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. Ziel und Zweck von den Wettbewerben erläutern evtl auf Netflix bezug nehmen 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. Das Team allerdings blieb 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 bereitstellt und das zu lösende 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 vom Veranstalter festgelegte Metrik ermittelt, wie beispielsweise die mittlere quadrierte Abweichung vom wahren Wert.

Zusätzlich gibt es für jede Challenge ein Forum, bei dem Mitglieder untereinander ihre Erkenntnisse und Fragen zur Aufgabenstellung diskutieren können und außerdem ein öffentliches Leaderboard, dass die Zwischenergebnisse der Teilnehmer anhand eines Teils 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]

Big Data in Practice mit einbauen

## 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 von bis zu 20.000$ gab es außerdem Bewerbungsgespräche mit dem Analytics Team von Winton zu gewinnen. Ziel des Wettbewerbs war es zukünftige Wertpapierrenditen vorherzusagen auf Basis deren vergangenen Performance, 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 gewähren und somit für Chancengleichheit zu Sorgen.

Gestartet wurde der Wettbewerb im Oktober 2015 und lief bis Januar 2016, anders als bei den meisten Kaggle-Competitions lag die Laufzeit somit nicht bei zwei sondern bei drei Monaten und auch das Teilen von Code oder veröffentlichen von Kernels im Zusammenhang mit dem Wettbewerb 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 die Problemen mit denen sie dort konfrontiert sind zu thematisieren. Explizit genannt werden das Umgehen mit großen verrauschten Datenmengen und daraus resultierend einem starken Hang zum overfitten.

Für die Teilnehmer bereitgestellt wurde ein Trainingsdatensatz mit 40.000 Einträgen, der abgesehen von den 25 Features, die beiden vergangene Tagesrenditen, die Minutenrenditen von eins bis 180 und darauf folgen dann die beiden zukünftigen Tagesrenditen enthielten. Außerdem je eine Gewichtung für die Tages- und Minutenrenditen.



Im Testdatensatz mit 120.000 Einträgen sind nur die Features, die vergangenen Tagesrenditen und die Minutenrenditen von eins bis 120 enthalten, die folgenden 60 Minuten, sowie die folgenden zwei Tage müssen von den Teilnehmern vorhergesagt werden. Als Evaluationsmetrik für die Vorhersagen wurde die gewichtete durchschnittliche absolute Abweichung vom dem wahren Wert der Renditen gewählt (Weighted Mean Absolute Error, kurz: WMAE) den es zu minimieren galt. Zwar gestand der Veranstalter ein, dass es sich dabei um eine nicht ganz perfekte Messgröße für den Erfolg handelte, allerdings wurde durch die Gewichtung versucht in einer sehr vereinfachten Form Trading-Kosten zu simulieren.

Ein Satz zur Benchmark

Wie für Kaggle Competitions üblich gab es ein für alle sichtbares öffentliches 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 letztendliche Platzierung ausschlaggebend war. (The Winton Stock Market Challenge 2015; Anderson 2016)

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 jeweils einige verschiedenen Datenpunkte, die Art wie die Testdaten in öffentliches Leaderboard und privates Leaderboard aufgeteilt wurde rück zu rechnen und somit waren sie auch in der Lage die wahren Werte des privaten Leaderboards zu entschlüsseln. Winton entschied sich nach dem Bekanntwerden dazu die Testdaten anzupassen, wohlwissen das dies ein markanter Eingriff in den Wettbewerb war, begründeten sie ihre Entscheidung damit, dass das Entschlüsseln der Werte keine statistischen Modelle darstellt und somit nicht zugelassen werden könnte, des weiteren seien Teilnehmer mit robusten statistischen Modellen auch weiterhin in der Lage den neuen Datensatz ähnlich gut vorherzusagen und diese somit auch nicht maßgeblich benachteiligt würden. (Anderson 2015)

## Was ist die Forschungslücke

Probleme der Challenge Cross validation und übertragung der daten aufs lb

Anzahl der Modelle die nichteinmal die Benchmark übertreffen

Drops von Private LB auf Public LB

Anteil der Leute die die Benchmark nicht geschlagen haben

# Ein Blick auf die Daten

## Daten erklären

Aus dem abschließendem Statement von Winton über die Challenge geht hervor, dass es sich bei den Daten um echte, wenn auch aus Gründen der Vertraulichkeit durch Transformation und leichtes Verrauschen unkenntlich gemachte Finanzmarktdaten handelt. Die Features sind Kennzahlen der Fundamental- und Chartanalyse, vergangene Performance oder teilweise auch ohne jegliche Bedeutung, trotzdem wird angenommen das mit Ausnahme von Feature\_5 und Feature\_7 die meisten keine besonders ausgeprägte Vorhersagekraft haben und somit auch nur die beiden im Weiteren tiefgehender betrachtet werden. Die Tages- und Minutenrenditen waren außerdem bereits logarithmiert.

