Stock Market prediction

Veranstaltung: „Neueste Trends in Big Data Analytics“
Betreuer: Julian Kunkel

CLEMENS BECKER

Agenda

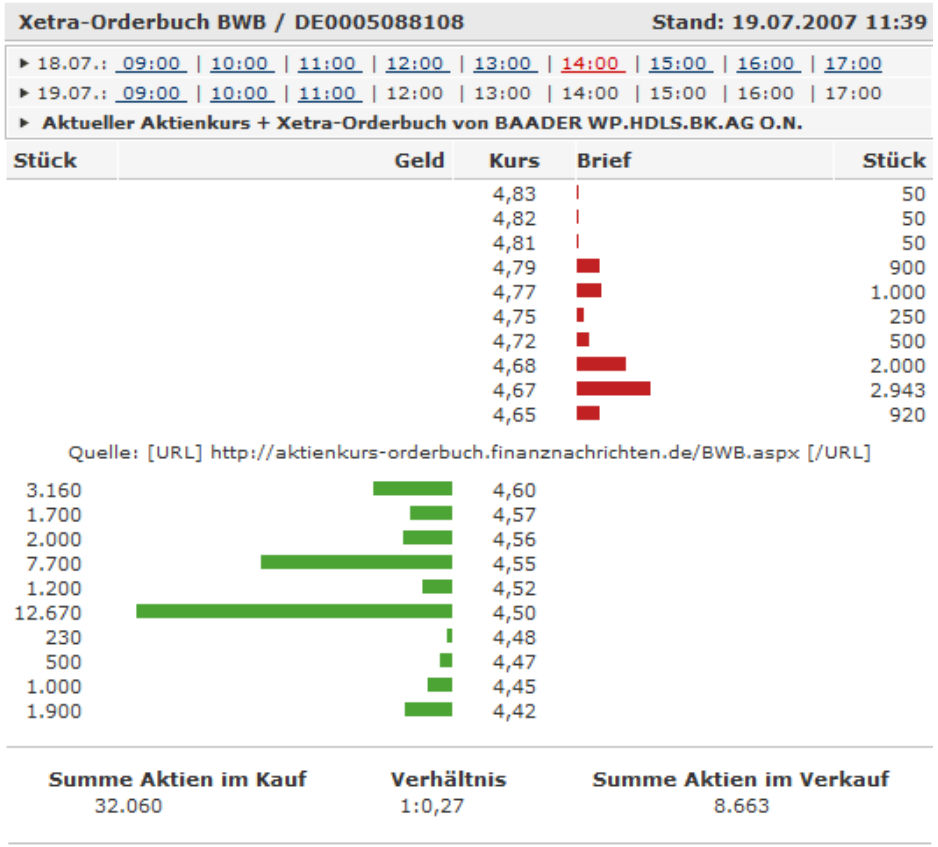
1. Allgemeine Erläuterung des Aktienmarktes
2. Zeitreihen & ARIMA-Modell
3. Vorhersage durch Finanzartikel
4. Aktuelles Beispiel
5. Zusammenfassung

Aktien

Quelle: Aktienerläuterung, Finanztraum, http://finanz-traum.de/wp-content/uploads/2015/07/so-funktionieren-anleihen_1435700885142_block_1.png

Quelle: Dividendenerläuterung, Finanztraum, http://finanz-traum.de/wp-content/uploads/2015/07/so-funktionieren-anleihen_1435700885142_block_2.png

Ankauf und Verkauf: Orderbuch



Xetra-Orderbuch BWB / DE0005088108 **Stand: 19.07.2007 11:39**

▸ 18.07.: 09:00 | 10:00 | 11:00 | 12:00 | 13:00 | 14:00 | 15:00 | 16:00 | 17:00

▸ 19.07.: 09:00 | 10:00 | 11:00 | 12:00 | 13:00 | 14:00 | 15:00 | 16:00 | 17:00

▸ **Aktueller Aktienkurs + Xetra-Orderbuch von BAADER WP.HDLS.BK.AG O.N.**

Stück	Geld	Kurs	Brief	Stück
		4,83		50
		4,82		50
		4,81		50
		4,79		900
		4,77		1.000
		4,75		250
		4,72		500
		4,68		2.000
		4,67		2.943
		4,65		920

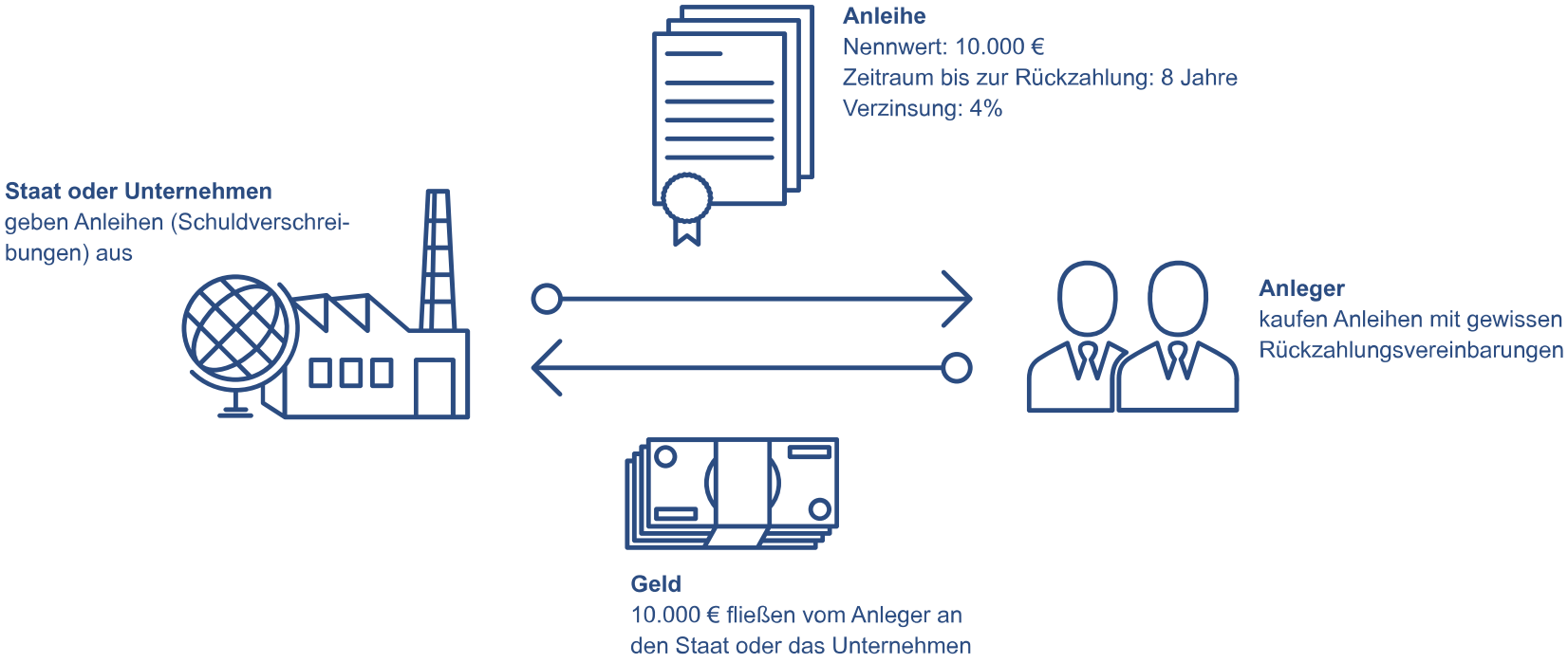
Quelle: [URL] http://aktienkurs-orderbuch.finanznachrichten.de/BWB.aspx [/URL]

3.160	4,60	
1.700	4,57	
2.000	4,56	
7.700	4,55	
1.200	4,52	
12.670	4,50	
230	4,48	
500	4,47	
1.000	4,45	
1.900	4,42	

Summe Aktien im Kauf	Verhältnis	Summe Aktien im Verkauf
32.060	1:0,27	8.663

Quelle: Beispiel für das Xetra Orderbuch, Finanznachrichten.de, http://fns1.de/img/xetra-orderbuch1.gif

Anleihen: Kapitalmarktkredite



Quelle: Anleihen Erläuterung, Deutsche Bank AG,
https://www.maxblue.de/static/assets/content/20151020_mb_infografiken_1140px_20_anleihen.svg

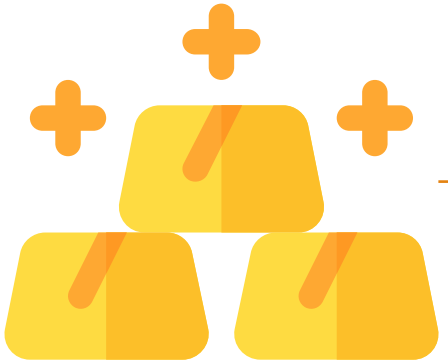
Terminkontrakt

KAUF

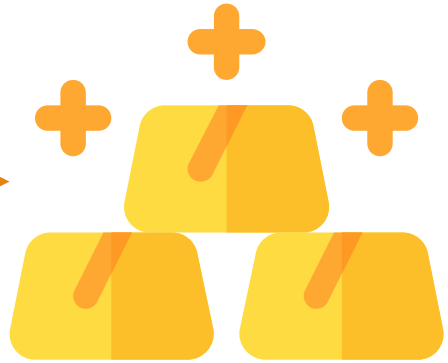
Januar

100 Stück

Dezember



40 $ * 100 = 4.000$



360 $ *100 = 36.000$

400 $ * 100 = 40.000$

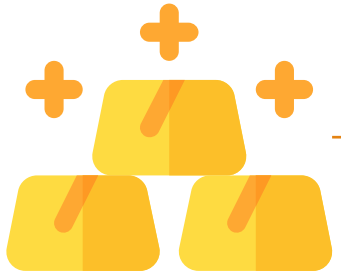
Quelle: Gold, Freepik.com, https://www.flaticon.com/free-icon/gold-ingot_677101#term=gold&page=1&position=7

