Gene Severity Analysis:

Predicting Rater Concurrence

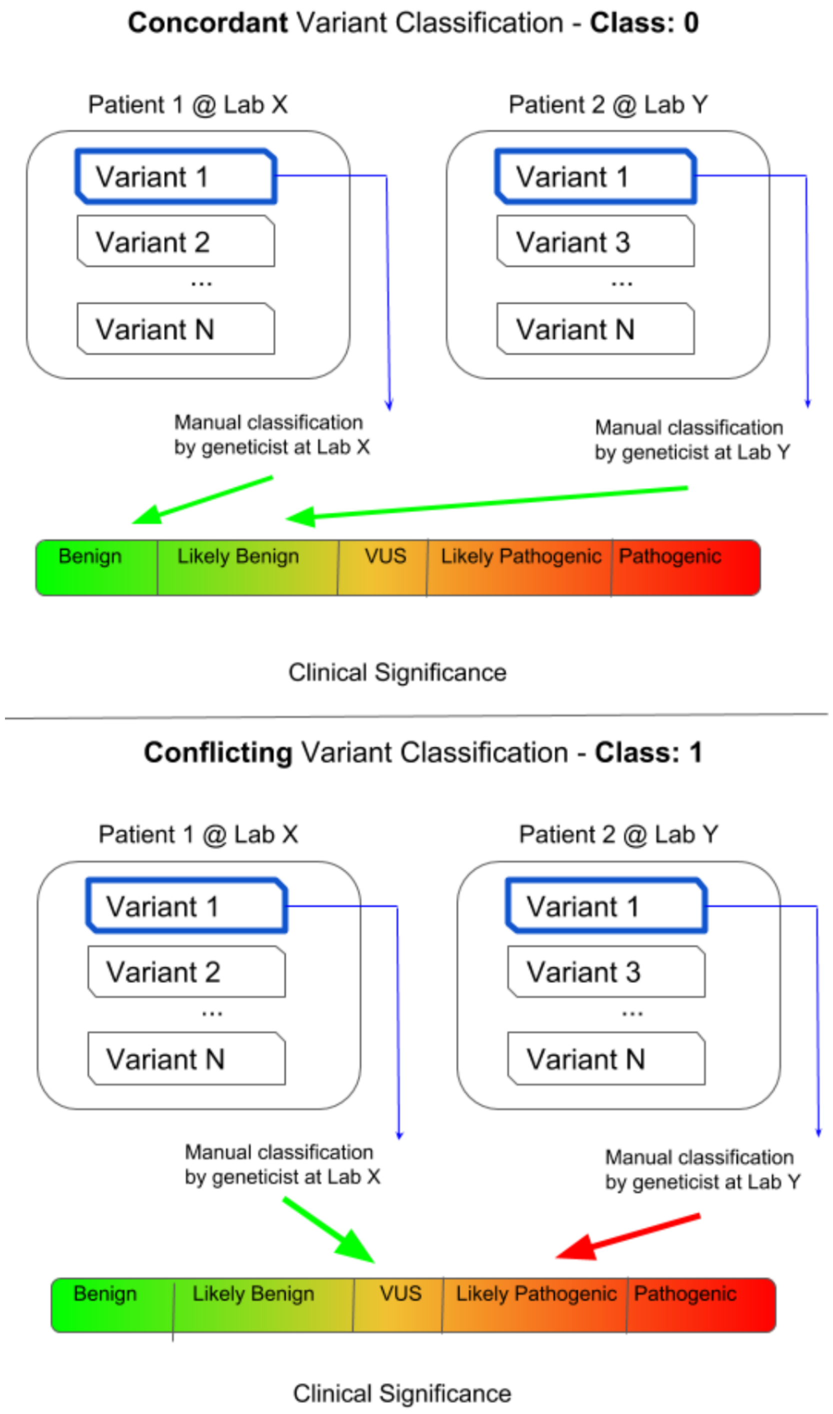
David Braslow

April 23, 2019

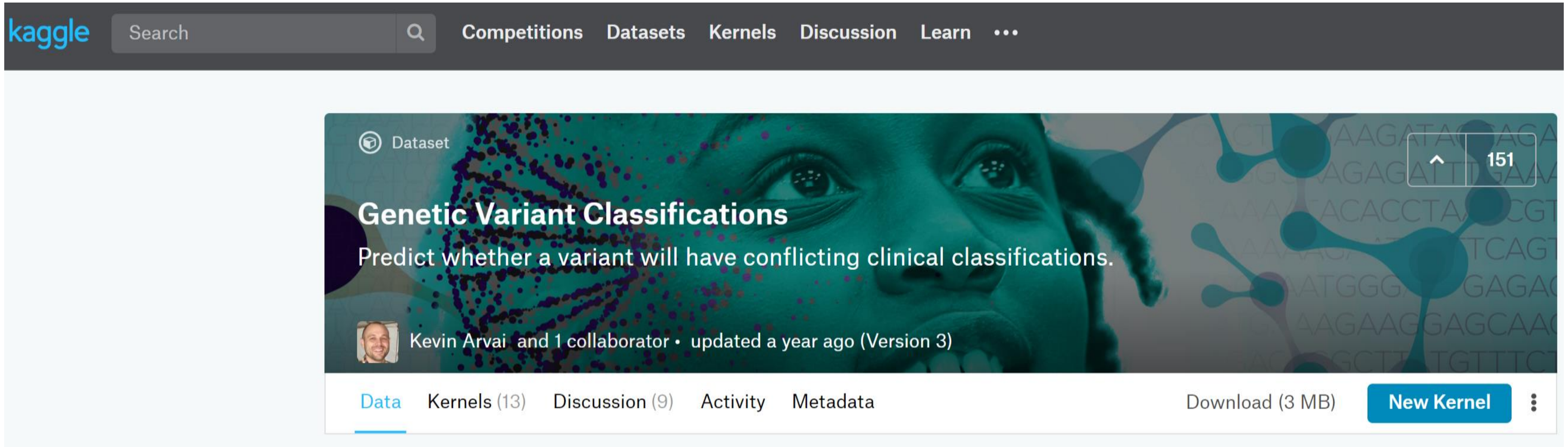
Introduction

Gene variants can have low or high clinical significance.