Allgemein zeigte sich allerdings, dass abgesehen von fehlenden Werten, die Datensätze von Haus aus relativ sauber waren, was darauf zurückzuführen ist, dass das Entwickeln einer funktionierenden Cross-Validation-Methode und von Modellen mit tatsächlicher Vorhersagekraft von Winton als anspruchsvoll genug betrachtet wurde und nicht zusätzlich durch Datenaufbereitung erschwert werden sollte Quelle. Trotzdem erfordern viele Machine Learning Algorithmen vollständige und komplette Inputvariablen, weshalb im ersten Schritt der Datensatz auf fehlende Werte untersucht und entsprechend bereinigt werden muss.

Ein Bild, das Text enthält.

Mit hoher Zuverlässigkeit generierte Beschreibung

Abbildung 1 Anteil der fehlenden Werte für Trainings- und Testdaten

Wie auf Abbildung 1 zu sehen ist, der Anteil der fehlenden Werte bei den Features über Trainings- und Testdaten ähnlich verteilt, während einzelne Variablen mit äußert hohen Fehlraten hervorstechen liegt sie bei dem Großteil der Features unterhalb von 10%. Außerdem fällt auf, dass Feature\_5 und Feature\_7 jeweils überhauptkeine fehlenden Werte enthalten.

Ebenfalls ohne fehlende Werte sind die vergangenen Tagesrenditen. Für die vergangenen Minutenrenditen verteilt sich der Anteil der fehlenden Werte wie folgt:

|  |  |  |
| --- | --- | --- |
|  | **Train** | **Test** |
| **Count** | 40000 | 120000 |
| **Mean** | 0.046859 | 0.048532 |
| **Min** | 0.000000 | 0.000000 |
| **25%** | 0.000000 | 0.000000 |
| **50%** | 0.000000 | 0.000000 |
| **75%** | 0.050000 | 0.050000 |
| **Max** | 0.500000 | 0.500000 |

Abbildung 2 Verteilung der fehlenden Werte bei den vergangenen Minutenrenditen nach Datensätzen

Man sieht das der überwiegende Teil der Datensätze in Bezug auf die Minutenrenditen relativ sauber ist. Die meisten Einträge sind komplett und enthalten für jede Minute eine Rendite und selbst bei denen, die unvollständig sind, liegt die Fehlerrate häufig nur bei wenigen Prozent. Trotzdem gibt es allerdings auch einige Einträge, bei denen bis zu der Hälfte der Renditen nicht enthalten sind.

Nach Kaiser (2014) lässt sich grundlegend nach 3 verschiedenen Ansätzen zum Umgang mit fehlenden Daten unterscheiden. Als erstes dem Reduzieren des Datensatzen um die Einträge, die unvollständig sind, zweitens das behandeln der fehlenden Werte als „special values“, also dem zuweisen eines neuen Wertes, der ersichtlich macht, das es sich dabei um fehlende Werte handeln, oder drittens mehrere Techniken zur Imputation, wobei hier versucht wird die nichtvorhandenen Informationen sinnvoll zu schätzen. Jede der drei Möglichkeiten hat verschiedene Vor- und Nachteile und ihre Anwendung ist abhängig von den gegebenen Daten oder der Charakteristik bzw. den Gründen für die Lücken im Datensatz. Nachdem allerdings nichts genaueres bekannt ist, worauf die fehlenden Werte zurückzuführen sind, wurde versucht mit einer Kombination aus Reduktion und Imputation zu arbeiten um den Datensatz zu säubern. Features die fehlende Werte enthalten wurden aus dem Datensatz entfernt, einmal, weil bei Features mit einer Fehlerrate von 50% bzw. 80% eine Schätzung der fehlenden Werte schwierig und ungenau wäre und außerdem, weil mit Ausnahme von den beiden vollständigen Features fünf und sieben, den anderen seitens Winton auch keine besondere Vorhersagekraft zugesprochen wurde. Lediglich bei den Base Models in Abschnitt 3.2 wird noch einmal das komplette Feature-Set herangezogen werden und sofern nötig fehlende Werte der Einfachheit mit den Mittelwerten gefüllt und vermeintlich kategoriale Features mittels One-Hot-Kodierung in Dummy-Variablen überführt.

