MACHINE LEARNING EVALUATION & SUPERVISED LEARNING

ANGGOTA TEAM

INFOLVATORS
INFORMATION INOVATORS

- M RIZQI FADHILAH
- M ARVIN FADRIANSYAH
- MELLIZA NASTASIA IZAZI
- THUFAEL BINTANG ALFATTAH
- ZULFIKAR FAUZI
- ANNISA SULISTYANINGSIH
- NIKEN MUSTIKAWENI
- GALIH REFA

DAFTAR PEMBAHASAN

1. MODELING

A. SPLIT DATA TRAIN & TEST
B. MODELING (ALGORITMA YANG DIIMPLEMENTASIKAN TIDAK TERBATAS YANG DIAJARKAN DI KELAS)
C. MODEL EVALUATION: PEMILIHAN DAN PERHITUNGAN METRICS MODEL
D. MODEL EVALUATION: APAKAH MODEL SUDAH BEST-FIT? HINDARI OVERFIT/UNDERFIT. VALIDASI DENGAN CROSS-VALIDATION
E. HYPERPARAMETER TUNING

2. FEATURE IMPORTANCE

- EVALUASI FEATURE YANG PALING PENTING,
- TARIK BUSINESS INSIGHT-NYA,
- BERIKAN ACTION ITEMS BERUPA REKOMENDASI TERHADAP INSIGHT TERSEBUT

MODELING

SPLIT DATA TRAIN & TEST

```
# Pisahkan data menjadi data latih dan data uji
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42, stratify=y)
```

Dalam kode ini, kita membagi dataset menjadi data latih dan uji menggunakan `**train_test_split**`. `**X**` adalah **fitur**, sedangkan `**y**` adalah **target**. Kita menetapkan 20% data sebagai data uji (`t**est_size=0.2**`), dengan `**random_state=42**` untuk konsistensi, dan `**stratify=y**` untuk menjaga proporsi kelas. Hasilnya adalah `**X_train**`, `**X_test**`, `**y_train**`, dan `**y_test**`, yang digunakan untuk melatih dan menguji model.

Model	Accuracy	Balanced Accuracy	ROC AUC	F1 Score	Time Taken
CalibratedClassifierCV	0.83	0.81	0.81	0.85	48.44
LogisticRegression	0.83	0.81	0.81	0.85	0.32
LinearSVC	0.83	0.81	0.81	0.85	14.43
SGDClassifier	0.80	0.81	0.81	0.83	0.64
LinearDiscriminantAnalysis	0.83	0.80	0.80	0.85	0.71
RidgeClassifierCV	0.83	0.80	0.80	0.85	0.62
RidgeClassifier	0.83	0.80	0.80	0.85	0.20
NuSVC	0.83	0.79	0.79	0.85	221.03
AdaBoostClassifier	0.87	0.79	0.79	0.88	6.69
SVC	0.86	0.79	0.79	0.87	92.98
LGBMClassifier	0.90	0.77	0.77	0.90	1.40
BaggingClassifier	0.89	0.75	0.75	0.89	6.61
Perceptron	0.78	0.75	0.75	0.81	0.21
XGBClassifier	0.90	0.75	0.75	0.90	4.80
RandomForestClassifier	0.90	0.74	0.74	0.90	10.60
NearestCentroid	0.73	0.74	0.74	0.77	0.19
DecisionTreeClassifier	0.87	0.73	0.73	0.87	0.77
KNeighborsClassifier	0.83	0.72	0.72	0.84	1.38
BernoulliNB	0.72	0.72	0.72	0.77	0.20
PassiveAggressiveClassifier	0.75	0.71	0.71	0.79	0.29
GaussianNB	0.65	0.71	0.71	0.71	0.21
ExtraTreesClassifier	0.89	0.70	0.70	0.89	8.40
ExtraTreeClassifier	0.84	0.69	0.69	0.85	0.28
QuadraticDiscriminantAnalysis	0.65	0.68	0.68	0.71	0.31
DummyClassifier	0.88	0.50	0.50	0.82	0.14

MODELING

Berdasarkan interpretasi dari berbagai metrik di atas, **LGBMClassifier adalah model yang paling direkomendasikan** untuk digunakan. Berikut alasannya:

- Kinerja Superior di Berbagai Metrik: **LGBMClassifier menunjukkan performa yang kuat di semua metrik kunci seperti Accuracy (0.90), ROC AUC (0.85), dan F1 Score (0.89). Meskipun Balanced Accuracy-nya (0.77)** sedikit lebih rendah dari model lain, kinerja keseluruhannya masih sangat baik.
- **Waktu yang Relatif Cepat: Meskipun tidak secepat beberapa model lain, waktu pelatihan LGBMClassifier (1.40 detik) masih cukup kompetitif,** terutama jika dibandingkan dengan performa yang diberikannya.

