

DataScience

Vorhersage des Stromverbrauchs im LKH Feldkirch

Fachhochschule Vorarlberg
Studiengang Energietechnik Energiewirtschaft

Betreut von
Dr. Klaus Rheinberger

Vorgelegt von
Mario Rogen

Dornbirn, 30.06.2017

Inhaltsverzeichnis

1.	Einleitung.....	2
1.1.	Ausgangslage.....	2
1.2.	Ziele der Analyse	2
1.3.	Fragestellung	2
1.4.	Literaturreferenzen	Fehler! Textmarke nicht definiert.
2.	Daten	2
2.1.	Quellen	2
2.2.	Qualität.....	3
2.3.	Evtl. Alternativen.....	3
2.4.	Begründung der Datenwahl	3
3.	Methoden.....	3
3.1.	Begründung der Annahmen	3
3.2.	Modelle, Methoden.....	6
4.	Resultate.....	6
4.1.	Grundlegende Tests.....	6
4.2.	Dummies für Tag, Kalenderwoche und Monat	7
4.3.	Autoregressiver Term	7
4.4.	Autoregressiver Term und Werktag/Samstag/Sonntag	9
5.	Diskussion.....	10
5.1.	Ziele erreicht?.....	10
5.2.	Defizite.....	10
5.3.	Verbesserungsvorschläge.....	10
5.4.	Vergleich mit Literatur.....	10

Tabellen- und Abbildungsverzeichnis

Abbildung 1 - Stromverbrauch im Jahr 2016	3
Abbildung 2 - Stromverbrauch März 2016.....	4
Abbildung 3 - Wärme, Kälte, Temperatur 2016	4
Abbildung 4 - Scores ohne zusätzliche Features	6
Abbildung 5 - Übersicht Scores inkl. Autoregressiver Term.....	7
Abbildung 6 - Lineares Modell mit Autoregressivem Term - Split 1	7
Abbildung 7 - Lineares Modell mit Autoregressivem Term - Split 2	8
Abbildung 8 - Lineares Modell mit Autoregressivem Term - Split 3	8
Abbildung 9 - Lineares Modell mit Autoregressivem Term - Split 5	8
Abbildung 10 - Übersicht Scores inkl. Autoregressiver Term und Werktag.....	9
Tabelle 1 - Daten Export aus dem EMC.....	2

1. Einleitung

1.1. Ausgangslage

Im LKH Feldkirch soll das Energiemanagement System (EMC) um vorhandene Stromzähler erweitert werden. Zukünftig werden bis zu 500 Zähler eingebunden. Aktuell ist nur der Gesamtstromverbrauch ersichtlich. Aus diesem Grund wurde mir von Seiten der technischen Leitung der Auftrag gestellt die vorhandenen Daten zu analysieren und die Vorgehensweise bei der Einbindung der zusätzlichen Zähler festzulegen. Anhand der Messwerte soll es in Zukunft möglich sein, die Stromverbräuche der einzelnen Gebäudeteile und Abteilungen bis hin zu den einzelnen Gewerken (Lüftung, Kälte, Wärme, Dampf, Druckluft, Beleuchtung, etc.) und Geräten (einzelne Lüftungsanlagen, Med. techn. Großgeräte, etc.) aufzuzeichnen. Mit den dann ermittelten Daten können Energieeffizienzmaßnahmen zielgerichtet vorbereitet und umgesetzt werden. Ebenfalls ist damit in Zukunft die Wirksamkeit der einzelnen Maßnahmen nachvollziehbar.

1.2. Ziele der Analyse

Um vorab einen Eindruck des Stromverbrauches zu bekommen, soll Anhand dieses Projektes der Gesamtstromverbrauch im LKH Feldkirch genauer untersucht werden. Ziel ist es, mögliche Einflüsse bereits vorab herauszufinden und dann die Vorgehensweise anhand dieser Priorisierten Einflussquelle umzusetzen.

