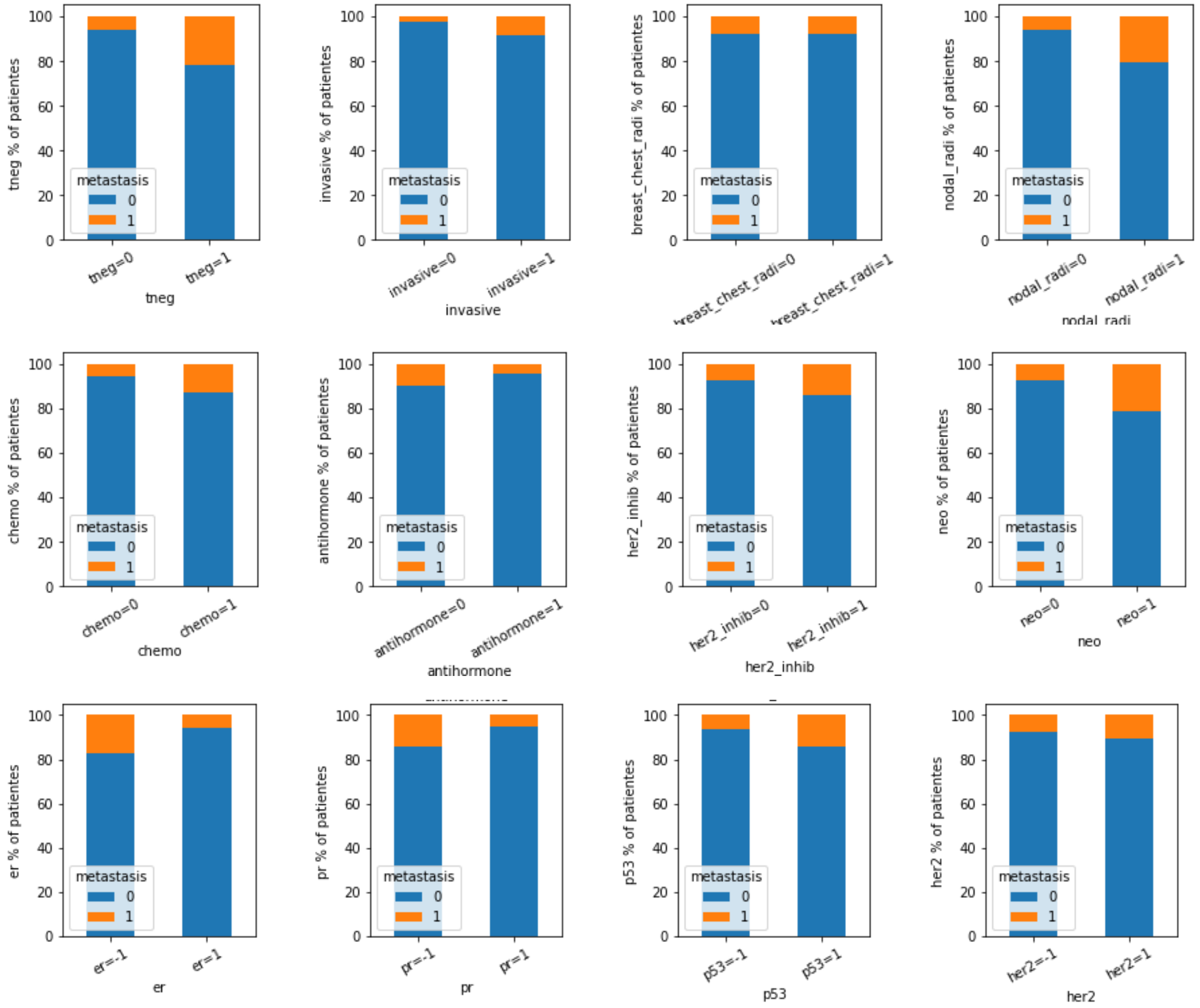
Cancer metastasis prediction tool

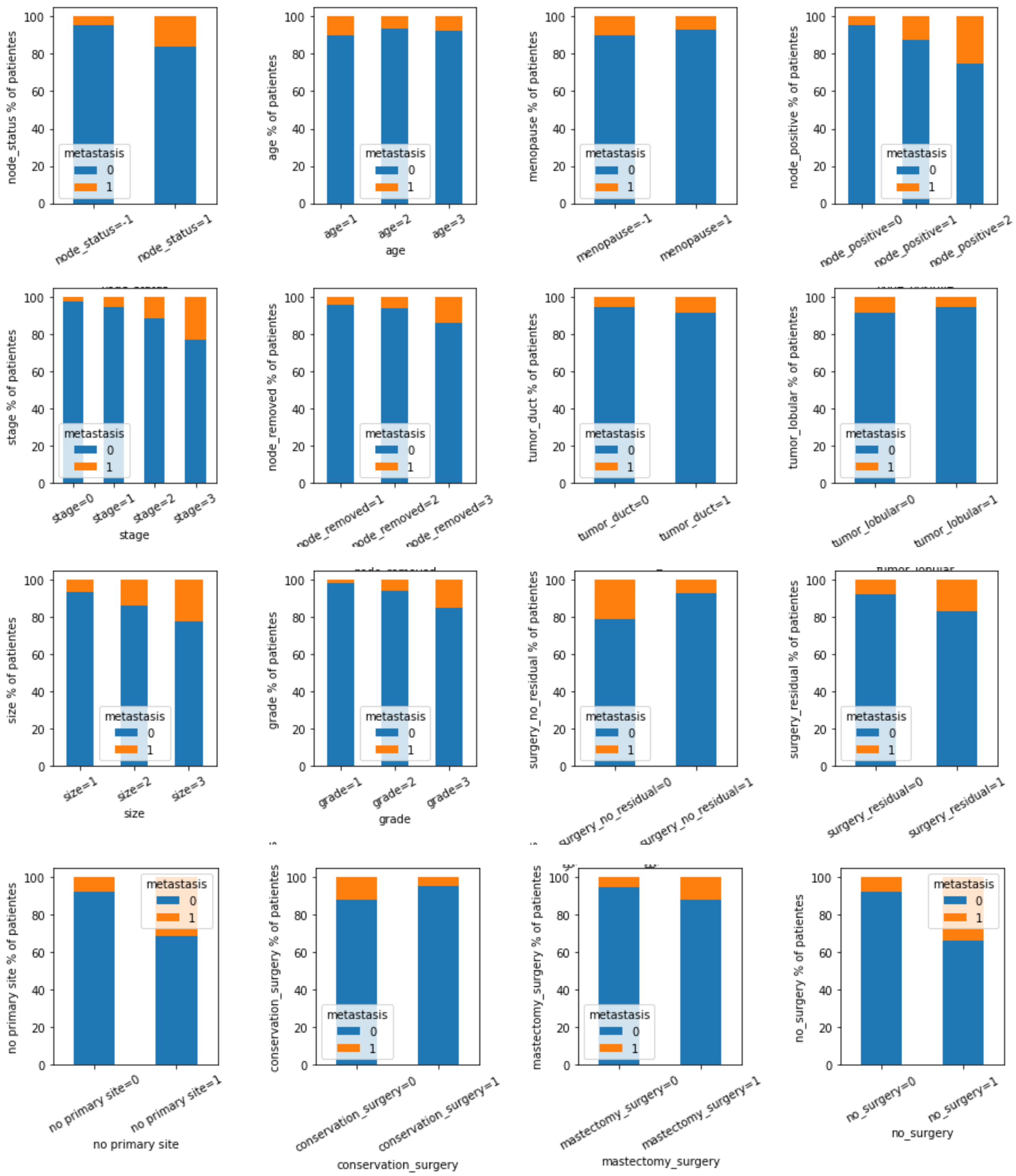
Dr Xia Jiang, in collaboration with Alan Wells, Adam Brufsky, and Richard Neapolitan, developed a Clinical Decision Support System (CDSS) to recommend the best treatment for breast cancer patients with the final goal of reducing metastasis in those patients in the following 5 years after completing the treatment. Such system (CDSS) is based on an algorithm called Treatment Feature Interactions (TFI) and a Bayesian network architecture called Causal Modeling with Internal Layers (CAMIL).

The CDSS recommendation system provides better chances for cancer patients. But even when the results were better with the CDSS recommendations, some results were controversial. Therefore, I created a ML model to analyze the data of those cancer patients to predict metastasis after treatment. Note, my prediction is after treatment, and not a system recommendation for treatment like CDSS.

I extracted the same dataset used for the treatment recommendation system CDSS, which contained information of 6,726 patients and 24 categorical features. The data was relatively clean and almost ready to use, except for the need of creating dummy variables to represent the categorical values.

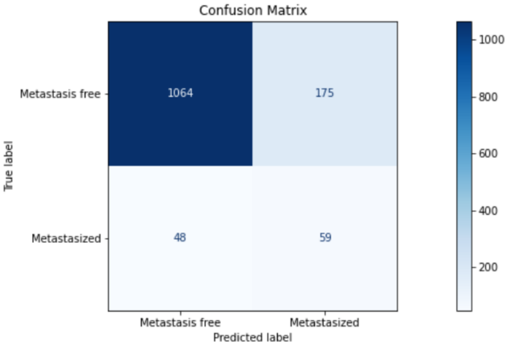
The data distribution for each variable contribution in the metastasis is represented by the following graphs:



