

Shipment Arrival Prediction

Supervised
Homework



Pembagian Tugas

Stage 3:

- Nur Cahyanti (Modelling and Evaluation)
- Utlia Rahma (Modelling and Evaluation)
- Indra Laksana (Feature Importance)
- Handika (Modelling and Evaluation)
- Fajar Nurdiono (Modelling and Evaluation)
- Refanie Fajrina (Feature Importance)

1. Modelling

1A. Split Training & Testing Data

```
X = df.drop(columns=['Reached.on.Time_Y.N', 'Unnamed: 0', 'prior_purchase_std', 'product_cost_norm', 'discount_norm', 'weight_norm'])  
y = df[['Reached.on.Time_Y.N']]  
  
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42, stratify=y)
```

Split data training dan testing data menggunakan rasio 70 : 30 dari jumlah 10,096 data, dengan distribusi sebagai berikut:

- data training sejumlah 7,067 data
- data testing sejumlah 3,029 data

1B. Modelling

Algoritma yang kami pakai dalam tugas ini adalah:

1. Logistic Regression
2. K-Nearest Neighbor
3. Decision Tree
4. Random Forest
5. Gradient Boosting
6. Gaussian Naive-Bayes
7. AdaBoost
8. XGBoost

1C. Pemilihan Model dan Perhitungan Metrik

Kami melakukan fitting model dengan data training dan testing dengan hyperparameter bawaan package scikit-learn.

Metrik yang akan kami cari nilainya adalah, di antara lain:

- Score
- Precision (Test)
- Recall (Test)
- AUC (test-proba)
- AUC (train-proba)

1C. Pemilihan Model dan Perhitungan Metrik

(Using Quantile Transformation)

Logistic Regression

```
Score: 0.65
Accuracy (Test Set): 0.65
Precision (Test Set): 0.6952
Recall (Test Set): 0.747
F1-Score (Test Set): 0.7202
roc-auc (test-proba): 0.7241
roc-auc (train-proba): 0.7302
```

K-Nearest Neighbors

```
Score: 0.6395
Accuracy (Test Set): 0.6395
Precision (Test Set): 0.7139
Recall (Test Set): 0.6709
F1-Score (Test Set): 0.6917
roc-auc (test-proba): 0.6969
roc-auc (train-proba): 0.8579
```

Decision Tree

```
Score: 0.6458
Accuracy (Test Set): 0.6458
Precision (Test Set): 0.7072
Recall (Test Set): 0.7037
F1-Score (Test Set): 0.7055
roc-auc (test-proba): 0.6307
roc-auc (train-proba): 1.0
```

Random Forest

```
Score: 0.654
Accuracy (Test Set): 0.654
Precision (Test Set): 0.7529
Recall (Test Set): 0.6342
F1-Score (Test Set): 0.6885
roc-auc (test-proba): 0.7359
roc-auc (train-proba): 1.0
```

Gradient Boosting

```
Score: 0.6758
Accuracy (Test Set): 0.6758
Precision (Test Set): 0.8505
Recall (Test Set): 0.5608
F1-Score (Test Set): 0.6759
roc-auc (test-proba): 0.7326
roc-auc (train-proba): 0.8406
```

Gaussian Naive Bayes

```
Score: 0.659
Accuracy (Test Set): 0.659
Precision (Test Set): 0.7638
Recall (Test Set): 0.6287
F1-Score (Test Set): 0.6897
roc-auc (test-proba): 0.7337
roc-auc (train-proba): 0.7416
```

AdaBoost

```
Score: 0.6741
Accuracy (Test Set): 0.6741
Precision (Test Set): 0.8301
Recall (Test Set): 0.5778
F1-Score (Test Set): 0.6813
roc-auc (test-proba): 0.7377
roc-auc (train-proba): 0.7907
```

XGboost

```
Score: 0.6434
Accuracy (Test Set): 0.6434
Precision (Test Set): 0.7266
Recall (Test Set): 0.655
F1-Score (Test Set): 0.6889
roc-auc (test-proba): 0.7321
roc-auc (train-proba): 0.9832
```


1C. Pemilihan Model dan Perhitungan Metrik

(Using Log Transformation)