Für die fehlenden Werte in den Minutenrenditen wurde Imputation durch dem Mittelwert der jeweiligen Zeitreihe angewandt, da es sich hierbei um ein sehr einfaches und gängiges Verfahren handelt, und bei den meisten Zeitreihen der Anteil der Fehler ohnehin sehr gering war. Somit ist von einem maßgeblicher Informationsverlust dadurch eher nicht auszugehen.

Abschließender Satz zum neuen gereinigtem Datensatz

Aufteilen in Minute und Daily Problem

## EDA/Descriptive Statistik

### Feature 7

Schon während der Competition ist vielen Teilnehmern Feature 7 im besonderen Maße ins Auge gestochen, wenn auch im Diskussionsforum häufig Unsicherheit herrschte, was die genaue Bedeutung anging, oder wie man damit am besten verfahren sollte. Auffällig war besonders die große Bedeutung, die diesem durch Desicion-Tree-Algorithmen zugesprochen wurde, wie hier auf Abbildung 3 beispielhaft zu sehen ist.

Ein Bild, das Screenshot enthält.

Mit sehr hoher Zuverlässigkeit generierte Beschreibung

Abbildung 3 Feature importance nach XGBRegressor auf die rohen Trainingsdaten

Trotz der vermeintlich großen Bedeutung ließen sich damit auf der anderen Seite allerdings kaum Modelle erstellen die auch auf dem Leaderboard gute Ergebnisse erzielten. Ein Grund dafür war, dass sich die Werte für Feature\_7 zwischen den Trainings- und Testdaten grundlegend unterschieden waren bzw. sich nicht überschnitten. Während innerhalb eines Datensatzes einzelne Ausprägungen im Schnitt um die 50 Mal auftreten, treten ebendiese Ausprägungen im anderen Datensatz wiederum überhaupt nicht auf. Außerdem sind die Werte für Feature\_7 nicht fortlaufend und springen in unregelmäßigen Abständen z.B. von 26 auf 64 auf 138 und so weiter.

Ersichtlich wurde das Ganze, nachdem Winton bekannt gab, dass es sich hierbei um eine Art Kennnummer für verschiedene Tage handelt, und weiter, dass die Wertpapiere Tendenzen aufweisen innerhalb eines Tages zu korrelieren. Dem entsprechen lernen die Desicion-Tree-Algorithmen quasi einfach auswendig was an den jeweiligen Tagen passiert, nachdem in den Testdaten dann allerdings komplett andere Tage enthalten sind können sie das Gelernte nicht mehr anwenden und liefern schlechte Prognosen. (Anderson 2016)

Diese Erkenntnisse werden später insbesondere beim Entwickeln einer Cross-Validation-Methode wieder von Bedeutung, da hier darauf geachtet werde

### Feature 5

Das zweite Feature das Winton als wichtig beschrieben hatte war Feature\_5. Hierbei handelt es sich um ein ganzzahliges Feature mit Ausprägungen von eins bis zehn. Die Werte sind zwischen den beiden Datensätzen in etwa gleich verteilt und reichen von ca. 2,0% bis ca. 17,5%.

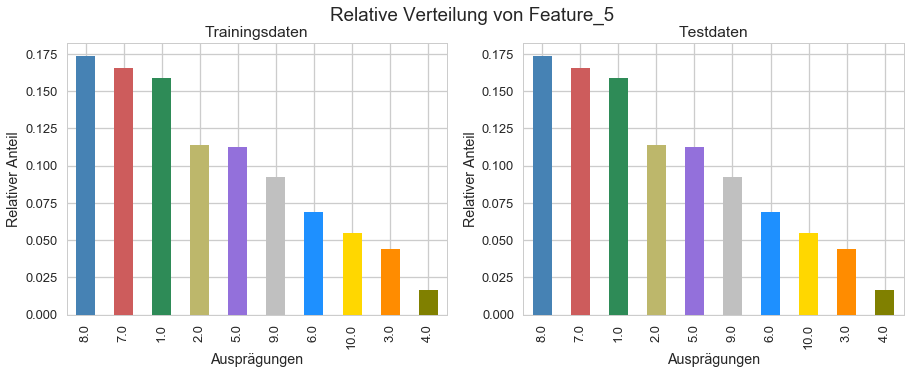


Abbildung 4

Eine offizielle Auflösung um was es sich genau handelt gab es auch nach der Challenge zwar nicht, allerdings wurde im Forum vermutet, dass es sich um die Monate Januar bis Oktober handeln könnte. Eine andere Möglichkeit wäre, dass es eine Art Branchenindex darstellt.

