Predicting Hypertensive Disorders of Pregnancy for Prioritized Screening

Rinat Kirshner, 04/06/2025

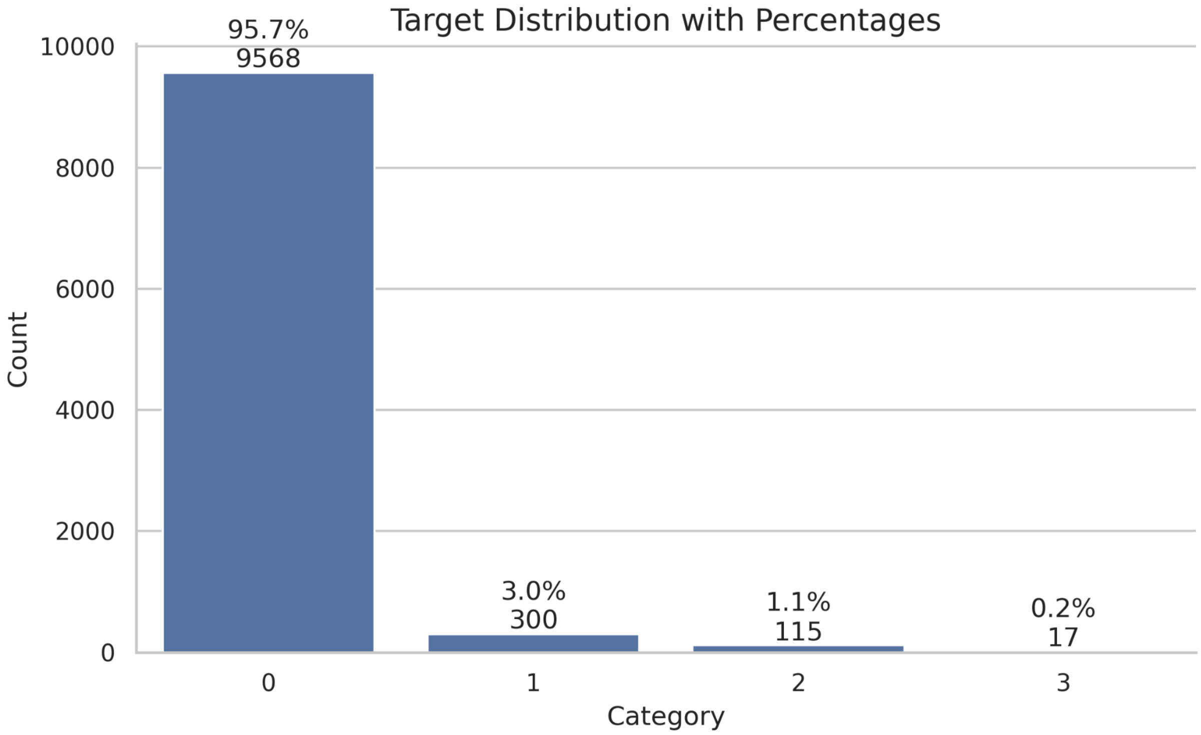
This project addresses a critical clinical challenge: identifying patients at the highest risk of developing hypertensive disorders of pregnancy around week 15. The core objective is to build a robust predictive model that enables efficient allocation of limited healthcare resources by prioritizing high-risk individuals for further examination.

The project consists of the following steps:

- Target design
- Data preparation
  (coping with missing values, removing zero-variance features, addressing highly correlated features, and extracting features from clinical text)
- Training and evaluation
  (designing multiple performance metrics, addressing imbalance data)
- Screening and prioritization under budget constraints

Target Design

I decided to introduce a new, multi-level **Hypertension_severity** target variable. It is categorized into four levels: **0 (No Hypertension), 1 (Gestational Hypertension), 2 (Preeclampsia), and 3 (Eclampsia)**. Specifically, any patient diagnosed with hypertension (Y=1) is assigned level 1, patients with Preeclampsia - level 2, and those with Eclampsia - level 3. The **Hypertension_severity** is the sole target information that is used for training; all other ground-truth columns are omitted from the data set. The data is highly imbalanced (see Figure), with merely 4% of the population suffering from **Hypertension_severity**.



