Predicting Overnight Hospitalizations with NHANES Data

Supervised Learning Capstone

By Heather Knudson

Image Credit: https://heartinsight.heart.org/images/

- 2015-2016
- Demographics
- Hospital Utilization & Access to Care

nhanes

The Nation's Mobile Health Survey

Image Credit: https://www.cdc.gov/nchs/images/nhanes/nhanes_apple_color_tagline.

Data Cleaning

Image Credit: https://cdn.datafloq.com/blog_pictures/data-science-methods-and-algorithms-for-big-data.

Data Cleaning

1. Removed columns with > 1000 null values
2. For columns with < 1000 nulls, imputed code ‘1000’ to stand in for missing values
3. Concatenated the 2 datasets
4. Removed columns that were copies of other columns
5. Re-coded all columns so their first code was 0, rather than 1
6. Made a new categorical variable for age

Total: 9954 observations & 27 columns

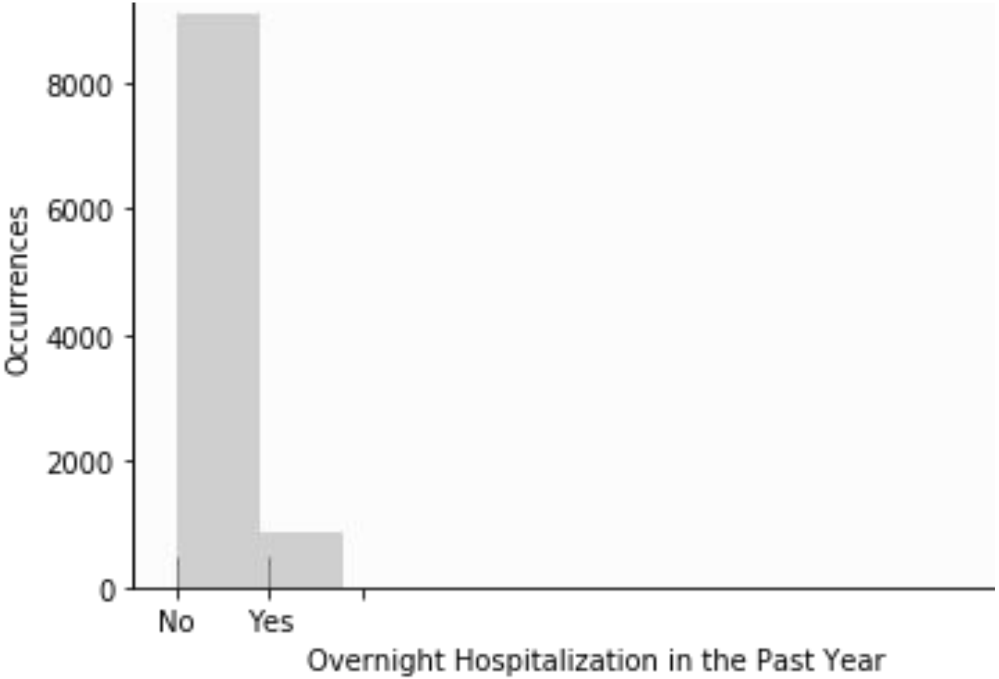
Research Question

To what extent can incidences of overnight hospitalization be predicted with NHANES demographic and hospital utilization features from the 2015-2016 year?

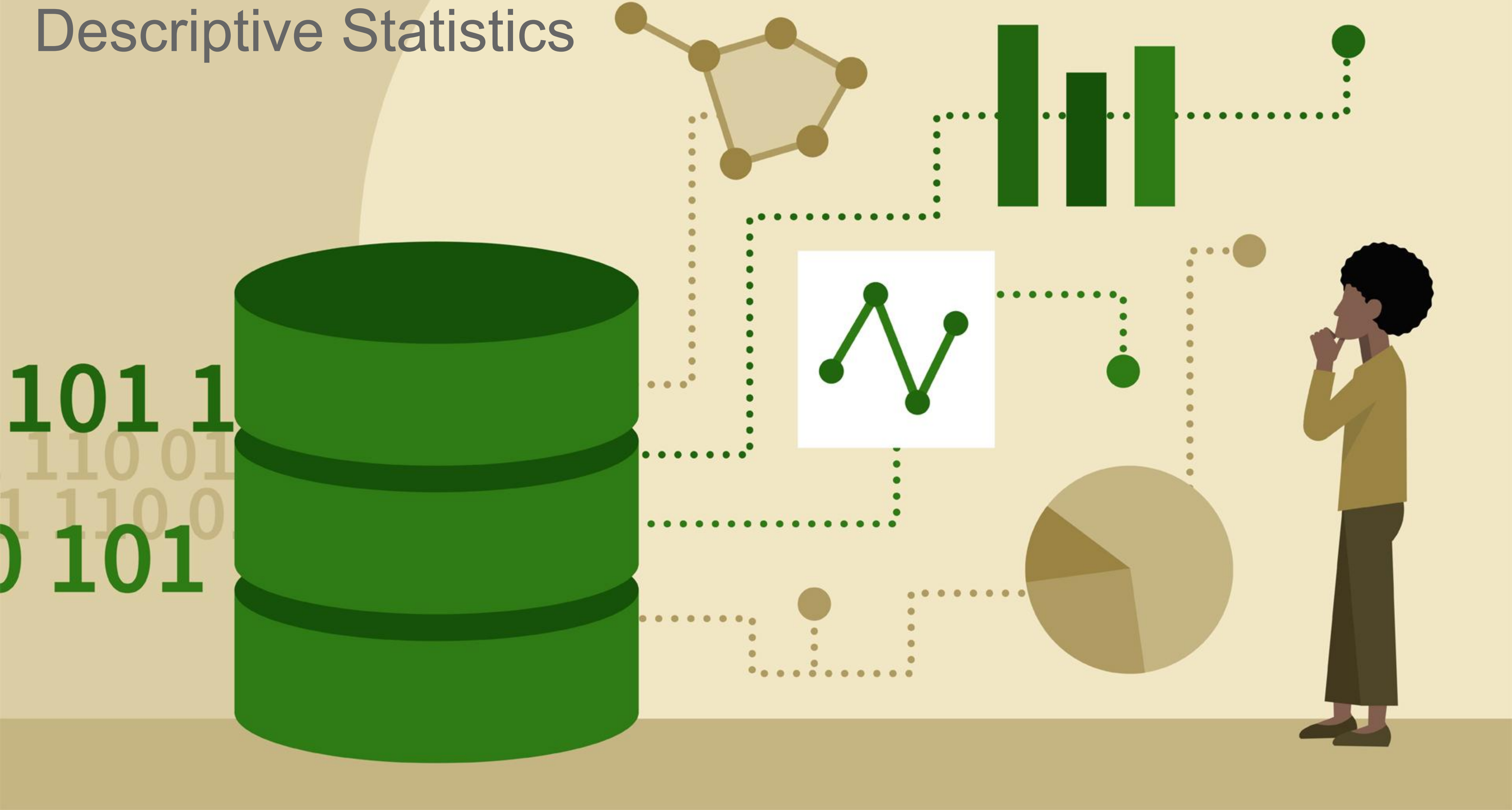
Target Variable - Overnight Hospitalization

