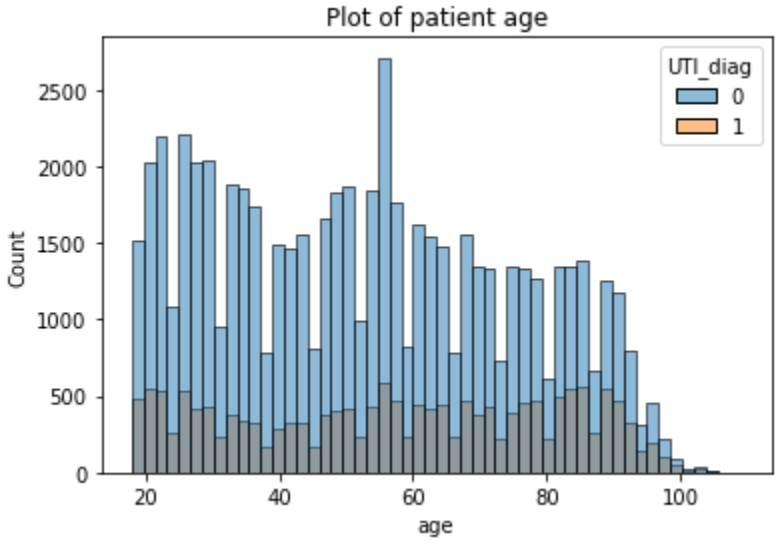
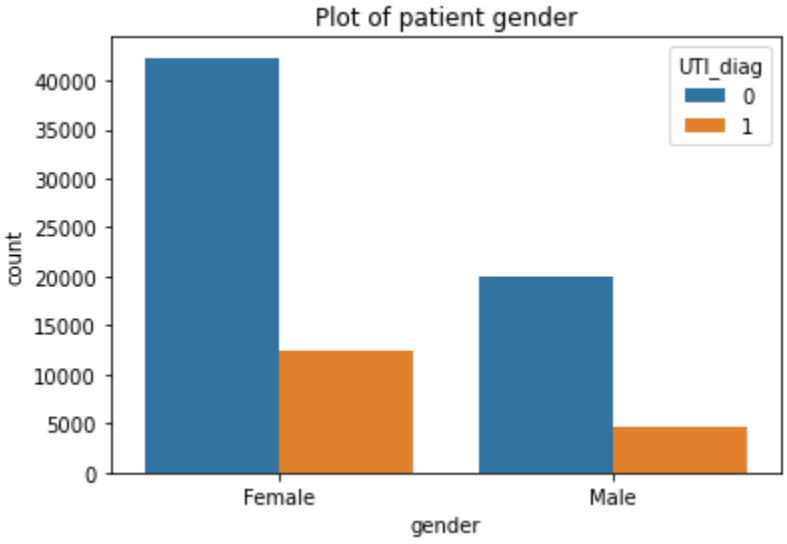
Peng Shen