Logistic Regression

```
Score: 0.6474
Accuracy (Test Set): 0.6474
Precision (Test Set): 0.7007
Recall (Test Set): 0.7245
F1-Score (Test Set): 0.7124
roc-auc (test-proba): 0.7306
roc-auc (train-proba): 0.7355
```

XGboost

```
Score: 0.6428
Accuracy (Test Set): 0.6428
Precision (Test Set): 0.7257
Recall (Test Set): 0.655
F1-Score (Test Set): 0.6885
roc-auc (test-proba): 0.7318
roc-auc (train-proba): 0.9832
```

K-Nearest Neighbors

```
Score: 0.6144
Accuracy (Test Set): 0.6144
Precision (Test Set): 0.6836
Recall (Test Set): 0.6709
F1-Score (Test Set): 0.6772
roc-auc (test-proba): 0.6533
roc-auc (train-proba): 0.8347
```

Decision Tree

```
Score: 0.651
Accuracy (Test Set): 0.651
Precision (Test Set): 0.7082
Recall (Test Set): 0.7163
F1-Score (Test Set): 0.7122
roc-auc (test-proba): 0.6341
roc-auc (train-proba): 1.0
```

Random Forest

```
Score: 0.6576
Accuracy (Test Set): 0.6576
Precision (Test Set): 0.753
Recall (Test Set): 0.6429
F1-Score (Test Set): 0.6936
roc-auc (test-proba): 0.7361
roc-auc (train-proba): 1.0
```

Gradient Boosting

```
Score: 0.6761
Accuracy (Test Set): 0.6761
Precision (Test Set): 0.8512
Recall (Test Set): 0.5608
F1-Score (Test Set): 0.6761
roc-auc (test-proba): 0.7324
roc-auc (train-proba): 0.8406
```

Gaussian Naive Bayes

```
Score: 0.6705
Accuracy (Test Set): 0.6705
Precision (Test Set): 0.8062
Recall (Test Set): 0.5969
F1-Score (Test Set): 0.686
roc-auc (test-proba): 0.7351
roc-auc (train-proba): 0.7437
```

AdaBoost

```
Score: 0.6741
Accuracy (Test Set): 0.6741
Precision (Test Set): 0.8301
Recall (Test Set): 0.5778
F1-Score (Test Set): 0.6813
roc-auc (test-proba): 0.7377
roc-auc (train-proba): 0.7907
```


1D. Apakah model sudah best-fit?

Dari model-model yang telah kami coba, belum ada model yang memiliki semua dari:

1. Nilai AUC-train dan AUC-test yang berbeda tipis,
2. Nilai AUC-train < 1.00 , dan
3. Nilai Recall yang besar.

Dengan kata lain, model-model yang telah kami coba masih overfitting.

1E. Hyperparameter Tuning (Using Quantile Transformation)

Logistic Regression

```
Score: 0.7247
Accuracy (Test Set): 0.6425
Precision (Test Set): 0.6881
Recall (Test Set): 0.7442
F1-Score (Test Set): 0.7151
roc-auc (test-proba): 0.7247
roc-auc (train-proba): 0.7301

Best penalty : 12
Best C : 0.03
Best solver : liblinear
```

K-Nearest Neighbors tuned

```
Score: 0.7191
Accuracy (Test Set): 0.6431
Precision (Test Set): 0.7365
Recall (Test Set): 0.6353
F1-Score (Test Set): 0.6822
roc-auc (test-proba): 0.7191
roc-auc (train-proba): 0.7927
```

Decision Tree tuned

```
Score: 0.7383
Accuracy (Test Set): 0.65
Precision (Test Set): 0.7415
Recall (Test Set): 0.644
F1-Score (Test Set): 0.6893
roc-auc (test-proba): 0.7383
roc-auc (train-proba): 0.829
```

Random Forest tuned

```
Score: 0.7409
Accuracy (Test Set): 0.6728
Precision (Test Set): 0.818
Recall (Test Set): 0.5882
F1-Score (Test Set): 0.6843
roc-auc (test-proba): 0.7409
roc-auc (train-proba): 0.8393
```

Gradient Boosting tuned

```
Score: 0.7369
Accuracy (Test Set): 0.6028
Precision (Test Set): 0.6028
Recall (Test Set): 1.0
F1-Score (Test Set): 0.7522
roc-auc (test-proba): 0.7369
roc-auc (train-proba): 0.7468
```