- During the past 12 months, were you a patient in a hospital overnight? Do not include an overnight stay in the emergency room.

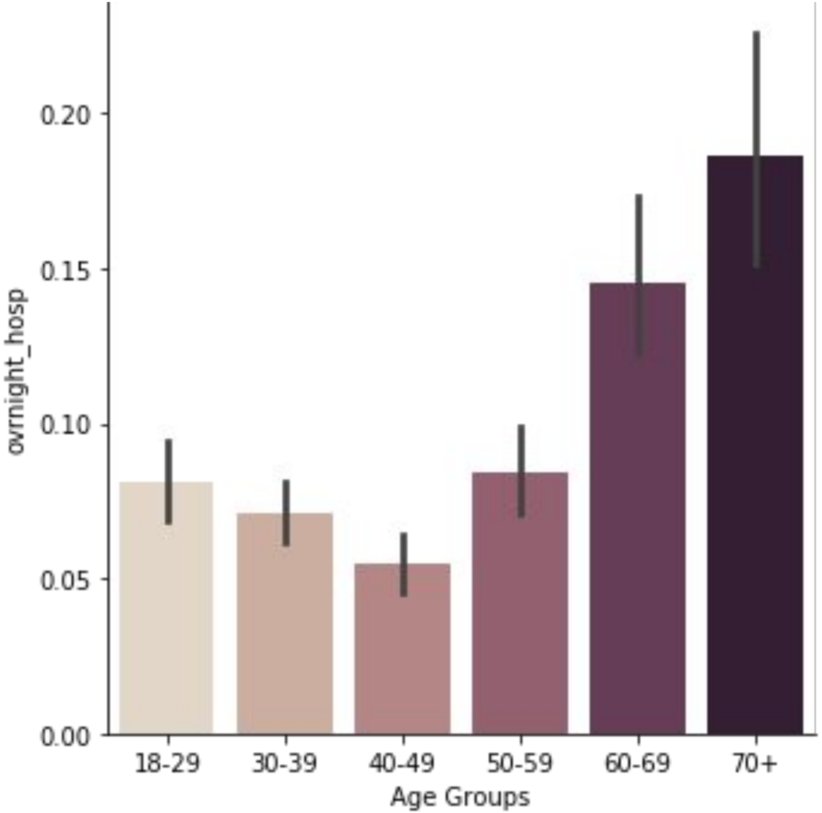
Distribution of overnight stays in hospital



