SMOKER DETECTOR

Classification Modeling for Smoking Status Assessment in Insurance:
Leveraging Health Data for Risk Evaluation

HEALTHIER, LONGER, BETTER LIVES

BY DEWA DWI AL-MATIN

Hai! My name is

DEWA DWI AL-MATIN

I am a Data Scientist at

OBJECTIVE:

- Develop a robust classification model to predict smoking status based on health report data.
- Accurately identify smokers to refine risk assessment processes and offer precise, fair premiums based on health profiles.

IMPACT:

- Enhance underwriting practices to better manage risks associated with smoking-related health issues.

ALIGNMENT:

- This initiative supports our commitment to providing personalized insurance solutions and promoting healthier lifestyles.

DATASET INFORMATION

- ID : index
- gender
- age : 5-years gap
- height(cm)
- weight(kg)
- waist(cm) : Waist circumference
- eyesight(left)
- eyesight(right)
- hearing(left)
- hearing(right)
- systolic : Blood pressure
- relaxation : Blood pressure
- fasting blood sugar
- Cholesterol : total
- triglyceride
- HDL : cholesterol type
- LDL : cholesterol type
- hemoglobin
- Urine protein
- serum creatinine
- AST : glutamic oxaloacetic transaminase type
- ALT : glutamic oxaloacetic transaminase type
- Gtp : γ-GTP
- oral : Oral Examination status
- dental caries
- tartar : tartar status
- **smoking**

PRELEMINARY DATA ANALYSIS

SHAPE:

- 55.692 rows
- 27 columns

MISSING:

- None

CLASS POPULATION:

- 35.237 non-smoker
- 20.455

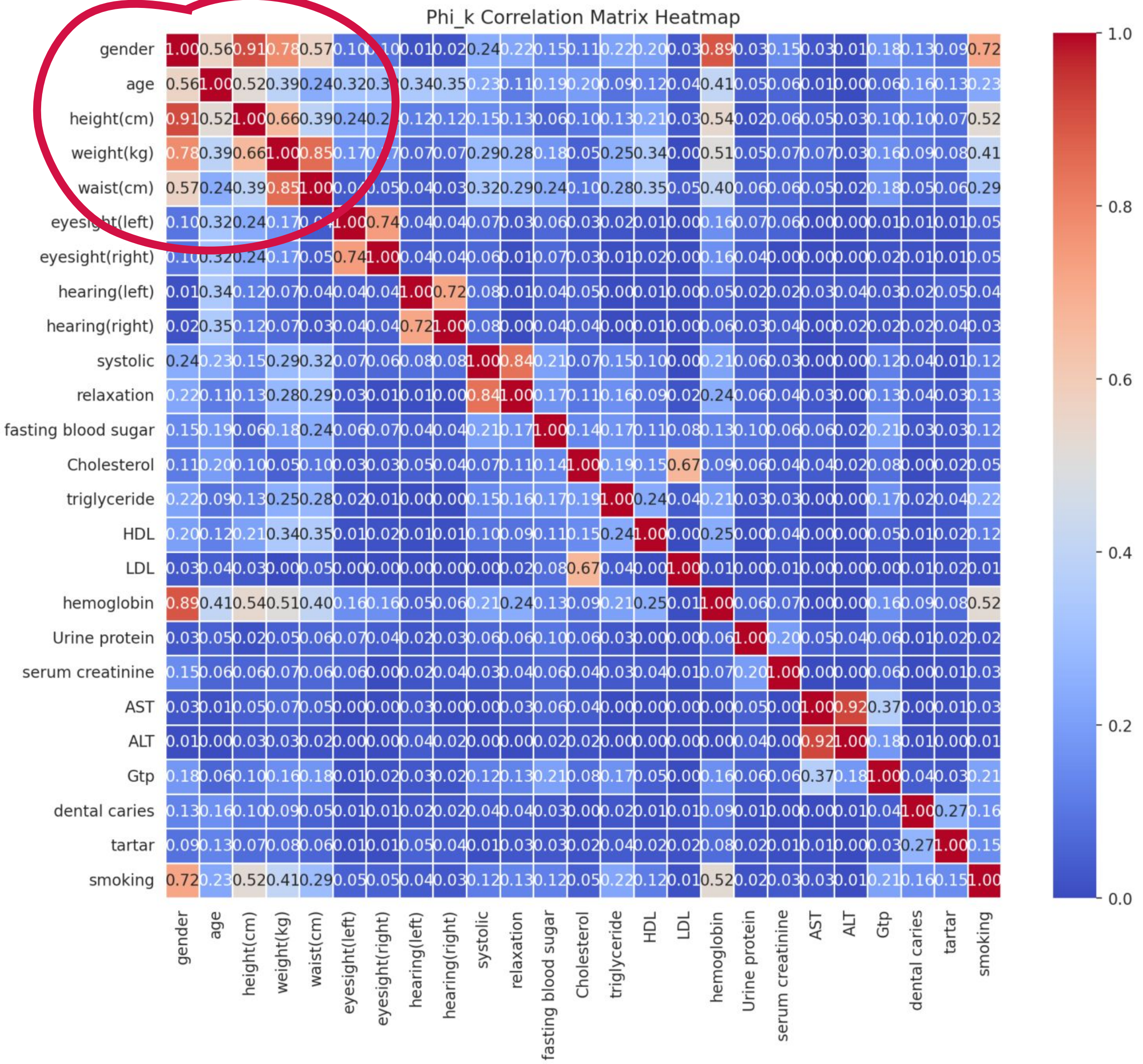
DUPLICATES:

- None

EXPLORATORY DATA ANALYSIS

CORRELATION ANALYSIS

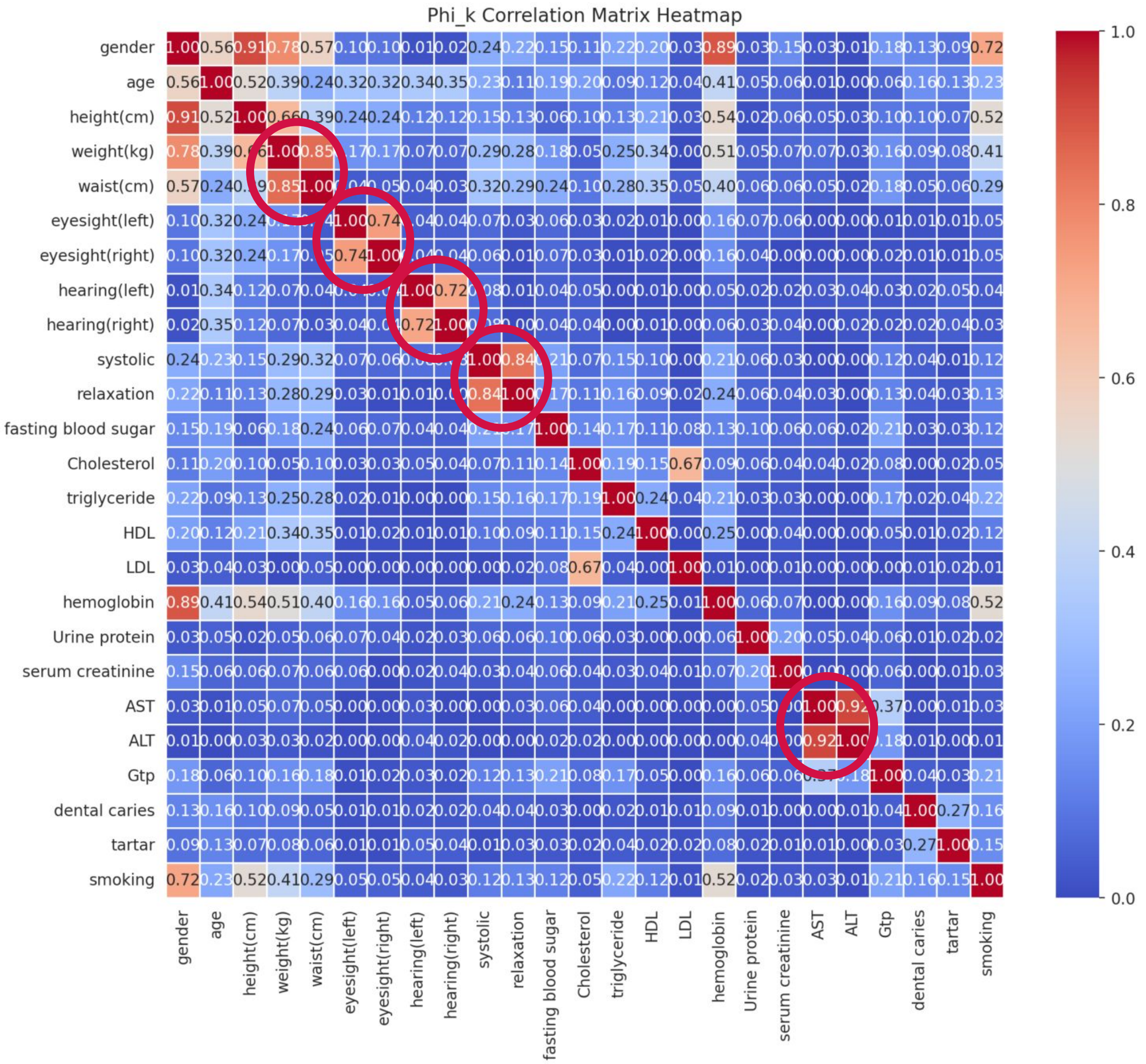
1. Strong correlation found between demographical data