Can I predict when geneticists will disagree?



Obtaining the Data



Genetic Variant Classifications

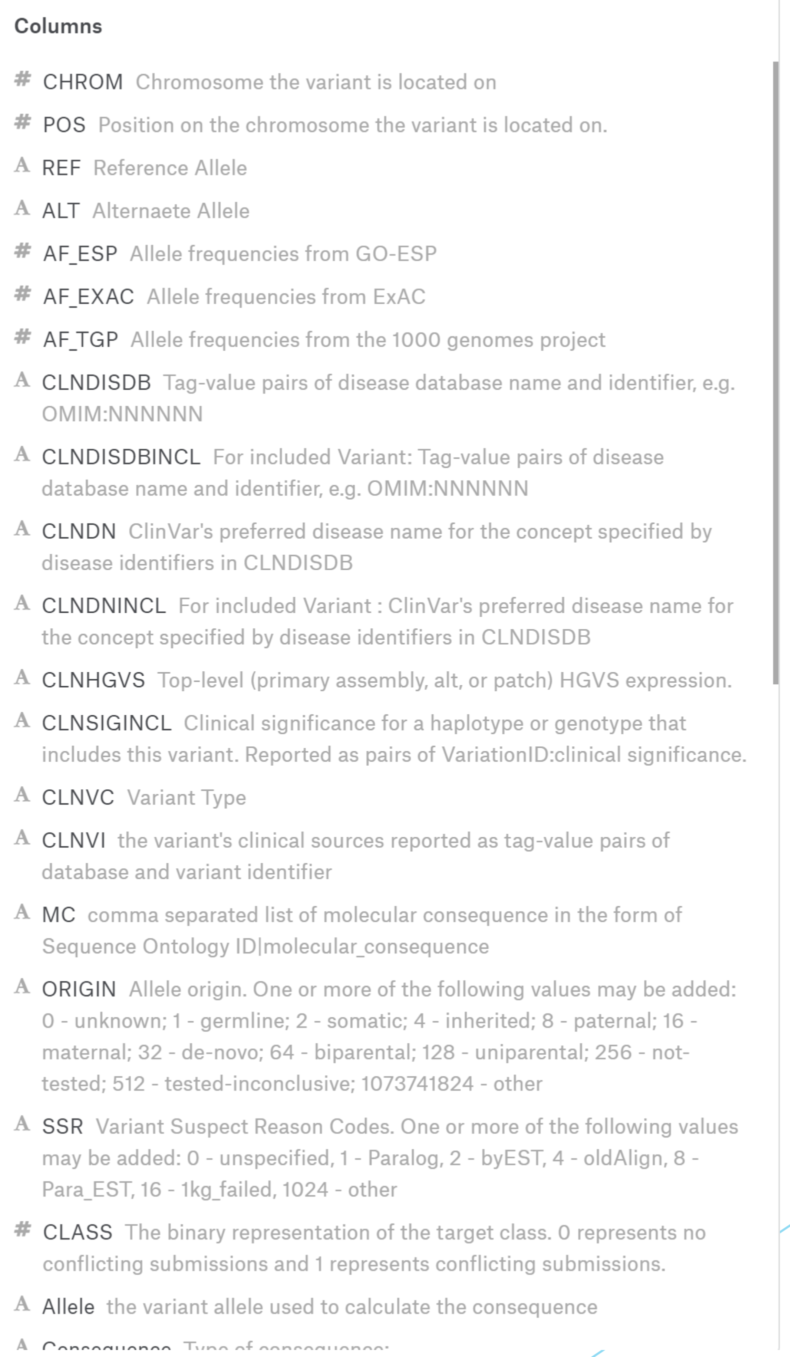
Predict whether a variant will have conflicting clinical classifications.