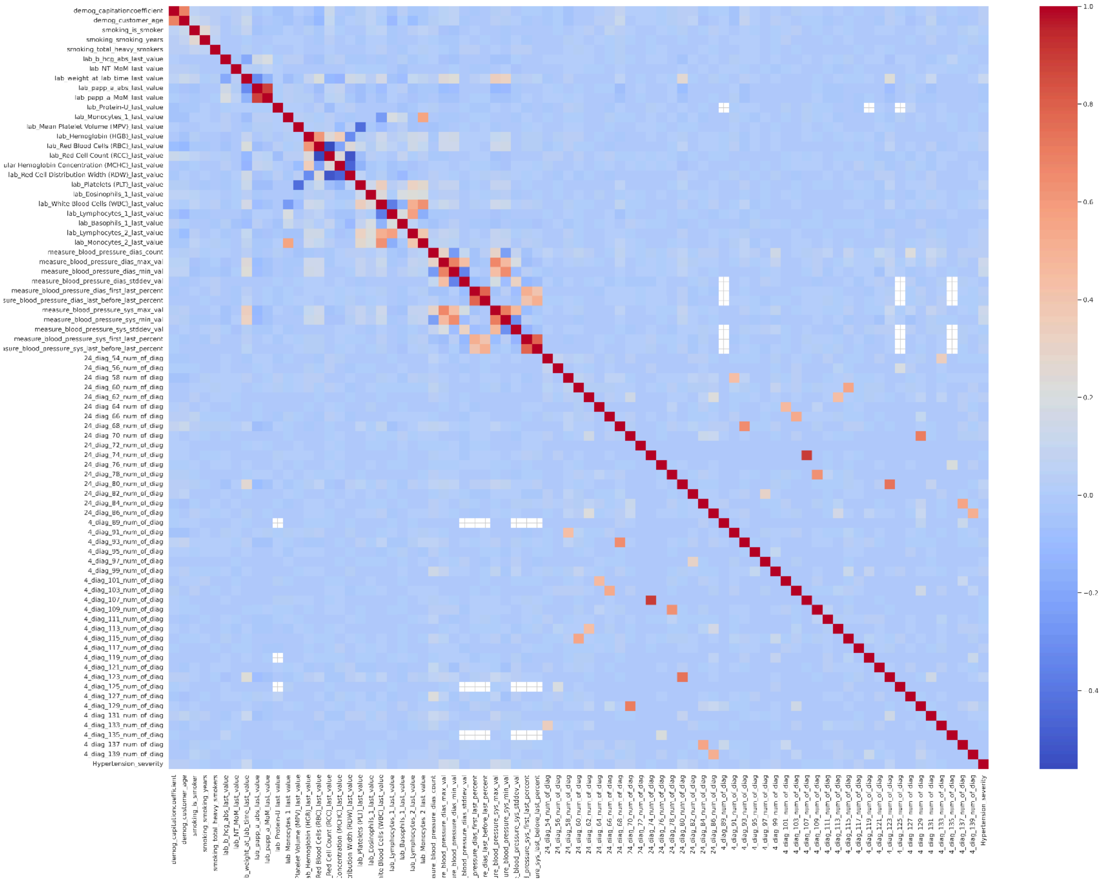
Data preparation

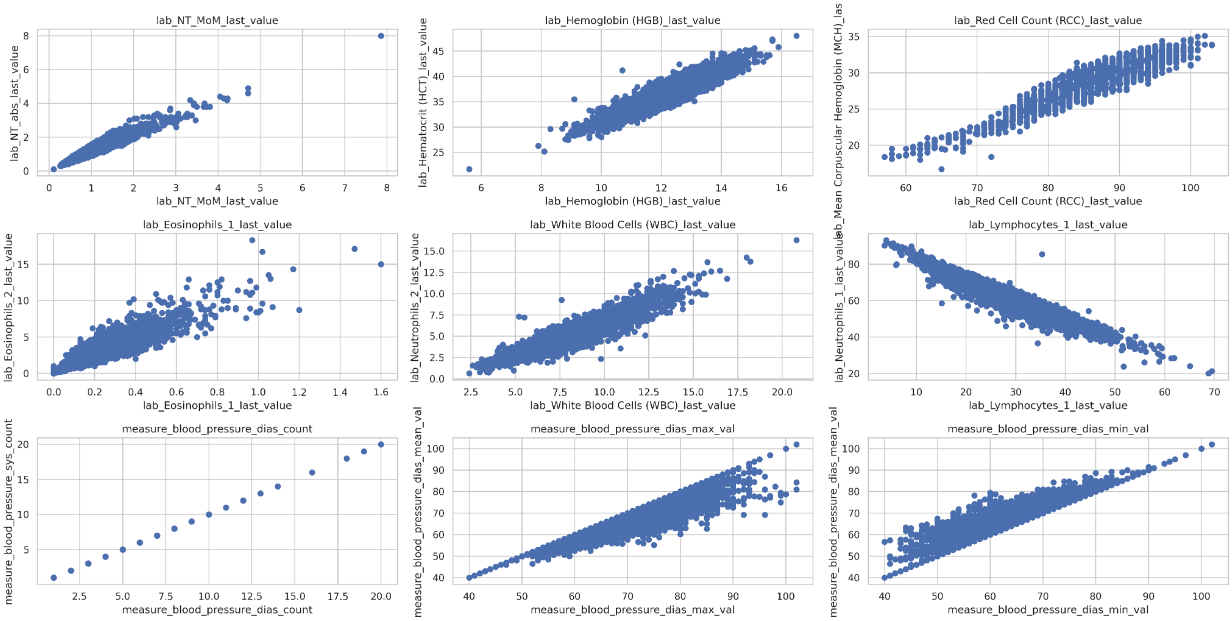
I carried out the following data cleanup operations:

- 45 features with over 70% missing values were omitted.
- 1 Feature with 0 variance was omitted.
- 16 feature with correlation of 90% between them were omitted
- Outliers rejection
  (I utilized IQR to find features with more than 10% of data below and above 5% and 75%, respectively.

