MODELING WITH REAL WORLD DATA

Behavioral Risk Factor Surveillance System dataset

Presented by:

Fahad & Faisal

PRESENTATION OUTLINE

Chosen dataset
Features selection
EDA
Modeling
Model comparison
Future work & Challenges
Conclusion

The dataset

BEHAVIORAL RISK FACTOR SURVEILLANCE SYSTEM DATASET CAN BE FOUND AT KAGGLE

Data objective

To reach preventive health practices and to pinpoint risk behaviors that are linked to chronic diseases.

Data collection

It was collected via telephone surveys in the U.S.

Data shape

(441456, 330)

Features

SELECTION CRITERIA

- Prior knowledge of existing indirect relationship
- Differentiating unavoidable feature
- Feature is within our scope
- Feature have manegable data
- Features are exlcluded if they have a direct relationship with targets or have very high correlation

Features	Possible targets
Sex	Diabetes
Age	High Blood Pressure
Marital status	High Cholesterol
Education level	Heart Attack
Employment	Coronary Heart DIS
Income level	Stroke
General health	Depression
Mental health	Anxiety
Poor health	
Medical cost	
Hours of work	
Aspirin intake	
Life satisfaction	
Smoke	
Phisical activity	
Heavy drinker	

PROBLEM STATEMENT

Can our selected features empower the models we choose?

- TARGETS

- **DEPRESSION**
- **HIGH BLOOD PREASURE**

Cleaning proccess

- Removing rows that are not meaningful or missing.
- Adjusting rows to become binary
- Imputing mean for some column's misssing values
- Unit conversion
- Rescaling some columns
- Dropping 4 columns with significant lost values

RESULT

- Clean df.shape = (203412, 20)

Exploratory data analysis

We conducted analyses of our features.
Briefly we will explore some OBSERVATIONS

INCOME

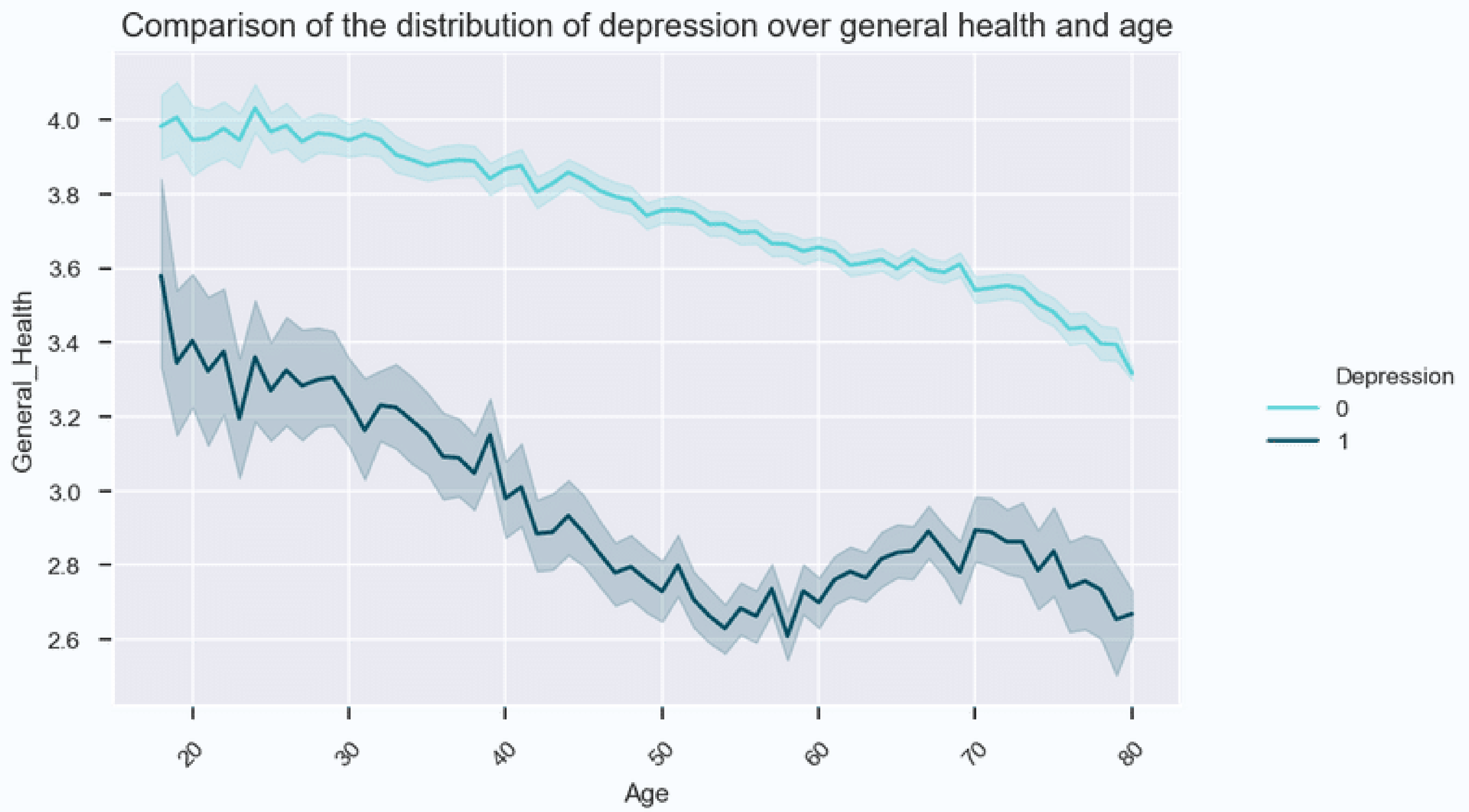
Improvment in income could mean improvment in other aspect such as health and mental health. Chances of being married are are higher with increasment in income. There was a positive relationship between income and education.

