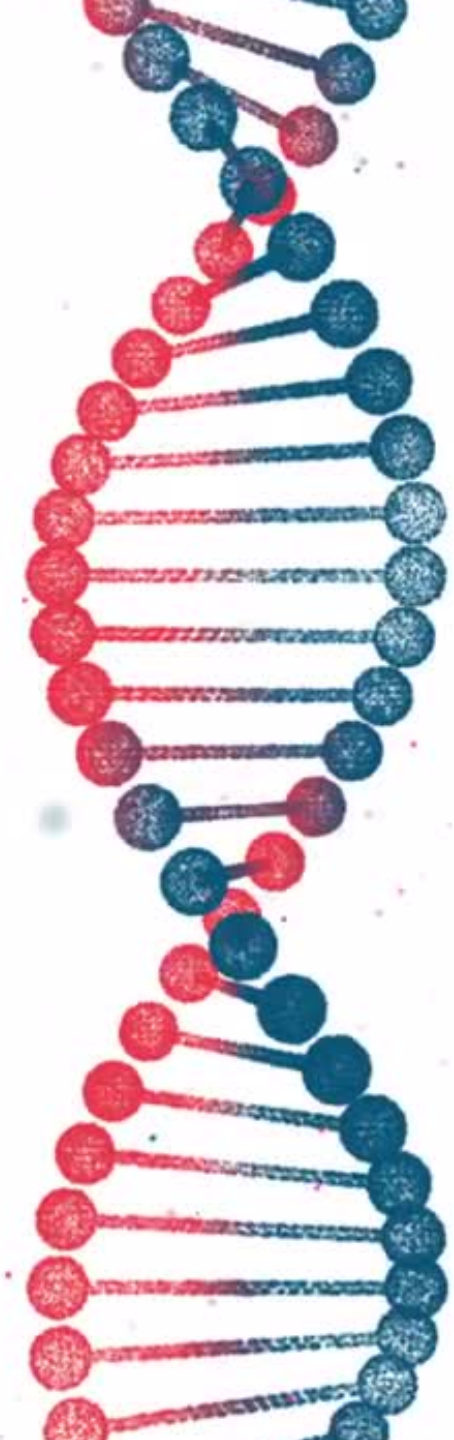
Genetic Variant Classification (Classification)

