Kelompok: Kelompok 5 – Anaconda
Stage: 3
Mentor: Fiqry R
Pukul/ Tanggal: 18 Juni 2022, 19.00 WIB

Pembagian tugas di stage 3:

1. Cross validation & Hyperparameter tuning: Muhammad Irfan Fadhlurrahman
2. Feature Importance: Ramado Dipradelana I
3. Feature Selection: Bima Sandi
4. Modelling: Deni Indra Permana
5. Insights from feature importance, Notulen stage 3, slide presentasi : Ni Putu Tasya
6. Laporan : Semuanya

Poin Pembahasan:

- Pemilihan metrics
- Modeling
- Feature Importance
- Feature Selection

Tindak Lanjut:

1. **Menambahkan alasan pemilihan metrics secara lebih detail**
2. **Mengatur ulang Pipeline untuk preprocessing agar hasil hyperparameter tuning dapat meningkatkan evaluation metrics**
3. **Menambahkan interpretasi feature importance sesuai hasil diskusi**

Kelompok: Kelompok 5 - Anaconda
Stage: 3
Mentor: Fiqry R
Pukul/ Tanggal: 18 Juni 2022, 19.00 WIB

Hasil Diskusi:

- Semakin banyak halaman yang dibuka, semakin besar kemungkinan seorang visitor e-commerce melakukan pembelian. Untuk meningkatkan jumlah pembelian, perlu menaikkan jumlah traffic, sehingga page value meningkat dan conversion rate juga meningkat.
- Nilai F1 masih kecil (< 70%). Argumen SMOTE perlu di-setting supaya nilai F1 meningkat. Tambahkan sampling_strategy='all'.
- Kolom month (feature engineering):
    - Tidak perlu, karena kesannya seperti bulan yg satu lebih rendah derajatnya daripada bulan yang lain. Lebih baik dikategorikan saja. Kelompokkan menjadi kelompok-kelompok tertentu, sehingga variable nominalnya jadi sedikit (e.g: kategorikan per quarter).
    - Parsing month seperti ini akan lebih tepat jika feature month dihubungkan dengan feature lain (misal: jumlah hari libur, etc), sehingga month tertentu lebih baik dari bulan lainnya.
    - Kategorikan bulan dengan menggunakan base tertentu: misal event selama bulan tersebut. Sehingga masuk akal.
- RFE sifatnya trial dan eror sampai stabil. Kalau tidak signifikan dibuang, lalu di train lagi. Pastikan lagi RFE emang membuang feature berdasarkan feature importance nya. Kalaupun mau pakai RFE kemungkinan di next step saja. Jadi sekarang di take down dulu.
- Sertakan landasan atau alasan saat membuang feature.
- SHAP sama partial dependence plot gak perlu masuk homework.
- Feature importance: melihat kontribusi dari setiap feature terhadap model.
- Dari feature-feature tersebut cek mana yang lebih masuk akal hubungannya dengan target dan sesuai dengan domain knowledge.
- Pemilihan Metrics:
    - Apakah model sudah sesuai dengan kemampuan prediksi? Apakah cukup akurat?
    - Apakah model yg di train sudah bisa belajar dengan adil dan sesuai kaidah atau tidak? Bila menggunakan data yang tidak seimbang maka model memperlakukan data secara timpang berdasarkan mayoritas datanya.
    - ROC-AUC untuk melihat apakah model sudah benar-benar belajar. Bisa juga pakai Gini Index (2 x ROC-AUC).
    - Boleh pakai precision dibandingkan F1 score. Karena resikonya false positive dan false negative timpang.
    - Boleh pakai F1 score kalau tujuannya bukan hanya meningkatkan jumlah transaksi tetapi juga menurunkan cost
    - Lebih masuk akal menggunakan pembobotan yang mana precisionnya lebih besar dibanding recall (F Beta Score), dengan konteks fokus dalam meningkatkan transaksi dan tetapi juga mempertimbangkan mengurangi cost.

Kelompok: Kelompok 5 - Anaconda
Stage: 3
Mentor: Fiqry R
Pukul/ Tanggal: 18 Juni 2022, 19.00 WIB

Hasil Diskusi (Lanjutan):

- Threshold overfitting tidak ada acuan yang bisa digunakan sebagai referensi dan bisa ditentukan sendiri, misalnya 10% difference between f1-train dan f1-test.
- Slide presentation:
- Masalah yang dihadapi; Revenue True lebih sedikit dibanding yang False. Masalah utama orang yang belanja jumlahnya sedikit.
- Tujuan: mencari insights untuk mengembangkan strategi perusahaan sehingga yang visitor e-commerce yang tidak belanja menjadi belanja.
- Ubah wording action nya, supaya terlihat kalau machine learning nya terpakai. Karena machine learning tidak hanya digunakan untuk meng-automasi pekerjaan manusia tetapi untuk mengekstrak insights dari feature importance suatu model.
- Jangan 2 grafik dalam 1 slide (tapi jika casenya pakai supporting grafik tidak masalah, asalkan dijelaskan dan diberi informasi tentang hubungannya