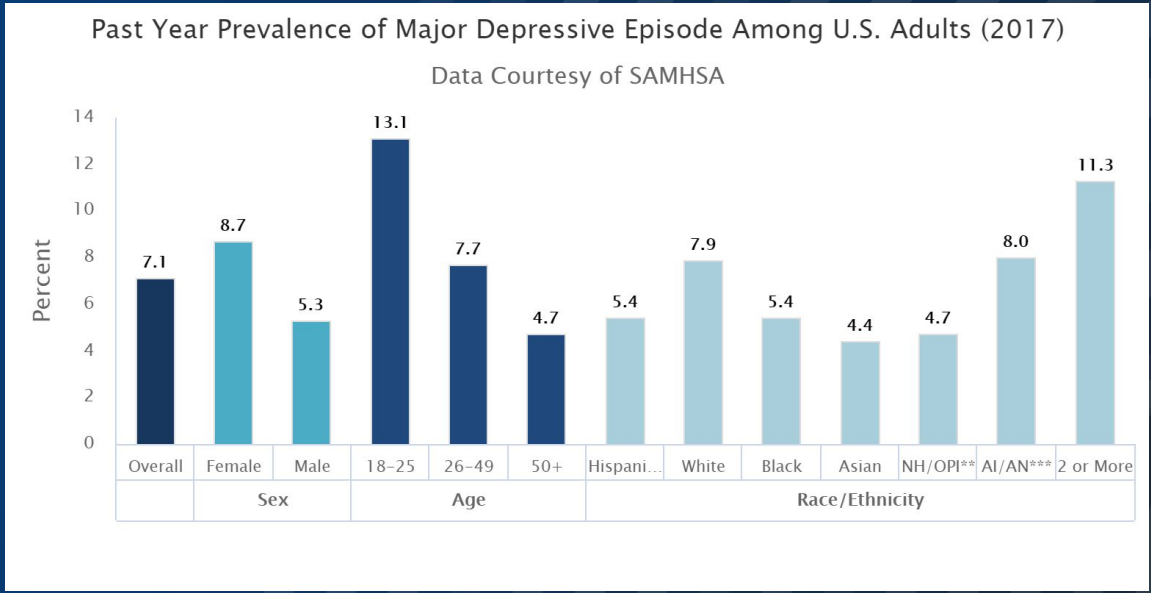
Predicting Depression Using Health Care Data

Amodh Udupa
Rahul V
Pothur Veda Vikas
Sharath Gowda

Depression

- 264 million people globally have depression - WHO
- 7.1% of US adults had a major depressive episode in 2017 with young adults being the most affected - NIMH



The Problem

- Physicians have become the front line for handling mental health disorders.
- The majority of physician visits are driven by mental disorders and few people follow up with a mental health professional.
- 90% of those who commit suicide have a mental health disorder and 40% had visited their doctor in the last month

Source: American Psychological Association

The Problem

- Study published in JAMA - Patients who receive more holistic care with doctors for mental health reduce healthcare costs and improve patient outcomes.
- This can be unfeasible for many reasons
  - Buy-in, logistics, training
- Machine learning could fill the gap to predict patients who are depressed.

Source: https://jamanetwork.chttps://jamanetwork.com/journals/jama/fullarticle/2545685om/journals/jama/fullarticle/2545685

Using machine learning to identify individuals with depression could connect patients with the help they need more quickly and easily while reducing healthcare costs and burden on physicians.

