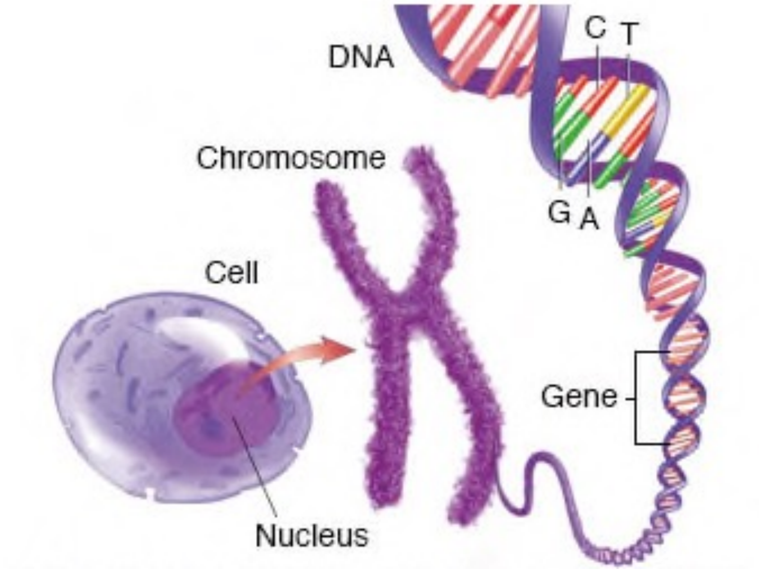
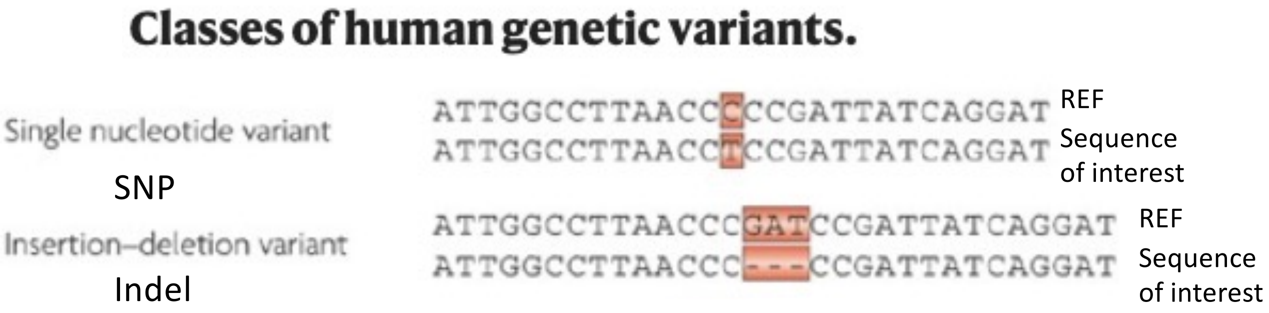
Genetic Variant Classifications

Ni-Ting Chiou

Genetic variants come from the changes of DNA sequences





Classes of human genetic variants.