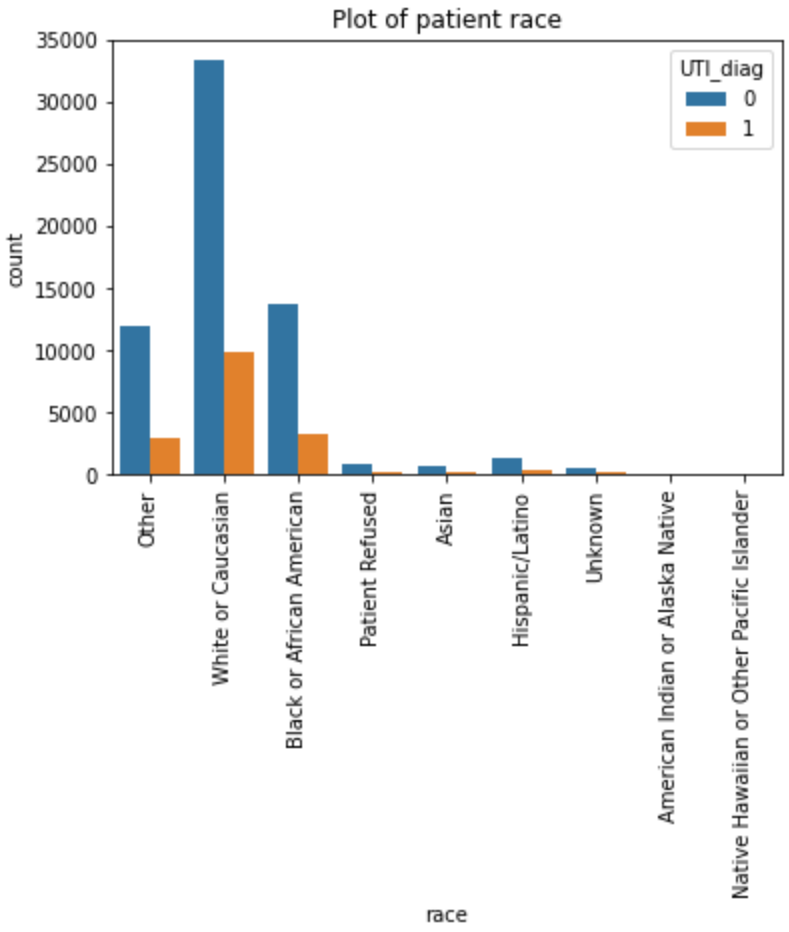
Visualizing Descriptive



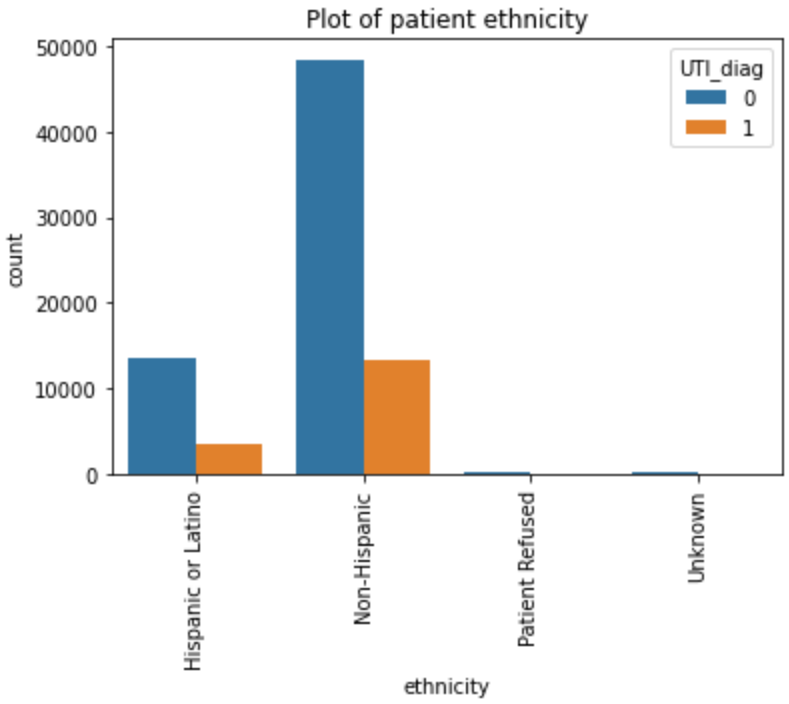
Most of the subjects are between 20 to 80. The proportion of having urinary tract infection is higher for subjects of older ages.