Terminkontrakt

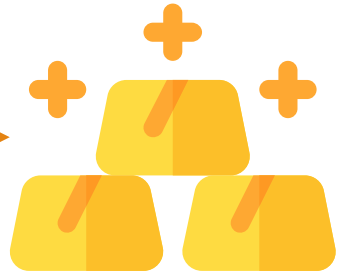
Juni

VERKAUF
100 Stück

Dezember

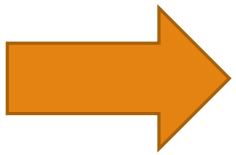


40 $ * 100 = 4.000$



400 $ * 100 = 40.000$

440 $ * 100 = 44.000$

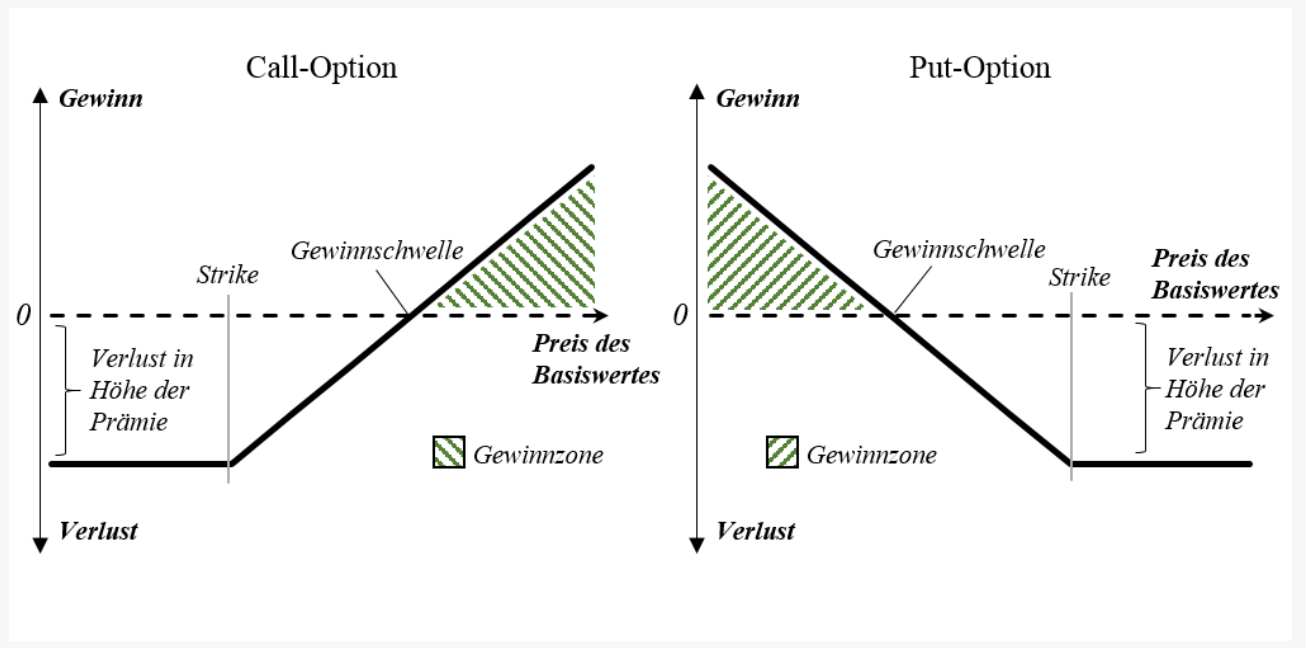


4.000 $ Gewinn

Quelle: Gold, Freepik.com, https://www.flaticon.com/free-icon/gold-ingot_677101#term=gold&page=1&position=7

Optionen

- Käufer erwirbt nur ein Recht zum Kaufen/Verkaufen



Quelle: So funktionieren Call- und Put-Optionen, Dr.Manuel Kay und Sara Zinnecker, http://www.finanztip.de/fileadmin/images/Geldanlage/Optionen/Call_-_Put.png

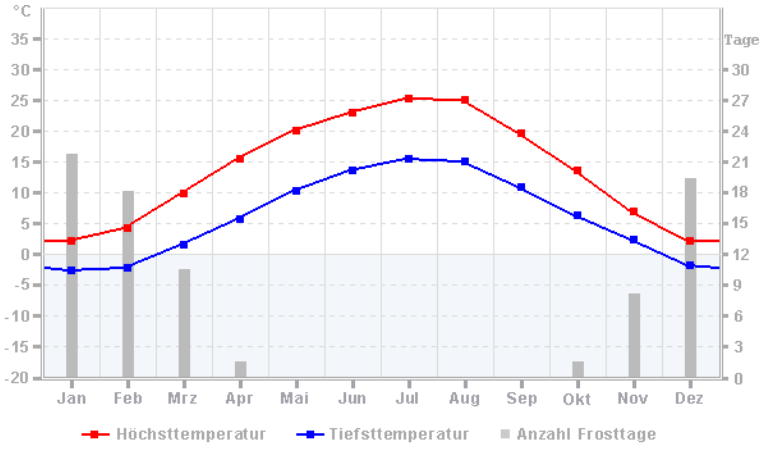
Zeitreihen

- gemessene Datenpunkte in bestimmtem Zeitraum

Quelle: Apple explodiert, Michael, http://www.newgadgets.de/uploads/2012/08/Aktienkurs-Apple-e1345546251378.png

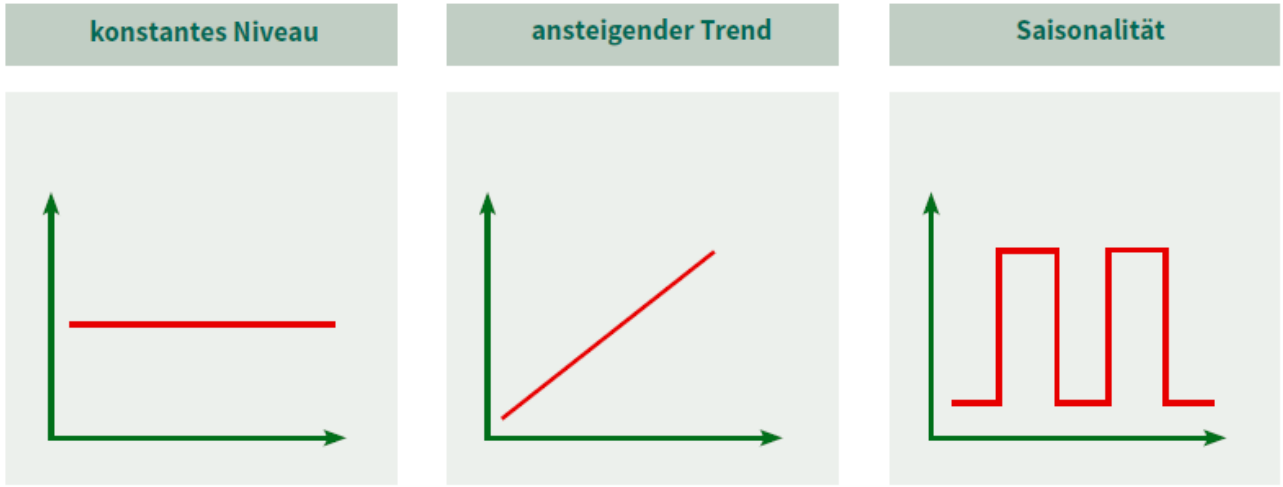
ARIMA-Modell

mögliche zu analysierende Zeitreihen



Quelle: Wetterstation Linz Flughafen, Wetteronline, https://www.wetteronline.de/?pid=p_rueckblick_climatediagram&src=rueckblick/vermarktung/wom/de/p_rueckblick_climatediagram/temperatur/klimadiagramm-11010-linz-flughafen-temperatur.gif

