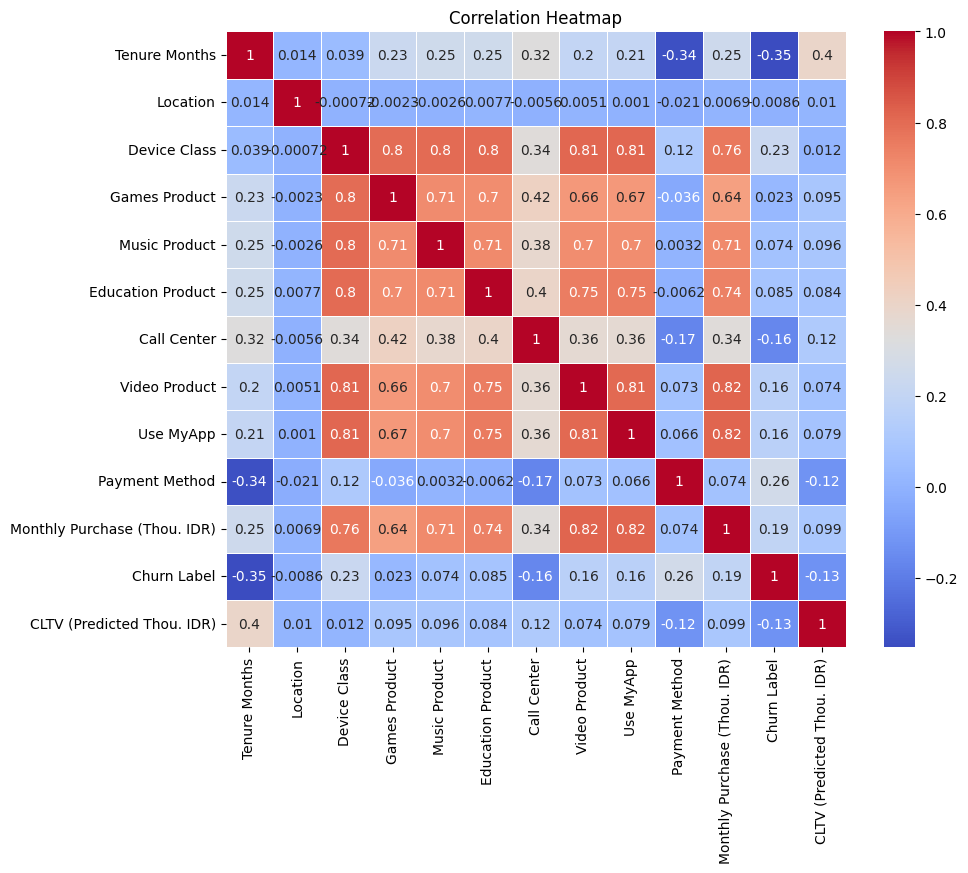
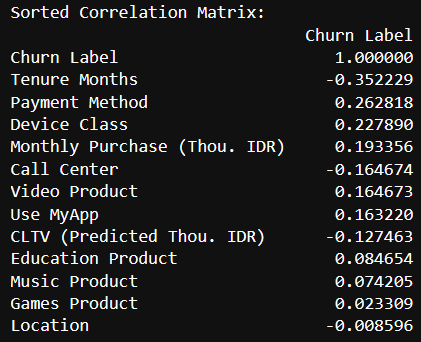
* Hasil correlation untuk raw dataset



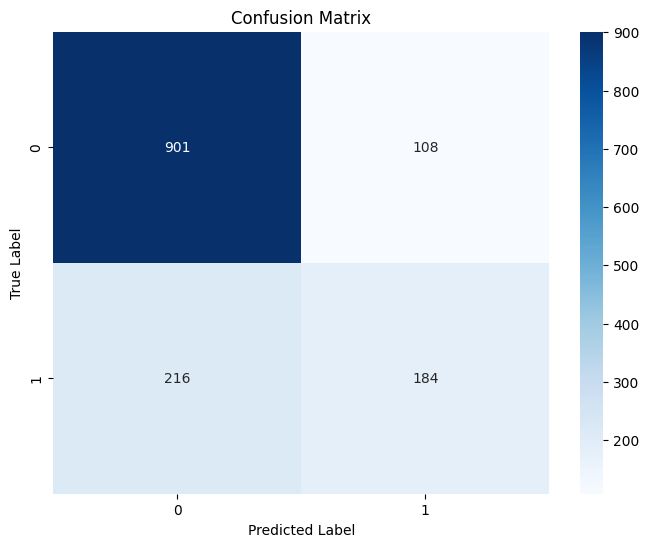
* Hasil sorting correlation matrix:



