ANALYSE DES WEINVERBRAUCHS IN ÖSTERREICH

Projektarbeit in Data Science

Verfasser

Michael Pölz

1610727026

Abgabedatum:

28. Juni 2017

Inhalt

ABBILDUNGSVERZEICHNIS

1 EINLEITUNG

Als Ausgangslage für das gewählte Thema dieses Projekts diente mein persönliches Interesse für Wein. Da wir selbst Wein produzieren und ich mich auch sonst dafür interessiere, wollte ich herausfinden, ob man bestimmte Größen bezüglich Wein mit der Hilfe von anderen Kennzahlen aus der Wirtschaft oder Bevölkerungsstatistik bestimmen kann.

Das Ziel der Analyse war somit ursprünglich die Bestimmung des Weinverbrauchs in Österreich anhand von Kennzahlen der Bevölkerungsstatistik und von volkswirtschaftlichen Größen. Die Frage war, ob der Weinverbrauch von den ausgewählten Faktoren abhängig ist und ob man ihn in einem gewissen Jahr bestimmen kann, wenn man die anderen Kennzahlen kennt. Die Bestimmung wäre aus Gründen der Einfachheit erstmals für das Jahr gewesen in dem die anderen Kennzahlen bereits bekannt sind und bei guten Resultaten eventuell auch für das zukünftige Jahr, da dies interessanter wäre.

Leider hat sich im Laufe der Datenanalyse herausgestellt, dass es mit den verwendeten Kennzahlen so gut wie unmöglich ist den Weinverbrauch halbwegs korrekt zu bestimmen. Aus diesem Grund hat sich das Ziel der Analyse dahingehend verändert, dass anstatt dem Weinverbrauch die Weineinfuhr nach Österreich bestimmt wurde. Auf die genaueren Gründe warum gerade die Weineinfuhr bestimmt wird, wird noch in Kapitel 4 eingegangen.

2 DATEN

Die für die Analyse verwendeten Daten stammen von der Internetseite der Statistik Austria (http://www.statistik.at/web_de/statistiken/index.html), der Statistikseite der OECD (http://stats.oecd.org/Index.aspx#) und der Homepage der Gemeinde Wien (https://www.wien.gv.at/). Da die verwendeten Quellen vertrauenswürdig sind, sollte die Qualität der Daten prinzipiell gut sein.

Ein großes Problem ist allerdings, dass eine große Anzahl an Features verwendet wird, denen nur wenige Samples gegenüber stehen. Insgesamt werden in dieser Analyse 46 Features verwendet, auf welche gleich noch genauer eingegangen wird, und nur 35 Samples, da nur für die Jahre von 1981 – 2015 alle Daten vorhanden waren und hier nur Jahresdaten. Besser wäre es daher gewesen, die Features eventuell bereits vor der Analyse zu reduzieren und mehr Samples zu bekommen.

Dafür wären monatliche Daten von Vorteil gewesen, da dies die Samples verzwölffacht hätte. Allerdings sind nicht für alle verwendeten Features Monatsdaten verfügbar.

Als Features wurden zuerst Daten aus dem Weingeschäft ausgewählt, wie zum Beispiel die Weinproduktion in Österreich, die Weininlandsverwendung, die industrielle Verwendung oder der Weinimport und –export. Zusätzlich wurde noch der Vorjahresverbrauch beziehungsweise später die Vorjahreseinfuhr hinzugefügt. Diese Features wurden ausgewählt, da alle zusammenhängen könnten. Weitere Features sind der Bierkonsum, die Bierproduktion und der Bierimport und –export in Österreich. Diese wurden verwendet, da die Möglichkeit bestehen könnte, dass ein Anstieg im Konsum des einen Genussmittels einen Rückgang des anderen bedeuten könnte. Als Faktoren der Bevölkerungsstatistik wurden die gesamte Bevölkerung sowie die nach Altersklassen, Geschlecht und Herkunft aufgeteilte Bevölkerung miteinbezogen. Zusätzlich wurden noch die Anzahl der Studienabschlüsse und ebenfalls eine Aufteilung nach Art des Studiums hinzugefügt. Diese Kennzahlen wurden verwendet um zu sehen, ob eine Veränderung in der Bevölkerungsstruktur eine Auswirkung auf das Konsumverhalten hat. Das BIP, die Arbeitslosenrate und die privaten Konsumausgaben wurden als volkswirtschaftliche Faktoren in die Features mitaufgenommen. Der Grund dafür war, dass diese Kennzahlen oft für Vergleiche oder Analysen verwendet werden. Die letzten ausgewählten Features waren verschiedene Wetterdaten aus Wien, wie zum Beispiel die Jahresmitteltemperatur, Sommertage oder Frosttage. Der Grund dafür war, dass die Möglichkeit bestehen könnte, dass zum Beispiel in außergewöhnlich heißen Jahren mehr beziehungsweise weniger konsumiert wird. Die verwendeten Wetterdaten stammen nur aus Wien, weil diese frei verfügbar waren. Allerdings sollten diese aussagekräftig genug für einen österreichweiten Trend sein.