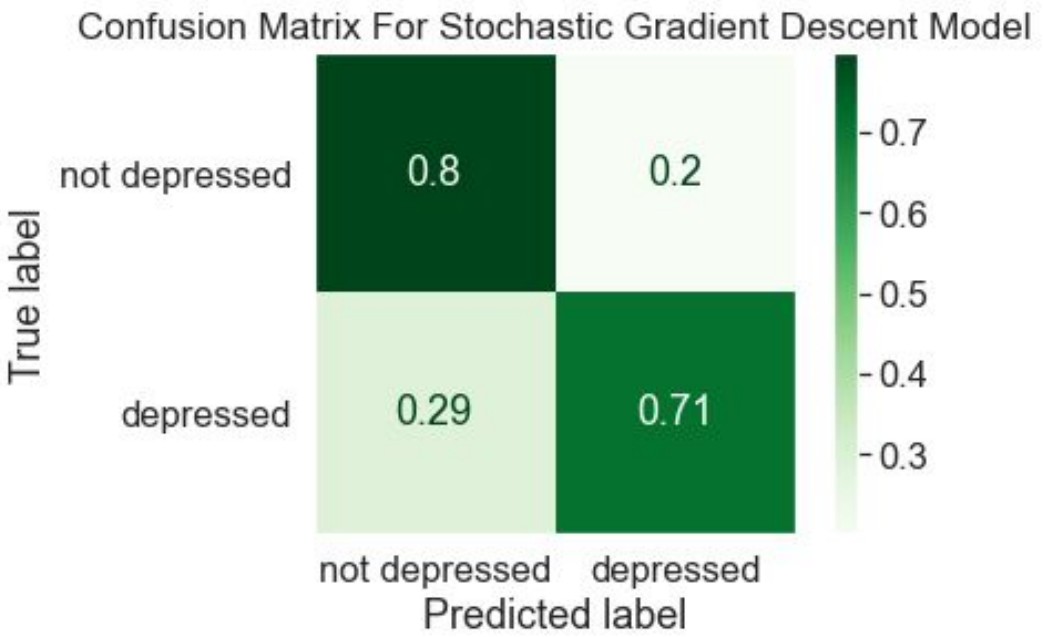
The Data

- CDC NHANES data
- 2005 to 2018 comprising 36259 entries total
- Tried to use data that was consistent across years and could reasonably be found in a patient's medical file
- Created labels of who was depressed and not depressed based on the "depression screener" in the data

Methods

- Classification of depressed or not depressed
- Roughly followed OSEMiN method
  - Obtain, scrub, explore, model, interpret
- Modeling was done from simple going to complex
- Wanted to use as few features as necessary
- Imbalanced data means accuracy was not a good metric
- Recall weighted more heavily to reduce false negatives