1.3. Fragestellung

Die oben beschriebene Sachlage führte dann zur Fragestellung „Ist es möglich den –Gesamt-Stromverbrauch im LKH Feldkirch vorherzusagen und welche Einflüsse sind dafür verantwortlich?“

2. Daten

2.1. Quellen

Die Daten für diese Projektarbeit stammen aus dem Energiemanagement System der Firma Siemens, welche für das Land Vorarlberg und im speziellen das LKH Feldkirch angeschafft wurde. Ein Großteil der darin vorhandenen Daten wird automatisch über Zähler im Leittechniksystem erfasst und dann in die Datenbank des Energiemanagementsystems übertragen. Es gibt jedoch auch noch ältere Zähler, welche periodisch von Hand eingetragen werden. Sämtliche gespeicherte Daten können als CSV Dateien mit unterschiedlichen Auflösungen exportiert werden. Für diese Auswertung wurde der Zeitraum 01.01.2016 bis 22.04.2017 im 15 Minuten Raster exportiert. In Tabelle 1 sind die exportierten Daten sichtbar.

Tabelle 1 - Daten Export aus dem EMC

	Tag	Strom Gesamt [kWh]	Erdgas VEG [m3]	Brauchwasser Gesamt [m3]	Waerme Gesamt [MWh]	Kaelte Gesamt Erzeugung [kWh]	Kaltwasser Quell + Stadt Gesamt [m3]	Wasser Gesamt (abzüglich BWW) [m3]	FELDKIRCH in C
Datum									
01.01.2016	5	39248.0	7379.1	20.20	51.0	903.33	149.73	129.53	4,2
02.01.2016	6	39232.0	7379.1	20.20	52.9	840.00	161.73	141.53	2,7
03.01.2016	7	37528.0	7379.1	28.88	48.5	856.67	149.73	120.86	4,7

2.2. Qualität

Die Daten sind Großteils von brauchbarer Qualität, jedoch gibt es auch einzelne Ausreißer die vorab korrigiert werden mussten. An einem Tag im September waren an drei Stunden Werte von über 10.000 kWh/h aufgezeichnet. Diese wurden mit dem Mittelwert der vorhergehenden und nachfolgenden Stunde ersetzt. Ebenfalls gab es anfangs Probleme mit den 1000er Trennzeichen und dem Kommazeichen. Dies konnte über einstellbare Optionen beim Export angepasst werden, damit hat der Python Import funktioniert.

Der Datensatz besteht aus 480 Zeilen und 9 Features (bzw. 10 mit dem Datum).

2.3. Evtl. Alternativen

Weitere Daten sind im Leittechniksystem vorhanden, diese müssten jedoch per Befehl aufgezeichnet und dann exportiert werden. Ein Rückblick in die Vergangenheit ist nur bei bestimmten Daten möglich. Die Grundlage dieser Daten ist jedoch die Selbe, weshalb ein Export nur begrenzt sinnvoll ist. Weitere Daten sind im jährlichen Energiebericht einsehbar, diese können jedoch aufgrund der geringen Auflösung nicht für Detailprognosen verwendet werden.

2.4. Begründung der Datenwahl

Die ausgewählten Datenreihen beziehen sich jeweils auf den Gesamtverbrauch eines Gewerks, eine bessere Aufteilung auf einzelne Bereiche ist für relativ neu installierte Anlagen vorhanden, jedoch nicht für alle. Aus diesem Grund kann man damit keine Sinnvollen Rückschlüsse auf den Gesamtstromverbrauch schließen.

3. Methoden

3.1. Begründung der Annahmen

Nach dem Einlesen der Daten ist es sinnvoll diese zu visualisieren. Die definierte Zielgröße, der Stromverbrauch über das Jahr schaut in der Abbildung 1 folgendermaßen aus:

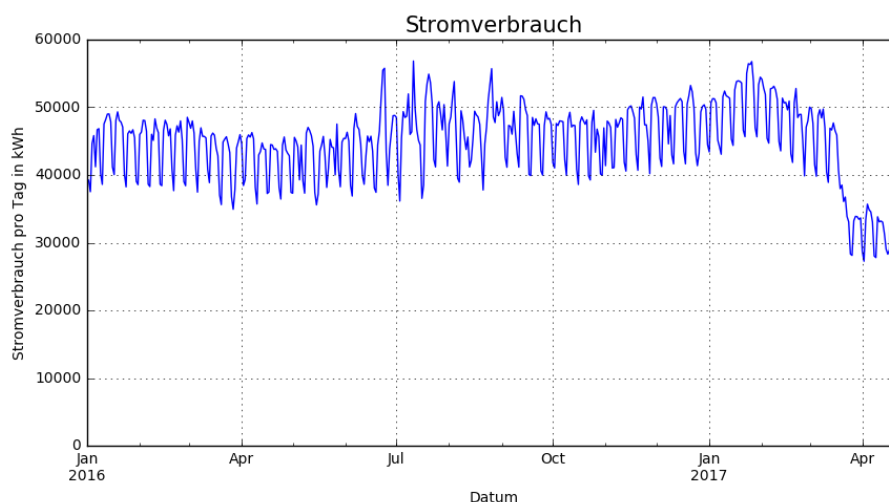


Abbildung 1 - Stromverbrauch im Jahr 2016

Man kann im Sommer starke Schwankungen sehen, der Grund ist aktuell nicht bekannt. Ein weiteres auffälliges Merkmal ist ab Ende März erkennbar. Der um ca. 10.000 kWh/Tag niedrigere Stromverbrauch ist auf eine fehlerhafte Einbindung der unterschiedlichen

Hauptstromzähler ins EMC zurückzuführen. Eine Trafostation wurde vermutlich beim Umbau nicht mehr eingebunden. Aus diesem Grund wird für die Bearbeitung nur der Zeitraum 01.01.2016 bis 31.12.2016 hergenommen.

Ein weiteres offensichtliches Merkmal der Daten sind die niederen Verbräuche an den Wochenenden. Im Detail ist dies im Abbildung 2 ersichtlich:

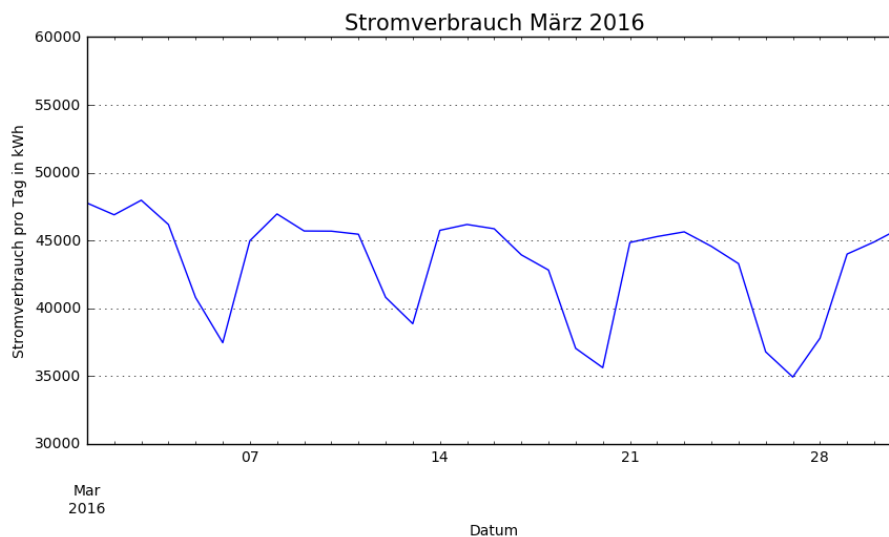


Abbildung 2 - Stromverbrauch März 2016

Weitere Visualisierungen wurden für die anderen Größen erstellt. Eine Kombination aus der bereitgestellten Wärme, Kälte und der Tagesdurchschnittstemperatur ist in Abbildung 3 sichtbar:

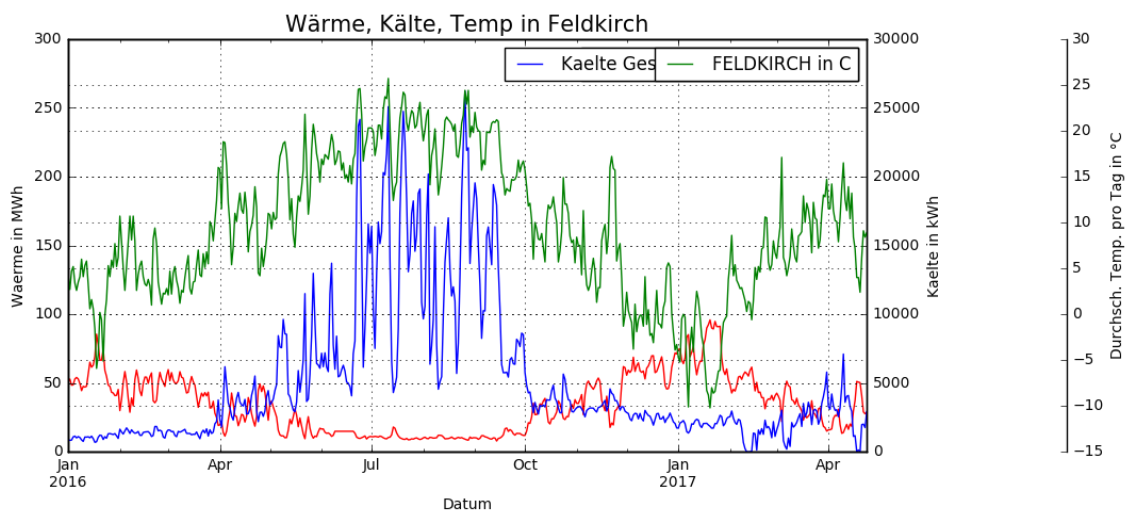


Abbildung 3 - Wärme, Kälte, Temperatur 2016

3.2. Zusammenhang der Daten

Anhand dieser Datenlage sind aktuell keine direkten Zusammenhänge, außer dem Wochentag, ersichtlich. Dies wurde mit Hilfe eines Scatterplots dann auch nochmals zusätzlich verifiziert und dargestellt.

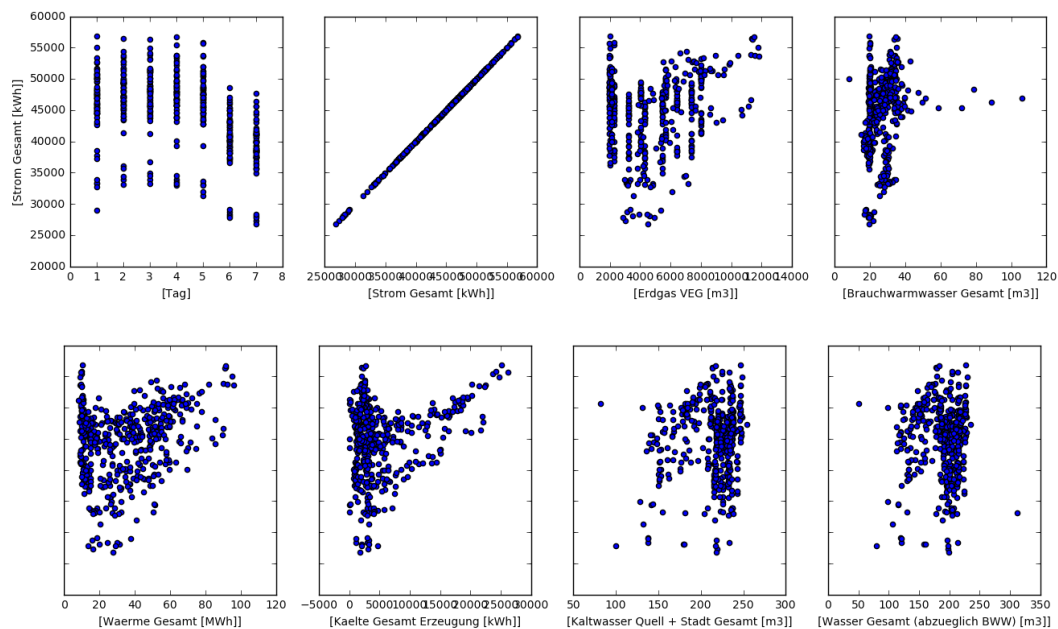


Abbildung 4 - Scatter Matrix der einzelnen Features

3.3. Modelle, Methoden

Da es sich um zeitkontinuierliche Daten handelt, kommen nur Regressionsalgorithmen in Frage. Aktuell werden folgende Regressoren eingesetzt:

- LinearRegressor
- KNeighborsRegressor
- DecisionTreeRegressor
- RandomForestRegressor

Um die Effektivität des Algorithmus zu evaluieren, wurde auf das sklearn Modul „TimeSeriesSplit“ zurückgegriffen.

Mögliche Defizite

Da die Daten aktuell nur für ein Jahr gespeichert sind, können saisonale Einflüsse nicht berücksichtigt werden. Somit sind auch Features wie die Kalenderwoche oder der Monat nur bedingt für die Vorhersage sinnvoll.

4. Resultate

4.1. Grundlegende Tests

Als erstes wurde eine Vorhersage mit allen verfügbaren Regressoren für die im Datensatz vorhandenen Features ausprobiert. Dabei hat sich die Vermutung aus der Betrachtung der Scatterplots bestätigt, es hat nur der Tag einen direkten Einfluss auf die Vorhersagequalität. Alle anderen Features (Wärme, Kälte, Temperatur, etc.) helfen dem Algorithmus nicht.

