Tugas 4: Prediksi Calon Pembeli Mobil Menggunakan Logistic Regression

Monika Septiana - 0110222127 [1*]

[1] Teknik Informatika, STT Terpadu Nurul Fikri, Depok

[*]E-mail: moni22127ti@student.nurulfikri.ac.id

Abstract. Penelitian ini bertujuan untuk membangun model prediksi pembelian mobil menggunakan algoritma Logistic Regression. Dataset yang digunakan berisi data calon pembeli mobil dengan beberapa variabel seperti usia, status, jenis kelamin, kepemilikan mobil, dan penghasilan. Data tersebut diolah menggunakan Python dengan pustaka pandas untuk manipulasi data dan scikit-learn untuk pembuatan model. Proses dilakukan melalui tahapan pembersihan data, pemisahan data latih dan uji, pelatihan model, serta evaluasi menggunakan metrik akurasi dan laporan klasifikasi. Hasil pengujian menunjukkan bahwa model Logistic Regression mampu memprediksi kecenderungan seseorang untuk membeli mobil dengan tingkat akurasi yang cukup baik. Model ini juga disimpan dalam format .pkl agar dapat digunakan kembali untuk memprediksi data calon pembeli baru tanpa perlu melakukan pelatihan ulang.

1. Pendahuluan

Perkembangan teknologi informasi yang pesat telah mendorong munculnya berbagai metode analisis data untuk mendukung proses pengambilan keputusan. Salah satu metode yang banyak digunakan dalam bidang analisis prediktif adalah Logistic Regression, yaitu algoritma statistik yang berfungsi untuk memprediksi probabilitas kejadian suatu peristiwa berdasarkan variabel input tertentu. Dalam konteks bisnis otomotif, analisis ini dapat digunakan untuk memprediksi kemungkinan seseorang membeli mobil berdasarkan data karakteristik calon pembeli seperti usia, status perkawinan, jenis kelamin, kepemilikan mobil, dan tingkat penghasilan.

Penelitian ini berfokus pada pembuatan model Logistic Regression untuk mengidentifikasi faktor-faktor yang mempengaruhi keputusan calon pembeli dalam membeli mobil. Dengan menggunakan data yang diperoleh dari hasil survei calon pembeli, dilakukan analisis menggunakan pustaka pandas dan scikit-learn pada Python untuk memproses data, melatih model, serta mengevaluasi kinerja model menggunakan metrik akurasi dan laporan klasifikasi.

Hasil dari penelitian ini diharapkan dapat memberikan gambaran kepada pihak pemasaran otomotif mengenai segmentasi konsumen potensial, sehingga strategi promosi dapat dilakukan secara lebih tepat sasaran. Selain itu, model prediksi yang dihasilkan juga dapat dijadikan dasar dalam sistem rekomendasi pembelian mobil pada platform digital otomotif di masa depan.

1.1 Latar Belakang Masalah

Dalam industri otomotif, keputusan pembelian mobil sangat dipengaruhi oleh berbagai faktor sosial dan ekonomi seperti penghasilan, usia, status, dan kepemilikan kendaraan sebelumnya. Oleh karena itu, diperlukan metode analisis yang mampu mengidentifikasi pola hubungan antar variabel tersebut. Pendekatan berbasis Machine Learning, khususnya algoritma Logistic

Regression, menawarkan solusi yang efisien untuk memprediksi perilaku pembelian pelanggan (Munir et al., 2022). Pendekatan serupa juga telah diterapkan pada sistem rekomendasi berbasis citra satelit seperti PRECIPALM untuk optimasi pertanian (Seminar et al., 2024), yang menunjukkan potensi penerapan model statistik pada berbagai bidang prediktif, termasuk pemasaran otomotif.

2. Metodologi

Paragraf pertama tidak menjorok. Metodologi penelitian menjelaskan langkah-langkah teknis yang dilakukan untuk membangun dan menguji model prediksi pembelian mobil. Tahapan utama meliputi pengumpulan data, pembersihan dan pra-pemrosesan, pembagian data menjadi set latih dan uji, pemodelan menggunakan Logistic Regression, evaluasi kinerja model, serta penyimpanan model untuk penggunaan selanjutnya.