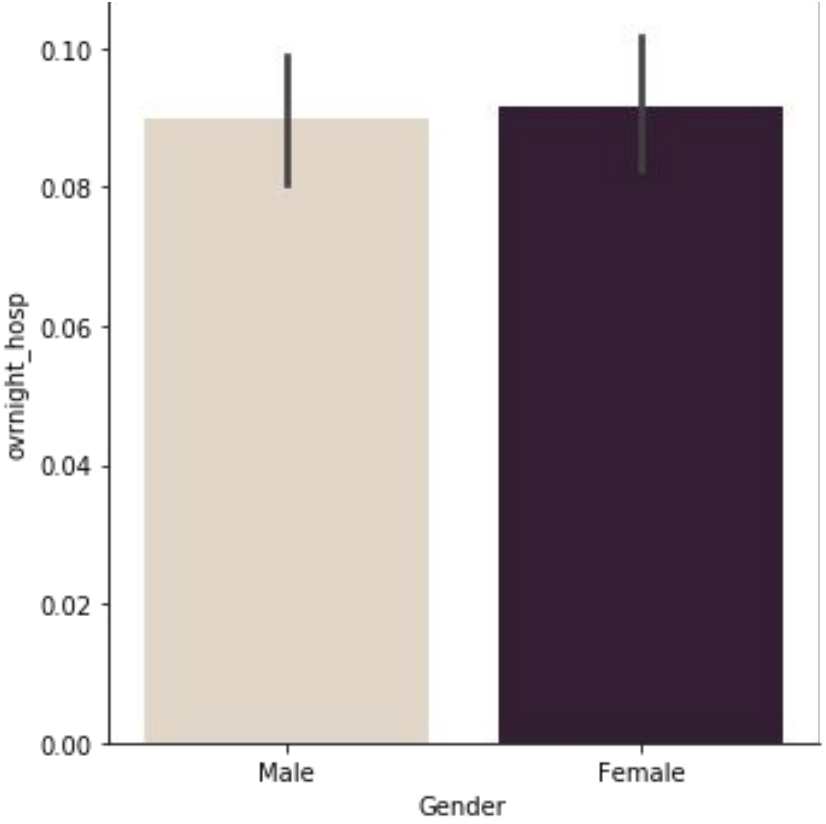
Descriptive Statistics



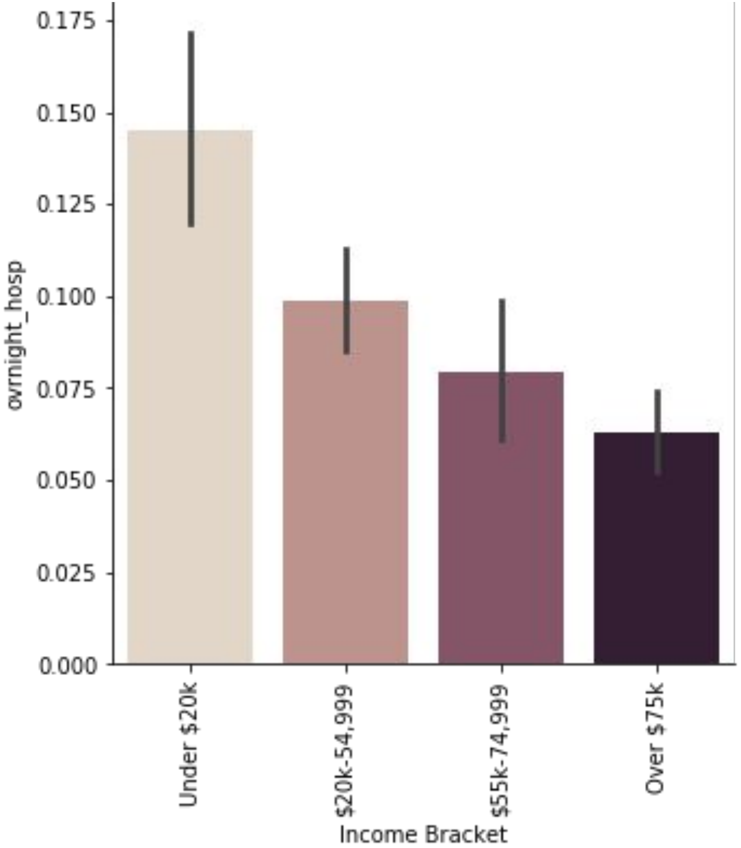
Age Group and Hospital Stays



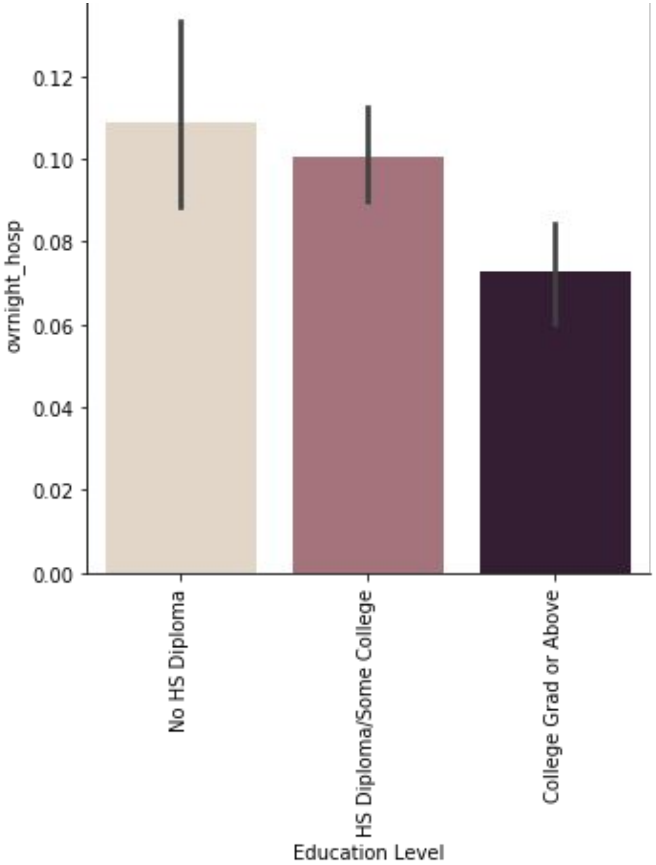
Gender and Hospital Stays



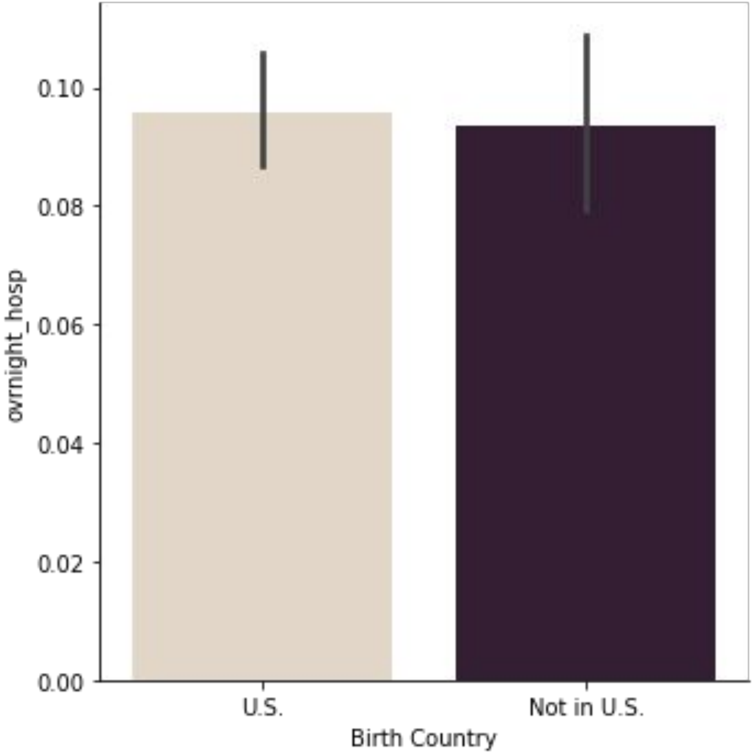
