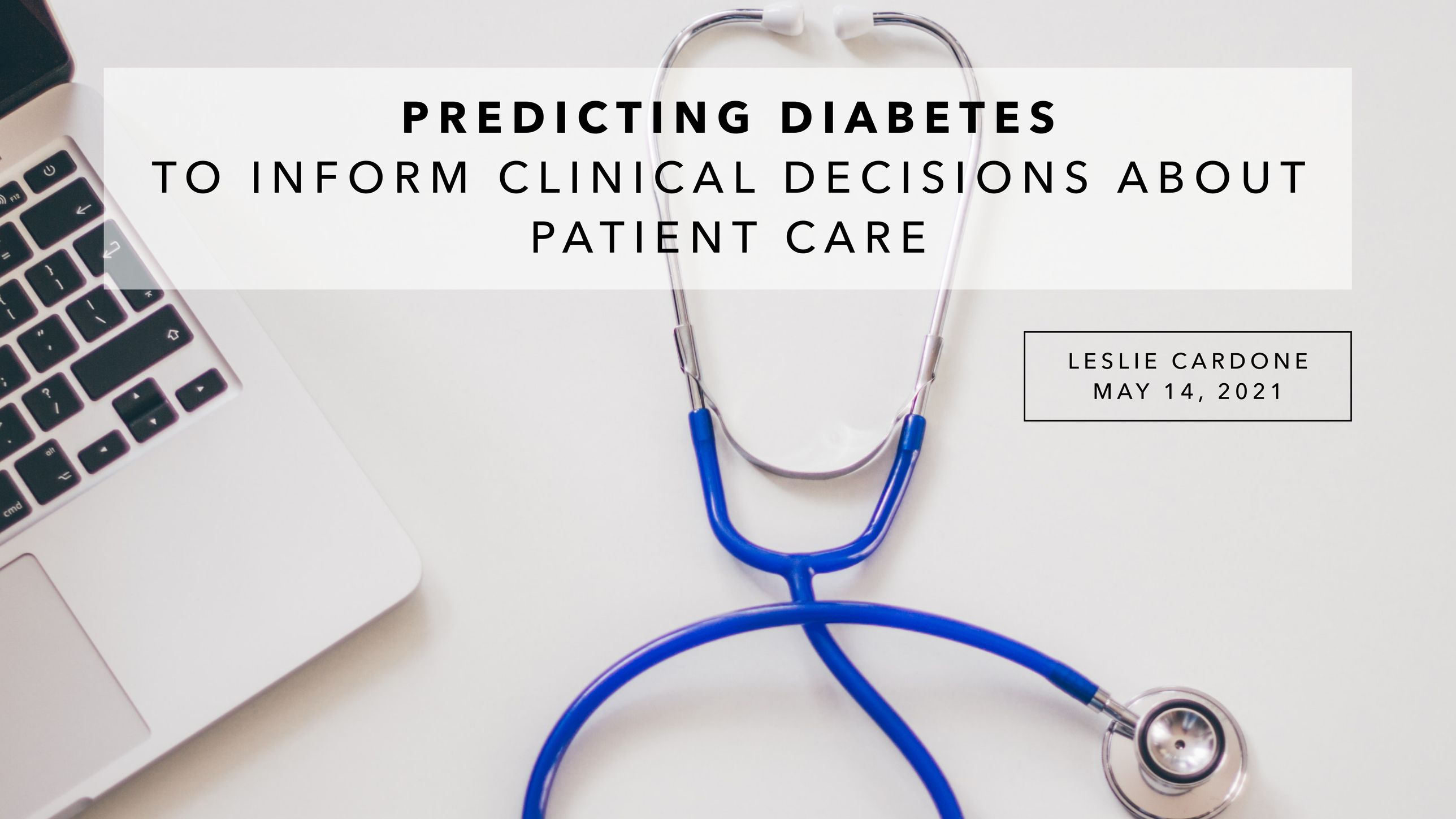


PREDICTING DIABETES

TO INFORM CLINICAL DECISIONS ABOUT PATIENT CARE

LESLIE CARDONE

MAY 14, 2021

GOALS

Predict diabetes (all types) using data collected during patients first 24 hours in the ICU.