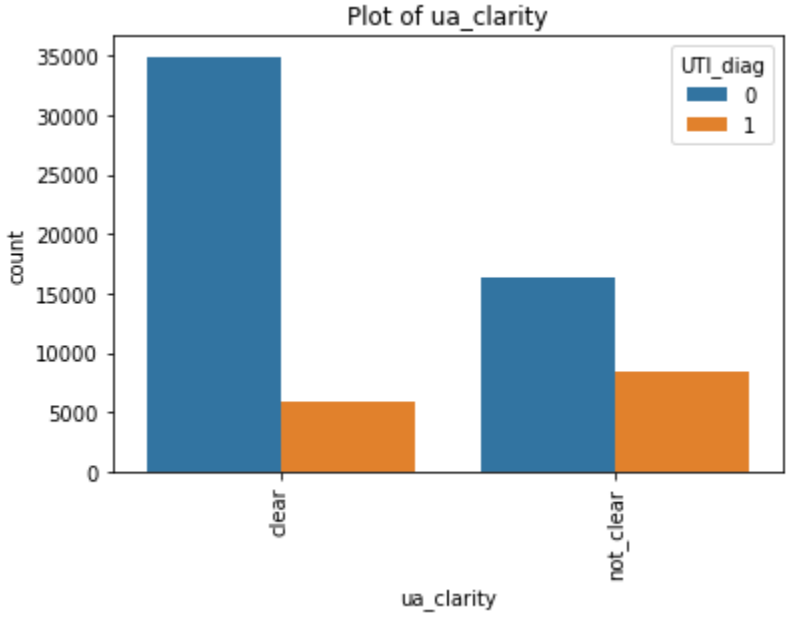
There are more female subjects than male subjects. The proportion of having urinary tract infection is higher for female than male.



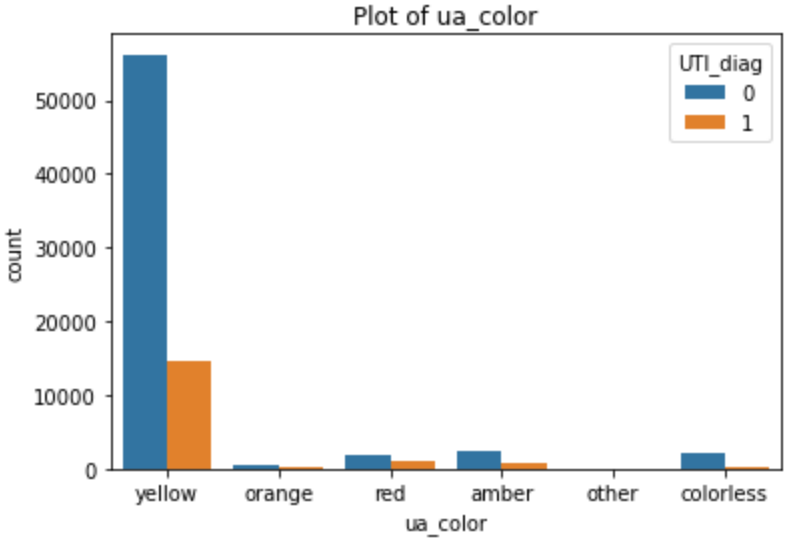
Most of the subjects are white and African American. The study has very few Asian, Latino and Native American. The proportion of having urinary tract infection is among the highest for white people.



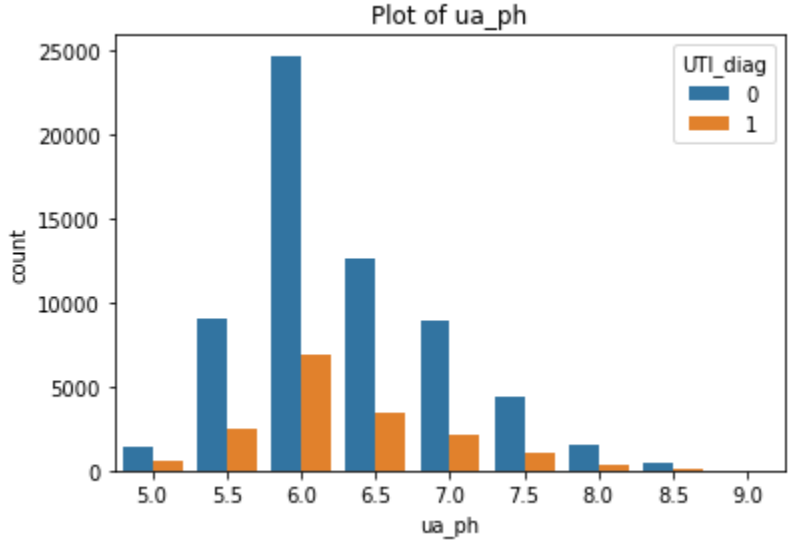
The proportion of having urinary tract infection is higher for non-Hispanic than for Hispanic.