Kevin Arvai and 1 collaborator • updated a year ago (Version 3)

https://www.kaggle.com/kevinarvai/clinvar-conflicting/version/3

Scrubbing the Data

45 gene properties, most are categorical



Columns

# CHROM Chromosome the variant is located on

# POS Position on the chromosome the variant is located on.

A REF Reference Allele

A ALT Alternaete Allele

# AF_ESP Allele frequencies from GO-ESP

# AF_EXAC Allele frequencies from ExAC

# AF_TGP Allele frequencies from the 1000 genomes project

A CLNDISDB Tag-value pairs of disease database name and identifier, e.g. OMIM:NNNNNN

A CLNDISDBINCL For included Variant: Tag-value pairs of disease database name and identifier, e.g. OMIM:NNNNNN

A CLNDN ClinVar's preferred disease name for the concept specified by disease identifiers in CLNDISDB

A CLNDNINCL For included Variant : ClinVar's preferred disease name for the concept specified by disease identifiers in CLNDISDB

A CLNHGVS Top-level (primary assembly, alt, or patch) HGVS expression.

A CLNSIGINCL Clinical significance for a haplotype or genotype that includes this variant. Reported as pairs of VariationID:clinical significance.

A CLNVC Variant Type

A CLNVI the variant's clinical sources reported as tag-value pairs of database and variant identifier

A MC comma separated list of molecular consequence in the form of Sequence Ontology ID|molecular_consequence

A ORIGIN Allele origin. One or more of the following values may be added: 0 - unknown; 1 - germline; 2 - somatic; 4 - inherited; 8 - paternal; 16 - maternal; 32 - de-novo; 64 - biparental; 128 - uniparental; 256 - not-tested; 512 - tested-inconclusive; 1073741824 - other

A SSR Variant Suspect Reason Codes. One or more of the following values may be added: 0 - unspecified, 1 - Paralog, 2 - byEST, 4 - oldAlign, 8 - Para_EST, 16 - 1kg_failed, 1024 - other