Single nucleotide variant

SNP

ATTGGCCTTAACCCCCGATTATCAGGAT REF

ATTGGCCTTAACCTCCGATTATCAGGAT Sequence of interest

Insertion–deletion variant

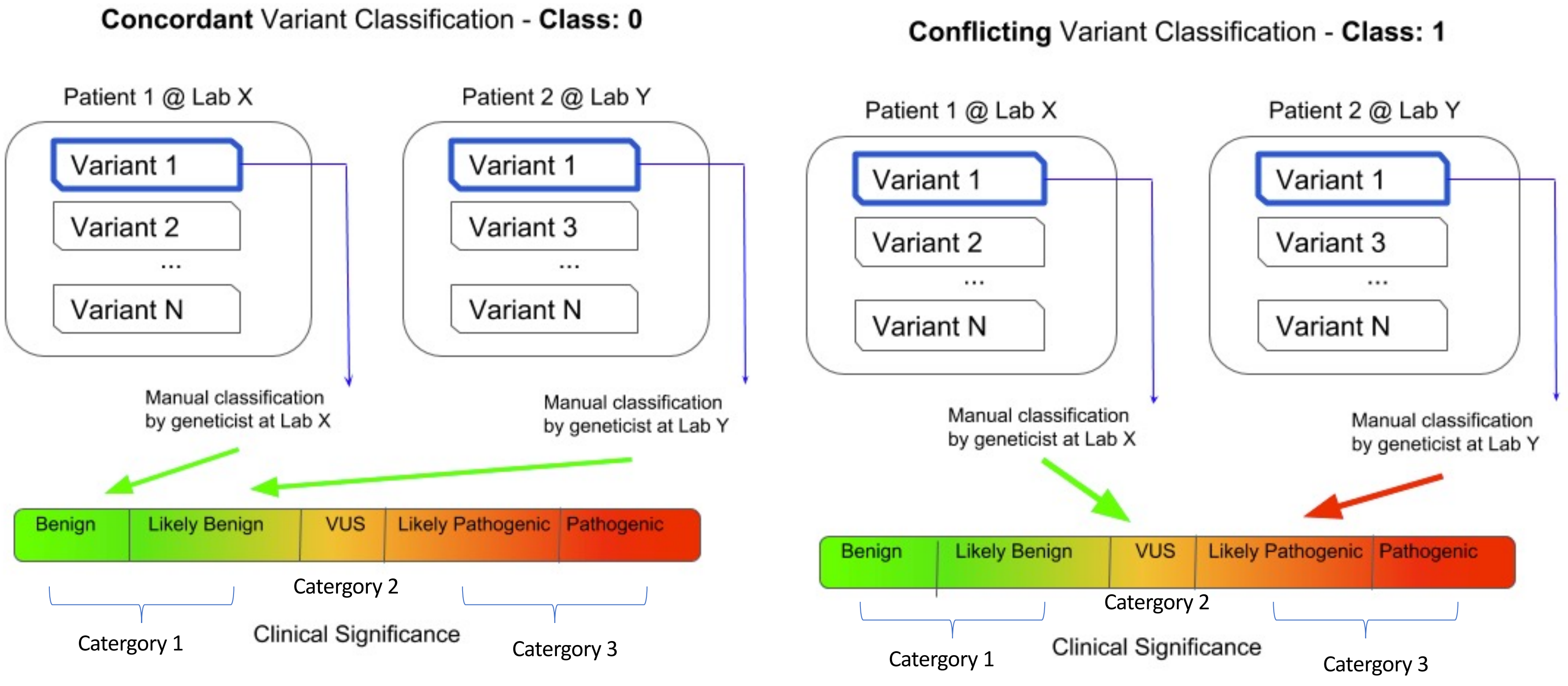
Indel

ATTGGCCTTAACCCGATCCGATTATCAGGAT REF

ATTGGCCTTAACCC---CCGATTATCAGGAT Sequence of interest

- Variants might have negative, little or no effect to diseases

Genetic variants are classified manually which resulting in conflicting classification



Concordant Variant Classification - **Class: 0**

Patient 1 @ Lab X

Variant 1

Variant 2

...

Variant N

Patient 2 @ Lab Y

Variant 1

Variant 3

...

Variant N

Manual classification by geneticist at Lab X

Manual classification by geneticist at Lab Y

Benign | Likely Benign | VUS | Likely Pathogenic | Pathogenic

Catergory 1

Catergory 2

Catergory 3

Clinical Significance

Conflicting Variant Classification - **Class: 1**

Patient 1 @ Lab X

Variant 1

Variant 2

...

Variant N

Patient 2 @ Lab Y

Variant 1

Variant 3

...