3 METHODEN

Die im folgenden Kapitel beschriebenen Methoden beziehen sich auf die finale Bearbeitung mit dem Ziel die Weineinfuhr zu bestimmen. Allerdings waren die vorher angewandten Methoden mit dem Ziel der Bestimmung des Weinverbrauchs fast ident. Die Vorgehensweise für die Analyse war folgendermaßen:

1. Einlesen der Daten und Aufbereitung:

   Der erste Schritt war, die Daten der verschiedenen Quellen in ein Excel –

Dokument zu bringen, damit diese ins Programm eingelesen werden können. Nach dem erfolgreichen Einlesen wurden die Daten unterteilt in ein Target (Weineinfuhr) und die Features. Beide Datensätze wurden nochmals in einen Trainings- und einen Test-Datensatz unterteilt. Da die Anzahl der Samples sehr beschränkt war, wurden 31 Samples für den Trainings und nur 4 Samples für den Test Satz verwendet, um die Fits besser trainieren zu können. Anschließend wurde zur besseren Veranschaulichung das Target geplottet, um mögliche Trends oder Schwankungen zu erkennen. Dann folgten noch die Bestimmung der verschieden Korrelationen, um zu sehen wie gut die Features passen und ein Plot der neun besten beziehungsweise für mich interessantesten Features.

2. Anwendung verschiedener Regressionsmodelle:

   Um einen ersten Überblick zu bekommen wie gut die Daten bereits ohne große Veränderungen und Anpassungen für eine Analyse verwendet werden können, wurden die verschiedenen Regressionsmodelle angewandt. Hier wurden meist rein zufällige Parameter angenommen, wobei insbesondere bei der Lasso – Regression bereits versucht wurde, mit Hilfe der einstellbaren Parameter einige Features zu eliminieren. Die Ergebnisse wurden dann auch zum Teil graphisch dargestellt.

3. Reduktion der Features:

   Da eigentlich bereits vor dem Anwenden der verschiedenen Regressionen bekannt war, dass die Anzahl der Features stark reduziert werden muss, folgte nun einer der wesentlichsten Teile. Um kein Over Fitting zu generieren wurde versucht die Anzahl der Features auf zirka vier zu reduzieren. Dafür wurden drei unterschiedliche Varianten der Feature Reduktion eingesetzt.

   3.1. Model-based Feature Selection:

   Zuerst wurde diese Variante versucht, bei der alle Features berücksichtigt werden. Als Modelle wurden einmal die Random-Forrest-Regression verwendet und einmal die Linear-Regression. Diese zwei Regressionsmodelle wurden anschließend natürlich auch für die Auswertung der Analyse verwendet.

   3.2. Iterative feature selection (Recursive feature elimination):

   Bei dieser Variante wird immer das schlechteste Feature eliminiert und anschießend ein neues Modell gebildet, bis nur mehr die Anzahl der gewünschten Features übrig ist. Als Modelle wurden wie bereits zuvor die Random-Forrest-Regression und die Linear-Regression verwendet.

3.3. Manuelle Feature Auswahl:

Hier habe ich die vier verwendeten Features selbst ausgewählt. Als Basis dienten mir die unterschiedlichen Korrelationswerte, die ich am Anfang berechnen ließ. Von diesen habe ich die vier Features mit den besten korrelierenden Werten zum ausgewählten Target bestimmt und als Features zur weiteren Analyse festgelegt.

