Recommender System for Corn Genotypes

Hamdan Khalid, Bhargavi Katuru, Amrish Nayak, Jason Paik, Varun Rege

Abstract

Bayer is a German multinational, life sciences and pharmaceutical company. We were given 2 data sheets: one that had the gene pairings and the average yield they produced, the other had genetic similarity between two gene pairs. They wanted us to be able to recommend to the commercial breeders and farmers the highest grain-yielding corn cross breeds based on genetic pairing for specific geographical locations.
Our first objective was to understand recommender algorithms and the concepts associated with it. We were then tasked to read the data and do explorative analysis on the data. After wards, we had to manipulate the data, and filling in the missing data values. Using R programming language, we found the weighted average to fill in the ENTRY_MEAN(which is the average yield) column. For the future, we have to design an algorithm to predict the grain yield for the corn crosses that have never been tested before. We believe that the recommender system could also be used find the best gene pairings based on location as certain crops likely grow better under specific geographical conditions.

Introduction

Bayer is
- German multinational company
- Pharmaceutical company
- Life sciences company

Goal:
- This will help the farmers and breeders with he best seed for their location
- Model of crop yield from genetic pairings
- Identify the best male and female gene pairing
  - Best = Highest Mean of corn yield

Important Stakeholders:
- Commercial breeders
- Farmers
- Data Scientists

Data:
- 250,000 observations
- Some male/female pairings
- Yield given in bushels
- Genetic Similarity between male pairs of genes, and Genetic Similarity between female genes

Methodology

Exploratory Analysis:

- About 0.13% missing values in the MALE column
  - Not al lot of missing values
- 2018 has more testing samples
- More samples in the P3 experimental stage
- About 18% of the values missing in Similarity column
  - These need to be imputed
- Most similarity values lie in 0.7-0.9 range
  - Accurate measurement for modeling

Study:
- To understand concepts better:
  - Read “Deep Learning based Recommender System: A Survey and New Perspectives” by Zhang, Shuai, et al. 2019

- Modeling Data:
  - Inefficient and expensive to cross breed all pairs
  - Use Genetic Similarity
    - Find weighted average yield for other gene pairs

Process:

- Find weighted average yield for each male/female pair (Found in Figure 3)
- Input weighted average yield into the missing values (See Figure 2)
- Example:
  - Utilize variables such as :
    - Experiment Stage
    - Harvest Year
    - Test ID Number
    - Number of Recorded Observations
    - Category of the Observation
    - Average Yield
    - Female genotype
    - Male genotype
    - ID of the Farm Field
    - Geographical location of the field

  To impute values such as 151.7013 into the N/A value for ENTRY_MEAN (See Figure 3, Row 5)

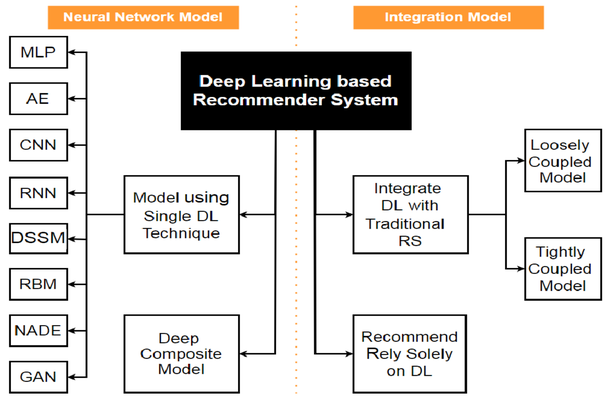

Figure 1: Concepts and understanding of a recommender system

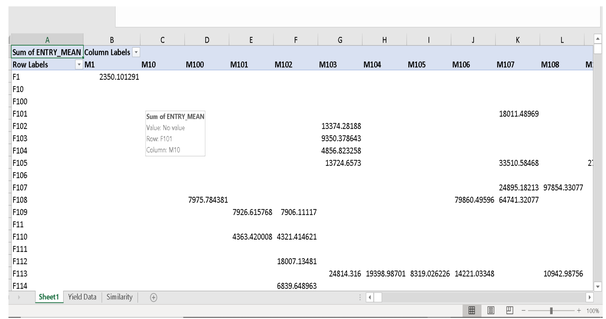

Figure 2: A small view of the missing values in the dataset given

	all_females	all_males	ENTRY_MEAN	Yield_Estimate_F	Yield_Estimate_M
	<fct>	<fct>	<dbl>	<dbl>	<dbl>
1	F1	M1	146.8813	150.7820	NaN
2	F1	M2	119.7950	143.6243	NaN
3	F1	M3	139.1489	147.6724	NaN
4	F1	M4	131.3756	154.9879	NaN
5	F1	M5	NA	151.7013	NaN
6	F1	M6	NA	144.0226	NaN

Figure 3:: The weighted data that we found: Yield_Estimate_F and Yield_Estimate_M

Discussion/Analysis

Which Data to use
- If Yield_Estimate_M value = null, use Yield_Estimate_F
  - Vice versa if Yield_Estimate_F = null
- If both M and F column are null;
  - Use a recommender algorithm to use genetic similarities to find the missing values .

Conclusion and Future Goals

- Process to finding missing values has been achieved, but there is much more that can be done.
- Design an algorithm to predict the grain yield for the corn crosses that have never been tested before.
- Find the best gene pairings based on location.
  - A drought land might produce better yield with a certain seed, but that seed might not work for a wetland area.

Acknowledgements

- We would like thank our corporate mentors: Dr. Chavali and Dr. Mirshekari.
- We would like to thank our data mine mentors: Dr. Ward, Ms Gundlach and Ms Betz

Sources

Zhang, Shuai, et al. “Deep Learning Based Recommender System: A Survey and New Perspectives.” *ACM Computing Surveys* (CSUR), 1 Feb. 2019, dl.acm.org/doi/10.1145/3285029.