Minimize false negatives / maximize true positives

Target is binary

1 Has been diagnosed with diabetes

0 Has not been diagnosed

DATA

kaggle

130,157 SAMPLES

180 FEATURES

CATEGORICAL

CONTINUOUS

> 3% CORRELATION

> 3% CORRELATION

UNDER 60% NAN

36 FEATURES

METHODS

LOGISTIC REGRESSION BASELINE #1

ONE FEATURE (HIGHEST GLUCOSE CONCENTRATION)

FILL ALL NAN WITH MEDIAN OF FEATURE

LOGISTIC REGRESSION BASELINE #2

15 FEATURES – ALL CONTINUOUS VARIABLES

SCALE AND REGULARIZE

FILL ALL NAN WITH MEDIAN OF FEATURE

METHODS

SCORING METRIC	MODEL #1	MODEL #2
ACCURACY		
PRECISION		
RECALL	0.2115	0.2562
F1	0.3115	0.3579

LOW RECALL

HIGH FALSE NEGATIVES

METHODS

LOGISTIC REGRESSION FINAL MODEL

ALL 36 FEATURES

FILL ALL NAN WITH KNN PREDICTION

SCALE AND REGULARIZE (L1 - LASSO)

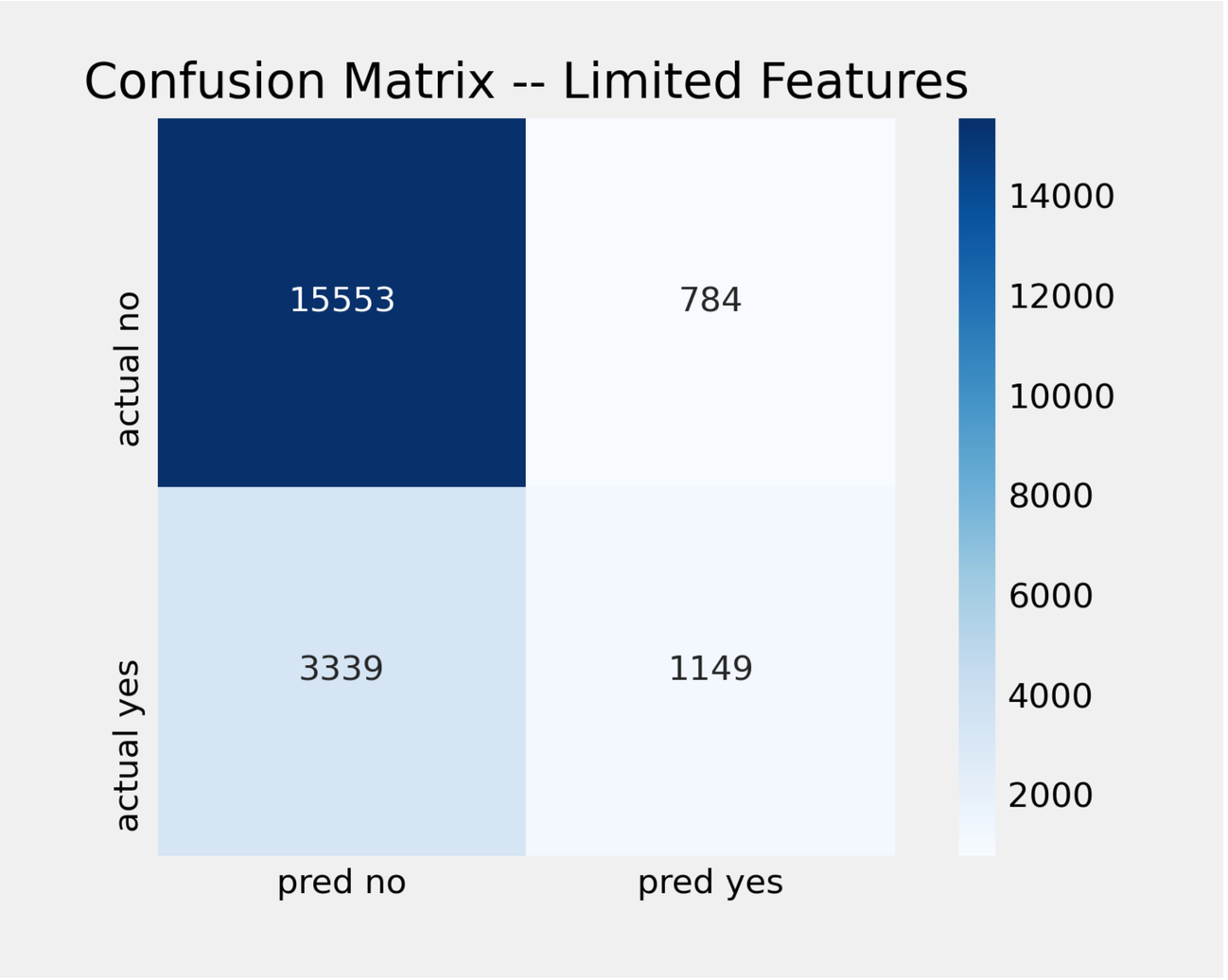
TUNED WEIGHTS AND C VALUE

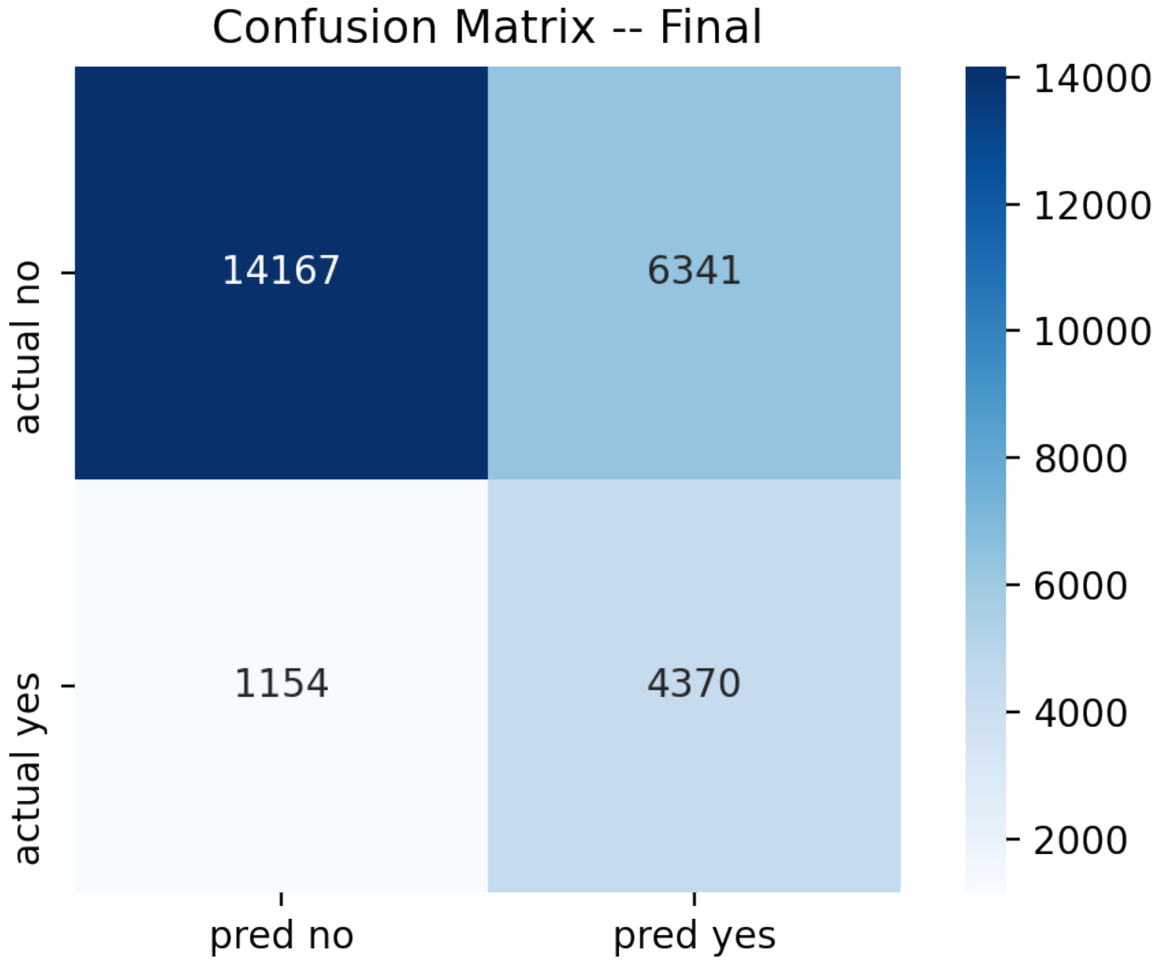
SCORE ON HOLD OUT SET

METHODS

