Rakamin Academy

E-Commerce Shipping Data

"Asklepios"

Awalsyah Rinanto Putra

Fathah Oscar

M Rizky Septiansyah

Hermawan Febrianto

Devi Puji Ayuningsih

Anggita Citanegara Lubis

Stage 3 (Supervised Learning)

1. Model Evaluation

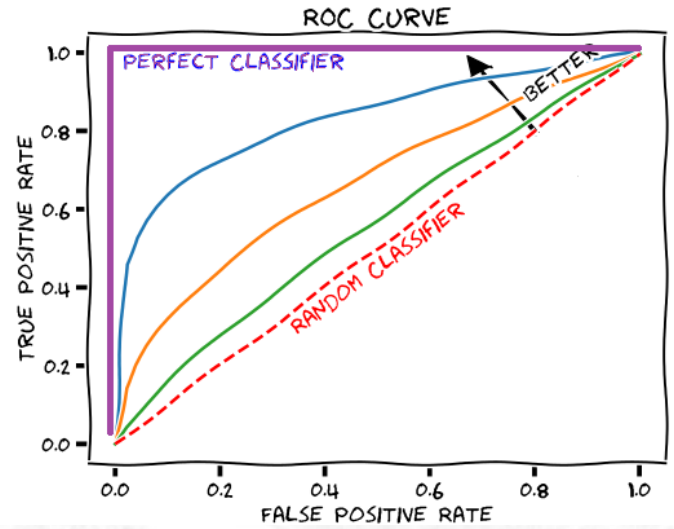
Algoritma yang digunakan :

- Logistic Regression
- K-Nearest Neighbor
- Decision Tree
- Random Forest
- AdaBoost
- XGBoost

Skor evaluasi yang dihitung :

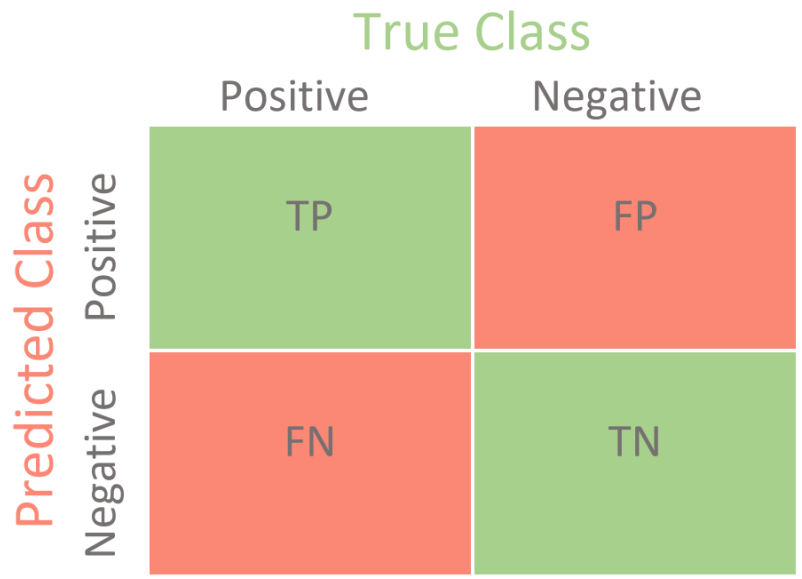
- Accuracy
- Precision
- Recall
- F1-Score
- ROC-AUC

Matrix Evaluasi yang digunakan: ROC-AUC



ROC-AUC memiliki sifat yang robust terhadap dataset yang imbalance pada target, sehingga ROC-AUC cocok digunakan pada dataset kami yang sedikit imbalanced pada target (59% sample late, 41% sample on time).

Matrix yang digunakan adalah **ROC-AUC**, karena dalam studi kasus ini penting untuk meminimalisir false positif dan false negative.



False Positive: keadaan dimana model memprediksi pengiriman terlambat (Late), namun kenyataannya datang on time.

Impact: **failed to meet customer's expectation.**

Tidak sefatal false negative, namun sebisa mungkin diminimalisir agar experience customer menggunakan layanan kita tetap baik.

False Negative: keadaan dimana model memprediksi pengiriman ontime, namun kenyataannya datang terlambat (Late).

Impact: **failed to meet customer's expectation.**

Hal ini akan membuat penilaian customer terhadap experience sangat buruk karena sudah berekspektasi produk dating on time namun datang terlambat.

ROC-AUC						
Dataset 1 (Removing Outlier Z-Score)						
Method	Logreg	kNN	Decision Tree	Random Forest	AdaBoost	XGBoost
Train	0.71	0.76	0.79	0.75	0.75	0.79
Test	0.72	0.72	0.72	0.73	0.74	0.73

Dari hasil perhitungan ROC-AUC score 2 dataset, yaitu dataset 1 yang dilakukan remove outlier berdasarkan Z-Score dan dataset 2 yang dilakukan remove oulier berdasarkan Z-Score dan IQR, didapatkan pemodelan boosting dengan **algoritma Adaboost pada dataset 1 menghasilkan model dengan performa terbaik.**

Hal ini dapat diliat pada ROC-AUC score pada metode algoritma Adaboost relatif lebih tinggi daripada ROC-AUC score pada metode yang lain, serta memiliki skor data training lebih besar daripada data test dengan gap yang tidak terlalu besar.

