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 dibandingkna 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 memiki 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 leaerning 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 leaerning 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 algrothma Logistic Regression

Hyperparameter Tuning

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

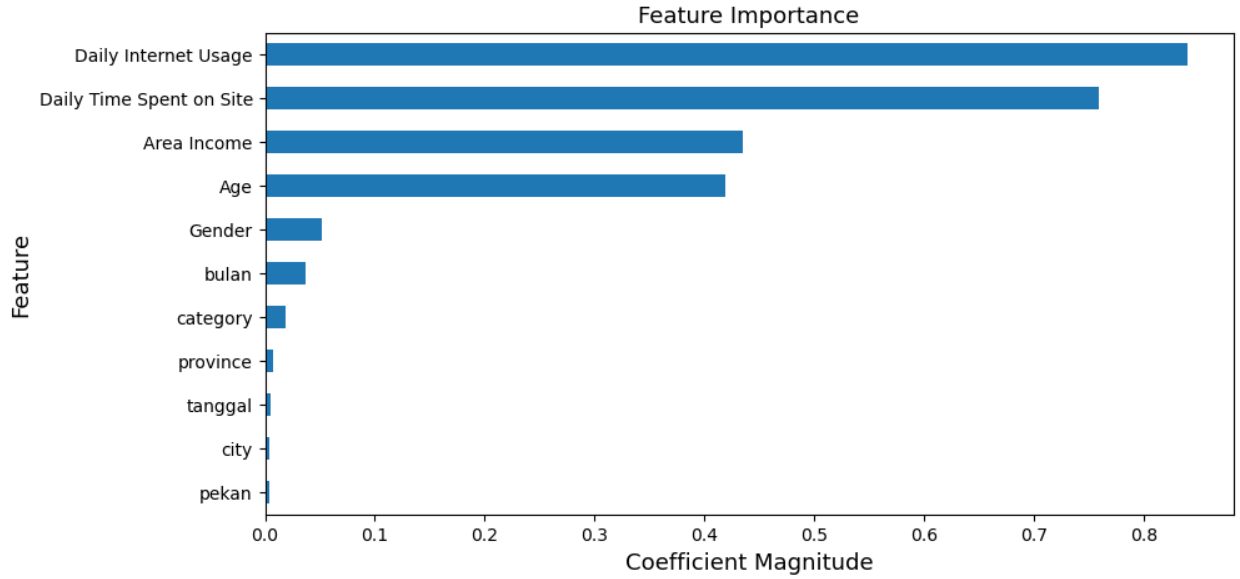
- Nilai regularization (C) : 0.01, 0.1, 1, 1.0
- Algorithma Solver : lbfgs, liblinear, sag, saga

Dari hasil proses learning didapat untuk best parameter nilai C adalah 0.01 dan algorithma 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.