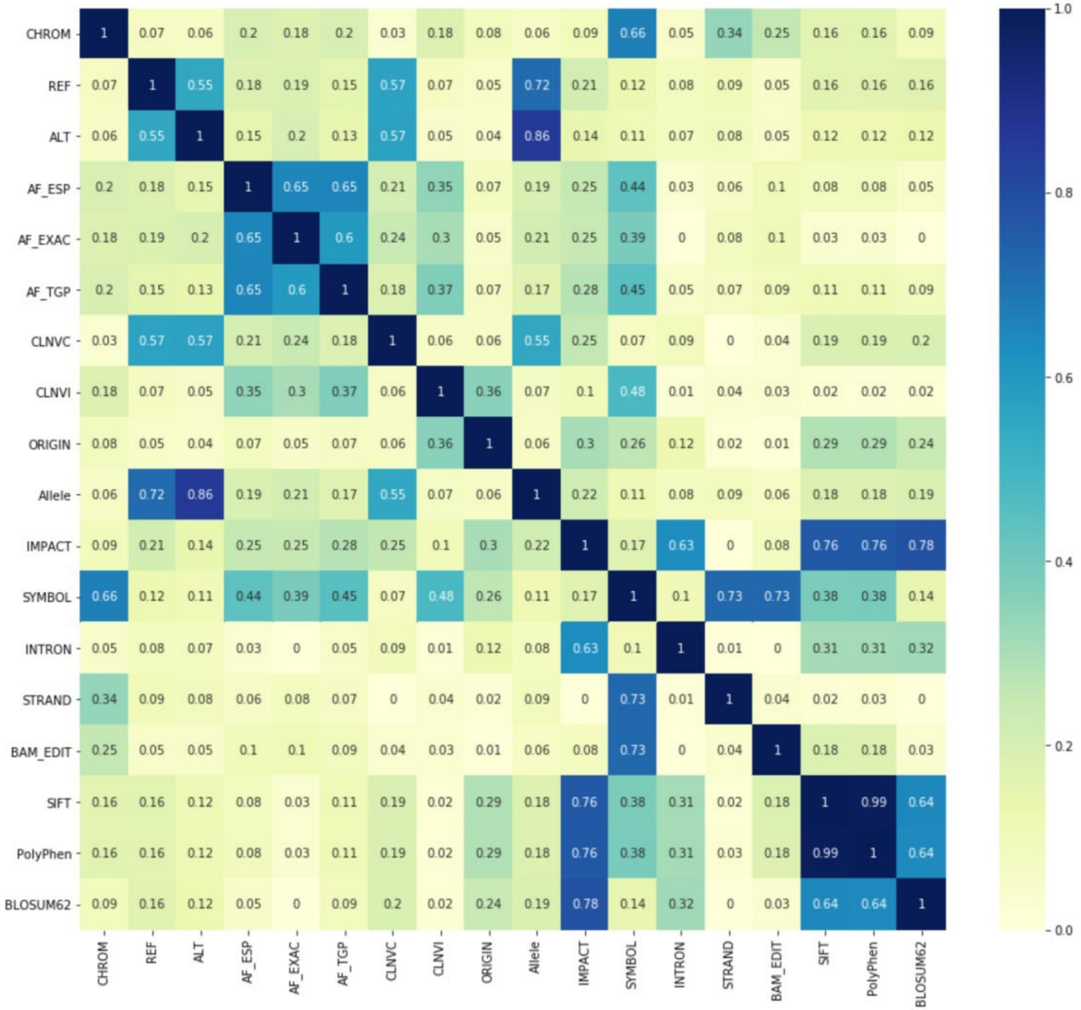
# CLASS The binary representation of the target class. 0 represents no conflicting submissions and 1 represents conflicting submissions.

A Allele the variant allele used to calculate the consequence

A Consequence Type of consequence:

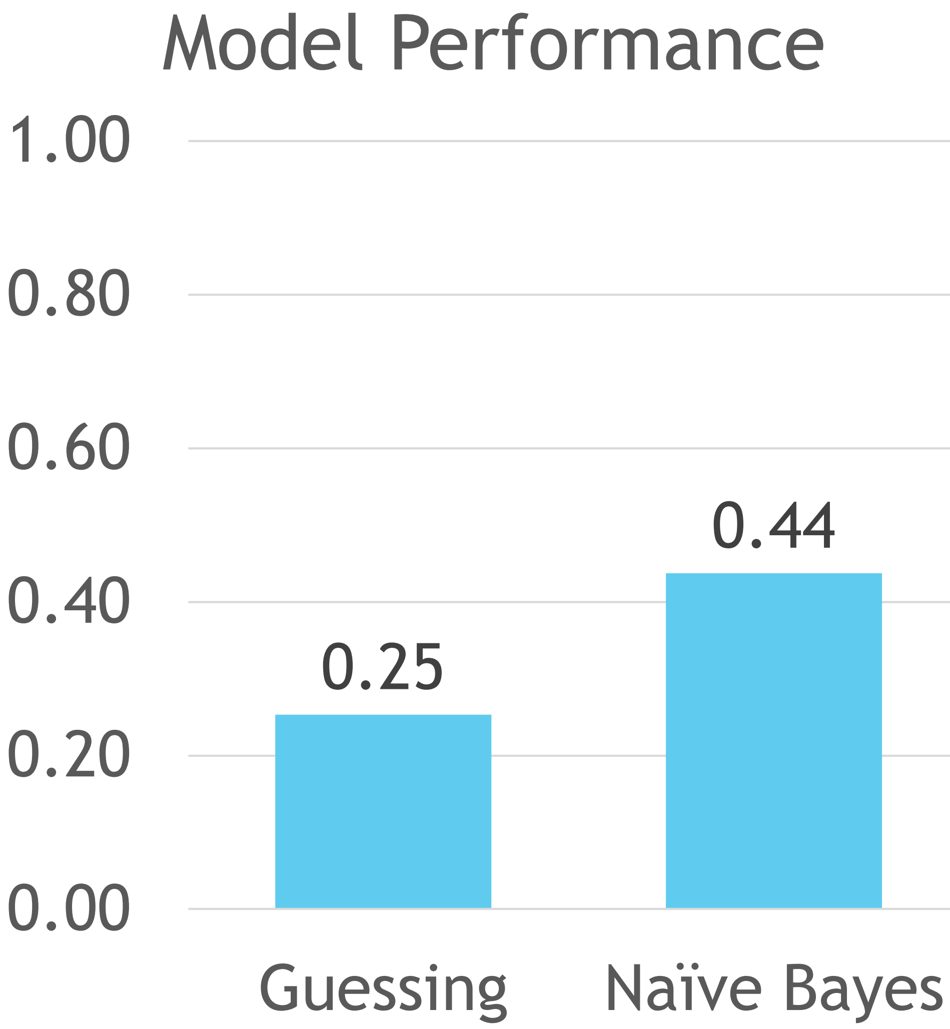
Exploring the Data

Feature Correlations

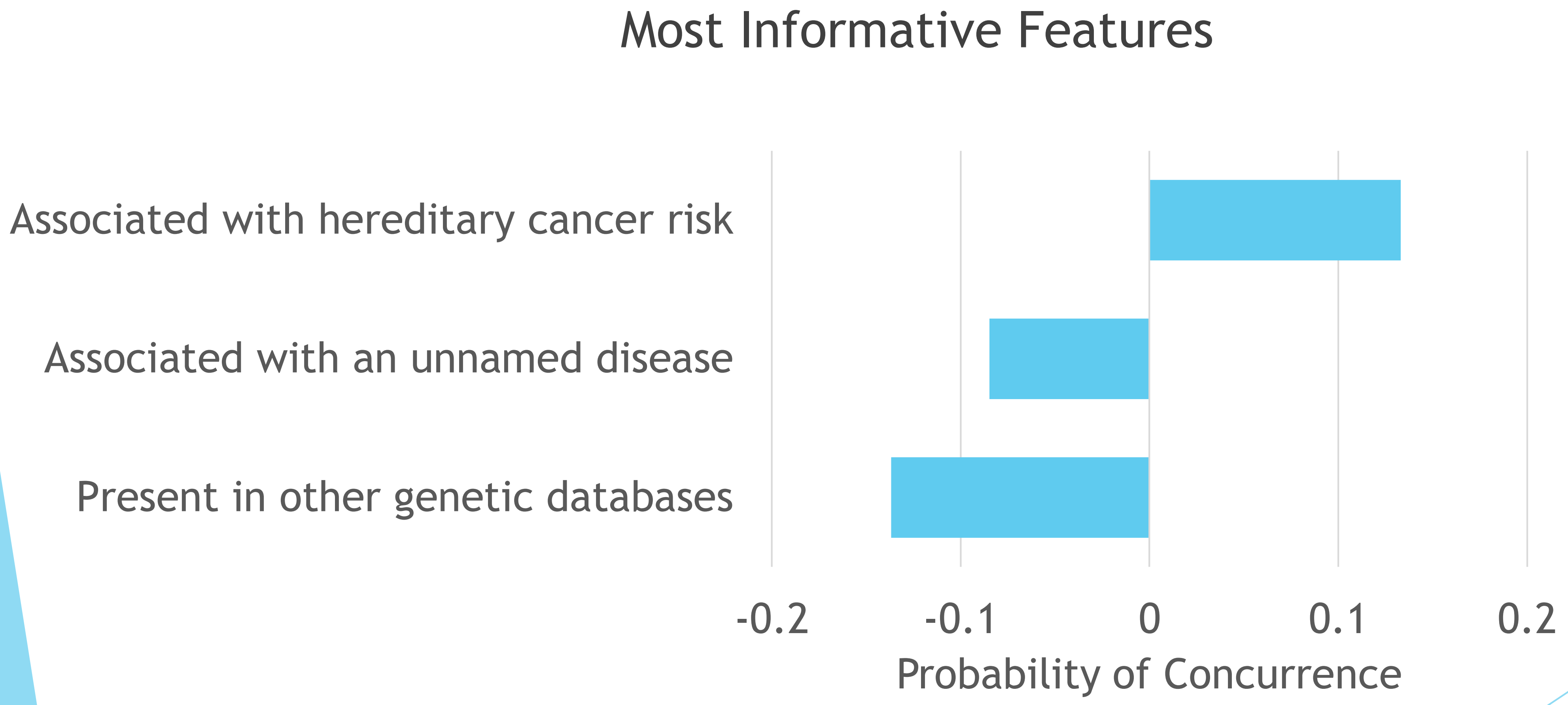


	CHROM	REF	ALT	AF_ESP	AF_EXAC	AF_TGP	CLNVC	CLNVI	ORIGIN	Allele	IMPACT	SYMBOL	INTRON	STRAND	BAM_EDIT	SIFT	PolyPhen	BLOSUM62
CHROM	1	0.07	0.06	0.2	0.18	0.2	0.03	0.18	0.08	0.06	0.09	0.66	0.05	0.34	0.25	0.16	0.16	0.09
REF	0.07	1	0.55	0.18	0.19	0.15	0.57	0.07	0.05	0.72	0.21	0.12	0.08	0.09	0.05	0.16	0.16	0.16
ALT	0.06	0.55	1	0.15	0.2	0.13	0.57	0.05	0.04	0.86	0.14	0.11	0.07	0.08	0.05	0.12	0.12	0.12
AF_ESP	0.2	0.18	0.15	1	0.65	0.65	0.21	0.35	0.07	0.19	0.25	0.44	0.03	0.06	0.1	0.08	0.08	0.05
AF_EXAC	0.18	0.19	0.2	0.65	1	0.6	0.24	0.3	0.05	0.21	0.25	0.39	0	0.08	0.1	0.03	0.03	0
AF_TGP	0.2	0.15	0.13	0.65	0.6	1	0.18	0.37	0.07	0.17	0.28	0.45	0.05	0.07	0.09	0.11	0.11	0.09
CLNVC	0.03	0.57	0.57	0.21	0.24	0.18	1	0.06	0.06	0.55	0.25	0.07	0.09	0	0.04	0.19	0.19	0.2
CLNVI	0.18	0.07	0.05	0.35	0.3	0.37	0.06	1	0.36	0.07	0.1	0.48	0.01	0.04	0.03	0.02	0.02	0.02
ORIGIN	0.08	0.05	0.04	0.07	0.05	0.07	0.06	0.36	1	0.06	0.3	0.26	0.12	0.02	0.01	0.29	0.29	0.24