EXPLORATORY DATA ANALYSIS

CORRELATION ANALYSIS

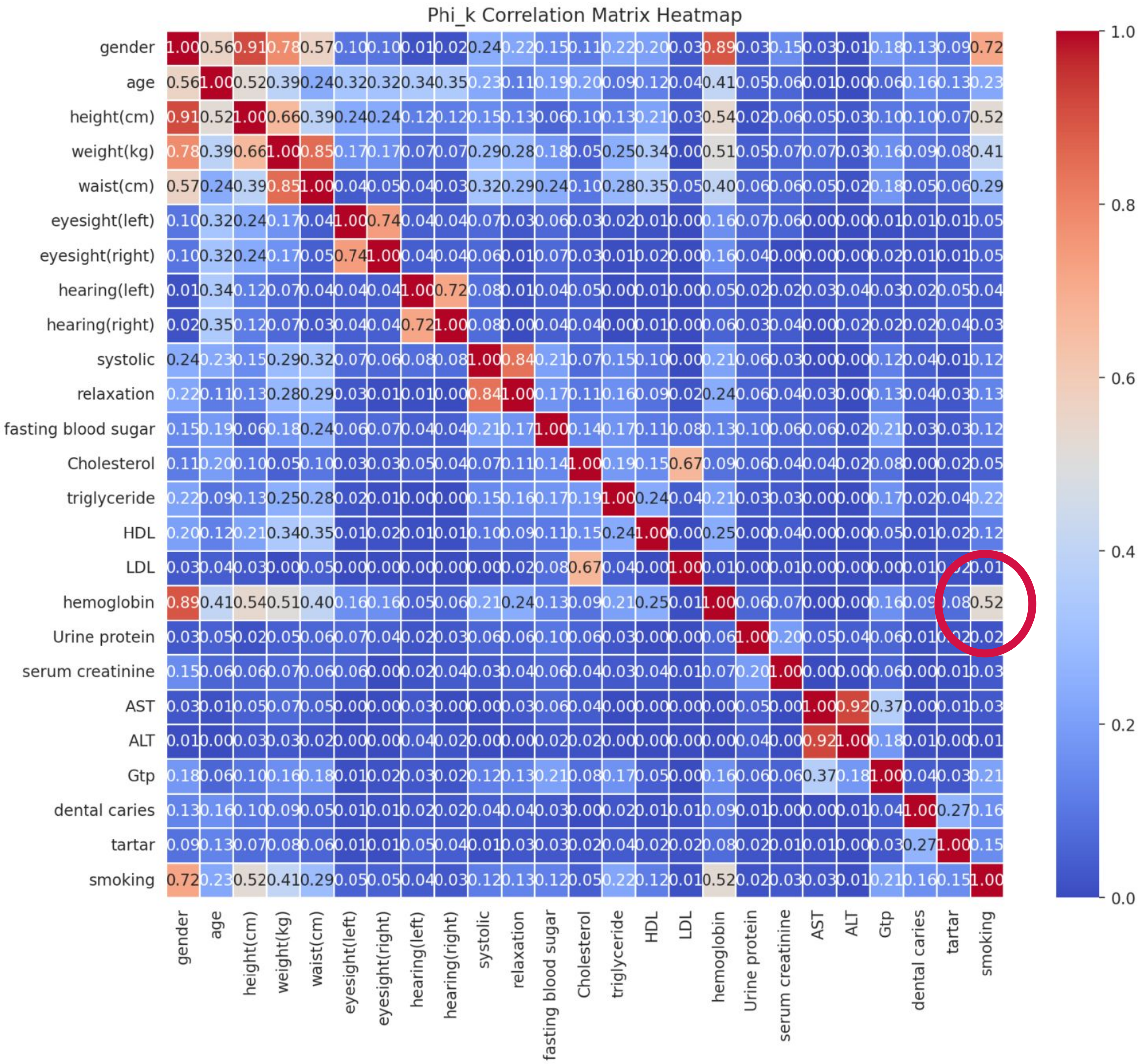
1. Strong correlation found between demographical data
2. Strong correlation found between paired data



