Bericht Feature Engineering zum Kanton Solothurn

In dieser Arbeit wurde das Ziel verfolgt, mittels Feature-Engineering zwei Machiene Learning Modelle so zu optimieren, dass sie in der Lage sind, dass im Assignment 1 erstellte Gemeindranking anhand neuer, unabhängiger Features abzubilden. Dabei wurden verschiedene Feature-Engineering-Techniken evaluiert, um die Performance zweier Modelle, der Linearen Regression und des Gradient Boostings, zu optimieren. Die Experimente umfassten diverse Schritte, welche in den Methoden aufgeführt sind. Zusätzlich wurde ein neues geographisches Feature, die Anzahl der Volg-Läden im Umkreis von 5 km, hinzugefügt.
Die Ergebnisse zeigen, dass Feature Engineering einen signifikanten Einfluss auf die Modellleistung hat. Insbesondere die Log-Normalisierung verbesserte das lineare Modell deutlich. Die besten Ergebnisse wurden durch die Kombination von Log-Normalisierung, Erstellung polynomialer Features und anschliessender Feature-Selektion mittels Kollinearitätsprüfung und RFE erzielt.

Methoden

Die Grundlage der Arbeit bildete ein Datensatz mit verschiedenen Merkmalen der Gemeinden des Kantons Solothurn sowie der im Assignment 1 ermittelte "Score" als Zielvariable.
Folgende Feature-Engineering-Methoden kamen zum Einsatz:

- Datenvorverarbeitung: Entfernung ursprünglicher und damit zusammenhängender Features (wie Landwirtschaftsfläche, Ausländeranteil, Kriminalitätsrate und Bevölkerungsdichte). Das Datum der Aufnahme als Tage seit dem frühesten Datum kodieren.
- Geographisches Feature: Ermittlung der Koordinaten jeder Gemeinde mittels der geo.admin.ch API und Zählung der Volg-Läden im Umkreis von 5 km über die Overpass API.
- Feature-Normalisierung: Anwendung und Vergleich von Log-Transformation (log1p), MinMax-Skalierung und Z-Normalisierung (Standardisierung).
- Feature-Transformation: One-Hot Encoding der Bezirksnummer zur Berücksichtigung kategorialer Information.
- Polynomiale Features: Erstellung von polynomialen Features (Grad 2) auf Basis des gesamten Datensatzes sowie auf ausgewählten Feature-Kombinationen, um nicht-lineare Zusammenhänge zu modellieren.
- Feature-Selektion: Entfernung kollinearer Features basierend auf einem Korrelationsschwellenwert. Anwendung von Recursive Feature Elimination (RFE) auf die one-hot-kodierten Features sowie auf die polynomialen Features nach Entfernung kollinearer Merkmale, um die informativsten Features für beide Modelle (Linear Regression und Gradient Boosting) zu identifizieren.

Ergebnis

Die Leistung der Modelle wurde anhand des Negative Mean Squared Error bewertet, wobei Werte näher bei 0 eine bessere Leistung anzeigen. Die Experimente zeigten deutliche Unterschiede in der Wirksamkeit der Feature-Engineering-Schritte für die beiden Modelle. Die Resultate sind in der folgenden Abbildung 1 gezeigt.

