

Genetic Variant Classifications

Ni-Ting Chiou

Genetic variants come from the changes of DNA sequences

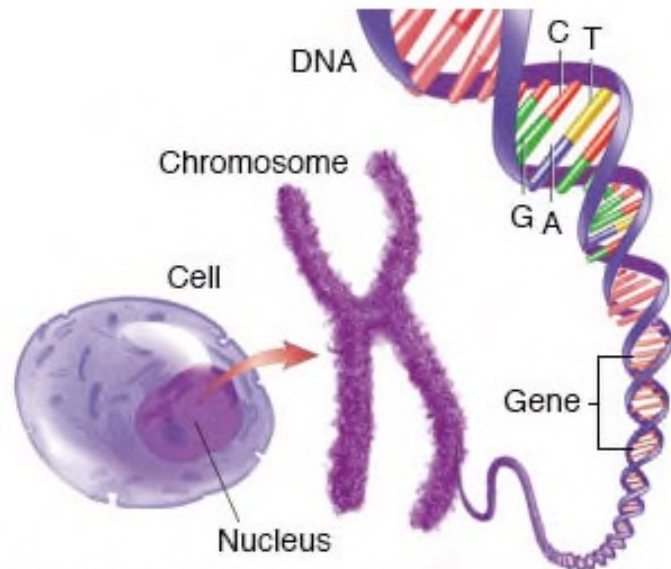


Figure 1: Classes of human genetic variants.

Single nucleotide variant

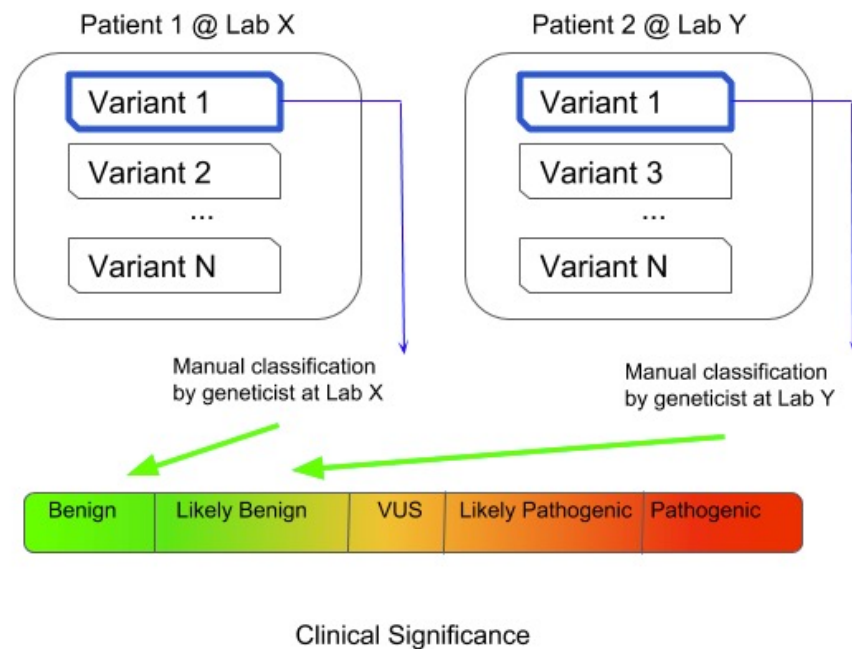
ATTGGCCTTAACCCCGATTATCAGGAT
ATTGGCCTTAACCCGATTATCAGGAT

Insertion-deletion variant

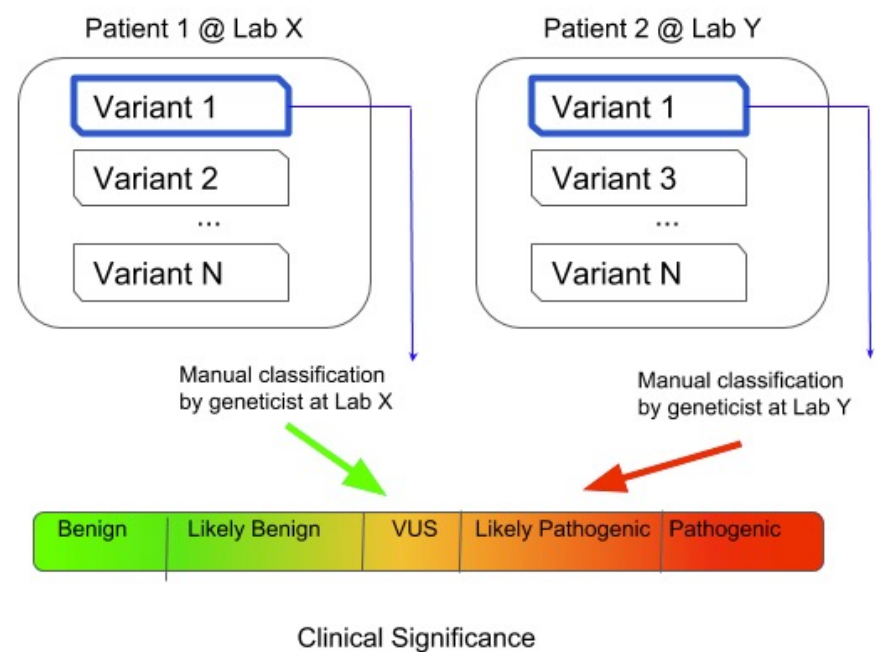
ATTGGCCTTAACCCGATTATCAGGAT
ATTGGCCTTAACCC---CCGATTATCAGGAT