EXPLORATORY DATA ANALYSIS

CORRELATION ANALYSIS

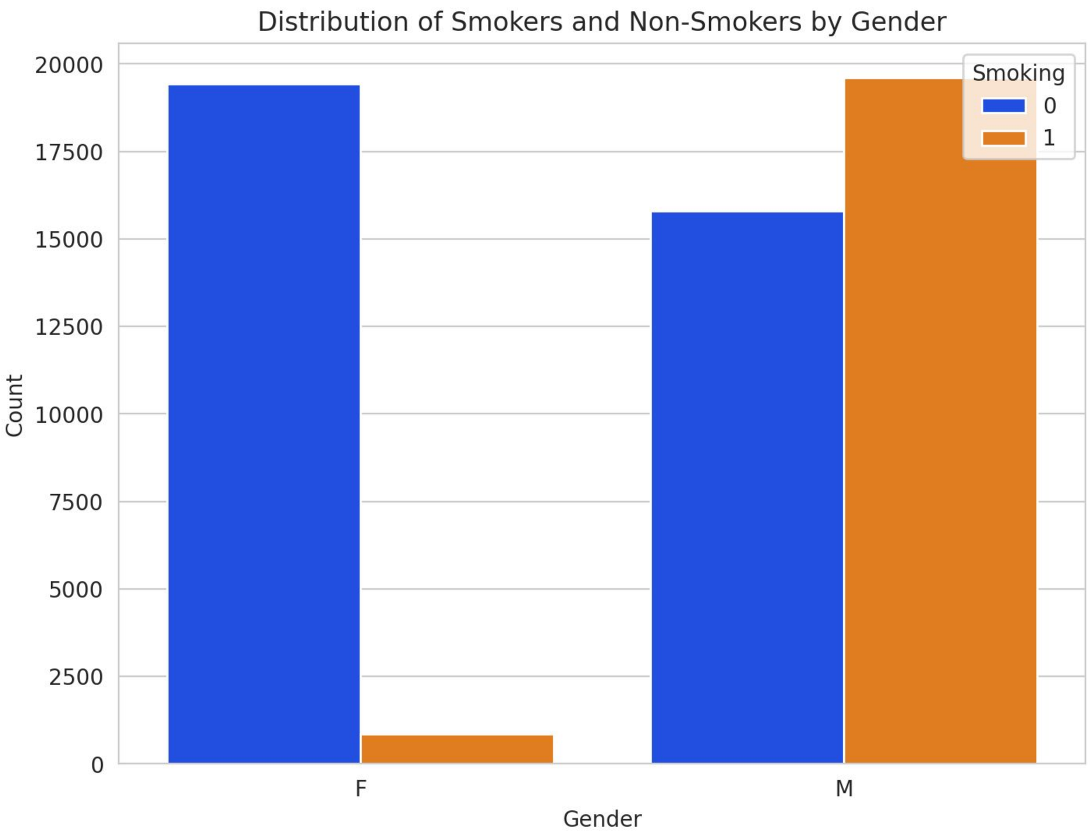
1. Strong correlation found between demographical data
2. Strong correlation found between paired data
3. Other than demographical data, hemogloobin correlates with target



