Homework Solution - Classification

I Gusti Ngurah Agung Hari Vijaya Kusuma Batch 57

August 8, 2025

Submission Links

- Repository: github.com/AgungHari

1 Pendahuluan

1.1 Latar Belakang

Perusahaan telekomunikasi di era digital saat ini menghadapi tantangan besar dalam mempertahankan pelanggan. Salah satu permasalahan utama yang sering dihadapi adalah fenomena *customer churn*, yaitu kondisi ketika pelanggan memutuskan untuk berhenti berlangganan suatu layanan. Tingginya tingkat churn dapat berdampak signifikan terhadap pendapatan perusahaan, mengingat biaya untuk memperoleh pelanggan baru umumnya lebih tinggi dibandingkan mempertahankan pelanggan yang sudah ada.

Memahami faktor-faktor yang mempengaruhi churn serta mampu memprediksi pelanggan mana yang berpotensi churn menjadi sangat penting bagi perusahaan. Dengan prediksi yang akurat, perusahaan dapat melakukan intervensi yang tepat, seperti memberikan penawaran khusus atau meningkatkan kualitas layanan, guna mencegah kehilangan pelanggan.

Melalui pemanfaatan data historis pelanggan, seperti data demografi, riwayat penggunaan layanan, dan perilaku pembayaran, teknologi *data science* dan *machine learning* memungkinkan perusahaan untuk membangun model prediksi churn yang efektif. Model ini dapat menjadi alat bantu pengambilan keputusan strategis, sehingga perusahaan dapat fokus pada segmen pelanggan yang paling berisiko churn dan merancang strategi retensi yang lebih efisien.

Tugas ini berfokus pada pembangunan model klasifikasi untuk memprediksi churn pelanggan pada perusahaan telekomunikasi berdasarkan berbagai fitur pelanggan. Diharapkan hasil dari proyek ini dapat memberikan insight dan rekomendasi bisnis yang relevan untuk meningkatkan retensi pelanggan.

1.2 Tujuan

Tujuan dari tugas ini adalah diantara lain untuk:

- Membangun model klasifikasi untuk memprediksi churn pelanggan pada perusahaan telekomunikasi.
- Menganalisis faktor-faktor yang berkontribusi terhadap churn pelanggan.
- Memberikan rekomendasi strategis untuk meningkatkan retensi pelanggan berdasarkan hasil analisis dan model yang dibangun.

1.3 Batasan atau Ruang Lingkup

Batasan masalah dalam tugas ini mencakup:

- Menggunakan dataset yang diberikan oleh rakamin.
- Penggunaan model klasifikasi untuk memprediksi churn.
- Analisis dilakukan pada fitur-fitur yang tersedia dalam dataset, tanpa melakukan pengumpulan data tambahan.

1.4 Manfaat

Manfaat dari tugas ini adalah membuat model klasifikasi yang dapat digunakan untuk memprediksi churn pelanggan.

2 Tinjauan Pustaka

2.1 Customer Churn

Customer churn, atau kehilangan pelanggan, adalah fenomena di mana pelanggan berhenti menggunakan layanan atau produk yang ditawarkan oleh suatu perusahaan. Hal ini menjadi perhatian utama bagi perusahaan, terutama di industri yang sangat kompetitif seperti telekomunikasi, karena churn dapat berdampak langsung pada pendapatan dan profitabilitas.

2.2 EDA (Exploratory Data Analysis)

Exploratory Data Analysis (EDA) adalah proses analisis data yang bertujuan untuk memahami struktur, pola, dan hubungan dalam dataset sebelum menerapkan model statistik atau machine learning. EDA melibatkan visualisasi data, statistik deskriptif, dan identifikasi anomali atau outlier. Proses ini penting untuk mendapatkan wawasan awal tentang data dan membantu dalam pengambilan keputusan selanjutnya.

2.3 Model Klasifikasi

Model klasifikasi adalah teknik dalam machine learning yang digunakan untuk mengelompokkan data ke dalam kategori atau kelas tertentu. Model ini dilatih menggunakan dataset yang berisi fitur-fitur input dan label output yang sesuai. Tujuan dari model klasifikasi adalah untuk memprediksi kelas dari data baru berdasarkan pola yang telah dipelajari dari data pelatihan.