Genetic variants are classified manually which resulting in conflicting classification

Concordant Variant Classification - Class: 0



Conflicting Variant Classification - Class: 1



Data exploration analysis

clinvar_conflicting.var (Kaggle)
(46 features)

Remove features

1. Redundant
2. Not correlated

<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

CLNVC - Variant Type

MC - Molecular consequence

IMPACT - the impact of the variants

SYMBOL - Gene Name

AF_ESP - Allele frequencies

LoFtool - Loss of Function tolerance score

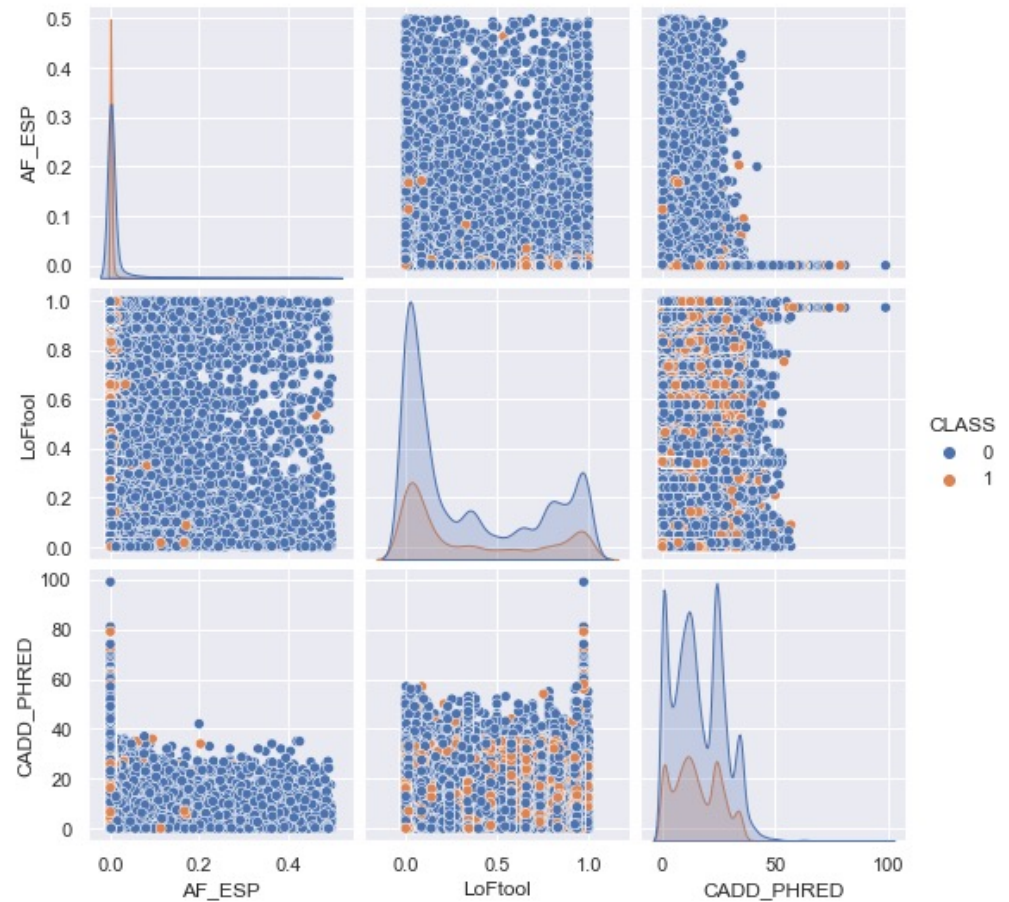
CADD_PHRED - Scoring the deleteriousness of the variants