Income and Hospital Stays



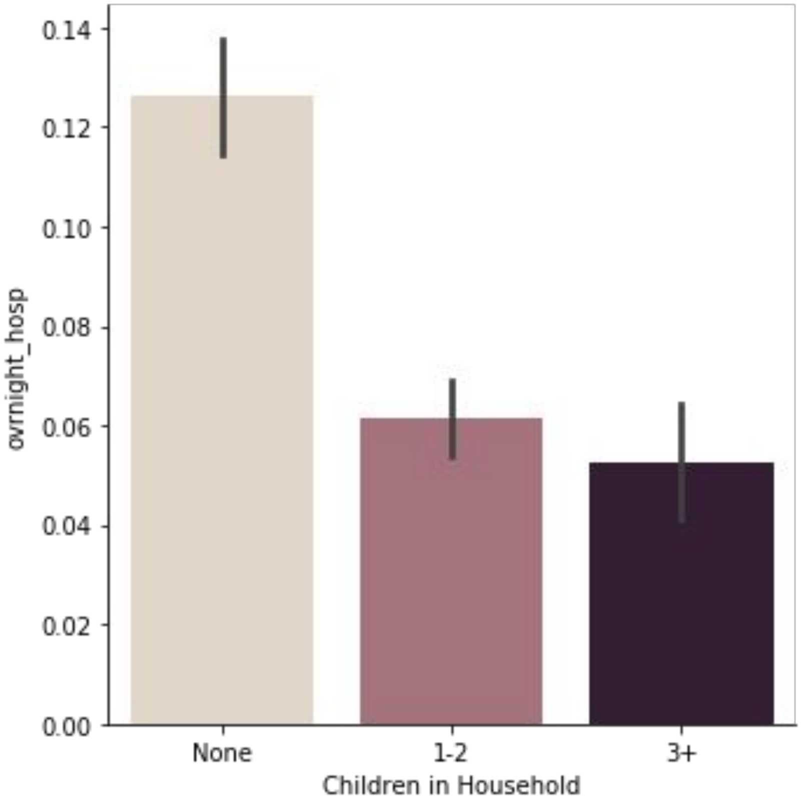
Education Level and Hospital Stays



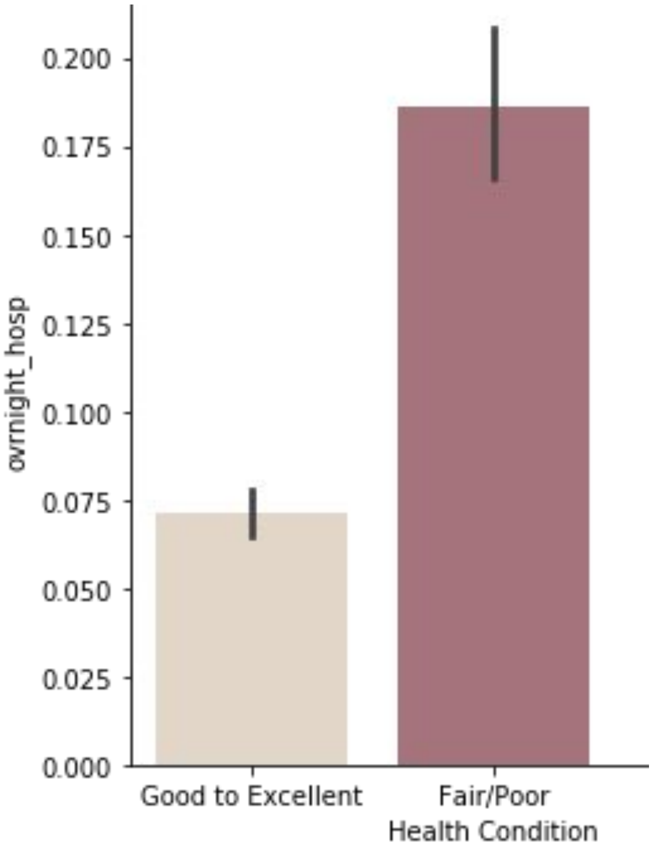
Country of Birth and Hospital Stays



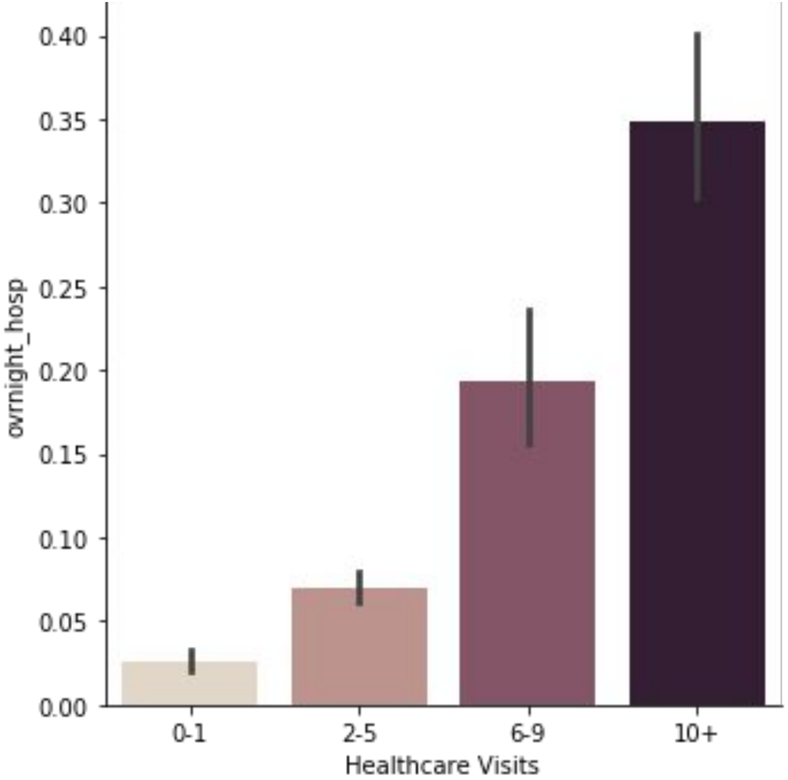
Children in Household and Hospital Stays