Variant N

Manual classification by geneticist at Lab X

Manual classification by geneticist at Lab Y

Benign | Likely Benign | VUS | Likely Pathogenic | Pathogenic

Catergory 1

Catergory 2

Catergory 3

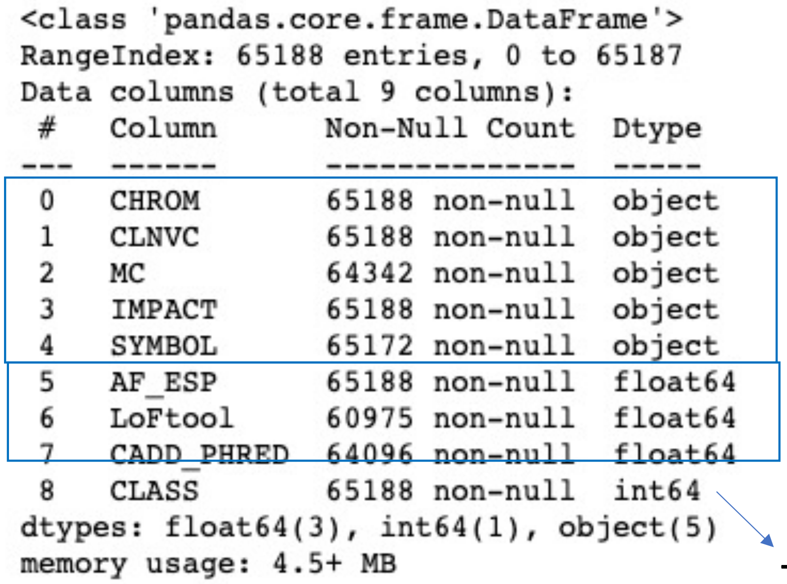
Clinical Significance

Data exploration analysis

clinvar_conflicting.csv (Kaggle)
(46 features)