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

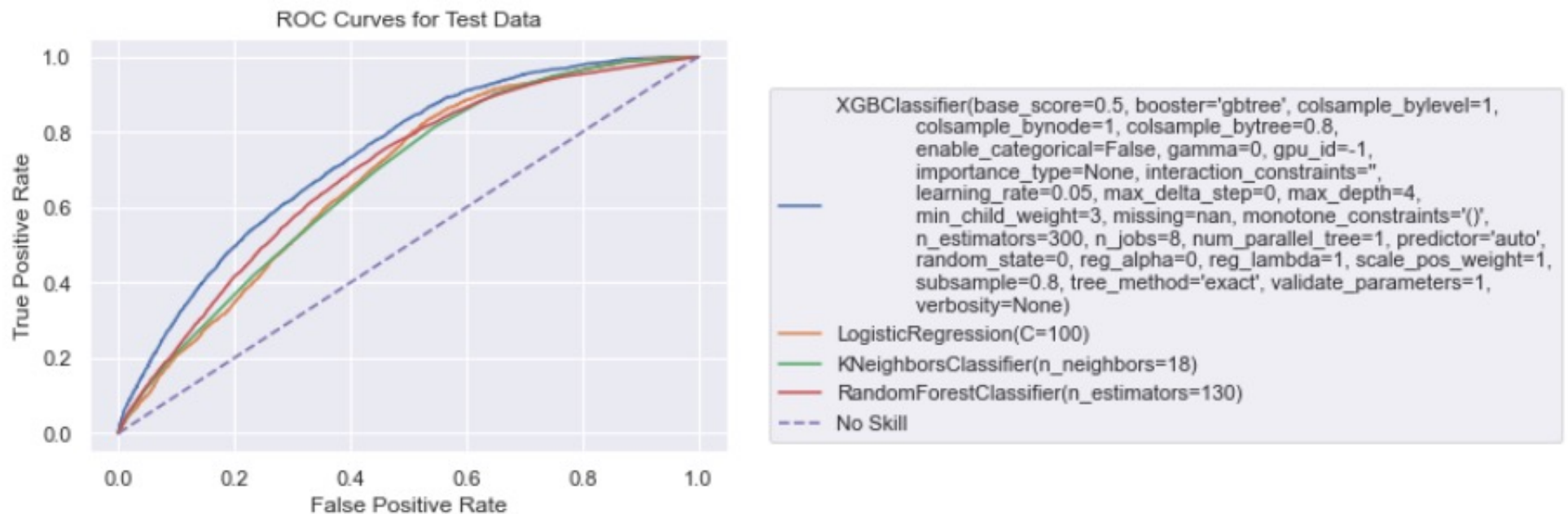
Pairplot for numerical features



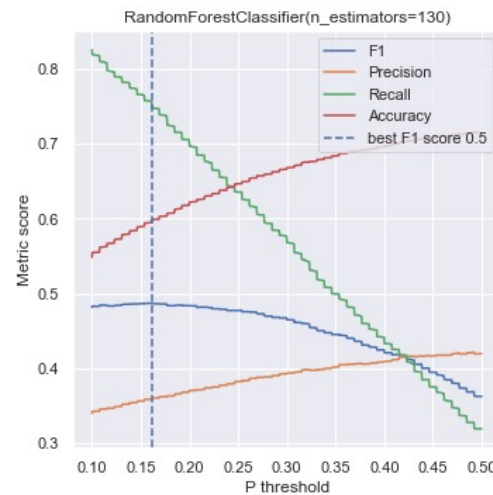
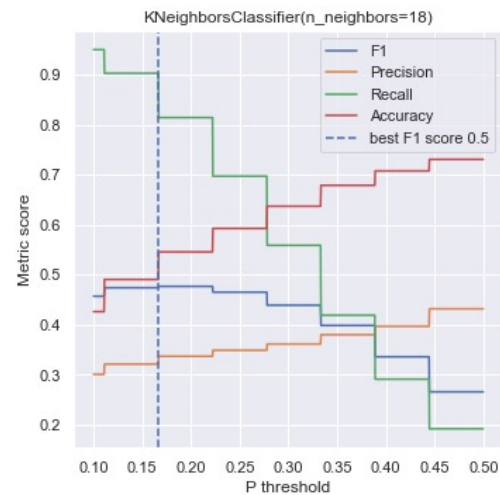
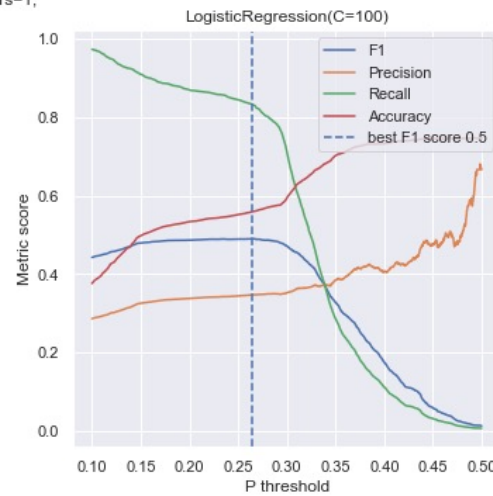
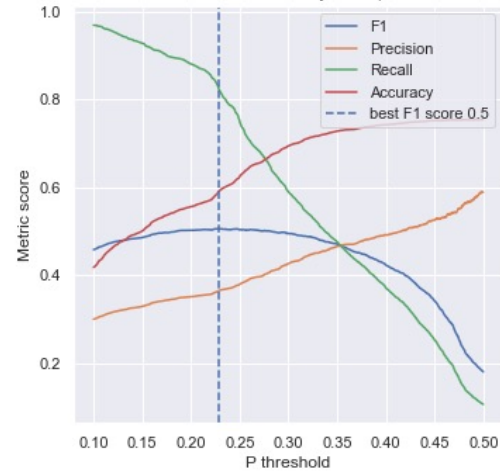
Data preprocessing