Email nochmal lesen und hier mit rein bringen

### Zeitreihen und Renditen

Betrachtet man die empirische Verteilung der Tagesrenditen im Vergleich zur entsprechenden Normalverteilung zeigen sich die als stilisierte Fakten von Renditen bekannte Spitzigkeit und bei genauerer Betrachtung der Extremwerte auch schwere Ränder. Folglich treten verhältnismäßig häufig sehr kleine Renditen auf oder stark positive bzw. negative. Die Verteilung ist außerdem leicht rechtschieg, was ebenfalls als charakteristisch gilt. Evtl jarque-Bera Test mit einbringen

Ein Bild, das Text, Karte enthält.

Mit hoher Zuverlässigkeit generierte Beschreibung

Abbildung 5

Volatilitätscluster sind nicht so deutlich erkennbar, was zum Teil dadurch zu begründen ist, dass die Werte für Feature\_7, also die Zeitpunkte, nicht durchgehen sind und auch nicht bekannt ist, wie viel Zeit jeweils zwischen zwei Zeitpunkten vergangen ist. Trotzdem äußern sich Cluster dahingehen, das sich um bestimmte Stellen herum die Ausreißer und Extremwerte eher zu konzentrieren scheinen, als bei den gleichen Tagesrenditen, die zufällig nochmal neu auf die Zeitpunkte verteilt, also gemischt, wurden. Beispiele hierfür liegen um 3000, 5000 oder 9000. (Schmid und Trede 2006, S. 12–15)

Ein Bild, das Screenshot enthält.

Mit sehr hoher Zuverlässigkeit generierte Beschreibung

Abbildung 6

Für die Minutenrenditen wurden zwei zufällig gezogene Datensätze aus dem Trainingsdatensatz zur Veranschaulichung hergenommen, insbesondere bei dem

Für folgende Korrelationsmatrix wurden aus den ersten 120 Minutenrendite des Trainingsdatensatzes die kummulierte Periodenrendite Ret\_MinutePast gebildet und für die darauffolgenden 60 Minuten Ret\_MinuteFut. Anschließend war auf der Matrix zu sehen, dass sich zwischen den einzelnen Features mitunter starke Korrelationen zeigen, zwischen den Features und den Renditen allerdings kaum. Gruppiert man aber den Datensatz nach Feature 7 und schaut sich nur die Korrelationen innerhalb eines Zeitpunktes an, ist zu erkennen, dass sich hier durchaus deutlichere Zusammenhänge zeigen. Höhere Werte wie in Abbildung 8 zwischen Feature\_25 und Ret\_PlusOne treten zwar hin und wieder auf, allerdings übertragen sich diese nicht zwischen den verschiedenen Zeitpunkt und jeder Gruppe von Feature\_7 korreliert mit Variablen besser oder schlechter.

Ein Bild, das schwarz, Gebäude, weiß, Fenster enthält.

Mit hoher Zuverlässigkeit generierte Beschreibung

Abbildung 7 Korrelationsmatrix für die 25 Features und die Renditen, Ret\_MinutePast ergibt sich aus den kummulierten ersten 120 Minutenrenditen, Ret\_MinuteFut ergibt sich aus den folgenden 60 Minutenrenditen



Abbildung 8 Korrelationen für die 114 Datensätze mit dem Wert 37168 für Feature 7

# Methodik

## Benchmarks/ZeroBenchmark

Bevor im Anschluss die ersten Modelle zur Vorhersage von den Renditen betrachtet und evaluiert werden, ist es wichtig zu verstehen wie diese Evaluation funktioniert. Als Maßeinheit ist wie bereits erwähnt von Winton die gewichtete mittlere absolute Abweichung vom wahren Wert gewählt worden und die Gewichte dienen als einfache Form von Trading Kosten. Im Trainingsdatensatz sind in den letzten beiden Spalten Gewichte für einmal die Intraday-Minutenrenditen und einmal für die Tagesrenditen, die Gewichte für den Testdatensatz sind unbekannt.

Ins Verhältnis gesetzt wird diese Score zu der sog. Zero-Benchmark, also dem Wert den man für den WMAE erhält, wenn man für jeden Datenpunkt eine Null vorhersage, und somit davon ausgehe, dass sich die Kurse der Wertpapiere nicht verändern.

