

Latar Belakang:

Dalam dunia digital marketing, iklan online telah menjadi salah satu cara utama untuk mempromosikan produk dan jasa. Untuk meningkatkan efektivitas iklan, perusahaan seringkali menggunakan alat berbasis machine learning untuk mengklasifikasikan customer berdasarkan perilaku dan preferensi mereka terhadap iklan. Dengan menganalisis perilaku customer, perusahaan dapat memahami lebih baik jenis iklan mana yang lebih cenderung menarik perhatian dan menarik para calon pembeli, serta dapat menyajikan iklan yang lebih relevan dan tepat sasaran.

Tujuan:

Tujuan dari proyek "Predict Clicked Ads Customer Classification by using Machine Learning" adalah untuk membangun model machine learning yang dapat mengklasifikasikan customer berdasarkan kemungkinan mereka untuk mengklik iklan tertentu. Dengan demikian, tujuan proyek ini adalah untuk memberikan wawasan yang lebih dalam tentang tipe-tipe customer yang berbeda dan bagaimana perilaku mereka terhadap iklan-iklan yang ditampilkan.

Masalah Utama:

Masalah utama yang ingin diselesaikan dalam proyek ini adalah bagaimana mengidentifikasi dan mengklasifikasikan pelanggan berdasarkan kemungkinan mereka untuk mengklik iklan tertentu.

Exploratory Data Analysis

Univariate Analysis :

1. Age

Dari gambar diketahui bahwa untuk user dengan rentang umur kurang dari 35 tahun memiliki kecenderungan untuk tidak klik iklan, sedangkan untuk user di atas 35 tahun memiliki kecenderungan untuk klik iklan tersebut. Dimana diketahui juga jumlah user terbanyak dimiliki kluster kelompok umur 32 tahun.

2. Daily Time Spent on Site

Dari gambar diketahui bahwa:

- User yang klik iklan jumlah nya lebih sedikit jika dibandingkan dengan user yang tidak klik iklan

- User yang tidak klik iklan, lebih lama menghabiskan waktu nya untuk mengunjungi situs web dibandingkan dengan user yang klik iklan dimana rentang dari user yang tidak klik iklan ada pada rata2 60 – 90 (s), sedangkan untuk user yang klik iklan ada pada rentang rata2 30 – 80 (s)

3. Daily Internet Usage

Dari gambar diketahui bahwa user yang tidak klik iklan, lebih banyak menggunakan internet nya jika dibandingkan dengan user yang klik iklan dimana diketahui bahwa untuk user yang tidak klik iklan menghabiskan kuota di range 200 – 250 MB (berdasarkan jumlah user terbanyak), sedangkan untuk user yang klik iklan menghabiskan kuota di range 100 – 150 MB (berdasarkan jumlah user terbanyak).

Bivariate Analysis :

Gambar ini merepresentasikan korelasi antara kolom umur dengan penggunaan internet dan waktu yang dihabiskan, dimana dari kedua plot tersebut dapat diambil kesimpulan bahwa relasi antara umur dengan penggunaan internet dan relasi antara umur dengan waktu yang dihabiskan kedua nya cenderung memiliki korelasi negative (berbanding terbalik)

Multivariate Analysis :

Berdasarkan gambar, masing2 feature memiliki korelasi yang cukup kuat terhadap feature target. Sementara itu ada beberapa feature yang memiliki korelasi yang cukup kuat terhadap feature lainnya seperti contoh, korelasi antara feature internet usage dengan time spent, internet usage dengan age dan age dengan time spent.

