Rakamin Academy

Homework

Final Project - Stage 3
Machine Learning Evaluation

1. Modelling

A. Data Splitting

```
# mendefinisikan X (features) dan y (target), dimana X dan y hasil dari imbalance handling sebelumnya
X_train = X_resampled
y_train = y_resampled
```

Data telah di split antara data training dan data testing di tahap Pre-Processing. Sebelum melakukan pemodelan, akan didefinisikan kembali data train yang sudah dilakukan imbalance handling yaitu oversampling dengan teknik SMOTE dikarenakan ingin meningkatkan jumlah kuantitas suatu label.

1. Modelling

Rakamin Academy

B. Modelling

Support Vector Machine (SVM), Adaboost, dan XGBoost akan menjadi algoritma yang digunakan. Pemilihan ketiga model ini disebabkan oleh rendahnya tingkat kesalahan dan tingginya tingkat presisi yang telah teruji berdasarkan percobaan dengan 8 model.

	CV Precision	Precision_Train	Precision_Test	Diff	Diff (%)
LogisticRegression	0.802285	0.805656	0.844000	-0.038343	-4.759285
KNeighborsClassifier	0.753420	0.812624	0.802228	0.010396	1.279259
DecisionTreeClassifier	0.850591	0.996430	0.818996	0.177434	17.806926
SVC	0.799256	0.856226	0.836876	0.019350	2.259904
RandomForestClassifier	0.665800	0.712551	0.780197	-0.067646	-9.493512
GaussianNB	0.810698	0.738874	0.802251	-0.063377	-8.577558
XGBClassifier	0.847233	0.864785	0.827637	0.037148	4.295622
AdaBoostClassifier	0.872463	0.879967	0.822993	0.056974	6.474576

1. Modelling

C. Model Evaluation

Metrics Evaluation yang digunakan adalah Precision dan Accuracy. Precision digunakan untuk meminimalkan False Positive karena pada kasus ini, kami perlu melakukan cost efficiency pada budget marketing. Setelah itu, kami juga mempertimbangkan nilai Accuracy sebagai parameter sekunder pada dua model yang menghasilkan model terbaik.

Berdasarkan hasil pemodelan dari ketiga model dibawah ini, model Adaboost menunjukkan nilai Precision dan Accuracy yang lebih tinggi dibandingkan dengan dua model lainnya. Namun, perbedaan nilai prediksi (diff) model Adaboost pada data training dan data test memiliki selisih yang lebih besar dibandingkan dengan dua model lainnya. Dengan demikian, model Adaboost memiliki performa yang lebih baik dalam mengklasifikasikan data dengan akurasi dan ketepatan yang lebih tinggi, tetapi memiliki perbedaan yang lebih besar dalam memprediksi data dibandingkan dengan dua model lainnya. Nilai diff tersebut akan diperkecil melalui hyperparameter tuning.

	CV Precision	Precision_Train	Precision_Test	Diff	Diff (%)
SVC	0.799378	0.856226	0.836876	0.019350	2.259904
XGBClassifier	0.847233	0.864785	0.827637	0.037148	4.295622
AdaBoostClassifier	0.872463	0.879967	0.822993	0.056974	6.474576

	CV Accuracy	Accuracy_Train	Accuracy_Test	Diff	Diff (%)
SVC	0.832502	0.852899	0.743119	0.109780	12.871371
XGBClassifier	0.840023	0.864710	0.807339	0.057371	6.634668
AdaBoostClassifier	0.861142	0.879742	0.816514	0.063229	7.187166

1. Modelling

D. Hyperparameter Tuning

Untuk menemukan nilai optimal untuk parameter yang digunakan pada model, akan dilakukan hyperparameter tuning. Dengan melakukan langkah ini, diharapkan dapat meningkatkan performa model dan menghasilkan model yang best-fit. Metode yang digunakan adalah Grid Search pada masing-masing model yang sudah dipilih.

```
# Define grid search
grid = dict(learning_rate=learning_rate, n_estimators=n_estimators, algorithm=algorithm)
grid_search = GridSearchCV(estimator=model, param_grid=grid, n_jobs=-1, cv=5, scoring='precision', error_score=0, return_
grid_result = grid_search.fit(X_train, y_train)
```

Metode Grid search dipilih karena Grid Search dapat mengeksplorasi seluruh kombinasi hyperparameter yang telah ditentukan dalam parameter yang didefinisikan sebelumnya. Ini memastikan bahwa kita tidak akan melewatkan setiap kemungkinan konfigurasi hyperparameter yang mungkin menghasilkan model yang optimal.