Fitting 5 folds for each of 10 candidates, totalling 50 fits
Gaussian Naive Bayes tuned

```
Score: 0.7337
Accuracy (Test Set): 0.6586
Precision (Test Set): 0.7636
Recall (Test Set): 0.6281
F1-Score (Test Set): 0.6893
roc-auc (test-proba): 0.7337
roc-auc (train-proba): 0.7416
```

XGboost tuned

```
Score: 0.7318
Accuracy (Test Set): 0.6471
Precision (Test Set): 0.7241
Recall (Test Set): 0.6698
F1-Score (Test Set): 0.6959
roc-auc (test-proba): 0.7318
roc-auc (train-proba): 0.9572
```

1E. Hyperparameter Tuning (Using Log Transformation)

Logistic Regression

```
Score: 0.73
Accuracy (Test Set): 0.6467
Precision (Test Set): 0.6933
Recall (Test Set): 0.7426
F1-Score (Test Set): 0.7171
roc-auc (test-proba): 0.73
roc-auc (train-proba): 0.7373

Best penalty : 12
Best C : 0.1
Best solver : liblinear
```

K-Nearest Neighbors tuned

```
Score: 0.6949
Accuracy (Test Set): 0.621
Precision (Test Set): 0.6747
Recall (Test Set): 0.7169
F1-Score (Test Set): 0.6952
roc-auc (test-proba): 0.6949
roc-auc (train-proba): 0.766
```

Decision Tree tuned

```
Score: 0.7271
Accuracy (Test Set): 0.6464
Precision (Test Set): 0.7289
Recall (Test Set): 0.6583
F1-Score (Test Set): 0.6918
roc-auc (test-proba): 0.7271
roc-auc (train-proba): 0.8396
```

Random Forest tuned

```
Score: 0.7269
Accuracy (Test Set): 0.6633
Precision (Test Set): 0.8292
Recall (Test Set): 0.5559
F1-Score (Test Set): 0.6656
roc-auc (test-proba): 0.7269
roc-auc (train-proba): 0.8564
```

Gradient Boosting tuned

```
Score: 0.7362
Accuracy (Test Set): 0.6718
Precision (Test Set): 0.8925
Recall (Test Set): 0.5181
F1-Score (Test Set): 0.6556
roc-auc (test-proba): 0.7362
roc-auc (train-proba): 0.7638
```

Fitting 5 folds for each of 10 candidates, totalling 50 fits
Gaussian Naive Bayes tuned

```
Score: 0.7352
Accuracy (Test Set): 0.6705
Precision (Test Set): 0.8062
Recall (Test Set): 0.5969
F1-Score (Test Set): 0.686
roc-auc (test-proba): 0.7352
roc-auc (train-proba): 0.7437
```

XGboost tuned

```
Score: 0.7423
Accuracy (Test Set): 0.6715
Precision (Test Set): 0.8924
Recall (Test Set): 0.5175
F1-Score (Test Set): 0.6551
roc-auc (test-proba): 0.7423
roc-auc (train-proba): 0.7571
```


1E. Evaluasi Model Setelah Tuning (Kesimpulan)

1. Quantile Transformation

- Sebelum tuning : Model terbaik pada logistik regression

```
Score: 0.65
Accuracy (Test Set): 0.65
Precision (Test Set): 0.6952
Recall (Test Set): 0.747
F1-Score (Test Set): 0.7202
roc-auc (test-proba): 0.7241
roc-auc (train-proba): 0.7302
```

- Setelah tuning : Model terbaik pada logistik regression

```
Score: 0.7247
Accuracy (Test Set): 0.6425
Precision (Test Set): 0.6881
Recall (Test Set): 0.7442
F1-Score (Test Set): 0.7151
roc-auc (test-proba): 0.7247
roc-auc (train-proba): 0.7301

Best penalty : 12
Best C : 0.03
Best solver : liblinear
```

2. Log Transformation

- Sebelum tuning : Model terbaik pada logistik regression

```
Score: 0.6474
Accuracy (Test Set): 0.6474
Precision (Test Set): 0.7007
Recall (Test Set): 0.7245
F1-Score (Test Set): 0.7124
roc-auc (test-proba): 0.7306
roc-auc (train-proba): 0.7355
```

