supper&supper

Yield Prediction of Hybrids

Project Goal

The project goal was to predict how corn hybrid variants would behave at new sites under varying environmental conditions, i.e. temperature, rain fall and soil condition. Based on the data provided, a model was created to predict crop yields and furthermore identify new favorable combinations of corn hybrid variants and locations.

Provided Data

Three datasets were provided in total: One dataset included information from 2001 to 2016 on a variety of hybrids, their yield and location as well as comparative species. The next one contained the hybrids and sites for 2017 for which yields were to be predicted. The final dataset consisted of genetic markers of all hybrids and soil parameters, represented by monthly values covering the entire time span of a year and every location. Altogether the data represented more than 2,000 hybrid types and locations over a time span of 15 years.

Challenges

The data provided needed to be analyzed and cleaned up to ensure that the data mining algorithms work properly. The matrix that has been produced by merging the three data sets contained more than 3 billion data points. Therefore, we had to find a solution to reduce the amount of data points in order to apply our algorithms without requiring absurd amounts of CPU processing power. As the final hurdle, assumptions on weather had to be made, for which no parameters were available yet. This required the creation of an additional forecast model for the weather, apart from the original model to forecast crop yields.

Applied Methods

Data Cleaning

For each data set provided a thorough analysis of the data was performed. Outliers within different locations were identified by overlapping geographic data, climate information and a variety of events. These outliers were removed.

Big Data Problem

Many instances had a large intersection regarding the genetic material. Therefore, the dimensionality of the data set was significantly reduced by using a multi-factor analysis, allowing for simpler analysis of the data while keeping information loss to a minimum.

Weather Prediction

An Autoregressive Integrated Moving Average Model (ARIMA) was adapted to the residuals of a Fourier Regression Model. A Wave-Type Covariance Model was used for the temporal dimension, an Exponential Covariance Model was used for the spatial dimension and a spatio-temporal kriging was applied to predict the space-time random fields. This model predicted the weather with an accuracy of 95%.

Yield Prediction

The simplified data set of the genetic material, as well as the weather, soil and yield data were combined into a single data set and used as the training data set for the model (3-Fold Cross-Validation). Multiple algorithms were applied and then compared, such as Artificial Neural Networks, Support Vector Regression and Decision Trees. The best predictive quality was achieved with a Random Forest model which was able to predict hybrid performance with an accuracy of 75%.

Project Outcome

The model was successfully applied to the new hybrids at 20,000 new sites. This enabled us to identify the most efficient species for 2017.