METIS®

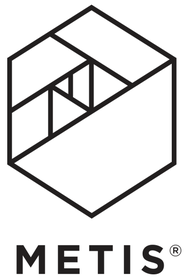
Classification on Heart Disease Indicators

classification model for identifying high-risk patients

project for Metis EDA Bootcamp

by Krystian Krystkowiak, 2022

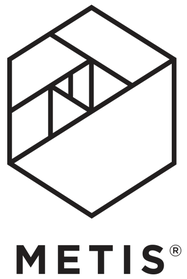
Introduction



- Early heart disease identification may not only be desired by doctors and medical institutions
- Also for others: insurances, medical apps, fitness or nutritionist, individuals conscious about health

- GOAL: Creating model that can help to detect heart disease and raise red flag during initial questionnaire.

Methodology

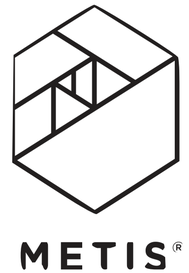


- Data from Behavioural Risk Factor Surveillance System, conducts annual telephone surveys to gather data on the **health status of U.S. residents** (initially cleaned by Kamil Pytlak at Kaggle).

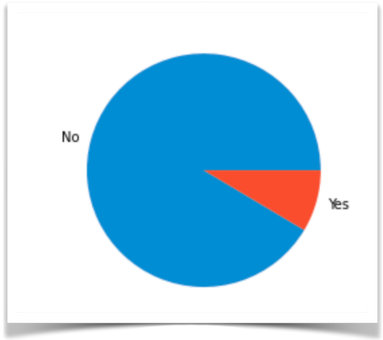
- **319k rows** of data, **19 columns**

- Features: **HeartDisease (target)**, BMI, Smoking, AlcoholDrinking, Stroke, PhysicalHealth, MentalHealth, DiffWalking, Sex, AgeCategory, Race, Diabetic, PhysicalActivity, GenHealth, SleepTime, Asthma, KidneyDisease, SkinCancer

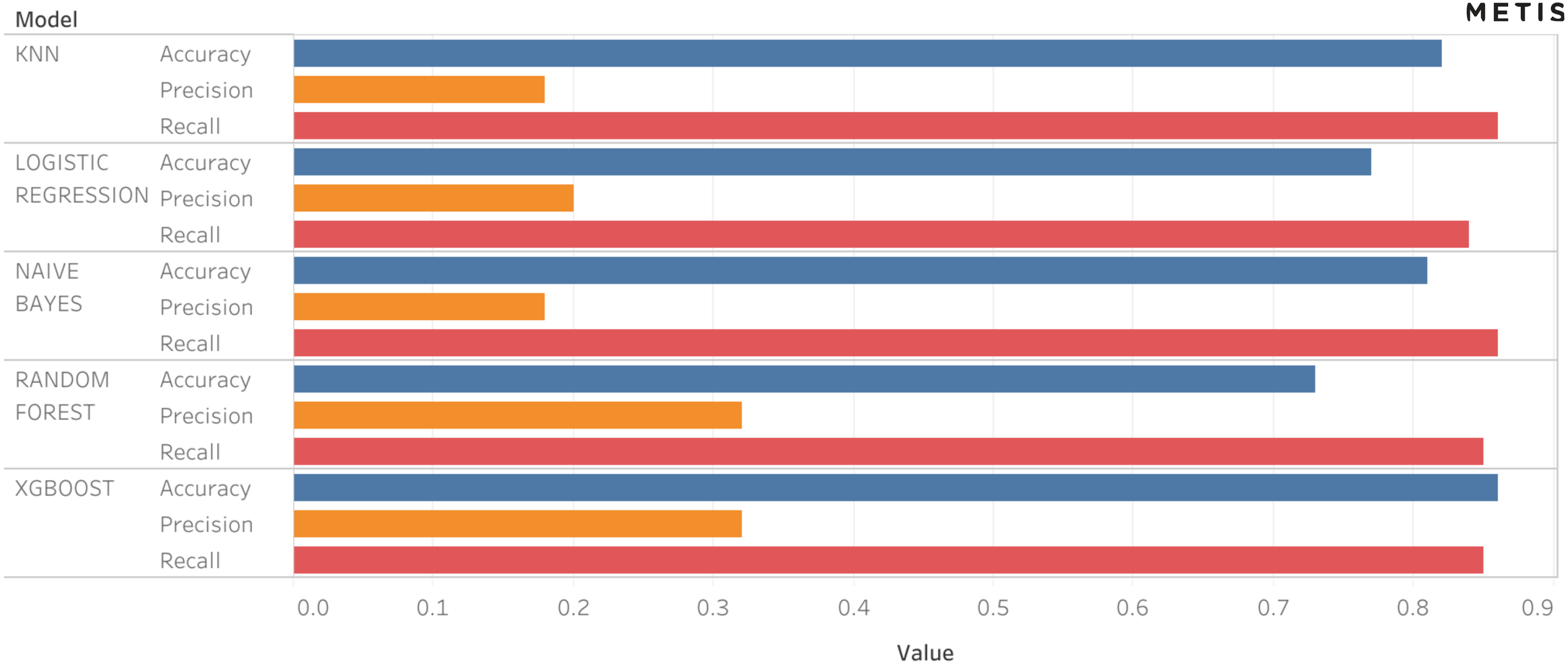
Methodology



- Important problem, **85% recall** score, precision score as second priority
- 8% of the target is positive - **class imbalance** - undersampling, class weights and threshold adjustment
- Categorical variables into **dummies**
- Data divided to **train/validation/test**
- **Tuning** with RandomizedSearchCV

