

PREDIKSI DURASI PERJALANAN DAN OPTIMISASI RUTE LOGISTIK MENGGUNAKAN MACHINE LEARNING DAN GOOGLE OR-TOOLS

Studi Kasus: NYC Yellow Cab Taxi

Abstract

Inefisiensi dalam logistik perkotaan, yang disebabkan oleh durasi perjalanan yang sulit diprediksi, telah menjadi tantangan operasional yang signifikan dan memakan biaya. Proyek ini bertujuan untuk mengatasi masalah tersebut dengan mengembangkan solusi dua tahap berbasis data. Tahap pertama adalah membangun model *machine learning* (pembelajaran mesin) **XGBoost** dengan akurasi tinggi untuk memprediksi durasi perjalanan taksi di New York City. Tahap kedua adalah menggunakan model prediksi ini sebagai "mesin penghitung biaya" di dalam *solver* **Google OR-Tools** untuk menyelesaikan *Vehicle Routing Problem* (VRP) atau Masalah Perutean Kendaraan. Hasil akhirnya adalah sebuah purwarupa aplikasi berbasis **Streamlit** yang mampu memberikan rekomendasi rute paling efisien untuk armada kendaraan, divisualisasikan pada peta interaktif **PyDeck** untuk mendukung pengambilan keputusan operasional yang lebih cerdas dan hemat biaya.

Proyek Untuk : Final Project Bootcamp at PT. Epam Digital Mandiri
User : Ir. Mega Bagus Herlambang, ST, MT, PhD, IPM, ASEAN.Eng
Data Scientist : Muhammad Ketsar Ali Abi Wahid
Tanggal : 07 Oktober 2025
Versi : 2.0

I. Latar Belakang Masalah (Background)

Perusahaan yang bergerak di bidang mobilitas dan logistik perkotaan seperti taksi atau kurir menghadapi tantangan operasional setiap hari di kota metropolitan.

- Masalahnya, efisiensi rute sangat bergantung pada estimasi waktu perjalanan yang akurat, namun lalu lintas kota sangat dinamis dan sulit ditebak.
- Konsekuensinya, terjadi pemborosan bahan bakar, waktu kerja pengemudi yang tidak efisien, dan ketidakpuasan pelanggan karena estimasi waktu tiba (ETA) yang sering meleset.
- Saat ini, perencanaan rute untuk armada seringkali dilakukan secara manual atau menggunakan perangkat lunak sederhana yang tidak mempertimbangkan pola lalu lintas berdasarkan waktu dan hari.
- Perusahaan membutuhkan sebuah sistem cerdas yang tidak hanya bisa memprediksi durasi perjalanan dengan akurat, tetapi juga bisa merangkai puluhan titik pengantaran menjadi rute-rute yang paling optimal untuk setiap kendaraan dalam armadanya.
- Dengan solusi ini, perusahaan dapat menekan biaya operasional secara signifikan, meningkatkan jumlah perjalanan yang bisa dilayani per hari, dan membangun kepercayaan pelanggan dengan ETA yang andal.

II. Pemahaman Bisnis (Business Understanding)

Mendefinisikan Masalah

Bagian	Deskripsi
Target User	Manajer Operasional atau Direktur Logistik di perusahaan taksi, kurir, atau layanan pengiriman.
Masalah	Bagaimana cara menentukan rute harian yang paling efisien (total waktu tempuh terendah) untuk sebuah armada kendaraan yang harus melayani banyak titik pelanggan yang tersebar, dengan kondisi lalu lintas yang selalu berubah?
Kenapa Penting	Pengurangan waktu tempuh armada sebesar 5-10% saja dapat menghemat miliaran rupiah per tahun dalam biaya bahan bakar dan gaji pengemudi, serta meningkatkan kapasitas layanan tanpa harus menambah jumlah kendaraan.
Kondisi Saat Ini	Perencanaan rute masih sangat bergantung pada pengalaman pengemudi atau menggunakan aplikasi navigasi standar yang hanya mengoptimalkan rute dari titik A ke B, bukan untuk serangkaian titik (A-B-C-D-...).

KPI (Key Performance Indicators)

Bagian	Deskripsi
Ekspektasi Manfaat	Mengurangi biaya operasional armada, meningkatkan produktivitas (jumlah perjalanan per hari), dan meningkatkan akurasi estimasi waktu tiba (ETA).
Metrics	<p>Outcome (Dampak Bisnis):</p> <ul style="list-style-type: none">Persentase penurunan total waktu perjalanan armada (%).Peningkatan jumlah pelanggan yang dilayani per kendaraan per hari. <p>Output (Performa Model):</p> <ul style="list-style-type: none">Model Prediksi: Nilai <i>error Root Mean Squared Logarithmic Error</i> (RMSLE).Model Optimisasi: Total waktu/jarak dari rute yang dihasilkan untuk sebuah skenario simulasi.
Targets	<p>Outcome:</p> <p>Mencapai simulasi penurunan total waktu tempuh armada sebesar >10% dibandingkan metode dasar (<i>baseline</i>).</p> <p>Output:</p> <ul style="list-style-type: none">Mencapai skor RMSLE di bawah 0.40 pada test set.

	<ul style="list-style-type: none"> Menghasilkan rekomendasi rute untuk skenario simulasi 5 kendaraan dan 50 titik pelanggan.
Batasan	<ul style="list-style-type: none"> Model prediksi dilatih menggunakan data tahun 2016; pola lalu lintas mungkin telah berubah. Rekomendasi rute tidak memperhitungkan kejadian real-time seperti kecelakaan atau penutupan jalan mendadak. Model VRP tidak mempertimbangkan batasan seperti jendela waktu pengiriman pelanggan (misal, "paket harus tiba antara jam 2-4 sore").