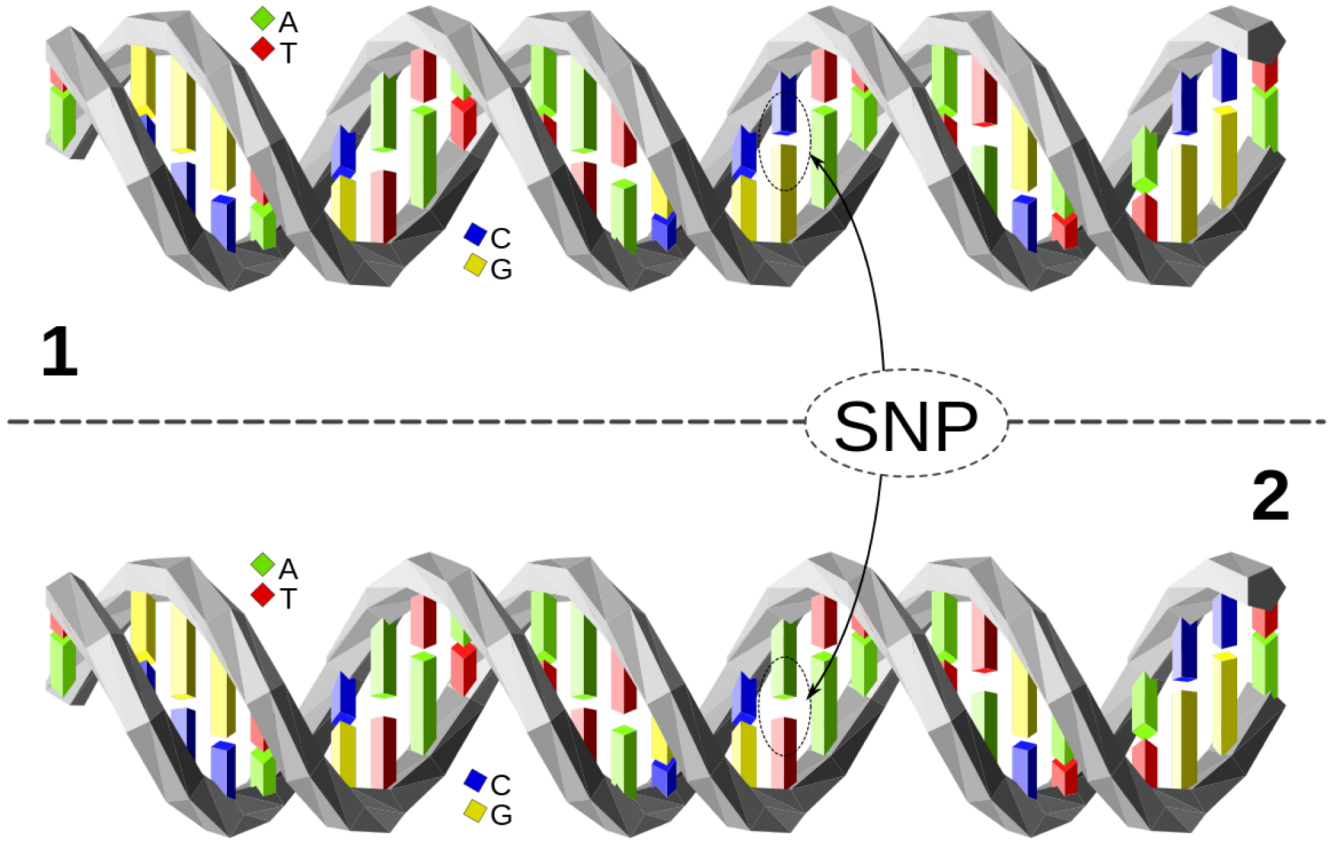
Nick Kinnaird

2/10/21



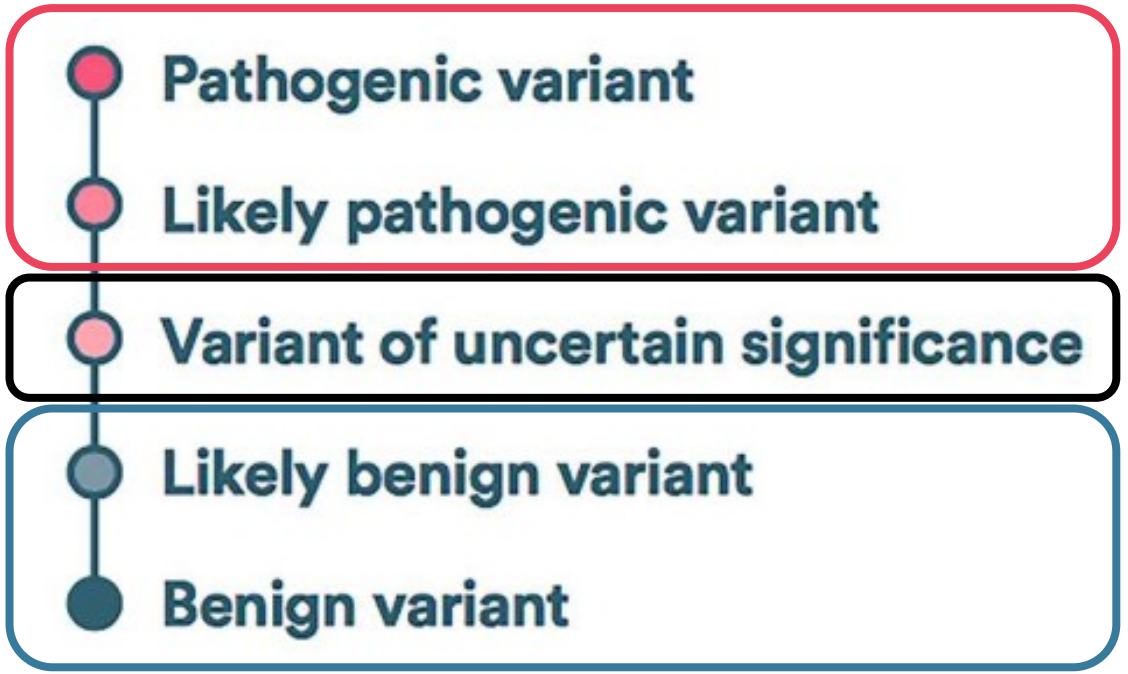
Gene Variant Primer

- We all have genes, some of which are variants
- These variants might be common (eg. SNP) or rare
- Variants might have little to no effect, or a negative effect – these can be related to various diseases



Variant Classification

- Genetic variants are classified at labs (usually manually) in one of 5 different ways in terms of clinical significance
- These are split into 3 categories
- When results from separate labs are classified as different categories, they are said to be “conflicting”



Purpose

- By identifying which gene variants are likely to be classified in a conflicting manner or not:
  - Biologists and genetic labs can better identify those variants which require further study and lab testing
  - Physicians can better plan patient treatment and respond to lab results

The Data

- Kaggle Dataset: https://www.kaggle.com/kevinarvai/clinvar-conflicting/
  - ClinVar – public archive of genetic data

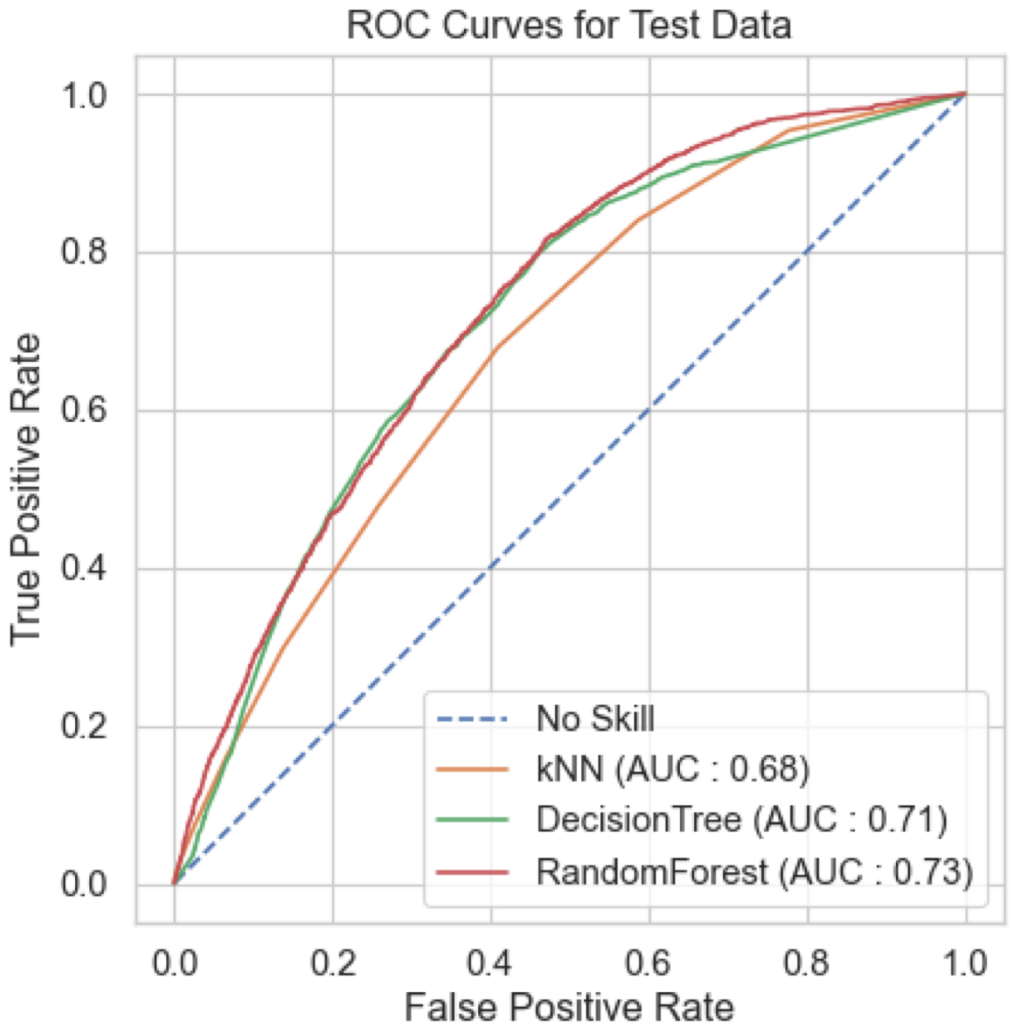
- **Prediction Target:** conflicting (1) or non-conflicting (0)

- ~60,000 entries of data
- 1:3 conflicting to non-conflicting

- Features:
  - Categorical: Name of the gene, variant type, ‘impact’ of the variant
  - Numeric: Variant frequency, ‘deleteriousness’ score, ‘loss-of-function’ score

