LAPORAN TEKNIS CAPSTONE PROJECT: PENAMBANGAN DATA

Implementasi Sistem Prediksi Churn Pelanggan E-commerce Menggunakan Ensemble Learning (Soft Voting)

Mata Kuliah: Penambangan Data **Disusun Oleh**: [Akbar Dwi Saputro / A11.2023.15371]
Tahun Akademik: Ganjil 2025/2026

BAB 1: PROBLEM DEFINITION & DATA ACQUISITION

1.1 Latar Belakang Masalah

Industri e-commerce modern menghadapi tantangan besar dalam mempertahankan loyalitas pelanggan. Biaya akuisisi pelanggan (Customer Acquisition Cost) yang terus meningkat membuat strategi retensi menjadi krusial bagi keberlanjutan bisnis. Fenomena *churn*—yaitu kondisi di mana pelanggan berhenti melakukan transaksi atau menghapus akun—merupakan indikator utama kegagalan retensi. Tanpa sistem deteksi dini, perusahaan sering kali terlambat memberikan penawaran atau insentif yang tepat untuk mencegah kepergian pelanggan.

Proyek ini bertujuan untuk membangun model prediktif berbasis *Machine Learning* yang mampu mengidentifikasi perilaku pelanggan yang berisiko tinggi untuk *churn*. Dengan memanfaatkan 13 fitur kunci perilaku pelanggan, sistem ini dirancang untuk memberikan skor risiko secara *real-time*. Penggunaan teknik *Ensemble Learning* (Soft Voting) dipilih untuk memberikan hasil yang lebih moderat dan stabil dibandingkan model tunggal, sehingga departemen CRM dapat mengambil keputusan yang lebih objektif dalam memberikan voucer atau program loyalitas tambahan.

1.2 Tujuan Bisnis dan Metrik Kesuksesan

- **Tujuan Bisnis**: Mengurangi tingkat kehilangan pelanggan melalui deteksi dini profil risiko tinggi.
- **Metrik Kesuksesan**: Mencapai nilai **F1-Score** minimal 0.85 untuk memastikan keseimbangan antara ketepatan prediksi (Precision) dan cakupan pelanggan yang berhasil dideteksi (Recall).

1.3 Statistik Deskriptif Dataset

- **Sumber Data**: Dataset pelanggan e-commerce dari repository publik.
- **Ukuran Data**: [Isi jumlah baris] baris dan 13 fitur.
- **Fitur Utama**: Terdiri dari 7 fitur numerik (Tenure, Jarak, Skor Kepuasan, dll.) dan 5 fitur kategorikal (Gender, Komplain, Status, dll.).

BAB 2: EXPLORATORY DATA ANALYSIS & PREPROCESSING

2.1 Analisis Kualitas Data (EDA)

Berdasarkan eksplorasi pada 01_eda.ipynb, ditemukan beberapa *insight* kunci:

1. **Missing Values**: Terdapat data kosong pada fitur Tenure yang ditangani menggunakan *Median Imputation*.
2. **Outliers**: Beberapa pelanggan memiliki nilai Cashback ekstrem yang tetap dipertahankan karena merupakan profil pelanggan premium.
3. **Korelasi Churn**: Fitur Komplain memiliki korelasi positif terkuat terhadap status *churn*.

2.2 Preprocessing Pipeline (Justifikasi Teknik)

Seluruh langkah prapemrosesan dibungkus dalam Scikit-learn Pipeline untuk mencegah *data leakage*:

- **Median Imputer**: Digunakan untuk data numerik agar tidak terpengaruh oleh *outlier*.
- **StandardScaler**: Melakukan standarisasi pada fitur numerik agar algoritma berbasis jarak (Logistic Regression) dapat konvergen lebih cepat.
- **One-Hot Encoding**: Mengubah data kategorikal menjadi biner agar dapat diproses secara matematis oleh XGBoost dan Logistic Regression.

BAB 3: MODELING & EVALUATION

3.1 Implementasi Model (Ensemble Learning)

Sesuai syarat Soal 3, proyek ini menerapkan minimal 2 model berbeda:

1. **XGBoost (Non-Linear)**: Sangat baik dalam menangkap pola interaksi fitur yang rumit. Dioptimasi menggunakan GridSearchCV untuk parameter max_depth dan learning_rate.
2. **Logistic Regression (Linear)**: Digunakan sebagai penyeimbang linear untuk menjaga stabilitas prediksi.
3. **Soft Voting**: Menggabungkan probabilitas kedua model untuk hasil yang lebih moderat.

3.2 Tabel Perbandingan Performa

Metrik	Logistic Regression	XGBoost (Tuned)	Ensemble (Final)
Accuracy	~84%	~91%	**~92%**

Metrik	Logistic Regression	XGBoost (Tuned)	Ensemble (Final)
F1-Score	~0.79	~0.88	**~0.90**
ROC-AUC	~0.86	~0.94	**~0.95**

3.3 Interpretasi Model (SHAP Values)

Interpretasi menggunakan SHAP menunjukkan bahwa fitur **Komplain**, **Tenure (Lama Berlangganan)**, dan **Skor Kepuasan** adalah tiga faktor terbesar yang menggerakkan model dalam menentukan risiko *churn*.

BAB 4: DEPLOYMENT & STREAMLIT APPLICATION

Aplikasi web yang di-deploy mencakup fungsionalitas berikut:

1. **Dashboard EDA**: Visualisasi interaktif hubungan komplain dan *churn*.
2. **Model Demo**: Antarmuka input 13 fitur untuk prediksi risiko secara instan.
3. **Evaluasi Model**: Tampilan *Confusion Matrix* dan metrik performa pelatihan.
4. **Auto-Initialization**: Logika pada app.py yang melatih model secara otomatis saat pertama kali dijalankan di cloud jika artefak belum tersedia.

BAB 5: KESIMPULAN & REKOMENDASI

5.1 Kesimpulan

Sistem *Ensemble Learning* berhasil diimplementasikan dengan peningkatan F1-Score sebesar ~2% dibandingkan model individu. Metrik evaluasi menunjukkan model sangat handal dalam menangani data tidak seimbang:

$$F1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

5.2 Rekomendasi Bisnis

Perusahaan disarankan untuk memprioritaskan penanganan keluhan pelanggan (Komplain), karena faktor ini meningkatkan risiko *churn* hingga berkali-kali lipat dibandingkan fitur lainnya.

LAMPIRAN TAUTAN

youtube

- https://youtu.be/YCmUE4dIMpc

github

- https://github.com/blackjack082703/UAS_Data_Mining