Duale Hochschule Baden-Württemberg Mannheim

Data Exploration Projekt:

‘Insight Explorers‘

House Price Prediction

Studiengang Wirtschaftsinformatik

Data Science

Gruppenmitglieder:	Felix Hüsgen	3583266
	Lennart Fertig	8602307
	Pascal Schmidt	8133405

Studiengang: WWI19DSB

Studiengangsleiter: Prof. Dr. Bernhard Drabant

Dozenten: Sebastian Schön, Simon Poll

Kurs: Data Exploration

Bearbeitungszeitraum: 09.06. – 12.07.2021

Inhaltsverzeichnis

1. Thema und Motivation

Seit einigen Jahren steigen die Immobilienpreise in den USA wie auch in Europa rasant an. Vor allem große Städte sind von den Preissteigerungen betroffen. Generell sind Hauspreise von hohem Interesse für verschiedenste Shareholder und somit gerade die Voraussage dieser Preise von hoher Bedeutung.

Ziel dieses Data Exploration Projektes ist es, Hauspreise auf Basis verschiedener Eigenschaften durch den Einsatz von Machine Learning Modellen vorauszusagen. Da keine Preisklasse, sondern ein genauer Preis prognostiziert werden soll, wird im Projekt auf Regressionsmodelle zurückgegriffen. Durch solche Vorhersagen können Shareholder eine bessere Informationslage erlangen und somit bessere Kaufentscheidungen treffen. Aspekte wie der optimale Kaufzeitpunkt oder auch Renovierungsvorhaben können so besser geplant werden.

Bei der Vorhersage sollen typische Aspekte der Bewertung einer Immobilie wie Einrichtung und Lage berücksichtigt werden. Beispielsweise kann die Größe der Wohnfläche oder auch die Anzahl der Badezimmer einen großen Einfluss auf den Hauspreis haben.

Folglich sollen in diesem Bericht zunächst die Grundlagen zum Projekt erläutert und anschließend die Ergebnisse präsentiert werden. Der Bericht wird mit einer kritischen Reflexion abgeschlossen.

2. Grundlagen

2.1. Auswahl des Datensatzes

Bei der Auswahl eines geeigneten Datensatzes für das beschriebene Ziel hat sich die Gruppe für einen Datensatz von der Online-Community Kaggle entschieden. Der Datensatz beinhaltet Hausdaten von 21600 Häusern inklusive Verkaufspreisen in King County bzw. Seattle in den USA. Es sind 20 verschiedene Attribute wie Wohnfläche, Baujahr oder Anzahl der Badezimmer für jedes Haus gegeben.

Trotz der eher kleinen Anzahl an Daten ist der Datensatz aufgrund seiner Vollständigkeit und der vielen Attribute sehr gut für das Projekt geeignet. Dadurch, dass die Häuser einer einzelnen großen Stadt betrachtet werden, erhofft sich die Gruppe bessere Ergebnisse als beispielsweise bei der Betrachtung von Häusern eines ganzen Landes oder Kontinents. In einer einzelnen Stadt sind bestimmte Korrelationen möglicherweise stärker vorhanden.

2.2. Herangehensweise und verwendete Technologien

Die Gruppe hat sich dazu entschieden, den Python-Code in Google Colaboratory zu entwickeln, da hier gemeinsam an Notebooks gearbeitet werden kann und einige benötigte Module wie scikit-learn schon vorinstalliert sind. Somit kann der Code darüber hinaus schnell und unkompliziert mit anderen Personen geteilt werden.

Zunächst wird der Datensatz als csv-Datei über Google Drive in die Colaboratory-Umgebung geladen und in ein pandas-Dataframe verpackt. Das anschließende Vorgehen beginnt mit einer Datenanalyse, bei der die Daten vorrangig auf Zusammenhänge untersucht werden. Danach werden die Daten für den Machine Learning Einsatz vorbereitet. Es sollen verschiedene Modelle zur Regression trainiert werden, die dann anschließend evaluiert und gegebenenfalls optimiert werden.

2.3. Related Work

Während der Bearbeitung des Projektes wird unteranderem Bezug auf das Buch „Praxiseinstieg Machine Learning mit Scikit-Learn und TensorFlow: Konzepte, Tools und Techniken für intelligente Systeme“ von Aurélien Géron genommen[1]. Dies eignet sich besonders gut für die Bearbeitung des Data Exploration Projekts, da es sich direkt an angehende Data Scientists richtet und den Einsatz von einfachen und effizienten Werkzeugen zum Implementieren eines Machine-Learning-Projekt vermittelt.

