Rakamin Academy

Online Shoppers Purchasing Intention

Summary Final Project Stage 3 - Anaconda

Summary

Split data train and test: Sebelum data preprocessing dan modelling, data di-split menjadi dua yaitu train set dan test set agar tidak terjadi data leakage. Rasio split yang digunakan 80% train set dan 20% test set. Kemudian agar distribusi target antara train dan test set tetap sama maka menggunakan stratified sampling. Model yang digunakan Logistic Regression (baseline model), Decision Tree, Random Forest, Extra Trees, Ada Boost, dan XGBoost. Alasan memilih tree-based model karena tepat digunakan untuk karakteristik dataset yang memiliki distribusi right-skewed dan terdapat banyak outliers.

Model evaluation: Evaluation metrics yang digunakan adalah precision dan ROC-AUC score. Alasan memilih precision karena kesalahan prediksi seorang visitor membeli padahal aktualnya tidak membeli (false positive) lebih beresiko mengurangi revenue dibandingkan kesalahan prediksi seorang visitor tidak membeli namun aktualnya membeli (false negative). Alasan memilih ROC-AUC score yaitu sebagai metric tambahan untuk mengevaluasi performa model pada dataset yang imbalance. Walaupun model yang sudah di train sudah melewati baseline model yang berarti model sudah cukup baik, namun model belum best-fit.

Summary

Hyperparameter tuning: Parameter yang digunakan tree-based model cenderung sama. Parameter yang umum digunakan untuk di-tuning sebagai berikut.

- n_estimators: jumlah subtree yang akan dibangun. Semakin banyak subtree, semakin meningkatkan waktu komputasi
- criterion: cara menghitung impurity pada feature (gini, entropy). melihat feature mana yang menjadi root/node
- max_depth: maksimal kedalaman tree untuk mencegah overfitting
- min_sample_split: berapa jumlah sample yg dibutuhkan pada node untuk membuat leaf baru (agar tidak terlalu sedikit sehingga mengakibatkan overfit).
- min_sample_leaf: berapa jumlah sample yg dibutuhkan pada leaf agar leaf terbentuk (agar tidak terlalu sedikit sehingga mengakibatkan overfit).

Eksperimen yang dilakukan dengan menambahkan feature-feature baru lalu mencoba men-train dengan berbagai model dan mencoba berbagai feature encoding. Kemudian, mencoba melakukan cross validation dan hyperparameter tuning. Setelah hyperparameter tuning, seluruh tree-based model hampir semuanya overfitted kecuali Decision Tree. Jika dibandingkan dengan sebelum tuning, performa model sebelum tuning lebih baik. Oleh karena itu, akan menggunakan model sebelum tuning untuk mengevaluasi feature importance. Model yang dipilih adalah Random Forest karena dibandingkan dengan tree-based model lainnya memiliki nilai ROC-AUC test paling tinggi dan cenderung tidak overfitted.

Summary

Feature Importance: 5 feature dengan importance score tertinggi yaitu

- page_values,
- exit_rates,
- month_Q2
- product_related_duration
- administrative_duration

Berdasarkan feature importance, dapat dilihat bahwa page values memiliki pengaruh yang signifikan terhadap konversi. Sesi dengan page values yang tinggi cenderung menghasilkan revenue, sehingga untuk menghasilkan peningkatan pada purchase/conversion rate, kita juga perlu meningkatkan page values. Salah satu caranya adalah dengan meningkatkan jumlah trafik yang berkualitas. Dalam strategi marketing, pemilihan waktu yang tepat untuk memberikan promo atau voucher diskon juga penting. Cek event-event tertentu dari suatu bulan yang memungkinkan peningkatan jumlah user yang belanja.

Summary

Sesi yang menghasilkan pendapatan cenderung memiliki exit rate yang lebih rendah dibandingkan yang tidak menghasilkan revenue. Sehingga untuk meningkatkan conversion rate/purchase rate, diperlukan aksi yang dapat membantu mengoptimalisasi exit rate. Lakukan pengecekan halaman mana dari ecommerce yang memiliki exit rates paling tinggi. Lalu lakukan optimalisasi. Bila dilihat dari seluruh transaksi yang menghasilkan revenue, sebagian besar adalah berasal dari visit ke product page. Durasi kunjungan ke product page cukup berpengaruh terhadap konversi. Namun, 90% total kunjungan product page dibandingkan dengan 15% sesi yang menghasilkan conversion, terbilang rendah. kita perlu melakukan pengecekan terhadap product related page dari segi interface, kemudahan akses, kejelasan informasi dan demonstrasi produk.

Dari sisi domain knowledge, durasi kunjungan ke administrative page seharusnya tidak berperan dalam peningkatan jumlah konversi. Namun dari feature importance, administrative duration termasuk salah satu feature yang memiliki pengaruh yang cukup tinggi ke model, dimana dari sesi yang menghasilkan transaksi, sebagian besar memiliki kunjungan ke administrative page. Perlu di cek apakah user yang melakukan konversi dan memiliki kunjungan ke administrative page merupakan proses UAT atau bukan. Asumsi: administrative page adalah admin user page (dashboard).