Lhutfia Ichsan's

Portfolio *Presentation*

>>

DATA SCIENCE BATCH 13

Content

Hello, I'm Lhutfia

I am a recent Information Systems Student who has interested in Data Processing. I enjoy taking on responsibilities, contributing to a team, and finishing tasks on my own.

Honest and on time is my work ethic. And I'am exited for my upcoming adventure!

BACK TO CONTENT PAGE

Mini Project 04

Predict Customer Clicked Ads by Using Machine Learning

BACK TO CONTENT PAGE

Workflow ››

EDA

Melakukan eksplorasi data saat ini

Data Preprocessing

Mempersiapkan data sebelum masuk kedalam machine learning

Modelling

- Implementasi terhadap data dengan menggunakan dua schema pemodellan
- Menggunakan business metric untuk menentukan model

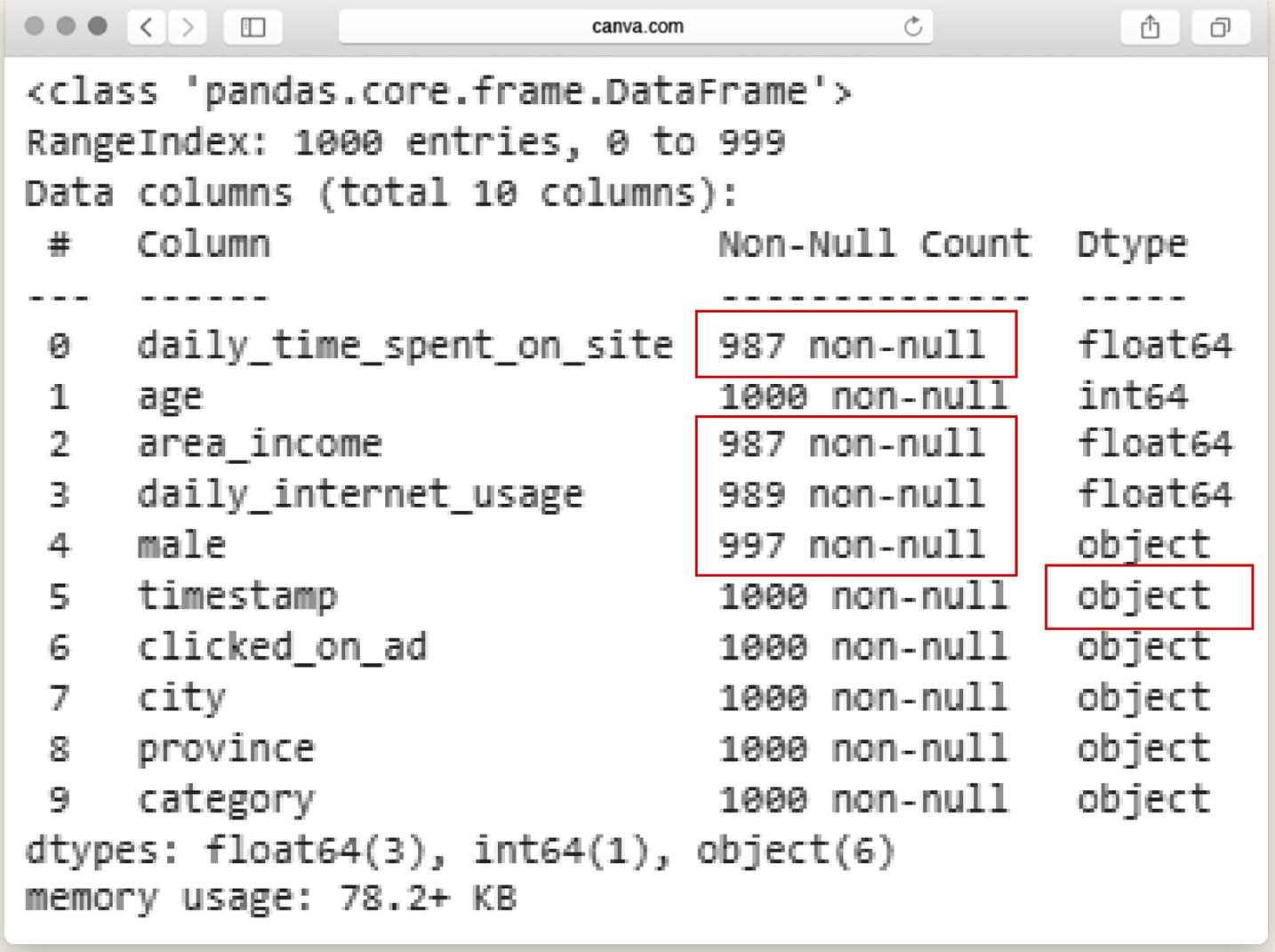
Evaluation

Simulation

Melakukan simulasi perbedaan sebelum dan sesudah penggunaan machine learning

Recommendation

Memberikan rekomendasi bisnis yang relevant untuk

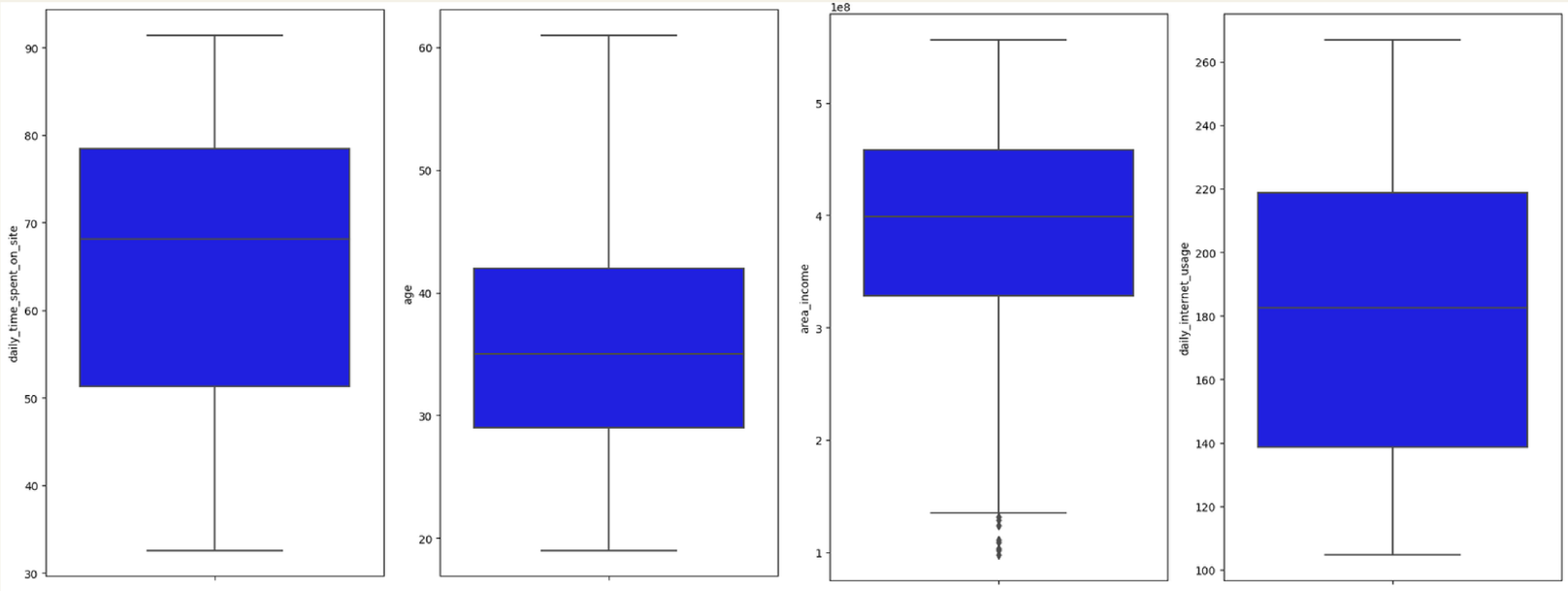


EDA

How was the Dataset?