2.4 Logistic Regression

Logistic regression adalah metode statistik yang digunakan untuk analisis regresi ketika variabel dependen bersifat kategorikal. Meskipun namanya mengandung kata "regression", logistic regression sebenarnya digunakan untuk klasifikasi. Metode ini memodelkan probabilitas suatu kejadian dengan menggunakan fungsi logit, yang mengubah output linear menjadi nilai antara 0 dan 1, sehingga cocok untuk prediksi kelas biner.

Untuk memahami logistic regression, kita perlu memahami konsep dasar dari regresi logistik. Regresi logistik digunakan untuk memprediksi probabilitas dari suatu kejadian yang bersifat biner (dua kelas), seperti churn atau tidak churn. Model ini mengasumsikan bahwa log odds dari probabilitas kejadian tersebut adalah linear terhadap fitur-fitur input.

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n)}} \tag{1}$$

di mana:

- $P(Y = 1|X)$ adalah probabilitas bahwa kelas target Y adalah 1 (misalnya, churn).
- β_0 adalah intercept dari model.
- $\beta_1, \beta_2, ..., \beta_n$ adalah koefisien regresi untuk fitur $X_1, X_2, ..., X_n$.

- e adalah basis dari logaritma natural.
- $X_1, X_2, ..., X_n$ adalah fitur-fitur input yang digunakan untuk memprediksi kelas target.

Dalam tugas ini apabila perhitungan probabilitas P(Y=1—X) lebih besar dari 0.5 maka akan dikategorikan sebagai churn, sebaliknya jika kurang dari 0.5 maka tidak churn.

2.5 Random Forest

Random Forest adalah algoritma ensemble learning yang menggabungkan beberapa pohon keputusan (decision trees) untuk meningkatkan akurasi prediksi. Setiap pohon dalam hutan dibangun menggunakan subset acak dari data pelatihan dan fitur, sehingga mengurangi risiko overfitting yang sering terjadi pada pohon keputusan tunggal. Random Forest dapat digunakan untuk klasifikasi maupun regresi, dan memiliki keunggulan dalam menangani data besar dengan banyak fitur.

$$\hat{Y} = \frac{1}{N} \sum_{i=1}^{N} T_i(X) \tag{2}$$

di mana:

- $\hat{Y}$ adalah prediksi akhir dari Random Forest.
- N adalah jumlah pohon dalam hutan.
- $T_i(X)$ adalah prediksi dari pohon keputusan ke-i untuk input X.
- X adalah fitur-fitur input yang digunakan untuk memprediksi kelas target.
- $\sum_{i=1}^{N} T_i(X)$ adalah jumlah prediksi dari semua pohon dalam hutan.

2.6 XGBoost/Gradient Boosting

XGBoost (Extreme Gradient Boosting) adalah algoritma machine learning yang merupakan implementasi dari teknik gradient boosting. XGBoost dirancang untuk efisiensi, fleksibilitas, dan kinerja yang tinggi. Algoritma ini membangun model prediksi secara bertahap dengan menambahkan pohon keputusan baru yang mengoreksi kesalahan dari pohon sebelumnya. XGBoost sangat populer dalam kompetisi data science karena kemampuannya dalam menangani data besar dan kompleks.

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \tag{3}$$

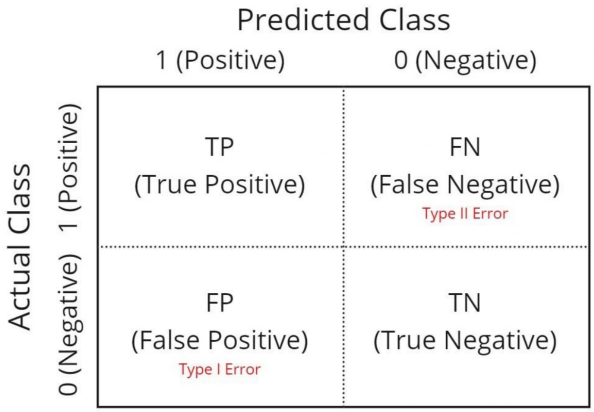
di mana:

- $\hat{y}_i$ adalah prediksi untuk data ke-i.
- K adalah jumlah pohon dalam model XGBoost.

