Rakamin Academy

Online Shoppers Purchasing Intention

Dokumen Laporan Final Project Stage 3 - Anaconda

Modelling

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

Model klasifikasi yang digunakan sebagai berikut. Hampir seluruh model merupakan tree-based model. Alasan memilih tree-based model karena tepat digunakan untuk karakteristik dataset yang memiliki distribusi right-skewed dan terdapat banyak outliers.

- Logistic Regression (baseline model)
- Decision Tree
- Random Forest
- Extra Trees
- Ada Boost
- XGBoost

Modelling

Pemilihan metrics

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.

Model evaluation

- Logistic Regression cenderung sudah best fit yang mana nilai ROC AUC score pada train dan test set hampir sama sebesar 0.910 dan 0.912.
- Tree-based model seperti Random Forest, Extra Trees, dan Ada Boost, cenderung tidak overfitting. Namun Decision Tree dan XGBoost masih overfitting.
- Tree-based model yang paling bagus adalah Random Forest dengan ROC AUC dan precision test sebesar 0.922 dan 0.618.

Cross validation & hyperparameter tuning

Cross validation setup

Cross validation menggunakan 5 fold dan stratified sampling untuk membentuk masing-masing fold agar distribusi target setiap fold tetap sama.

Hyperparameter tuning setup

Hanya tree-based model saja yang di-tuning karena di awal menganggap Logistic Regression sebagai baseline model untuk mengukur seberapa bagus performa tree-based model jika dibandingkan dengan Logistic Regression.

Hyperparameter tuning untuk berbagai tree-based model berada di notebook terpisah dan menggunakan Optuna agar proses tuning lebih singkat dibandingkan RandomizedSearchCV atau GridSearchCV. Berikut ini link untuk notebook tersebut.

- exp-hyperparameter-tuning.ipynb

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).

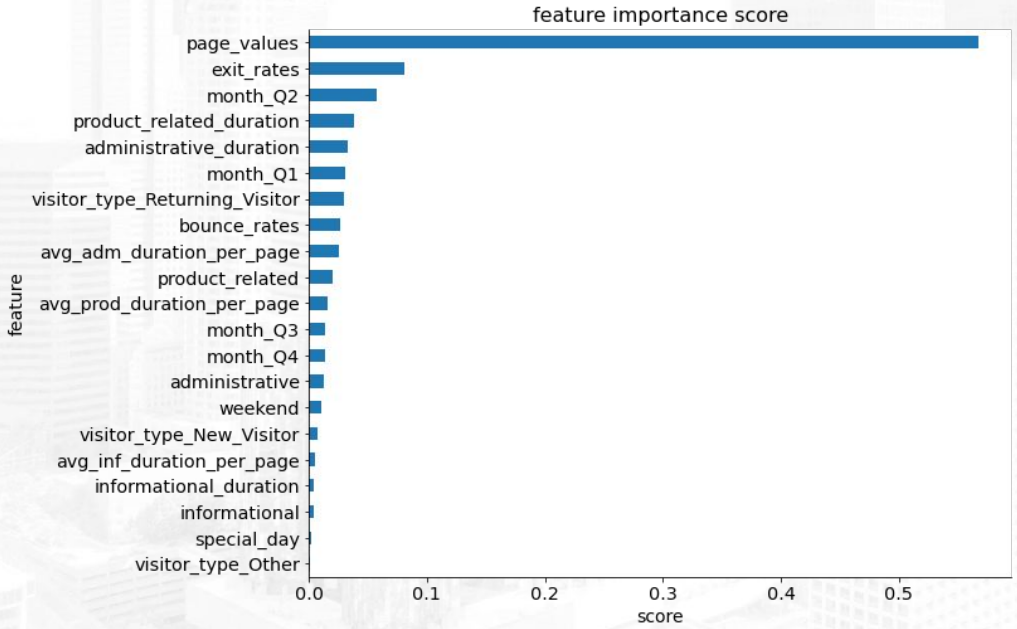
Cross validation & Hyperparameter tuning

Hasil 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 tidak overfitted.

Feature Importance



Pengamatan

5 feature dengan importance score tertinggi yaitu

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

page_values sangat berpengaruh ke model dan nilainya sangat tinggi (> 0.5) jika dibandingkan dengan feature-feature lain.