Besonders wird sich dabei auf Kapitel 2 „Ein Machine-Learning-Projekt von A bis Z“ bezogen, welches der Gruppe anhand eines Beispiels die wichtigsten Schritte in einem Data Science Projekt vermittelte. Dabei beschreibt Aurélien Géron, angefangen von der Beschaffung der Daten über ihre Erkundung und Visualisierung, sowie der Vorbereitung der Daten für den Machine-Learning Algorithmus, die einzelnen Etappen für den erfolgreichen Ablauf eines Machine-Learning-Projektes.

[1] Buch : Aurélien Géron. (2020). Praxiseinstieg Machine Learning mit Scikit-Learn, Keras und TensorFlow, 2nd Edition. Dpunkt Verlag.

3. Ergebnisse

3.1. Datenanalyse durch Visualisierung und Vorverarbeitung

Nachdem der Datensatz erfolgreich über Google Drive in die Colaboratory-Umgebung geladen und in ein pandas-Dataframe verpackt wurde, wird zunächst die Datenstruktur analysiert. Dazu nutzt die Gruppe, neben der Methode *info()*, die eine schnelle Beschreibung des Datensatzes liefert, die Methode *describe()*, um die numerischen Merkmale zusammenfassen zu können. Des Weiteren wird durch das Erstellen von Histogrammen, tiefere Erkenntnisse über die Mengenverteilung jedes nummerischen Attributes, bezogen auf die einzelnen Immobilien, gewonnen (siehe Abbildung 1). Im nächsten Schritt wird mithilfe der Methode *corr()* der Korrelationskoeffizient berechnet, um die Korrelation jedes Merkmals mit dem Preis der Immobilien zu erhalten. Dieser wird neben einer Heatmap (siehe Abbildung 2), auch mithilfe weiterer Scatter Plots visualisiert. Durch das zur Verfügung stehen von geografischen Informationen (Breite und Länge), wird weiter die geografische Lage einzelner Immobilien dargestellt. Jede Immobilie wird durch einen Datenpunkt beschrieben, wobei die Farbe eines jeden Datenpunktes den Preis der Immobilie verdeutlicht. Dafür wird die vordefinierte Farbskala (seismic) verwendet, wobei niedrigere Hauspreise mithilfe der Farbe Blau, höhere Hauspreise durch die Farbe Rot, visualisiert werden (siehe Abbildung 3).

Nachdem die Gruppe den Datensatz durch Analysen und Visualisierungen erkundet hat, stehen weite Erkenntnisse zur Verfügung, um einen Trainings- und Testdatensatz zu generieren. Dabei werden zunächst alle Immobilien, die einen Preis von über 2 Millionen Dollar haben aus dem Datensatz entfernt, um einen zu großen Einfluss von Ausreißern auf die Modell-Performance zu verhindern. Um einen Testdatensatz zu erstellen, wählt die Gruppe 20 Prozent des reduzierten Datensatzes aus, um einen „Train-Test“ Split durchzuführen.

Im nächsten Schritt werden die Daten für den Machine-Learning-Algorithmus vorbereitet. Dabei sind durch den gegebenen Datensatz alle Datenpunkte je Merkmal gegeben, wodurch ein setztes des Datenpunktes auf einen bestimmten Wert oder ein gar löschen des Merkmals aus dem Datensatz nicht notwendig ist. Da zusätzlich in dem Datensatz keine kategorischen Merkmale gegeben sind, ist das Konvertieren von Einträgen zu nummerischen Einträgen nicht erforderlich. Dennoch wird der Datensatz durch das herausnehmen von den Merkmalen ‚date‘, ‚zipcode‘, ‚id‘ und ‚price‘ (da

‚price' das Label ist) weiter angepasst, bevor dieser skaliert wird. Eine Skalierung des Datensatzes ist erforderlich, damit der Machine-Learning-Algorithmus besser mit diesem umgehen und aus diesem lernen kann. Dazu wird eine Pipeline initialisiert, um die Daten mithilfe des ‚StandardScalers' aus der Scikit-Learn Bibliothek skalieren zu können. Dieser organisiert die Abfolge der nötigen Transformationsschritte, um einen skalierten Datensatz zu erhalten.

Am Ende dieses Schrittes steht der Gruppe ein vorbereiteter Trainingsdatensatz ohne ‚Labeling' zur Verfügung.