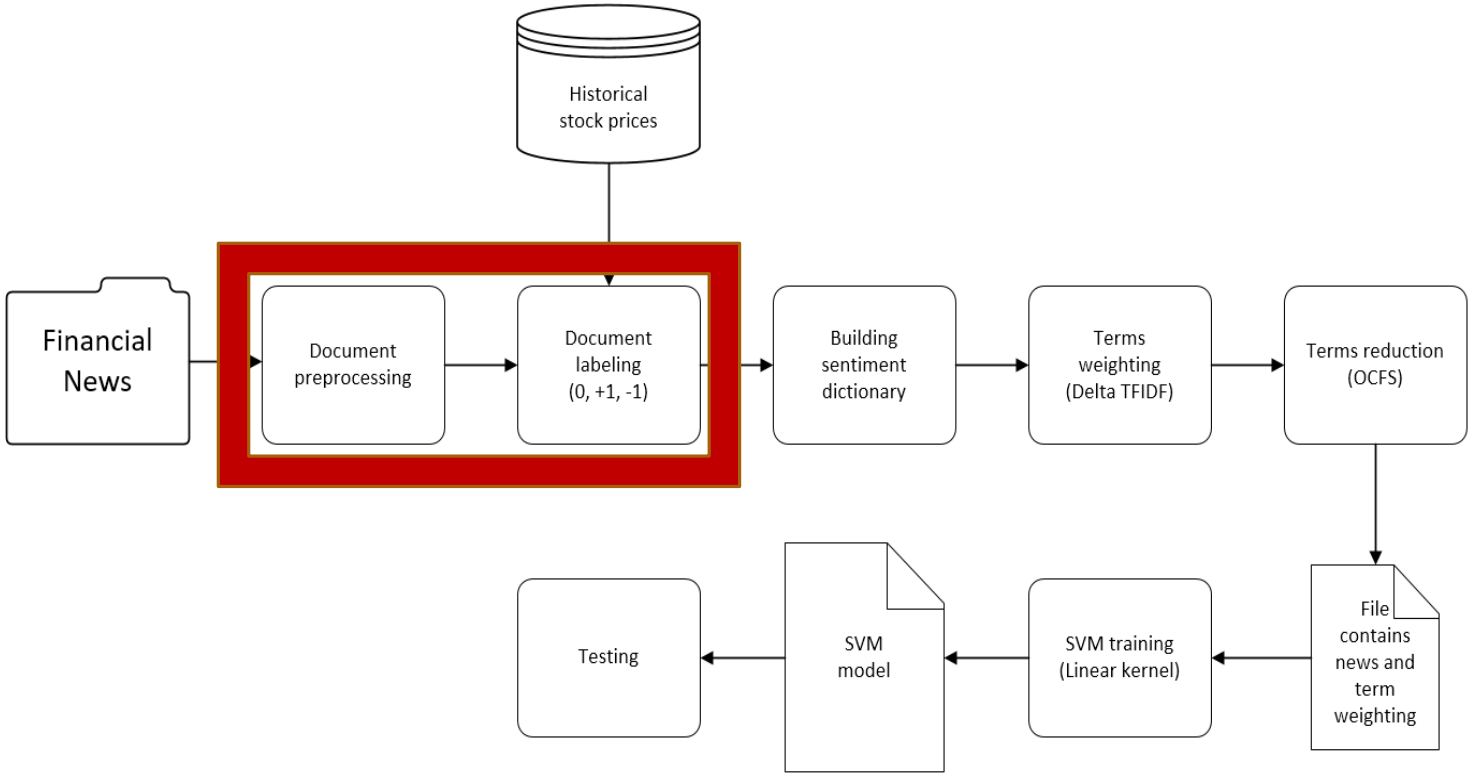
Zeitreihenmuster



Quelle: Muster, die auf Zeitreihen bei der Analyse angewendet werden können, Isabell Tran, https://wr.informatik.uni-hamburg.de/_media/teaching/sommersemester_2016/pir-16-isabella_tran-report.pdf

Vorhersage durch Finanzartikel: Artikelvorbereitung

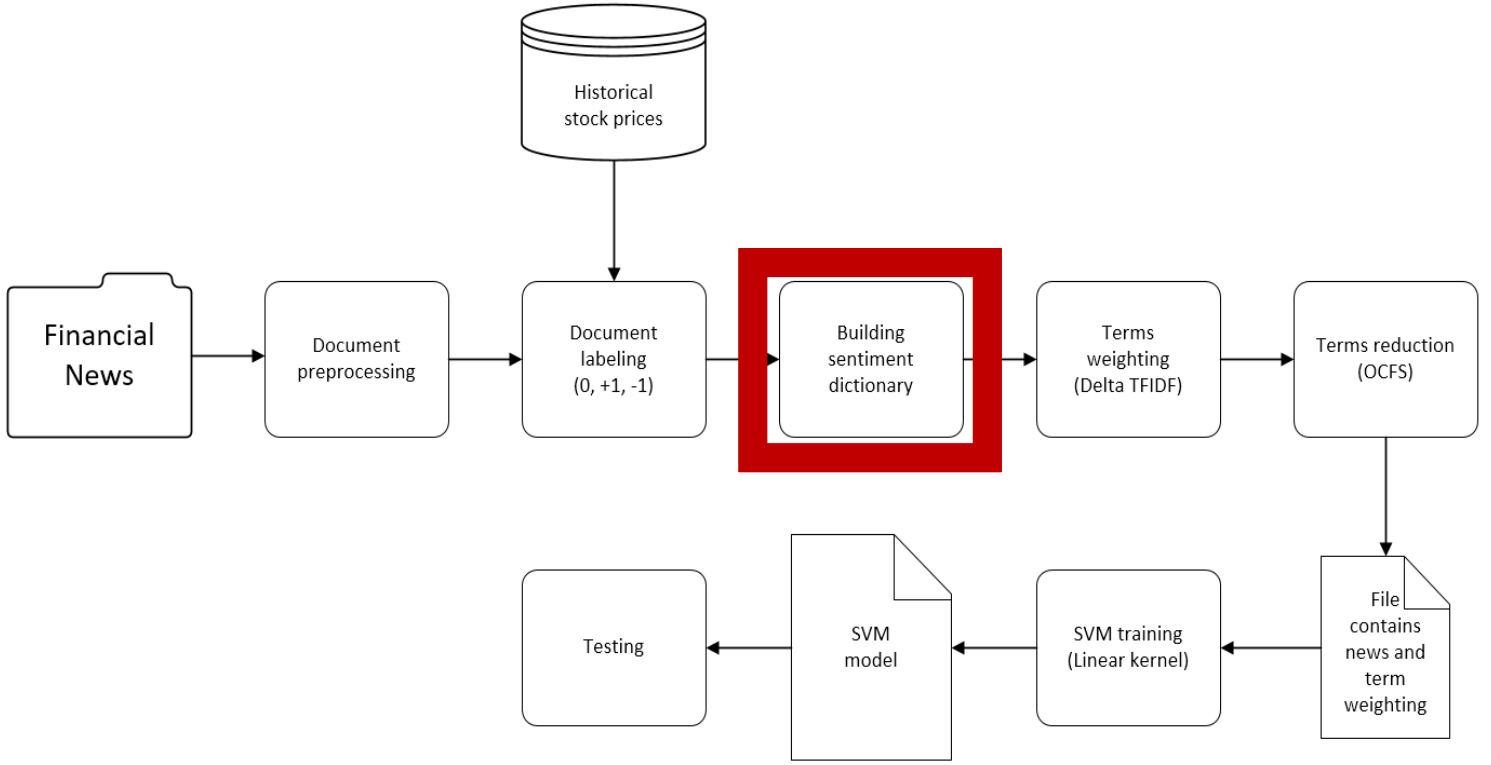
- Verarbeitung und Einteilung der Artikel



Quelle: An overview of our stock trend prediction process, Duc Duong, Toan Nguyen, Minh Dang, https://dl.acm.org/citation.cfm?id=3457619

Sentiment Dictionary

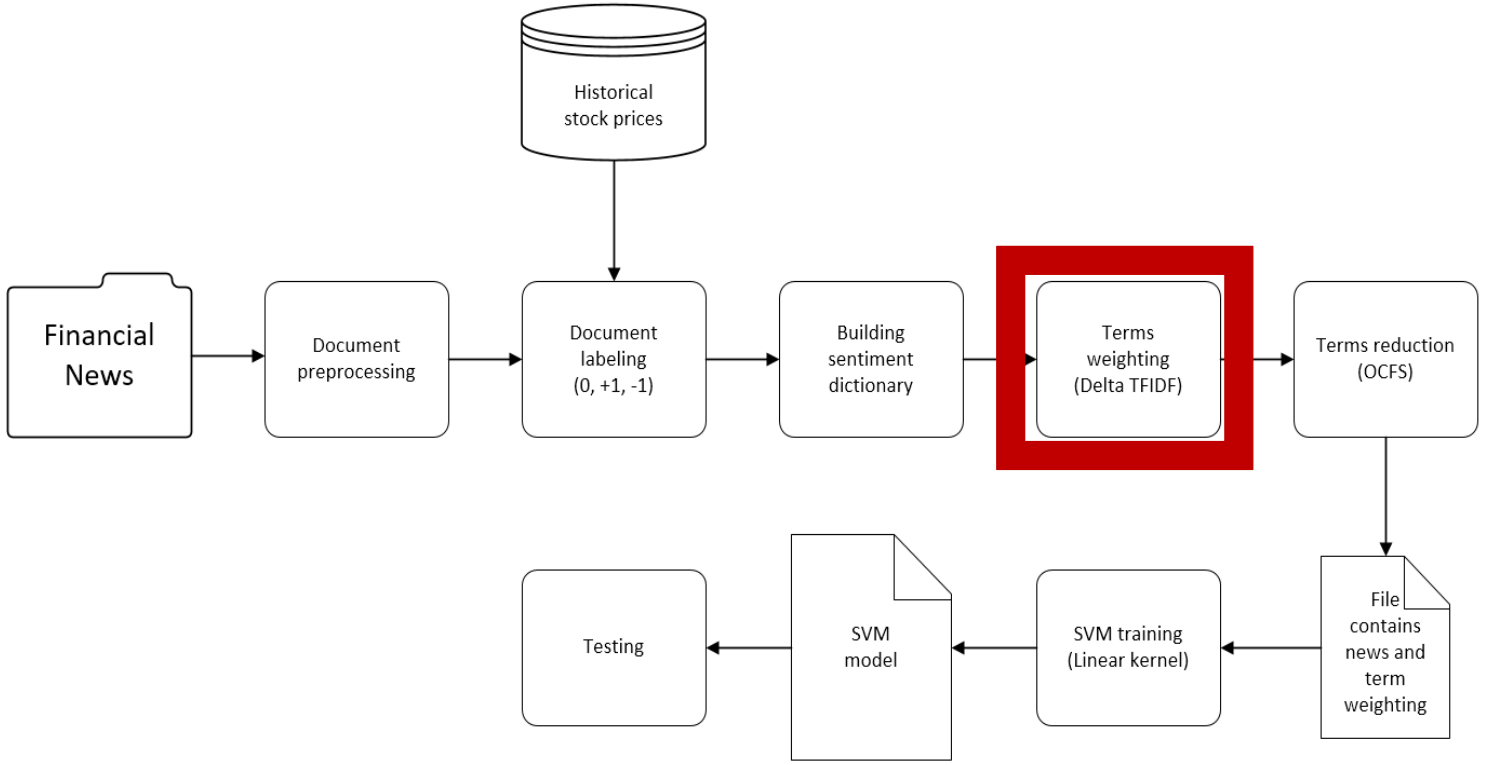
- Enthält nur tatsächlich vorkommende Wörter