- 100 rows, 10 columns
- Terdapat missing values
- Terdapat unnormal data type
- Tidak terdapat data yang aneh untuk masing-masing columns

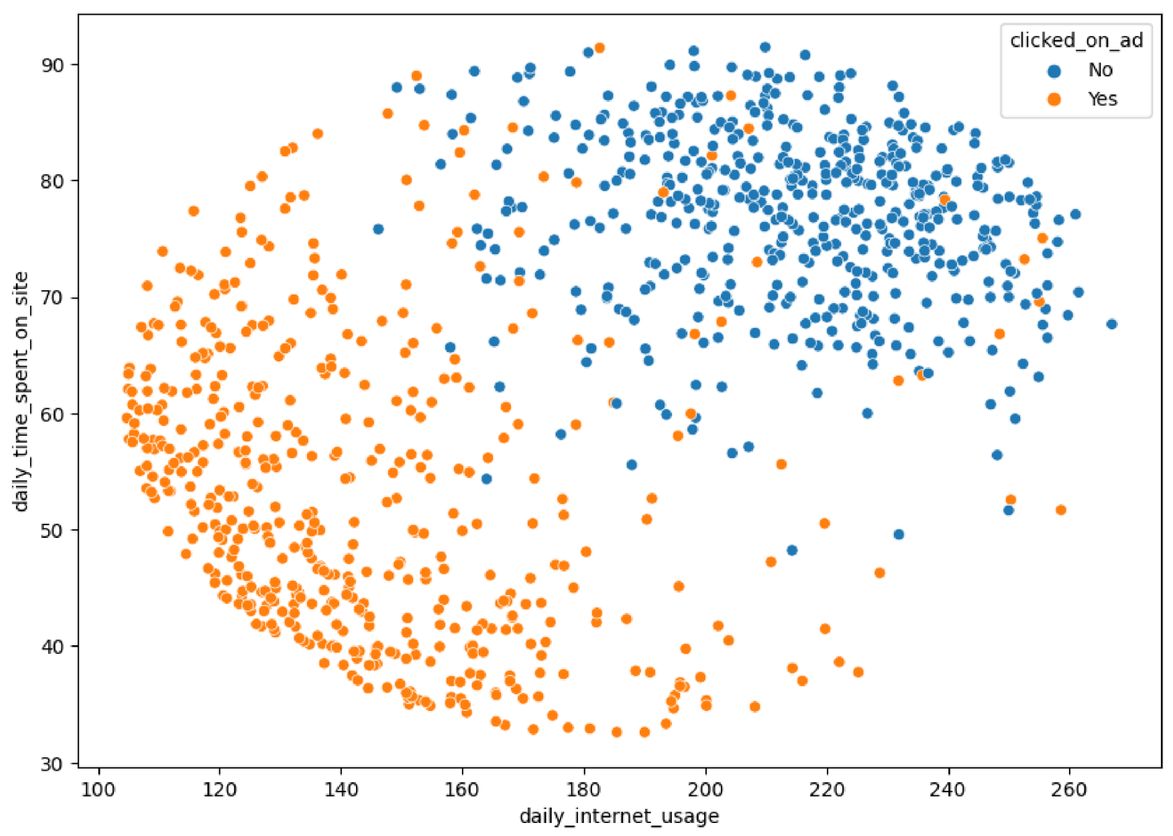
Univariate Analysis - EDA