Allele	0.06	0.72	0.86	0.19	0.21	0.17	0.55	0.07	0.06	1	0.22	0.11	0.08	0.09	0.06	0.18	0.18	0.19
IMPACT	0.09	0.21	0.14	0.25	0.25	0.28	0.25	0.1	0.3	0.22	1	0.17	0.63	0	0.08	0.76	0.76	0.78
SYMBOL	0.66	0.12	0.11	0.44	0.39	0.45	0.07	0.48	0.26	0.11	0.17	1	0.1	0.73	0.73	0.38	0.38	0.14
INTRON	0.05	0.08	0.07	0.03	0	0.05	0.09	0.01	0.12	0.08	0.63	0.1	1	0.01	0	0.31	0.31	0.32
STRAND	0.34	0.09	0.08	0.06	0.08	0.07	0	0.04	0.02	0.09	0	0.73	0.01	1	0.04	0.02	0.03	0
BAM_EDIT	0.25	0.05	0.05	0.1	0.1	0.09	0.04	0.03	0.01	0.06	0.08	0.73	0	0.04	1	0.18	0.18	0.03
SIFT	0.16	0.16	0.12	0.08	0.03	0.11	0.19	0.02	0.29	0.18	0.76	0.38	0.31	0.02	0.18	1	0.99	0.64
PolyPhen	0.16	0.16	0.12	0.08	0.03	0.11	0.19	0.02	0.29	0.18	0.76	0.38	0.31	0.03	0.18	0.99	1	0.64
BLOSUM62	0.09	0.16	0.12	0.05	0	0.09	0.2	0.02	0.24	0.19	0.78	0.14	0.32	0	0.03	0.64	0.64	1

Modeling the Data



Interpreting the Results



Summary

- Naïve Bayes modeling can improve prediction of conflicting ratings
- This model can help prioritize research
- There is still room for improvement with the model