1. Modelling

E. Pemilihan model terbaik

Setelah melakukan hyperparameter tuning, berikut ini adalah hasil metrics evaluation dari SVM (CV Precision > 0.8 dan Diff < 10%). Model ini menghasilkan nilai Precision dan Accuracy lebih rendah dibanding model lainnya.

Precision SVM

Parameter Set	Model	CV Precision	Precision_Train	Precision_Test	Diff	Diff (%)	Parameters
param_16	SVC	0.874959	0.844516	0.807559	0.036957	4.376101	{'C': 3, 'gamma': 'auto', 'kernel': 'poly'}
param_13	SVC	0.874557	0.840083	0.810913	0.029170	3.472247	{'C': 3, 'gamma': 'scale', 'kernel': 'poly'}
param_10	SVC	0.873749	0.837537	0.814110	0.023427	2.797151	{'C': 2, 'gamma': 'auto', 'kernel': 'poly'}

Accuracy SVM

Parameter Set	Model	CV Accuracy	Accuracy_Train	Accuracy_Test	Diff	Diff (%)	Parameters
param_5	SVC	0.834291	0.854331	0.740826	0.113505	13.285841	{'C': 1, 'gamma': 'auto', 'kernel': 'rbf'}
param_2	SVC	0.832502	0.852899	0.743119	0.109780	12.871371	{'C': 1, 'gamma': 'scale', 'kernel': 'rbf'}
param_28	SVC	0.817103	0.849320	0.802752	0.046568	5.482937	{'C': 5, 'gamma': 'auto', 'kernel': 'poly'}

1. Modelling

E. Pemilihan model terbaik

Setelah melakukan hyperparameter tuning pada model XGBoost, berikut ini adalah hasil metrics evaluationnya (CV Precision > 0.8 dan Diff < 10%). Untuk model ini, kami melakukan perhitungan dengan max depth=1 untuk menghindari overfit.

Precision XGBoost

Parameter Set	Model	CV Precision	Precision_Train	Precision_Test	Diff	Diff (%)	Parameters
param_140	XGBClassifier	0.912467	0.922357	0.818055	0.104302	11.308190	{'colsample_bytree': 0.8, 'gamma': 1, 'learnin...
param_143	XGBClassifier	0.912467	0.922357	0.818055	0.104302	11.308190	{'colsample_bytree': 0.8, 'gamma': 1, 'learnin...
param_1147	XGBClassifier	0.912229	0.924031	0.835093	0.088938	9.625034	{'colsample_bytree': 1.0, 'gamma': 2, 'learnin...

Accuracy XGBoost

Parameter Set	Model	CV Accuracy	Accuracy_Train	Accuracy_Test	Diff	Diff (%)	Parameters
param_1003	XGBClassifier	0.879048	0.924123	0.850917	0.073206	7.921638	{'colsample_bytree': 1.0, 'gamma': 1, 'learnin...
param_1006	XGBClassifier	0.879048	0.924123	0.850917	0.073206	7.921638	{'colsample_bytree': 1.0, 'gamma': 1, 'learnin...
param_247	XGBClassifier	0.878325	0.918754	0.837156	0.081599	8.881427	{'colsample_bytree': 0.8, 'gamma': 2, 'learnin...

1. Modelling

E. Pemilihan model terbaik

Setelah melakukan hyperparameter tuning pada model Adaboost, berikut ini adalah hasil metrics evaluationnya (CV Precision > 0.8 dan Diff < 10%).