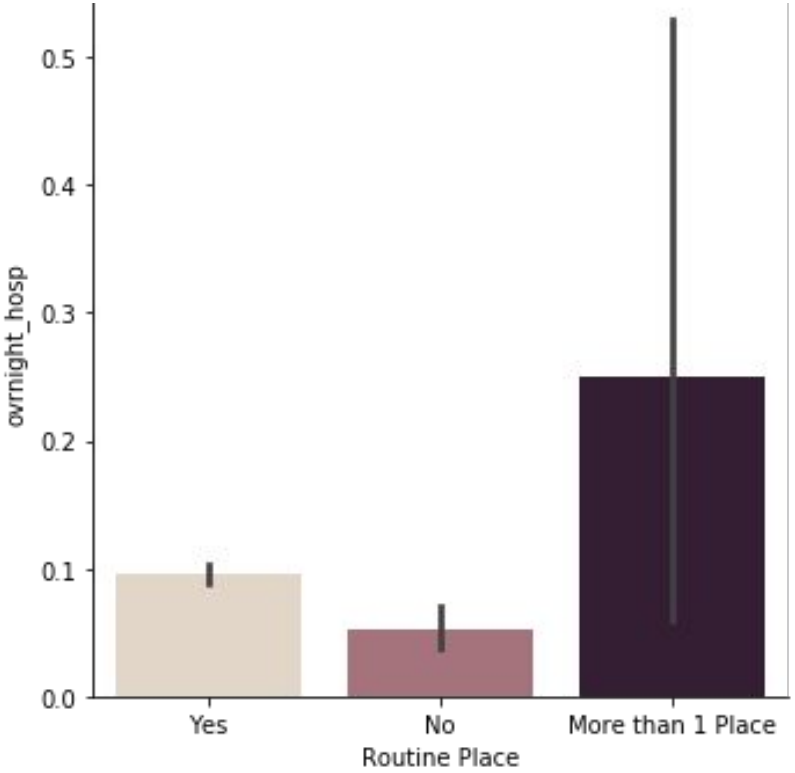
Health Condition and Hospital Stays



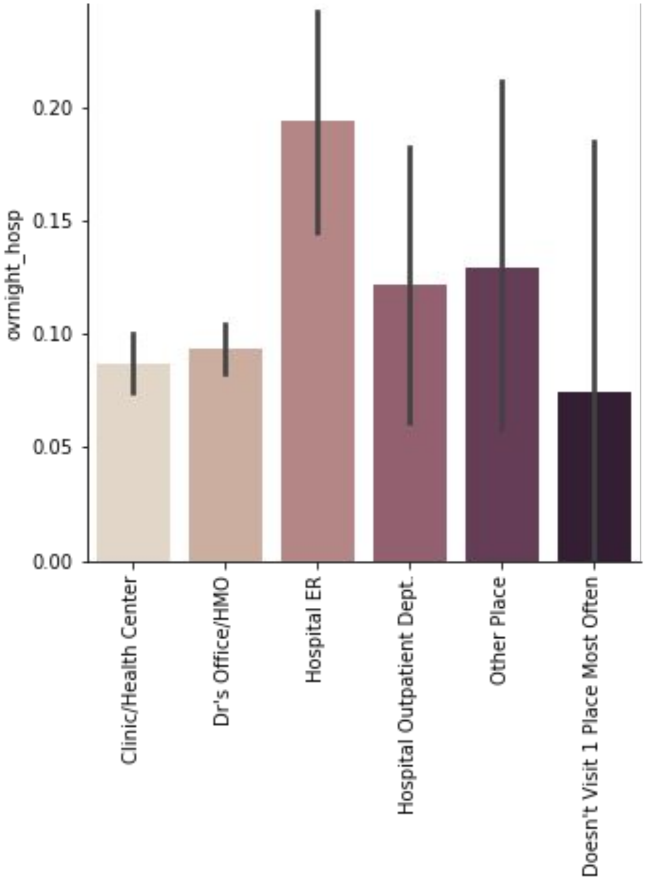
Number of Healthcare Visits and Hospital Stays



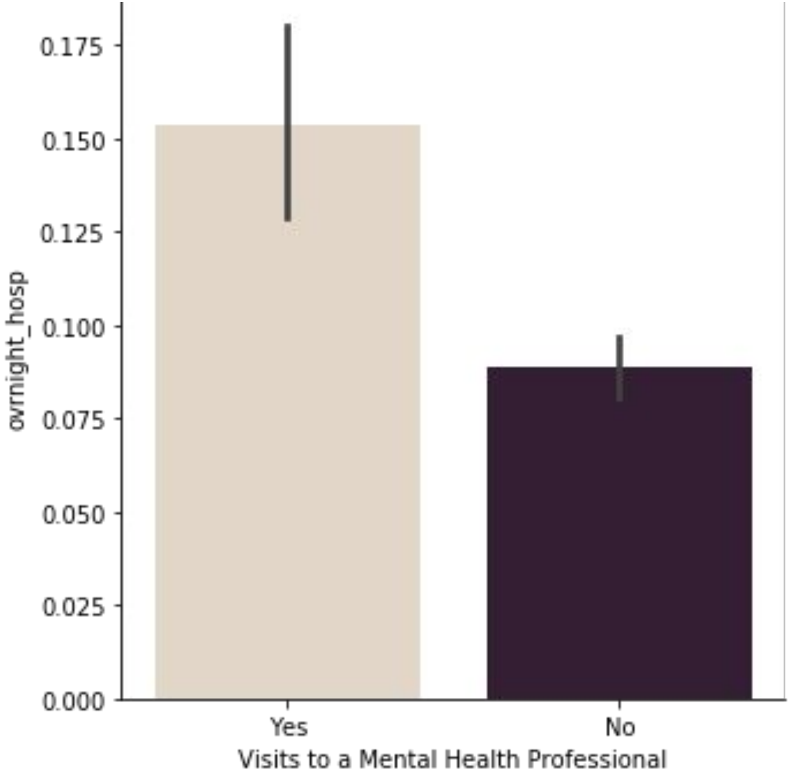
Routine Place for Healthcare and Hospital Stays