The proportion of having urinary tract infection is higher for subjects with cloudy urine than for subjects with clear urine.



Most of the subjects have yellow urine. But the proportion of having urinary tract infection is highest for subjects with red urine.



Most of the urine PH is in the range of 5.5 to 7.0. The proportion of having urinary tract infection is seemingly not affected by urine's PH.

Short Answer Questions

Q1. There are a total of 85 features selected to include in my ML models. There are 'ua_bacteria', 'ua_blood', 'ua_clarity', 'ua_epi', 'ua_leuk', 'ua_nitrite', 'ua_ph', 'ua_protein', 'ua_rbc', 'ua_spec_grav', 'ua_urobili', 'ua_wbc', 'abd_tenderness', 'fever', 'abd_pain', 'dysuria', 'chief_complaint', 'age', 'race', 'maritalStatus', 'employStatus', 'insurance_status', 'arrival', 'Temp_First', 'Temp_Last', 'Temp_Max',

'Temp_Min', 'Temp_Mean', 'HR_First', 'HR_Last', 'HR_Max', 'HR_Min', 'HR_Mean', 'SBP_First', 'SBP_Last', 'SBP_Max', 'SBP_Min', 'SBP_Mean', 'DBP_First', 'DBP_Last', 'DBP_Max', 'DBP_Min', 'DBP_Mean', 'RR_First', 'RR_Last', 'RR_Max', 'RR_Min', 'O2_Sat_First', 'O2_Sat_Last', 'O2_Sat_Max', 'O2_Sat_Min', 'O2_Sat_Mean', 'GCS_First', 'GCS_Last', 'Absolute_Lymphocyte_Count', 'Alanine_Aminotransferase', 'Alkaline_Phosphatase', 'anc', 'Anion_Gap', 'Aspartate_Aminotransferase', 'Basophils', 'bun', 'Calcium', 'Chloride', 'co2', 'Creatinine', 'Eosinophils', 'Glucose', 'Hematocrit', 'Hemoglobin', 'Lymphocytes', 'mch', 'mchc', 'mcv', 'Monocytes', 'mpv', 'Neutrophils', 'Platelets', 'Potassium', 'rbc', 'rdw', 'Sodium', 'wbc', 'abx', 'antibiotics'.

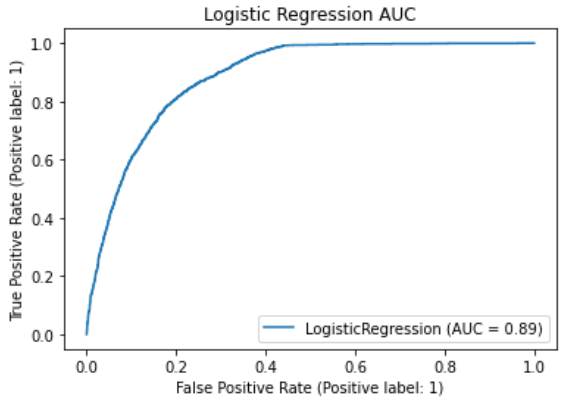
The features are selected with random forest. Random forest generate feature importance and the features whose absolute importance value is greater or equal than the mean importance are kept while the others are discarded.

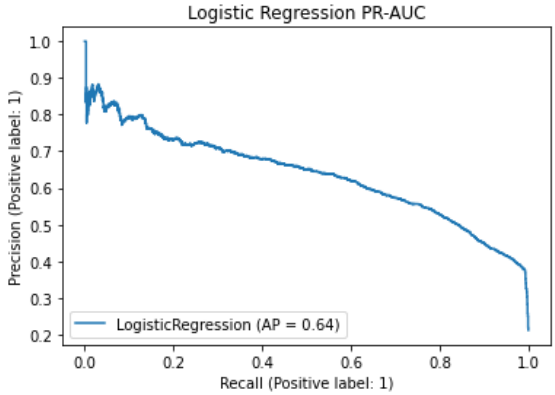
Q2. The dataset is breaking up into training, validation, and testing dataset with a proportion of 80%, 10% and 10%. The training dataset is used to build the model. The validation dataset is used to tune the parameters of the model such as random forest and xgboost to maximize the model performance on the validation set. The test set is used to understand the model performance on future data.