Der massive Einfluss der Gewichtung wird deutlich, wenn man die ungewichtete absolute Abweichung mit dem WMAE vergleicht. Während sich für die Zero-Benchmark im Trainingsdatensatz ungewichtet lediglich ein Wert von 0.00111 ergibt, beträgt die der WMAE 1773.9244. Da sich die Gewichte für jedes Wertpapier bzw. jede Zeile in einem Datensatz unterscheiden, ergeben sich für Trainingsbereich, öffentliches- und privates Leaderboard somit auch jeweils unterschiedliche Zero-Benchmarken, insbesondere zwischen den beiden Leaderboards weichen die Scores deutlich voneinander ab.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Trainingsset | Public LB | Private LB |
| Zero-Benchmark | 1773.92440 | 1770.03211 | 1728.62346 |

Abbildung 9 die jeweiligen Zero-Benchmarks im Vergleich

Gleiches zeigt sich außerdem wenn man den Trainingsdatensatz weiter in Trainings und Validationsdaten unterteilt. Da jede Unterteilung der Daten andere Werte für die Gewichte enthält, errechnet sich somit auch ein neuer WMAE und ein guter Wert bei der Validierung ergibt sich unter Umständen lediglich aus ‚besseren‘ Gewichten und nicht aus besseren Prognosen.

|  |  |  |
| --- | --- | --- |
|  | Trainingsset | Validationsset |
| Split 1 | 1774.320947 | 1772.734739 |
| Split 2 | 1772.171982 | 1779.181636 |
| Split 3 | 1778.381342 | 1760.553556 |

Abbildung 10 Zero-Benchmarks für drei zufällige Splits in Trainings und Validationsdaten mit jeweils 75% Trainingsdaten und 25% Validationsdaten

Veranschaulicht wurde dies in Abbildung 10 auf der für jeweils drei zufällige Unterteilungen die beiden Zero-Benchmarks abgebildet sind. Zu erkennen ist, dass nicht zur zwischen den einzelnen Splits, sondern auch innerhalb eines Splits die Werte der beiden Sets voneinander abwichen. Es sollte also unbedingt darauf geachtet werden, Verbesserungen immer relativ im Vergleich zur entsprechenden Score zu betrachten.

Dieses Phänomän wirkte sich für manche Teilnehmer auch auf ihre Platzierung auf den beiden Leaderboards mehr oder weniger glücklich aus. Quelle

Einmal jupms auf dem leadearboar und unterschiedliche benchmarks je score

Zweimal ergeben sich für subsamples auch wiederum neue scores abhänig von den gewichten und sind u.u nicht gut vergleichbar

## Base Models

Um zu erkunden, wie viel Vorhersagekraft in den von Winton bereitgestellten Daten tatsächlich enthalten ist, wird zu Beginn versucht die beiden zukünftigen Tagesrenditen auf Basis der 25 Features und der vergangenen Tagesrenditen zu prognostizieren. Die zukünftigen Minutenrenditen werden hier außer Acht gelassen, einmal weil im Winton-Statement bereits erklärt wurde, dass sich einzelne Minutenrenditen eigentlich nicht vorhersagen lassen (Anderson 2016) und außerdem weil im Forum-Thread ‚Solution Sharing‘ einige Teilnehmer erkennen ließen, dass sie sich auch darauf beschränkt und trotzdem oder gerade deswegen gute Ergebnisse erzielten. Darunter zum Beispiel auch der Zweitplatzierte Humberto Brandão, der nach eigenen Angaben nur 74 Werte für Ret\_PlusOne prognostizierte und den Rest mit Nullern auffüllte. (The Winton Stock Market Challenge 2016).

Verglichen werden im Folgenden die Ergebnisse für einen XGBoost-Regressor und eine Ridge-Regression. Gewählt wurden diese Algorithmen, da XGBoost aufgrund von Boosting und Ridge durch Shrinkage

als robust gegenüber Overfitting gelten. Nachdem bereits bekannt ist das Feature\_7 besonders behandelt werden sollte, wurden die Algorithmen jeweils einmal mit und einmal ohne trainiert.

Da XGBoost auch mit fehlenden Werten umgehen kann wird dieser auf die rohen Trainingsdaten angewendet, für die Ridge Regression werden zunächst die Features eins, und zehn entfernt, da der Anteil der fehlenden Werte teilweise weit größer als über 40% bzw. über 80% ist und eine Imputation mittels des Mittelwerts vermutlich eher ungenau. Daneben werden die Features fünf, 16 und 20 als kategorial angenommen, da sie maximal 10 unterschiedliche Ausprägungen enthalten und durch One-Hot-Kodierung in Dummy-Variablen überführt. Anschließen werden die fehlenden Werte mit den Mittelwerten angefüllt um so einen lückenlosen Datensatz zu erhalten.