3.2. Umsetzung mit Machine Learning und dessen Ergebnisse

Nach der explorativen Analyse und der erfolgten Vorverarbeitung der Daten können nun Machine Learning Modelle für die Hauspreis-Prognose trainiert werden. Hierfür wird auf Module der Scikit-Learn Bibliothek zurückgegriffen.

Bevor die Modelle trainiert werden, müssen zunächst die relevanten Gütemaße für die Evaluation geklärt werden. Für die Regression eignet sich hierbei der Root Mean Square Error (kurz: RMSE) sehr gut. Dieser gibt die mittlere Abweichung der Prognose von den realen Daten an. Als zweites ergänzendes Gütemaß kann der R2-Wert fungieren. Der R2-Wert als Bestimmtheitsmaß gibt an, wie angepasst die Regressionsgerade an die realen Daten ist. Bei einem Wert von 1 herrscht eine perfekte Modellanpassung bei einem linearen Zusammenhang.[2]

Als erstes Regressionsmodell für die Hauspreis-Vorhersage wird eine lineare Regression trainiert. Der RMSE ist hierbei allerdings auf den Trainingsdaten wie auch nach Kreuzvalidierung mit ca. 155000 recht hoch. Auch der R2-Wert von 0,7 ist nicht zufriedenstellend. Hieraus lässt sich schließen, dass die lineare Regression die Daten underfittet. Dies liegt vor allem daran, dass viele Attribute nur eine sehr geringe Korrelation zum Hauspreis haben. Es wird also für unseren Use Case ein deutlich komplexeres Modell benötigt.

Als nächstes Modell wird ein Decision Tree trainiert, hierbei wird jedoch schnell klar, dass der Decison Tree viel zu stark overfittet und dass die Performance auf den

[2] Vgl. INWT Statistics. (2014). Bestimmtheitsmaß R^2 - Teil 2: Was ist das eigentlich, ein R^2? Abgerufen am 05.07.2021 von https://www.inwt-statistics.de/blog-artikel-lesen/Bestimmtheitsmass_R2-Teil2.html

Testdaten deutlich zu gering ist. Daher wird dieses Modell wieder aus dem Notebook herausgelöscht. Um jedoch trotzdem weiterhin mit Entscheidungsbäumen zu experimentieren, wird als passenderes Modell dann ein Random Forest Regressor trainiert. Der Random Forest ist ein Ensemble Modell, das mithilfe von Bagging die Ergebnisse vieler einzelner Entscheidungsbäume kombiniert.[3] Auf unseren Trainingsdaten hat der Random Forest einen RMSE von ca. 39000, während er auf den Testdaten einen RMSE von ca. 104000 hat. Der Random Forest passt also deutlich besser zu den Daten, unterliegt aber trotzdem auch noch einem Overfitting.

Nicht zuletzt wird als weiteres Ensemble-Learning Modell noch ein Gradient Boosting Regressor trainiert. Dieser wird gewählt, da Ensemble-Methoden scheinbar gut passen, wir aber noch einen Boosting statt Bagging Algorithmus ausprobieren wollen. Die Performance ist mit einem RMSE von ca. 102000 auf den Trainingsdaten und 109000 auf den Testdaten zwar etwas schlechter als beim Random Forest, allerdings liegt beim Gradient Boosting auch weniger Overfitting vor. Die R2-Werte liegen mit 0.86 beim Random Forest und 0.85 beim Gradient Boosting fast gleich hoch.

Insgesamt hat sich die Gruppe dazu entschieden, den Random Forest Regressor trotz dessen Overfitting als passendstes Modell zu wählen. Der Random Forest Regressor schneidet mit dem besten RMSE und R2-Wert ab.

3.3. Kritische Bewertung

In diesem Data Exploration Project konnten die Daten erfolgreich vorverarbeitet und auf verschiedenen Regressionsmodellen trainiert werden. Die Preise von Häusern in Seattle können nun mithilfe der Ergebnisse ungefähr prognostiziert werden. Der Random Forest Regressor ermöglicht diese Prognose am besten. Es kann jedoch durchaus kritisch angemerkt werden, dass die trainierten Modelle keine perfekten Ergebnisse darstellen. Die mittleren quadratischen Abweichungsfehler sind noch zu hoch, um Hauspreise verlässlich genau vorhersagen zu können. Als Grund hierfür kann angeführt werden, dass das Datenset für Machine Learning Zwecke eher klein ist und dass einige Features im Datenset kaum einen Zusammenhang mit dem Hauspreis haben. Außerdem ist kritisch zu bemerken, dass der Kaufzeitpunkt in den