EXPLORATORY DATA ANALYSIS

SMOKER BY GENDER

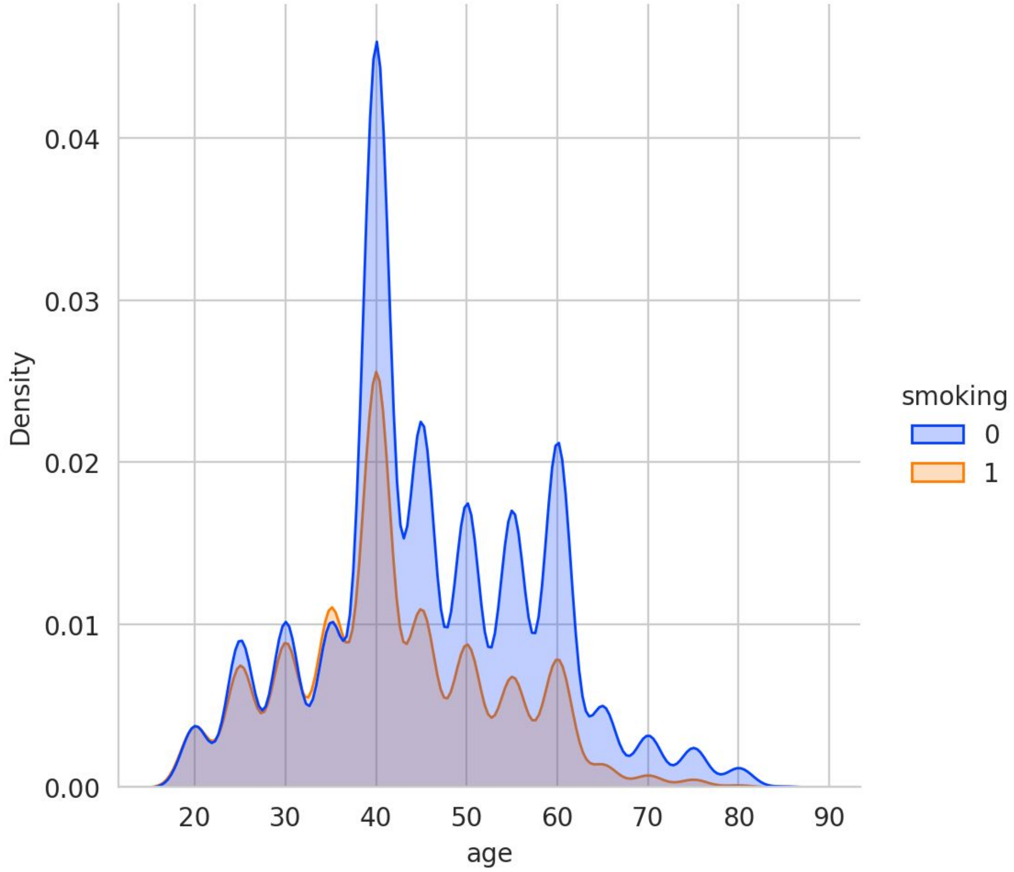
1. Almost no female smoke
2. Most of the population are males
3. Proof that demographical data are highly biased