Remove features

1. Redundant
2. Not correlated
3. Have > 90% nan



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 65188 entries, 0 to 65187
Data columns (total 9 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   CHROM       65188 non-null  object
 1   CLNVC       65188 non-null  object
 2   MC          64342 non-null  object
 3   IMPACT      65188 non-null  object
 4   SYMBOL      65172 non-null  object
 5   AF_ESP      65188 non-null  float64
 6   LoFtool     60975 non-null  float64
 7   CADD_PHRED  64096 non-null  float64
 8   CLASS       65188 non-null  int64
dtypes: float64(3), int64(1), object(5)
memory usage: 4.5+ MB
```

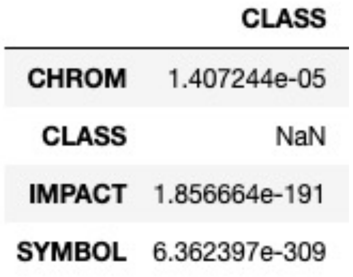
Categorical features :
CHROM- chromosome
CLNVC - Variant Type
MC - Molecular consequence
IMPACT - the impact of the variants
SYMBOL - Gene Name

Numerical features:
AF_ESP - Allele frequencies of variants
LoFtool - Loss of Function tolerance score
CADD_PHRED - Scoring the deleteriousness of the variants

Target:
class 0 (concordant variant classification)
class 1 (conflicting variant classification)

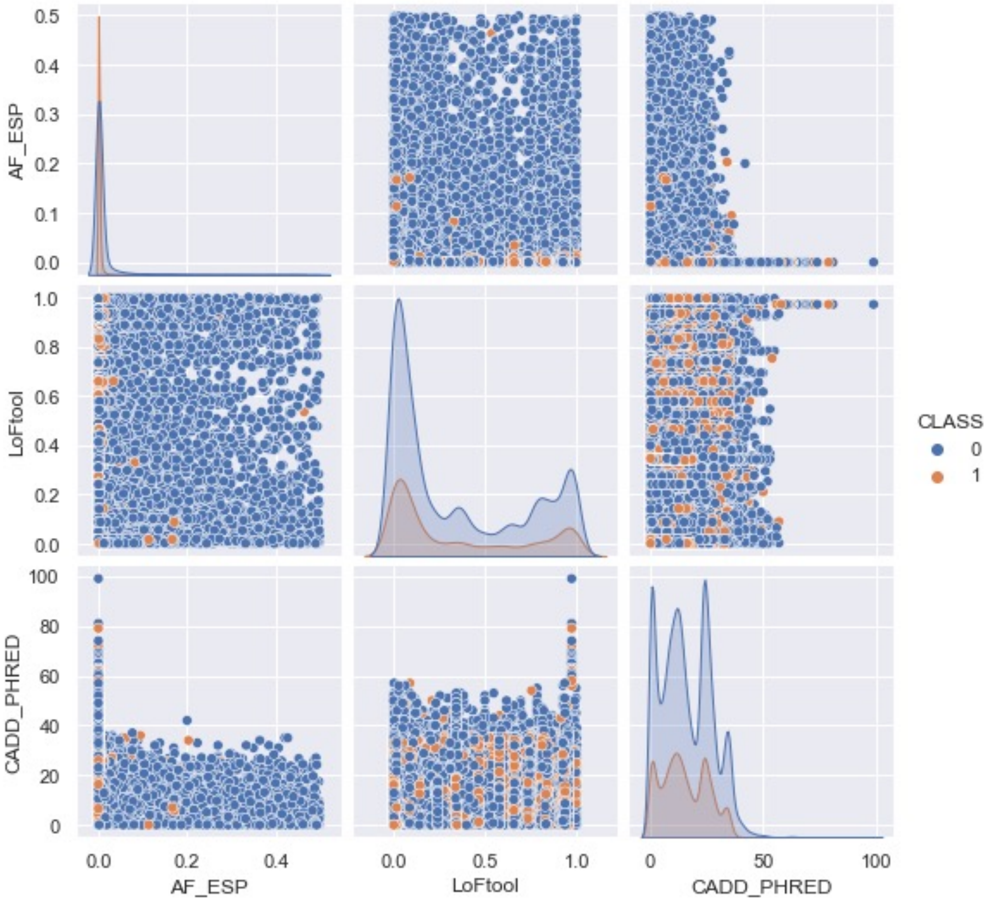
AF_ESP and SYMBOL are more distinguishable among 2 classes

Chi2 test for **categorical features** (p-value)



	CLASS
CHROM	1.407244e-05
CLASS	NaN
IMPACT	1.856664e-191
SYMBOL	6.362397e-309

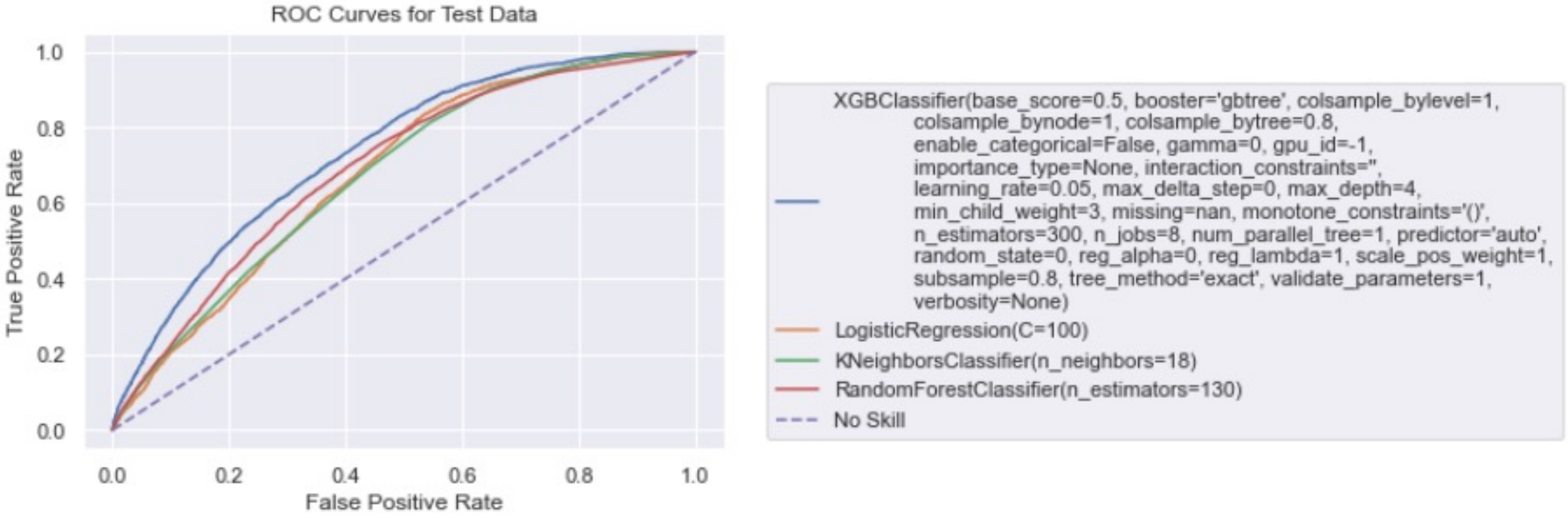
Pairplot for **numerical features**

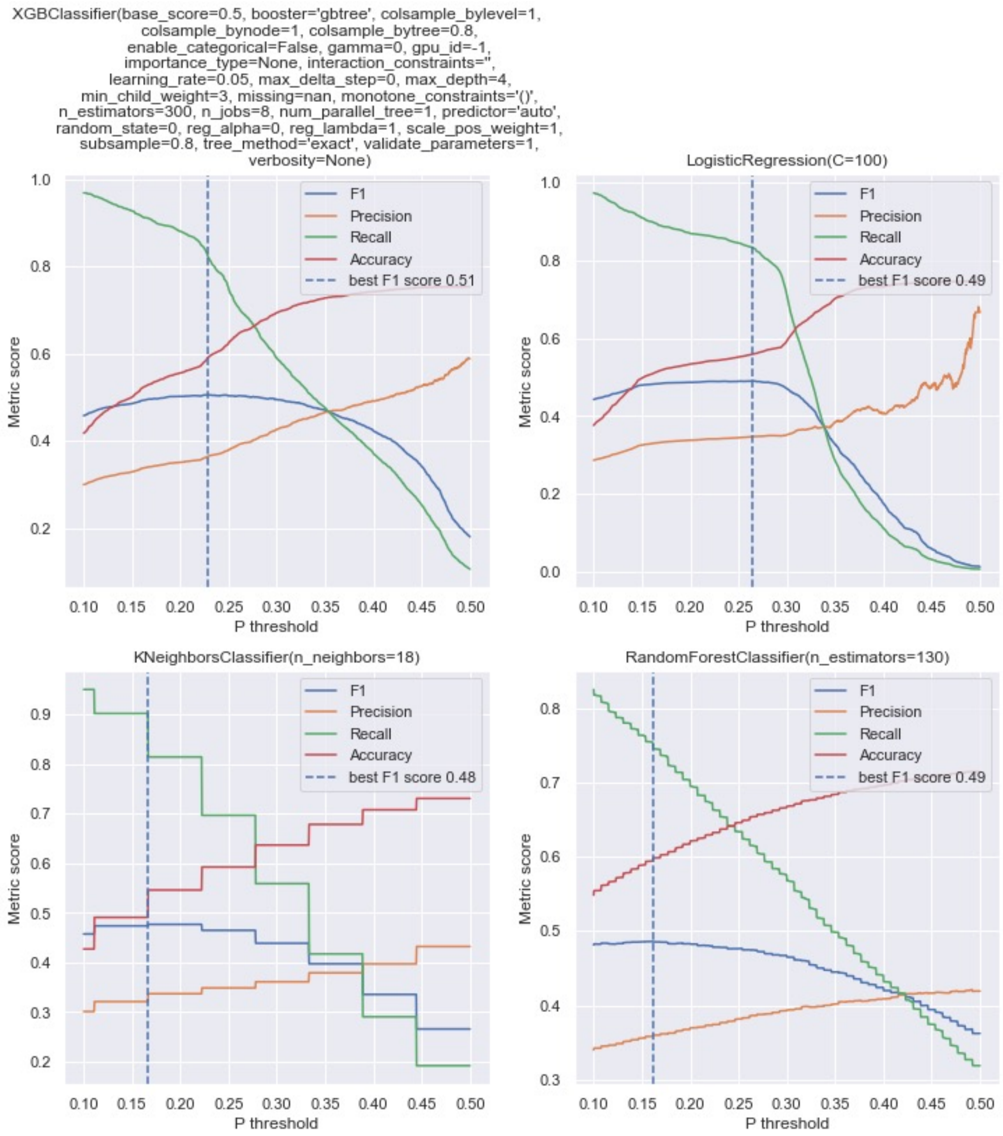


Data preprocessing

- Convert categorical features into dummies variables
- Scale the numerical data
- Class rebalance
- Model fitting (GridSearch for hyperparameter optimalization)

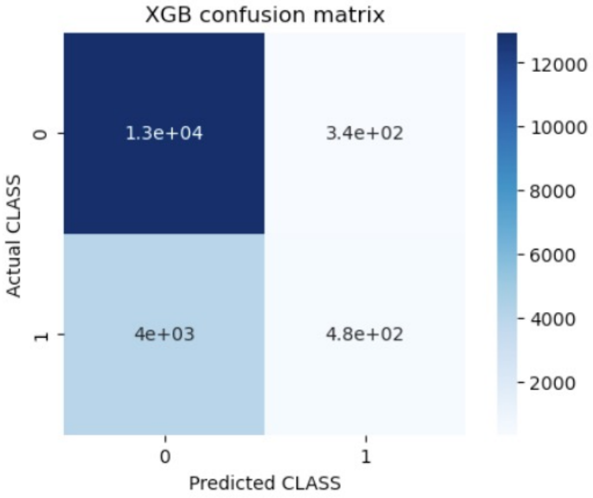
XGBoost tree has better performance for the prediction