4. Bestimmung der besten Parameter und Cross – Validation:

   Nach der Auswahl der Features folgt nun die Bestimmung der besten Parameter für die Regressionsmodelle und die Cross – Validation. Dies wurde für alle drei Varianten der Feature Selektion gleich angewandt. Als Regressionsmodelle dienen jetzt nur noch die Random-Forrest-Regression, da diese bereits zuvor die besten Ergebnisse lieferte und die Linear-Regression, da der Plot des Targets den Anschein vermittelte, dass diese ganz gut passen könnte.

   Um die beste Anzahl der verwendeten Bäume für die Random-Forrest-Regression zu finden und gleichzeitig die Cross-Validation durchzuführen, wurde der Befehl GridSearchCV benutzt. Hier wurde die Anzahl der Bäume zwischen 1 und 100 variiert und der beste Wert ausgewählt.

   Die Linear-Regression wurde ganz standardmäßig durchgeführt und die Cross-Validation erfolgte danach mit einem eigenen Befehl.

   Die Ergebnisse wurden zur Veranschaulichung wieder geplottet.

5. Anwendung eines neuen Splits:

   Da die vorherigen Ergebnisse nicht ganz zufriedenstellend waren, wurde versucht die Daten nochmals zu analysieren, aber einen anderen Split durchzuführen. Es wurden wieder 31 Samples für die Trainingsdaten und 4 Samples für die Testdaten ausgewählt. Allerdings erfolgte der Split diesmal zufällig und nicht gezielt nach dem Datum. Die anschließende Feature Selektion und die angewandten Regressionsmodelle blieben gleich wie vorhin.

4 RESULTATE

Hier werden nur die Resultate der Simulation mit der Weineinfuhr als Target besprochen. Warum diese dem Weinverbrauch schlussendlich vorgezogen wurde folgt im abschließenden Kapitel.

In Abbildung 1 ist der Verlauf der Weineinfuhr über den betrachteten Zeitraum zu sehen.

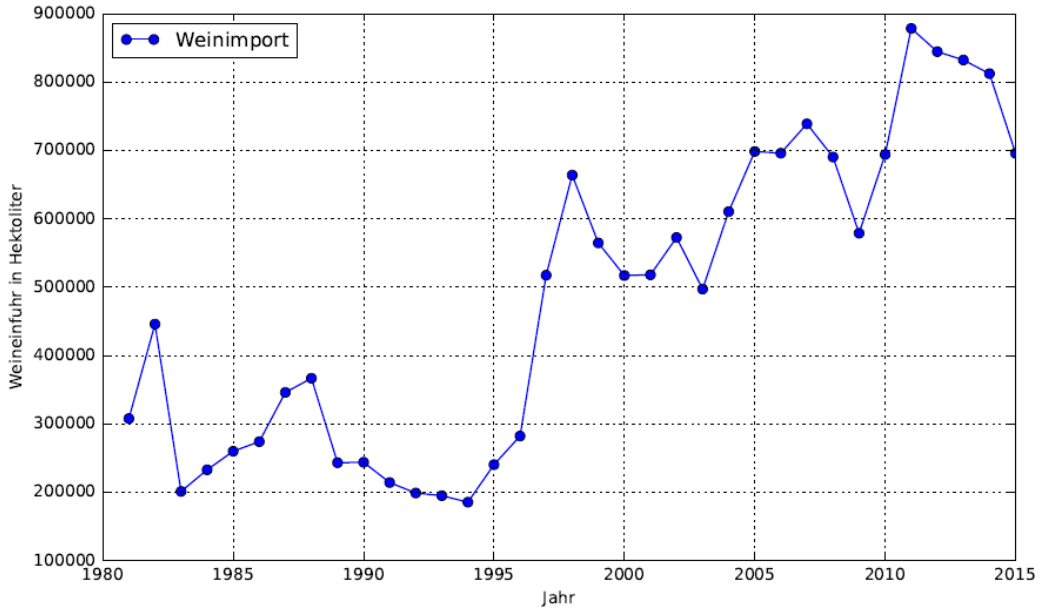

Abbildung 1: Weineinfuhr nach Österreich von 1981 bis 2015

Hier ist gut zu erkennen, dass der generelle Trend ein steigender Import ist, wobei es immer wieder zu Schwankungen kommt. Anhand dieses Trends kann man schließen, dass eine lineare Regression ein gutes Modell sein könnte. Allerdings ist gerade zum Schluss ein relativ starker Rückgang zu erkennen.