- Setelah tuning : Model terbaik pada logistik regression

```
Score: 0.73
Accuracy (Test Set): 0.6467
Precision (Test Set): 0.6933
Recall (Test Set): 0.7426
F1-Score (Test Set): 0.7171
roc-auc (test-proba): 0.73
roc-auc (train-proba): 0.7373

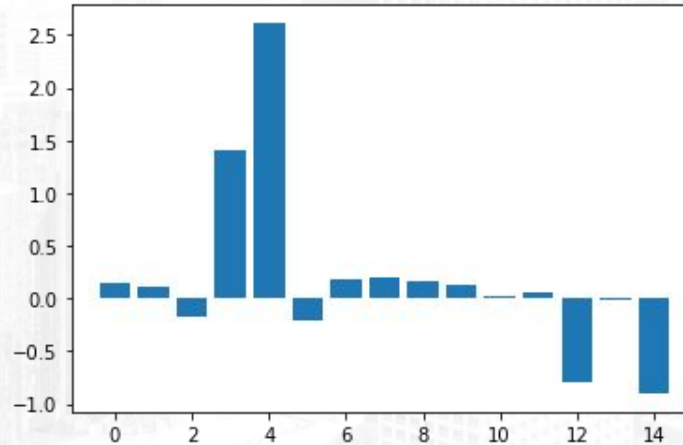
Best penalty : 12
Best C : 0.1
Best solver : liblinear
```

Kesimpulan : Model terbaik saat ini menggunakan transformasi logaritmik, dengan persentase roc-auc pada data test dan roc-auc pada data training sebesar 73% dan 73.73% serta persentase recall sebesar 74.26%.

2. Feature Importance

2A. Feature Importance

	Feature	Score
0	Product_importance	0.138431
1	Gender	0.103564
2	prior_purchase_std	-0.180974
3	product_cost_norm	1.409413
4	discount_norm	2.610242
5	weight_norm	-0.210436
6	Warehouse_block_B	0.172538
7	Warehouse_block_C	0.194155
8	Warehouse_block_D	0.158731
9	Warehouse_block_F	0.134506
10	Mode_of_Shipment_Road	0.017114
11	Mode_of_Shipment_Ship	0.056444
12	Cost_Per_Weight	-0.788251
13	Cost_After_Disc	-0.010331
14	Weight_level	-0.900502



Feature discount_norm (discount_offered yang telah dinormalisasi) merupakan fitur yang paling berpengaruh pada permodelan machine learning dengan nilai > 2,61.

Business Insight yang dapat digali:

1. Semakin tinggi diskon yang ditawarkan ternyata semakin tinggi juga potensi keterlambatan, asumsinya adalah perusahaan mengabaikan service kepada customer karena sudah diberikan diskon.
2. Selain discount_norm, model lainnya adalah cost_per_weight. Dalam case ini semakin tinggi cost_per_weight (rate) suatu barang, juga meningkatkan potensi keterlambatan, asumsinya adalah dalam bisnis logistik, salah satu hal yang mempengaruhi rate adalah jarak pengiriman, sehingga semakin panjang kota tujuannya maka potensi keterlambatan juga semakin tinggi

2B. Feature Selection

Berdasarkan tabel Future Importance pada slide sebelumnya, kita dapat memakai kolom 'discount_norm', 'product_cost_norm', 'Weight_level', dan 'Cost_Per_Weight' saja untuk iterasi model.

```
Score: 0.6349  
Accuracy (Test Set): 0.6349  
Precision (Test Set): 0.676  
Recall (Test Set): 0.7574  
F1-Score (Test Set): 0.7144  
roc-auc (test-proba): 0.7294  
roc-auc (train-proba): 0.7322
```

Setelah melakukan iterasi model sesuai dengan best parameters, nilai AUC-test turun dari 73% menjadi 72.94%, AUC-train turun dari 73.73% menjadi 73.22%, dan nilai Recall naik dari 74.26% ke 75.74%.

Walaupun Feature Selection ini hanya menggunakan 4 fitur, penurunan nilai kedua score AUC tidak terlalu signifikan dan peningkatan nilai Recall cukup besar.