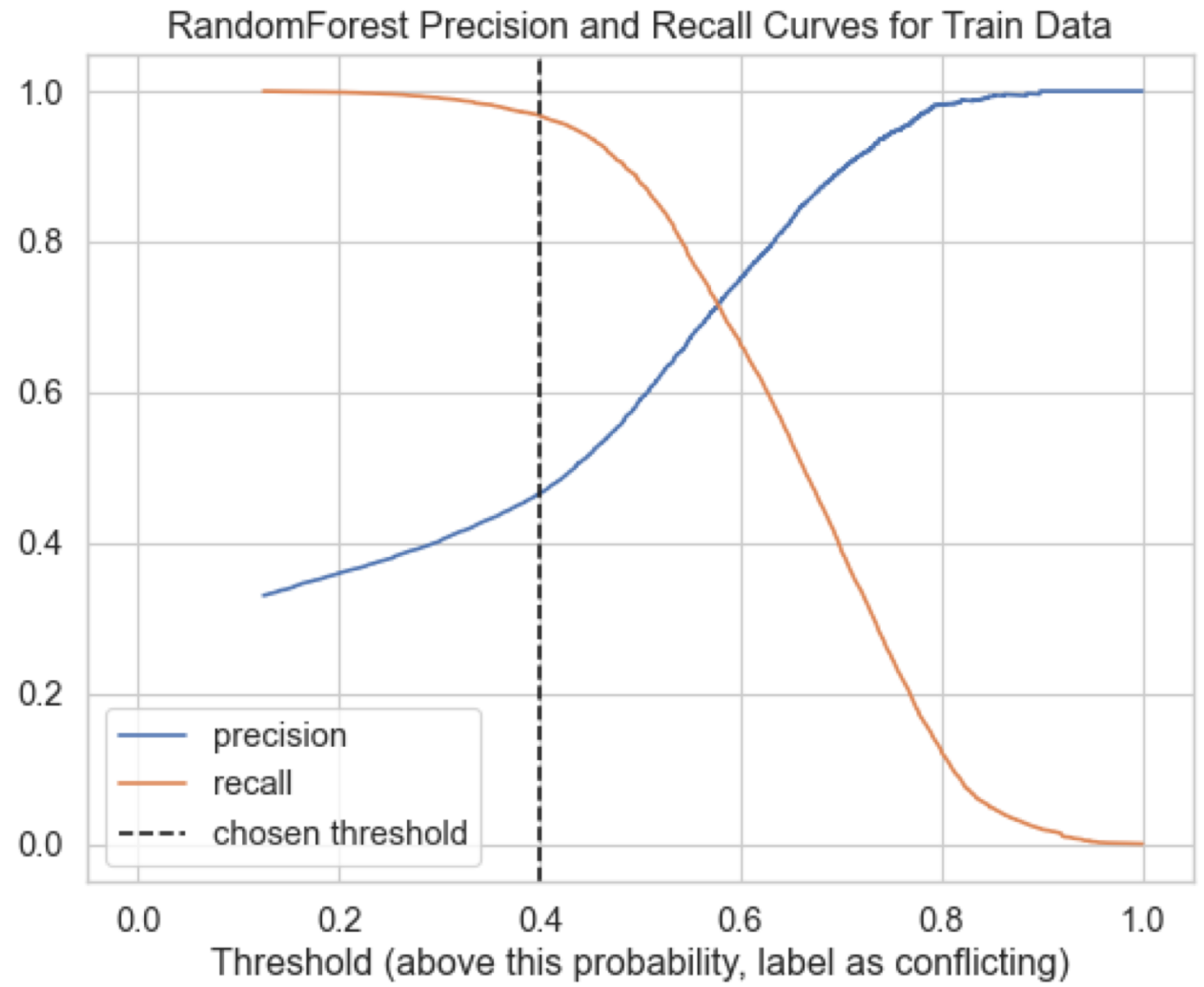
Modeling

- Tried kNN, DecisionTree, and RandomForest
- Used GridSearchCV to tune parameters
- Ultimately chose RandomForest since it performed the best (and most consistently) on the training and validation data