In der Korrelationsmatrix, die im Python – File zu finden ist, kann man gut erkennen, dass zwar viele Features nicht wirklich mit der Weineinfuhr korrelieren, aber es durchaus einige Features gibt, die eine hohe Korrelation mit ihr haben. So gibt es zwei Features mit einer Korrelation von über 0,9 und 14 Features mit einer Korrelation von über 0,8. Ein Scatterplot von neun ausgewählten Features mit dem Target ist in Abbildung 2 zu sehen. Hier wurden die Features mit der besten Korrelation ausgewählt und zusätzlich noch welche die für interessant gehalten wurden. Bei diesem ist die Korrelation gut zu erkennen und für mich war es auch gut zu sehen, dass die Features, von denen ich vorher dachte, dass sie eine Beeinflussung ausüben, für die weitere Simulation leider nicht hilfreich sein werden.

Bei den nun durchgeführten ersten Regressionen ist gut zu erkennen, dass das bereits zuvor prognostizierte Problem eintritt, dass zu viele Features für zu wenig Samples vorhanden sind. Dies führt zum klassischen Over Fitting und in diesem Fall auch zu sehr schlechten Testergebnissen. Die Ergebnisse der verschiedenen Fits auf den Trainingsdaten ergeben meistens eins oder fast eins und die Anwendung der Fits auf die Testdaten ergibt meist leicht negative bis stark negative Werte.