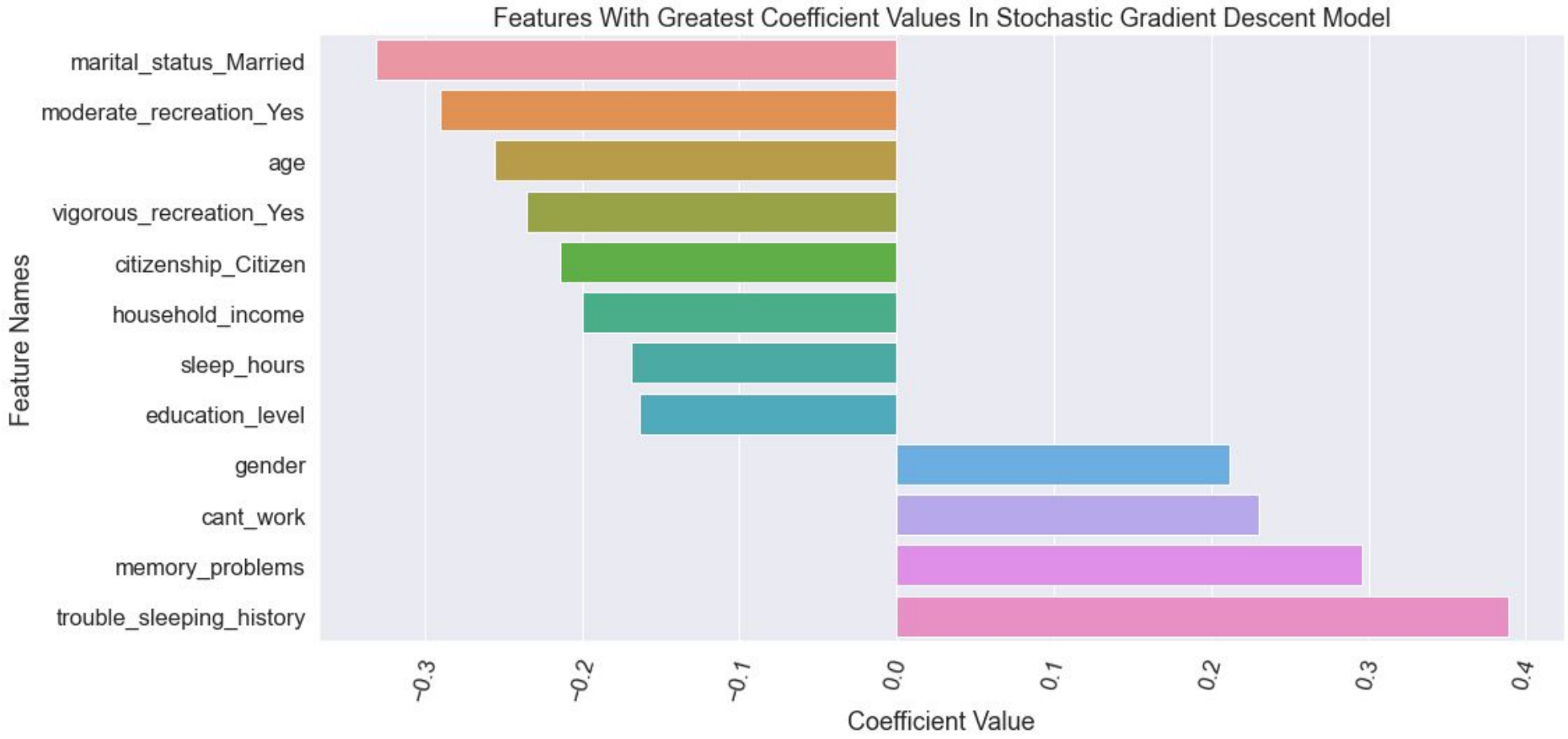
Best Model Results



Logistic regression is a model that calculates probabilities of entries being in one or the other class then uses a 50% threshold to make the prediction.

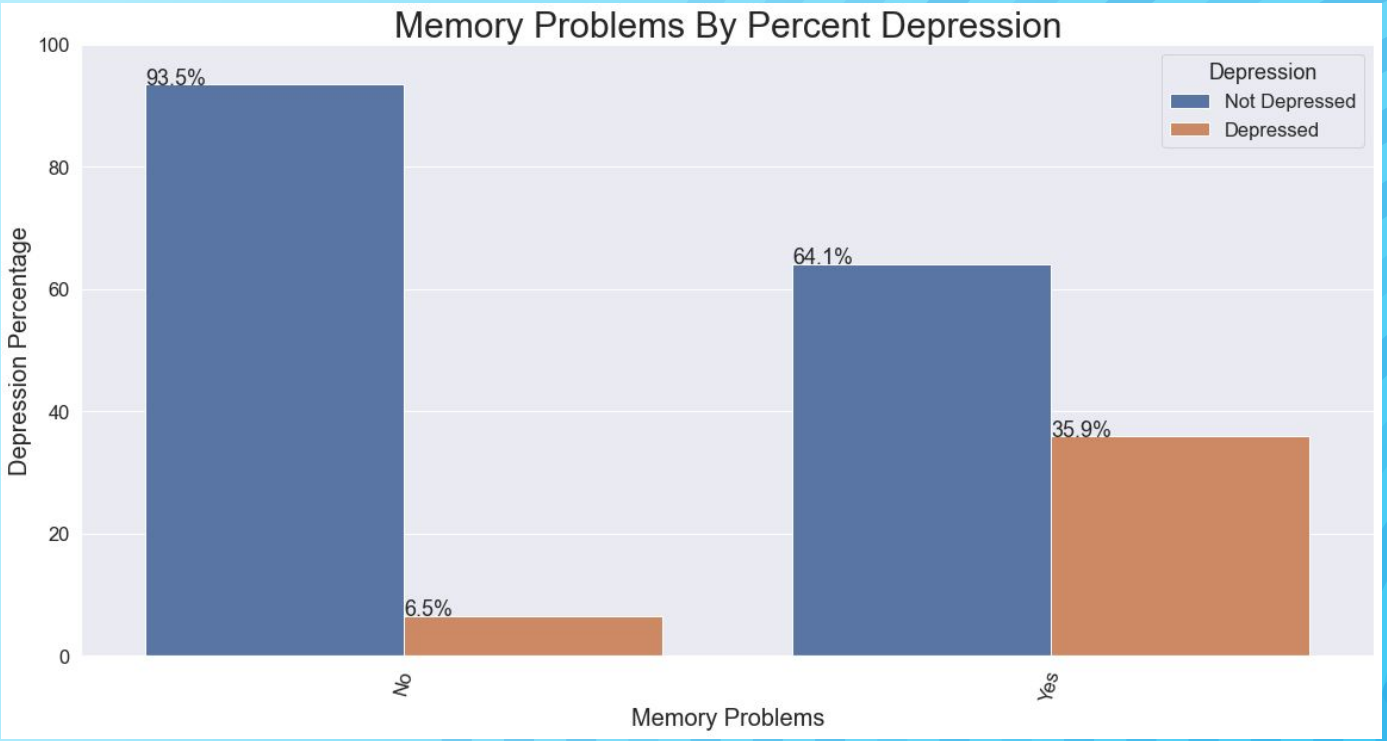
Gradient descent was used to maximize on a logistic regression.