I examined correlation between the features:

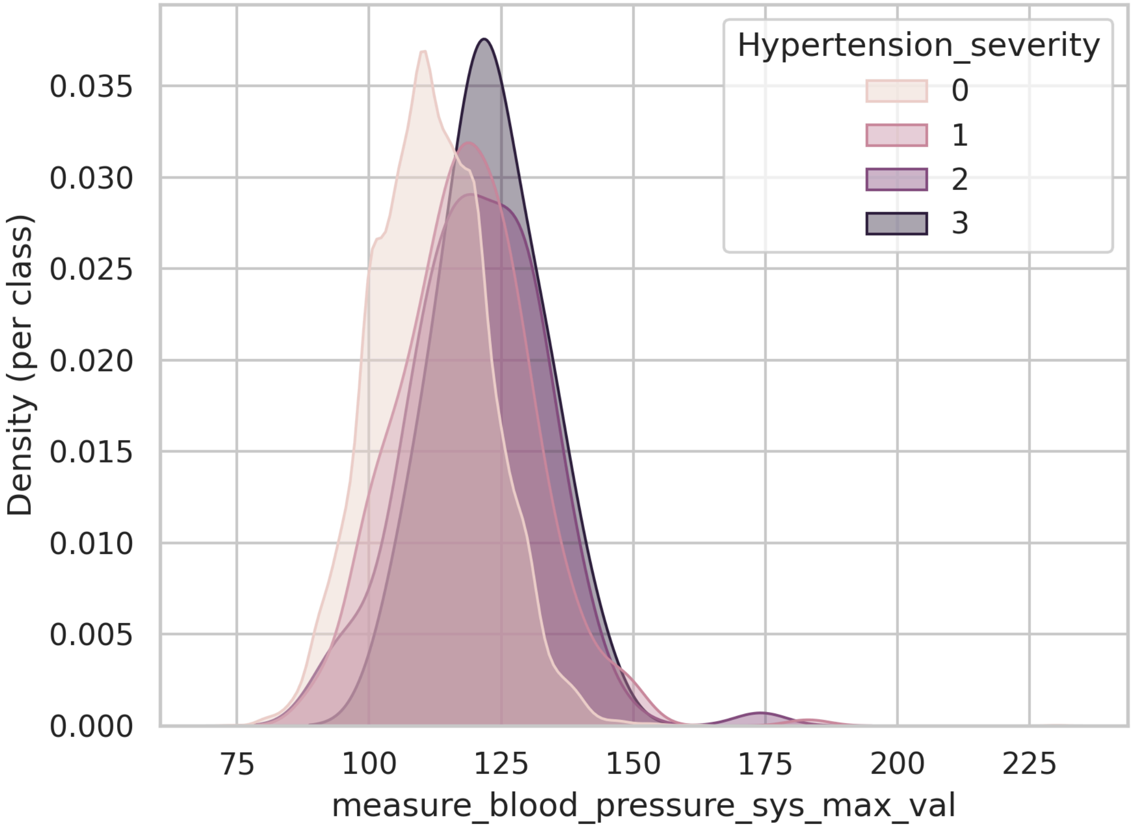
- correlation matrix (see Figure).
- specific feature pairs (see Figure).



