supper&supper

Automatisierte und nutzerfreundliche Pipeline zur Analyse landwirtschaftlicher Feldversuche

Projektziel

Die Extraktion von Rohdaten aus einer Datenbank, Daten Wrangling, Erstellung von Datenvisualisierungen und die Anwendung statistischer Schlussfolgerungen sind die wichtigsten wiederkehrenden Schritte eines Analyseprojekts. Für die Analyse von Feldversuchsdaten zur Bestimmung der Wirksamkeit von Produkten und zur Ableitung von Maßnahmen für deren Verwendung werden sowohl ein Domänenexperte als auch Analysten oder Data Scientisten benötigt. Sie helfen dem Business aus den Daten Erkenntnisse zu gewinnen und Antworten auf ihre Forschungsfragen zu finden.

Viele der unternommenen Schritte wiederholen sich ohnehin und sind für verschiedene Analysen anwendbar. Ziel dieses Projekts war es daher, ein Tool zu entwickeln, das eine allumfassende automatische Pipeline für die Analyse landwirtschaftlicher Daten beinhaltet. Diese Pipeline soll verschiedene Datenbanksysteme, Visualisierungswerkzeuge, Training von maschinellen Lernmodellen und statistische Inferenz integrieren. Das Werkzeug ermöglicht es Personen mit nicht-technischem Hintergrund, neue Analysen zu initialisieren und erleichtert das Treffen von Geschäftsentscheidungen.

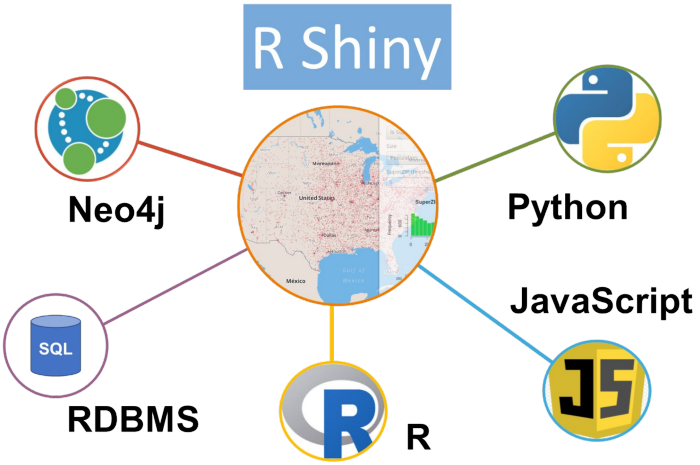
Herausforderungen

Die größte Herausforderung bestand darin, die verschiedenen Datenverarbeitungswerkzeuge zu kombinieren. Die Realisierung erforderte ein breites Wissen über Datenspeicherung, Datenextraktion, Datenladen und Datenanalysetechniken. Die zweite Herausforderung war das Interface Design. Das Tool war in erster Linie für Nicht-Technik-Anwender konzipiert. Daher musste es gut dokumentiert sein und erforderte eine intuitive, benutzerfreundliche Schnittstelle.

Angewandte Methoden

Interface Design

Die Benutzerschnittstelle wurde in R-shiny erstellt, einer interaktiven Webanwendung (App) direkt aus R. Sie bietet außerdem die Flexibilität, Datenvisualisierungsfunktionen mit JavaScript zu entwickeln.
Die Ergebnisse können sowohl in einem Webbrowser angezeigt als auch als pdf/docx-Format gespeichert werden. Über die Webschnittstelle können Benutzer die Daten visuell erforschen und grafische und statistische Ausgaben für verschiedene Anwendungsfälle erzeugen, ohne eine einzige Zeile Code schreiben zu müssen.



Datenextraktion– und Wrangling

Die Datenextraktion ist der erste Schritt der Pipeline. Da die Rohdaten strukturiert gespeichert sind, kann das entwickelte Tool die benötigten Daten automatisch aus verbundenen relationalen DBMS (z.B. SQLite, MySQL, PostgreSQL usw.) sowie aus einem Neo4j-Knowledgegraph und von verschiedenen Webdiensten laden.

Der Zieldatensatz wird durch Aufruf der verschiedenen Datenbanken auf der Grundlage von Benutzerfiltern erstellt. Diese Filter enthalten Schlüsselinformationen wie Bewertungstyp, Produktbestandteile, spezifische Versuche, Kulturen, Zeitrahmen und mehr. Nach der Extraktion werden diese Datensätze dann in das R-Backend importiert, wobei die Möglichkeit besteht, sie als csv-Datei herunterzuladen. Das Data Wrangling und Zusammenführen der Daten wurde in R mit Paketen wie dplyr, tidyr und stringr realisiert.

