Zusammenfassung:
In diesem Bericht fassen wir die Ergebnisse unseres Feature-Engineering-Prozesses zusammen, mit dem Ziel, die Leistung eines Machine-Learning-Modells zur Schätzung der Qualität von Gemeinden zu verbessern. Wir haben Datenbereinigung, Visualisierung, Feature-Transformation und Feature-Selektion durchgeführt, um die relevantesten Features zu identifizieren und irrelevante zu eliminieren. Unsere 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. Wir haben auch festgestellt, dass bestimmte Features, wie 'Total Aktiven', 'Finanzvermögen¹', 'Verwaltungsvermögen', 'Total Passiven', 'Fremdkapital', 'Eigenkapital' und 'Aufwertungsreserven', stark mit dem Score korrelieren und signifikant zur Leistung des Modells beitragen. Schließlich haben wir die Log-Funktion auf bestimmte Spalten angewendet, um rechtsschiefe Verteilungen anzugehen und potenzielle Ausreißer zu reduzieren. Unsere Ergebnisse zeigen, dass geeignete Feature-Engineering-Techniken zu signifikanten Verbesserungen in der Genauigkeit von Machine-Learning-Modellen zur Schätzung der Qualität von Gemeinden führen können.

Einleitung:
Das Ziel dieser Studie war es, die Genauigkeit eines Machine-Learning-Modells zur Schätzung der Qualität von Gemeinden zu verbessern. In Assignment 1 haben wir Daten aus verschiedenen Quellen gesammelt und eine erste Datenbereinigung durchgeführt, was zu einem Dataframe mit 48 Spalten führte. Jedoch sind nicht alle dieser Spalten gleichermaßen relevant für die Vorhersage des Qualitätsscores von Gemeinden. Daher haben wir verschiedene Feature-Engineering-Techniken angewendet, um die informativsten Features zu identifizieren und irrelevante zu eliminieren.

Methoden:
Wir begannen mit der Visualisierung der Daten mittels Boxplots, Barplots und Heatmaps, um Muster und Korrelationen zwischen Features und Score zu identifizieren. Wir haben dann die MinMaxScaler()-Normalisierungsmethode auf alle Spalten im Dataframe angewendet, um sie vergleichbar zu machen und potenzielle Ausreißer zu reduzieren. Wir haben auch Feature-Selektion mittels der Recursive Feature Elimination-Methode für lineare Modelle durchgeführt und die informativsten Features für das Boosting-Modell ermittelt. Schließlich haben wir die Log-Funktion auf bestimmte Spalten angewendet, um rechtsschiefe Verteilungen anzugehen und potenzielle Ausreißer zu reduzieren.

Ergebnisse:
Unsere 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. Wir haben auch festgestellt, dass bestimmte Features, wie 'Total Aktiven', 'Finanzvermögen¹', 'Verwaltungsvermögen', 'Total Passiven', 'Fremdkapital', 'Eigenkapital' und 'Aufwertungsreserven', stark mit dem Score korrelieren und signifikant zur Leistung des Modells beitragen. Darüber hinaus haben wir bestimmte Spalten identifiziert, die rechtsschief waren, und die Log-Funktion angewendet, um potenzielle Ausreißer zu reduzieren und die Leistung zu verbessern. Insgesamt zeigen unsere Ergebnisse die Bedeutung von sorgfältigem Feature Engineering im Machine Learning und demonstrieren dessen Potenzial zur Verbesserung der Genauigkeit von Modellen für verschiedene Anwendungen.

Schlussfolgerung:
Zusammenfassend hat unser Feature-Engineering-Prozess zu signifikanten Verbesserungen in der Genauigkeit von Machine-Learning-Modellen zur Schätzung der Qualität von Gemeinden geführt. Wir haben die informativsten Features identifiziert, irrelevante eliminiert und geeignete Normalisierungs- und Transformations-Techniken angewendet, um potenzielle Ausreißer und rechtsschiefe

Verteilungen zu reduzieren. Unsere Ergebnisse betonen die Bedeutung von sorgfältigem Feature Engineering im Machine Learning und zeigen dessen Potenzial zur Verbesserung der Genauigkeit von Modellen für verschiedene Anwendungen.