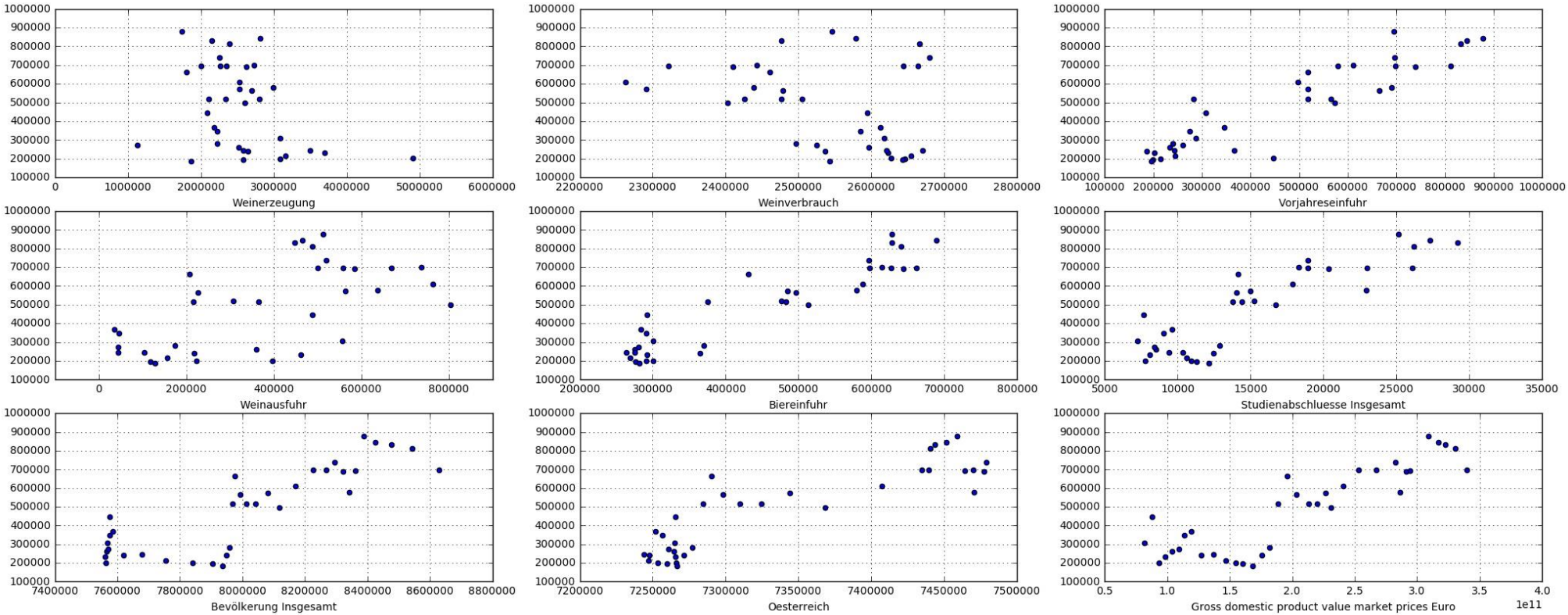

Abbildung 2: Scatter Plot der besten Features

Auch bei der Lasso-Regression, die durch einen hohen Alpha-Wert bereits selbstständig einige Features eliminiert, trifft dies zu. Die Ergebnisse der Lasso-Regression sind in Abbildung 3 und Abbildung 4 dargestellt. Hier ist gut zu erkennen, dass der vorhergesagte Wert bei den Trainingsdaten exakt mit dem tatsächlichen Wert übereinstimmt, also Over Fitting besteht.

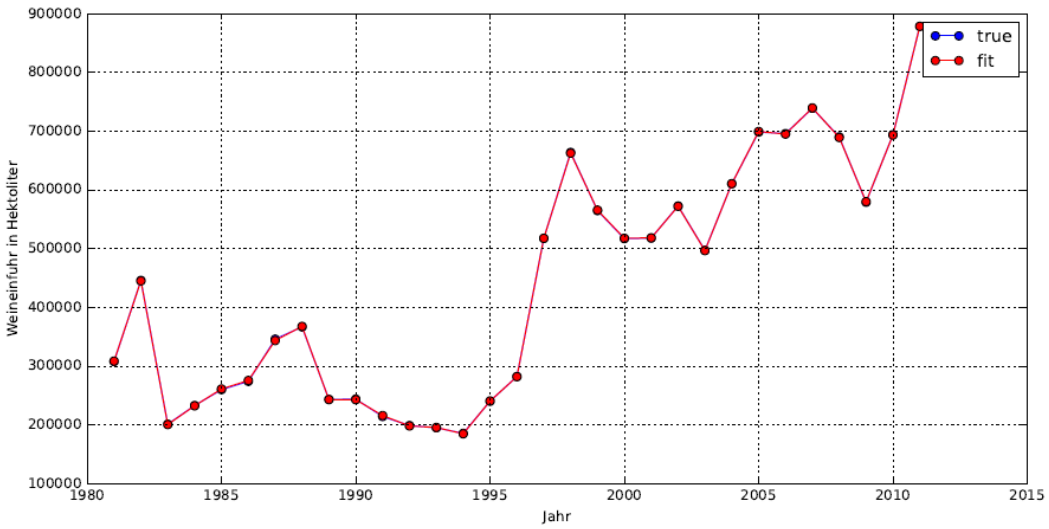

Abbildung 3: Ergebnis Lasso-Regression auf Trainingsdaten

Hier sieht man, dass der vorhergesagt Wert der Testdaten weit von den tatsächlichen Daten abweicht, was auch am negativen Test score von -10,22 zu erkennen war.