Best Model Results



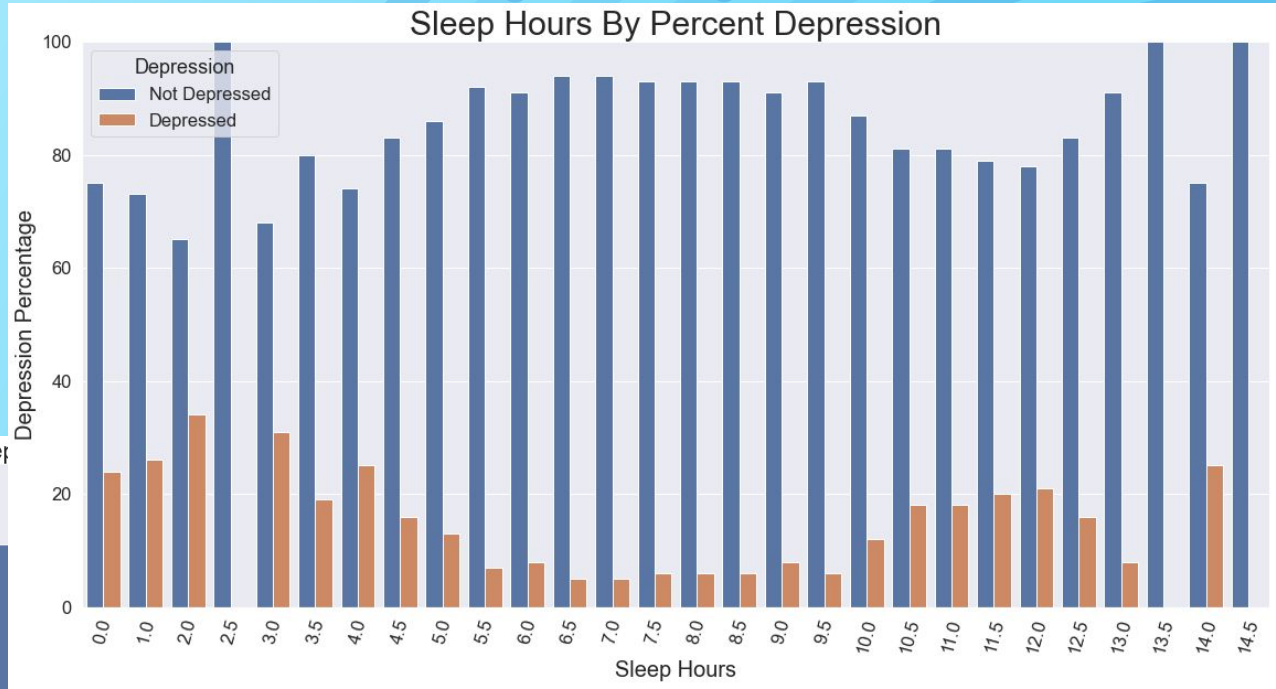
Recommendation 1:

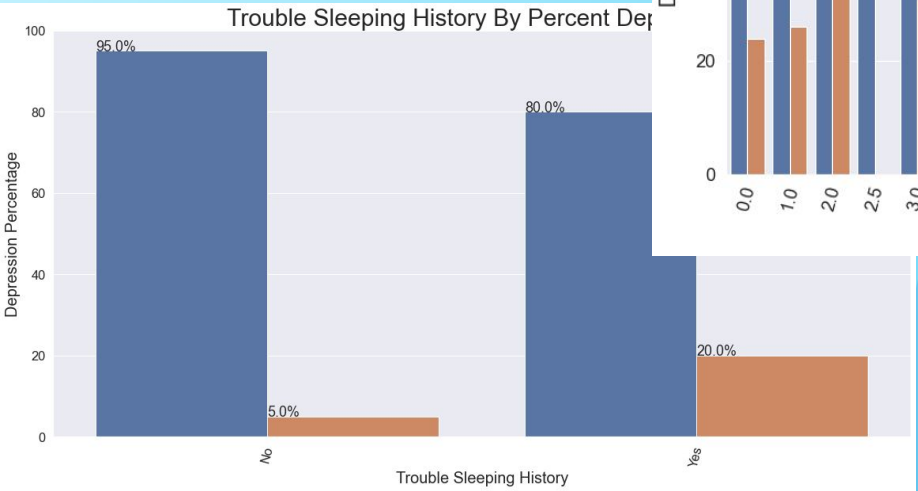
Watch for patients with memory problems



Recommendation 2:

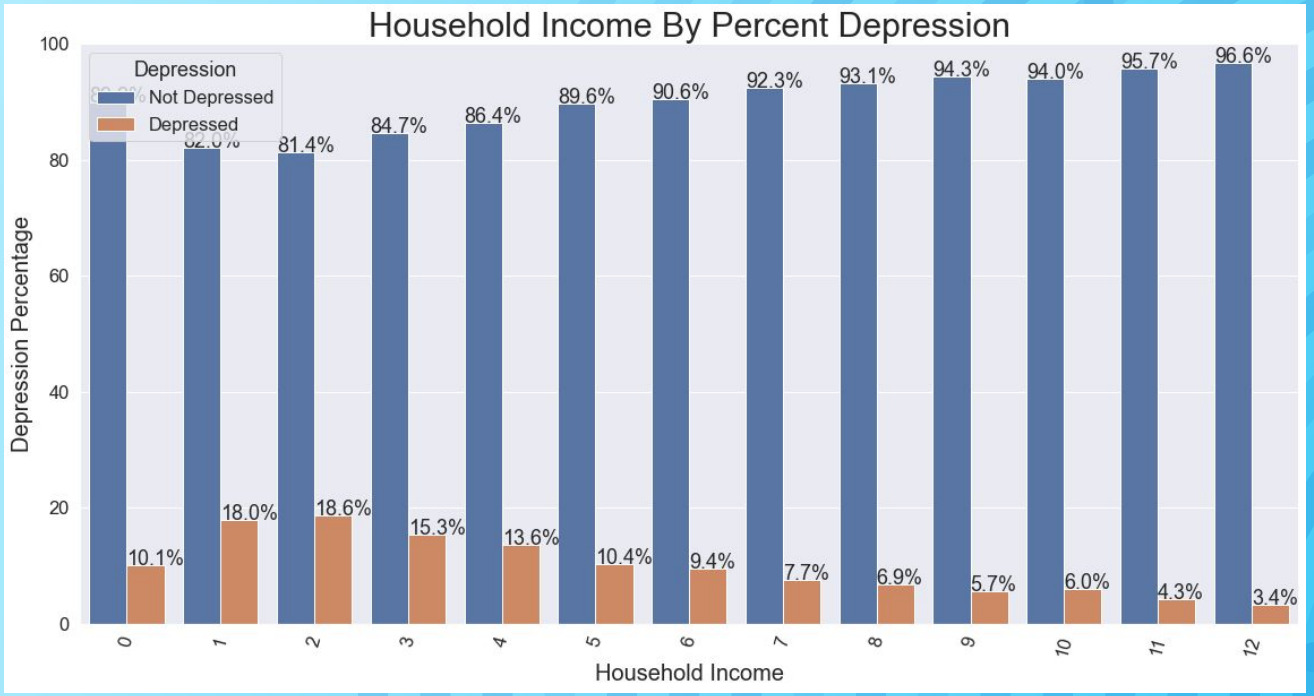
Watch for patients who have trouble sleeping and sleep too much or too little