Precision Adaboost

Parameter Set	Model	CV Precision	Precision_Train	Precision_Test	Diff	Diff (%)	Parameters
param_56	AdaBoostClassifier	0.908271	0.918675	0.843316	0.075359	8.203003	{'algorithm': 'SAMME.R', 'learning_rate': 1.0,...
param_55	AdaBoostClassifier	0.903253	0.910859	0.831003	0.079856	8.767125	{'algorithm': 'SAMME.R', 'learning_rate': 1.0,...
param_52	AdaBoostClassifier	0.900182	0.912634	0.822355	0.090279	9.892101	{'algorithm': 'SAMME.R', 'learning_rate': 0.8,...

Accuracy Adaboost

Parameter Set	Model	CV Accuracy	Accuracy_Train	Accuracy_Test	Diff	Diff (%)	Parameters
param_55	AdaBoostClassifier	0.872244	0.910523	0.834862	0.075660	8.309532	{'algorithm': 'SAMME.R', 'learning_rate': 1.0,...
param_56	AdaBoostClassifier	0.871889	0.918397	0.850917	0.067479	7.347494	{'algorithm': 'SAMME.R', 'learning_rate': 1.0,...
param_52	AdaBoostClassifier	0.870808	0.912312	0.830275	0.082037	8.992193	{'algorithm': 'SAMME.R', 'learning_rate': 0.8,...

1. Modelling

E. Pemilihan model terbaik

Dapat dilihat pada gambar dibawah, bahwa model XGB memiliki tingkat presisi yang paling tinggi dibandingkan SVC dan AdaBoost. Namun, XGB memiliki Diff yang paling tinggi sehingga potensi overfit nya paling tinggi. Sehingga, kami memutuskan untuk memilih **Adaboost Classifier dengan param_56** karena model ini menunjukkan nilai Precision dan Accuracy yang relatif moderate (tidak paling rendah presisinya dan tidak paling tinggi Diff-nya).

Parameter Set	Model	CV Precision	Precision_Train	Precision_Test	Diff	Diff (%)	Parameters
param_16	SVC	0.874959	0.844516	0.807559	0.036957	4.376101	{'C': 3, 'gamma': 'auto', 'kernel': 'poly'}
param_13	SVC	0.874557	0.840083	0.810913	0.029170	3.472247	{'C': 3, 'gamma': 'scale', 'kernel': 'poly'}
param_10	SVC	0.873749	0.837537	0.814110	0.023427	2.797151	{'C': 2, 'gamma': 'auto', 'kernel': 'poly'}

Parameter Set	Model	CV Precision	Precision_Train	Precision_Test	Diff	Diff (%)	Parameters
param_140	XGBClassifier	0.912467	0.922357	0.818055	0.104302	11.308190	{'colsample_bytree': 0.8, 'gamma': 1, 'learnin...
param_143	XGBClassifier	0.912467	0.922357	0.818055	0.104302	11.308190	{'colsample_bytree': 0.8, 'gamma': 1, 'learnin...
param_1147	XGBClassifier	0.912229	0.924031	0.835093	0.088938	9.625034	{'colsample_bytree': 1.0, 'gamma': 2, 'learnin...

Parameter Set	Model	CV Precision	Precision_Train	Precision_Test	Diff	Diff (%)	Parameters
param_56	AdaBoostClassifier	0.908271	0.918675	0.843316	0.075359	8.203003	{'algorithm': 'SAMME.R', 'learning_rate': 1.0,...
param_55	AdaBoostClassifier	0.903253	0.910859	0.831003	0.079856	8.767125	{'algorithm': 'SAMME.R', 'learning_rate': 1.0,...
param_52	AdaBoostClassifier	0.900182	0.912634	0.822355	0.090279	9.892101	{'algorithm': 'SAMME.R', 'learning_rate': 0.8,...

