Rakamin Academy

Laporan Final Project Byte Blazers

- Muhamad Faiz Widagdo
- Robiatul Adawiyah
- Chianti Ridhwan
- Lulu Safira
- Retno Debbi Yulisya
- Imam Luthfi
- Melliza Nastasia Izazi

Rakamin Academy

STAGE 3

Machine Learning Modeling

Modelling Experiments

1.) Split Data Train and Test

X_train, X_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=42)

2.) Train Model

Random Forest

Accuracy (Test Set): 0.6680
Precision (Test Set): 0.5977
Recall (Test Set): 0.5200
F1-Score (Test Set): 0.5561
ROC AUC (Test-proba): 0.6610
ROC AUC (Train-proba): 0.9485

Decision Tree

Accuracy (Test Set): 0.6320
Precision (Test Set): 0.5541
Recall (Test Set): 0.4100
F1-Score (Test Set): 0.4713
ROC AUC (Test-proba): 0.5811
ROC AUC (Train-proba): 0.9514

Logistic Regression

Accuracy (Test Set): 0.6960
Precision (Test Set): 0.7308
Recall (Test Set): 0.3800
F1-Score (Test Set): 0.5000
ROC AUC (Test-proba): 0.6722
ROC AUC (Train-proba): 0.7041

Gradient Boosting

Accuracy (Test Set): 0.7640
Precision (Test Set): 0.8596
Recall (Test Set): 0.4900
F1-Score (Test Set): 0.6242
ROC AUC (Test-proba): 0.7468
ROC AUC (Train-proba): 0.8368

Kesimpulan: Penanganan class imbalance membantu meningkatkan recall pada beberapa model seperti Random Forest, tetapi tidak memberikan perubahan yang signifikan pada model lainnya seperti Logistic Regression dan Decision Tree. Model Gradient Boosting menunjukkan kinerja yang stabil tanpa banyak perubahan setelah penanganan *class imbalance*.

Analisis dan kesimpulan hasil evaluasi metrics recall Model:

Random Forest:
- Recall : 0.52
- Dengan penanganan class imbalance, terjadi peningkatan signifikan pada recall dari model ini. Penanganan class imbalance dapat membantu model untuk lebih baik dalam mengidentifikasi kasus positif.

Logistic Regression:
- Recall: 0.38
- Model ini menunjukkan bahwa penanganan class imbalance tidak memberikan dampak yang signifikan pada kemampuan model untuk mengidentifikasi kasus positif.

Decision Tree:
- Recall: 0.41
- Penanganan class imbalance tidak menghasilkan perubahan yang signifikan dalam recall model Decision Tree.

Gradient Boosting:
- Recall: 0.49
- Model ini menunjukkan hasil yang stabil, tanpa perubahan yang signifikan setelah penanganan class imbalance. Recall yang relatif tinggi menunjukkan bahwa model ini mungkin sudah cukup baik dalam mengidentifikasi kasus positif.

Alasan kami mengamati Recall dikarenakan kami memiliki Goal untuk membuat prediksi customer yang akan membeli travel insurance dari faktor-faktor yang dimiliki oleh customer penerbangan.

Modelling Experiments

3.) Hyperparameter Tuning

```
param_grid = {
    'bootstrap': [True],
    'max_depth': [100, 200, 300],
    'max_features': [2, 3],
    'min_samples_leaf': [3, 4, 5],
    'min_samples_split': [8, 10, 12],
    'criterion' : ['gini', 'entropy'],
    'n_estimators': [100, 200, 300]
}

cv = GridSearchCV(estimator = rf, param_grid = param_grid, cv = 3)
cv.fit(X_train, y_train)
best_params = cv.best_params_
print('The best params are:',best_params)
```

```
tuned_rf = RandomForestClassifier(bootstrap=True, criterion='gini', max_depth=100, max_features=2,
                                  min_samples_leaf=3, min_samples_split=10, n_estimators=100)
tuned_rf.fit(X_train, y_train)
y_pred = tuned_rf.predict(X_test)
eval_classification(tuned_rf)
```

Random Forest

Accuracy (Test Set) : 0.7640
Precision (Test Set) : 0.8475
Recall (Test Set) : 0.5000
F1-Score (Test Set) : 0.6289
ROC AUC (Test-proba) : 0.7088
ROC AUC (Train-proba) : 0.8723