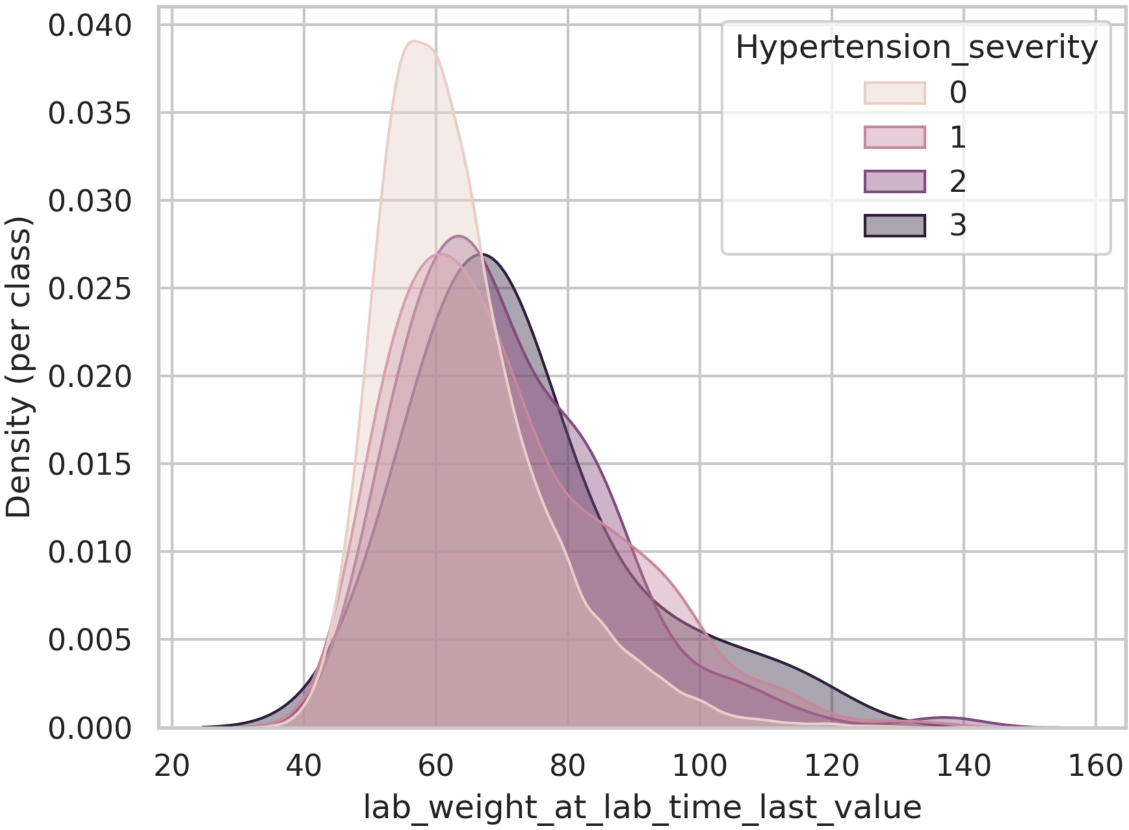

lab_NT_MoM_last_value
lab_NT_abs_last_value
0
2
4
6
8
0
1
2
3
4
5
6
7
8
lab_NT_MoM_last_value
lab_Hemoglobin (HGB)_last_value
lab_Hematocrit (HCT)_last_value
25
30
35
40
45
6
8
10
12
14
16
lab_Hemoglobin (HGB)_last_value
lab_Red Cell Count (RCC)_last_value
lab_Mean Corpuscular Hemoglobin (MCH)_las
20
25
30
35
60
70
80
90
100
lab_Red Cell Count (RCC)_last_value
lab_Eosinophils_1_last_value
lab_Eosinophils_2_last_value
0
5
10
15
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
lab_Eosinophils_1_last_value
lab_White Blood Cells (WBC)_last_value
lab_Neutrophils_2_last_value
0.0
2.5
5.0
7.5
10.0
12.5
15.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
lab_White Blood Cells (WBC)_last_value
lab_Lymphocytes_1_last_value
lab_Neutrophils_1_last_value
20
40
60
80
10
20
30
40
50
60
70
lab_Lymphocytes_1_last_value
measure_blood_pressure_dias_count
measure_blood_pressure_sys_count
5
10
15
20
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
measure_blood_pressure_dias_count
measure_blood_pressure_dias_max_val
measure_blood_pressure_dias_mean_val
40
50
60
70
80
90
100
40
50
60
70
80
90
100
measure_blood_pressure_dias_max_val
measure_blood_pressure_dias_min_val
measure_blood_pressure_dias_mean_val
40
50
60
70
80
90
100
40
50
60
70
80
90
100
measure_blood_pressure_dias_min_val

The following figures describe histogram probability density functions of selected features per **Hypertension_severity** category:

- Measure blood_preassure_sys_max_val - when plotting this feature against target variable, it can be observed that for severities above 0 the distribution is slightly right skewed than the distribution when severity is 0. There is no difference in distribution for target variables 1,2,3.