Type of Healthcare Facility Typically Visited and Hospital Stays



Visited Mental Health Professional and Hospital Stays



Methods

- Model Creation
- Model Optimization
- Model Evaluation

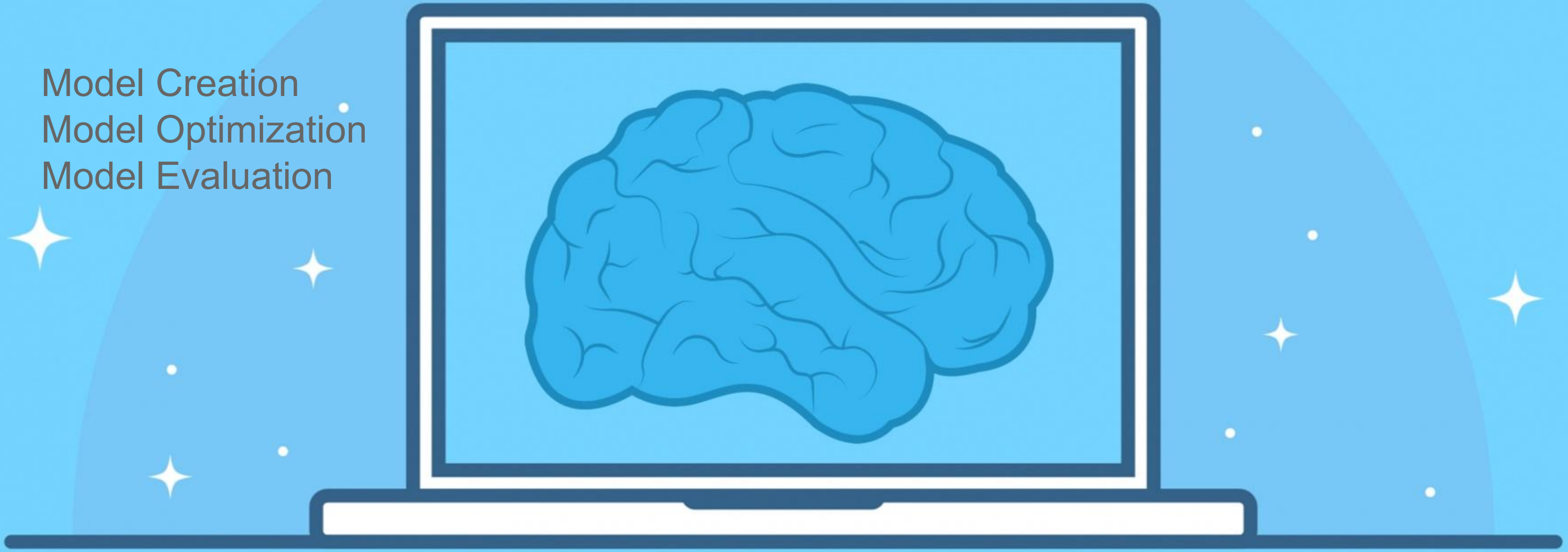


Image Credit: https://www.dataliftoff.com/wp-content/uploads/2019/10/machine-learning.jpg

Spearman Correlations

Variable	**Spearman Correlation with ovrnight_hosp**
Number of healthcare visits	.245
Health condition	.150
Elderly people in household	.113
Children in household	- .104
Proxy in person’s interview	.092
Health now vs. 1 year ago	-.071
Family size	-.056
Household size	-.054
Household income	-.053
Routine place for healthcare	-.052

Variable	**Spearman Correlation with ovrnight_hosp**
Family income	-.050
Education level	-.029
Mental health visits	-.026
Citizenship status	- .023
Interpreter in fam interview	.018
Marital status	.017
Interpreter in person’s interview	.014
Race/ethnicity	.013
Age group	.013

Logistic Regression

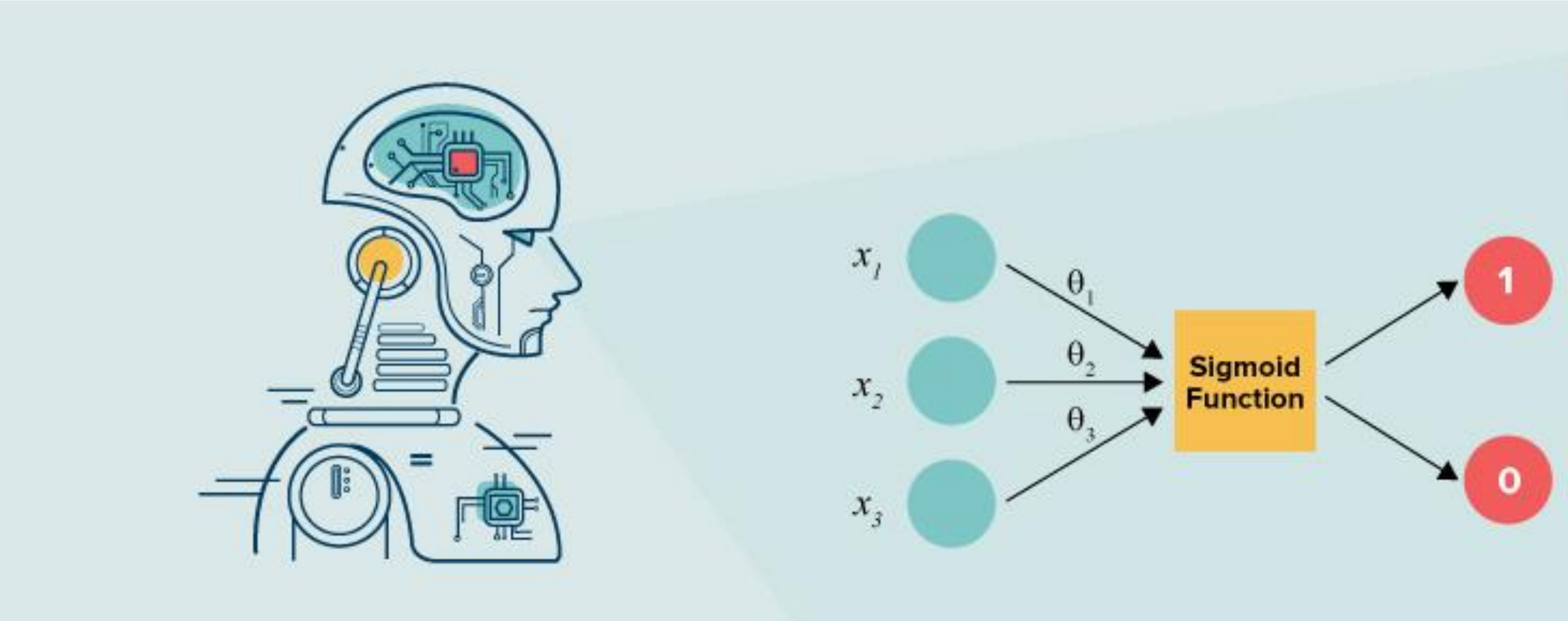


Image Credit: https://d2o2utebsixu4k.cloudfront.net/media/images/9a57ce9a-b10c-4ed0-9729-50d979af0a6f.jpg