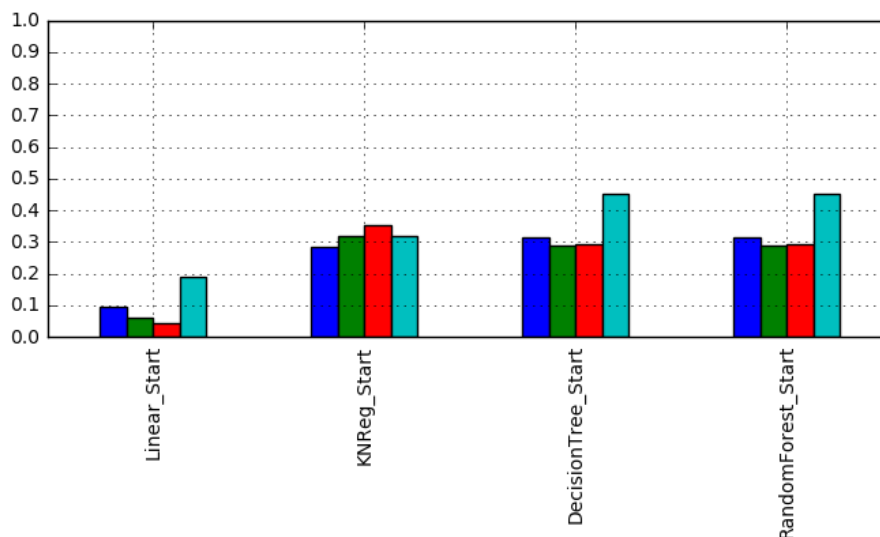


Abbildung 5 - Scores ohne zusätzliche Features

Nur mit Hilfe des Tages als Nummer von 1 bis 7 ergeben sich R^2 Werte zwischen 0,1 und 0,45. Dabei kann man gut erkennen, dass die Vorhersagequalität abhängig vom Split ist.

4.2. Dummies für Tag, Kalenderwoche und Monat

Um den Regressoren mehr Features zur Verfügung zu stellen, wurden sogenannte Dummie-Variablen für den Tag, die Kalenderwoche und den Monat eingefügt. Mit Hilfe von diesen wurde die Anzahl der Features von 1 auf 72 erhöht. Die Qualität der Vorhersage konnte damit jedoch nicht gesteigert werden. Die große Anzahl an Features hat sogar teilweise zu Verschlechterungen der Ergebnisse geführt.

4.3. Autoregressiver Term

Als nächster Versuch wurde der Tagesstromverbrauch vom Vortag als zusätzliches Feature hinzugefügt. Dies würde in der Praxis die Vorhersage jedoch stark einschränken, da man nur noch einen Tag in die Zukunft eine Prognose machen könnte.

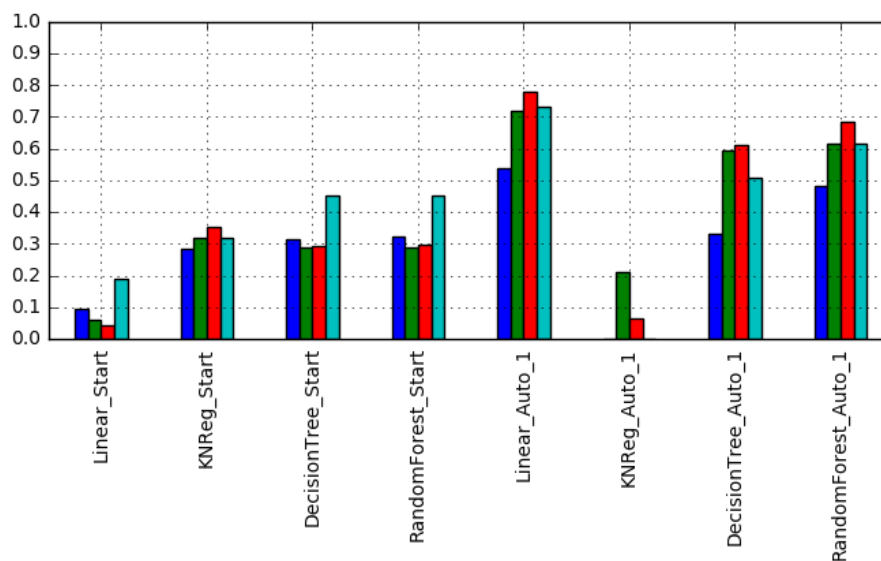


Abbildung 6 - Übersicht Scores inkl. Autoregressiver Term

Der R^2 Wert für die Regressoren „Linear“, „DecisionTree“ und „RandomForest“ steigen sehr stark an. Die Features sind für die Unterstützung der Prognose geeignet. Es gibt immer noch Unterschiede zwischen den einzelnen Splits, jedoch liegt die Vorhersage beim linearen Modell immer über 54% und steigt sogar bis auf 78%.