[3] Vgl. Big Data Insider. (2020). Was ist Random Forest? Abgerufen am 05.07.2021 von https://www.bigdata-insider.de/was-ist-random-forest-a-913937/

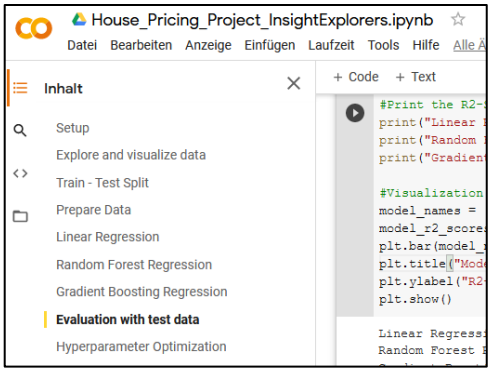
trainierten Modellen nicht berücksichtigt wird. Gerade in den letzten Jahren sind die Immobilienpreise stark gestiegen, weshalb auch der zeitliche Aspekt einen wichtigen Einfluss auf den Preis haben sollte.

Insgesamt konnte die erlernte Theorie zu Machine Learning aus vergangenen Vorlesungen durch das Data Exploration Projekt erfolgreich in die Praxis umgesetzt werden. Die Gruppe hat wertvolle Erfahrungen gesammelt und sich im Bereich Data Science deutlich weiterentwickelt.

Anhang

Anmerkungen zum Quellcode:

- Der Code kann unter folgendem Link in Google Colab eingesehen und ausgeführt werden:
  https://colab.research.google.com/drive/13zeyTUhW77VeARpavi0vZ1Z_OHBxl8A9?usp=sharing
  (Da die Google-Laufzeitumgebung benutzt wird, muss nichts lokal installiert werden)



- Außerdem ist die Datei in unserem GitHub-Repository zu finden:
  https://github.com/LennartFertig/DataExploration_HousePriceSeattle

Abbildungen:

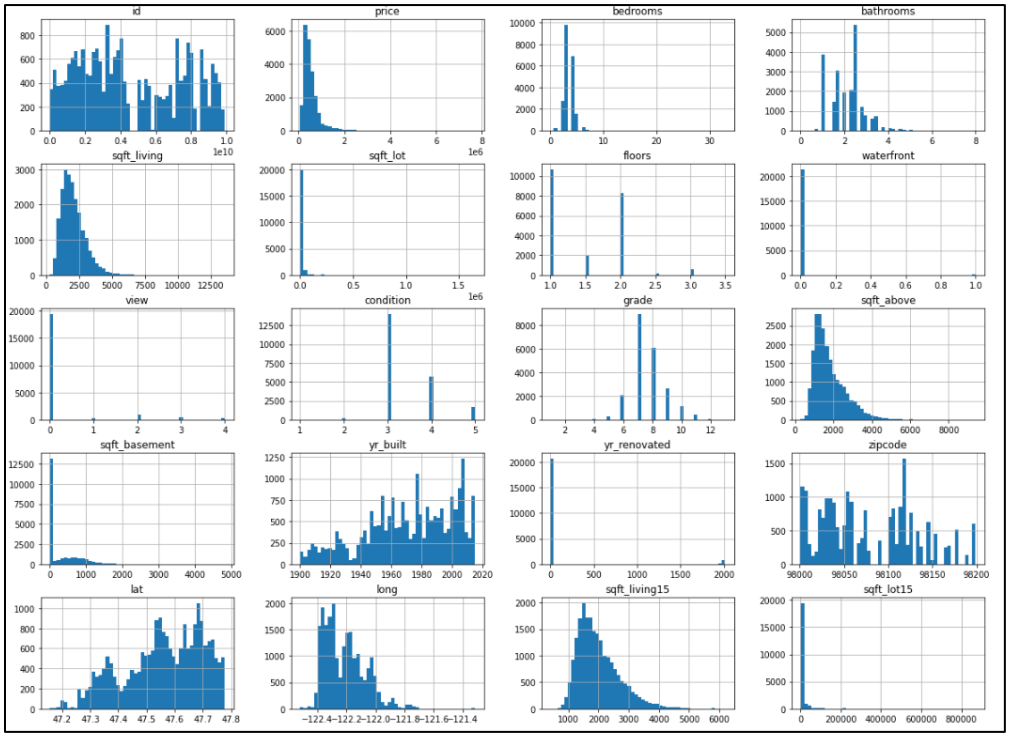


Abbildung 1: Histogramme zur Visualisierung der Mengenverteilung der einzelnen Merkmale

