Jacob Krol BIOS 7747 Oct 25, 2023 Classification and Model Selection

Assignment

Motivation

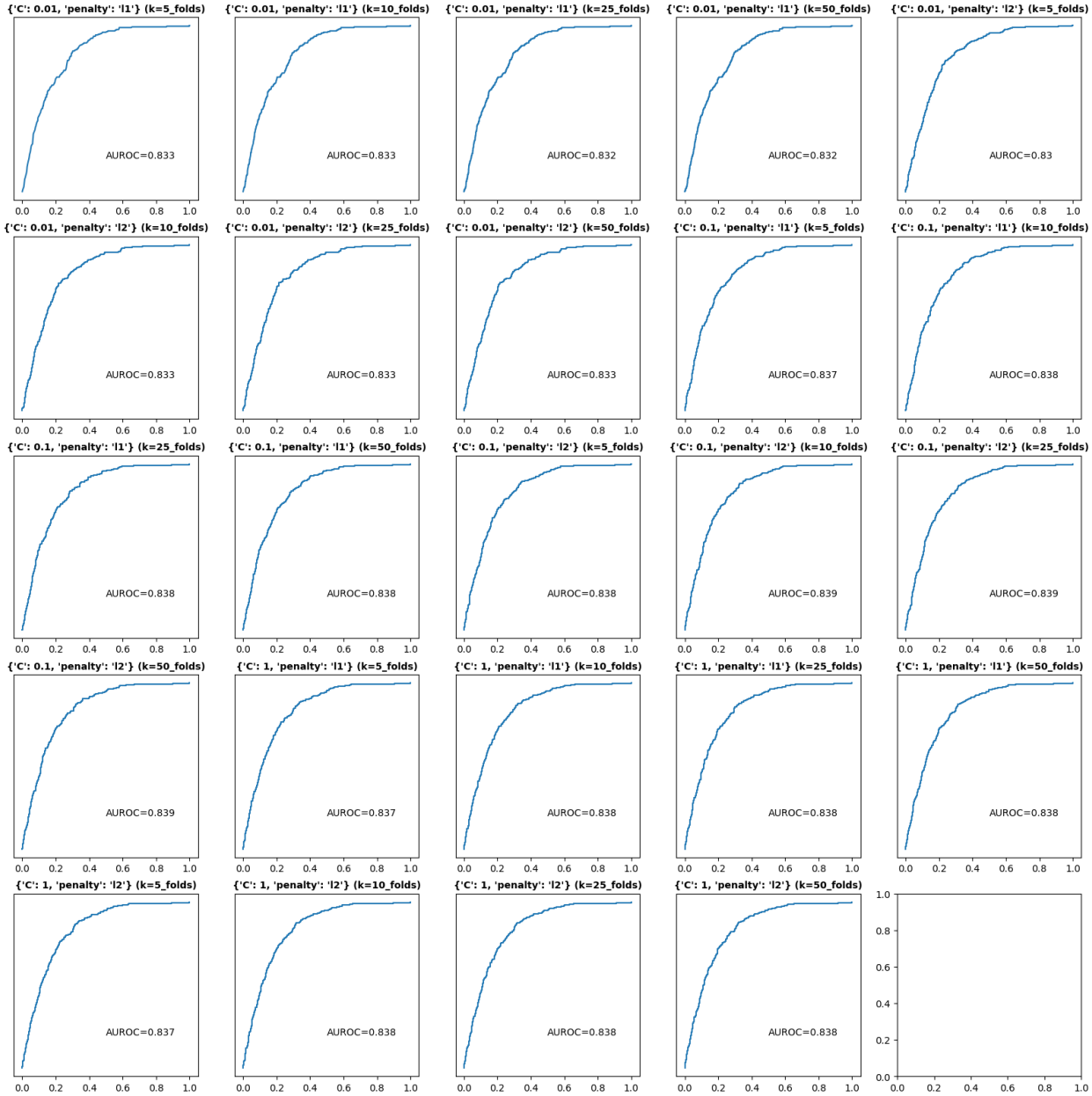
To understand classifier outcomes and select appropriate classification models

Study summary

The Departments of Neurology and Neurosurgery would like to study the possibility of predicting which of their patients is at risk of presenting a brain stroke. Since it is known that several factors such as smoking, hypertension or obesity are related to an increased risk of stroke, they designed a multi-institutional study collecting that information from a large number of patients who were followed up to record whether they had a stroke or not.