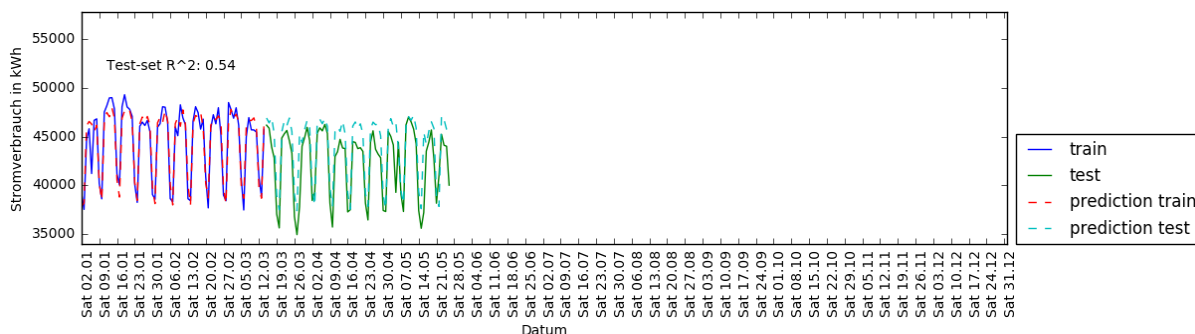


Abbildung 7 - Lineares Modell mit Autoregressivem Term - Split 1

Die Vorhersage für einen relativ kurzen Zeitraum (ca. 5 Monate) schafft bei einem 50% split für die Training- und Testdaten einen R^2 Wert von 54%. (Abbildung 7)

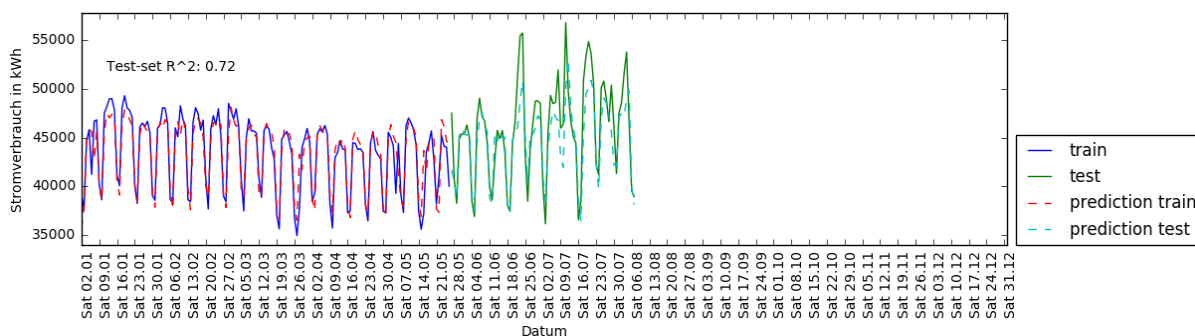


Abbildung 8 - Lineares Modell mit Autoregressivem Term - Split 2

Der Split mit 66% Training und 34% Testdaten kommt immerhin auf 72% Score. (Abbildung 8)