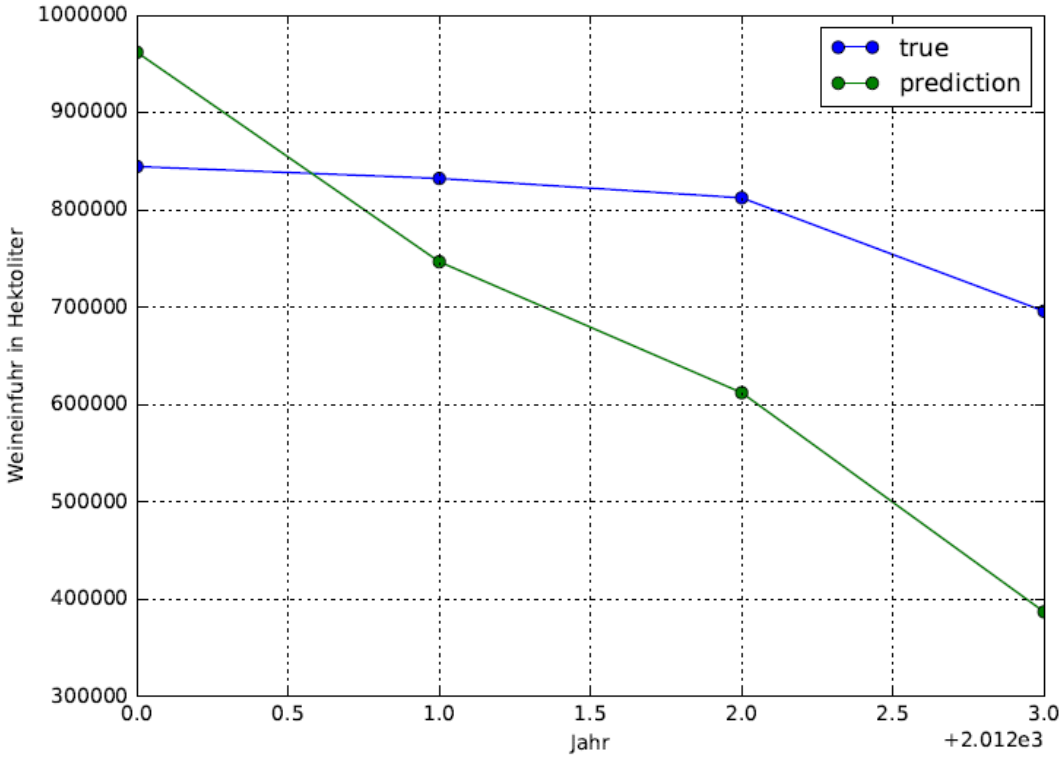

Abbildung 4: Ergebnis Lasso-Regression auf Testdaten

Die besten Ergebnisse liefert in diesem ersten Versuch der Regressionsmodelle der Random-Forrest-Regressor, weshalb dieser auch für die späteren Regressionen nach der Feature Selektion ausgewählt wurde. Allerdings war auch hier der Cross-Validation Score leicht im negativen Bereich.

Nach der Feature Selektion, die entweder genau vier Features oder meist um die vier Features liefert, wurden nun wieder die Random-Forrest-Regression und die Linear-Regression verwendet. Die Trainingsergebnisse sind immer noch sehr hoch bei allen drei Varianten der Feature Selektion mit ca. 0,98. Der Test Score variiert teils recht stark, wobei er generell höher ist als zuvor. Im für den Bericht ausgewerteten Fall beträgt er zwischen 0,00 und 0,34, was leider nicht hoch ist. Positiv ist, dass nun auch der Cross – Validation Score mit ca. 0,24 leicht im positiven Bereich ist. Eine graphische Darstellung der Ergebnisse ist in Abbildung 5 und Abbildung 6 zu sehen.