GENERAL HEALTH

Percentage of unemployment decreases with the improvment of health. Distribution of depression was skewed towards lower levels of general health. General health seemed to be a better indicator of mental health than income.

AGE

General health declines over age. In contrast, mental health was shown to improve over age in our sample.

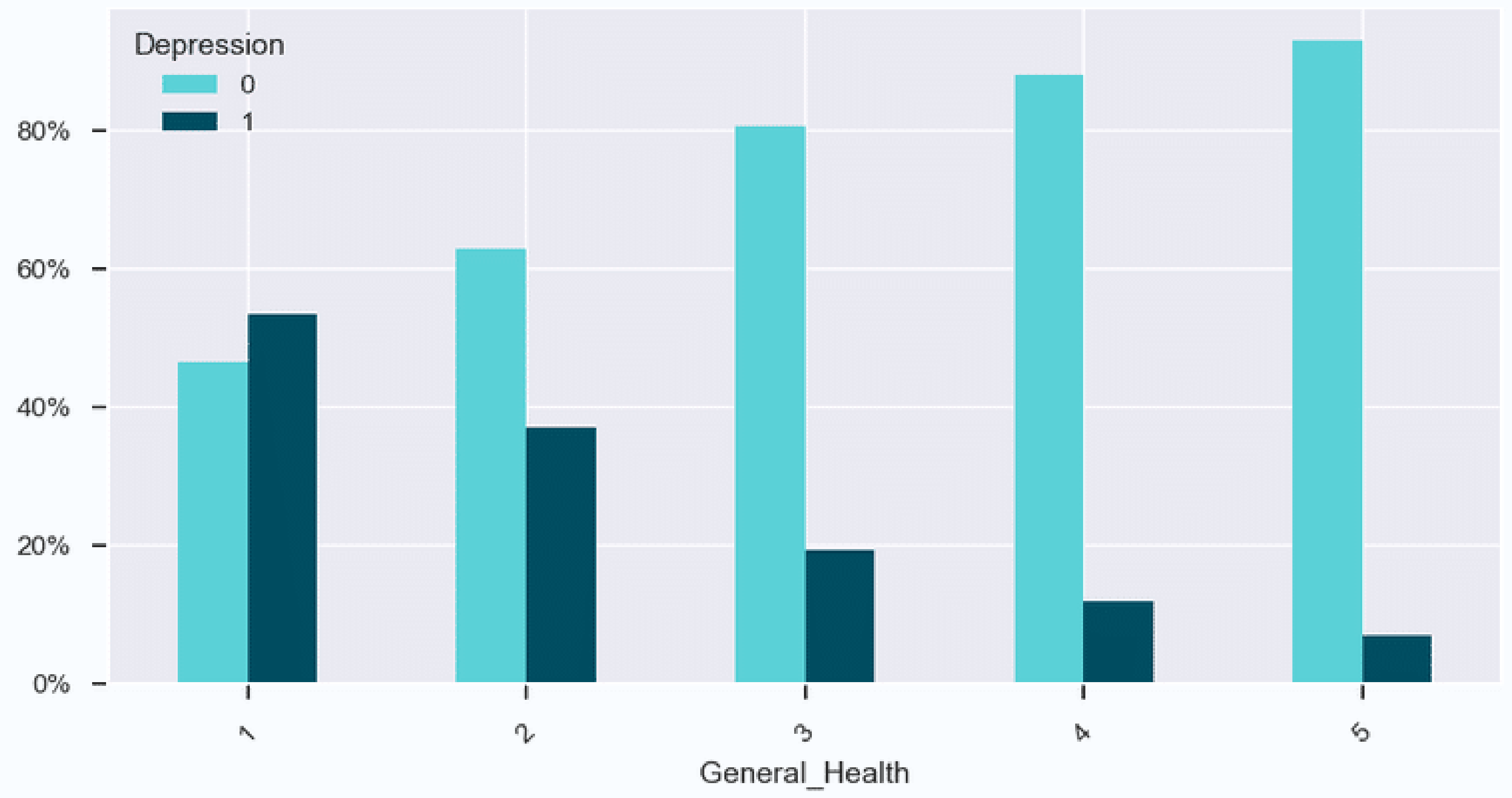


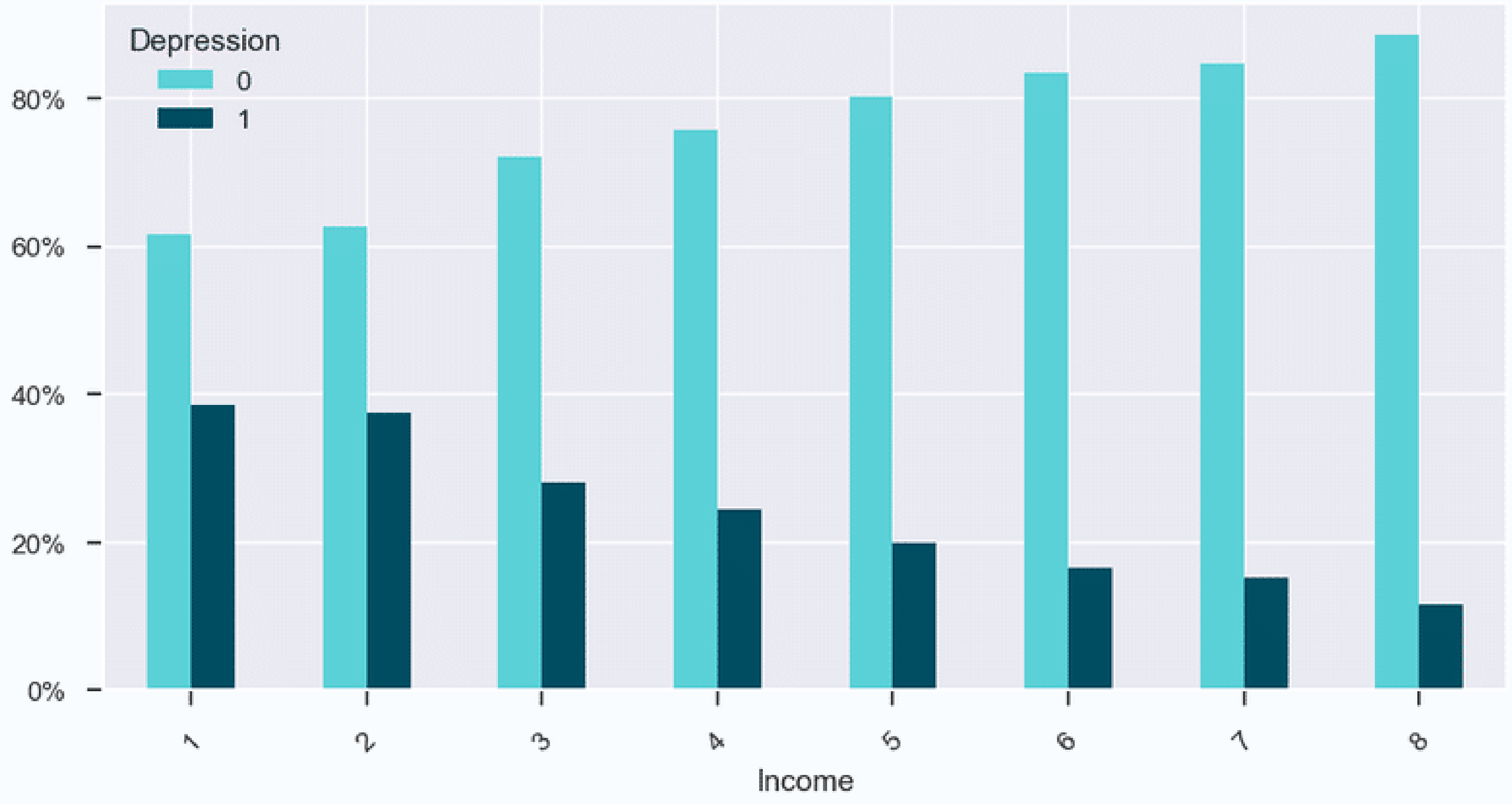
LEFT

Distribution is not clearly separated

RIGHT

Distinctive distribution



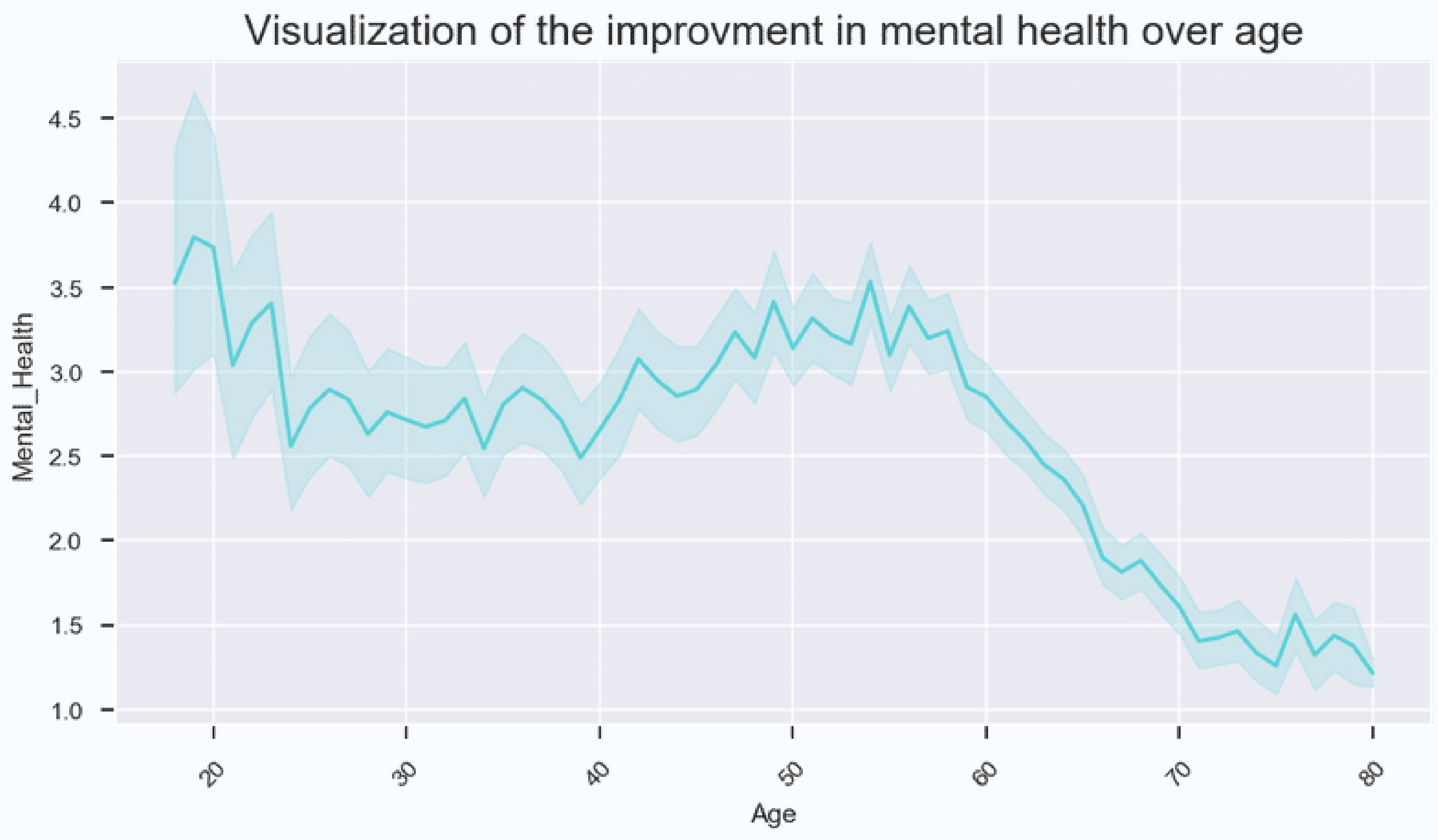


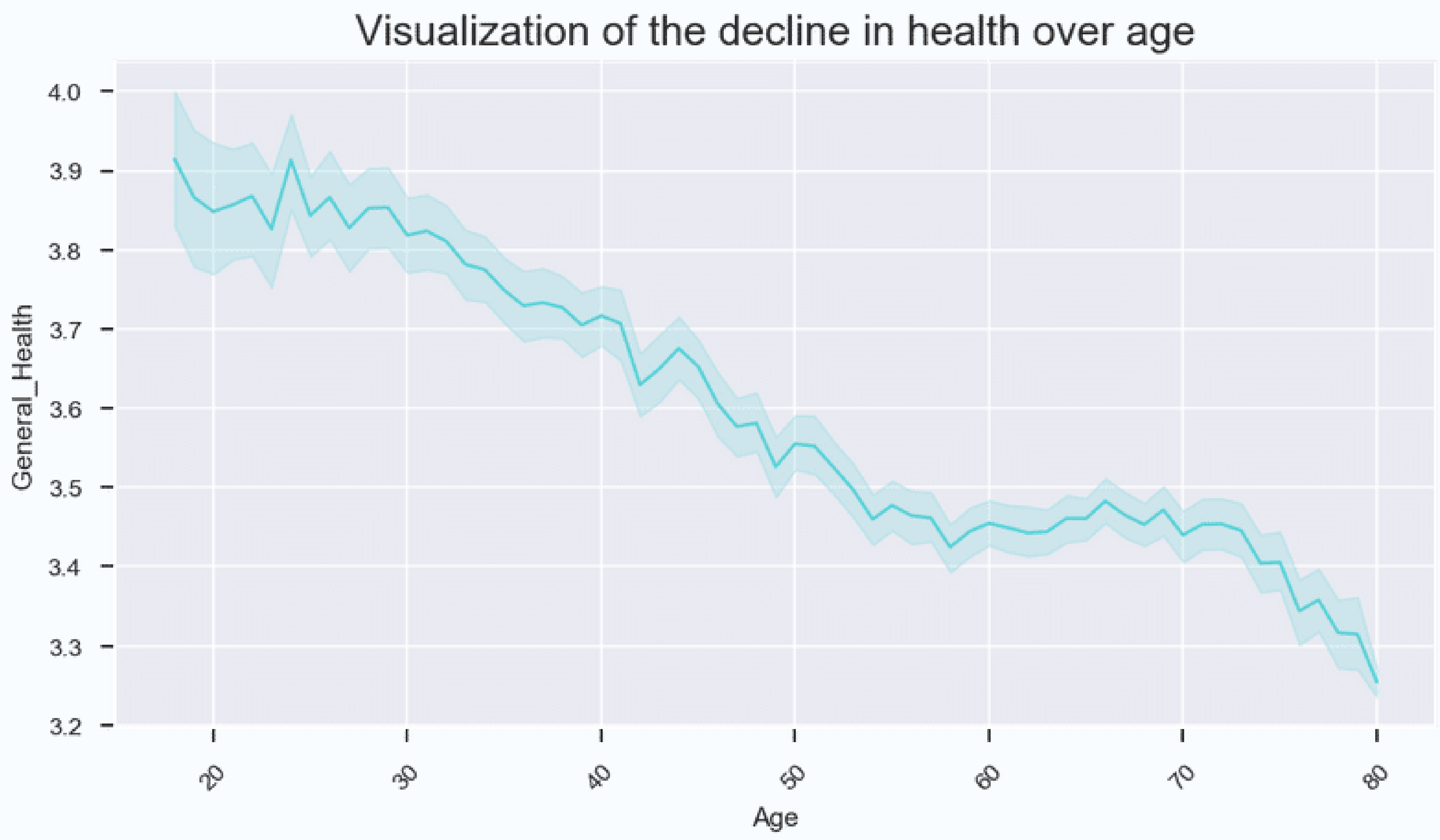