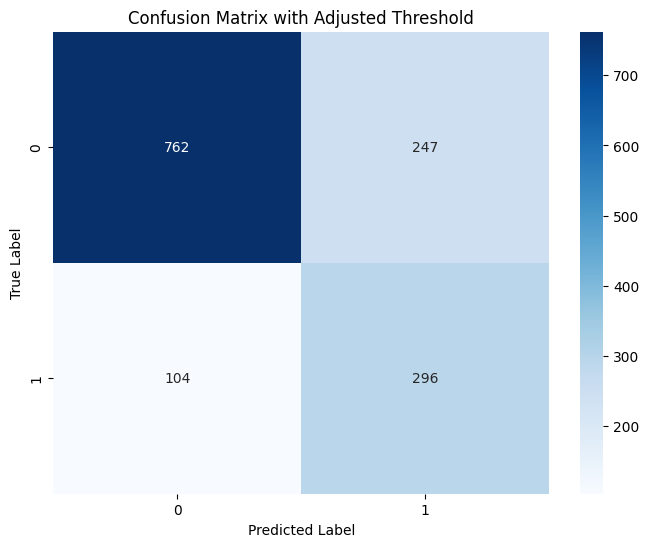
* Dari hasil sorting kita tahu yang berkorelasi kuat itu: tenure months, payment method, sama device class 🡪 ini akan dijadikan patokan conditional statements.
* Feature engineering supaya memudahkan random forest mengklasifikasi
* Tenure months
  + pelanggan baru: 0 -12
  + pelanggan lama: 13 – 48
  + pelanggan setia: > 48
* Monthly purchase
  + low: < 50
  + medium: 50 -100
  + high: > 100
* CLTV
  + low: < 4000
  + medium: 4000 - 6000
  + high: > 6000

Penentuan angka2 ini ke dalam kategorikal dari hasil perkiraan distribusinya

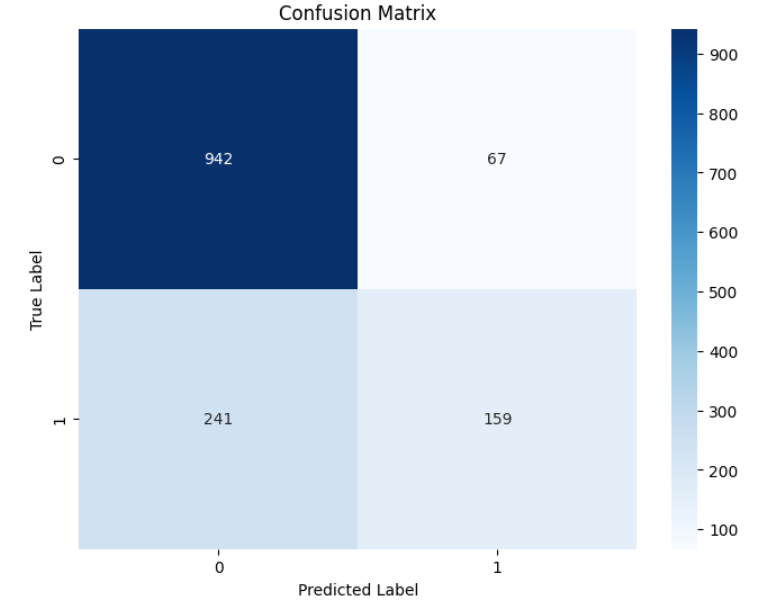
* Hasil confusion matrix random forest baseline