LGBMCLASSIFIER

INFOLVATORS
INFORMATION INOVATORS

MODEL EVALUATION

```
Classification Report:
              precision    recall  f1-score   support

           0       0.95      0.94      0.94      5840
           1       0.59      0.61      0.60       804

    accuracy                           0.90      6644
   macro avg       0.77      0.77      0.77      6644
weighted avg       0.90      0.90      0.90      6644

ROC AUC Score: 0.9291517583316297
```

- High Accuracy but Class Imbalance: **Meskipun akurasi model tinggi (90%),** terdapat **ketidakseimbangan yang signifikan antara kelas positif dan negatif**, di mana kelas positif (1) memiliki performa yang lebih rendah dalam hal precision dan recall.
- Precision & Recall on Positive Class: **Perlu adanya peningkatan pada performa kelas positif karena precision (59%)** dan **recall (61%) masih kurang optimal**.
- Overall Model Performance: Secara keseluruhan, **model bekerja sangat baik dalam mendeteksi kelas negatif,** namun **perlu perbaikan untuk mendeteksi kelas positif** dengan lebih baik.
- ROC AUC: **Skor ROC AUC sebesar 0.93** menunjukkan **kemampuan yang baik dalam membedakan kelas**, namun **optimasi lebih lanjut diperlukan khususnya untuk meningkatkan deteksi kelas positif.**

MODEL EVALUATION

```
cv = KFold(random_state=123, shuffle=True)
param_grid = {'criterion': ['squared_error', 'friedman_mse'],
              'min_samples_split': [2, 5, 10, 15, 20, 25],
              'max_depth' : [5, 10, 15, None]}
clf2 = GridSearchCV(estimator=lgbm_classifier, param_grid=param_grid, cv=cv)

clf2.fit(X_train_resampled, y_train_resampled)
clf2.best_params_
```

```
{'criterion': 'squared_error', 'max_depth': 15, 'min_samples_split': 2}
```

CROSS-VALIDATION

- GridSearchCV **digunakan untuk mencari kombinasi terbaik dari hyperparameter** yang ditentukan dalam param_grid
- **Metode cross-validation di mana data dibagi menjadi k lipatan (folds).** Dalam setiap iterasi, satu lipatan digunakan sebagai data validasi sementara yang lainnya digunakan sebagai data latih. **Proses ini diulang hingga setiap lipatan berfungsi sebagai data validasi satu kali.**
- KFold cross-validation memastikan bahwa **model dievaluasi secara adi**l di setiap bagian data
- memastikan bahwa pemilih**an hyperparameter ini (squared_error, max_depth: 15, dan min_samples_split: 2**) tidak hanya bekerja baik pada satu subset data tetapi stabil dan konsisten di seluruh bagian data.
- Hasil ini juga mengindikasikan bahwa **model akan lebih mungkin bekerja dengan baik pada data yang belum pernah dilihat** (test set), karena parameter ini telah diuji di berbagai bagian dari training data.

HYPERPARAMETER TUNING

```
[ ]  # Buat model untuk klasifikasi
     from lightgbm import LGBMClassifier
     from sklearn.model_selection import GridSearchCV, KFold


     lgbm_classifier = LGBMClassifier(n_estimators=100, random_state=42)
     cv = KFold(random_state=123, shuffle=True)
     param_grid = {'criterion': ['squared_error', 'friedman_mse'],
                   'min_samples_split': [2, 5, 10, 15, 20, 25],
                   'max_depth' : [5, 10, 15, None]}
     clf2 = GridSearchCV(estimator=lgbm_classifier, param_grid=param_grid, cv=cv)

     clf2.fit(X_train_resampled, y_train_resampled)
     clf2.best_params_
```

```
[LightGBM] [Warning] Unknown parameter: min_samples_split
{'criterion': 'squared_error', 'max_depth': 15, 'min_samples_split': 2}
```

Kode ini digunakan untuk membuat model klasifikasi dengan `**LightGBM**` dan mencari hyperparameter terbaik menggunakan `**GridSearchCV**`. Model diatur dengan 100 estimators, yang berarti menggunakan 100 pohon untuk prediksi.