- $f_k(x_i)$ adalah output dari pohon ke-k untuk input x_i.
- x_i adalah fitur-fitur input yang digunakan untuk memprediksi kelas target.
- $\sum_{k=1}^{K} f_k(x_i)$ adalah jumlah prediksi dari semua pohon dalam model XGBoost.

2.7 Confusion Matrix

Confusion matrix merupakan salah satu pengukuran yang paling mudah dilakukan dalam mencari nilai tingkat kebenaran dan juga akurasi dari model. *Confusion matrix* adalah sebuah tabel berbentuk dua dimensi yang terdiri dari data aktual dan data prediksi yang masing-masing memiliki kelas. Data aktual terletak pada bagian kolom tabel, sedangkan data prediksi terletak pada bagian baris dari tabel. Gambar 1 merupakan representasi visual dari perhitungan confusion matrix.



Gambar 1: Contoh Confusion Matrix

Confusion matrix memberikan informasi tentang jumlah prediksi yang benar dan salah untuk setiap kelas. Dari confusion matrix, kita dapat menghitung berbagai metrik evaluasi model seperti akurasi, presisi, recall, dan F1-score. Metrik-metrik ini membantu dalam menilai kinerja model klasifikasi secara keseluruhan.

2.8 Recall

Recall, juga dikenal sebagai sensitivitas atau true positive rate, adalah metrik evaluasi yang mengukur kemampuan model dalam mengidentifikasi kelas positif. Recall didefinisikan sebagai rasio antara jumlah prediksi benar positif (true positives) dengan jumlah total kasus positif aktual (true positives + false negatives). Rumusnya adalah:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{4}$$

di mana:

- True Positives (TP): Jumlah kasus positif yang benar-benar terdeteksi oleh model.

- False Negatives (FN): Jumlah kasus positif yang tidak terdeteksi oleh model (salah diklasifikasikan sebagai negatif).

Recall sangat penting dalam konteks di mana kita ingin meminimalkan jumlah kasus positif yang terlewatkan, seperti dalam deteksi penyakit atau pencegahan churn pelanggan. Metrik ini memberikan gambaran tentang seberapa baik model dapat menangkap semua kasus positif yang sebenarnya ada.

2.9 ROC-AUC

ROC (Receiver Operating Characteristic) curve adalah grafik yang menunjukkan kinerja model klasifikasi biner pada berbagai ambang batas (threshold). ROC curve memplot true positive rate (TPR) terhadap false positive rate (FPR) pada berbagai nilai threshold. Area di bawah kurva ROC (AUC - Area Under the Curve) memberikan ukuran kinerja model secara keseluruhan.

AUC adalah nilai antara 0 dan 1, di mana nilai 1 menunjukkan model yang sempurna (mampu memisahkan kelas positif dan negatif dengan sempurna), sedangkan nilai 0.5 menunjukkan model yang tidak lebih baik dari tebakan acak. AUC yang lebih tinggi menunjukkan bahwa model memiliki kemampuan yang lebih baik dalam membedakan antara kelas positif dan negatif.

$$\text{AUC} = \int_0^1 \text{TPR}(FPR)\, dFPR \tag{5}$$

di mana:

- TPR (True Positive Rate) adalah rasio antara jumlah prediksi benar positif dengan jumlah total kasus positif aktual.
- FPR (False Positive Rate) adalah rasio antara jumlah prediksi salah positif dengan jumlah total kasus negatif aktual.
- Integral ini menghitung area di bawah kurva ROC, yang memberikan nilai AUC.
- Nilai AUC yang lebih tinggi menunjukkan bahwa model memiliki kemampuan yang lebih baik dalam membedakan antara kelas positif dan negatif.

3 Desain dan Implementasi

3.1 Deskripsi sistem

4 Pengujian dan analisis

4.1 Pengujian Sistem

5 Kesimpulan dan Saran