End-to-end Machine Learning Project

Petunjuk Pengerjaan:

- Silahkan membuat tugas machine learning menggunakan Python sesuai arahan yang diberikan di bawah ini
- Teman-teman dimohon untuk save hasil dalam .ipynb dan memberikan komentar/jawaban tertulis di luar syntax dalam bentuk comment ataupun dalam bentuk Power Point apabila dirasa lebih nyaman menggunakan PPT

Informasi terkait Dataset

Data ini adalah data mengenai pelanggan-pelanggan yang melakukan churn pada perusahaan telekomunikasi atau telco (churn adalah kehilangan pelanggan atau pelanggan yang “pergi” atau tidak berlangganan lagi). **Perusahaan ingin mengurangi jumlah pelanggan yang churn dengan memberikan penawaran menarik, bagi pelanggan yang berpotensi besar akan melakukan churn.**

Dataset ini terdiri dari 21 kolom, penjelasan setiap kolomnya secara berurutan sebagai berikut:

1. customerID: ID konsumen
2. Gender: jenis kelamin konsumen (laki-laki atau perempuan)
3. SeniorCitizen: konsumen merupakan warga lanjut usia atau bukan (1, 0)
4. Partner: konsumen memiliki pasangan atau tidak (Ya, Tidak)
5. Dependents: konsumen memiliki tanggungan atau tidak (Ya, Tidak)
6. Tenure: berapa bulan konsumen telah berlangganan di perusahaan
7. PhoneService: konsumen memiliki layanan telepon atau tidak (Ya, Tidak)
8. MultipleLines: konsumen memiliki banyak saluran atau tidak (Ya, Tidak, Tidak ada layanan telepon)
9. InternetService: layanan internet konsumen (DSL, Fiber optic, Bukan keduanya)
10. OnlineSecurity: konsumen memiliki keamanan online atau tidak (Ya, Tidak, Tidak ada layanan internet)
11. OnlineBackup: konsumen memiliki backup online atau tidak (Ya, Tidak, Tidak ada layanan internet)
12. DeviceProtection: konsumen memiliki perlindungan perangkat atau tidak (Ya, Tidak, Tidak ada layanan internet)
13. TechSupport: konsumen memiliki dukungan teknis atau tidak (Ya, Tidak, Tidak ada layanan internet)
14. StreamingTV: konsumen memiliki TV streaming atau tidak (Ya, Tidak, Tidak ada layanan internet)
15. StreamingMovies: konsumen memiliki streaming film atau tidak (Ya, Tidak, Tidak ada layanan internet)
16. Contract: jangka waktu kontrak konsumen (Bulan-ke-bulan, Satu tahun, Dua tahun)

17. PaperlessBilling: konsumen memiliki tagihan tanpa kertas (*paperless*) atau tidak (Ya, Tidak)
18. PaymentMethod: metode pembayaran konsumen (Cek elektronik, Cek pos, Transfer bank (otomatis), Kartu kredit (otomatis)
19. MonthlyCharges: jumlah yang ditagihkan kepada konsumen setiap bulan
20. TotalCharges: total yang ditagihkan kepada konsumen
21. Churn: konsumen keluar atau tidak (Ya, Tidak)

Tujuan Pengolahan Data

Mengetahui customer yang berpotensi churn (atau ganti provider) sejak dini, sehingga bisa segera dicegah.

Soal Penugasan

Data understanding

1. Berapa banyak baris dan kolom yang ada pada data
2. Kolom mana yang disebut sebagai fitur, dan mana yang disebut sebagai label?
3. Apakah tipe data sudah sesuai dengan deskripsi setiap fitur?
4. Explore setiap fitur dengan melihat statistika deskriptif atau value counts nya
5. Apakah ditemukan hal yang menarik atau hal yang janggal pada data?
6. Berapa banyak (bisa dalam %) customer yang churn pada data ini?

Data cleaning

1. Apakah ada missing values pada data ini?
2. Jika ada missing values, lakukan imputasi pada data
3. Apakah ada nilai unknown atau nilai lain yang tidak konsisten?
4. Lakukan cleaning jika ada nilai unknown atau nilai yang tidak konsisten
5. Apakah ada data yang duplikasi?
6. Jika ada data duplikat, bisa dilakukan cleaning juga
7. Jika ada yang janggal pada data, dan ingin melakukan drop fitur maupun modifikasi, silakan dilakukan
8. Setelah melalui tahapan cleaning, ada berapa banyak fitur dan baris yang tersisa?

EDA

1. Lakukan visualisasi setiap fitur yang ditinjau berdasarkan kolom label
2. Berdasarkan hasil perhitungan maupun visualisasi, berapa perbandingan proporsi kategori/kelas pada label? apakah balance atau imbalance?
3. Adakah yang menarik dari hasil visualisasi per fitur? fitur mana saja yang dirasa signifikan dan kurang signifikan?

Data preprocessing before modeling

1. Berdasarkan EDA yang telah dilakukan, fitur apa saja yang ingin di drop (tidak dipakai) dalam modeling

2. Definisikan input (X) dan output (y) untuk dimasukan ke dalam model
3. Lakukan encoding pada data kategorikal (boleh label encoding atau one hot encoding)
4. Lakukan normalisasi atau standarisasi pada data (berikan juga alasan kenapa menggunakan salah satunya)
5. Bagi data menjadi data train dan test dengan proporsi 75:25

Data modeling

1. Buatlah model regresi logistik dan xgboost untuk data ini
2. Lakukan validasi berdasarkan hasil kfold. apakah hasil setiap fold nya konsisten?
3. Berdasarkan confusion matrix, berapa precision dan recall yang dihasilkan masing-masing model?
4. Model manakah yang dirasa lebih baik untuk memodelkan data ini?
5. Lakukan modifikasi hyperparameter, tuning mana yang memberikan hasil paling baik?
6. Lakukan pula modifikasi threshold. Pada threshold berapa yang menghasilkan hasil prediksi paling baik untuk data ini?

Interpretasi bisnis

1. Bagaimana hasil model dapat diinterpretasikan ke permasalahan nyata?
2. Faktor apa saja yang mempengaruhi customer churn?
3. Apa suggestion yang bisa diberikan kepada perusahaan telco, berdasarkan model yang dihasilkan?