supper&supper

Automated and user-friendly Data Analysis Pipeline for Agricultural Field Trials

Project Goal

Extracting raw data from a database, conducting data wrangling, creating data visualizations and applying statistical inference are the main reoccurring steps of an analytics project. For the analytics of field trial data to determine efficacy of products and derive actions for the use of it, a domain expert as well as business analysts or data scientists are needed. They help the business people to gain insights from the data and find answers to their questions.

Anyways a lot of the taken steps are repetitive and applicable for different analyses. Therefore, the goal of this project was to develop a tool incorporating an all-in-one automatic pipeline for agricultural data analysis, that integrates different database systems, visualization tools, training of machine learning models and statistical inference. The tool enables people from a non-technical background to initialize new analyses and help them to make business decisions easily.

Challenges

The major challenge was to combine the diverse set of data processing tools. The realization required broad knowledge of data storage, data extraction, data loading and data analysis techniques.

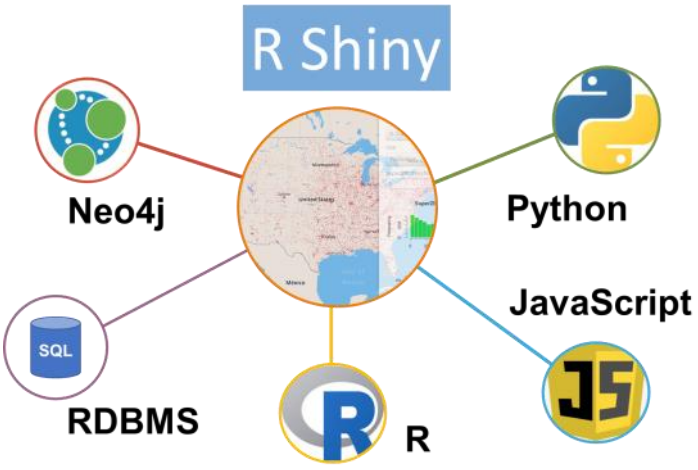
The second challenge was the interface design. The tool was designed primarily for non-tech users. Therefore the tool needed to be well documented and required an intuitive user-friendly interface.

Applied Methods

Interface Design

The user interface was built in R-shiny which is an interactive web application (app) straight from R. It also offers flexibility to develop data visualization functions with JavaScript.

The results can be displayed in a web browser as well as be saved as pdf/docx format. Through the web interface users can explore the data visually and produce graphical and statistical output for different use cases without writing a single line of code.



Data Extraction and Wrangling

Data extraction is the first step of the pipeline. As the raw data is stored in a structured way, the developed tool can fetch the needed data automatically, from connected relational database manage systems (e.g. SQLite, MySQL, PostgreSQL, etc.), a Neo4j knowledge graph and different web services.

The target dataset is obtained through calling the different databases based on filters set by the user. These filters include key information such as assessment type, product ingredients, specific trials, crops, timeframe and more. After extraction, these datasets are then imported into the R backend, with the option to be downloaded as a csv file. Data wrangling and merging in R was realized using packages like dplyr, tidyr and stringr.

Data Visualization

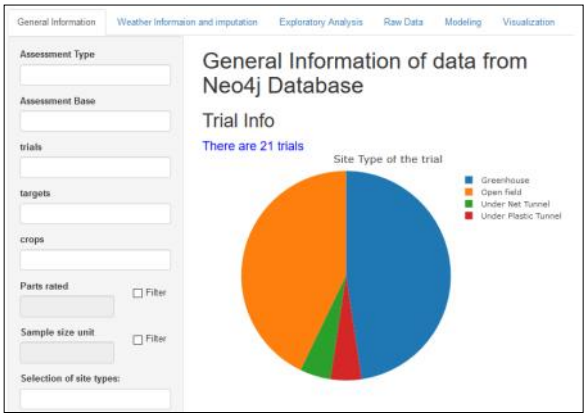
With the integration of a JavaScript library, the tool generates interactive maps showing the trial locations and related information, such as weather data, soil information and more meta data. Users can also zoom-in and -out to select specific trials on the map for more detailed information.

Statistical Modeling

Random forest and XGBoosting are used for the modeling. The user gets an overview of the model performance (R-square), the importance of different factors and influences of individual factors as well as the interaction strength between factors.

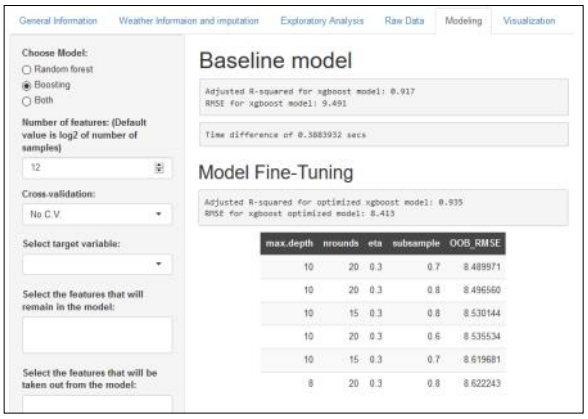
Project Outcome

The developed tool was successfully implemented and will support product managers in their operational work. It provides an easy accessible overview of all trial information regarding a product and makes it possible to discover the most important factors for successful product applications and a high level of efficacy. The data is presented in interactive user-friendly dashboards. Further functions can be implemented to extend the functionalities depending on new business needs and use cases.

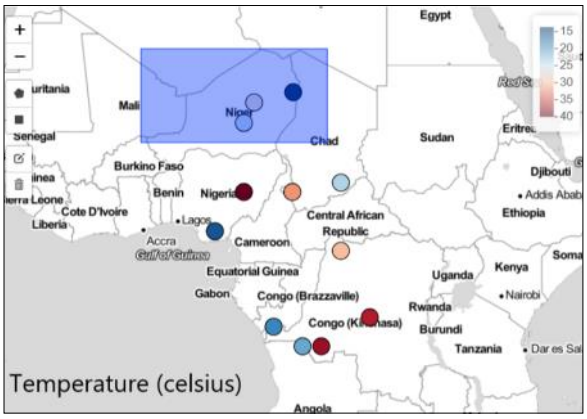


General information of the trials

Exploratory analysis of the features



Machine learning model training



Geographic location of the trials