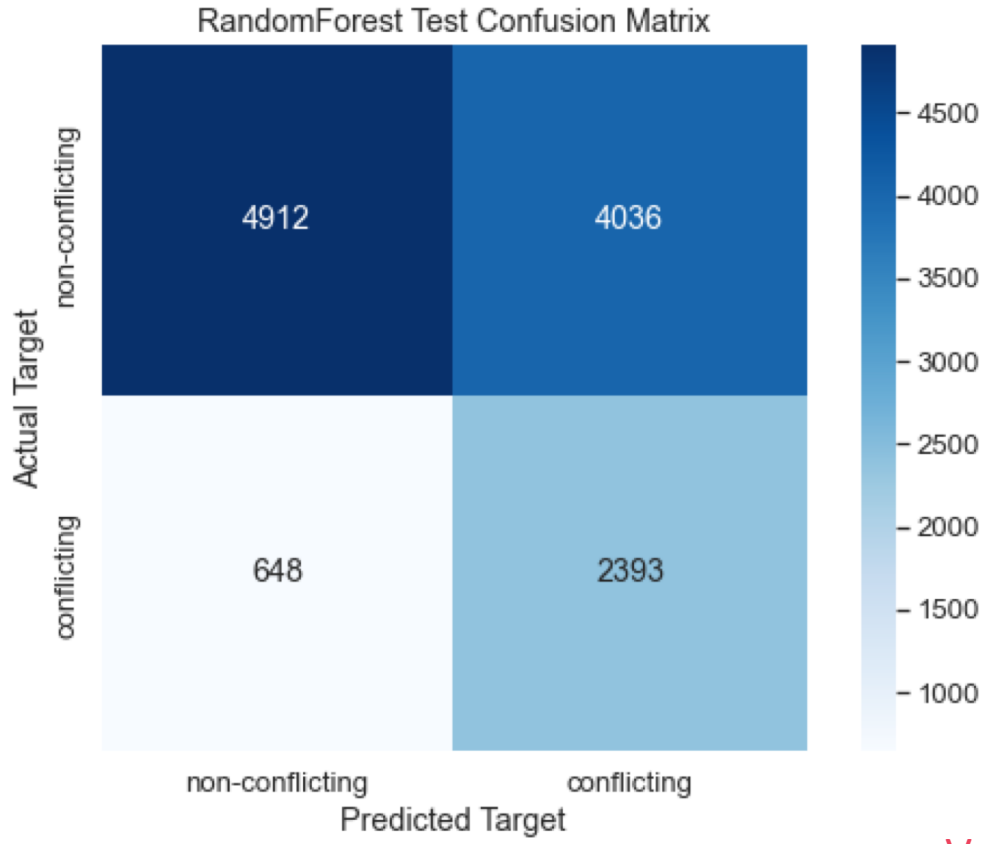


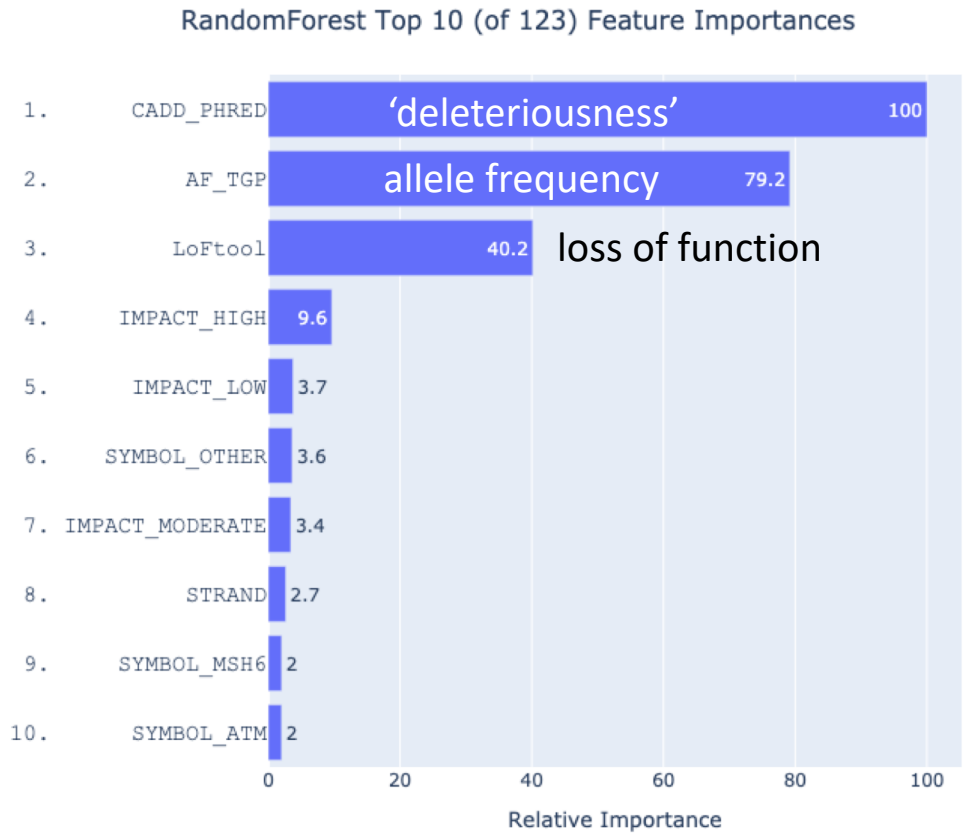
Modeling

- Optimized on recall
- More true positives at the cost of false positives which is acceptable
- Chose threshold of 0.4



Results





‘deleteriousness’

allele frequency

loss of function

Recall: 0.78

Variants which are more common or more damaging have been studied more and are more likely to be classified in a non-conflicting manner (and vice versa)

Future Work

- There was one or two omitted features which encoded some interesting information which could potentially be used

- The project could be changed to a multi-label classification problem on the pathogenicity of the variant
  - This would be a more useful application
  - (Indeed there are companies who are working on this stuff)

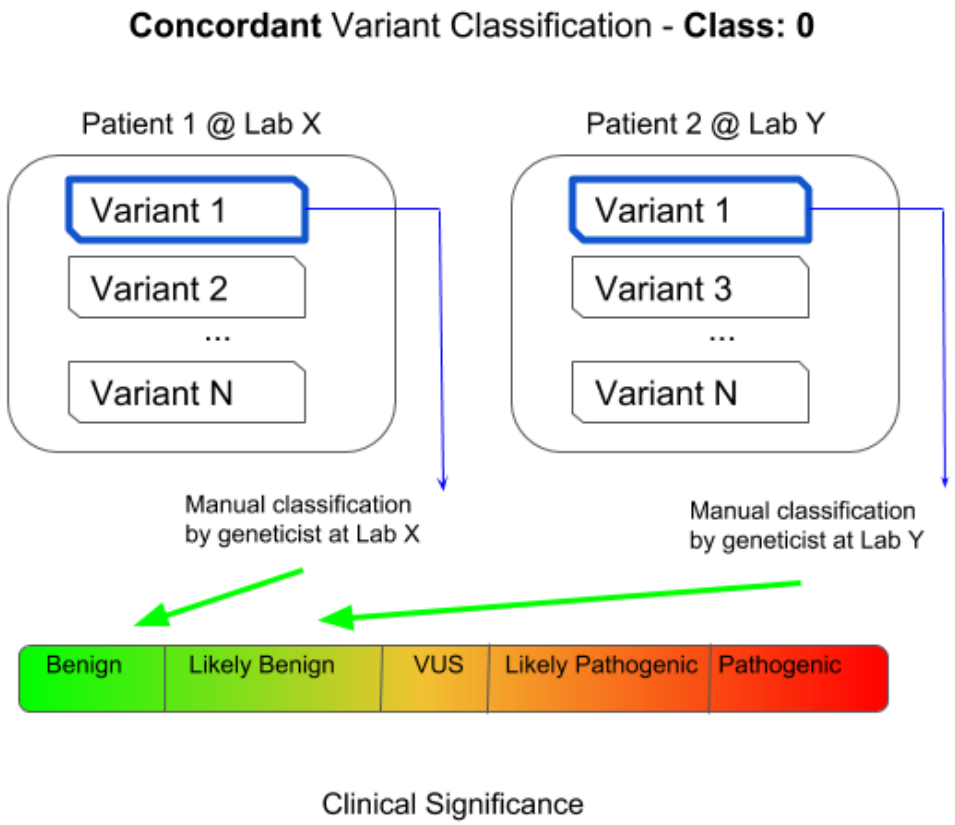
End

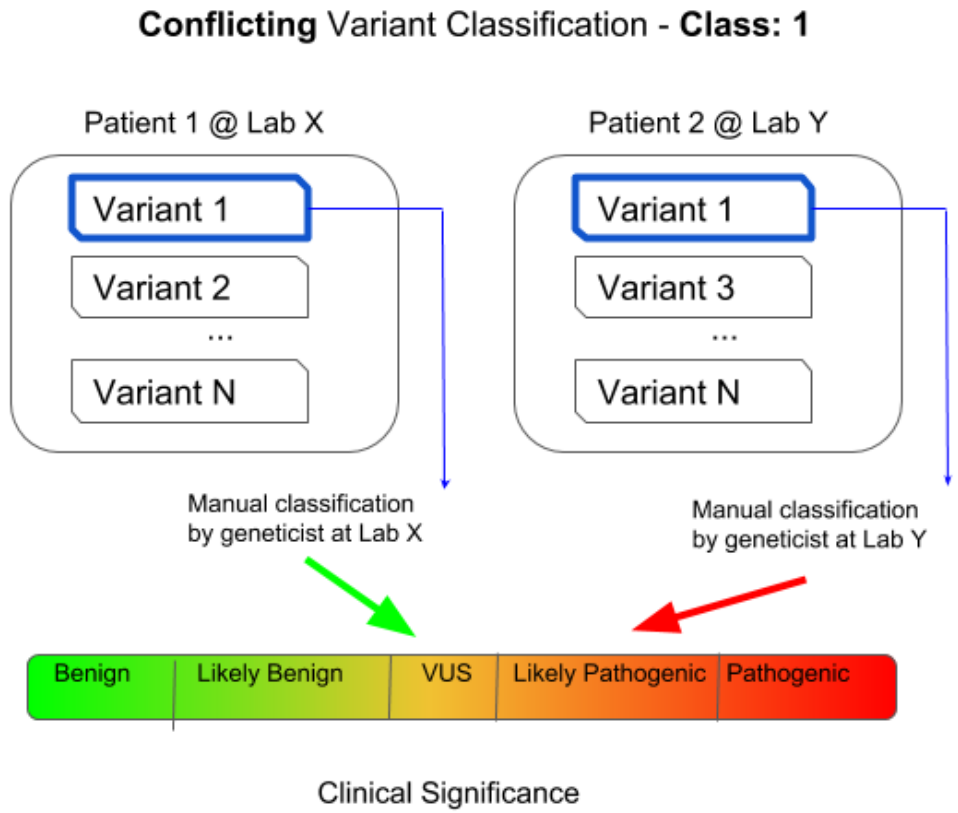
"The human genome is only on the order of a gigabyte of data, which is a tiny little database. If you take the entire living biosphere…about one petabyte…that's still very small compared with Google or the Wikipedia. And somehow mother nature manages to create…this incredibly rich environment with this amazingly small amount of data."