SCORING METRIC	MODEL #1	MODEL #2	FINAL
ACCURACY	0.798	0.802	0.712
PRECISION	0.5913	0.5943	0.4080
RECALL	0.2115	0.2562	0.791
F1	0.3115	0.3579	0.538

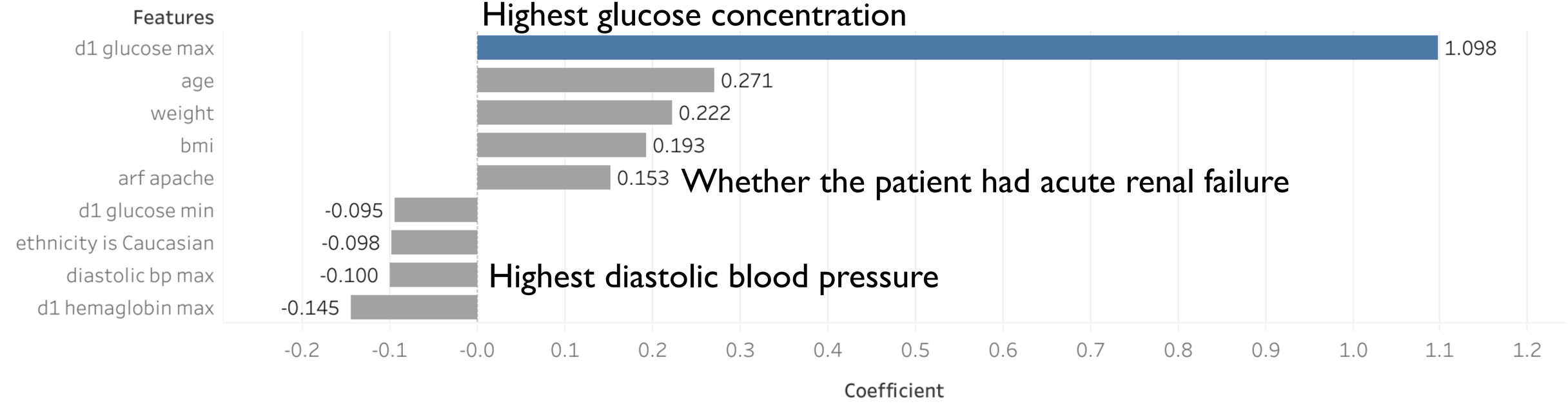
FINDINGS


Confusion Matrix -- Limited Features
actual no
actual yes
15553
784
3339
1149
pred no
pred yes
14000
12000
10000
8000
6000
4000
2000


Confusion Matrix -- Final
actual no
actual yes
14167
6341
1154
4370
pred no
pred yes
14000
12000
10000
8000
6000
4000
2000