ROC-AUC						
Dataset 1 (Removing Outlier Z-Score)						
Method	Logreg	kNN	Decision Tree	Random Forest	AdaBoost	XGBoost
Train	0.71	0.76	0.79	0.75	0.75	0.79
Test	0.72	0.72	0.72	0.73	0.74	0.73

ROC-AUC						
Dataset 2 (Removing Outlier Z-Score & IQR)						
Method	Logreg	kNN	Decision Tree	Random Forest	AdaBoost	XGBoost
Train	0.58	0.58	0.61	0.62	0.63	0.88
Test	0.58	0.64	0.57	0.60	0.61	0.60

Secara umum, skor AUC pada dataset 1 terlihat jauh lebih bagus daripada dataset 2, hal ini terjadi karena banyaknya baris data yang dihapus pada dataset 2 dengan metode remove outlier dengan IQR mempunyai pengaruh yang cukup besar terhadap performa model.

Hyperparameter Terbaik (ROC-AUC Dataset 1)					
Logistic Regression	**kNN**	**Decision Tree**	**Random Forest**	**AdaBoost**	**XGBoost**
Penalty = 12	Algorithm = auto	Criterion' = 'gini'	Max_depth = 5	n_estimator = 225	max_depth = 90
C = 0.028	Leaf_size': 30	Max_depth': 79	Criterion = 'entropy	learning_rate = 0.08358	min_child_weight = 6
Solver = lbfgs	Metric = minkowski	Max_features = 'sqrt'	n_estimators = 2	algorithm = SAMMER. R	gamma = 0.4
Intercept scaling = 1	n_neeighbors = 83	'min_samples_leaf': 83	Min_samples_split = 6		tree_method = his’
Multi_class = auto	p = 1	'min_samples_split': 100	Min_samples_leaf = 2		eta = 0.1313
tol = 0.0001	weights = uniform	'splitter': 'best'	Max_features = sqrt		Lambda = 0
					alpha = 0.3

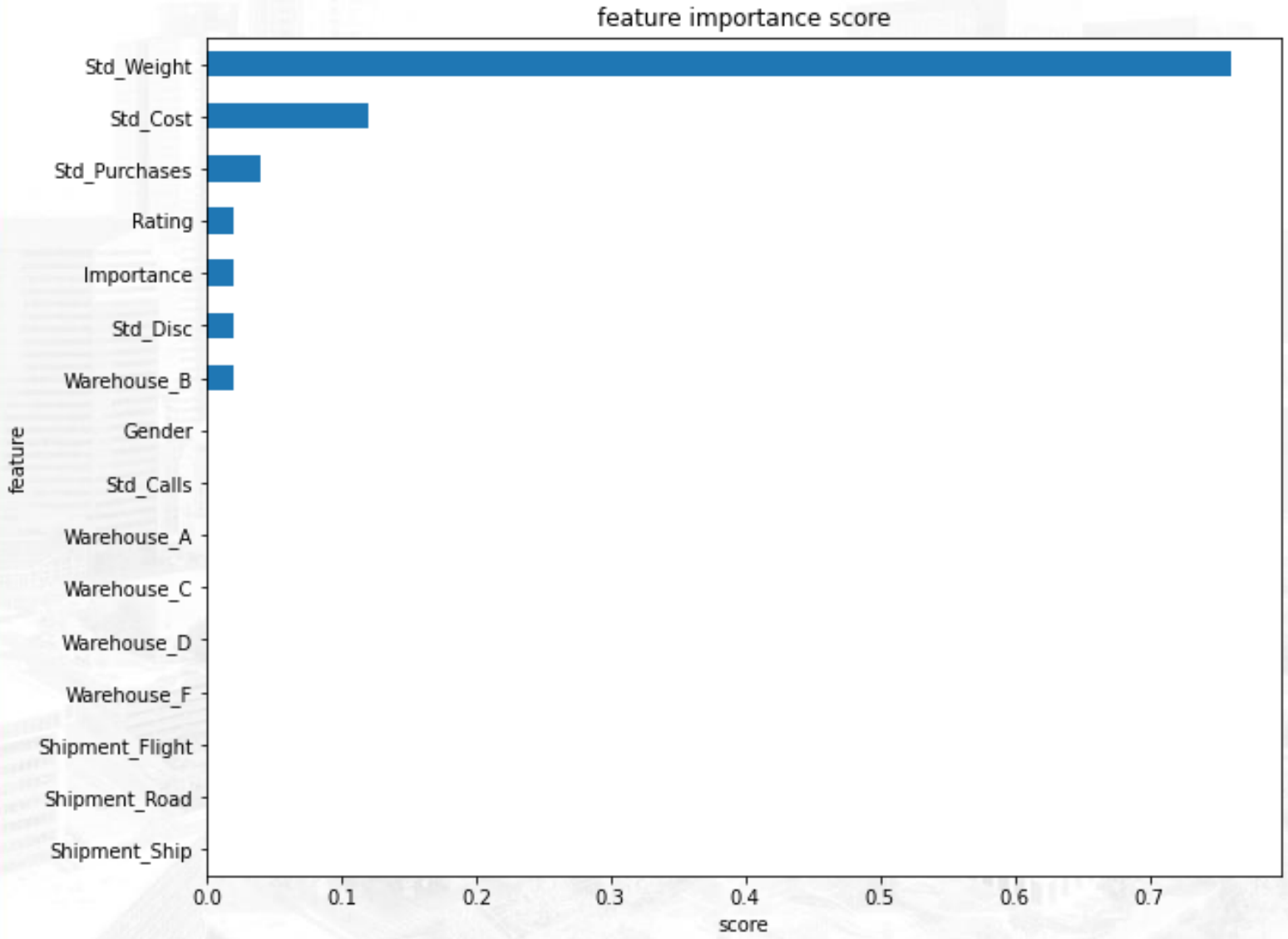
Tuning hyperparameter dilakukan dengan list parameter diatas agar menemukan ROC-AUC score sebaik mungkin yaitu, ROC-AUC score pada data train lebih besar daripada data test namun dengan gap yang tidak terlalu besar.

