In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
from sklearn.utils.class_weight import compute_class_weight
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split, GridSearchCV, c
from sklearn.metrics import accuracy_score, recall_score, precision_s
from sklearn.metrics import ConfusionMatrixDisplay
from sklearn.metrics import classification_report
from sklearn.pipeline import Pipeline
from imblearn.pipeline import Pipeline as ImbPipeline
from sklearn.decomposition import PCA
from imblearn.over_sampling import SMOTE, BorderlineSMOTE
from google.colab import files
uploaded = files.upload()
```

Choose Files No file chosen

Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.

```
Saving pivoted_df.csv to pivoted_df.csv
```

Load the dataframe in, inspect the data.

In [2]:
```
pivoted_df = pd.read_csv('pivoted_df.csv', index_col=0)
```

In [3]:
```
pivoted_df.head()
```

Out[3]:

	season	Age	Throws	Surgery	AB_release_speed_weighted_avg	CH_release_speed_weig
0	2008	37.0	1	0.0	0.0	
1	2009	38.0	1	0.0	0.0	
2	2010	39.0	1	0.0	0.0	
3	2011	40.0	1	0.0	0.0	
4	2012	41.0	1	0.0	0.0	

5 rows × 130 columns

In [4]:
```
pivoted_df.shape
```

Out[4]:
```
(3688, 130)
```

In [5]:
```
pivoted_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 3688 entries, 0 to 3687
Columns: 130 entries, season to SV_vz0_weighted_avg
dtypes: float64(128), int64(2)
memory usage: 3.7 MB
```

In [6]:
```
pivoted_df['Surgery'].value_counts()
```

Out[6]:
```
0.0    2772
1.0     916
Name: Surgery, dtype: int64
```

Time to start modeling! Split target and features and make a baseline model.

In [7]:
```
y = pivoted_df['Surgery']
X = pivoted_df.drop('Surgery', axis=1)
```

In [8]:
```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0
```

In [9]:

```
#Make a pipeline to simplify process
logreg_pipeline = Pipeline([
    ('scale', StandardScaler()),
    ('logreg', LogisticRegression(solver='liblinear'))
])

# Define parameter grid to search
param_grid = {
    'logreg__C': [0.001, 0.01, 0.1, 1, 10, 100],  # Regularization st
    'logreg__penalty': ['l1', 'l2']  # Norm used in the penalization
}

# Initialize GridSearchCV with pipeline, parameter grid, and scoring 
grid_search = GridSearchCV(logreg_pipeline, param_grid, cv=5, scoring

# Assuming X_train and y_train are already defined
grid_search.fit(X_train, y_train)

# Best parameters found
print("Best parameters: ", grid_search.best_params_)

# Best cross-validation score
print("Best cross-validation score: {:.2f}".format(grid_search.best_s

# Test set score using the best parameters
print("Test set score: {:.2f}".format(grid_search.score(X_test, y_tes
```

```
/usr/local/lib/python3.10/dist-packages/sklearn/svm/_base.py:1244: Conve
rgenceWarning: Liblinear failed to converge, increase the number of iter
ations.
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/svm/_base.py:1244: Conve
rgenceWarning: Liblinear failed to converge, increase the number of iter
ations.
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/svm/_base.py:1244: Conve
rgenceWarning: Liblinear failed to converge, increase the number of iter
ations.
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/svm/_base.py:1244: Conve
rgenceWarning: Liblinear failed to converge, increase the number of iter
ations.
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/svm/_base.py:1244: Conve
rgenceWarning: Liblinear failed to converge, increase the number of iter
ations.
  warnings.warn(