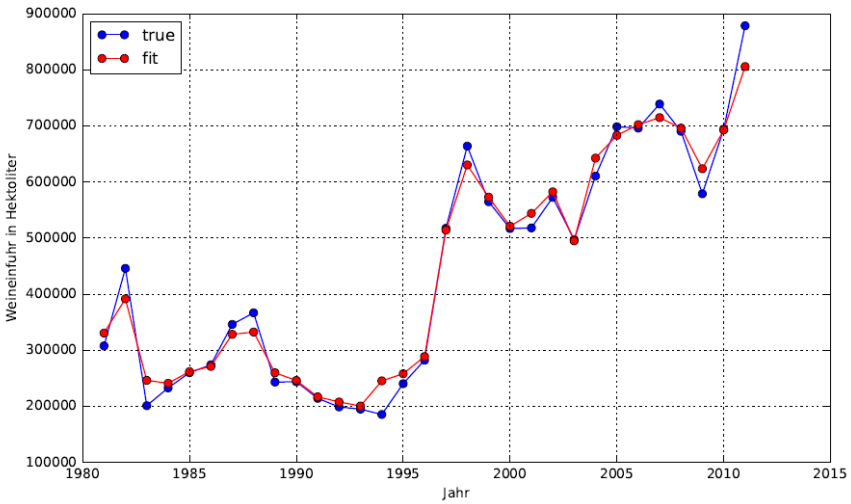

Abbildung 5: Ergebnis Random-Forrest-Regression auf Trainingsdaten

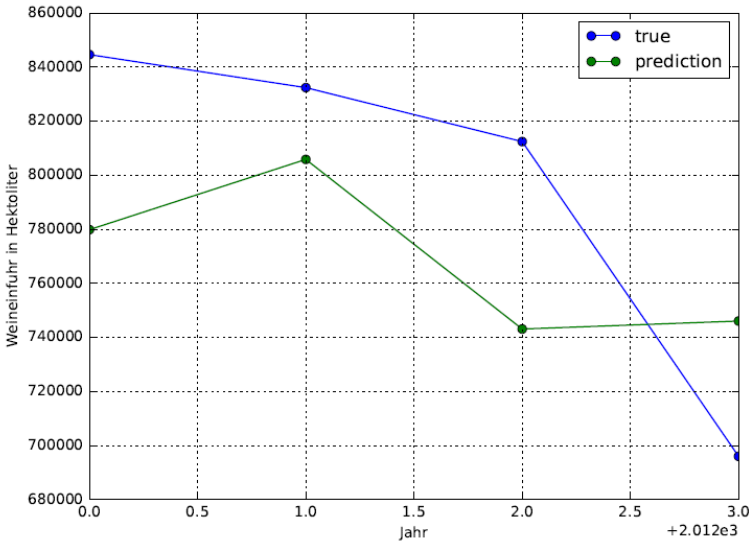

Abbildung 6: Ergebnis Random-Forrest-Regression auf Testdaten

Bei Verwendung der Linear Regression nach der Feature Selektion sind die Ergebnisse zwar besser als vorher, allerdings immer noch sehr schlecht. So variieren die Testergebnisse, je nach Art der Selektion, zwischen -0,11 und -11,12. Das ist zwar eine deutliche Steigerung im Gegensatz zum Score von -43,34 vor der Reduzierung der Features, aber immer noch nicht brauchbar.

Auf Grund der schlechten Ergebnisse wird jetzt noch ein Zufallssplit durchgeführt und die Features werden wie vorhin reduziert. Bei Verwendung der Random-Forrest-Regression ist der Score auf die Trainingsdaten wieder ähnlich wie vorher. Die Testergebnisse sind aber deutlich besser mit Werten bis zu 0,75. Auch der Wert der Cross-Validation ist mit über 0,8 deutlich besser als zuvor.

Die Linear Regression bringt nun bei den automatischen Feature Selektionen bessere Werte als zuvor, die mit einem Test score von 0,27 leicht im positiven Bereich liegen. Der Cross – Validation Wert ist aber immer noch negativ. Bei der manuellen Feature Auswahl hingegen liefert auch die Linear Regression gute Ergebnisse. So liegen die Testergebnisse meistens über dem Wert von 0,7 und auch der Cross-Validation Wert ist größer als 0,8.

5 DISKUSSION

Die Ziele der Analyse wurden in zweifacher Hinsicht leider nicht erreicht. Erstens wurde es nicht geschafft den Weinverbrauch in Österreich zu bestimmen, was das ursprüngliche Ziel war. Dieser musste im Laufe der Bearbeitung durch die Weineinfuhr ersetzt werden. Der Grund dafür war, dass die erzielten Ergebnisse nicht brauchbar waren. Ein Grund dafür konnte durch die Korrelationen gefunden werden, wo man sah, dass die verwendeten Features nur minimal mit dem Weinverbrauch korrelierten und es nur wenige Ausnahmen mit einer etwas besseren Korrelation gab. Wie in Abbildung 7 zu sehen ist, kann man dies eventuell darauf zurückführen, dass der Weinverbrauch über die Jahre betrachtet keine wirkliche Tendenz zeigt, weder fallend noch steigend. Man kann gut erkennen, dass er sehr stark variiert und es von einem Jahr auf das nächste teils große Sprünge gibt, auch wenn man sich von der verwendeten Skalierung nicht täuschen lassen darf.