LEFT

Visualization of the percentage of depression in each general health level

RIGHT

Visualization of the percentage of depression in each income level





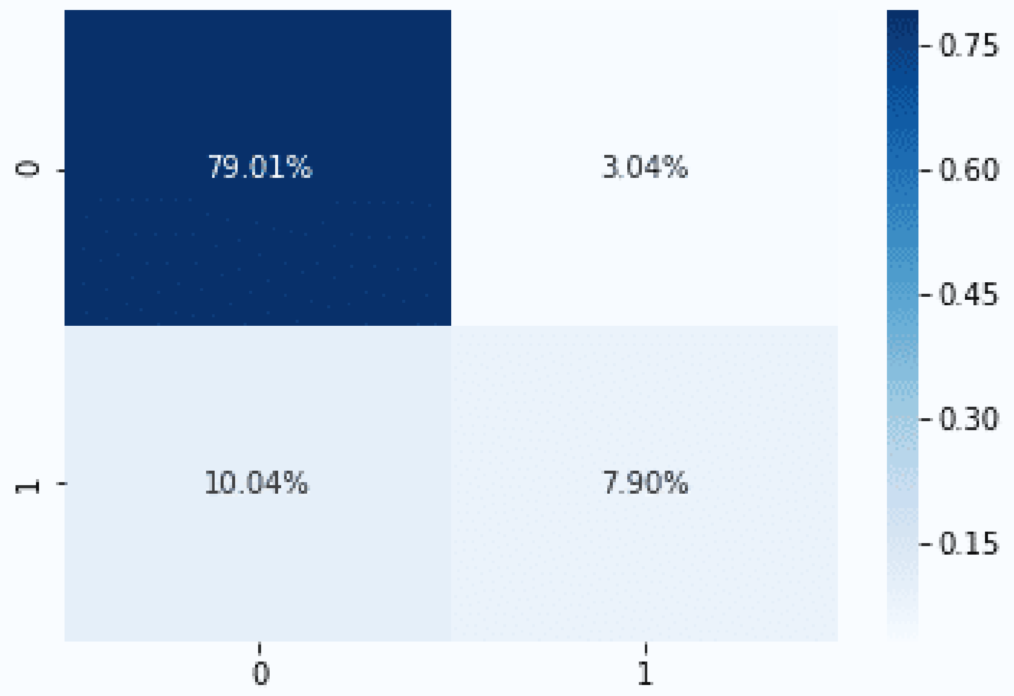
LEFT

Sharp drop in the amount of mental ilness days after 60

RIGHT