Best parameters:  {'logreg__C': 10, 'logreg__penalty': 'l1'}
Best cross-validation score: 0.76
Test set score: 0.77
```

In [10]:
```
logreg_pipeline = Pipeline([
    ('scale', StandardScaler()),
    ('logreg', LogisticRegression(penalty='l1', C=10.0, solver='libli
])
```

In [11]:
```
logreg_pipeline.fit(X_train, y_train)
```

Out[11]:
```
Pipeline(steps=[('scale', StandardScaler()),
                ('logreg',
                 LogisticRegression(C=10.0, penalty='l1', solver='liblin
ear'))])
```

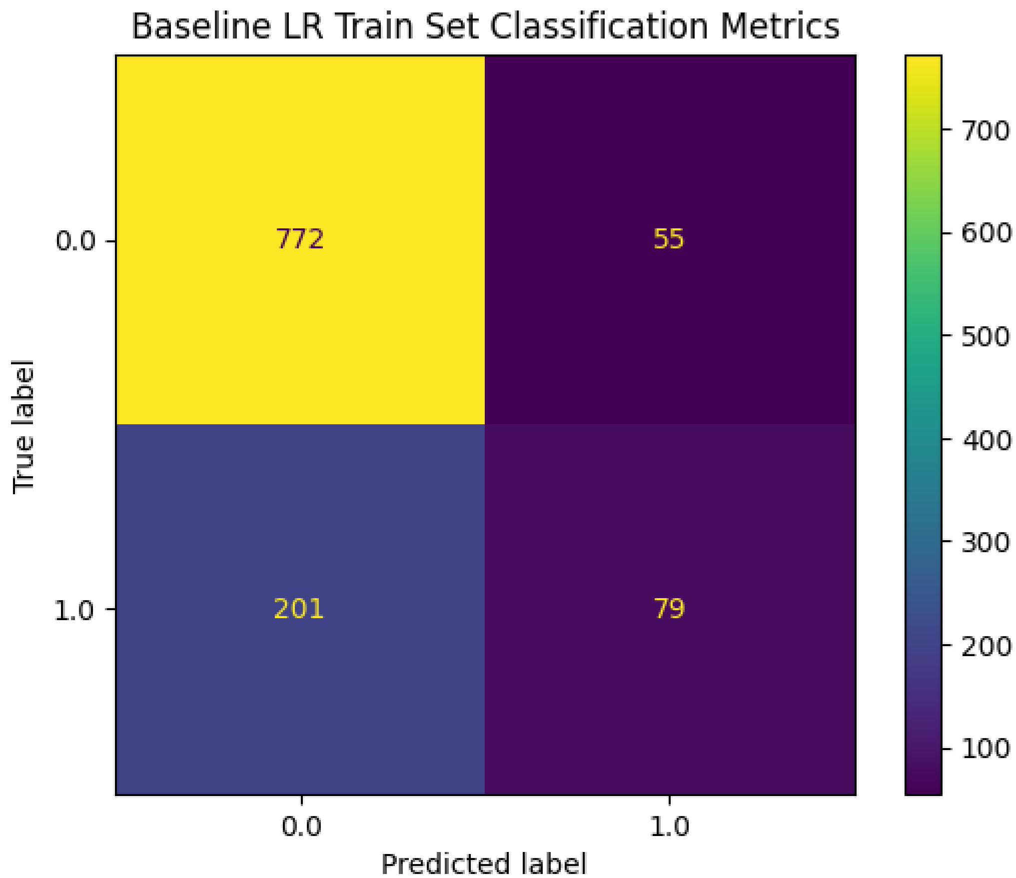
In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

In [12]:
```
logreg_pipeline.score(X_test, y_test)
```

Out[12]:
```
0.7687443541102078
```

In [13]:
```
y_pred = logreg_pipeline.predict(X_test)
```

In [14]:
```
ConfusionMatrixDisplay.from_predictions(y_test, y_pred)
plt.title('Baseline LR Train Set Classification Metrics')
plt.show()
print(classification_report(y_test, y_pred))
```




```
              precision    recall  f1-score   support

         0.0       0.79      0.93      0.86       827
         1.0       0.59      0.28      0.38       280

    accuracy                           0.77      1107
   macro avg       0.69      0.61      0.62      1107
weighted avg       0.74      0.77      0.74      1107
```

Dataset is imbalanced, need to adjust. Should also focus on Recall score since this is a medical issue (better to have False Positive than True Negative!)

In [15]:
```
y = pivoted_df['Surgery']
X = pivoted_df.drop('Surgery', axis=1)
```

In [16]:
```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0
```

In [17]:

```
# Set up pipeline
weight_logreg_pipeline = Pipeline([
    ('scale', StandardScaler()),
    ('logreg', LogisticRegression(solver='liblinear'))
])

# Define the parameter grid to search over, including class weights
param_grid = {
    'logreg__C': [0.01, 0.1, 1, 10],
    'logreg__penalty': ['l1', 'l2'],
    'logreg__class_weight': [None, 'balanced', {0: 1, 1: 2}, {0: 1, 1
    'logreg__max_iter': [5000],
    'logreg__tol': [0.01]
}

# Create a scoring function that focuses on recall for the positive c
recall_scorer = make_scorer(recall_score, pos_label=1)

# Initialize GridSearch with pipeline, param grid, and recall
grid_search = GridSearchCV(weight_logreg_pipeline, param_grid, cv=5,