- Convert categorical feature into dummies variables
- Scale the numerical data
- Class rebalance
- Model fitting

XGBoost tree has better performance for the prediction

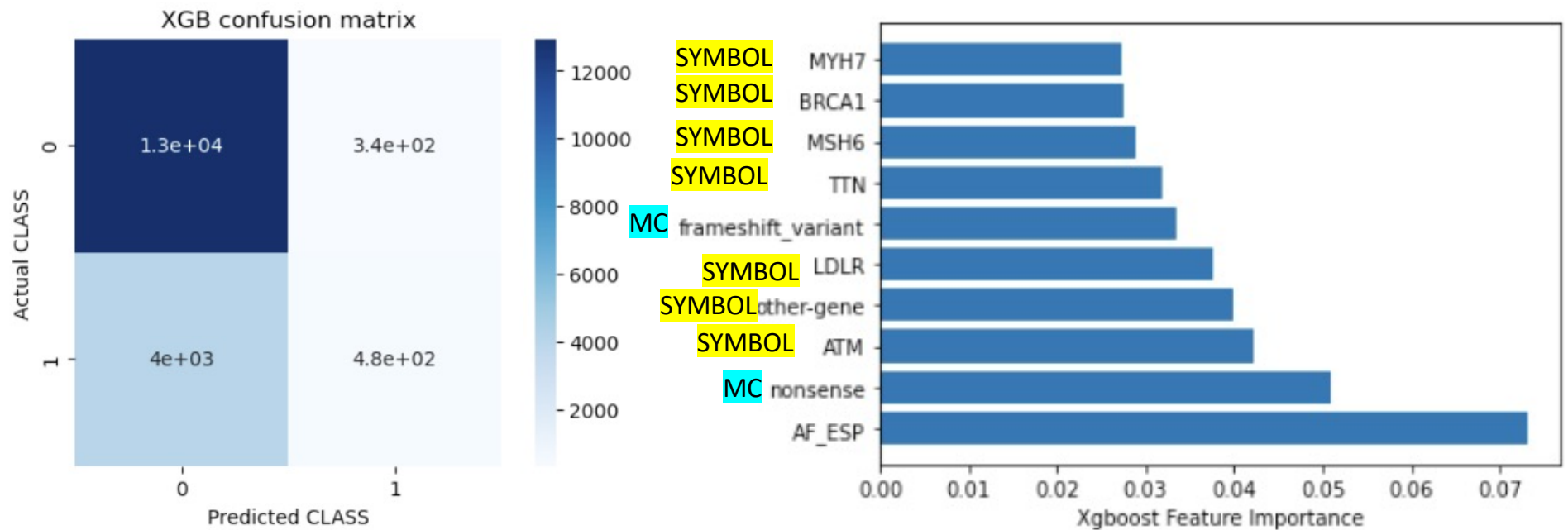


```
XGBClassifier(base_score=0.5, booster='gbtree', colsample_bylevel=1,
              colsample_bynode=1, colsample_bytree=0.8,
              enable_categorical=False, gamma=0, gpu_id=-1,
              importance_type=None, interaction_constraints='',
              learning_rate=0.05, max_delta_step=0, max_depth=4,
              min_child_weight=3, missing=nan, monotone_constraints=()),
              n_estimators=300, n_jobs=8, num_parallel_tree=1, predictor='auto',
              random_state=0, reg_alpha=0, reg_lambda=1, scale_pos_weight=1,
              subsample=0.8, tree_method='exact', validate_parameters=1,
              verbosity=None)
```



All 4 models
have similar
F1 scores

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



Questions