Da man bei der Weineinfuhr doch eine bessere Tendenz erkennen konnte und diese auch stärker mit den Features korrelierte wurde schließlich diese für die weitere Analyse verwendet. Allerdings konnte leider auch hier das gesteckte Ziel nicht wirklich erreicht werden. Die Weineinfuhr kann nicht wirklich gut für ein Jahr

bestimmt werden anhand der Verwendung der vergangenen Werte. Die Ergebnisse bei Verwendung des ersten Splits waren leider schlecht.

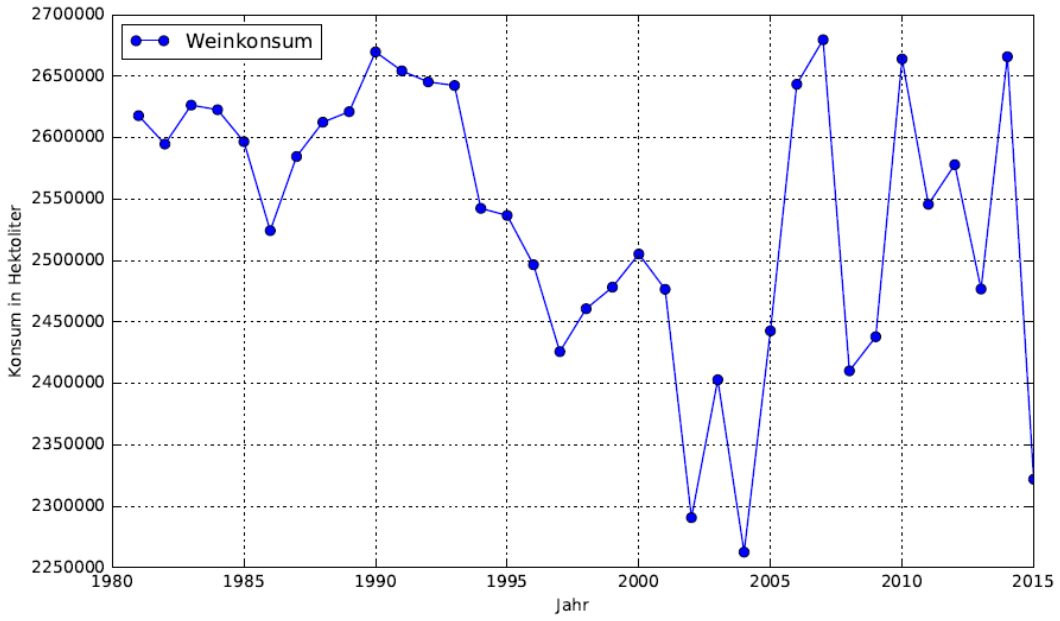

Abbildung 7: Weinverbrauch in Österreich von 1981 bis 2015

Erst bei Verwendung eines zufälligen Splits und nicht mehr eines Splits nach Jahren wurden die Ergebnisse brauchbar. Das war zwar nach den vorherigen Ergebnissen sehr erfreulich, allerdings nicht sehr hilfreich für das ursprüngliche Ziel und vor allem nicht wenn man darüber hinausgehen wollte und die Weineinfuhr für ein zukünftiges Jahr bestimmen. Nun kann man bei Daten über alle Jahre und eine zufällige Aufteilung die Testwerte dazwischen ganz gut bestimmen.

Eine Verbesserungsmöglichkeit wäre es bestimmt, wenn man mehr Samples zur Verfügung hätte. Dies würde die gesamte Simulation erleichtern und verbessern. Vielleicht könnte man auch noch bessere Features finden, die für das gewünschte Ziel aussagekräftiger sind. Eine weitere Möglichkeit zur Verbesserung könnte auch noch in der Auswahl der Regressionsmodelle oder der Modelle zur Feature Selektion beziehungsweise generell in der Datenaufbereitung vorhanden sind. Die verwendeten Methoden wurden nach bestem Wissen ausgewählt, aber es ist durchaus möglich, dass es bessere Modelle gibt.