FINDINGS

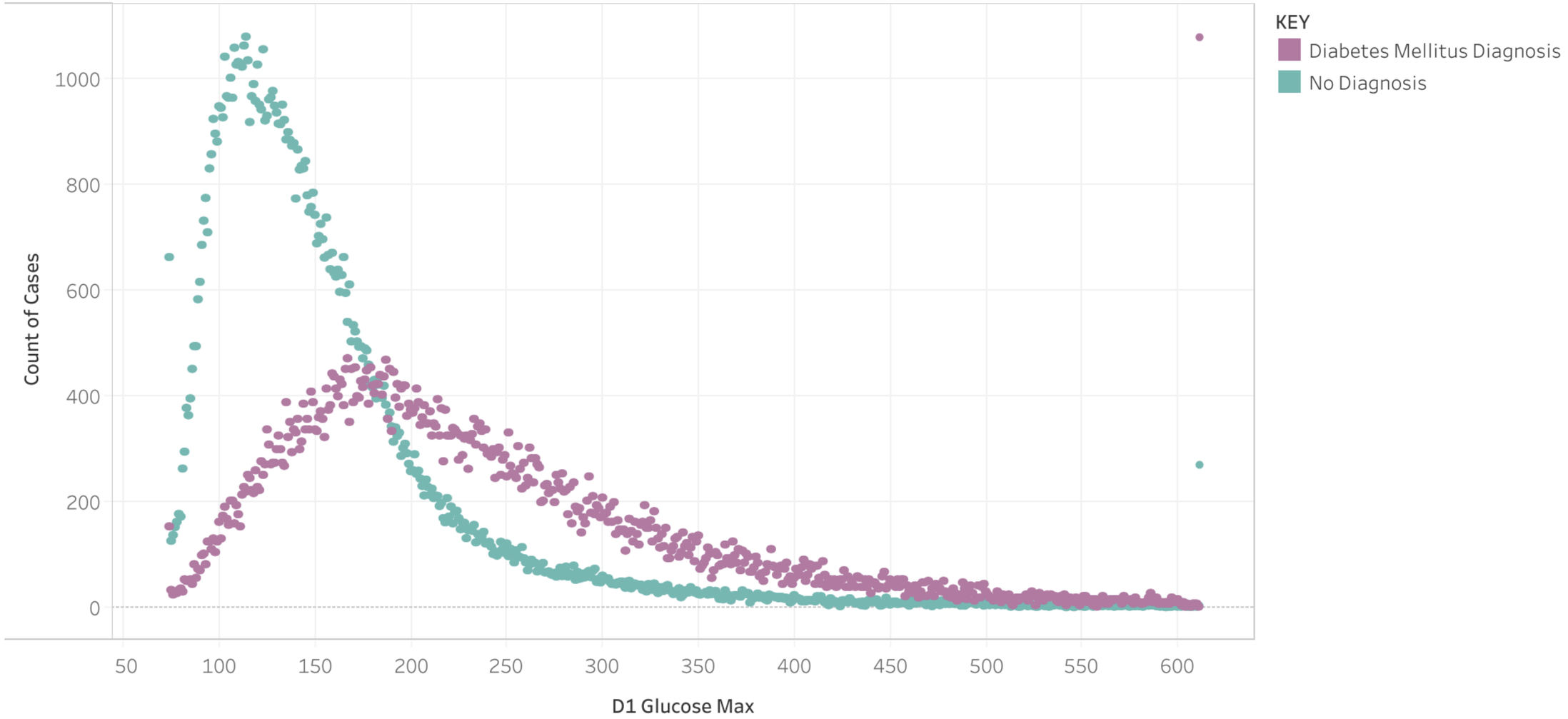
Feature "Importance"