Korelasi antara feature independent dengan feature dependent (feature target = 'Clicked on Ad'), dapat dilihat lebih jelas nya pada gambar disamping. Dimana dari ke 4 feature, hanya feature age saja yang memiliki korelasi positive dimana semakin tua umur user, kemungkinan klik iklan nya lebih besar. Untuk feature yang lain sebaliknya, semakin besar value nya maka kemungkinan untuk klik iklan juga semakin kecil

Data Cleansing and Preprocessing

1. Handle missing value

Dari dataset, diketahui terdapat 4 feature yang memiliki missing value yaitu :

- Daily time spent on site

Handle dengan menggunakan rata2 waktu yang dihabiskan oleh user saat mengunjungi website setiap hari

- Area income

Handle dengan menggunakan rata2 income user

- Daily internet usage

Handle dengan menggunakan rata2 penggunaan internet harian user

- Gender

Handle menggunakan modus dari feature gender

2. Handle data duplikat

Tidak ada data duplikat

3. Feature Engineering

Membuat feature baru : extract feature timestamp menjadi tahun, bulan, pekan, tanggal

4. Feature encoding

Melakukan encoding terhadap feature kategorikal, diantaranya : gender, clicked on ad, city, province, category

5. Split feature and target

Membagi dataset menjadi :

- X (feature) : drop kolom timestamp (sudah di extract), tahun (karena hanya ada tahun 2016), clicked on ad (menjadi kolom target)
- Y (target) : kolom clicked on ad menjadi kolom target untuk proses learning

Data Modeling

Dalam tahap ini saya melakukan 2 experiment model machine learning, dimana experiment pertama tanpa menggunakan normalisasi data dan yang kedua menggunakan normalisasi data. Decision Tree dan Random Forest.

- Experiment 1

Berikut adalah tahapan pemodelan machine learning pada experiment 1 :

- Split Data Train dan Test

Data dibagi menjadi 2 (data train dan test) dengan perbandingan 80:20

- Learning Process

Membuat 3 model algorithma machine learning, yaitu logistic regression, decision tree dan random forest. Hasil dari proses learning sebagai berikut :

Algorithm ML	Accuracy	Precision	Recall	F1 Score	ROC AUC
Logistic Regression	47.5%	0	0	0	75%
Decision Tree	92.5%	93.3%	92.4%	92.8%	92.5%
Random Forest	95.5%	95.3%	96.2%	95.7%	98.6%

- Experiment 2

- Feature Transform

Melakukan Standard Scaler pada feature numerical sebelum dataset di split

- Split Data Train dan Test

Data dibagi menjadi 2 (data train dan test) dengan perbandingan 80:20

- Learning Process

Membuat 3 model algorithma machine learning, yaitu logistic regression, decision tree dan random forest. Hasil dari proses learning sebagai berikut :

Algorithm ML	Accuracy	Precision	Recall	F1 Score	ROC AUC
Logistic Regression	97.5%	98.1%	97.1%	97.6%	99.1%
Decision Tree	92.5%	93.3%	92.4%	92.8%	92.5%
Random Forest	95.5%	95.3%	96.2%	95.7%	98.6%

Evaluation metrics score berfokus kepada nilai precision (meminimalisir angka FP), dan dari kedua experiment tersebut, precision score tertinggi ada pada experiment kedua dengan model algothma Logistic Regression

Hyperparameter Tuning

Selanjutnya untuk meningkatkan performa jauh lebih baik lagi, dilakukan hyperparameter tuning dengan setting parameter :