EXPLORATORY DATA ANALYSIS

SMOKER BY AGE

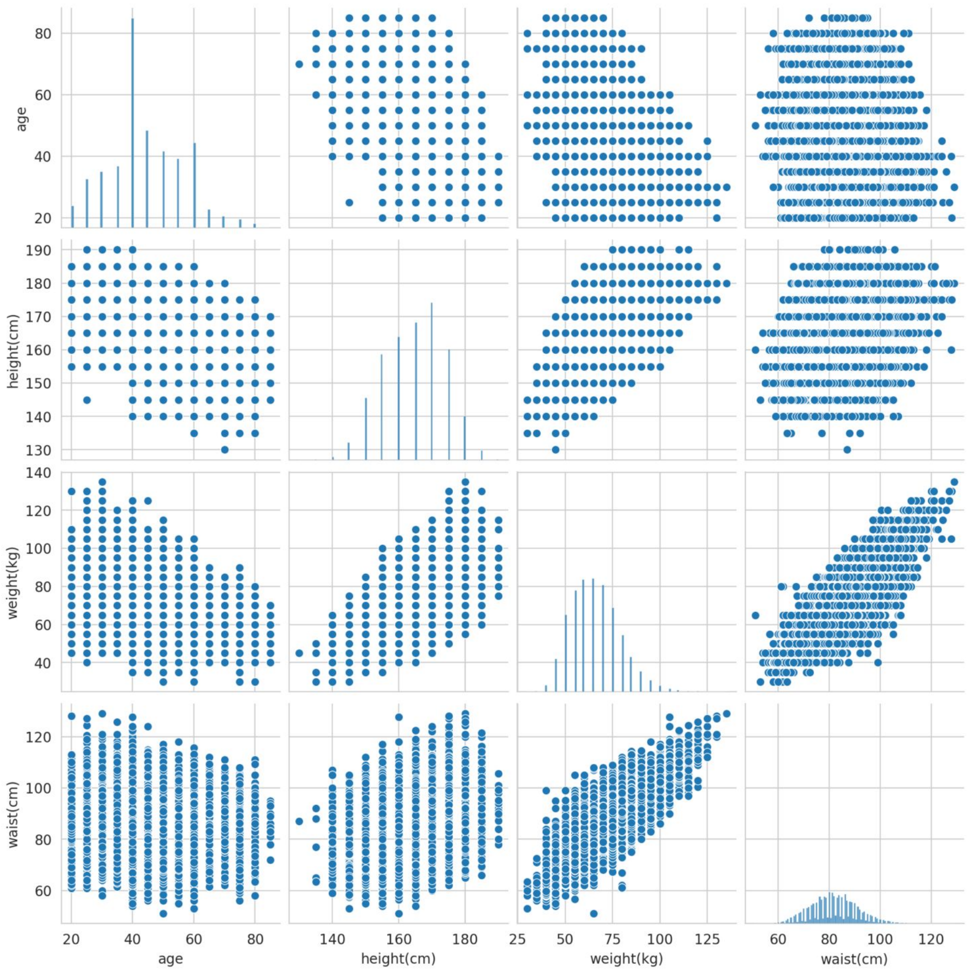
1. Grouped in 5 years bin
2. Most of the population are 40s
3. Smoker count is higher than non-smoker in mid 30s



EXPLORATORY DATA ANALYSIS

DEMOGRAPHY BUILT

1. Older → Thinner
2. Taller → Heavier
3. Heavier → Bigger



FEATURE SELECTION

DOMAIN KNOWLEDGE

DROPPED COLUMNS

- ID
- Gender
- Age
- Height
- Weight
- Waist
- Oral

FEATURE IMPORTANCE

SELECTED COLUMNS

- Hemoglobin
- Gtp
- Dental caries
- Serum creatinine
- AST
- Triglyceride
- LDL
- HDL
- ALT
- Hearing