19 OF 36 FEATURES AFTER REGULARIZATION

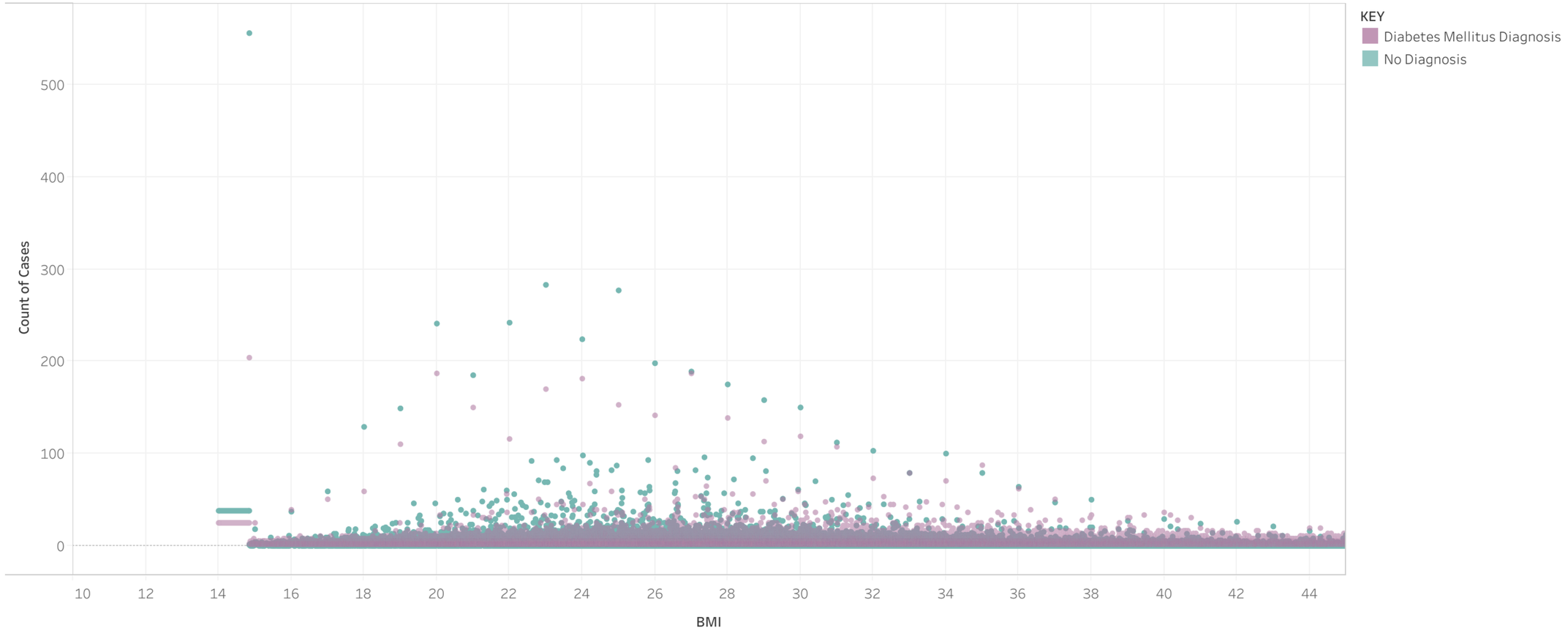
FINDINGS

Distribution of Diabetes Diagnosis



FINDINGS

Distribution of Diabetes Diagnosis



CONCLUSIONS

Class imbalance significantly affects the predictive capabilities of the logistic regression classification model

