Initial Model Report: Sassafras

Capstone

Samantha Harper

24 November 2024

Contents

Background

I haven't seen any obvious need to adjust these so far

Research Question

The research question is: Can we use machine learning algorithms to mine SNPs to find gene or gene regions of interest between natural cultivars (strains) of Sassafras?

Hypothesis

Hypothesis: Underlying genes, as identified by SNPs, in Sassafras are influenced by environmental factors because environmental pressure can cause mutations to persist in a population that is unique to each area.

Prediction

Prediction: Populations of Sassafras that are under high environmental pressure are more likely to have many predictive SNPs due to evolutionary influences.

Introduction

first draft of introduction for final paper

Methods

- plan for initial mode;
- results of EDA last week
- Explain initial model
- Explain initial model choice
- Explain cross validation
- Explain assumptions and testing

Testing

Algorithm

Assumptions

Overfitting

Discussion

- key takeaways and revised plan (refer to plan)
- Next steps for model tuning and selection
- Additional models and validation
- How to tune hyperparameters
- How did the initial model change plans?

Appendix

- data dictionary

https://github.com/samanthaharper/sassafras_capstone