1. Modelling

E. Pemilihan model terbaik

Setelah melakukan hyperparameter tuning pada model AdaBoost, didapatkan hasil presisi yang lebih baik dibandingkan sebelum hyperparameter tuning. Dapat dilihat bahwa terjadi kenaikan presisi sebesar 3.6% dari 87.2% menjadi 90.8%. Kenaikan tersebut diimbangi oleh suatu tradeoff yaitu peningkatan yang terjadi pada nilai selisih antara nilai presisi pada data train dan data test yaitu sekitar 1.7% dari 6.5% menjadi 8.2%. Selisih tersebut masih dapat ditoleransi karena masih dibawah 10% untuk dikategorikan sebagai best-fit.

Sebelum tuning

	CV Precision	Precision_Train	Precision_Test	Diff	Diff (%)
AdaBoostClassifier	0.872463	0.879967	0.822993	0.056974	6.474576

+1.7% Diff ↓ +3.6% Precision

Setelah tuning

Parameter Set	Model	CV Precision	Precision_Train	Precision_Test	Diff	Diff (%)
param_56	AdaBoostClassifier	0.908271	0.918675	0.843316	0.075359	8.203003

Berikut pertimbangan dalam pemilihan model ini:

1. Modelling

E. Pemilihan model terbaik

CV Precision yang Tinggi:

param_56 - CV Precision: 0.908271: Model ini memiliki nilai CV Precision yang tinggi selama proses cross validation. Ini menunjukkan bahwa model memiliki kemampuan yang baik untuk memberikan prediksi positif yang benar pada data yang belum pernah dilihat sebelumnya.

Perbedaan (Diff) yang Relatif Kecil:

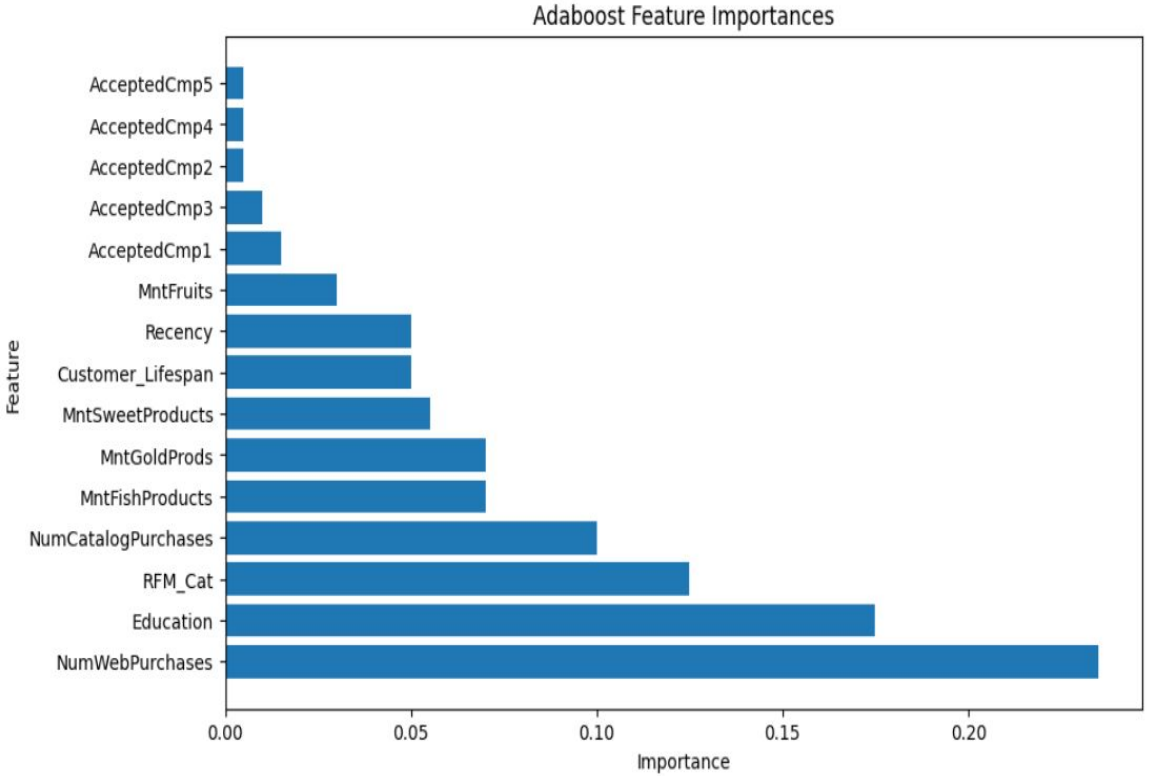
param_56 - Diff: 0.075359 (8.2%): Meskipun terdapat perbedaan antara Precision pada data train dan data test, perbedaannya relatif kecil (8.2%). Hal ini menunjukkan bahwa model ini cenderung tidak overfitting secara signifikan pada data train dan masih dapat menggeneralisasi dengan baik pada data test.