- Lab_weight_at_lab_time_last_value - same observation as above



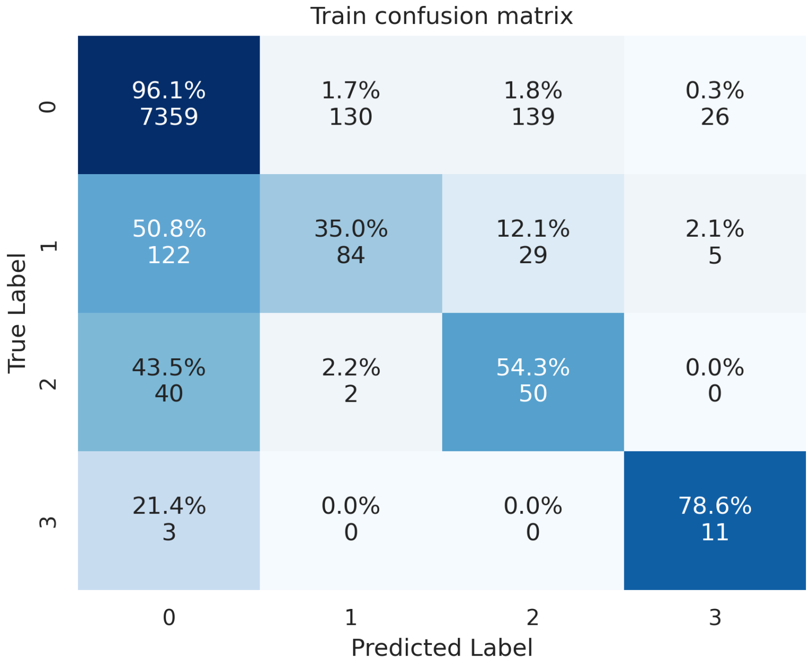
- Lab white lab cells - very difficult to discern between the groups
- It is very hard to distinguish between severity levels based on these features.

Textual Data:

To augment the dataset for predicting hypertensive disorders of pregnancy, features were extracted from unstructured Hebrew clinical notes. This was done by utilizing the Google Gemini large language model through a LangChain pipeline. For each clinical sheet, a prompt was used to instruct the Gemini model to identify and list relevant risk factors, ensuring the output provided standardized Hebrew names (e.g., 'סוכרת' ,'עישון') and excluded any explicitly negated or unmentioned factors. This process converted narrative text into structured risk factor indicators, which were then integrated as additional features for the predictive model.This process resulted with additional 16 risk features to the data.

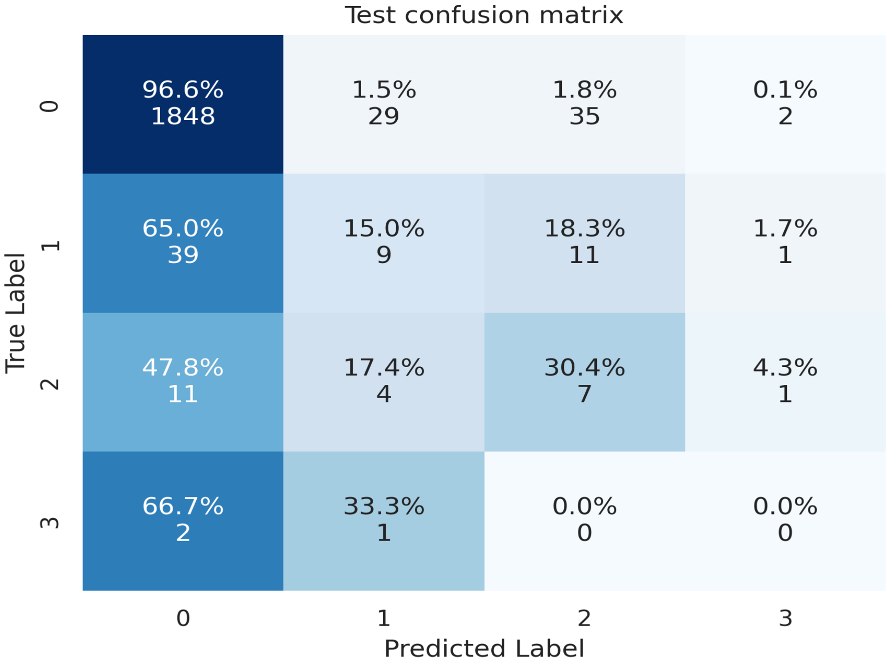
Model Training and Evaluation

- I decided to use a Random Forest classifier for this supervised classification problem. This type of classifier required no feature standardization
- I used the **StratifiedKFold** variation of the K-Fold cross-validation technique, to ensure that each fold maintains the same proportion of samples from each class as the overall dataset.
- Model selection:
  I firstly used the Random Forest model with the class_weight configuration to address the issue of imbalanced target variables. However, the model was found to overfit the training data (ideal confusion matrix) and to provide poor performance for the test data (all classes were classified to 0). To cope with the overfitting, I used the SMOT approach, which is an oversampling technique for generating *synthetic* new samples for the minority classes i.e classes 1,2,3, instead. I further modified the Random Forest parameters by lowering its max_depth parameter and by increasing the min_sample_leaf parameter.
- Confusion Matrix:
  The Train Confusion matrix does not overfit the train data (see Figure).
  The Test Confusion matrix does not overfit the test data, either (see Figure). Yet, the model does not predict class 3 at all, and misclassifies class 0 as class 3 in a few cases.
- Lift metric:
  - Class 0 - 1.02, the model is the same as picking randomly class 0 patients.
  - Class 1 - 6.98, the model is 7 times better than picking up randomly class 1 patients.
  - Class 2 - 11.48, model is 12 times better than picking up randomly class 2 patients.
  - Class 3 - no classification.