Quelle: An overview of our stock trend prediction process, Duc Duong, Toan Nguyen, Minh Dang, https://dl.acm.org/citation.cfm?id=3457619

Gewichtung

- Unterscheidung ob positive oder negative Auswirkung



Quelle: An overview of our stock trend prediction process, Duc Duong, Toan Nguyen, Minh Dang, https://dl.acm.org/citation.cfm?id=3457619

TF-IDF: Term frequency-inverse document frequency

TFIDF Berechnung

I *think* this phone is *good*. [POS]

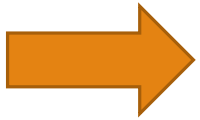
I *think* this phone is *bad*. [NEG]

This is a *good* phone. [POS]

It has a very *bad* user experience [NEG]

Term frequency: 1

Document frequency: 2



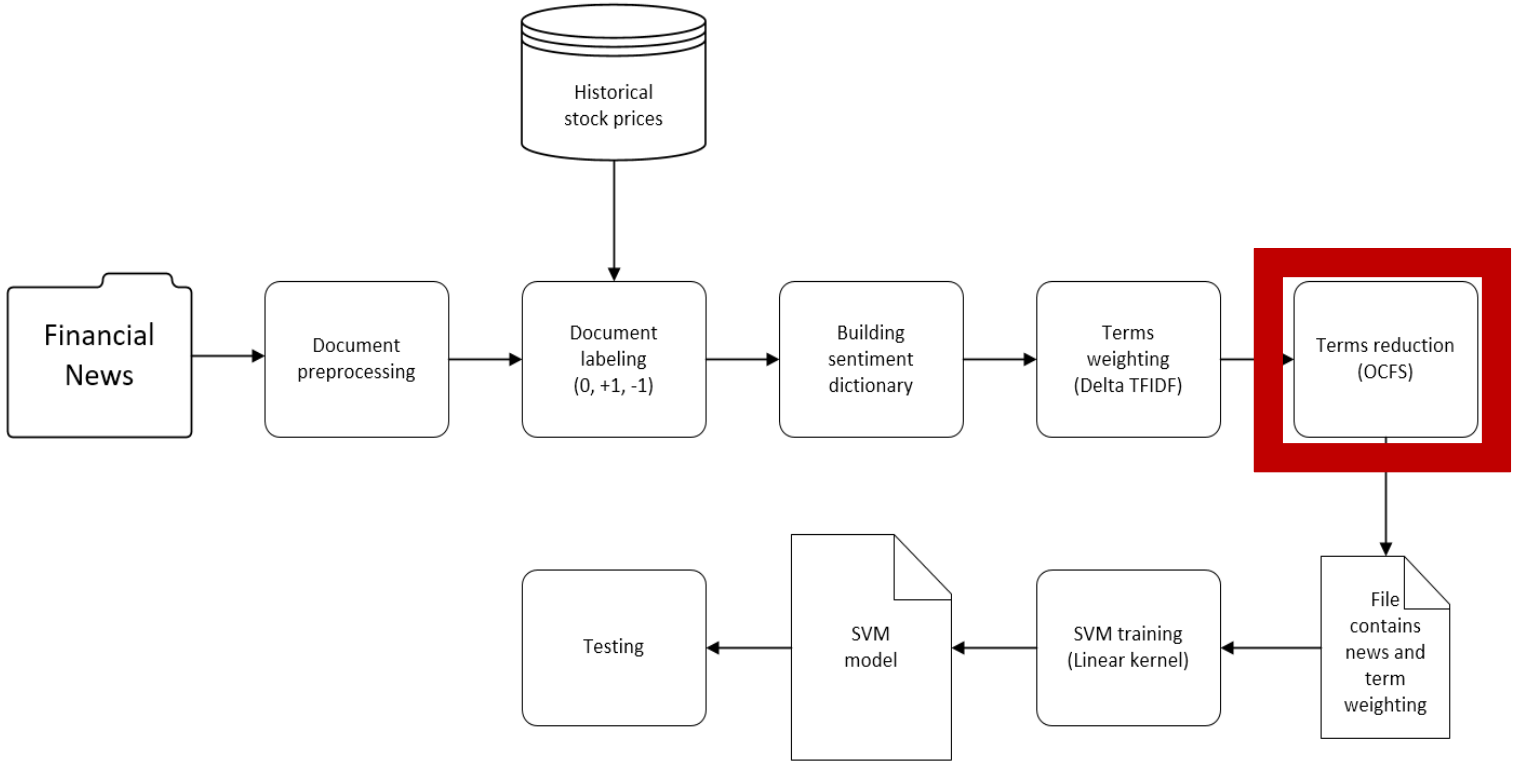
Lösung des Problems

Delta TFIDF = TFIDF_POS - TFIDF_NEG

Quelle: How does Delta TF-IDF work?, Hady Elsahar, https://www.quora.com/How-does-Delta-TF-IDF-work

Reduzierung

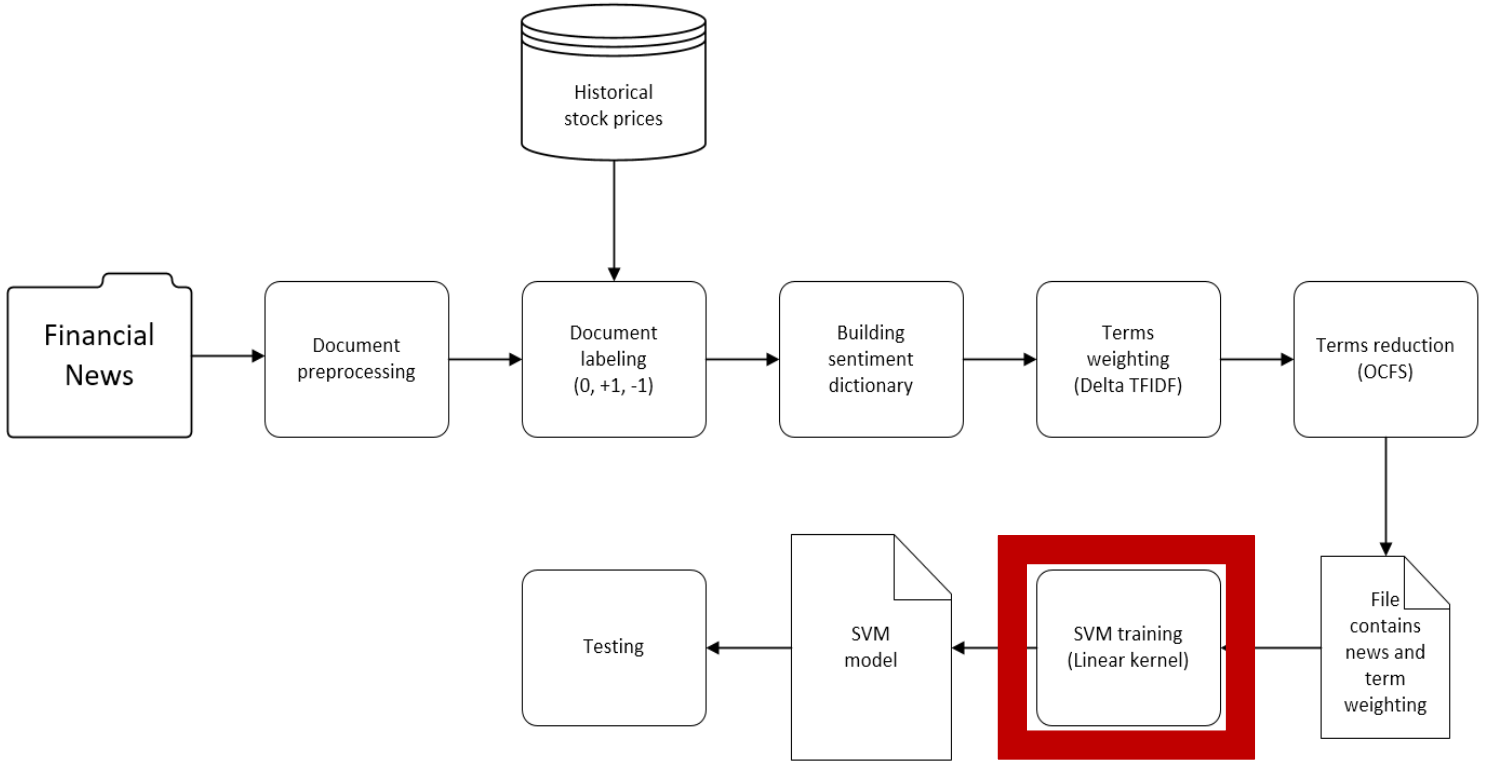
- Rechenzeit-verkürzung



Quelle: An overview of our stock trend prediction process, Duc Duong, Toan Nguyen, Minh Dang, https://dl.acm.org/citation.cfm?id=3457619