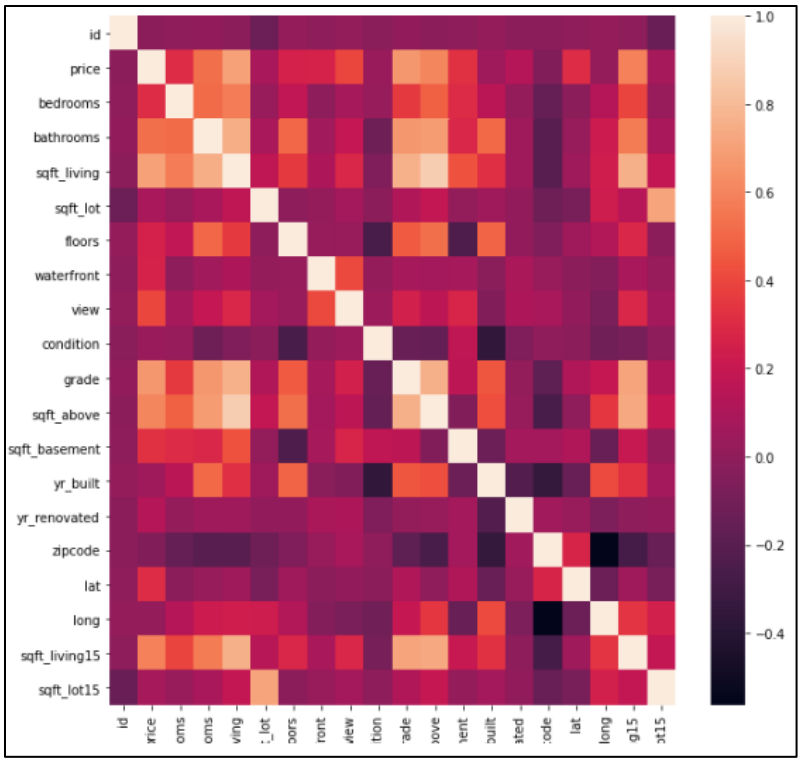


Abbildung 2: Heatmap zur Visualisierung der Korrelation der einzelnen Merkmale.

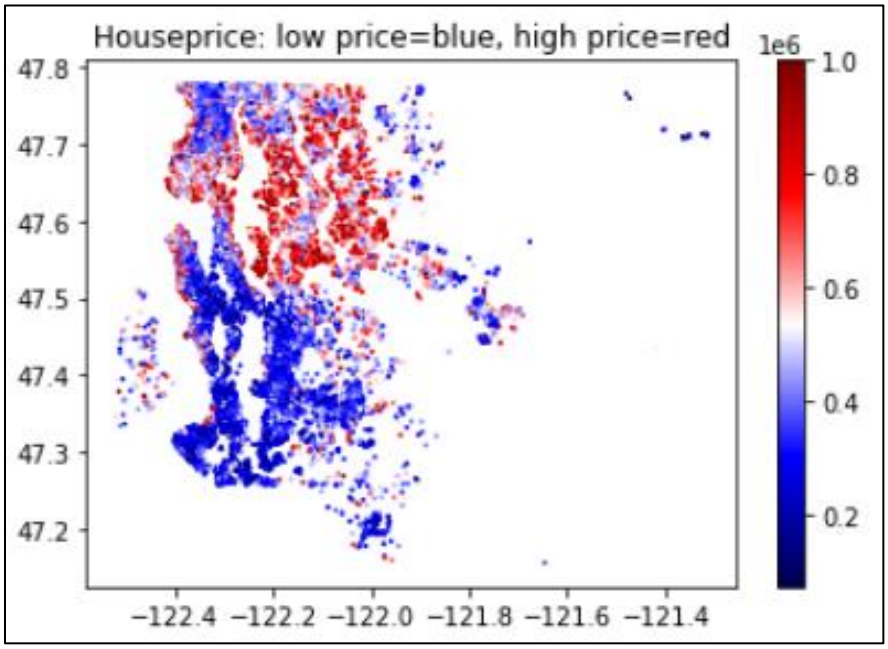


Abbildung 3: Visualisierung der geografischen Lage und des Preises einzelner Immobilien in Seattle.