- Freeman Dyson

Appendix

Variant Classification

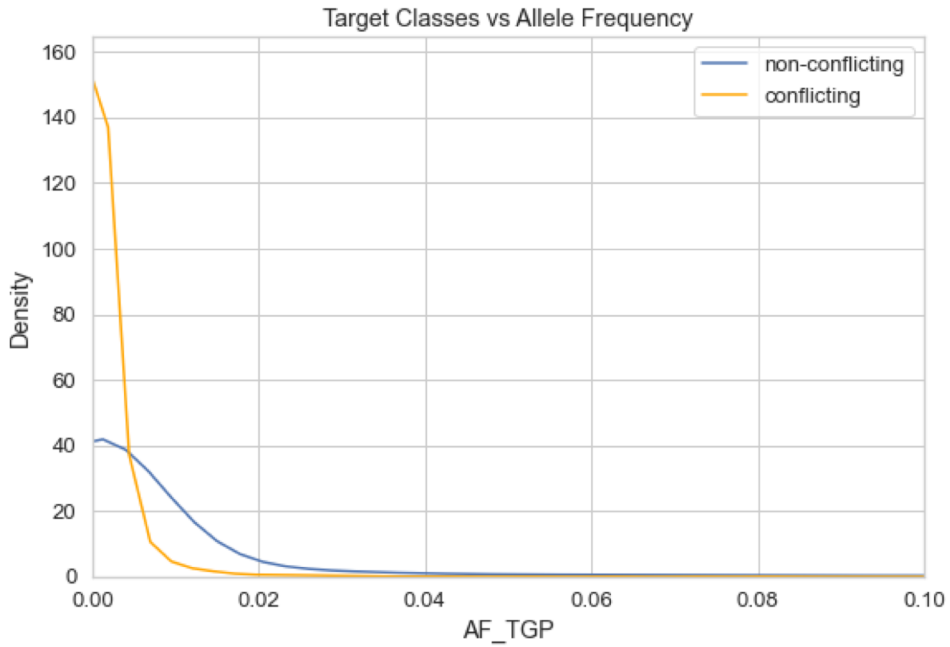


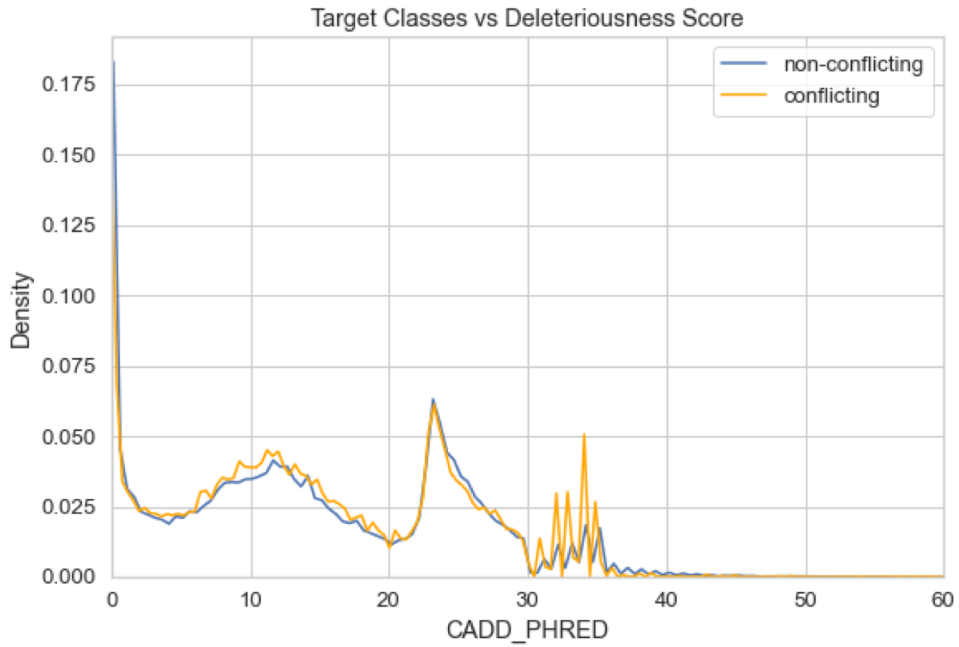


Model Parameters

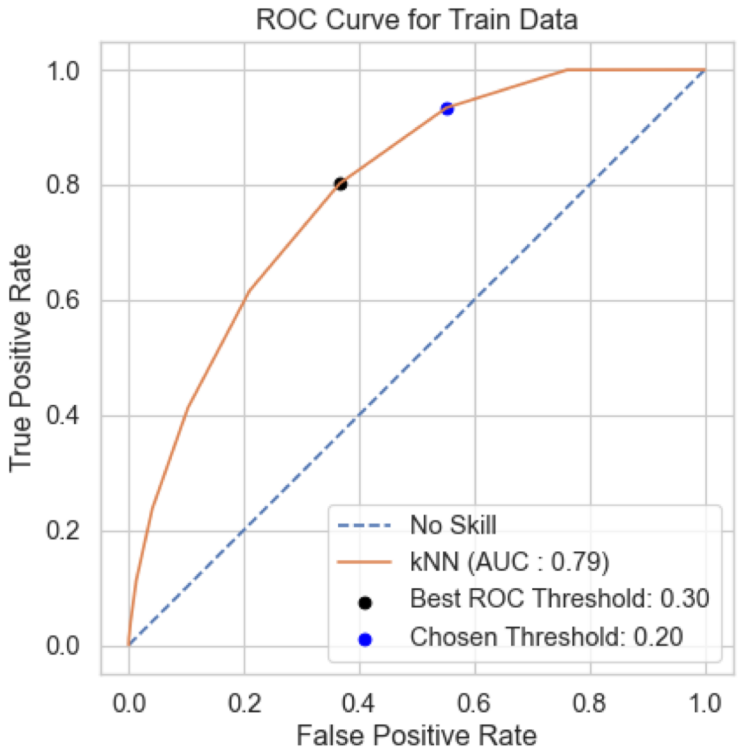
- kNN : KNeighborsClassifier(n_neighbors=10)
- DecisionTree : DecisionTreeClassifier(max_depth=25, max_features=20, random_state=121, class_weight={0:1,1:3})
- RandomForest : RandomForestClassifier(n_estimators=10, max_depth=25, max_features=20, random_state=121, class_weight={0:1,1:3})