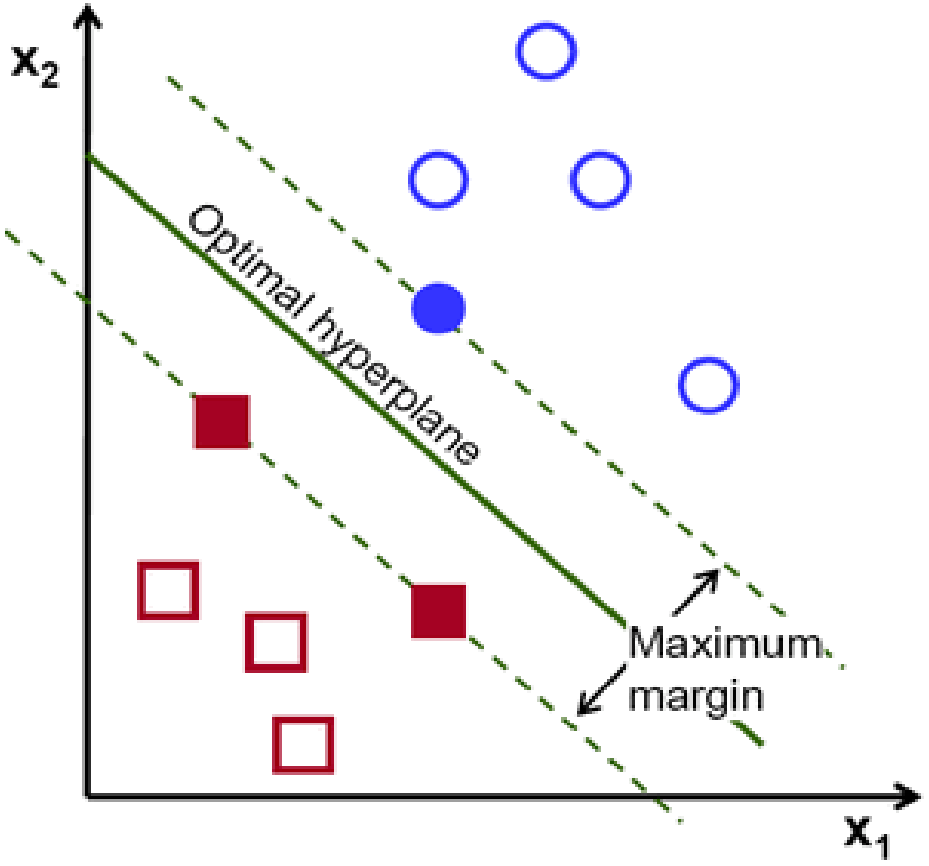
Support Vector Machine

- Binäre Artikel-klassifizierung



Quelle: An overview of our stock trend prediction process, Duc Duong, Toan Nguyen, Minh Dang, https://dl.acm.org/citation.cfm?id=3457619

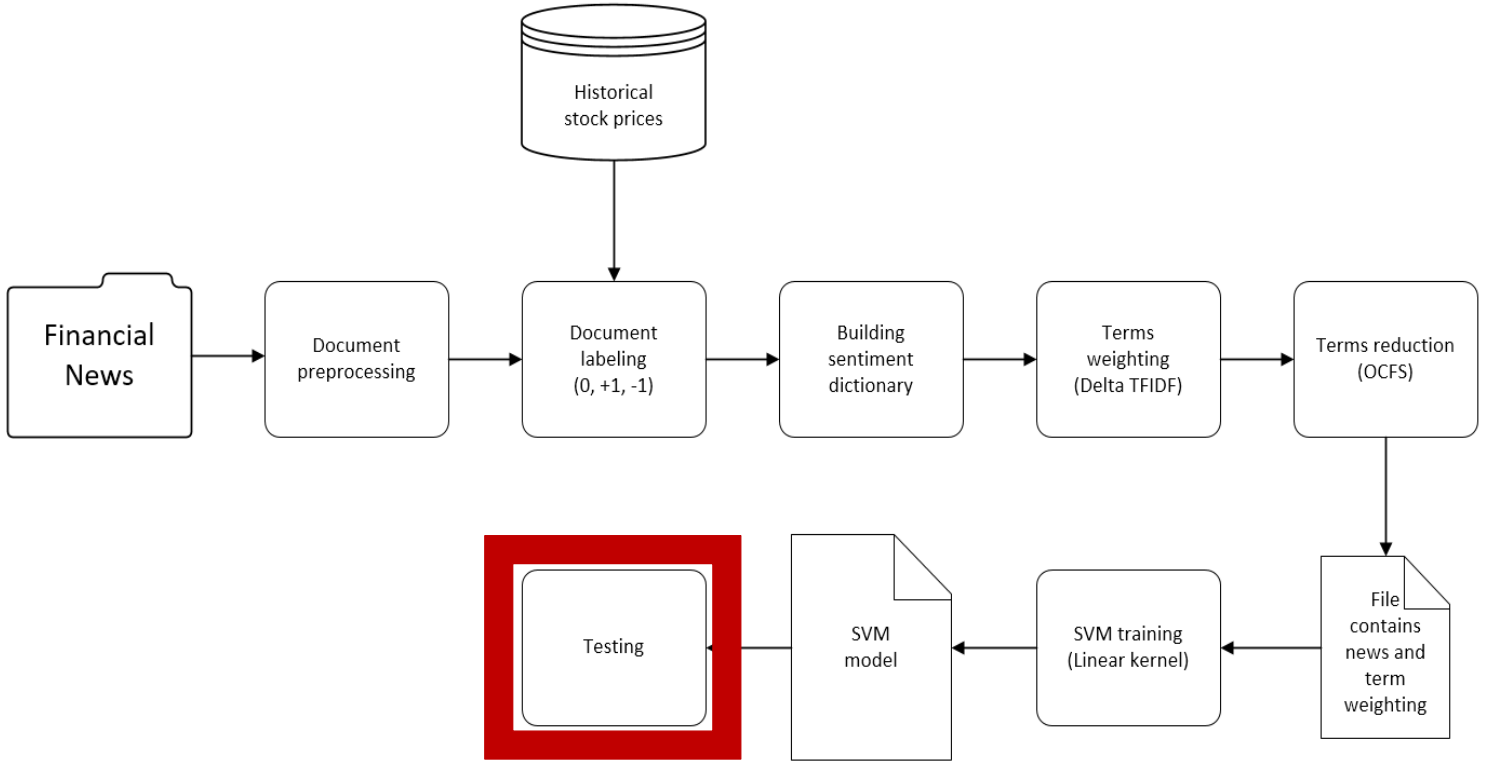
Support Vector Machine



Quelle: What is a SVM?, OpenCV, https://docs.opencv.org/2.4/_images/optimal-hyperplane.png

Datensammlung und Testing

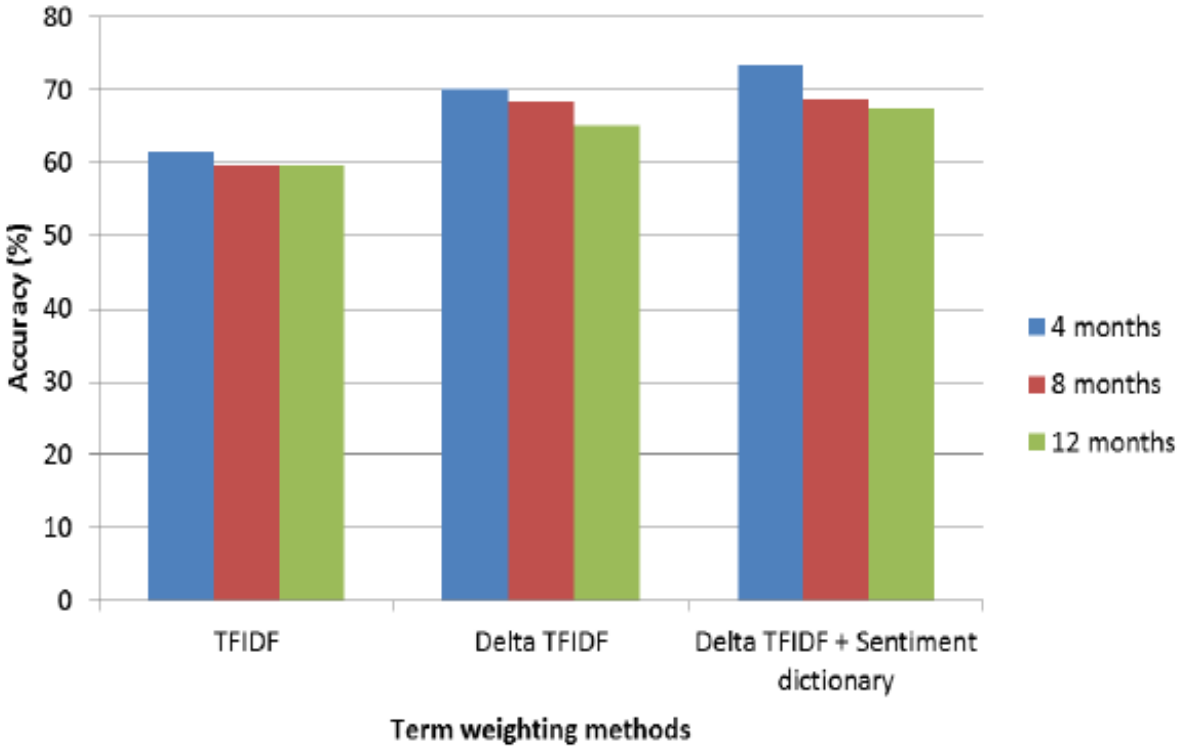
- Sammlung von Artikeln und Aktienkursen