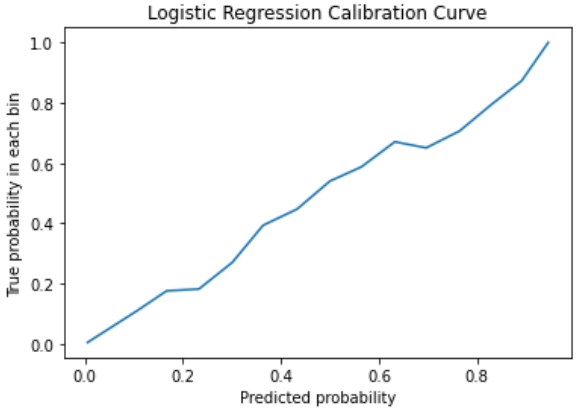
Q3. Data leakage refers to the use of data or information during the model training which should not be provided, causing the performance metric to be overestimated. In our case, if during the model training part, we include both training set and testing test. Then, when we implement the model later, on the test set, the AUC may be higher than it should be.

Model performance plots

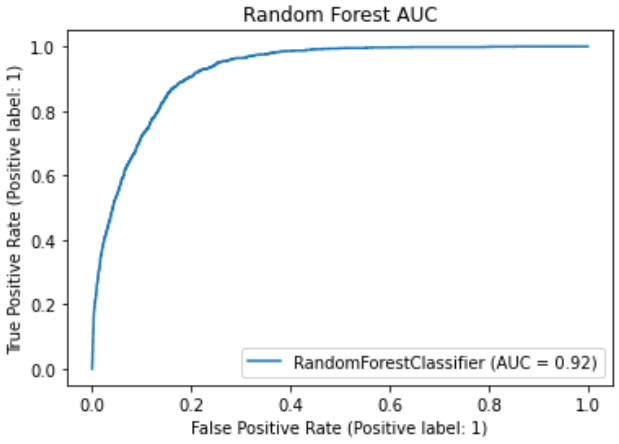
For logistic regression, the AUC is 0.89, the PR-AUC is 0.64 and the Caliberation curve is shown below. The AUC is high indicating the model has a decent fit.