Rekomendasi Aksi dari hasil interpretasi model

- **Rekomendasi Aksi**
  - **a. page_values**

    Page values merupakan nilai rata-rata untuk halaman yang dikunjungi pengguna sebelum menyelesaikan konversi atau transaksi eCommerce (ref). 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. ***Lakukan pengecekan: halaman mana yang memiliki traffic yang tinggi, tetapi page values nya rendah, lalu optimalkan konten dan interface dari halaman tersebut, sehingga memperbesar kemungkinan customer akan melakukan pembelian.***

  - **b. month_Q2**

    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. Dari dataset dengan asumsi ecommerce ini aktivitasnya berkegiatan di US, bulan pada Q2 yang memiliki conversion terbanyak adalah May (Mother's day) - source. Dengan memprediksi mana customer yang akan membeli dan mana yang tidak, kita bisa melakukan aksi berupa ***pemberian voucher diskon khusus event tertentu kepada customer yang diprediksi tidak membeli, sehingga jadinya membeli***. Untuk lebih meminimalisir cost bisa juga ***memberi voucher untuk pembelian produk yang kemungkinan banyak dibutuhkan pada event tersebut*** atau pada bulan tersebut (misal: produk coklat saat valentine, kotak kado di libur natal, etc)

Rekomendasi Aksi dari hasil interpretasi model

c. exit_rates

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.*** Jika ternyata halaman yang memiliki exit rates tinggi adalah halaman produk, atau halaman check out, lakukan lagi pengecekan kenapa user tidak melanjutkan proses conversi. Apakah karena harga, atau waktu tertentu yang tidak tepat untuk membeli suatu barang.

d. product_related_duration

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. Sehingga sama dengan rekomendasi pada exit rates, kita perlu **melakukan pengecekan terhadap product related page dari segi interface, kemudahan akses, kejelasan informasi dan demonstrasi produk, dll. Apakah high product page visit tapi low conversion disebabkan oleh hal-hal tersebut.**

Rekomendasi Aksi dari hasil interpretasi model

e. Administrative_duration

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).

Feature Selection

```
all features
{'model': 'randomforestclassifier',
 'Accuracy (Test set)': 0.89,
 'Precision (Test set)': 0.618,
 'Recall (Test set)': 0.78,
 'F1 (Test set)': 0.69,
 'F0.5 (Test set)': 0.645,
 'ROC AUC (Test set)': 0.922}

use 5 features:['page_values', 'exit_rates', 'month_Q2', 'product_related_duration', 'administrative_duration']
{'model': 'randomforestclassifier',
 'Accuracy (Test set)': 0.884,
 'Precision (Test set)': 0.602,
 'Recall (Test set)': 0.772,
 'F1 (Test set)': 0.677,
 'F0.5 (Test set)': 0.63,
 'ROC AUC (Test set)': 0.911}

use 9 features:['page_values', 'exit_rates', 'month_Q2', 'product_related_duration', 'administrative_duration', 'month
_Q1', 'visitor_type_Returning_Visitor', 'bounce_rates', 'avg_adm_duration_per_page']
{'model': 'randomforestclassifier',
 'Accuracy (Test set)': 0.888,
 'Precision (Test set)': 0.613,
 'Recall (Test set)': 0.772,
 'F1 (Test set)': 0.684,
 'F0.5 (Test set)': 0.64,
 'ROC AUC (Test set)': 0.92}
```

Feature selection berdasarkan feature importance

Apabila menggunakan 5 atau 9 features dengan importance paling tinggi, nilai precision dan ROC AUC lebih rendah dibandingkan dengan memakai semua features

Git

Repository for Final Project

https://github.com/irfan-fadhlurrahman/online-shoppers-purchasing-intention

Kontribusi

1. Cross validation & Hyperparameter tuning: Muhammad Irfan Fadhlurrahman
2. Feature Importance: Ramado Dipradelana I
3. Feature Selection: Bima Sandi
4. Modelling: Deni Indra Permana
5. Insights from feature importance, Notulen stage 3, slide presentasi : Ni Putu Tasya
6. Laporan : Semuanya