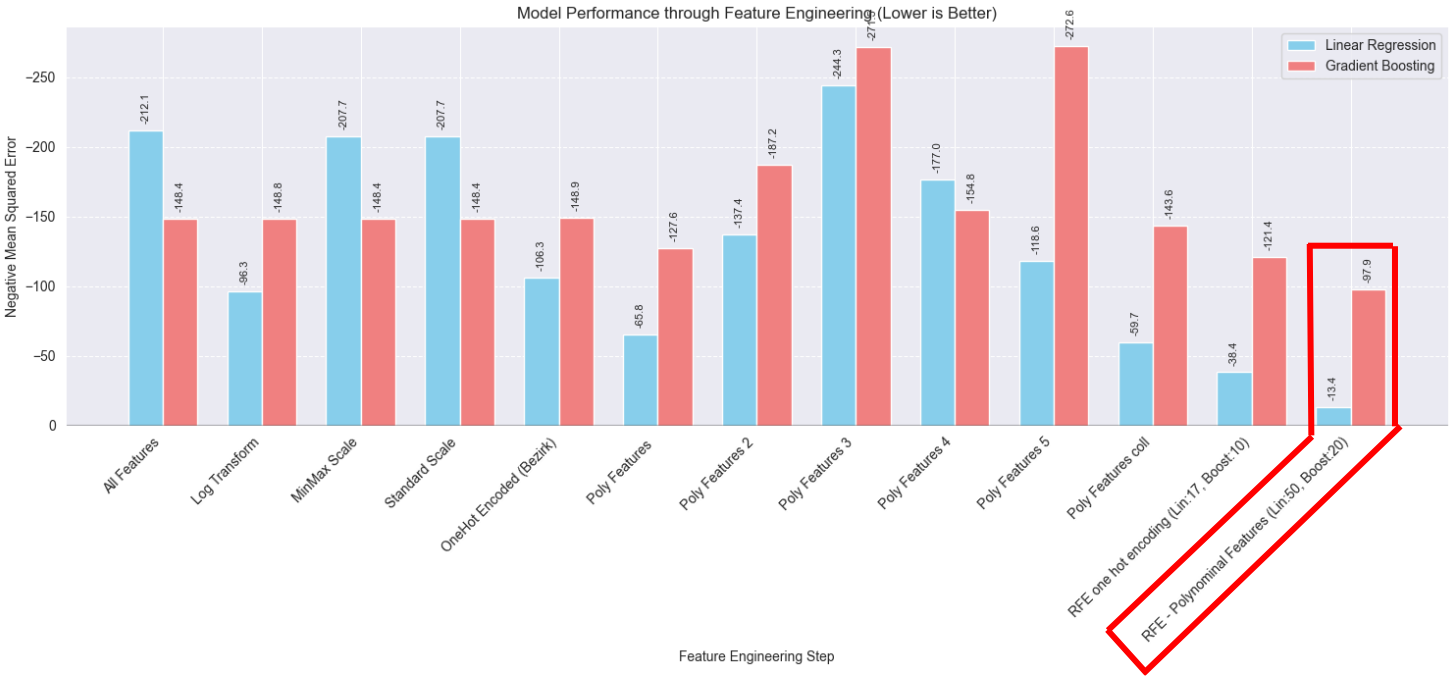

Abbildung 1: Model Performance durch Feature Engineering Schritte

Die Ergebnisse zeigen, dass die Log-Transformation eine signifikante Verbesserung für das lineare Modell brachte. Was auf die Effektivität dieser Methode zur Adressierung von Schiefe und zur Stabilisierung der Varianz in den Daten für lineare Modelle hindeutet. Das Hinzufügen polynomialer Features auf dem gesamten Datensatz verbesserte beide Modelle weiter. Die anschliessende Entfernung kollinearer Features führte zu einer leichten Verbesserung des linearen Modells und verschlechterte das Boosting Modell leicht. Jedoch wurden durch diesen Schritt genug Features eliminiert, sodass RFE innerhalb von akzeptabler Zeit durchlaufen konnte. Die Anwendung von RFE, insbesondere auf dem Datensatz mit polynomialen Features und nach Entfernung kollinearer Features, resultierte in den besten Scores für beide Modelle. Das lineare Modell profitierte am stärksten von der Kombination aus Log-Transformation, polynomialen Features und RFE, während das Boosting-Modell ebenfalls Verbesserungen zeigte, wenn auch weniger drastisch durch den fehlenden Einfluss der Normalisierung.

Schlussfolgerung

Die Arbeit demonstriert eindrucksvoll, wie gezieltes Feature Engineering die Leistung von Machine Learning Modellen verbessern kann. Das lineare Regressionsmodell zeigte die grösste Sensibilität gegenüber den angewandten Feature-Engineering-Schritten. Das Gradient Boosting Modell zeigte fast keine Abhängigkeit von der Normalisierung der Features. Dies ist konsistent mit der Funktionsweise baumbasierter Modelle, die auf Schwellenwerten und nicht auf Distanzen zwischen Datenpunkten operieren.

Die Auswahl von 50 Features durch RFE und den dadurch sehr tiefen MSE für das lineare Modell könnten Indikatoren für Overfitting sein. Angesichts der Grösse des Datensatzes könnte eine derart hohe Anzahl von Prädiktoren dazu führen, dass das Modell spezifische Muster in den Trainingsdaten lernt, anstatt die zugrundeliegende Datenstruktur zu generalisieren. Obwohl der ermittelte NMSE auf den Validierungsdaten vielversprechend erscheint, ist eine abschliessende Beurteilung des Generalisierungsvermögens des Modells ohne Evaluation auf einem unabhängigen Testdatensatz limitiert.