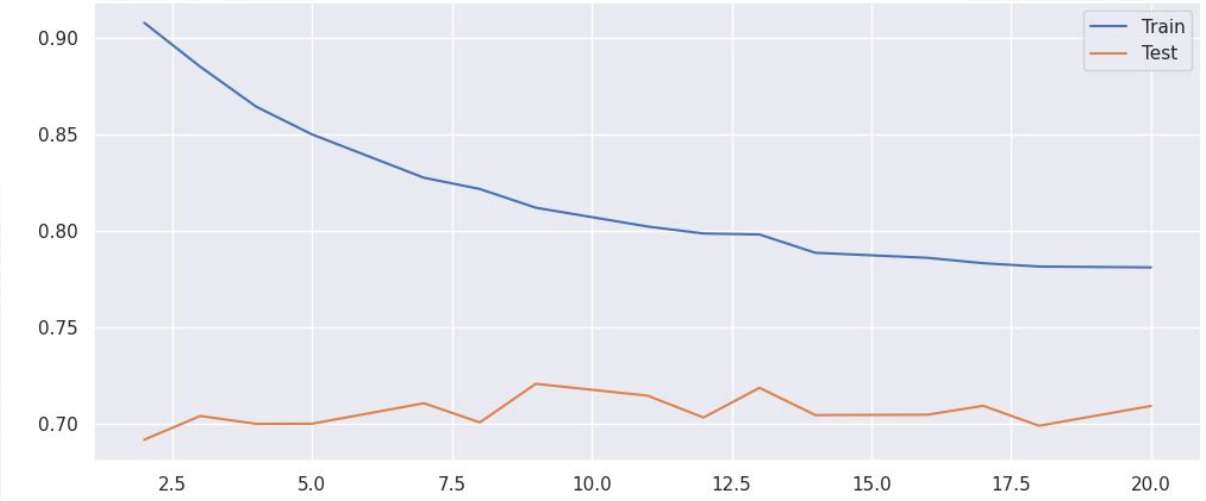
Mempertimbangkan faktor-faktor ini, **Random Forest** memiliki **kinerja keseluruhan terbaik** berdasarkan matriks evaluasi yang dilakukan.

Dari **hyperparameter tuning** yang telah dilakukan, didapatkan hasil**:**

- **Random Forest** menunjukkan peningkatan dalam beberapa metrics evaluasi, termasuk **akurasi (0.7640), presisi (0.8475), dan F1-Score (0.6289), dan recall (0.5000)**
- **Setelah tuning**, model Random Forest menunjukkan **peningkatan** dalam beberapa metrik evaluasi, termasuk **akurasi (0.7640)**, **presisi (0.8475), F1-score (0.6289), dan recall (0.5000)**.
- Hasil ini menunjukkan bahwa penanganan hyperparameter berhasil meningkatkan kinerja model Random Forest secara signifikan, terutama dalam hal mengidentifikasi **kasus positif**.

Learning Curve

Dilakukan learning curve untuk mengetahui nilai optimal dari fungsi model yang telah dibuat




```
param value: 2; train: 0.9077811874809263; test: 0.6917666666666666
param value: 3; train: 0.8850327218473433; test: 0.7040333333333333
param value: 4; train: 0.864339968125869; test: 0.6999666666666667
param value: 5; train: 0.8499415075785833; test: 0.7000333333333334
param value: 7; train: 0.827502458377132; test: 0.7106333333333333
param value: 8; train: 0.8216616933979859; test: 0.7007
param value: 9; train: 0.8119638194703469; test: 0.7207
param value: 11; train: 0.8022150825675631; test: 0.7145666666666666
param value: 12; train: 0.7985995727510088; test: 0.7032333333333334
param value: 13; train: 0.7980951815808213; test: 0.7186333333333333
param value: 14; train: 0.7885965209725001; test: 0.7045
param value: 16; train: 0.7860003899494761; test: 0.7047
param value: 17; train: 0.7831965684446102; test: 0.7092999999999999
param value: 18; train: 0.7814820623240989; test: 0.6989666666666667
param value: 20; train: 0.7810306534196874; test: 0.7092333333333334
```

- **Akurasi training** data **meningkat** drastis pada **titik awal**, walaupun berkurang sedikit sehingga **stabil** pada **titik** sekitar **0,90** hal ini menunjukkan bahwa **model dapat mengambil konsisten dari data train**.
- Sedangkan **akurasi test** data **meningkat** secara perlahan **seiring** dengan **peningkatan akurasi training data**, tetapi kemudian **stabil** pada titik sekitar **0,85**. Hal ini menunjukkan bahwa model memiliki performa yang baik pada data uji, tetapi dapat menunjukkan tanda overfitting pada titik lebih tinggi.
- **Selisih antara akurasi training data dan test data** menunjukkan bahwa **model memiliki sedikit overfitting** pada **titik lebih tinggi**. Namun, selisih tersebut tidak terlalu besar, sehingga **model masih memiliki performa yang baik pada data uji.**

Rakamin Academy

Website Url : https://github.com/BYTE-BLAZERS/Stage_3.git

GitHub CLI: ***gh repo clone BYTE-BLAZERS/Stage_3***