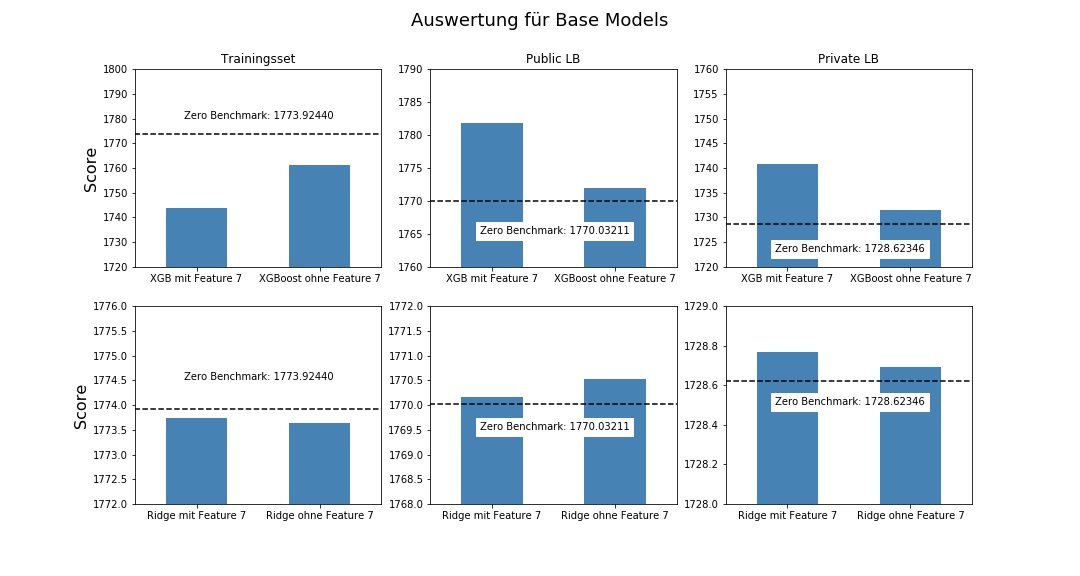


Abbildung 11

Die Ergebnisse für die beiden Algorithmen sind auf Abbildung 11 zu sehen. Beide Algorithmen scheinen prinzipiell zu overfitten, da die Ergebnisse in der Trainingsumgebung die Zerobenchmark mit Leichtigkeit schlagen, auf den Leaderboards allerdings nicht und somit nicht generalisieren. Um ein Gefühl für die Größenordnung zu bekommen, die es im Vergleich zur Benchmark zu erreichen gilt sei gesagt, dass die Score des Erstplatziertem mit 1727.53575 auf dem privatem Leaderboard gerade einmal eine Verbesserung von ca. 1,1 darstellt und ein Wert von 1728.04 bereits gereicht hätte um sich in den Top 10 zu platzieren. Daran wird ersichtlich, dass vor allem die Ergebnisse für xgBoost mehr als bescheiden abschneide und stark variieren. Die Ridge-Regression liefert stabilere Ergebnisse, die zwar auf den Trainingssets nicht ganz so gut wie der XGB abschneiden, allerdings dem entsprechen auf den Leaderboards auch nicht so schlecht. Abgesehen davon scheint Feature\_7 für den XGB tatsächlich Überanpassung zu begünstigen, hier sind die Differenzen zwischen den Trainings- und Test-Scores mit Abstand am größten.

Nun ist es nicht verwunderlich, dass das Vermeiden von Overfitten eine der Hauptaufgaben der Challenge ist, ist es schließlich auch eins der großen Themen im Bereich Machine Learning. Um dem Überanpassen entsprechend zu begegnen ist es erst einmal wichtig zu verstehen, was genau es ist und wo die Ursachen dafür liegen könnten. Domingos (2012) zieht hierfür den Bias-Varianz-Trade-Off herbei und unterteilt Overfitting in „high bias“ und „high variance“. Modelle mit großem Bias neigen dazu aus verschiedenen Trainingsdatensätzen die gleichen Fehler zu lernen und so konsistentere aber verzerrte Schätzungen zu liefern Große Varianz bedeutet abgesehen vom eigentlichem Signal außerdem zufällige Informationen zu lernen und so für verschiedene Datensätze stark abweichende Prognosen zu liefern. Allgemein tendieren lineare Modelle eher zu verzerrten aber konsistenten Schätzungen, und flexiblere Modelle wie z.B. Decision-Trees zu geringerem Bias dafür aber größerer Varianz. Grundlegendes Ziel wäre Modelle zu finden die in beiden Bereichen gut abschneide.