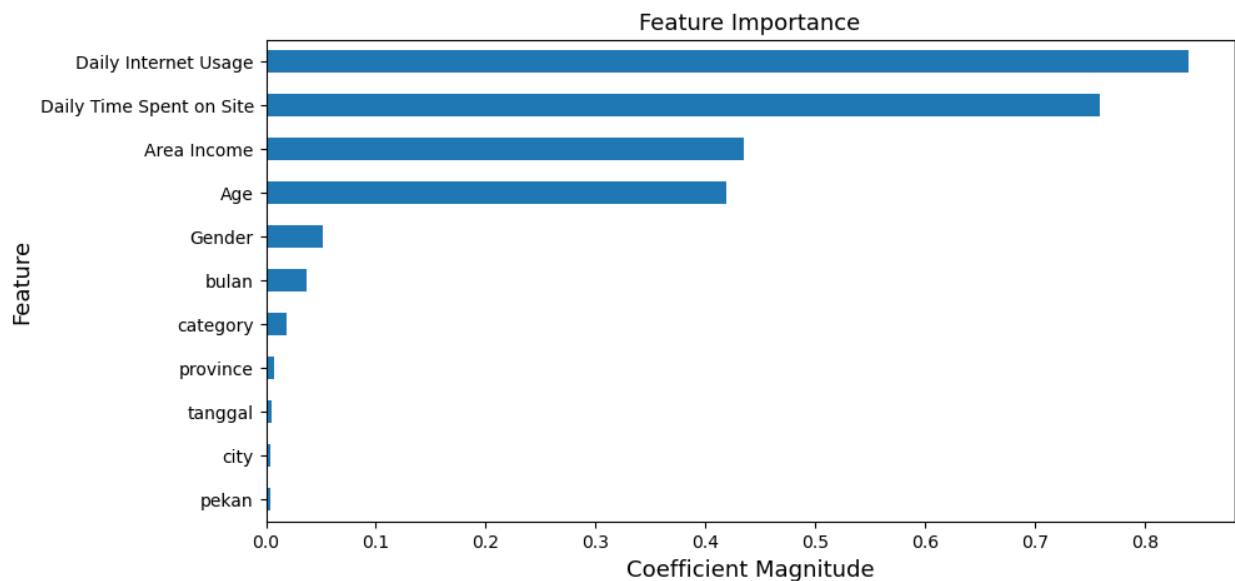
- Nilai regularization (C) : 0.01, 0.1, 1, 1.0
- Algorithm Solver : lbfgs, liblinear, sag, saga

Dari hasil proses learning didapat untuk best parameter nilai C adalah 0.01 dan algorithm solver adalah lbfgs. Berikut untuk hasil evaluasi metrics score :

Algorithm ML	Accuracy	Precision	Recall	F1 Score	ROC AUC
Logistic Regression	97%	100%	94.3%	97.1%	99.2%

Feature Importance

Dari hasil yang diperoleh dari proses hyperparameter tuning sebelumnya, diketahui terdapat beberapa feature yang sangat mempengaruhi proses learning. Feature ini bisa dijadikan bahan pertimbangan oleh tim marketing dalam menyusun marketing plan kedepannya, feature tersebut ditampilkan pada gambar berikut :



Semakin besar nilai coef nya, maka semakin besar pula pengaruh feature tersebut terhadap model saat melakukan proses learning.

Bisnis Rekomendasi

Berdasarkan hasil EDA dan feature importance, kita dapat memberikan beberapa rekomendasi bisnis yang relevan. Dengan menghubungkan hasil dari EDA dengan feature important untuk menyusun rekomendasi bisnis:

1. Target Pasar Berdasarkan Usia:

- Mengingat kecenderungan user di atas 35 tahun lebih cenderung untuk mengklik iklan, maka kita dapat memfokuskan campaign iklan pada kelompok usia ini.
- Mengingat kluster kelompok umur terbanyak adalah 30 – 40 tahun, fokuskan iklan dan produk yang sesuai dengan preferensi dan kebutuhan kelompok usia ini.

2. Optimalkan Iklan Berdasarkan Waktu:

- Mengetahui bahwa user yang tidak mengklik iklan menghabiskan waktu lebih lama di situs web, pertimbangkan untuk mengoptimalkan tampilan dan penempatan iklan agar lebih menarik perhatian bagi user tersebut.
- Selain itu, karena user yang mengklik iklan cenderung menghabiskan waktu lebih sedikit di situs web, pertimbangkan untuk menghadirkan tawaran yang menarik dan informatif secara singkat di iklan.

3. Segmentasi User Berdasarkan Penggunaan Internet:

- Dengan mengetahui bahwa user yang tidak mengklik iklan cenderung menghabiskan lebih banyak kuota internet, kita dapat mempertimbangkan untuk menyajikan konten yang lebih menarik dan relevan bagi user dengan pola penggunaan internet yang lebih tinggi.
- Untuk user yang cenderung mengklik iklan dan menggunakan kuota internet lebih sedikit, fokuskan iklan pada produk atau layanan yang sesuai dengan kebutuhan mereka dan menarik perhatian dalam waktu singkat.