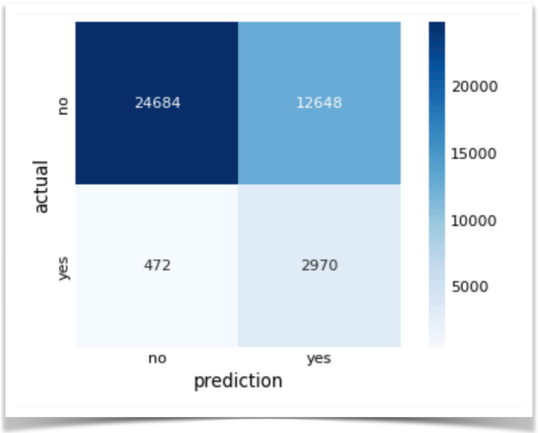


Results

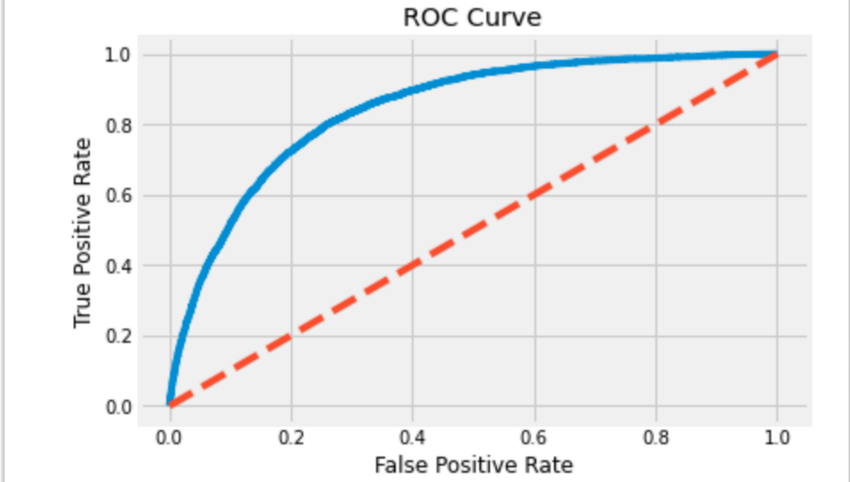

Model
KNN
Accuracy
Precision
Recall
LOGISTIC REGRESSION
Accuracy
Precision
Recall
NAIVE BAYES
Accuracy
Precision
Recall
RANDOM FOREST
Accuracy
Precision
Recall
XGBOOST
Accuracy
Precision
Recall
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Value

Results

- **XGBoost**

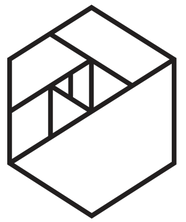



```
model: XGBClassifier(base_score=0.5, booster='gbtree', colsample_bylevel=1,
              colsample_bynode=1, colsample_bytree=1, enable_categorical=False,
              eval_metric='rmse', gamma=1.3, gpu_id=-1, importance_type=None,
              interaction_constraints='', learning_rate=0.04, max_delta_step=0,
              max_depth=6, min_child_weight=2, missing=nan,
              monotone_constraints='()', n_estimators=850, n_jobs=6,
              num_parallel_tree=1, predictor='auto', random_state=0,
              reg_alpha=0, reg_lambda=1, scale_pos_weight=10, subsample=0.5,
              tree_method='exact', use_label_encoder=False,
              validate_parameters=1, verbosity=None)
 accuracy on training set: 0.6510392883116004
 accuracy on validation set: 0.49931328787953105
 accuracy on test set: 0.8552637064832187
 precission: 0.1901651940069151
 precission on test set: 0.19358682699599178
 recall: 0.862870424171993
 recall on test set: 0.8465182378019895
 F1: 0.3116474291710388
 F1 on test set: 0.31511197319696704
```



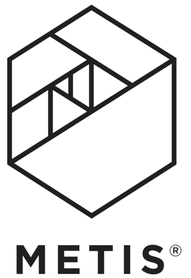
- Important features: **BMI**, **DiffWalking**, **Age**Category_80 or older, **Stroke**, AgeCategory_70-74, Diabetic_Yes, GenHealth_Poor, AgeCategory_65-69

Conclusions/Recommendations



- **XGBoost** model
- Systematic preventive medical **examination** to decrease chance of heart decease
- **Healthy** life style
- Constructed **models are ready** base for similar search

Future Work



- Deeper **EDA**, searching for feature relations
- Fine tuning of **XGBoost**
- **More data** (next year)

Thank you!

Questions?

project for Metis EDA Bootcamp

by Krystian Krystkowiak, 2022