EDA

See Skewed Data Using Boxplot

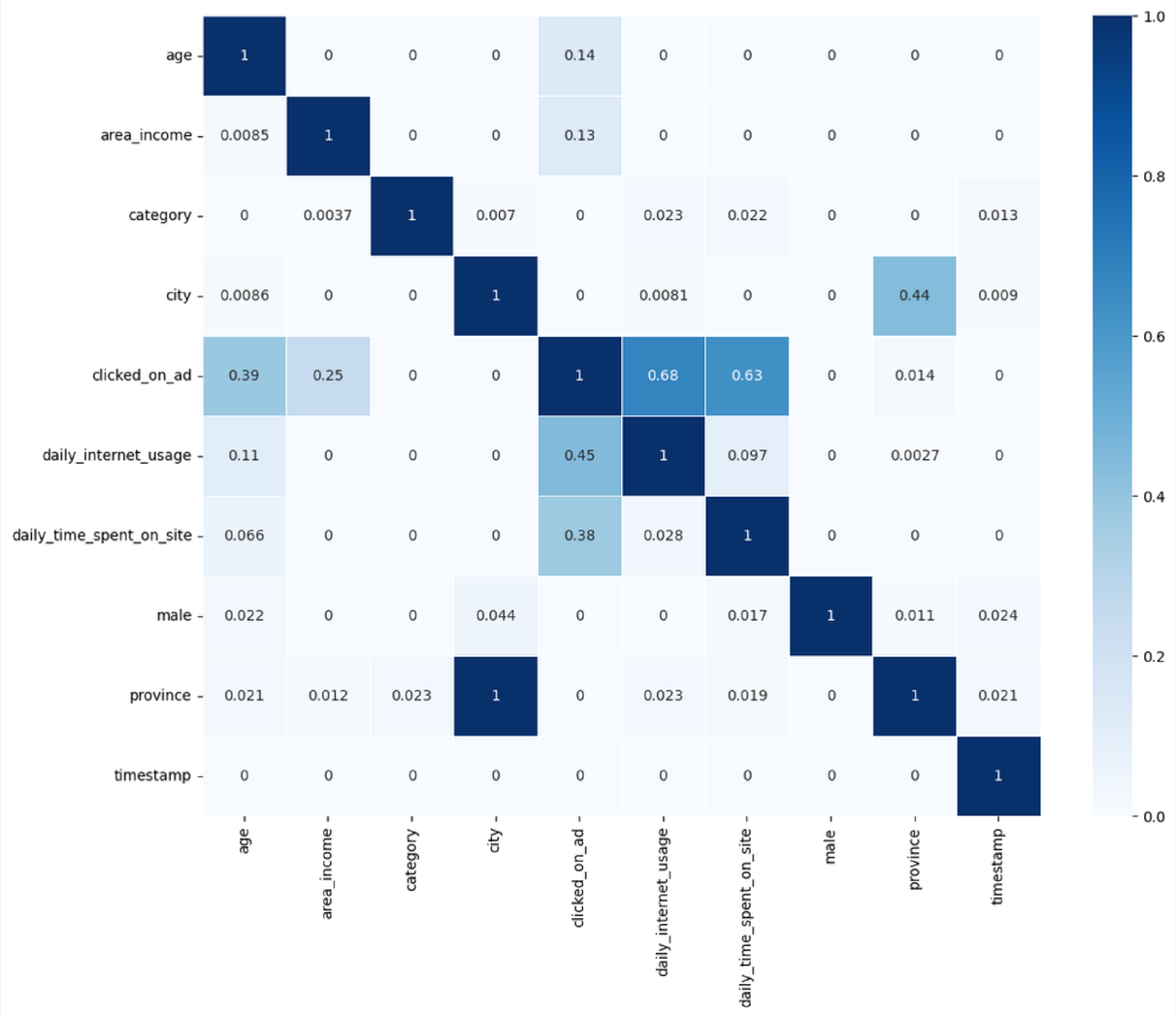
Bivariate Analysis - EDA



EDA

Using Scatter plot for *Crowd Zero*

Multivariate Analysis - EDA



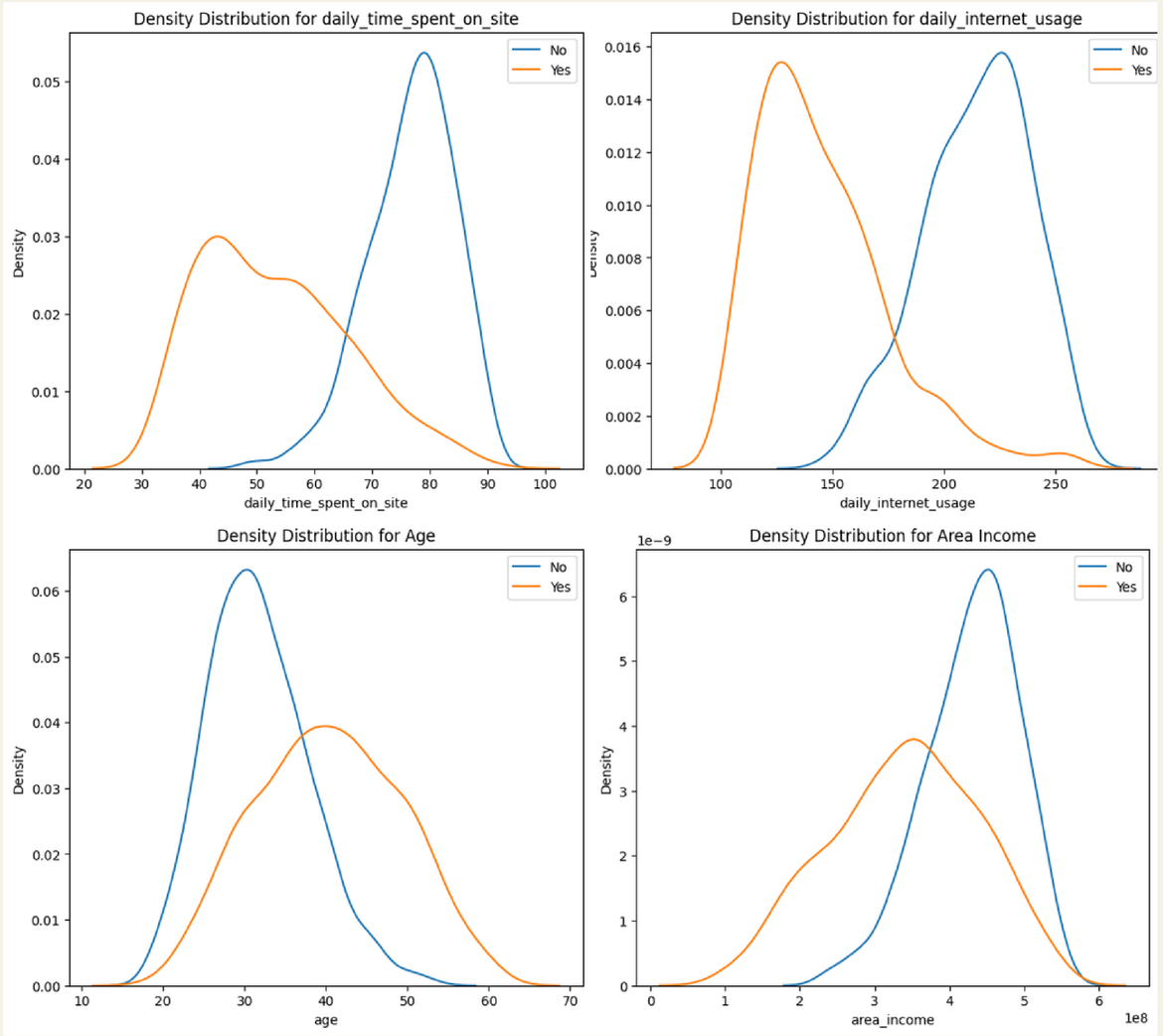
EDA

Using PPScore for correlation

Diketahui bahwa feature yang berhubungan dengan target (clicked_ad_on) adalah : age, area_income, daily_usage_internet, daily_time_spent_on_site dan province.