Having features that are highly correlated with the target are also important

Complicated imputation did not improve any scoring metrics by much

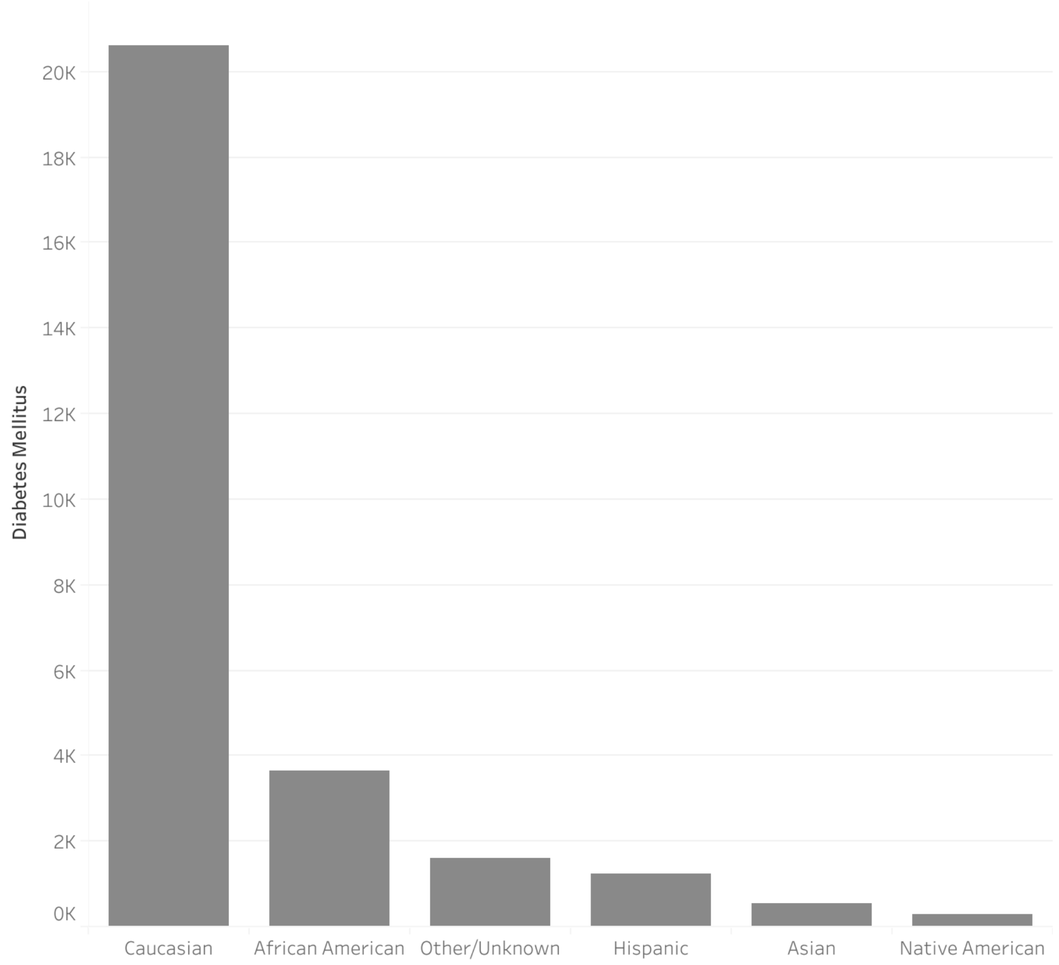
FURTHER STUDIES

More feature engineering, create feature interactions

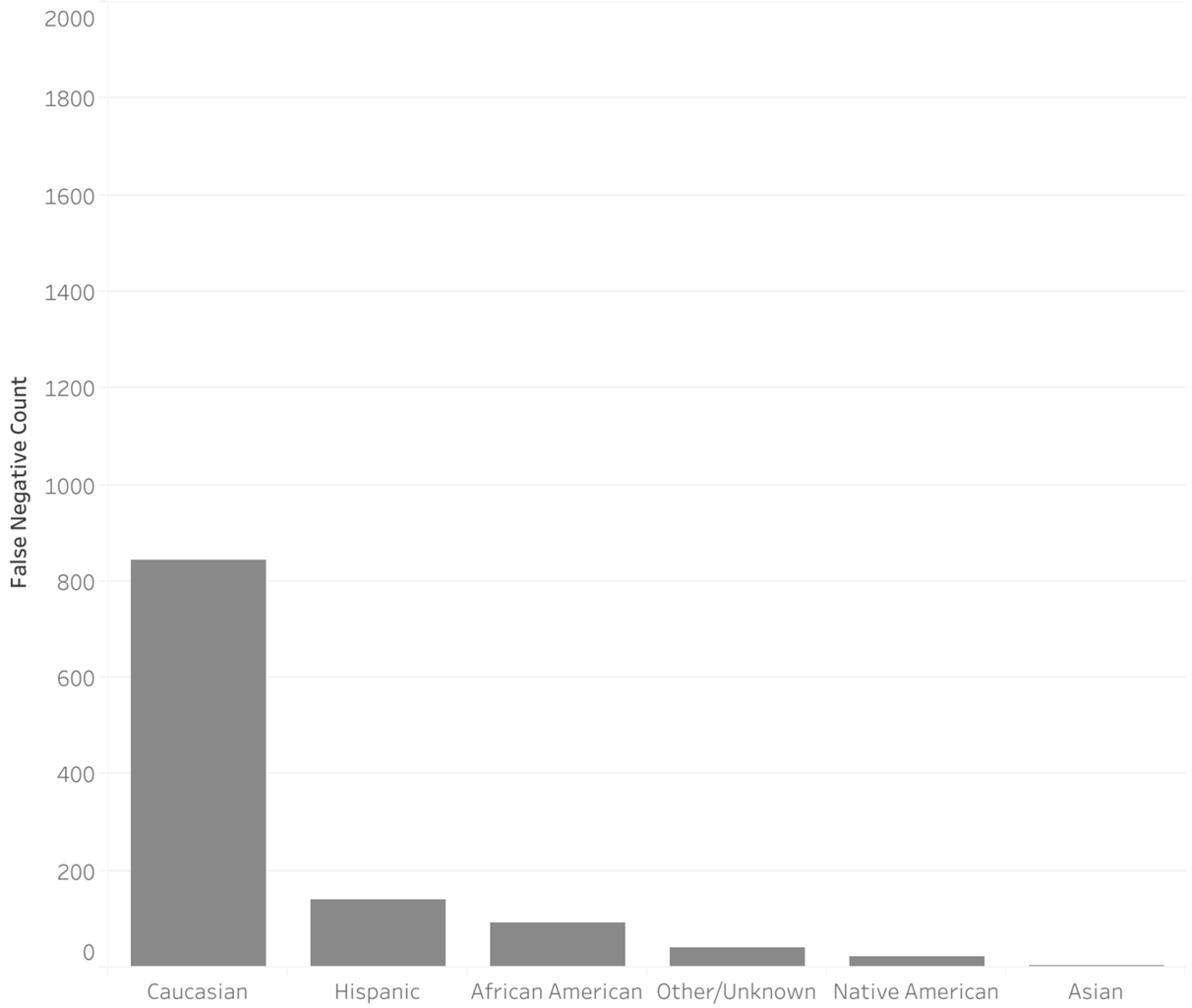