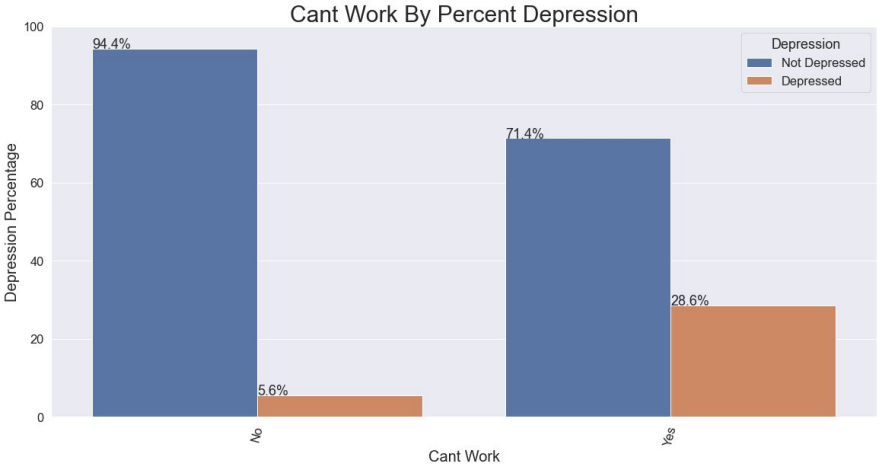
Recommendation 3:

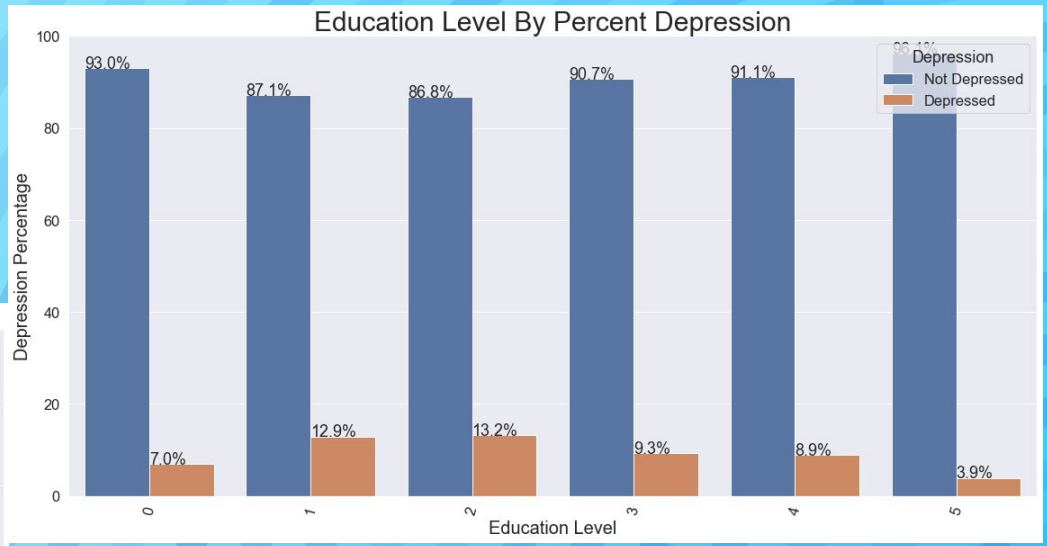
Watch for patients who can't work, have low household income, and/or low education



Recommendation 3:

Watch for patients who can't work, have low household income, and/or low education





Recommendations Recap

1. Watch for patients with memory problems
2. Watch for patients with sleep troubles
3. Watch for patients who can’t work, have low income, or low education

Future Work

- Try different models - perhaps neural networks
- Add more entries and evaluate valuable features
- Tuning and testing of parameters

Thank you for your time

Any questions?