Goal: To identify if the risk of stroke can be identified using the collected dataset.

Jacob Krol BIOS 7747 Oct 25, 2023 Classification and Model Selection



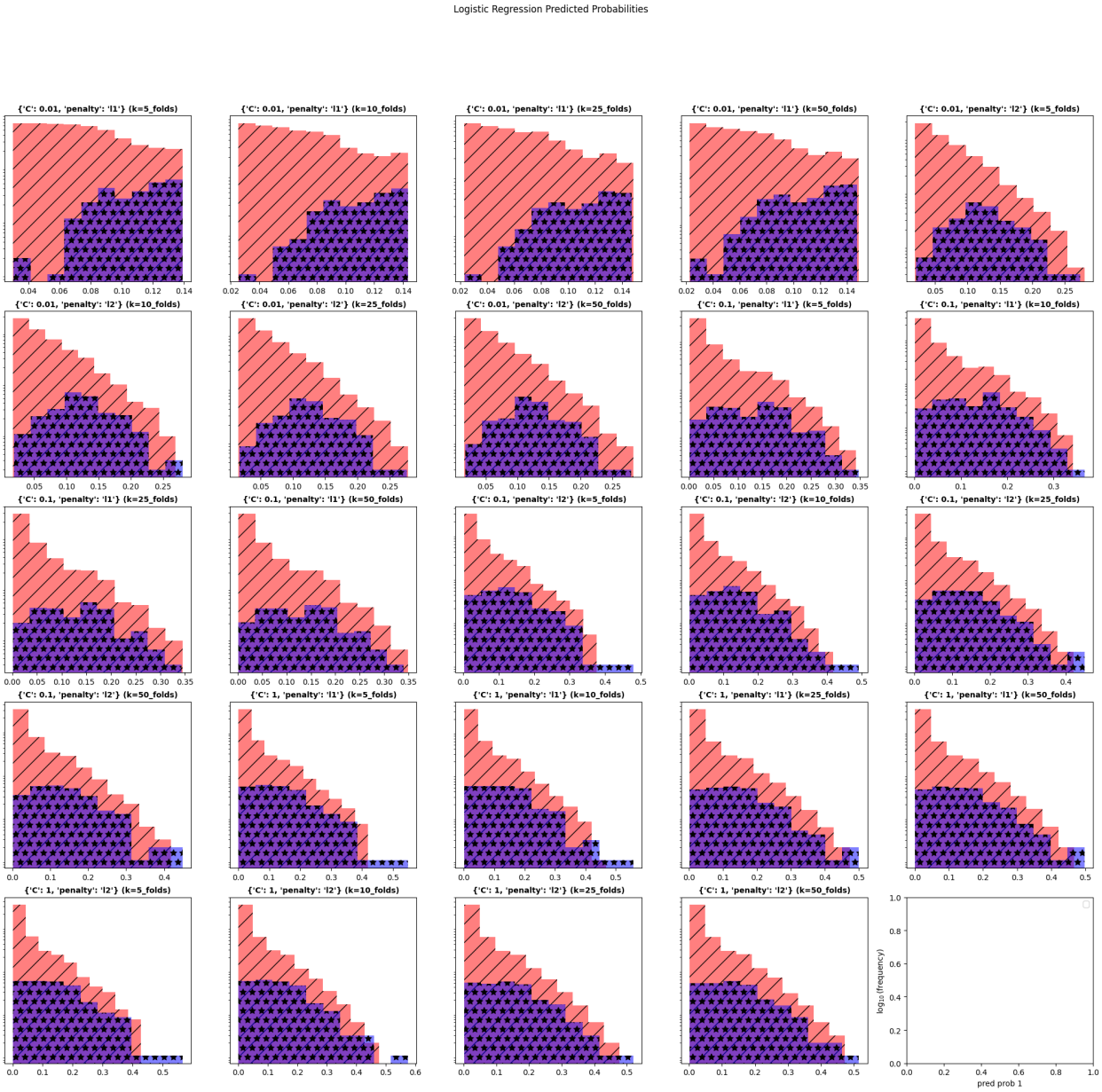
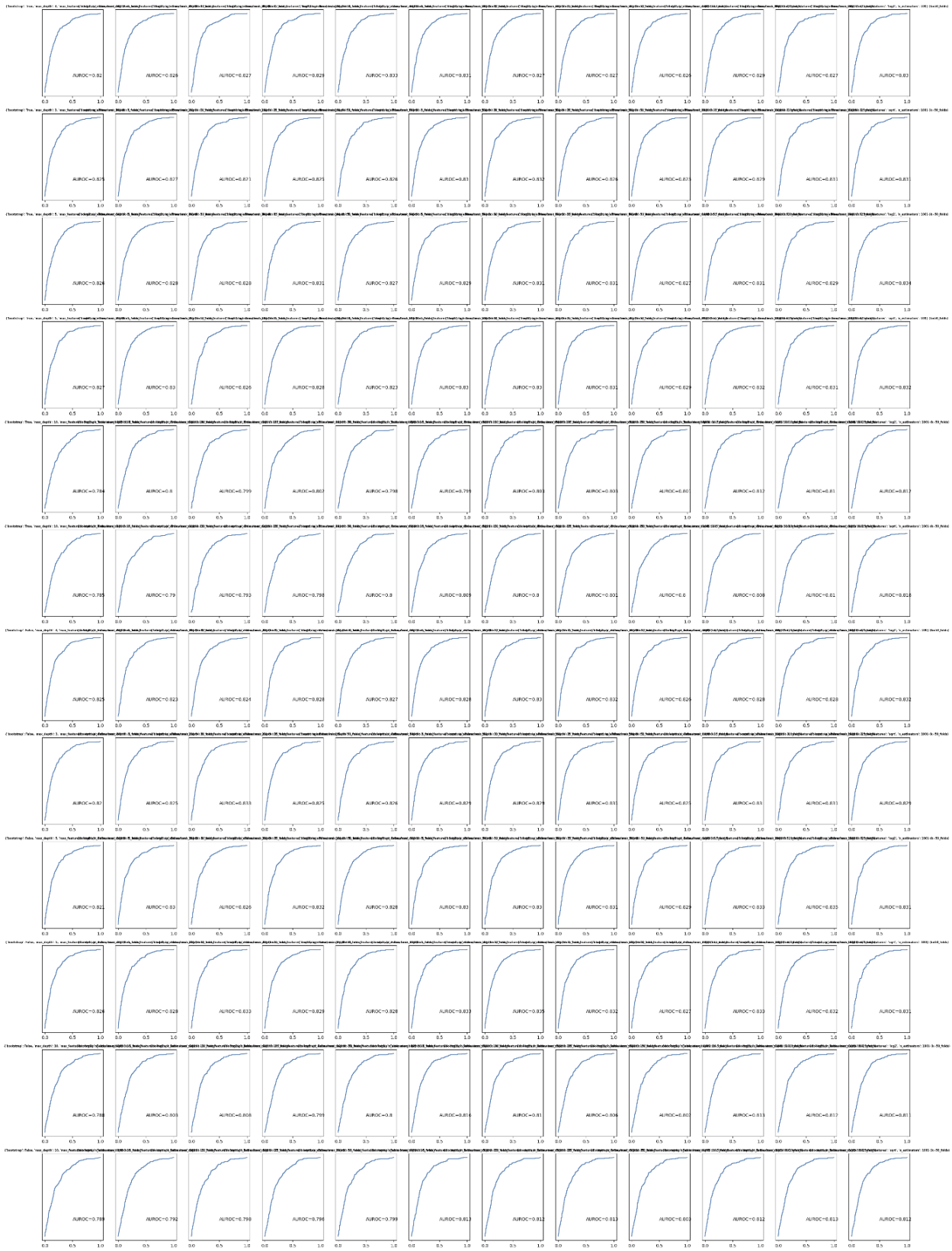