Train confusion matrix

	0	1	2	3
0	96.1% 7359	1.7% 130	1.8% 139	0.3% 26
1	50.8% 122	35.0% 84	12.1% 29	2.1% 5
2	43.5% 40	2.2% 2	54.3% 50	0.0% 0
3	21.4% 3	0.0% 0	0.0% 0	78.6% 11

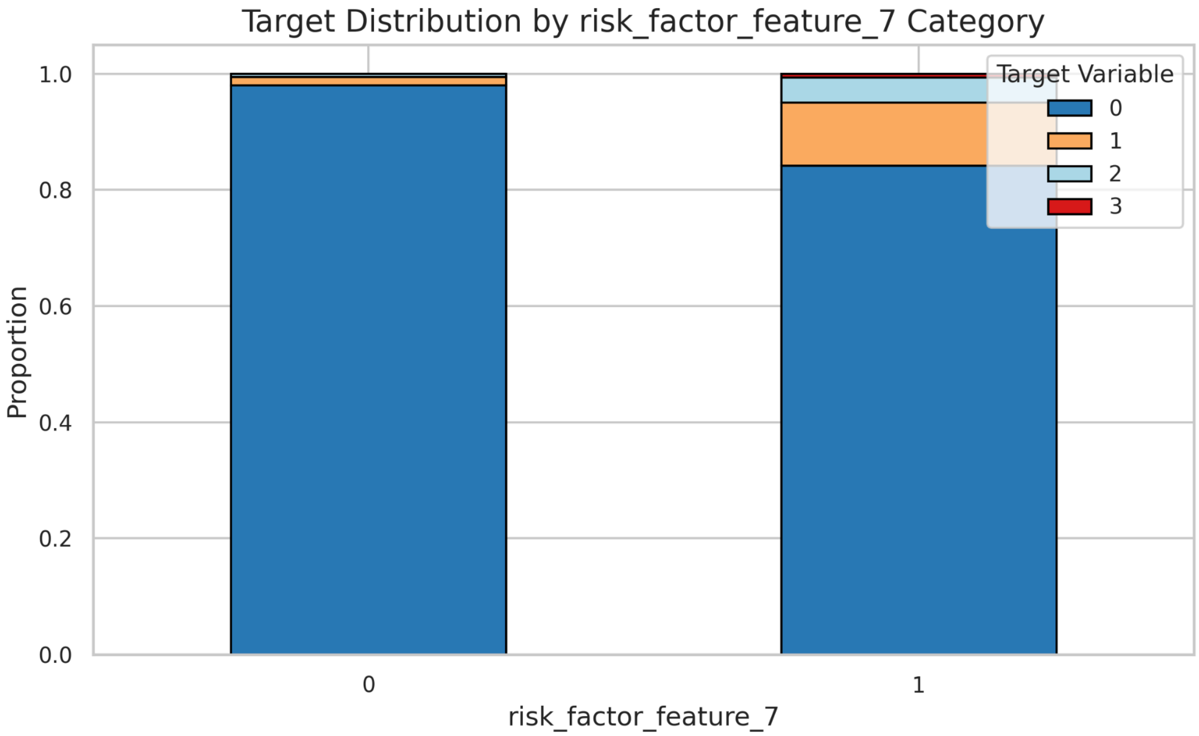


Test confusion matrix

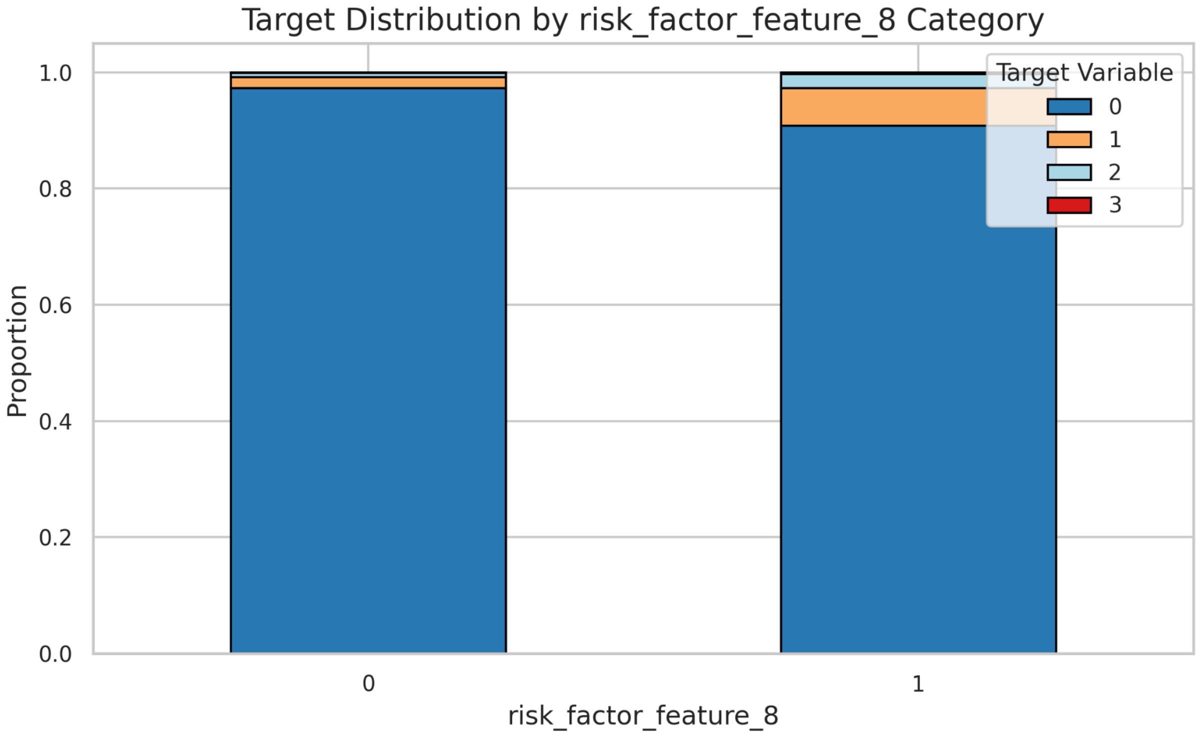
	0	1	2	3
0	96.6% 1848	1.5% 29	1.8% 35	0.1% 2