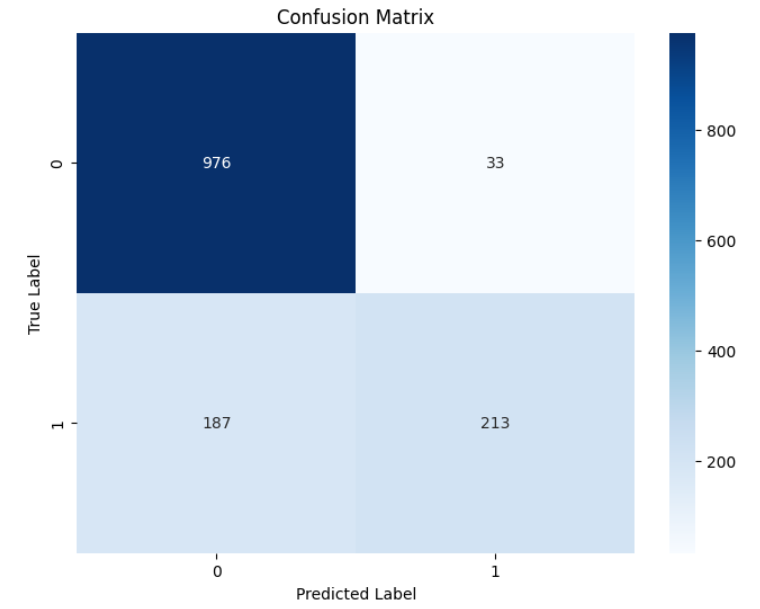
* Hasil confusion matrix random forest baseline + SMOTE



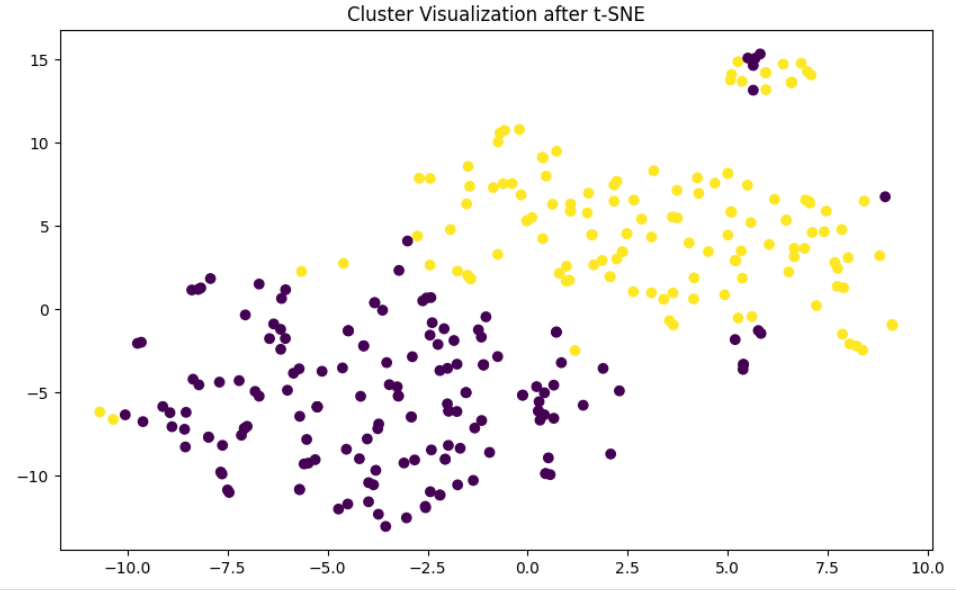
* Hasil confusion matrix random forest + feature engineering



* Hasil confusion matrix random forest + feature engineering + conditional statement



* Penentuan conditional statement:
  + Kalau dia pelanggan lama, pulsa, high end 🡪 kemungkinan churn / keluar
  + Kalau dia pelanggan baru, pulsa, high end 🡪 kemungkinan dia tidak churn / bertahan
* Penentuan conditional statement diambil dari cluster.xlsx. Dari file excel ini kita tahu



Akan terbentuk 2 cluster yang salah prediksi.

Bagian ungu misal cluster 0

Bagian kuning cluster 1

|  |  |  |  |
| --- | --- | --- | --- |
| Cluster | Prediksi | Aslinya | Jumlah data |
| 0 (pelanggan lama/setia) | Tidak keluar | keluar | 136 |
|  | keluar | Tidak keluar | 28 |
| 1 (pelanggan baru) | Tidak keluar | Keluar | 39 |
|  | keluar | Tidak keluar | 105 |

Dari sini dicoba2 filter di excel dari 3 fitur (tenure months, payment method, device class) mana value fitur yang paling banyak salah prediksi.

Ternyata keliatan bahwa:

* Kalau dia pelanggan lama, pulsa, high end 🡪 kemungkinan churn / keluar
* Kalau dia pelanggan baru, pulsa, high end 🡪 kemungkinan dia tidak churn / bertahan