Fig 1: **Logistic regression AUROC and positive label predicted probability distributions across various hyperparameters**: **A** (top) AUROC. **B** (bottom) predicted probability for stroke (1). Y-axis is log10 scaled. Red histogram with diagonal lines represents the ground truth negative samples. Blue histogram with stars represents the ground truth positive samples.


Random Forest AUROC

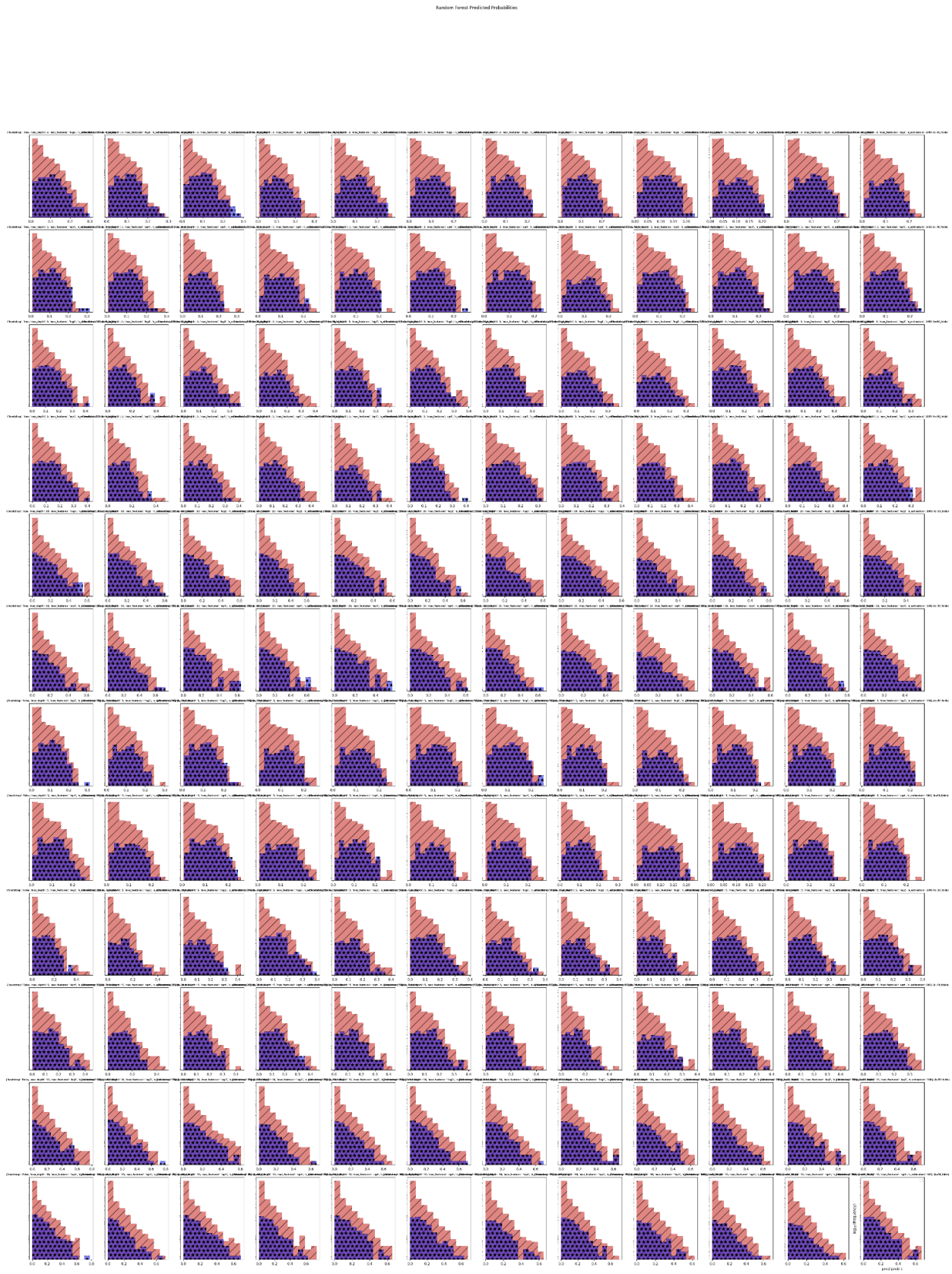
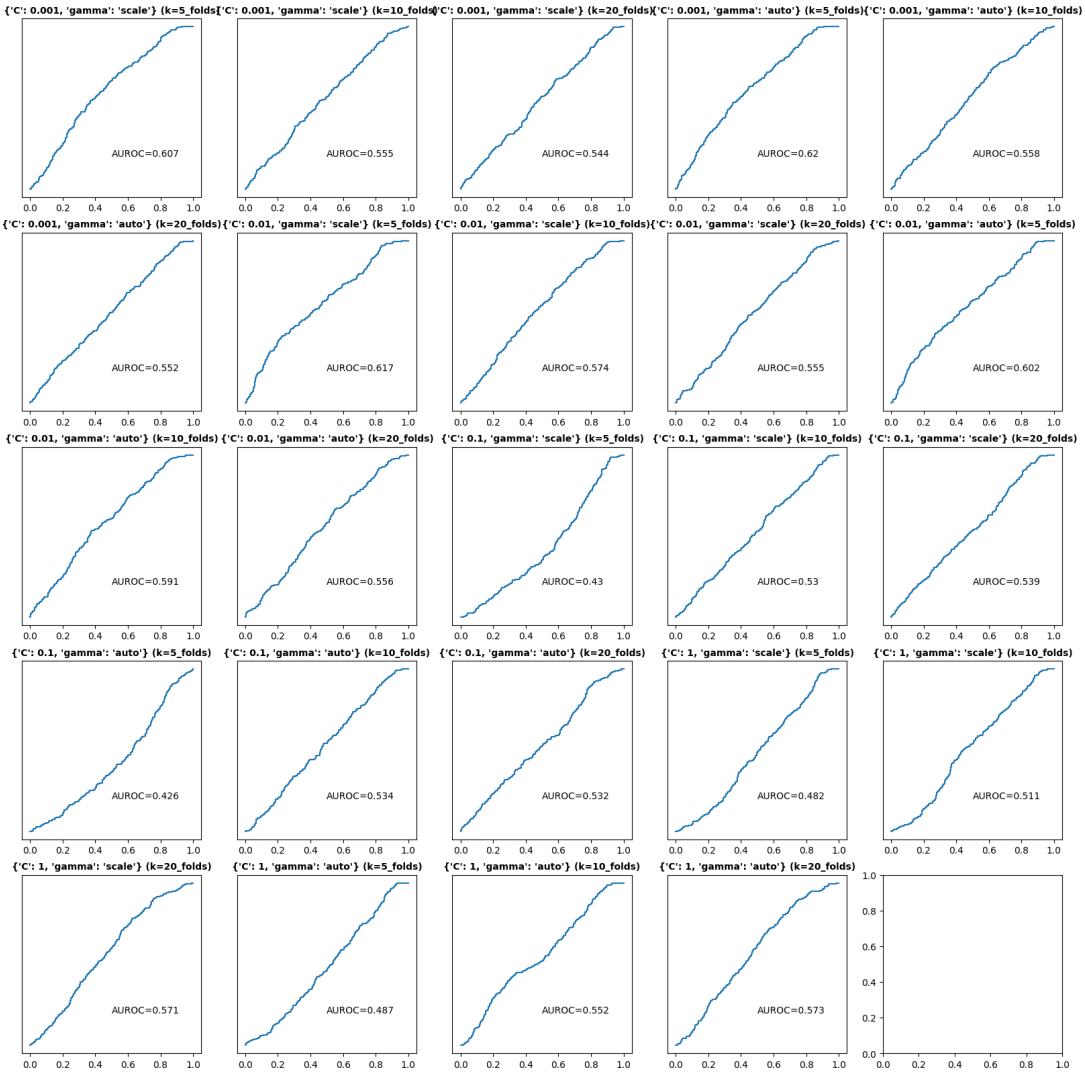