Quelle: An overview of our stock trend prediction process, Duc Duong, Toan Nguyen, Minh Dang, https://dl.acm.org/citation.cfm?id=3457619

Ergebnisse

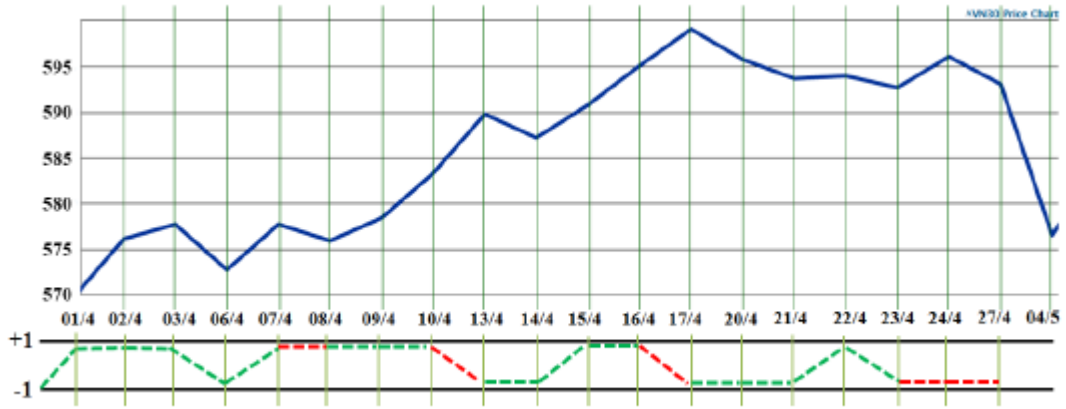
- Genauigkeit = richtig eingeordnete Nachrichten / Gesamtanzahl an Nachrichten



Quelle: Comparison of term weighting techniques, Duc Duong, Toan Nguyen, Minh Dang, https://dl.acm.org/citation.cfm?id=3457619

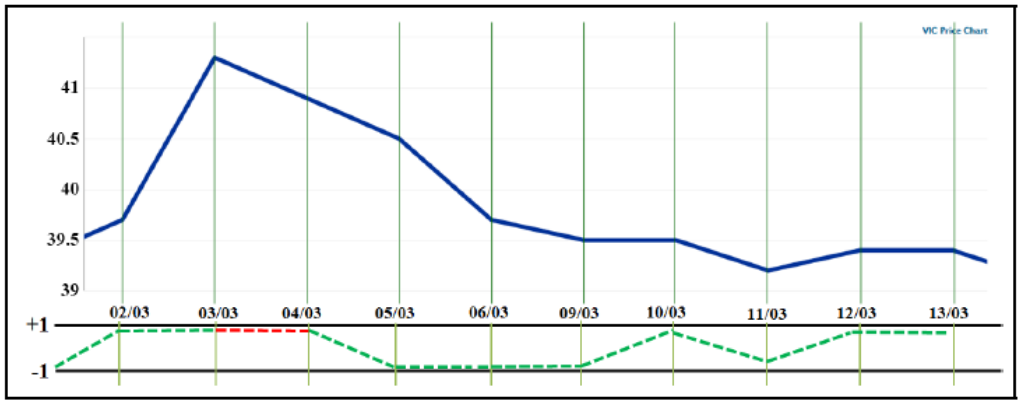
Ergebnisse

Vorhersage des vietnamesischen Aktienindex



Quelle: VN30 price chart with our model, Duc Duong, Toan Nguyen, Minh Dang, https://dl.acm.org/citation.cfm?id=3457619

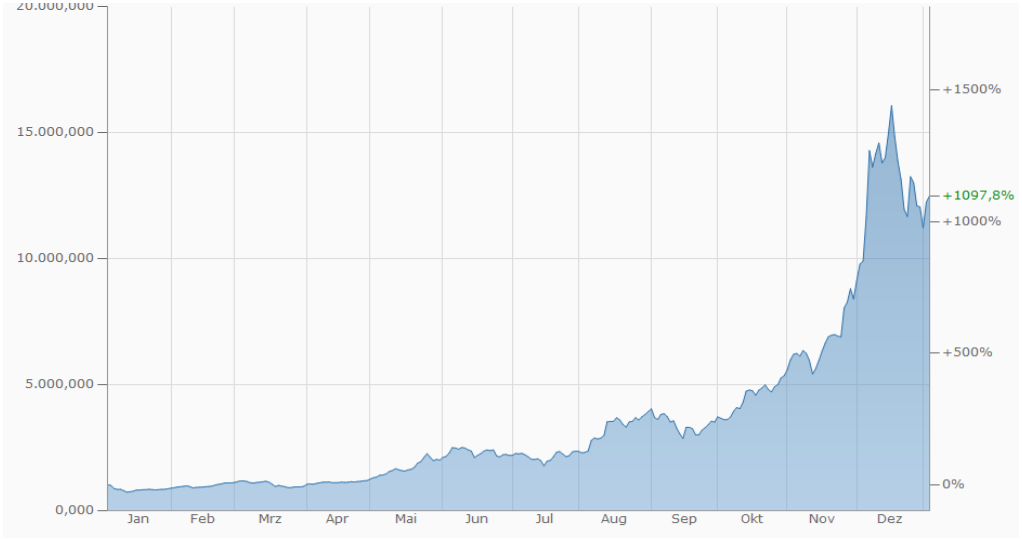
Beste Vorhersage Einzelaktie mit 90% Genauigkeit



Quelle: Trend prediction and VIC price chart, Duc Duong, Toan Nguyen, Minh Dang, https://dl.acm.org/citation.cfm?id=3457619