Steep decline in genral health with aging

LR Targeting depression



	precision	recall	f1-score	support
0	0.90	0.72	0.80	49944
1	0.33	0.62	0.43	11080
accuracy			0.70	61024
macro avg	0.61	0.67	0.61	61024
Base line auc	0.81783768902523			

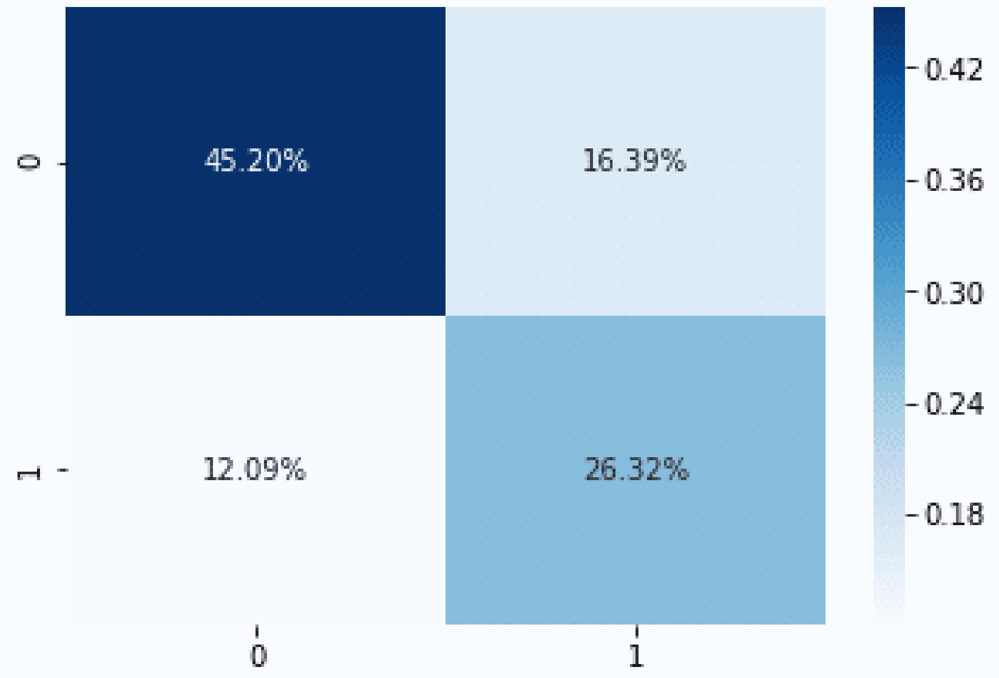
RF Targeting depression

	precision	recall	f1-score	support
0	0.89	0.96	0.92	41729
1	0.72	0.44	0.55	9124
accuracy			0.87	50853
macro avg	0.80	0.70	0.73	50853
Base line auc	0.81783768902523			

18%

DEPRESSED

LR Targeting HBP



	precision	recall	f1-score	support
0	0.73	0.79	0.76	58200
1	0.69	0.61	0.65	43506
accuracy			0.72	101706
macro avg	0.71	0.70	0.71	101706
Base line auc	0.573034039289717			

RF Targeting HBP

	precision	recall	f1-score	support
0.0	0.71	0.74	0.72	58406
1.0	0.62	0.59	0.61	43300
accuracy			0.68	101706
macro avg	0.67	0.66	0.67	101706
Base line auc	0.573034039289717			