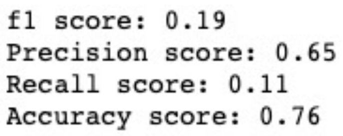


XGB model has the highest F1 scores at the P threshold of 0.22

AF (variant frequency), symbol (gene names) and MC(molecular consequences) are the important features

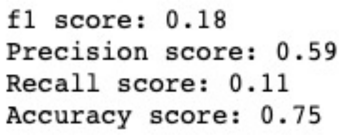


Training set

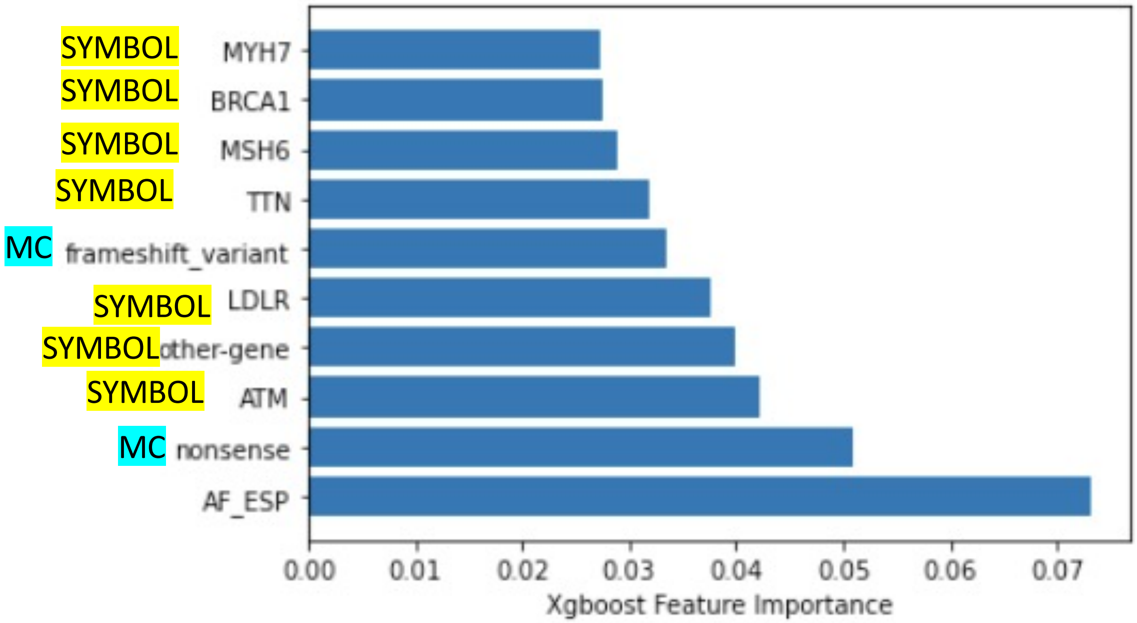


```
f1 score: 0.19
Precision score: 0.65
Recall score: 0.11
Accuracy score: 0.76
```

Test set



```
f1 score: 0.18
Precision score: 0.59
Recall score: 0.11
Accuracy score: 0.75
```



Discussion and future work

- The variants with the low allele frequency (AF), do not have the known deleterious molecular consequence (MC) and located at the cancer genes (SYMBOL) tend to have the conflicting classification.

- The variants within the conflicting classification can be compared with the cancer variant databases to be classified better.