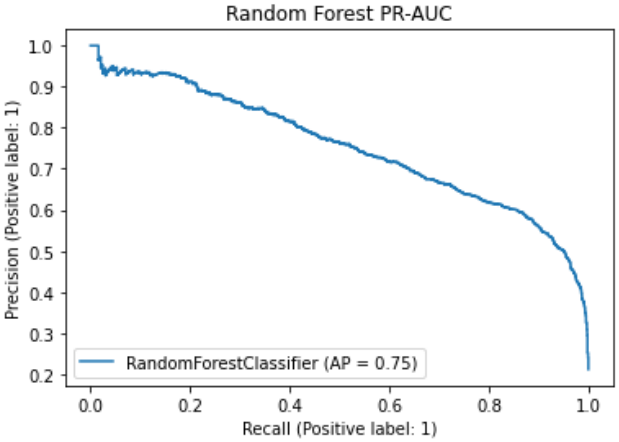


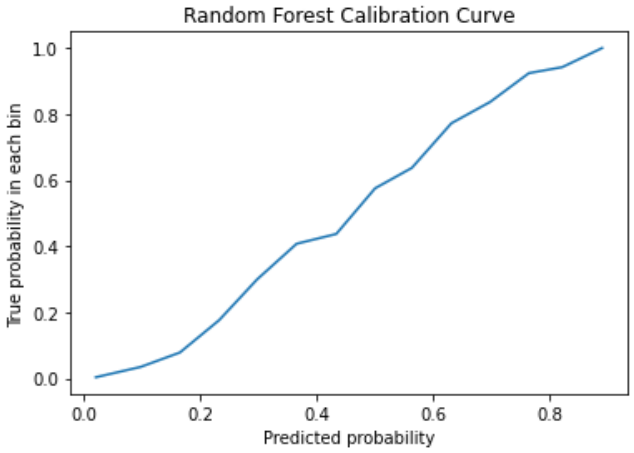


For random forest, the AUC is 0.92, the PR-AUC is 0.75 and the Caliberation curve is shown below. The random forest's AUC and the PR-AUC is higher for logistic regression, indicating random forest is a better model for this dataset.

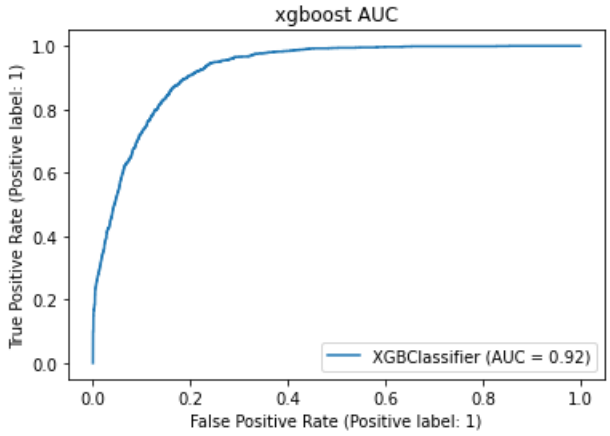


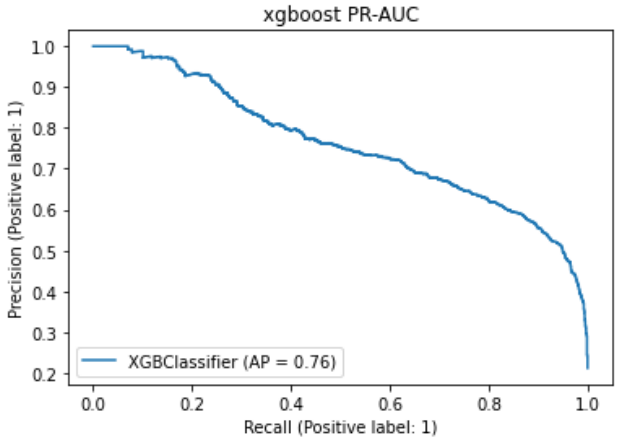
‘

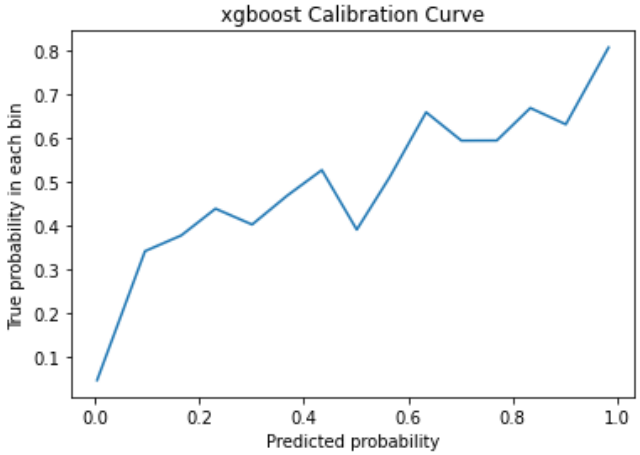




For xgboost, the AUC is 0.92, the PR-AUC is 0.76 and the Caliberation curve is shown below. The xgboost’s AUC and the PR-AUC is very similar to random forest’s.


xgboost AUC
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
XGBClassifier (AUC = 0.92)
0.0
0.2
0.4
0.6
0.8
1.0


xgboost PR-AUC
Precision (Positive label: 1)
Recall (Positive label: 1)
XGBClassifier (AP = 0.76)
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0


xgboost Calibration Curve
True probability in each bin
Predicted probability
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8