Feature Engineering Gemeinderanking

Ausgangslage

Im vorgängigen Assignment erstellte ich aus verschiedensten Features drei Ranglisten der Gemeinden des Kanton Waadt. Eine dieser Ranglisten galt es durch ein Machine Learning Modell und eine neue Auswahl von Features möglichst genau nachzubilden. Die Basis bestand dabei aus dem vom BFS veröffentlichten Gemeindeportrait (2021) mit Kennzahlen zu Schweizer Gemeinden sowie der Anzahl Restaurants, welche über die OpenStreetMap API abgefragt wurden.

Feature Engineering

In einem ersten Schritt mussten die Daten bereinigt werden. Dazu gehörte unter anderem die Behandlung von fehlenden Werten und die Berücksichtigung von Gemeindefusionen, welche zwischen den beiden Datenständen durchgeführt wurden.

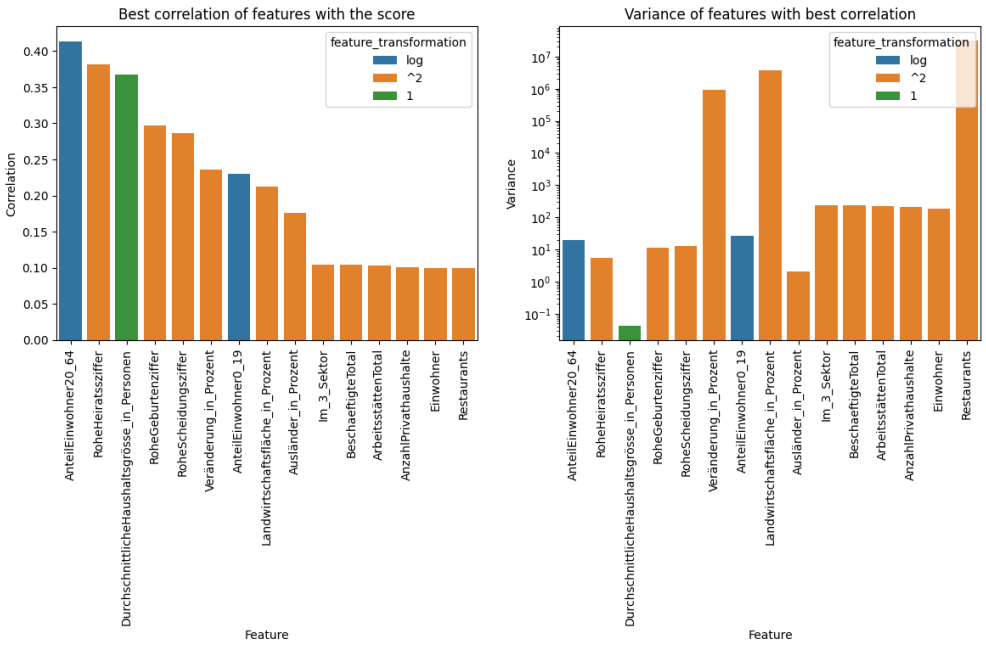
Um die Features insbesondere für das Linear-Regression Modell zu optimieren, wurden die Features mit dem Standard-Scaling Verfahren normalisiert. Die originalen Werte wurden dabei nicht überschrieben, da diese für das Gradient-Boosting Modell bessere Resultate liefern.

Zudem wurden die Features auch transformiert. Dabei verwendete ich die Quadrierung, die Wurzel und die Logarithmierung. Durch solche Transformationen können die Features u. U. das ursprüngliche Ranking besser nachbilden.

Es wurde zusätzlich versucht, durch die Kombination von Features die Resultate zu optimieren. Konkret berechnete ich aus den prozentualen Anteilen der Altersklassen die absoluten Werte. Ebenfalls berechnete ich die Anzahl Restaurant pro Einwohner und pro Einwohner Kategorie. Diese neuen Features brachten allerdings keine Optimierung.

Schliesslich wählte die 10 Features aus, welche die höchsten Korrelationen mit dem nachzubildenden Ranking erreichten. Automatisiert testete ich für beide Machine Learning Modelle alle Kombinationen dieser Features aus, um den besten Score von den Modellen zu erhalten.

Hier sind die 15 Features abgebildet, welche die höchste Korrelation aufgewiesen haben. Die Streuung der Varianz dieser Features ist sehr hoch.



Resultat

Beim Linear-Regression Modell konnte ich den Score von anfänglich **-204.45** (Second Baseline mit originalen Features aus Assignment 1) auf -**109.92** verringern. Diese Verbesserung erreichte ich mit folgenden Features:

- AnteilEinwohner20_64 (normalized|log)
- RoheHeiratssziffer (normalized|^2)
- DurchschnittlicheHaushaltsgrösse_in_Personen
- RoheGeburtenziffer (normalized|^2)
- Veränderung_in_Prozent (^2)
- AnteilEinwohner0_19 (normalized|log)
- Ausländer_in_Prozent (normalized|^2)

Beim Gradient-Boostring Modell reduzierte sich der Score von **-218.51** (selbe Ausgangslage wie Linear-Regression Modell) auf **-155.69**. Folgende Features wurden dafür verwendet:

- Einwohner0_19 ()2)
- Einwohner65_plus (^2)
- ranking1_criteria1
- ranking3_criteria3 (^2)

Fazit

Mittels Feature Engineering konnte ich den Score deutlich verbessern. Es ist aufgefallen, dass durch die unterschiedliche Funktionsweise beider Modelle auch auf verschiedene Features sowie Aufarbeitungs-Schritte notwendig sind. So ist z.B. die Normalisierung für Gradient-Boosting überhaupt keine geeignete Methode.