Logistic Regression

Best: lbfgs, L2 regularization, 2000 max_iter

Mean CV score: .916

AUC score: .756

Random Forest

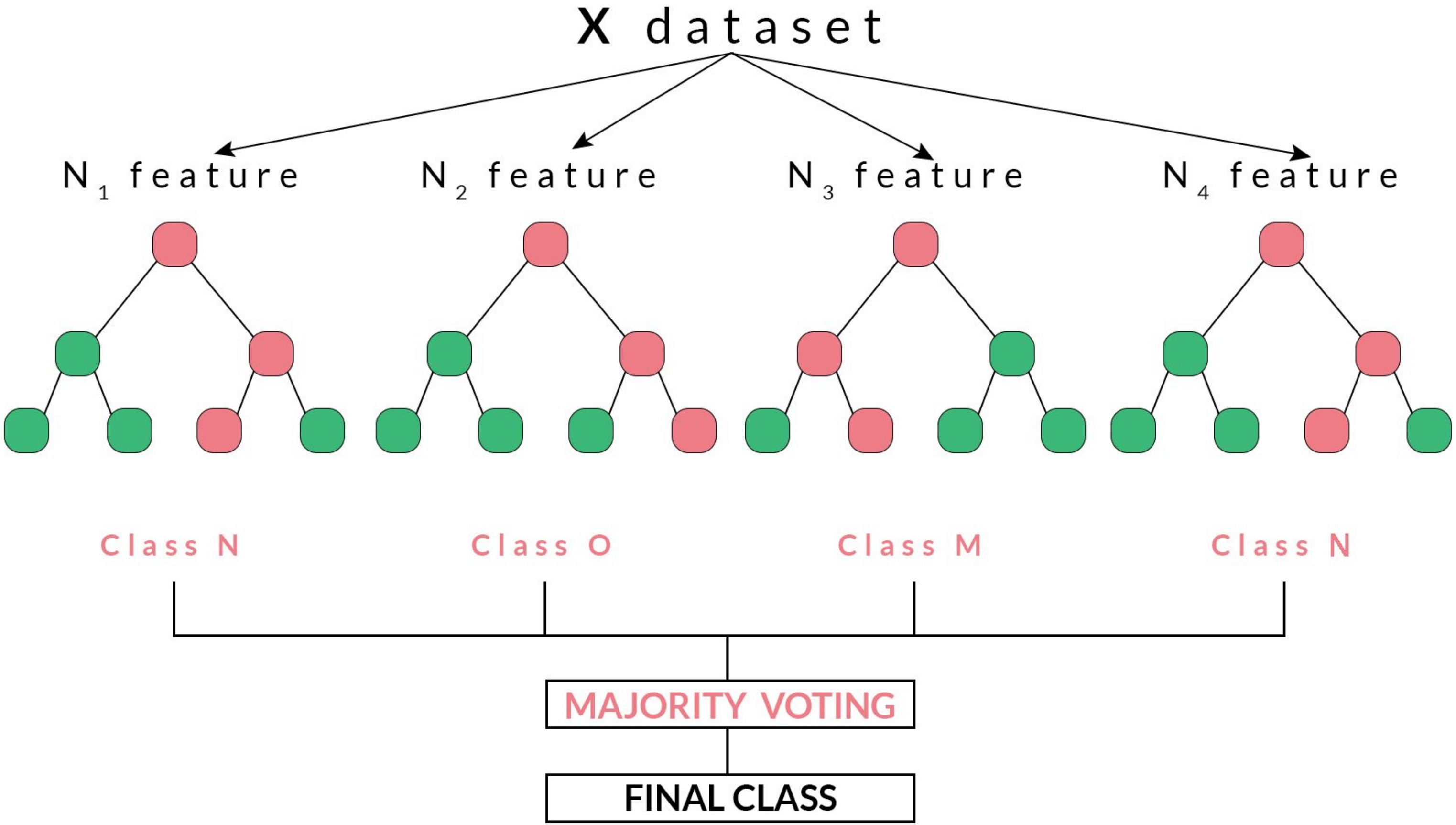


Image Credit: https://d1rwhvwstyk9gu.cloudfront.net/2019/03/Random-Forest-Algorithm.jpg

Random Forest

Best: n_estimators: 250, max_depth: 20

Mean CV score: .901

AUC score: .992

Gradient Boosting Model

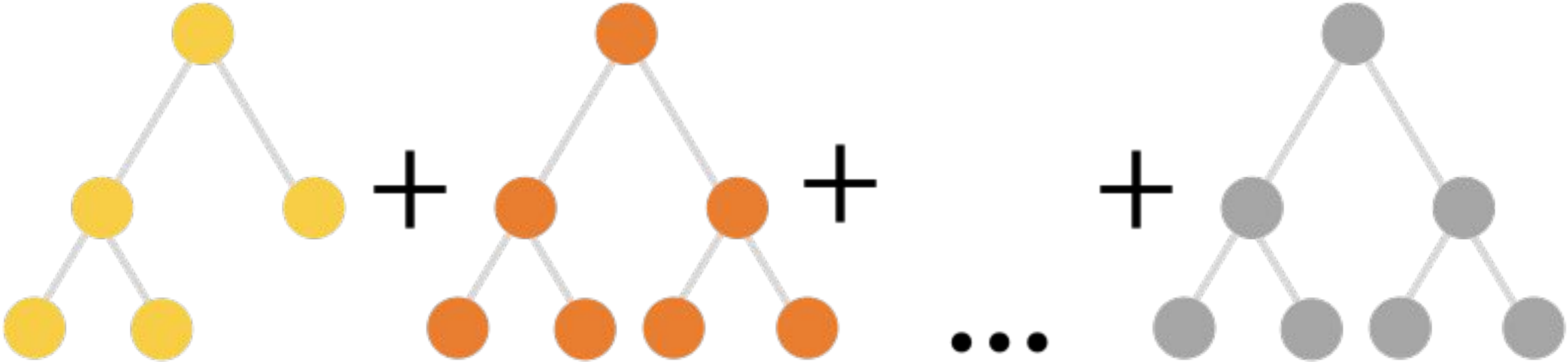


Image Credit: https://avatars.mds.yandex.net/get-yablogs/49865/file_1548410211030/orig