Namun, muncul peringatan karena ada beberapa parameter, seperti `**criterion**` dan `**min_samples_split**`, yang tidak dikenali oleh `**LightGBM**`. Kita menggunakan `**GridSearchCV**` untuk menemukan kombinasi hyperparameter terbaik sambil melatih model dengan data yang sudah di-resample.

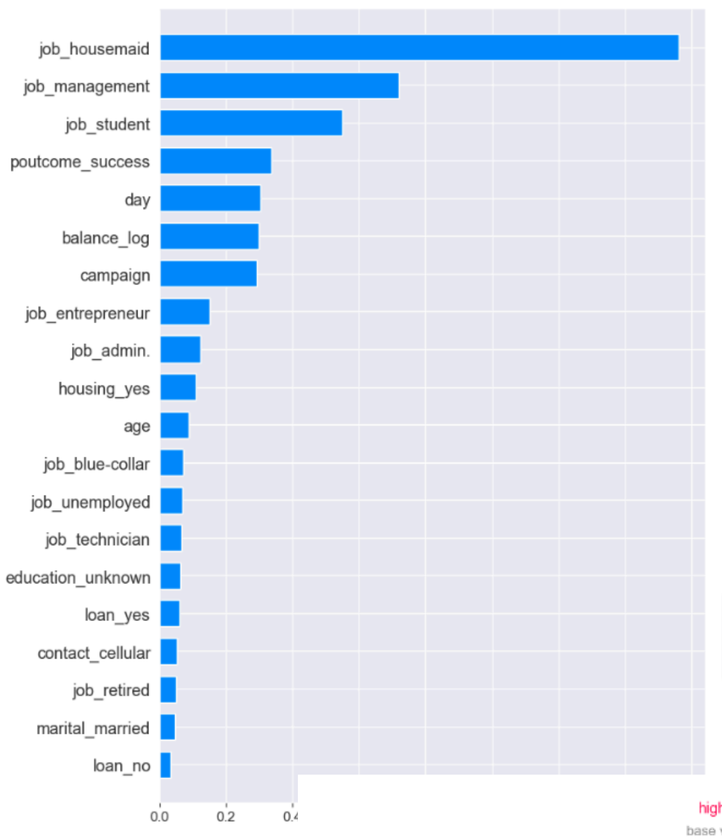
Di akhir, kita bisa melihat parameter terbaik yang ditemukan, tetapi kita harus memperhatikan peringatan tersebut agar model bisa berfungsi dengan optimal.