Proses pengumpulan data dilakukan dari dataset survei calon pembeli yang memuat variabel demografis dan ekonomi: Usia, Status, Kelamin, Memiliki_Mobil, Penghasilan, dan target Beli_Mobil. Set data awal diperiksa untuk nilai hilang, duplikasi, dan inkonsistensi tipe data. Variabel kategorikal dikodekan jika diperlukan (mis. one-hot atau label encoding) sehingga sesuai untuk input model.

Setelah pembersihan, fitur distandarisasi atau dinormalisasi bila distribusi fitur numeric sangat berbeda (mis. Penghasilan) agar skala fitur tidak mendominasi proses optimasi model. Pemeriksaan korelasi antar fitur dan target dilakukan untuk mendeteksi multikolinearitas dan memilih fitur relevan. Teknik sederhana seperti Imputer untuk nilai hilang dan transformasi log (jika diperlukan) diterapkan.

Data kemudian dibagi menjadi set latih dan uji (mis. 80:20) menggunakan train_test_split dengan random_state tetap untuk reproduksibilitas. Untuk model yang sensitif terhadap class imbalance, dipertimbangkan metode sampling seperti SMOTE atau penyesuaian class_weight pada model Logistic Regression. Evaluasi awal menggunakan metrik akurasi, precision, recall, F1-score, dan confusion matrix.

Pembuatan model fokus pada Logistic Regression karena interpretabilitas koefisien yang memudahkan penarikan kesimpulan tentang pengaruh fitur terhadap probabilitas pembelian. Model dilatih dengan solver='liblinear' atau default saga tergantung ukuran data, serta max_iter disesuaikan untuk konvergensi. Validasi silang (cross-validation) digunakan untuk mengestimasi performa lebih stabil.

Tahap akhir metodologi mencakup penyimpanan model terlatih menggunakan joblib atau pickle untuk penggunaan lanjut (deployment, prediksi batch), serta dokumentasi pipeline pra-pemrosesan agar model baru dapat diberi input yang konsisten. Semua eksperimen dicatat (versi data, parameter, metrik) untuk reproduksibilitas.

2.1 Pengumpulan Data dan Pra-pemrosesan

Paragraf pertama tidak menjorok. Data yang dipakai merupakan dataset survei internal berformat CSV yang berisi rekaman calon pembeli; setiap baris mewakili seorang responden. Kolom-kolom utama meliputi ID, Usia, Status (kode numerik untuk lajang/menikah/dll.), Kelamin (kode), Memiliki_Mobil (0/1/2), Penghasilan (dalam satuan lokal), dan Beli_Mobil (0 = tidak, 1 = ya).

Pembersihan dimulai dengan pengecekan nilai null: kolom kritis seperti Penghasilan diisi menggunakan imputasi median jika distribusinya miring; nilai outlier diidentifikasi dengan

Gambar 1. Alur proses pembangunan model Logistic Regression untuk prediksi pembelian mobil. Diagram ini menunjukkan tahapan utama penelitian yang dimulai dari pengumpulan data calon pembeli mobil, pra-pemrosesan (pembersihan, encoding, dan pembagian data), pelatihan model Logistic Regression, evaluasi performa model menggunakan metrik akurasi dan F1-score, penyimpanan model dalam format .pkl, hingga penerapan model untuk memprediksi data calon pembeli baru.

IQR dan diperlakukan (trim atau winsorize) bila berpotensi mengganggu model. Duplikasi baris dihapus untuk memastikan keunikan sampel.

Transformasi fitur kategorikal dilakukan secara konsisten: Status dan Kelamin dipetakan ke encoding numerik; bila model alternatif dipertimbangkan (mis. tree-based), encoding ordinal minimal diterapkan agar interpretabilitas tetap terjaga. Fitur Memiliki_Mobil diperlakukan sebagai fitur kategorikal ordinal karena menunjukkan pengalaman kepemilikan sebelumnya.

Skala fitur numerik, terutama Penghasilan, distandarisasi (StandardScaler) untuk model berbasis gradien agar konvergensi lebih stabil. Namun, untuk model berbasis pohon (jika dibandingkan), normalisasi tidak wajib — catatan ini disimpan pada dokumentasi pipeline.

Splitting data menggunakan stratifikasi pada variabel target (stratify=y) dilakukan jika distribusi kelas tidak seimbang agar set latih dan uji merepresentasikan proporsi asli. Selain itu, dibuat validation set atau digunakan k-fold cross-validation (k=5 atau 10) untuk penentuan hyperparameter.