Fig 2: Random Forest AUROC and positive label predicted probability distributions

across various hyperparameters: **A** (top) AUROC. **B** (bottom) predicted probability for stroke (1). Y-axis is log10 scaled. Red histogram with diagonal lines represents the ground truth negative samples. Blue histogram with stars represents the ground truth positive samples.

SVM AUROC


{'C': 0.001, 'gamma': 'scale'} (k=5_folds)
AUROC=0.607
{'C': 0.001, 'gamma': 'scale'} (k=10_folds)
AUROC=0.555
{'C': 0.001, 'gamma': 'scale'} (k=20_folds)
AUROC=0.544
{'C': 0.001, 'gamma': 'auto'} (k=5_folds)
AUROC=0.62
{'C': 0.001, 'gamma': 'auto'} (k=10_folds)
AUROC=0.558
{'C': 0.001, 'gamma': 'auto'} (k=20_folds)
AUROC=0.552
{'C': 0.01, 'gamma': 'scale'} (k=5_folds)
AUROC=0.617
{'C': 0.01, 'gamma': 'scale'} (k=10_folds)
AUROC=0.574
{'C': 0.01, 'gamma': 'scale'} (k=20_folds)
AUROC=0.555
{'C': 0.01, 'gamma': 'auto'} (k=5_folds)
AUROC=0.602
{'C': 0.01, 'gamma': 'auto'} (k=10_folds)
AUROC=0.591
{'C': 0.01, 'gamma': 'auto'} (k=20_folds)
AUROC=0.556
{'C': 0.1, 'gamma': 'scale'} (k=5_folds)
AUROC=0.43
{'C': 0.1, 'gamma': 'scale'} (k=10_folds)
AUROC=0.53
{'C': 0.1, 'gamma': 'scale'} (k=20_folds)
AUROC=0.539
{'C': 0.1, 'gamma': 'auto'} (k=5_folds)
AUROC=0.426
{'C': 0.1, 'gamma': 'auto'} (k=10_folds)
AUROC=0.534
{'C': 0.1, 'gamma': 'auto'} (k=20_folds)
AUROC=0.532
{'C': 1, 'gamma': 'scale'} (k=5_folds)
AUROC=0.482
{'C': 1, 'gamma': 'scale'} (k=10_folds)
AUROC=0.511
{'C': 1, 'gamma': 'scale'} (k=20_folds)
AUROC=0.571
{'C': 1, 'gamma': 'auto'} (k=5_folds)
AUROC=0.487
{'C': 1, 'gamma': 'auto'} (k=10_folds)
AUROC=0.552
{'C': 1, 'gamma': 'auto'} (k=20_folds)
AUROC=0.573