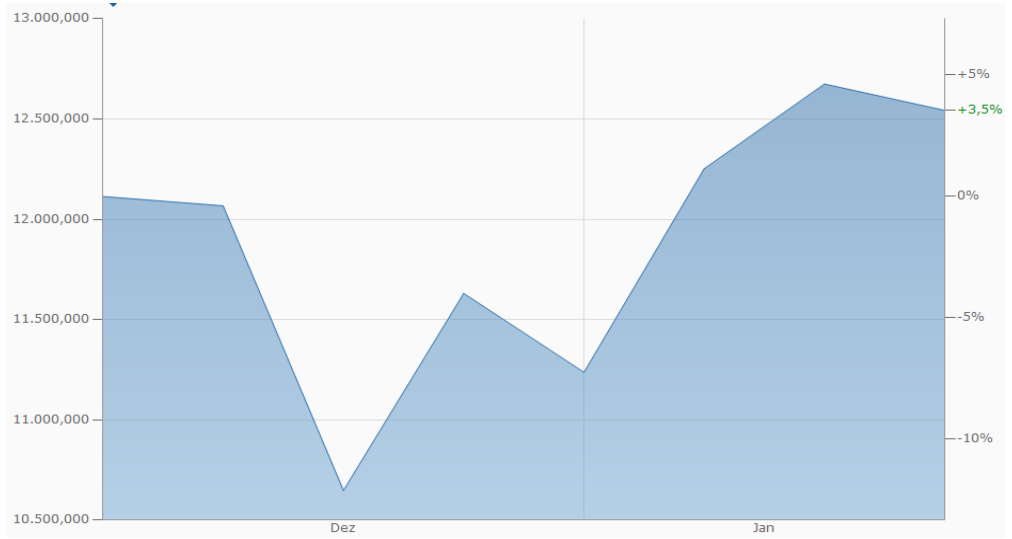
Aktuelles Beispiel

Entwicklung Bitcoin 5 Jahre



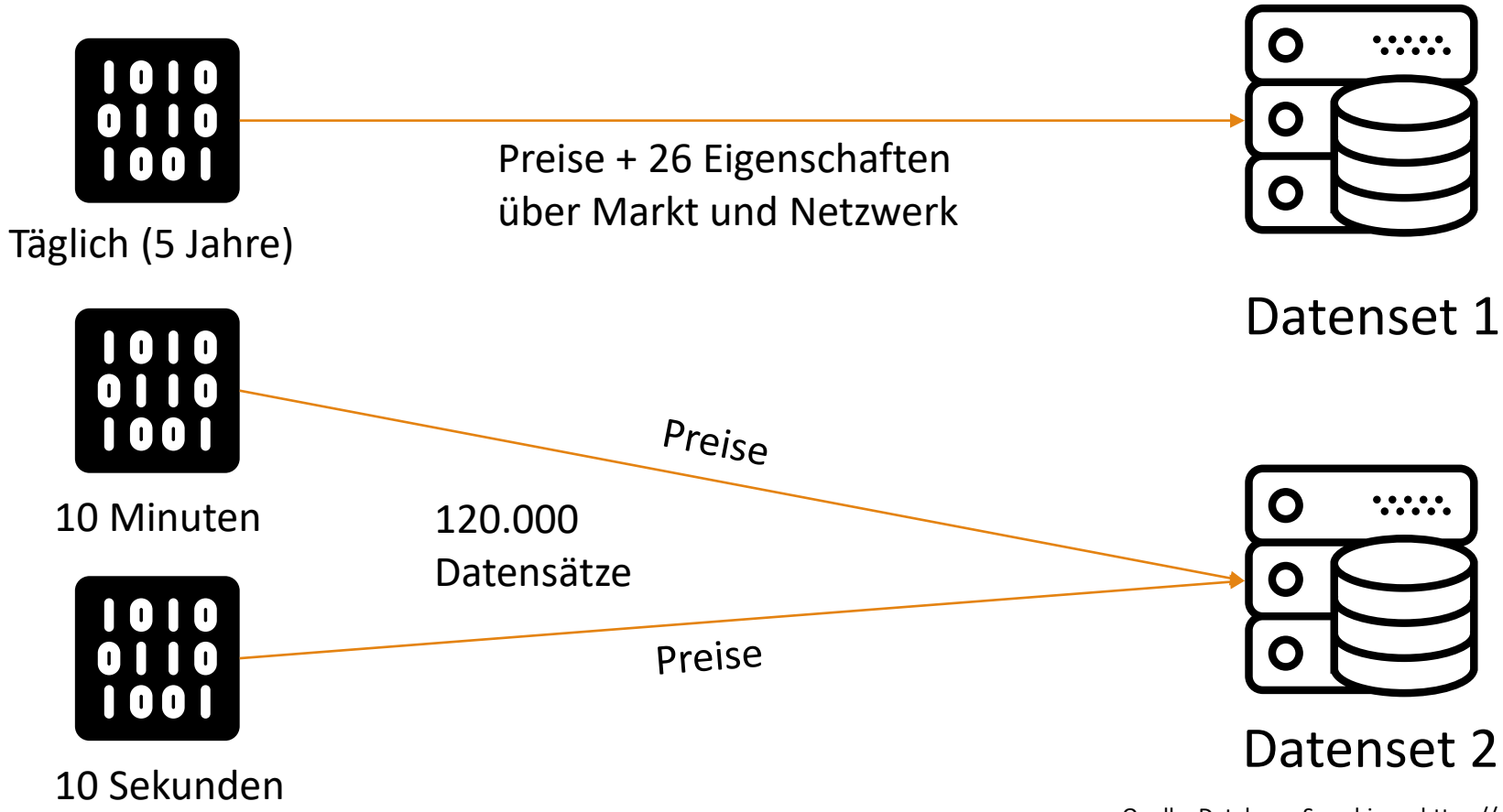
Quelle: Bitcoin – Euro Chart – 1 Jahr, Finanzen.net, https://www.finanzen.net/devisen/bitcoin-euro/chart

Entwicklung Bitcoin 1 Woche



Quelle: Bitcoin – Euro Chart – 1 Woche, Finanzen.net, https://www.finanzen.net/devisen/bitcoin-euro/chart

Datensetzusammenstellung



Quelle: Code, Freepik.com, https://www.flaticon.com/free-icon/binary-code_673159#term=code&page=1&position=37

Quelle: Database, Smashicon, https://www.flaticon.com/free-icon/database_149205#term=database&page=1&position=3

Feature Selection

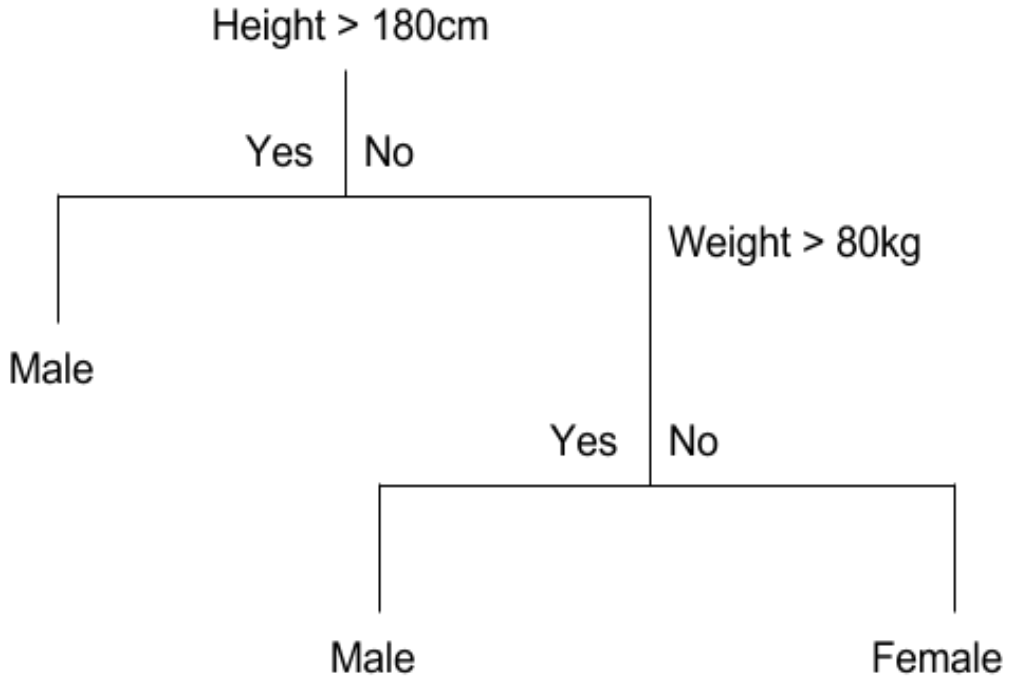
- 16 von 26 Features ausgewählt

FEATURE	DEFINITION
Average Confirmation Time	Ave. time to accept transaction in block
Block Size	Average block size in MB
Cost per transaction percent	Miners revenue divided by the number of transactions
Difficulty	How difficult it is to find a new block
Estimated Transaction Volume	Total output volume without change from value
Hash Rate	Bitcoin network giga hashes per second
Market Capitalization	Number of Bitcoins in circulation * the market price
Miners Revenue	(number of BTC mined/day * market price) + transaction fees
Number of Orphaned Blocks	Number of blocks mined / day not off blockchain
Number of TXN per block	Average number of transactions per block
Number of TXN	Total number of unique Bitcoin transactions per day
Number of unique addresses	Number of unique Bitcoin addresses used per day
Total Bitcoins	Historical total Number of Bitcoins mined
TXN Fees Total	BTC value of transaction fees miners earn/day
Trade Volume	USD trade volume from the top exchanges
Transaction to trade ratio	Relationship of BTC transaction volume and USD volume

Quelle: Feature Selection, Issac Man, Shaurya Saluja, Aojia Zhao, http://cs229.stanford.edu/proj2014/Isaac%20Madan,%20Shaurya%20Saluja,%20Aojia%20Zhao,Automated%20Bitcoin%20Trading%20via%20Machine%20Learning%20Algorithms.pdf

Random Forest

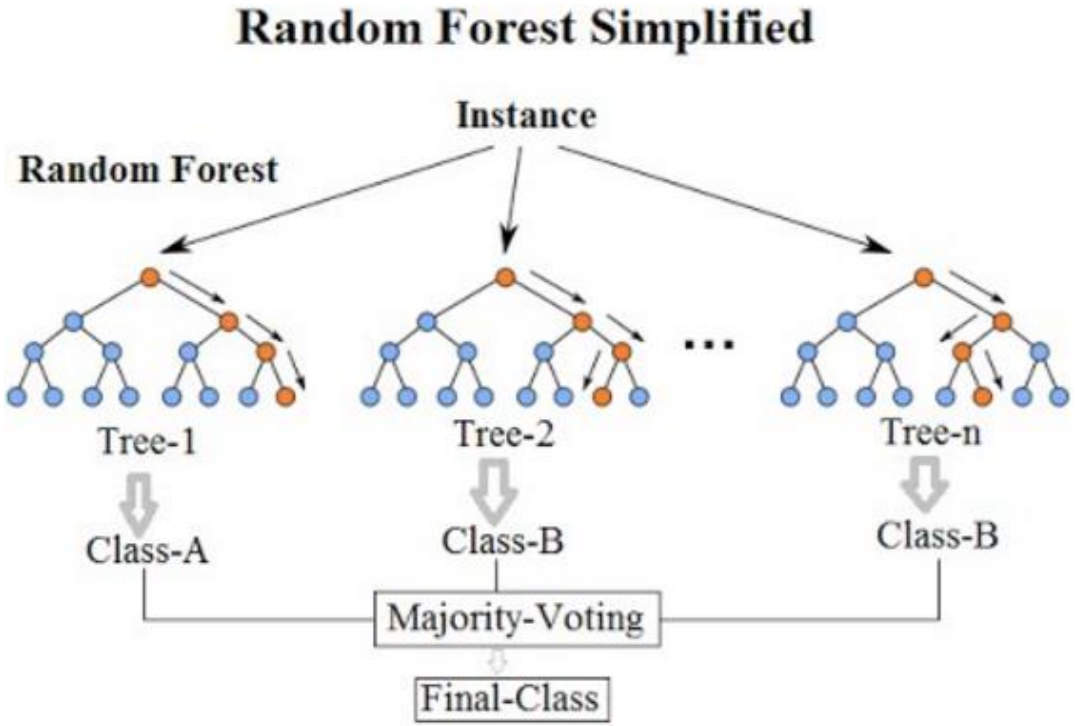
- Random Forests bestehen aus vielen einzelnen CART-Trees



Quelle: Example Decision Tree, Dr.Jason Brownlee, https://3qeqpr26caki16dnhd19sv6by6v-wpengine.netdna-ssl.com/wp-content/uploads/2016/02/Example-Decision-Tree.png

Random Forest

- Classification: abschließendes Voting



Quelle: Random Forest Simplified, Venkata Jagannath, https://d2wh20haedxe3f.cloudfront.net/sites/default/files/random_forest_diagram_complete.png

