

!! Catatan: Diharapkan seluruh pengerjaan Assignment tidak sepenuhnya mengandalkan penggunaan AI **!!**

"Proses belajar ibarat menanam pohon. Jika hanya mengandalkan AI tanpa memahami esensinya, yang berkembang bukan kompetensimu, melainkan ketergantungan yang melemahkan."

- Learning Design Dibimbing

Assignment Guidance: Case Study Feature Importance Analysis & Model Interpretation

Data Science dan Data Analyst Bootcamp

Periode Pembelajaran

Feature Importance Analysis & Model Interpretation

Objectives

1. Student mampu memahami dan menganalisis data pelanggan yang diberikan, termasuk distribusi usia, korelasi antara tingkat pendidikan dan pendapatan, serta faktor-faktor lain yang dapat mempengaruhi churn.
2. Student mampu membangun model prediksi churn menggunakan machine learning untuk membantu organisasi dalam mengidentifikasi pelanggan yang kemungkinan besar akan meninggalkan platform atau layanan.
3. Student mampu melakukan evaluasi kinerja model menggunakan metrik yang sesuai untuk memastikan keefektifan dan keakuratan prediksi, dan membandingkannya dengan baseline atau model yang ada sebelumnya.
4. Student mampu menghasilkan rekomendasi berdasarkan hasil analisis yang dapat membantu organisasi dalam mengurangi tingkat churn dan mempertahankan pelanggan.

Deskripsi Assignment

Anda adalah seorang konsultan data dari salah satu firma konsultan yang diminta oleh divisi Payment dari sebuah lembaga pinjaman online untuk memprediksi user mana yang kemungkinan besar akan churn.

Sebelumnya, model yang digunakan oleh perusahaan memiliki akurasi yang tinggi tetapi tidak dapat menjelaskan mengapa seorang customer diprediksi churn atau tidak. Oleh karena itu, Anda diminta untuk membangun model yang tidak hanya akurat, tetapi juga

mudah diinterpretasikan sehingga dapat memberikan insight bisnis yang dapat ditindaklanjuti.

Detail Assignment

Link Dataset: [Bank Churn Data](#)

Metadata pada Dataset sebagai berikut:

1. *user_id*: customer account number.
2. *attrition_flag*: customer status (Existing and Attrited).
3. *customer_age*: age of the customer.
4. *gender*: gender of customer (M for male and F for female).
5. *dependent_count*: number of dependents of customers.
6. *education_level*: customer education level (Uneducated, High School, Graduate, College, Post-Graduate, Doctorate, and Unknown).
7. *marital_status*: customer's marital status (Single, Married, Divorced, and Unknown).
8. *income_category*: customer income interval category (Less than \$40K, \$40K-\$60k, \$60K-\$80K, \$80K-\$120K, \$120K+, and Unknown).
9. *card_category*: type of card used (Blue, Silver, Gold, and Platinum).
10. *months_on_book*: period of being a customer (in months).
11. *total_relationship_count*: the number of products used by customers in the bank.
12. *months_inactive_12_mon*: period of inactivity for the last 12 months.
13. *contacts_count_12_mon*: the number of interactions between the bank and the customer in the last 12 months.
14. *credit_limit*: credit card transaction nominal limit in one period.
15. *total_revolving_bal*: total funds used in one period.
16. *avg_open_to_buy*: the difference between the credit limit set for the cardholder's account and the current balance.
17. *total_amt_chng_q4_q1*: increase in customer transaction nominal between quarter 4 and quarter 1.
18. *total_trans_amt*: total nominal transaction in the last 12 months.
19. *total_trans_ct*: the number of transactions in the last 12 months.
20. *total_ct_chng_q4_q1*: the number of customer transactions increased between quarter 4 and quarter 1.
21. *avg_utilization_ratio*: percentage of credit card usage.

Untuk menyelesaikan assignment ini, lakukan langkah-langkah berikut:

1. Eksplorasi Data

- Analisis distribusi usia pelanggan dan identifikasi kelompok usia yang memiliki kemungkinan churn terbesar.
- Analisis hubungan antara tingkat pendidikan dan pendapatan, serta bagaimana korelasinya dengan churn.

- Identifikasi perbedaan churn berdasarkan gender.

Goals: Memahami pola dalam data yang dapat membantu dalam pembuatan model prediktif.

2. Pemodelan Machine Learning

- Bangun model prediksi churn menggunakan algoritma machine learning yang sesuai.
- Gunakan teknik validasi yang tepat untuk mengukur performa model.
- Bandingkan performa model yang Anda buat dengan baseline model.

Goals: Membangun model yang akurat untuk memprediksi churn.

3. Evaluasi Model

- Gunakan metrik evaluasi yang sesuai untuk menilai performa model.
- Visualisasikan hasil evaluasi untuk membuktikan bahwa model yang dibangun lebih baik dibandingkan tanpa model.

Goals: Menilai efektivitas model yang dibangun dan memahami apakah model memberikan prediksi yang lebih baik daripada asumsi dasar.

4. Analisis Interpretabilitas Model

- Jika model yang digunakan bersifat interpretable, gunakan teknik interpretasi yang sesuai.
- Jika model yang digunakan bersifat non-interpretable, gunakan metode model-agnostic seperti:
 - **Permutation Feature Importance + Partial Dependence Plot**, atau
 - **Shapley Additive Value Feature Importance + Shapley Dependence Plot**

Goals: Memahami faktor-faktor yang berkontribusi terhadap churn dan memastikan model memberikan insight yang dapat digunakan untuk pengambilan keputusan bisnis.

