Einleitung:
Das Ziel dieses Assignments war es, die Genauigkeit eines Machine-Learning-Modells zur Schätzung der Qualität von Gemeinden zu verbessern. In Assignment 1 habe ich Daten aus verschiedenen Quellen gesammelt und eine erste Datenbereinigung durchgeführt, was zu einem Dataframe mit über 45 Spalten führte. Jedoch sind nicht alle dieser Spalten gleichermaßen relevant für die Vorhersage des Qualitäts-Scores von Gemeinden. Daher haben ich verschiedene Feature-Engineering-Techniken angewendet, um die informativsten Features zu identifizieren und irrelevante zu eliminieren.

Methoden:
Nachdem die Daten aufbereitet wurden, begann ich mit der Visualisierung der Daten mittels Boxplots, Barplots und Heatmaps, um Muster und Korrelationen zwischen Features und Score zu identifizieren. Für Ausgewählte Features habe ich Transformationen durchgeführt und nur diejenigen behalten, welche beim Plot am nächsten an einer linearen Funktion dran waren. Mit Hilfe der Boxplots habe ich Ausreiser visualisiert. Anschliessend habe ich die Ausreisser eliminieret. Danach wurde die MinMaxScaler()-Normalisierungsmethode auf alle Spalten im Dataframe angewendet, um sie vergleichbar zu machen und die übrig gebliebenen Ausreißer zu reduzieren. Zudem wurde auch Feature-Selektion mittels der Recursive Feature Elimination-Methode für lineare Modelle durchgeführt und die informativsten Features für das Boosting-Modell ermittelt. So kam ich zu den Resultaten. Zuletzt habe ich noch mit Feature Agglomeration und Random Projection versucht das Resultat zu verbessern.

Ergebnisse:
Meine Ergebnisse zeigen, dass die Anwendung der MinMaxScaler()-Normalisierungsmethode auf alle Spalten im Dataframe zu den besten Leistungen sowohl für lineare als auch für Boosting-Modelle führt. Bei der Feature Selection war es notwendig verschiedene Anzahl auszuprobieren bis das Score am besten da stand. Ebenfalls habe ich festgestellt das Features die eine Korrelation um 0 haben, sehr wenig Einfluss auf das Resultat haben. Das Eliminieren der Ausreisser hatte nicht den gwünschten Effekt, erst als Min/Max Normalisierung angewandt wurde konnte das Score effizient gesenkt werden. Random Projection und Agglomeration brachten auch nicht mehr die gewünschten Verbesserungen.

Schlussfolgerung:
Zusammenfassend hat mein Feature-Engineering-Prozess zu signifikanten Verbesserungen in der Genauigkeit von Machine-Learning-Modellen zur Schätzung der Qualität von Gemeinden geführt. Ich haben die informativsten Features identifiziert, irrelevante eliminiert und geeignete Normalisierungs- und Transformations-Techniken angewendet, um potenzielle Ausreißer und rechtsschiefe Verteilungen zu reduzieren. Meine Ergebnisse betonen die Bedeutung von sorgfältigem Feature Engineering im Machine Learning und zeigen dessen Potenzial zur Verbesserung der Genauigkeit von Modellen für verschiedene Anwendungen.

Beste Features linear(7):
Total Aktiven, Finanzvermögen, Verwaltungsvermögen, total Passiven, Fremdkapital, Eigenkapital und Aufwertungsreserven

Beste Features boosting (3)

Lenker/innen, Fussgänger/innen und Fremdkapital