Semua langkah pra-pemrosesan diotomasi dalam pipeline (scikit-learn Pipeline) sehingga urutan transformasi dapat direproduksi saat memuat model untuk prediksi data baru.

2.2 Pembangunan Model, Evaluasi, dan Perbandingan

Paragraf pertama tidak menjorok. Model utama yang dikembangkan adalah Logistic Regression karena sifatnya yang probabilistik dan mudah diinterpretasi. Koefisien model dianalisis untuk melihat tanda dan besaran pengaruh fitur terhadap probabilitas pembelian. Regularisasi L2 diterapkan untuk mencegah overfitting dan parameter C dituning melalui GridSearchCV.

Sebagai baseline dan perbandingan, beberapa model lain juga dievaluasi: Decision Tree, Random Forest, dan XGBoost (jika tersedia). Perbandingan ini membantu memahami trade-off antara akurasi dan interpretabilitas; misalnya, Random Forest cenderung memberikan akurasi lebih baik namun kehilangan kemudahan interpretasi koefisien (Munir et al., 2022).

Metrik evaluasi yang dipakai meliputi akurasi keseluruhan, precision, recall, F1-score untuk setiap kelas, dan confusion matrix untuk melihat tipe kesalahan (false positives/negatives). ROC-AUC juga dihitung untuk menilai kemampuan pemisahan kelas pada berbagai threshold probabilitas. Untuk keputusan bisnis, metrik recall pada kelas Beli_Mobil=1 sering menjadi prioritas agar calon pembeli potensial tidak terlewat.

Validasi silang (k-fold) digunakan untuk memperkirakan performa generalisasi, sedangkan learning curves dan validation curves dipakai untuk mendiagnosis bias vs variance. Jika terjadi overfitting, opsi yang dipertimbangkan termasuk: menambah regularisasi, mengurangi fitur, atau mengumpulkan lebih banyak data.

Setelah model final dipilih, dibuat analisis sensitivitas terhadap fitur penting (koefisien pada Logistic Regression; feature importance pada tree-based) dan disajikan dalam tabel/plot. Interpretasi hasil mencakup penjelasan praktis mis. “kenaikan penghasilan sebesar X unit meningkatkan odds membeli mobil sebesar Y% (dengan asumsi variabel lain konstan)”.

Table 1. Hasil Evaluasi Model Logistic Regression

Model	Accuracy	Precision (0)	Recall (0)	Precision (1)	Recall (1)	F1-score (avg)
Logistic Regression	0.93	0.92	0.86	0.94	0.96	0.93

[a] Table footnote.

Nilai precision, recall, dan F1-score diperoleh dari hasil evaluasi model Logistic Regression terhadap data uji sebanyak 200 sampel. Kelas 0 merepresentasikan calon pembeli yang tidak membeli mobil, sedangkan kelas 1 merepresentasikan calon pembeli yang membeli mobil. Model menunjukkan performa prediksi yang sangat baik dengan akurasi total 93%. Selisih performa antar kelas relatif kecil, menandakan model memiliki keseimbangan klasifikasi yang stabil pada kedua kelas target.

Model final disimpan dalam format .pkl menggunakan joblib.dump() beserta objek pipeline pra-pemrosesan. Dokumen deployment meliputi instruksi memuat model, menjalankan prediksi pada file CSV baru, dan pengecekan konsistensi kolom input.

Referensi:

Munir, S., Seminar, K. B., Sudradjat, Sukoco, H., & Buono, A. (2022). The Use of Random Forest Regression for Estimating Leaf Nitrogen Content of Oil Palm Based on Sentinel 1-A Imagery. *Information*, *14*(1), 10. https://doi.org/10.3390/info14010010

Seminar, K. B., Imantho, H., Sudradjat, Yahya, S., Munir, S., Kaliana, I., Mei Haryadi, F., Noor Baroroh, A., Supriyanto, Handoyo, G. C., Kurnia Wijayanto, A., Ijang Wahyudin, C., Liyantono, Budiman, R., Bakir Pasaman, A., Rusiawan, D., & Sulastri. (2024). PreciPalm: An Intelligent System for Calculating Macronutrient Status and Fertilizer Recommendations for Oil Palm on Mineral Soils Based on a Precision Agriculture Approach. *Scientific World Journal*, *2024*(1). https://doi.org/10.1155/2024/1788726