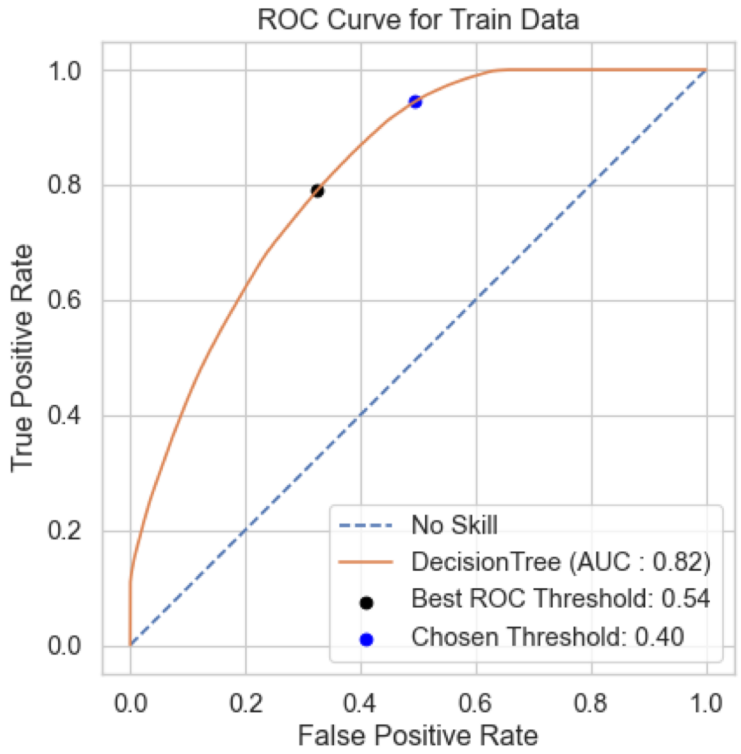
Some EDA

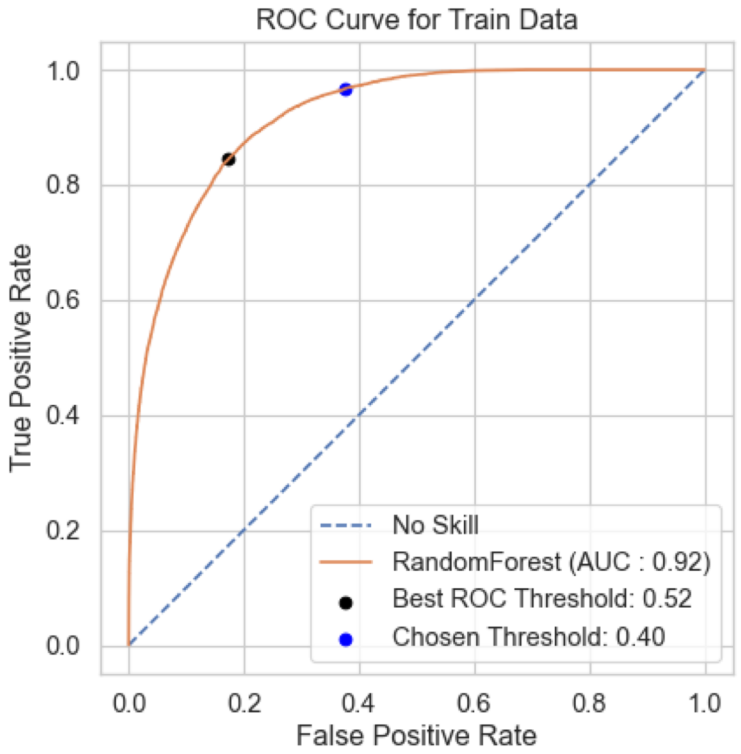

Target Classes vs Allele Frequency
non-conflicting
conflicting
Density
AF_TGP
0
20
40
60
80
100
120
140
160
0.00
0.02
0.04
0.06
0.08
0.10


Target Classes vs Deleteriousness Score
non-conflicting
conflicting
Density
CADD_PHRED
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0
10
20
30
40
50
60