MODEL PIPELINE STEPS

1. TRANSFORMER
    a. StandardScaler
    b. OneHotEncoding
2. BALANCING
    a. RandomUnderSapler
3. MODEL
    a. Classifier Models:

        SVM, KNN, DT, RF, XGB

BASE MODEL EVALUATION USING CROSS-VAL

```
svm
Recall- All - Cross Validation  :  [0.79443255 0.80055063 0.80881003 0.80446756 0.80507956]
Recall- Mean - Cross Validation :  0.8026680665110814
Recall- Std - Cross Validation  :  0.004880257412938665
Recall- Range of Test-Set       :  0.7977878090981427 - 0.80754832392402
=========================================================

knn
Recall- All - Cross Validation  :  [0.73784032 0.72499235 0.73447537 0.73684211 0.75367197]
Recall- Mean - Cross Validation :  0.7375644242322354
Recall- Std - Cross Validation  :  0.009248244480395088
Recall- Range of Test-Set       :  0.7283161797518404 - 0.7468126687126305
=========================================================

dt
Recall- All - Cross Validation  :  [0.71428571 0.72682778 0.72132151 0.72399021 0.71970624]
Recall- Mean - Cross Validation :  0.7212262891679675
Recall- Std - Cross Validation  :  0.004231006769985912 5
Recall- Range of Test-Set       :  0.7169952823979816 - 0.7254572959379534
=========================================================

rf
Recall- All - Cross Validation  :  [0.82349342 0.82716427 0.84429489 0.83323133 0.83506732]
Recall- Mean - Cross Validation :  0.832650247699823 2
Recall- Std - Cross Validation  :  0.007151517440053046 5
Recall- Range of Test-Set       :  0.8254987302597702 - 0.8398017651398763
=========================================================

xgb
Recall- All - Cross Validation  :  [0.80146834 0.80911594 0.80666871 0.80813953 0.80385557]
Recall- Mean - Cross Validation :  0.8058496173205285
Recall- Std - Cross Validation  :  0.002819762586177679 7
Recall- Range of Test-Set       :  0.8030298547343508 - 0.8086693799067062
```

Selected model: **XGBClassifier**

HYPERPARAMETER TUNING RESULT

BEST PARAMETER:

subsample: 0.8

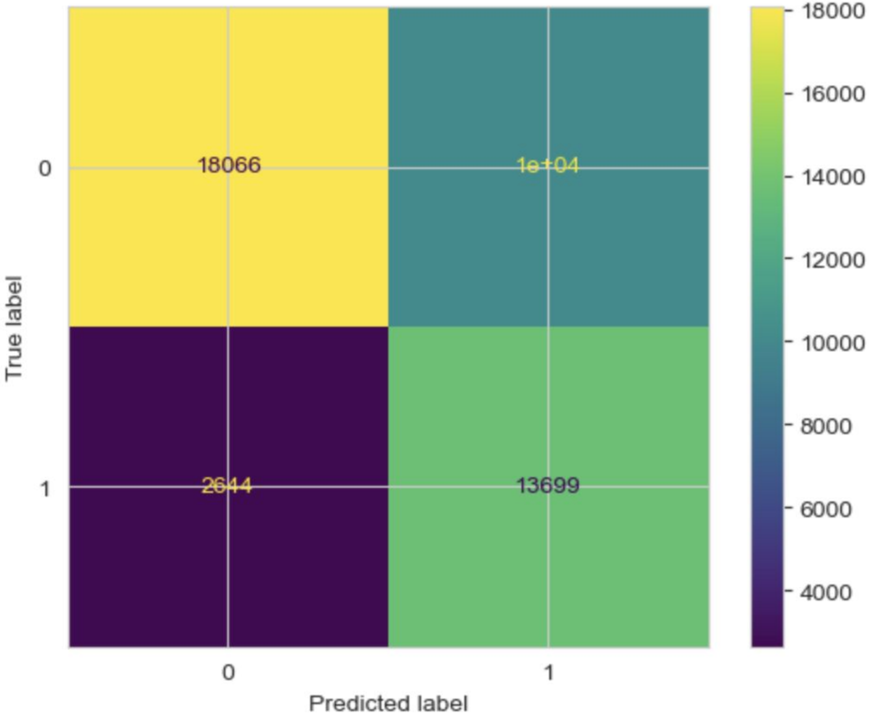
n_estimators: 100

max_depth: 5

learning_rate: 0.01

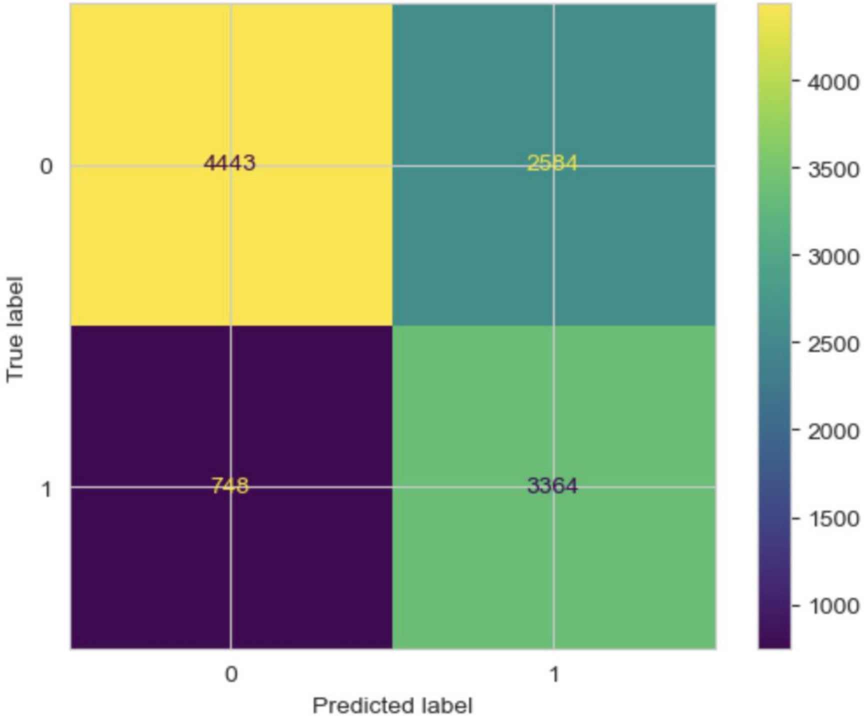
colsample_bytree: 1.0

XGB – Train



	precision	recall	f1-score	support
0	0.87	0.64	0.74	28210
1	0.57	0.84	0.68	16343
accuracy			0.71	44553
macro avg	0.72	0.74	0.71	44553
weighted avg	0.76	0.71	0.72	44553

XGB – Test



	precision	recall	f1-score	support
0	0.86	0.63	0.73	7027
1	0.57	0.82	0.67	4112
accuracy			0.70	11139
macro avg	0.71	0.73	0.70	11139
weighted avg	0.75	0.70	0.71	11139

RESULTS AND IMPLICATIONS

Final model achieved high recall (and accuracy) in identifying smokers, crucial for insurance risk assessment.

Implications include informed decisions on policy premiums and coverage, leading to more equitable and reliable insurance offerings for customers.

CHALLENGES AND SOLUTIONS

Addressed slight overfitting through feature selection techniques, focusing on informative variables while eliminating noise.

Balanced class imbalance (60-40 split between non-smokers and smokers) to ensure equitable representation during model training, enhancing prediction accuracy.

FUTURE DIRECTIONS

Continuous model refinement and data collection essential for further improvement.

Explore advanced modeling techniques and gather more comprehensive datasets to enhance predictive accuracy.

Ongoing monitoring of model performance critical for ensuring effectiveness in insurance risk assessment over time.

THANK YOU