Die Ursachen für Overfitting können vielfältig sein und sind nicht immer nur auf Noise zurück zu führen. Demensprechend gibt es auch keine one-fits-all Lösung um es zu vermeiden, sondern es gilt eher problemspezifisch unterschiedliche Ansätze zu testen und zu kombinieren. Neben Cross-Validation wird von Domingo (2012) Regularisierung angesprochen, also dem Einführen eines Strafterms der Überanpassung vermeiden soll. Bei der Ridge Regression soll das mit dem dem bereits erwähntem Shrinkage-Koeffizienten erreicht werden, der die Regressionskoeffizienten Richtung Null drückt, während XGBoost-Regression versucht das Problem mit Boosting anzugehen und der Algorithmus kombiniert hier viele einfache Modelle zu einem robusterem Komplexen. (Hastie et al. 2017, 61-68 ,337-341). Trotzdem scheint hier keiner der beiden Ansätze recht zu funktionieren allerdings wurden auch jeweils die Standardparameter der in Sklearn bzw xgboost implementierten Algorithem beibehalten, auf Cross-Validation verzichtet und außerdem ist nicht einmal gesichert ob in den Features überhaupt geeignete Informationen enthalten sind.

Demensprechend geht es im nächsten Schritt nun darum eine geeignete Cross-Validation-Methode zu entwickeln, anhand der dann die jeweiligen Modellparameter optimiert werden können, anschließend wird überprüft ob sich durch neue aus den Daten berechnete Features die Modelle weiter verbessern können.

## Time sensitive CV

Das Entwickeln einer funktionierenden Cross-Validation Methode, war das große Problem an dem bei der Challenge so viele gescheitert sind, während sich einige damit abgefunden haben, dass sich die Trainingsergebnisse nicht aufs Leaderboard übertragen und voll und ganz ihrem Cross-Validation-Set vertraut haben, gab es auch einige die statt Cross-Validation jedes Mal ihre Ergebnisse hochgeladen und auf dem Leaderboard überprüft haben. Eine sehr aufwändige Methode die allerdings auch vom drittplatzierten Mendrika Ramarlina angewandt wurde, oder um es in seinen Worten zu sagen: „My conclusion, local CV was useless in this competition.“ (The Winton Stock Market Challenge 2016).

Der Veranstalter der Challenge räumte klassischer Cross-Validation tatsächlich keine großen Erfolgsaussichten ein, stattdessen empfahl er in jedem Fall die Daten so zu teilen, dass sich zwischen Trainings und Validationsdaten keine gemeinsamen Werte für Feature\_7 finden und die Daten so nach Zeitpunkten zu gruppieren. (Anderson 2016)

Die einfachste Form der Cross-Validation ist das einfach aufteilen der Trainingsdaten in Trainings- und Validationsset. Normalerweise werden die Daten zufällig einer der beiden Gruppen zugeteilt, da allerdings soweit möglich keine Überschneidungen für Feature\_7 gewünscht sind, werden die Daten zuerst danach sortiert und dann die ersten 10.000 Einträge der Trainingsgruppe zugeteilt und die nächsten 30.000 dem Validationsset. Auch hier ist wieder zu beachten, dass die Gewichte des WMAE für die Score eine große Rolle spielen, demnach müssen nun für die Trainings- und Validationsdaten wiederum neue Zero-Benchmarken errechnet werden.

Ein Bild, das Screenshot enthält.

Mit sehr hoher Zuverlässigkeit generierte Beschreibung

Abbildung 12

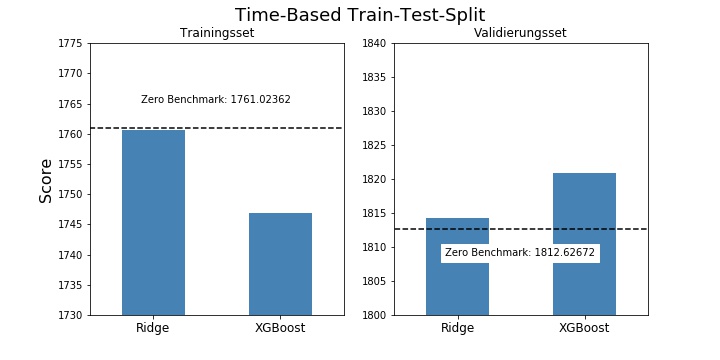


Abbildung 13

Im Vergleich zu einer zufälligen Aufteilung der Daten, bei der der XGBoost die Zeitpunkte auswendig lernen kann, liefert das Validationsset mit Berücksichtigung der Zeitachse ein ähnliches Bild wie für die Testdaten und Feature\_7 scheint tatsächlich der Schlüssel für eine repräsentative Cross-Validation-Methode zu sein. Die Ridge-Regression ist davon zwar nicht ganz so stark betroffen, da es Feature\_7 als diskretes Merkmal wahrnimmt und demnach kaum in einzelne Zeitpunkte separieren kann, trotzdem trat auch hier eine Annäherung an das Bild der Testscore ein.