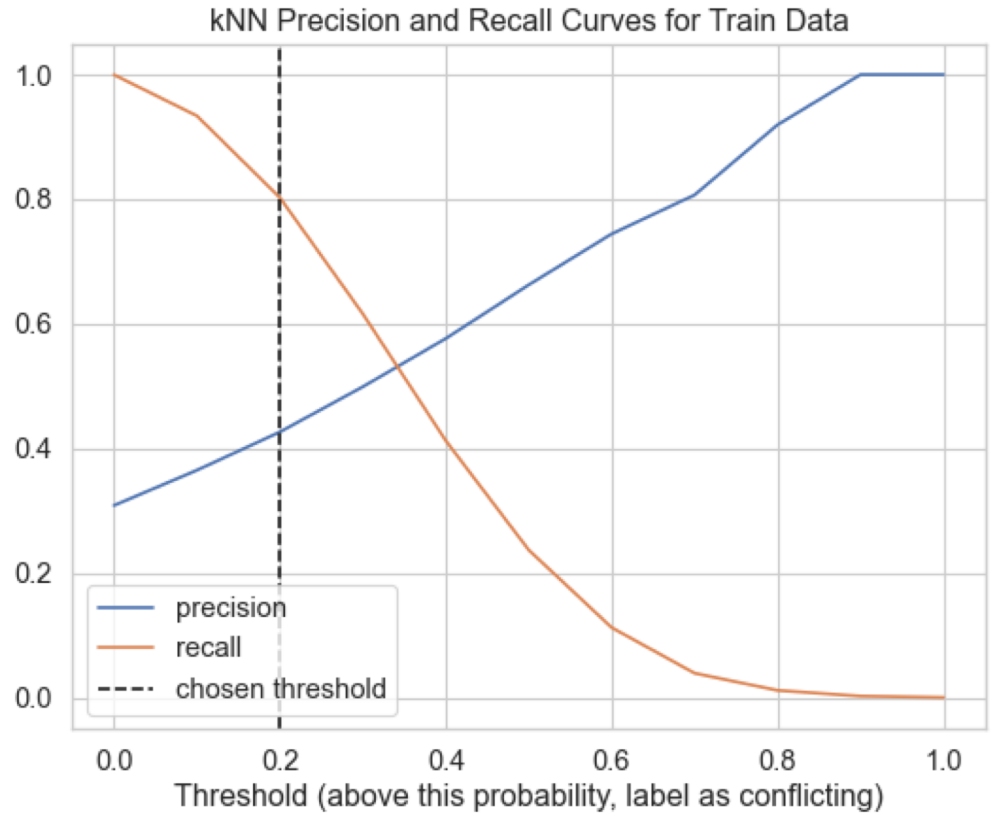
ROC Curves on Train Data

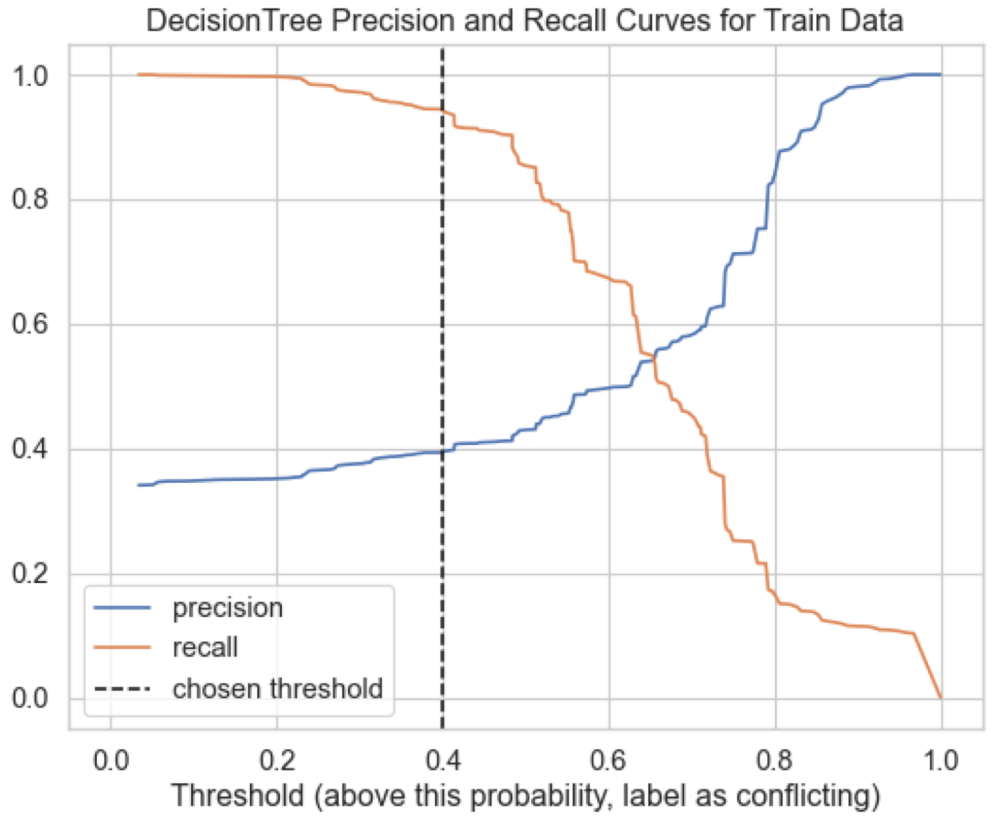




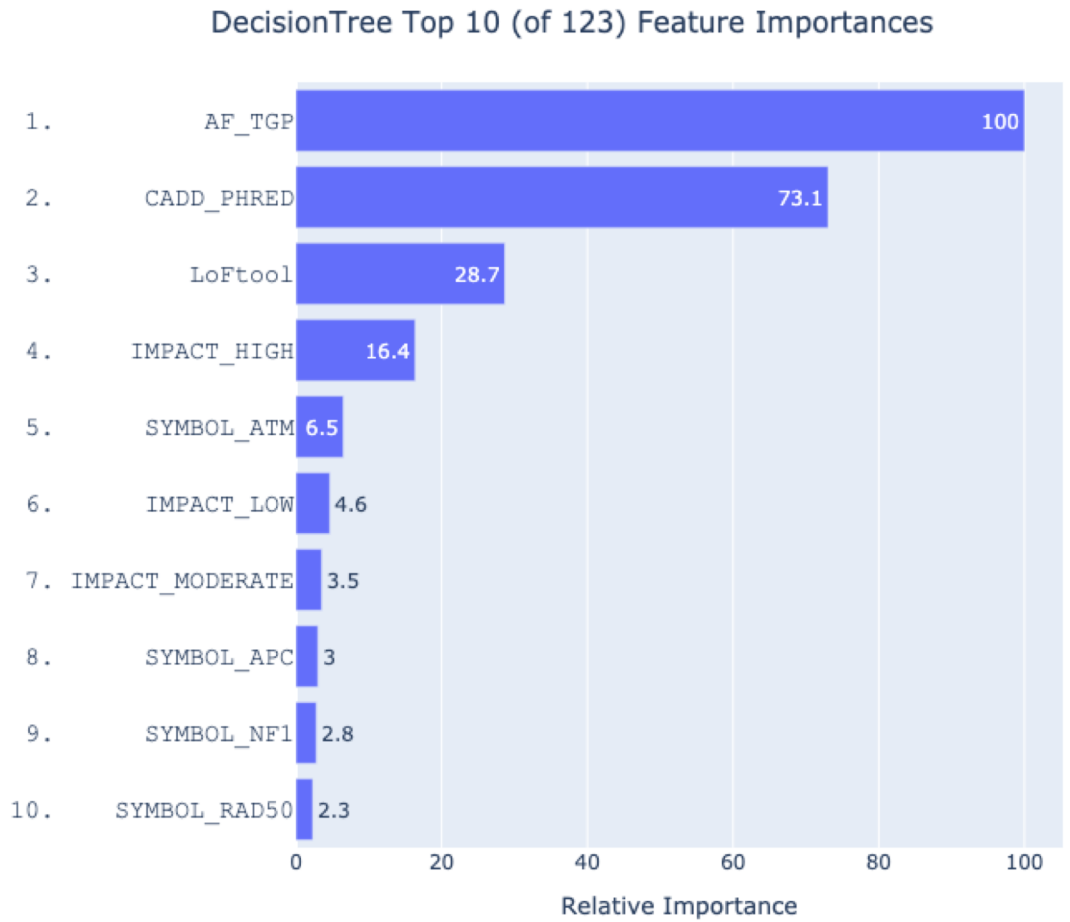


Other Precision and Recall Curves

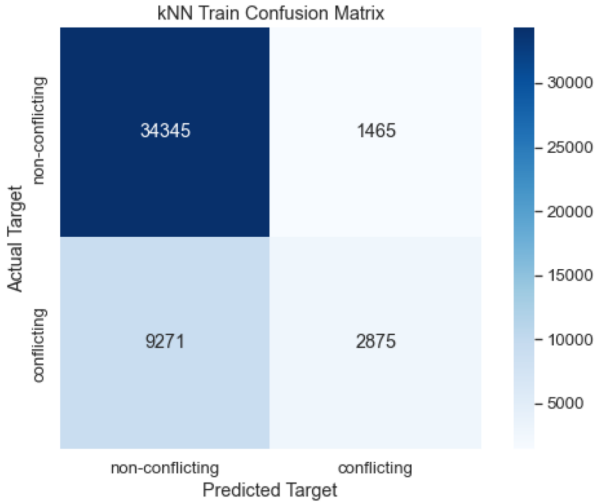


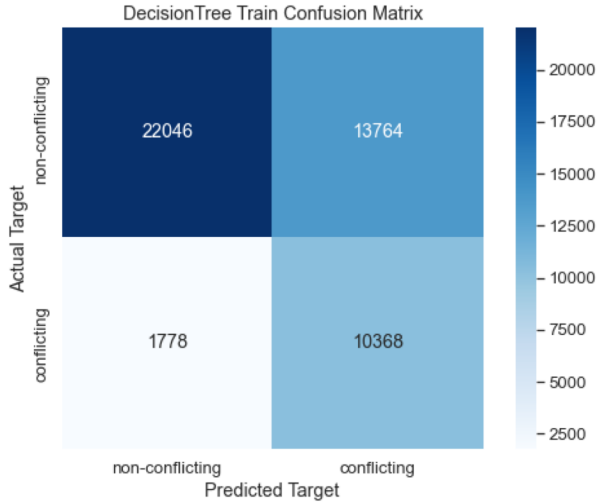


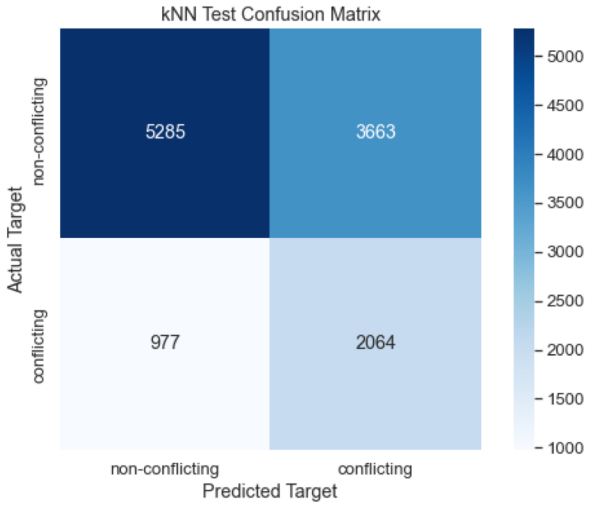
DecisionTree Feature Importances

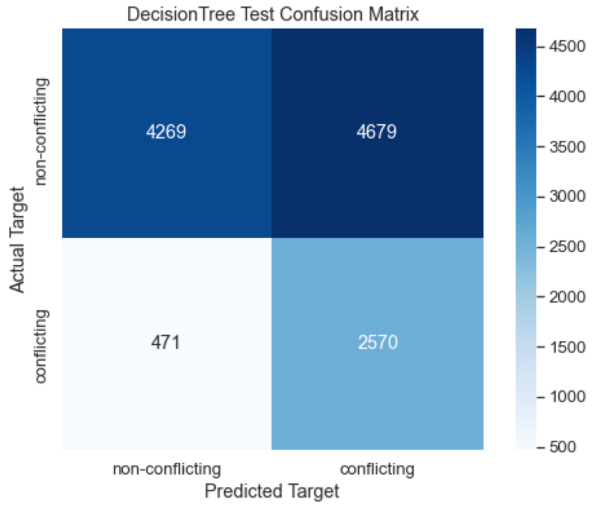


Other Confusion Matrices