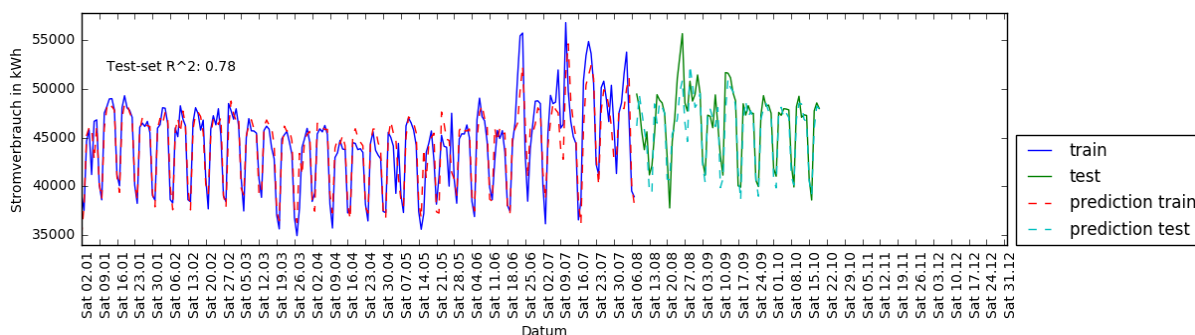


Abbildung 9 - Lineares Modell mit Autoregressivem Term - Split 3

Die Vorhersage mit Lerndaten vom Jänner bis Anfang August kann die Werte für August bis September mit einer Wahrscheinlichkeit von 78% bestimmen. (Abbildung 9)

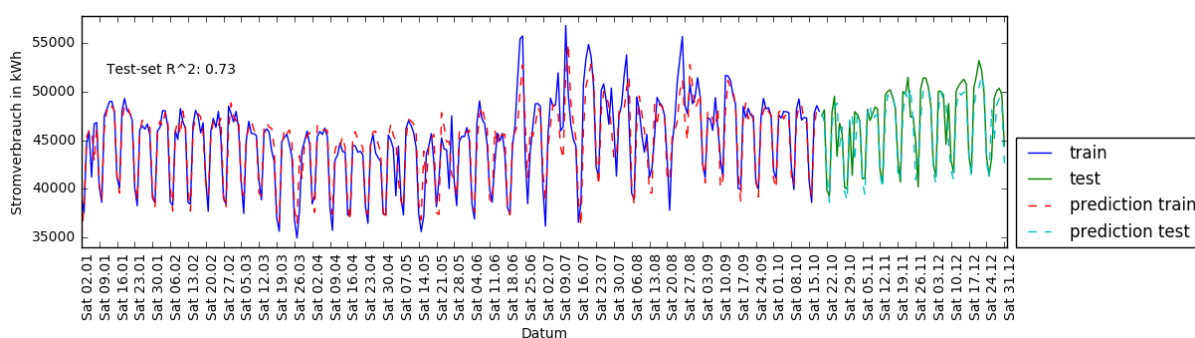


Abbildung 10 - Lineares Modell mit Autoregressivem Term - Split 5

Schlussendlich bekommt der Algorithmus 80% der Daten zum Lernen und 20% für den Test, hier schafft er noch einen Score von 73%. (Abbildung 10)

Weitere Autoregressive Terme unterstützen die Regressoren nicht, eine Verbesserung der Vorhersagequalität ist nicht zu erkennen.

4.4. Autoregressiver Term und Werktag/Samstag/Sonntag

Um weitere sinnvolle Features für die Vorhersage zu bekommen, wurde ein Feature mit den Zahlen 1 für Wochentag, 2 für Samstag und 3 für Sonntag hinzugefügt.