Sumber Data & Faktor (Features)

- Data Utama:** "2016 NYC Yellow Cab Trip Record Data" yang tersedia di kompetisi Kaggle. Dataset ini mencakup train.csv (1.45 juta baris) dan test.csv (0.62 juta baris).

Kategori	Nama Faktor	Deskripsi
ID & Vendor	<i>id, vendor_id</i>	Pengenalan unik untuk setiap perjalanan dan kode penyedia layanan taksi.
Waktu	<i>pickup_datetime</i>	Tanggal dan waktu penjemputan. Kolom target <i>trip_duration</i> (dalam detik) dan <i>dropoff_datetime</i> hanya ada di data latih.
Spasial/Geografis	<i>pickup/dropoff_longitude, pickup/dropoff_latitude</i>	Koordinat geografis titik penjemputan dan pengantaran.
Atribut Perjalanan	<i>passenger_count</i>	Jumlah penumpang dalam kendaraan.
	<i>store_and_fwd_flag</i>	Flag yang menandakan apakah data perjalanan disimpan dulu di memori kendaraan sebelum dikirim ke server.
Fitur Buatan	<i>distance_km</i>	Jarak Haversine (garis lurus di permukaan bumi) antara titik <i>pickup</i> dan <i>dropoff</i> .
(Engineered)	<i>pickup_hour, day_of_week, month</i>	Fitur yang diekstrak dari <i>pickup_datetime</i> untuk menangkap pola temporal.
	<i>speed_kmh</i>	Kecepatan rata-rata perjalanan, dihitung dari <i>distance_km</i> dan <i>trip_duration</i> . Berguna untuk deteksi anomali.

III. Persiapan Data (Data Preparation)

- **Pembersihan Data:** Menghapus data yang tidak logis, yaitu perjalanan dengan durasi di bawah 60 detik atau di atas 6 jam, serta perjalanan dengan jarak 0 km.
- **Feature Engineering:** Membuat fitur-fitur baru yang relevan seperti yang telah disebutkan di tabel atas (*distance_km*, *bearing*, *pickup_cluster*, dll.) untuk meningkatkan daya prediksi model.
- **Transformasi Data:** Menerapkan transformasi logaritmik $\log_{1p}()$ pada *trip_duration* untuk menormalkan distribusinya, sesuai dengan metrik evaluasi RMSLE.

IV. Modelling

Tahap 1: Model Prediksi Durasi

- Melakukan perbandingan performa sistematis terhadap 7 algoritma regresi (termasuk Random Forest, LightGBM, **XGBoost**, dan model linear) menggunakan metrik RMSLE melalui *K-Fold Cross-Validation*.
- Melakukan *hyperparameter tuning* menggunakan *RandomizedSearchCV* pada model-model terbaik.
- Berdasarkan hasil evaluasi, **XGBoost** dipilih sebagai model final karena memberikan keseimbangan terbaik antara akurasi dan kecepatan.

Tahap 2: Model Optimisasi Rute

- Menggunakan model **XGBoost** dari Tahap 1 sebagai "mesin penghitung biaya" untuk menghitung estimasi waktu tempuh antar setiap pasang titik lokasi.
- Mengimplementasikan solusi *Vehicle Routing Problem* (VRP) menggunakan library **Google OR-Tools** (modul *Constraint Programming*), bukan algoritma metaheuristik manual. Ini adalah *solver* canggih yang dirancang khusus untuk masalah optimisasi.

V. Evaluasi

- **Performa Model Prediksi:** Mengevaluasi model XGBoost final pada *test set* dan berhasil mencapai skor RMSLE yang ditargetkan (< 0.40).
- **Validasi Model Optimisasi:** Membandingkan total durasi rute yang dihasilkan oleh Google OR-Tools dengan metode dasar untuk sebuah skenario simulasi, menunjukkan adanya potensi efisiensi.
- **Interpretasi Model:** Menganalisis **Feature Importance** dari model XGBoost dan menggunakan plot **SHAP** (*SHapley Additive exPlanations*) untuk memahami secara mendalam faktor-faktor apa yang paling berpengaruh terhadap durasi perjalanan (misalnya, *distance_km* dan *pickup_cluster*).

VI. Deployment & Monitoring

Deployment

- **Visualisasi Hasil:** Membuat *dashboard* interaktif menggunakan **Streamlit** yang di-hosting secara lokal, yang memiliki dua fungsi utama sesuai implementasi kode:
 1. **Kalkulator Durasi:** Pengguna memasukkan titik awal, tujuan, dan waktu, lalu aplikasi menampilkan prediksi durasi perjalanan.

2. **Perencana Rute:** Pengguna memilih jumlah kendaraan, pelanggan, dan lokasi depot, lalu aplikasi menjalankan optimisasi dan menampilkan visualisasi rute paling optimal di peta interaktif menggunakan **PyDeck**.

Monitor (Rencana Jangka Panjang):

- **Penurunan Performa Model:** Menyadari bahwa performa model prediksi akan menurun seiring waktu karena perubahan pola lalu lintas atau infrastruktur baru.
- **Rencana Pelatihan Ulang (Re-training):** Merencanakan untuk melatih ulang model secara berkala (misal, setiap 1-2 tahun) dengan data perjalanan yang lebih baru agar prediksi dan rekomendasi rute tetap akurat dan relevan.