Datenvisualisierung

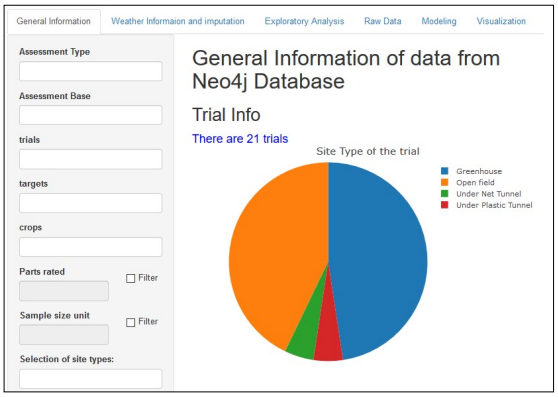
Durch die Integration einer JavaScript-Bibliothek generiert das Tool interaktive Karten, die die Versuchsstandorte und damit verbundene Informationen wie Wetterdaten, Bodeninformationen und weitere Metadaten anzeigen. Die Benutzer können auch hinein- und herauszoomen, um bestimmte Versuche auf der Karte für detailliertere Informationen auszuwählen.

Statistische Modellbildung

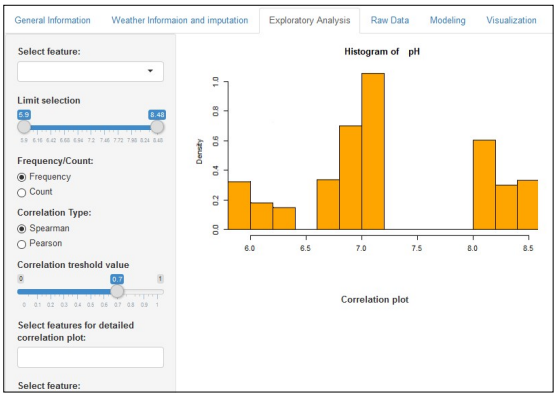
Für die Modellierung werden Random Forest und XGBoosting verwendet. Der Benutzer erhält einen Überblick über die Leistungsfähigkeit des Modells (R^2), die Bedeutung der verschiedenen Faktoren und Einflüsse einzelner Faktoren sowie die Stärke der Wechselwirkung zwischen den Faktoren.

Projektergebnis

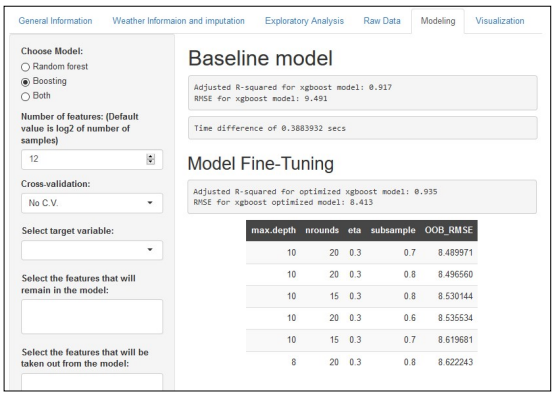
Das entwickelte Tool wurde erfolgreich implementiert und wird Produktmanager in ihrer operativen Arbeit unterstützen. Es bietet einen leicht zugänglichen Überblick über alle Studieninformationen zu einem Produkt und ermöglicht es, die wichtigsten Faktoren für eine erfolgreiche Produktanwendung und eine hohe Wirksamkeit herauszufinden. Die Daten werden in interaktiven, benutzerfreundlichen Dashboards präsentiert. Je nach neuen Geschäftsanforderungen und Anwendungsfällen können weitere Funktionen implementiert werden, um die Funktionalitäten zu erweitern.



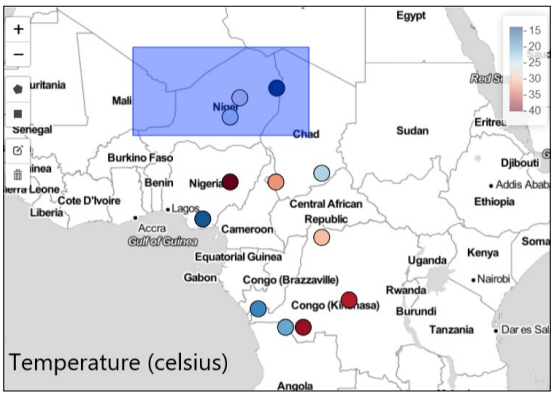
Allgemeine Informationen zu den Versuchen



Explorative Analyse der Merkmale



Machine Learning Modelltraining



Geographische Lage der Versuche