Precision_Train dan Precision_Test yang Tinggi:

Model ini juga memiliki nilai Precision yang tinggi pada data Train dan data test, menunjukkan kemampuan baik pada keduanya.Dengan mempertimbangkan kombinasi CV Precision yang tinggi, perbedaan yang relatif kecil antara data train dan test, serta nilai Precision yang tinggi pada kedua dataset.

2. Feature Importance

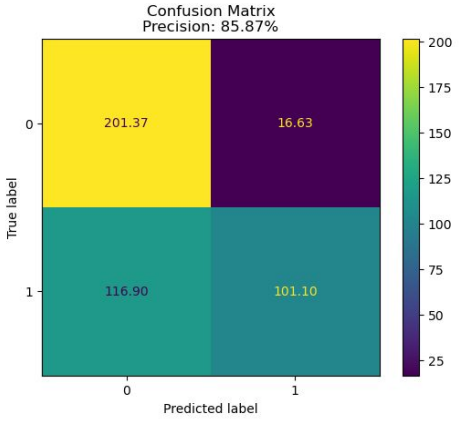
A. Grafik hasil feature importance dari model Adaboost Classifier

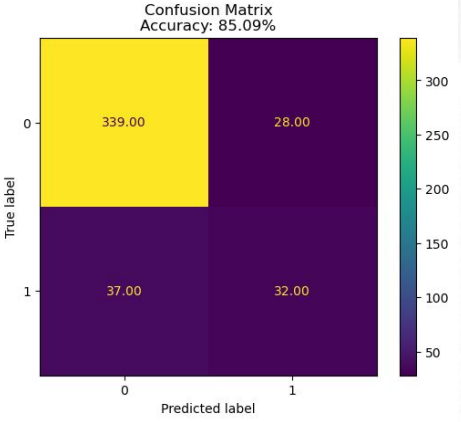


	Feature	Importance
12	NumWebPurchases	0.235
6	Education	0.175
13	RFM_Cat	0.125
11	NumCatalogPurchases	0.100
7	MntFishProducts	0.070
9	MntGoldProds	0.070
10	MntSweetProducts	0.055
5	Customer_Lifespan	0.050
14	Recency	0.050
8	MntFruits	0.030

2. Feature Importance

A. Confusion matrix dari model Adaboost Classifier





Classification Report on Test Data

	precision	recall	f1-score	support
0	0.90	0.92	0.91	367
1	0.53	0.46	0.50	69
accuracy			0.85	436
macro avg	0.72	0.69	0.70	436
weighted avg	0.84	0.85	0.85	436

Classification Report on Train Data

	precision	recall	f1-score	support
0	0.91	0.93	0.92	1397
1	0.93	0.91	0.92	1397
accuracy			0.92	2794
macro avg	0.92	0.92	0.92	2794
weighted avg	0.92	0.92	0.92	2794

Perhitungan confusion matrix pada precision menggunakan average weighted, karena jumlah label pada data test tidak seimbang.