Random Forest

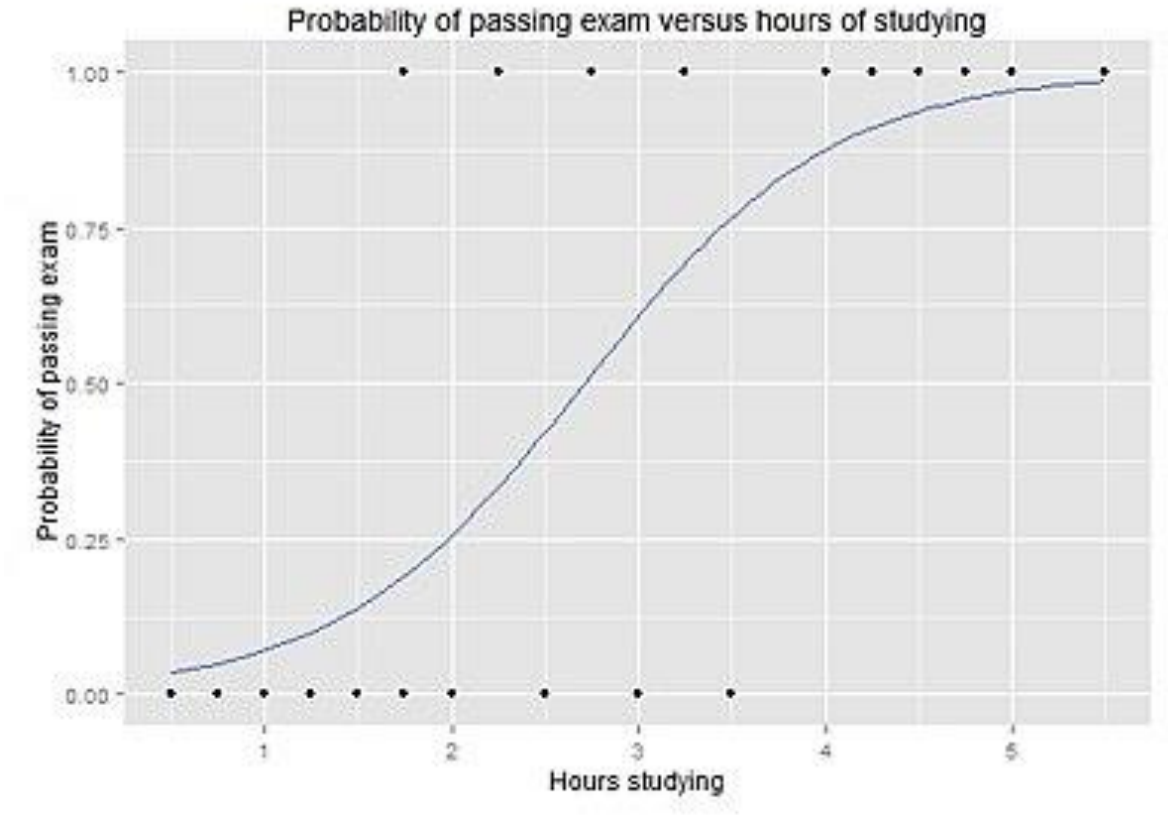
- Regression: Bildung von Durchschnitt

CART	Band	1	2	3
Age	28-40	70%	23%	7%
Gender	Male	70%	27%	3%
Education	Diploma	80%	14%	6%
Industry	Manufacturing	60%	35%	5%
Residence	Metro	70%	20%	10%
Final probability		70%	24%	6%

Quelle: Distribution about salary bands, Tavish Srivastava, https://www.analyticsvidhya.com/blog/2014/06/introduction-random-forest-simplified/

GLM: Generalized Linear Model

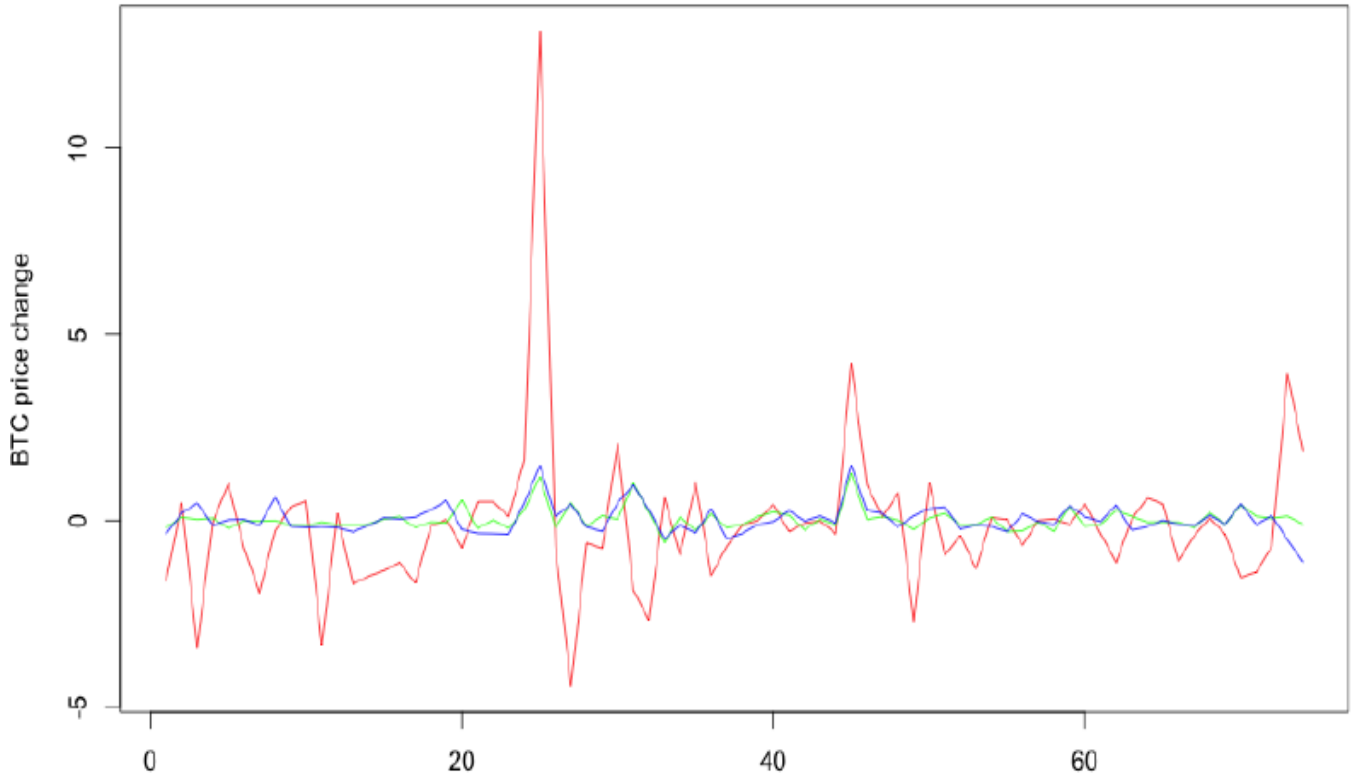
- Logistic Regression



Quelle: probability of passing exams versus hours of studying, Wikipedia, https://upload.wikimedia.org/wikipedia/commons/thumb/6/6d/Exam_pass_logistic_curve.jpeg/400px-Exam_pass_logistic_curve.jpeg

Ergebnisse

- Voraussage GLM (grün) und Random Forest (blau) im Vergleich zu realer Entwicklung (rot)



Quelle: Results on 10 Minute interval Bitcoin data, Issac Man, Shaurya Saluja, Aojia Zhao, http://cs229.stanford.edu/proj2014/Isaac%20Madan,%20Shaurya%20Saluja,%20Aojia%20Zhao,Automated%20Bitcoin%20Trading%20via%20Machine%20Learning%20Algorithms.pdf

Zusammenfassung

- Unterschiedliche Produkte vorausgesagt
- Nicht ein State-of-the-art Ansatz
- Hundert prozentige Voraussage wird schwierig bleiben,

  da Emotionen der Anleger im Markt eine große Rolle

  spielen

Quellen

Tavish Srivastava „Introduction to Random forest – Simplified“ 01.01.2018

https://www.analyticsvidhya.com/blog/2014/06/introduction-random-forest-simplified/

Hady Elsahar „How does Delta TF-IDF work?“ 29.12.2017

https://www.quora.com/How-does-Delta-TF-IDF-work

„Was ist ein Futures-Kontrakt?“ 17.12.2017

https://www.boerse.de/grundlagen/eurex/Was-ist-ein-Futures-Kontrakt-6

Isabella Tran „Einführung und Verarbeitung von Zeitserien“ 2016

https://wr.informatik.uni-hamburg.de/_media/teaching/sommersemester_2016/pir-16-isabella_tran-report.pdf

Quellen

Isaac Madan, Shaurya Saluja, Aojia Zhao „Automated Bitcoin Trading via Machine Learning Algorithms“ 2014

http://cs229.stanford.edu/proj2014/Isaac%20Madan,%20Shaurya%20Saluja,%20Aojia%20Zhao,Automated%20Bitcoin%20Trading%20via%20Machine%20Learning%20Algorithms.pdf

Duc Duong, Toan Nguyen, Minh Dang „ Stock Markt Prediction using Financial News Articles on Ho Chi Minh Stock Exchange“ 2016

https://dl.acm.org/citation.cfm?id=3457619

Logistic Regression, 06.01.2018

https://en.wikipedia.org/wiki/Logistic_regression

Feature Selection using Singular Value Decomposition and orthogonal centroid feature selection for text classification 05.01.2018

https://de.scribd.com/document/334772804/FEATURE-SELECTION-USING-SINGULAR-VALUE-DECOMPOSITION-AND-ORTHOGONAL-CENTROID-FEATURE-SELECTION-FOR-TEXT-CLASSIFICATION-pdf