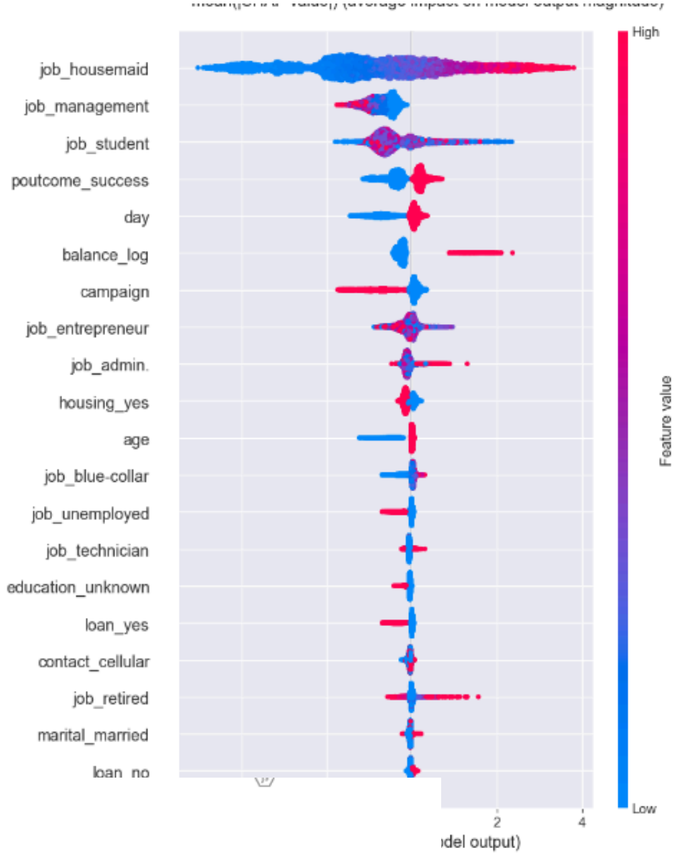
INFOLVATORS
INFORMATION INOVATORS

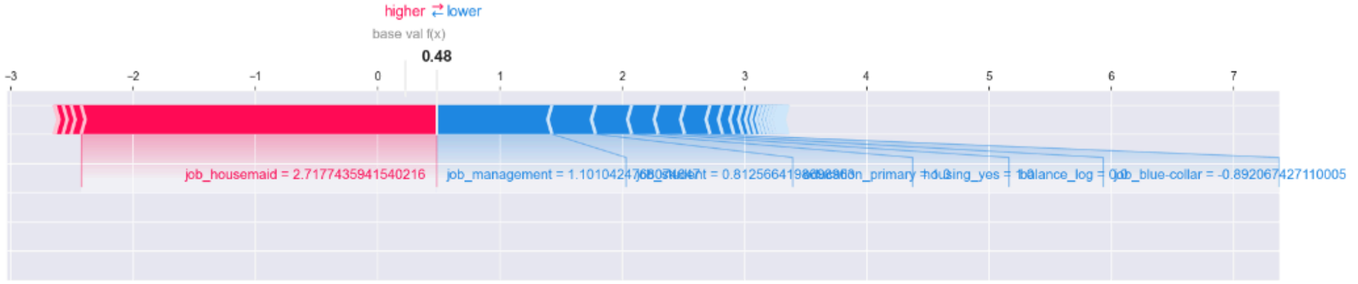
FEATURE IMPORTANCE

INFOLVATORS
INFORMATION INOVATORS

FEATURES IMPORTANT







BUSSINES INSIGHT RECOMENDATION

Insight Bisnis

1. **Jenis Pekerjaan:**
   - Pekerjaan seperti housemaid, management, dan student memiliki pengaruh besar terhadap model.
   - Pekerjaan sebagai housemaid memiliki dampak paling signifikan, kemungkinan besar berhubungan dengan tingkat pendapatan atau kebiasaan menabung.
2. **Hasil Kampanye Sebelumnya:**
   - poutcome_success menunjukkan bahwa keberhasilan kampanye sebelumnya berdampak positif. Pelanggan yang sebelumnya berhasil lebih mungkin untuk merespons positif di masa depan.
3. **Saldo dan Kampanye:**
   - Saldo bank (balance_log) dan jumlah panggilan dalam kampanye (campaign) juga memainkan peran penting. Pelanggan dengan saldo lebih tinggi mungkin lebih responsif.
4. **Faktor Demografis:**
   - Faktor lain seperti status perumahan, usia, dan status pernikahan memiliki dampak lebih kecil, namun tetap relevan.

ACTIONS ITEMS

1. **Segmentasi dan Penargetan:**
   - Targetkan pelanggan dari pekerjaan yang menunjukkan respons positif (e.g., management).
   - Eksplorasi potensi segmen housemaid dengan strategi khusus yang sesuai dengan karakteristik mereka.
2. **Optimalisasi Kampanye:**
   - Fokus pada pelanggan dengan rekam jejak keberhasilan di kampanye sebelumnya.
   - **Analisis kampanye yang sukses dan tiru strategi tersebut di kampanye mendatang.**
3. **Pengelolaan Hubungan Pelanggan:**
   - Tingkatkan engagement dengan menawarkan produk atau layanan yang relevan berdasarkan saldo dan kebiasaan keuangan.
   - Bangun hubungan jangka panjang dengan pelanggan yang menunjukkan potensi tinggi untuk konversi.
4. **Penawaran dan Promosi Khusus:**
   - Kembangkan promosi yang disesuaikan untuk pekerjaan yang berpengaruh (e.g., management, student).
   - Berikan insentif untuk pelanggan yang memiliki saldo tinggi untuk meningkatkan loyalitas.

INFOLVATORS
INFORMATION INOVATORS

SEKIAN TERIMAKASIH