Auch wenn der Train-Test-Split relativ leicht und umgänglich zu implementieren ist und im ersten Resultat auch vielversprechende Ergebnisse erzielt hat, hat er auf der anderen Seite zwei entscheidende Nachteile: Erstens können die Ergebnisse auf dem Validierungsset stark von den dort enthaltenen Daten abhängig sein, verschiedene Aufteilungen in die jeweiligen Gruppen stark abweichende Ergebnisse erzielen bzw. die Ergebnisse des Validationsset lassen sich ggf. schlecht auf andere Daten übertragen. Zweitens kann auf einem Viertel der Daten nun nicht mehr trainiert werden, weil das für die Validierung reserviert ist. Unter Umständen gehen so wertvolle Informationen verloren. (James et al. 2017, 176.178)

Besser ist meistens die Verwendung einer k-Fold Cross-Validierung. Hier wird der Datensatz in k aufeinanderfolgende Gruppen aufgeteilt und so entsprechend durchiteriert das jeweils auf k-1 Gruppen trainiert und die verbleibende geschätzt, bis für alle Folds eine Score vorliegt, dadurch kann der gesamte Datensatz zum Trainieren benutzt werden. Die Scores der verschiedenen Folds werden am Ende gemittelt. Um auch hier die zeitliche Ordnung zu beachten, muss eine Überschneidung von Feature\_7 zwischen den Folds vermieden werden. Sklearn bietet verschiedene Cross-Validation Methoden an, die die Folds so zusammenstellen, dass sich die Daten entsprechend eines Gruppenparameters nicht überschneiden, darunter GroupKFold,, der einfach nur sich nicht überlappende in etwa gleich große Folds erstellt und GroupShuffleSplit, bei dem nicht in Folds aufgeteilt wird,.

Fehlt noch die beurteilung verschiedeneer verfahren und überleitung, dass in den bestehenden features keine Informationen enthalten sind und demnach feature engineering die lösung sein sollte

## Advanced Feature Engineering

## 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

Anderson, James (2015): The Winton Stock Market Challenge. New Holiday Data from Winton. Online verfügbar unter https://www.kaggle.com/c/the-winton-stock-market-challenge/discussion/18006.

Anderson, James (2016): The Winton Stock Market Challenge. Congratulations, Thoughts on the Problem. Online verfügbar unter https://www.kaggle.com/c/the-winton-stock-market-challenge/discussion/18645.

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).

Domingos, Pedro (2012): A few useful things to know about machine learning. In: *Commun. ACM* 55 (10), S. 78–87. DOI: 10.1145/2347736.2347755.

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.

Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome H. (2017): The elements of statistical learning. Data mining, inference, and prediction. Second edition, corrected at 12th printing 2017. New York, NY: Springer (Springer series in statistics).

James, Gareth; Witten, Daniela; Hastie, Trevor; Tibshirani, Robert (2017): An introduction to statistical learning. With applications in R. Corrected at 8th printing. New York, Heidelberg, Dordrecht, London: Springer (Springer texts in statistics).

Kaiser, Jiří (2014): Dealing with Missing Values in Data. In: *Journal of Systems Integration* 5 (1), S. 42–51. Online verfügbar unter http://si-journal.org/index.php/JSI/article/viewFile/178/255, zuletzt geprüft am 09.08.2018.

Marr, Bernard (2016): Big data in practice. How 45 successful companies used big data analytics to deliver extraordinary results. Chichester, West Sussex: Wiley.

Schmid, Friedrich; Trede, Mark (2006): Finanzmarktstatistik. Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg. Online verfügbar unter http://dx.doi.org/10.1007/3-540-29795-2.

The Winton Stock Market Challenge. Overview (2015). Online verfügbar unter https://www.kaggle.com/c/the-winton-stock-market-challenge.

The Winton Stock Market Challenge. Solution Sharing (2016). Online verfügbar unter https://www.kaggle.com/c/the-winton-stock-market-challenge/discussion/18584.