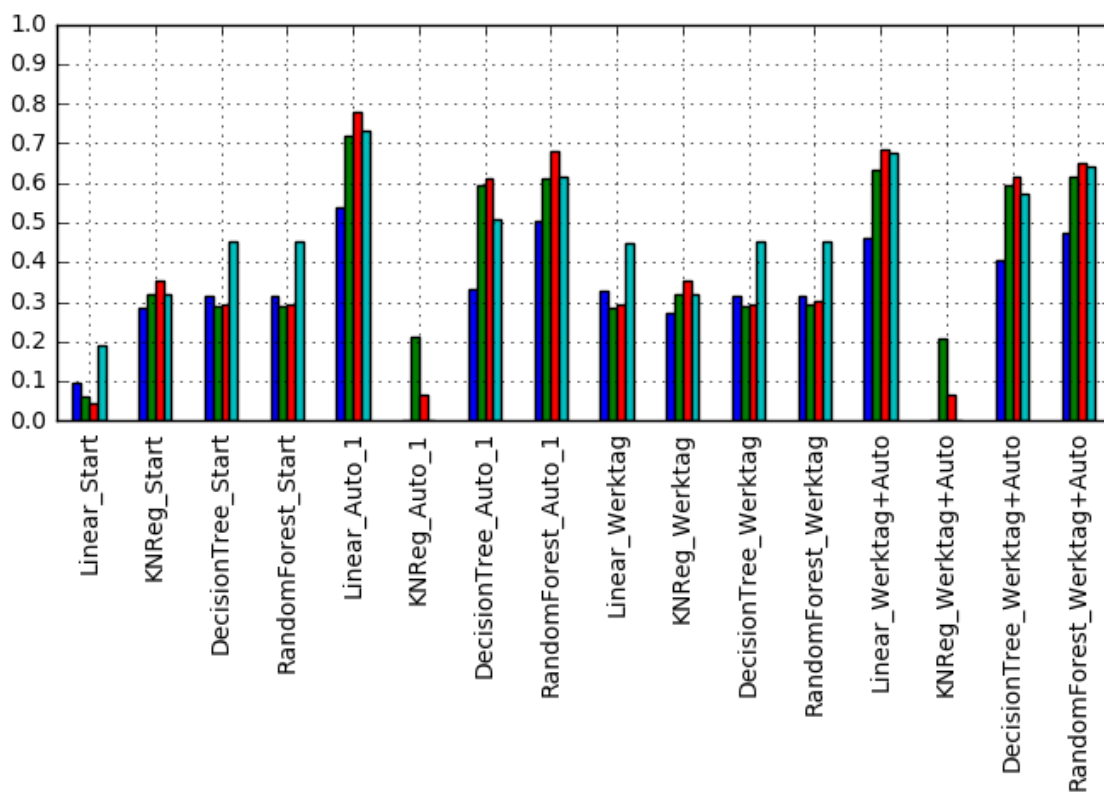


Abbildung 11 - Übersicht Scores inkl. Autoregressiver Term und Werktag

Einen Unterschied zum bereits vorhandenen Feature „Tag“ ist nicht zu erkennen. Die Kombination des Werktag-Features mit einem Autoregressiven Term verhilft den Algorithmen auch nicht zu einer großen Verbesserung.

5. Diskussion

5.1. Ziele erreicht?

Das Ziel einer Vorhersage des Stromverbrauchs für einen bestimmten Tag wurde nur bedingt erreicht. Erst der Einsatz von Autoregressiven Termen verhalf den Algorithmen zu einem Score von über 50%. Dies ist für die Praxis nicht wirklich sinnvoll. Somit wurde das Ziel einer Vorhersage nicht erreicht. Der Umgang mit dem Energiemanagement System und die Analyse der dort gespeicherten Daten war dennoch sehr lehrreich und spannend. Die Daten können eventuell für eine Masterarbeit sinnvoll sein. Aus diesem Grund hat sich das Projekt dennoch gelohnt.

5.2. Defizite

Wie schon bei der Betrachtung des Stromverbrauches sichtbar, sind außer dem Wochentag nicht wirklich Merkmale erkennbar. Der Verlauf ist zudem über das Jahr verteilt nicht wirklich gleichmäßig und hat teilweise größere Ausreißer. Um eine bessere Vorhersehbarkeit herbeizuführen, müsste man die Daten von mehreren Jahren vergleichen. Eventuell rühren die Sprünge von einem Verbraucher her, welcher immer unter gleichen Bedingungen anläuft. (z.B. Kältemaschinen) Dies konnte anhand der vorhandenen Daten jedoch nicht nachgewiesen werden.

5.3. Verbesserungsvorschläge

Der durchschnittliche Stromverbrauch von ca. 45.000 kWh bei einer Schwankungsbreite von ca. ± 10.000 kWh ist für eine Beurteilung einzelner Verursacher viel zu ungenau. Weitere Features wie z.B. der Gesamtstromverbrauch einzelner Gebäude oder Gewerke würde die Schwankungsbreite reduzieren und die Vorhersagequalität verbessern. Diese Features werden in Zukunft automatisch über das EMC erfasst. Eine weitere Auswertung sollte dann in ca. 1-2 Jahren möglich werden.

5.4. Vergleich mit Literatur

Einer ähnlichen Aufgabenstellung widmet sich die Seite <http://cs109-energy.github.io/>. Dort wird ebenfalls mittels Machine-Learning Algorithmen versucht die Energieverbräuche vorherzusagen. Der Score für die Vorhersage mittels linearer Regression erreicht dort einen R^2 -Wert von knapp über 60%. Mittels SupportVectorRegression wird 68% und mit Hilfe von Gauß'scher Regression sogar über 80%. Eine Anwendung der dort beschriebenen Algorithmen auf diese Daten wäre sicher ebenfalls noch sinnvoll.