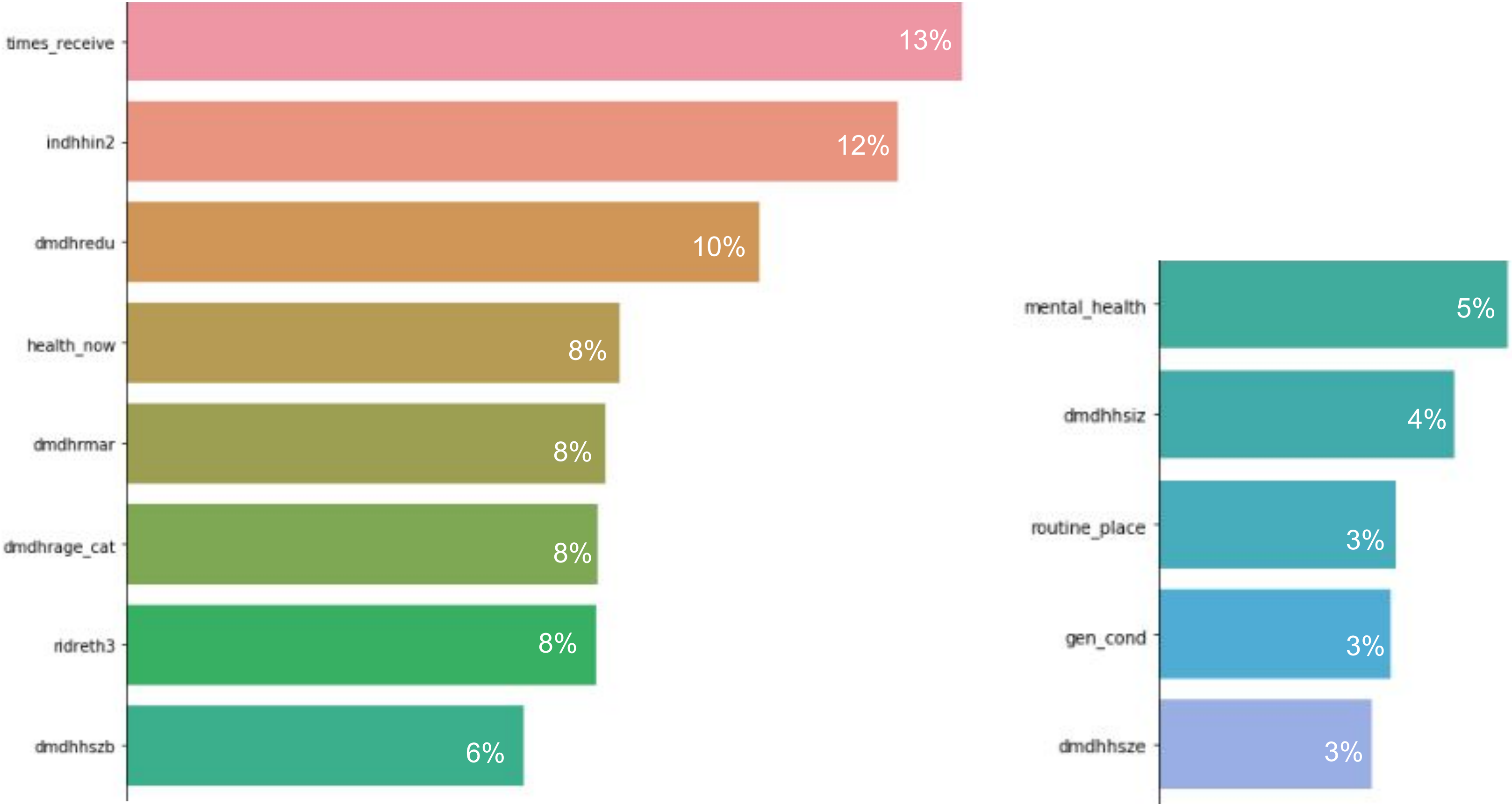
GBM

Final model

Best: n_estimators: 200, max_depth: 15, learning_rate: .007, max_features: 7

- Mean CV score: .905
- AUC score: .993

GBM Feature Importance



Statistical Logistic Regression

- Ran statistical logistic regression (using statsmodels.Logit)

  Removed *general health condition* and *race/ethnicity*, which improved the model:
  - AIC score < by **.41**
  - BIC score < by **7.62**
  - Log-likelihood < by **1**

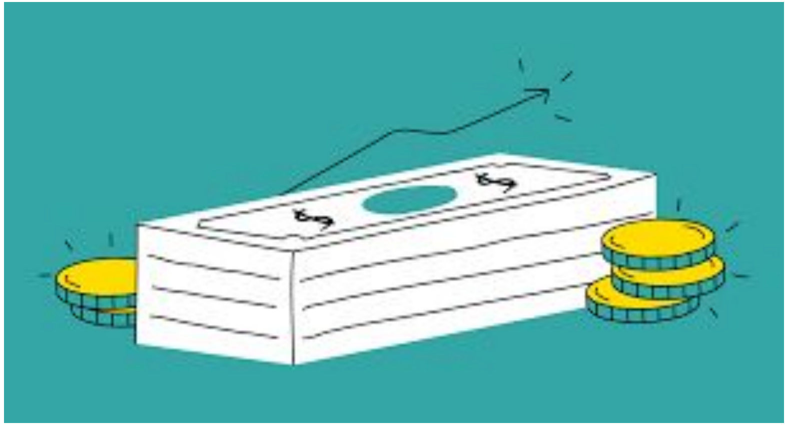
Re-Running GBM
Mean CV score: .898
AUC score: .983

Odds Ratios

Number of healthcare visits: 1.52
For each 1 point increase in the number of times participants received healthcare during the year, they are **52% more likely** to have stayed overnight in the hospital in the last year.

Odds Ratios

Income Bracket: .68
For each 1 point increase in participants' household income, they are **32% less likely** to have stayed overnight in the hospital in the last year.



Odds Ratios

Education Level: .58
For each 1 point increase in participants' level of education, they are **42% less likely** to have stayed in the hospital overnight in the last year.

Odds Ratios

Health now vs. 1 year ago: 1.00
For each 1 point increase in participants' rating of their health now as compared to 1 year ago, they are **equally as likely** to have stayed overnight in the hospital in the last year.

Odds Ratios

Marital Status: .70
For each 1 point increase in participants' marital status, they are **30% less likely** to have stayed overnight in the hospital in the last year.

Limitations

- Keeping the variables the way NHANES had them coded, which is many cases is backwards (e.g., 0 means ‘yes’, 1 means ‘no’); makes interpretation somewhat more confusing

- Condensing the variables into smaller code ‘buckets’ rather than using dummy variables for each specific research code most likely affected accuracy of at least some models

- The race/ethnicity variable and the categorical age group variable were multicollinear, which most likely affected the accuracy of the logistic regression

Conclusion & Implications

Conclusion

- Yes, NHANES data can predict overnight hospitalization fairly well.

Findings/Implications

- Underlying or recurrent health conditions are the major influence on hospitalizations
- Class and/or social status are major predictors as well
- Hospitals/insurance companies should develop programs that target those demographics most at risk of overnight hospitalization, especially when hospitalization may be rendered unnecessary through preventive care

Future Research Possibilities

- Model using only health science variables as opposed to characteristic variables
- See how well this model predicts overnight hospitalization in other NHANES years