1	65.0% 39	15.0% 9	18.3% 11	1.7% 1
2	47.8% 11	17.4% 4	30.4% 7	4.3% 1
3	66.7% 2	33.3% 1	0.0% 0	0.0% 0

- Feature importance

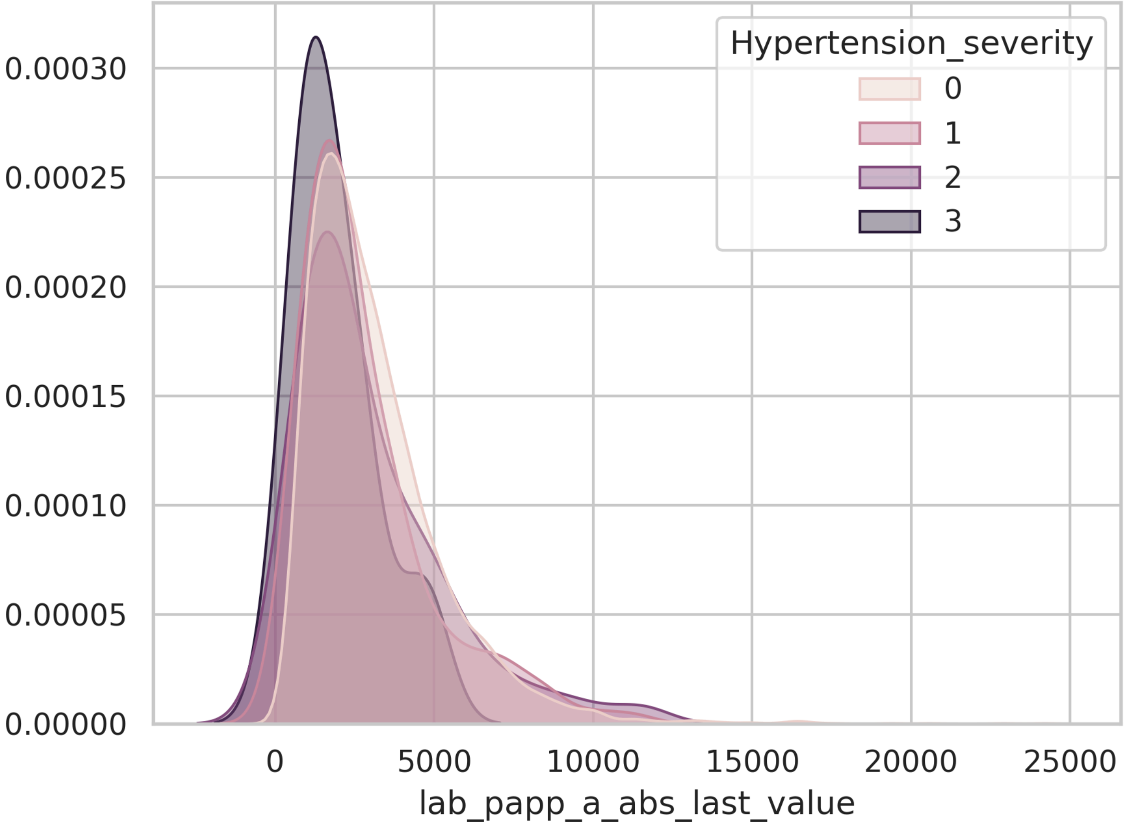
  Risk Factor Feature_7 (BMI) helps distinguish between the target variables. If there is a BMI issues, the patient has higher chances to have severe hypertension.