Insight ››

Pelanggan yang cenderung meresponse adalah pelanggan dengan kriteria daily internet usage dan daily time spent pada website kecil, dengan usia yang relative lebih tua dan pendapatan yang relative lebih rendah. Sedangkan untuk pelanggan yang cenderung tidak mereponse adalah yang sebaliknya.



Data Preprocessing

Missing values, Data type, and Duplicated data

- Delete missing Values 1.3 %
- Change data type columns = 'timestamp' >> datetime
- Duplicated Data 0.0%

Feature Encoding

- Feature encoding untuk columns yang bersifat binary Yes/No.
- Feature one hot encoding untuk columns yang non-binary

Feature One Hot Encoding

male	timestamp	clicked_on_ad	city	province	category	is_Bali	is_Banten	is_Daerah Khusus Ibukota Jakarta	is_Jawa Barat	is_Jawa Tengah
0	2016-05-30 02:34:00	0	Semarang	Jawa Tengah	Furniture	0	0	0	0	1
1	2016-03-19 08:00:00	0	Medan	Sumatra Utara	Electronic	0	0	0	0	0
0	2016-04-29 07:49:00	0	Semarang	Jawa Tengah	Travel	0	0	0	0	1

Feature Encoding

BACK TO CONTENT PAGE

Business Metrics for Model Evaluation

Recall, dipertimbangkan untuk mengurangi False Negative yaitu keadaan dimana Machine Learning memprediksi pelanggan tidak akan meresponse namun ternyata meresponse >> Lost Cost

Precision, dengan tujuan untuk mengurangi False Positive yaitu keadaan dimana Machine Learning memprediksi pelanggan tidak meresponse namun ternyata meresponse >> Lost Revenue

Revenue vs. Cost

Modelling

WITHOUT FEATURE SCALLING

	model_name	model	accuracy	recall	precision	duration
0	K-Nearest Neighbor	KNeighborsClassifier()	0.671280	0.630137	0.691729	0.009359
1	Logistic Regression	LogisticRegression()	0.494810	0.000000	0.000000	0.024757
2	Decision Tree	DecisionTreeClassifier()	0.930796	0.952055	0.914474	0.020319
3	Random Forest	(DecisionTreeClassifier(max_features='sqrt', r...	0.948097	0.945205	0.951724	0.549195
4	Gradient Boosting	([DecisionTreeRegressor(criterion='friedman_ms...	0.944637	0.958904	0.933333	0.793807

WITH FEATURE SCALLING

	model_name	model	accuracy	recall	precision	duration
0	K-Nearest Neighbor	KNeighborsClassifier()	0.750865	0.732877	0.764286	0.008675
1	Logistic Regression	LogisticRegression()	0.948097	0.924658	0.971223	0.023625
2	Decision Tree	DecisionTreeClassifier()	0.910035	0.924658	0.900000	0.007551
3	Random Forest	(DecisionTreeClassifier(max_features='sqrt', r...	0.951557	0.952055	0.952055	0.582976
4	Gradient Boosting	([DecisionTreeRegressor(criterion='friedman_ms...	0.937716	0.952055	0.926667	0.942897