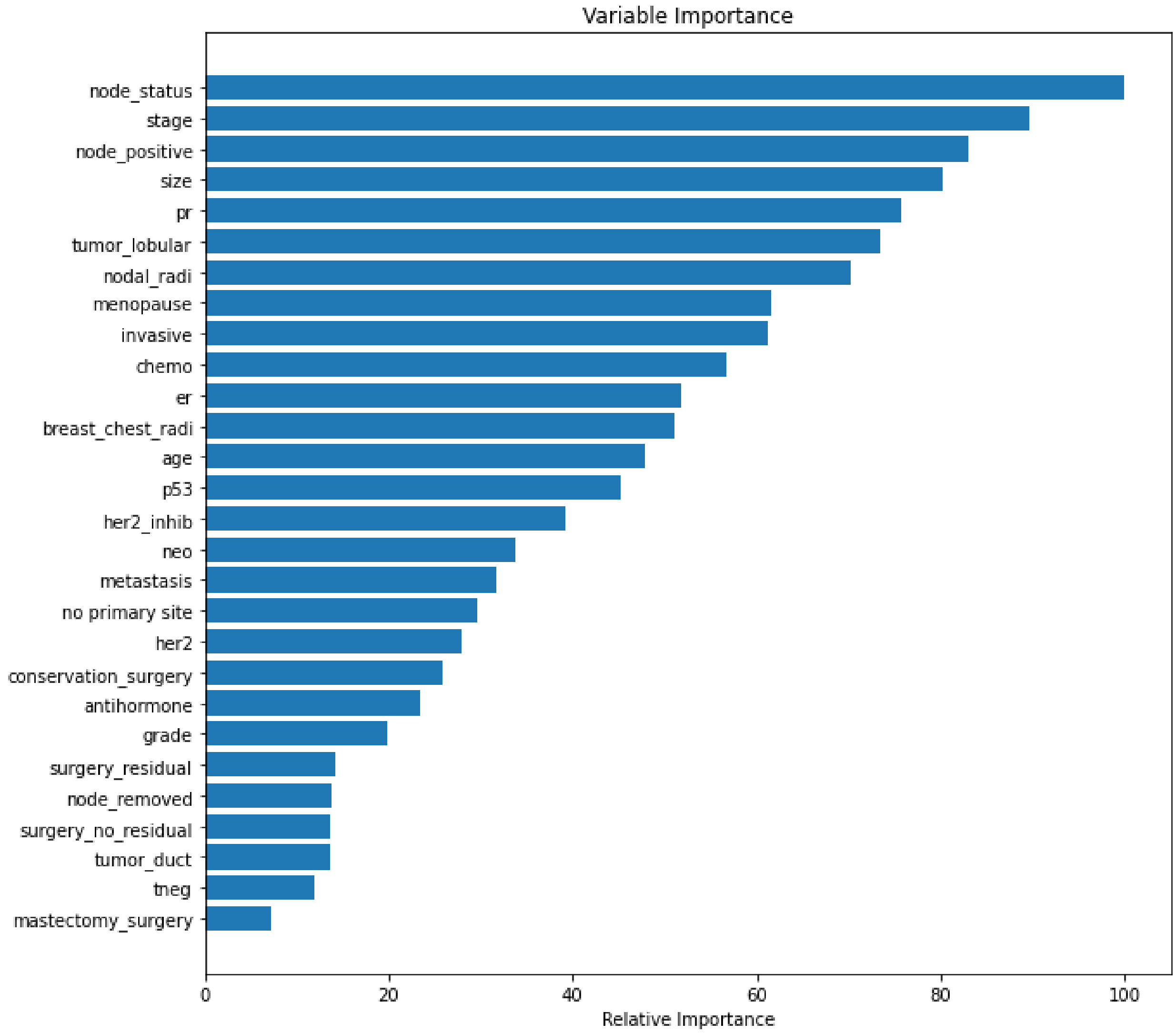


I also evaluated the chi2 metric to identify which variables were independent and which were dependent against metastasis. The two most dependent variables were breast_chest_radi and histology.

I evaluated the Logistic Regression model together with the Random Forest for the predictions, using 20% for testing and 80% of the data for training the model. The accuracy of both models were very similar in 0.92. However, the f1-score for the metastasized class was lower in the logistic regression model with a 0.15 value, while the f1-score for the Random Forest was 0.22. Based on this, I tuned the RandomForest model to increase the f1-score for the metastasized class to 0.55. The parameters tuned were 'class_weight', 'criterion', 'max_depth', 'min_samples_leaf', 'min_samples_split' and 'n_estimators'. Here is the confusion matrix for the test data:



My model final prediction accuracy was 0.83, which is above the initial target goal of 0.80. The relevance of the features in the model prediction is shown in the graph below, being the five most relevant features, the node status (if the patient had any positive lymph nodes), the stage (composite of size and # positive nodes), the node positive (number of positive lymph nodes), the size of the tumor in millimeters and PR (the progesterone receptor expression).


Variable Importance
node_status
stage
node_positive
size
pr
tumor_lobular
nodal_radi
menopause
invasive
chemo
er
breast_chest_radi
age
p53
her2_inhib
neo
metastasis
no primary site
her2
conservation_surgery
antihormone
grade
surgery_residual
node_removed
surgery_no_residual
tumor_duct
tneg
mastectomy_surgery
0
20
40
60
80
100
Relative Importance