2. Feature Importance

A. Prediksi Peningkatan Response Rate dan ROI Rate

```
total customer: 2240
total response: 334
rate accept: 14.91%
total cost: 2240
total revenue: 1670
total profit: -570
roi rate before pemodelan: -34.13%
```

```
total customer: 118
total response: 101
rate accept: 85.59%
total cost: 118
total revenue: 505
total profit: 387
roi rate after pemodelan: 76.63%
```

Setelah melalui proses pemodelan dan evaluasi, maka didapatkan hasil adanya peningkatan response rate dari 14.91% menjadi 85.59%.

Menambahkan asumsi untuk cost per customer/campaign adalah $1 dan revenue per customer/campaign adalah $5, maka didapatkan peningkatan roi rate dari -34.13% menjadi 76.63%.

2. Feature Importance

B. Business Insight & Recommendation

Dari hasil feature importance, berikut adalah insight yang didapat dan rekomendasi untuk meningkatkan efektivitas campaign dan memaksimalkan keuntungan bisnis:

a. **Number Purchase on Website**:
Business Insight: Penjualan melalui website merupakan faktor terpenting dalam marketing campaign.
Business Recommendation: Tingkatkan fokus pada pengoptimalan website, meningkatkan pengalaman pembelian online, gunakan data segmentasi pelanggan untuk menawarkan insentif atau penawaran khusus melalui platform web, dan perkuat strategi pemasaran online, untuk menawarkan insentif atau penawaran khusus sesuai dengan preferensi masing-masing kelompok pelanggan.

b. **Education**:
Business Insight: Tingkat pendidikan Phd adalah tingkat pendidikan dengan response rate tertinggi.
Business Recommendation: Pertimbangkan untuk mengadakan kolaborasi atau acara yang menghadirkan konten berkualitas tinggi, yang dapat menarik perhatian pelanggan Ph.D. dan memberikan nilai tambah dalam konteks pendidikan atau pengetahuan.

c. **RFM_Cat**:
Business Insight: Kategori RFM memainkan peran penting dalam respons pelanggan. Champions dan Loyal Customer mempunyai response rate tertinggi.
Business Recommendation: Tingkatkan layanan pelanggan untuk pelanggan dalam kategori Champions dan Loyal Customer, Kembangkan inovasi produk yang dapat memenuhi harapan tinggi dari pelanggan dalam kategori ini, Tawarkan produk atau layanan tambahan yang relevan dengan preferensi mereka.

2. Feature Importance

B. Business Insight & Recommendation

d. Number Purchase on Catalog:

Business Insight: Pembelian melalui katalog juga memiliki dampak yang signifikan

Business Recommendation: Pastikan bahwa produk yang ditampilkan di katalog sesuai dengan kebutuhan dan preferensi target pelanggan, sediakan penawaran eksklusif atau diskon khusus yang hanya dapat diakses melalui pembelian melalui katalog, dan pastikan bahwa inventaris katalog selalu diperbarui dan mencerminkan stok aktual.

e. Amount Spent on Fish, Gold, Sweet, dan Fruit in Last 2 Years:

Berdasarkan tingkat pendidikan customer, customer yang merespon campaign cenderung memiliki pengeluaran yang lebih besar baik pada produk ikan, buah, emas, dan sweets (permen dan coklat).

Business Recommendation: Buat penawaran bundle atau paket khusus dan berikan diskon bagi pelanggan yang membeli paket ini. Sediakan informasi yang mendalam tentang kualitas dan sumber produk.

f. Customer Lifespan & Recency:

Business Insight: Rata-rata Customer lifespan berdasarkan tingkat pendidikan maupun segmentasi pelanggan yang lebih tinggi (> 400 hari) dan rata-rata recency berdasarkan tingkat pendidikan yang lebih rendah (35 hari) cenderung merespon marketing campaign.

Business Recommendation: Fokuskan kampanye pemasaran pada pelanggan dengan recency rendah, perbarui program loyalitas untuk pelanggan dengan customer lifespan tinggi, sertakan kampanye edukasi yang memberikan nilai tambah bagi pelanggan dengan tingkat pendidikan tinggi, tawarkan konten yang berfokus pada kecerdasan kepada pelanggan dengan tingkat pendidikan tinggi.