Rakamin Academy

Homework

Machine Learning Evaluation & Supervised Learning
Stage 3 - Final Project

By: ec-Team (Kelompok 9)

1. Modeling

Split Data Train & Test

Sebelum dilakukan modeling perlu dilakukan split data yang sudah melalui preprocessing (df_final) dengan stratified sampling menjadi 80% data train dan 20% data test agar tidak terjadi leakage.

Modeling (Algoritma yang diimplementasikan)

Berikut model klasifikasi yang digunakan untuk dataset yang memiliki banyak fitur dan kompleksitas yang cukup tinggi:

- Logistic Regression
- Decision Tree
- Random Forest
- kNN
- SVC
- Ada Boost
- Gradient BoostingClassifier
- XGBoost

1. Modeling

Model Evaluation

A. Pemilihan metrics

Metrics yang cocok digunakan untuk modeling dari dataset ini adalah ROC-AUC score metriks tersebut digunakan untuk menghitung / memprediksi class yang minoritas (True Revenue).

B. Perhitungan metrics model

- Berdasarkan ROC AUC score, Gradient Boosting merupakan model yang lebih baik dibandingkan model lainnya karena dapat dikatakan best fit. ROC-AUC score pada data train (0.948592) tidak berbeda jauh dengan data train (0.936877).
- Decision Tree dan kNN masih overfitting

1. Modeling

Model Evaluation: Apakah model sudah best-fit? Hindari Overfit/Underfit. Validasi dengan cross-validation

Cross validation perlu dilakukan untuk mendapatkan model terbaik yang lebih akurat dan cenderung tidak overfit. Cross validation yang dilakukan menggunakan 5 fold, yang umum digunakan karena ukuran foldnya dikatakan optimal.

Berdasarkan hasil ROC AUC score pada cross validation, model dengan Gradient Boosting, Random Forest, dan XG Boost memiliki score CV ROC AUC tertinggi dibandingkan model yang lain dengan score masing-masing model 0.930925, 0.930825, dan 0.919733.

Ketiga model tersebut masih belum bisa dikatakan best fit karena gap antara train dan test masih cukup besar, sehingga perlu dilanjutkan ke proses hyperparameter tuning untuk mencari kombinasi nilai terbaik dari hyperparameter pada sebuah model machine learning sehingga performa model dapat ditingkatkan.

1. Modeling

Hyperparameter Tuning

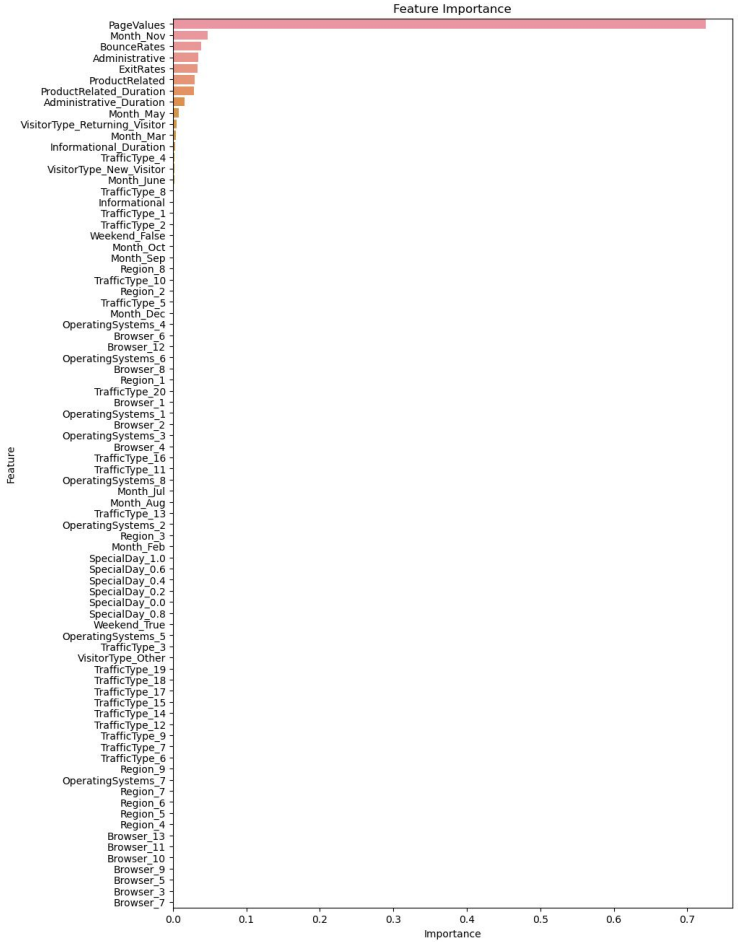
Pada Hyperparameter Tuning hanya tiga model terbaik yang di-tuning, yaitu Gradient Boosting Classifier, Random Forest Classifier, dan XGB Classifier.

Hasil Hyperparameter Tuning

- Setelah dilakukan hyperparameter tuning, seluruh tree-based model tidak ada yang overfitted.
- Jika dibandingkan dengan sebelum tuning, performa model setelah tuning lebih baik

Oleh karena itu, akan digunakan model setelah tuning untuk feature importance. Model yang dipilih adalah Gradient Boosting Classifier karena dengan dibanding model lainnya nilai ROC-AUC tesnya paling tinggi.

