**Data Science Career Track**

**Model Metrics Exercise**

1. Look at the table below. If the goal is to optimize the True Positives which model would you choose and why?

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Recall** | **Precision** | **Accuracy** | **F1** |
| **Logistic** | 0.746 | 0.775 | 0.999 | 0.761 |
| **Logistic with auto threshold** | 0.891 | 0.061 | 0.976 | 0.114 |
| **Logistic with class weights** | 0.878 | 0.110 | 0.988 | 0.195 |
| **Hinge with auto threshold** | 0.905 | 0.014 | 0.890 | 0.028 |
| **Hinge with class weights** | 0.878 | 0.103 | 0.987 | 0.185 |

1. Calculate the F-1 scores for each model and identify the best model based on the F1 score.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Recall** | **Precision** | **F1** | **Auc/Roc** |
| **Deep NN** | 0.79 | 0.82 |  | 0.92 |
| **Logistic Regression** | 0.75 | 0.79 |  | 0.90 |
| **Random Forest** | 0.80 | 0.66 |  | 0.90 |
| **LinearSVC** | 0.74 | 0.75 |  | 0.82 |

1. Identify the best parameter values for ‘alpha’ and ‘L1-ratio’ based on the above comparison.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Parameter** | **Parameter** | **Metric** | **Metric** | **Metric** |
|  | **Alpha** | **L1-ratio** | **MAE** | **R-squared** | **RMSE** |
| **Linear Regression** | 0.5 | 0.2 | 84.27 | 0.277 | 158.1 |
| **Linear Regression** | 0.2 | 0.5 | 84.08 | 0.264 | 159.6 |
| **Linear Regression** | 0.5 | 0.5 | 84.12 | 0.272 | 158.6 |
| **Linear Regression** | 0 | 0 | 84.49 | 0.249 | 161.2 |

**1.** Maximizing the True Positives means maximizing TP and minimizing FN, which means that we can directly use the Recall = TP/(TP+FN).

The higher the Recall, the higher is the TP.

Answer: I'd choose the "Hinge with auto threshold" model based on its highest Recall = 0.905.

**2.** We need to express the F1 in terms of the given scores: Recall, Precision, and AUC/ROC.

Their definitions are:

F1 = 2\*TP/(2\*TP+FP+FN);

Recall = TP/(TP+FN);

Precision = TP/(TP+FP);

AUC/ROC = FPR\*TPR = FP/(TN+FP)\*TP/(TP+FN);

Solving these equations, we obtain

F1 = 2\*Recall\*Precision/(Recall + Precision);

Model F1

--------------------------------------------

Deep NN 0.8048

Logistic Regression 0.7694

Random Forest 0.7232

LinearSVC 0.7450

Answer: "Deep NN" model has the highest F1 score.

**3.** For linear regression, the best measure of accuracy is the goodness of fit R^2, followed by the RMSE.

The choice of the model has to be done based on the highest R^2 together with smallest RMSE.

Answer: the best values of parameters are Alpha = 0.5, L1-ratio = 0.2.