Modelling With and Without Feature Scalling

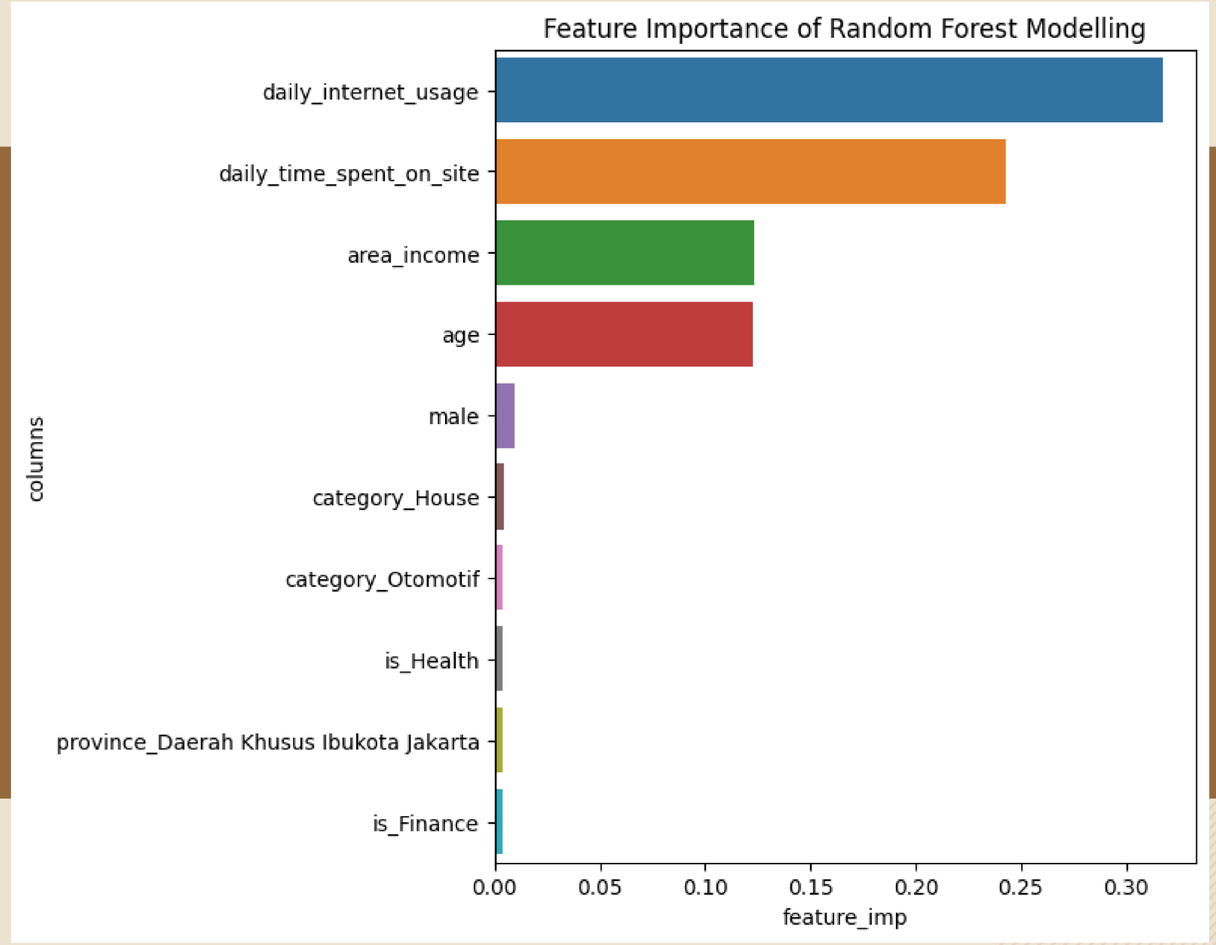
- Feature scalling mampu meningkatkan performa khususnya pada model Logistic Regression
- Model dengan performa terbaik adalah Random Forest dengan Recall 0.95 dan Precision 0.95

Feature Importance

Berikut ini merupakan feature Importance dari penggunaan model Random Forest yaitu :

1. Daily internet usage
2. Daily time spent on site
3. Income
4. Age

Feature Importance sejalan dengan hubungan correlation yang telah disebutkan sebelumnya.