4. Focus on Marketing Cost:

- Diketahui bahwa daily internet usage dan daily time spent on site adalah dua fitur yang sangat mempengaruhi proses 'learning' (feature importance tinggi). Maka, kita dapat memprioritaskan alokasi dana pemasaran pada pengembangan campaign yang efektif berdasarkan fitur-fitur ini.
- Analisis lebih lanjut juga dapat dilakukan untuk mengoptimalkan marketing cost dengan fokus pada segmen user yang paling berpotensi untuk mengklik iklan dan berinteraksi dengan web site.

5. A/B Testing:

- Selalu lakukan uji coba A/B test untuk menguji berbagai strategi pemasaran, konten iklan, dan penempatan iklan. Dengan demikian, kita dapat mengidentifikasi apa yang paling efektif dalam menarik minat dan mengklik iklan user.

6. Personalisasi Konten:

- Gunakan data yang dimiliki untuk menyajikan konten yang lebih personal dan relevan untuk user. Dengan memahami preferensi dan kebiasaan user, kita dapat

menyampaikan pesan yang lebih persuasif dan meningkatkan peluang untuk mengklik iklan.

Simulasi Bisnis

1. Tanpa Penggunaan Model Machine Learning:

- Tentukan jumlah total user yang akan menjadi target pemasaran dalam simulasi ini.
- Asumsikan biaya pemasaran yang akan dikeluarkan untuk mencapai target audience ini.

➤ Asumsi 1000 User

➤ Asumsi biaya Rp 500/reach

Total biaya marketing : $1000 * 500 = \text{Rp } 500.000$

- Berdasarkan data historis atau asumsi konversi dari kampanye sebelumnya, tentukan tingkat konversi (conversion rate) dari user yang mengklik iklan menjadi pelanggan atau melakukan aksi tertentu yang menghasilkan pendapatan (misalnya membeli produk atau menggunakan layanan).

CTR = 50% (data dari dataset)

Conversion Rate = 1.5% (asumsi)

- Hitung jumlah user yang akan mengklik iklan dan melakukan transaksi berdasarkan tingkat konversi.

➤ $1.000 * 50\% * 1.5\% = 7.5 \text{ user}$

- Hitung pendapatan yang diharapkan berdasarkan jumlah user yang mengklik iklan dan nilai rata-rata dari aksi yang diambil (misalnya rata-rata nilai pembelian).

➤ Asumsi per transaksi 100.000. Total pendapatan yang diharapkan : $100.000 * 7.5 = 750.000$

- Hitung total biaya pemasaran dan total pendapatan untuk mendapatkan profit.

➤ Profit = $\text{Rp } 750.000 - \text{Rp } 500.000 = \text{Rp } 250.000$

2. Dengan Penggunaan Model Machine Learning:

- Lakukan proses prediksi menggunakan model machine learning yang telah Anda latih sebelumnya pada dataset yang telah Anda persiapkan untuk simulasi ini.

➤ Accuracy = 97%

➤ Recall = 94.3%

➤ F1 Score = 97.1%

➤ ROC AUC = 99.2%

➤ Precision = 100%

- Berdasarkan hasil prediksi, tentukan jumlah user yang diprediksi akan mengklik iklan.

➤ $\text{Prediksi Jumlah user klik iklan} : 1000 \text{ user} * 97\% = 970 \text{ user}$

- Hitung pendapatan yang diharapkan berdasarkan prediksi jumlah user yang mengklik iklan dan nilai rata-rata dari aksi yang diambil (misalnya rata-rata nilai pembelian).

➤ **Revenue :**

$970 \text{ user} * 1.5\% (\text{conversion rate}) * \text{Rp } 100.000 = \text{Rp } 1.455.000$

- Hitung total biaya pemasaran dan total pendapatan untuk mendapatkan profit.

➤ $\text{Profit} = \text{Rp } 1.455.000 - \text{Rp } 500.000 = \text{Rp } 955.000$