5. Dokumentasi dan Rekomendasi

- Catat seluruh proses analisis, pemodelan, dan interpretasi model.
- Buat laporan yang mencakup rekomendasi kepada tim bisnis berdasarkan hasil analisis.

Goals: Menghasilkan laporan yang dapat digunakan oleh stakeholder untuk memahami hasil prediksi dan mengambil keputusan yang lebih baik.

Pertanyaan yang Perlu Dijawab

- Seperti apa distribusi usia para nasabah? Usia manakah yang cenderung memiliki kemungkinan churn terbesar?
- Apakah terdapat korelasi antara tingkat pendidikan dan income category? Bagaimana korelasi dari kedua prediktor tersebut terhadap churn?
- Apakah terdapat perbedaan yang signifikan antara gender dengan kemungkinan nasabah untuk churn?
- Apakah semakin lama customer menjadi nasabah maka semakin kecil kemungkinan nasabah tersebut untuk churn?
- Apakah semakin sering customer berinteraksi dengan bank dalam 1 tahun terakhir mempengaruhi besar atau kecilnya kemungkinan nasabah untuk churn?
- Buatlah model prediksi dengan menggunakan model machine learning (bebas, tidak dibatasi) dan evaluasi seberapa baik model yang anda buat.
- Anda ingin meyakinkan stakeholder anda bahwa model yang anda gunakan akan lebih baik dalam mengidentifikasi customer yang akan churn dibandingkan tanpa menggunakan model. Apa metrik yang dapat anda gunakan untuk membuktikan hal ini? Visualkan!
- Karena stakeholder anda membutuhkan knowledge dari model, analisislah model yang anda buat dengan menggunakan explanatory model analysis dan buatlah rekomendasi berdasarkan hasil dari analisis tersebut.

Catatan: Gunakan hanya salah satu dari 3 metode ini tergantung jenis modelnya:

- Jika model yang dipilih adalah model yang interpretable, gunakan teknik interpretasi sesuai dengan model yang dipilih.
- Jika model yang dipilih adalah model yang non-interpretable, gunakan metode model-agnostic seperti:
 - **Permutation Feature Importance + Partial Dependence Plot**, atau
 - **Shapley Additive Value Feature Importance + Shapley Dependence Plot**

Reflection Questions

1. Mengapa eksplorasi data (mis. analisis usia, pendidikan, gender) penting sebelum membangun model prediksi churn, dan bagaimana temuan EDA dapat memengaruhi pemilihan fitur serta performa model? Jawab dengan contoh konkret dari dataset (mis. kelompok usia atau tingkat pendidikan yang berisiko tinggi).
2. Apa keuntungan bisnis dari memiliki model prediktif churn yang akurat dibandingkan hanya mengandalkan asumsi atau aturan bisnis sederhana? Jelaskan dampak langsung terhadap strategi retensi, pengalokasian anggaran pemasaran, atau ROI.

Tools

Google Colaboratory

Pengumpulan Assignment

Deadline :

Maksimal H+10 Kelas (Pukul 23.30 WIB)

Details :

Dikumpulkan dalam bentuk link Google Colab. Jangan lupa input link github di Google Colab masing-masing, secara INDIVIDU, di LMS.

Seluruh jawaban Reflection Questions dikumpulkan dalam Google Docs terpisah (khusus untuk refleksi saja) dan juga dikumpulkan secara INDIVIDU melalui LMS.

Indikator Penilaian

No.	Aspek Penilaian	Parameter	Bobot Maksimal
1	Customer Churn Modelling	Mampu melakukan analisis data dengan baik dan menjawab pertanyaan	20
2	Model Evaluasi	Mampu membuat model evaluasi yang sesuai serta dapat mengevaluasi model tersebut	30
3	Pemilihan Metrik	Mampu menentukan metrik yang digunakan untuk mengevaluasi model	10
4	Interpretasi Model	Mampu menganalisis model yang dibuat dan memberikan rekomendasi dari hasil analisis	35
5	Reflection Questions		5
Total			100

Referensi/Template

1. <https://www.analyticsvidhya.com/blog/2020/03/support-vector-regression-tutorial-for-machine-learning/>
2. <https://towardsdatascience.com/unlocking-the-true-power-of-support-vector-regression-847fd123a4a0>

Sanksi Penggunaan AI:

Apabila student terdeteksi 100% menggunakan AI, maka hasil assignment akan diberikan skor 0

Ketentuan Pencapaian Nilai:

Nilai minimum Lulus Penyaluran Kerja: 75

Nilai minimum Lulus Bootcamp: 65

Ketentuan Penilaian:

Mengumpulkan Assignment tepat waktu: Sesuai dengan nilai yang diberikan mentor

Mengumpulkan Assignment 12 jam setelah deadline: - 3 dari nilai yang diberikan mentor

Mengumpulkan Assignment 1 x 24 Jam setelah deadline: - 6 dari nilai yang diberikan mentor

Mengumpulkan Assignment 2 x 24 Jam setelah deadline: - 12 dari nilai yang diberikan mentor

Mengumpulkan Assignment 3 x 24 Jam setelah deadline: - 18 dari nilai yang diberikan mentor

Mengumpulkan Assignment 4 x 24 Jam setelah deadline: - 24 dari nilai yang diberikan mentor

Mengumpulkan Assignment 5 x 24 Jam setelah deadline: - 30 dari nilai yang diberikan mentor

Mengumpulkan Assignment 6 x 24 Jam setelah deadline: - 36 dari nilai yang diberikan mentor

Mengumpulkan Assignment 7 x 24 Jam setelah deadline: - 42 dari nilai yang diberikan mentor