44%

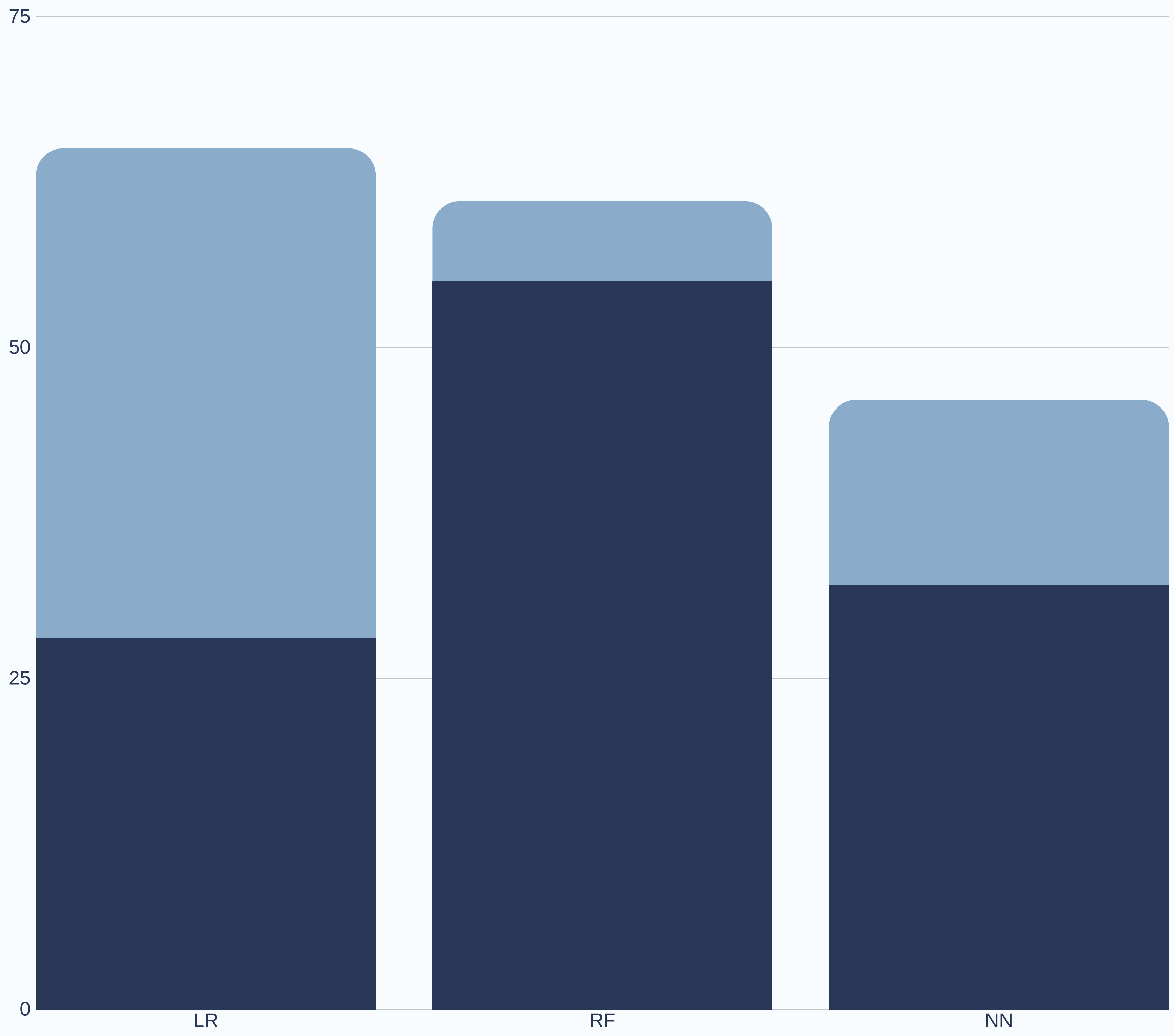
HBP

Comparison of modeles performance

HBP OUTPERFORMANCE

We had very good performance when we targeted HBP. Two reasons for that:

- Impact of imbalced class
- Better features for HBP



Future work & Challenges

WIDEN FEATURES SCOPE

RUN FEATURE SELECTION ALGORITHMS

UTILIZE CLOUD COMPUTING

CONCIDER TENSORFLOW

Thanks