Risk Factor Feature_8 (high blood pressure) helps distinguish between the target variables. If there is an issue the patient has higher chances to have severe hypertension.



Lab_papp_a_abs_ast value - while this feature is a prominent variable, its histograms across classes are difficult to distinguish. Only higher severity levels exhibit a slight left skew compared to no severity, with no discernible differences in distributions among severity levels above zero.

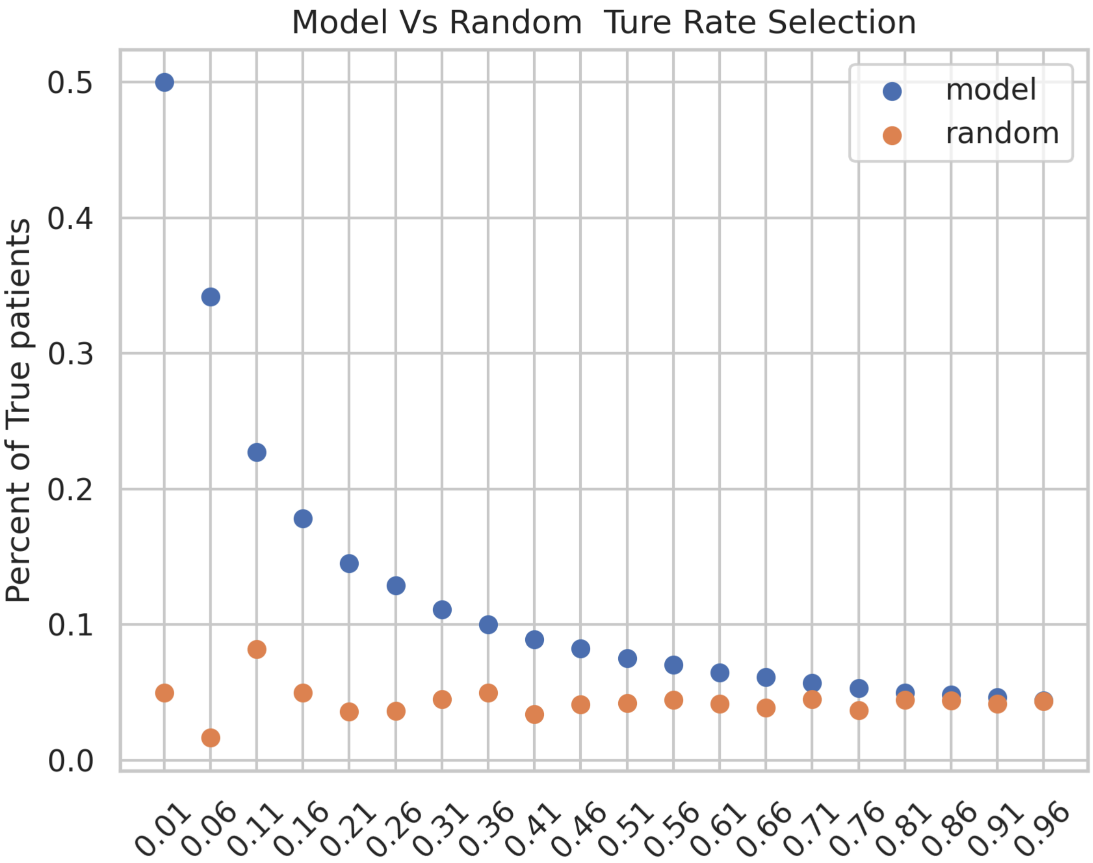


○

Screening and Prioritization:

After model selection, we can calculate probability scores for each patient across all classes. For patient allocation within budget limitations, we begin by addressing the most severe class, assigning patients with the highest probabilities first. If resources permit, we then proceed to the next most severe class, continuing this approach until the budget is fully utilized. Should there be remaining budget for Class 0, patients in this group are allocated in ascending order of their predicted probability for that class.

Following this, we determine the rate of patients whose allocation aligns with the ground truth, and this is compared to a benchmark established by random patient selection. This process is repeated to evaluate performance under various budget constraints.



The figure indicates that as the budget increases our rate decreases, as expected, since it is possible to allocate more patients to the additional blood tests. Note that the proposed model and screening approach yield better precision over random selection.

Possible Improvements:

- Consult a medical specialist for the ground-truth severity classification.
- Data preparation threshold values tuning.
- Utilization of additional models such as XGBOOST.
- Manually analysing misclassifications events, understanding their root-cause and update the algorithmic approach.