# Fit the grid search to the data
grid_search.fit(X_train, y_train)

# Print the best parameters found and the best recall score
print("Best parameters: ", grid_search.best_params_)
print("Best cross-validation recall score: {:.2f}".format(grid_search

# Evaluate the best model on the test set
best_model = grid_search.best_estimator_
y_pred = best_model.predict(X_test)
print("Test set recall score: {:.2f}".format(recall_score(y_test, y_p
```

```
Best parameters:  {'logreg__C': 0.01, 'logreg__class_weight': {0: 1, 1:
5}, 'logreg__max_iter': 5000, 'logreg__penalty': 'l1', 'logreg__tol': 0.
01}
Best cross-validation recall score: 0.87
Test set recall score: 0.85
```

In [18]:

```
best_model.fit(X_train, y_train)
```

Out[18]:

```
Pipeline(steps=[('scale', StandardScaler()),
                ('logreg',
                 LogisticRegression(C=0.01, class_weight={0: 1, 1: 5},
                                    max_iter=5000, penalty='l1',
                                    solver='liblinear', tol=0.01))])
```

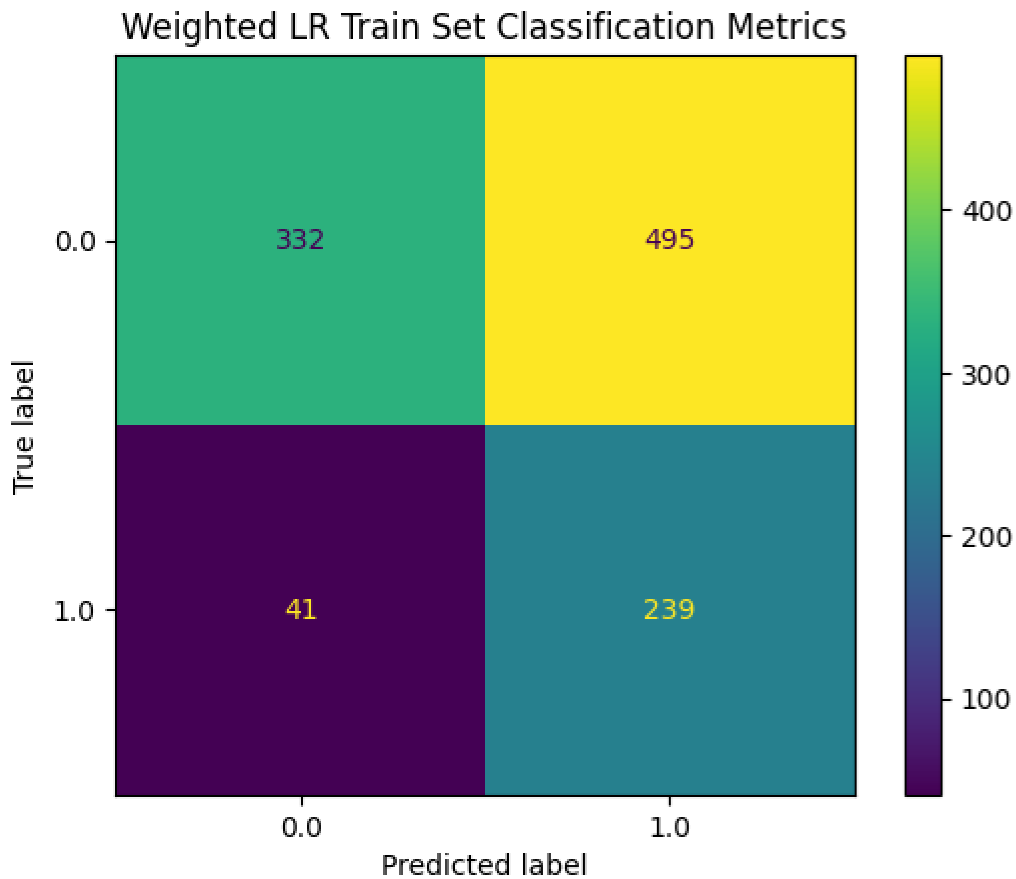
In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

In [19]:
```
best_model.score(X_test, y_test)
```

Out[19]: 0.5158084914182475

In [20]:
```
y_pred = best_model.predict(X_test)
```

In [21]:
```
ConfusionMatrixDisplay.from_predictions(y_test, y_pred)
plt.title('Weighted LR Train Set Classification Metrics')
plt.show()
print(classification_report(y_test, y_pred))
```