Business Simulation

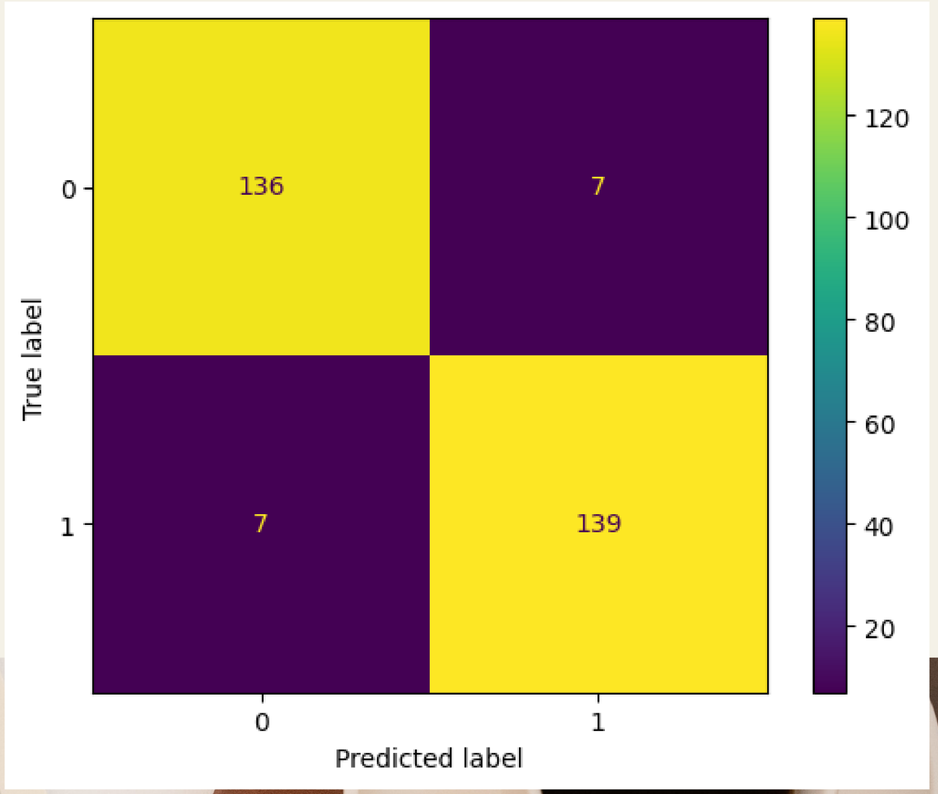
Untuk melakukan simulasi, diasumsikan bahwa biaya yang perusahaan keluarkan adalah $3/iklan. Kemudian, revenue yang didapatkan perusahaan adalah $11/click. Maka :

cost = $3 × (289 iklan) = $ 867
revenue = $11 × (146 iklan) = $ 1.606
revenue - cost = 1.606 - 867 = $ 739

cost = $3 × (145 iklan) = $435
revenue = $11 × (139 iklan) = $1.529
revenue - cost = $1.094

Sebelum (+ 185%)

Sesudah (+ 351%)



Dengan adanya model, perusahaan bisa lebih memfokuskan terhadap pelanggan yang diprediksi akan meresponse meskipun tidak mengurangi skala cost tsb.

Customer Segmentation Result

Potential Customer	Target Customer
Kecenderungan tidak Meresponse	Kecenderungan Meresponse
Rentang usia relative lebih muda dan income yang tinggi	Rentang usia lebih tua dan income rendah
Daily internet usage dan time spent pada website yang cenderung tinggi	Daily internet usage dan time spent pada website yang cenderung rendah

Soft Selling Strategy

Karena diasumsikan segment potential customer lebih melek digital sehingga tidak mudah tergiur iklan

Topic and Content Strategy

Kemudahan untuk dipahami dan topik yang mampu menarik minat segment kelompok pelanggan

BACK TO CONTENT PAGE