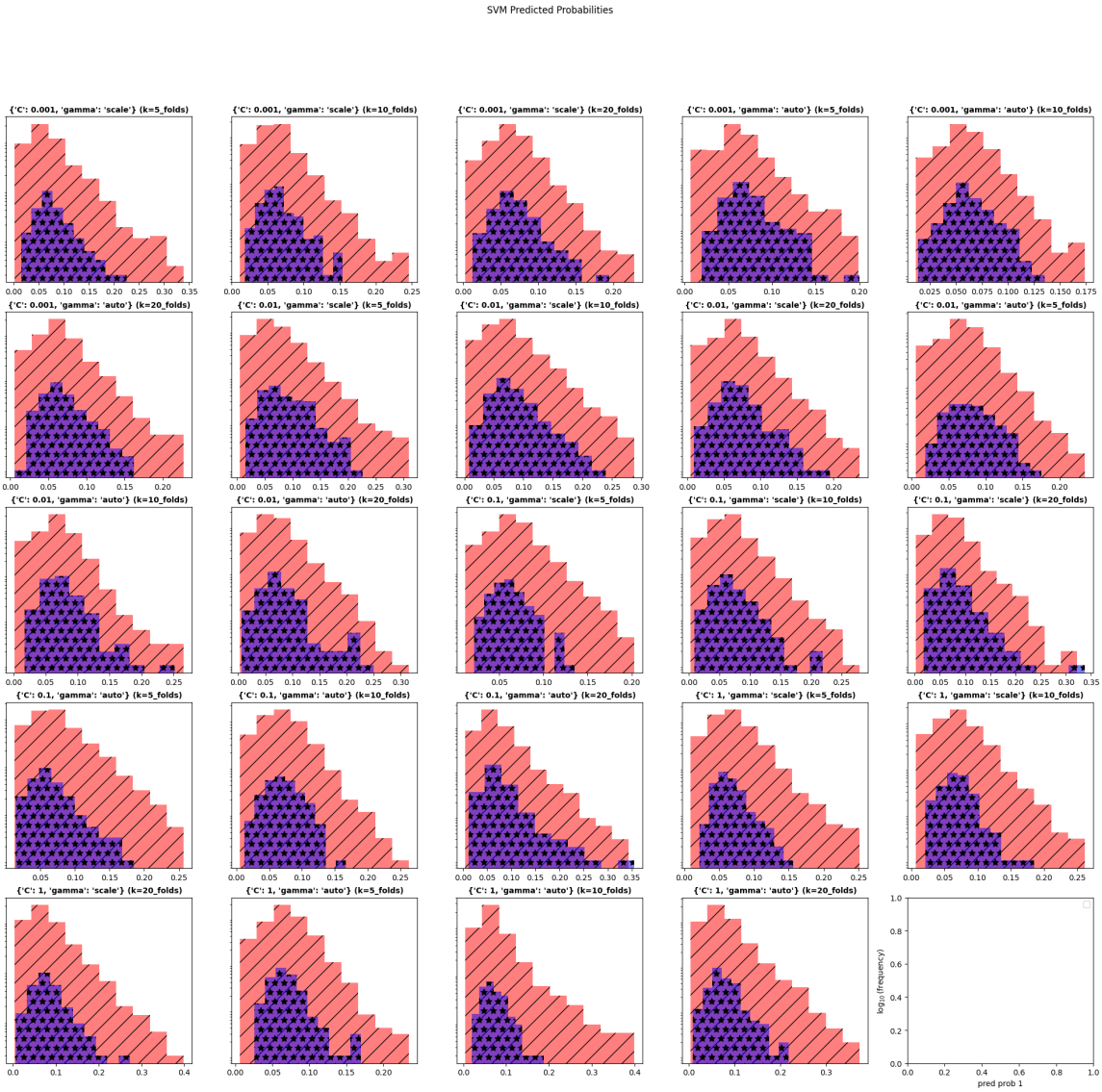


Fig 3: SVM AUROC and positive label predicted probability distributions across various hyperparameters. **A** (top) AUROC. **B** (bottom) predicted probability for stroke (1). Y-axis is log10 scaled. Red histogram with diagonal lines represents the ground truth negative samples. Blue histogram with stars represents the ground truth positive samples.

Build and evaluate the performance of: a random forest classifier, a linear support vector machine classifier, and a logistic regression classifier. Feel free to evaluate non-linear support vector machines if curious.

Three classification models (Logistic Regression [LR], Random Forest [RF], and Support Vector Machines [SVM]) were trained and tested using a grid search across various stratified k-folds using data from previous patients with or without stroke in their medical history (**Fig 1-3**). The model with the maximum AUROC score was Logistic regression with an AUROC of 0.839.

Random Forest models performed similar to logistic regression and SVM measured relatively low AUROC values across all hyperparameters (**Fig 1-3**).

The predicted probability distributions (**Fig 1-3**) indicate there's no classifier achieving exceptional class separation; an exceptional classifier would have no overlap between the stroke negative (red) and stroke positive (blue) distributions.

Also, **Fig 1-3** indicate that the number of stratified K-folds used for evaluation has marginal to no effect on the measured performance (AUROC).

Discuss the similarities and differences between the performance of the three trained classifiers in terms of ROC curve and the distributions of their probabilistic predictions.

Good general performance and consistency (AUROC range of 0.83-0.839 [**Fig 1-3**]) is demonstrated by logistic regression across various hyperparameters while Random forest performs similarly well with less consistency (AUROC range of 0.79-0.83) and linear SVM often does not perform better than a random classifier (AUROC range of 04.-0.6) (**Fig 1-3**).

For predicted probability separation, logistic regression's L1 penalizer and a low regularization constant produces the best class separation compared to various other hyperparamter permutations and the RF/SVM models (**Fig 1**, top left).

Explain which approach you would deploy clinically and justify the reasons for it.

To classify stroke risk, a Logistic Regression model with a **L1 regularization penalty and regularization coefficient of 0.01** (**LRL1C0.01**) will be used since this model demonstrated the best class separation. Out of all the models (LR, SVM, RF) LRL1C0.01 demonstrated the best class separation in the predicted probabilities distribution plots (**Fig 1**, top left) which allows for a decision threshold that can retrieve nearly all of true stroke cases while excluding at least some of the negative cases.

For a risk model, it is important to maximize the retrieval of true positives (sensitivity), since early-risk detection is the aim, while the ability to classify negative samples (specificity) is less imperative; however, there must be some amount of specificity for a useful model. Based on the comparison with all other predicted probability distributions, the **LRL1C0.01** model is the **best** classifier where a threshold can be placed (e.g., at 0.07) to yield almost all of the stroke history positive samples while correctly classifying a portion of the negative samples.