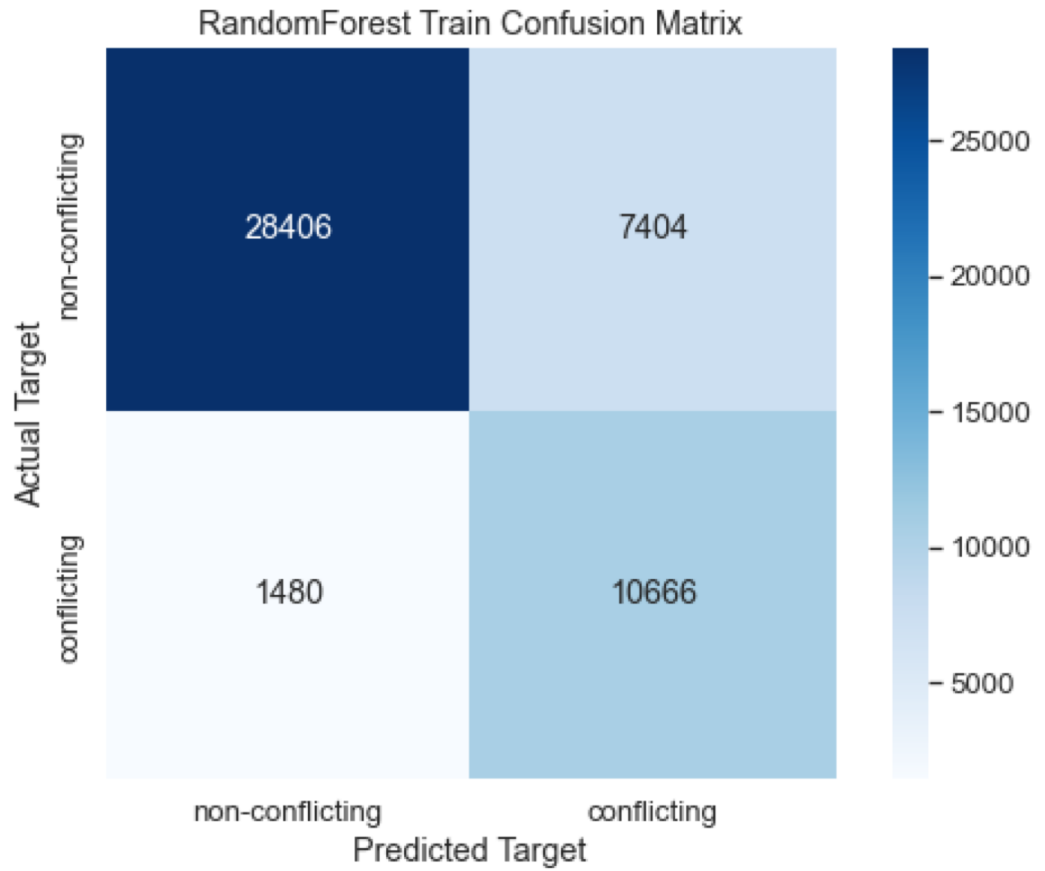








RandomForest Train Confusion Matrix



Feature Descriptions

- Categorical features
  - SYMBOL – Name of the gene
  - CLNVC – Variant type
  - IMPACT – Impact of the variant
- Numeric features
  - AF_TGP – Frequency of allele
  - CADD_PHRED – ‘Deleteriousness’ score
  - LoFtool – Loss of function score
  - Strand – Forward or backward DNA strand