Try Gradient Boosting / XGBoost

APPENDIX

Diabetes By Ethnicity

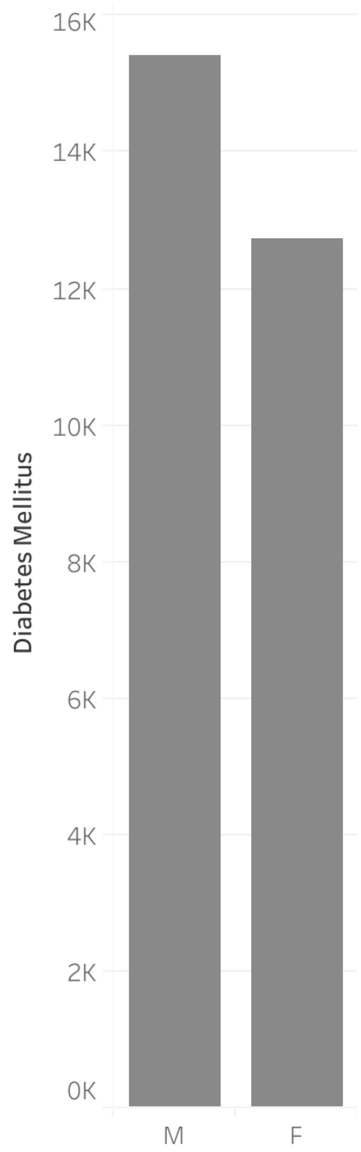


Diabetes Misclassification By Ethnicity

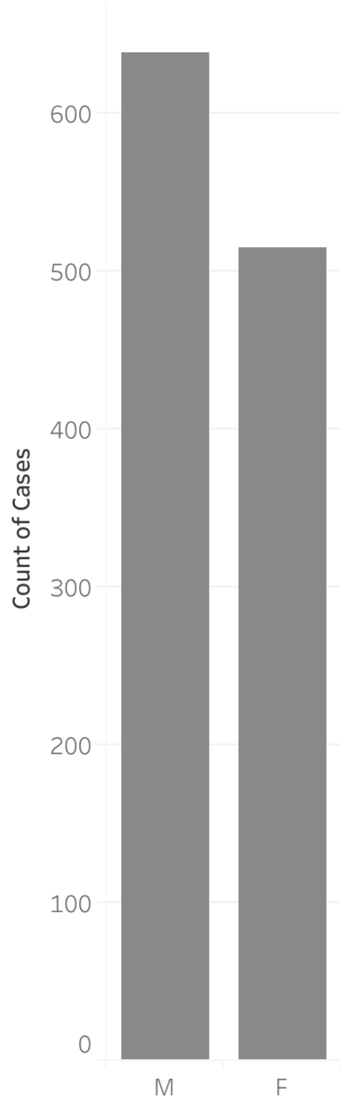


APPENDIX

Count of Diabetes Cases by Gender



Count of False Negatives by Gender



APPENDIX

Distribution of False Negatives by Age