2. Feature Importance



Pengamatan

5 Feature dengan imporatance score tertinggi, yaitu
-PageValues
-Month_Nov
-BounceRates
-Administrative
-ExitRates

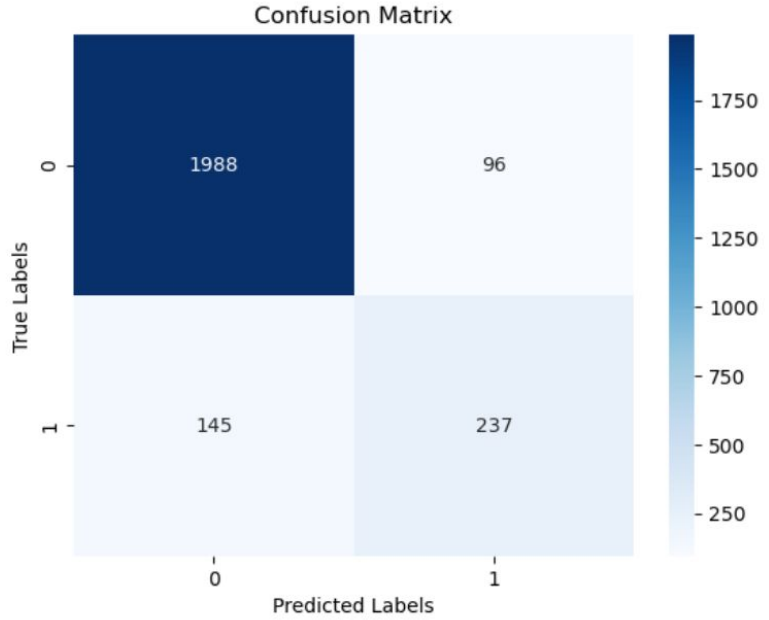
page_values sangat berpengaruh ke model dan dengan nilai yang sangat tinggi(>0.7) jika dibandingkan feature-feature lain.

2. Feature Importance

Rekomendasi Aksi

- Page_values adalah hasil dari nilai rata-rata untuk halaman yang dikunjungi pengguna sebelum menyelesaikan konveksi atau transaksi eCommerce. Berdasarkan feature importance, dapat dilihat page values memiliki pengaruh yang signifikan terhadap konversi. Sesi dengan page values yang tinggi cenderung menghasilkan revenue, sehingga untuk menghasilkan peningkatan pada purchase/conversion rate, kita juga perlu meningkatkan revenue
-

Confusion Matrix

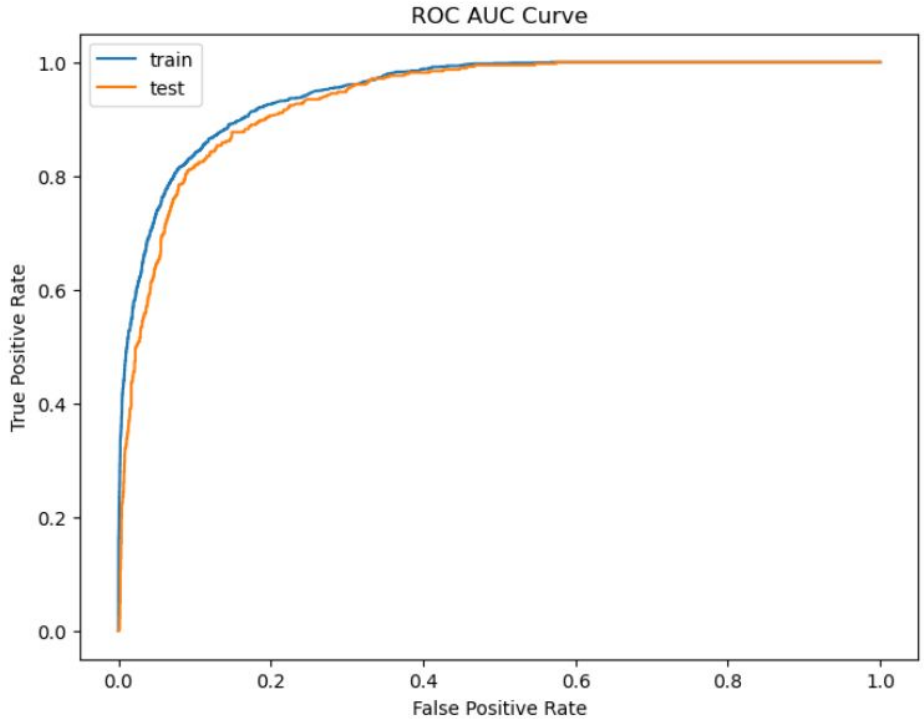


Classification Report:

	precision	recall	f1-score	support
0	0.93	0.95	0.94	2084
1	0.71	0.62	0.66	382
accuracy			0.90	2466
macro avg	0.82	0.79	0.80	2466
weighted avg	0.90	0.90	0.90	2466

- False Negative (FN) terjadi ketika model memprediksi bahwa seseorang tidak akan membeli (negative), padahal kenyataannya orang tersebut membeli (positive).
- Dalam kasus prediksi apakah seseorang akan membeli atau tidak, FN sangat berbahaya karena dapat menyebabkan perusahaan kehilangan pelanggan potensial.
- Dalam kata lain, FN dapat menyebabkan perusahaan melewatkan peluang untuk menjual produknya pada orang yang sebenarnya berminat.
- Oleh karena itu, FN dapat dikurangi dengan meningkatkan recall sehingga pelanggan yang sebenarnya akan membeli produk dapat terdeteksi dengan lebih baik.

ROC AUC Curve




```
Optimal Threshold: 0.1227
Optimal Precision: 0.2251
Optimal Recall: 1.0000
Optimal F1-Score: 0.3675
```

Link Git ec-Team:

https://github.com/EC-Teams/Final-Project-Online-Shopping-Intention