Confussion Matrix	Predicted Label	
Actual Label	**True Positive** **1175** **(36.68 %)**	**False Negative** **697** **(21.82%)**
	False Positive **363** **(11.36 %)**	**True Negative** **958** **(30.00%)**

Positive = Late

Negative = On Time

2. Feature Importance



Feature yang paling penting :

Terdapat 3 Feature yang paling berpengaruh terhadap target berdasarkan grafik di samping, yaitu :

- Std_Weight
- Std_Cost
- Std_Purchases

Business Insight :

- Untuk Feature dengan Importance tertinggi adalah Std_Weight. Jika kita mundur ke stage 2 yaitu Data Pre-Processing, terlihat di Grafik Heatmap bahwa Std_Weight memiliki korelasi negative dengan target, yaitu Late. Jadi dapat disimpulkan bahwa semakin ringan berat suatu barang, semakin berpotensi barang tersebut mengalami keterlambatan pengiriman.
- Untuk Feature dengan Importance tertinggi kedua adalah Std_Cost. Terlihat di Grafik Heatmap bahwa Std_Cost memiliki korelasi negative dengan target. Jadi dapat disimpulkan bahwa semakin rendahnya cost pembelanjaan barang, semakin berpotensi barang akan mengalami keterlambatan pengiriman.
- Untuk Feature dengan Importance terakhir adalah Std_Purchases. Terlihat di Grafik Heatmap bahwa Purchases memiliki korelasi negative dengan target. Jadi dapat disimpulkan bahwa semakin rendah nilai Purchase, semakin berpotensi barang akan mengalami keterlambatan pengiriman.

Jadi, dari Ketiga Feature Importance yang sudah dijelaskan di atas, dapat disimpulkan untuk Rootcause Problemnya adalah dari sisi Traffic yang Tinggi dan manajemen pengiriman yang kurang baik. Merujuk dari 3 Feature di atas, keterlambatan barang lebih tertuju kepada barang yang ringan, murah, dan memiliki nilai purchase rendah.

Action Item :

- Untuk menjaga Customer Satisfaction, sebaiknya ada sistem baru yang dibuat Perusahaan, yaitu sistem notifikasi dan estimasi waktu pengiriman barang. Sistem Notifikasi, ketika barang berpotensi mengalami keterlambatan, customer akan mendapatkan pemberitahuan bahwa pengiriman akan mengalami keterlambatan. Kemudian untuk Sistem Estimasi Pengiriman Barang, customer dapat memilih waktu pengiriman yang sesuai dengan jenis shipment yang dipilih. Contoh customer memilih pengiriman menggunakan kapal maka estimasi pengiriman yang ter-*create* oleh sistem yaitu pengiriman membutuhkan waktu 5 hari.
- Karena Traffic dan Frekuensi Pengiriman yang tinggi, Maka harus menjaga Manajemen Pengiriman yang baik. Perusahaan harus membuat atau memperbaiki SOP dari Pengiriman. Seperti Pengelolaan dan Packing pengiriman barang yang lebih ketat (Misalnya memperhatikan antrian barang berdasarkan waktu pembelian, agar pengiriman barang sesuai dengan urutan antrian), Pengawasan atau Monitoring Pengiriman secara Realtime, Peningkatan Sumberdaya Pengiriman baik dari Armada ataupun Manusia.
- Karena Traffic dan Frekuensi Pengiriman yang tinggi juga, Maka harus dilakukan Manajemen Armada yang baik, jika dilihat dari Dataset, pengiriman paling banyak menggunakan kapal. Bisa dibuat atau ditambahkan SOP untuk itu, misalnya load pengiriman harus sama rata antara satu kapal dan kapal lainnya (Load Balancing).