```
              precision    recall  f1-score   support

         0.0       0.89      0.40      0.55       827
         1.0       0.33      0.85      0.47       280

    accuracy                           0.52      1107
   macro avg       0.61      0.63      0.51      1107
weighted avg       0.75      0.52      0.53      1107
```

Much better model. False Negatives is low, other classes much higher.

In [22]:

```
coef = best_model['logreg'].coef_
```

In [23]:

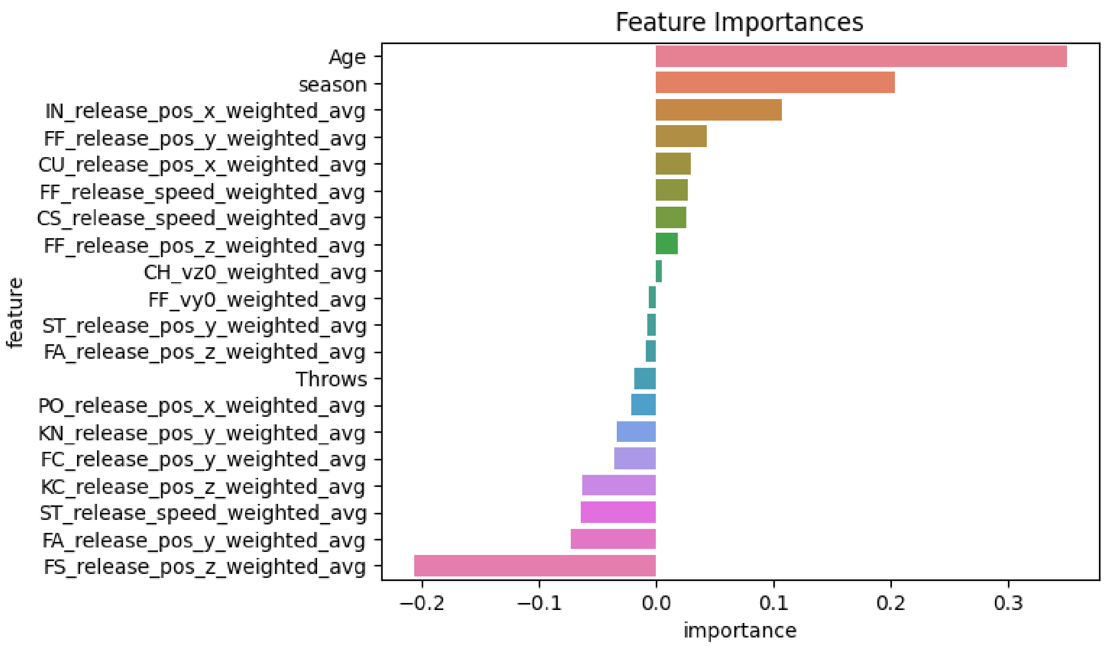
```
features = pivoted_df.columns

zipped = zip(features, coef[0])
sorted_pairs = sorted(zipped, key=lambda x: x[1], reverse=True)
sorted_pairs

feature_importances = pd.DataFrame(sorted_pairs, columns=['feature',
feature_importances = feature_importances[abs(feature_importances['im
```

In [24]:

```
sns.barplot(x='importance', y='feature', data=feature_importances, hu
plt.title('Feature Importances')
plt.show()
```



Need to update feature names so they can be understood more easily.

In [25]:
```
feature_importances
```

Out[25]:

	feature	importance
0	Age	0.350326
1	season	0.204089
2	IN_release_pos_x_weighted_avg	0.106698
3	FF_release_pos_y_weighted_avg	0.042848
4	CU_release_pos_x_weighted_avg	0.028877
5	FF_release_speed_weighted_avg	0.026781
6	CS_release_speed_weighted_avg	0.026057
7	FF_release_pos_z_weighted_avg	0.017854
8	CH_vz0_weighted_avg	0.004592
118	FF_vy0_weighted_avg	-0.005897
119	ST_release_pos_y_weighted_avg	-0.007521
120	FA_release_pos_z_weighted_avg	-0.008562
121	Throws	-0.018381
122	PO_release_pos_x_weighted_avg	-0.020677
123	KN_release_pos_y_weighted_avg	-0.033399
124	FC_release_pos_y_weighted_avg	-0.036234
125	KC_release_pos_z_weighted_avg	-0.063403
126	ST_release_speed_weighted_avg	-0.064187
127	FA_release_pos_y_weighted_avg	-0.072684
128	FS_release_pos_z_weighted_avg	-0.206804

In [26]:

```
# This will rename the index if 'feature' is actually set as the inde
feature_importances = feature_importances.set_index('feature')  # Mak
feature_importances = feature_importances.rename(index={
    'season': 'Season',
    'IN_release_pos_x_weighted_avg': 'Intentional Ball Release Positi
    'FF_release_pos_y_weighted_avg': 'Fastball Release Position Y-dim
    'CU_release_pos_x_weighted_avg': 'Curveball Release Position X-di
    'FF_release_pos_z_weighted_avg': 'Fastball Release Position Z-dim
    'CU_release_pos_z_weighted_avg': 'Curveball Release Position Z-di
    'CS_release_speed_weighted_avg': 'Slow Curve Release Speed avg.',
    'FF_release_speed_weighted_avg': 'Fastball Release Speed avg.',
    'CU_release_speed_weighted_avg': 'Curveball Release Speed avg.',
    'ST_vy0_weighted_avg': 'Sweeper Velocity Y-dimension avg.',
    'CH_vz0_weighted_avg': 'Changeup Velocity Z-dimension avg.',
    'FF_vy0_weighted_avg': 'Fastball Velocity Y-dimension avg.',
    'ST_release_speed_weighted_avg': 'Sweeper Release Speed avg.',
    'CS_vy0_weighted_avg': 'Slow Curve Velocity Y-dimension avg.',
    'PO_release_pos_x_weighted_avg': 'Pick-off Release Position X-dim
    'ST_release_pos_z_weighted_avg': 'Sweeper Release Position Z-dime
    'ST_release_pos_y_weighted_avg': 'Sweeper Release Position Y-dime
    'KN_release_pos_y_weighted_avg': 'Knuckleball Release Position Y-
    'FC_release_pos_y_weighted_avg': 'Cutter Release Position Y-dimen
    'KC_release_pos_z_weighted_avg': 'Knucklecurve Release Position Z
    'FA_release_pos_y_weighted_avg': '4-seam Fastball Release Positio
    'FS_release_pos_z_weighted_avg': 'Split-finger Release Position Z

})
```

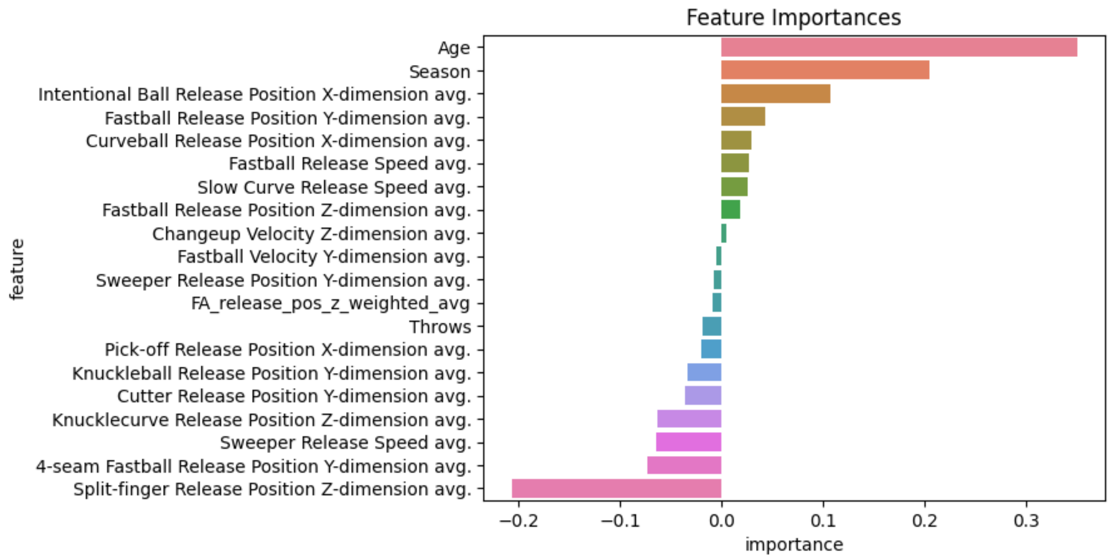
In [27]:

```
feature_importances
```

Out[27]:

feature	importance
Age	0.350326
Season	0.204089
Intentional Ball Release Position X-dimension avg.	0.106698
Fastball Release Position Y-dimension avg.	0.042848
Curveball Release Position X-dimension avg.	0.028877
Fastball Release Speed avg.	0.026781
Slow Curve Release Speed avg.	0.026057
Fastball Release Position Z-dimension avg.	0.017854
Changeup Velocity Z-dimension avg.	0.004592
Fastball Velocity Y-dimension avg.	-0.005897
Sweeper Release Position Y-dimension avg.	-0.007521
FA_release_pos_z_weighted_avg	-0.008562
Throws	-0.018381
Pick-off Release Position X-dimension avg.	-0.020677
Knuckleball Release Position Y-dimension avg.	-0.033399
Cutter Release Position Y-dimension avg.	-0.036234
Knucklecurve Release Position Z-dimension avg.	-0.063403
Sweeper Release Speed avg.	-0.064187
4-seam Fastball Release Position Y-dimension avg.	-0.072684
Split-finger Release Position Z-dimension avg.	-0.206804

In [28]:
```
sns.barplot(x='importance', y='feature', data=feature_importances, hu
plt.title('Feature Importances')
plt.show()
```



This shows the features that have the most impact in predicting 1.0 surgery (positive and negative)

Decision Tree Classifier, baseline model.

In [29]:
```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0
```

In [30]:
```
tree_clf = DecisionTreeClassifier(criterion='gini', max_depth=5)
tree_clf.fit(X_train, y_train)
```

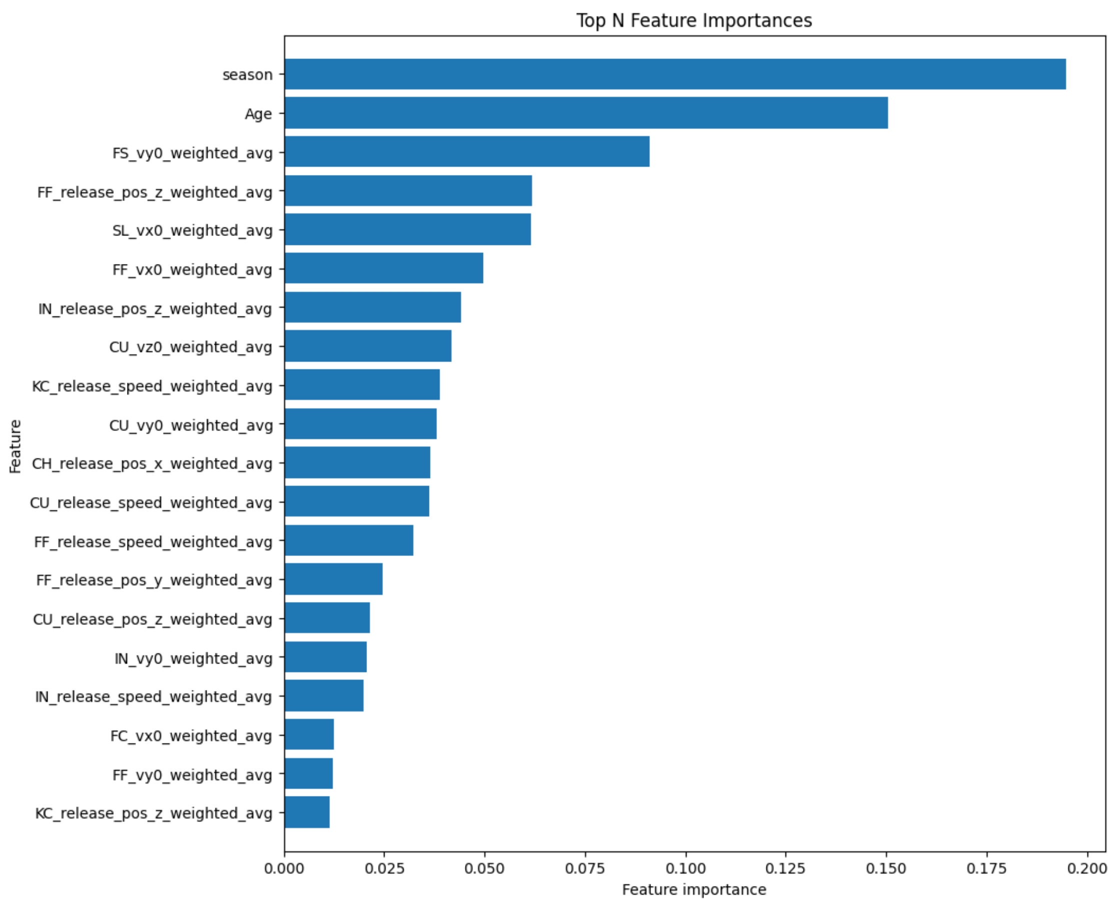
Out[30]:
```
DecisionTreeClassifier(max_depth=5)
```

In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

In [31]:

```
def plot_feature_importances(model, n_top_features=20):
    importances = model.feature_importances_
    indices = np.argsort(importances)[-n_top_features:]
    plt.figure(figsize=(10,10))
    plt.title('Top N Feature Importances')
    plt.barh(range(n_top_features), importances[indices], align='cent
    plt.yticks(range(n_top_features), [X_train.columns[i] for i in in
    plt.xlabel('Feature importance')
    plt.ylabel('Feature')
    plt.ylim(-1, n_top_features)

plot_feature_importances(tree_clf, n_top_features=20)
plt.show()
```



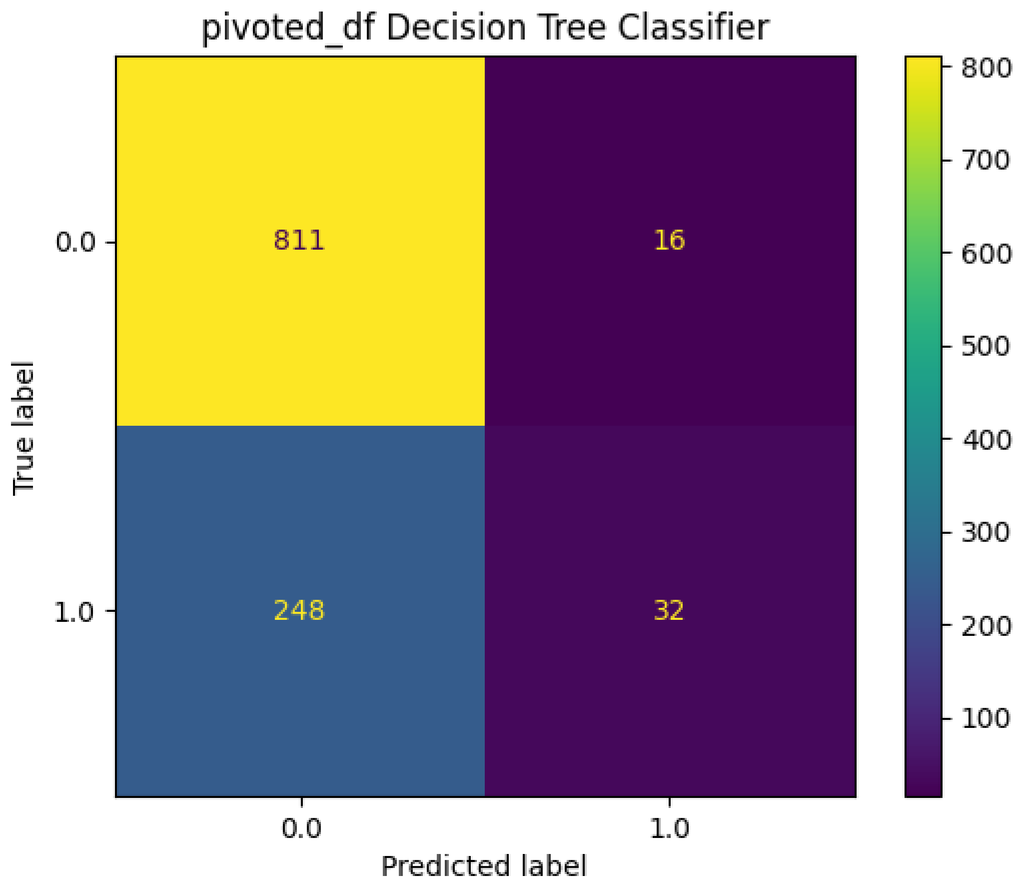
In [32]:

```
pred = tree_clf.predict(X_test)
```

In [33]:

```
pred = tree_clf.predict(X_test)

ConfusionMatrixDisplay.from_predictions(y_test, pred)
plt.title('pivoted_df Decision Tree Classifier')
plt.show()
print(classification_report(y_test, pred))
```




```
              precision    recall  f1-score   support

         0.0       0.77      0.98      0.86       827
         1.0       0.67      0.11      0.20       280

    accuracy                           0.76      1107
   macro avg       0.72      0.55      0.53      1107
weighted avg       0.74      0.76      0.69      1107
```

Terrible for TP and FP. Need to adjust. Features are interesting. Mostly fastball, curveball, some slider and split-finger.

In [34]:

```
y = pivoted_df['Surgery']
X = pivoted_df.drop('Surgery', axis=1)
```

In [35]:
```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0
```

In [36]:
```
param_grid = {
    'criterion': ['gini', 'entropy'],
    'max_depth': [5, 10, 15, 20],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4],
    'class_weight': ['balanced', {0:1, 1:2}, {0:1, 1:3}]
}

tree_clf = DecisionTreeClassifier()
scorer = make_scorer(recall_score)
grid_search = GridSearchCV(estimator=tree_clf, param_grid=param_grid,
grid_search.fit(X_train, y_train)

print("Best parameters:", grid_search.best_params_)
print("Best score:", grid_search.best_score_)

best_tree = grid_search.best_estimator_
y_pred = best_tree.predict(X_test)
print("Test recall score:", recall_score(y_test, y_pred))
```

```
Best parameters: {'class_weight': 'balanced', 'criterion': 'gini', 'max_
depth': 5, 'min_samples_leaf': 2, 'min_samples_split': 10}
Best score: 0.6540600393700788
Test recall score: 0.7607142857142857
```

In [37]:
```
best_tree.fit(X_train, y_train)
```

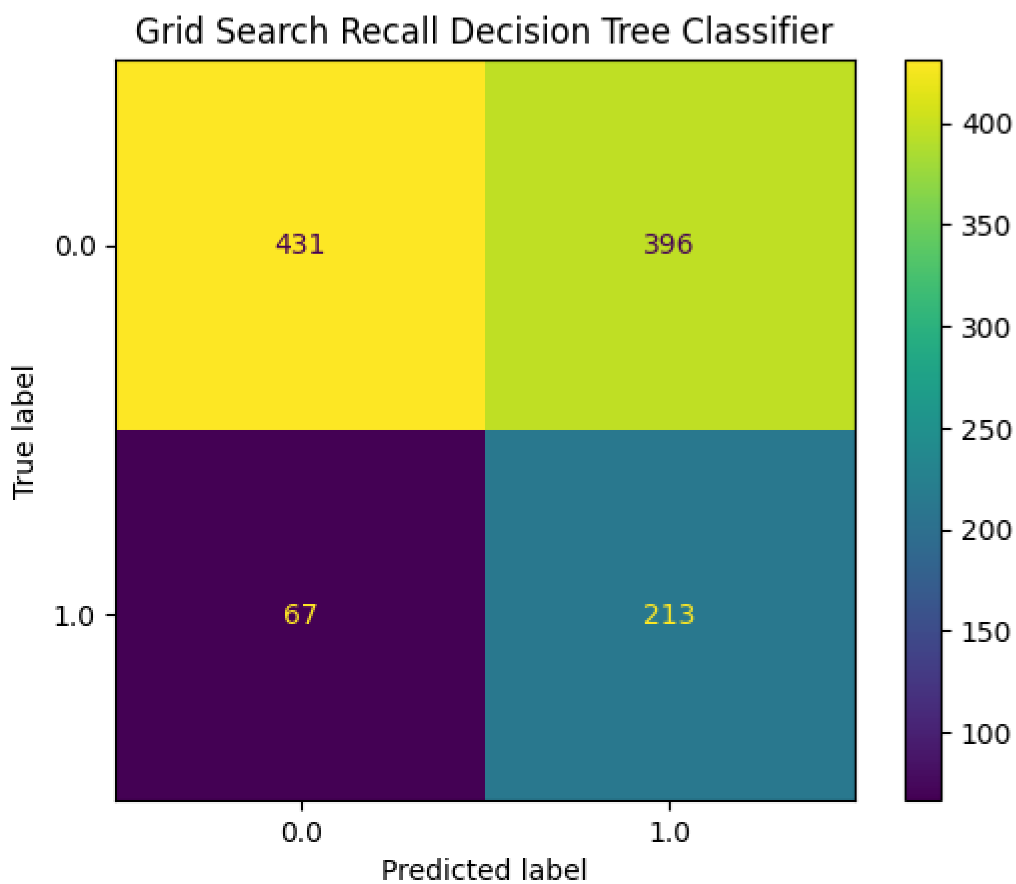
Out[37]:
```
DecisionTreeClassifier(class_weight='balanced', max_depth=5, min_samples
_leaf=2,
                       min_samples_split=10)
```

In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

In [38]:
```
pred = best_tree.predict(X_test)
```

In [39]:

```
ConfusionMatrixDisplay.from_predictions(y_test, pred)
plt.title('Grid Search Recall Decision Tree Classifier')
plt.show()
print(classification_report(y_test, pred))
```




```
              precision    recall  f1-score   support

         0.0       0.87      0.52      0.65       827
         1.0       0.35      0.76      0.48       280

    accuracy                           0.58      1107
   macro avg       0.61      0.64      0.56      1107
weighted avg       0.74      0.58      0.61      1107
```

The Logistic Regression model with adjusted class weights performed the best.

In [ ]:

```

```